nature communications

Article

https://doi.org/10.1038/s41467-024-49457-w

Intracellular spatial transcriptomic analysis toolkit (InSTAnT)

Received: 27 October 2023

Accepted: 4 June 2024

Published online: 06 September 2024

Check for updates

Anurendra Kumar[1], Alex W. Schrader[2], Bhavay Aggarwal[3], Ali Ebrahimpour Boroojeny[4], Marisa Asadian[2], JuYeon Lee[2], You Jin Song[5], Sihai Dave Zhao[6,7] ✉, Hee-Sun Han[2,7] ✉ & Saurabh Sinha[8,9] ✉

Imaging-based spatial transcriptomics technologies such as Multiplexed error-robust fluorescence in situ hybridization (MERFISH) can capture cellular processes in unparalleled detail. However, rigorous and robust analytical tools are needed to unlock their full potential for discovering subcellular biological patterns. We present Intracellular Spatial Transcriptomic Analysis Toolkit (InSTAnT), a computational toolkit for extracting molecular relationships from spatial transcriptomics data at single molecule resolution. InSTAnT employs specialized statistical tests and algorithms to detect gene pairs and modules exhibiting intriguing patterns of co-localization, both within individual cells and across the cellular landscape. We showcase the toolkit on five different datasets representing two different cell lines, two brain structures, two species, and three different technologies. We perform rigorous statistical assessment of discovered co-localization patterns, find supporting evidence from databases and RNA interactions, and identify associated subcellular domains. We uncover several cell type and region-specific gene co-localizations within the brain. Intra-cellular spatial patterns discovered by InSTAnT mirror diverse molecular relationships, including RNA interactions and shared sub-cellular localization or function, providing a rich compendium of testable hypotheses regarding molecular functions.

A grand challenge in biology is to understand how molecules and cells cooperatively perform higher-level processes and how these processes are coordinated. An emerging involves using single-cell sequencing technologies to profile cellular composition and states at unprecedented resolution[1,2]. Spatial omics technologies further bolster this approach by characterizing the spatial organization of molecules and cells, providing insights into their functional organization. Most existing analytic tools for extracting biological insights from spatial data have focused on cell-level or coarser resolution analyses (Fig. 1a). These include detecting spatially variable genes[3,4], identifying cell types and spatial domains[5–7], and inferring cell-cell interaction[8–10]. This is true also for grid-based spatial encoding technologies[11,12], where the grid size limits resolution to be super-cellular. Even with single-molecule resolution technologies[13–17], tissue-scale analyses mostly set the unit of analysis to be a cell[18]. The focus on cell-level analyses is likely due to the straightforward interpretations they provide, such as

[1]College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA. [2]Department of Chemistry, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. [3]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA. [4]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. [5]Department of Cell and Developmental Biology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. [6]Department of Statistics, University of Illinois Urbana-Champaign, Urbana, IL 61820, USA. [7]Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. [8]H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30318, USA. [9]The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.
✉e-mail: sdzhao@illinois.edu; hshan@illinois.edu; saurabh.sinha@bme.gatech.edu

cellular arrangements around diseased phenotypes[19], cellular interactions[20,21], and spatial context-dependent cell functions[7,18,22,23].

Despite this initial progress, the focus on cell-level analyses has left higher-resolution analyses of spatial omics data relatively unexplored. Analyzing subcellular expression patterns can add new dimensions to our understanding of cell functions. For example, decades of work has highlighted the importance of RNA localization in transcriptional regulation[24–26], translational regulation[27,28], protein localization[29,30] and protein complex assembly[26]. While traditional in situ hybridization methods have revealed the functional implications of RNA localization, those studies are mostly limited to a handful of genes. New single-molecule resolution spatial transcriptomics technologies offer an unprecedented window into this world of sub-cellular organization, at far greater scale than before. For instance, Xia et al. [31] profiled the locations of transcripts from roughly 10,000 genes to infer the pseudotime ordering of cells using transcript distribution in nuclei versus cytoplasm[32,33] while the Bento[34] toolkit identifies subcellular domains of RNA localization, revealing molecular interactions involving RNA Binding Proteins (RBP)[35,36].

A natural next step after localization of individual RNA species is to consider spatial relationships between pairs of RNA molecules, because molecular interactions and functional relationships are mediated by physical proximity: direct interactions[37], interactions with common mediators[34,38], interactions with organelles[39], shared compartment localization[40,41] etc. Indeed, colocalization frequencies are used to characterize molecular interactions in super-resolution imaging studies[42], and directly measuring intermolecular distance is common in the protein literature[43]. However, most of these studies probe colocalization frequencies of a limited number of targets due to technical constraints. One solution to this limitation is to analyze the similarity of localization patterns instead of colocalization profiles. Battich et al.[44] inferred the spatial relationships of a large number of RNA pairs from regular in situ hybridization experiments by characterizing "localization features" of individual genes, showing genes with similar spatial localization to have similar functions. The recent development of transcriptome-scale single molecule-resolution spatial transcriptomics technologies allows direct measurement of distances between transcripts, affording us a more accurate view of spatial relationships than indirect inference from similar localization patterns of genes.

However, there have been no rigorous large-scale studies of RNA colocalization and subcellular spatial relationships, and necessary analytical tools do not exist. The rare tools that could be used[34] or adapted[43] for the purpose have major limitations: one is not accompanied by significance tests that can produce valid p-values[34] and another relies on assumptions of high molecular counts and homogeneous spatial distributions[43] that are typically violated in RNA data (Table 1). Moreover, existing colocalization metrics apply to individual cells and do not reveal spatial patterns that repeat across cells. Chen et al.[13] had considered such intercellular persistence of spatial relationships, but their approach is limited to a very coarse form of colocalization. Finally, the few existing statistical approaches are often not available as easy-to-use software, limiting their use in reconstructing colocalization maps.

Here, we introduce Intracellular Spatial Transcriptomics Analysis Toolkit (InSTAnT), a set of robust methods for extracting subcellular localization patterns of RNA. Its rigorous statistical foundation allows us to detect reproducible results with low false positive rates. InSTAnT identifies gene pairs whose transcripts tend to appear within distance d significantly more than by chance ("d-colocalized pairs") and reports the cellular domains where they appear. It employs formal statistical procedures to account for various sources of confounding such as overall transcript abundance, which is critical for highlighting spatial patterns of biological significance. We present a statistical test for persistent subcellular co-localization in many cells, which increases specificity of co-localization discovery in the face of cell-to-cell variabilities. The InSTAnT toolkit offer two analyses that are not offered in existing frameworks: reporting colocalized pairs that are specific to cell types, predetermined tissue regions or phenotypes, and testing if a pair's sub-cellular colocalization exhibits a tissue-level spatial pattern. The latter combines sub-cellular and multi-cellular analyses of spatial transcriptomics data. Both of these tools have the potential to reveal new insights about mechanisms and functions related to sub-cellular distribution of RNA species.

Demonstrative applications of the InSTAnT toolkit to a variety of data sets representing three different technologies (MERFISH, SeqFISH +, Xenium), five different sources (including in-house MERFISH data), and four biological contexts (cell lines or tissues) from human and mouse identifies hundreds of d-colocalized gene pairs with low estimated false positive rates and high reproducibility between replicates and data sources. The identified gene pairs exhibit biologically relevant higher order characteristics such as specificity to cell types and brain regions as well as non-random spatial variability in the tissue sample. We also find evidence of their possible relationship to RNA-RNA or RNA-protein interactions, pathway-level co-functionality, and localization to domains such as nuclear speckles. We also note a significant tendency of extracellular matrix-related genes to exhibit d-colocalization, suggesting widespread role for local translation. Our results suggest that InSTAnT can recover known biology and generate previously uncharacterized hypotheses about the functional role of RNA spatial localization. We believe that the statistical concept of d-colocalization introduced in this work will serve as a fundamental unit of subcellular spatial transcriptomics analyses, similar to how co-expression analysis has served as a core concept of transcriptomics analysis.

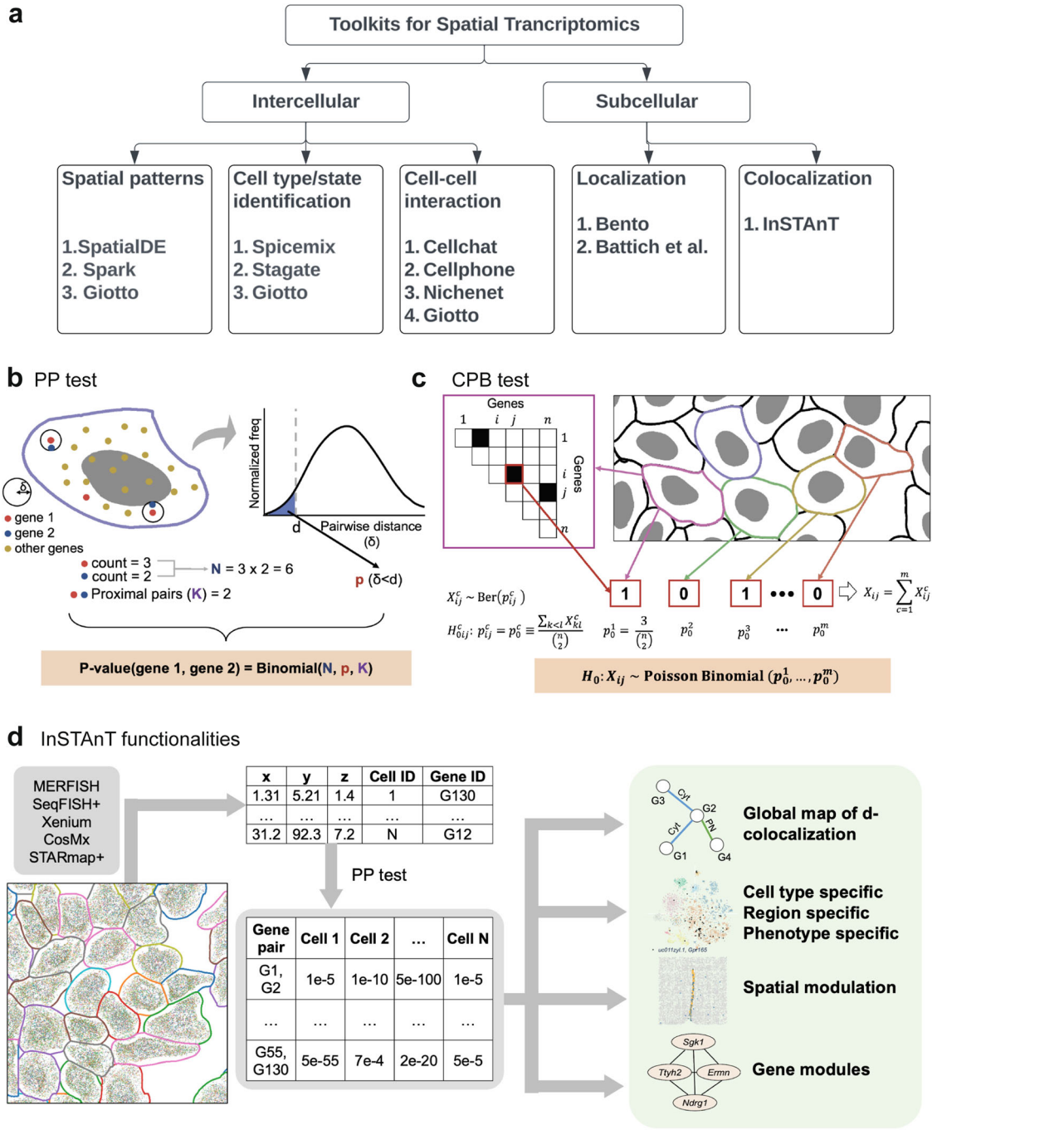
Results

Overview of InSTAnT

InSTAnT is a suite of statistical tools for spatial transcriptomics analysis at sub-cellular resolution. It can discover intracellular spatial patterns involving transcripts of multiple genes, leading to hypotheses regarding their functional relationships. At its heart is a statistical test to detect "proximal pairs" of genes by analyzing the spatial coordinates of transcripts within that cell, available from single-molecule resolution spatial transcriptomics technologies[13–17]. Specifically, the "Proximal Pairs" (PP) test determines if transcripts of a gene pair, in a given cell, are located within a distance threshold d significantly more often than expected by chance (Fig. 1b). Proximal pairs may represent various phenomena, e.g., direct or indirect interactions (detected at small d), or shared transcript localization in organelles or subcellular compartments (large d). The chance expectation may vary from cell to cell, depending on cell size and RNA density, so it is calculated empirically based on the distances between all pairs of transcripts in a cell regardless of gene identities. The test provides a p-value for each gene pair, representing its departure from this expectation (Methods). The scale parameter d is user-configurable, allowing the user to probe the spatial texture at different scales, though in practice it is ultimately limited by spatial resolution of the data. The PP test can be performed in either two- or three-dimensional mode (PP-3D), depending on whether or not data are available from multiple z-planes (Methods).

We define a *"d*-colocalized" gene pair to be a pair that is detected as proximal pair by the PP test in significantly many cells. This gives us increased confidence in a spatial relationship between the two genes. To detect d-colocalization, InSTAnT provides a test called "Conditional Poisson Binomial" (CPB) test that assigns a p-value to a gene pair based on the number of cells in which it is found to be a proximal pair. This test is based on a Poisson Binomial distribution and allows for different cells having varying numbers of proximal pairs due to varying transcript counts and spatial distributions (Fig. 1c, Methods). Initially, we noticed certain highly expressed genes to feature among d-colocalized

https://doi.org/10.1038/s41467-024-49457-w



x	y	z	Cell ID	Gene ID
1.31	5.21	1.4	1	G130
…	…	…	…	…
31.2	92.3	7.2	N	G12

Gene pair	Cell 1	Cell 2	…	Cell N
G1, G2	1e-5	1e-10	5e-100	1e-5
…	…	…	…	…
G55, G130	5e-55	7e-4	2e-20	5e-5

Fig. 1 | Schematic of InSTAnT. a Categories of existing analytical toolkits and methods for spatial transcriptomics (ST) datasets. Fewer methods perform subcellular analysis by focusing on gene localization patterns. In contrast, InSTAnT extracts colocalization patterns with statistical rigor. **b** Schematic of Proximal Pair (PP) test to detect if transcripts of a gene pair (gene 1, gene 2) tend to occur near each other (within distance d) in a single cell. A histogram of distances (δ) between transcript pairs (regardless of gene identity) in the cell is used to calculate the background probability of a transcript pair being near each other $p(\delta<d)$, and the number of such proximal pairs (K) of the pair (gene 1, gene 2) is assessed using a Binomial test. **c** Simplified schematic of Conditional Poisson Binomial (CPB) test. For a gene pair i,j, the random variable X_{ij}^c indicates if it is significant under the PP test and follows Bernoulli distribution with parameter p_0^c, estimated as the fraction of all pairs that are significant in that cell. The sum X_{ij} of X_{ij}^c over all cells follows a Poisson Binomial distribution. The CPB test further adjusts p_0^c to be dependent on the genes i,j (not illustrated here). **d** Schematic showing functionalities of the InSTAnT toolkit. The input is spatial transcriptomics data with spatial coordinates and gene identifier of each transcript. At the core of the toolkit is the PP test, which reports a p-value for each gene pair in each cell, and these results can then be utilized for various subsequent analysis, shown on the right. The CPB test can be applied on the collection of cells, resulting in the global d-colocalization map; significant pairs are also annotated with the cellular region where they tend to colocalize: Perinuclear (PN) region, Cell Periphery (CP), Cytosol (Cyt) or Nucleus (Nuc). The Differential colocalization routine can be employed to find cell type-specific, region-specific or phenotype-specific colocalization patterns. Other routines can be used to test if a gene pair's subcellular colocalization is spatially modulated at the tissue level or to identify modules of genes that colocalize with each other.

pairs far more frequently (Supplementary Fig. 1). The CPB test de-emphasizes pairs involving such genes by adjusting the null distribution of each pair to account for the global *d*-colocalization frequency of the involved genes (Methods). The InSTAnT suite is available as a python package with routines that return PP test results for every cell and CPB test results across all cells, for each gene pair. To assist with biological interpretation of the detected spatial relationships, it can annotate each *d*-colocalized gene pair with the cellular regions where its proximal transcripts tend to be found: nuclear, peri-nuclear, cytosolic and peri-membrane (Fig. 1d, Methods). In addition to reporting *pairs* of *d*-colocalized genes, InSTAnT allows us to identify *modules* of genes that are frequently colocalized across multiple cells (details below).

InSTAnT provides a routine called "differential colocalization" that determines if the cells exhibiting colocalization of a gene pair (significant PP test p-value) are statistically enriched for a user-provided cellular attribute. This functionality can be used for example to detect if a spatial relationship is specific to a cell type, a particular spatial region of intact tissue or even to cells from one experimental condition versus another. It thus provides a window into the complex biological factors that may influence, or are influenced by, RNA-RNA proximity. Another InSTAnT routine that aids analysis of colocalization in intact tissue, called "spatial modulation", tests if a gene pair's subcellular colocalization is a spatially variable phenomenon at the tissue level, analogous to current tools for detecting spatially variable genes but at the level of gene pair colocalization rather than individual gene expression.

Like other statistical phenomena such as differential expression of a gene or co-expression of a gene pair, the different kinds of spatial patterns recovered by InSTAnT, such as *d*-colocalization, differential colocalization and spatial modulation, may serve as a starting point for discovery of underlying biological relationships.

InSTAnT finds gene-gene relationships with high accuracy

We first applied InSTAnT to MERFISH data on human osteosarcoma cells (U2-OS), which profiles 130 genes in 3237 cells[45] (Methods), identifying 'proximal pairs' within each cell and '*d*-colocalized pairs' across all cells. An example of a highly significant *d-colocalized* pair is *THBS1-COL5A1*, which appears as a proximal pair (PP test p-value < 0.001, at $d = 4$ μm) in ~67% of the 3,147 cells where both genes were detected (Supplementary Fig. 3). We calculated false positive rates (FPRs) by applying InSTAnT to a random baseline dataset established by permuting the gene labels of all transcripts within each cell. We used FPR estimates to select p-value thresholds for PP and CPB tests. Overall, our tests suggested that hundreds of gene pairs exhibit the *d*-colocalization phenomenon, out of all ~8,500 pairs possible with 130 genes.

We compared the PP test with the only existing method for subcellular colocalization detection, the Colocalization Quotient used by Bento[34], finding the latter to exhibit significantly greater FPR estimated as above (>90% Fig. 2a) and lower reproducibility (<10%, Supplementary Fig. 5). As shown in Fig. 2a (blue), at $d = 4$ μm the PP test identifies sixty significant proximal pairs per cell with an estimated FPR below 10%. Smaller values of the scale parameter *d* yielded larger FPR values (pink and orange, Fig. 2a), suggesting lower sensitivity of the test and/or lesser frequency of proximal pairs in this regime. We arrived at similar estimates of FPR through an entirely different approach that exploits presence of "blank" gene probes in the data (Methods and Supplementary Fig. 2). We found significantly lower FPR for the CPB test (Fig. 2b). The only alternative approach for aggregating colocalization information across cells is the bin-based based approach of Chen et al.[13] (Methods). This approach calculates the correlation coefficient between transcript counts of a gene pair in four subcellular regions (bins), and aggregates correlations across cells. The low sample count (four) used in correlation calculation may result in less

Table 1 | Table compares InSTAnT with other methods that directly or indirectly examine colocalization

	Type	Off the shelf toolkit	Nuclear and cell membrane needed	Minimum number of molecules needed in a cell	Scalable?	Statistical significance for subcellular colocalization	Statistical significance of colocalization across cells
Battich et al.[a]	Spatial features of each molecule based (first order)	Not available for gene-gene relationship	Yes	No threshold	Yes	No	No
SODA[b]	Distance between molecules based (second order)	Yes, not usable, not executable	No	100	No	Yes (Null model based on complete spatial randomness)	No
Bento[c] v1 inspired (June 13, 2022)	Spatial features of each molecule based (first order)	Yes	Yes	No threshold	Yes	No	No
Bento[3] v2 (Colocation Quotient, Apr 13,2023)	Distance between molecules based (second order)	Yes	No	10	Yes	No	No
InSTAnT	Distance between molecules based (second order)	Yes	No	No threshold	Yes	Yes (Null model conditional on spatial distribution of transcript of all genes)	Yes

Battich et al.[a] and SODA[b] do not provide an off-the shelf toolkit readily available to be used for single molecule resolution spatial transcriptomic data. Bento[c] provides such a toolkit for analyzing subcellular transcriptomic data. Battich et al.[a] and Bento[c] v1 relied on spatial features of each RNA that needs nucleus and cell masks. SODA[b] and Bento[c] v2 extracts statistics based on distance between molecules. InSTAnT is the only tool generating valid p-values for subcellular colocalization across cells for single molecule resolution spatial transcriptomic data.

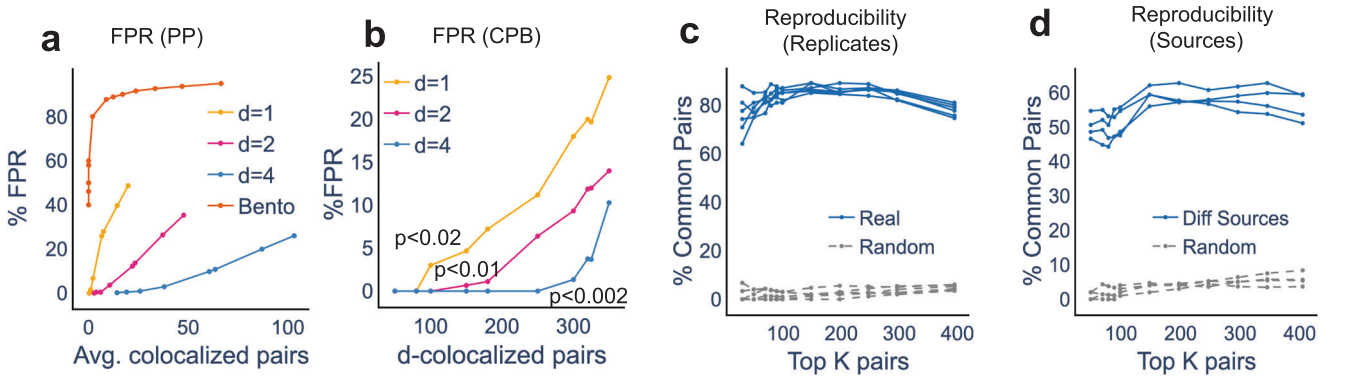


Fig. 2 | Assessment of InSTAnT on U2OS MERFISH data. a Estimates of false positive rates (FPR) on U2OS MERFISH data, at varying *p*-value thresholds for PP test (at three different values of distance threshold *d*) and at varying colocation quotient scores used by Bento. (Bento was set to use number of neighbor $K = 10$; this corresponds to $d \sim 5.5$ mu.) FPR is calculated by comparing the average number of significant pairs per cell on randomized data to the average number on real data. The estimated FPR (y) is plotted against the average number of significant pairs detected per cell (x). **b** Estimates of FPR of the CPB test plotted against number of detected gene pairs, at varying p-value thresholds (*p* value < 0.02 for d = 1, *p* value < 0.01 for d = 2, *p* value < 0.002 for d = 4). Results are shown for three different values of the distance threshold *d*. The number of significant pairs on randomized data is compared to the number (at the same p-value threshold) on real data to obtain an FPR estimate at that threshold. **c** Reproducibility of CPB test results across replicates of a dataset. For each pair of replicates (out of four), the K most significant pairs (by CPB test) in either replicate are compared, and the percentage of shared pairs (out of K) reported (blue). The exercise was repeated for randomized versions of the replicates to obtain random baselines (grey). **d** Reproducibility of CPB test results across different datasets. Each replicate of the Moffit et al. MERFISH data set was compared to our MERFISH data for U2OS to obtain percentages of common *d*-colocalized pairs (blue). Corresponding random baselines are shown in grey.

reliable colocalization quantification compared to the rigorous p-values of the CPB test. Furthermore, the coarse binning may result in missed colocalized pairs at finer spatial resolutions, e.g., ~4 micron, while InSTAnT robustly handles such resolution. These considerations underscore the importance of InSTAnT's rigorous statistical testing procedures for reliable detection of spatial patterns.

Two previous studies have examined gene pair relationships based on pre-defined localization features of individual genes[34,44]. This is fundamentally different from our approach, since gene pairs with similar localization are not necessarily the same as gene pairs whose transcripts are located close to each other. We compared *d*-colocalized pairs with those obtained using localization features of each gene (Methods) and found <2% overlap (Supplementary Fig. 6, Supplementary Fig. 7). *d*-colocalization is also distinct from the commonly analyzed tissue-level phenomenon of spatially varying genes. Supplementary Fig. 13 shows that some spatially variable genes are also significantly colocalized with other genes inside cells, but this is not true of the majority of genes, supporting a clear distinction between the two phenomena.

Our next assessment focused on the replicability of *d*-colocalization findings across four biological replicates of the U2OS data set[45]. We identified the most significant gene pairs (CPB test, $d = 4$ μm) in each replicate and observed that ~80% of the top 50–400 gene pairs are common between replicates (Fig. 2c), supporting the reproducibility of the reported pairs. The same assessment on randomized versions of the replicates yielded ~5% or less replicability expected by chance. We also tested the extent to which *d*-colocalization phenomena persist across independent MERFISH experiments. For this, we generated the spatial transcriptome map of U2OS cells using our home built MERFISH platform (Methods). We used InSTAnT to identify *d*-colocalized gene pairs from our dataset and compared the top K gene pairs (for varying K) between the Moffitt et al. and our data. As shown in Fig. 2d, about 50–60% of the identified gene pairs are shared between these two studies, with the chance expectation (established via randomized versions of the two datasets) being < 10%. As another reference point, a similar comparison of the top co-expressed gene pairs (detected using correlation of cellular transcript counts) also shows ~50–55% of commonality between the two studies (Supplementary Fig. 4b). Taken together, these reproducibility analyses suggest that the *d*-colocalized gene pairs reported by InSTAnT capture real biological phenomena or relationships.

InSTAnT constructs global *d*-colocalization maps

The CPB test identified 304 *d*-colocalized gene pairs at an FPR of < 2% ($p < 0.0001$), with $d = 4$ μm (~5% of the diameter of an average cell) (Supplementary Data 1). These gene pairs constitute the global *d*-colocalization map. InSTAnT provides annotations of the cellular regions where each gene pair tends to colocalize, revealing perinuclear and nuclear colocalization as most frequent (Fig. 3a–c, Supplementary Fig. 8). We also noted a few gene pairs to colocalize in the cytosolic or cell periphery regions (see Fig. 3d, e for examples).

A *d*-colocalization map is expected to capture different biology at different values of *d*. The maps created at $d = 1$ μm (Supplementary Data 2) and 4 μm are substantially different (Fig. 3f): while 94 pairs were common to the top 304 significant pairs of either map, 190 of the pairs in the $d = 4$ map had CPB test *p* value > 0.1 in the $d = 1$ map, and 178 gene pairs were similarly exclusive to the $d = 1$ map. Two examples of such scale-specific pairs are *SPTBN1-TLN1* (detected with $d >= 4$) and *LUZP1-SAMD12* (only with $d = 1$). (Also see Supplementary Fig. 9). These results illustrate scale-dependence of the colocalization phenomenon and suggests multiple types of underlying biological relationships, though some part of the exclusivity is likely to be due to varying sensitivity of the test at different d values.

Reconstructing gene-gene co-expression networks is a common analysis performed with non-spatial single cell RNA-seq data[46]. To test if the global *d*-colocalization map reflects such co-expression networks or if it reveals a different type of relationship, we derived a co-expression network from cell-level transcript counts in the same MERFISH data and found over 70% of the pairs in either "co-expressed" or "colocalized" set to be exclusive to that set (Supplementary Fig. 10, Supplementary Data 3). This shows that *d*-colocalization relationships are not revealed through conventional co-expression analysis, and probe a new type of information.

In addition to constructing a basic global map, InSTAnT can run the PP test in a "intra-nucleus" mode where the analysis, including null distribution estimation, is limited to subnuclear transcripts. The default (whole-cell) mode disregards the selective enrichment of a gene in certain subcellular regions, and nucleus-enriched genes may

https://doi.org/10.1038/s41467-024-49457-w

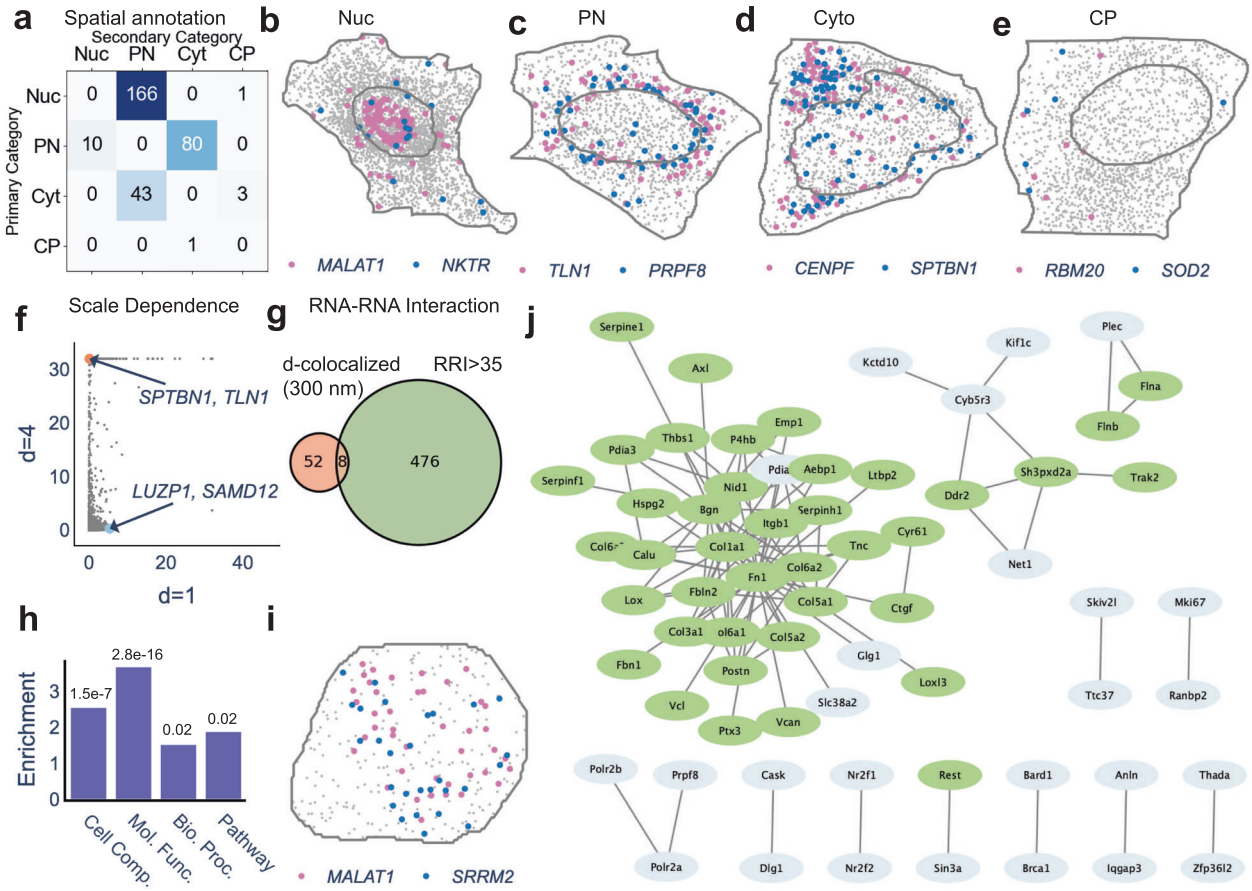


Fig. 3 | Characterization and validation of d-colocalization maps. a Regional annotation (nuclear, perinuclear, cytoplasm or peri-membrane) of all *d*-colocalized gene pairs detected in U2OS MERFISH *d*ata. Proximally located transcripts of a d-colocalized gene pair across all cells are recorded and aggregated over all cells to obtain the most and second-most frequent regional annotations. **b**–**e** Examples of *d*-colocalized gene pairs annotated as nuclear (**b**), perinuclear (**c**), cytosolic (**d**) and cell periphery (**e**), respectively. Shown is one of many cells in which the respective gene pair was significant by the PP test. **f** Negative log p-value from the CPB test for all gene pairs, at $d = 1$ micron and $d = 4$ micron. An example of a gene pair specific to each d is highlighted. **g** Overlap of the set of *d*-colocalized gene pairs ($d = 300$ nm) with gene pairs with high RNA-RNA interaction (RRI > 35) scores (Hypergeometric test p-value of 0.02, due to 8 gene pairs common to both sets). **h** Hypergeometric test shows enrichment of set of d-colocalized pairs with set of functionally related gene pairs. A gene pair is functionally related if both genes are annotated with same GO terms (Cellular Component, Molecular Function, Biological Process) or Kegg pathways. **i** Nucleus of a cell showing transcripts of *MALAT1* and *SRRM2*. The PP test *p*-value for this nucleus is 4.3e-19. **j** Cytoscape visualization of top 109 d-colocalized gene pairs (CPB p-value < 1e-10) detected at $d = 2$ on SeqFISH+ data on NIH/3T3 cell line. We noted a large module of genes related to extracellular matrix (green nodes), encoding proteins that are either components of the ECM or known for remodeling ECM or mediating ECM-cell interactions.

dominate detected co-localized pairs (Supplementary Fig. 11). The intra-nucleus mode effectively removes such bias. We observed many gene pairs with greater statistical significance in the intra-nucleus analysis compared to whole-cell analysis, despite smaller numbers of transcripts examined. Such pairs promise to reveal biologically meaningful spatial patterns that arise from colocalization to sub-nuclear structures, organelles and domains.

d-colocalization maps suggest functional relationships

One plausible mechanism for *d*-colocalization is direct or indirect interaction between two RNAs. To test this, we computed an RNA interaction score ("RRI score") for all gene pairs using RNAplex[47]. We recreated the *d*-colocalization map with *d* set to 300 nm (MERFISH resolution between pixels is 167 nm) to capture the greater proximity expected of interacting RNAs, and using a stricter CPB test p-value (1e-5) to control FPR (FPR = 0%). We found significant overlap between gene pairs having high RRI scores and *d*-colocalized pairs (Hypergeometric p-value 0.02, Fig. 3g, Methods, Supplementary Data 4). This analysis suggests that RNA-RNA interactions may underlie some of the relationships in a global *d*-colocalization map at a suitably small value of the scale parameter.

Furthermore, we found that the d-colocalized gene pairs were enriched with functionally related gene pairs as defined based on KEGG pathway or Gene Ontology (GO) annotations (Methods) (Fig. 3h). The highest enrichment happened with molecular function GO terms, where 461 functionally related pairs and 303 *d*-colocalized pairs had an overlap of 56 pairs (p-value 2.8E-16), all of which were annotated with the term "protein binding". These results suggest that *d*-colocalization of a gene pair may have biological consequences such as colocalization of their protein products or protein binding to form a ribonucleoprotein complex.

Intriguingly, the map indicates that some colocalized RNA pairs proxy for protein-RNA interactions. The most prominent pair in the intra-nucleus analysis at d=2 μm is *MALAT1-SRRM2*, with a CPB test p-value of 1.05e-18 (see Fig. 3i, Supplementary Fig. 11). It is detected as a proximal pair in 6.2% of the nuclei, the most for any pair involving either *SRRM2* or *MALAT1*. Notably, SRRM2 protein is a key marker of nuclear speckles (NS), organizing NS formation via liquid condensation[48], and lncRNA *MALAT1* is localized to NS[49], suggesting that the detected intra-nuclear *d*-colocalization of these two RNAs may be related to their colocalization in NS. This is an intriguing possibility, since NS localization of SRRM2 protein does not imply or necessitate a

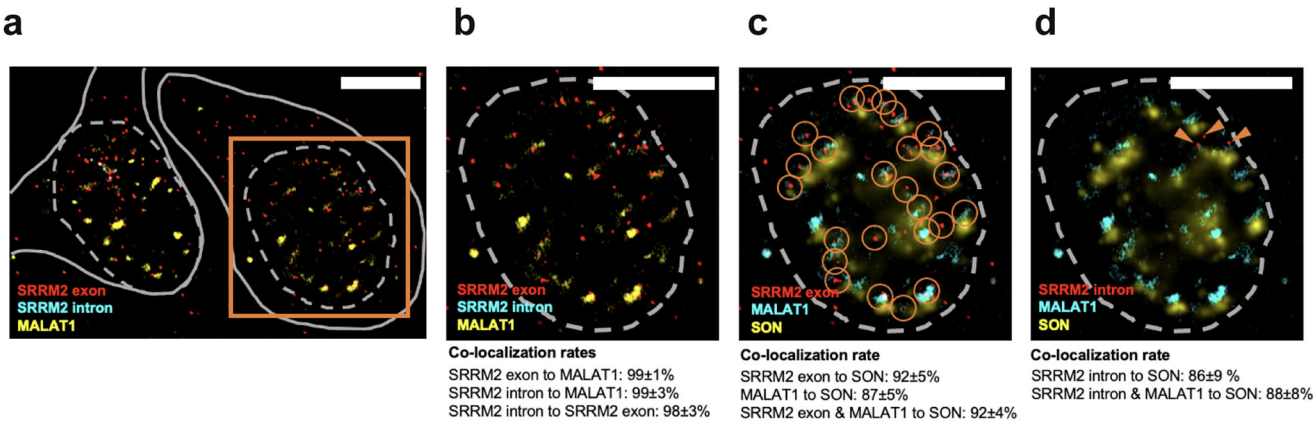


Fig. 4 | Colocalization of SRRM2 and MALAT1 in Nuclear Speckles. a *SRRM2* exon (red), *SRRM2* intron (cyan), and *MALAT1* (yellow). RNAs labeled with smFISH probes in fixed U-2 OS cells. Dashed gray lines indicate the nuclear boundaries, and solid gray lines indicate cytosolic boundaries. **b** Selected nuclear region shown in the orange box in **a**, showing high co-localization rate of *SRRM2* exon mRNAs with *MALAT1* lncRNAs in the nucleus. As expected, the *SRRM2* intron puncta co-localize with *SRRM2* exon puncta. **c** *SRRM2* exon mRNA (red), *MALAT1* lncRNA (cyan), and SON protein (yellow) labeled in the same nucleus as **b**. Orange circles indicate co-localization of *SRRM2* exon puncta and *MALAT1* puncta, most of the SRRM2 exons co-localize with MALAT1. SON protein was selected to label nuclear speckles. Many co-localized RNA pairs are nearby to SON protein. **d** Similar as (**c**), plotting *SRRM2* intron (red) with *MALAT1* lncRNA and SON protein, orange arrows indicate *SRRM2* intron puncta that co-localize with *MALAT1* puncta. Similar to *SRRM2* exon puncta, *SRRM2* intron puncta tend to be near SON protein. The experiment was performed once and co-localization rate was calculated using 13 cells. Scale bars for (**a**–**d**) is 10 μm.

similar localization of its mRNA. To see whether lncRNA *MALAT1* and mRNA *SRRM2* colocalize near NS, we co-stained *MALAT1*, *SRRM2* RNA, and SON in U2-OS cells using single-molecule FISH and immunostaining (Fig. 4). Consistent with the InSTAnT result, most *SRRM2* RNAs are d-colocalized with *MALAT1* in *SRRM2* positive cells (99±1%, N=13 cells). The overlaid SON signals show that most d-colocalized *MALAT1-SRRM2* pairs are within 1 μm distance from NS (92±4% for *SRRM2* exon and 88 ± 8% for *SRRM2* intron). It is well known that SRRM2 protein signals overlaps with SON signals[48]; thus, our result shows the d-colocalization of *SRRM2* mRNA and pre-mRNA with SRRM2 proteins in nucleus. Further, these results suggest that d-colocalization maps can be used to study biomolecular condensates such as NS.

To demonstrate InSTAnT's applicability to data from diverse technologies, we used it to construct a global d-colocalization map (d = 2 μm) from SeqFISH+ data on a mouse fibroblast cell line[14] (Methods). The dataset consists of 3726 genes and 179 cells after preprocessing. Among the most significant pairs in this map, we noted a remarkable enrichment for functions related to extracellular matrix (ECM) (Fig. 3j, Supplementary Data 5, Supplementary note A) and cell adhesion, consistent with reports of localized translation of ECM/adhesion proteins[50–52] in perinuclear or peri membrane regions as well as localization and non-coding functions of mRNAs at focal adhesions[53]. Our findings suggest such mRNA localization may play a widespread role in ECM-cell interactions.

InSTAnT analysis of brain data reveals cell type- and behavior-specific colocalization

We next used InSTAnT to analyze MERFISH data[54] on 5149 cells from the hypothalamic preoptic region in mouse. The data feature seven z-planes and were thus analyzed with the PP-3D test of proximal pairs (d set to 2 μm). This brain dataset includes nine different cell types (Fig. 5d), so we used the Differential Colocalization module of InSTAnT for insights into cell type differences in colocalization. Given a binary (yes/no) annotation of each cell – in this case whether it belongs to a cell type or not – this module uses a sequence of statistical tests (Methods, Fig. 5b) to find gene pairs specific to a cell type: those that appear as proximal pairs (PP test) more frequently in the one cell type than others. These are further divided into two classes – those where cell type specificity may arise simply because one of the genes in the pair is expressed specifically in that cell type (Category 1) and those whose association goes beyond what would be expected from the cell type-specificity of either gene's expression (Category 2) (Methods). Between these two categories we found more than fifty gene pairs specific to one of the six most abundant cell types (Fig. 5a, Supplementary Data 6). Many of these top pairs have plausible mechanisms for being proximal (examples below), such as localized translation of proteins for relevant molecular pathways, reiterating the potential of d-colocalization to capture underlying phenomena.

The top gene pairs in Category 1 specific to astrocytes involve the genes *Aqp4* (Aquaporin 4), *Ttyh2* (Tweety Family Member 2), *Cxcl14* (CXC motif chemokine ligand 14) and *Mlc1* (Modulator of VRAC current 1) (p-value of cell type association < 1.1E-20). Figure 5c illustrates for the pair *Cxcl14-Mlc1*, which is a proximal pair in cells of most types, but with a higher frequency in astrocytes, leading to the statistically detected specificity. *Cxcl14* transcripts are known to be enriched in and possibly locally translated in peripheral astrocyte processes (PAPs)[55], and MLC1 protein is localized in PAPs[56]. We speculate that *Mlc1* transcripts are also subject to local translation in PAPs, leading to the observed colocalization of *Cxcl14* and *Mlc1* in astrocytes. Additionally, MLC1 protein forms a complex with AQP4 in cultured astrocytes[57] and localizes to the cell membrane[55,58] providing a functional implication of *Mlc1-Aqp4* RNA differential colocalization.

The top pair in Category 2 (Fig. 5e, Supplementary Data 7) consists of transmembrane proteins *Gpr165* (G protein-coupled receptor 165) and *uc011zyl.1* (adhesion molecule with Ig-like domain 2) and is significantly associated with inhibitory neurons, while *Gpr165* and *Omp* (Olfactory marker protein[59]) form a colocalized pair specific to excitatory neurons (Supplementary Fig. 12). This example illustrates that different colocalized pairs involving a common gene (*Gpr165*) can statistically mark different cell types.

We also identified differentially colocalized gene pairs that mark cellular function (Supplementary Data 6), e.g., *Esr1* (estrogen receptor 1) and *Npy2r* (Neuropeptide Y receptor Y2) are colocalized specifically in inhibitory neurons compared to excitatory neurons (p-value 1.28E-8, see Fig. 5f, Supplementary Data 7). Prior work shows that the expression of these two genes underlies a social behavioral switch in virgin mice via activation of a specific subtype of neurons[60], suggesting a functional implication of *Esr1-Npy2r* colocalization.

In search of colocalization patterns related to phenotypic variation, we used the Differential Colocalization routine on MERFISH data[54]

https://doi.org/10.1038/s41467-024-49457-w

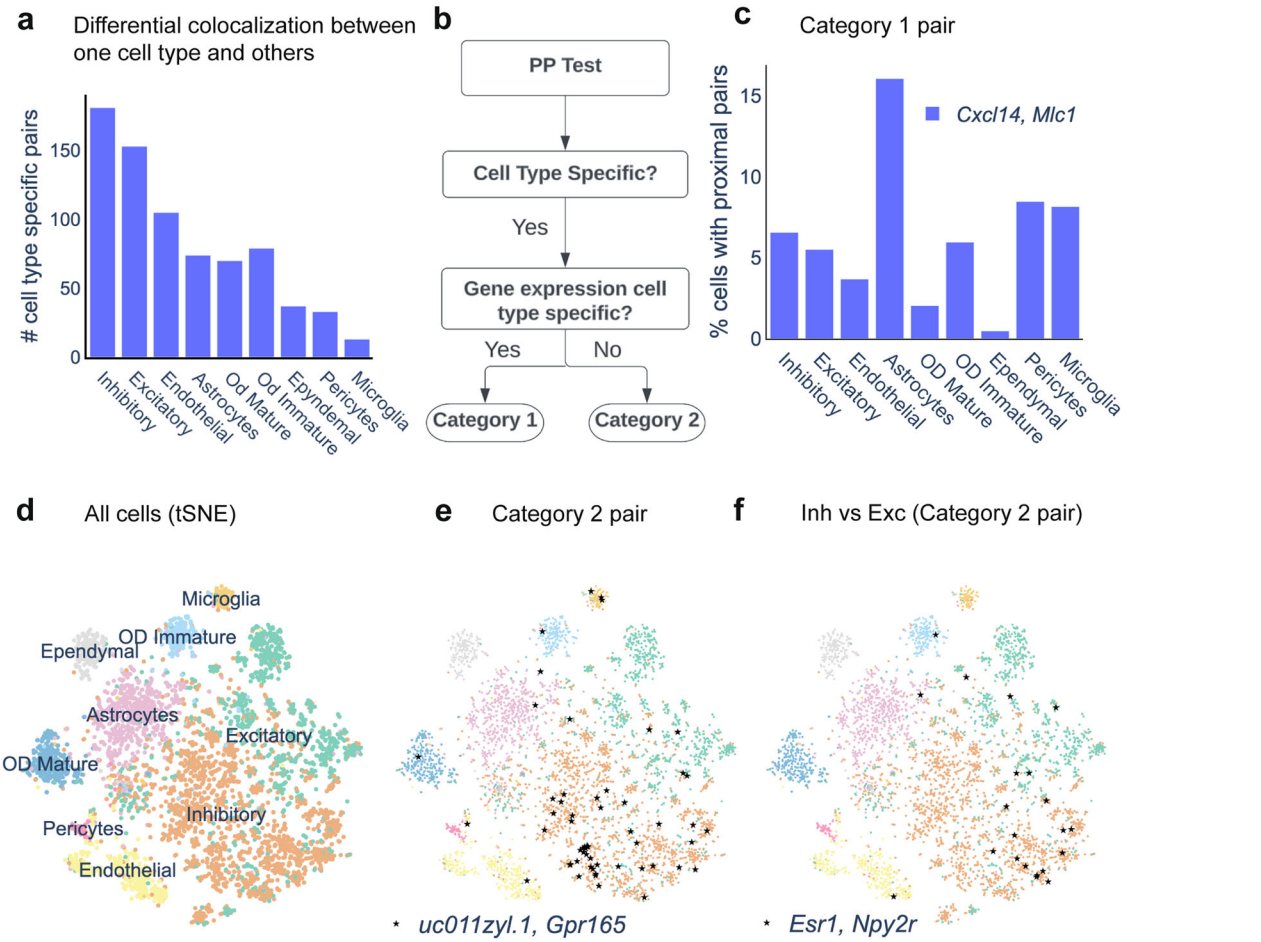


Fig. 5 | Cell type-specificity of *d*-colocalized pairs in mouse hypothalamus preoptic region. a Bar plot showing number of cell type-specific pairs for each cell type using Differential Colocalization routine. ("Od" stands for oligodendrocytes.) **b** Flow chart showing how a differentially colocalized pair is classified into one of the two categories depending on whether either gene is a marker of that cell type. **c** Example of a category 1 pair, found to be a proximal pair in many cells of different types but significantly more frequently in astrocytes. Shown is the percentage of cells of each type where the gene pair is significant in the PP test. The gene pair is of category 1 because both genes are marker genes. **d** t-SNE plot of all cells annotated with cell type assignments obtained from Moffit et al. The gene count for each cell is aggregated by summing their transcript count across seven z-slices. **e** Example of a category 2 pair, specific to inhibitory neurons. Each black star is a cell where the pair was significant under PP test. **f** Example of a category 2 gene pair specific to inhibitory neurons compared to excitatory neurons. (Cell type- specificity was defined based on a two-way comparison here, in contrast to the one-versus-all comparison used for examples in **a**, **c**, **e**.).

from brains of mice with behavioral differences and identified gene pairs that colocalize specifically to male mice exhibiting aggressive behavior, compared to naïve mice (Methods, Supplementary note B). The top reported pair in category 2 is *Cbln2-Pak3* (Supplementary Data 8). Cbln2 (cerebellin-2) is functionally associated with aggressive behavior[61] and PAK3 (protein-activated kinase 3) has been linked to aggressive behavior in humans[62]. Both proteins are localized to dendrites and involved in dendritic spine formation[63,64], and contributes to synapse formation or transmission[65,66]. Moreover, *Pak3* mRNA has also been found to be enriched in dendrites compared to somata[67], suggesting local synthesis of the protein for its dendritic functions. These observations suggest that the recorded function of the corresponding proteins in aggression may manifest at the sub-cellular level through their respective tendencies to localize in and be synthesized in dendrites.

InSTAnT reveals tissue-level spatial variations of colocalization patterns

Brain tissue is well-known to be spatially heterogeneous, so we applied InSTAnT's Differential Colocalization module to explore if colocalization is specific to certain brain regions. We analyzed Xenium data[68] on the mouse brain (d = 750 nm), focusing on cells from three adjacent regions – the dentate gyrus (DG) and areas CA3 and CA1 – of the hippocampus (Fig. 6a). We identified six category 2 gene pairs whose colocalization is specific to one of these three regions versus others (Methods, Supplementary Data 9). These included the pair *Gad1-Pvalb* that colocalizes specifically to CA3 and CA1 relative to the Dentate gyrus (Fig. 6b,c). Gad1 encodes glutamate decarboxylase 1 and is a marker of inhibitory GABAergic neurons[69] and interneurons[70]. Pvalb (Parvalbumin) encodes a calcium-binding protein that is often associated with a subclass of inhibitory interneurons (PVALB+ interneurons)[71]. Deficit of the Gad1 product in hippocampal PVALB+ interneurons has functional consequences and disease associations[72], suggesting an underlying molecular relationship for the observed region-specific colocalization.

Complementary to the differential colocalization module that reports colocalization specific to pre-annotated regions, InSTAnT provides a "Spatial Modulation" routine to identify colocalized pairs with tissue-level spatial variation in an unbiased manner. This functionality is similar to discovery of spatially varying genes[3,4], but specialized for gene *colocalization* instead of individual gene expression. It is based on a probabilistic model for calculating data likelihood under the hypothesis of spatial modulated colocalization, for a gene pair (Fig. 6d). Pairs with log likelihood ratio above a threshold

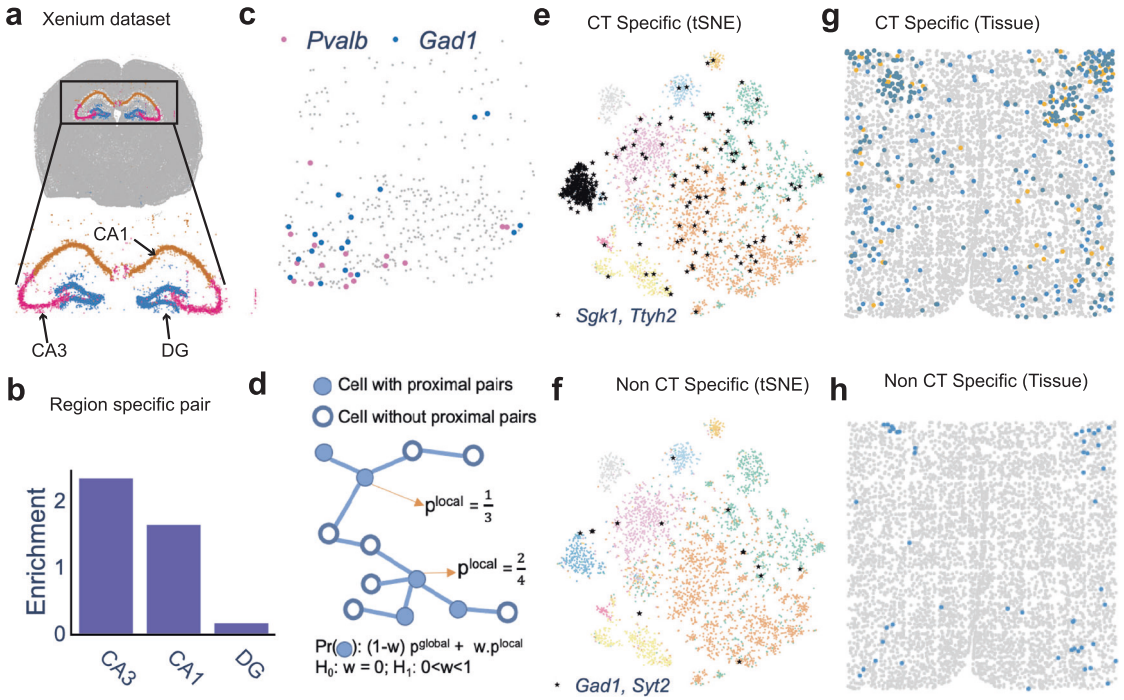


Fig. 6 | InSTAnT detects *d*-colocalization patterns with tissue-level spatial variation in mouse hypothalamus preoptic region. a Xenium data from mouse brain, with cells in analyzed regions - CA1 (orange), CA3(pink) and Dentate Gyrus (blue) in hippocampus – shown in color. **b** Enrichment of a category 2 gene pair (*Pvalb, Gad1*) in CA3 and CA1 cells. Enrichment is obtained as ratio of fraction of cells with proximal pairs in one region vs other two regions. **c** A sample cell showing the colocalization of the pair *Pvalb, Gad1* (z axis not shown). **d** Probabilistic graphical model to detect spatially modulated gene pair. In a graph where nodes represent cells and edges represent spatial proximity, each cell is first flagged based on whether the gene pair is significant by PP test in that cell. The likelihood function is a product over all cells of a weighted sum of p^{local}, the local density of flagged cells in cell's neighborhood, and p^{global}, a free parameter. The weight w is also a free parameter. A likelihood ratio score is computed to compare this model to a null model where the local (spatial) information is not used. **e** t-SNE plot of a spatially modulated *d*-colocalized gene pair (*Sgk1, Ttyh2*) showing that it is a proximal pair (black stars) significantly more often in Mature Oligodendrocytes (OD) though it is detected in other cell types as well. (See Fig. 5d for cell type annotations.) **g** Cells in spatial coordinates, shown in blue if the gene pair of (**e**) – *Sgk1, Ttyh2* – is a proximal pair, in orange if the cell is Mature OD but Sgk*1, Ttyh2* is not a proximal pair, and in grey otherwise. (**f**, **h**) t-SNE plot (**f**) and spatial plot (**h**) of a gene pair (*Gad1, Syt2*) that is spatially modulated but not specific to any cell type.

(obtained using randomization of data) are designated as spatially modulated.

We used the Spatial Modulation routine to study how *d*-colocalization varies across the mouse hypothalamic region, using MERFISH data[54], finding 45 spatially modulated pairs (Supplementary Data 10). Thirty eight of these pairs exhibited colocalization in a cell type-specific manner (p-value 5E-6, Bonferroni corrected p value < 0.05). For instance, the gene pair *Sgk1-Ttyh2* – the strongest spatially modulated pair (LLR 228, Fig. 6g) – colocalizes far more frequently in mature oligodendrocytes than others (Hypergeometric test p-value 2.8e-178, Fig. 6e). Prior work suggests both of these genes to function in oligodendrocyte response to stress[73–75] and this co-functional relationship may underlie their oligodendrocyte-specific colocalization. We found one spatially modulated gene pair *Gad1-Syt2* (LLR 26, Fig. 6f, h) whose colocalization is not specific to any cell type (Hypergeometric test p-value > 0.05 for every cell type). GAD1 is responsible for producing GABA[69], while SYT2 facilitates the release of neurotransmitters (including GABA) into the synaptic cleft and has been seen colocalized with the vesicular GABA transporter VGAT[76]. Their spatially modulated colocalization may be pointing to their cooperation in neurotransmission processes and synchronized release of neurotransmitters[77]. In summary, the above examples of brain region-specific and spatially modulated *d*-colocalization provide a rich pool of potential functional relationships for future exploration.

InSTAnT reveals modules of genes colocalizing with each other

We noted above multiple instances of "gene modules", i.e., sets of genes exhibiting pair-wise *d*-colocalization. Drawing inspiration from these observations and from the popular concept of co-expression modules[78], we implemented routines that systematically retrieve *d*-colocalization gene modules.

InSTAnT provides two complementary "Module Discovery" routines. The first routine, called Global Colocalization Clustering (GCC), identifies modules by representing the CPB test results as a matrix of gene-gene *d*-colocalization strengths and clustering rows and columns of this matrix (Methods). Figure 7a shows the results of such clustering for U2OS data, revealing two modules (top left) whose compositions are shown in Fig. 7b. Module M1 (Fig. 7d,e) consists of 14 genes, with 84 of 91 pairs being significantly *d*-colocalized, almost always with perinuclear region annotation. Gene Ontology enrichment analysis of the module revealed shared annotations (p-value < 0.05, Fig. 7c) related to cytoskeleton and ribonucleoprotein complexes. mRNA-cytoskeletal associations have been long known to play a key role in mRNA transport and targeting to specific subcellular locations, partly mediated by RBPs and ribonucleoprotein complexes[79,80].

A module reported by GCC comprises gene pairs whose *d*-colocalization is supported by many cells, but these supporting cells differ for different gene pairs and very few cells may have the entire module colocalized. Motivated by this, InSTAnT includes a second module discovery routine, called "Frequent Subgraph Mining" (FSM)[81], that seeks a network of genes "colocalized" in many cells. (Colocalization of a network in a cell means that every edge in that network is a proximal gene pair in that cell (Fig. 7f).) FSM can be used to find networks with a pre-specified minimum size (numbers of nodes and edges) that are supported by a large number of cells (Methods). For illustration, we used FSM on the brain MERFISH data[54] to search for fully connected

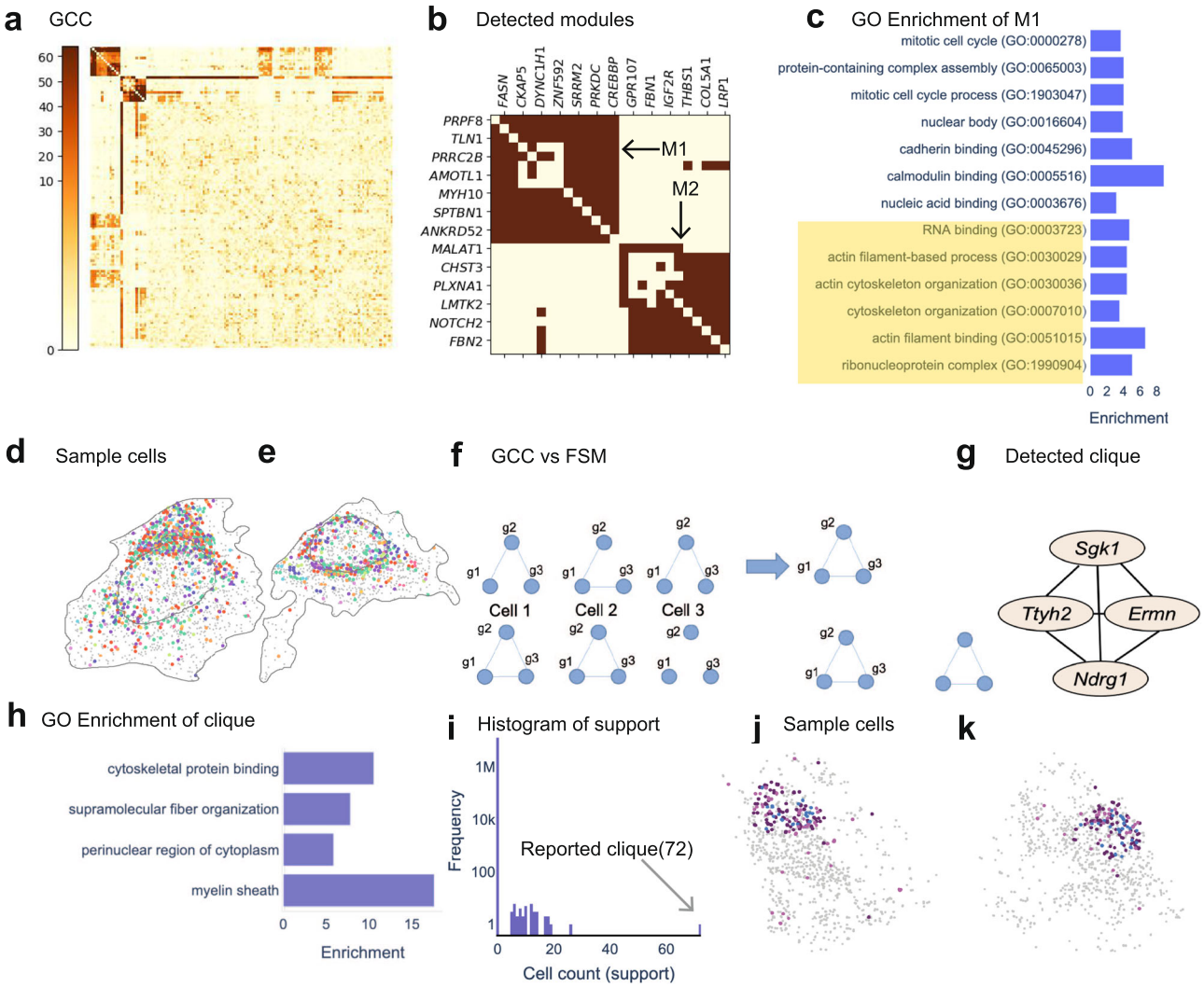


Fig. 7 | Gene module discovery. a Global Colocalization Clustering (GCC): Global *d*-colocalization map for U2OS data, represented as a matrix of -log(p-value) of CPB test for gene pairs, is subjected to hierarchical clustering to reveal two gene modules. **b** Closer view of the two modules (M1, M2) discovered by GCC, shown after thresholding p-values at 1e-4 (FPR < 2%). **c** Gene Ontology (GO) terms enriched in gene module M1, shown with the fold enrichment over random expectation. (Criterion for selection: Fisher exact *p* value < 0.03) **d**, **e** Two cells illustrating spatial distribution of transcripts of M1 genes (colored dots) along with all other transcripts (grey). Each color corresponds to a gene. **f** Schematic illustration of difference between Global Colocalization Clustering (GCC) and Frequent Subgraph Mining (FSM). In each row, the three graphs on the left show proximal pair relationships (edges) involving genes g1, g2, g3, in three different cells. In either case, GCC reports the 3-gene module as the global map includes each of the three gene pairs. FSM, on the other hand, finds the 3-gene clique to occur frequently in the bottom scenario but not in the top scenario. **g** A 4-gene module detected using FSM on brain data. **h** Gene ontology terms enriched in the 4-gene module of **g**. (Criterion of selection: Fisher exact *p* value < 0.03). **i** Histogram of "support" of all possible 4-gene cliques. Support refers to the number of cells where all pairwise relationships in the 4-gene set are significant by the PP test. The clique of **g** has a support of 72, far greater than all other cliques. **j, k** Example of two cells supporting the 4-gene module of **g**. Each color represents a transcript of one of the four genes, grey represents all other transcripts.

networks ("cliques") with at least four genes and found a single module – *Sgk1, Ttyh2, Ndrg1* and *Ermn* (Fig. 7g) – that is colocalized in 72 cells, far greater than the support of the next most frequent four-gene clique (26 cells) (Fig. 7i–k). We also found that the six gene pairs comprising this module are differentially colocalized in mature oligodendrocytes and the module is significantly associated (p-value 8.3e-3) with myelin sheath[82–85] (Fig. 7h). We speculate that their co-localization in specific partitions inside cells reflects coordinated transport and translation in mature oligodendrocytes, which are known for their role in myelination[86].

Discussions

In this work, we present the InSTAnT toolkit to screen for subcellular colocalization patterns of RNA pairs and modules in an unbiased manner, through rigorous statistical analysis of single-molecule resolution spatial transcriptomics data. We define these patterns as new statistical phenomena that may point to biological relationships such as RNA-RNA interactions, formation of condensates, co-transportation, and shared subcellular localization. InSTAnT is a suite of statistical tests, at the heart of which lie the PP test that finds colocalized gene pairs in each cell individually. Its findings are then analyzed further by multiple routines to derive more global patterns, such as (1) *d-colocalization* (CPB test) that represents patterns seen persistently across cells, (2) differential colocalization between two pre-annotated groups of cells, (3) tissue-level spatial modulation of subcellular colocalization patterns, and (4) colocalized gene modules (GCC and FSM). We emphasize that the co-localization analyses of InsTAnT probes distinct phenomena compared to the localization profiles and co-expression profiles, as evidenced by small overlap between the identified pairs from each analysis.

We employed InSTAnT to detect hundreds of colocalized gene pairs on human U2OS cell line, mouse fibroblast cell line (NIH/3T3) and regions of mouse brain data, and their further examination suggested a variety of underlying molecular relationships: RNA-RNA interactions (Fig. 3g), protein-protein interaction or shared pathway membership (Fig. 3h), ECM-cell interactions (Fig. 3j) and RNA-protein interactions (Fig. 4). We have observed different types of biological relationships when using different values of "d". For instance, small d (< 1 um) was used to find an enrichment of RNA-RNA physical interactions (predicted using RNAPlex) among colocalized pairs, a medium "d" (=2 um) was used for the analysis that led us to identify MALAT1-SRRM2 as colocalized in nuclear speckles, while d=4um in u2-os data identified RNAs that tend to be around perinuclear space or ER (which are larger features). Our brain data analysis shows that some RNA colocalized pairs may be cell type- or brain region-specific, show spatial modulation, and share functional annotation with other colocalizing pairs. We noted examples where InSTAnT-derived colocalization patterns provide specific, testable biological insights that are not available from other methods. For instance, multiple data sets revealed genes related to extracellular matrix and cell adhesion to colocalize at transcript level, leading to a hypothesis of widespread local translation of corresponding proteins.

We found that the rigorous statistical foundation of InSTAnT has a dramatic impact on the accuracy (low FPR and high reproducibility) of detecting subcellular patterns. The detected patterns can also include rare events. For instance, in spatial modulation analysis, some detected gene pairs were specific to ependymal cells, which is a rare cell type (~ 1% of cells) in our dataset. InSTAnT may be used to represent a cell as a graph where nodes represent genes and edges represent proximal pairs. Such a graph, along with the transcript count vector commonly used to represent an individual cell, may prove powerful in single cell analytics, allowing us to discover cell types through a more nuanced clustering of cells than possible using count vectors alone. Overall, InSTAnT may be used to generate a robust compendium of testable hypotheses that enhance our understanding of molecular functions, opening new horizons for experimentation and discovery.

Methods

Source data

Source data are provided with this paper.

Statistics and reproducibility

MERFISH data were generated on one sample of U2OS cell line (described above), and the data included thousands of cells as samples for INSTANT analysis. Experiment for SRRM2-MALAT1 interaction (Fig. 4a) was performed once and the colocalization rate was aggregated over cells.

InSTAnT user guide

InSTAnT tools have tunable parameters that can be selected based on the user's requirement. We selected the scale parameter d based on the average cell's diameter and CPB False Positive Rate (2%) estimates. The user can also obtain region annotations of a gene pair's colocalization if the data include masks for cell and nucleus boundaries. Similarly, they may run cell type/region specificity analysis if the data include cell type or region information. We advise caution when using InSTAnT with small distance thresholds, such as 1 μm or less, as the false positive rates in this regime can be high. This is due to the fact that colocalization with small distance is relatively rare and the estimate of null probability of a pair of transcripts being proximal, a key aspect of the PP test, is error-prone in such cases. We believe that higher number of transcripts and improved optical resolution[17],e.g., through expansion microscopy, may alleviate this problem.

Comparison with Bento

We used Bento[34] from the official GitHub repository (version 90f4ab4). We used the function for colocation- "coloc_quotient()" at https://github.com/ckmah/bento-tools/blob/master/bento/tools/_colocation.py. The original function uses an AnnData object to load data. However, we wrote scripts to load input data in csv format. The rest of the function was used as it is. We used K=10 in our experiments (d-5.5 mu). We also explored setting distance d=4 mu but noted worse FPR in this setting.

Comparison of Proximal pairs with gene pairs having similar localization features

We sought to compare the results of the PP test with the approach of Battich et al.[44] that is based on localization features. However, since their code is not easy to use, we used a function from Bento[34], that uses a very similar approach based on pre-defined localization features of transcripts. The features used were- *nucleus_inner_proximity, nucleus_outer_proximity, l_half_radius, l_max, l_max_gradient, l_min_gradient, l_monotony, cell_inner_asymmetry, nucleus_inner_asymmetry, nucleus_outer_asymmetry, point_dispersion*, and *nucleus_dispersion*. The authors of Bento[34] had made these features available for a SeqFISH+ data set on a mouse fibroblast cell line, hence we performed our comparison on this data set. Note that this function and the method of Battich et al.[44] do not aim to identify gene pairs whose transcripts tend to be near each other; rather, they separately characterize the location of (transcripts of) each gene and report gene pairs with similar localizations in a cell. We found that this approach reports very different gene pairs compared to Instant (PP test), with <2% or fewer of the top 500 pairs in any one cell being common. For illustration purposes, we present one example of a gene pair deemed significantly d-colocalized by Instant PP test but not identified by the localization feature-based method, and one converse example (Supplementary Fig. 7).

Implementation of Chen et al. (Bin based method)

We followed the method described in Chen et al.[13]. First, we divided each cell into 2 by 2 regions (bins). For each gene, fraction of occurrence in each bin was calculated. Enrichment of a gene in a bin was calculated as ratio of the observed fraction in a given region to average fraction of all genes in the same region. Next, Pearson correlation of region-to-region variation in enrichment of a gene pair was calculated for each cell. For each gene pair, we removed the cells that had one of the genes as constant across four bins (and therefore correlation is not defined). Finally, for each gene pair, we take median of correlation across all cells as a global measure of their colocalization. Note that (a) the spatial resolution of this approach is quite low (about a quarter of a cells area/volume), and (b) correlation coefficients are calculated from four samples at a time, leading to unreliable estimates, which are then averaged.

False positive rate (FPR)

We generate random baseline dataset established by permuting the gene labels of all transcripts within each cell, which recapitulates the spatial patterns of the original data but not the gene-gene relationships. The gene pairs obtained with InSTAnT on real data comprise true positives (TP) and false positives (FP). The gene pairs found under the random data are assumed to be false positives (FP). FPR is obtained by comparing the number of detected pairs obtained on randomized data with number of detected pairs on real data. Since thousands of cells are independently randomized (gene labels of transcripts in an individual cell are shuffled), this procedure makes use of an extensive level of randomization. Ten of the 140 genes probed in the U2OS MERFISH data set were "blanks", meaning that they do not represent any particular RNA or other molecule. Any gene pair involving such blank "genes", if found to d-colocalize, is clearly a false positive. This provided us another opportunity to assess the false positive errors in our global co-localization map. We recorded the fraction of such false

positives among predicted pairs at varying levels of significance (Supplement Fig. 3).

Hyperparameter selection

U2OS MERFISH dataset: d was chosen to be 4 microns, which corresponded to ~5% of average diameter of a cell. The p-value threshold was chosen to be 0.001 for PP test and 0.0001 for CPB test that resulted in CPB FPR < 2%.

Hypothalamus brain MERFISH dataset: d was chosen to be 2 microns, which corresponded to ~5% of average diameter of a cell. The p-value threshold was chosen to be 0.001 for PP test. For differential colocalization, we use p-value threshold of 5e-6 (Bonferroni corrected hypergeometric p-value 0.05) for unconditional p-value obtained from Hypergeometric test. Same parameters were used for behavior analysis.

NIH/3T3 SeqFISH+ dataset: d was chosen to be 2 microns, which corresponds to ~0.5% of average diameter of a cell. The p-value threshold was chosen to be 0.01 for PP test. CPB estimated FPR is ~0% for top 109 pairs (Fig. 3j).

Hippocampus brain Xenium dataset: d was chosen to be 0.75 micron that corresponds to ~3.5% of average diameter. The p-value threshold was chosen to be 0.01 for PP test. For differential colocalization, we use p-value threshold of 5e-6 (Bonferroni corrected hypergeometric p-value 0.05) for unconditional p-value obtained from Hypergeometric test.

Proximal pair (PP) test

PP test reports proximal pairs of genes in a particular cell. A gene pair g_i, g_j is a proximal pair in a cell if their transcripts are proximally located (separated by distance d or less) significantly more often than expected by chance. The null probability p is estimated from the distances between all pairs of transcripts (regardless of gene identities) in the cell, by calculating the fraction of transcript pairs that are proximally located. Let t_i and t_j denote the transcript counts of genes g_i, g_j respectively in the cell, let $T = t_i t_j$ and let K be the number of proximally located transcript pairs of these genes. The PP test performs a Binomial test providing a p-value (one-sided) for g_i, g_j as

$$\mathrm{p-value}(g_i, g_j) = Binomial(T, p, K) \tag{1}$$

PP-3D test

PP-3D is an extension of PP test to handle three-dimensional data in the form of 2D (x-y) locations of transcripts in each of multiple z-planes. We assume that data from different planes are independent and identically distributed. The new distribution is the sum of independent Binomial distributions (with the same parameter), which is also a Binomial distribution. The null probability of two transcripts being proximal is estimated as a weighted combination of estimated null probability for each of the z-planes,

$$p \equiv \frac{\sum_z l_z p_z}{\sum_z l_z} \tag{2}$$

where, p_z denotes the null probability for z-th plane, l_z denotes the total number of transcripts in z-th slice. T and K are also aggregated across z-planes:

$$T = \sum_z T_z$$

$$K = \sum_z K_z$$

where K_z is total number of proximal transcript pairs and T_z is total number of transcript pairs (of g_i, g_j) in z-th plane. PP-3D calculates a p-value for each gene pair as $\text{p-value}(g_i, g_j) = Binomial(T, p, K)$.

Conditional poisson binomial (CPB) test

CPB test detects a d-colocalized gene pair, i.e., a gene pair that is a proximal pair in significantly many cells. It assigns a p-value (one-sided) to the number of cells in which a gene pair is found to be proximal pair detected using PP test. We first describe a simpler version of the test ("unconditional Poisson Binomial" or UPB) test that assumes that all gene pairs are equally likely to be proximal pair in a cell but allows for the fact that different cells may have different number of proximal pairs. Let X_{ij}^c be a binary variable denoting if g_i, g_j are a proximal pair in c-th cell. X_{ij}^c is assumed to follow a Bernoulli distribution with parameter p_0^c, which is estimated as the fraction of proximal gene pairs in the cell:

$$p_0^c = \equiv \frac{\sum_{k \le l} X_{k,l}^c}{\sum_{k \le l} 1} = \frac{\sum_{k \le l} X_{k,l}^c}{\binom{n}{2}} \tag{3}$$

where n denotes total number of genes. This estimate of p_0^c assumes that all gene pairs can be a proximal pair. To incorporate the fact that a gene pair cannot be a proximal pair if either of the genes is not expressed in the cell, the above estimate is modified as,

$$p_0^c \equiv \frac{\sum_{k \le l} X_{k,l}^c}{\sum I_{k \le l}(g_k, g_l)} \tag{4}$$

where $I(g_k, g_l)$ is an indicator function that equals to 1 iff both g_k and g_l are expressed.

CPB test is a modified version of the UPB test that accounts for the possibility that all gene pairs are not equally likely to be colocalized in a cell and sets the Bernoulli parameter (p_0^c above) to be gene pair-dependent. Let z_i denote total number of proximal pairs having gene i as one of the genes, aggregated across all cells, i.e.,

$$z_i = \sum_{j \le c} X_{ij}^c \tag{5}$$

We use these global summary statistics to model the prior probability Π_{ij} that a proximal pair detected in a cell is the gene pair g_i, g_j, as follows:

$$\Pi_{ij} \equiv \frac{z_i z_j}{\sum_{i \le j} z_i z_j} \tag{6}$$

This model de-emphasizes gene pairs comprising genes that are frequently found to be in proximal pairs across cells. Now, the Bernoulli parameter for variable X_{ij}^c is estimated as

$$p_{ij}^c \equiv 1 - \left(1 - \Pi_{ij}\right)^{\sum_{i \le j} X_{ij}^c} \tag{7}$$

The total number of cells where g_i, g_j is a proximal pair follows a Poisson Binomial distribution

$$\sum_{c=1}^{m} X_{ij}^c \sim Poisson\ Binomial\left(p_{ij}^1, \ldots, p_{ij}^m\right) \tag{8}$$

Subcellular annotation of a d-colocalized gene pair

A d-colocalized pair is annotated by cellular region where the gene pair's proximal pairs tend to be found. We define four categories –

Nucleus (Nuc), Peri-Nucleus (PN), Cytosol (Cyto) and Cell Periphery (CP). Proximal pairs in each cell are annotated by cellular region and is aggregated across cells to yield primary and secondary category. Perinuclear (PN) region is defined as including x microns on either side of the nuclear membrane, while Cell Periphery (CP) is defined as regions in cytoplasm within y microns of the cell membrane. Remaining regions are designated as Cytosol (Cyt) or Nucleus (Nuc). We chose x = 2 micron which corresponded to ~35% of nucleus transcripts being annotated as perinuclear, and y = 4 micron which corresponds to ~35% cytosolic transcripts being annotated as cell periphery.

RNA-RNA interaction (RRI)

For RRI, we set distance d to be 300 nm (resolution of MERFISH data is 200 nm). The small distance was chosen to capture gene pairs whose d-colocalization may be explained due to the binding of their transcripts. To control FPR at small distance, we used stricter p-value threshold of 1e-5 that resulted in 60 d-colocalized pairs (FPR-0%). We used RNAplex[47] to compute the RRI scores. For this, we retrieved the nucleotide sequences from the Ensembl database[87] and got the specific transcript id to get the correct spliced form. RNAplex has been shown to be among the most accurate tools while being fast enough to compute the scores for gene pairs with their full transcripts. Finally, we performed a hypergeometric test on d-colocalized pairs and pairs with RRI score greater than a fixed threshold (RRI > 35). 8 out of 60 d-colocalized pairs had high RRI scores that led to significant overlap (p-value = 0.02).

Enrichment analysis

To understand the biological mechanism or consequences of d-colocalization, we tested if the compendium of d-colocalized gene pairs has significant overlap with functionally related gene pairs. We define a gene pair to be functionally related if both genes are present in same KEGGpathway or are annotated with same GO terms more than K times. K was chosen such thatnumber of gene pairs is similar across d-colocalized and functionally related set. This partially offsets the confounding impact of set size variations when performing multiple gene set enrichment tests. In our analysis, K (MF) = 2, K (BP) =1, K (CC) = 3, K (pathway) = 1. We performed a hypergeometric test between d-colocalized pairs and functionally related set.

Differential colocalization of a gene pair

InSTAnT employs a series of statistical tests to categorize a pair based on its specificity to a cell type, region or phenotype. First, it tests the association between cells where a gene pair was deemed a significant proximal pair and cells of a particular type (e.g., inhibitory neurons), using a Hypergeometric test. (This process is repeated for every cell type). If such an association is found to be statistically significant, it is subjected to further tests to determine if the cell type specificity arises simply because one of the genes in the pair is expressed specifically in that cell type. For this, InSTAnT utilizes a version of the generalized Hypergeometric test that tests for an association between two sets conditional on a third set[88], as described below. In this case, the third set comprises the cells with high expression of one of the genes in the pair.

Let U be the set of all cells, M be the set of cells of a particular cell type, O be the set of cells where a gene pair is deemed a proximal pair and E be the set of cells with high expression of one of the genes in the pair. M, O and E are subsets of U. The threshold for high gene expression used in defining E is chosen such that size (E) = size (M). Let $|M \cap E| = \gamma, |M \cap O| = \lambda, |E \cap O| = \alpha$ |. The Hypergeometric test p-value of association between M and O is given by the probability that a random set of size $|O|$ has an overlap (intersection) of size greater than or equal to λ with M. However, we wish to test if the overlap between M and O is significant beyond what is expected not from a random set of size $|O|$ but a random set of this size that respects the known overlap between M and E and between E and O. For this, we calculate probability of the overlap between M and a random set of $|O|$ being greater than or equal to λ conditional on the observed overlap between M and E and that between E and O, as follows:

$$\frac{\sum_{k=\lambda}^{\min(|M|,|O|)} \sum_{\beta=0}^{k} \binom{\gamma}{\beta}\binom{m-\gamma}{k-\beta}\binom{n_1-\gamma}{\alpha-\beta}\binom{|U|-|M|-|E|+\gamma}{|O|-\alpha-k+\beta}}{\binom{|E|}{\alpha}\binom{|U|-|E|}{|O|-\alpha}} \quad (9)$$

This is an example of multivariate hypergeometric distribution. We use *scipy.stats.multivariate_hypergeom* package for multivariate hypergeometric distribution.

For each gene pair that is associated with a cell type, InSTAnT performs the above test twice, each time conditioning on a set E defined by the high expression cells for one of the genes of the pair. Significant p-values in both tests thus performed indicate that the cell type-specificity of the d-colocalized gene pair is significant beyond what is expected from the specificity of either gene's expression. Furthermore, InSTAnT tests if either gene of the pair is a marker of the cell type, defined as any gene among the top 10 by association between their expression and the cell type. A marker gene is found by conducting Hypergeometric test of overlap between O and E.

Using the above tests, InSTAnT categorizes a gene pair vis-à-vis specificity as follows: If the gene pair is significantly associated with a cell type/region/phenotype (first test above), then it belongs to Category 2 if the association is significant by the Hypergeometric test conditional on high expression cells of both genes and neither gene is a marker of the cell type, otherwise it belongs to Category 1.

Probabilistic graphical model for spatial modulation

InSTAnT uses a likelihood ratio test to determine if sub-cellular colocalization of a d-colocalized gene pair is spatially modulated at the tissue level. Informally, this means that the cells in which the gene pair is deemed to be a proximal pair are non-randomly distributed in the physical space.

The probabilistic model is formulated around a graph with a node for each cell and edges between neighboring cells. Two cells are neighboring cells if they are located within a configurable distance (set to 100 micron in our tests). Each node is associated with a binary variable s_c that indicates whether the specific gene pair (say g_i, g_j) is a proximal pair in the corresponding cell c, as detected by the PP test. The variable s_c is assumed to be a Bernoulli-distributed variable. The null hypothesis is that the Bernoulli parameter is a global constant p^{global} shared across all cells, i.e., it does not depend on the cell c and thus on its spatial location:

$$H_0: s_c \sim Ber\left(p^{global}\right) \quad (10)$$

p_{global} is estimated as the fraction of cells where the gene pair g_i, g_j is a proximal pair, which is its maximum likelihood estimate. In the alternative hypothesis, the model assumes that the distribution of variable s_c depends on the fraction of cells c' in the neighborhood of c for which $s'_c = 1$. Let p^{local} be the fraction of cells c' in the neighborhood of c for which $s'_c = 1$.

$$H_1: s_c \sim Ber\left(wp^{local} + (1-w)p^{global}\right) \quad (11)$$

$$0 < w < 1$$

The parameters p^{global}, p^{local}, w are learnt by maximizing likelihood. Weight w controls the contribution of local neighborhood. InSTAnT calculates the log likelihood ratio (LLR) for each gene pair in

the d-colocalization map and pairs with LLR above a threshold are designated as spatially modulated. The threshold is obtained by random permutation of the of s_c values of cells, repeating the above test and selecting the highest LLR score (over all gene pairs) seen on the randomized data. This allows us to detect spatially clustered distributions of cells supporting g_i, g_j colocalization.

Module discovery: global colocalization clustering (GCC)

GCC is a procedure to analyze a d-colocalization map to identify subsets of genes that exhibit a high frequency of pairwise d-colocalization relationships. To this end, it represents the d-colocalization map as an n x n matrix (n = number of genes) whose entries are the negative logarithm of p-values of gene pairs from the CPB test and performs a hierarchical clustering of rows and columns using Euclidean distance with Ward criterion. (The constant 1e-64 is added to all the p-values to handle zero p-values prior to taking logarithms).

Module discovery: frequent subgraph mining (FSM)

FSM seeks a network of genes that is "colocalized" in many cells, where colocalization of a network in an individual cell means that every gene pair connected by an edge in that network is a proximal pair in that cell. It constructs a *colocalization graph* for each cell with genes as nodes and edges representing proximal gene pairs from PP test. It then uses an efficient graph mining tool called gSPAN[54] to detect subgraphs with a pre-specified minimum size (numbers of nodes and edges) that are supported by a pre-specified minimum number of cells.

MERFISH imaging and analysis

Cell line Source and Authentication: U2OS Cell lines were purchased from ATCC, original donor white female. Cell lines were authenticated by Cancer center at Illinois using the following method: Amplified with AmpFISTR Identifiler Plus PCR Amplification Kit and analyzed on the Applied Biosystems 3730/ GeneMapper 6.

General cell culture conditions. U2 OS cells were cultured in minimal essential medium (MEM) from ATCC with 1 mM sodium pyruvate, 10% fetal bovine serum (FBS), and 1% penicillin-streptomycin (Pen-Strep). The cells were obtained from ATCC and maintained using the recommended protocol.

MERFISH sample preparation. U2 OS MERFISH samples were prepared using a previously published method[89]. MERFISH encoding probe sequences were originally from the Zhuang lab(ref. 89), and can be found in the "Zhuang U2OS probes" Source Data file. In brief, U2 OS cells were plated on a salinized 40mm #1.5 coverslip (Fisher Scientific). Plated cells were transferred to a 37 °C and 5% CO_2 incubator overnight to grow. Cells were then fixed with 4% paraformaldehyde (Electron Microscopy Sciences) and permeabilized with 0.5% (vol/vol) Triton X-100 (Sigma Aldrich). Samples were stained with encoding probes (10nM/probe) and anchor probes (1μM) for 36 hours in a humidified incubator at 37 °C. To stabilize the cells during clearing, the stained cells were embedded in a thin, 4% polyacrylamide (PA) gel. Fiducial beads (Spherotech, FP-0245-2) were also included in the gel to align rounds of MERFISH images.

Commonly used imaging solutions. The following solutions were used during imaging experiments described in this work. Readout wash buffer was adapted from Moffit et al.[89] and contained 10% (v/v) ethylene carbonate (Sigma Aldrich), 0.1% Triton X-100 in 2x SSC. Imaging buffer adapted from Moffit et al.[89] and contained 5mM 3,4-dihydroxybenzoic acid (PCA; Sigma Aldrich), 2 mM trolox (Sigma Aldrich), 50 μM trolox quinone, 1:500 of recombinant protocatechuate 3,4-dioxygenase (rPCO; OYC Americas), adjusted to a pH of 7-7.2 using 1 N NaOH (VWR International) in 2x SSC. Cleavage buffer was adapted from[89] and contained 0.05 M TCEP HCl, adjusted to a pH of 7-7.2 using 1 N NaOH, in 2x SSC. Stripping buffer was adapted from Eng. et al.[14] and contained 55% formamide, and 0.1% Triton X-100 in 2x SSC.

MERFISH imaging. All images were acquired using a Zeiss Axiovert-200m widefield microscope (Carl Zeiss AG) located in the IGB core imaging facility. The sample was placed into a flow cell (Bioptechs, FCS2), filled with RNAse free 2x SSC, and connected to a lab built automated flow system. Briefly, computer-controlled valves (Hamilton, MVP/4, 8-5 valve) are used to select which solution was pulled across the sample by a computer controlled pump (Gilson, Minipuls 3). All systems are controlled by a custom designed Python script that can communicate with the microscope to start imaging or start flowing after an imaging round is done. In brief, a single round of imaging involves staining with fluorescently labeled readout probes (0.4 mL/min for 6 minutes, and 0.34 mL/min for 6 minutes), washing with readout wash buffer (0.23 mL/minute for 9 minutes) to remove unbound probes, and imaging buffer was flowed into the flow cell prior to imaging (0.34 mL/minutes for 6 minutes) to reduce photobleaching. MERFISH readout probe sequences were originally from the Zhuang lab (ref. 89), and can be found in the "16 bit U2OS RO probes" Source Data file. A single quad band excitation filter (Chroma, ZET402/468/555/638x) and dichroic (Chroma, ZT405/470/555/640rpc-UF1) were used to image all samples. Excitation was provided by a 7 laser system (LDI WF, 89 North). Alexa Fluor 647 (Fisher scientific) labeled probes were excited using a 647 nm laser (0.5 W) with a ET700/75m (Chroma) emission filter, and 1.5 second exposure time. Atto 565 (Atto tec) labeled probes were excited using a 555 nm laser (1 W) with a ET610/75 m (Chroma) emission filter, and a 0.75 second exposure time. Fiducial beads were imaged with a 405 nm laser (0.3 W) with a ET440/40 m emission filter, and a 1-second exposure time. Samples were imaged with a 63x oil immersion objective (Carl Zeiss AG, 420782-9900-000), and focus was maintained between imaging rounds using Definite Focus (Carl Zeiss AG). 9 z planes with 0.7 μm steps were taken for each FOV, and a total of 100 FOVs were acquired. After imaging is complete, a cleavage buffer (0.2 mL/minute for 15 minutes) was flowed across the sample to remove the fluorophores from the probes. The cleavage buffer was washed away using RNAse free 2x SSC (0.5 mL/minute for 10 minutes). This process was repeated for a total of 8 rounds of imaging. PolyA probes were stained after the final imaging round using the same method as described above.

MERFISH data processing. Individual FOVs were exported from czi format into 16 bit tiff format using Zen (Carl Zeiss AG) using the image export method. Images then were reformatted into image stacks by FOV and round. A modified copy of MERLIN[90] was used to decode MERFISH spots. In brief, for each FOV, images from different rounds are aligned using fiducial beads that were imaged in each round. Aligned images are then normalized, decoded, and identified spots filtered using previously published methods[31]. Cell segmentation was done separately from MERLIN using Cellpose[91] on PolyA and DAPI images for each FOV. To improve FOV alignment to neighboring FOVs, the DAPI channel was used with the restitching function found in Zen (Edge detection: on, minimal overlap: 5%, maximal shift: 15%, comparer: best, Global optimizer: best). Using the aligned images, segmented cells that cross FOV boundaries were merged into single cells, and global positions were generated for each spot. Spots are then assigned to cells based on their spatial coordinates. Spots were then filtered to remove any spot smaller than 3 pixels in size.

smFISH probe design. All smFISH probes were designed using the Stellaris probe designer (Biosearch technologies). Probes were designed using the following settings: Masking level: 5, max number of probes: 48, oligo length: 20, minimum spacing length: 2. SRRM2 exon probes were designed against SRRM2 isoform ENST00000301740 (GRCh38.p13). SRRM2 intron probes were randomly selected from

Table 2 | Imaging parameters used for smFISH experiment

Channel	Target	Laser line (power)	Exposure time	Emission filter
DAPI	Fiducial beads, nuclei	405 nm (0.3 W)	0.075 seconds	ET440/40 m
Alexa Fluor 488	MALAT1 lncRNA	470 nm (1 W)	2 seconds	ET525/50 m
Cy3	SRRM2 intron RNA	555 nm (1 W)	2 seconds	ET610/75 m
Cy5	SRRM2 exon mRNA	640 nm (0.5 W)	3 seconds	ET700/75 m

Emission filters were purchased from Chroma.

probes designed for three different introns defined by ensemble (SRRM2-230 intron 1, SRRM2-230 intron 2, and SRRM2-230 intron 10) (GRCh38.p13). MALAT1 probes were designed against MALAT1 isoform ENST00000534336 (GRCh38.p13). All probes were purchased from Biosearch modified with mdC (TEG-Amino) at the 3′ terminus. All probe sequences corresponding to MALAT1, SRRM2 exon and SRRM2 intron can be found in "smFISH probes" Source Data file. The probes were dissolved in TE buffer and labeled using AF488/Cy3/Cy5 NHS esters for MALAT1, SRRM2 intron, and SRRM2 exon, respectively. The labeled probes were purified using the Bio-Rad Bio-Spin P-6 purification columns (Cat # 732-6221).

smFISH sample preparation. Approximately 1.5-1.8 million U2OS cells were plated on a #1.5, 40 mm coverslip (Fisher Scientific) that has been UV treated before plating. The cells were then transferred to an incubator at 37 °C and 5% CO2, overnight for 12-16 hours.

Modified from Fei et al.[92], the sample was rinsed with 1x PBS (Corning), followed by fixation using 4% paraformaldehyde (PFA; Electron Microscopy Sciences) in 1x PBS for 10 minutes at room temperature (RT). The sample was then washed three times with 1x PBS and permeabilized with 0.5% Triton X-100 (Sigma Aldrich), 2 mM vanadyl ribonucleoside complexes (VRC; Sigma Aldrich) in 1x PBS for 10 minutes on ice, followed by three quick washes with 1x PBS. At this point, the sample can be stored in 70% Ethanol at 4 °C if the experiment needs to be paused temporarily.

To prepare for smFISH hybridization, sample was rinsed with 10% formamide (Sigma Aldrich) in 2x saline sodium citrate (SSC; Fisher Scientific). smFISH probe hybridization buffer was prepared with 0.2 mg/mL of bovine serum albumin (BSA; Fisher Scientific), 2 mM VRC, 10% dextran sulfate (Sigma Aldrich), 1 mg/mL yeast tRNA (Fisher Scientific), 10% formamide, 1% murine RNase inhibitor (New England BioLabs) in 2x SSC. Avoid light exposure from this point forward. smFISH probes were then added to the FISH hybridization buffer at a final concentration of 14 nM for each targeted RNA (MALAT1, SRRM2 intron, and SRRM2 exon).

A humidified chamber was made using an empty pipette box filled halfway with nuclease-free water (Corning) at the base and a UV-treated glass slide covered with a parafilm layer on top. A 100 µl drop of the FISH probe hybridization buffer was then added on top of the parafilm layer and the sample was casted over the drop with the cell side facing down. The chamber was then placed in an incubator in dark and wrapped entirely with aluminum foil overnight at 37 °C for at least 16 hours. The sample was quickly rinsed two times with 10% formamide in 2x SSC then stained with 4′,6-diamidino-2-phenylindole (DAPI; Invitrogen by Fisher Scientific) 1:1000 of 1 mg/mL stock solution and 1:5000 of Fluoro-Max Blue Aqueous Fluorescent Particles (fluorescent beads; Fisher Scientific) in 2x SSC. The sample was incubated with the DAPI and fluorescent beads solution for 5 minutes while rocking at RT, followed by a quick wash with 2x SSC, then stored in 2x SSC at 4 °C until ready for imaging.

Protein staining. After smFISH imaging, the sample can be stored in 1x PBS at 4 °C for up to a week before protein staining. Samples were fixed a second time with 4% PFA in 1x PBS for 5 minutes at RT, then rinsed three times with 1x PBS. This was followed by incubation with a blocking solution of 1% BSA in 1x PBS for three consecutive times with 10 minutes each time at RT.

The SON primary antibody (Anti-SON, Sigma Aldrich, HPA023535) was kept at −20 °C until ready for use. The primary antibody stock solution of 1:1000 was prepared with 1x PBS and kept on ice. A 1:5000 primary antibody dilution was prepared in blocking solution and the sample was incubated with 200 µl of the primary antibody solution for approximately 1 hour at RT in the dark.

The sample was washed with blocking solution three consecutive times with a 10-minute incubation each time at RT, followed by three washes with 1x PBS, for 10 minutes each time at RT.

The secondary antibody was conjugated to Alexa Fluor 647 (Goat anti-rabbit, Invitrogen, A21245). The concentrated secondary antibody was kept at 4 °C until ready for use. Sample staining was accomplished by 1:1000 dilution of the secondary antibody in blocking solution and casting of the sample on a 200 µl drop of the secondary antibody solution, with the cell side facing down. The sample was then incubated for 1 hour in the dark at RT. The sample was re-stained with DAPI in 1x PBS with the same concentration and incubation time described in smFISH staining section. This was followed by a quick rinse with 1x PBS and the sample was stored in 1x PBS at 4 °C until ready for imaging.

smFISH image acquisition. smFISH and protein imaging were done on the same MERFISH imaging and fluidic system described above (MERFISH imaging). After placing the sample into the flow cell, imaging buffer was flowed through the system (0.34 mL/minute for 5 minutes). Excitation and dichroic filters were the same as used above. Table 2 contains the dyes, lasers, and emission filters were used for smFISH imaging.

Samples were imaged with the same 63x oil immersion objective as above, and focus was maintained between imaging rounds using Definite Focus. 9 z planes were imaged with a step size of 0.7 µm. After imaging, smFISH probes were removed using a stripping buffer that was flowed through the system (0.34 mL/minutes for 5 minutes) without removing the sample from the microscope. After stripping the sample was washed with 2x SSC (0.5 mL/minutes for 5 minutes). The sample was imaged a second time using the same settings as above. After imaging the sample was removed from the flow cell and placed into 1x PBS prior to protein staining (Protein staining).

After protein staining was complete, the sample was placed into the flow cell and filled with imaging buffer. The same region imaged during the smFISH experiment was found and reimaged using the same objective and z-stack settings as above. Table 3 contains the dyes, lasers, and emission filters were used for protein imaging.

SRRM2 image registration and alignment. Individual FOVs were exported from czi format into 16 bit tiff format using Zen's (Carl Zeiss AG) image export method. To align images from the same FOV across multiple rounds of imaging or experiment, blue fluorescent beads imaged in the DAPI channel were used as fiducial markers. We found that aligning images from the same experiment required a simple translation. To align protein images with mRNA images, an iterative rotation and translation process was developed. For each iterative round of alignment, the protein DAPI channel was rotated, then translated to best align with the mRNA image, this warped image was

Table 3 | Imaging parameters used for SON protein labeling experiment

Channel	Target	Laser line (power)	Exposure time	Emission filter
DAPI	Fiducial beads, nuclei	405 nm (0.3 W)	0.05 seconds	ET440/40 m
Alexa Fluor 647	SON protein	640 nm (0.5 W)	1.5 seconds	ET700/75 m

Emission filters were purchased from chroma.

then used as the starting protein DAPI image for the next round of alignment. We found that it took between 2 and 5 rounds of alignment to align protein images to mRNA images. Chromatic aberration was corrected by aligning all channels to the Cy5 channel. Multicolor beads (Multi-speck bead slide, Carl Zeiss AG, 1783-455) that included dyes in the Alexa Fluor 488, Cy3, and Cy5 channels were used to correct Alexa Fluor 488 and Cy3 channels. The DAPI channel was corrected to the Cy5 channel using the fiducial bead cross talk between the DAPI and Alexa Fluor 488 channels. This was done by calculating the shift between non-nuclear regions of the DAPI and Alexa Fluor 488 channels, then adding the Alexa Fluor 488 to Cy5 shift to the DAPI to Alexa Fluor 488 shift.

SRRM2 image preprocessing. To remove cross talk in DAPI and Alexa Fluor 488 channels caused by the fiducial beads, stripped Alexa Fluor 488 mRNA channel was subtracted from the stained Alexa Fluor 488 channel. As fiducial beads are not affected by the mRNA stripping conditions, any spots that remain in the stripped Alexa Fluor 488 channel would be from the beads, not from MALAT1 mRNA. In order to reduce background in other images, round subtraction was also done on the other channels of the mRNA FOV.

SRRM2 co-localization analysis. Co-localization analysis was done on a single z plane from each experiment stack. Images were then filtered using a high pass filter (5 pixel sigma) and Lucy–Richardson deconvolution (10 iterations, 9 pixel filter size, 1.4 pixel sigma). Filtered images are then converted to binary masks with manually defined thresholds. To remove false positives in the MALAT1 channel, the MALAT1 mask was multiplied with the inverse of the stripped MALAT1 mask. Cell nuclei were identified using the DAPI channel and segmented using a manually defined threshold.

The co-localization rate was calculated for each nucleus defined from the DAPI channel. To calculate the co-localization rate between two channels, each channel is multiplied against the nuclei mask. For each spot in the first mask, the spot was dilated by 2 μm and then compared against the second mask. If the dilated spot overlaps any spot in the second mask, it is considered to be colocalized. The colocalization rate was then calculated to be the following:

$$colocalization\ percent = \frac{Co-localized\ spots\ ct}{Total\ spots\ ct} * 100\% \quad (12)$$

The colocalization percent was averaged across 13 cells.

Software used. We used Merlin software for our U2OS data(v0.1.6,-Zenodo, https://doi.org/10.5281/zenodo.3758540).

SRRM2 figure generation (Fig. 4). SRRM2 exon and intron images were filtered using a high pass filter with 2 pixel sigma, while MALAT1 was filtered using high pass filter with 5 pixel sigma. Raw SON images were used in panels Fig. 4a–d.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

U2OS Dataset: We obtained MERFISH data[45] on a human osteosarcoma cell line (U2-OS) from http://zhuang.harvard.edu/MERFISHData/data_for_release.zip. We used the authors' Matlab code to extract and output the data in table format. We filtered the data to retain transcripts having minimum area of 3 and intensity of $10^{0.75}$. The dataset had 7 replicates. We were able to extract data for four replicates – *rep2, rep3, rep4, rep5*; the other replicates presented severe memory management challenges and were not analyzed. Most of the reported results are from analysis of *rep3*, which profiles 130 genes in 3237 cells with an average of 1243 transcripts per cell. Global *d*-colocalization maps were constructed for all four replicates and compared to assess reproducibility. Brain Dataset: Data reported in Moffit et al.[54] were obtained through personal communication with Dr. Jeffrey Moffitt. The dataset contained 6325 cells with 553 average number of transcripts across 7 z-planes. We obtained cell type assignment from Supplememtary Data 1 from Moffit et al.[54]. We removed ambiguous cells leading to 5149 cells with 9 cell types. Proximal pairs were detected in cells that have at least one z-plane with 20 or more transcripts. Brain Dataset for behavior analysis: Data reported in Moffit et al.[54] were obtained through personal communication with Dr. Jeffrey Moffitt. We used a naïve animal (animal ID=5) and an aggressive animal (animal ID=31) for the analysis. Seqfish+ Data: We used data from NIH/3T3 mouse embryonic fibroblast cells[14] spatially profiled with seqFISH+. The data is available at - https://zenodo.org/record/2669683 and can be accessed using Bento[34] tool. We used Bento[34] for filtering preprocessing that resulted in 3726 genes and 179 cells. Xenium Data: We downloaded whole mouse brain spatial transcriptomic data from https://www.10xgenomics.com/resources/datasets/fresh-frozen-mouse-brain-replicates-1-standard. We used data from replicate 1. Xenium data, in addition to providing the subcellular spatial transcriptome of each cell, provides a cluster identifier that refers to the cluster of cells (obtained based on transcriptomic similarity) that this cell belongs to. We used Allen brain atlas[93] to identify clusters that approximate CA1 (cluster 27), CA3 (cluster 42) and Dentate gyrus (cluster 13) region of Hippocampus. We found a total of 7915 cells across these three clusters. The dataset consists of x,y and z (fine-grained) positions. To analyze Xenium data, we used InSTAnT PP test using all of the x,y,z information instead of PP-3D test since data consisted of fine-grained z positions (not multiple z-planes). Inhouse U2OS Dataset: The MERFISH U2-OS dataset is available at https://doi.org/10.13012/B2IDB-2930842_V1. Source data are provided with this paper.

Code availability

The code is available at https://github.com/bhavaygg/InSTAnT[94]. The package uses anndata object and can be installed as "pip install scinstant". The runtime for InSTAnT on the U2-OS data consisting of 3237 cells and 140 genes on a computing cluster consisting of 16 threads was ~465 seconds for PP test and ~560 seconds for CPB test.

References

1. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
2. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. methods* **18**, 9–14 (2021).
3. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. methods* **15**, 343–346 (2018).

4. Zhu, J., Sun, S. & Zhou, X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biol.* **22**, 1–25 (2021).
5. Chidester, B., Zhou, T., Alam, S. & Ma, J. SpiceMix enables integrative single-cell spatial modeling of cell identity. *Nat. Genet.* **55**, 78–88 (2023).
6. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 1–31 (2021).
7. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).
8. Jin, S. et al. Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).
9. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
10. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. methods* **17**, 159–162 (2020).
11. Rao, N., Clark, S. & Habern, O. Bridging genomics and tissue pathology: 10x genomics explores new frontiers with the visium spatial gene expression solution. *Genet. Eng. Biotechnol. N.* **40**, 50–51 (2020).
12. Liu, Y. et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* **183**, 1665–1681.e18 (2020).
13. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
14. Eng, C.-H. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).
15. Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
16. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
17. Alon, S. et al. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science* **371**, eaax2656 (2021).
18. Hu, J. et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. methods* **18**, 1342–1351 (2021).
19. Chen, W.-T. et al. Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* **182**, 976–991.e19 (2020).
20. Liu, Z., Sun, D. & Wang, C. Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information. *Genome Biol.* **23**, 1–38 (2022).
21. Li, D., Ding, J. & Bar-Joseph, Z. Identifying signaling genes in spatial single-cell expression data. *Bioinformatics* **37**, 968–975 (2021).
22. Pham, D. et al. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv*, 2020.05. 31.125658 (2020).
23. Hildebrandt, F. et al. Spatial transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. *Nat. Commun.* **12**, 7046 (2021).
24. Doyle, F. et al. Bioinformatic Tools for Studying Post-Transcriptional Gene Regulation. *Post-Trans. Gene Regul.* **419**, 39–52 (2008).
25. Parton, R. M., Davidson, A., Davis, I. & Weil, T. T. Subcellular mRNA localisation at a glance. *J. cell Sci.* **127**, 2127–2133 (2014).
26. Kloc, M., Zearfoss, N. R. & Etkin, L. D. Mechanisms of subcellular mRNA localization. *Cell* **108**, 533–544 (2002).
27. Besse, F. & Ephrussi, A. Translational control of localized mRNAs: restricting protein synthesis in space and time. *Nat. Rev. Mol. cell Biol.* **9**, 971–980 (2008).
28. Bourke, A. M., Schwarz, A. & Schuman, E. M. De-centralizing the central dogma: mRNA translation in space and time. *Mol. Cell* **83**, 452–468 (2023).
29. Martin, K. C. & Ephrussi, A. mRNA localization: gene expression in the spatial dimension. *Cell* **136**, 719–730 (2009).
30. Blower, M. D. Molecular insights into intracellular RNA localization. *Int. Rev. cell Mol. Biol.* **302**, 1–39 (2013).
31. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl Acad. Sci.* **116**, 19490–19499 (2019).
32. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
33. Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. RNA velocity—current challenges and future perspectives. *Mol. Syst. Biol.* **17**, e10282 (2021).
34. Mah, C. K. et al. Bento: A toolkit for subcellular analysis of spatial transcriptomics data. *BioRxiv* (2022).
35. Engel, K. L., Arora, A., Goering, R., Lo, H. Y. G. & Taliaferro, J. M. Mechanisms and consequences of subcellular RNA localization across diverse cell types. *Traffic* **21**, 404–418 (2020).
36. Fazal, F. M. & Chang, H. Y. Subcellular spatial transcriptomes: Emerging frontier for understanding gene regulation. in *Cold Spring Harbor symposia on quantitative biology* 84 31–45 (Cold Spring Harbor Laboratory Press, 2019).
37. Liu, Y. et al. Proximity Chemistry in Living Systems. *CCS Chem.* **5**, 802–813 (2023).
38. Jansen, R.-P. & Niessing, D. Assembly of mRNA-protein complexes for directional mRNA transport in eukaryotes-an overview. *Curr. Protein Pept. Sci.* **13**, 284–293 (2012).
39. Ryder, P. V. & Lerit, D. A. RNA localization regulates diverse and dynamic cellular processes. *Traffic* **19**, 496–502 (2018).
40. Quinodoz, S. A. et al. RNA promotes the formation of spatial compartments in the nucleus. *Cell* **184**, 5775–5790.e30 (2021).
41. Katz, Z. B. et al. β-Actin mRNA compartmentalization enhances focal adhesion stability and directs cell migration. *Genes Dev.* **26**, 1885–1890 (2012).
42. Lelek, M. et al. Single-molecule localization microscopy. *Nat. Rev. Methods Prim.* **1**, 39 (2021).
43. Lagache, T. et al. Mapping molecular assemblies with fluorescence microscopy and object-based spatial statistics. *Nat. Commun.* **9**, 698 (2018).
44. Battich, N., Stoeger, T. & Pelkmans, L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat. methods* **10**, 1127–1133 (2013).
45. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci.* **113**, 11046–11051 (2016).
46. Van Dam, S., Vosa, U., van der Graaf, A., Franke, L. & de Magalhaes, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinforma.* **19**, 575–592 (2018).
47. Tafer, H. & Hofacker, I. L. RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics* **24**, 2657–2663 (2008).
48. Ilik, İ. A. et al. SON and SRRM2 are essential for nuclear speckle formation. *Elife* **9**, e60579 (2020).
49. Miyagawa, R. et al. Identification of cis-and trans-acting factors involved in the localization of MALAT-1 noncoding RNA to nuclear speckles. *Rna* **18**, 738–751 (2012).
50. Chicurel, M. E., Singer, R. H., Meyer, C. J. & Ingber, D. E. Integrin binding and mechanical tension induce movement of mRNA and ribosomes to focal adhesions. *Nature* **392**, 730–733 (1998).

51. Adekunle, D. A. & Wang, E. T. Transcriptome-wide organization of subcellular microenvironments revealed by ATLAS-Seq. *Nucleic Acids Res.* **48**, 5859–5872 (2020).
52. Stefanovic, B., Stefanovic, L. & Manojlovic, Z. Imaging of type I procollagen biosynthesis in cells reveals biogenesis in highly organized bodies; Collagenosomes. *Matrix Biol.* **12**, 100076 (2021).
53. Theisen, U., Straube, E. & Straube, A. Directional persistence of migrating cells requires Kif1C-mediated stabilization of trailing adhesions. *Dev. cell* **23**, 1153–1166 (2012).
54. Moffitt, J. R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
55. Sakers, K. et al. Astrocytes locally translate transcripts in their peripheral processes. *Proc. Natl Acad. Sci.* **114**, E3830–E3838 (2017).
56. Kater, M. S. et al. A novel role for MLC1 in regulating astrocyte–synapse interactions. *Glia* **71**, 1770–1785 (2023).
57. Lanciotti, A. et al. Megalencephalic leukoencephalopathy with subcortical cysts protein 1 functionally cooperates with the TRPV4 cation channel to activate the response of astrocytes to osmotic stress: dysregulation by pathological mutations. *Hum. Mol. Genet.* **21**, 2166–2180 (2012).
58. Hwang, J., Vu, H. M., Kim, M.-S. & Lim, H.-H. Plasma membrane localization of MLC1 regulates cellular morphology and motility. *Mol. brain* **12**, 1–14 (2019).
59. Fleischer, J., Schwarzenbacher, K., Besser, S., Hass, N. & Breer, H. Olfactory receptors and signalling elements in the Grueneberg ganglion. *J. Neurochem.* **98**, 543–554 (2006).
60. Liu, M., Kim, D.-W., Zeng, H. & Anderson, D. J. Make war not love: The neural substrate underlying a state-dependent switch in female social behavior. *Neuron* **110**, 841–856.e6 (2022).
61. Seigneur, E., Wang, J., Dai, J., Polepalli, J. & Südhof, T. C. Cerebellin-2 regulates a serotonergic dorsal raphe circuit that controls compulsive behaviors. *Mol. psychiatry* **26**, 7509–7521 (2021).
62. Almutairi, O., Almutairi, H. A. & Al Rushood, M. Protein-activated kinase 3 (PAK3)-related intellectual disability associated with combined immunodeficiency: a case report. *Am. J. Case Rep.* **22**, e930966–1 (2021).
63. Shibata, M. et al. Hominini-specific regulation of CBLN2 increases prefrontal spinogenesis. *Nature* **598**, 489–494 (2021).
64. Dubos, A. et al. Alteration of synaptic network dynamics by the intellectual disability protein PAK3. *J. Neurosci.* **32**, 519–527 (2012).
65. Seigneur, E. & Südhof, T. C. Cerebellins are differentially expressed in selective subsets of neurons throughout the brain. *J. Comp. Neurol.* **525**, 3286–3311 (2017).
66. Meng, J., Meng, Y., Hanna, A., Janus, C. & Jia, Z. Abnormal long-lasting synaptic plasticity and cognition in mice lacking the mental retardation gene Pak3. *J. Neurosci.* **25**, 6641–6650 (2005).
67. Perez, J. D. et al. Subcellular sequencing of single neurons reveals the dendritic transcriptome of GABAergic interneurons. *Elife* **10**, e63092 (2021).
68. Marco Salas, S. et al. Optimizing Xenium In Situ data utility by quality assessment and best practice analysis workflows. *bioRxiv*, 2023.02. 13.528102 (2023).
69. Dicken, M. S., Hughes, A. R. & Hentges, S. T. Gad1 mRNA as a reliable indicator of altered GABA release from orexigenic neurons in the hypothalamus. *Eur. J. Neurosci.* **42**, 2644–2653 (2015).
70. Le, T. N. et al. GABAergic interneuron differentiation in the basal forebrain is mediated through direct regulation of glutamic acid decarboxylase isoforms by Dlx homeobox transcription factors. *J. Neurosci.* **37**, 8816–8829 (2017).
71. Brown, J. A. et al. Inhibition of parvalbumin-expressing interneurons results in complex behavioral changes. *Mol. psychiatry* **20**, 1499–1507 (2015).
72. Lewis, D. A., Curley, A. A., Glausier, J. R. & Volk, D. W. Cortical parvalbumin interneurons and cognitive dysfunction in schizophrenia. *Trends Neurosci.* **35**, 57–67 (2012).
73. Miyata, S. et al. Plasma corticosterone activates SGK1 and induces morphological changes in oligodendrocytes in corpus callosum. *PloS one* **6**, e19859 (2011).
74. Miyata, S. et al. Sgk1 regulates desmoglein 1 expression levels in oligodendrocytes in the mouse corpus callosum after chronic stress exposure. *Biochem. Biophys. Res. Commun.* **464**, 76–82 (2015).
75. Cathomas, F. et al. Oligodendrocyte gene expression is reduced by and influences effects of chronic social stress in mice. *Genes, Brain Behav.* **18**, e12475 (2019).
76. Sommeijer, J.-P. & Levelt, C. N. Synaptotagmin-2 is a reliable marker for parvalbumin positive inhibitory boutons in the mouse visual cortex. *PloS one* **7**, e35323 (2012).
77. Turecek, J. & Regehr, W. G. Neuronal regulation of fast synaptotagmin isoforms controls the relative contributions of synchronous and asynchronous release. *Neuron* **101**, 938–949.e4 (2019).
78. Lemoine, G. G., Scott-Boyer, M.-P., Ambroise, B., Périn, O. & Droit, A. GWENA: gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. *BMC Bioinforma.* **22**, 1–20 (2021).
79. Jansen, R.-P. RNA–cytoskeletal associations. *FASEB J.* **13**, 455–466 (1999).
80. Singer, R. H. The cytoskeleton and mRNA localization. *Curr. Opin. cell Biol.* **4**, 15–19 (1992).
81. Yan, X. & Han, J. gspan: Graph-based substructure pattern mining. in *2002 IEEE International Conference on Data Mining, 2002. Proceedings*. 721–724 (IEEE, 2002).
82. Okura, A. et al. SGK1 in Schwann cells is a potential molecular switch involved in axonal and glial regeneration during peripheral nerve injury. *Biochem. Biophys. Res. Commun.* **607**, 158–165 (2022).
83. King, R. H. et al. Ndrg1 in development and maintenance of the myelin sheath. *Neurobiol. Dis.* **42**, 368–380 (2011).
84. Dugas, J. C., Tai, Y. C., Speed, T. P., Ngai, J. & Barres, B. A. Functional genomic analysis of oligodendrocyte differentiation. *J. Neurosci.* **26**, 10967–10983 (2006).
85. Ziaei, A. et al. Ermin deficiency leads to compromised myelin, inflammatory milieu, and susceptibility to demyelinating insult. *Brain Pathol.* **32**, e13064 (2022).
86. Kuhn, S., Gritti, L., Crooks, D. & Dombrowski, Y. Oligodendrocytes in development, myelin generation and beyond. *Cells* **8**, 1424 (2019).
87. Hubbard, T. et al. The Ensembl genome database project. *Nucleic acids Res.* **30**, 38–41 (2002).
88. Kazemian, M., Zhu, Q., Halfon, M. S. & Sinha, S. Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. *Nucleic acids Res.* **39**, 9463–9472 (2011).
89. Moffitt, J. R. et al. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl Acad. Sci.* **113**, 14456–14461 (2016).
90. Emanuel, G., Eichhorn, S. & Zhuang, X. MERlin-Scalable and extensible MERFISH analysis software, v0. 1.6. *Zenodo* **10** (2020).
91. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. methods* **18**, 100–106 (2021).
92. Fei, J. et al. Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *J. cell Sci.* **130**, 4180–4192 (2017).
93. Jones, A. R., Overly, C. C. & Sunkin, S. M. The Allen brain atlas: 5 years and beyond. *Nat. Rev. Neurosci.* **10**, 821–828 (2009).
94. Kumar, A. & Sinha, S. Intracellular Spatial Transcriptomic Analysis Toolkit (InSTAnT). *Zenodo*, https://doi.org/10.5281/zenodo.10994621 (2024).

Acknowledgements
We thank Dr. Jeffrey Moffitt for sharing the data from Moffit et al.[35], Zijun Wu for assistance in formatting of figures, Alton S. Barbehenn for helpful discussion for the statistical analysis, Prof. Prasanth V. Kannanganattu for advice on sample preparation. Funding: This work was supported by the National Institutes of Health (R35GM131819 to S.S., R35GM147420 to H.-S.H and A.W.S, R21HG013180 to SDZ, and T32-842 GM136629 to M.A.), Johnson & Johnson (WiSTEM2D Award for Science to H.-S.H.), Cancer Center at Illinois (Seed grant to H.-S.H), and Georgia Institute of Technology (Wallace H. Coulter Distinguished Faculty Chair: S.S.) Facilities: We acknowledge Core Facilities at the Carl R. Woese Institute for Genomic Biology for their microscope and staff support.

Author contributions
A.K. and S.S. came up with the conceptual and mathematical formulation. A.K. implemented the algorithm and tested on U2OS and MERFISH Brain dataset. B.A. refactored the codebase and ran experiments on Xenium and SeqFISH datasets. A.E.B. performed RRI and GO analysis for Fig. 3g and Fig. 3h. A.W.S, M.A., J.L. and Y.S. generated the inhouse U2OS MERFISH dataset. A.K., S.D.Z., H.-S.H and S.S. worked together to refine analysis methods and scope, and to write the paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-49457-w.

Correspondence and requests for materials should be addressed to Sihai Dave Zhao, Hee-Sun Han or Saurabh Sinha.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.