Bento: A toolkit for subcellular analysis of spatial transcriptomics data

Authors: Clarence K. Mah[1,2]*, Noorsher Ahmed[2]*, Dylan Lam[2,3], Alexander Monell[4], Colin Kern[5], Yuanyuan Han[5], Anthony J. Cesnik[6], Emma Lundberg[6,7,8], Quan Zhu[5], Hannah Carter[1], Gene W. Yeo[2,9,10]**

1 Division of Medical Genetics, Department of Medicine, University of California San Diego, La Jolla, CA, USA

2 Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA

3 Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA

4 Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

5 Center for Epigenomics, University of California San Diego, La Jolla, CA, USA

6 Department of Bioengineering, Stanford University, Stanford, CA, USA

7 Department of Pathology, Stanford University, Stanford, CA, USA

8 Chan-Zuckerberg Biohub, San Francisco, CA, USA

9 Stem Cell Program, University of California San Diego, La Jolla, CA, USA

10 Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, 92093, USA

*Co-first authors/ These authors contributed equally
**Corresponding (geneyeo@ucsd.edu)

Abstract

Advances in spatial transcriptomics technologies produce RNA imaging data at increasingly higher throughput and scale. Current computational methods identify and measure the spatial relationships between cell-types, but do not leverage the spatial information of individual RNA molecules to reveal subcellular spatiotemporal dynamics of RNA processing. Here, we developed Bento, a computational framework for subcellular analysis of high-throughput spatial transcriptomics datasets. Bento handles single-molecule data generated by diverse spatial transcriptomics technologies and computes spatial statistics of subcellular RNA molecular distributions, compartmental expression, and cell morphology to build multidimensional feature sets for exploratory analysis. We also developed a multi-label ensemble model for generalizable classification of subcellular localization of every gene in every cell. To demonstrate Bento's utility, we applied it to analyze spatial transcriptomics datasets generated by seqFISH+ (10k genes in ~200 fibroblast cells) and MERFISH (130 genes quantified in ~1000 U2-OS cells) to understand the interplay between gene function, cellular morphology and RNA localization. To understand the role of RNA localization in RNA processing, we integrated spatial data with RNA binding protein (RBP) binding data to explore the spatiotemporal dynamics of RBP-RNA interactions at unprecedented scale (3,165 RNA species x 148 RBPs). We found RNA targets of individual RBPs to be enriched in specific subcellular compartments – such as Splicing Factor 3a Protein Complex (SF3A3), and that preferential localization is influenced by RBP binding of genomic regions for RBPs including Staufen homolog 2 (STAU2). Bento builds on the existing ecosystem of single-cell and spatial analysis toolkits, to ensure accessibility and community-driven tool development. We provide Bento as an open-source tool for the community to further expand our understanding of subcellular biology.

Introduction

Cells are the smallest organizational unit in living organisms, and how their internal components are organized is critical for homeostatic function. Protein-coding genes encoded in the genome are expressed as RNA and after maturation are translated into proteins. Whereas protein localization is well studied[1], and protein mislocalization is a known driver of disease[2,3], these same principles for RNA are less well appreciated. The spatiotemporal dynamics of RNA play a crucial role in localized cellular processes such as cell migration and cell division[4,5], as well as specialized cell functionalities like synaptic plasticity[6–8]. Mislocalization of RNA has been associated with diseases such as Huntington's disease (HD), where defects in axonal mRNA transport and subsequent translation in human spiny neurons lead to cell death and neurodegeneration[9–12].

The study of subcellular RNA localization necessitates single-molecule measurements. Since the development of single-molecule fluorescent *in situ* hybridization (smFISH), recent advances to develop multiplexed methods such as MERFISH[13], seqFISH+[14], HybISS[15], and Ex-Seq[16] have enabled RNA localization measurements at near transcriptome scales. A number of computational tools, such as Squidpy[17], stLearn[18], Giotto[17], and single-cell expression tools like Seurat[19] and Scanpy[20], have enabled characterization of tissue architecture, quantify cell-cell interactions, and identify spatial expression patterns. While these tools get at spatial variation in tissues, they lack the ability to investigate subcellular resolution, which is necessary to study the function of RNA localization in normal cell function and disease states.

Meanwhile, methods such as FISH-quant and FISHFactor have been developed to analyze fluorescent *in situ* hybridization (FISH)-based imaging data for subcellular patterns describing the spatial distribution of RNA species[21,22]. While these methods take advantage of subcellular-resolution spatial data, they cannot be applied to entire spatial transcriptomics datasets. More recently, the ClusterMap method was developed to identify cellular and subcellular regions from transcript locations alone, but only demonstrated the ability to identify cells and nuclei. There is clearly a need for a toolkit that scales to entire spatial transcriptomics datasets and has the flexibility to accommodate cutting-edge analysis approaches.

To address these shortcomings, we present Bento, a toolkit for exploring spatial transcriptomics data with an emphasis on subcellular biology. Bento ingests single-molecule resolution data from highly multiplexed spatial transcriptomics imaging experiments, enabling visualization, exploration and analysis of subcellular biology. The toolkit has an accessible programming interface (API) in Python, and opens the doors to testing hypotheses about subcellular phenotypes by providing methods to compute spatial statistics and measure spatial phenotypes for RNA localization, compartmental expression, and cell morphology to build multidimensional feature sets for exploratory analysis. These methods are inspired and adapted from existing work[21,23–25] and other scientific domains[26]. To demonstrate the utility of Bento, we developed a scalable method to quantify subcellular "localization signatures;" we define a localization signature as the weighted combination of localization patterns simultaneously associated with groups of genes and cells. We show that localization signatures link RNA subcellular localization to function. Bento can be used to further discover evidence of novel RNA interactions, and identify localization mechanisms of post-transcriptional gene regulation. These contributions are important for understanding RNA processing from transcription to translation and how dysregulated localization may be a primary mechanism of neurodegenerative diseases.

Results

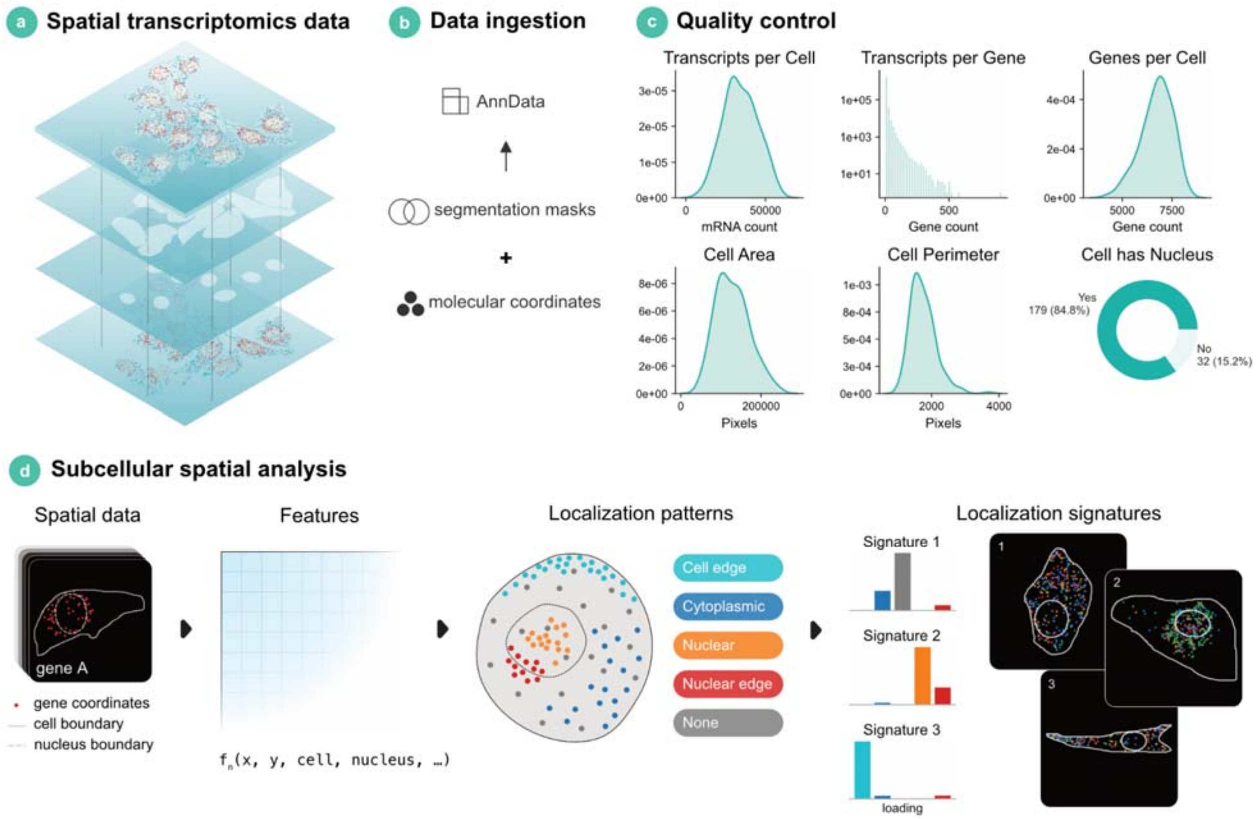

Fig 1. Workflow and functionality of Bento, subcellular spatial analysis toolkit. A) Molecular coordinates and segmentation masks from spatial transcriptomics data are required for analysis and visualization. **B)** Input data is stored in the AnnData data format, which can be manipulated with Bento as well as a wide ecosystem of single-cell omics tools. **C)** Quality control metrics are illustrated for the seqFISH+ dataset, where the top row shows transcript frequency distributions and the bottom row shows distributions of simple physical measures of cells. Visualization of distributions can be visually inspected to identify and remove outliers, which may indicate low quality cells. **D)** Bento has a standard interface to calculate subcellular features, identify subcellular localization patterns and infer localization signatures from spatial transcriptomics datasets.

Overview of Bento data infrastructure for subcellular analysis

Bento is an open-source Python toolkit for scalable analysis of multiplexed spatial transcriptomics data at subcellular spatial resolution. It utilizes computational geospatial tools (GeoPandas[27]) to enable spatial analysis of molecular imaging data, and data science tools including SciPy[28], Dask[29,30], PyTorch[31] and Tensorly[32] to enable scalable analysis of high-dimensional data. We build on the AnnData data format[33] to store both expression and spatial information, enabling integration of subcellular spatial analysis with the vast ecosystem of single-cell expression analysis.

In order to facilitate a flexible workflow, Bento is generally compatible with molecule-level resolution spatial transcriptomics data (**Fig. 1A**), such as datasets produced by MERFISH[13], seqFISH+[14], CosMx (NanoString)[34], Xenium (10x Genomics)[15,35], and Molecular Cartography (Resolve Biosciences)[36]. Bento's workflow takes as input 1) 2D spatial coordinates of transcripts annotated by gene, 2) cell segmentation boundaries and 3) nuclear segmentation boundaries (**Fig. 1B**). If available, Bento can also handle arbitrary sets of segmentations for other subcellular structures or regions of interest. These inputs are stored in the AnnData data format[33], which links cell and gene metadata to a standard count matrix, providing compatibility with standard single-cell RNA-seq quality control and analysis tools. Bento provides additional spatial metrics to augment quality control (**Fig. 1C**). With a data structure for segmentation boundaries and transcript coordinates in place, Bento can easily compute spatial statistics and measure spatial phenotypes for every gene in every cell to build flexible multidimensional feature sets for exploratory subcellular analysis (**Fig. 1D**).

To predict subcellular localization, Bento uses a multilabel classifier across five broad categories: nuclear, cytoplasmic, nuclear edge, cell edge, and none of the above (**Fig. 1D**). To reveal variation in subcellular localization patterns, we identify groups of cells and genes exhibiting similar RNA subcellular localization patterns via tensor decomposition. In addition, Bento supports quantifying enrichment in subcellular regions of interest (ROIs) defined by segmentation masks, enabling transcriptome-scale colocalization studies of RNA in organelles or other segmentable cellular structures[37].

Bento also tackles the challenge of visualizing datasets of hundreds to thousands of cells each with thousands of single RNA molecules with a suite of plotting options such as rendering single cells or fields of view, as well as individual molecules or rasterized representations. Furthermore, its use of the AnnData data format enables complementary spatial analysis at the tissue resolution with other tools in the ecosystem[17,19,38].

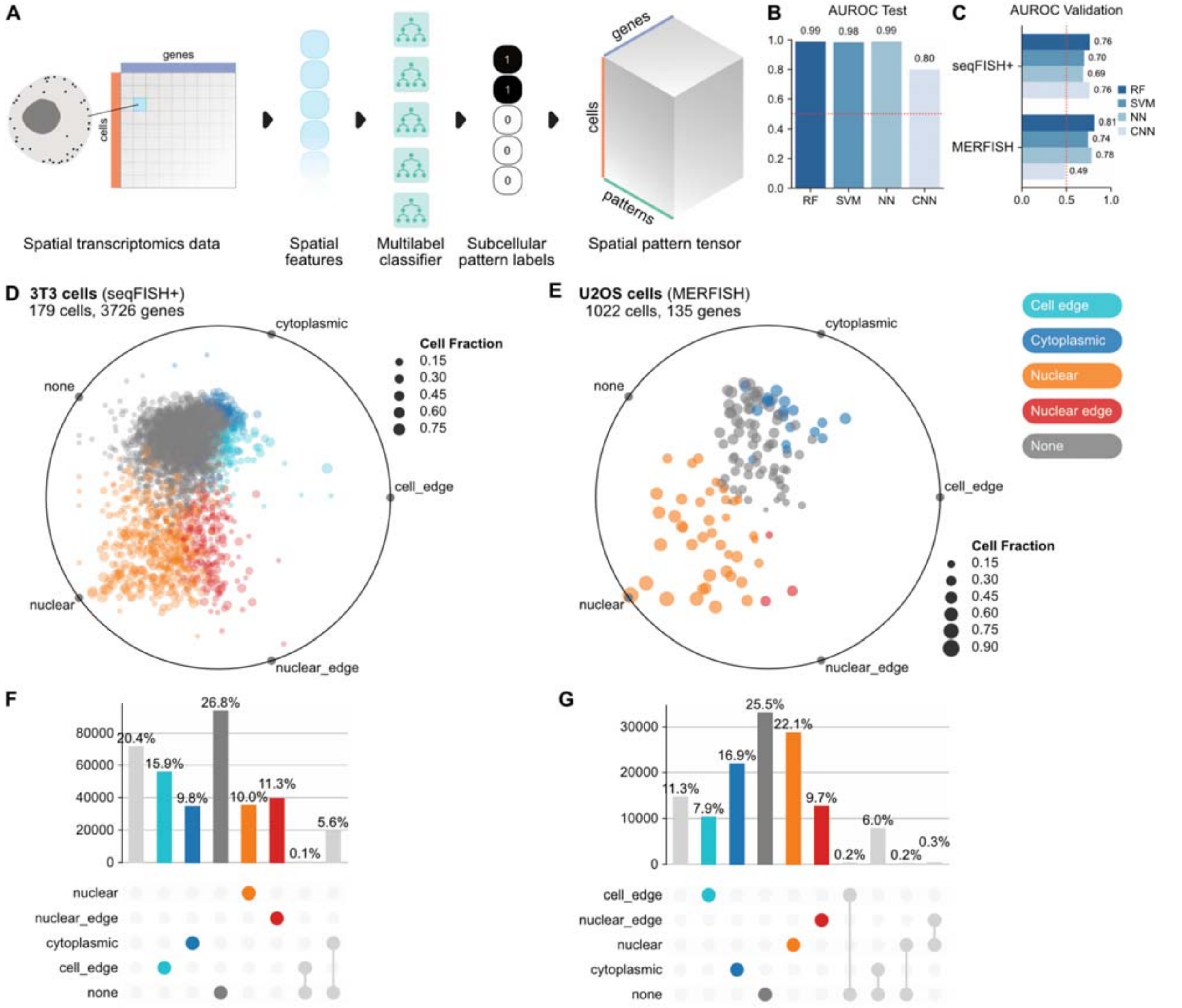

Fig 2. Scalable prediction of subcellular localization patterns in spatial transcriptomics data. A) A single sample shown as the set of transcripts for a gene in a particular cell. Spatial features are computed for all samples across every combination of cell and gene, which are used as input for a 5-class multilabel classifier. The output is represented as a 3-dimensional tensor. **B)** Performance of 4 models trained on simulated data and evaluated on

neural network, and CNN for convolutional neural network. **C)** Validation performance of the same 4 models on manually annotated seqFISH+ and MERFISH data; RadViz projection of genes for the **D)** seqFISH+ dataset and **E)** MERFISH dataset, where the point position denotes the balance between subcellular localization pattern frequencies, color denotes most frequent pattern, and size denotes cell fraction. Upset plot[39] showing relative proportions of all **F)** seqFISH+ samples and **G)** MERFISH samples classified across patterns. Light gray columns correspond to samples predicted as multiple patterns, denoted by connected dots under the bar graph.

Bento annotates subcellular localization patterns

We built a multilabel classifier using a set of binary random forest classifiers to assign labels to each gene in every cell across five categories: (i) nuclear (contained in the volume of the nucleus), (ii) cytoplasmic (diffuse throughout the cytoplasm), (iii) nuclear edge (near the inner/outer nuclear membrane), (iv) cell edge (near the cell membrane), and (v) none (complete spatial randomness). These categories are a consolidation of those observed in several high-throughput smFISH imaging experiments in HeLa cells[40–43]. We used the FISH-quant simulation framework to generate realistic ground-truth images using empirically derived parameters from the aforementioned high-throughput smFISH HeLa cell imaging experiments[42]. Each sample is defined as a set of points with coordinates in two dimensions, representing the set of observed transcripts for a gene in a particular cell. In total, we simulated 2,000 samples per class for a total of 10,000 samples (**Methods**). We used 80% of the simulated data for training and held out the remaining 20% for testing. Each sample is encoded by a set of 13 input features, describing characteristics of each sample's point distribution, including proximity to cellular compartments and extensions (features 1-3), measures of symmetry about a center of mass (features 4-6), and measures of dispersion and point density (feature 7-13) (**Fig. 2A, Supp. Table 1**).

We evaluated several base models for the multilabel classifier including random forests (RF), support vector machines (SVM), feed-forward fully-connected neural networks (NN), and convolutional neural networks (CNN) (**Methods, Supp. Table 2**). To evaluate their performance on simulated data, we compared the model macro area under receiving operator curve (macro-AUROC) values on the hold-out test data. All models except the CNN were able to generalize well to the simulated test data (**Fig. 2B**).

To evaluate the ability of each model to generalize to real world data in contrast to simulated data, we identified datasets with a large number of target genes or a large number of cells. We manually annotated a random subset of samples across these datasets for benchmarking model performance (**Methods**). The seqFISH+ dataset met the first criteria, targeting 10,000 genes across 227 cultured NIH/3T3 mouse embryonic fibroblast cells spatially profiled with seqFISH+[14]. We also generated a MERFISH dataset capturing 2,716 cultured U2-OS osteosarcoma cells, profiling 130 genes and 5 non-targeting. The RF classifier consistently had the best agreement with manual annotations on a held-out test set, with a macro-AUROC of 0.76 and 0.81 in the seqFISH+ and MERFISH datasets respectively, in contrast to a macro-AUROC of 0.99 in the held-out test data (**Fig. 2C**).

The multilabel classifier was used to assign one or more labels to every cell-gene pair in each dataset. We first preprocessed datasets to remove genes with low expression and cells without annotated nuclei (**Methods**). Genes commonly showed a wide range of variability in localization across cells (**Fig. 2D&E**). Of the localization patterns besides "none", "cell edge" was the most common (15.9%) in the 3T3 fibroblast cells while "nuclear" was the most common (22.1%) in the U2-OS osteosarcoma cells (**Fig. 2F&G**). We compared 63 orthologous genes measured in both cell types and found that 28 genes shared the same most frequent localization pattern and the rest had different localizations, possibly reflecting cell type specific localization behavior (**Methods, Supp. Fig 1**). With the ability to annotate localization patterns for entire spatial transcriptomics datasets, we can now interrogate how pattern frequencies vary within a gene across cells and between genes for the entire transcriptome.

Tensor decomposition identifies subcellular localization signatures

Just as transcription shows cell-to-cell variation in expression, even within the same cell type, the subcellular spatial distribution of RNA molecules is not deterministic. To understand how RNA subcellular localization varies across cells and between genes simultaneously, we employed tensor decomposition — specifically, non-negative parallel factor analysis — a data-driven, unsupervised approach for discovering substructure in high-dimensional data[32,44]. We used tensor decomposition to decompose spatial transcriptomics datasets into a set of “localization signatures'' representing combinations of cells and genes that share similar localization behavior. First, we represent a single spatial transcriptomics dataset as a third-order tensor of size **P**x**C**x**G** where **P**, **C**, and **G** represent the number of patterns, cells, and genes, respectively, in our data. In this approach, **P** = 5 for the 5 subcellular localization patterns (i.e., as defined previously, this is nuclear, cytoplasmic, nuclear edge, cell edge, and none) while **C** and **G** represent the number of genes and cells measured in the dataset. Missing and low-expression data are removed and ignored in downstream calculations that utilize the dataset tensor. This allows the tensor to hold the output of the spatial localization classifier as a five-digit binary vector for every sample (**Methods**).

Tensor decomposition factors the dataset tensor into **k** localization signatures using the elbow method heuristic to determine the least number of signatures needed to reconstruct the original tensor, optimizing for the mean squared error reconstruction loss function. Missing values are ignored when calculating the loss. Unlike matrix dimensionality reduction methods, such as PCA, the order of the components (signatures) is unassociated with the amount of variance explained. Each of the **k** signatures resulting from tensor decomposition is composed of 3 loading vectors, corresponding to the pattern, cell, and gene dimensions. Higher values denote a stronger association with that signature. We interpret a particular signature’s pattern loading as the weighted combination of patterns characterizing that signature. Similarly, a signature’s cell loading and gene loading denote the strength of association of cells and genes respectively. The localization signatures derived from tensor decomposition are not necessarily mutually exclusive and can share overlapping sets of patterns, cells and genes.

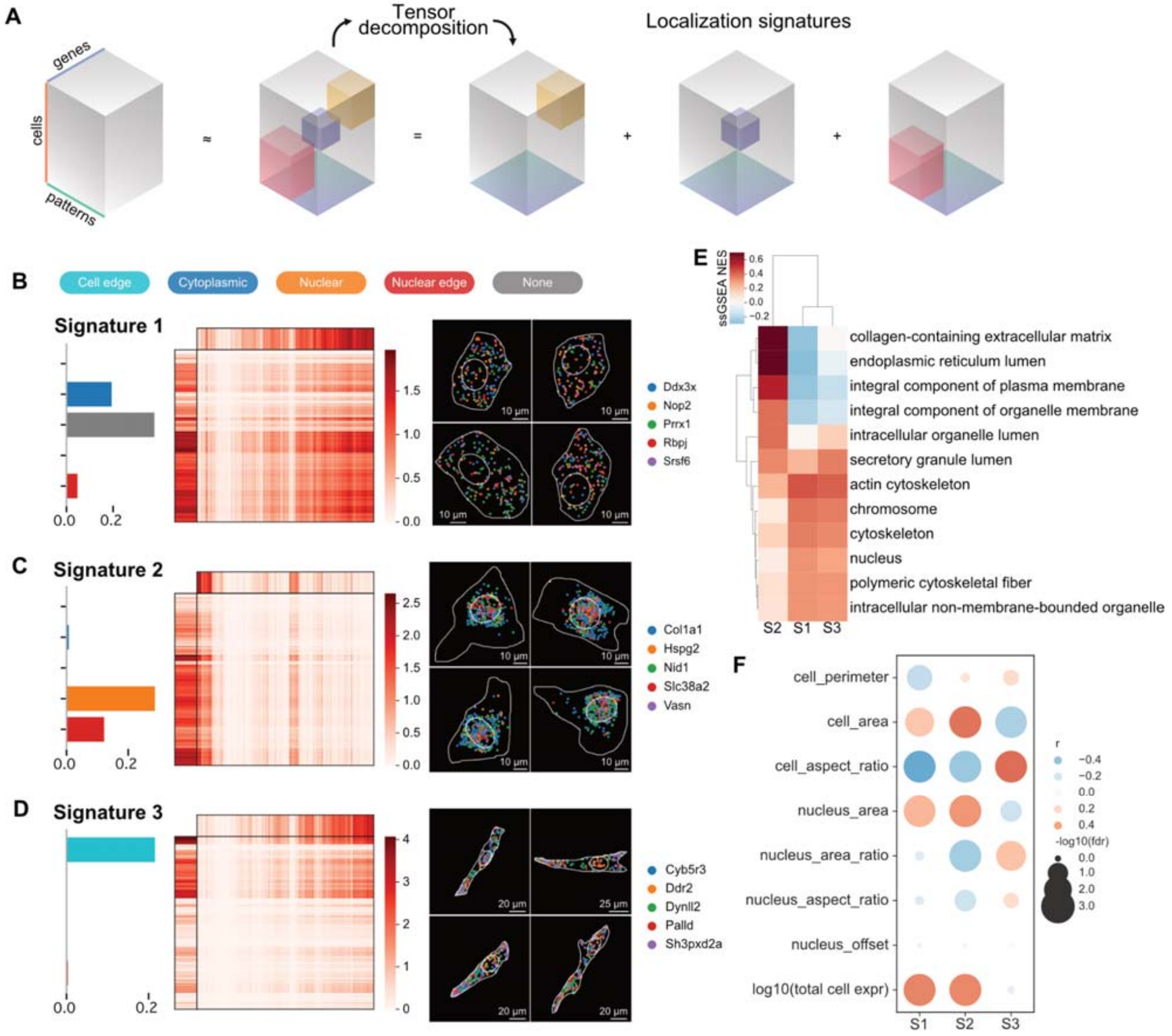

Fig 3. Identification of seqFISH+ subcellular localization signatures via tensor decomposition. A) Schematic of the tensor decomposition procedure applied to the dataset tensor, producing localization signatures. **B-D)** Localization signatures of the seqFISH+ dataset. In each row, the bar plots show associations of localization pattern loadings with the respective signatures colored by pattern class. Heatmaps show log-scaled cell and gene loadings along the top and left while associated cell-gene pairs are shown in the center heatmap as a dot product of the log-scaled cell and gene loadings. The top 4 genes are shown in the top-ranked cell for each signature under the heatmaps. **E)** ssGSEA scores for GO Ontology (Cellular Component) terms calculated using gene loadings of each signature. Heatmap shows the top 5 terms for each signature aggregated and hierarchically clustered. **F)** Dot plot shows correlation of signature cell loadings with cell morphology measures. Larger dots denote higher significance while color saturation indicates a stronger magnitude of correlation.

Functional characterization of localization signatures

We applied the tensor decomposition strategy to the seqFISH+ dataset and found 3 distinct signatures (**Fig. 3A**). The first signature's pattern loading is dominated by "none" and "cytoplasmic", while signature 2 is a combination of "nuclear" and "nuclear edge" and signature 3 is primarily "cell edge" (**Fig. 3B-D**). These signatures recapitulate patterns found in the original seqFISH+ study, in which three major clusters of spatially co-occurring genes were observed and manually annotated as protrusion, nuclear/perinuclear, and cytoplasmic[14]. This demonstrates the ability of tensor decomposition to extract meaningful biological structure from localization patterns in a data-driven manner. We also applied the tensor decomposition results suggest that subcellular localization is systematically regulated regardless of cell type and that identifying localization signatures is robust to measurement platforms.

To determine if gene signatures are associated with specific subcellular compartments corresponding to their localization preferences, we used single-sample Gene Set Enrichment Analysis (ssGSEA) to identify enriched terms from the GO Cellular Component domain[45–48] (**Fig. 3E, Methods**). For example, we found that organelles involved in cotranslational translocation e.g. the endoplasmic reticulum, golgi apparatus and the extracellular matrix (ECM), characterize gene signature 2, which was dominated by the nuclear and nuclear edge patterns. Additionally, we found gene signature 3 (cell edge) was enriched for cytoskeleton components. While RNA localization patterns in these fibroblasts cells reflected stereotypical organelle organization, we found that localization also reflected cell type-specific function. Secreted vesicle terms were associated with gene signatures 1 and 3 and ECM terms were associated with gene signature 2, reflecting the primary function of mouse embryonic fibroblasts to synthesize and secrete ECM proteins. These results show concordance of RNA localization to their protein counterparts regarding subcellular organelles and cellular function.

We next characterized the morphological features of cells with different localization signatures. To do so, we compared cell signatures to spatial measures of cell morphology, nucleus morphology and total cell expression (**Fig. 3F, Methods**). Cell and nucleus morphology measures include perimeter, area, aspect ratio (i.e. minor to major axis ratio), the ratio of nucleus to cell area, and the distance between nucleus and cell centroids. These measures correlated only moderately to weakly with localization signatures (Spearman correlation analysis); all correlation coefficient magnitudes were less than or equal to 0.4. Cell signature 2 was positively correlated with cell area and nuclear area while cell signature 3 was positively correlated with elongated cells. Conversely, cell signature 1 was negatively correlated with elongated cells, preferring more evenly shaped cells. Visually inspecting the top scoring cells of each signature confirms these associations (**Fig. 3B-D**). In conclusion, we found evidence that cell morphology is a source of heterogeneity in localization signatures but is insufficient to fully explain it. These results show that our tensor decomposition approach can systematically identify biological factors explaining RNA localization patterns considering cells and genes simultaneously.

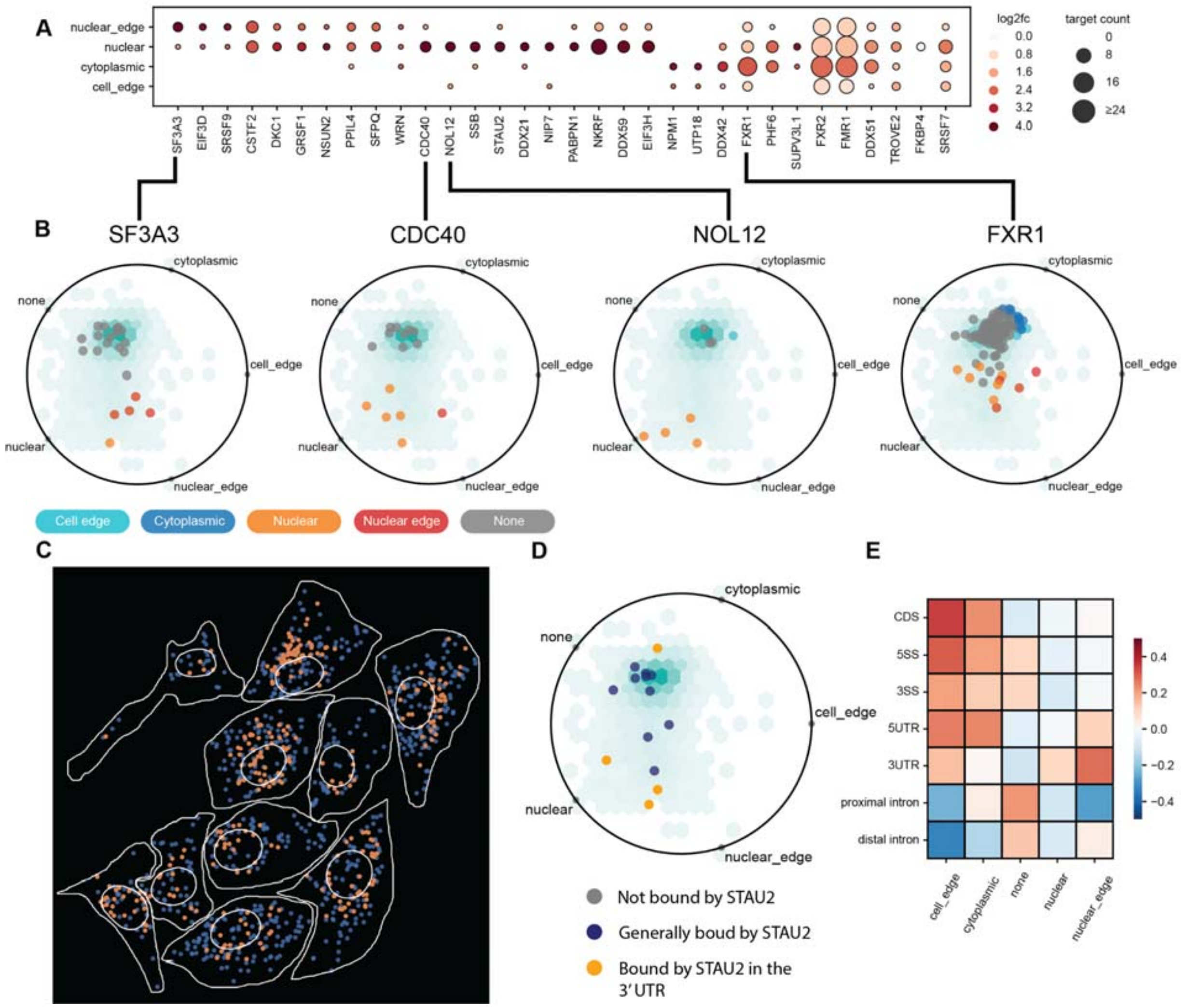

Fig 4. Complex RBP-RNA interactions drive RNA subcellular localization. A) Distribution of RNA targets across localization patterns for RBPs with significant eCLIP-seq binding peaks (RBP-Bind) overlapping genomic regions of 3,165 genes measured by seqFISH+. The top 10 enriched RBPs (sorted by log2 fold-change) for each localization pattern are shown. Color denotes log2 fold-change of one localization category versus mean of rest. Size denotes RBP target count capped at 24 for clarity. **B)** All RNA binding partners of RBPs SF3A3, CDC40, NOL12 and FXR1 are shown as points and colored by their most frequent localization pattern. RNA targets are enriched for different localization patterns, reflecting RBP function. **C)** Visualization of single molecules of STAU2 RNA targets based on binding to their 3' UTR or other region in a field of view. **D)** STAU2 targets show preferential localization throughout the 3T3 dataset based on 3' UTR binding. **E)** Correlations between RBP binding to genomic regions and localization patterns of RNA targets.

RNA binding protein interactions with RNA influence subcellular localization

RNA subcellular localization is governed by an elaborate choreography of RNA binding, transport, and degradation mechanisms that all depend on RNA binding proteins (RBPs). However, the complexity of multiple binding partners and promiscuous RBPs has made it difficult to discern exactly which RBPs and which sequence features govern the rules of RNA subcellular transport and localization[49]. With our ability to annotate RNA subcellular localization with Bento, we can systematically explore the complex RBP-RNA relationships that influence RNA localization.

We sought to identify RBP-RNA interactions associated with subcellular localization patterns in the seqFISH+ 3T3 fibroblast cell line dataset as previously identified with Bento. For RBP-RNA interactions, we used publicly available cross-linking immunoprecipitation high-throughput sequencing (CLIP-seq) data from subcellular localization pattern annotations, we were able to quantify the subcellular enrichment of RNA targets for each RBP across genomic regions. We observe a number of RBPs preferentially binding to targets in particular localization patterns (**Fig 4A**). This preferential binding reflects RBP function. For example, Splicing Factor 3a Protein Complex (SF3A3), a component of the U2 snRNP splicing machinery[51], and Cell Division Cycle 40 (CDC40), another RBP involved in RNA splicing[52], both bind transcripts in the nucleus and nuclear edge (**Fig 4B**). Nucleolar Protein 12 (NOL12) is a multifunctional RBP regulating RNA metabolism in nucleoli, nucleoplasm, paraspeckles, as well as GW/P-bodies in the cytoplasm[53]. Mirroring the multifaceted nature of its function, NOL12 targets localize to the nucleus, nuclear edge, as well as the cell edge (**Fig 4B**). Fragile X Mental Retardation, Autosomal Homolog 1 (FXR1) is an RBP that shuttles between the nucleus and cytoplasm and associates with polyribosomes[54]. Reflecting the nuclear-cytoplasmic shuttling, FXR1 targets are present throughout the cell.

While RBP binding state can inform RBP-RNA spatiotemporal associations, the location in the target RNA's sequence an RBP binds can also influence the RBP-RNA functional relationship. We explore this sequence-feature centered relationship in Staufen homolog 2 (STAU2), an RBP involved in the transport and localization of mRNAs to subcellular compartments and organelles[55]. In neurons, STAU2 is a marker for dendritic transport of ribonucleoprotein particles (RNPs) transporting target transcripts by 3'UTR binding[56]. In retinal ganglion cells (RGCs), STAU2 has been demonstrated to accumulate asymmetrically during prophase and metaphase stages of mitosis during corticogenesis, driving asymmetry in mRNA localization and leading to cell fate differentiation of daughter cells[57]. Furthermore, proteins in the larger Staufen family are found in ribosome and endoplasmic reticulum (ER) containing granules[58]. For the transcripts in the seqFISH+ 3T3 fibroblast dataset, STAU2 binds to a variety of regions for a number of genes, resulting in no discernable spatial relationship (**Fig 4C&D**). However, by highlighting transcripts that are only bound by STAU2 in their 3' UTRs, spatial relationships emerge that reflect the prior understandings of STAU2 spatiotemporal behavior. 3' UTR-bound transcripts localize preferentially to the nucleus and nuclear edge, a localization characteristic previously demonstrated for HSPA5 and other ER-enriched RNAs as measured by MERFISH[59] (**Fig 4C&D**). Furthermore, an asymmetric polarization in target localization can be observed in a majority of cells across the nuclear membrane. To test if RBP binding to specific regions in mRNA more broadly affects RNA localization, we looked at the enrichment of region-specific binding events in targets localized to each subcellular region (**Fig 4E**). While patterns emerge, such as an enrichment of CDS binding events in transcripts localized to the cell edge or 3' UTR binding events and nuclear edge localization, the correlations are weak. This reflects the complexity of factors driving RNA localization and RBP-RNA spatiotemporal relationships, such as compartment-specific localization and function of RBPs and their interactors[60].

We conclude that Bento enables exploration of links between RBP-RNA interactions and RNA localization. Using Bento's ability to distinguish subcellular localization patterns and visualization capabilities, we enable exploration of how RBPs and the sequence features they bind to preferentially localize transcripts.

Discussion

Bento seeks to interrogate subcellular biology via its "subcellular first" approach to spatial analysis, complementary to "cell-type or tissue first" spatial analysis methods. It is a platform to explore the spatial relationships between the transcriptome and subcellular structures. Bento implements a classification strategy to label five RNA subcellular localization patterns relative to the cell membrane and nuclear membrane. Bento also enables discovery of biological structure with an unsupervised approach for decomposing high-dimensional spatial feature sets. Finally, Bento provides tools to study enrichment of RNA vis-a-vis subcellular location. We demonstrate its versatility by successfully applying it to two publicly available datasets despite differences in technology, detection efficiency, and upstream image processing. While Bento can be utilized for any subcellular resolution spatial transcriptomic datasets, image segmentation will be a significant factor for studying subcellular localization, especially in noisier contexts such as tissue. Nevertheless, image segmentation is an active area of computational research and improving continuously. With that in mind, Bento can be applied to reveal transcript localization patterns across datasets including spatial tissue atlases, maps of the tumor immune microenvironment, developmental systems, and more.

Notably, we found that the CNN-based model had difficulty generalizing across datasets; we speculate that its learned features are highly dependent on detection efficiency. Because many more molecules were detected per gene in the MERFISH dataset, the input images had a different distribution of pixel intensities compared to the seqFISH+ input images. In contrast, the other models mitigated this effect using features that aggregate across the set of molecules rather than encoding their presence or absence. As a result, we chose to use the random forest-based model for our task. Our adaptation reflects how the relative sparsity and presence/absence nature of multiplexed transcriptomics data poses a challenge to pixel-based analysis approaches, in comparison to their success in analyzing proteomics imaging data [61]. We anticipate unsupervised deep learning methods will be useful for spatial transcriptomics as datasets with sufficient data become available.

As the major role of coding RNA is as an intermediary information molecule for proteins, we expected RNA localization to correspond to that of their functional protein counterparts. In our analysis, we found that localization signatures are dominated by genes informative of subcellular compartments and cell-type specific function. Additionally, we found the localization signatures grouped cells with similar morphology suggesting early links between RNA subcellular localization and cell state. We expect there are many more sources of heterogeneity that play a role in subcellular localization and should be interrogated, including cell type effects, cell cycle effects, noncoding RNA interactions and more.

At the center of RNA processing lie multifaceted spatiotemporal RBP-RNA interactions which we explore at unprecedented scale leveraging Bento's classification and visualization capabilities. Many models have been proposed involving RBP interactions with target “zipcode” sequences that drive RNA transport and localized RNA stability and result in subcellular RNA patterning[49,62]. While specific instances of RBPs and sequence features have been demonstrated to influence RNA localization, a global understanding of RNA localization remains elusive. We explored the subcellular localization phenotypes of the RNA targets of 148 RBPs and found most RBPs to preferentially bind to RNA targets in specific subcellular compartments. Furthermore, the subcellular localization of RNA targets reflects well characterized RBP function, of RBPs such as RNA splicing, RNA transport, and cell cycle. We also demonstrated that RNA targets of STAU2 preferentially localize dependent on 3’ UTR binding. When examining the relationship across the whole dataset between genomic regions of RBP binding and RNA subcellular localization, we found modest correlations to exist between specific sequence features and subcellular compartments. The weakness of the correlations is partially attributed to the lack of isoform-specificity in the seqFISH+ and MERFISH data. Prior studies using APEX-labeling approaches to RNA subcellular localization found splicing to be an important driver of RNA localization[63,64]. The ability to resolve isoforms in spatial transcriptomics data will be necessary to push forward our understanding of the mechanisms governing RBP-RNA spatiotemporal dynamics.

In summary, we developed Bento, a toolkit for analyzing subcellular RNA organization from spatial transcriptomics data. We demonstrate that Bento identifies biologically relevant localization signatures leveraging spatial relationships between molecules and subcellular segmentation boundaries, scaling to entire datasets by utilizing open-source scientific computing tools. Bento fills the need for accessible subcellular spatial analysis in the existing ecosystem of single-cell and spatial analysis toolkits. We provide Bento as an open-source tool for the community to further expand our understanding of subcellular biology.

Methods

Simulating subcellular RNA localization patterns

We trained a multilabel classifier to assign each gene in every cell labels from five categories: (i) nuclear (contained in the volume of the nucleus), (ii) cytoplasmic (diffuse throughout the cytoplasm), (iii) nuclear edge (near the inner/outer nuclear membrane), (iv) cell edge (near the cell membrane), and (v) none (complete spatial randomness). These categories are a consolidation of those observed in several high-throughput smFISH imaging experiments in HeLa cells [40–43]. We used the FISH-quant simulation framework to generate realistic ground-truth images using empirically derived parameters from the mentioned high-throughput smFISH imaging experiments in HeLa cells [42]. In total, we simulate 2,000 samples per class for a total of 10,000 training samples.

1. **Cell shape**: Cell morphology varies widely across cell types and for classifier generalizability, it is important to include many different morphologies in the training set. We use a catalog of cell shapes for over 300 cells from smFISH images in HeLa cells that captures nucleus and cell membrane shape [42]. Cell shapes were obtained by cell segmentation with CellMask and nuclear segmentation was obtained from DAPI staining.
2. **mRNA abundance**: We simulated mRNA abundance at three different expression levels (40, 100, and 200 mRNA per average sized cell) with a Poisson noise term. Consequently, total mRNA abundance per cell was between 5 and 300 transcripts.
3. **Localization pattern**: We focused on 5 possible 2D localization patterns, including cell edge, cytoplasmic, none, nuclear, and nuclear edge. Each pattern was further evaluated at 3 different degrees - weak, moderate, and strong. Moderate corresponds to a pattern typically observed in a cell, whereas weak is close to spatially random. These 5 classes aim to capture biologically relevant behavior generalizable to most cell types; there is room for additional classes describing other biologically relevant localization patterns so long as they can be accurately modeled.

A model for predicting subcellular localization

We evaluated 4 base models for the multilabel classifier including random forests (RF), support vector machines (SVM), feed-forward fully-connected neural networks (NN), and convolutional neural networks (CNN). Each multilabel classifier consists of 5 binary classifiers with the same base model. We used the labeled 10,000 simulated samples for training, stratifying 80% of the simulated data for training and holding out the remaining 20% for testing. To select the best hyperparameters for each multilabel classifier, we sampled from a fixed hyperparameter space with the Tree-structured Parzen Estimator algorithm, and evaluated performance with 5-fold cross validation (**Supp. Table 2**). We retrained the final model (random forest base model) on all training data with the best performing set of hyperparameters.

Manually annotated validation data

Using 3 individual annotators, we annotated the same 600 samples across both datasets, keeping samples with 2 or more annotator agreements as true annotations, resulting in 165 annotated seqFISH+ samples and 238 annotated MERFISH samples (403 total). We used Cohen's kappa coefficient[65] to calculate agreement between pairs of annotators for each label, yielding an overall coefficient of 0.602. We found that pairwise agreement between annotators across labels was fairly consistent ranging between 0.588 and 0.628, while label-specific agreement varied more, ranging between 0.45 and 0.72 (**Supplementary Table 3**).

Comparison of localization patterns across datasets

To compare the subcellular localization of genes between the MERFISH dataset and the seqFISH+ dataset, we first found the set of orthologous genes measured in both datasets with MyGene.info (http://mygene.info). The MERFISH genes were first matched to their mouse orthologs, since the U2-OS cell line is human while the seqFISH+ 3T3 cell line is mouse. 63 genes were present in both datasets. Each gene was assigned to its most frequent pattern within each dataset. We found 28 of the 63 genes were found to share the same label across both datasets while 35 genes had different patterns.

Data preprocessing and filtering

For the seqFISH+ dataset, we limited the scope of our analysis to the set of genes for which at least 10 molecules were detected in at least one cell. This helped reduce sparsity in the data, resulting in 3726 genes remaining. Because pattern classification requires nuclear segmentation masks, we removed all cells lacking annotated nuclei for a remainder of 179 cells. Because the MERFISH data had a much higher number of molecules detected per gene, no gene filtering was performed. Again, cells without annotated nuclei were removed, leaving 1022 cells for pattern analysis.

Tensor decomposition for identifying localization signatures

To understand the heterogeneity in subcellular localization across the transcriptome, we employed non-negative parallel factor analysis, which seeks to represent our dataset tensor $\mathbf{X}$ in a lower dimensional space of $\mathbf{R}$ signatures by decomposing $\mathbf{X}$ as the sum of $\mathbf{R}$ rank-one 3-way tensors. Each of these tensors is described as the outer product of 3 vectors, x_c^r, y_g^r and z_p^r. The collection of vectors across **R** signatures we denote as x^r (cell loadings), y^r (gene loadings) and z^r (pattern loadings) respectively. We find the optimal rank-$\mathbf{R}$ decomposition of $\mathbf{X}$ by minimizing reconstruction error as a function of the number of signatures $\mathbf{R}$ and use the elbow function heuristic to choose the best-fit across the range of 2-10 factors.

$$X_{cgp} = \sum_{r=1}^{R} x_c^r y_g^r z_p^r$$

Characterizing subcellular compartments associated with localization signatures

To identify subcellular compartments associated with localization signatures, we performed single-sample Gene Set Enrichment Analysis (ssGSEA) on each signature's gene loadings to compute normalized enrichment scores. ssGSEA was performed with the GSEApy Python package and the "GO_Cellular_Component_2021" gene set library curated by Enrichr. For the seqFISH+ dataset, gene sets with a minimum size of 50 and a maximum size of 500 were analyzed. For the MERFISH dataset, gene sets with a minimum size of 15 and maximum size of 500 were analyzed to account for fewer genes measured.

Cell morphology associated with localization signatures

We investigated the relationship of localization signatures with cell morphology by correlating cell loadings with spatial measures. These spatial measures were calculated by converting the cell's segmentation mask and nuclear segmentation mask to closed polygons with GeoPandas. By treating each mask as separate polygons, GeoPandas was used to calculate cell perimeter, cell area and cell aspect ratio (major to minor axis ratio) for each cell shape. It was also used to calculate nucleus area, the ratio of nucleus to cell area, nucleus aspect ratio, and absolute distance from the nucleus centroid to the cell centroid. Finally, pairwise Spearman correlations and p-values were calculated between these measures and each signature's cell loadings. P-values were adjusted for multiple-hypothesis testing with the Holm-Bonferroni method to measure significance.

RBP-RNA interaction analyses

Cross-linking immunoprecipitation high-throughput sequencing (CLIP-seq) data from the ENCODE database was accessed using RBP-Bind. High confidence reproducible peaks between replicates (IDR peaks) showing a fold enrichment greater than or equal to 8 and with a p-value of less than 0.0001 across two replicates in HepG2 cells across all RBPs were gathered for each gene in the 3T3 seqFISH+ dataset.

Genes with no IDR peaks were removed from the analysis. In total, 3,165 genes had IDR peaks across 148 different RBPs (**Supplementary Table 4**).

We then calculated correlations between localization pattern proportions and sequence regions proportions. The 7 sequence regions include coding sequence, 5-prime splice site, 3-prime splice site, 5-prime untranslated region (UTR), 3-prime UTR, proximal intron and distal intron as defined in the RBP-Bind database. For localization pattern proportions, we first labeled each gene with its most common localization pattern in the seqFISH+ dataset. Then for each RBP, we counted the number of target genes per pattern and normalized counts by the number of target genes. This yielded pattern proportions, summing to 1 for each RBP. To get sequence region proportions, we counted the number of target genes bound by each RBP within each sequence region. Similarly, these values were normalized by the number of target genes yielding sequence region proportions summing to 1 for each RBP. With these RBP-specific binding proportions, we performed pairwise correlations between the 5 localization pattern proportions and the 7 sequence region proportions.

MERFISH of U2-OS cells

MERFISH sample preparation. MERFISH measurements of 130 genes with five non-targeting blank controls were done as previously described, using the published encoding[66] and readout probes[67]. Briefly, U2-OS cells were cultured on 40 mm #1.5 coverslips that are silanized and poly-L-lysine coated[66] and subsequently fixed in 4% (vol/vol) paraformaldehyde in 1x PBS for 15 minutes at room temperature. Cells were then permeabilized in 0.5% Triton X-100 for 10 minutes at room temperature and washed in 1x PBS containing Murine RNase Inhibitor (NEB M0314S). Cells were preincubated with a hybridization wash buffer (30% (vol/vol) formamide in 2x SSC) for ten minutes at room temperature with gentle shaking. After preincubation, the coverslip was moved to a fresh 60 mm petri dish and residual hybridization wash buffer was removed with a Kimwipe lab tissue. In the new dish, 50 uL of encoding probe hybridization buffer (2X SSC, 30% (vol/vol) formamide, 10% (wt/vol) dextran sulfate, 1 mg ml^{-1} yeast tRNA, and a total concentration of 5 uM encoding probes and 1 μM of anchor probe: a 15-nt sequence of alternating dT and thymidine-locked nucleic acid (dT+) with a 5′-acrydite modification (Integrated DNA Technologies). The sample was placed in a humidified 37C oven for 36 to 48 hours then washed with 30% (vol/vol) formamide in 2X SSC for 20 minutes at 37C, 20 minutes at room temperature. Samples were post-fixed with 4% (vol/vol) paraformaldehyde in 2X SSC and washed with 2X SSC with murine RNase inhibitor for five minutes. The samples were finally stained with an Alexa 488-conjugated anchor probe-readout oligo (Integrated DNA Technologies) and DAPI solution at 1 μg/ml.

MERFISH imaging. MERFISH measurements were conducted on a home-built system as described in Huang et al. 2021[67].

MERFISH analysis. Individual RNA molecules were decoded in MERFISH images using MERlin v0.1.6[68]. Images were aligned across hybridization rounds by maximizing phase cross-correlation on the fiducial bead channel to adjust for drift in the position of the stage from round to round. Background was reduced by applying a high-pass filter and decoding was then performed per-pixel. For each pixel, a vector was constructed of the 16 brightness values from each of the 16 rounds of imaging. These vectors were then L2 normalized and their euclidean distances to each of the L2 normalized barcodes from the MERFISH codebook was calculated. Pixels were assigned to the gene whose barcode they were closest to, unless the closest distance was greater than 0.512, in which case the pixel was not assigned a gene. Adjacent pixels assigned to the same gene were combined into a single RNA molecule. Molecules were filtered to remove potential false positives by comparing the mean brightness, pixel size, and distance to the closest barcode of molecules assigned to blank barcodes to those assigned to genes to achieve an estimated misidentification rate of 5%. The exact position of each molecule was calculated as the median position of all pixels consisting of the molecule.

Cellpose v1.0.2[69] was used to perform image segmentation to determine the boundaries of cells and nuclei. The nuclei boundaries were determined by running Cellpose with the ‘nuclei’ model using default parameters on the DAPI stain channel of the pre-hybridization images. Cytoplasm boundaries were segmented with the 'cyto' model and default parameters using the polyT stain channel. RNA molecules identified by MERlin were assigned to cells and nuclei by applying these segmentation masks to the positions of the molecules.

Data Availability

Preprocessed datasets have been deposited at https://doi.org/10.6084/m9.figshare.15109236.v2 and are accessible through the Bento Python package. These include the seqFISH+[14] and the generated MERFISH datasets. Raw MERFISH data is available upon request.

Code Availability

The source code for Bento is available on the GitHub repository: https://github.com/ckmah/bento-tools. Documentation for Bento can be found here: http://bento-tools.readthedocs.io/.

Acknowledgements

C.K.M. is supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE-2038238). N.A. was partially supported by NIH Training Grant T32 GM008666. This work was partially supported by National Institutes of Health grants NS103172, MH107367, AI132122, AI123202, AG069098, HG004659, and HG009889 to G.W.Y. G.W.Y. is also supported by an Allen Distinguished Investigator Award, a Paul G. Allen Frontiers Group advised grant from the Paul G. Allen Family Foundation. A.J.C. and E.L. acknowledge support from the Chan Zuckerberg Initiative (CZF2019-002448) and the Knut and Alice Wallenberg Foundation (KAW 2021.0346) to E.L. We thank members of the Yeo lab, Carter lab, Michelle Franc Ragsac, Erick Armingol, and Nate Lewis for helpful discussions and feedback on the manuscript.

Author Contributions

C.K.M, N.A., and G.W.Y. conceptualized the project. C.K.M. and N.A. co-developed the software. C.K.M. and D.L. trained the classification model for subcellular localization. C.K.M., N.A., and D.L. manually annotated data for benchmarking model performance. C.K.M. implemented the tensor decomposition. N.A. conducted the multimodal RBP-RNA localization analysis. A.M., C.K., Y.H., and Q.Z. generated the MERFISH dataset. A.C. and E.L. aided multimodal spatial analyses. C.K.M., N.A., H.C., and G.W.Y. wrote the manuscript. H.C. and G.W.Y supervised the project.

Competing Interests

G.W.Y. is a co-founder, member of the board of directors, equity holder, and paid consultant for Locanabio and Eclipse Bioinnovations, and a Scientific Adviser and paid consultant to Jumpcode Genomics. G.W.Y. is a Distinguished Visiting Professor at the National University of Singapore. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict-of-interest policies. The authors declare no other competing interests.

References

1. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, (2017).

2. Laurila, K. & Vihinen, M. Prediction of disease-related mutations affecting protein localization. *BMC Genomics* **10**, 122 (2009).

3. Park, S. *et al.* Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Mol. Syst. Biol.* **7**, 494 (2011).

4. Chin, A. & Lécuyer, E. RNA localization: Making its way to the center stage. *Biochim. Biophys. Acta Gen. Subj.* **1861**, 2956–2970 (2017).

5. Bovaird, S., Patel, D., Padilla, J.-C. A. & Lécuyer, E. Biological functions, regulatory mechanisms, and disease relevance of RNA localization pathways. *FEBS Lett.* **592**, 2948–2972 (2018).

6. Das, S., Singer, R. H. & Yoon, Y. J. The travels of mRNAs in neurons: do they know where they are going? *Curr. Opin. Neurobiol.* **57**, 110–116 (2019).

7. Sahoo, P. K., Smith, D. S., Perrone-Bizzozero, N. & Twiss, J. L. Axonal mRNA transport and translation at a glance. *J. Cell Sci.* **131**, (2018).

8. von Kügelgen, N. & Chekulaeva, M. Conservation of a core neurite transcriptome across neuronal types and species. *Wiley Interdiscip. Rev. RNA* e1590 (2020).

9. Culver, B. P. *et al.* Huntington's Disease Protein Huntingtin Associates with its own mRNA. *J Huntingtons Dis* **5**, 39–51 (2016).

10. Romo, L., Mohn, E. S. & Aronin, N. A Fresh Look at Huntingtin mRNA Processing in Huntington's Disease. *J Huntingtons Dis* **7**, 101–108 (2018).

11. White, J. A., 2nd *et al.* Huntingtin differentially regulates the axonal transport of a sub-set of Rab-containing vesicles in vivo. *Hum. Mol. Genet.* **24**, 7182–7195 (2015).

12. Fernandopulle, M. S., Lippincott-Schwartz, J. & Ward, M. E. RNA transport and local translation in neurodevelopmental and neurodegenerative disease. *Nat. Neurosci.* (2021) doi:10.1038/s41593-020-00785-2.

13. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. 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. Gyllborg, D. *et al.* Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkaa792.

16. Alon, S. *et al.* Expansion Sequencing: Spatially Precise In Situ Transcriptomics in Intact Biological Systems. *Cold Spring Harbor Laboratory* 2020.05.13.094268 (2020) doi:10.1101/2020.05.13.094268.

17. Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data.

Genome Biol. **22**, 78 (2021).

18. 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) doi:10.1101/2020.05.31.125658.
19. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
20. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15–15 (2018).
21. Imbert, A. *et al.* FISH-quant v2: a scalable and modular analysis tool for smFISH image analysis. *bioRxiv* 2021.07.20.453024 (2021) doi:10.1101/2021.07.20.453024.
22. Walter, F. C., Stegle, O. & Velten, B. FISHFactor: A probabilistic factor model for spatial transcriptomics data with subcellular resolution. *bioRxiv* (2021) doi:10.1101/2021.11.04.467354.
23. Park, H. Y., Trcek, T., Wells, A. L., Chao, J. A. & Singer, R. H. An unbiased analysis method to quantify mRNA localization reveals its correlation with cell motility. *Cell Rep.* **1**, 179–184 (2012).
24. Stueland, M., Wang, T., Park, H. Y. & Mili, S. RDI Calculator: An Analysis Tool to Assess RNA Distributions in Cells. *Sci. Rep.* **9**, 8267 (2019).
25. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
26. Ripley, B. D. The second-order analysis of stationary point processes. *J. Appl. Probab.* **13**, 255–266 (1976).
27. Jordahl, K. *et al. geopandas/geopandas: v0.9.0*. (2021). doi:10.5281/zenodo.4569086.
28. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
29. Team, D. D. Dask: Library for dynamic task scheduling. (2016).
30. Rocklin, M. Dask: Parallel computation with blocked algorithms and task scheduling. in *Proceedings of the 14th python in science conference* vol. 130 136 (Citeseer, 2015).
31. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv [cs.LG]* (2019).
32. Kossaifi, J., Panagakis, Y., Anandkumar, A. & Pantic, M. TensorLy: Tensor Learning in Python. *J. Mach. Learn. Res.* **20**, 1–6 (2019).

33. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Alexander Wolf, F. anndata: Annotated data. *bioRxiv* 2021.12.16.473007 (2021) doi:10.1101/2021.12.16.473007.

34. He, S. *et al.* High-plex multiomic analysis in FFPE at subcellular level by spatial molecular imaging. *bioRxiv* (2021) doi:10.1101/2021.11.03.467020.

35. Lee, J. H. *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10**, 442–458 (2015).

36. Hu, S. *et al.* Dynamic control of metabolic zonation and liver repair by endothelial cell Wnt2 and Wnt9b revealed by single cell spatial transcriptomics using Molecular Cartography. *bioRxiv* 2022.03.18.484868 (2022) doi:10.1101/2022.03.18.484868.

37. Chen, J. *et al.* The Allen Cell and Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images. *bioRxiv* 491035 (2020) doi:10.1101/491035.

38. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* (2022) doi:10.1038/s41592-021-01358-2.

39. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).

40. 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).

41. Stoeger, T., Battich, N., Herrmann, M. D., Yakimovich, Y. & Pelkmans, L. Computer vision for image-based transcriptomics. *Methods* **85**, 44–53 (2015).

42. Samacoits, A. *et al.* A computational framework to study sub-cellular RNA localization. *Nat. Commun.* **9**, 4584 (2018).

43. Chouaib, R. *et al.* A Dual Protein-mRNA Localization Screen Reveals Compartmentalized Translation and Widespread Co-translational RNA Targeting. *Dev. Cell* **54**, 773–791.e5 (2020).

44. Shashua, A. & Hazan, T. Non-negative tensor factorization with applications to statistics and computer vision. in *Proceedings of the 22nd international conference on Machine learning* 792–799 (Association for Computing Machinery, 2005).

45. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

46. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*

49, D325–D334 (2021).

47. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
48. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90 (2021).
49. 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).
50. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
51. Cieśla, M. *et al.* Oncogenic translation directs spliceosome dynamics revealing an integral role for SF3A3 in breast cancer. *Mol. Cell* **81**, 1453–1468.e12 (2021).
52. Petasny, M. *et al.* Splicing to Keep Cycling: The Importance of Pre-mRNA Splicing during the Cell Cycle. *Trends Genet.* **37**, 266–278 (2021).
53. Scott, D. D. *et al.* Nol12 is a multifunctional RNA binding protein at the nexus of RNA and DNA metabolism. *Nucleic Acids Res.* **45**, 12509–12528 (2017).
54. Tamanini, F. *et al.* Differential expression of FMR1, FXR1 and FXR2 proteins in human brain and testis. *Hum. Mol. Genet.* **6**, 1315–1322 (1997).
55. Miki, T., Takano, K. & Yoneda, Y. The role of mammalian Staufen on mRNA traffic: a view from its nucleocytoplasmic shuttling function. *Cell Struct. Funct.* **30**, 51–56 (2005).
56. Heraud-Farlow, J. E. *et al.* Staufen2 regulates neuronal target RNAs. *Cell Rep.* **5**, 1511–1518 (2013).
57. Kusek, G. *et al.* Asymmetric segregation of the double-stranded RNA binding protein Staufen2 during mammalian neural stem cell divisions promotes lineage progression. *Cell Stem Cell* **11**, 505–516 (2012).
58. Mallardo, M. *et al.* Isolation and characterization of Staufen-containing ribonucleoprotein particles from rat brain. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 2100–2105 (2003).
59. 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. U. S. A.* **116**, 19490–19499 (2019).
60. Gasparski, A. N., Mason, D. E., Moissoglu, K. & Mili, S. Regulation and outcomes of localized RNA translation. *Wiley Interdiscip. Rev. RNA* e1721 (2022).
61. Spitzer, H., Berry, S., Donoghoe, M., Pelkmans, L. & Theis, F. J. Learning consistent subcellular

landmarks to quantify changes in multiplexed protein maps. *bioRxiv* 2022.05.07.490900 (2022) doi:10.1101/2022.05.07.490900.

62. Das, S., Vera, M., Gandin, V., Singer, R. H. & Tutucci, E. Intracellular mRNA transport and localized translation. *Nat. Rev. Mol. Cell Biol.* **22**, 483–504 (2021).
63. Wu, K. E., Parker, K. R., Fazal, F. M., Chang, H. Y. & Zou, J. RNA-GPS predicts high-resolution RNA subcellular localization and highlights the role of splicing. *RNA* **26**, 851–865 (2020).
64. Fazal, F. M. *et al.* Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell* **178**, 473–490.e26 (2019).
65. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
66. Moffitt, J. R. *et al.* High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11046–11051 (2016).
67. Huang, H. *et al.* CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nat. Genet.* **53**, 1064–1074 (2021).
68. Emanuel, G., seichhorn, Babcock, H., leonardosepulveda & timblosser. *ZhuangLab/MERlin: MERlin v0.1.6*. (2020). doi:10.5281/zenodo.3758540.
69. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).