UNIVERSITY OF CALIFORNIA SAN DIEGO

Computational Frameworks for Functional Subcellular Analysis of Spatial Transcriptomics Data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy

in

Bioinformatics & Systems Biology

by

Clarence K. Mah

Committee in charge:

Professor Hannah Carter, Chair
Professor Gene Yeo, Co-Chair
Professor Nate Lewis
Professor Samara Reck-Peterson
Professor Bing Ren

2023

The Dissertation of Clarence K. Mah is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

...

This dissertation is dedicated to my parents and sister,
for believing in my journey into the unknown.

EPIGRAPH

Moral courage is a rarer commodity than bravery in battle or great intelligence.
Yet it is the one essential, vital quality for those who seek to change a world
that yields most painfully to change.

- *Robert Kennedy*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank the following people for their support throughout my graduate school journey. Hannah Carter and Gene Yeo for being a pair of trusting, insightful and inspiring advisors and providing me as many as four separate offices in my graduate career. Nate Lewis for pushing me to always think about the bigger picture in terms of both my research and career. Samara Reck-Peterson and Bing Ren for their thoughtful suggestions. My partner, Michelle Ragsac, for always being my number one advocate, grounding me in reality, and for sharing my enthusiasm for food. My dog, Yuuki, for accompanying me for the entirety of my graduate education (she deserves a "PhDoggo" of her own). My parents and sister for providing me a home and environment to get away and decompress. Noorsher Ahmed for all the wild brainstorming sessions, last minute sprints to materialize results on crazy deadlines, and late night milk tea runs. Andrea Castro for keeping me company in our almost windowless offices. Kivilcim Ozturk for being my long time office buddy and Michelle Dow for being an inspirational role model, for all the invaluable scientific and life advice she gave me at the start of my PhD, for bestowing her desk to me and for taking Yuuki on walks when I got stuck in meetings all day. Adam Klie, James Talwar, and David Laub for entertaining my philisophical rants on software design and machine learning. Erick Armingol and Cameron Martino whose work on tensor decomposition inspired my own research. Bruce Hamilton for showing me the importance of sticking to my scientific and moral principles. The Yeo lab for being an amazing collaborative research environment. The Carter lab for a supporting my work life balance. The Bioinformatics and Systems Biology program for being my support network when I needed it most and fighting for my rights as a graduate student researcher.

VITA

2012	High School Diploma Lynbrook High School
2012–2016	B.S. in Bioengineering: Bioinformatics University of California San Diego
2016–2018	Associate Computational Biologist University of California San Diego
2018–2023	Ph.D. in Bioinformatics & Systems Biology University of California San Diego

PUBLICATIONS

Author names marked with † indicate shared first co-authorship.
Publications marked with △ are included in this text.

Avery Pong, **Clarence K. Mah**, Gene Yeo, Nathan E. Lewis. "Computational cell-cell interaction technologies drive mechanistic and biomarker discovery in the tumor microenvironment." *Accepted, Current Opinion in Biotechnology*, 2023.

Clarence K. Mah†, Noorsher Ahmed†, Nicole Lopez, Dylan Lam, Alexander Monell, Colin Kern, Yuanyuan Han, Gino Prasad, Anthony J. Cesnik, Emma Lundberg, Quan Zhu, Hannah Carter, Gene W. Yeo. "Bento: A toolkit for subcellular analysis of spatial transcriptomics data." *bioRxiv*, 2022. △

Sydney C. Morgan†, Stefan Aigner†, Catelyn Anderson†, Pedro Belda-Ferre†, Peter De Hoff†, Clarisse A. Marotz†, Shashank Sathe†, Mark Zeller†, Noorsher Ahmed, Xaver Audhya, Nathan A. Baer, Tom Barber, Bethany Barrick, Lakshmi Batachari, Maryann Betty, Steven M. Blue, Brent Brainard, Tyler Buckley, Jamie Case, Anelizze Castro-Martinez, Marisol Chacón, Willi Cheung, LaVonnye Chong, Nicole G. Coufal, Evelyn S. Crescini, Scott DeGrand, David P. Dimmock, J. Joelle Donofrio-Odmann, Emily R. Eisner, Mehrbod Estaki, Lizbeth Franco Vargas, Michele Freddock, Robert M Gallant, Andrea Galmozzi, Nina J. Gao, Sheldon Gilmer, Edyta M. Grzelak, Abbas Hakim, Jonathan Hart, Charlotte Hobbs, Greg Humphrey, Nadja Ilkenhans, Marni Jacobs, Christopher A. Kahn, Bhavika K. Kapadia, Matthew Kim, Sunil Kurian, Alma L. Lastrella, Elijah S. Lawrence, Kari Lee, Qishan Liang, Hanna Liliom, Valentina Lo Sardo, Robert Logan, Michal Machnicki, Celestine G. Magallanes, **Clarence K. Mah**, Denise Malacki, Ryan J. Marina, Christopher Marsh, Natasha K. Martin, Nathaniel L. Matteson, Daniel J. Maunder, Kyle McBride, Bryan McDonald, Daniel McDonald, Michelle McGraw, Audra R. Meadows, Michelle Meyer, Amber L. Morey, Jasmine R. Mueller, Toan T. Ngo, Julie Nguyen, Viet Nguyen, Laura J. Nicholson, Alhakam Nouri, Victoria Nudell, Eugenio Nunez, Kyle O'Neill, R. Tyler Ostrander, Priyadarshini Pantham, Samuel S. Park, David Picone, Ashley Plascencia, Isaraphorn Pratumchai, Michael Quigley, Michelle Franc Ragsac, Andrew C. Richardson, Refugio Robles-Sikisaka, Christopher A. Ruiz, Justin Ryan, Lisa Sacco, Sharada Saraf, Phoebe Seaver,

Leigh Sewall, Elizabeth W. Smoot, Kathleen M. Sweeney, Chandana Tekkatte, Rebecca Tsai, Holly Valentine, Shawn Walsh, August Williams, Min Yi Wu, Bing Xia, Brian Yee, Jason Z. Zhang, Kristian G. Andersen, Lauge Farnaes, Rob Knight, Gene W. Yeo, Louise C. Laurent. "Automated, miniaturized, and scalable screening of healthcare workers, first responders, and students for SARS-CoV-2 in San Diego County." *medRxiv*, 2021.

Shashank Sathe, **Clarence K. Mah**, Noorsher Ahmed, Michelle Franc Ragsac, John Williams, Gene Yeo. "INSPECT Sample Tracking System." *protocols.io*, 2020.

ABSTRACT OF THE DISSERTATION

Computational Frameworks for Functional Subcellular Analysis of Spatial Transcriptomics Data

by

Clarence K. Mah

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California San Diego, 2023

Professor Hannah Carter, Chair
Professor Gene Yeo, Co-Chair

Emerging genomic technologies that measure spatial information about RNA molecules promise to shed light on cell biology and function. However, most analytical techniques have primarily concentrated on spatial relationships at the multicellular and cellular scale without fully tapping into single-molecule spatial information. To address this gap, I introduce Bento, a toolkit designed for discerning spatial relationships at the subcellular scale. Bento incorporates a suite of statistical and machine learning methods within an intuitive Python programming interface, emphasizing the FAIR data management principles. To showcase its capabilities, I utilized Bento to study RNA localization changes in doxorubicin-treated cardiomyocytes profiled with spatial transcriptomics. Our findings reveal that doxorubicin-induced stress leads to the depletion of

disease-associated genes in the endoplasmic reticulum, along with expression changes previously associated with doxorubicin-induced cardiotoxicity. This places the endoplasmic reticulum as a pivotal subcellular structure in the response to doxorubicin treatment. In essence, Bento emerges as a potent toolkit for the subcellular analysis of spatial transcriptomics data, paving the way for the discovery of new spatial relationships between subcellular structures and molecules. Furthermore, I have created a framework tailored to streamline image processing for spatial transcriptomics data called spotfish. Similar to Bento's ethos, spotfish is built in alignment with FAIR principles and leverages open-source standards like the Nextflow workflow language and Open Microscopy Environment file formats. Collectively, Bento and spotfish empower researchers to harness spatial transcriptomics technologies, enabling more comprehensive exploration of the spatial and molecular organization of cells at an unprecedented throughput.

Introduction

Introduction

0.1 The importance of RNA localization

The fundamental unit of life is the cell, from unicellular organisms like bacteria to complex multicellular organisms like humans. While it is convenient to think of cells as amorphous liquid bags of lipids, proteins and sugars, cells are highly structured and and regulated. The genome serves as the template for RNAs; they are synthesized then modified by tightly orchestrated processes such as splicing, localization, translation, and degradation so a cell can function. We can measure the abundance of a cell's transcriptome, the complete repertoire of RNA, to loosely quantify cellular activity. But what do these molecules physically interact with? Where do these interactions occur in the cell and what causes them to interact? This layer of regulation, RNA localization, plays an important role in cell processes such as protein synthesis, signaling pathways and RNA degradation. For example, mRNAs exhibit asymmetric distributions in developing *Drosophila melanogaster* embryos, compartment-specific localization in the neurites of neurons, and colocalization with the actin cytoskeleton in fibroblasts[1]. The prevalence of RNA localization across diverse cell types and organisms indicate that it is a highly conserved process. Abnormal RNA localization has also been associated with many neurodegenerative 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[2]. Despite these repeated observations, the determinants of localization are not well understood.

0.2 Spatial transcriptomics technology

While we can easily quantify RNA expression with sequencing, RNA imaging techniques have traditionally been limited to visualizing a handful of species per experiment. However,

recent multiplexed imaging technologies have unlocked much higher experimental throughput at hundreds to thousands of species, enabling nearly transcriptome-scale analysis of spatial RNA distributions. Single-molecule fluorescent in situ hybridization[3] (smFISH) was one of the first popularized techniques able to image RNAs by species using synthesized complementary DNA (cDNA) sequences with fluorochromes. The cDNA probes hybridize to RNA targets and emit light upon excitation, which is captured by microscope cameras as dots of light, less than a micron wide. By designing specific probes for each RNA species of interest, it is possible to image multiple unique species at a time in the same cells. In order to scale to target hundreds to tens of thousands of unique RNA species, recent combinatorial FISH techniques compress the number of imaging rounds needed to identify each target by designing sets of barcodes that fluoresce in a specific sequence of images for individual RNA species. For example, MERFISH[4] is one technique using a barcoding scheme that allows detection of 10,000 unique RNA targets in 69 rounds of sequential images. At such a scale, we can begin to study the RNA life cycle from a new perspective, by observing the spatial organization of the transcriptome and uncovering principles of RNA regulation linked to localization. The set of technologies able to capture the spatial organization of RNA in cells and tissue is termed spatial transcriptomics.[i]

0.3 Current analysis trends

As spatial transcriptomics assays reaches the scientific main stream[5], there is a growing need for scalable analysis software and computational infrastructure. The most robust and enduring tools adhere to FAIR principles[6] — Findability, Accessibility, Interoperability, and Reusability—a set of standards proposed by researchers to maximize reuse of research objects for advancing scientific discovery. Data management strategies such as version control, software containerization, and pipeline management are often second priority in academic research. Consequentially, many academic tools do not see use outside of their initial projects and collaborations due to low adoption. Mainstream media attention around the "reproducibility crisis" and high profile cases of academic fraud have demonstrated the clear value of enforcing

[i] In addition to imaging-based methods, there exists a host of slide-based methods that are just as prevalent but not in the scope of this work. In summary these methods use a grid of barcoded wells on slides to capture and sequence transcripts. The location of each well is used to spatially map transcripts.

FAIR principles for academic research, both to researchers and for building trust with the average citizen.

The field of spatial transcriptomics has witnessed an evolution of various tools and platforms, each specializing in different aspects of analysis and data handling. For image processing, tools like *multi-fish*[7], *mcmicro*[8], and *MERlin*[9] are tailored specifically for certain technologies. In contrast, *starfish/PIPEFISH*[10;11] and *spotfish* (described in this work) attempt to be platform agnostic and focus on pipeline building infrastructure instead of task-specific algorithms. In terms of data structures, *AnnData*[12] specifically supports single-cell data matrices, while *SpatialData*[13], *SpatialExperiment*[14] offer more complex representations, attempting supporting a spectrum of data modalities and the relationships between them. Single-cell analysis is the *de facto* approach to analyze spatial transcriptomics, and tools such as *Giotto*[15], *Squidpy*[16], *Stereopy*[17], *stLearn*[18], and *Voyager*[19] are equipped to handle cell-centric functional analyses. Subcellular analysis, which delves deeper into spatial interactions at the molecular level, features tools like *INSTANT*[20], *SpaGNN*[21], and *FISHfactor*[22]. *Bigfish*[23] and *Bento*[24] in this category are examples of software packages developed with FAIR principles in mind. Overall, this brief listing highlights the growing interest in the budding field of spatial transcriptomics and the need for FAIR tools to support the maturation of the field.

Table 0.1. Assessment of FAIR principles in spatial transcriptomics tools

Category	Tool	Spatial-Tx Compatible	Findability	Accessibility	Interoperability	Reusability
Image Processing	easi-fish		x	x	x	x
	MERlin	MERFISH only	x	x	x	x
	mcmicro		x	x	x	x
	starfish/PIPEFISH	x	x			x
	spotfish	x	x	x	x	x
Data Structure	AnnData		x	x	x	x
	SpatialData	x	x	x	x	x
	SpatialExperiment	x	x	x	x	x
Single-Cell Analysis	Giotto	x	x	x	x	x
	Squidpy	x	x	x	x	x
	Stereopy	x	x	x	x	x
	stLearn	x	x	x	x	x
	Voyager	x	x	x	x	x
Subcellular Analysis	INSTANT	x				
	SpaGNN	x				
	FISHfactor	x				x
	Bigfish		x	x		x
	Bento	x	x	x	x	x

Chapter 1

Bento: A toolkit for subcellular analysis of spatial transcriptomics data

1.1 Introduction

The spatial organization of molecules in a cell is essential for performing their functions. While protein localization[25] and disease-associated mislocalization are well appreciated[26;27], the same principles for RNA have begun to emerge. For instance, the spatial and temporal regulation of RNA play a crucial role in localized cellular processes such as cell migration and cell division[28;29], as well as specialized cell functionalities like synaptic plasticity[30–32]. 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[2;33–35].

The study of subcellular RNA localization necessitates single-molecule measurements. Since the development of single-molecule fluorescent in situ hybridization (smFISH), recent advances in multiplexed methods such as MERFISH[36], seqFISH+[37], HybISS[38], and Ex-Seq[39] have enabled RNA localization measurements at near transcriptome scales, while maintaining single-molecule resolution. A number of computational toolkits, such as Squidpy[16], stLearn[18], Giotto[15], Seurat[40], and Scanpy[41] enabled the characterization of tissue architecture, cell-cell interactions, and spatial expression patterns. Despite the single-molecule measurements in spatial transcriptomics, these analytical approaches are limited to investigating spatial variation at the multicellular scale and lack the ability to investigate subcellular organization. To further our understanding of RNA localization and its function in normal and abnormal cell activity, we

need to expand our analytical capacity to the subcellular scale.

Recent methods such as FISH-quant-v2[23] and FISHFactor[22] identify subcellular patterns describing the spatial distribution of RNA species, but are unable to annotate more than a single gene per cell or are limited to analyze at most 20,000 molecules on accessible computing resources. In contrast, a single spatial transcriptomics experiment measures at least hundreds to thousands of genes across hundreds of thousands of cells. Additionally, methods such as ClusterMap[42] and Baysor[43] highlight the potential for transcript locations alone to inform meaningful domains such as cell and nuclear regions Using spatial proteomics data, CAMPA[44] and Pixie[45] utilize subcellular spatial variation in protein abundance to identify subcellular regions and annotate pixel-level features.

Building on these promising approaches, we present Bento, an open-source Python toolkit for scalable analysis of spatial transcriptomics data at the subcellular resolution. Bento ingests single-molecule resolution data and segmentation masks, utilizing geospatial tools (GeoPandas[46], Rasterio[47]) for spatial analysis of molecular imaging data, and data science tools including SciPy[48], and Tensorly[49] for scalable analysis of high-dimensional feature matrices. Furthermore, Bento is a member of the Scverse ecosystem, enabling integration with Scanpy[41], Squidpy[16], and more than thirty other single-cell omics analysis tools.

1.2 Results

1.2.1 Overview of Bento data infrastructure for subcellular 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[4], seqFISH+[37], CosMx (NanoString)[50], Xenium (10x Genomics)[51;52], and Molecular Cartography (Resolve Biosciences)[53]. Bento's workflow takes as input 1. 2D spatial coordinates of transcripts annotated by gene, and 2. segmentation boundaries (e.g. cell membrane, nuclear membrane, and any other regions of interest) (Fig. 1B). While 3D molecular coordinates are commonly included, 3D segmentation information is limited to z-stacked 2D segmentation, limiting its usability. 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[12], which links cell

and gene metadata to standard count matrices, providing compatibility with standard single-cell RNA-seq quality control and analysis tools in the Scverse ecosystem[41]. With a data structure for segmentation boundaries and transcript coordinates in place, Bento can easily compute spatial statistics and measure spatial phenotypes to build flexible multidimensional feature sets for exploratory subcellular analysis and utilize these spatial metrics to augment quality control (Fig. 1C).

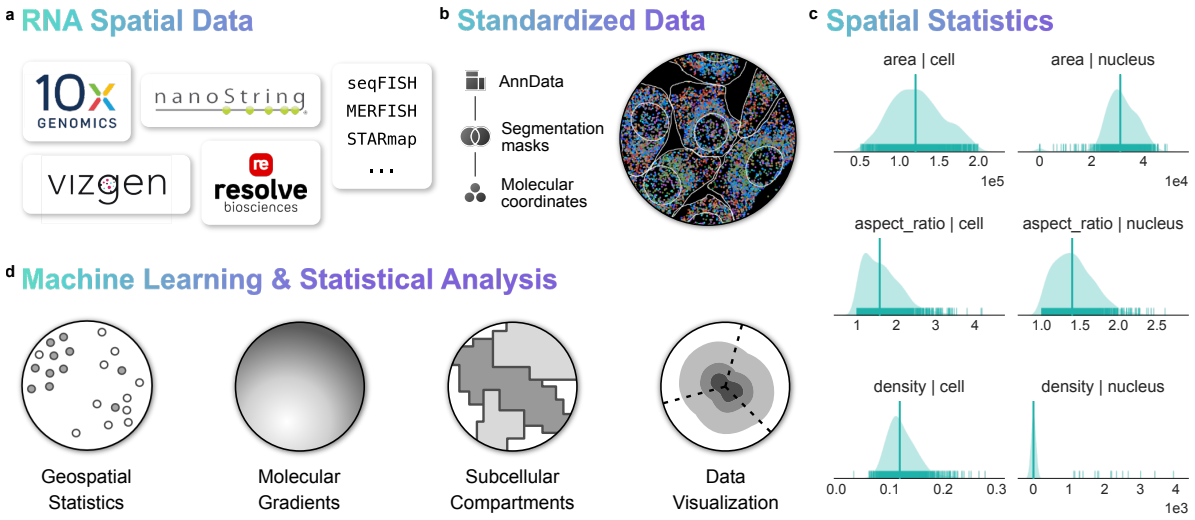

Figure 1.1. Workflow and functionality of the Bento toolkit. A. Single-molecule resolved spatial transcriptomics data from commercial or custom platforms are ingested into Bento where it is converted to the AnnData format (B.), where it can be manipulated with Bento as well as a wide ecosystem of single-cell omics tools. C. Geometric statistics are illustrated for the seqFISH+ dataset, including metrics describing cell and nuclear geometries and cell density to assess overall data quality. D. Bento has a standard interface to perform a wide variety of subcellular analyses.

Bento offers a precise yet flexible palette of novel complementary subcellular analyses (Fig. 1D). We present RNAforest, a multilabel approach for annotating RNA localization patterns adapted from FISH-quant v2[54]. We find that many RNAs are spatially distributed according to gene function. We then implement RNAcoloc, a context-specific approach to quantify colocalization to characterize how genes colocalize with each other in a compartment-specific manner. Having established systematic patterning and organization of RNA transcripts, we demonstrate RNAflux, an unsupervised method for semantic segmentation of subcellular domains. RNAflux first quantifies subcellular expression gradients at pixel resolution before identifying consistent subcellular domains via unsupervised clustering. We demonstrate the utility of Bento's

tools characterizing subcellular organization in two spatial transcriptomics datasets, a 10k gene MERFISH dataset of U2-OS cells and a 130 gene seqFISH+ dataset of 3T3 cells. We find that RNA localization patterns are associated with known gene function, and that genes with similar localization patterns are functionally related. We also find that genes with similar localization patterns are co-regulated at the transcriptional level. Finally, we find that RNAflux identifies subcellular domains that are consistent across cells and are associated with known subcellular structures.

1.2.2 RNAforest: Utilizing subcellular landmarks to predict RNA subcellular localization

In computer vision, key points or landmarks are commonly used for tasks like facial recognition[55] and object detection. Analogous to these classical applications, we derive spatial features using cell and nucleus boundaries as landmarks to predict RNA localization patterns from spatial summary statistics. Building on the summary statistics used for classifying smFISH data in FISH-quant v2[23], RNAforest consists of an ensemble of five binary random forest classifiers rather than a single multi-classifier model to assign one or more labels. These pattern labels, adapted from several high-throughput smFISH imaging experiments in HeLa cells[56–59], are broadly applicable to eukaryotic cells: (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). It is important to note, as was done previously in FISH-quant v2[23] that because of the 2D nature of the dataset, RNA that is in truth cytoplasmic but above or below the nucleus will still appear as though in the nucleus when collapsed in the z-dimension. As we use the FISH-quant v2 pattern simulation framework, this is accounted for in the training dataset.

We used the FISH-quant v2 simulation framework to generate realistic ground-truth data[58]. 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 its spatial 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). These features are normalized to morphological properties of the cell to control for variability in cell shape. A detailed description of every feature is described in Supp. Table 1.

We applied RNAforest on the MERFISH dataset measuring 130 genes (low plexity) in U2-OS cells and high detection efficiency per gene (111 molecules per gene per cell on average), and on the seqFISH+ dataset measuring 10,000 genes (very high plexity) but lower detection efficiency (8 molecules per gene per cell on average) (Fig. 2B-C, Supp. Fig. 1). In agreement with previous work characterizing RNA localization of 411 genes[59], we find that genes commonly exhibit variability in localization across cells. This suggests that heterogeneity in localization likely generalizes to the entire transcriptome. Of the localization patterns besides "none", "nuclear" was the most common (22.1%) in the U2-OS osteosarcoma cells (Fig. 2D & 2F), while "cell edge" was the most common (15.9%) in the 3T3 fibroblast cells (Fig. 2E & 2G).

In the U2-OS cells, we found many genes to have preferential localization in different subcellular compartments (Fig. 2H). In agreement with our RNAflux findings, we find genes known to localize to the nucleus[20;60] to be frequently labeled "nucleus" (MALAT1, SOD2) and genes encoding secreted extracellular proteins[36] to be frequently labeled "nuclear edge" (FBN1, FBN2). As expected, we find genes preferentially "nuclear" and "nuclear edge" localized to mirror nucleus and endoplasmic reticulum genes found in a 10k genes MERFISH study of U2-OS cells that included ER staining[61] (Supp. Fig. 2, Methods). Leveraging the 3T3 seqFISH+ dataset's higher plexity, we were able to ask whether genes with similar localization preference are functionally related. We applied gene set enrichment analysis to gene localization frequencies to identify enriched gene ontology terms[62] (Fig. 2I, Methods). Secretory processes were enriched in the nucleus and nuclear edge, which may be linked to increased transcription of fibroblast-related functions. Cell edge enriched pathways consisted of those with the cell membrane as their site of function (e.g. endocytosis and tight junction suggesting local translation of these genes). Additionally, the term for cell cycle was significantly enriched in the cytoplasm only. Genes without strong localization preference (most frequently "none") were not significantly associated with any pathways. These genes likely do not undergo active transport and are functionally

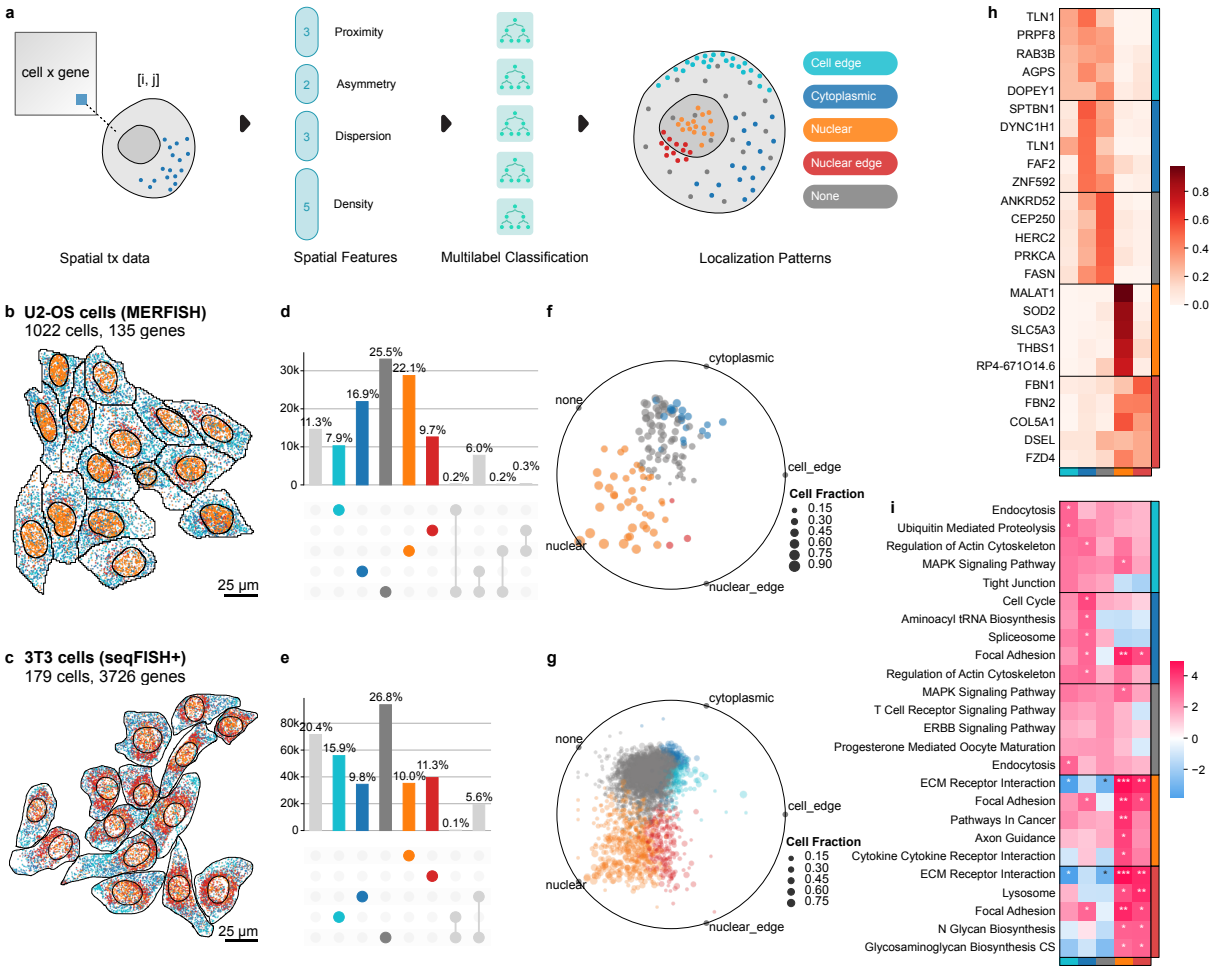

Figure 1.2. Subcellular localization pattern identification with RNAforest. A. Thirteen spatial summary statistics are computed for every gene-cell pair describing the spatial arrangement of molecules and boundaries in relation to one another. The features are inputs for RNAforest, a multilabel ensemble classifier which assigns one or more subcellular localization labels: cell edge, cytoplasmic, nuclear, nuclear edge, and none. Top 10 genes for each label visualized for each label other than "none" in B. U2-OS cells, and C. 3T3 cells. D. and E. show the proportion of measured transcripts assigned to each label. F. and G. show the relative label proportion across cells for each gene and is colored by the majority label (F and G). H. Top 5 consistent genes for each label. I. ssGEA identifies enrichment of GO cellular component domains for each label in the 3T3 cell dataset.

independent of local translation. RNAforest gives a user a facile method for annotating RNA localization patterns and quantifying heterogeneity in a transcriptome-wide manner independent of RNA abundance. Beyond known RNA localizations, we find that transcript location is generally associated with known gene function, alluding to the systematic spatial regulation of RNA transport. We foresee RNAforest will be a valuable addition to characterize RNA localization across diverse spatial transcriptomics datasets.

1.2.3 RNAcoloc: An approach for context-specific RNA colocalization

In geospatial information processing, a fundamental feature that is often gleaned from large datasets is the colocation of objects (e.g. gleaning socialization metrics from cell phone colocation data in Singapore[63]). Colocation is similarly valuable in understanding co-translation and interaction networks of genes in a biological context[64]. Recent spatial transcriptomics approaches have used a number of colocalization metrics from the geographic information systems and ecology fields e.g. the bivariate versions of the Ripley's K function (also known as cross-k-function)[65], Moran's I[66], and the join count statistic[67]. These metrics are designed to measure spatial associations between two populations i.e. gene A transcripts and gene B transcripts. However, it is more appropriate to think of all transcripts in a single cell from a single population; after all, RNA transcription and localization is not completely stochastic. We have shown that the subcellular distribution of RNA is highly structured with RNAforest. As such, we developed RNAcoloc, an approach that combines the Colocation Quotient (CLQ)[68] metric and tensor decomposition for context-specific RNA colocalization (Methods). The CLQ is a colocalization statistic that is capable of accounting for the biophysical properties of RNA spatial distributions. First, the CLQ considers how clustered the overall RNA population is in a cell and measures whether specific pairs of genes are more clustered than expected given the spatial pattern of the overall population. Second, the CLQ is inherently asymmetric, and captures the direction of attraction i.e. the attraction of gene A to gene B is not the same as the attraction of gene B to gene A. This is most common when gene A and gene B have very different expression levels, which is prevalent due to overdispersion in gene expression data.

RNAcoloc calculates CLQ scores for each gene per cell in a compartment-specific manner,

such that each sample has 2 scores, a nucleus and cytoplasm CLQ score. An initial comparison of global colocalization between nuclear and cytoplasmic fractions unsurprisingly found that transcripts from the same gene tend to cluster more tightly with themselves than with transcripts from other genes (Fig. 3B). Additionally, self-colocalization is significantly stronger in the cytoplasm than in the nucleus. In conjunction with our findings from RNAforest analysis that genes of the same localization pattern tend to have similar functions, this suggests that the RNAs are more tightly spatially regulated once exported from the nucleus.

By calculating CLQ scores for every gene-gene pair across compartments, RNAcoloc constructs a tensor of shape P x C x S where P, C, and S represent the number of gene-gene pairs, cells, and compartments, respectively (Fig. 3A, Methods).

RNAcoloc then applies tensor decomposition — specifically, non-negative parallel factor analysis — a data-driven, unsupervised approach for discovering substructure in high-dimensional data[49;69] to decompose the U2-OS dataset colocalization tensor into k = 4 "colocalization factors". The number of factors was determined using the elbow method heuristic, optimizing for the root mean squared error (RMSE) reconstruction loss (Methods). Unlike matrix dimensionality reduction methods, such as PCA, the order of the components (factors) is unassociated with the amount of variance explained. Each of the 4 colocalization factors is composed of 3 loading vectors, which correspond to the compartments, cells and gene pairs. Higher values denote a stronger association with that factor. Crucially for interpretation, factors derived from tensor decomposition are not mutually exclusive and share overlapping sets of associated compartments, cells, and gene pairs.

These trends are broken down into unique combinations of colocalization behavior (Fig. 3C). Factor 0 captures gene pairs in a subpopulation of cells that tend to colocalize across the entire cell, with pairs including SLC38A1 showing the strongest signal. Factor 3 describes gene pairs in mostly the same cell subpopulation, that colocalize specifically in the cytoplasm. Pairs including PIK3CA dominate this behavior. Interestingly, PIK3CA and DYNC1H1 transcripts colocalize cytoplasmically. While little is known about their RNA interactions, PIK3CA and other members of the PI3K pathway are known regulators of mitotic organization, including the regulation of dynein and dynactin motor proteins. DYNC1H1 specifically encodes cytoplasmic

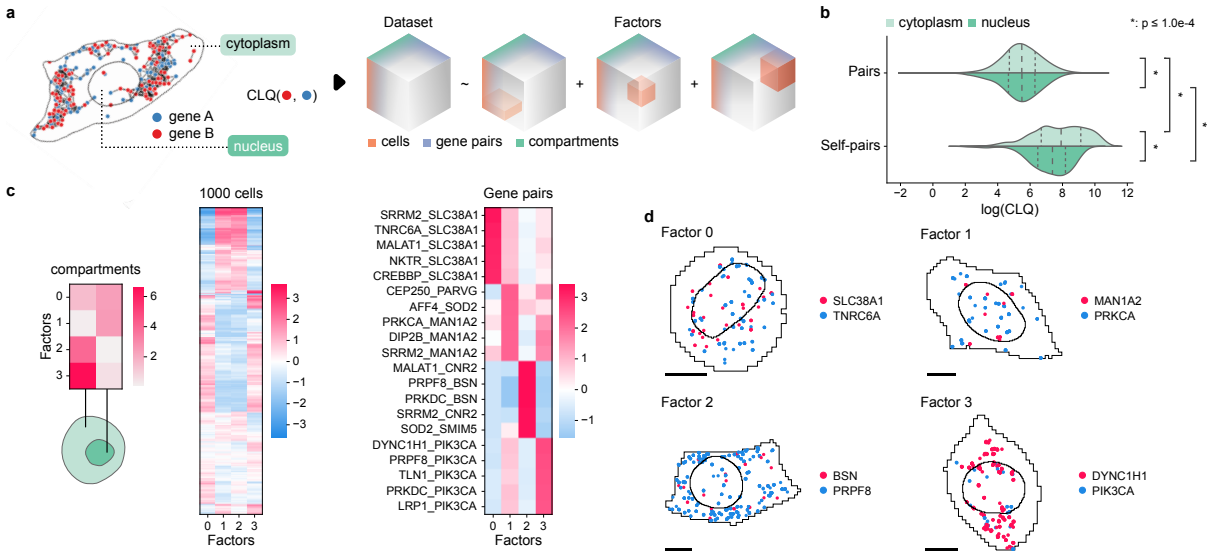

Figure 1.3. Compartment-specific RNA colocalization with RNAcoloc. A. Transcripts are separated by compartment (nucleus and cytoplasm) before CLQ scores are calculated for every gene pair across all cells. This yields a cell x gene pair x compartment tensor. B. Comparison of log CLQ distributions for gene pairs and self-pairs, further categorized by compartment. C. Tensor decomposition yields 4 factors. From left to right, the three heatmaps show the loadings of each factor for each dimension – compartments, cells, and gene pairs. Only the top 5 associated gene pairs for each factor are shown. D. Top examples of compartment-specific colocalized gene pairs. Black scale bars denote 10 um.

dynein, a motor protein critical for spindle formation and chromosomal segregation in mitosis[70]. In the complementary cell population, Factors 1 and 2 highlight colocalized gene pairs in the nucleus and cytoplasm respectively. Notably, Factor 2 associated cells have high loadings for MALAT1 and CNR2 in the cytoplasm and low loadings in the nucleus. Even though MALAT1 is abundantly localized to the nucleus, this demonstrates that the CLQ score identifies gene pairs colocalizing more than expected given the overabundance of MALAT1 relative to CNR2, whereas other approaches seem confounded by large differences in expression[20].

We demonstrate the ability of RNAcoloc to quantify compartment-specific gene-pair colocation by exploring cytoplasmic vs. nuclear colocalization. As we found separately with RNAforest, RNAcoloc analysis finds evidence that RNA transport is spatially regulated, especially after nuclear export. We highlight several examples of colocalization suggesting how RNA localization allows the same gene to have multiple functions in a spatially-dependent fashion i.e. depending on its molecular neighbors and local environment[71;72]. We foresee RNAcoloc will be increasingly relevant as many spatial technologies are beginning to image proteins along with RNA, which can be used to delineate more granular compartments, such as cell organelles or distinct regions e.g. neuron cell bodies vs dendrites.

1.2.4 RNAflux: Unsupervised semantic segmentation of subcellular domains in single cells

To build on RNAforest, we overcame the restricted number of localization patterns defined by the supervised method by framing RNA localization as an unsupervised embedding problem. RNAflux looks at local neighborhoods within the space of a cell and builds a normalized gene composition per neighborhood. Differences in neighborhood compositions can be leveraged to identify distinct subcellular domains in a manner that is entirely unsupervised and independent of cell geometry.

We applied this embedding procedure to compute a gene composition vector for every pixel in 2D coordinate space, generating a spatial composition gradient across entire cells (Fig. 4A, Methods).

Applied to a MERFISH dataset with a target panel of 130 genes across over 1153 U2-OS cells, we demonstrate that RNAflux embeddings can detect transcriptionally distinct subcellular

domains. Performing dimensional reduction of the embeddings showed that the top sources of variation spatially correspond to the nucleus, the nuclear periphery, and cytoplasmic regions consistently across cells (Fig. 4B, Methods) confirming that RNAflux measures intracellular transcriptional variation, as opposed to intercellular variation. To delineate compositionally similar domains in a data-driven manner, we cluster pixel embeddings using self-organizing maps (SOMs), effectively performing unsupervised semantic segmentation (Methods). We denote the resulting clusters as "fluxmap domains". We found that this assigned pixels to 5 fluxmap domains, consistently highlighting spatial regions across every cell (e.g. fluxmap 2 is always nuclear while the remaining domains constitute the cytoplasm) (Fig. 4B). By considering the spatial distribution of molecules across fluxmap domains, we can quantify the composition of molecules for each gene across fluxmaps (Fig. 4C) e.g. nuclear-localized MALAT1 [20;60].

Finally, we sought to characterize the fluxmap domains with known information about RNA localization. We used data from a previous study that measured gene expression at "distinct subcellular locales" via APEX-seq, a technique for proximity labeling and sequencing of RNA [73]. Of the 3288 genes differentially enriched to one or more locales, 63 overlapped with the 130 MERFISH genes. The location enrichment score for each pixel is calculated by taking the weighted sum of its RNAflux embedding and the measured relative enrichment i.e. log fold change measured by APEX-seq loadings for a given organelle-specific geneset (Methods). Visualizing each pixel's location-specific enrichment scores from the APEX-seq dataset highlights the subcellular localization of these compartments, including the cytosol, nucleus, nucleolus, nuclear pore, nuclear lamina, endoplasmic reticulum lumen (ER lumen), ER membrane (ERM), and the outer mitochondrial membrane (OMM) (Fig. 4D). We find the nuclear compartments have high scores in domain 2, while the cytoplasm scores rank highest in domains 4 and 5. Both the ERM and OMM scores are the strongest in domain 1 (Fig. 4E).

In summary, RNAflux finds distinct subcellular domains with consistent spatial organization and local gene composition. As an unsupervised method, RNAflux can be applied to any cell type for inferring subcellular domains from transcript locations and functionally annotated with biological enrichment analysis.

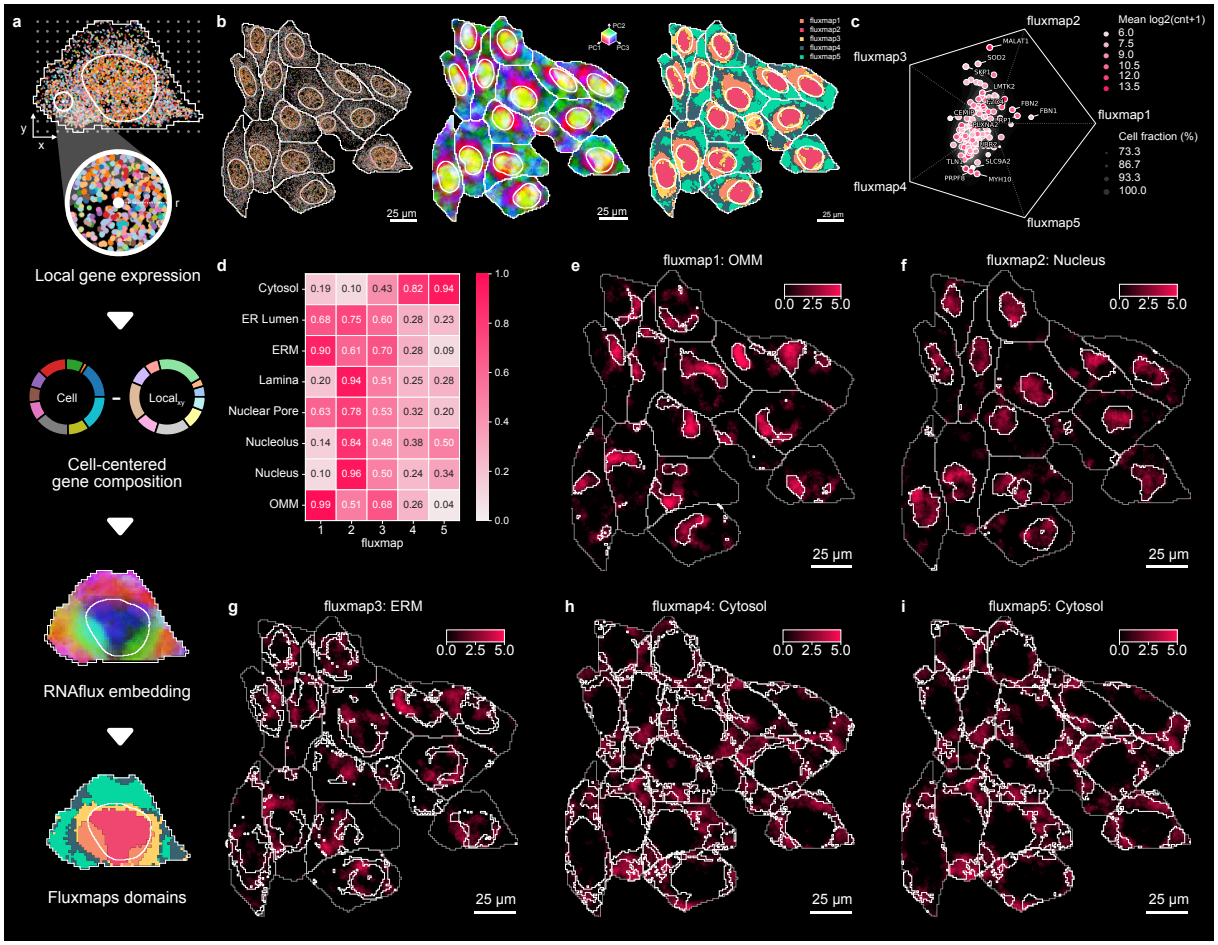

Figure 1.4. RNAflux finds distinct subcellular domains with consistent spatial organization and local gene composition. A. Flowchart of RNAflux and fluxmap computation. Local neighborhoods of a fixed radius are arrayed across a cell and a normalized gene composition is computed for each pixel coordinate, producing an RNAflux embedding. The first three principal components of the RNAflux embedding are visualized for U2-OS cells coloring RGB values by PC1, PC2, and PC3 values respectively for each pixel. Fluxmap domains are computed from each RNAflux embedding to create semantic segmentation masks of each subcellular domain. B. The left panel shows a field of view of U2-OS cells, dots denoting individual molecules colored by gene species, nuclei and cell boundaries outlined in white. For the same field of view of cells, the center panel shows RNAflux embeddings and the right panel shows fluxmap domains. C. The scatter plot shows how the composition of each gene is distributed across fluxmap domains. The position of each point denotes the relative bias of a given gene's composition across fluxmaps. D. Heatmap showing the fraction of pixels with a positive enrichment value for each APEX-seq location for each fluxmap domain. E-I. The most highly enriched location is shown for each fluxmap domain. Domain boundaries are denoted by white lines within each cell.

1.3 Discussion

Bento seeks to interrogate biology via its "subcellular first" approach to spatial analysis, complementary to "cell-type or tissue first" spatial analysis methods. The toolkit enables quantitative, reproducible, and accessible analysis agnostic of spatial technology platforms in a standardized framework. We implement three novel methods to interrogate subcellular RNA organization: RNAforest for supervised annotation of localization patterns, RNAcoloc for compartment-aware colocalization analysis, and RNAflux for identifying transcriptionally distinct subcellular domains. We showed that with RNAflux, we were able to quantify RNA localization in a variety of contexts, including domain-specific gene localization, drug-induced changes in localization, and cell type specific localization. With both RNAflux and RNAforest, we find that subcellular mRNA localization reflects gene function. With RNAcoloc, we explore the use of CLQ scores to quantify pairwise gene colocalization with the context of asymmetric associations.

From these results, we found three main factors to limit the effectiveness of subcellular-resolution analysis: molecule density, segmentation quality, and target panel composition. In particular, RNAflux becomes uninformative if too few molecules are detected per cell or if the number of molecules per gene is too sparse. We found that datasets with higher density i.e. molecules per micrometer2 are less noisy and inform more coherent gradients and domains, such as the U2-OS dataset. In contrast, RNAforest performs reliably beyond a minimum of 5-10 molecules per sample, but is sensitive to accurate segmentation for calculating cell morphology-dependent features. The 3T3 cells were manually segmented and the U2-OS cells had relatively accurate segmentation, and were therefore amenable to applying RNAforest. In the case of RNAcoloc, the limiting factor to identify relevant biology is target panel composition. The current focus of most target panels typically include cell type markers and highly expressed genes, whereas it would be more informative to identify colocalizing members of protein complexes, functional pathways, or ligand-receptor pairs.

A dimensional limitation of Bento is its current inability to process three-dimensional spatial transcriptomic data. While some commercially available spatial transcriptomic methods yield RNA molecular coordinates in 3D, the nuclear and cell segmentation is inevitably still

two dimensional making it difficult to interpret z-dimensional positions lacking the context of cellular geometry in 3D. However, the algorithms behind RNAforest, RNAcoloc, RNAflux, and the plethora of feature calculation functions in Bento are inherently extensible to leveraging three dimensionality. When three dimensional cell segmentation improves, we intend to extend Bento to support three dimensional analysis.

1.4 Conclusion

Conventionally, RNA is treated as an intermediary vehicle encoding genomic information for protein synthesis. We began our investigation of RNA localization with the hope of understanding how the spatial organization of RNA functions as a mechanism for post-transcriptional regulation. However, RNAflux conceptually introduces using RNA molecular coordinates as a latent layer of information encoding cellular space-time. Here, we used that latent layer of information to identify subcellular domains. As spatial omic technologies improve to capture more and more information, the potential applications of such latent embeddings will grow as well. Indeed at the tissue level, this concept is already being leveraged with a recent tool, TensionMap, using RNA localization information to predict mechanical tension[74]. As applications for spatial transcriptomics grow in popularity and complexity, we hope Bento is a platform for the tools needed to quantify the complex molecular dynamics governing normal and abnormal cellular processes.

1.5 Methods

1.5.1 MERFISH and seqFISH+ data preprocessing

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 needed to be removed. Again, cells without annotated nuclei were removed, leaving 1022 cells for downstream analysis.

1.5.2 RNAforest: model selection and training

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). While all other models use the 13 spatial features for input (Supp. Table 1), the CNN takes 64x64 image representations of each sample as input. 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 3). We retrained the final model (random forest base model) on all training data with the best performing set of hyperparameters (Supp. Fig. 1E).

1.5.3 RNAforest: Image rasterization of molecules and segmentation masks for CNN

To generate an image for a given sample, point coordinates, the cell segmentation mask and nuclear segmentation mask are used. The area of the cell is tiled as a 64 x 64 grid, where each bin corresponds to a pixel in the final image. Values are stored in a single channel to render a grayscale image. Pixels inside the cell are encoded as 20, inside the nucleus encoded as 40. Bins with molecules are encoded as (40 + 20 x n) where n is the number of molecules. Finally values are divided by 255 and capped to be between 0 and 1.

1.5.4 RNAforest: Simulating training data

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[56–59]. 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[58]. 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[58]. 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.

4. RNAforest: Manual annotation of validation data

    (a) 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).

    (b) We used Cohen's kappa coefficient[75] to calculate agreement between pairs of annotators for each label yielding an overall coefficient of 0.602.

    (c) 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 (Supp. Table 4).

1.5.5 RNAforest: Functional enrichment of gene pattern distributions

For enrichment of compartment-specific expression from Xia et al 2019[61], scores are calculated by taking the weighted sum of gene pattern frequencies and published compartment log fold-change values (Supp. Fig. 2). The Benjamini-Hochberg correction was used to correct p-values for multiple hypothesis testing. For the seqFISH+ dataset, we performed single-sample Gene Set Enrichment Analysis[76;77] on gene pattern frequencies to compute enrichment scores (Fig. 2I). ssGSEA was performed with the GSEApy Python package and the "GO_Cellular_Component_2021" gene set library curated by Enrichr[78]. Gene sets with a minimum size of 50 and a maximum size of 500 were analyzed.

1.5.6 Colocation quotient for RNA colocalization analysis

Pairwise colocalization of genes was determined for each compartment of every cell separately. In this case, each cell was divided into compartments, cytoplasm and nucleus. The colocation quotient (CLQ) was calculated for every pair of genes A and B. The CLQ is defined as an odds ratio of the observed to expected proportion of B transcripts among neighbors of A for a fixed radius r; it is formulated as:

$$CLQ_{A\to B} = \frac{C_{A\to B}/N_A}{N_B^A/N - 1}$$

Here $C_{A\to B}$ denotes the number of A transcripts of which B transcripts are considered a neighbor. N_A denotes the total number of A transcripts, while N_B stands for the total number of B transcripts. In the case that $A = B$, N_B equals the total number of B transcripts minus one. N denotes the total number of transcripts in the cell. Following statistical recommendations from the original formulation of the colocation quotient (CLQ), genes with fewer than 10 transcripts were not considered to reduce sparsity and improve testing power[68].

1.5.7 Tensor decomposition for compartment-specific colocalization

For tensor decomposition, we employed non-negative parallel factor analysis as implemented in Tensorly[49], which seeks to represent our dataset tensor X in a lower dimensional space of R signatures by decomposing X as the sum of R rank-one 3-way tensors. Each of these tensors

is described as the outer product of 3 vectors, xr, yr and zr. The collection of vectors across R signatures we denote as x^r (compartment loadings), y^r (cell loadings) and z^r (gene pair loadings) respectively. We find the optimal rank-R decomposition of X by minimizing reconstruction error as a function of the number of signatures R and use the elbow function heuristic to choose the best-fit across the range of 2-12 factors. Missing values are ignored when calculating the loss.

$$X = \sum_{r=1}^{R} x^r y^r z^r$$

1.5.8 RNAflux: Unsupervised spatial embedding and subcellular domain quantization

To generate RNAflux embeddings, first a set of query coordinates are generated tiling across the cell area on a uniform grid. This effectively downsamples the original data units (pixels) resulting in much fewer samples to compute embeddings. For the MERFISH U2-OS dataset, a step size of 10 data units (pixels) was used to generate the uniform grid. Each query point is assigned an expression vector, counting the abundance of each gene within a fixed radius of 40 and 50 data units respectively. Each expression vector is normalized to sum to one, converting the expression vector to a composition vector. Similarly, the cell composition vector is calculated by normalizing the total cell expression to sum to one. The RNAflux embedding at a given query coordinate is defined as the difference between the query composition and its corresponding cell composition, divided by the standard deviation of each feature within each cell.

The RNAflux embedding serves as an interpretable spatial gene embedding that quantifies highly local fluctuations in gene composition. Dimensional reduction of the embeddings is performed using truncated singular value decomposition (SVD). Truncated SVD was chosen over PCA to better handle large but sparse data. Embeddings were reduced to the top 10 components. To assign domains, self-organizing maps (SOM) were used for low-rank quantization of query embeddings. In analysis of the MERFISH dataset, SOMs of size 1 x k were fit across a range of 2 to 12; the best model was determined using the elbow method heuristic to evaluate quantization error. Similarly, domains were determined for the cardiomyocytes spatial transcriptomics data by fitting the vehicle and treatment samples separately, for k across a range of 2 to 8. The

elbow method heuristic determined an optimal k of 6; subsequently a k of 4 was used for further analysis for ease of interpretation.

1.5.9 RNAflux: Visualizing spatial embeddings

The top 3 principal components of the RNAflux embeddings are transformed to map to red, green and blue values respectively. Embeddings are first quantile normalized and scaled to a minimum of 0.1 and 0.9 to avoid mapping extreme quantiles to white and black. These values are then used for red, green, and blue color channels. To map the downsampled grid back to the original data units, linear interpolation was used to rescale the computed color values and fill the space between the uniform grid points.

1.5.10 RNAflux: Enrichment of locale-specific transcriptomes derived by APEX-seq

The enrichment score for each pixel is calculated by first taking the weighted sum of its RNAflux embedding and locale-specific log fold-change values as implemented by the decoupler tool[79]. Scores for pixels within a given cell are normalized against a null distribution constructed via random permutations of the input embeddings, to produce z-scaled enrichment scores. Fluxmap domain enrichment scores are simply obtained by taking the mean score of all pixels within the boundary of each domain. Fluxmap domain overlaps are computed by counting the fraction of pixels within the boundary of each domain with a positive enrichment score.

1.5.11 MERFISH of U2-OS cells

MERFISH sample preparation. MERFISH measurements of 130 genes with five non-targeting blank controls was done as previously described, using the published encoding[60] and readout probes[80]. Briefly, U2-OS cells were cultured on 40 mm #1.5 coverslips that are silanized and poly-L-lysine coated[60] 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 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 uM 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 werZe finally stained with a Alexa 488-conjugated anchor probe-readout oligo (Integrated DNA Technologies) and DAPI solution at 1 ug/ml.

MERFISH imaging. MERFISH measurements were conducted on a home-built system as described in Huang et al. 2021[80].

MERFISH spot detection. Individual RNA molecules were decoded in MERFISH images using MERlin v0.1.6[9]. 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 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.

MERFISH image segmentation. Cellpose v1.0.2[81] 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.

1.5.12 Data Availability

Preprocessed and raw datasets have been deposited at https://doi.org/10.6084/m9.figshare.c.6564043.v1 and are accessible through the Bento Python package. These include the seqFISH+[37], MERFISH, and Molecular Cartography datasets. Raw MERFISH and Molecular Cartography data is available upon request.

1.5.13 Code Availability

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

1.5.14 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 of 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.

1.5.15 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., N.A. and G.P. performed data preprocessing and analysis. A.M., C.K., Y.H., and Q.Z. generated the MERFISH experiment. N.L. designed the gene panel and cultured the cardiomyocytes. 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.

1.5.16 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.

Chapter 2

Doxorubicin-induced stress in cardiomyocytes results in RNA localization changes

2.1 Introduction

Doxorubicin (DOX) was once one of the most effective broad-spectrum anti-cancer anthracycline antibiotics[82;83] with particular efficacy against solid malignancies such as lung and breast cancer, as well as hematologic neoplasia[84;85]. However, DOX's propensity to cause cardiac damage in patients has led to significant limitations in its clinical use[86]. There are two known mechanisms of action by which DOX acts in cells[87]: generation of reactive oxygen species via potential interactions with oxidation reaction pathways which then damage lipid membranes, disrupt mitochondrial function, induce DNA damage and triggers apoptotic pathways; and direct interaction with DNA topoisomerase II to induce single-stranded and double-stranded breaks. The exact mechanism by which DOX induces heart failure is unclear, but significant evidence suggests cardiomyocyte injury driven by oxidative stress as a major factor[84;88–91]. Specifically, DOX causes stress and dysfunction in multiple cellular compartments in cardiomyocytes such as mitochondria, Sarco/endoplasmic reticulum (SER), deficiencies in calcium signaling, and lipid degradation at the cellular membrane[92]. There is growing evidence that DOX not only interacts with DNA, but also with some affinity to double-stranded RNAs[93], rRNAs[94] and RNA aptamers[95].

Having established Bento's utility to characterize RNA localization in cell lines (see Chapter 1), we applied Bento to doxorubicin-treated and untreated cardiomyocytes, a cell line model for these cardiomyopathies. We performed single molecule spatial transcriptomics

(Molecular Cartography) on doxorubicin-treated and untreated cardiomyocytes to measure consequential differences across multiple classes of phenotypes in a single experiment: RNA localization, gene expression, cell morphology.

2.2 Results

We designed a panel of 100 genes to profile with spatial transcriptomics, capturing pathways for cardiomyocyte health and function[24]. These include genes involved in cardiomyocyte contraction and conduction; cellular cytoskeletal pathways including myofibril assembly and cytoskeleton components; and also mitochondrial function to capture perturbations to oxidative metabolism. We reasoned that we could recapitulate known dysfunction of subcellular domains in cardiomyocytes upon DOX stress and measure novel RNA localization phenotypes that are not explained by expression changes alone.

We utilized a chemically-defined protocol to differentiate human induced pluripotent stem cells (iPSCs) into beating cardiomyocytes and treated them with either DMSO (vehicle) or 2.5 uM DOX for twelve hours, 24 hours, or 48 hours before fixation (Methods). Each treatment had 2 replicates. Single molecule spatial transcriptomes were measured by Resolve Bioscience using Molecular Cartography. The resulting data was segmented using ClusterMap[42] for cell boundaries and Cellpose[81] for nuclei boundaries. Non-myocytes were filtered out using SLC8A1 as a canonical marker for cardiomyocytes (Methods, Supp. Fig. 1A).

Comparing vehicle and DOX treated cardiomyocytes, we found vehicles cells to cluster distinctly from all DOX treated cells (Fig. 1) and DOX treated cells forming a duration-dependent expression gradient from 12-48 hours. Notably, transcript density i.e. transcript count dividedby cell area, decreases with treatment duration. Differential expression analysis of each timepoint relative to vehicle indicate that DOX induces cellular stress as expected. NPPA and NPPB are important biomarkers in clinical cardiology that become upregulated during cardiac stress[96;97]. Elevated levels of NPPB have been used to diagnose patients with doxorubicin induced cardiotoxicity and elevated levels of NPPB also correlate with severity of heart failure. An increase in NPPA and NPPB levels upon Doxorubicin exposure at 24 and 48 hours indicates that the cardiomyocytes have transitioned to a state of cellular stress (Fig. 2A,B).

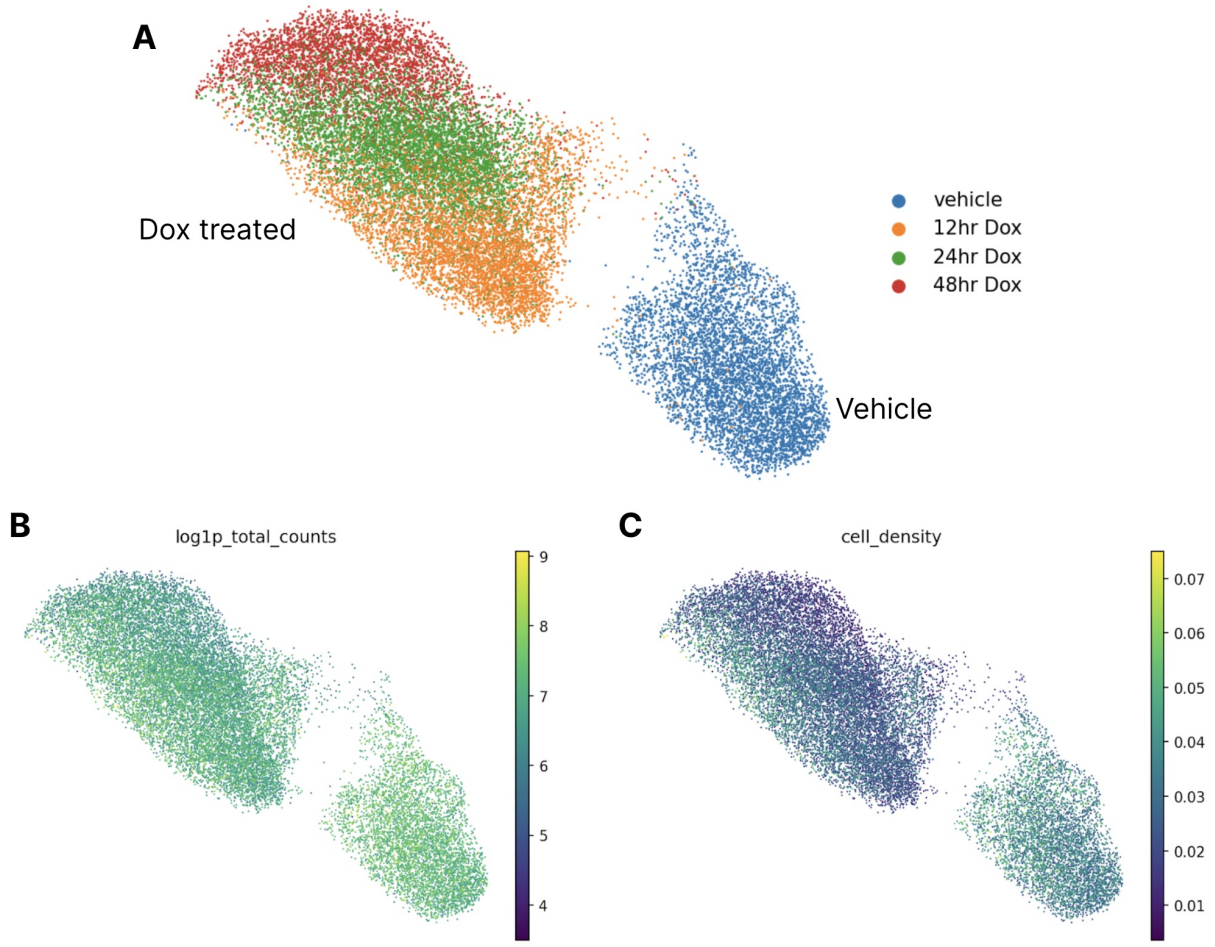

Figure 2.1. Single-cell expression of doxorubicin-treated caradiomyocytes time points A. UMAP projection of single-cell expression of vehicle and doxorubicin-treated cardiomyocytes across vehicle, 12 hour, 24 hour, and 48 hour time points. Two replicates per time point. B. UMAP projection of single-cells colored by log-scaled total RNA expression and C. transcript density (transcript count divided by cell area).

We then restricted our analysis to focus on the 12 hour treatment and vehicle for spatial analyses via Bento. Poor segmentation quality for the other timepoints limited the precision and accuracy of 2-dimensional spatial analysis algorithms. We identified subcellular domains in vehicle and 12 hour DOX treated cardiomyocytes using RNAflux, clustering the domains into four fluxmap domains (Fig. 2C). Enrichment of location-specific gene expression aligned domains to the nucleus (nuclear pore, nucleolus, and nucleus), ERM and OMM, ER lumen, and cytosol respectively (Fig. 2C & D, Supp. Fig. 1C). Comparing the gene composition in each domain, we observe an overall localization bias towards both the nucleus and ERM/OMM in vehicle treated cells (**Fig. 2E top**), in agreement to prior poly(A) smFISH studies[98]. However, RNA in the DOX treated cardiomyocytes demonstrated a shift in average RNA localization away from the ERM/OMM and towards the nucleus (**Fig. 2C bottom**). There is evidence that 90% of genes have a half life of less than 260 minutes[99], far less than the 12 hour DOX treatment, indicating that the shift in RNA localization is likely due to reduced nuclear export of newly synthesized RNA from the nucleus to the ERM/OMM. Indeed, even low concentrations of DOX have been demonstrated to alter structural fibrous proteins as well as mitochondrial depolarization and fragmentation[100]. Of particular note, the RNA binding protein RBM20 – a critical regulator of mRNA splicing of genes encoding key structural proteins associated with cardiac development and function – had a pronounced depletion of RNA transcripts outside of the nucleus upon DOX treatment (Fig. 2F). With further validation, this may indicate nuclear retention and or degradation of nuclear exported RBM20 mRNA as a potential mechanism of DOX induced cardiomyopathy. Similarly, we found the mRNA of calcium voltage-gated channel subunit CACNB2 to also deplete outside of the nucleus (Fig. 2G). The loss of CACNB2 translation outside of the nucleus may impact calcium signaling crucial to cardiomyocyte function[101].

2.3 Discussion

In this study of DOX-induced stress in cardiomyocytes, we utilized single-molecule spatial transcriptomics to identify changes in both gene expression and subcellular RNA localization resulting from DOX treatment. Of particular interest was the RNA binding protein RBM20, whose extranuclear depletion in mRNA represents a potential target for therapeutic intervention.

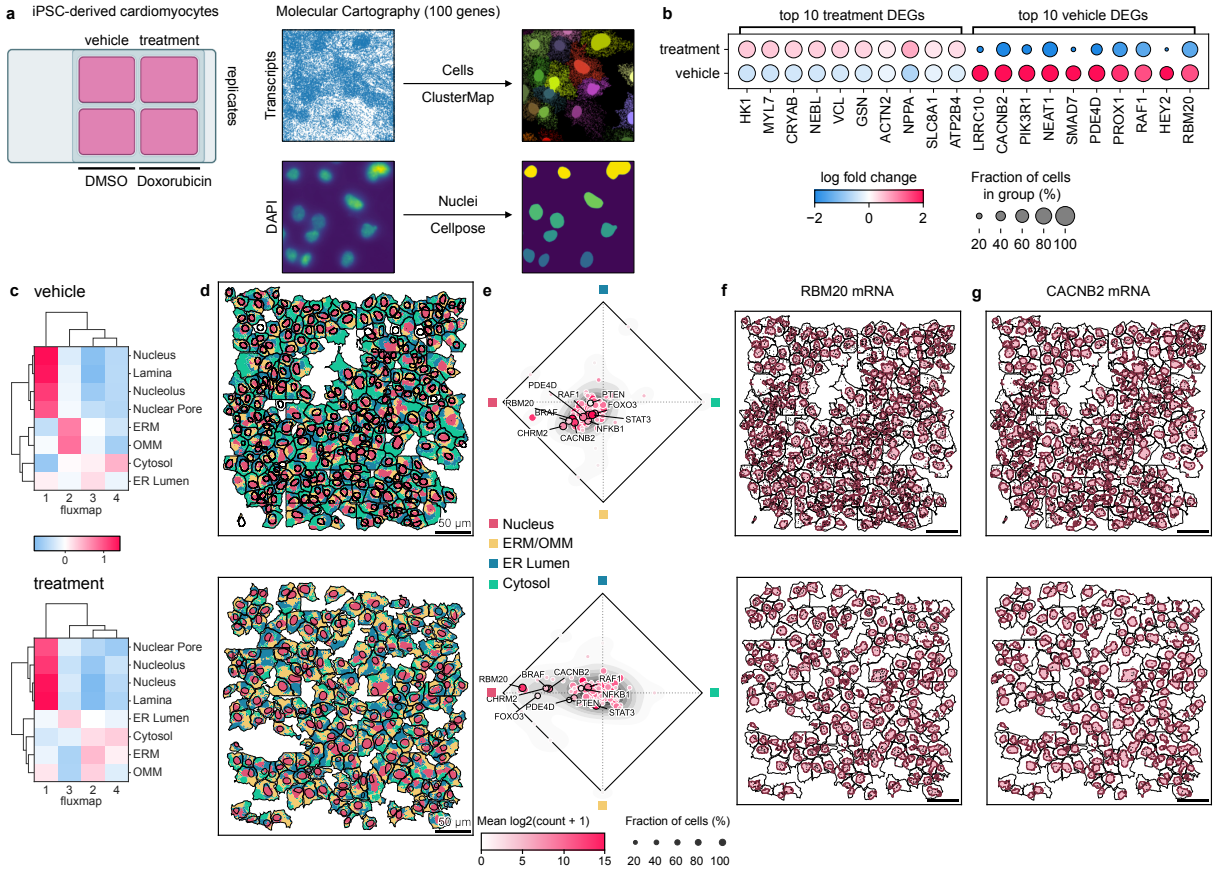

Figure 2.2. Subcellular RNA localization changes upon Doxorubicin treatment in iPSC-derived cardiomyocytes A. Cardiomyocytes derived from human iPSCs were treated with DMSO or 2.5 uM DOX for 12 hours. The localizations of 100 genes relevant to cardiomyocyte health and function were measured using Molecular Cartography. Cell boundaries were determined using ClusterMap and nuclei were segmented using Cellpose. B. Top 10 upregulated and downregulated genes in vehicle versus treatment. C. APEX-seq location-specific gene enrichment of fluxmap domains for the cytosol, endoplasmic reticulum membrane (ERM), endoplasmic reticulum lumen (ER Lumen), nuclear lamina, nucleus, nucleolus, nuclear pore and outer mitochondrial matrix (OMM). D. Fluxmap domains visualized for a representative field of view of cardiomyocytes for vehicle and treatment respectively highlighting cellular nuclei, ERM/OMM, ER Lumen, and cytosol. E. RNAflux fluxmap enrichment of each gene averaged across vehicle and treatment cardiomyocytes captures changes in subcellular RNA localization. Visualization of RBM20 F. and CACNB2 G. confirms the depletion of transcripts from the perinuclear and cytosolic compartments of cardiomyocytes upon DOX treatment.

This localization behavior may be an early consequence leading to the functional mis-splicing of RBM20's cardiomyopathy-associated targets. Sequestration of the mRNA may be an indirect mechanism of down-regulation. The ERM-associated fluxmap seems to be relatively larger in DOX treated cells compared to vehicle, suggesting that remodeling of organelles may drive movement of molecules or vice versa.

We found that the 2D spatial resolution of molecular coordinates and segmentation data to limit the clarity of our analyses. While the Clustermap based cell segmentation was sufficient to approximate subcellular domains with RNAflux in the vehicle and 12 hour treatment samples, many regions of the 24 hour and 48 hour treatment samples have denser cells that sit on top of one another due to tighter cell seeding densities. As a result, molecular coordinates and segmentations were flattened to two dimensions, making it impossible to disambiguate expression patterns from overlapping cells. We foresee that better resolution and 3D compatible segmentation algorithms will alleviate this in the future. This is likely required for analysis to achieve subcellular resolution in more complex systems e.g. tissue slices and organoids.

Due to the targeted nature of the particular spatial transcriptomics platform, the 100 gene panel is biased for genes annotated for cardiac function, limiting discovery of novel targets. Expanding the panel size would allow us to capture a better picture of spatial perturbations to the transcriptomic landscape. Additionally, generalizing spatial analyses in Bento from 2D to 3D would enable finer segmentation of subcellular compartments and cells, in turn improving RNA localization analysis. Overcoming these challenges will be useful not only for enabling spatial analysis to other cell lines and conditions, but also to even more heterogeneous systems such as tissue.

2.4 Methods

2.4.1 Preprocessing cardiomyocytes datasets

Single-cell expression matrices of both vehicle replicates and both DOX treatment samples were concatenated as a single expression matrix. Cells were projected into two dimensions with UMAP dimensional reduction. No significant batch effects were detected. Leiden clustering was performed at resolution=0.5 to isolate and filter out a non-myocyte population depleted

in SLC8A1 expression (Supp. Fig. 1A). All described preprocessing steps were performed in Scanpy[41].

2.4.2 RNAflux: Unsupervised spatial embedding and subcellular domain quantization

For the iPSC-derived cardiomyocytes, a step size of 5 data units was used to compute RNAflux embeddings. Visualization and enrichment of locale-specific transcriptomes derived by APEX-seq were performed as described in Chapter 1.

2.4.3 Molecular Cartography

Cultured cell processing. After Doxorubicin treatment, cardiomyocytes were washed with PBS (1x) twice and fixed in Methanol (-20°C) for 10 min. After fixation, Methanol was aspirated and cells were dried and stored at -80°C until use. The samples were used for Molecular CartographyTM (100-plex combinatorial single molecule fluorescence in-situ hybridization) according to the manufacturer's instructions Day 1: Molecular Preparation Protocol for cells, starting with the addition of buffer DST1 followed by cell priming and hybridization. Briefly, cells were primed for 30 minutes at 37°C followed by overnight hybridization of all probes specific for the target genes (see below for probe design details and target list). Samples were washed the next day to remove excess probes and fluorescently tagged in a two-step color development process. Regions of interest were imaged as described below and fluorescent signals removed during decolorization. Color development, imaging and decolorization were repeated for multiple cycles to build a unique combinatorial code for every target gene that was derived from raw images as described below. Probe Design. The probes for 100 genes were designed using Resolve's proprietary design algorithm. Briefly, the probe-design was performed at the gene-level. For every targeted gene, all full-length protein coding transcript sequences from the ENSEMBL database were used as design targets if the isoform had the GENCODE annotation tag 'basic'[102;103]. To speed up the process, the calculation of computationally expensive parts, especially the off-target searches, the selection of probe sequences was not performed randomly, but limited to sequences with high success rates. To filter highly repetitive regions, the abundance of k-mers was obtained from the background transcriptome using Jellyfish[104]. Every target sequence was scanned once

for all k-mers, and those regions with rare k-mers were preferred as seeds for full probe design. A probe candidate was generated by extending a seed sequence until a certain target stability was reached. A set of simple rules was applied to discard sequences that were found experimentally to cause problems. After these fast screens, the remaining probe candidates were mapped to the background transcriptome using ThermonucleotideBLAST[105] and probes with stable off-target hits were discarded. Specific probes were then scored based on the number of on-target matches (isoforms), which were weighted by their associated APPRIS level[106], favoring principal isoforms over others. A bonus was added if the binding-site was inside the protein-coding region. From the pool of accepted probes, the final set was composed by picking the highest scoring probes. Probes with catalog numbers can be found in Supp. Table 1[24].

Imaging. Samples were imaged on a Zeiss Celldiscoverer 7, using the 50x Plan Apochromat water immersion objective with an NA of 1.2 and the 0.5x magnification changer, resulting in a 25x final magnification. Standard CD7 LED excitation light source, filters, and dichroic mirrors were used together with customized emission filters optimized for detecting specific signals. Excitation time per image was 1000 ms for each channel (DAPI was 20 ms). A z-stack was taken at each region with a distance per z-slice according to the Nyquist-Shannon sampling theorem. The custom CD7 CMOS camera (Zeiss Axiocam Mono 712, 3.45 um pixel size) was used. For each region, a z-stack per fluorescent color (two colors) was imaged per imaging round. A total of 8 imaging rounds were done for each position, resulting in 16 z-stacks per region. The completely automated imaging process per round was realized by a custom python script using the scripting API of the Zeiss ZEN software (Open application development).

Image Processing and Spot Segmentation. As a first step all images were corrected for background fluorescence. A target value for the allowed number of maxima was determined based upon the area of the slice in um^2 multiplied by the factor 0.5. This factor was empirically optimized. The brightest maxima per plane were determined, based upon an empirically optimized threshold. The number and location of the respective maxima was stored. This procedure was done for every image slice independently. Maxima that did not have a neighboring maximum in an adjacent slice (called z-group) were excluded. The resulting maxima list was further filtered in an iterative loop by adjusting the allowed thresholds for (Babs-Bback) and (Bperi-Bback)

to reach a feature target value (Babs: absolute brightness, Bback: local background, Bperi: background of periphery within 1 pixel). This feature target values were based upon the volume of the 3D-image. Only maxima still in a zgroup of at least 2 after filtering were passing the filter step. Each z-group was counted as one hit. The members of the z-groups with the highest absolute brightness were used as features and written to a file. They resemble a 3D-point cloud. To align the raw data images from different imaging rounds, images had to be registered. To do so the extracted feature point clouds were used to find the transformation matrices. For this purpose, an iterative closest point cloud algorithm was used to minimize the error between two point-clouds. The point clouds of each round were aligned to the point cloud of round one (reference point cloud). The corresponding point clouds were stored for downstream processes. Based upon the transformation matrices the corresponding images were processed by a rigid transformation using trilinear interpolation. The aligned images were used to create a profile for each pixel consisting of 16 values (16 images from two color channels in 8 imaging rounds). The pixel profiles were filtered for variance from zero normalized by total brightness of all pixels in the profile. Matched pixel profiles with the highest score were assigned as an ID to the pixel. Pixels with neighbors having the same ID were grouped. The pixel groups were filtered by group size, number of direct adjacent pixels in group, number of dimensions with size of two pixels. The local 3D-maxima of the groups were determined as potential final transcript locations. Maxima were filtered by the number of maxima in the raw data images where a maximum was expected. Remaining maxima were further evaluated by the fit to the corresponding code. The remaining maxima were written to the results file and considered to resemble transcripts of the corresponding gene. The ratio of signals matching to codes used in the experiment and signals matching to codes not used in the experiment were used as estimation for specificity (false positives). The algorithms for spot segmentation were written in Java and are based on the ImageJ library functionalities. Only the iterative closest point algorithm is written in C++ based on the libpointmatcher library (https://github.com/ethz-asl/libpointmatcher).

Image segmentation. Cellpose v1.0.2[81] was used to perform image segmentation to determine the boundaries of 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 determined with ClusterMap[42] using spot coordinates.

2.4.4 iPSC Cardiac Differentiation and Doxorubicin Treatment

Matrigel (Corning, cat # 354277) coated plates were used to culture iPSCs with mTESR Plus human iPSC medium (StemCell Technologies, cat # 100-0276) in a humidified incubator at 37°C with 5% CO2. iPSCs were dissociated with Gentle Cell Dissociation Reagent (StemCell Technologies, cat # 100-0485) and passaged with mTESR Plus medium and 10uM ROCK inhibitor (Tocris, cat #1254) at a ratio of 1:12. mTESR plus medium was replaced every other day until the cells reached 80% confluency for maintenance and replating, or 90% confluency for cardiac differentiation utilizing a chemically defined protocol[107]. On day 0 of cardiac differentiation, cells were treated with 6uM CHIR99021 (Selleck Chem, cat # S1263) in RPMI 1640 media (Gibco, cat # 11875) and B27 minus insulin supplement (Thermo Fisher, cat # A1895601). On day 2, CHIR was removed, and cells were cultured with RPMI 1640 media and B27 minus insulin supplement (Thermo Fisher, cat # A18956). On day 3, media was replaced with RPMI media containing B27 minus insulin supplement and 5 uM Wnt-C59 (Cellagen Technologies, cat # C7641-2s). On days 5, 7, and 9, media was replaced with RPMI media containing B27 insulin supplement (Thermo Fisher, cat # 17504). On days 11 and 13, media was replaced with RPMI 1640 media without glucose (Thermo Fisher, cat # 11879020) containing B27 insulin supplement for purification of cardiomyocytes. From days 15 onward, the cells were cultured in RPMI 1640 media containing B27 supplement which was changed every other day until the cells reached day 30 for replating. For replating, cells were incubated in 10X TrypLE (Thermo Fisher, cat # A1217701) for 12 minutes at 37 C, neutralized with equal volumes of RPMI 1640 media containing B27 supplement with 20% FBS (Gibco, cat # 26140-079), gently dissociated by pipetting, then spun down and resuspended for replating in RPMI 1640 media containing B27 supplement with 20% FBS. The next day, the cell media was replaced with RPMI 1640 media containing B27 supplement which was replaced with fresh media every other day. On day 48 the cells were replated onto chamber slides (Ibidi, cat # 80826) as described above and recovered for 10 days before doxorubicin treatments began (MedChemExpress, cat # HY-15142). On day 60, doxorubicin treatments concluded, and the cells underwent methanol fixation.

2.4.5 Data Availability

Preprocessed data for Molecular Cartography profiled cardiomyocytes is deposited at https://doi.org/10.6084/m9.figshare.c.6564043.v1 and is accessible through the Bento Python package.

2.4.6 Code Availability

Analysis code for generating figures can be found at: https://github.com/ckmah/bento-manuscript.

2.4.7 Acknowledgements

As this work was derived from the same manuscript, see 1 for corresponding details.

2.4.8 Author Contributions

As this work was derived from the same manuscript, see 1 for corresponding details.

2.4.9 Competing Interests

As this work was derived from the same manuscript, see 1 for corresponding details.

Chapter 3

Spotfish: A modular framework for decoding spatial imaging data

3.1 Background

Image-based spatial transcriptomics is a rapidly evolving field that seeks to map the spatial distribution of RNA molecules in situ. These technologies have enabled researchers to study the spatial organization of cells and tissues at unprecedented resolution, with the potential to uncover novel biological insights.The field has seen a surge in interest in recent years, with the development of several novel technologies, including MERFISH[4], seqFISH[108], STARmap[109], ISS[110], and Slide-seq[111]. Despite their varied underlying technologies, they consistently share the same backbone with a unified objective: reporting the location and identity of individual RNA molecules. A significant challenge in this domain is the difficulty in validating the quality of image analysis outputs. This issue is confounded by the lack of standardized data quality metrics accepted by researchers. While existing pipeline development tools for spatial transcriptomics image analysis are functional, they often suffer from limited scalability, a lack of interoperability with newer methodologies, and restricted portability across different computing environments. Recognizing these limitations, there emerges a clear need for an unbiased framework to build spatial transcriptomics pipelines.

To address these challenges, I developed spotfish, a modular pipeline building framework that abstracts the series of tasks for processing spatial transcriptomics data by standardizing inputs and outputs between workflow tasks. This allows swapping in new tools as new alternatives are published frequently, by wrapping the chosen tool for compatible data formats. It also encourages

reporting quality metrics for diagnosing data quality and evaluating the performance of chosen tool/parameter combinations. The framework is built using Nextflow, a robust workflow language specifically built for and heavily adopted by bioinformatics researchers[112]. Nextflow also abstracts how the pipelines are executed on different computing environments, e.g. locally, on compute clusters, or various cloud services, meaning spotfish is inherently usable for researchers regardless of computational environment. In contrast to starfish's approach to programmatic Python-based pipeline construction, spotfish modules are programming language agnostic through the use of containerization for each step of the process. Additionally, the pipeline prioritizes usage of open file formats supported by the Open Microscopy Environment[113] (OME) to ensure transparency and compatibility with the rich ecosystem of bio-imaging analysis tools. Spotfish is guided by FAIR principles[6] and utilize modern open-source standards, ensuring accessibility for a broad spectrum of users, from researchers and core facilities to technologists.

3.2 Properties of multiplexed transcriptomics imaging data

The raw data consists of a large mosaic of microscopy images. To resolve individual fluorescent molecules, images are taken at roughly to 10-100 times magnification. A single image, or field of view, contains roughly 1-100 cells depending on the magnification and cell type. The same field of view is then imaged multiple times, in which the spectrum of light is limited to specific frequency bands at each iteration. This allows us to assign a unique combination of fluorescent probes that emit light at unique frequency bands to each target. The focal plane is much narrower than the height of cells as a consequence of the high magnification, which requires imaging the same field of view multiple times at different z-planes to maximize the volume interrogated. This process is usually repeated for a grid of positions to capture a larger cumulative area of the sample, which can produce upwards of a terabyte of data per experiment. The challenge arises due to the multi-dimensional nature of these measurements, across spatial dimensions x, y and z, channels (multiple laser wavelengths), and rounds (repeated imaging with different combinations of probes).

3.3 Framework Design

Imaging-based acquisition of spatial transcriptomics data requires coordinating a series of tasks, including image stitching, registration, background correction, spot detection, and finally barcode decoding to produce labeled molecular coordinates corresponding to a predesigned set of gene targets. Every task can be accomplished with existing tools, but it remains difficult to perform an end-to-end analysis outside of tailored pipelines without significant data wrangling. Even though many studies utilize commercial platforms for spatial transcriptomics, only processed data is usually available to the customer. Their analysis pipelines remains proprietary and blackbox, forcing users to rely on arbitrarily defined quality metrics. In contrast, starfish is a open-source unified pipeline framework implemented in Python, abstracting processing steps using a object-oriented programming design[10]. It is extremely flexible and has accommodated data processing for 7 different technologies. However, the API's steep learning curve, many parameters, and lack of maintenance makes it difficult to build pipelines and integrate cutting-edge tools without significant refactoring of the tool or starfish itself. There has only been one prominent third-party contribution by a recent development that integrated their novel barcode decoding algorithm, CheckAll[11].

Spotfish abstracts pipelines into two subworkflows: image registration and spot analysis. Image registration is a common image processing step not unique to spatial transcriptomics and is usually required for acquisitions across multiple fields of view. By decoupling this step, it is convenient to adapt tools outside of the immediate domain. The spot analysis subworkflow encompasses two tasks, spot detection and barcode decoding to produce molecular coordinate tables with target i.e. gene annotations. This can then be combined with cell and nuclear segmentation data for functional analysis in other software, such as Bento[24], Squidpy[16] and Scanpy[41].

3.4 Case Study: 69-bit MERFISH of U2-OS Cells

To demonstrate the utility of spotfish, I applied it to previously published 69-bit MERFISH dataset of U2-OS cells designed to target 10,000 genes with a minimum hamming distance 4

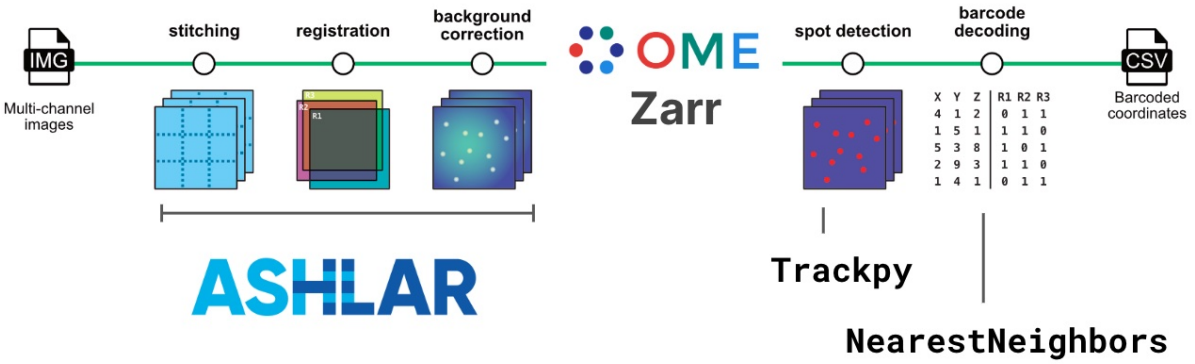

Figure 3.1. Spotfish workflow Overview of the analysis tasks in the spotfish workflow from left to right. Image registration encompasses stitching, registration and background correction (as needed). Intermediate output format of OME-ZARR. Spot analysis encompasses spot detection and barcode decoding outputing coordinate tables with annotations. Tools used to implement each subworkflow indicated below.

(HD4) encoding scheme[61]. Because the experiment had a total of 72 rounds of imaging, it was suitable for testing the scalability of the the spotfish framework. Commercial platforms reportedly use 16 rounds (CosMx[50]) and 15 rounds (Xenium[51]) to decode 980 and 313 targets respectively. This meant the 10k MERFISH dataset uses roughly 4.5 times more imaging rounds than the current largest commercial platforms. The specific tools implemented in the pipeline were chosen based on their performance on the 10k MERFISH dataset. The image registration step was performed using Ashlar[114], a software package originally designed for stitching and aligning multiplexed immunofluorescent samples acquired via cyclical imaging and tile-scanning. The spot detection step was performed using trackpy[115], a Python library for particle tracking in 2D and 3D. The barcode decoding step was performed using the nearest neighbor approach implemented in the scikit-learn analysis package[116].

Spotfish allowed parallelization of the image registration step across all 69 rounds with Ashlar. This step was ultimately limited by memory not computing speed, using 30 GB to stitch each round in 16 cpu hours. The output of the image registration step was a 1.5 TB OME-ZARR file storing a multi-scale representation of the image data registered to the same coordinate system. By standardizing the output format, this eliminates the need for storing additional metadata defining tile positions and channel orders. The spot detection step was performed using trackpy, which was able to detect 477 million spot coordinates in 3 dimensions.

The barcode decoding step was performed using the nearest neighbor approach implemented in the scikit-learn analysis package. Without spotfish, the total compute time is estimated to be 1248 hours whereas parallelization reduced runtime over 26-fold, to 39 hours with 32 parallel processes for spot detection and 8 parallel processes for spot calling. The output was then visualized using Napari[117] to assess the quality of the data interactively.

3.5 Conclusion

All together, I demonstrate spotfish's flexibility in integrating heterogeneous set of tools to build a scalable image analysis pipeline for spatial transcriptomics. By adhering to open file formats and containerization, spotfish addresses existing technologies and is well positioned to adapt to future advancements for image registration, spot detection and barcode decoding. Future work will focus on creating quality control modules that provide quantitative metrics for assessing the quality of the data at each step of the pipeline. This will allow researchers to identify the optimal tool and parameter combinations for their data. To facilitate open discussion of spotfish's development, I aim to collaborate with the nf-core community, a consortium of bioinformatics researchers that develop and maintain a collection of high quality modular bioinformatics pipelines. This will ensure that spotfish is well maintained and accessible to the community. Finally, I will continue to develop spotfish to support additional spatial transcriptomics technologies, such as seqFISH and STARmap. This will allow researchers to compare the performance of different technologies on the same dataset, and to integrate data from different technologies for meta-analysis.

Epilogue

4.1 Conclusion

In this dissertation, I have presented a series of computational methods to analyze spatial transcriptomics data. I began by developing Bento, a computational framework for subcellular analysis of spatial transcriptomics data. This work is one of the first to leverage the spatial resolution of imaging-based spatial transcriptomics data to study subcellular RNA localization. I demonstrated the utility of Bento by applying it to a variety of spatial transcriptomics datasets, including cardiomyocytes to study changes in RNA localization as scale. Unexpectedly, we found several genes mislocalized to the nucleus as a result of doxorubicin treatment, including RBM20 and CACNB2, suggesting that RNA localization is an underappreciated cell phenotype that has the potential to uncover functional biology. To lower the barrier to functional analysis of spatial transcriptomics datasets, I also created spotfish, a modular framework for decoding spatial imaging data. This framework is designed to be flexible, scalable, and interoperable with existing tools. I demonstrated the utility of spotfish by applying it to a 69-bit MERFISH dataset of U2-OS cells. The framework is aimed to be a community resource for building spatial transcriptomics pipelines, and I hope to eventually collaborate with the nf-core community to ensure that spotfish is well maintained and accessible to the community.

4.2 Limitations and Future Directions

There are a great deal of challenges with the current generation of spatial transcriptomics data. During my graduate work, it was important to me that I focus on core problems that reveal fundamental biology, not a transient technical property of any one technology. For example, the most popular commercial spatial methods are slide-based capture assays paired with traditional

sequencing for comprehensive transcriptome profiling. However, compared to imaging-based approaches which have single-molecule resolution, each spatial location on the assay captures several to tens of cells depending on the technology. This loss in fidelity has spawned an entire subfield of deconvolution methods, specifically to estimate properties of each spatial location such as the proportion of cell types, the expression of cells given predicted cell types, technical dropout of expression, etc. These techniques may be useful now, but are ultimately tied to a specific iteration of rapidly evolving spatial transcriptomic technologies. Instead, I chose to tackle problems initially hampered by the lack of tangible datasets; the recent availability of public datasets has indeed lowered the barrier to method development. The increasing throughput of new technologies such as Xenium from 10x Genomics and STOmics from BGI Genomics will only improve our ability to draw biological insights at the molecular resolution. Similarly, the imminent move towards multi-omics spatial imaging will enable us to capture more snapshots of the RNA life cycle than ever before.

While the current functionality of Bento is limited to 2-dimensional spatial analysis, the obvious extension to 3 dimensions will enable subcellular analysis in biological systems more complex than monolayer cell cultures such as tissue slices and organoids. This will also open the door to exploring true physical molecular gradients in their natural 3 dimensions. While we showed its value to discover spatial subcellular domains, one can imagine gradient shifts between cells e.g. at the cell membrane, tight junctions, synapses etc. or even in the extracellular matrix characterizing cell signaling molecules. These are some of the functional biology questions waiting to be explored with spatial transcriptomics data. Bento is also capable of measuring morphological phenotypes; paired with the appropriate experimental design, spatial transcriptomics will be a powerful tool to interrogate the relationship between cell states, RNA localization, and cell morphology. These are especially relevant in developmental biology, neurological diseases, and cancer where changes to cell shapes and molecular condensates are frequently measured already. Algorithmic improvements will also require computational scalability, which I hope to address by interfacing with the global research community, including the Scverse Foundation[118] developers and the nf-core project members. Looking forward, this will manifest as integration with new open-source data standards, such as SpatialData and distributed computing with Dask.

4.3 Closing thoughts

The field of spatial transcriptomics is still in its infancy, and there are many exciting opportunities for future computational work. I believe the most impactful innovations will come from other fields, such as computer vision and genomics. Deep learning has already made its mark in both fields to accomplish everything from self-driving cars to functional genomics with DNA large language models, with the potential to bridge the gap between imaging and sequencing. I am hopeful for the creativity in this field and am excited to see new applications beyond tissue atlases and drug screening platforms. I hope that my work will contribute to the growing body of open-source tools and resources for spatial transcriptomics, and that it will inspire others to do the same.

Appendix A

Supplemental Material for Chapter 1

All Supplemental Tables can be found in the version 2 bioRxiv preprint at the following location: https://www.biorxiv.org/content/10.1101/2022.06.10.495510v2.supplementary-material.

A.1 Supplementary Figures

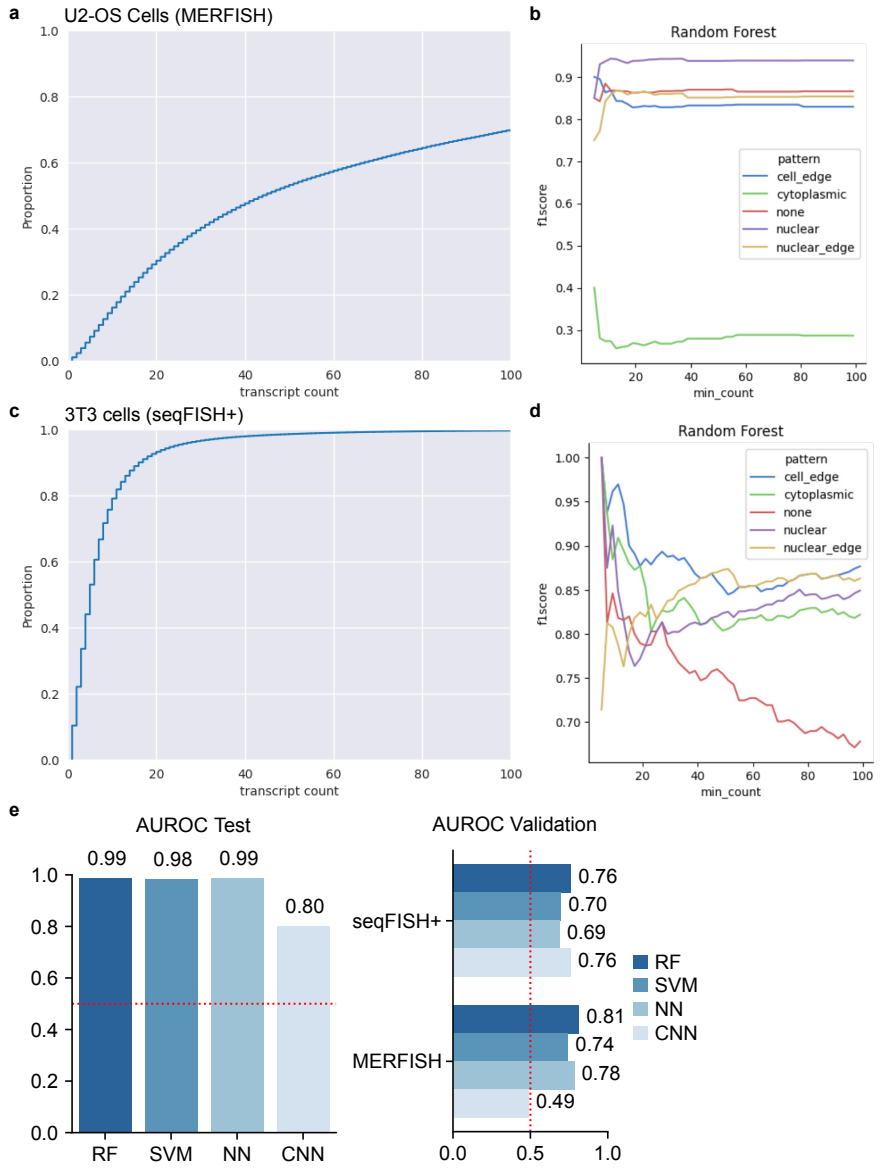

Figure A.1. RNAforest performance evaluation. A. Cumulative distribution of sample molecule copy number in U2-OS cells MERFISH dataset. B. Validation F1-score of each binary classifier in RNAforest as a function of sample molecule copy number for MERFISH dataset. C. Cumulative distribution of sample molecule copy number in 3T3 cells seqFISH+ dataset. D. Validation F1-score of each binary classifier in RNAforest as a function of sample molecule copy number for seqFISH+ dataset. E. Benchmarking performance of the 4 base models (RF - random forest, SVM - support vector machine, NN - fully connected neural network, CNN - convolutional neural network), showing AUROC in test and validation data.

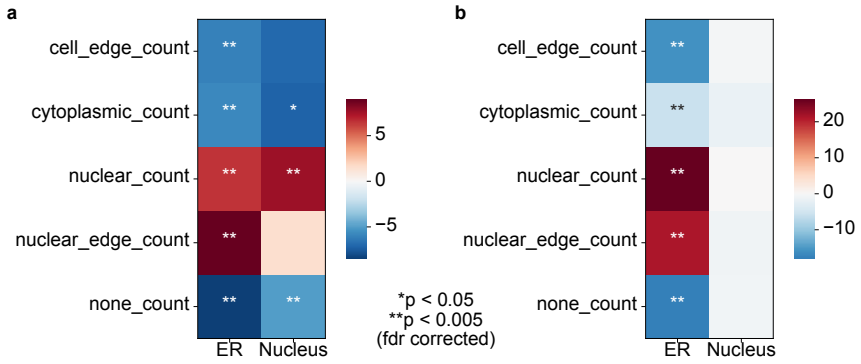

Figure A.2. Enrichment of compartment-specific expression for RNAforest gene pattern frequencies. Compartment-specific enrichment of endoplasmic reticulum (ER) and nucleus gene expression – from Xia et al 2019[61] – relative to RNAforest gene pattern frequencies in the A. MERFISH dataset and B. seqFISH+ dataset.

Appendix B

Supplemental Material for Chapter 2

B.1 Supplementary Figures

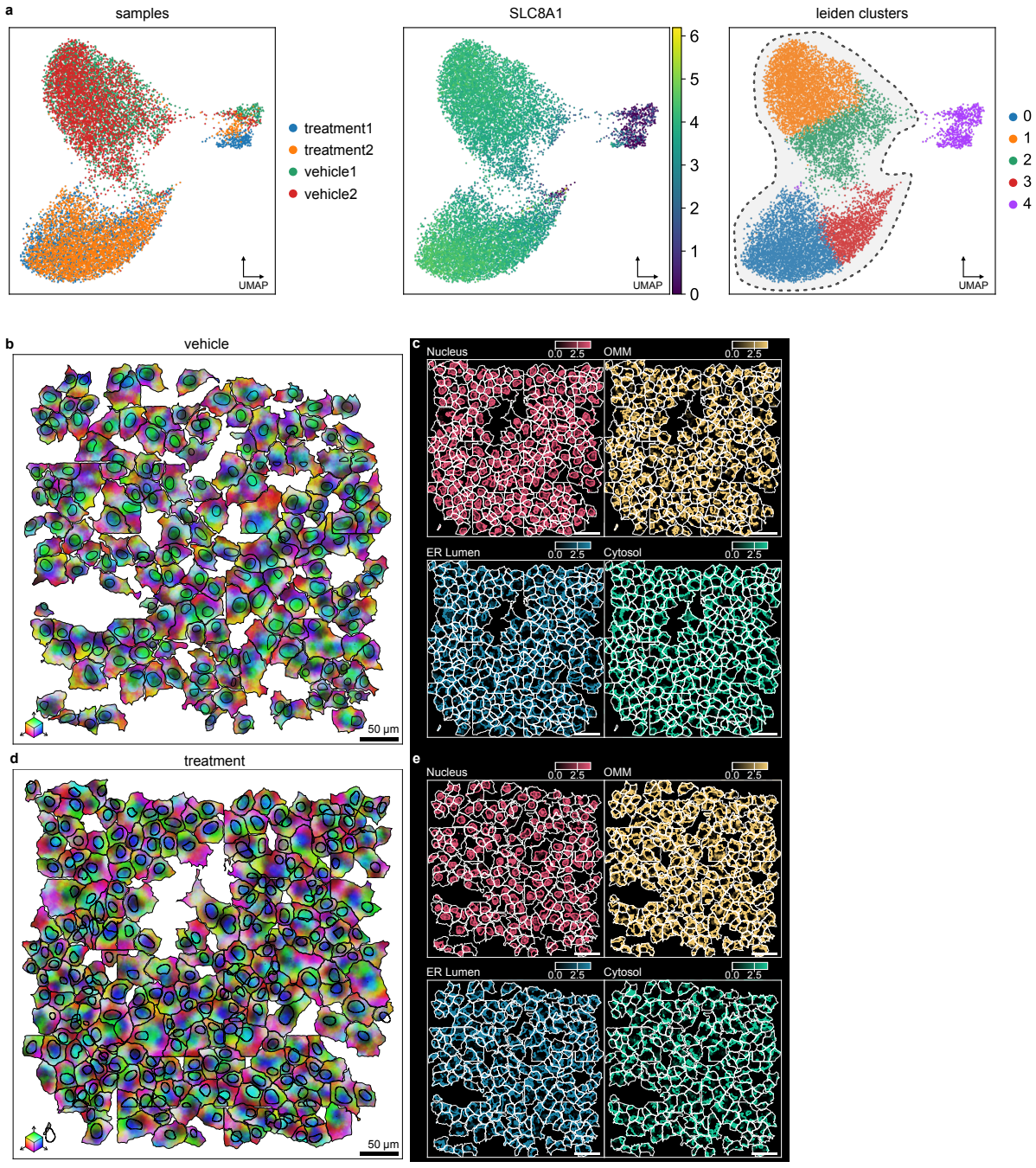

Figure B.1. Filtering and RNAflux analysis of DOX treated cardiomyocytes. A. Left: UMAP of all 4 cardiomyocyte samples, colors denote different samples. Center: Cells are colored by log-scaled SLC8A1 RNA expression. Right: Leiden clustering identifies 5 clusters, separating low expression SLC8A1 into cluster 4. Representative crop of B. vehicle and D. treatment samples, colored by the first 3 principal components of its RNAflux embedding. Relative enrichment of transcripts enriched for location-specific expression in C. vehicle and E. treatment samples. Red, yellow, blue and green enrichment correspond to nuclear, OMM, ER lumen, and cytosol genesets respectively.

Bibliography

[1] Adina R Buxbaum, Gal Haimovich, and Robert H Singer. In the right place at the right time: Visualizing and understanding mRNA localization. *Nature Reviews Molecular Cell Biology*, 16(2):95–109, February 2015. ISSN 1471-0072, 1471-0080. doi: 10.1038/nrm3918.

[2] Michael S. Fernandopulle, Jennifer Lippincott-Schwartz, and Michael E. Ward. RNA transport and local translation in neurodevelopmental and neurodegenerative disease. *Nature Neuroscience*, 24(5):622–632, May 2021. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-020-00785-2.

[3] Arjun Raj, Patrick van den Bogaard, Scott A. Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, October 2008. ISSN 1548-7105. doi: 10.1038/nmeth.1253.

[4] Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233): aaa6090, April 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa6090.

[5] Vivien Marx. Method of the Year: Spatially resolved transcriptomics. *Nature Methods*, 18 (1):9–14, January 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01033-y.

[6] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18.

[7] Yuhan Wang, Mark Eddison, Greg Fleishman, Martin Weigert, Shengjin Xu, Tim Wang, Konrad Rokicki, Cristian Goina, Fredrick E. Henry, Andrew L. Lemire, Uwe Schmidt, Hui Yang, Karel Svoboda, Eugene W. Myers, Stephan Saalfeld, Wyatt Korff, Scott M.

Sternson, and Paul W. Tillberg. EASI-FISH for thick tissue defines lateral hypothalamus spatio-molecular organization. *Cell*, 184(26):6361–6377.e24, December 2021. ISSN 00928674. doi: 10.1016/j.cell.2021.11.024.

[8] Denis Schapiro, Artem Sokolov, Clarence Yapp, Yu-An Chen, Jeremy L. Muhlich, Joshua Hess, Allison L. Creason, Ajit J. Nirmal, Gregory J. Baker, Maulik K. Nariya, Jia-Ren Lin, Zoltan Maliga, Connor A. Jacobson, Matthew W. Hodgman, Juha Ruokonen, Samouil L. Farhi, Domenic Abbondanza, Eliot T. McKinley, Daniel Persson, Courtney Betts, Shamilene Sivagnanam, Aviv Regev, Jeremy Goecks, Robert J. Coffey, Lisa M. Coussens, Sandro Santagata, and Peter K. Sorger. MCMICRO: A scalable, modular image-processing pipeline for multiplexed tissue imaging. *Nature Methods*, 19(3):311–315, March 2022. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-021-01308-y.

[9] ZhuangLab/MERlin: MERlin v0.1.6. doi: 10.5281/zenodo.3758540.

[10] Others. Starfish: Open Source Image Based Transcriptomics and Proteomics Tools.

[11] Cecilia Cisar, Nicholas Keener, Mathew Ruffalo, and Benedict Paten. A unified pipeline for FISH spatial transcriptomics. *Cell Genomics*, 3(9):100384, September 2023. ISSN 2666-979X. doi: 10.1016/j.xgen.2023.100384.

[12] Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. Anndata: Annotated data. Preprint, Bioinformatics, December 2021.

[13] Luca Marconato, Giovanni Palla, Kevin A. Yamauchi, Isaac Virshup, Elyas Heidari, Tim Treis, Marcella Toth, Rahul B. Shrestha, Harald Vöhringer, Wolfgang Huber, Moritz Gerstung, Josh Moore, Fabian J. Theis, and Oliver Stegle. SpatialData: An open and universal data framework for spatial omics, May 2023.

[14] Dario Righelli, Lukas M Weber, Helena L Crowell, Brenda Pardo, Leonardo Collado-Torres, Shila Ghazanfar, Aaron T L Lun, Stephanie C Hicks, and Davide Risso. SpatialExperiment: Infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. *Bioinformatics*, 38(11):3128–3131, May 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac299.

[15] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, Rani E. George, Nico Pierson, Long Cai, and Guo-Cheng Yuan. Giotto: A toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22(1):78, December 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02286-2.

[16] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. Squidpy: A scalable framework for spatial single cell analysis. Preprint, Bioinformatics, February 2021.

[17] STOmics/Stereopy. STOmics, October 2023.

[18] Duy Pham, Xiao Tan, Jun Xu, Laura F. Grice, Pui Yeng Lam, Arti Raghubar, Jana Vukovic, Marc J. Ruitenberg, and Quan Nguyen. stLearn: Integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. Preprint, Bioinformatics, May 2020.

[19] Lambda Moses, Pétur Helgi Einarsson, Kayla Jackson, Laura Luebbert, A. Sina Booeshaghi, Sindri Antonsson, Nicolas Bray, Páll Melsted, and Lior Pachter. Voyager: Exploratory single-cell genomics data analysis with geospatial statistics, August 2023.

[20] Anurendra Kumar, Alex Schrader, Ali Boroojeny, Marisa Asadian, Juyeon Lee, You Song, Sihai Zhao, Hee-Sun Han, and Saurabh Sinha. Intracellular Spatial Transcriptomic Analysis Toolkit (InSTAnT). Preprint, In Review, January 2023.

[21] Zhou Fang, Adam J Ford, Thomas Hu, Nicholas Zhang, Athanasios Mantalaris, and Ahmet F Coskun. Subcellular spatially resolved gene neighborhood networks in single cells.

[22] Florin Walter, Oliver Stegle, and Britta Velten. FISHFactor: A Probabilistic Factor Model for Spatial Transcriptomics Data with Subcellular Resolution.

[23] Arthur Imbert, Wei Ouyang, Adham Safieddine, Emeline Coleno, Christophe Zimmer, Edouard Bertrand, Thomas Walter, and Florian Mueller. FISH-quant v2: A scalable and modular tool for smFISH image analysis. *RNA*, 28(6):786–795, June 2022. ISSN 1355-8382, 1469-9001. doi: 10.1261/rna.079073.121.

[24] Clarence K. Mah, Noorsher Ahmed, Nicole Lopez, Dylan Lam, Alexander Monell, Colin Kern, Yuanyuan Han, Gino Prasad, Anthony J. Cesnik, Emma Lundberg, Quan Zhu, Hannah Carter, and Gene W. Yeo. Bento: A toolkit for subcellular analysis of spatial transcriptomics data. Preprint, Bioinformatics, June 2022.

[25] Peter J Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M Breckels, Anna Bäckström, Frida Danielsson, Linn Fagerberg, Jenny Fall, Laurent Gatto, Christian Gnann, Sophia Hober, Martin Hjelmare, Fredric Johansson, Sunjae Lee, Cecilia Lindskog, Jan Mulder, Claire M Mulvey, Peter Nilsson, Per Oksvold, Johan Rockberg, Rutger Schutten, Jochen M Schwenk, Åsa Sivertsson, Evelina Sjöstedt, Marie Skogs, Charlotte Stadler, Devin P Sullivan, Hanna Tegel, Casper Winsnes, Cheng Zhang, Martin Zwahlen, Adil Mardinoglu, Fredrik Pontén, Kalle von Feilitzen, Kathryn S Lilley, Mathias Uhlén, and Emma Lundberg. A subcellular map of the human proteome. *Science*, 356(6340):eaal3321, May 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal3321.

[26] Kirsti Laurila and Mauno Vihinen. Prediction of disease-related mutations affecting protein localization. *BMC Genomics*, 10:122, March 2009.

[27] Solip Park, Jae-Seong Yang, Young-Eun Shin, Juyong Park, Sung Key Jang, and Sanguk Kim. Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Mol. Syst. Biol.*, 7:494, May 2011.

[28] Eric Lécuyer, Hideki Yoshida, Neela Parthasarathy, Christina Alm, Tomas Babak, Tanja Cerovina, Timothy R Hughes, Pavel Tomancak, and Henry M Krause. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 131(1):174–187, October 2007.

[29] Samantha Bovaird, Dhara Patel, Juan-Carlos Alberto Padilla, and Eric Lécuyer. Biological functions, regulatory mechanisms, and disease relevance of RNA localization pathways. *FEBS Letters*, 592(17):2948–2972, September 2018. ISSN 0014-5793, 1873-3468. doi: 10.1002/1873-3468.13228.

[30] Sulagna Das, Robert H Singer, and Young J Yoon. The travels of mRNAs in neurons: Do they know where they are going? *Curr. Opin. Neurobiol.*, 57:110–116, August 2019.

[31] Pabitra K Sahoo, Deanna S Smith, Nora Perrone-Bizzozero, and Jeffery L Twiss. Axonal mRNA transport and translation at a glance. *Journal of Cell Science*, 131(8):jcs196808, April 2018. ISSN 1477-9137, 0021-9533. doi: 10.1242/jcs.196808.

[32] Nicolai Kügelgen and Marina Chekulaeva. Conservation of a core neurite transcriptome across neuronal types and species. *WIREs RNA*, 11(4), July 2020. ISSN 1757-7004, 1757-7012. doi: 10.1002/wrna.1590.

[33] Brady P Culver, Josh DeClercq, Igor Dolgalev, Man Shan Yu, Bin Ma, Adriana Heguy, and Naoko Tanese. Huntington's Disease Protein Huntingtin Associates with its own mRNA. *J Huntingtons Dis*, 5(1):39–51, 2016.

[34] Lindsay Romo, Emily S Mohn, and Neil Aronin. A Fresh Look at Huntingtin mRNA Processing in Huntington's Disease. *J Huntingtons Dis*, 7(2):101–108, 2018.

[35] Joseph A White, 2nd, Eric Anderson, Katherine Zimmerman, Kan Hong Zheng, Roza Rouhani, and Shermali Gunawardena. Huntingtin differentially regulates the axonal transport of a sub-set of Rab-containing vesicles in vivo. *Hum. Mol. Genet.*, 24(25): 7182–7195, December 2015.

[36] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, April 2015.

[37] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, and Long Cai. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*, 568(7751): 235–239, April 2019.

[38] Daniel Gyllborg, Christoffer Mattsson Langseth, Xiaoyan Qian, Eunkyoung Choi, Sergio Marco Salas, Markus M Hilscher, Ed S Lein, and Mats Nilsson. Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. *Nucleic Acids Research*, 48(19):e112, November 2020. ISSN 0305-1048. doi:

10.1093/nar/gkaa792.

[39] Shahar Alon, Daniel R Goodwin, Anubhav Sinha, Fei Chen, Evan R Daugharthy, Yosuke Bando, and Atsushi Kajita. Expansion Sequencing: Spatially Precise In Situ Transcriptomics in Intact Biological Systems.

[40] Andrew Butler and Rahul Satija. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *bioRxiv*, pages 164889–164889, 2017. ISSN 0780351355. doi: 10.1101/164889.

[41] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15–15, February 2018. doi: 10.1186/s13059-017-1382-0.

[42] Yichun He, Xin Tang, Jiahao Huang, Jingyi Ren, Haowen Zhou, Kevin Chen, Albert Liu, Hailing Shi, Zuwan Lin, Qiang Li, Abhishek Aditham, Johain Ounadjela, Emanuelle I. Grody, Jian Shu, Jia Liu, and Xiao Wang. ClusterMap for multi-scale clustering analysis of spatial gene expression. *Nature Communications*, 12(1):5909, October 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26044-x.

[43] Viktor Petukhov, Ruslan A. Soldatov, Konstantin Khodosevich, and Peter V. Kharchenko. Bayesian segmentation of spatially resolved transcriptomics data. Preprint, Bioinformatics, October 2020.

[44] Hannah Spitzer, Scott Berry, Mark Donoghoe, Lucas Pelkmans, and Fabian J Theis. Learning consistent subcellular landmarks to quantify changes in multiplexed protein maps.

[45] Candace C. Liu, Noah F. Greenwald, Alex Kong, Erin F. McCaffrey, Ke Xuan Leow, Dunja Mrdjen, Bryan J. Cannon, Josef Lorenz Rumberger, Sricharan Reddy Varra, and Michael Angelo. Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering. *Nature Communications*, 14(1):4618, August 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-40068-5.

[46] Geopandas/geopandas: V0.9.0. doi: 10.5281/zenodo.4569086.

[47] S Gillies, B Ward, and A S Petersen. Rasterio: Geospatial raster I/O for Python programmers. *URL https://github. com/mapbox/rasterio*.

[48] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. Van Der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul Van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony

Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius De Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0686-2.

[49] Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. TensorLy: Tensor Learning in Python. *J. Mach. Learn. Res.*, 20(26):1–6, 2019.

[50] Shanshan He, Ruchir Bhatt, Carl Brown, Emily A. Brown, Derek L. Buhr, Kan Chantranuvatana, Patrick Danaher, Dwayne Dunaway, Ryan G. Garrison, Gary Geiss, Mark T. Gregory, Margaret L. Hoang, Rustem Khafizov, Emily E. Killingbeck, Dae Kim, Tae Kyung Kim, Youngmi Kim, Andrew Klock, Mithra Korukonda, Alecksandr Kutchma, Zachary R. Lewis, Yan Liang, Jeffrey S. Nelson, Giang T. Ong, Evan P. Perillo, Joseph C. Phan, Tien Phan-Everson, Erin Piazza, Tushar Rane, Zachary Reitz, Michael Rhodes, Alyssa Rosenbloom, David Ross, Hiromi Sato, Aster W. Wardhani, Corey A. Williams-Wietzikoski, Lidan Wu, and Joseph M. Beechem. High-plex Multiomic Analysis in FFPE at Subcellular Level by Spatial Molecular Imaging. Preprint, Genomics, November 2021.

[51] Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Morgane Rouault, Ghezal Beliakoff, Michelli Faria De Oliveira, Andrew Kohlway, Jawad Abousoud, Carolyn A. Morrison, Tingsheng Yu Drennon, Seayar H. Mohabbat, Stephen R. Williams, 10x Development Teams, and Sarah E.B. Taylor. High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. Preprint, Cancer Biology, October 2022.

[52] Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M Church. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.*, 10(3):442–458, March 2015.

[53] Shikai Hu, Silvia Liu, Yu Bian, Minakshi Poddar, Sucha Singh, Jackson McGaughey, Aaron Bell, Levi L Blazer, Jarret J Adams, Sachdev S Sidhu, Stephane Angers, and Satdarshan P Monga. Dynamic control of metabolic zonation and liver repair by endothelial cell Wnt2

and Wnt9b revealed by single cell spatial transcriptomics using Molecular Cartography.

[54] Arthur Imbert, Wei Ouyang, Adham Safieddine, Edouard Bertrand, Thomas Walter, and Florian Mueller. FISH-quant v2: A scalable and modular analysis tool for smFISH image analysis.

[55] P Viola and M Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, December 2001.

[56] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature Methods*, 10(11): 1127–1133, November 2013. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2657.

[57] Thomas Stoeger, Nico Battich, Markus D Herrmann, Yauhen Yakimovich, and Lucas Pelkmans. Computer vision for image-based transcriptomics. *Methods*, 85:44–53, September 2015.

[58] Aubin Samacoits, Racha Chouaib, Adham Safieddine, Abdel-Meneem Traboulsi, Wei Ouyang, Christophe Zimmer, Marion Peter, Edouard Bertrand, Thomas Walter, and Florian Mueller. A computational framework to study sub-cellular RNA localization. *Nat. Commun.*, 9(1):4584, November 2018.

[59] Racha Chouaib, Adham Safieddine, Xavier Pichon, Arthur Imbert, Oh Sung Kwon, Aubin Samacoits, Abdel-Meneem Traboulsi, Marie-Cécile Robert, Nikolay Tsanov, Emeline Coleno, Ina Poser, Christophe Zimmer, Anthony Hyman, Hervé Le Hir, Kazem Zibara, Marion Peter, Florian Mueller, Thomas Walter, and Edouard Bertrand. A Dual Protein-mRNA Localization Screen Reveals Compartmentalized Translation and Widespread Co-translational RNA Targeting. *Developmental Cell*, 54(6):773–791.e5, September 2020. ISSN 15345807. doi: 10.1016/j.devcel.2020.07.010.

[60] Jeffrey R Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P Babcock, and Xiaowei Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U. S. A.*, 113(39): 11046–11051, September 2016.

[61] Chenglong Xia, Jean Fan, George Emanuel, Junjie Hao, and Xiaowei Zhuang. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, 116(39):19490–19499, September 2019.

[62] The Gene Ontology Consortium, Seth Carbon, Eric Douglass, Benjamin M Good, Deepak R Unni, Nomi L Harris, Christopher J Mungall, Siddartha Basu, Rex L Chisholm, Robert J Dodson, Eric Hartline, Petra Fey, Paul D Thomas, Laurent-Philippe Albou, Dustin Ebert, Michael J Kesling, Huaiyu Mi, Anushya Muruganujan, Xiaosong Huang, Tremayne Mushayahama, Sandra A LaBonte, Deborah A Siegele, Giulia Antonazzo, Helen Attrill, Nick H

Brown, Phani Garapati, Steven J Marygold, Vitor Trovisco, Gil Dos Santos, Kathleen Falls, Christopher Tabone, Pinglei Zhou, Joshua L Goodman, Victor B Strelets, Jim Thurmond, Penelope Garmiri, Rizwan Ishtiaq, Milagros Rodríguez-López, Marcio L Acencio, Martin Kuiper, Astrid Lægreid, Colin Logie, Ruth C Lovering, Barbara Kramarz, Shirin C C Saverimuttu, Sandra M Pinheiro, Heather Gunn, Renzhi Su, Katherine E Thurlow, Marcus Chibucos, Michelle Giglio, Suvarna Nadendla, James Munro, Rebecca Jackson, Margaret J Duesbury, Noemi Del-Toro, Birgit H M Meldal, Kalpana Paneerselvam, Livia Perfetto, Pablo Porras, Sandra Orchard, Anjali Shrivastava, Hsin-Yu Chang, Robert Daniel Finn, Alexander Lawson Mitchell, Neil David Rawlings, Lorna Richardson, Amaia Sangrador-Vegas, Judith A Blake, Karen R Christie, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry M Sitnikov, Midori A Harris, Stephen G Oliver, Kim Rutherford, Valerie Wood, Jaqueline Hayles, Jürg Bähler, Elizabeth R Bolton, Jeffery L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Cody Plasterer, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lisa Matthews, James P Balhoff, Suzi A Aleksander, Michael J Alexander, J Michael Cherry, Stacia R Engel, Felix Gondwe, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Matt Simison, Marek S Skrzypek, Shuai Weng, Edith D Wong, Marc Feuermann, Pascale Gaudet, Anne Morgat, Erica Bakker, Tanya Z Berardini, Leonore Reiser, Shabari Subramaniam, Eva Huala, Cecilia N Arighi, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Alex Bateman, Marie-Claude Blatter, Emmanuel Boutet, Emily Bowler, Lionel Breuza, Alan Bridge, Ramona Britto, Hema Bye-A-Jee, Cristina Casals Casas, Elisabeth Coudert, Paul Denny, Anne Estreicher, Maria Livia Famiglietti, George Georghiou, Arnaud Gos, Nadine Gruaz-Gumowski, Emma Hatton-Ellis, Chantal Hulo, Alexandr Ignatchenko, Florence Jungo, Kati Laiho, Philippe Le Mercier, Damien Lieberherr, Antonia Lock, Yvonne Lussi, Alistair MacDougall, Michele Magrane, Maria J Martin, Patrick Masson, Darren A Natale, Nevila Hyka-Nouspikel, Sandra Orchard, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Sangya Pundir, Catherine Rivoire, Elena Speretta, Shyamala Sundaram, Nidhi Tyagi, Kate Warner, Rossana Zaru, Cathy H Wu, Alexander D Diehl, Juancarlos N Chan, Christian Grove, Raymond Y N Lee, Hans-Michael Muller, Daniela Raciti, Kimberly Van Auken, Paul W Sternberg, Matthew Berriman, Michael Paulini, Kevin Howe, Sibyl Gao, Adam Wright, Lincoln Stein, Douglas G Howe, Sabrina Toro, Monte Westerfield, Pankaj Jaiswal, Laurel Cooper, and Justin Elser. The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, January 2021. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkaa1113.

[63] Yang Xu, Alexander Belyi, Iva Bojic, and Carlo Ratti. How friends share urban space: An exploratory spatiotemporal analysis using mobile phone data. *Trans. GIS*, 21(3):468–487, June 2017.

[64] Hu Zeng, Jiahao Huang, Jingyi Ren, Connie Kangni Wang, Zefang Tang, Haowen Zhou, Yiming Zhou, Hailing Shi, Abhishek Aditham, Xin Sui, Hongyu Chen, Jennifer A. Lo, and Xiao Wang. Spatially resolved single-cell translatomics at molecular resolution. *Science*, 380(6652):eadd3067, June 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.add3067.

[65] B D Ripley. The second-order analysis of stationary point processes. *J. Appl. Probab.*, 13 (2):255–266, June 1976.

[66] Michael Tiefelsdorf. *Modelling Spatial Processes: The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I*. Springer, April 2006.

[67] Andrew David Cliff and J K Ord. *Spatial Processes: Models & Applications*. Pion, 1981.

[68] Timothy F Leslie and Barry J Kronenfeld. The colocation quotient: A new measure of spatial association between categorical subsets of points. *Geogr. Anal.*, 43(3):306–326, July 2011.

[69] Amnon Shashua and Tamir Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, ICML '05, pages 792–799, New York, NY, USA, August 2005. Association for Computing Machinery. ISBN 978-1-59593-180-1. doi: 10.1145/1102351.1102451.

[70] Reto Gassmann. Dynein at the kinetochore. *Journal of Cell Science*, 136(5):jcs220269, March 2023. ISSN 0021-9533, 1477-9137. doi: 10.1242/jcs.220269.

[71] Munishwar Nath Gupta and Vladimir N. Uversky. Moonlighting enzymes: When cellular context defines specificity. *Cellular and Molecular Life Sciences*, 80(5):130, May 2023. ISSN 1420-682X, 1420-9071. doi: 10.1007/s00018-023-04781-0.

[72] Christian Gnann, Anthony J. Cesnik, and Emma Lundberg. Illuminating Non-genetic Cellular Heterogeneity with Imaging-Based Spatial Proteomics. *Trends in Cancer*, 7(4):278–282, April 2021. ISSN 24058033. doi: 10.1016/j.trecan.2020.12.006.

[73] Furqan M Fazal, Shuo Han, Kevin R Parker, Pornchai Kaewsapsak, Jin Xu, Alistair N Boettiger, Howard Y Chang, and Alice Y Ting. Atlas of subcellular RNA localization revealed by APEX-seq. *Cell*, 178(2):473–490.e26, July 2019.

[74] Adrien Hallou, Ruiyang He, Benjamin D. Simons, and Bianca Dumitrascu. A computational pipeline for spatial mechano-transcriptomics, August 2023.

[75] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20(1):37–46, April 1960.

[76] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, October 2005.

[77] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Stefan Fröhling, Edmond M Chan, Martin L Sos, Kathrin Michel, Craig Mermel, Serena J Silver, Barbara A Weir, Jan H Reiling, Qing Sheng, Piyush B Gupta, Raymond C Wadlow, Hanh Le, Sebastian

Hoersch, Ben S Wittner, Sridhar Ramaswamy, David M Livingston, David M Sabatini, Matthew Meyerson, Roman K Thomas, Eric S Lander, Jill P Mesirov, David E Root, D Gary Gilliland, Tyler Jacks, and William C Hahn. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–112, November 2009. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature08460.

[78] Zhuorui Xie, Allison Bailey, Maxim V. Kuleshov, Daniel J. B. Clarke, John E. Evangelista, Sherry L. Jenkins, Alexander Lachmann, Megan L. Wojciechowicz, Eryk Kropiwnicki, Kathleen M. Jagodnik, Minji Jeon, and Avi Ma'ayan. Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, 1(3):e90, March 2021. ISSN 2691-1299, 2691-1299. doi: 10.1002/cpz1.90.

[79] Pau Badia-i-Mompel, Jesús Vélez Santiago, Jana Braunger, Celina Geiss, Daniel Dimitrov, Sophia Müller-Dott, Petr Taus, Aurelien Dugourd, Christian H Holland, Ricardo O Ramirez Flores, and Julio Saez-Rodriguez. decoupleR: Ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, 2(1):vbac016, January 2022. ISSN 2635-0041. doi: 10.1093/bioadv/vbac016.

[80] Hui Huang, Quan Zhu, Adam Jussila, Yuanyuan Han, Bogdan Bintu, Colin Kern, Mattia Conte, Yanxiao Zhang, Simona Bianco, Andrea M. Chiariello, Miao Yu, Rong Hu, Melodi Tastemel, Ivan Juric, Ming Hu, Mario Nicodemi, Xiaowei Zhuang, and Bing Ren. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nature Genetics*, 53(7):1064–1074, July 2021. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-021-00863-6.

[81] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: A generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, January 2021. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-020-01018-x.

[82] Balaraman Kalyanaraman. Teaching the basics of the mechanism of doxorubicin-induced cardiotoxicity: Have we been barking up the wrong tree? *Redox Biology*, 29:101394, January 2020. ISSN 22132317. doi: 10.1016/j.redox.2019.101394.

[83] R C Young, R F Ozols, and C E Myers. The anthracycline antineoplastic drugs. *N. Engl. J. Med.*, 305(3):139–153, July 1981.

[84] Mohammad Sheibani, Yaser Azizi, Maryam Shayan, Sadaf Nezamoleslami, Faezeh Eslami, Mohammad Hadi Farjoo, and Ahmad Reza Dehpour. Doxorubicin-Induced Cardiotoxicity: An Overview on Pre-clinical Therapeutic Approaches. *Cardiovascular Toxicology*, 22(4): 292–310, April 2022. ISSN 1559-0259. doi: 10.1007/s12012-022-09721-1.

[85] Jie Yu, Changxi Wang, Qi Kong, Xiaxia Wu, Jin-Jian Lu, and Xiuping Chen. Recent progress in doxorubicin-induced cardiotoxicity and protective potential of natural products. *Phytomedicine*, 40:125–139, February 2018.

[86] Atiar M Rahman, Syed Wamique Yusuf, and Michael S Ewer. Anthracycline-induced

cardiotoxicity and the cardiac-sparing effect of liposomal formulation. *International Journal of Nanomedicine*, 2(4):567–583, 2007.

[87] K. M. Tewey, T. C. Rowe, L. Yang, B. D. Halligan, and L. F. Liu. Adriamycin-induced DNA damage mediated by mammalian DNA topoisomerase II. *Science (New York, N.Y.)*, 226(4673):466–468, October 1984. ISSN 0036-8075. doi: 10.1126/science.6093249.

[88] Mari C Asensio-López, Fernando Soler, Domingo Pascual-Figal, Francisco Fernández-Belda, and Antonio Lax. Doxorubicin-induced oxidative stress: The protective effect of nicorandil on HL-1 cardiomyocytes. *PLoS One*, 12(2):e0172803, February 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0172803.

[89] Tomáš Šimůnek, Martin Štěrba, Olga Popelová, Michaela Adamcová, Radomír Hrdina, and Vladimír Geršl. Anthracycline-induced cardiotoxicity: Overview of studies examining the roles of oxidative stress and free cellular iron. *Pharmacol. Rep.*, 61(1):154–171, January 2009.

[90] Chen Xiong, Yan-Zhao Wu, Yu Zhang, Zi-Xiao Wu, Xue-Yan Chen, Ping Jiang, Hui-Cai Guo, Ke-Rang Xie, Ke-Xin Wang, and Su-Wen Su. Protective effect of berberine on acute cardiomyopathy associated with doxorubicin treatment. *Oncol. Lett.*, 15(4):5721–5729, April 2018.

[91] M F Xu, P L Tang, Z M Qian, and M Ashraf. Effects by doxorubicin on the myocardium are mediated by oxygen free radicals. *Life Sci.*, 68(8):889–901, January 2001.

[92] Pushkar Singh Rawat, Aiswarya Jaiswal, Amit Khurana, Jasvinder Singh Bhatti, and Umashanker Navik. Doxorubicin-induced cardiotoxicity: An update on the molecular mechanism and novel therapeutic strategies for effective management. *Biomedicine & Pharmacotherapy*, 139:111708, July 2021. ISSN 07533322. doi: 10.1016/j.biopha.2021.111708.

[93] Ana R. Rubio, Natalia Busto, José M. Leal, and Begoña García. Doxorubicin binds to duplex RNA with higher affinity than ctDNA and favours the isothermal denaturation of triplex RNA. *RSC Advances*, 6(103):101142–101152, October 2016. ISSN 2046-2069. doi: 10.1039/C6RA21387A.

[94] Ryan J. Marcheschi, Kathryn D. Mouzakis, and Samuel E. Butcher. Selection and characterization of small molecules that bind the HIV-1 frameshift site RNA. *ACS chemical biology*, 4(10):844–854, October 2009. ISSN 1554-8937. doi: 10.1021/cb900167m.

[95] Vaishali Bagalkot, Omid C. Farokhzad, Robert Langer, and Sangyong Jon. An Aptamer–Doxorubicin Physical Conjugate as a Novel Targeted Drug-Delivery Platform. *Angewandte Chemie International Edition*, 45(48):8149–8152, 2006. ISSN 1521-3773. doi: 10.1002/anie.200602251.

[96] Joyce Man, Phil Barnett, and Vincent M Christoffels. Structure and function of the

Nppa–Nppb cluster locus during heart development and disease. *Cell. Mol. Life Sci.*, 75 (8):1435–1444, April 2018.

[97] Wei Song, Hao Wang, and Qingyu Wu. Atrial natriuretic peptide in cardiovascular biology and disease (NPPA). *Gene*, 569(1):1–6, September 2015. ISSN 03781119. doi: 10.1016/j.gene.2015.06.029.

[98] Yair E Lewis, Anner Moskovitz, Michael Mutlak, Joerg Heineke, Lilac H Caspi, and Izhak Kehat. Localization of transcripts, translation, and degradation for spatiotemporal sarcomere maintenance. *J. Mol. Cell. Cardiol.*, 116:16–28, March 2018.

[99] Brendan M. Smalec, Robert Ietswaart, Karine Choquet, Erik McShane, Emma R. West, and L. Stirling Churchman. Genome-wide quantification of RNA flow across subcellular compartments reveals determinants of the mammalian transcript life cycle. Preprint, Genomics, August 2022.

[100] Vilma A Sardão, Paulo J Oliveira, Jon Holy, Catarina R Oliveira, and Kendall B Wallace. Morphological alterations induced by doxorubicin on H9c2 myoblasts: Nuclear, mitochondrial, and cytoskeletal targets. *Cell Biol. Toxicol.*, 25(3):227–243, June 2009.

[101] Marcel Meissner, Petra Weissgerber, Juan E Camacho Londoño, Jean Prenen, Sabine Link, Sandra Ruppenthal, Jeffery D Molkentin, Peter Lipp, Bernd Nilius, Marc Freichel, and Veit Flockerzi. Moderate calcium channel dysfunction in adult mice with inducible cardiomyocyte-specific excision of the cacnb2 gene. *J. Biol. Chem.*, 286(18):15875–15882, May 2011.

[102] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, 47(D1):D766–D773, January 2019.

[103] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, José Carlos Marugán, Carla Cummins, Claire Davidson, Kamalkumar Dodiya, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, Thomas Maurel, Mark McDowall, Aoife McMahon,

Shamika Mohanan, Benjamin Moore, Michael Nuhn, Denye N Oheh, Anne Parker, Andrew Parton, Mateus Patricio, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Mira Sycheva, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth IIsley, Myrto Kostadima, Nick Langridge, Jane E Loveland, Fergal J Martin, Joannella Morales, Jonathan M Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, Stephen J Trevanion, Fiona Cunningham, Kevin L Howe, Daniel R Zerbino, and Paul Flicek. Ensembl 2020. *Nucleic Acids Research*, page gkz966, November 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkz966.

[104] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, March 2011.

[105] Jason D Gans and Murray Wolinsky. Improved assay-dependent searching of nucleic acid sequence databases. *Nucleic Acids Res.*, 36(12):e74, July 2008.

[106] Jose Manuel Rodriguez, Juan Rodriguez-Rivas, Tomás Di Domenico, Jesús Vázquez, Alfonso Valencia, and Michael L Tress. APPRIS 2017: Principal isoforms for multiple gene sets. *Nucleic Acids Research*, 46(D1):D213–D217, January 2018. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkx997.

[107] Xiaojun Lian, Cheston Hsiao, Gisela Wilson, Kexian Zhu, Laurie B Hazeltine, Samira M Azarin, Kunil K Raval, Jianhua Zhang, Timothy J Kamp, and Sean P Palecek. Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling. *Proc. Natl. Acad. Sci. U. S. A.*, 109(27):E1848–57, July 2012.

[108] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. SeqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron*, 94(4): 752–758.e1, May 2017.

[109] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science (New York, N.Y.)*, 361(6400):eaat5691–eaat5691, 2018. doi: 10.1126/science.aat5691.

[110] Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, 10(9):857–860, September 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2563.

[111] Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, March 2019. doi: 10.1126/science.

aaw1219.

[112] Paolo Di Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, April 2017. ISSN 1546-1696. doi: 10.1038/nbt.3820.

[113] Ilya G. Goldberg, Chris Allan, Jean-Marie Burel, Doug Creager, Andrea Falconi, Harry Hochheiser, Josiah Johnston, Jeff Mellen, Peter K. Sorger, and Jason R. Swedlow. The Open Microscopy Environment (OME) Data Model and XML file: Open tools for informatics and quantitative analysis in biological imaging. *Genome Biology*, 6(5):R47, May 2005. ISSN 1474-760X. doi: 10.1186/gb-2005-6-5-r47.

[114] Jeremy L Muhlich, Yu-An Chen, Clarence Yapp, Douglas Russell, Sandro Santagata, and Peter K Sorger. Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. *Bioinformatics*, 38(19):4613–4621, September 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac544.

[115] Soft-matter/trackpy: V0.6.1. doi: 10.5281/zenodo.7670439.

[116] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(Oct):2825–2830, 2011.

[117] Napari: A multi-dimensional image viewer for Python. doi: 10.5281/zenodo.8115575.

[118] Isaac Virshup, Danila Bredikhin, Lukas Heumos, Giovanni Palla, Gregor Sturm, Adam Gayoso, Ilia Kats, Mikaela Koutrouli, Bonnie Berger, Dana Pe'er, Aviv Regev, Sarah A. Teichmann, Francesca Finotello, F. Alexander Wolf, Nir Yosef, Oliver Stegle, and Fabian J. Theis. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology*, 41(5):604–606, May 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01733-8.