Briefings in Bioinformatics, 2026, 27, bbag177
https://doi.org/10.1093/bib/bbag177
Published: 30 April 2026
Problem Solving Protocol

OXFORD

Reconstructing 3D transcriptional organization from spatial transcriptomics reveals consistent oncogenic translocations and developmental dynamics

Yifei Sheng[1,2,3,4,‡], Shiying Li[1,2,‡], Zhengyang Xue[1,2], Miaozhe Huo[1,2], Yiqi Jiang[1,2], Shuai Cheng Li[1,2,*]

[1]Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong 999077, China
[2]City University of Hong Kong Shenzhen Research Institute, Nanshan District, Shenzhen, 518057, China
[3]Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Shenzhen University Medical School, 1066 Xueyuan Avenue, Nanshan District, Shenzhen 518060, China
[4]Shenzhen Nanshan People's Hospital, Affiliated Nanshan Hospital of Shenzhen University, 89 Taoyuan Road, Nanshan District, Shenzhen 518052, China

*Corresponding author. E-mail: shuaicli@cityu.edu.hk
‡Yifei Sheng and Shiying Li contributed equally to this work

Abstract

Reconstructing the 3D spatial organization of transcription from 2D spatial data is a fundamental challenge in genomics. Here, we introduce Cytocraft, a computational framework that addresses this problem by inferring a shared, cell-type-specific 3D configuration of transcription centers from subcellular spatial transcriptomics profiles. After validating Cytocraft's robust accuracy in simulations (median relative error: 0.0346), we applied it to diverse datasets to explore spatial patterns that may reflect underlying transcriptional organization. In human nonsmall cell lung cancer, we observed consistent spatial repositioning patterns, suggesting that the marker *MALAT1* may undergo directional translocation during malignant transformation across eight patient samples. In the developing axolotl brain, we found that transcription center reorganization exhibits conserved developmental dynamics across distinct cell types, suggesting coordinated developmental dynamics. By enabling 3D reconstruction of transcription center configurations from 2D data, Cytocraft provides a powerful tool to explore the spatial organization of transcription in development and disease.

Keywords **subcellular spatial transcriptomics, transcription center, spatial configuration, 3D transcriptional organization**

Introduction

Genes do not transcribe at random positions; instead, they occupy preferred nuclear neighborhoods that foster coordinated regulation [1, 2]. Hi-C and related methods have characterized the structural landscape of chromatin folding based on DNA–DNA contacts [3], while RNA–FISH reveals where transcription actively occurs [4, 5]. These two views are related but distinct: genes may co-localize at shared transcriptional hubs without direct chromatin contact, and vice versa [6]. Understanding transcription thus demands methods that can capture both perspectives.

Subcellular spatial transcriptomics (SST) technologies, such as multiplexed error-robust fluorescence *in situ* hybridization (MERFISH) [7], CosMx SMI [8], Xenium [9], and Stereo-seq [10], provide 2D coordinates for thousands of RNA molecules within single cells. However, tissue sectioning and imaging collapses the cell's intricate 3D structure into a flattened 2D projection. This "crushed dimension" problem obscures true z-axis distances between transcripts, limiting our ability to infer spatial relationships among transcription sites.

To address this problem and to specifically capture the transcriptional landscape from RNA data, we operationally define a "transcription center" for each gene as the weighted geometric centroid of its detected RNA transcripts within a cell. Critically, this RNA-based definition captures where transcription occurs and where transcripts localize, rather than where chromatin physically resides—a distinction from Hi-C-derived 3D genome structures. The full set of these centers forms a gene-specific 3D spatial configuration that, if reconstructed, could provide estimates of inter-gene distances and clustering patterns at the transcriptional level. Reconstructing these configurations may offer insight into cell-type diversity and the spatial reorganization of transcription during development or disease.

We introduce Cytocraft, a structure-from-motion framework that reconstructs a shared, cell-type–specific 3D configuration of these

Received: November 18, 2025. **Revised:** February 6, 2026. **Accepted:** March 20, 2026

transcription centers from 2D SST profiles. Cytocraft treats each cell as an orthographic projection of a common underlying configuration, assuming rigid rotations across cells of a type and accurate molecule-to-cell assignment. Through iterative factorization and optimization, the algorithm infers the latent spatial arrangement, up to inherent ambiguities in scale, rotation, and chirality. Extensive simulations demonstrate robustness across varying cell numbers, imaging resolutions, detection efficiencies, dropout rates, and noise levels, while real-world tests on MERFISH [7], CosMx SMI [8], Xenium [9], and Stereo-seq [10] datasets confirm broad applicability (Supplementary Tables 1–3).

By reconstructing 3D transcriptional organization in human non-small cell lung cancer, we observe consistent spatial repositioning patterns, suggesting a conserved, directional movement of the *MALAT1* transcription center during tumorigenesis. Furthermore, by charting 3D configurations across the developing axolotl brain, we find that transcription center reorganization exhibits conserved patterns across distinct cell types, consistent with coordinated developmental dynamics. By restoring the 3D, Cytocraft transforms static 2D snapshots into 3D spatial maps, offering a new perspective on the spatial organization of transcription in development and disease.

Results

Cytocraft: an iterative algorithm for reconstructing spatial transcription center configurations

Cytocraft addresses the fundamental challenge of reconstructing 3D spatial configurations of transcription centers from inherently 2D SST data. The algorithm operates on the principle that cells of the same type share a common 3D spatial arrangement of their transcription centers, with individual cells representing different angular projections of this shared configuration onto the 2D observation plane (Fig. 1a).

The reconstruction process begins with identifying transcription centers for each gene within individual cells. Cytocraft defines the transcription center of any given gene as the weighted geometric centroid of its transcriptional spots (as detailed in Methods), as illustrated in Fig. 1b. This approach transforms the complex spatial distribution of transcripts into discrete, quantifiable coordinates that serve as the foundation for 3D reconstruction. The method assumes that while individual cells may exhibit rotational variations and observational noise, the underlying spatial configuration remains conserved among cells of the same type.

The core of Cytocraft lies in its iterative optimization framework that alternates between two complementary steps until convergence (Fig. 1c). The "F-step" updates the 3D configuration matrix to best fit the observed 2D transcription centers given current estimates of cell orientations. Conversely, the "R-step" refines the rotation matrices representing individual cell orientations to optimally align the inferred configuration with transcription center observations. This iterative factorization and optimization approach [11] ensures that both the shared spatial configuration and cell-specific projection angles are simultaneously optimized to minimize discrepancies between predicted and observed transcription center locations.

Upon convergence, Cytocraft produces two key outputs that enable spatial analysis. The optimal spatial configuration represents an inferred 3D arrangement of transcription centers shared among cells of the same type, providing estimate spatial distances and relationships between genes that are otherwise obscured in 2D projections. Additionally, the algorithm generates rotation matrices for each cell, which may reflect biological features such as tissue organization or developmental constraints. Together, these outputs transform limited planar observations into reconstructions up to gauge ambiguities of gene organization within cells.

The 3D configurations reconstructed by Cytocraft enable sophisticated analyses of gene spatial relationships that are impossible with conventional 2D data.

Robust performance of Cytocraft in reconstructing spatial configurations from simulated data

To rigorously evaluate Cytocraft's reconstruction capabilities, we developed a comprehensive simulation framework that generates synthetic SST data with known ground truth configurations (Fig. 2a). The pipeline uses reference 3D configurations predicted from Hi-C data by the FLAMINGO model [12], applies random rotations, and incorporates various sources of noise and data deficiencies typical of real SST experiments. The controlled simulation environment enables direct quantitative assessment of reconstruction accuracy by comparing Cytocraft's output against the known original configuration. The ground truth refers to computationally derived conformations from Hi-C, serving as structured reference points for assessing algorithmic accuracy. This establishes Cytocraft's performance bounds and sensitivity to noise parameters, complementing biological validation in subsequent sections.

The simulation parameters were designed to encompass and exceed the challenging conditions present in real-world SST data. Detection efficiency was varied from 10% to 50%, reflecting the range across different platforms where MERFISH achieves ∼80% efficiency while SMI detects only one or two copies per cell [13]. Gene dropout rates were set between 30% and 60% to simulate incomplete gene expression matrices. To model positional uncertainties, we introduced Gaussian noise with standard deviations σ ranging from 0 to 5 pixels, where each pixel corresponds to one spatial unit in the simulation grid. Here, $\sigma = 0$ represents perfect localization (no noise), $\sigma = 2$ represents low noise conditions typical of well-calibrated imaging systems, and $\sigma = 5$ represents high noise conditions that may arise from suboptimal imaging or challenging sample preparations. Additional parameters included cell numbers (100–1000), resolutions (333–1000 nm), ensuring comprehensive coverage of experimental conditions that could potentially challenge the reconstruction algorithm.

Despite these stringent simulation conditions, Cytocraft demonstrated consistently robust performance across ∼10 000 parameter combinations (Fig. 2b). The algorithm achieved a median relative error of 0.0346 (IQR: 0.0138–0.1577) and median Spearman correlation coefficient of 0.9748 (IQR: 0.9157–0.9895) across all tested scenarios. Visual comparison between original and reconstructed configurations showed excellent structural preservation (Supplementary Fig. 1a), with gene–gene distance matrices demonstrating strong correspondence between simulated and recovered spatial arrangements. This robust performance validates Cytocraft's ability to handle the inherent challenges of SST data.

We compared Cytocraft against two simple yet representative baselines (Fig. 2c). The Average baseline collapses depth by taking the per-gene mean of all observed 2D transcription centers (setting $z = 0$), while the Isomap baseline applies nonlinear manifold embedding (3D, 10 neighbors) to genes treated as samples with concatenated per-cell 2D coordinates. Across ∼30 000 simulations stratified by noise (none,

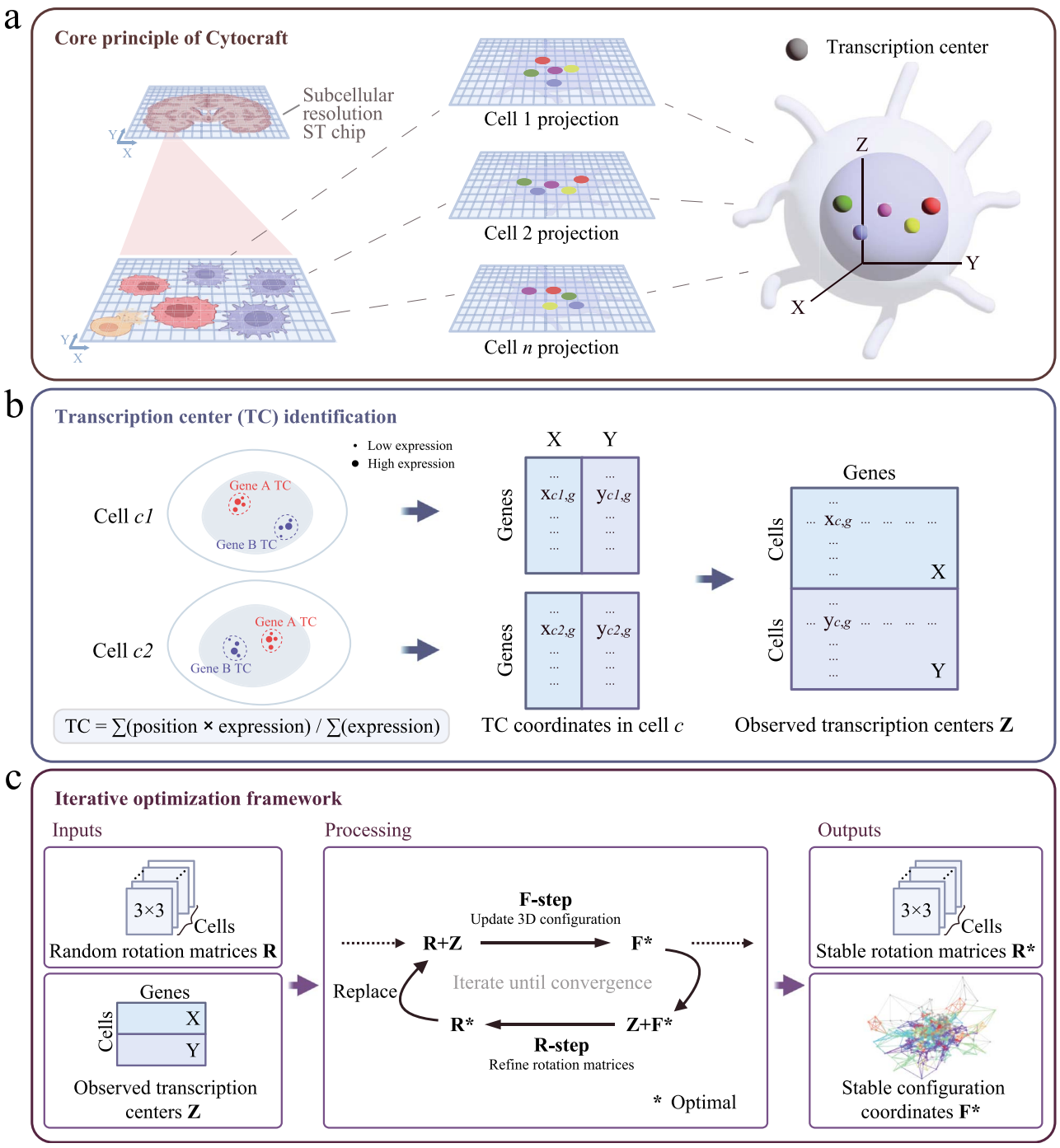

Figure 1 Overview of Cytocraft. (a) The basic assumption of Cytocraft: a single cell in a 2D slice serves as a projection of a shared spatial configuration of transcription centers from a specific angle. The diagram illustrates how the observed 2D transcription centers **Z** result from the projection of the 3D configuration matrix **F** through cell-specific orientation matrices **O**. This projection-based model forms the foundation of our reconstruction approach. (b) Illustration of the transcription center identification process. For each gene in cell, Cytocraft locates its transcription center by calculating the geometric center of transcriptional spots weighted by expression level. (c) The iterative workflow of Cytocraft for inferring the 3D spatial configuration. The algorithm initializes with random rotation matrices **R** and observed transcription centers **Z**, then alternates between the F-step (configuration update) and R-step (rotation update) until convergence. This iterative optimization produces a stable configuration matrix $\mathbf{F}^*$ and stable rotation matrices $\mathbf{R}^*$ that together minimize the difference between projected and observed transcription centers across all cells.

low, and high), Cytocraft consistently outperformed Isomap, reducing median relative error by 1.4–11.3-fold and increasing median Spearman correlation by an absolute 0.07–0.08. Compared with the Average baseline, Cytocraft showed substantial improvements under no-noise (12.7-fold lower error) and low-noise (3.8-fold lower error) conditions. However, under high noise, the Average baseline's simple aggregation proved more robust, resulting in a lower median error than Cytocraft (Fig. 2c, left panel). Under the no-noise condition ((Fig. 2c, right panel

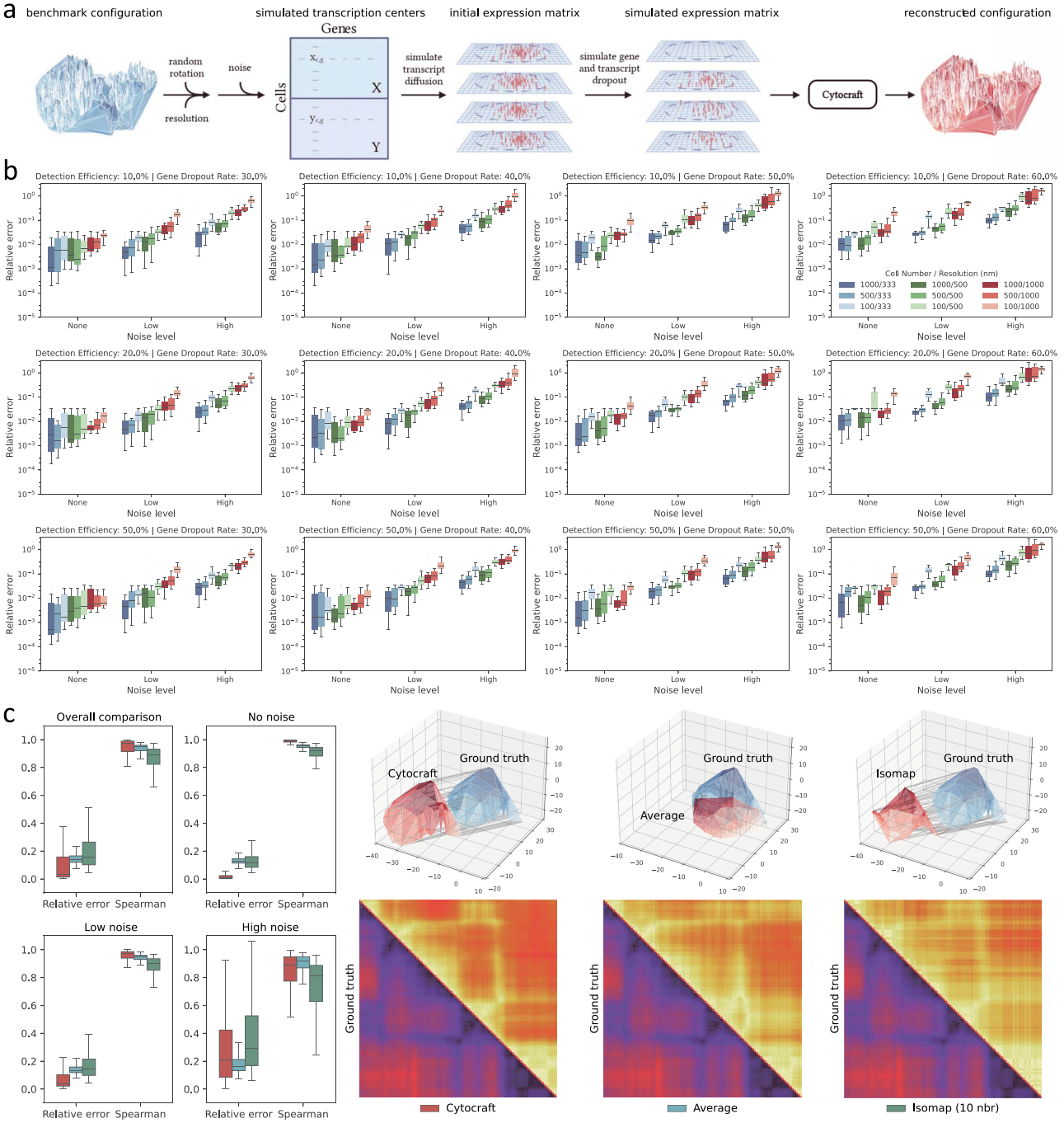

Figure 2 The performance of Cytocraft on simulated SST data. (a) Overview of the simulation pipeline. The diagram shows how Cytocraft generates simulated data by first creating a reference 3D configuration, then applying random rotations to produce 2D projections. These projections are further processed with varying levels of noise, detection efficiency, and gene dropout rates to model realistic SST data. Cytocraft then attempts to recover the original configuration from this simulated data, allowing direct assessment of reconstruction accuracy. (b) Comprehensive performance evaluation across multiple simulation parameters. The panel plots display relative errors from ∼10 000 simulation experiments with varying cell numbers (100–1000), resolutions (333–1000 nm), detection efficiencies (10%–50%), gene dropout rates (30%–60%), and noise levels (Gaussian noise with standard deviations 0–5). The results demonstrate Cytocraft's robust performance even under challenging data conditions, with particularly strong accuracy when noise levels are controlled. (c) Performance comparison of Cytocraft against baseline methods. Left: Bar plots summarize aggregate performance (overall and stratified by noise: none, low, high) comparing Cytocraft with two baseline approaches (Average, Isomap) across ∼30 000 simulation runs using relative error (lower better) and Spearman correlation (higher better). Average baseline: per-gene mean of 2D transcription centers across cells with z set to zero. Isomap baseline: nonlinear manifold embedding (10 neighbors) mapping genes (concatenated per-cell 2D coordinates) into a 3D space. Right: Representative reconstruction example under a no-noise setting showing ground-truth 3D configuration versus reconstructions from each method (top) and the corresponding gene–gene distance matrices (bottom); Cytocraft most faithfully preserves both global geometry and pairwise distance structure.

example), Cytocraft closely reproduces both global geometry and fine-grained pairwise distances, whereas baseline reconstructions exhibit shape distortion (Average) or manifold warping (Isomap). These results indicate that explicit joint modeling of shared 3D configuration and per-cell orientations substantially outperforms generic manifold unfolding and, except in high-noise regimes, naive averaging.

Under optimal conditions with no noise and highest resolution, Cytocraft achieved near-perfect reconstruction accuracy (median relative error=0.0062, median correlation=0.9952), establishing the method's theoretical performance ceiling. Importantly, the algorithm maintained satisfactory accuracy even under high-noise scenarios (median relative error=0.0802, median correlation=0.9520), demonstrating exceptional noise tolerance. When high-noise scenarios were excluded from analysis, overall performance improved substantially (median relative error=0.0228, median correlation=0.9841), confirming that Cytocraft's robust performance extends across the full spectrum of realistic experimental conditions encountered in spatial transcriptomics applications.

Stratified sensitivity analyses confirmed robust performance across noise levels ($\sigma = 0$–5), detection efficiencies (10%–50%), and gene dropout rates (30%–60%), with correlations remaining >0.95 in all conditions (Supplementary Result 2.1.3 and Table 5).

Furthermore, simulations of cell segmentation errors (5%–30% transcript misassignment) demonstrated that Cytocraft maintains robust accuracy (correlation >0.98, error <0.18) even under combined high noise and segmentation leakage conditions (Supplementary Fig. 2 and Table 6).

Cells of the same type share conserved spatial configurations of highly expressed genes

We applied Cytocraft to analyze a mouse embryo dataset at stage E16.5 generated by Stereo-seq [10] (Fig. 3a). The original study annotated 25 distinct cell types, highlighting their spatial heterogeneity. With Cytocraft, we effectively reconstructed the spatial configurations for all 25 cell types (Supplementary Fig. 3), demonstrating that cells of the same type indeed share a consistent spatial layout of genes.

Our analysis reveals that highly expressed genes exhibit the strongest spatial conservation within cell types. We defined highly expressed genes as those with an average count of at least five per cell and assessed their conservation by measuring the deviation between their predicted 3D positions and observed 2D transcription centers. Across all 25 cell types, highly expressed genes consistently exhibited significantly lower deviations compared with permuted controls (Fig. 3c). This finding provides direct evidence that the spatial configurations of highly expressed genes are indeed shared among cells of the same type, with observed deviations likely reflecting natural cell diversity and the dynamic nature of chromosomes.

The conservation principle is particularly evident in cell-type marker genes, which represent both high expression levels and functional specificity. To demonstrate conservation, we compared the spatial positioning of marker genes against random permutations within each cell type. Marker genes such as *Actc1* (Cardiomyocyte) [14], *Bpifa1* (Olfactory epithelial cell) [15], and *Gphn* (Thalamus neuron) [16] showed significantly lower positional deviations in their original transcription centers compared with permuted configurations (Fig. 3b). This conservation of marker gene positions further supports the shared spatial configuration hypothesis, as these functionally critical and highly expressed genes maintain consistent spatial arrangements within cells of the same type.

The observed conservation of spatial configurations among highly expressed genes suggests that transcriptional activity levels and spatial organization are intimately linked within cell types. This relationship likely reflects the functional importance of maintaining consistent spatial arrangements for genes that are critical to cell-type identity and function. The shared spatial configurations reconstructed by Cytocraft thus capture fundamental organizational principles that distinguish different cell types while remaining conserved within each type.

Spatial configuration analysis enables estimation of cell and nucleus dimensions

Cytocraft's reconstructed spatial configurations enable novel estimation of cellular dimensions through analysis of the positional relationship between transcription centers and cell boundaries. We developed a bias quantification method that estimates cell and nucleus sizes by measuring the distribution of distances between the nucleus centroid (centroid of reconstructed transcription centers) and the cell centroid (centroid of observed transcription centers) (Supplementary Method 1.2). This approach provides an alternative to traditional computational segmentation methods that rely on histological staining or immunofluorescence, offering insights into cellular architecture directly from spatial transcriptomics data.

Our dimension estimation method accurately reflects known biological characteristics of different cell types (Fig. 3d). For cell size estimation, we observed that cell types such as limb fibroblast, facial fibroblast, endothelial cell, and keratinocyte exhibited larger estimated cell sizes, which aligns with established knowledge of these cell types' morphological characteristics. Conversely, epithelial cells, forebrain radial glial cells, radial glial cells, and immune cells showed smaller estimated cell sizes, consistent with their known compact cellular architecture and specialized functions requiring efficient molecular transport.

The nucleus size estimations similarly correspond to established biological knowledge, providing validation for our dimensional analysis approach. Erythrocytes demonstrated the smallest estimated nucleus size among all cell types in the mouse embryo, which is consistent with the findings of Fraser *et al.* [17], who reported that between developmental stages E9.5 and E14.5 in mice, the nuclear diameter of peripheral blood cells decreased by nearly 60%, with cross-sectional areas and nuclear sizes reduced by over 90%. In contrast, endothelial cells exhibited the largest estimated nuclei, which correlates with reports that larger nuclei in endothelial cells are more mechanically compliant and can deform at compression sites, suggesting these cells can detect and react to mechanical stimuli within the vasculature [18].

The positional bias analysis reveals systematic patterns in nuclear positioning relative to cellular boundaries across different cell types. The positional biases between nucleus centroid and cell centroid consistently exhibit left-skewed distributions with negative skewness coefficients across all 25 cell types (Fig. 3e), and biases within the same cell type fall within a single characteristic distribution. This distribution pattern indicates that nuclei are typically positioned away from the cell centroid, reminiscent of the peripheral nuclear positioning observed in skeletal muscle cells [19]. The negative skewness coefficients serve as quantitative indicators of the relative distance

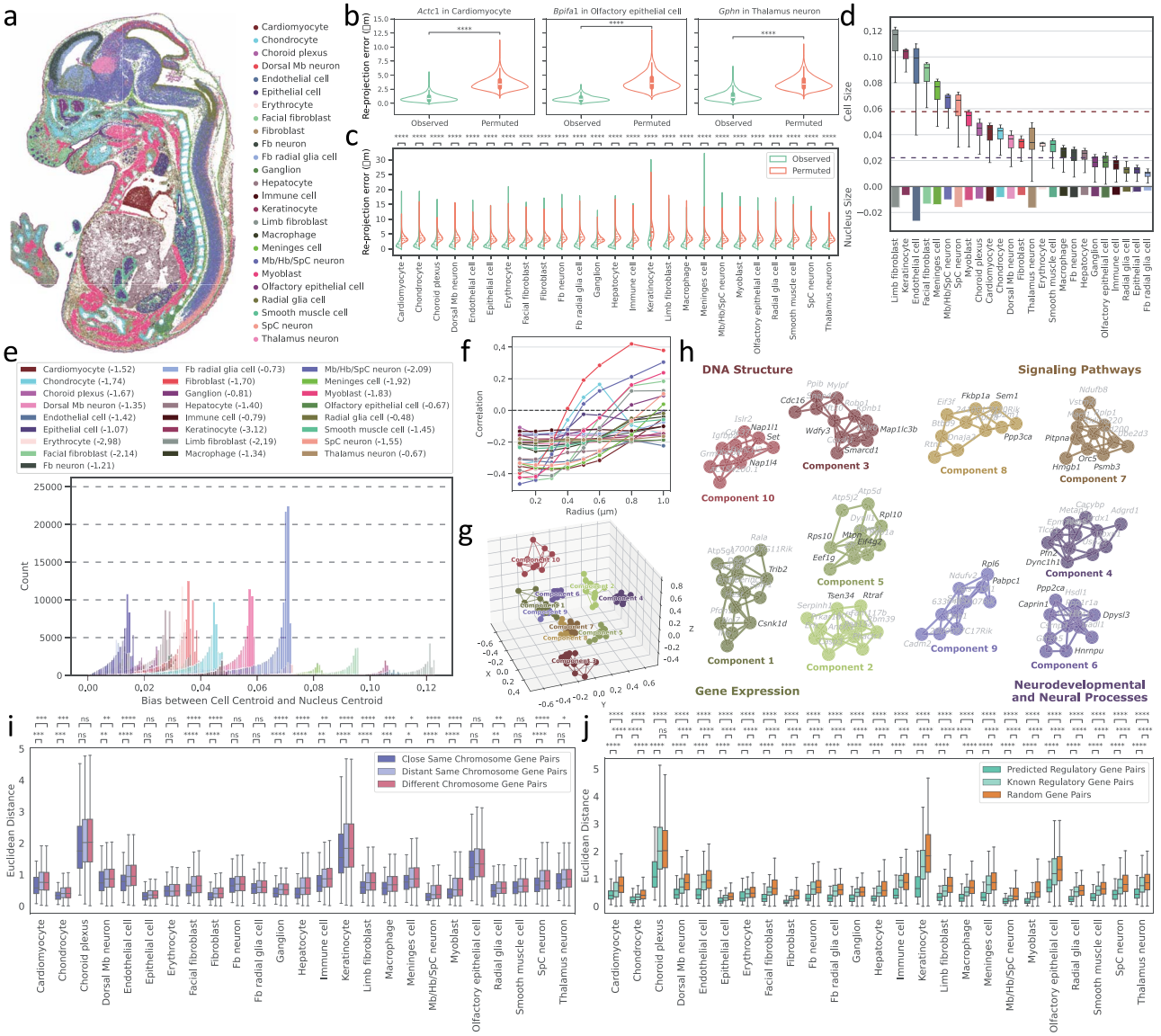

Figure 3 Cell-type spatial configurations reveal cell sizes and gene relationships in a mouse embryo [10]. (a) Spatial distribution of 25 distinct cell types in an E16.5 mice embryo section visualized using t-SNE dimensionality reduction. Different colors represent different cell types, with their anatomical locations labeled (Mb, midbrain; Fb, forebrain; Hb, hindbrain; SpC, spinal cord). (b) Quantitative assessment of marker gene localization conservation. The violin plot compares deviation distributions for three marker genes (*Actc1*, *Bpifa1*, and *Gphn*) between original and permuted transcription centers. Deviations measure the distance between a gene's projected 3D coordinates and its observed 2D locations across cells. (c) Comprehensive analysis of spatial conservation across all highly expressed genes. The violin plot shows deviation distributions for genes with average counts $\geq$5 per cell across all 25 cell types. (d) Cell and nucleus size estimations derived from spatial configurations. The bar plot displays estimated sizes for all 25 cell types, ordered by cell size. Dashed lines indicate upper and lower quartiles of all cell sizes, respectively. (e) Analysis of bias distributions between cell centroids and nucleus centroids. The left-skewed distributions (with negative skewness coefficients indicated in the legend) suggest that nuclei are typically positioned away from cell centroids. (f) Correlation between gene length and spatial neighborhood. The plot shows how the correlation between cube root of gene length and number of neighboring genes varies across different radius thresholds for all cell types. The consistent negative correlation at small radii ($<$0.3 μm) indicates that longer genes have fewer neighbors, likely due to spatial constraints, while this relationship diminishes at larger distances. (g) Visualization of the top ten gene components by size in dorsal midbrain neuron. The 3D spatial arrangement shows how transcription centers cluster into distinct components (labeled 1–10), revealing the spatial organization of gene groups within this specific cell type. (h) Force-directed layout of gene components grouped by Gene Ontology (GO) categories. The network visualization reveals functional relationships within components, with genes associated with specific GO terms highlighted in contrast to the others. (i) Analysis of spatial distances between gene pairs based on their chromosomal locations. The violin plot compares Euclidean distances for gene pairs that are close on the same chromosome ($<$100 000 bp apart), distant on the same chromosome ($>$100 000 bp apart), or on different chromosomes. (j) Relationship between gene regulatory relationships and spatial proximity. The violin plot shows the distribution of Euclidean distances for three categories of gene pairs: those predicted to have regulatory relationships by GENIE3, known regulatory pairs from databases, and random gene pairs.

between cell centroid and nucleus, providing a metric for nuclear eccentricity that varies systematically among cell types.

The robustness of our dimensional estimation approach is demonstrated by consistent distribution patterns observed across multiple SST platforms (Supplementary Figs 4f and 5d). Importantly, the observed left-skewed bias distributions cannot be attributed to RNA diffusion artifacts, as evidenced by the presence of near-zero bias instances and the cell-type-specific nature of these distributions. If RNA diffusion were the primary factor, we would expect uniform distributions across cell types rather than the distinct, cell-type-specific patterns we observe. This validation confirms that our method captures genuine biological differences in cellular and nuclear architecture rather than technical artifacts of the spatial transcriptomics workflow.

Spatial proximity in configurations suggests functional organization of genes

The spatial proximity of genes within reconstructed configurations may reflect their functional organization and regulatory relationships. This organization manifests in multiple ways: through direct spatial constraints imposed by gene characteristics, through functional clustering of related genes, and through regulatory networks that maintain spatial coherence.

Physical characteristics of genes directly influence their spatial organization within configurations. Gene length demonstrates a significant impact on transcription center proximity, with a negative correlation between the number of neighboring genes within a specified radius and the cube root of gene length (Fig. 3f). This relationship suggests that longer genes tend to have fewer neighboring genes due to their spatial occupation, which limits the available space for nearby genes. When the radius exceeds 0.3 μm, the space occupied by individual genes becomes negligible, and this correlation diminishes, indicating that spatial constraints operate primarily at short distances.

Spatially proximal genes cluster into functional components that reflect their biological roles. To identify these functional clusters, we constructed K nearest neighbor (KNN) graphs of genes based on their 3D spatial coordinates and applied the Leiden community detection algorithm, resulting in groups of neighboring genes we term "gene components" (Supplementary Figs 3 and 6). These gene components may indicate functional relationships through spatial proximity [20]. Analysis of dorsal midbrain neuron cells exemplifies this organization, where the 10 largest gene components display distinct functional categories (Fig. 3g; Supplementary Table 7).

GO analysis of these spatially defined components revealed four main functional categories in dorsal midbrain neurons (Fig. 3h): DNA structure-related processes (component 10), gene expression regulation (component 5), intracellular signaling including interleukin [21], and NFAT pathways [22] (components 7–8), and neurodevelopmental processes (component 6). Detailed GO enrichment statistics are provided in Supplementary Table 7.

The functional organization revealed by spatial proximity partially reflects chromosomal relationships, at short genomic scales. Comparing spatial distances among gene pairs revealed that those located closer together on the same chromosome (within 100 kb) exhibit significantly shorter median Euclidean distances in transcription center configurations compared with distant pairs on the same chromosome (>100 kb apart) or pairs on different chromosomes (Fig. 3i). This likely reflects the tendency of nearby genes to share local regulatory environments and be co-transcribed [23, 24], rather than direct chromatin contact—consistent with our observation that Cytocraft distances show near-zero correlation with Hi-C contacts at larger genomic scales (see Discussion). Thus, while short-range genomic neighbors tend to co-localize transcriptionally, the configurations primarily capture functional transcriptional relationships rather than chromatin polymer structure [4].

Spatial proximity in configurations may be associated with regulatory relationships between genes, suggesting that 3D organization could support gene regulation. Using GENIE3 [25] to identify potential regulatory gene pairs, we found that these computationally inferred regulatory pairs exhibit significantly closer spatial proximity compared with both random gene pairs and known regulatory pairs from databases (Fig. 3j). This spatial–regulatory relationship suggests that the reconstructed configurations can assist in identifying cell-type-specific regulatory networks through spatial analysis [20].

Configuration transformation analysis identifies consistent gene translocation patterns during tumorigenesis

To systematically investigate gene translocation patterns during tumorigenesis, we applied Cytocraft to a human nonsmall cell lung cancer (NSCLC) dataset [8] generated using the SMI platform. This dataset contains eight slices from five NSCLC samples (Lung5 Rep1–3, Lung6, Lung9 Rep1–2, Lung12, and Lung13), enabling comprehensive analysis of spatial configuration changes from epithelial to tumor cells across multiple biological contexts.

Configuration similarity analysis revealed that epithelial cells represent the closest cellular origin to tumor cells, providing the foundation for studying translocation patterns (Fig. 4a and b). In Lung5's three biological replicates, we observed strong configuration conservation within the same cell types, with configurations clustering into six distinct groups based on RMSD analysis. Notably, the predominant tumor cell type in Lung5 (tumor 5) exhibited the highest configuration similarity to epithelial cells (mean RMSD: 2.79 $\pm$ 0.32), confirming the established understanding that lung tumor cells originate from epithelial cells [26] and validating our approach for tracking translocation patterns during malignant transformation.

The magnitude of configuration transformation between epithelial and tumor cells correlates with cancer progression stages (Fig. 4c). Stage IIB NSCLC demonstrated significantly larger configuration differences (RMSD: 8.09) compared with stage IIIA samples (mean RMSD: 2.45 across seven samples, P-value$<$.0001, one-sample t-test). This association suggests that the extent of gene translocation patterns may reflect disease progression, though validation with larger sample sizes across different stages is needed.

To quantify gene translocation patterns systematically, we measured displacement distances for genes during epithelial-to-tumor transformation and categorized them based on movement magnitude (Fig. 4d and e). We defined "short-distance translocation genes" as the top 10% with smallest displacements and "long-distance translocation genes" as those with largest displacements. Short-distance translocation genes showed significant overlap across different samples (Jaccard similarity analysis) and were strongly enriched in cytokine signaling pathways within the immune system, indicating that certain immune-related gene positions remain relatively conserved during tumorigenesis.

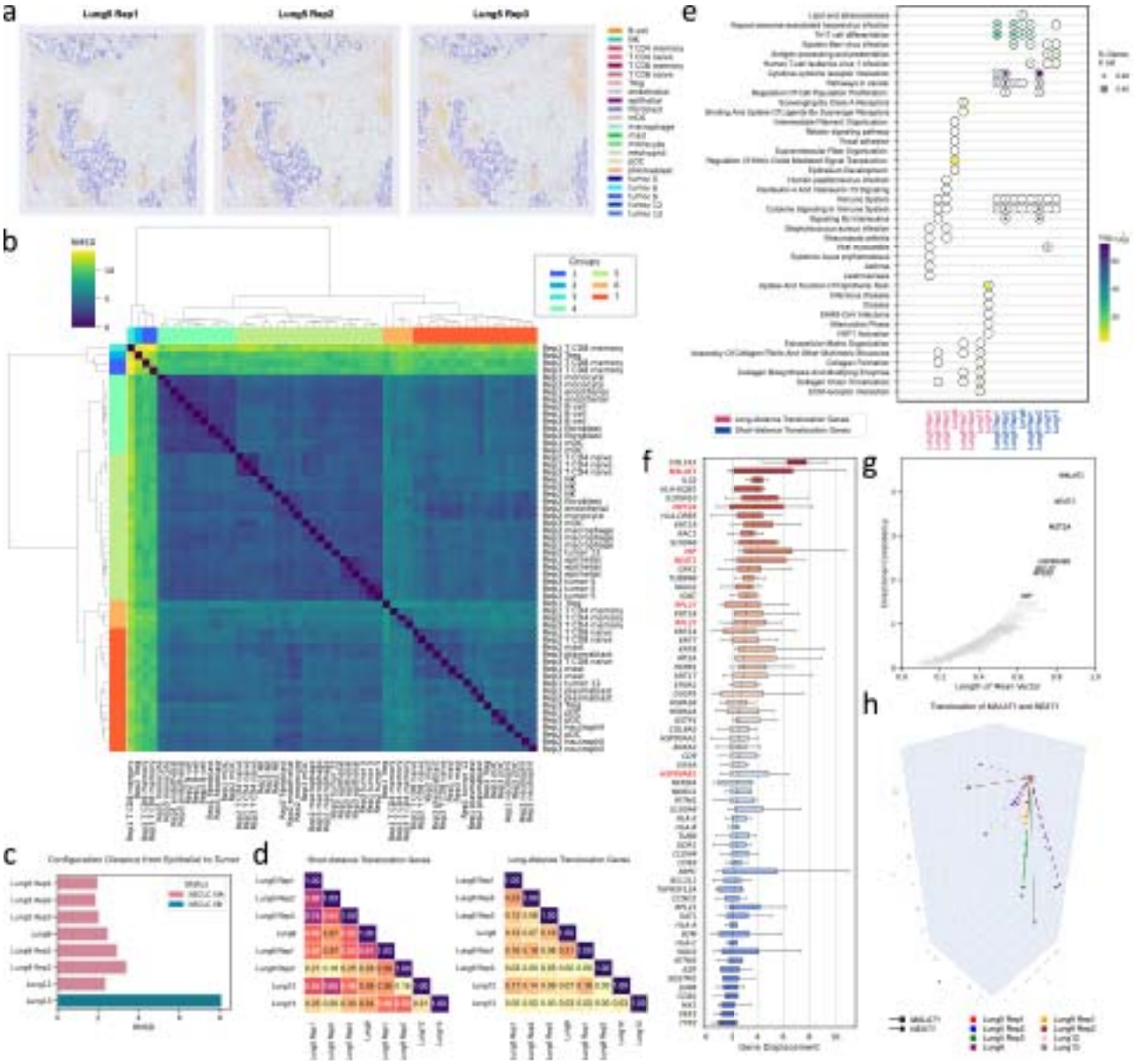

Figure 4 Spatial configuration changes from epithelial to tumor cells in human nonsmall cell lung cancer (NSCLC) [8]. (a) Spatial visualization of cell types across three biological replicates (Rep1–3) of NSCLC sample Lung5. Different colors represent distinct cell populations including epithelial cells, tumor cells, and various immune cell types (NK, natural killer cell; Treg, regulatory T cell; mDC, myeloid dendritic cell; pDC, plasmacytoid dendritic cell). (b) Configuration similarity analysis between cell types. The heatmap displays root-mean-square deviation (RMSD) values between configurations of different cell types across the three replicates, with hierarchical clustering revealing six distinct configuration groups (indicated by the annotation bars). Lower RMSD values (darker blue) indicate greater configuration similarity, with tumor cells showing highest similarity to epithelial cells (mean RMSD 2.79 ± 0.32), supporting their epithelial origin. (c) Quantitative analysis of configuration changes from epithelial to tumor cells across eight NSCLC samples. The bar chart shows RMSD values categorized by cancer staging. The significantly larger configuration difference in stage IIB (RMSD 8.09) compared with stage IIIA (mean RMSD 2.45) suggests a potential relationship between configuration changes and disease progression. (d) Conservation analysis of translocation gene sets. The heatmap shows Jaccard similarity indices for short-distance and long-distance translocation genes across eight samples, with higher values (darker shading) indicating greater overlap. The stronger conservation observed in short-distance translocation genes suggests consistent preservation of certain gene positions during epithelial-to-tumor transformation. (e) Functional enrichment of translocation genes. The diagram shows overrepresented GO terms for short-distance and long-distance translocation genes, with the significant enrichment of immune system-related terms in short-distance genes highlighting the conservation of immune signaling components during tumor development. (f) Distribution of gene displacement distances between epithelial and tumor cell configurations. The box plot shows the displacement distribution for 64 shared genes across eight samples, with highlighted genes (including *MALAT1* and *NEAT1*) exhibiting high directional consistency in their movement. (g) Quantitative assessment of directional consistency in gene repositioning. The scatter plot displays the mean vector length and directional consistency values for the repositioning unit vectors of genes from panel (f). Genes with higher values in both metrics (especially *MALAT1*) exhibit more consistent directional movement across samples, suggesting potentially nonrandom spatial reorganization during tumorigenesis. (h) Detailed analysis of *MALAT1* and *NEAT1* repositioning. The vector diagram shows the displacement distance and direction for these two long noncoding RNAs across eight samples, showing their consistent co-repositioning pattern in seven samples and divergent behavior in Lung13.

Among genes exhibiting translocation patterns, *MALAT1* emerged as having the most consistent directional movement across all samples (Fig. 4f and g). Analysis of 64 shared translocation genes revealed that *MALAT1* demonstrated both the second-largest translocation distance and the highest directional consistency (mean vector length: 0.81, directional consistency: 4.36). This observation is noteworthy given *MALAT1*'s established role as a long noncoding RNA closely associated with lung cancer metastasis [27], suggesting that consistent spatial repositioning of certain oncogenes may be associated with malignant transformation.

The translocation analysis revealed coordinated movement patterns between functionally related genes, exemplified by the relationship between *MALAT1* and *NEAT1* (Fig. 4h). These two genes, located within 70 kb on human chromosome 11 [28], maintained consistent co-translocation patterns in seven of eight samples, preserving their spatial distance during epithelial-to-tumor transformation (mean modulus difference: 0.83 ± 0.59). The notable exception in Lung13 (modulus difference: 7.16, P-value<.0001) corresponded to the sample showing the largest overall configuration change, suggesting that disruption of coordinated gene movements may indicate more severe malignant transformation. Given that both genes function as oncogenes in NSCLC—with *NEAT1* regulating gene expression through paraspeckle formation [29, 30]—their coordinated translocation patterns may represent a fundamental mechanism of spatial gene regulation during cancer progression.

Discussion

Decades of studies in 3D genome biology have established two complementary views of nuclear organization: the structural landscape of chromatin folding captured by Hi-C and related methods [3, 31], and the transcriptional landscape of RNA localization that reflects where genes are actively transcribed [4, 32]. Cytocraft addresses the latter, bridging spatial transcriptomics and the study of 3D transcriptional organization by reconstructing the spatial configuration of transcription centers—defined by RNA positions—from SST data. The method leverages iterative factorization and optimization to infer the spatial arrangement of transcription centers, based on the assumption that cells of the same type share a common 3D configuration obscured by observational noise and rotational variations. Our study validates Cytocraft's performance through simulations and applications across multiple SST platforms, demonstrating its robustness and potential for exploring spatial relationships among genes at the transcriptional level.

We reconstructed configurations for 134 cell types across 19 slices from five datasets using four platforms (Supplementary Tables 1–3). Targeted technologies (SMI and MERFISH) achieved high detection efficiency [13], while whole-transcriptome technologies (Stereo-seq) enabled localization of thousands of transcription centers despite lower efficiency.

Cytocraft may produce the true configuration or its chiral counterpart, as depth information is lost in 2D projections. However, biological structures such as DNA and proteins consistently exhibit right-handedness [33]. Thus, we assumed consistent chirality when comparing their configurations. Simulation experiments showed that parameter changes had little effect on chirality distribution, with each configuration stabilized at ∼ 50% (Supplementary Fig. 1d). Importantly, chirality ambiguity does not affect biological conclusions because all downstream analyses rely on chirality-invariant metrics (pairwise distances, RMSD after alignment). For cross-sample and developmental trajectory comparisons, we resolve chirality by aligning configurations to a common reference (Supplementary Method 1.3). Thus, all reported biological patterns remain valid regardless of which enantiomer is reconstructed.

Cytocraft also calculates cell orientations, which affect cell behavior and environment interaction. Cell alignment influences mechanical responses and communication [34], and may determine interaction probability if membrane receptor distributions are biased.

Cytocraft assumes cells of the same type share a common configuration, which may partially obscure heterogeneity from cell-cycle effects or transcriptional bursting (Supplementary Result 2.1.6). However, the weighted centroid approach dampens stochastic fluctuations, and residual deviations (Fig. 3b and c) capture cell-to-cell variability. Future extensions could incorporate mixture models to explicitly model subpopulations.

Like all structure-from-motion approaches, Cytocraft's solution is subject to inherent gauge ambiguities. The reconstructed configuration is determined only up to an arbitrary global rotation, uniform scaling, translation, and reflection (chirality). We mitigate these ambiguities as follows: (i) translation is resolved by centering transcription centers within each cell; (ii) scale ambiguity is addressed by normalizing configurations to unit variance, enabling relative (but not absolute) distance comparisons; (iii) rotation ambiguity is handled via Procrustes alignment when comparing configurations across cell types or samples; (iv) chirality is resolved by aligning to a common reference for directional analyses (Supplementary Method 1.3), while chirality-invariant metrics (pairwise distances, RMSD) are used for most biological interpretations. Importantly, these ambiguities affect only the coordinate system, not the intrinsic geometry of the configuration—all pairwise gene–gene distances and clustering relationships are preserved regardless of gauge choice.

Cytocraft's computational complexity scales as $O(m \cdot n \cdot k)$ per iteration, where m is the number of cells, n is the number of genes, and k is the number of iterations to convergence (typically 10–40). For the datasets analyzed here (586–42 011 cells, 980–25 136 genes), runtime ranged from 20 min to 100 h per cell type on a standard workstation (Supplementary Fig. 1b). Memory requirements scale as $O(m \cdot n)$ for storing transcription center coordinates, ranging from 0.7 GB for smaller cell types with ∼20 GB for the largest. While sufficient for current SST technologies, future platforms with substantially higher cell counts (>50 000 cells per type) or gene panels (>20 000 genes) may require algorithmic optimizations such as stochastic gradient descent, mini-batch processing, or GPU acceleration. We note that the per-cell-type analysis strategy naturally parallelizes across cell types, enabling efficient scaling for multi-cell-type datasets.

Although Cytocraft has yielded promising results, additional improvements merit consideration. First, more applications regarding spatial configurations need development. For example, integrating Cytocraft's transcriptional configurations with Hi-C-derived chromatin structures could reveal how the "transcriptional landscape" (RNA localization) relates to the "structural landscape" (DNA contacts) within the same cell types. Second, extending Cytocraft to a fully probabilistic framework would enable uncertainty quantification for reconstructed coordinates and formal model comparison.

To address whether Cytocraft recovers chromatin architecture, we compared Cytocraft-inferred distances with Hi-C contact frequencies from A549 lung cancer cells (ENCODE ENCFF689CUX [35]). Across 18

sample-cell type combinations, Cytocraft distances showed negligible correlation with Hi-C contacts (mean Spearman ρ = –0.06), whereas genomic distance exhibited the expected strong negative correlation (ρ = –0.54), consistent with polymer physics models of chromatin [3]. Furthermore, Cytocraft distances showed near-zero correlation with linear genomic distance (ρ = 0.05), confirming that genes close in the inferred configuration are not simply genomically proximal. This apparent discrepancy is not a methodological failure but rather confirms that Cytocraft captures a fundamentally different aspect of nuclear organization. Hi-C measures physical DNA–DNA contacts reflecting chromatin folding and chromosomal territories [31], whereas Cytocraft reconstructs transcription center configurations—the spatial arrangement of active transcription sites based on nascent RNA localization [4, 32]. These represent complementary views: the "structural landscape" (Hi-C) versus the "transcriptional landscape" (Cytocraft). Genes co-localized in Cytocraft configurations share transcriptional neighborhoods (e.g. nuclear speckle proximity [5, 36]) without necessarily being genomically proximal or forming direct chromatin contacts [6]. This distinction underscores that Cytocraft should be interpreted as revealing transcriptional co-localization patterns rather than chromatin 3D structure (Supplementary Fig. 7).

Highly expressed genes such as *MALAT1* may be subject to quantification biases in single-cell technologies [37], but we found no evidence of saturation (broad dynamic range, 74.3% detection rate, and high variability). Importantly, Cytocraft depends on transcript positions rather than counts, and the consistent directional translocation across all eight samples argues against systematic bias (Supplementary Fig. 8 and Result 2.3.4).

In summary, modeling the 3D configuration of transcription centers with Cytocraft opens new possibilities for exploring how the spatial organization of transcription may relate to cellular function. Future studies applying Cytocraft to more extensive datasets may provide new insights into complex genetic diseases and developmental processes.

Materials and methods

Framework of Cytocraft

Cytocraft infers 3D spatial configurations of transcription centers from 2D SST data. It identifies 2D transcription centers as expression-weighted geometric centers of transcriptional spots per gene, treating these as noisy projections of the underlying 3D configuration.

Cytocraft iteratively alternates between the F-step (updating configuration given rotations) and R-step (refining rotations given configuration), minimizing distance between projected and observed transcription centers to recover the latent spatial configuration.

Estimating transcription centers

From SST data of m cells and n genes, Cytocraft utilizes *priori* knowledge of cell-spot membership to estimate the coordinates of transcription centers for each cell.

A cell c contains l_c spots with expression profiles $\mathbf{E_c} = (e_1, e_2, \cdots, e_{l_c})$ and coordinates $\mathbf{P} = (\mathbf{p_1}, \mathbf{p_2}, \cdots, \mathbf{p_{l_c}})$. Within a cell c, Cytocraft locates the transcription center for each gene g as the geometric center of spots weighed by expression, i.e. $(x_{c,g}, y_{c,g}) = \frac{\sum_s^{l_c} e_{s,g}\mathbf{p}_s}{\sum_s^{l_c} e_{s,g}}$. Cytocraft further normalizes the transcription center among all genes in the cell c as $(\hat{x}_{c,g}, \hat{y}_{c,g}) = (x_{c,g} - \overline{x}_c, y_{c,g} - \overline{y}_c)$, where $\overline{x}_c$ and $\overline{y}_c$ denote the average coordinates. Consequently, for cell c, the normalized transcription centers are summarized as a vector $\mathbf{z_c} = (\hat{\mathbf{x}}_c | \hat{\mathbf{y}}_c) \in \mathbb{R}^{2n}$.

Collecting the transcription centers for all cells, Cytocraft constructs a matrix $\mathbf{Z} = (\mathbf{z_1}, \mathbf{z_2}, \cdots, \mathbf{z_m}) \in \mathbb{R}^{2n \times m}$. To enhance the precision of the configuration, genes expressed in <10% of cells are excluded.

Inferring spatial configurations with iterative factorization and optimization

Assuming that each cell on the slice is a 2D projection of its 3D configuration from an arbitrary angle, our primary objective is to recover the 3D configuration from these observed 2D projections. Specifically, given rotation matrices that represent cell orientations, Cytocraft infers a shared 3D configuration with the observed transcription centers of cells and the rotation matrices, represented by the configuration matrix $\mathbf{F} \in \mathbb{R}^{n \times 3}$, where the vector $\mathbf{f_g}$ denotes the 3D coordinates for the gene g. This process is termed the F-step. Conversely, given $\mathbf{F}$, Cytocraft refines each cell's rotation matrix $\mathbf{R_c} \in \mathbb{R}^{3 \times 3}$ to best align the 2D projection of the rotated configuration with the observed transcription centers. The rotation matrices are collectively represented as a rotation tensor $\mathbf{R} \in \mathbb{R}^{m \times 3 \times 3}$. This process is termed the R-step.

Initially, Cytocraft starts with a set of random rotation matrices to generate the initial configuration matrix in the F-step. The algorithm then alternates between the R-step and the F-step until both the rotation tensor $\mathbf{R}$ and the configuration matrix $\mathbf{F}$ converge. A high-level pseudocode summarizing these steps is provided in Algorithm 1. The following sections offer a detailed description of each step. In this work, we apply Cytocraft to cells of the same type, generating a single common configuration for each cell type.

F-step: inferring the configuration matrix from the transcription centers

Given a rotation tensor $\mathbf{R}$ of all cells, Cytocraft infers the optimal $\mathbf{F}$ by requiring that the configuration matrix projected by each rotation matrix $\mathbf{R_c}$ should approximate the observed $2 \times n$ transcription centers for all cells. In the initial epoch, a random rotation tensor is used. Let $\mathbf{R_c}$ denotes the rotation matrix for cell c, we can formulate the objective function to find the optimal $\mathbf{F}^*$ from $\mathbf{Z}$ and $\mathbf{R}$ as:

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} h(\mathbf{F}|(\mathbf{R}, \mathbf{Z})) = \arg\min_{\mathbf{F}} \sum_c ||\mathbf{F} \cdot \mathbf{R_c} \cdot \mathbf{T} - \mathbf{Z_c}||_F^2, \quad (1)$$

where $\mathbf{T}$ is an orthogonal projection operator that sets the last row of a matrix to zeros, and $||\cdot||_F$ denotes the Frobenius norm.

Considering that $\mathbf{f_g}$ are independent among genes, the objective function can be written as

$$\mathbf{F}^* = \arg\min_{\mathbf{f_g}, \forall g} h_g(\mathbf{f_g}|(\mathbf{R}, \mathbf{Z})) = \arg\min_{\mathbf{f_g}, \forall g} \sum_c \mathbf{f_g} \cdot \begin{bmatrix} \mathbf{R_{c1}} \\ \mathbf{R_{c2}} \\ \mathbf{R_{c3}} \end{bmatrix} \cdot \mathbf{T} - (x_{c,g}, y_{c,g}, 0)$$

$$= \arg\min_{\mathbf{f_g}, \forall g} \sum_c (\mathbf{f_g} \cdot \mathbf{R_{c1}} - x_{c,g})^2 + (\mathbf{f_g} \cdot \mathbf{R_{c2}} - y_{c,g})^2. \quad (2)$$

We solve for $\mathbf{f_g^*}$ by setting the partial derivatives of (f_{g1}, f_{g2}, f_{g3}) as zero, yielding the linear system below

$$\begin{cases} \partial_{f_{g1}} h_g(\mathbf{f_g}|(\mathbf{R},\mathbf{Z})) = \sum_c R_{c1,1}(\mathbf{f_g}\cdot\mathbf{R_{c1}} - x_{c,g}) + R_{c1,2}(\mathbf{f_g}\cdot\mathbf{R_{c2}} - y_{c,g}) = 0; \\ \partial_{f_{g2}} h_g(\mathbf{f_g}|(\mathbf{R},\mathbf{Z})) = \sum_c R_{c2,1}(\mathbf{f_g}\cdot\mathbf{R_{c1}} - x_{c,g}) + R_{c2,2}(\mathbf{f_g}\cdot\mathbf{R_{c2}} - y_{c,g}) = 0; \\ \partial_{f_{g3}} h_g(\mathbf{f_g}|(\mathbf{R},\mathbf{Z})) = \sum_c R_{c3,1}(\mathbf{f_g}\cdot\mathbf{R_{c1}} - x_{c,g}) + R_{c3,2}(\mathbf{f_g}\cdot\mathbf{R_{c2}} - y_{c,g}) = 0. \end{cases} \tag{3}$$

In particular, when the linear system is unsolvable in the case of a singular matrix or a near-singular matrix, the gene g will be dropped.

R-step: estimating a rotation tensor from transcription centers with factorization and optimization

From the observed transcription centers $\mathbf{Z}$, Cytocraft estimates the cell rotation tensor with factorization [11] based on singular value decomposition (SVD), followed by further optimization of the estimated rotation tensor given the configuration matrix from the F-step. The projected transcription centers for two cells c and c' are $(\mathbf{i_c^T f}, \mathbf{j_c^T f})$ and $(\mathbf{i_{c'}^T f}, \mathbf{j_{c'}^T f})$, respectively. Here, $\mathbf{i_c}$ and $\mathbf{j_c}$ are orthogonal 3D unit vectors that represent the projection angle, or the *orientation* of cell c. Across observed cells, we denote the orientation matrix as $\mathbf{O} = (\mathbf{I}|\mathbf{J})^T \in \mathbb{R}^{2n\times 3}$, where $\mathbf{I} = (\mathbf{i_1}, \mathbf{i_2}, \cdots, \mathbf{i_n})$ and $\mathbf{J} = (\mathbf{j_1}, \mathbf{j_2}, \cdots, \mathbf{j_n})$. Under orthography, let the configuration matrix $\mathbf{F} \in \mathbb{R}^{3\times m}$ denote the coordinates of transcription configurations, and the projection process can be formulated as $\mathbf{Z} = \mathbf{OF}$.

Since the observations in $\mathbf{Z}$ could be noisy, we approximate $\mathbf{Z}$ with $\hat{\mathbf{Z}} = \mathbf{U\Sigma V}$ from the SVD by keeping the three largest singular values. Given the projection process $\mathbf{Z} = \mathbf{OF}$, we can approximate the orientation matrix $\mathbf{O}$ with $\hat{\mathbf{O}} = \mathbf{U}[\mathbf{\Sigma}]^{\frac{1}{2}}$. Furthermore, we need to rotate the approximated matrix to meet the given 3D-to-2D projection condition, for which we introduce a linear transformation matrix $\mathbf{Q}$ satisfying $\hat{\mathbf{i}_c^T}\mathbf{Q}\mathbf{Q^T}\hat{\mathbf{i}_c} = \hat{\mathbf{j}_c^T}\mathbf{Q}\mathbf{Q^T}\hat{\mathbf{j}_c} = 1$ and $\hat{\mathbf{i}_c^T}\mathbf{Q}\mathbf{Q^T}\hat{\mathbf{j}_c} = 0$ for any cell c. After transformation, we obtain the orientation matrix $\mathbf{O} = \hat{\mathbf{O}}\mathbf{Q}$. From the orientation matrix, Cytocraft derives the rotation matrix $\mathbf{R_c^{fac}} \in \mathbb{R}^{3\times 3} = (\mathbf{i_c}, \mathbf{j_c}, \mathbf{i_c}\times\mathbf{j_c})$ for each cell c.

On the other hand, factorization on $\mathbf{Z}$ also yields a spatial configuration matrix $\mathbf{F^{fac}} = \mathbf{Q^{-1}}[\mathbf{\Sigma}]^{\frac{1}{2}}\mathbf{V}$. To maintain the consistency of the spatial configuration direction, Cytocraft finds a rotation matrix $\mathbf{R_c^{rot}}$ that minimizes the RMSD between $\mathbf{R_c^{rot}F^{fac}}$ and $\mathbf{F}$ using the Kabsch–Umeyama algorithm [38]. The final rotation matrix is therefore $\mathbf{R_c} = \mathbf{R_c^{rot}R_c^{fac}}$ for each cell c. In practice, we select a subset of the most highly expressed genes as anchor genes to perform the R-step, ensuring efficient and robust modeling.

Algorithm 1 Recover 3D Transcription Configuration and Rotation Tensor from 2D Observations

Require: Observed transcription centers $\mathbf{Z}$, number of genes n, number of cells m
Ensure: Configuration matrix $\mathbf{F}$, Rotation tensor $\mathbf{R}$

Initialize random rotation tensor $\mathbf{R} \in \mathbb{R}^{m\times 3\times 3}$
 # Initial F-step: Obtain initial configuration matrix F from random R
for gene $g = 1, 2, \ldots, n$ **do**
  Solve the linear system in Eq. (3) for $\mathbf{f_g}$ using random $\mathbf{R}$
  if linear system is unsolvable **then**
    Remove gene g from $\mathbf{Z}$ and $\mathbf{F}$
  else
    Initialize $\mathbf{F_g}$ with solution $\mathbf{f_g}$
  end if
end for
for epoch = 1, 2, ..., max_epochs **do**
  # R-step: Estimate the rotation tensor R from the transcription centers Z and the configuration F
  Compute singular value decomposition: $\mathbf{Z} \approx \hat{\mathbf{Z}} = \mathbf{U\Sigma V^T}$
  Approximate orientation matrix $\hat{\mathbf{O}} = \mathbf{U}[\mathbf{\Sigma}]^{\frac{1}{2}}$
  Find the transformation matrix $\mathbf{Q}$ satisfying projection conditions
  Obtain the orientation matrix $\mathbf{O} = \hat{\mathbf{O}}\mathbf{Q}$
  Derive factorized rotation matrices $\mathbf{R_c^{fac}}$ from $\mathbf{O}$ for each cell c
  Compute factorized configuration $\mathbf{F^{fac}} = \mathbf{Q^{-1}}[\mathbf{\Sigma}]^{\frac{1}{2}}\mathbf{V^T}$
  Find optimal rotation matrices $\mathbf{R_c^{rot}}$ minimizing RMSD between $\mathbf{R_c^{rot}F^{fac}}$ and $\mathbf{F}$
  Update rotation tensor $\mathbf{R}$ with $\mathbf{R_c} = \mathbf{R_c^{rot}R_c^{fac}}$ for each cell c
  # F-step: Optimize configuration matrix F given updated rotation tensor R
  for gene $g = 1, 2, \ldots, n$ **do**
    Solve the linear system in Eq. (3) for $\mathbf{f_g}$ using updated $\mathbf{R}$
    if linear system is unsolvable **then**
      Remove gene g from $\mathbf{Z}$ and $\mathbf{F}$
    else
      Update $\mathbf{F_g}$ with the solution $\mathbf{f_g}$
    end if
  end for
end for

return F, R

Gene spatial relationships revealed from configurations

We analyze gene spatial relationships using pairwise Euclidean distances from reconstructed coordinates.

Gene components and gene ontology analysis

A gene component is a group of spatially proximal genes in the configuration that potentially represent biological functions and regulatory relationships. To identify gene components, we build the KNN graph with $K = 5$ from the spatial configuration. Subsequently, we extract gene components by the Louvain community detection method implemented in NetworkX [39] with default parameters. We further sort and index the identified components by their number of genes, from largest to smallest.

We examine the possible biological functions of gene components (>5 genes) through GO analysis, assessing the 10 largest for enrichments in KEGG, Reactome, and Biological Process databases (adjusted P <.05, ≥2 genes).

Compare spatial distances of regulation pairs inferred by GENIE3

To identify the regulatory gene pairs, we employed the Gene Network Inference with Ensemble of Trees (GENIE3) algorithm [25]. The expression matrix is subjected to a GENIE3 algorithm analysis to reconstruct the co-expressed gene network for each transcription center. We filter regulation pairs with a weight threshold of 0.05. Then, we compare the spatial distances of the regulation pairs inferred by

GENIE3 with the distances of all known regulation pairs and all gene pairs.

Compare gene chromosome length with number of radius-defined neighbors

To analyze the relationship between the chromosomal length of a gene and the number of its neighboring genes, we first construct a radius-nearest neighbor (RNN) graph, using predefined thresholds to identify neighboring genes, which we refer to as proximal genes. Then, we extract the lengths of the gene chromosomes from the gene annotation files for specific species. Next, we perform a linear regression to determine the correlation between the number of proximal genes and the cube root of the gene chromosome lengths.

Benchmark performance using simulated configuration

We applied simulated spatial configurations with knowledge of genome structures to assess the precision of the reconstruction by Cytocraft. We simulate transcription centers for cells by projecting the spatial configuration onto the 2D plane with randomized cell orientations. Specifically, we simulate a chromosome-like configuration matrix $\mathbf{F^{sim}}$. Next, we apply a set of random rotation matrices $\mathbf{R^{random}}$ on the simulated configuration matrix, each corresponding to a cell. For each rotated cell, we simulate the transcription centers with consideration of sequencing characteristics, namely the resolution r and the noise level σ. The larger the configuration scaled with resolution r, the greater the number of grids within the cell boundaries, as a way to simulate the high-resolution case. Given a resolution r, we locate the transcription centers as $\mathbf{Z^{sim}} = r\mathbf{F^{sim}} \cdot \mathbf{R^{random}} \cdot \mathbf{T} + \mathcal{N}(0,\sigma)$, where $\mathcal{N}(0,\sigma)$ represents independent Gaussian noise with zero mean and standard deviation σ (in pixels) added to both the x and y coordinates of each transcription center. This noise models the positional uncertainty arising from experimental sources such as optical aberrations, image registration errors, and molecular localization imprecision. A noise level of $\sigma = 0$ represents ideal localization, while $\sigma = 2$ and $\sigma = 5$ represent low and high levels of positional uncertainty, respectively.

From transcription centers, we generate gene expression matrices at various noise levels Σ, detection efficiencies λ, and gene dropout rates θ based on SST data characteristics. SST technologies generally characterize low detection efficiency and different resolutions [13]. In addition, positional offsets between the detected and actual transcript positions can be introduced during experiments. Assuming gene g in cell c follows a Gaussian distribution around its center, the probability density function is $f(x,y|Z^{\text{sim}}_{c,g})$ of mean $\mu = Z_{c,g}$ and covariance $\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$. Detection efficiency λ follows binomial distribution $B(m|n,\lambda)$, which represents the probability of the exact m appearance of the gene g in a total n spots within the cell c. Following this, the transcription centers for certain genes are lost based on a provided gene dropout rate θ, for which we define an indicator $L \in \{0,1\}$ where $P(L=0) = \theta$ and $P(L=1) = 1-\theta$. Thus, the final spatial expression matrix $\mathbf{M}$ is calculated by

$$M_{c,g} = \lfloor \alpha \cdot f(x,y|Z^{\text{sim}}_{c,g}, \Sigma) \rceil \cdot B(m|n,\lambda) \cdot L, \tag{4}$$

where α represents the number of all current transcripts of a gene g in a cell c and $\sigma_1 = \sigma_2 = 8$ by default.

For an overall evaluation of Cytocraft's robustness, various parameter combinations were tested (Supplementary Table 4) with 10 repetitions each. We evaluated the performance by comparing the reconstructed spatial configuration with the original spatial configuration by the relative error and Spearman correlation. The relative error between the reconstructed 3D coordinates ($\mathbf{C_{re}}$) and the benchmark coordinates ($\mathbf{C_{benchmark}}$) of transcription centers was calculated as: $RE = ||\mathbf{C_{re}} - \mathbf{C_{benchmark}}||_2^2 / ||\mathbf{C_{benchmark}}||_2^2$.

Root-mean-square deviation as a metric to evaluate configuration similarity

RMSD is minimized using the Kabsch–Umeyama algorithm [38]. This algorithm achieves optimal superposition by minimizing RMSD through the rotation and translation of the structures. We calculate RMSD using the top k shared genes with highest expression to ensure reliable comparisons.

Configuration alignment

Configurations are aligned using Kabsch–Umeyama [38] to enable cross-sample comparison of gene displacement directions (Supplementary Method 1.3 and 1.4).

Identifying conserved subconfigurations

To identify subconfigurations that are conserved across samples with similar biological conditions or successive developmental stages, we first obtained the set of genes shared by all samples and then took the distance matrix of these genes as input. Let n denotes the number of genes shared by all m samples, we use $\mathbf{A_k} \in \mathbb{R}^{n\times n}$ to denote the gene–gene distance matrix in the kth sample. We calculate the standard deviation matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ across all samples as: $\mathbf{S} = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(\mathbf{A_k} - \overline{\mathbf{A}})^2}$, where $\overline{\mathbf{A}} = \frac{1}{m}\sum_{k=1}^{m}\mathbf{A_k}$. We then used hierarchical clustering to identify sets of conserved genes whose standard deviations of the distances between genes were all small, signifying that their configurations were relatively conserved across samples. These consistent configurations across samples are called subconfigurations.

Key Points

- Cytocraft reconstructs 3D gene architecture from 2D spatial transcriptomics data by treating each cell as a different angular projection of a shared 3D configuration of transcription centers, solving the fundamental "crushed dimension" problem in current spatial genomics approaches.
- Consistent oncogenic translocations are revealed in human lung cancer, with the cancer-associated gene *MALAT1* showing highly directional and conserved spatial repositioning during epithelial-to-tumor transformation across eight patient samples, suggesting nonrandom spatial reorganization during tumorigenesis.
- Conserved developmental dynamics emerge across different cell types in the developing axolotl brain, where gene spatial reorganization follows similar temporal patterns regardless of cell-type identity, indicating fundamental principles governing spatial gene organization during development.

- Functional gene organization is linked to spatial proximity, with genes clustering into components that reflect their biological roles, chromosomal relationships preserved in 3D space, and regulatory gene pairs maintaining closer spatial distances than random pairs.
- The method demonstrates robust performance across multiple platforms (MERFISH, CosMx SMI, Xenium, and Stereo-seq) with median relative error of 0.0346 in simulations, enabling broad application to study the spatial logic of the genome in development and disease.

Acknowledgements

Figures 1 and 2a were created with BioRender.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Conflicts of interest

None declared.

Funding

We express our gratitude for the support provided by the National Natural Science Foundation of China (Grant No. 32270687), and the National Key R&D Program of China (Grant No. 2023YFC3403200).

Data availability

The mouse embryo and axolotl development datasets generated by Stereo-seq are available under STOmicsDB [40] with accession numbers STDS0000058 and STDS0000056. The NSCLC dataset generated by the CosMx SMI platform is available at http://nanostring.com/CosMx-dataset. The mouse ileum dataset generated by the MERFISH platform is available at Dryad [41]. The human breast cancer dataset generated by Xenium *in situ* is downloaded from the GEO database with accession number GSE243168. Cytocraft's open-source code is available on GitHub and published under the MIT license. In addition, all scripts and notebooks required to reproduce the analysis and figures presented in this manuscript have been deposited in the same repository under the reproducibility directory. Cytocraft is released through the Pythonpackageindex.

References

1. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. *Nature* 2007; **447**:413–7. https://doi.org/10.1038/nature05916
2. Misteli T. The self-organizing genome: principles of genome architecture and function. *Cell* 2020; **183**:28–45. https://doi.org/10.1016/j.cell.2020.09.014
3. McCord RP, Kaplan N, Giorgetti L. Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Mol Cell* 2020; **77**:688–708. https://doi.org/10.1016/j.molcel.2019.12.021
4. Papantonis A, Cook PR. Transcription factories: genome organization and gene regulation. *Chem Rev* 2013; **113**:8683–705. https://doi.org/10.1021/cr300513p
5. Chen Y, Belmont AS. Genome organization around nuclear speckles. *Curr Opin Genet Dev* 2019; **55**:91–9. https://doi.org/10.1016/j.gde.2019.06.008
6. Su J-H, Zheng P, Kinrot SS *et al.* Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell* 2020;**182**:1641–1659.e26. https://doi.org/10.1016/j.cell.2020.07.032
7. Xia C, Fan J, Emanuel G *et al.* Spatial transcriptome profiling by merfish reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc Natl Acad Sci USA* 2019;**116**:19490–9. https://doi.org/10.1073/pnas.1912459116
8. He S, Bhatt R, Brown C *et al.* High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol* 2022;**40**:1794–806. https://doi.org/10.1038/s41587-022-01483-z
9. Janesick A, Shelansky R, Gottscho AD *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun* 2023;**14**:8353. https://doi.org/10.1038/s41467-023-43458-x
10. Chen A, Liao S, Cheng M *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* 2022;**185**:1777–1792.e21. https://doi.org/10.1016/j.cell.2022.04.003
11. Tomasi C, Kanade T. Shape and motion from image streams under orthography: a factorization method. *Int J Comp* 1992;**9**:137–54. https://doi.org/10.1007/BF00129684
12. Wang H, Yang J, Zhang Y *et al.* Reconstruct high-resolution 3D genome structures for diverse cell-types using flamingo. *Nat Commun* 2022;**13**:2645. https://doi.org/10.1038/s41467-022-30270-2
13. Yue L, Liu F, Hu J *et al.* A guidebook of spatial transcriptomic technologies, data resources and analysis approaches. *Comput Struct Biotechnol J* 2023;**21**:940–55. https://doi.org/10.1016/j.csbj.2023.01.016
14. Yap L, Wang JW, Moreno-Moral A *et al.* In vivo generation of post-infarct human cardiac muscle by laminin-promoted cardiovascular progenitors. *Cell Rep* 2019;**26**:3231–3245.e9. https://doi.org/10.1016/j.celrep.2019.02.083
15. Clifton C, Niemeyer BF, Novak R *et al.* BPIFA1 is a secreted biomarker of differentiating human airway epithelium. *Front Cell Infect Microbiol* 2022;**12**:1035566. https://doi.org/10.3389/fcimb.2022.1035566
16. Kneussel M, Betz H. Clustering of inhibitory neurotransmitter receptors at developing postsynaptic sites: the membrane activation model. *Trends Neurosci* 2000;**23**:429–35. https://doi.org/10.1016/S0166-2236(00)01627-1
17. Fraser ST, Isern J, Baron MH. Maturation and enucleation of primitive erythroblasts during mouse embryogenesis is accompanied by changes in cell-surface antigen expression. *Blood* 2007;**109**:343–52. https://doi.org/10.1182/blood-2006-03-006569
18. Rosso G, Liashkovich I, Shahin V. In situ investigation of interrelationships between morphology and biomechanics of endothelial and glial cells and their nuclei. *Adv Sci* 2019;**6**:1801638. https://doi.org/10.1002/advs.201801638
19. Roman W, Gomes ER. Nuclear positioning in skeletal muscle. In: Cadot B (ed.), *Seminars in Cell & Developmental Biology*, Vol. **82**, pp. 51–6. Netherlands: Elsevier, 2018.
20. Rao SS, Huntley MH, Durand NC *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80. https://doi.org/10.1016/j.cell.2014.11.021
21. Hewett SJ, Jackman NA, Claycomb RJ. Interleukin-1β in central nervous system injury and repair. *Eur J Neurodegener Dis* 2012;**1**:195–211.

22. Nguyen T, Di Giovanni S. NFAT signaling in neural development and axon growth. *Int J Dev Neurosci* 2008;**26**:141–5. https://doi.org/10.1016/j.ijdevneu.2007.10.004
23. Fraser J, Ferrai C, Chiariello AM *et al.* Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol* 2015;**11**:852. https://doi.org/10.15252/msb.20156492
24. Oudelaar AM, Higgs DR. The relationship between genome structure and function. *Nat Rev Genet* 2021;**22**:154–68. https://doi.org/10.1038/s41576-020-00303-x
25. Huynh-Thu VA, Irrthum A, Wehenkel L *et al.* Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 2010;**5**:e12776. https://doi.org/10.1371/journal.pone.0012776
26. Sainz de Aja J, Dost A, Kim C. Alveolar progenitor cells and the origin of lung cancer. *J Intern Med* 2021;**289**:629–35. https://doi.org/10.1111/joim.13201
27. Gutschner T, Hämmerle M, Eißmann M *et al.* The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res* 2013;**73**:1180–9. https://doi.org/10.1158/0008-5472.CAN-12-2850
28. Hutchinson JN, Ensminger AW, Clemson CM *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* 2007;**8**:1–16. https://doi.org/10.1186/1471-2164-8-39
29. Hirose T, Virnicchi G, Tanigawa A *et al.* Neat1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Mol Biol Cell* 2014;**25**:169–83. https://doi.org/10.1091/mbc.e13-09-0558
30. Jen J, Tang YA, Lu YH *et al.* Oct4 transcriptionally regulates the expression of long non-coding RNAs NEAT1 and MALAT1 to promote lung cancer progression. *Mol Cancer* 2017;**16**:1–12. https://doi.org/10.1186/s12943-017-0674-z
31. Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. *Cell* 2016;**164**:1110–21. https://doi.org/10.1016/j.cell.2016.02.007
32. Martin KC, Ephrussi A. mRNA localization: gene expression in the spatial dimension. *Cell* 2009;**136**:719–30. https://doi.org/10.1016/j.cell.2009.01.044
33. Huber G, Knecht C, Trump W *et al.* Entropy and chirality in sphinx tilings. *Phys Rev Res* 2024;**6**:013227. https://doi.org/10.1103/PhysRevResearch.6.013227
34. Tamiello C, Buskermolen AB, Baaijens FP *et al.* Heading in the right direction: understanding cellular orientation responses to complex biophysical environments. *Cell Mol Bioeng* 2016;**9**:12–37. https://doi.org/10.1007/s12195-015-0422-7
35. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. https://doi.org/10.1038/nature11247
36. Quinodoz SA, Ollikainen N, Tabak B *et al.* Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* 2018;**174**:744–757.e24. https://doi.org/10.1016/j.cell.2018.05.024
37. Squair JW, Gautier M, Kathe C *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* 2021;**12**:5692. https://doi.org/10.1038/s41467-021-25960-2
38. Umeyama S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans Pattern Anal Mach Intell* 1991;**13**:376–80. https://doi.org/10.1109/34.88573
39. Hagberg A, Swart PJ, Schult DA. Exploring Network Structure, Dynamics, and Function Using Networkx. In: Gäel Varoquaux, Travis Vaught, Jarrod Millman (eds.), *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA USA, pp. 11–15, Aug 2008.
40. Xu Z, Wang W, Yang T *et al.* STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. *Nucleic Acids Res* 2024;**52**:D1053–61. https://doi.org/10.1093/nar/gkad933
41. Petukhov V, Xu RJ, Soldatov RA *et al.* Cell segmentation in imaging-based spatial transcriptomics. *Nat Biotechnol* 2022;**40**:345–54. https://doi.org/10.1038/s41587-021-01044-w

Briefings in Bioinformatics, 2026, **27**, bbag177
https://doi.org/10.1093/bib/bbag177
Problem Solving Protocol