MOLECULAR CARTOGRAPHY UNCOVERS EVOLUTIONARY AND MICROENVIRONMENTAL DYNAMICS IN SPORADIC COLORECTAL TUMORS

By

Cody Nicholas Heiser

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Chemical and Physical Biology

May 12, 2023

Nashville, Tennessee

Approved:

Vito Quaranta, M.D.

Robert J. Coffey, M.D.

Jacob J. Hughey, Ph.D.

Simon Vandekar, Ph.D.

Ken S. Lau, Ph.D.

ACKNOWLEDGMENTS

I would first like to thank the numerous mentors that have shaped my career both before and during my time at Vanderbilt. Rob Taylor, Ted Markulin, Andrew Hadd, and Robyn Cardwell, among others, were pivotal in teaching and supporting me as I amassed the early research and work experience that afforded me the opportunity to pursue my Ph.D.

Since beginning my graduate studies in 2018, the number of mentors surrounding me increased significantly, with countless valuable professional relationships that I've formed within and beyond my thesis committee and collaborators. I'd like to particularly thank John Wikswo for the most thought-provoking courses in my academic career, and Chase Spurlock and the Decode Health team for providing me with an enriching internship experience that allowed me to translate my academic knowledge to real-world machine learning applications in healthcare.

To round out my Vanderbilt mentors, I have been tremendously lucky to have an incredibly supportive and engaged thesis committee, and I need to thank Vito Quaranta, Bob Coffey, Simon Vandekar, and Jake Hughey for providing strong technical expertise, acting as a sounding board for professional development and career advice, and motivating and guiding me during the writing process. Finally, I would like to thank Ken Lau for all the time, money, and effort he invested in my training over the past four years. He truly pushed me to be the best scientist I can be while affording me the independence to pursue the systems biology questions I was drawn to, and has stood firmly behind me and my work every step of the way.

Next, I need to acknowledge the scientists who do the work that generates the rich datasets from which I reap the benefits. It took a village to build the spatial atlas described herein, however, Joey Simmons and Frank Revetta, with enormous support from the core facilities at Vanderbilt University and VUMC, were absolutely instrumental in executing the multimodal experimental design that I envisioned in early 2019. Additionally, I had the opportunity to mentor some incredible undergraduate students who contributed greatly to my published work and will undoubtedly go on to do great things: Victoria Wang, Molly Bingham, Justin Shao, and Janney Wang.

Finally, I'd like to thank my friends and family, old and new, for their undying support. My closest friends who predate my graduate school career, now scattered coast-to-coast, always make me feel like a day hasn't passed since we've seen each other. And the great friends I have made in the past four years have made Nashville home. My wife, Taylor, and her family in Florida and New Jersey have provided defining and lasting memories from my Ph.D. years. Along with my family in Texas who have been encouraging and supportive from day one, Taylor and I are so lucky to have such a strong support system to carry us into the next phase of our lives and beyond.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Background and Motivation

1.1 Big data and systems biology

Big data has become increasingly important in the field of biology, particularly in research and development of precision medicine. Advances in technology have enabled researchers to generate large amounts of data from genomic, transcriptomic, and proteomic profiling of human tissue (Gerdes et al., 2013; Klein et al., 2015; Ståhl et al., 2016; Cao et al., 2018; Wang et al., 2018; Rodriques et al., 2019; Black et al., 2021). With the help of powerful computational tools, these large datasets can be analyzed and mined for insights into disease mechanisms, potential drug targets, and patient outcomes (Suvà and Tirosh, 2019; Yoosuf et al., 2020).

One of the most promising applications of data science methods to the field of medicine is patient stratification and targeted drug discovery, which manifest themselves in precision diagnostics and individualized treatments, respectively. By analyzing large datasets of patient information, including genomic and clinical metadata, researchers and clinicians can identify specific biomarkers and genetic mutations that can be targeted with precision therapies (Mckenna et al., 2018). This approach has the potential to improve patient outcomes and reduce healthcare costs, as treatments can be tailored to the specific needs of each patient. Furthermore, the use of big data in precision medicine could lead to the development of new drugs and therapies that are more effective and have fewer side effects, as researchers gain a deeper understanding of the underlying biology of diseases.

1.1.1 Data quality and reproducibility

In the field of biomedical data science and machine learning, quality control and reproducibility are critical for ensuring accurate and reliable results. In order to draw meaningful conclusions from large datasets, it is essential to ensure that the data is of high quality and that any errors or biases are identified and corrected (Simmons and Lau, 2017, 2022). This requires rigorous quality control measures, including careful data cleaning and normalization, as well as the use of appropriate statistical methods to control for confounding variables.

Reproducibility is also key to the validity of scientific findings in this field. The ability to reproduce results from different datasets and across multiple modalities is essential for confirming the robustness of scientific findings and for building confidence in the accuracy of the results and models for which they serve as inputs (Luecken and Theis, 2019; Harris et al., 2022; Schapiro et al., 2022). In the context of machine

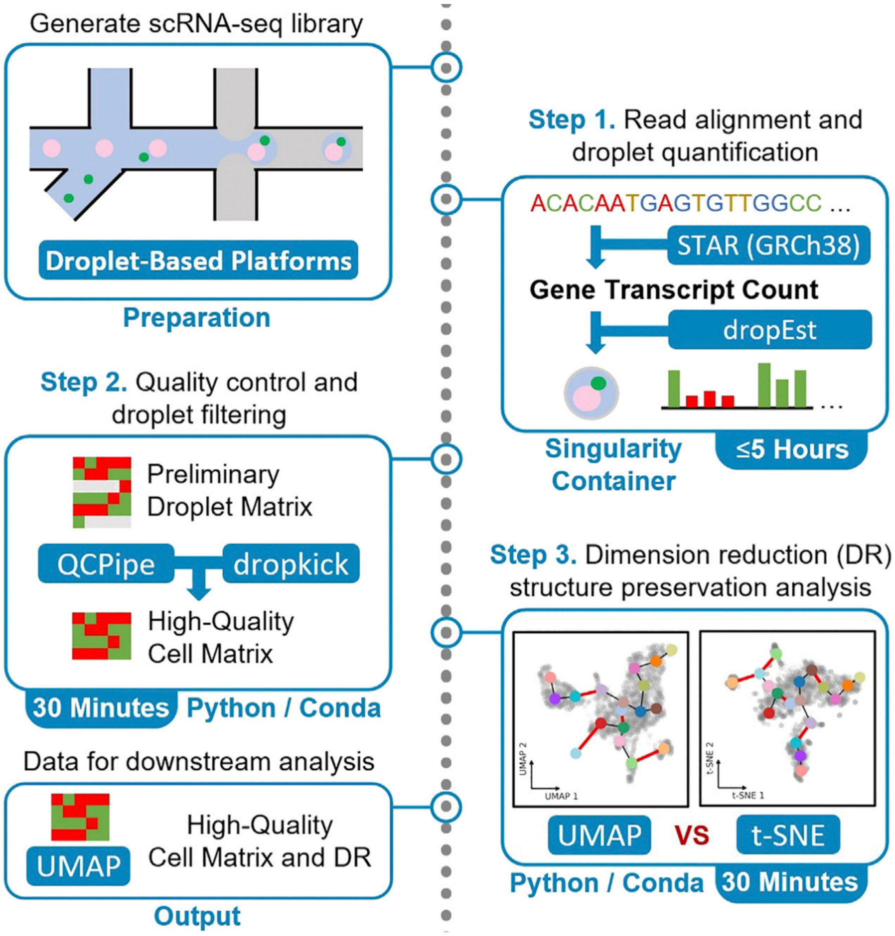

Generate scRNA-seq library
Droplet-Based Platforms
Preparation
Step 1. Read alignment and droplet quantification
ACACAATGAGTGTTGGCC ...
STAR (GRCh38)
Gene Transcript Count
dropEst
Singularity Container
≤5 Hours
Step 2. Quality control and droplet filtering
Preliminary Droplet Matrix
QCPipe
dropkick
High-Quality Cell Matrix
30 Minutes
Python / Conda
Step 3. Dimension reduction (DR) structure preservation analysis
UMAP 2
UMAP 1
t-SNE 2
t-SNE 1
UMAP
VS
t-SNE
Python / Conda
30 Minutes
Data for downstream analysis
UMAP
High-Quality Cell Matrix and DR
Output

Figure 1.1: Processing single-cell RNA-seq data for dimension reduction-based analyses using open-source tools. Adapted from Chen et al., 2021a.

learning, reproducibility is critical for validating different algorithms and models, and for comparing their performance against one another. By prioritizing quality control and reproducibility, researchers and clinicians can ensure that their findings are accurate, reliable, and meaningful, which is essential for driving progress and innovation in this rapidly evolving field (Gutierrez et al., 2018; Chen et al., 2021a; Bao et al., 2021; Figure 1.1).

1.2 Systems-level analysis uncovers key patterns underlying biological complexity

Systems biology is an interdisciplinary field that seeks to understand complex biological systems by studying their interactions and dynamics rather than solely focusing on individual components. Unlike traditional reductionist approaches, which aim to break down biological systems into smaller parts for analysis, systems biology considers the system as a whole and seeks to understand how its components interact with each other and with the environment (Huang, 2012; Beger et al., 2016; Marabita et al., 2022). Systems biology often involves the integration of multiple types of data, including genomics, transcriptomics, proteomics, and metabolomics, and the development of computational models to simulate and predict system behavior (Leduc et al., 2011).

Emerging high-dimensional single-cell and spatial technologies enable systems biology by providing a more comprehensive view of biological systems. Single-cell technologies allow researchers to study the behavior and interactions of individual cells, revealing cell-to-cell heterogeneity and enabling the identification of rare cell types and states in human health and disease (Han et al., 2018; Nagle et al., 2021). Spatial technologies provide information on the organization and context of cells within a tissue, allowing researchers to study the relationships between different cell types and their microenvironment (Moor and Itzkovitz, 2017; Wu et al., 2022a). By integrating these high-dimensional datasets, systems biologists can construct detailed models of complex biological systems that can be used to predict how they will respond to perturbations or treatments.

Large, atlas-style systems biology studies of human tissues and diseases such as cancer are providing unprecedented insights into the underlying mechanisms of disease and potential therapeutic targets (Regev et al., 2017; Rozenblatt-Rosen et al., 2020). These studies are generating comprehensive maps of the cellular and molecular landscapes of tissues and diseases, revealing previously unknown cell types, signaling pathways, and genetic alterations. By analyzing these maps, systems biologists are identifying new biomarkers for disease diagnosis and prognosis and developing personalized treatment strategies (Stunnenberg et al., 2019; Lukowski et al., 2019; Aizarani et al., 2019; Luca et al., 2021; Cheng et al., 2021). Additionally, these studies are providing a framework for understanding how different diseases are related and how they can be classified based on underlying biological mechanisms rather than traditional clinical criteria. Overall, systems biology offers a suite of powerful computational approaches to understand complex biological systems and developing more effective treatments for a range of diseases.

1.2.1 Integrative modeling of cancer progression across resolutions

The advent of single-cell and spatial molecular assays has provided tools for unbiased interrogation of tissue, lending insight to cellular activity and signaling in the context of populational heterogeneity and tissue morphology. These methods have disadvantages in efficiency and depth compared to traditional bulk measurements, but may complement one another through modality-specific strengths. Mutual and independent information offered by features of fluorescence microscopy and single-cell and spatial transcriptomics can be used to augment and enhance the aggregate dataset through linear and mixture models and matrix decomposition methods. Using regional genomic, transcriptomic, and proteomic measurements, we can map co-evolution of tumor and immune cells along spatially informed pseudotime of colorectal cancer (CRC)

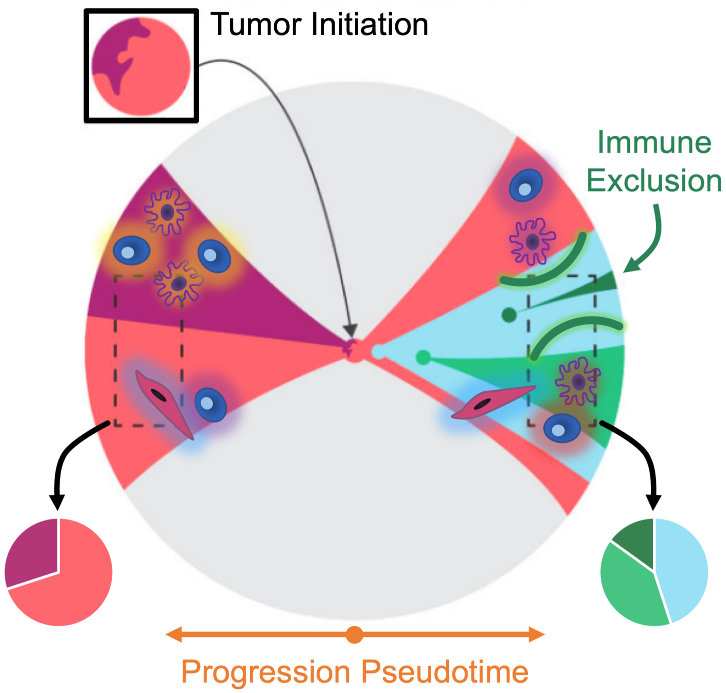
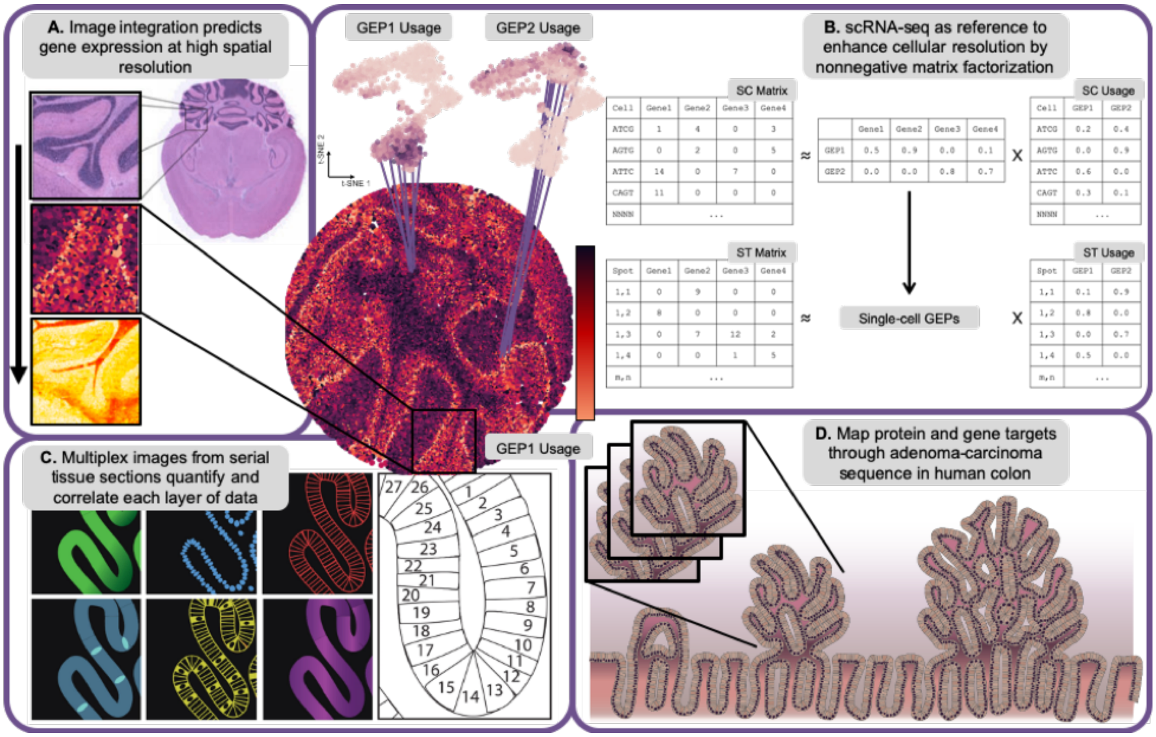


Figure 1.2: Integrated spatial and single-cell data will allow for characterization of co-evolving tumor and immune cells in CRC. Clonal mutations shown as colored fractions of tumor through space and time. Adapted from Sottoriva et al., 2015.

progression (Figure 1.2). A resulting understanding of the mutual progression of tumor and immune compartments has the potential to impact patient treatment decisions and contribute to cancer immunotherapy development (Rao et al., 2021; Lewis et al., 2021; Palla et al., 2022a).

An early proposal (ca. 2019) for the current work is detailed in Figure 1.3. Since the conception of this project, many technologies have emerged to address gaps in the field that our original aims focused on. For instance, image fusion models improve the resolution of spatial transcriptomics (ST) using registered brightfield and fluorescence data (Bergenstråhle et al., 2020), and deconvolution tools have been developed to distinguish cellular admixtures in ST (Andersson et al., 2020; Elosua-Bayes et al., 2021; Cable et al., 2021; Kleshchevnikov et al., 2022). For the purpose of this phylogeographical atlas, the general experimental design remained relatively unchanged, and successful integration of multiple spatially resolved modalities has yielded a valuable resource in the characterization of tumor evolution and microenvironmental interactions in CRC (Chapter 5; Figure 5.1).

A major challenge in building models across ST, multiplex immunofluorescence (MxIF), histology, and multiregional sequencing is accounting for differences in spatial resolution (Figure 1.4). Untargeted ST captures mRNA from up to dozens of single cells per microwell, necessitating the deconvolution of the


A. Image integration predicts gene expression at high spatial resolution
GEP1 Usage
GEP2 Usage
t-SNE 2
t-SNE 1
B. scRNA-seq as reference to enhance cellular resolution by nonnegative matrix factorization
SC Matrix
≈
X
SC Usage
ST Matrix
Single-cell GEPs
ST Usage
C. Multiplex images from serial tissue sections quantify and correlate each layer of data
GEP1 Usage
27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
D. Map protein and gene targets through adenoma-carcinoma sequence in human colon

Cell	Gene1	Gene2	Gene3	Gene4
ATCG	1	4	0	3
AGTG	0	2	0	5
ATTC	14	0	7	0
CAGT	11	0	0	0
NNNN	...			

	Gene1	Gene2	Gene3	Gene4
GEP1	0.5	0.9	0.0	0.1
GEP2	0.0	0.0	0.8	0.7

Cell	GEP1	GEP2
ATCG	0.2	0.4
AGTG	0.0	0.9
ATTC	0.6	0.0
CAGT	0.3	0.1
NNNN	...	

Spot	Gene1	Gene2	Gene3	Gene4
1,1	0	9	0	0
1,2	8	0	0	0
1,3	0	7	12	2
1,4	0	0	1	5
m,n	...			

Spot	GEP1	GEP2
1,1	0.1	0.9
1,2	0.8	0.0
1,3	0.0	0.7
1,4	0.5	0.0
m,n	...	

Figure 1.3: Graphical abstract for original data integration aims proposal.
(A) Schematic of image fusion to improve spatial resolution of one modality through modeling against another.
(B) Matrix representation of reference NMF (refNMF) for deconvolution of spatial transcriptomics to single-cell resolution.
(C) Diagram of single-cell segmentation from multiplex protein imaging.
(D) Application of multimodal spatial experiment detailed in A-C to map tissue dynamics in human colorectal tumors through serial sectioning of FFPE blocks.

resulting cell-type admixures. On the other hand, paired MxIF images offer subcellular resolution, and can be segmented into single-cell masks in order to validate ST deconvolution (McKinley et al., 2022; Figure 1.3C). Likewise, multiregional exome sequencing via laser capture microdissection (LCM-WES) requires large swaths of tissue for genome-scale profiling of somatic mutations and copy number variations (CNVs). However, ST offers the opportunity to confirm cancer cell fraction and CNVs within LCM regions of interest (ROIs; Satas et al., 2021). In the present studies, data integration approaches take advantage of mutual and independent information between modalities with varying spatial and cellular resolution in order to learn relationships between genes, proteins, and mutational drivers (Heiser et al., 2023; Chapter 5).

1.2.2 Impact on digital pathology and precision oncology

Spatially resolved technologies have particular relevance to digital pathology, which involves the use of high-dimensional images and computational tools to comprehensively profile tissues and diagnose disease

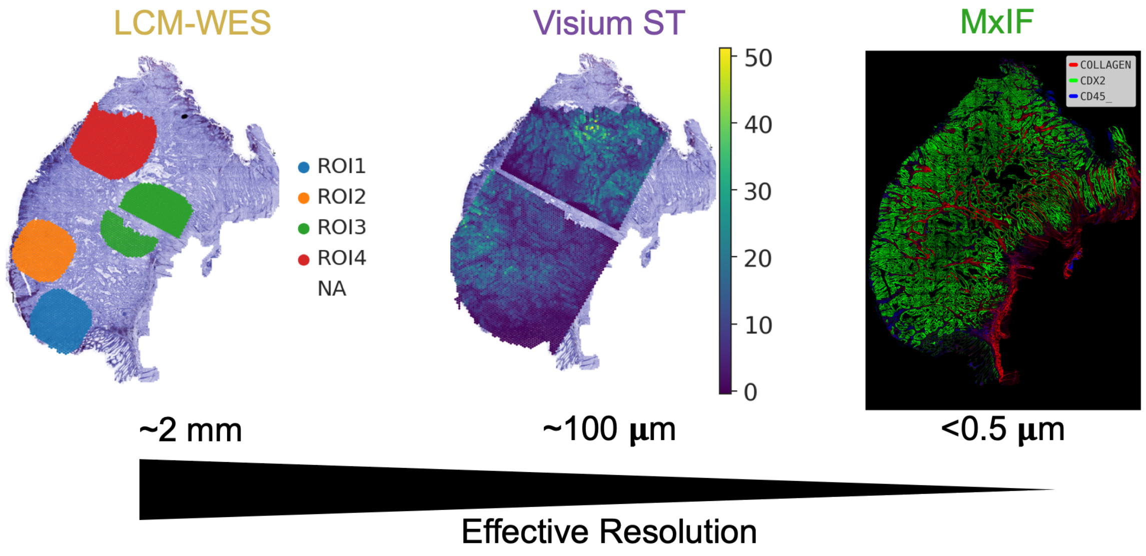


Figure 1.4: Effective spatial resolutions of various molecular assays used in these studies.

(Bankhead et al., 2017; Javed et al., 2020). With these assays, researchers can generate detailed molecular profiles of tumors, which can be used to identify specific biomarkers for diagnosis, prognosis, and patient stratification (Rana et al., 2018; Moehlin et al., 2021; Lin et al., 2023). Additionally, the information provided by these assays can be used to develop new diagnostic tools and therapies that are tailored to the specific molecular characteristics of individual tumors.

The work presented herein culminates in a phylogeographical atlas of CRC (Heiser et al., 2023; Chapter 5). Building on the big data principles and methods discussed in Chapters 2-4, we collected and integrated spatial multi-omic data, constructing patient-specific phylogeographic landscapes of tumor evolution and global trajectories from normal colonic tissue to malignancy. In doing so, we revealed microenvironmental and clonal dynamics along tumor progression and identified key genes and cell states associated with immune exclusion in CRC. We expect this atlas dataset and the associated multi-scale modeling techniques to serve as a valuable resource for future data mining efforts related to stratification and targeted treatment of CRC. Overall, we demonstrated the potential of computational systems biology approaches to drive precision oncology.

As genomic, transcriptomic, and proteomic profiling technologies continue to advance in sensitivity and scale, the volume and complexity of generated data describing human health and disease will expand exponentially. This enormous challenge necessitates the development, validation, and continual refinement of mathematical and computational techniques for automated, reproducible, and interpretable modeling of molecular dynamics at multiple resolutions. Overall, big data and systems biology approaches offer valuable tools for advancing our understanding of cancer biology and developing more effective cancer treatments. By providing detailed information about the molecular characteristics of tumors and their microenvironments, these approaches help researchers identify new therapeutic targets and prognostic biomarkers. As such, they have the potential to revolutionize cancer diagnosis and treatment and to improve patient outcomes.

CHAPTER 2

A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques

Adapted from:

Heiser, C. N. and Lau, K. S. (2020). A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques. *Cell Reports*, 31(5):107576

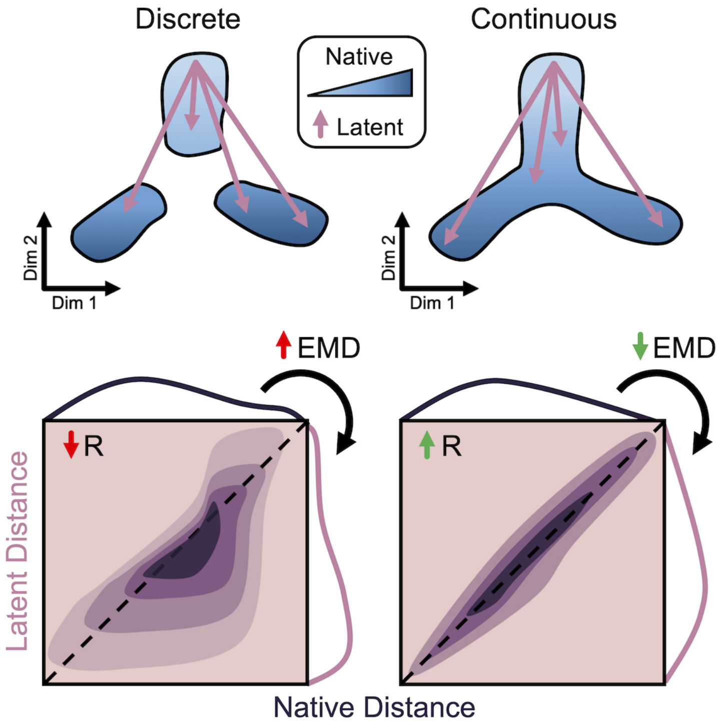


Figure 2.1: Heiser and Lau, 2020 graphical abstract

2.1 Summary

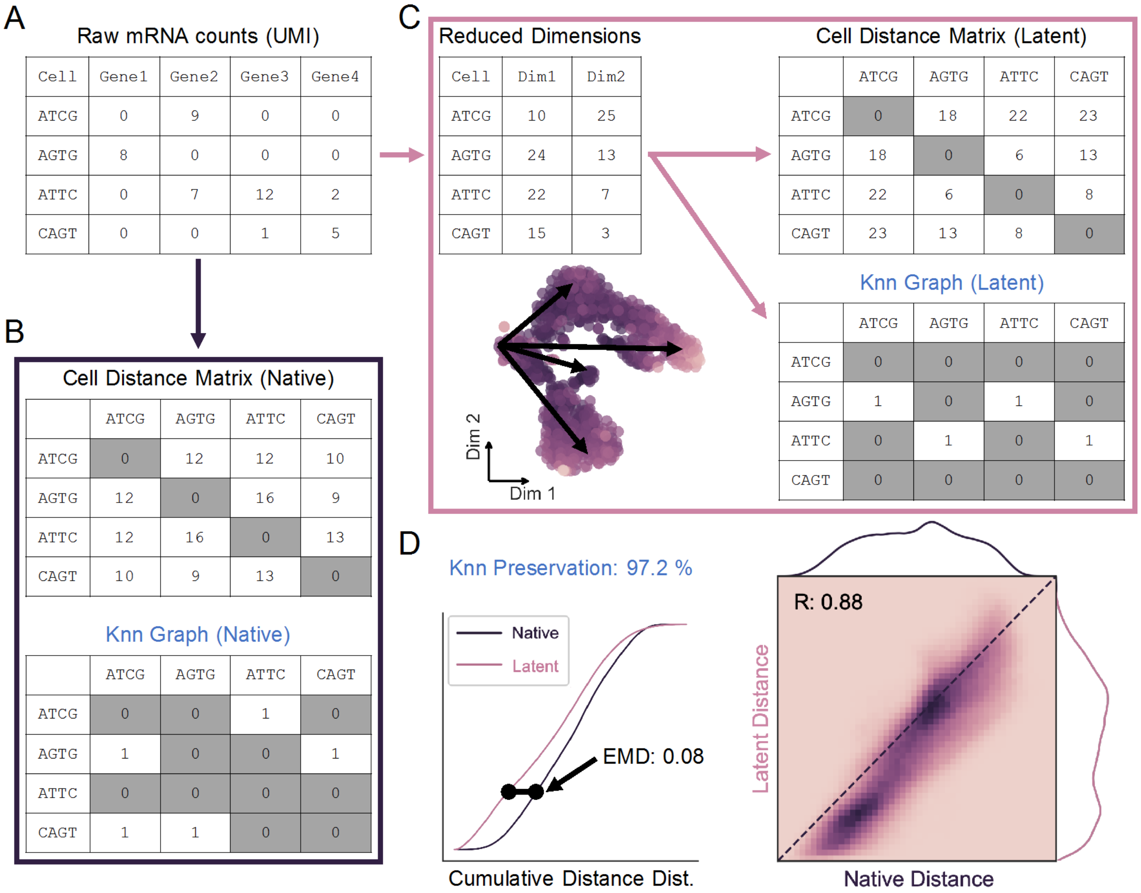
High-dimensional data, such as those generated by single-cell RNA sequencing (scRNA-seq), present challenges in interpretation and visualization. Numerical and computational methods for dimensionality reduction allow for low-dimensional representation of genome-scale expression data for downstream clustering, trajectory reconstruction, and biological interpretation. However, a comprehensive and quantitative evaluation of the performance of these techniques has not been established. We present an unbiased framework that defines metrics of global and local structure preservation in dimensionality reduction transformations. Using

discrete and continuous real-world and synthetic scRNA-seq datasets, we show how input cell distribution and method parameters are largely determinant of global, local, and organizational data structure preservation by 11 common dimensionality reduction methods.

2.2 Introduction

Single-cell RNA sequencing (scRNA-seq) offers parallel, genome-scale measurement of tens of thousands of transcripts for thousands of cells (Klein et al., 2015; Macosko et al., 2015). Data of this magnitude provide powerful insight toward cell identity and developmental trajectory – states and fates – which are used to interrogate tissue heterogeneity and characterize disease progression (Regev et al., 2017; Wagner et al., 2018). Yet, extracting meaningful information from such high-dimensional data presents a massive challenge. Numerical and computational methods for dimensionality reduction have been developed to reconstruct underlying distributions from native "gene space" and provide low-dimensional representations of single-cell data for more intuitive downstream interpretation. Basic linear transformations such as principal component analysis (PCA) have proven to be valuable tools in this field (Sorzano et al., 2014; Tsuyuzaki et al., 2020). However, given the distribution and sparsity of scRNA-seq data, complex, nonlinear transformations are often required to capture and visualize expression patterns. Unsupervised machine learning techniques are being rapidly developed to assist researchers in single-cell transcriptomic analysis (Van der Maaten and Hinton, 2008; Pierson and Yau, 2015; Wang et al., 2017; Linderman et al., 2017; Becht et al., 2018; Ding et al., 2018; Lopez et al., 2018; McInnes et al., 2018; Risso et al., 2018; Eraslan et al., 2019; Townes et al., 2019). Because these techniques condense cell features in the native space to a small number of latent dimensions, lost information can result in exaggerated or dampened cell-cell similarity. Furthermore, depending on input data and user-defined parameters, the structure of resulting embeddings can vary greatly, potentially altering biological interpretation (Kobak and Linderman, 2019).

With a deluge of computational techniques for dimension reduction, the field is lacking a comprehensive assessment of native organizational distortion consequential to such methods. We present an unbiased, quantitative framework for evaluation of data structure preservation by dimensionality reduction transformations. We propose metrics for broad characterization of these methods based on cell-cell distance in native, high-dimensional space. Initial benchmarking of eleven published software tools on discrete and continuous cell distributions shows global, local, and organizational data structure conservation under different parameter and input conditions. Applying our framework to additional data types underscores the modality- and dataset-specific nature of dimension reduction performance.



Raw mRNA counts (UMI)

Cell	Gene1	Gene2	Gene3	Gene4
ATCG	0	9	0	0
AGTG	8	0	0	0
ATTC	0	7	12	2
CAGT	0	0	1	5

Cell Distance Matrix (Native)

	ATCG	AGTG	ATTC	CAGT
ATCG	0	12	12	10
AGTG	12	0	16	9
ATTC	12	16	0	13
CAGT	10	9	13	0

Knn Graph (Native)

	ATCG	AGTG	ATTC	CAGT
ATCG	0	0	1	0
AGTG	1	0	0	1
ATTC	0	0	0	0
CAGT	1	1	0	0

Reduced Dimensions

Cell	Dim1	Dim2
ATCG	10	25
AGTG	24	13
ATTC	22	7
CAGT	15	3

Cell Distance Matrix (Latent)

	ATCG	AGTG	ATTC	CAGT
ATCG	0	18	22	23
AGTG	18	0	6	13
ATTC	22	6	0	8
CAGT	23	13	8	0

Knn Graph (Latent)

	ATCG	AGTG	ATTC	CAGT
ATCG	0	0	0	0
AGTG	1	0	1	0
ATTC	0	1	0	1
CAGT	0	0	0	0

Figure 2.2: Cell distance distributions describe global structure of high-dimensional data.
(A) Representation of scRNA-seq counts matrix.
(B) Cell-cell distances in native gene space are calculated to generate an m x m matrix, where m is the total number of cells. K nearest-neighbors (Knn) graph is constructed from these distances as a binary m x m matrix.
(C) Upon transformation to low-dimensional space, a distance matrix and Knn graph can be calculated as in B.
(D) Distance matrices from native (B) and latent (C) spaces are used to build cumulative probability density distributions, which can be compared to one another by Earth Mover's Distance (EMD, left). Unique cell-cell distances are correlated (right), and Knn preservation represents element-wise comparison of nearest-neighbor graph matrices in each space. See also Figure S2.

2.3 Results

2.3.1 Cell distance distributions describe global structure of high-dimensional data

In order to evaluate dimensionality reduction techniques, Euclidean cell-cell distance in native, high-dimensional space is used as a quantitative standard. In scRNA-seq, counts of unique molecular identifiers (UMI) for each gene make up the features of the dataset, while every observation represents a single cell (Figure 2.2A). In this way, transcriptomic data is represented as an m x n matrix (observations x features).

Global data structure in the native space can be constructed by first calculating an m x m matrix containing the pairwise distances between all observations in n dimensions (Figure 2.2B, top). The upper triangle of this symmetric distance matrix contains unique cell distances in the dataset, which can then be represented by a probability density distribution as in Figure 2.2D. From these distances, local "neighborhoods" can be defined in the form of a K nearest-neighbor (Knn) graph. The Knn graph is represented as a binary m x m matrix that defines the K cells with the shortest distances to each cell in the dataset (Figure 2.2B, bottom). Similarly, a distance distribution and Knn graph can be constructed from a low-dimensional latent space resulting from dimensionality reduction (Figure 2.2C).

Preservation of unique distances following dimension reduction is measured by direct Pearson correlation, while structural alteration of the cell distance distribution is quantified by the Wasserstein metric or Earth-Mover's Distance (EMD; Figure 2.2D). Widely applied to image processing, EMD determines the energy cost associated with shifting one distribution to another (Werman et al., 1985; Rubner et al., 1998; Levina and Bickel, 2001). This metric is ideal for our application as it scales linearly with separation of the means of two continuous distributions – in contrast to similar Cramér-von Mises or Kolmogorov-Smirnov distances – and therefore captures maximum variability (Cramér, 1928; Kolmogorov, 1933). Finally, preservation of a Knn graph before and after low-dimensional embedding can also be quantified as the percentage of total matrix elements conserved in order to describe maintenance of local substructures in the data.

2.3.2 Discrete and continuous cell distributions exemplify common biological patterns

A major consideration for testing dimensionality reduction techniques is the true structure of the input data in native, high-dimensional space. For the scope of our evaluation, we identify two overarching classes of scRNA-seq data for proof-of-principal: discrete and continuous. Discrete single-cell data are comprised of differentiated cell types with unique, highly discernable gene expression profiles. These data include classic PBMC experiments and neuronal datasets which can be easily clustered into distinct cell types (Zeisel et al., 2015; Rheaume et al., 2018). Conversely, continuous data contain multi-faceted expression gradients present during cell development and differentiation, and are commonly associated with dynamic systems such as erythropoiesis or embryonic development (Tusi et al., 2018; Wagner et al., 2018).

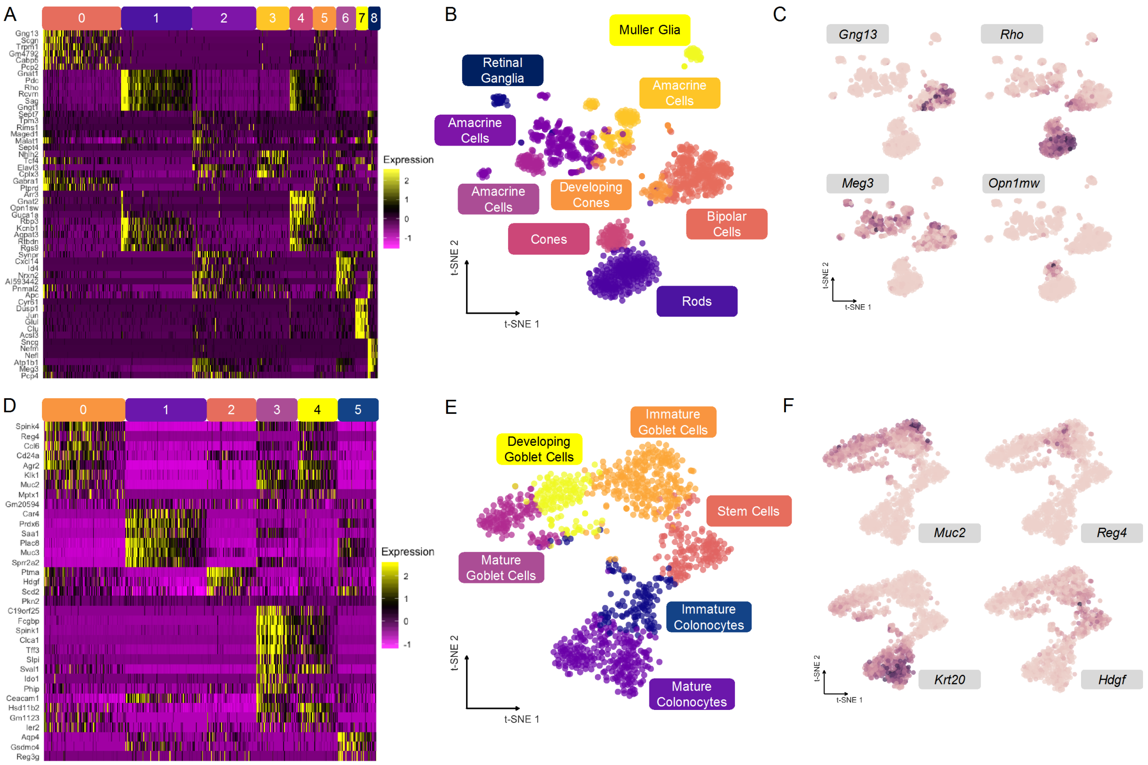


Figure 2.3: Discrete and continuous cell distributions exemplify common biological patterns.
(A) Relative expression of top genes in each cluster for mouse retina dataset.
(B) t-SNE embedding primed with 100 principal components of retina dataset with overlay of consensus clusters.
(C) t-SNE projection from B with overlay of marker genes used to identify cell types in A.
(D-F) Same as in A-C, for mouse colonic epithelium dataset. See also Figure S3.

Mouse retina cells, analyzed using Drop-seq by Macosko and coworkers, provide a discrete cell distribution for our analysis (Macosko et al., 2015). Counts data from 20,478 genes for 1,326 cells were analyzed using Louvain clustering to determine cell clusters (Figure 2.3A; Levine et al., 2015). We performed relatively coarse clustering, ignoring subtype heterogeneity in favor of clusters reflecting principal cell identity amenable to our downstream analyses (see Appendix A2.1). A t-SNE projection primed with 100 principal components (PCs) of all transcript counts allows for visualization of the data structure and represented cell types (Figure 2.3B). As evident from the 2D embedding, these data are highly discrete, and constituent cell clusters are easily distinguished by expression of marker genes identified in Macosko et al., 2015 (Figure 2.3C, Figure S3A).

Mouse colon data, representing a continuous distribution of actively differentiating cells along the crypt axis of the colonic epithelium, were generated using inDrops scRNA-seq (Herring et al., 2018). Counts data from 25,504 genes for 1,117 cells were similarly clustered and embedded using t-SNE to visualize continuous data structure (Figure 2.3D,E). The six clusters form a branching continuum of cell states identified by expression markers (Figure 2.3F, Figure S3B), resolving two major lineages in the colon: absorptive and secretory cells (Lepourcelet et al., 2005; Tamura et al., 2007; Larsson et al., 2012). These clusters are linked together by pseudotemporal trajectories and thus their arrangement is expected to be conserved upon low-dimensional embedding.

2.3.3 Input cell distribution determines performance of global structure preservation

Using metrics outlined in Figure 2.2, we compared eleven dimensionality reduction techniques applied to continuous and discrete datasets. To allow for direct input to these tools and comparison with linear PCA in the following analyses, raw counts for both datasets were feature-selected to the 500 most variable genes. Alternatively, a common preprocessing approach is initial dimension reduction with PCA, and we compare 500 PCs to our 500 variable genes (VGs) to demonstrate how this may affect downstream structure preservation (Figure S4A). Though our framework measures structure preservation relative to the input cell distribution, performance of dimension reduction methods is expected to vary under different preprocessing conditions, and we encourage the use of our metrics to evaluate not only the tools themselves, but also upstream handling of the data.

Calculating our metrics on all cells in the dataset, we first assess global structure preservation following transformation to a latent space. Representative examples of 2D projections and their corresponding distance distributions and correlations using SIMLR for the retina dataset and UMAP for the colon dataset are shown in Figure 2.4A and Figure 2.4H, respectively. Notably, the largest discrepancy in structural preservation is between the two datasets, highlighting the significance of input cell distribution to overall method performance. For example, Knn preservation is intuitively higher for most methods when applied to the colon dataset, reflecting the notion of continuous neighborhoods – a moving window of expression gradients – connecting all cells through developmental pseudotime. Another important observation regarding the dimension-reduced spaces involves the directionality of the cell distance distribution shift. A compression of distances from native to latent space is indicated by a shift left in the cumulative distance distribution (Figure 2.4B,J, Figure S2A) or below the identity line in the unique distance correlation (Figure 2.4D,L, Figure S2B). Alternatively, a shift right in the cumulative distance distribution or above the identity line of the distance correlation signifies an exaggeration of native distances. These phenomena are important in the context of global versus local structure preservation. For example, UMAP appears to compress small, local distances to a greater extent than t-SNE, while both methods maintain relative global structure as indicated by a favorable correlation of large distances. Although this characteristic of UMAP embeddings causes greater information loss reflected in less favorable preservation metrics (Figure 2.4C,K), clusters within the resulting projections tend to be visually condensed and perhaps more easily interpreted (Figure S4B,C).

These findings are particularly important when considering datasets and data types beyond the scope of this study. For instance, other single-cell technologies such as assay for transposase-accessible chromatin (scATAC-seq) and mass cytometry (CyTOF) have expectedly diverse distributions of cell-cell distances due to technical differences in dynamic range, dropout rate, and noise. Applying our structural preservation framework to two datasets used to benchmark UMAP against t-SNE, FIt-SNE, and scvis (Becht et al., 2018; Figure S5), we identify a clear distinction between these CyTOF and scRNA-seq datasets that is deterministic of method performance (see Appendix A2.1: Theoretical basis for difference in dimension reduction performance across single-cell modalities). Indeed, we assert that input data structure is highly variable across single-cell technologies and biological samples (Figure S4D), and we recommend evaluating dimensionality reduction tools in the context of their intended application.

2.3.4 Parameter optimization plays key role in structural preservation

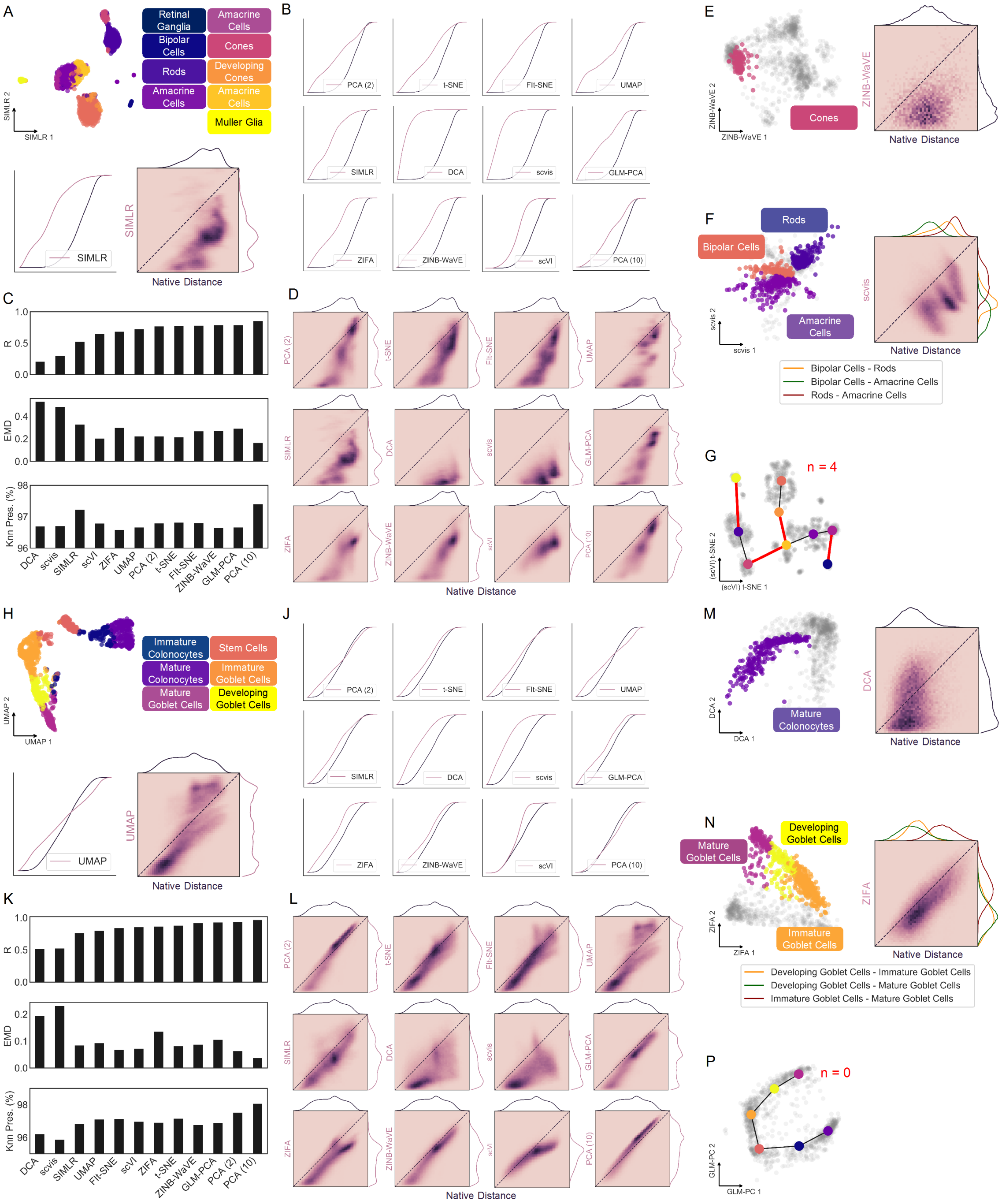
User-defined parameters for unsupervised algorithms often present themselves as “black-box” knobs with unknown consequences. Tuning these parameters can be a daunting task for the single-cell analyst, but is known to be crucial to algorithm performance (Belkina et al., 2018; Kobak and Linderman, 2019; Tsuyuzaki et al., 2020). Using our proposed metrics, we evaluated global structure preservation across a range of perplexity (n-neighbors) values for t-SNE and UMAP applied to both discrete and continuous data. Through a balance of distance correlation, EMD, and Knn preservation, we can identify an initial range of optimal perplexity values between 3 and 10 percent of the total number of cells in the dataset (Figure S4E).

Additionally, as our framework is agnostic to the distance metric and neighborhood size (K) chosen for evaluation, we can perform cursory comparisons of possible alternatives to Euclidean distance (Figure S4F) and titrate the value of K to determine its effect on observed preservation values (Figure S4G). Here we observe optimal K between 3 and 10 percent of the dataset size to reliably discriminate between methods, in accordance with the perplexity parameter.

2.3.5 Substructure analysis elucidates contribution to global performance

To corroborate results of global structure preservation and dissect contribution of local (within cluster) and organizational (between cluster) distances to overall dimension reduction performance, clusters were isolated for targeted substructure quantification. Here, we can measure distance preservation of individual clusters as well as distances between clusters to emphasize local arrangement (Figure S2C,D).

Retinal cone cells (Figure 2.3A, cluster 4, n = 94) were used as an example of local distances in the discrete dataset, while mature colonocytes (Figure 2.3D, cluster 1, n = 273) were isolated in the colon dataset


A
Retinal Ganglia
Amacrine Cells
Bipolar Cells
Cones
Rods
Developing Cones
Amacrine Cells
Amacrine Cells
Muller Glia
SIMLR 2
SIMLR 1
SIMLR
SIMLR
Native Distance
B
PCA (2)
t-SNE
FIt-SNE
UMAP
SIMLR
DCA
scvis
GLM-PCA
ZIFA
ZINB-WaVE
scVI
PCA (10)
C
R
1.0
0.5
0.0
EMD
0.4
0.2
0.0
Knn Pres. (%)
98
97
96
DCA
scvis
SIMLR
scVI
ZIFA
UMAP
PCA (2)
t-SNE
FIt-SNE
ZINB-WaVE
GLM-PCA
PCA (10)
D
PCA (2)
t-SNE
FIt-SNE
UMAP
SIMLR
DCA
scvis
GLM-PCA
ZIFA
ZINB-WaVE
scVI
PCA (10)
Native Distance
E
ZINB-WaVE 2
ZINB-WaVE 1
Cones
ZINB-WaVE
Native Distance
F
Rods
Bipolar Cells
Amacrine Cells
scvis 2
scvis 1
scvis
Native Distance
Bipolar Cells - Rods
Bipolar Cells - Amacrine Cells
Rods - Amacrine Cells
G
n = 4
(scVI) t-SNE 2
(scVI) t-SNE 1
H
Immature Colonocytes
Stem Cells
Mature Colonocytes
Immature Goblet Cells
Mature Goblet Cells
Developing Goblet Cells
UMAP 2
UMAP 1
UMAP
UMAP
Native Distance
J
PCA (2)
t-SNE
FIt-SNE
UMAP
SIMLR
DCA
scvis
GLM-PCA
ZIFA
ZINB-WaVE
scVI
PCA (10)
K
R
1.0
0.5
0.0
EMD
0.2
0.1
0.0
Knn Pres. (%)
98
96
DCA
scvis
SIMLR
UMAP
FIt-SNE
scVI
ZIFA
t-SNE
ZINB-WaVE
GLM-PCA
PCA (2)
PCA (10)
L
PCA (2)
t-SNE
FIt-SNE
UMAP
SIMLR
DCA
scvis
GLM-PCA
ZIFA
ZINB-WaVE
scVI
PCA (10)
Native Distance
M
DCA 2
DCA 1
Mature Colonocytes
DCA
Native Distance
N
Mature Goblet Cells
Developing Goblet Cells
Immature Goblet Cells
ZIFA 2
ZIFA 1
ZIFA
Native Distance
Developing Goblet Cells - Immature Goblet Cells
Developing Goblet Cells - Mature Goblet Cells
Immature Goblet Cells - Mature Goblet Cells
P
n = 0
GLM-PC 2
GLM-PC 1

(Figure 2.4E,M). Local distance compression represents the overarching trend for the eleven evaluated tools, indicated by a correlation shift below the identity line (Figure S4H,J). The latent spaces from scVI and 10-component PCA are notable exceptions, yielding the two lowest EMD values for each dataset (Figure S4M). This most likely results from the 10-dimensional latent spaces of these methods capturing more cellular variability than 2D projections, and these two embeddings should be considered with this caveat in the context of our larger analysis. Added noise in the SIMLR latent space of mouse retina cells indicates a disagreement with Louvain cluster membership, and may be attributed to the truncated, 500-feature input used for our analysis (Figure S4H). Moreover, this observation suggests that discrete, "on-off" expression patterns are less robust to dropouts that cause mis-assignment of cell type than continuous gradients of gene expression.

Besides maintenance of local structure, dimensionality reduction methods are also tasked with preserving cellular neighborhoods, or relationships between clusters. By calculating the distribution of distances from cells in one cluster to those in another, we can evaluate these associations to investigate organization of data substructures (Figure S2C, Figure 2.4F,N). In the mouse retina dataset, distances between bipolar cells, rod cells, and amacrine cells (Figure 2.3A, clusters 0, 1, 2, n = 309, 281, 258) are marked largely by compression, with some tools altering the arrangement of the three clusters (Figure S4K, red boxes). For example, the bipolar and amacrine clusters are closest to one another in the native gene space, but bipolar cells are closer to rod cells in the UMAP embedding, indicated by the ordering of each distribution. Conversely, relative distances between three adjacent clusters along the goblet cell lineage (Figure 2.3D, clusters 0, 3 and 4, n = 274, 140 and 135) are more highly conserved by all embeddings. These results confirm that related cells in continuous scRNA-seq data are tethered to their neighbors through intermediate expression states, resulting

Figure 2.4 *(preceding page)*: Global and local structure preservation analysis.
(A) Example 2D projection of mouse retina data using SIMLR with cluster overlay (top). Cumulative distance distributions for native and latent spaces (bottom left) and 2D histogram representing correlation between unique distances (bottom right).
(B) Cumulative distance distributions of evaluated projections of retina data.
(C) Summary of structure preservation metrics for retina data.
(D) 2D histograms of cell distance correlations for retina data.
(E) Example 2D projection of retina data using ZINB-WaVE and overlay of cone cells (left) and 2D histogram representing correlation between the two sets of unique distances (right).
(F) Same as in E for distances between bipolar, amacrine, and rod cell clusters, using scvis projection.
(G) Example graph representation of cluster topology for retina dataset, using t-SNE projection primed with scVI latent space. Red edges represent those not present in minimum spanning tree of native graph.
(H) Same as in A, with UMAP projection of mouse colon data.
(J-L) Same as in B-D, for colon data.
(M) Same as in E, with DCA projection of mature colonocytes.
(N) Same as in F for distances between immature, developing, and mature goblet cell clusters, using ZIFA projection.
(P) Same as in G for colon dataset, using GLM-PCA projection.

in improved structure preservation upon latent projection (Figure S4L,N).

To further capture substructure rearrangement in low-dimensional embeddings, we construct a coarse graphical representation of our native and latent spaces, with minimum spanning trees (MSTs) connecting nearest neighbor cluster centroids (Figure 2.4G,P). Comparing the edges of each graph allows us to evaluate latent cluster topology relative to the native space, as permuted edges indicate rearrangement of substructures following dimension reduction. Once again, we see a global increase in topological preservation of continuous versus discrete data, corroborating previous observations (Figure S4P-R).

2.3.6 Simulated datasets with defined topology validate observations

Single-cell data with expected global topology were simulated using Splatter (Zappia et al., 2017). Three distinct lineages, equally separated in high dimensional space, originate from a common state. A discrete simulation was generated by removing the central shared state (Figure 2.5A), while the continuous dataset maintains complete connectivity between the three developmental paths (Figure 2.5F). Pseudotime (PT) values, assigned to each cell by the simulation, should correlate directly to distance in the embedding and can thus be used as an alternative ground-truth native structural distribution for our framework. Both simulations were processed by previously evaluated dimensionality reduction tools (Figure 2.5B,C,G,H), and latent distances between cells from the three defined paths were correlated to pairwise sums of corresponding PT values (Figure 2.5D,E,J,K). In this way, large PT sums between cells at the ends of each simulated path and small PT sums between cells near the shared central state should have the largest and smallest distances from one another in an ideal latent embedding, respectively. Again, dimension reduction of discrete data performs poorly compared to the entirely continuous simulation. All embeddings generally cluster each path properly (Figure 2.5B), but misorientation of these clusters from their shared center results in negative structural correlation for some embeddings including t-SNE and UMAP (Figure 2.5C,E).

2.4 Discussion

As high-dimensional data become increasingly pervasive in systems biology, computational tools for reliable and reproducible analysis of these data are tremendous assets to discovery. Dimensionality reduction techniques allow for embedding cellular observations with tens of thousands of features into a low-dimensional space for visualization and downstream processing. We present an unbiased, quantitative framework based on native cell distance to evaluate data structure preservation by these tools.

We identified dispersion trends in local and global distance distributions that denote expansion and contraction of native cell distances. This allowed us to evaluate general performance of dimensionality reduction methods on entire single-cell datasets and take a deeper dive to examine how distances within or between

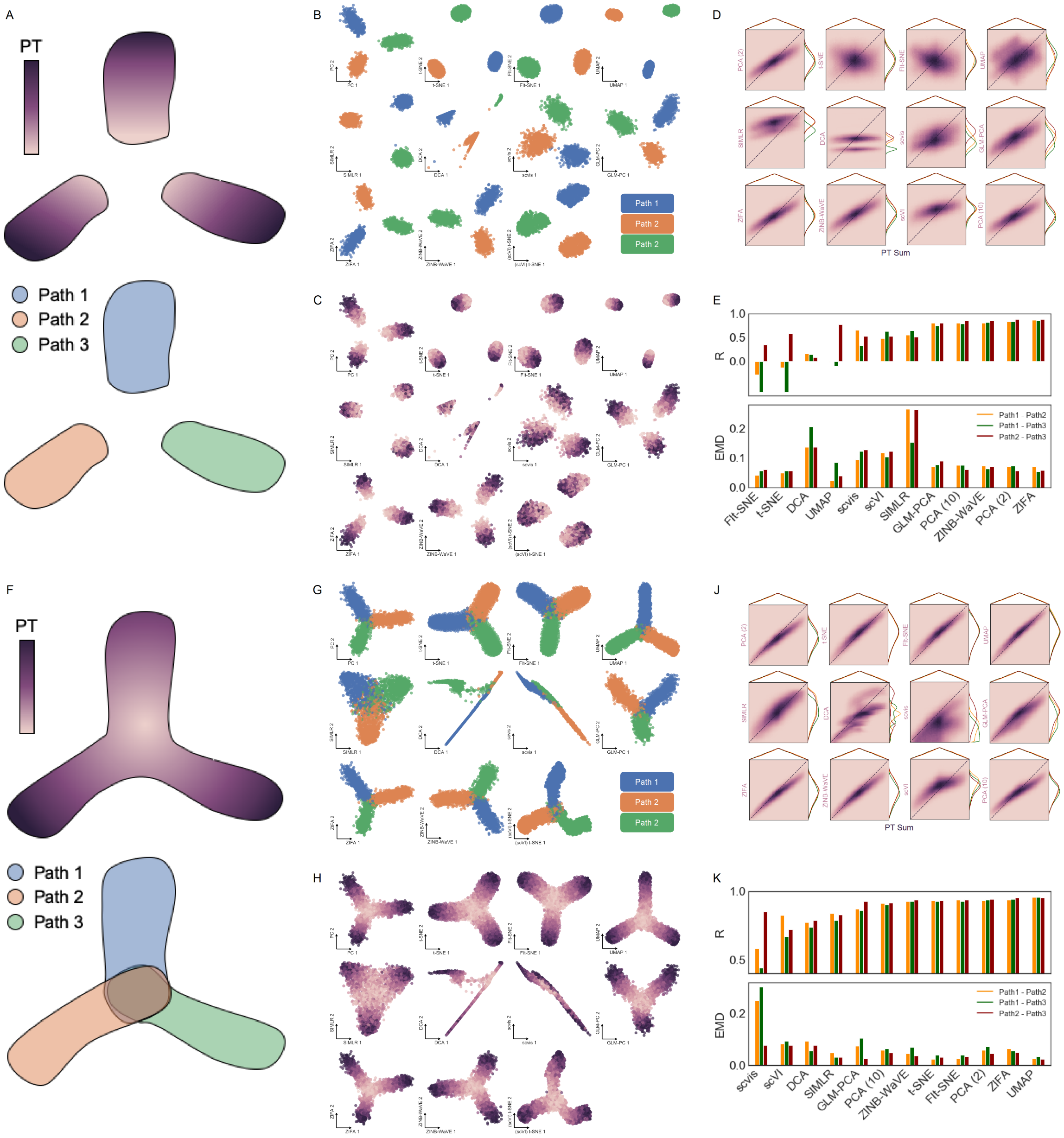


Figure 2.5: Simulated datasets with defined topology validate observations.

(A) Diagram of discrete synthetic data with ground-truth topology defined by three, equally spaced developmental paths along directional pseudotime (PT) from a common source state (removed to discretize paths).

(B) 2D embeddings by 11 dimensionality reduction tools showing unique paths defined in simulation.

(C) Same as in B, with overlay of PT values for each cell as defined in simulation.

(D) 2D histograms showing correlation of pairwise distances between cells in each of the three developmental paths with the sum of PT values between each pair of cells as ground-truth topology.

(E) Summary of correlation and EMD values between cells in each path for all dimensionality reduction methods.

(F-K) Same as in A-E for continuous simulation. See also Figure S5.

clusters contribute to the global structure of a low-dimensional embedding (Figure 2.4, Figure S4). With a goal of grouping cells by their gene expression profiles, most dimension reduction tools evaluated herein compress local distances, embellishing cluster similarity, while maintaining or expanding global distances, exaggerating cluster distinction. These characteristics of dimensionality reduction methods are desirable for most applications. However, resolution of rare cell types and sub-cluster heterogeneity may be lost, stressing the importance of preprocessing, feature selection, and user-defined parameters (Figure S4).

Discrete scRNA-seq data are more susceptible to structural perturbation by downstream dimension reduction, as indicated by larger EMD values and lower distance correlations in the retina dataset than colonic epithelial data (Figure 2.4). We also observed cluster rearrangement within the retina dataset, suggesting that relative substructure organization is poorly defined for discrete datasets while continuous cell distributions are more robust to these effects (Figure S4). Cursory exploration of perplexity and K parameters in t-SNE and UMAP – as well as alternative preprocessing approaches – reveals a range of optimal values that yield favorable structure preservation metrics, endorsing the need for parameter and preprocessing optimization for dimensionality reduction of single-cell datasets (Figure S4E,G). The above observations were confirmed using simulated datasets with defined global topology that could be quantified in place of native cell distances (Figure 2.5).

Finally, a careful look at additional synthetic and real-world data confirms that behavior of dimensionality reduction methods is primarily driven by the input cell distance distribution that is modality- and dataset-specific (Appendix A2.1, Figure S5). Our findings challenge the context in which dimensionality reduction methods are benchmarked and indicate that performance characterization is often not universally extensible. Consequently, we encourage evaluation of such tools on data types, datasets, and preprocessing approaches specific to the user's intended application.

CHAPTER 3

Automated quality control and cell identification of droplet-based single-cell data using dropkick

Adapted from:

Heiser, C. N., Wang, V. M., Chen, B., Hughey, J. J., and Lau, K. S. (2021). Automated quality control and cell identification of droplet-based single-cell data using dropkick. *Genome Research*, 31(10):1742–1752

3.1 Summary

A major challenge for droplet-based single-cell sequencing technologies is distinguishing true cells from uninformative barcodes in data sets with disparate library sizes confounded by high technical noise (i.e., batch-specific ambient RNA). We present dropkick, a fully automated software tool for quality control and filtering of single-cell RNA sequencing (scRNA-seq) data with a focus on excluding ambient barcodes and recovering real cells bordering the quality threshold. By automatically determining data set–specific training labels based on predictive global heuristics, dropkick learns a gene-based representation of real cells and ambient noise, calculating a cell probability score for each barcode. Using simulated and real-world scRNA-seq data, we benchmarked dropkick against conventional thresholding approaches and EmptyDrops, a popular computational method, showing greater recovery of rare cell types and exclusion of empty droplets and noisy, uninformative barcodes. We show for both low- and high-background data sets that dropkick's weakly supervised model reliably learns which genes are enriched in ambient barcodes and draws a multidimensional boundary that is more robust to data set–specific variation than existing filtering approaches. dropkick provides a fast, automated tool for reproducible cell identification from scRNA-seq data that is critical to downstream analysis and compatible with popular single-cell Python packages.

3.2 Introduction

Single-cell RNA sequencing (scRNA-seq) allows for untargeted profiling of genome-scale expression in thousands of individual cells, providing insights into tissue heterogeneity and population dynamics. Droplet-based platforms that involve microfluidic encapsulation of cells in water-oil emulsions (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017) have grown widely popular for their robustness and throughput. The use of barcoded poly-thymidine capture oligonucleotides provides information for assigning eventual sequencing reads to each droplet downstream of bulk library preparation. Due to the low cellular density required to avoid doublets (i.e., two or more cells captured in the same droplet), the vast majority of droplets are empty, ideally containing only tissue dissociation buffer and a barcoded RNA-capture bead with no cellular RNA. However, during the tissue dissociation process, cell death, lysis, and leakage result in the shedding of ambient mRNA into the supernatant solution, which is then captured as background in droplets containing individual cells and so-called "empty droplet" reactions. Ultimately, a droplet-based scRNA-seq dataset contains up to hundreds of thousands of barcodes that correspond to these "empty droplets" which include sequenced material from ambient RNA alone.

In order to prepare these data for downstream analysis, empty droplets and other uninformative barcodes with little to no molecular information must be removed. Often, computational biologists will define manual thresholds on global heuristics such as total counts of unique molecular identifiers (UMI) or the total number of genes detected in each barcode in order to isolate high-quality cells. While these hard cutoffs may generally yield expected cell populations and remove the bulk of populational noise in low-background samples, they are highly arbitrary, batch-specific, and generally biased against cell types with low RNA content or genetic diversity (Lun et al., 2019). Furthermore, lenient thresholds often yield filtered datasets with populations of dead and dying cells or empty droplets with high ambient RNA content, especially in encapsulations with high background resulting from tissue-specific cell viability and dissociation protocols. These cell clusters may be gated out manually by the experienced single-cell biologist, but they will distort dimension-reduced embeddings and alter statistical testing for differential gene expression if left unchecked.

Here we introduce dropkick, a fully automated machine learning software tool for data-driven filtering of droplet-based scRNA-seq data. dropkick provides a quality control (QC) module for initial evaluation of global distributions that define barcode populations (real cells vs. empty droplets) and quantifies the batch-specific ambient gene profile. The dropkick filtering module establishes initial thresholds on predictive global heuristics using an automated gradient-descent method, then trains a gene-based logistic regression model to assign confidence scores to all barcodes in the dataset. dropkick model coefficients are sparse and biologically informative, identifying a minimal number of gene features associated with empty droplets and low-quality cells in a weakly supervised fashion. The following study aims to show how dropkick outperforms basic threshold-based filtering and a similar data-driven model (Lun et al., 2019) in recovery of expected cell types and exclusion of empty droplets, with robustness and reproducibility across encapsulation platforms, samples, and varying degrees of noise from ambient RNA.

3.3 Results

3.3.1 Evaluating dataset quality with the dropkick QC module

Global data quality and predominance of ambient RNA affect both reliable cell identification as well as downstream analyses including clustering, cell type annotation, and trajectory inference in scRNA-seq data (Young and Behjati, 2018; Fleming et al., 2019; Yang et al., 2020). Single-cell data with a low signal-to-noise

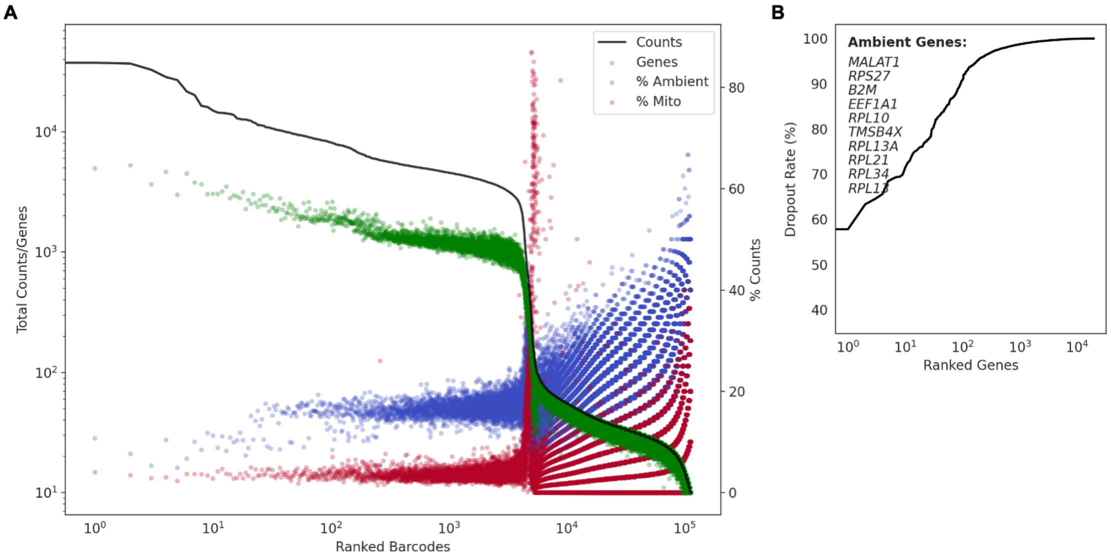


Figure 3.1: Evaluating dataset quality with the dropkick QC module.
(A) Profile of total counts (black trace) and genes (green points) detected per ranked barcode in the 4k pan-T cell dataset (10x Genomics). Percentage of mitochondrial (red) and ambient (blue) reads for each barcode included to denote quality along dataset profile.
(B) Profile of dropout rate per ranked gene. Ambient genes are identified by dropkick and used to calculate ambient percentage in A.

ratio due to high ambient background can result in information loss that may ultimately confound cell type and cell state identification and related statistical analyses (Zhang et al., 2019). For instance, a scRNA-seq encapsulation with a high degree of cell lysis can cause marker genes from abundant cell types to be present in the ambient RNA profile that contaminates all cell barcodes. In this scenario, global differences between cell populations would be diminished by the common detection of ambient noise, leading to loss of resolution in inference of cell identity and state.

In order to quantify ambient contamination that reduces this batch-specific signal-to-noise ratio, we have developed a comprehensive quality control report for unfiltered, post-alignment UMI counts matrices. Figure 1 provides an example dropkick QC report for a human T cell dataset encapsulated using the 10x Genomics Chromium platform (Zheng et al., 2017). This sample is exemplary of a low-background dataset, as the cells isolated from human blood do not require dissociation that causes cell stress and lysis in other tissues (Figure S6). Barcodes are ranked by total counts to yield a profile that describes the expected number of high-quality cells, empty droplets, and uninformative barcodes (Figure 3.1A; Fleming et al., 2019). The number of genes detected per barcode follows a similar distribution to total counts, which informs our choice of dropkick training thresholds in the following sections. The first plateau in the total counts profile of the T cell dataset indicates approximately 4,000 high-quality cells, followed by a sharp drop in the distribution (Figure 3.1A).

This drop-off in total UMI content signifies an estimated location for a manual cutoff as seen in the 10x CellRanger version 2 analysis software (Lun et al., 2019).

dropkick next defines a subset of ambient genes using the dropout rate, or the fraction of barcodes in which each gene is not detected. Ranking genes in ascending order by dropout rate (Figure 3.1B), dropkick labels those with dropout rates lower than the top ten as “ambient”. High-background datasets may have many (¿ 10) genes that are detected in nearly every barcode (dropout rate 0; Figure S6). The dropkick definition of an ambient profile thus ensures that all relevant genes are included. The contribution of this ambient subset to the total counts of each barcode can then be calculated, shown as blue points in the dropkick QC report (Figure 3.1A). Similarly, an overlay of mitochondrial read percentage indicates dead or dying cells undergoing apoptosis (Tait and Green, 2010). Indeed, the ambient and mitochondrial contributions to the empty droplets in the second plateau of the total counts log-rank curve is markedly higher than the first plateau (Figure 3.1A). Another noteworthy observation is that dropkick defines an ambient profile that is distinct from the subset of mitochondrial genes. This is important for assessing cell quality in downstream clustering and dimension reduction, as any empty droplets that remain in the dataset post-filtering often cluster together in low-dimensional embeddings and can be highlighted by their enrichment in ambient genes. As stated previously, marker genes from abundant cell types may show up in the ambient gene set due to excessive lysis of these common cells during tissue preparation (Young and Behjati, 2018; Fleming et al., 2019; Yang et al., 2020; Figure S6). Accordingly, analysts should be cognizant of background expression levels that contaminate adjacent cell populations and confound cell type identification during subsequent analysis.

As each scRNA-seq dataset has unique, batch-specific ambient RNA profiles and barcode distributions, the dropkick QC module allows for estimation of global data quality. Mouse colonic mucosa dissociated and encapsulated in parallel using inDrop and 10x Genomics platforms (Figure S6) exemplifies high-background scRNA-seq data, as indicated by elevated RNA levels in the second plateau of the total counts and genes curves. Moreover, marker genes Car1 and Muc2 from abundant colonocytes and goblet cells, respectively, are identified by dropkick as ambient genes for these data. This signifies lysis of common epithelial cell populations during tissue preparation and dissociation. Given the dropkick QC report, the user should thus expect background expression across all barcodes, which could prove pivotal to downstream processing and biological interpretation. Taken together, dropkick can estimate the number of high-quality cells in our dataset, determine average background noise from ambient RNA, and thus predict performance of filtering and ensuing analysis based on global data quality.

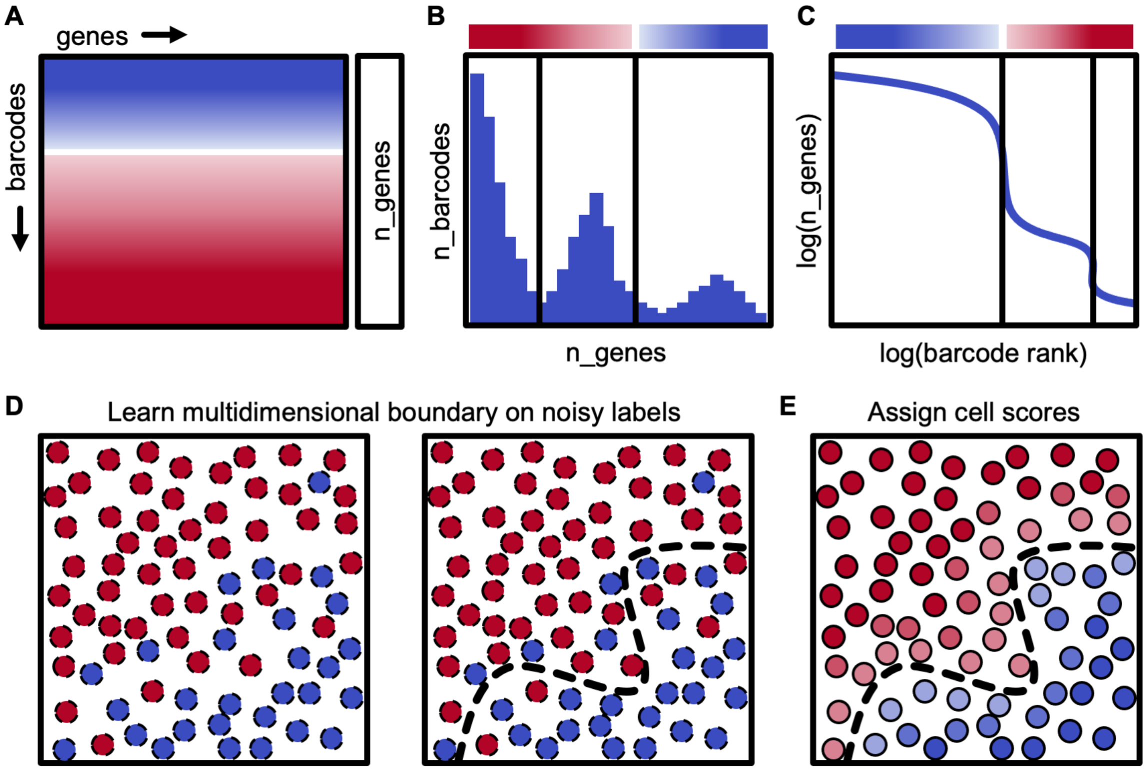


Figure 3.2: Description of dropkick filtering method.
(A) Diagram of scRNA-seq counts matrix with initial cell confidence for each barcode based solely on total genes detected (n-genes), depicted by color (red = empty droplet, blue = real cell).
(B) Histogram showing the distribution of barcodes by their n-genes value. Black lines indicate automated thresholds for training the dropkick model.
(C) log(n-genes) vs. log(rank) representation of barcode distribution as in dropkick QC report (Figure 1A). Thresholds from B are superimposed.
(D) Thresholds in heuristic space (B-C) are used to define initial training labels for logistic regression.
(E) dropkick chooses an optimal regularization strength through cross-validation, then assigns cell probabilities and labels to all barcodes using the trained model in gene space.

3.3.2 Description of dropkick filtering method

dropkick uses weakly supervised machine learning to build a model of single-cell gene expression in order to score and classify barcodes as real cells or empty droplets within individual scRNA-seq datasets. To construct a training set for this model, dropkick begins by calculating batch-specific global metrics that are generally predictive of barcode quality, such as the total number of genes detected (n-genes; Figure 3.2A) which was chosen as the default training heuristic for dropkick by testing concordance with three alternative cell labels across 46 scRNA-seq samples (Figure S7). A dataset similar to the 10x Genomics human T cell encapsulation (Figure 3.1) will exhibit a multimodal distribution of n-genes across all barcodes (Figure 3.2B) where the peaks of the distribution match the plateaus seen in the log-rank representation (Figure 3.2C). Next, dropkick performs multi-level thresholding on the n-genes histogram using Otsu's method (Otsu, 1979; Figure 3.2B,C).

This automated gradient-descent technique divides the barcode distribution into three levels in this “heuristic space”: a lower level containing uninformative barcodes (which are thrown away), an upper level containing barcodes with very high cell probability based on n-genes, and an intermediate level that consists of both high-RNA empty droplets and relatively low-RNA cells. The upper and intermediate barcode populations are labeled as real cells and putative empty droplets, respectively, for initial dropkick model training. These weakly self-supervised labels based on threshold cutoffs in “heuristic space” are expected to be noisy, and the goal of the next step in the dropkick pipeline is to re-draw these rough boundaries in “gene space” using logistic regression in order to recover real cells from the intermediate barcode cohort while removing ambient barcodes from the upper plateau (Figure 3.2D,E).

The logistic regression model employed by dropkick uses elastic net regularization (Zou and Hastie, 2005), which balances feature selection and grouping by preserving or removing correlated genes from the model in concert. The motivation for choosing this regularization method is two-fold. First, the resulting model exists in “gene space”, maintaining the relative dimensionality of the dataset and providing biologically interpretable coefficients that describe barcode quality. Second, the model is penalized for complexity, which yields the simplest model (sparse coefficients) that adjusts the noisy initial labels while compensating for expected collinearities and errors in measurement.

3.3.3 Evaluating dropkick filtering performance with synthetic data

We tested dropkick filtering on single-cell data simulations that define both empty droplets and real cells, providing ground-truth labels for comparison to dropkick outputs (Fleming et al., 2019). These synthetic datasets modeled ambient RNA noise in the cell populations to confound filtering, as seen in real-world datasets. We simulated both low (Figure 3.3A,B) and high (Figure 3.3C,D) background scenarios (see Appendix A3.1: Synthetic scRNA-seq data simulation).

To demonstrate the utility of the dropkick model over one-dimensional thresholding and an analogous data-driven filtering model, we ran dropkick, 10x Genomics CellRanger version 2 (CellRanger-2) and the EmptyDrops R package (Lun et al., 2019) on ten iterations of low and high-background simulations. An example UMAP embedding of all barcodes kept by dropkick-label (dropkick score ¿= 0.5) and the two analogous methods shows that all three methods excluded empty droplets (assigned cluster 0 from the simulation), with a single false negative (FN) barcode highlighted in the EmptyDrops label set (Figure 3.3A). An UpSet plot (Figure 3.3B; Lex et al., 2014) tabulating shared barcode sets across ten low-background simulations reveals nearly perfect specificity, sensitivity, and area under the receiver operating characteristic curve (AU-ROC) for all three methods in the low-background scenario (Figure S8A,B,D).

Conversely, the high-background simulations produced a large number of false positives (FP) in the

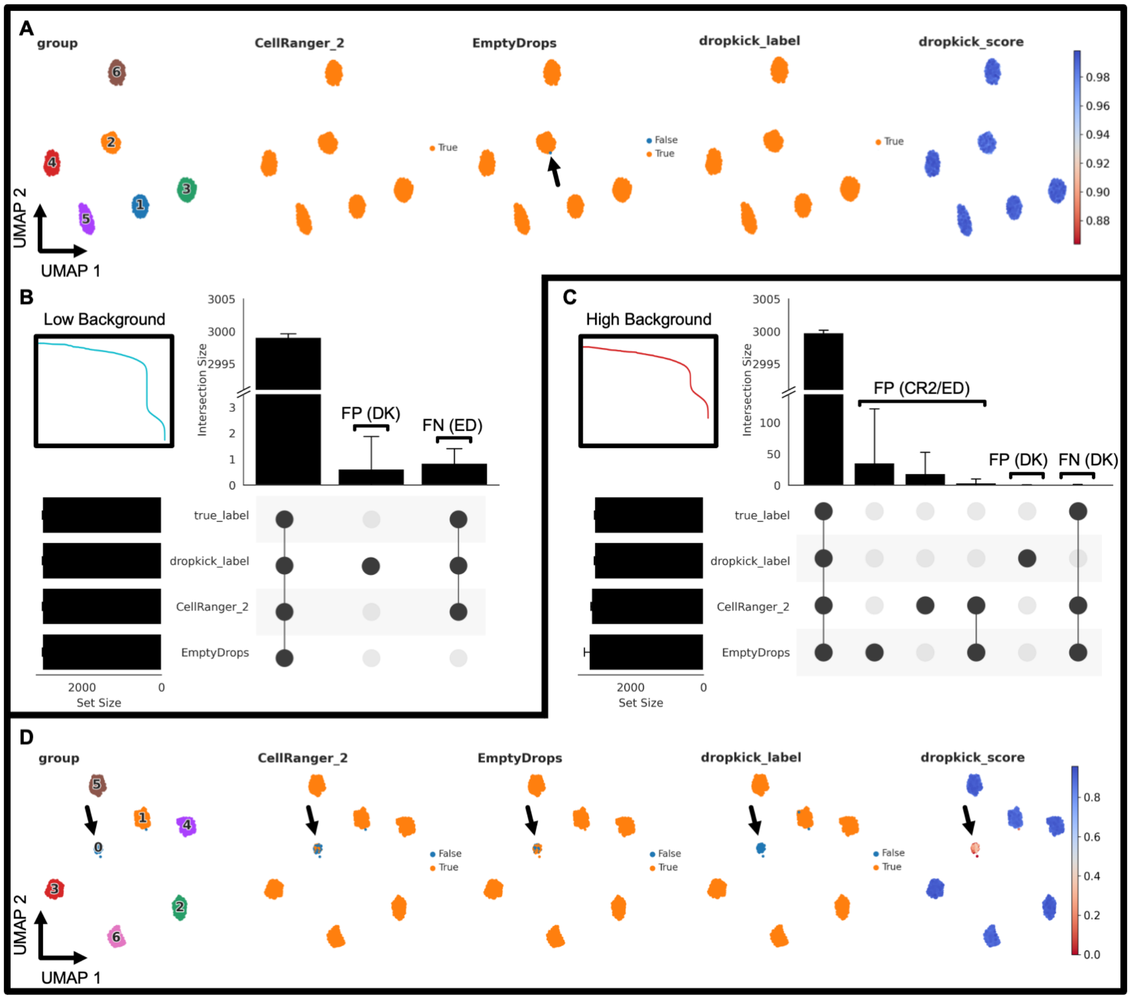


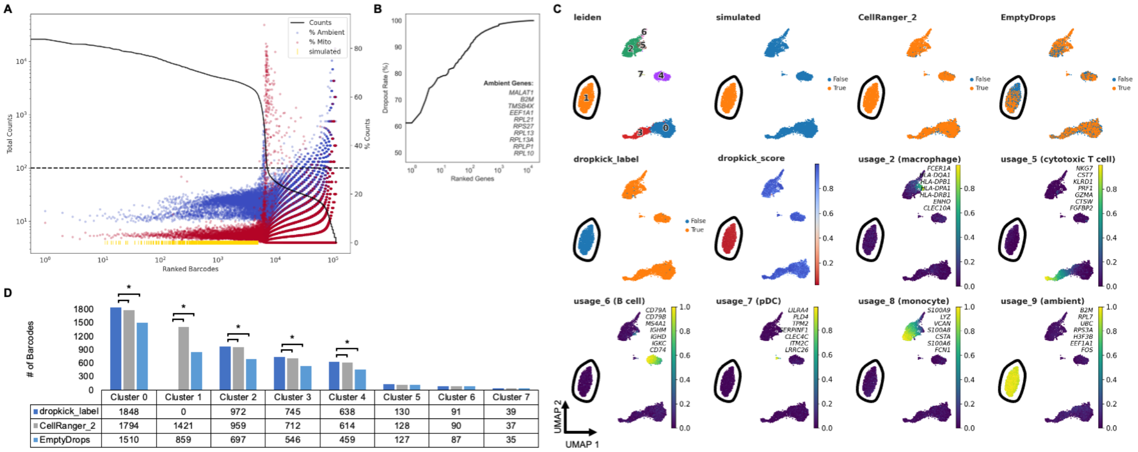
Figure 3.3: Evaluating dropkick filtering performance with synthetic data.
(A) UMAP embedding of all barcodes kept by dropkick-label, CellRanger-2 and EmptyDrops for an example low-background simulation. Points colored by each of the three filtering labels, as well as ground-truth clusters determined by the simulation and dropkick score (cell probability). Arrow highlights a single false negative (FN) barcode in the EmptyDrops label set for this replicate.
(B) UpSet plot showing mean size of shared barcode sets across dropkick-label, CellRanger-2, EmptyDrops, and true labels for ten simulations. Error bars represent standard deviation. Unique sets show false positive (FP) barcodes labeled by dropkick and false negative (FN) barcodes excluded by EmptyDrops. Inset shows log-rank representation of the low-background simulation in A.
(C) Same as in B, for ten high-background simulations. Inset shows log-rank representation of the high-background simulation in D.
(D) Same as in A, for an example high-background simulation. Arrow highlights cluster 0, designated as “empty droplets” by simulation (see Appendix A3.1: Synthetic scRNA-seq data simulation).

CellRanger-2 and EmptyDrops labels (Figure 3.3C), as ambient barcodes with high RNA content lie above the total counts threshold identified by CellRanger and the inflection point used as a testing cutoff by EmptyDrops (Lun et al., 2019). A UMAP embedding of an example high-background simulation reveals a large population of empty droplets (assigned cluster 0 by the simulation) that dropkick-label removes from the final dataset (Figure 3.3D). Accordingly, dropkick displayed overall specificity and AUROC of 0.9999 +/- 0.0002 and 0.9998 +/- 0.0002 for the high-background simulations compared to 0.9910 +/- 0.0018 and 0.9955 +/- 0.0009 for CellRanger-2 and 0.9838 +/- 0.0133 and 0.9917 +/-0.0071 for EmptyDrops, respectively (Figure S8E,F,H).

We also compared outputs from the trained model (dropkick-label) to automated dropkick training labels (thresholding on n-genes) in both low- and high-background scenarios to further demonstrate the utility of dropkick's machine learning model over heuristic cutoffs alone. Similar to CellRanger-2, the dropkick threshold performed favorably for the low background simulation, where real cells are separated distinctly from empty droplets in heuristic space – indicated by a sharp drop-off in total counts and genes in the dropkick QC log-rank plot (Figure 3.3B, inset). This one-dimensional thresholding resulted in sensitivity, specificity, and AUROC of 0.9986 +/- 0.0007, 0.997 +/- 0.0006, and 0.9978 +/- 0.0005, respectively for ten low-background simulations (Figure S8C). The trained dropkick model, on the other hand, recovered all real cells (sensitivity 1.0), with a perfect average AUROC of 1.0 +/- 0.0 (Figure S8D). This modest improvement indicates the utility of the dropkick model for sensitively discerning real cells from ambient barcodes over simple heuristic thresholding, even in a relatively low-background sample. In the high-background simulations, sensitivity of dropkick training labels fell to 0.8762 +/- 0.0092 with an average AUROC of 0.9074 +/- 0.0043 (Figure S8G). Following model training, dropkick's sensitivity and AUROC once again improved to 0.9995 +/- 0.0004 and 0.9998 +/- 0.0002, respectively (Figure S8H). These data further signify that the dropkick logistic regression model results in enhanced performance over one-dimensional heuristic thresholding, especially in the presence of high ambient noise in the training set.

3.3.4 Benchmarking dropkick performance on simulated high-background data

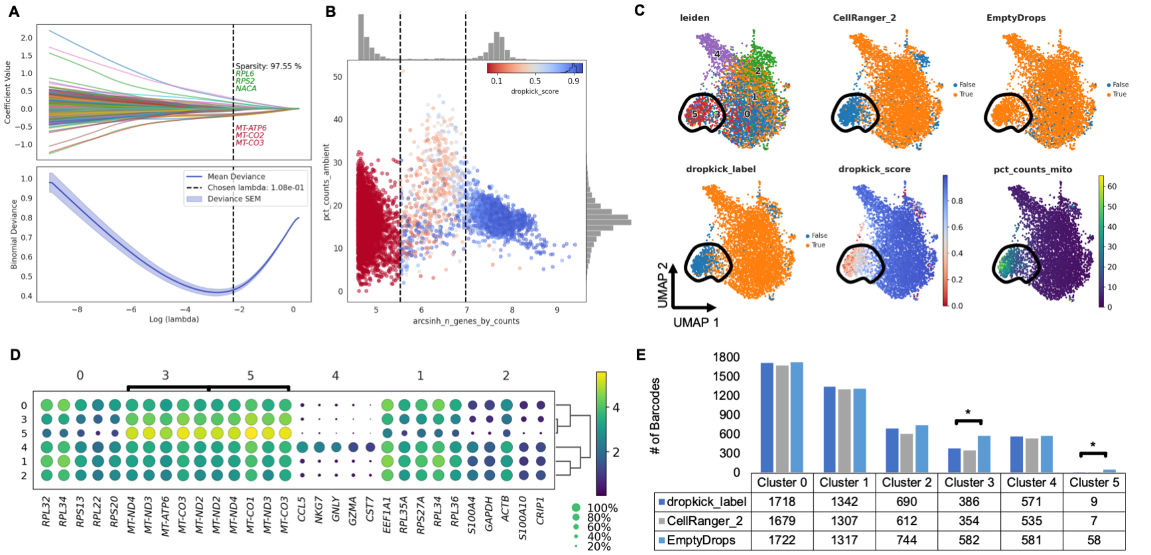
Next, we aimed to further confirm dropkick's utility in filtering high-background data by simulating extremely high-ambient droplets to overlay on the 10x Genomics human PBMC dataset. This data is particularly clean and easy to filter in its raw state, as the suspended cells from human blood were minimally agitated prior to encapsulation. In order to imitate empty droplets with high mRNA content, we combined all reads in barcodes with less than 100 total UMI counts and used the resulting pseudo-bulk as weightings for a random generation of count vectors from a multinomial distribution with UMI sums between 10 and 5,000 total counts. We added 2,000 of these count vectors back to the original matrix, modeling high-background empty droplets (Figure



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
■ dropkick_label	1848	0	972	745	638	130	91	39
■ CellRanger_2	1794	1421	959	712	614	128	90	37
■ EmptyDrops	1510	859	697	546	459	127	87	35

Figure 3.4: Benchmarking dropkick performance on simulated high-background data.
(A) Log-rank total counts curve for the high-background PBMC simulation. The horizontal dashed line indicates the threshold below which ground-truth empty droplets were used to build simulated barcodes from a multinomial distribution (100 total counts). Gold rug plot indicates the location along the total counts curve of 2,000 simulated high-UMI droplets (see Appendix A3.1: High-background PBMC simulation).
(B) Genes in PBMC simulation ranked by dropout rate. Top 10 ambient genes are listed, defining ambient profile used to calculate percentage in A.
(C) UMAP embedding of all barcodes kept by dropkick-label, CellRanger-2 and EmptyDrops. Points colored by each of the three filtering labels, Leiden clusters determined by NMF analysis, dropkick score (cell probability), and select cell-type metagene usages from NMF. Top seven gene loadings for each NMF factor are printed on their respective plots, in axis order from top to bottom. Circled area shows independent cluster of simulated empty droplets.
(D) Table and bar graph enumerating the total number of barcodes detected by each algorithm in all NMF clusters. Significant cluster enrichment as determined by sc-UniFrac is denoted by brackets.

3.4A). Upon filtering with dropkick, CellRanger version 2, and EmptyDrops, a large subset of the simulated ambient barcodes remained in the latter two label sets, while discarded entirely by dropkick (Figure 3.4C,D). We jointly processed all barcodes kept by the three filtering tools using nonnegative matrix factorization (NMF; Kotliar et al., 2019) to define cell clusters and corresponding cell type metagene scores (Figure 3.4C; Figure S9). dropkick recovered significantly more lymphoid progenitors, monocytes, and T and B cells than both EmptyDrops and CellRanger according to sc-UniFrac (Liu et al., 2018) analysis, indicating that it successfully parsed the noise introduced by the simulated droplets (Figure 3.4D). dropkick also completely excluded Leiden cluster 1, the simulated barcodes with high NMF scores for usage 9, which contained high loadings for several ambient genes (Figure 3.4B,C; Figure S9B). This result both confirmed the effectiveness of the pseudo-bulk multinomial simulation, and further established dropkick's robustness in filtering high-background data.



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
■ dropkick_label	1718	1342	690	386	571	9
■ CellRanger_2	1679	1307	612	354	535	7
■ EmptyDrops	1722	1317	744	582	581	58

Figure 3.5: dropkick recovers expected cell populations and eliminates low-quality barcodes in experimental data.
(A) Plot of coefficient values for 2,000 highly variable genes (top) and mean binomial deviance and SEM (bottom) for five-fold cross-validation along the lambda regularization path defined by dropkick. Top and bottom three coefficients are shown, in axis order, along with total model sparsity representing the percentage of coefficients with values of zero (top). Chosen lambda value indicated by dashed vertical line.
(B) Joint plot showing scatter of percent ambient counts versus arcsinh-transformed genes detected per barcode, with histogram distributions plotted on margins. Initial dropkick thresholds defining the training set are shown as dashed vertical lines. Each point (barcode) is colored by its final dropkick score after model fitting.
(C) UMAP embedding of all barcodes kept by dropkick-label, CellRanger-2 and EmptyDrops. Points colored by each of the three filtering labels, as well as Leiden clusters determined by NMF analysis, dropkick score (cell probability), and percent counts mitochondrial. Circled area shows high mitochondrial enrichment in a population discarded by dropkick.
(D) Dot plot showing top differentially expressed genes for each NMF cluster. The size of each dot indicates the percentage of cells in the population with nonzero expression for the given gene, while the color indicates the average normalized expression value in that population. Bracketed genes indicate significantly enriched populations in EmptyDrops compared to dropkick-label as shown in E.
(E) Table and bar graph enumerating the total number of barcodes detected by each algorithm in all NMF clusters. Significant cluster enrichment as determined by sc-UniFrac is denoted by brackets.

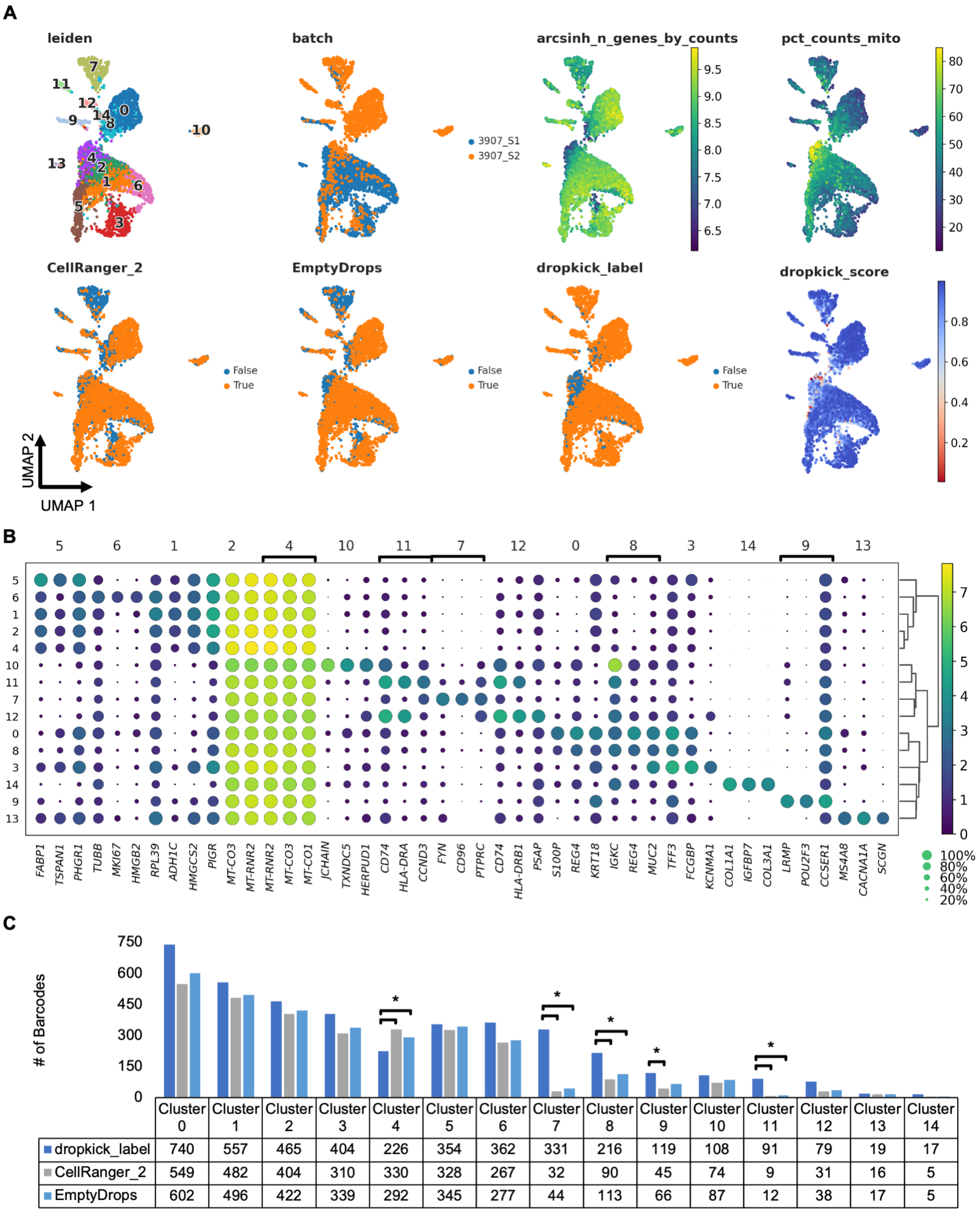
3.3.5 dropkick recovers expected cell populations and eliminates low-quality barcodes in experimental data

To evaluate dropkick's performance against existing scRNA-seq filtering algorithms with real-world data, we processed a human T cell dataset from 10x Genomics (Figure 3.1) and again compared default dropkick results (dropkick-label) to CellRanger version 2 and EmptyDrops. The final dropkick coefficients and chosen regularization strength (lambda; Figure 3.5A) reveal that the model is sparse – with nearly 98 percent of all coefficient values equal to zero – offering an interpretable gene-based output. Without prior training or supervision, dropkick identified higher counts of mitochondrial genes, which are markers of cell death and poor barcode quality (Tait and Green, 2010), as predictive of empty droplets (Figure 3.5A). To visualize heuristic distributions within the T cell dataset, the number of detected genes and the percentage of ambient counts per barcode are shown along with dropkick's automatic training thresholds (Figure 3.5B). Uninformative barcodes below the lower n-genes threshold were discarded before model training and assigned a dropkick score of zero. Barcodes between the two thresholds were initially assigned a label indicating putative empty droplets, while those above the upper threshold were labeled as real cells for model training. The dropkick score overlay illustrates how dropkick re-drew label boundaries in gene space (Figure 3.5B). dropkick scores are noticeably lower for barcodes with high ambient RNA content, while some putative empty droplets with lower background are "rescued" and labeled as real cells by the trained dropkick model. It is important to note that this high-dimensional boundary was learned by dropkick with no prior labeling of "ambient" transcripts. Rather, dropkick's weakly-supervised algorithm excluded barcodes with high ambient content based solely on their transcriptional similarity to the least informative barcodes (lower n-genes) in the training set.

We again jointly processed all barcodes kept by dropkick-label (dropkick score ¿= 0.5), CellRanger-2, and EmptyDrops using nonnegative matrix factorization (NMF; Kotliar et al., 2019) to define cell clusters, and sc-UniFrac (Liu et al., 2018) to determine population differences across labeled barcode sets. A UMAP embedding of these barcodes reveals a population of cells with high mitochondrial content that is mostly excluded by dropkick (Figure 3.5C). This area is enriched in clusters 3 and 5 from NMF analysis, which carry exclusively mitochondrial genes as their top differentially expressed features (Figure 3.5D). Based on sc-UniFrac, these two clusters constitute the only statistically significant differences between EmptyDrops and dropkick (Figure 3.5E). These data indicate that dropkick recovers as many or more real cells in expected populations than previous algorithms, while also identifying and excluding low-quality dead or dying cells with high mitochondrial RNA content.

3.3.6 dropkick outperforms analogous methods on challenging datasets

To challenge the robustness of the model, we next used dropkick to filter real-world samples with more complex cell types and higher noise. Human colorectal carcinoma (3907-S2) and adjacent normal colonic mucosa (3907-S1) samples were dissociated and encapsulated using the inDrop scRNA-seq platform (Klein et al., 2015). In contrast to the 10x Genomics pan-T cell dataset (Figure 3.1; Figure 3.5), these samples exhibited high levels of background, containing empty droplets with thousands of UMI counts detected per barcode and up to 40 percent ambient RNA in expected cell barcodes (Figure S11A,D). Because of this dominant ambient profile, infiltrating immune populations with lower mRNA content than epithelial cells can be lost among empty droplets. Indeed, CellRanger-2 and EmptyDrops show depletion in T cells (cluster 7) and macrophages (cluster 11) compared to dropkick (Figure 3.6A,B). Prevalence of high-RNA empty droplets also yields a population with low genetic diversity and mitochondrial gene enrichment (cluster 4;



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12	Cluster 13	Cluster 14
■ dropkick_label	740	557	465	404	226	354	362	331	216	119	108	91	79	19	17
■ CellRanger_2	549	482	404	310	330	328	267	32	90	45	74	9	31	16	5
■ EmptyDrops	602	496	422	339	292	345	277	44	113	66	87	12	38	17	5

Figure 3.6A) that is kept by the one-dimensional thresholding of CellRanger-2 but discarded by dropkick. sc-UniFrac analysis confirmed that dropkick recovers significantly more cells from rare populations than both CellRanger-2 and EmptyDrops in this pair of high-background datasets dominated by ambient RNA from dead and dying colonic epithelial cells (Figure 3.6C; Figure S11). Meanwhile, dropkick also identified and removed significantly more dead cells (cluster 4) than both CellRanger-2 and EmptyDrops (Figure 3.6C) by designating mitochondrial and ambient genes as negative coefficients (Figure S11B,E).

3.3.7 dropkick filters reproducibly across scRNA-seq batches

We also applied dropkick to a combined human placenta dataset from six patients to show robustness of the model to batch-specific variation. dropkick learned the distribution of genes and ambient RNA specific to each dataset and filtered them accordingly (Figure S12A), with a resulting AUROC of 0.9956 +/- 0.0051 across all six replicates compared to EmptyDrops labels. We also performed two types of manual cell labeling as well as the CellBender remove-background model (Fleming, Marioni and Babadi 2019) to provide additional alternative filtering labels to compare to dropkick (Figure S12B,D,E,F,G,H,J; see Appendix A3.1: CellRanger 2, EmptyDrops, CellBender, and manual filtering of real-world scRNA-seq datasets). The CellBender remove-background package primarily aims to subtract ambient background from single-cell expression datasets rather than filter alone. This resulted in the addition of a large population of high-ambient barcodes unique from those labeled by dropkick, EmptyDrops, and CellRanger 2, warranting further assessment of the efficacy of background-removal methods in the context of consensus cell labels beyond the scope of this paper (Figure S12B-E).

Extending this analysis to a larger cohort of scRNA-seq samples from both 10x Genomics (n = 13) and inDrop (n = 33) encapsulation platforms, we see that dropkick is highly concordant with CellRanger version 2 (AUROC 0.9656 +/- 0.0271) and EmptyDrops (AUROC 0.9817 +/- 0.012), suggesting global recovery

Figure 3.6 *(preceding page)*: dropkick outperforms analogous methods on challenging datasets.
(A) UMAP embedding of all barcodes kept by dropkick-label (dropkick score ¿= 0.5), CellRanger-2 and EmptyDrops for human colorectal carcinoma inDrop samples. Points colored by each of the three filtering labels, as well as clusters determined by NMF analysis, dropkick score (cell probability), arcsinh-transformed total genes detected, percent counts mitochondrial, and original batch. 3907-S1 is normal human colonic mucosa and 3907-S2 is colorectal carcinoma from the same patient.
(B) Dot plot showing top differentially expressed genes for each NMF cluster. The size of each dot indicates the percentage of cells in the population with nonzero expression for the given gene, while the color indicates the average expression value in that population. Bracketed genes indicate significantly enriched or depleted populations in dropkick compared to CellRanger-2 and/or EmptyDrops labels as shown in C.
(C) Table and bar graph enumerating the total number of barcodes detected by each algorithm in all NMF clusters for the combined dataset. Significant cluster enrichment as determined by sc-UniFrac is denoted by brackets.

of major cell populations (Figure S13A,B,E,F). dropkick filtering for 33 inDrop samples yielded an AUROC of 0.9729 +/- 0.0335 compared to manually curated labels using an inflection point cutoff followed by dimension-reduced cluster gating (Chen et al., 2021b; Figure S13C). For all 46 scRNA-seq samples, we also performed bivariate thresholding on total UMI counts and percent mitochondrial transcripts per droplet, mimicking another popular preprocessing technique. Again, dropkick's AUROC averaged 0.9805 +/- 0.0194, confirming the model's utility for robust filtering across several unique datasets (Figure S13D,H). Finally, we measured the total run time of dropkick, which was appreciably faster than both CellBender remove-background and the EmptyDrops R package on average, running to completion in 40.56 +/- 25.97 seconds across ten replicates of all 46 samples when utilizing five CPUs with dropkick's built-in parallelization (Figure S13J).

3.4 Discussion

Barcode filtering is a key preprocessing step in analyzing droplet-based single-cell expression data. Reliable filtering is confounded by distributions of global heuristics such as total UMI counts, total genes, and ambient RNA that can be highly variable across batches and encapsulation platforms. We have developed dropkick, a fully automated machine learning software tool that assigns confidence scores and labels to barcodes from unfiltered scRNA-seq counts matrices. By automatically curating a training set using predictive heuristics and training a gene-based logistic regression model, dropkick ensures that ambient barcodes ("empty droplets") are removed from the filtered dataset while recovering rare, low-RNA cell types that may be lost in ambient noise. We showed that unlike previous filtering approaches including one-dimensional thresholding (CellRanger 2) and a Dirichlet-multinomial model (EmptyDrops), dropkick is robust to the level of ambient RNA, performing favorably in both low and high-background scenarios across simulated and real-world datasets.

Although we have demonstrated that dropkick is more robust to varying degrees of ambient background than existing filtering methods, the dropkick model is still limited by the input dataset. As stated previously (see results: Evaluating dataset quality with dropkick QC module), the profile of ranked total counts/genes and the global contribution of ambient reads are vital to analysis of single-cell sequencing data, including cell filtering. Data with weak separation between high-quality cells and empty droplets (i.e. a unimodal distribution of n-genes lacking distinct plateaus in the log-rank curve) will perform poorly in inflection-point thresholding as well as data-driven models such as EmptyDrops and dropkick due to the similarity between theoretically "high-confidence" barcodes and ambient background droplets. Moreover, datasets dominated by expression of ambient genes (40 percent average ambient counts across all barcodes) will also perform poorly in automated filtering. While such data artifacts may be handled by dropkick's heavy feature selection conferred by HVG calculation and elastic net regularization, there will also be circumstances that cause dropkick – as well as CellRanger and EmptyDrops – to return an over- or under-filtered dataset. Scenarios such as those described should be considered QC failures, and further analysis should not be performed. For this reason, the dropkick QC module is extremely beneficial in post-alignment evaluation of scRNA-seq data quality and should be applied to all datasets prior to filtering.

The dropkick Python package provides a fast, user-friendly interface that integrates seamlessly with the `scanpy` (Wolf et al., 2018) single-cell analysis suite for ease of workflow implementation. dropkick is available for installation through the Python Package Index (pypi.org/project/dropkick/), and source code is hosted on GitHub (github.com/KenLauLab/dropkick).

CHAPTER 4

Consensus tissue domain detection in spatial multi-omics data using MILWRM

Adapted from:

Kaur, H., Heiser, C. N., McKinley, E. T., Antunes, L. V., Harris, C. R., Roland, J. T., Shrubsole, M. J., Coffey, R. J., Lau, K., and Vandekar, S. N. (2023). Consensus tissue domain detection in spatial -omics data using MILWRM. *bioRxiv*, page 2023.02.02.526900

4.1 Summary

Spatially resolved molecular assays provide high dimensional genetic, transcriptomic, proteomic, and epigenetic information in situ and at various resolutions. Pairing these data across modalities with histological features enables powerful studies of tissue pathology in the context of an intact microenvironment and tissue structure. Increasing dimensions across molecular analytes and samples require new data science approaches to functionally annotate spatially resolved molecular data. A specific challenge is data-driven cross-sample domain detection that allows for analysis within and between consensus tissue compartments across high volumes of multiplex datasets stemming from tissue atlasing efforts. Here, we present MILWRM – multiplex image labeling with regional morphology – a Python package for rapid, multi-scale tissue domain detection and annotation. We demonstrate MILWRM's utility in identifying histologically distinct compartments in human colonic polyps and mouse brain slices through spatially-informed clustering in two different spatial data modalities. Additionally, we used tissue domains detected in human colonic polyps to elucidate molecular distinction between polyp subtypes. We also explored the ability of MILWRM to identify anatomical regions of mouse brain and their respective distinct molecular profiles.

4.2 Introduction

The advent of spatially resolved molecular assays has enabled access to high dimensional genetic, transcriptomic, proteomic, and even epigenetic information in situ while preserving the spatial information lost in single-cell or bulk molecular assays (Black et al., 2021; Gerdes et al., 2013; Moses and Pachter, 2022; Schapiro et al., 2022). Spatially resolved data can provide powerful insight into interactions between cell types, progressive changes in tissue architecture in diseases such as cancer, or interactions between different structures in tissue such as lymphoid follicles and blood vessels (Ruddle, 2016; Sipos and Muzes, 2011; Hickey et al., 2021). Biological insights can be derived from recurring spatial patterns extracted using quantitative analysis on spatial data.

Many current methods attempt to complement single-cell analyses, essentially taking a bottom-up approach to reconstruct tissue domains, architectures, and communities from individual cells. In general, individual cells can be identified from high dimensional imaging data by segmentation. Cellular segmentation and annotation are the most challenging step in this kind of approach. There are various methods available for cellular segmentation (McKinley et al., 2022; Greenwald et al., 2021), annotation (Liu et al., 2022) and neighborhood analysis (Warchol et al., 2022; Kim et al., 2022; Chen et al., 2020). Widely used lower resolution imaging data such as spatial transcriptomics (ST) and imaging mass spectrometry data are analyzed using cellular deconvolution algorithms to approximate single-cell composition. Most of these algorithms require a parallel single-cell dataset for use as reference (Cable et al., 2020; Andersson et al., 2020). Different cell types are then arranged into interaction networks based on their spatial distributions and/or molecular interactions, and these networks are assembled into larger spatial structures that identify tissue- or organ-level domains. This type of analysis has been used for identifying cellular communities in various cancer types associated with patient prognosis (Keren et al., 2018; Schürch et al., 2020; Andersson et al., 2021; Jackson et al., 2020).

Another perspective comes from the pathology field, where spatial domains and architectures are first identified, followed by instances of cell identification by morphology, which is known as the top-down approach (Lin et al., 2023). Since this approach focuses directly on pixel-level information instead of reconstruction from single-cell data, it can identify both extracellular structures and cellular communities over a range of micro- to macro-scale. Pixel-based analysis also forms the basis of modern artificial intelligence learning from imaging, and thus paves the way for more complex learning algorithms to be applied to multiplex tissue data (Piccinini et al., 2017; Amitay et al., 2022; Wu et al., 2022b).

Various methods are currently available for pixel-based spatial domain detection from ST data (Alexandrov and Kobarg, 2011; Zhao et al., 2021; Townes and Engelhardt, 2021). However, they lack the scalability to work across batches and samples. Attempts to apply these methods across samples fail to yield global consensus domains, and instead identify regional domains that are sample-specific or confounded by batch effects. To decipher true emergent properties within spatial tissue domains, it is imperative that findings can be generalized across many samples at different resolutions. Here, we present multiplex image labeling with regional morphology (MILWRM) that is designed specifically for consensus tissue domain characterization across large sample sets with potential differing orientations and resolutions.

4.3 Results

4.3.1 The MILWRM pipeline generates consensus tissue domains across specimens

Whereas most spatial analysis algorithms focus on individual specimens, MILWRM aims to identify consensus tissue domains across samples with spatially resolved molecular data (e.g., multiplexed immunoflu-

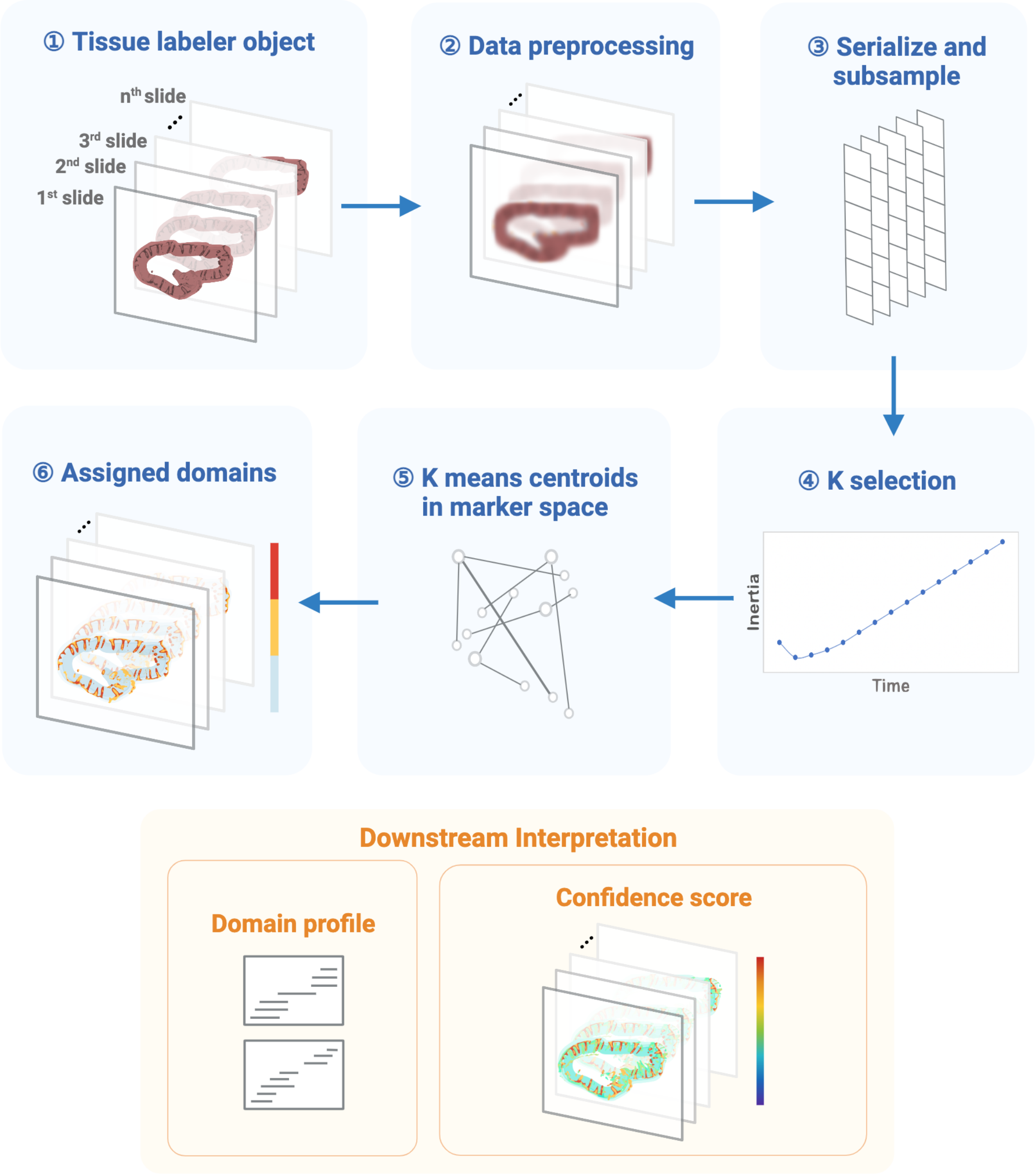


Figure 4.1: The workflow of MILWRM pipeline.
(A) MILWRM begins with constructing a tissue labeler object from all the sample slides that undergoes data preprocessing, serialization and subsampling to create a randomly subsampled dataset used for Kmeans model construction. This subsampled data is used to find optimal number of tissue domains – K selection using adjusted inertia method. Finally, a Kmeans model is constructed, and each pixel is assigned a tissue domain. Each tissue domain has a distinct domain profile describing the molecular feature. MILWRM also provides quality control metrics such as confidence score.

orescence [mIF] and ST). The MILWRM pipeline can be broadly categorized into three major steps: data preprocessing, tissue domain identification, and tissue domain analysis (Figure 4.1). To generalize pixel neighborhood information across batches, data preprocessing incorporates down-sampling, normalization, data smoothing, and dimensionality reduction. Preprocessing steps differ slightly for mIF and ST (Appendix A4.1). After preprocessing, tissue domains are identified using unsupervised K-means clustering by subsampling data across all samples (Appendix A4.1). The number of tissue domains is adjusted by inertia analysis (Clarke and Greenacre, 1985). Each pixel is assigned a tissue domain based on the nearest centroid. Domain profiles are calculated by MILWRM from the initial feature space to molecularly describe each tissue domain, which is useful for downstream annotation. Finally, MILWRM computes a variety of metrics to assess the quality of identified tissue domains (Appendix A4.1). Overall, MILWRM is a comprehensive, easy to use pipeline for tissue domain detection, providing interpretable results for biological analysis and quality assessment.

4.3.2 MILWRM identifies canonical tissue layers of the colonic mucosa

We applied MILWRM to mIF data generated for the Human Tumor Atlas Network (HTAN) consisting of human normal colon and different colonic pre-cancer subtypes (conventional adenomas – AD and serrated polyps – SER) (Chen et al., 2021b). These data comprised multichannel fluorescent images from 37 biospecimens consisting of tissues with different morphologies and pathological classification confirmed by two pathologists (Table S1). We performed low resolution application of MILWRM using a smoothing parameter (σ) of 2 after downsampling the images to an isotropic resolution of 5.6 µm/pixel and the penalty parameter of 0.05 that resulted in three tissue domains according to adjusted inertia, as illustrated by three representative samples (Figure 4.2A-B; Figure S15A). According to domain profiles (Figure 4.2C), the epithelial monolayer compartment was identified by markers such as CDX2, β-catenin, Na+-K+ ATPase, and proliferative marker PCNA, consistent with a high turnover hind-gut epithelium (McKinley et al., 2017; Herring et al., 2018). The mucus layer was enriched in MUC2, a secreted mucin (Tytgat et al., 1994; Allen et al., 1998; Karlsson et al., 1996). The lamina propria region, where stromal cells are prominent, was identified by Vimentin and Collagen (Vega et al., 2022). The results from MILWRM analysis are consistent with the tissue architecture of the colonic mucosa, as well as other mucosal tissues in the body.

MILWRM consistently identified these regions across the 37 tissue samples (Figure 4.2D), and pixel level data over the samples intermixed in UMAP-embedded space illustrating removal of batch effects between images (Figure 4.2E), demonstrating the ability of MILWRM to identify consensus regions over multiple samples. MILWRM was able to capture about 80% of the variance in the multidimensional imaging data without any notable outliers, demonstrating that information within the imaging data is retained after MIL-

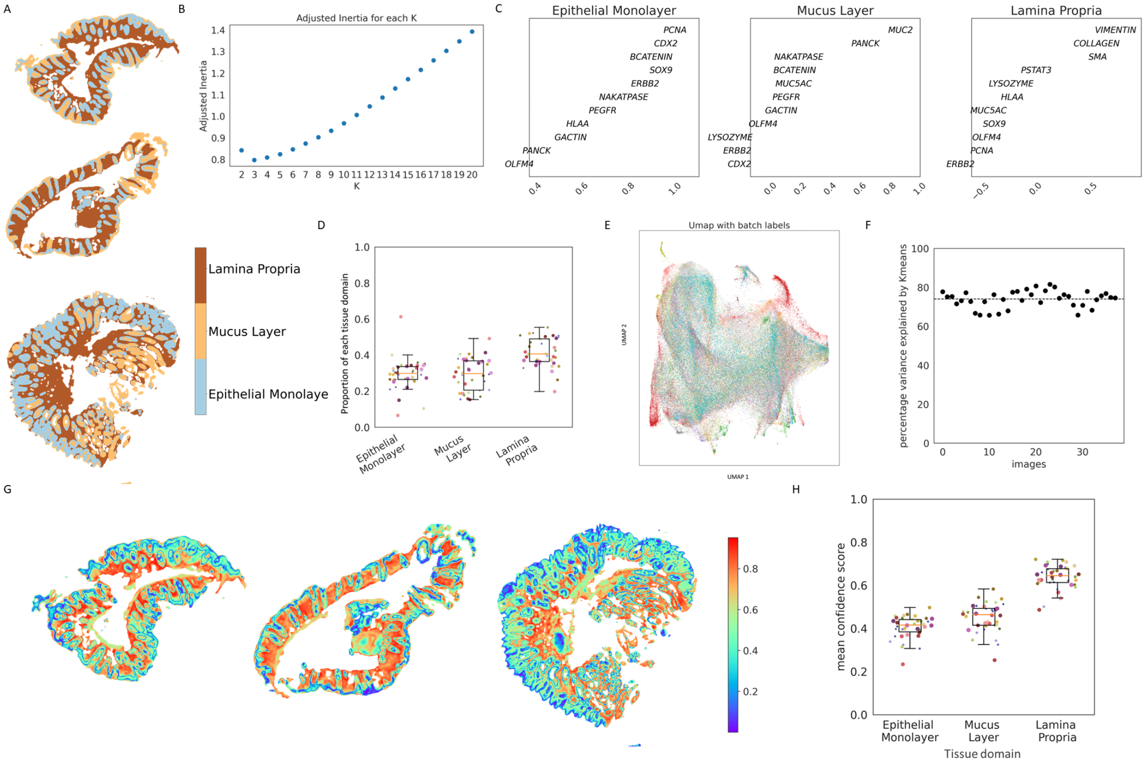


Figure 4.2: MILWRM detects canonical tissue domains in human colon mIF data.
(A) Three representative colon mIF images with labelled tissue domains ($\alpha = 0.05$)
(B) Estimated number of tissue domains in Adjusted inertia plot
(C) Domain profile describing marker composition of each tissue domain
(D) Proportion of each tissue domain across 38 samples
(E) UMAP of pixel data used for model building with batch labels
(F) Percentage variance explained by Kmeans
(G) Three representative colon mIF images with confidence score overlayed
(H) mean confidence score in each image for each tissue domain.

WRM analysis (Figure 4.2F). To assess the quality of tissue domain identification, MILWRM calculates a modified silhouette-based confidence score per pixel, which evaluates the deviation of each pixel from the centroid of the matched tissue domain relative to the closest Kmeans centroid. Most pixels across all samples have high confidence scores apart from a few in the epithelial and mucus tissue domains (Figure 4.2G-H; Figure S16A-B). Low confidence scores can be attributed to inherent biological heterogeneity within epithelial domains, as the analysis is performed over samples from mixed pathological categories (normal, AD, SER). Thus, MILWRM performed on a cohort of 37 biospecimens was able to provide physiologically relevant tissue domains with high confidence.

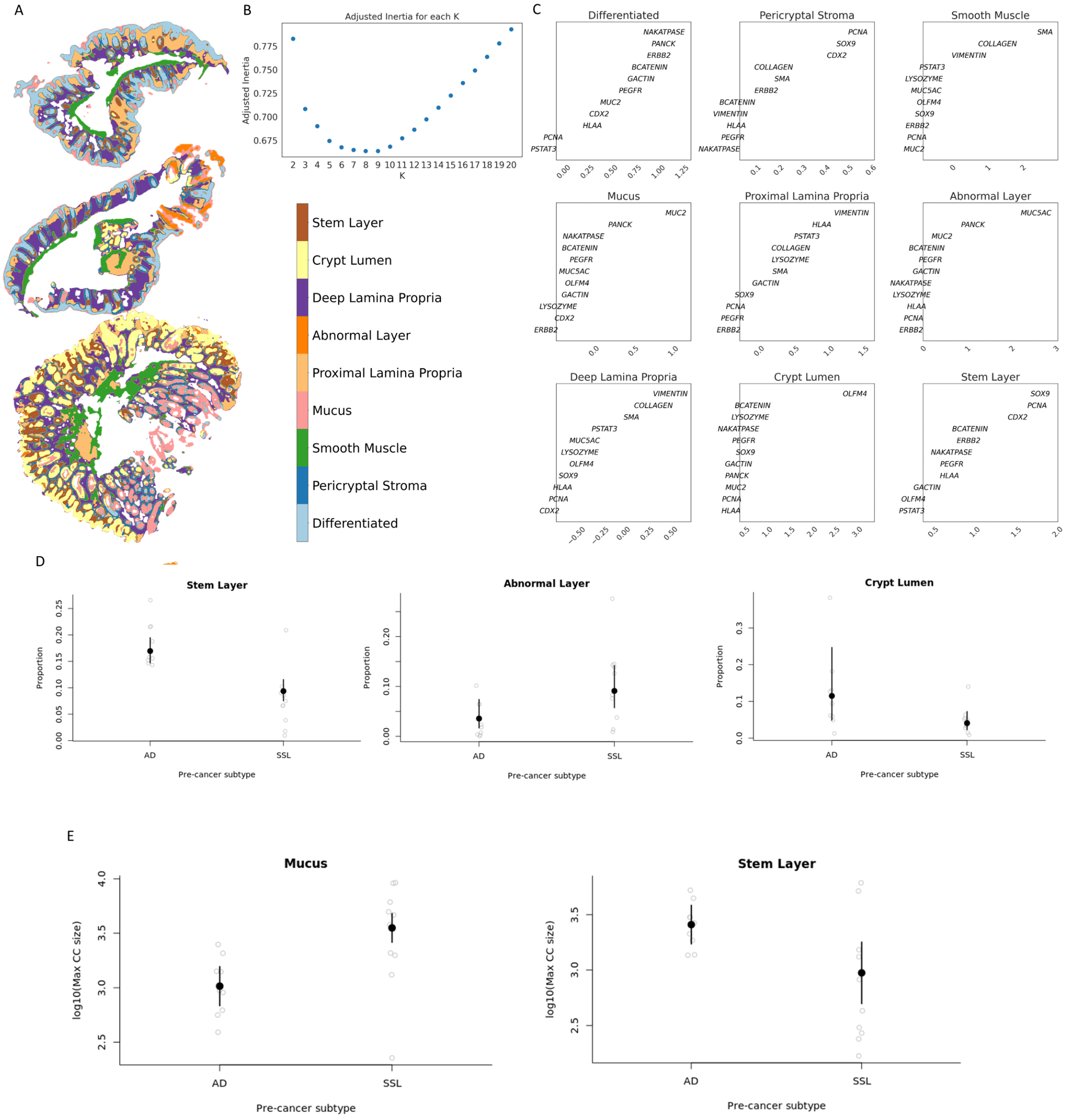


Figure 4.3: MILWRM tissue domains describe the molecular distinction between human colon adenoma pre-cancer subtypes.
(A) Three representative colon mIF images with labelled tissue domains ($\alpha = 0.02$)
(B) Estimated number of tissue domains in Adjusted inertia plot
(C) Domain profile describing marker composition of each tissue domain
(D) GEE model results for association between tissue domains and pre-cancer subtype
(E) GEE model results for association between size of connected components in tissue domains and pre-cancer subtype.

4.3.3 MILWRM identifies tissue domains that molecularly distinguish disease subtypes

To obtain more refined tissue domains that appropriately stratify the heterogenous pathological categories of our samples (normal, AD, SER), we next performed MILWRM with a reduced penalty parameter (α = 0.02). We obtained nine tissue domains that further broke down the epithelial compartment into stem (SOX9, PCNA, CDX2), differentiated (Na+-K+ ATPase, PANCK, β-catenin), mucus (MUC2), abnormal (MUC5AC+/PANCK+), and crypt lumen (OLFM4+), and the non-epithelial compartment into smooth muscle, pericryptal stroma, and proximal and deep lamina propria (Figure 4.3A-C; Figure S17A). These refined tissue domains were spatially localized appropriately. For instance, the stem and crypt lumen regions were located at the crypt base while the differentiated regions were located at the colonic surface. Interestingly, pericryptal stroma was identified with a mixture of epithelial and stromal markers and labeled a thin layer of fibroblasts that comprise telocytes constituting the stem cell niche (Figure 4.3A-C; Figure S18A-E; Aoki et al., 2016; Shoshkes-Carmel et al., 2018).

Table 4.1: Coefficient summary table for association between pixel size of tissue domain and pre-cancer subtype

Pixel size of tissue domain	Estimate	SE	Chi-square	p-value	RESI	pFDR
Differentiated tissue domain	0.412	2.12e-01	3.774	0.052	0.430	0.130
Pericryptal Stroma	0.240	1.31e-01	3.366	0.067	0.397	0.133
Smooth Muscle	-0.341	2.83e-01	1.449	0.229	0.173	0.381
Mucus	0.036	1.68e-01	0.046	0.830	0.000	0.922
Deep Lamina Propria	0.010	2.10e-01	0.002	0.962	0.000	0.962
Abnormal Layer	0.996	4.68e-01	4.534	0.033	0.485	0.130
Proximal Lamina Propria	0.145	1.87e-01	0.604	0.437	0.000	0.581
Crypt Lumen	-1.116	5.65e-01	3.904	0.048	0.440	0.130
Stem Layer	-0.680	1.48e-01	21.220	0.000	1.161	0.000
Total	50593.575	6.92e+04	0.535	0.464	0.000	0.581

We then asked whether the two pre-cancer subtypes, AD and SER, have any differences in organization of MILWRM tissue domains. We used generalized estimating equations (GEE) to model the association of MILWRM tissue domain proportions with tumor type and found a significant association between MILWRM proportions for crypt lumen, abnormal, and stem classes (Table 4.1; Figure 4.3D). Specifically, ADs were associated with higher proportions of pixels labelled as stem and crypt lumen classes, consistent with their characteristic increased stemness driven by WNT-signaling (Chen et al., 2021b; Becker et al., 2022). In contrast, serrated polyps were associated with increased pixel proportions of the abnormal class marked by MUC5AC; MUC5AC is a foregut endoderm mucin characteristic of metaplasia associated with serrated polyps (Sakamoto et al., 2017). AD arises from stem cell expansion which inevitably fill the entirety of abnormal crypts (Chen et al., 2021b). Thus, we also hypothesized that the stem MILWRM domain will be significantly more connected compared with SER tissues. We again used GEE to estimate the population average effect of pre-cancer subtype on MILWRM the maximum size of tissue-connected components (Table 4.2; Figure 4.3E) and found a significant association between connectedness of stem and mucus tissue domains and pre-cancer type. Stem domain was expectedly more connected in AD subtype whereas higher connectedness in mucus domain was associated with SER pre-cancer type. ADs have defects in differentiation of goblet cells that inherently depletes the mucus layer (Leow et al., 2004; Yang et al., 2008; Femia et al., 2007; Pretlow and Pretlow, 2005; Pietro Femia et al., 2004; Blache et al., 2004). This aligns with association of AD with decreased connected mucus components (Figure 4.3E). There was no such association observed for connectedness of the abnormal MUC5AC+ domain since it comprises sporadic abnormal cells associated with secretion. These results align with recent atlas results demonstrating that ADs arose from stem cell expansion and serrated polyps from pyloric metaplasia (Chen et al., 2021b; Sakamoto et al., 2017).

Table 4.2: Coefficient summary table for association between size of maximum connected component of each tissue domain and pre-cancer type

Maximum pixel size of connected components in tissue domain	Estimate	SE	Chi-square	p-value	RESI	pFDR
Differentiated tissue domain	0.045	0.178	0.065	0.798	0.000	0.898
Pericryptal Stroma	0.012	0.136	0.008	0.929	0.000	0.929
Smooth Muscle	-0.164	0.144	1.299	0.254	0.141	0.357
Mucus	0.535	0.105	25.849	0.000	1.287	0.000
Deep Lamina Propria	0.114	0.084	1.836	0.175	0.236	0.357
Abnormal Layer	0.275	0.254	1.177	0.278	0.109	0.357
Proximal Lamina Propria	-0.142	0.112	1.595	0.207	0.199	0.357
Crypt Lumen	-0.262	0.193	1.850	0.174	0.238	0.357
Stem Layer	-0.435	0.175	6.164	0.013	0.587	0.059

4.3.4 MILWRM applied to spatial transcriptomics reliably identify tissue domains across different mouse brain cross-sections

We applied MILWRM to a publicly available 10X Genomics Visium dataset comprising seven mouse brain samples including three coronal, two sagittal anterior, and two sagittal posterior slices (Figure 4.4A; Figure S19A-B) (10x Genomics, 2022b,a,g,f,e,d,c). We used the penalty parameter 0.02 for high resolution domain detection similar to the above for mIF data to distinguish functionally relevant brain regions (Figure 4.3B). We identified thirteen tissue domains in the brain ST data, and manually annotated them using histological information with a reference atlas from the Allen Brain Institute (Figure 4.4A – middle column) (Sunkin et al., 2013). Confidence score overlays demonstrate high quality and robust identification of most tissue domains (Figure 4.4A – right column). Notably, MILWRM identified consensus domains despite differences in the orientations and cuts of brain slices. For example, MILWRM was able to capture tissue domains that are unique only to certain slices, such as cerebellum specific to sagittal-posterior cut, as well as domains with

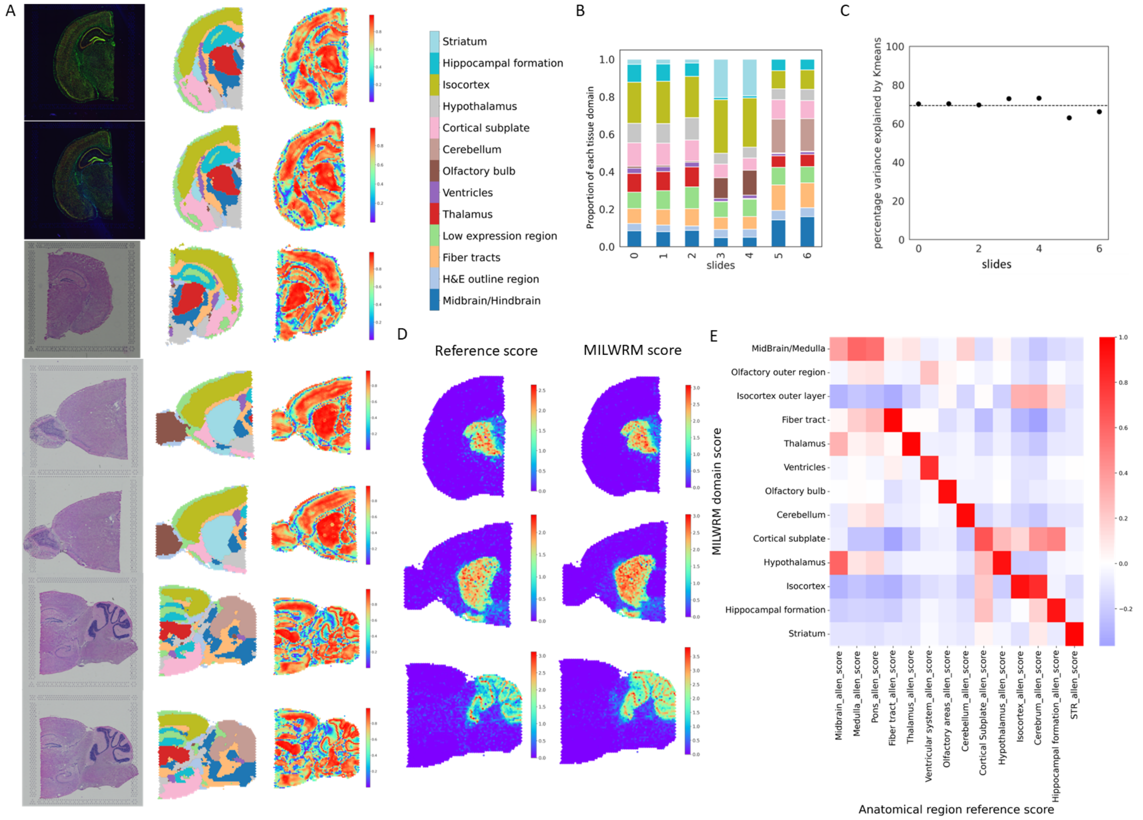


Figure 4.4: MILWRM detects consensus tissue domains in ST data from different mouse brain cross-sections.
(A) MILWRM detected tissue domains (at $\alpha = 0.02$, middle) in mouse brain ST data (H&E, left) and confidence scores (right)
(B) Proportion of tissue domains in slides
(C) Percentage variance explained by Kmeans
(D) Reference (left) and MILWRM scores (right) for Thalamus, Striatum and Cerebellum (top to bottom respectively)
(E) Overall correlation between MILWRM and reference scores for each tissue domain and anatomical region across all spots.

diverse shapes and sizes due to orientation differences, such as the striatum that is small in the coronal slice, large in the sagittal-anterior slice, and absent in the sagittal-posterior cut (Figure 4.4A-B). The MILWRM model captures approximately 70% of variance in ST data, similar to mIF results (Figure 4.4C; Figure S19C).

After histologically annotating the tissue domains using the reference atlas, we evaluated the ability of MILWRM to identify known domain distinguishing genes for potential use in unsupervised analysis. To achieve that, we curated a reference gene list for the corresponding histological regions from differential expression lists available at the Allen Brain Atlas obtained from ISH data 52. Reference lists for histological regions not available in Allen Brain Atlas were curated from the molecular atlas of mouse brain (Ortiz et al., 2020), which was generated from ST data. We first compared MILWRM domain-specific gene lists to curated reference gene lists for the corresponding histological regions. To validate the reference gene lists, we computed a signature score for each curated reference gene list per brain region, and then overlaid these signatures onto ST data. Reference gene signatures were expectedly highly specific to their respective brain regions (Figure 4.4D – left column; Figure S20A-G). In a similar vein, we also computed and overlaid MILWRM domain-specific signature scores and found that they were highly specific and accurately marked each histological brain region (Figure 4.4D – right column; Figure S21A-G). To quantify the performance between reference gene signatures and MILWRM signatures, we calculated a spot-by-spot correlation of the two sets of signature scores across all slides. High correlation between the MILWRM and reference scores was observed on a brain region-specific basis (Figure 4.4E). These results illustrate that the MILWRM approach can be effectively applied to genome-scale ST data for extracting tissue domain-specific molecular information.

4.3.5 MILWRM performs favorably when compared to SpaGCN

While there is a paucity of methods to identify and enumerate spatial domains across samples, we compared MILWRM to recently published SpaGCN, which is one of the only algorithms that can detect spatial domains on ST data over multiple samples (Hu et al., 2021). MILWRM and SpaGCN were performed on five brain ST slides analyzed above with effectively the same resolution (MILWRM α = 0.01, SpaGCN res = 0.52, p = 0.5) (Figure 4.5A). MILWRM was able to further sub-classify previously detected tissue domains into subregions. For instance, isocortex was divided into three additional layers and cerebellum into two layers, which corresponded to brain anatomy in the reference atlas (Figure 4.5A - middle column, Figure 4.5B). These finer sub-classifications were detected across multiple slices by MILWRM. In contrast, SpaGCN was unable to detect consensus spatial domains across all slides. Only common domains detected in similarly oriented cuts were identified, whereas the same domain across uniquely sliced slides were identified separately (Figure 4.5A – right column). SpaGCN, when performed at varying resolutions, was also unable to identify consensus domains across replicates slides (Figure 4.5C - Colored arrows). Although a consensus can be reached by searching for the right parameters, it is not consistent for all domains. These results further illustrate the ability of MILWRM as one of the only algorithms to robustly identify consensus tissue domains across slides with pixel information.

While SpaGCN was only designed for ST data, we still compared its performance on mIF data, as there are currently no existing algorithms for domain detection for more than one sample in mIF data. To enable SpaGCN which only works on low resolution ST data, we performed SpaGCN spatial detection on two mIF slides downsampled to 1/32 resolution with p = 0.5, res = 0.5. While the entire tissue in pixel space was classified into MILWRM domains, there were missing tissue portions in domains detected by SpaGCN (Figure 4.5D – left column). Additionally, SpaGCN detected twelve consensus spatial domains across MxIF slides, but many of these domains were spurious; only three domains predominantly represented real tissue

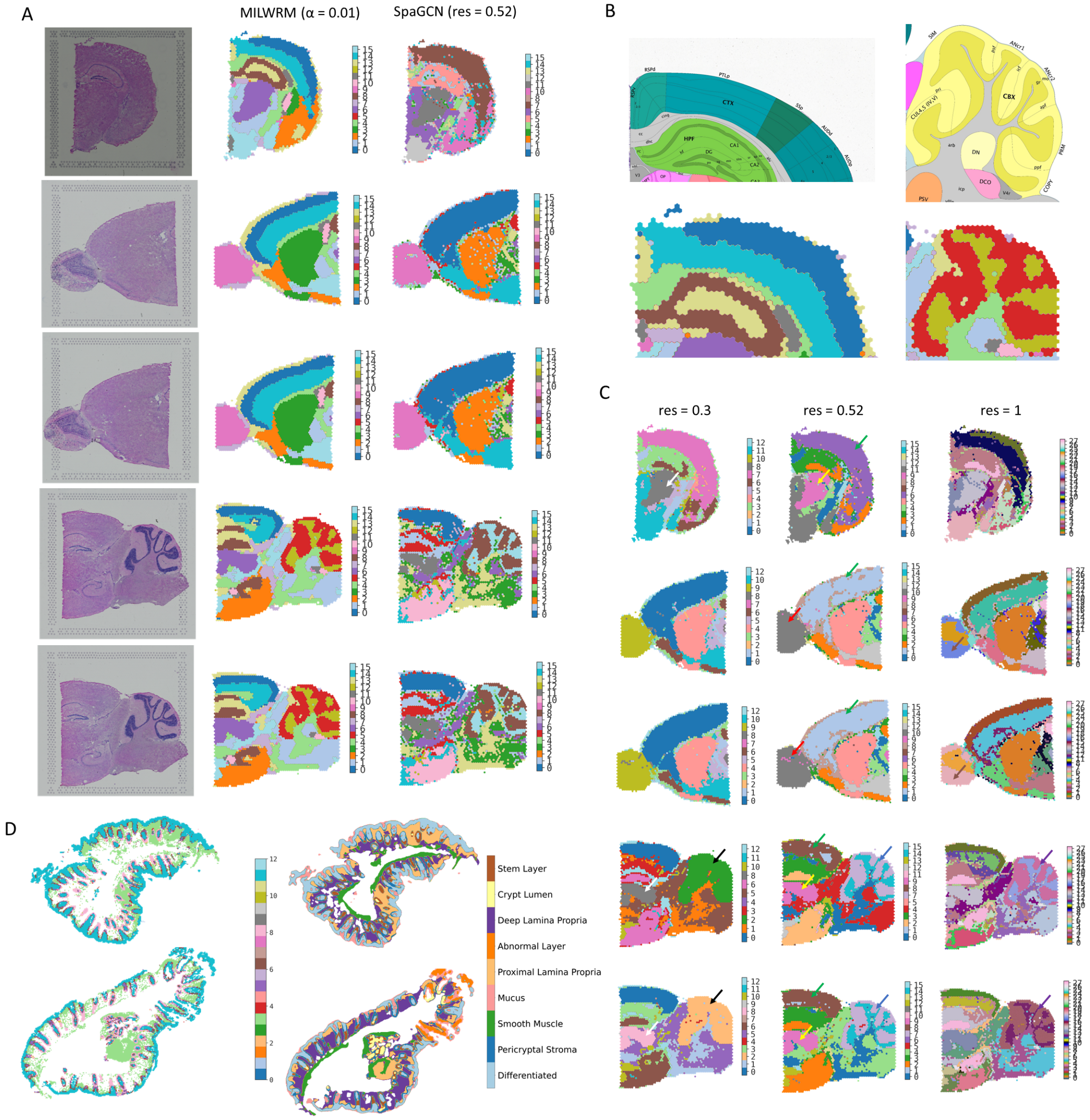


Figure 4.5: MILWRM performs better than SpaGCN.
(A) MILWRM (middle) and SpaGCN (right) detected tissue domains in mouse brain ST data
(B) Stratified layers in Isocortex and Cerebellum from Allen brain atlas (top) and MILWRM domains (bottom)
(C) SpaGCN domains at different resolutions (res 0.3, 0.52 and 1), colored arrows point the domains that should be consensus
(D) SpaGCN (left) and MILWRM domains (right) in two colon mIF specimens

regions. Furthermore, SpaGCN was unable to capture tissue domains that classify disease tissue subtypes (SER versus normal, for example), which was apparent in the MILWRM analysis across the two slides (Figure 4.5D – right column). Overall, MILWRM offers a robust and flexible approach for consensus domain identification across specimens that is generally applicable to different spatial molecular data types.

4.4 Discussion

Pixel-based tissue domain detection forms the basis of the top-down approach to spatial data analysis. Current methods of tissue domain detection are either based on a bottom-up approach, that is, building cellular neighborhoods using segmented single-cell data (Warchol et al., 2022; Kim et al., 2022; Chen et al., 2020) and/or lack scalability across samples (Alexandrov and Kobarg, 2011; Zhao et al., 2021; Townes and Engelhardt, 2021; Hu et al., 2021). Here, we addressed this gap by developing MILWRM, an algorithm to detect spatial domains across samples through a top-down, pixel-based approach. We demonstrated applicability of MILWRM to find relevant biological phenotypes in multiple data modalities (MxIF and ST) in an unsupervised way without manual thresholding and annotation.

An important demonstration of MILWRM is its ability to discern organizational differences in tissue domains related to disease subtypes. While abnormal tissues can be distinguished from normal tissues within a slide using other methods, MILWRM application across slides has significant value. There are specimens that are completely composed of abnormal tissues. In those circumstances, comparison between specimens (normal vs abnormal) is the only way to distinguish between disease states. In addition, MILWRM's ability to identify consensus tissue domains across specimens makes it possible to classify patients into disease subtypes based on tissue features. Finally, MILWRM was able to identify consensus domains and gene lists that match with organ anatomical features despite different cuts and orientations. This is important because the tissue structure from individual cuts may appear morphologically different but is functionally identical. These examples showed the real-world application of MILWRM in pathological diagnosis of disease subtypes and anatomic classification and characterization.

mIF data present additional pixel analysis obstacles. First, due to lower marker dimensionality, marker selection and management is of utmost importance. Unlike ST data where vectors of genes define programs and phenotypes, mIF phenotypes are usually defined by single markers. Highly expressed markers may mask lower expression markers if suboptimal preprocessing is performed, thus preventing some tissue domains from being detected. Secondly, high resolution microscopy data are generally incompatible with pixel-based algorithms built for low resolution ST data, such as SpaGCN. Creation of image tiles or large-scale downsampling is needed to satisfy speed and memory requirements. In contrast to most state-of-the-art methods for pixel-based analysis that are data type-specific, MILWRM is adaptable to multiple imaging data types and is scalable to many samples. While SpaGCN is a scalable method for spatial domain detection, it failed to discern disease-specific differences in tissues from the tissue domains it identified. Additionally, the domains identified by SpaGCN in ST data were not robust since they failed to reach consensus at different clustering resolutions. Additionally, MILWRM also provides various QC metrics which can be used to assess the quality of domain detection and ability to perform tissue clustering at different levels of smoothing, downsampling, and cluster resolution.

CHAPTER 5

Molecular cartography uncovers evolutionary and microenvironmental dynamics in sporadic colorectal tumors

Adapted from:

Heiser, C. N., Simmons, A. J., Revetta, F., McKinley, E. T., Ramirez-Solano, M. A., Wang, J., Shao, J., Ayers, G. D., Wang, Y., Glass, S. E., Kaur, H., Rolong, A., Chen, B., Vega, P. N., Drewes, J. L., Saleh, N., Vandekar, S., Jones, A. L., Washington, M. K., Roland, J. T., Sears, C. L., Liu, Q., Shrubsole, M. J., Coffey, R. J., and Lau, K. S. (2023). Molecular cartography uncovers evolutionary and microenvironmental dynamics in sporadic colorectal tumors. *bioRxiv*, page 2023.03.09.530832

5.1 Summary

Colorectal cancer exhibits dynamic cellular and genetic heterogeneity during progression from precursor lesions toward malignancy. Leveraging spatial molecular information to construct a phylogeographic map of tumor evolution can reveal individualized growth trajectories with diagnostic and therapeutic potential. Integrative analysis of spatial multi-omic data from 31 colorectal specimens revealed simultaneous microenvironmental and clonal alterations as a function of progression. Copy number variation served to re-stratify microsatellite stable and unstable tumors into chromosomally unstable (CIN+) and hypermutated (HM) classes. Phylogeographical maps classified tumors by their evolutionary dynamics, and clonal regions were placed along a global pseudotemporal progression trajectory. Cell-state discovery from a single-cell cohort revealed recurring epithelial gene signatures and infiltrating immune states in spatial transcriptomics. Charting these states along progression pseudotime, we observed a transition to immune exclusion in CIN+ tumors as characterized by a novel gene expression signature comprised of *DDR1*, *TGFBI*, *PAK4*, and *DPEP1*. We demonstrated how these genes and their protein products are key regulators of extracellular matrix components, are associated with lower cytotoxic immune infiltration, and show prog- nostic value in external cohorts. Through high-dimensional data integration, this atlas provides insights into co-evolution of tumors and their microenvironments, serving as a resource for stratification and targeted treatment of CRC.

5.2 Introduction

The genetic model of colorectal cancer (CRC) progression defines a sequence of cumulative mutational burden that drives dysplasia and malignancy in human colonic epithelium (Fearon and Vogelstein, 1990). This conventional adenoma-carcinoma trajectory involves alteration of driver genes *APC*, *KRAS*, and *TP53*, resulting in chromosomal instability (CIN) (Stoler et al., 1999). Alternatively, a subset of CRCs which arise from the so-called serrated pathway, are more likely to be *BRAF*-driven and microsatellite unstable (MSI-H) due to hypermethylation of *MLH1* and other mismatch repair (MMR) genes, causing hypermutation (Rhee et al., 2017; Nouri Nojadeh et al., 2018). Ensuing decades of investigation have yielded additional CRC subtyping that elucidates alternative pathways to invasion and metastasis, as well as characterization of pre-malignant lesions and their clinical prognoses (Conteduca et al., 2013; Obuch et al., 2015; Guinney et al., 2015). More recently, the advent of single-cell and spatial molecular assays has uncovered various degrees of intratumoral heterogeneity at high resolution, suggesting that previously proposed linear tumor progression along the conventional or serrated pathway cannot fully explain the evolutionary dynamics of the second leading cause of cancer-related mortality worldwide (Bray et al., 2018; Chen et al., 2021b; Joanito et al., 2022; Gil Vasquez et al., 2022). Moreover, layered molecular information from spatially resolved assays can be used to build models relating gene and protein expression, or cell ”state”, to clonal identity across tumor regions (Ji et al., 2020; Moncada et al., 2020; Wu et al., 2022a; Risom et al., 2022). Exploration of tumor phylogeography in this way allows for deeper profiling of evolutionary relationships while accounting for regional heterogeneity (Shibata, 2020).

Beyond, and perhaps fundamental to, genetic evolution of malignant cells themselves lies additional complexity introduced by interactions between tumor epithelium and infiltrating immune cells, which provide immunogenic selection pressure that profoundly impacts tumor evolution and prognosis (Rooney et al., 2015; Milo et al., 2018; Lee et al., 2020). Indeed, characterization of the immune compartment of solid tumors has been shown to be a better predictor of patient prognosis than traditional pathological staging, and tumor immunophenotype is a valuable measure in forecasting response to immune checkpoint inhibition (ICI) in several cancers (Galon et al., 2006; Charoentong et al., 2017). Furthermore, distinct pathways from initiation to malignancy confound these tumor characteristics, as serrated lesions and MSI-H CRCs are more immunogenic than their conventional adenoma and microsatellite stable (MSS) counterparts on average (Nouri Nojadeh et al., 2018; Chen et al., 2021b). Importantly, several types of advanced solid tumors have been shown to exhibit immune exclusion or evasion, by mechanisms both intrinsic to cancer cells and observed microenvironmentally, which shortens overall patient survival and confers ICI resistance (Feig et al., 2013; Roh et al., 2017; Lazarus et al., 2018; Luke et al., 2019; Abril-Rodriguez et al., 2019; Sun et al., 2021; Pelka et al., 2021; Baldominos et al., 2022). In CRC, an observed suppression of cytotoxic immunity that trends with an increased stem cell signature in late-stage carcinoma raises a critical question surrounding immune exclusion and the potential connection to tumor progression (Chen et al., 2021b).

In order to map the co-evolution of colorectal tumor cells and their microenvironments, we leveraged spatial multi-omics to build a phylogeographical atlas of CRC progression from pre-cancer to adenocarcinoma. We have generated a novel, spatially resolved dataset from a heterogeneous set of sporadic colorectal tumors, wherein distinct tumor regions represent snapshots of cancer evolution. Multiregional mutational profiling, untargeted spatial transcriptomics (ST), and multiplexed protein imaging offer high-dimensional paired measurements of layered molecular information while maintaining tissue contexts (Ryser et al., 2018; Kather et al., 2018; Barkley et al., 2022).

Combining spatial data with single-cell RNA sequencing (scRNA-seq) to enumerate consensus cell states, we projected tumor programs and microenvironmental features onto a generalized progression pseudotime (PPT) derived from regional copy number variants (CNVs) and somatic mutational profiles amongst a cohort of tumors. These efforts enabled discovery of multiple pathways that are distinctly altered during the progression of chromosomally unstable (CIN+) and hypermutated (HM) CRCs. This study presents a patient-centric roadmap of CRC evolution and progression arising from integrative, atlas-wide analyses across a unique and heterogeneous set of human CRC specimens.

5.3 Results

5.3.1 Spatial atlas queries layers of molecular heterogeneity in sporadic colorectal tumors

To model CRC progression through spatial heterogeneity, we selected human colonic specimens with regional morphologies representing transitions between tumor progression stages. Samples with concurrent pre-malignant, malignant, and invasive regions were identified by a pathologist (Appendix A5.1: Sample procurement). A diversity of tumor stages, grades, and locations was represented (Figure 5.1A-B) with fairly equal selection of MSS (*n* = 12) and MSI-H (*n* = 10) CRCs and pre-cancerous polyp subtypes (*n* = 8; 4 SS-L/HP and 4 TA/TVA). Along with a normal colon sample as control, these specimens provide a representative spectrum of disease states along the two major pathways to malignancy in the colon (Figure 5.1A-B).

For each of these 31 formalin-fixed, paraffin-embedded (FFPE) specimens, we collected serial tissue sections for parallel processing by molecular profiling assays. Multiplex immunofluorescence (MxIF) analysis with a 33-marker panel provided whole-slide protein expression at subcellular resolution (Gerdes et al., 2013; Figure 5.1C). We cut serial sections into one or more capture areas of 10X Genomics Visium spatial transcriptomics (ST) slides (Ståhl et al., 2016; Vickovic et al., 2019). Large tumors (¿7 mm diameter) were trimmed to regions of interest (ROIs) targeting dysplastic epithelium and minimizing stromal areas. H&E images were collected to align gene expression with tissue morphology and provide fiducial markers for spatial registration to MxIF. We employed laser capture microdissection followed by whole-exome sequencing (LCM-WES) on additional serial sections to genetically profile $\sim$2 mm ROIs in spatially distinct regions of individual tumors (Figure 5.1C). Dissociative single-cell RNA sequencing (scRNA-seq) for a subset of specimens provided a reference for expected cellular composition (Klein et al., 2015; Chen et al., 2021b).

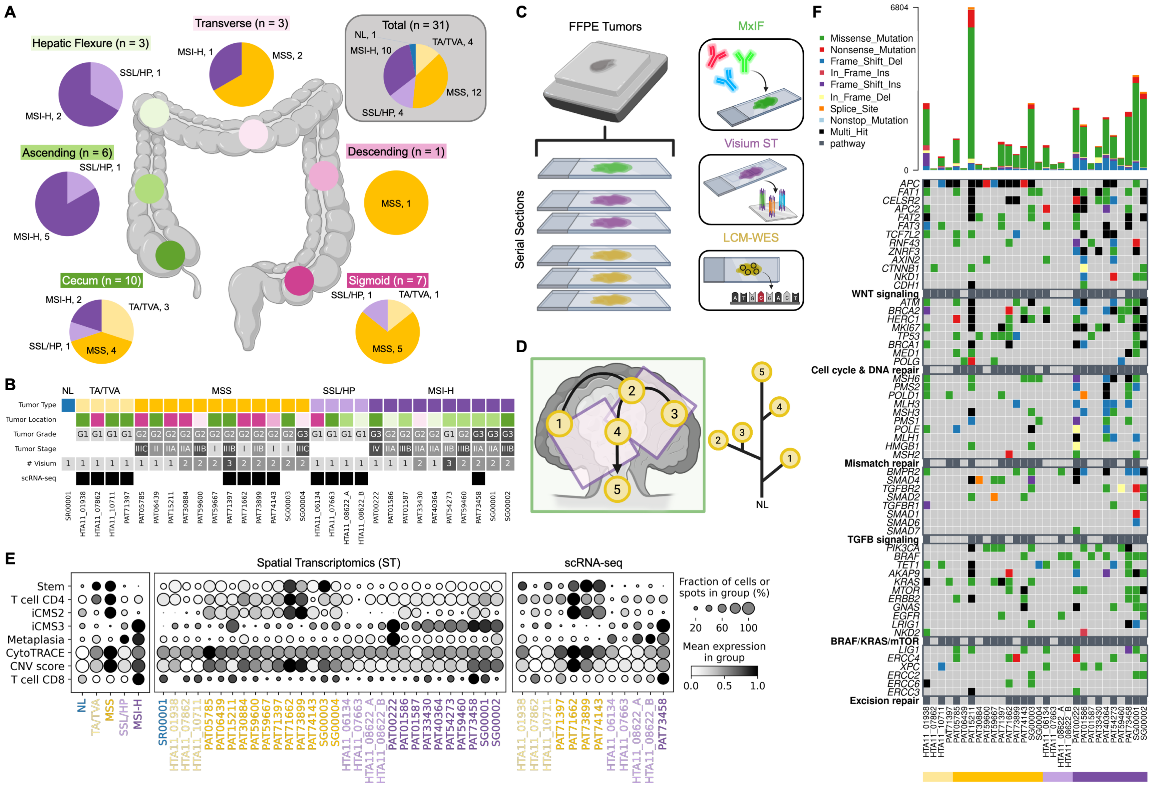


Figure 5.1: Spatial atlas queries layers of molecular heterogeneity in sporadic colorectal tumors.
(A) Diagram detailing colorectal specimens chosen for atlas experiments.
(B) Patient-level information from A in table format. Large CRCs (¿7 mm diameter; 12 MSS, 10 MSI-H) have at least 1 spatial transcriptomics (ST) replicate tiling the tissue. TA/TVAs (4), HP/SSLs (4), and NL (1) have 1 ST replicate each. All specimens have whole-slide MxIF imaging and bulk or multiregional WES.
(C) Experimental design consisting of layered spatial molecular assays from serial sections of FFPE tissue blocks.
(D) Diagram of phylogeographical cartography from multiregional sequencing (LCM-WES) data, layered with ST and MxIF images. Black arrows represent progression pseudotime (PPT) inferred from the phylogenetic relationships between ROIs.
(E) Summary of gene signature scores by tumor type (left), ST patient (middle), and matched scRNA-seq patient (right). Patient ID colors represent tumor type (MMR status). Mean signature expression scaled across groups. In this dotplot and hereafter, size of dots represents expression frequency, while shade represents intensity.
(F) Somatic mutations detected in LCM-WES samples, summarized by patient and grouped by biological pathway. Top barplot represents overall TMB breakdown by mutation class per patient.

The data procured from multi-modal analysis of adjacent sections present trade-offs across large ranges of spatial resolution and molecular specificity: MxIF offers subcellular (¡0.5 µm) imaging of dozens of molecular features, ST captures up to 19,000 mRNA transcripts at 100 µm spatial resolution, and LCM-WES profiles somatic mutations for large (1-2 mm) regions of tissue. We used image processing software to register genetic, transcriptomic, and proteomic organizational layers, modeling these molecular mixtures at varying spatial scales (Appendix A5.1: Spatial registration). We use these spatial features to infer relationships of tumor progression and evolution (Figure 5.1D), and thus provide a ”scaffold” for phylogenetic cartography and subsequent modeling of gene, protein, and cellular features (Shibata, 2020).

RNA expression data from 48 ST samples and a subset of 13 matched scRNA-seq samples largely exhibited expected gene expression trends across tumor class and grade when analyzed at the bulk level. Conventional TA/TVA polyps and MSS CRCs were enriched for stemness, intrinsic consensus molecular subtype 2 (iCMS2) epithelial signature, and an inflammatory immune response signature (CD4+ T lymphocytes), while SSL/HP and MSI-H CRCs were comprised of metaplastic and iCMS3 signatures accompanied by cytotoxic (CD8+ T cell) immunity (Chen et al., 2021b; Joanito et al., 2022; Figure 5.1E). Additionally, aggregating multiregional exome sequencing into bulk analyses per patient revealed *APC*, *KRAS*, and *TP53* mutations in TA/TVA/MSS tumors, consistent with the conventional adenoma-carcinoma sequence, and *BRAF* variants in SSL/HP/MSI-H samples (Figure 5.1F). Similar to previous observations, MSI-H CRCs exhibit hypermutation that is absent in their SSL/HP counterparts, delineating that this transition occurs after serrated pre-malignancy (Chen et al., 2021b). Moreover, a small subset of conventional pathway tumors (1 TVA - HTA11_01938 and 1 MSS - PAT15211) can be identified as hypermutated (HM), consistent with previous observations, likely due to deficiency in proofreading polymerases (*POLE*, *POLD1*; Bourdais et al., 2017; Figure 5.1F).

5.3.2 CNV inference establishes spatially resolved tumor clones and their phylogenetic relationships

A hallmark of the conventional adenoma-carcinoma sequence that gives rise to MSS CRC is chromosomal instability (CIN) which results in somatic gains, losses, and rearrangements of large segments of DNA heritable to cellular progeny (Nouri Nojadeh et al., 2018; Drews et al., 2022). Thus, cumulative increases in CNVs can be used to order tumor progression events amongst tumor regions if measured spatially (Erickson et al., 2022).

We inferred CNVs from ST data, quantifying levels of CIN across the atlas with support from orthogonal measurements including scRNA-seq, WES, and WGS data (Puram et al., 2017). Dimension reduction and embedding of inferred CNV profiles from epithelial ST microwells yields tumor-specific clustering indicative of unique somatic CNVs (Figure 5.2A). A subset of patients with matched scRNA-seq ($n = 11$) exhibited sim-

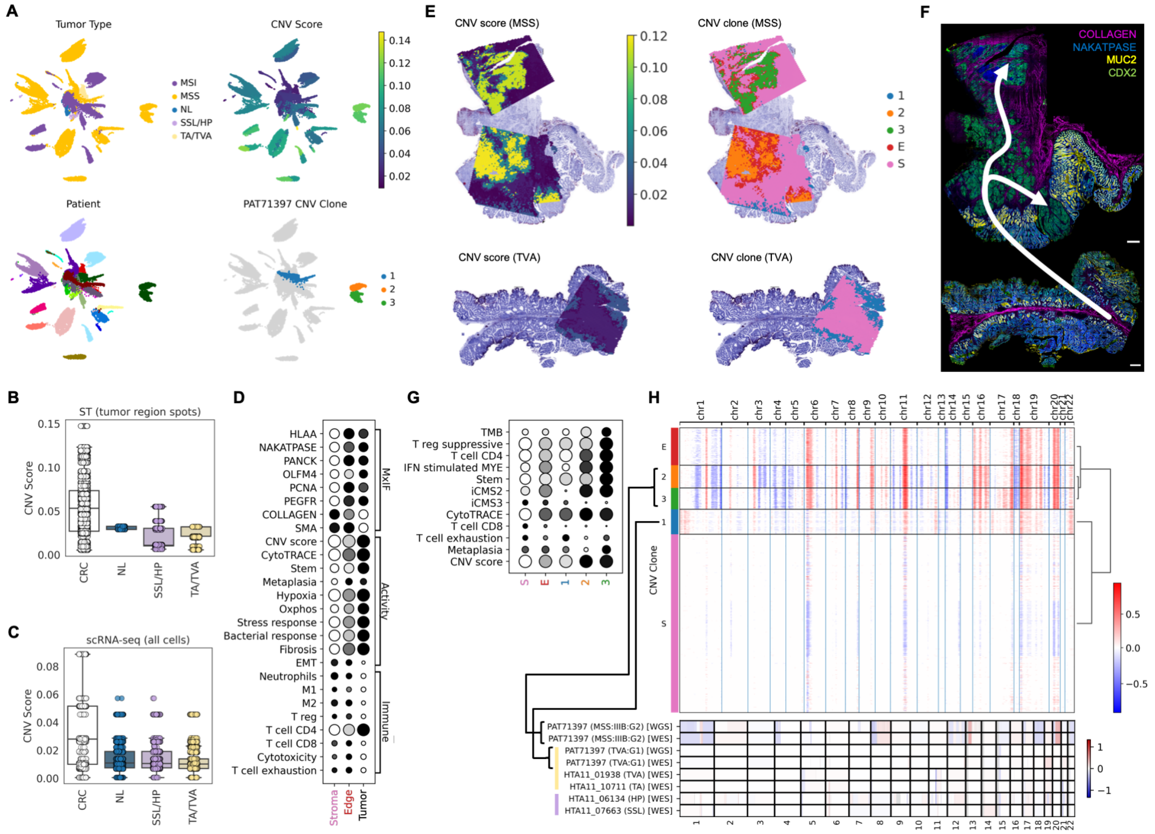


Figure 5.2: CNV inference establishes spatially resolved tumor clones and their phylogenetic relationships.
(A) UMAP embeddings generated from inferred CNV profiles of all ST samples colored by tumor type, CNV score, Patient, and PAT71397 CNV clone to accompany panels E-H. Individual points represent ST microwells, which were subsetted to major clone regions prior to embedding.
(B) Boxplots of CNV scores for all ST microwells in major CNV clone regions across atlas, grouped by sample type. CRC $n = 48{,}439$; NL $n = 1{,}067$; SSL/HP $n = 1{,}735$; TA/TVA $n = 1{,}951$.
(C) Boxplots of CNV scores for epithelial cells from Chen, *et al.* cohort, grouped by sample type. CRC $n = 11{,}982$; NL $n = 31{,}917$; SSL/HP $n = 11{,}896$; TA/TVA $n = 21{,}275$.
(D) Summary of MxIF intensities, cell activity, and immune gene signatures by major tissue domains determined through CNV inference.
(E) CNV scores (left) and tumor clone regions (right) for PAT71397.
(F) MxIF with inferred progression trajectory for PAT71397. Scale bars 500 µm.
(G) Summary of TMB, CNV score, and gene signatures for CNV clone regions of PAT71397.
(H) Heatmap of inferred CNVs for PAT71397 ST, corresponding to E-G (top), as well as CNVs measured by WGS and WES for PAT71397 blocks and additional selected pre-malignant tumors (bottom). Brackets connect WES and WGS from PAT71397 malignant (MSS) and benign (TVA) blocks to dominant CNV clones in respective ST to show similarity of measured and inferred CNV profiles.

ilar CNV profiles inferred from both scRNA-seq and ST, with exception of CIN-low pre-malignant tumors whose inferred copy number changes consisted of mostly background noise (Figure S22A-D). To quantify this copy number validation between gene expression modalities, we calculated pairwise cosine similarities between inferred CNV profiles of all cells (scRNA-seq) and microwells (ST) assigned to major clones in each tumor (Appendix A5.1: CNV inference from ST and scRNA-seq). The resulting distributions of these CNV similarities describes the confidence of inferred somatic copy number calls, as chromosomally unstable (CIN+) tumors have cosine similarity distributions approaching 1.0, while CIN- specimens have cosine similarities centered around 0.0 or less (Figure S22B-D). These results derive from the fact that CIN- specimens have so few somatic CNVs that the attempted inference of such rearrangements by expression yields low-confidence, noisy results in both modalities (Figure S22D).

To further validate these results using direct measurements of genomic alterations, we performed whole-genome (WGS) and/or whole-exome (WES) sequencing on a subset of tumors from this atlas, as well as additional pre-malignant lesions and CRCs from a larger cohort (Chen et al., 2021b). CNVs called in these data confirmed somatic copy number alteration patterns inferred by ST and scRNA-seq in overlapping samples (Figure S22E; Appendix A5.1: CNV calling from WES and WGS).

Several studies have demonstrated that APC dysfunction causes CIN due to the protein's interaction with microtubules in the spindle and contractile ring during cytokinesis (Tighe et al., 2001; Rusan and Peifer, 2008). Thus, CIN has been thought to arise as an early event in tumorigenesis, potentially when APC function is lost during initial adenoma formation, as implicated in mouse models (Alberici et al., 2007). However, human studies using limited numbers of specimens do not provide a clear confirmation (Sieber et al., 2002; Cardoso et al., 2006). To address the relationship between APC and CIN on a broader basis in humans, we summarized CNV scores across our ST atlas and a large cohort of scRNA-seq derived from CRCs and pre-cancers (Chen et al., 2021b; n = 85), demonstrating that conventional adenomas (TA/TVA) harboring *APC* mutations exhibited low CIN comparable to serrated polyps (SSL/HP) and baseline normal epithelium (Figure 5.2B-C). We performed WGS (n = 35) and WES (n = 18) on a selected subset of tumors and similarly calculated total CNV scores from these data, which validated the lack of CIN in TA/TVAs inferred from gene expression (Figure 5.2H; Figure S22E-F; Appendix A5.1: CNV calling from WES and WGS).

Taken together, these data suggest that the onset of CIN occurs later in carcinoma development than previously assumed, and that MSS tumors are more likely to become CIN+ than MSI-H carcinomas are (Stoler et al., 1999; Woodford-Richens et al., 2001; Sieber et al., 2002; Sheffer et al., 2009; Pino and Chung, 2010). We do, however, observe some MSI-H tumors that gain CIN, likely coincident with an observed transformation to a stem-like, iCMS2 phenotype (Chen et al., 2021b; Figure S22F-H). In fact, three MSI-H tumors in this atlas (SG00001, SG00002, PAT73458) exhibited high CNV scores, and were thus classified as CIN+. These exceptions are analogous to the hypermutated (HM) TA/TVA/MSS tumors (Figure 5.1F), suggesting that alternative pathways to CIN and HM will emerge in some cases.

In ST data, CNV inference provides tumor clone regions that spatially align with tissue domains annotated by histology as well as gene signatures enriched in dysplastic epithelium (Figure 5.2D). Specifically, ST CNV profiles cluster into major tumor clones based on similarity, while normal mucosa and stromal regions are aggregated into a background cluster (”S”; Figure 5.2D-H). Additionally, CNV inference identifies a tumor edge domain (”E”) resulting from a mixture of epithelium and surrounding stroma in ST microwells that dampens the CNV signal. These regions are useful for characterizing epithelial activity occurring at tumor borders such as epithelial-mesenchymal transition (EMT) and tumor cell invasion, as well as interactions with the tumor microenvironment (TME) including antigen presentation, lymphocyte exhaustion, cytotoxicity, and neutrophil recruitment (Figure 5.2D,G). Major clone regions can be ordered and labeled according to their inferred progression stages within each tumor, determined by a combination of CNV profile and mutational burden (TMB), as well as their relationships to one another (PAT71397 clone regions ”1”, ”2”, ”3”; Figure 5.2E-H). Finally, ordering of CNV clones on an individual tumor basis is validated by tissue characteristics that trend with malignancy and progression, exemplified in PAT71397 by an increase in TMB, CD4+ T cell and suppressive T reg infiltration, and gene signature scores including iCMS2, stem, and CytoTRACE (Figure 5.2G).

5.3.3 Multiregional somatic mutational profiles provide phylogeographical topology

Phylogenetic reconstruction provides insight into tumor progression from a snapshot in time, namely the moment of tumor resection and fixation. Our analysis revealed spatially informed clonal heterogeneity within tumors. We first called somatic mutations using bulk germline WES from each patient as a baseline (Appendix A5.1: Somatic mutational profiling with LCM-WES). Results from 22 patients with LCM-WES exhibited observable phylogeny across three or more regions of interest (ROIs), deciphered using public, shared, and private mutations detected in spatially distinct tumor regions (Figure 5.3A-C; Figure S23A-W). Common and unique genetic alterations between ROIs allowed for the reconstruction of evolutionary relationships between tumor regions, which can follow one of several proposed evolutionary models (Sottoriva et al., 2015; Venkatesan and Swanton, 2016; Turajlic et al., 2019; Ryser et al., 2020; West et al., 2021. Our data were consistent with three major models of tumor evolution (Figure 5.3D).

Linear or punctuated evolution consists of stepwise increases in fitness that result in periodic ”clonal sweeps” that replace the dominant makeup of the tumor (Gould and Eldredge, 1977). These tumors typically exhibit low degrees of heterogeneity across regions with many shared or public mutations (Turajlic et al., 2019). The neutral or ”big bang” model of tumor evolution consists of many subclones with near-equal fitness

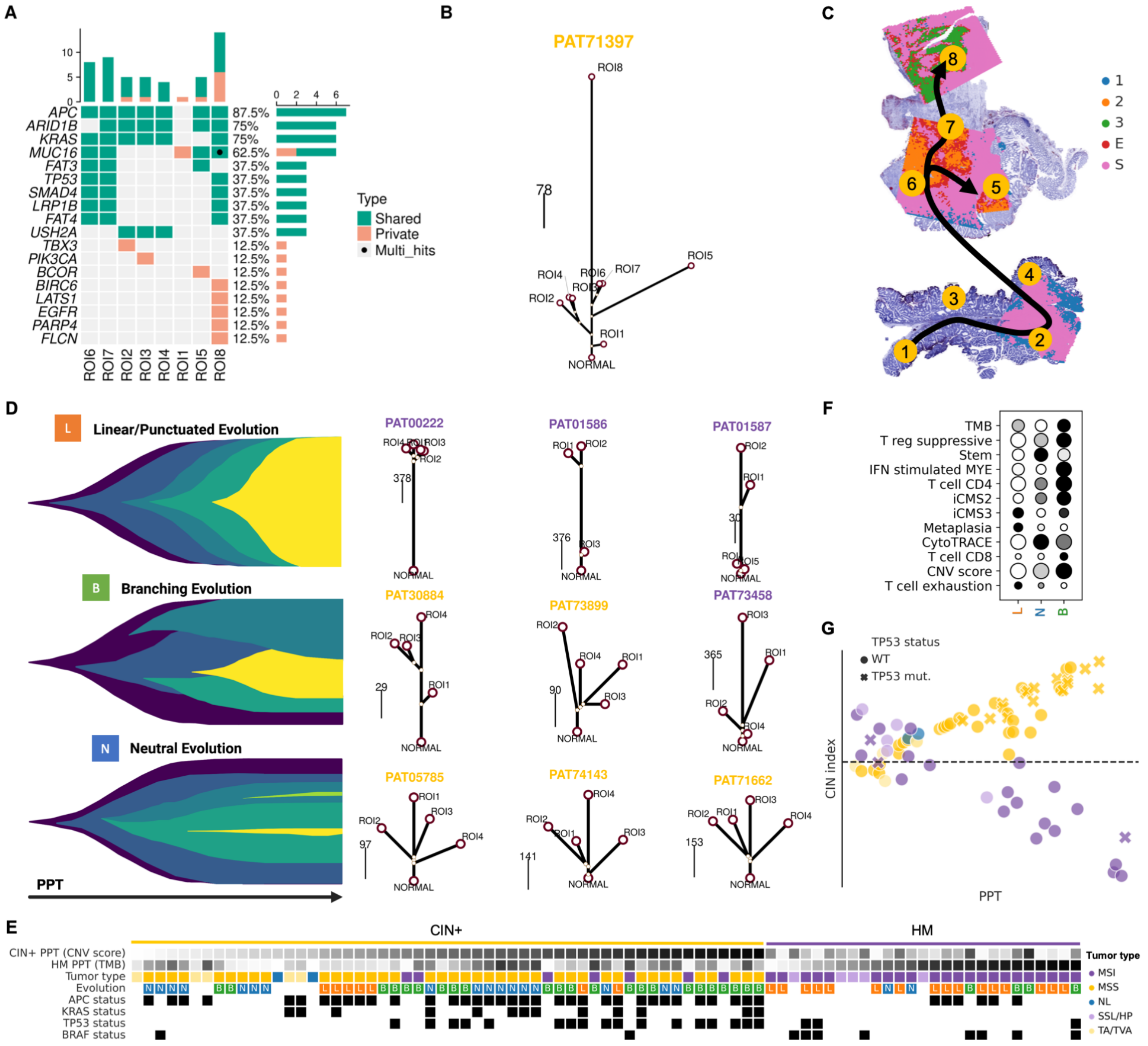


Figure 5.3: Multiregional somatic mutational profiles provide phylogeographical topology.
(A) Oncoplot of detected driver mutations within spatially sampled LCM ROIs of PAT71397.
(B) Phylogenetic tree for PAT71397. Length of branches are proportional to the number of shared or private somatic mutations in each LCM ROI.
(C) Diagram of LCM ROIs in PAT71397 blocks, overlaid on CNV clone regions identified in ST. Black arrow represents inferred progression trajectory from CNVs and mutational phylogeny in B.
(D) Diagram of observed modes of tumor evolution. Example phylogenetic trees from representative atlas samples shown to the right of each diagram. Patient ID colors represent tumor type (MMR status).
(E) Tumor regions and their clinical and mutational metadata divided by class and ordered left-to-right by corresponding PPT (CNV score for CIN+, TMB for HM).
(F) Summary of gene signatures across all tumor regions grouped by evolutionary mode from D.
(G) CIN index versus PPT for tumor regions from E. Points are colored by tumor class, except pre-malignant and normal regions, which are colored according to tumor type as in Figure 5.1A-B. Points are colored by tumor class. Point shape corresponds to regions with detected *TP53* mutation.

due to early bursts of mutational events that persist throughout the life of the lesion (Sottoriva et al., 2015). Similarly, branching evolution is characterized by co-existing subclones that exhibit hierarchical relationships between one another.

When categorizing carcinomas in the atlas by phylogenetic structure, we observed that 8/22 tumors exhibited linear evolution characterized by many public mutations and low regional heterogeneity. Of the remaining 14 carcinomas, we classified six as neutral - effectively the opposite of linear, with few public mutations and high regional heterogeneity - and eight as branching. Interestingly, we classified the evolution of only 2/15 CIN+ tumors as linear. Conversely, linear evolution dominated the HM cohort (6/8 cases; Figure 5.3E; Figure S23X; Appendix A5.1: Global PPT ordering and classification of tumor regions). Indeed, CIN+ signatures such as iCMS2, stem, CD4 T cells, and high CNV score are enriched in tumors with neutral and branching evolution, while HM markers including iCMS3, metaplasia, and T cell exhaustion are highly expressed by tumors undergoing linear evolution (Figure 5.3F). Despite a limited sample size, we speculate that hypermutation confers stronger and more protracted stepwise gains of clonal fitness, resulting in linear or punctuated evolution, whereas CIN promotes high-frequency, continual genetic alterations that yield minor, hierarchical differences between co-existing subclones in a branching or neutral evolutionary pattern (Grist et al., 1992; Jackson and Loeb, 1998; De Nooij-van Dalen et al., 2001; Nowak et al., 2002).

Given two distinct indicators of regional tumor progression, CNVs and somatic mutations, we next distinguished the major classes of colorectal tumors whose regional progression pseudotime (PPT) are best quantified by these indicators, respectively. Chromosomally unstable (CIN+) tumors, most likely MSS arising from the conventional adenoma pathway, should exhibit CNV profiles that reliably describe regional clonal relationships that recapitulate tumor progression topology. Hypermutated (HM) tumors, most likely arising from the serrated pathway, would conversely be MSI-H and chromosomally stable. Therefore, clonal progression is best described by regional TMB in the HM case.

To demonstrate these principles, we established PPT ordering and calculated a CIN index (quantifying the comparative degree of CIN and hypermutation) for tumor regions defined as a combination of LCM ROIs and CNV clones (Figure 5.3E,G; Appendix A5.1: Global PPT ordering and classification of tumor regions). PPT ranking allowed for atlas-level integration and modeling of spatial information along a global indicator of progression from normal mucosa and pre-malignant adenoma to invasive adenocarcinoma. Advanced tumor regions (PPT ¿ 0.4) with positive and negative CIN indices were classified as CIN+ and HM, respectively, uncovering spatial heterogeneity that describes transitions from MSI-H to CIN+ and MSS to HM. Importantly, we note that *TP53* mutations were more enriched than *APC* mutations in tumor regions with high CIN index, corroborating our observation that CIN emerges later in CRC development (Fearon and Vogelstein, 1990; Figure 5.3G; Figure S23Y).

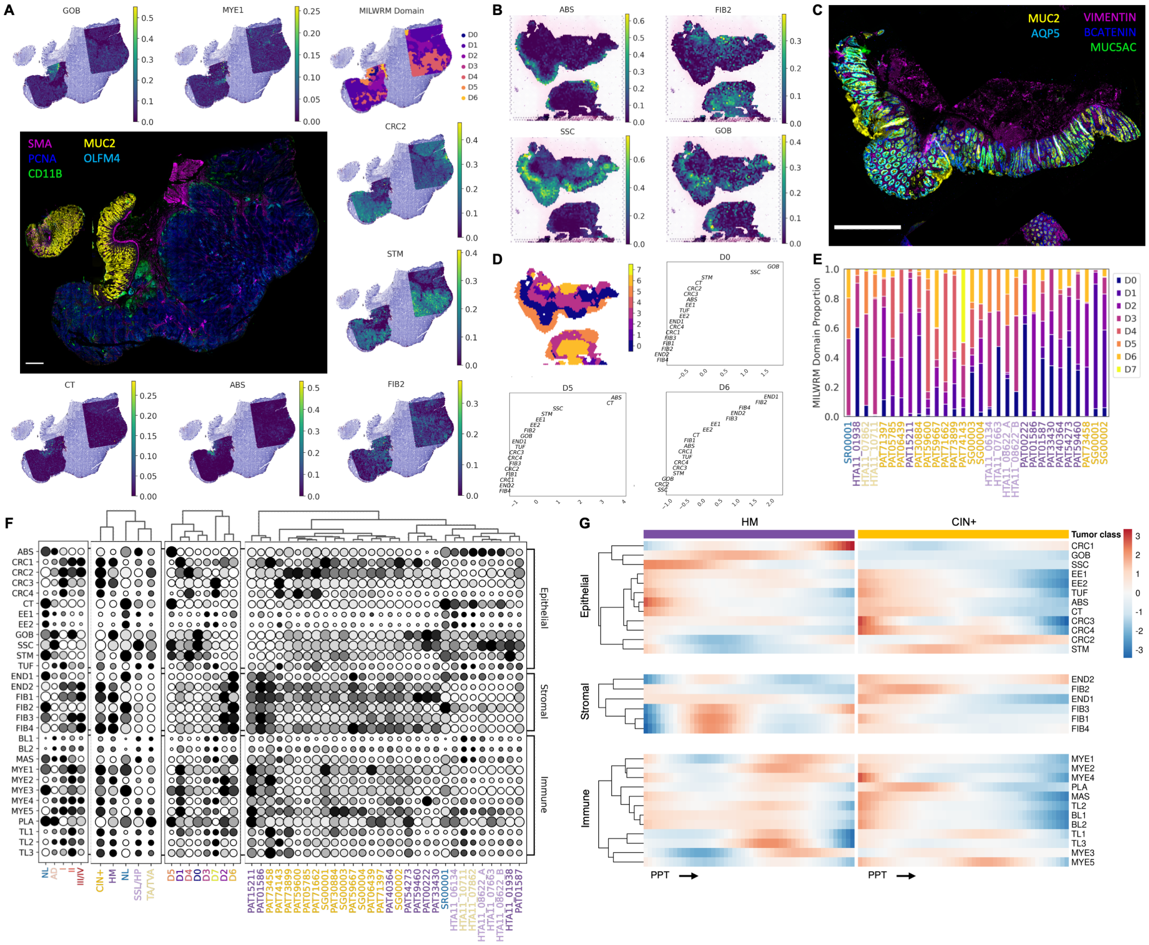


Figure 5.4: Cell-state deconvolution reveals pseudotemporal tissue dynamics.

(A) Example whole slide MxIF surrounded by refNMF usages for seven cell states as well as MILWRM tissue domain projected onto PAT30884 histology. Scale bar 500 µm.

(B) refNMF usages of normal absorptive colonocyte (ABS), normal fibroblast (FIB2), serrated-specific cell (SSC), and goblet cell (GOB) states for HTA11_08622_A.

(C) MxIF image of HTA11_08622_A. Scale bar 500 µm.

(D) MILWRM tissue domains for HTA11_08622_A, surrounded by top cell-state loadings for SSL (D0), normal epithelium (D5), and submucosa (D6) domains.

(E) Proportions of MILWRM domains detected in ST from each patient. Patient ID colors represent tumor class.

(F) refNMF states grouped by compartment and summarized across tumor stage, tumor class, MILWRM domain, and patient for all ST samples. Patient ID colors represent tumor class.

(G) Heatmap of GAM fits for refNMF states in all ST tumor regions ordered by PPT for HM (left) and CIN+ (right) tumors. Color represents scaled expression within each tumor class.

5.3.4 Cell-state deconvolution reveals pseudotemporal tissue dynamics

Using a previously published scRNA-seq dataset from 128 specimens consisting of CRCs, pre-malignant lesions, and adjacent normal mucosa (Chen et al., 2021b), we used non-negative matrix factorization (NMF; Kotliar et al., 2019; Figure S24A-B) to construct a consensus reference of 30 distinct cell states found in epithelial and stromal compartments and characterized by top-loaded genes and their associated pathway enrichment terms (Figure S24C-D). These cell states include normal mucosal cells such as tuft (TUF), enteroendocrine (EE1-2), and goblet (GOB), tumor-specific states derived from serrated lesions (SSC) and carcinomas (CRC1-4), and infiltrating immune populations including helper T cells (TL1), cytotoxic T cells (TL2), and neutrophils (MYE4), amongst others (Figure S24C-D).

To calculate cell-state contributions to each ST pixel, we used reference (ref) NMF to extract consensus states from ST expression matrices, effectively inferring fractional abundance of these states in each ST microwell (Appendix A5.1: refNMF cell-state discovery and deconvolution). Deconvolved cell-state fractions correlated with selected MxIF cell-type marker expression in registered serial tissue sections, as well as expression of literature-based cell-type and cell-activity gene signatures (Nirmal et al., 2018; Gulati et al., 2020; Chen et al., 2021b; Combes et al., 2022; Joanito et al., 2022; Barkley et al., 2022; Gil Vasquez et al., 2022; Figure 5.4A-C; Figure S24E-G; Appendix A5.1: refNMF validation). For example, GOB, ABS (absorptive colonocytes), and CT (crypt-top colonocyte) enrichment aligns with high MUC2 IF staining in normal epithelium, MYE1 (M1 macrophages) coincides with CD11B staining, and CRC2/STM-rich regions (CIN+ tumor cells and stem-like cells) express PCNA and OLFM4 in PAT30884 (Figure 5.4A). The SSL, HTA11_08622_A, further validates refNMF states such as ABS, FIB2, and serrated-specific cells (SSCs) by their spatial distribution in the mucosa and submucosa layers and correspondence to MUC5AC and AQP5 IF staining (Figure 5.4B-C).

Application of cell-cell interaction community reconstruction algorithms to ST data provided inaccurate results due to low spatial resolution and large distances between microwells. Focusing on pixel-based community detection, we employed refNMF usages as predictors for a MILWRM model to divide tissue into consensus domains based on cell-state makeup (Kaur et al., 2023; Figure 5.4D). The MILWRM model yielded eight domains (D0-D8) that correspond to CIN+ epithelium (D4 - high in CRC2 and STM), normal mucosa (D5 - enriched in ABS and CT), and sessile-serrated epithelium (D0 - high in SSC and GOB), amongst others, all of which spatially align with regional histology (Figure 5.4A-E; Figure S24H).

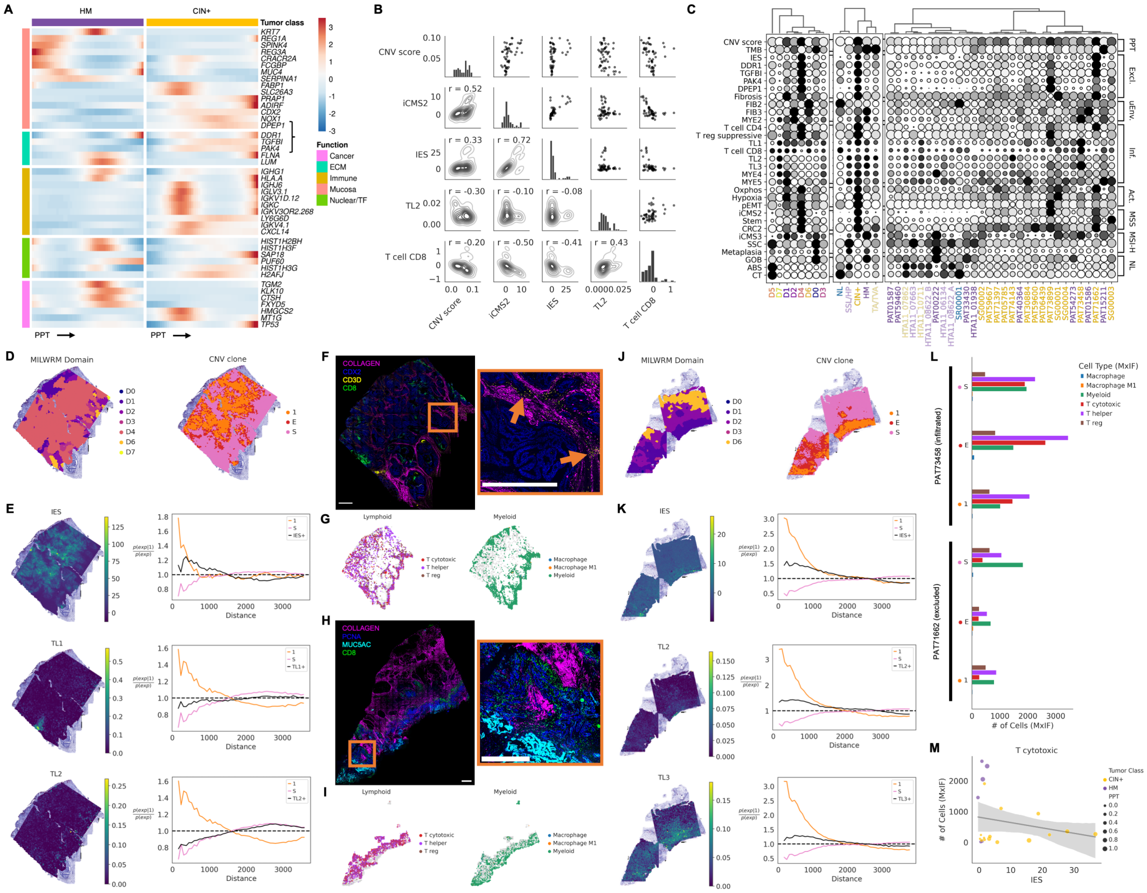
When summarizing refNMF abundances across MILWRM domains and tumor classes, we can identify epithelial cell states enriched by unique populations (CRC1 = HM; CRC2 = CIN+) and gain a coarse understanding of microenvironmental makeup of each tumor (TL2 [cytotoxic] = SSL/HP/HM; MYE4 [neutrophils] and MYE5 [DCs] = CIN+; Figure 5.4F). Finally, we note that MILWRM domain D1 is characteristic of HM, while D4 is characteristic of CIN+ tumors (Figure 5.4E). Applying these approaches to ST data enables validated mapping of consensus tumor and non-tumor cell populations across the atlas, and specifically provided the spatial distributions of low-abundance cell states such as infiltrating immune cells.

Using PPT ordering from CNV scores and TMB in CIN+ and HM tumors, respectively, we tracked changes in refNMF states across a global indicator of tumor progression and built generalized additive models (GAMs) that ascribe statistical significance to cell-state dynamics during CRC evolution (Figure 5.4G; Appendix A5.1: Modeling expression dynamics along PPT). We observed a replacement of normal cells (ABS, CT, STM, TUF, and EE) with carcinoma-specific epithelial states (CRC1-4) following the transition from pre-cancer to CRC in both the CIN+ and HM pathways. Along CIN+ PPT, STM-dominated early lesions give rise to CRC2, while HM tumors progress from metaplastic SSCs to mucinous CRC (GOB and CRC1), likely due to the respective cells-of-origin of CIN+ and HM tumors (Chen et al., 2021b). Consistent with prior knowledge, we observed an increase in most infiltrating immune populations with development of HM tumors, but a striking decrease in immune states in the tumor epithelium of CIN+ CRCs, coincident with the strong increase in CRC2 abundance (Figure 5.4G). This observation, in the context of global PPT, suggests that an epithelial-intrinsic program emerges in CIN+ CRCs to exclude or evade the host immune system.

5.3.5 Gene expression features of CIN+ CRCs predict immune exclusion

Immune exclusion correlates with poor patient outcomes in multiple cancer types and negatively predicts immune checkpoint inhibitor (ICI) response (Lazarus et al., 2018; Zhou et al., 2019; Gunnarsson et al., 2020; Lee et al., 2020; Bortolomeazzi et al., 2021; Pelka et al., 2021; Combes et al., 2022). Immune exclusion mechanisms provide potential therapeutic targets that may open the door for MSS CRCs to respond to ICI (Llosa et al., 2015; Lang et al., 2022). Biomarkers of immune exclusion may likewise act as predictors of the roughly 30 - 60 % of ICI non-responders in the MSI-H group that are currently immunotherapy candidates (Motta et al., 2021).

To further investigate the link between CIN+ tumor progression and immune exclusion, we extended GAM analysis to untargeted gene expression from ST atlas data. We identified genes with statistically significant dynamics within and between PPT trajectories for HM and CIN+ CRCs and grouped them by biological pathway or function (Figure 5.5A). Specifically, we identified genes that were significantly elevated in late CIN+ PPT, which formed a functional module for extracellular matrix (ECM) signaling and organization (*DDR1*, *TGFBI*, and *PAK4*). These genes originate from epithelial cells, which implicated a mechanism of microenvironmental modulation by the tumor itself. Each of these genes has also been shown to associate


A
HM
CIN+
Tumor class
KRT7
REG1A
SPINK4
REG3A
CRACR2A
FCGBP
MUC4
SERPINA1
FABP1
SLC26A3
PRAP1
ADIRF
CDX2
NOX1
DPEP1
DDR1
TGFBI
PAK4
FLNA
LUM
IGHG1
HLA.A
IGHJ6
IGLV3.1
IGKV1D.12
IGKC
IGKV3OR2.268
LY6G6D
IGKV4.1
CXCL14
HIST1H2BH
HIST1H3F
SAP18
PUF60
HIST1H3G
H2AFJ
TGM2
KLK10
CTSH
FXYD5
HMGCS2
MT1G
TP53
PPT
PPT
Function
Cancer
ECM
Immune
Mucosa
Nuclear/TF
B
CNV score
iCMS2
IES
TL2
T cell CD8
r = 0.52
r = 0.33
r = 0.72
r = -0.30
r = -0.10
r = -0.08
r = -0.20
r = -0.50
r = -0.41
r = 0.43
C
CNV score
TMB
IES
DDR1
TGFBI
PAK4
DPEP1
Fibrosis
FIB2
FIB3
MYE2
T cell CD4
T reg suppressive
TL1
T cell CD8
TL2
TL3
MYE4
MYE5
Oxphos
Hypoxia
pEMT
iCMS2
Stem
CRC2
iCMS3
SSC
Metaplasia
GOB
ABS
CT
PPT
Excl.
uEnv.
Inf.
Act.
MSS
MSI-H
NL
D
MILWRM Domain
CNV clone
E
IES
TL1
TL2
Distance
F
COLLAGEN
CD3D
CD8
G
Lymphoid
Myeloid
T cytotoxic
T helper
T reg
Macrophage
Macrophage M1
Myeloid
H
COLLAGEN
MUC5AC
CD8
I
Lymphoid
Myeloid
J
MILWRM Domain
CNV clone
K
IES
TL2
TL3
Distance
L
Cell Type (MxIF)
Macrophage
Macrophage M1
Myeloid
T cytotoxic
T helper
T reg
PAT73458 (infiltrated)
PAT71662 (excluded)
# of Cells (MxIF)
M
T cytotoxic
# of Cells (MxIF)
IES
Tumor Class
CIN+
HM
PPT

with immune exclusion in several solid tumor types, providing an interesting, concerted signature that potentially creates an immune-tolerant environment in CIN+ CRC (Ween et al., 2012; Abril-Rodriguez et al., 2019; Sun et al., 2021; Duan et al., 2022).

Additionally, we observed a significant increase in *DPEP1* expression with CIN+ PPT, a gene which was also highly enriched in the CRC2 epithelial state (Figure 5.5A; Figure 5.4H; Figure S24C-D). We added *DPEP1* to the signature with *DDR1*, *TGFBI*, and *PAK4* for analysis of immune exclusion in this atlas as it is implicated in neutrophil recruitment, which could lead to lymphocyte exclusion in the TME (Germann et al., 2020; Wang, 2022), and has been shown to be secreted by tumor cells in extracellular vesicles, thus offering a promising circulating biomarker in CRC patients (Zhang et al., 2021).

These four genes proved to be highly co-expressed by tumor epithelium (Figure S25A-B), and when combined into an Immune Exclusion Signature (IES), exhibited spatial correlation with iCMS2 and CNV score, and were enriched in the CIN+-specific MILWRM domain D4, corroborating prior studies (Roh et al., 2017; Luke et al., 2019; Joanito et al., 2022; Figure 5.5B-C). Surveying other gene signatures, we identified additional programs that coincided with IES such as pEMT and oxidative metabolism, which are hallmarks of advanced cancer and upregulated during tumor progression, validating the emergence of IES in late PPT (Fig-

Figure 5.5 *(preceding page)*: Gene expression features of CIN+ CRCs predict immune exclusion.
(A) Heatmap of GAM fits for top genes summarized across all ST tumor regions. Color represents scaled expression within each tumor class. Genes are grouped by biological function. Bracket denotes IES genes.
(B) Pairwise Pearson correlations between progression indicators (CNV score and iCMS2), IES, cytotoxic T cell refNMF state (TL2), and CD8 T cell gene signature in all CIN+ tumor regions.
(C) Genes, gene signatures, and refNMF states grouped into pseudotime indicators (”PPT”), immune exclusion markers (”Excl.”), microenvironmental cells (”uEnv.”), infiltrating immune cells (”Inf.”), tumor activity (”Act.”), and epithelial-specific markers of MSS, MSI-H, and normal mucosa summarized by MILWRM domain, tumor class, and patient for all ST samples. Patient ID colors represent tumor class.
(D) PAT71662 ST with annotated tissue domains from MILWRM (left) and CNV clone regions (right).
(E) Expression overlay and spatial co-occurrence analysis for IES, helper T cells (TL1), and cytotoxic T cells (TL2) in PAT71662. Line plots at right indicate the conditional probability of high signature or cell-state expression as a function of distance from CNV clone 1 microwells (Appendix A5.1: Spatial co-occurrence analysis from ST).
(F) PAT71662 MxIF showing collagen, CDX2 (marking MSS epithelium), and lymphocytes (CD3 and CD8). Inset highlights CD3/CD8+ cells sequestered to stroma. Scale bars 500 µm.
(G) Centroids of segmented single cells from PAT71662 MxIF plotted in whole-slide space, split into lymphoid and myeloid compartments.
(H) Same as F for PAT73458. PCNA and MUC5AC mark tumor epithelium. Inset highlights CD8+ cells invading epithelium. Scale bars 500 µm.
(I) Same as in G, for PAT73458.
(J-K) Same as in D-E, for PAT73458. TL3 represents $\gamma\delta$IELs.
(L) Census of infiltrating immune cells in PAT71662 (bottom) and PAT73458 (top) from G and I summarized by CNV clone region (Appendix A5.1: MxIF immune-exclusion analysis).
(M) Number of infiltrating CD8+ T cells detected in MxIF plotted against IES score for all tumor regions. Points are colored by tumor class and sized according to PPT ranking.

ure 5.5C; Figure S25B). Moreover, fibrosis and hypoxia signatures correlated with IES at the patient-level, lending credence to the involvement of constituent genes in ECM signaling and regulation as these characteristics are implicated in microenvironmental immunosuppression (Ween et al., 2012; Abril-Rodriguez et al., 2019; Sun et al., 2021; Duan et al., 2022; Zeng et al., 2022; Figure 5.5C; Figure S25D). Most importantly, this four-gene IES had a negative correlation with cytotoxic T cells (TL2) and with the CD8 T cell signature score across all tumor regions in the dataset, indicating its value as a predictor of immune exclusion (Combes et al., 2022; Figure 5.5B).

An example from this atlas of immune-excluded CRC is PAT71662, which contains a single major CNV clone region that delineates the epithelial compartment in ST (clone region ”1”; Figure 5.5D). We note that the majority of the area of this MSS tumor belongs to MILWRM D4, and is correspondingly high in CRC2, stem, fibrosis, and oxidative metabolism (Figure 5.5C-D). Spatial co-occurrence analysis revealed that this tumor exhibited enrichment of the epithelial-intrinsic IES in the tumor core, while excluding T lymphocytes from the CNV clone 1 region (Figure 5.5E). We confirmed the exclusion of CD8+ cells using spatially registered MxIF data, demonstrating how immune infiltrates are sequestered to the collagen-rich stroma (Figure 5.5F-G). Conversely, immune-infiltrated tumors such as PAT73458 exhibited much lower epithelial expression of IES while clear infiltration of cytotoxic (TL2) and $\gamma\delta$ (TL3) T cells is seen in ST and MxIF data (Figure 5.5C,H-K).

We next expanded immune-exclusion analysis using MxIF data to validate ST findings. Following single-cell segmentation (McKinley et al., 2022), we identified immune cell subsets by the presence of various marker protein stains quantified in each cell. We then masked the MxIF slides with CNV clone regions so we could enumerate the abundance and distribution of infiltrating immune cells in each major tumor compartment (Appendix A5.1: MxIF immune-exclusion analysis). We note that immune-excluded PAT71662 has more immune cells in its stroma (”S”) than tumor epithelium (”E” and ”1”), while the infiltrated tumor, PAT73458, exhibits not only higher levels of cytotoxic (CD3+/CD8+) and helper (CD3+/CD4+) T cells overall, but also a larger proportion of all infiltrating immune cells in the tumor edge and tumor core regions (Figure 5.5F-I,L). Extending these analyses to the entire spatial atlas, we observe a negative correlation between IES and T cells (CD8+ and CD4+) in tumor clone regions as identified by high-resolution MxIF imaging (Figure 5.5M; Figure S25C).

In HM/iCMS3 tumors with low IES, we observed a corresponding increase in other microenvironmental cell states that have been implicated in immune tolerant microenvironments, suggesting that HM CRCs follow an alternative path to immune evasion and potential ICI non-response compared with CIN+ CRCs. Cancer-associated fibroblasts (CAFs) expressing *FAP* and *CXCL12* (FIB3), and *SPP1*+ myeloid cells (MYE2) have been shown to foster an immunosuppressive niche in CRC and other solid tumors such as pancreatic cancer

(Feig et al., 2013; Calon et al., 2015; Qi et al., 2022; Figure 5.5C; Figure S25E). Moreover, *CXCL14*+ CAFs (FIB2) are suspected to counteract the immune-silencing effect of *CXCL12*+/*FAP*+ fibroblasts and are more prominent in MSI-H/iCMS3/HM tumors (Pelka et al., 2021). Since these three microenvironmental cell states emerge in immune infiltrated tumors (Figure 5.5C) and seem to trend positively with regional HM progression (Figure 5.6A), we hypothesize that this emergent struggle between immune evasive and immunogenic TMEs represents intracellular signaling complexity in late-stage CRC, and that tipping the scales of such interactions may explain the subset of MMR-deficient HM tumors that do not respond to ICI therapy (Qi et al., 2022).

5.3.6 IES trends with tumor progression and predicts poor patient outcomes

To further explore the link between tumor progression and immune exclusion, we modeled scoring of CIN+ epithelial-intrinsic IES, HM-enriched FIB2 and FIB3 CAFs and MYE2 macrophages, as well as infiltrating immune cell states and cancer progression signatures against PPT for all tumor regions across the atlas, confirming that IES is enriched in CIN+ tumors and trends proportionally to CNV-informed PPT, iCMS2 and CRC2 enrichment (Figure 5.6A; Appendix A5.1: Modeling expression dynamics along PPT).

We next set out to validate this expression signature as a predictor of immune exclusion by highlighting its translational utility in predicting patient outcomes in external cohorts with larger sample sizes. The IES score was significantly enriched in MSS (CIN+) vs. MSI-H (HM) TCGA COAD and READ samples (p = 6.04×10^{-8}; Figure 5.6B; Figure S26A), consistent with the distinct immunosuppressive mechanisms between the two tumor subtypes. High IES expression yielded a statistically significant drop in progression-free survival (PFS) for patients with high-scoring tumors compared to those with low-scoring tumors from the entire TCGA cohort (p = 0.015; Figure 5.6C) as well as the subset of MSS tumors in TCGA (p = 0.035; Figure S26B). Furthermore, only *TGFBI* exhibited a similar reduction in PFS by itself (Figure S26C-D), indicating that the epithelial-intrinsic IES score has prognostic value in identifying immune-cold CRCs with poor disease-free survival, and that the aggregate signature is more informative than the sum of its parts.

Finally, to increase the translational value of this signature, we investigated whether protein expression corroborated results obtained by gene expression in CRC such that immunohistochemistry (IHC) can be reliably used to predict immune exclusion. From IHC staining of a tumor microarray (TMA) consisting of 163 colorectal adenocarcinoma samples, we observed higher overall DDR1 and TGFBI protein expression in MSS versus MSI-H tumors, consistent with findings above (Figure 5.6D-F). Survival analysis on this TMA IHC cohort revealed significantly lower PFS (p = 0.00034; Figure 5.6G) and overall survival (OS; p = 0.0011; Figure S26E) for tumors with high IHC staining for both DDR1 and TGFBI, corroborating our TCGA query. Once more, this combination of IES markers yielded a statistically greater stratification of patient survival than the individual proteins (Figure S26F-G), confirming the utility of this expression

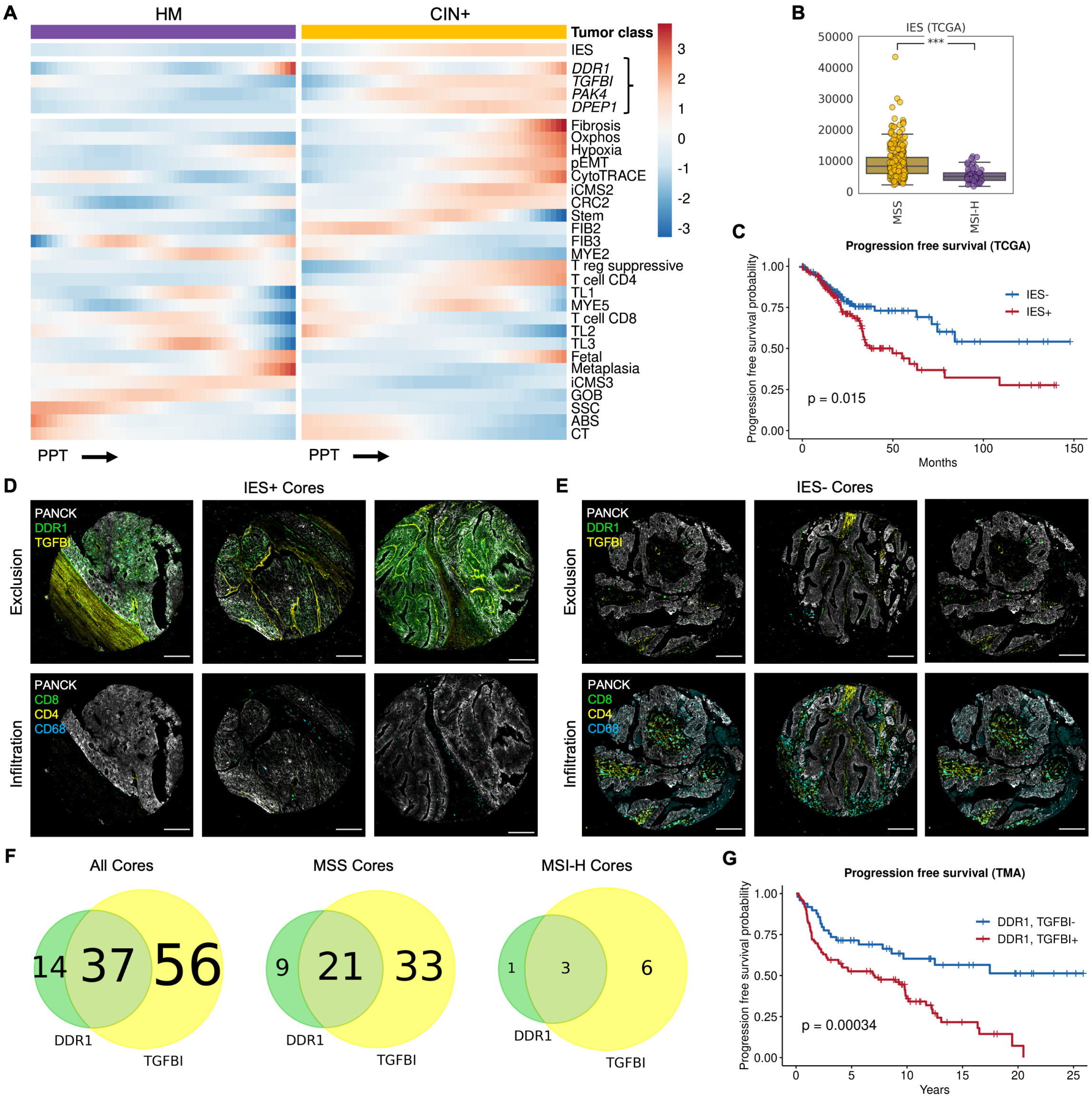


Figure 5.6: IES trends with tumor progression and predicts poor patient outcomes.
(A) Heatmap of GAM fits for genes, gene signatures, and refNMF cell states summarized across all ST tumor regions. Color represents scaled expression within each tumor class. Bracket denotes constituent genes in IES.
(B) Boxplots of IES scores in TCGA COAD and READ samples, stratified by MMR status (MSS n = 301; MSI-H n = 45). Student's T-test with Bonferroni correction yielded p = 6.04×10^{-8}
(C) Kaplan-Meier PFS curves for TCGA COAD and READ samples from B with high (+) and low (-) IES scores.
(D) MxIF images showing epithelial (top) and immune (bottom) markers from representative IES+ cores from the CRC TMA. Scale bars 100 µm.
(E) Same as in D, for representative IES- cores.
(F) Venn diagram of the number of CRC TMA cores with high-scoring DDR1 and TGFBI IHC staining (left; total n = 163), and stratified by MMR status (middle, n = 86; right, n = 22).
(G) Kaplan-Meier PFS curves for CRC TMA cores with high (+) and low (-) IHC staining of both DDR1 and TGFBI.

signature as a prognostic indicator in both mRNA and protein assays.

5.4 Discussion

Evolution of spatially resolved tumor clones in human CRC presents an opportunity to use spatial technologies to investigate cellular and microenvironmental heterogeneity along a trajectory of cancer progression. Whereas the probability of any particular polyp progressing to malignancy is inherently low (Conteduca et al., 2013; Colom et al., 2021), this spatial strategy of mapping evolution in pre-selected carcinomas with less advanced components offers a glimpse into the biology of malignant transitions that can inform diagnostic stratification and early intervention (Sottoriva et al., 2015; Ryser et al., 2018; Shibata, 2020; Househam et al., 2022; Lomakin et al., 2022). The phylogeographical atlas presented herein comprises spatially resolved genomic, transcriptomic, and protein profiling of 31 patients spanning normal mucosa, pre-cancerous lesions, and invasive adenocarcinoma. Spatial assays in each of these molecular domains recapitulated findings related to TA/TVA and SSL/HP progression to MSS and MSI-H CRCs from initial *WNT* activation and gastric metaplasia, respectively, to more advanced iCMS states (Chen et al., 2021b; Joanito et al., 2022).

Using CNV and LCM-WES analysis, we classified tumors as CIN+ or HM, and confirmed that the latter group displays a cytotoxic immune microenvironment while the former are more likely to be immune-cold (Guinney et al., 2015; Nouri Nojadeh et al., 2018; Luke et al., 2019. While previous work documented a mechanism by which *APC* loss-of-function leads to aneuploidy and chromosomal instability (Tighe et al., 2001; Alberici et al., 2007; Rusan and Peifer, 2008, our multi-modal CNV analysis supports that most CIN occurs as a late-onset characteristic within the adenoma-carcinoma sequence, likely following *TP53* loss and transition to malignancy (Sigurdsson et al., 2000; Dalton et al., 2010; Foijer et al., 2014; Bronder et al., 2021). In this regard, some MSI-H tumors (SG00001, SG00002, PAT73458) gain CIN through *APC*-independent mechanisms. Likewise, some MSS cases display an HM phenotype (HTA11_01938, PAT15211), likely driven by mutations in proofreading polymerases such as *POLE* and *POLD1* (Bourdais et al., 2017). Phenotypic manifestations are more consistent with CIN+ and HM classifications than microsatellite status, exemplified by MSI-H/CIN+ tumors transitioning to a stem-like, iCMS2 state (Chen et al., 2021b).

Multiregional somatic mutational profiling matched CNV phylogeography in CIN+ CRCs, provided high-confidence PPT for HM tumors, and allowed us to stratify patients based on evolutionary dynamics. In this atlas, we observed CRCs that underwent neutral or hierarchical evolution, exhibiting high degrees of regional heterogeneity, as well as tumors that displayed linear or punctuated evolution with many public driver mutations and relatively low clonal divergence (Sottoriva et al., 2015; Turajlic et al., 2019). We found that neutral and branching evolution dominated CIN+ tumors, while the majority of HM CRCs exhibited linear or punctuated evolution. Although our dataset lacks sufficient sample size to draw confident conclusions, we speculate that this dichotomy of evolutionary dynamics stems from the discrepancy between somatic mutation rate in HM CRCs and genomic structural alterations due to CIN, resulting in large, stepwise increases in clonal fitness and smaller, more frequent subclonal branching, respectively (Grist et al., 1992; Jackson and Loeb, 1998; De Nooij-van Dalen et al., 2001; Nowak et al., 2002). Furthermore, tumors driven by distinct mutational processes may recruit unique TMEs that further restrict clonal evolution, such as distinct immune cells or CAFs that suppress or promote tumor progression (Gerlinger et al., 2012; Calon et al., 2015; Sun et al., 2017; Lazarus et al., 2018). Nevertheless, observed evolutionary dynamics in this atlas serve to characterize overall tumor heterogeneity and enable pseudotemporal placement along a global progression trajectory.

Generalized additive modeling (GAM) along global PPT elucidated tumor and microenvironmental programs such as altered metabolism that drive or result from tumor progression (Brahimi-Horn et al., 2007; Ashton et al., 2018). Of specific interest are gene programs that modulate tumor-microenvironment interactions. We focused on an immune-exclusion mechanism specific to CIN+ tumors with an iCMS2/stem-like epithelial phenotype marked by the CRC2 refNMF state. Previous work from our group and others has demonstrated a transition into a stem-like tumor state as a function of progression, resulting in immunosuppression (Becker et al., 2021; Chen et al., 2021b). PPT models enabled the unbiased identification of gene programs relevant to tissue dynamics along tumor development, and allowed us to assemble an epithelial-intrinsic Immune Exclusion Signature (IES) defined by expression of *DDR1*, *TGFBI*, *PAK4*, and *DPEP1*. This aggregated signature is highly informative, as these genes have been implicated separately as immune modulators but not always in the context of CRC (Abril-Rodriguez et al., 2019; Lecker et al., 2021; Chen et al., 2021c; Duan et al., 2022).

DDR1 has been reported to align collagen fibers in a way that excludes T cells from the breast TME (Sun et al., 2021). These findings, coupled with *DDR1*'s role in fibrotic kidney disease and breast cancer, validate the observation that IES coincides with fibrosis and hypoxia in advanced CIN+ CRCs and suggest that microenvironmental signaling and collagen remodeling in the ECM are involved directly in immune evasion (Takai et al., 2018; Borza et al., 2022). *TGFBI* is associated with an immunosuppressive microenvironment in ovarian cancer where it is released from macrophages, but mechanistic details are lacking (Steitz et al., 2020). *DPEP1* is an endothelial adhesion receptor for neutrophils and monocytes during inflammation, and while a direct role for *DPEP1* in CRC immune evasion has not been reported, it has been shown that neutrophil infiltration can lead to T cell exclusion in the colon (Germann et al., 2020; Wang, 2022). We note that neutrophil infiltration increased with PPT in our dataset, highlighting the importance of *DPEP1* in CRC progression.

A confluence of these immunomodulatory processes has also been observed in a recent tumor microbiome study, where advanced, microbially-infiltrated CRC regions are aneuploid, hypoxic, and immunosuppressive in a neutrophil-dependent manner (Galeano Niño et al., 2022). Altogether, these molecules play distinct yet complementary roles in immune exclusion that will be the focus of future studies.

HM tumors, on the other hand, exhibited a stromal dichotomy of immune surveillance escape driven by *CXCL12*+ CAFs and *SPP1*+ myeloid cells (Qi et al., 2022), and *CXCL14*+ iCAFs that have been shown to be MSI-H-specific and immunogenic (Pelka et al., 2021). Importantly, we note that the prevalence of these tumor-immune interaction networks trends directly with HM PPT, suggesting a link to CRC development analogous to IES in the CIN+ cohort.

The immune exclusion biomarkers identified herein carry several clinically relevant implications for detection, prognosis, and potential treatment of CRC. IES genes and their protein products in aggregate offer a robust prognostic indicator of progression-free survival in CRC. Additionally, this expression signature may stratify patients by ICI response potential: identifying ICI responders in the CIN+/MSS cohort as well as ICI non-responders in the HM/MSI-H group. Moreover, these biomarkers may act as potential targets for adjuvant therapy to ICI, depending on tumor stage and CIN status, as the breakdown of immune evasive mechanisms could expose tumors to more effective immunotherapy.

Here we present a rich, spatially resolved dataset comprised of genetic, transcriptomic, and proteomic layers of molecular information. Integrated analyses herein aimed to map recurring cell states and signatures across patient-specific phylogeographical landscapes to uncover global pseudotemporal dynamics of tumor progression from a snapshot in time. However, there is yet much more to be learned regarding CRC progression, and these data will therefore prove to be a valuable resource in the further characterization of tumor-microenvironmental co-evolution.

CHAPTER 6

Discussion and future directions

6.1 Contributions to the computational biology ecosystem through open-source tool development

6.1.1 Understanding high-dimensional data structure for informing dimension reduction techniques

My early work addressed several technical issues in the field of computational systems biology from a tool-development perspective. First, we presented a comprehensive and unbiased framework that defines metrics of global and local structure preservation in dimensionality reduction transformations (Heiser and Lau, 2020; Chapter 2). We found that the input cell distribution and dynamic range of the underlying measurements are paramount to the preservation of data structure, which can greatly impact downstream clustering and trajectory reconstruction (Herring et al., 2018; Traag et al., 2019; Van den Berge et al., 2020). Moreover, we offered a valuable framework and code base for unbiased comparison and selection of appropriate methods and parameters for specific datasets, data types, and applications, informing the development of single-cell data processing and visualization pipelines.

6.1.2 Quality control and cell optimization in scRNA-seq using linear models

Next, we built dropkick, a fully automated software tool for quality control and filtering of scRNA-seq data with a focus on excluding ambient barcodes and recovering real cells bordering the quality threshold (Heiser et al., 2021; Chapter 3). We defined quality control metrics for evaluating batch-specific background, which is imperative to the integration of large, atlas-style datasets (Regev et al., 2017; Rozenblatt-Rosen et al., 2020). Using simulated and real-world data, we benchmarked dropkick against conventional thresholding approaches and EmptyDrops, a popular computational method, demonstrating greater recovery of rare cell types and exclusion of empty droplets and noisy, uninformative barcodes. We showed that for both low and high-background datasets that dropkick's weakly supervised model reliably learns which genes are enriched in ambient barcodes and draws a multidimensional boundary that is more robust to dataset-specific variation than existing filtering approaches. Moreover, dropkick is computationally efficient and compatible with popular single-cell Python packages (Wolf et al., 2018, 2019). As such, dropkick has the potential to become a valuable tool for researchers working with scRNA-seq data, can facilitate more accurate and reliable downstream analysis through automated, reproducible, and unbiased preprocessing, and has gained traction in the field as a preprocessing and QC method.

6.1.3 Tissue domain detection in spatial multi-omics for digital pathology

Most recently, we shifted our method development efforts from single-cell to spatial technologies. We developed MILWRM – multiplex image labeling with regional morphology – a Python package for rapid, multi-scale tissue domain detection and annotation (Kaur et al., 2023; Chapter 4). Here, we addressed the challenge of data-driven cross-sample domain detection, and demonstrated the utility of MILWRM in identifying histologically distinct tissue compartments through an untargeted clustering model. We were able to exhibit MILWRM modeling on two distinct data modalities and showcase its utility in the analysis of high volumes of multiplex spatial datasets from tissue atlasing efforts (Chen et al., 2021b). The ability to identify distinct molecular profiles of different tissue domains *in situ* is essential for understanding tissue organization and pathology. MILWRM has demonstrated utility in identifying histologically distinct compartments in human colonic tumor data (Heiser et al., 2023; Chapter 5), and will undoubtedly prove to be a valuable resource in future spatial -omics projects.

6.1.4 Open-source code ecosystems enable computational biology

The importance of building, validating, and sharing code-based tools and packages for biomolecular data analysis and visualization cannot be overstated. When consensus methods and pipelines for reproducible and robust analyses of unique data modalities and applications are built and distributed via open-source programming ecosystems, the effects on the field of biology as a whole are immeasurable (Wolf et al., 2018; Chen et al., 2021a; Schapiro et al., 2022). Collaboration and version-control tools such as Git and their associated online platforms (e.g. GitHub: https://github.com/) allow computational biologists to share and modify code at scale, providing a powerful resource for modular and iterative improvement upon prior work. Indeed, tool development and distribution will continue to be imperative to the advancement of systems biology when coupled with the emergence of new molecular technologies.

6.2 Phylogeographical cartography reveals immune exclusion mechanisms in CRC

Colorectal cancer exhibits dynamic cellular and genetic heterogeneity during progression from precursor lesions toward malignancy. Leveraging spatial molecular information to construct a phylogeographic map of tumor evolution can reveal individualized growth trajectories with diagnostic and therapeutic potential. Integrative analysis of spatial multi-omic data from 31 colorectal specimens revealed simultaneous microenvironmental and clonal alterations as a function of progression (Heiser et al., 2023; Chapter 5). Copy number variation served to re-stratify microsatellite stable and unstable tumors into chromosomally unstable (CIN+) and hypermutated (HM) classes. Phylogeographical maps classified tumors by their evolutionary dynamics, and clonal regions were placed along a global pseudotemporal progression trajectory from normal tissue to invasive adenocarcinoma.

Cell-state discovery from a single-cell cohort revealed recurring epithelial gene signatures and infiltrating immune states in spatially resolved transcriptomics. Charting these states along progression pseudotime, we observed a transition to immune exclusion in CIN+ tumors as characterized by a novel gene expression signature comprised of *DDR1*, *TGFBI*, *PAK4*, and *DPEP1* (IES). We demonstrated how these genes and their protein products are key regulators of extracellular matrix components, are associated with lower cytotoxic immune infiltration, and show prognostic value in external cohorts (Figure 6.1A; Figure 5.6; Figure S26). Alternatively, HM tumors exhibited an increase in stromal cell states that have been implicated in immune tolerant microenvironments, indicating that HM CRCs have an alternative path to immune exclusion and ICI non-response compared with CIN+ tumors (Figure 6.1B). *FAP*+ and *CXCL12*+ cancer-associated fibroblasts (CAFs), along with *SPP1*+ myeloid cells have been shown to foster an immunosuppressive niche in CRC and other solid tumors such as pancreatic cancer (Feig et al., 2013; Calon et al., 2015; Pelka et al., 2021; Qi et al., 2022). Importantly, we note that the prevalence of these stromal cell states trends directly with HM PPT in our atlas, suggesting a link to CRC development analogous to IES in the CIN+ cohort. Thus, we hypothesize that colorectal tumors developing along CIN+ and HM pathways have distinct mechanisms of immune exclusion and evasion at late stages.

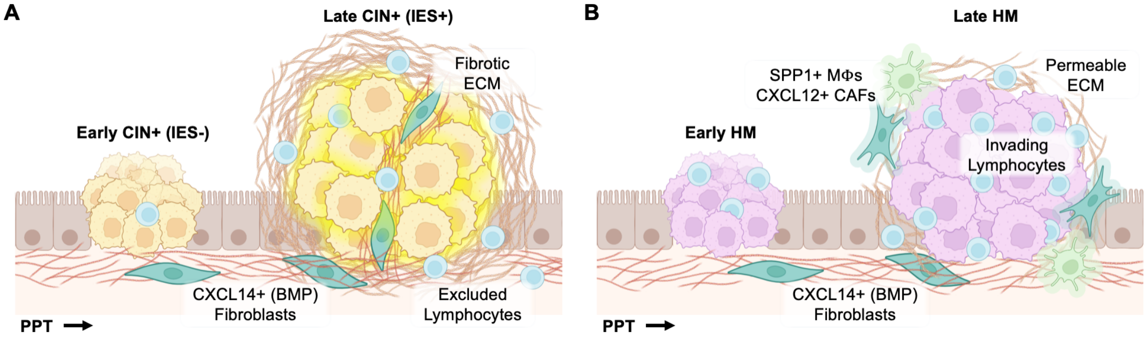


Figure 6.1: Major classes of CRC exhibit unique paths to immune exclusion.
(A) Diagram of early-to-late CIN+ CRC development, highlighting proposed IES mechanism.
(B) Diagram of early-to-late HM CRC.

6.2.1 Open questions around IES and development of clinical biomarkers

Future directions include functional validation studies to determine the mechanisms underlying this immune exclusion phenotype in CRC, with potential extension to other solid tumors. For instance, how is the iCMS2/stem-like epithelial phenotype of CIN+ CRC linked to IES, through epigenetics or cell signaling (Roh et al., 2017; Luke et al., 2019; Pinyol et al., 2019)? What is the mechanism of *DDR1*, *TGFBI*, and *PAK4* in ECM remodeling that renders tumors immune-excluded (Tumbarello et al., 2012; Lee et al., 2019;

Abril-Rodriguez et al., 2019; Sun et al., 2021; Chen et al., 2021c; Duan et al., 2022; Su et al., 2022)? What other factors might influence immunogenicity or immune evasion in late-stage cancer, and how might they be linked to a metastatic niche (Tu et al., 2019; Nikolos et al., 2022; Krishnamurty et al., 2022)?

There is also a large amount of interest in the development of these biomarkers into diagnostic tools for translation to the clinic. As an example, can DPEP1 or TGFBI protein released by CRCs in extracellular vesicles (EVs) be detected in the plasma as a liquid biopsy or treatment monitoring device (Zhang et al., 2021)? Finally, IES and its constituent markers suggest possible therapeutic targets for CRC and other tumors prone to immune exclusion. Once mechanism is more clearly defined, perhaps drugs can be designed to neutralize the barriers to cytotoxic immunity, enhancing survival and allowing for more effective immunotherapy (Angelova et al., 2015; Luoma et al., 2020; Bortolomeazzi et al., 2021; Shi et al., 2022).

6.2.2 Atlas resource for systems-level investigation and impact on precision medicine

These findings carry clinically relevant implications for detection, prognosis, and targeted treatment of CRC. The identified gene expression signature (IES) in CIN+ tumors and immunosuppressive cell states in HM CRCs may be used to stratify patients by ICI response potential: identifying ICI responders in the CIN+/MSS cohort as well as ICI non-responders in the HM/MSI-H group. Moreover, these biomarkers offer potential targets for adjuvant cancer therapy, as the breakdown of immune evasive mechanisms could expose tumors to more effective immuno- or chemotherapy.

These studies have significant impact on the fields of computational systems biology and precision oncology by demonstrating the potential of spatial multi-omic data integration to construct phylogeographic maps of tumor-microenvironmental co-evolution. In this work, we provide new insights into the tissue dynamics underlying tumor progression, and demonstrate the utility of atlas-style analyses in identifying potential targets for oncology. Overall, these studies highlight the capacity of computational systems biology approaches to drive precision medicine, and serve as an important resource for future research in this area.

References

10x Genomics (2022a). Adult mouse brain section 1 (coronal). stains: Dapi, anti-neun - 10x genomics.

10x Genomics (2022b). Adult mouse brain section 2 (coronal). stains: Dapi, anti-gfap, anti-neun - 10x genomics.

10x Genomics (2022c). Mouse brain section (coronal) - 10x genomics.

10x Genomics (2022d). Mouse brain serial section 1 (sagittal-anterior) - 10x genomics.

10x Genomics (2022e). Mouse brain serial section 1 (sagittal-posterior) - 10x genomics.

10x Genomics (2022f). Mouse brain serial section 2 (sagittal-anterior) - 10x genomics.

10x Genomics (2022g). Mouse brain serial section 2 (sagittal-posterior) - 10x genomics.

Abril-Rodriguez, G., Torrejon, D. Y., Liu, W., Zaretsky, J. M., Nowicki, T. S., Tsoi, J., Puig-Saus, C., Baselga-Carretero, I., Medina, E., Quist, M. J., Garcia, A. J., Senapedis, W., Baloglu, E., Kalbasi, A., Cheung-Lau, G., Berent-Maoz, B., Comin-Anduix, B., Hu-Lieskovan, S., Wang, C. Y., Grasso, C. S., and Ribas, A. (2019). PAK4 inhibition improves PD-1 blockade immunotherapy. *Nature Cancer*, 1(1):46–58.

Aizarani, N., Saviano, A., Sagar, Mailly, L., Durand, S., Pessaux, P., Baumert, T. F., and Grün, D. (2019). A Human Liver Cell Atlas: Revealing Cell Type Heterogeneity and Adult Liver Progenitors by Single-Cell RNA-sequencing. *bioRxiv*, page 649194.

Alberici, P., De Pater, E., Cardoso, J., Bevelander, M., Molenaar, L., Jonkers, J., and Fodde, R. (2007). Aneuploidy Arises at Early Stages of Apc-Driven Intestinal Tumorigenesis and Pinpoints Conserved Chromosomal Loci of Allelic Imbalance between Mouse and Human. *The American Journal of Pathology*, 170(1):377–387.

Alexandrov, T. and Kobarg, J. H. (2011). Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, 27(13):i230–i238.

Allen, A., Hutton, D. A., and Pearson, J. P. (1998). The MUC2 gene product: a human intestinal mucin. *The International Journal of Biochemistry & Cell Biology*, 30(7):797–801.

Amitay, Y., Bussi, Y., Feinstein, B., Bagon, S., Milo, I., and Keren, L. (2022). CellSighter – A neural network to classify cells in highly multiplexed images. *bioRxiv*, page 2022.11.07.515441.

Andersson, A., Bergenstråhle, J., Asp, M., Bergenstråhle, L., Jurek, A., Fernández Navarro, J., and Lundeberg, J. (2020). Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology*, 3(1):565.

Andersson, A., Larsson, L., Stenbeck, L., Salmén, F., Ehinger, A., Wu, S. Z., Al-Eryani, G., Roden, D., Swarbrick, A., Borg, Å., Frisén, J., Engblom, C., and Lundeberg, J. (2021). Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12(1):1–14.

Angelova, M., Charoentong, P., Hackl, H., Fischer, M. L., Snajder, R., Krogsdam, A. M., Waldner, M. J., Bindea, G., Mlecnik, B., Galon, J., and Trajanoski, Z. (2015). Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biology*, 16(1):64.

Aoki, R., Shoshkes-Carmel, M., Gao, N., Shin, S., May, C. L., Golson, M. L., Zahm, A. M., Ray, M., Wiser, C. L., Wright, C. V., and Kaestner, K. H. (2016). Foxl1-Expressing Mesenchymal Cells Constitute the Intestinal Stem Cell Niche. *Cellular and Molecular Gastroenterology and Hepatology*, 2(2):175–188.

Ashton, T. M., Gillies McKenna, W., Kunz-Schughart, L. A., and Higgins, G. S. (2018). Oxidative phosphorylation as an emerging target in cancer therapy. *Clinical Cancer Research*, 24(11):2482–2490.

Baldominos, P., Barbera-Mourelle, A., Barreiro, O., Huang, Y., Wight, A., Cho, J.-W., Zhao, X., Estivill, G., Adam, I., Sanchez, X., McCarthy, S., Schaller, J., Khan, Z., Ruzo, A., Pastorello, R., Richardson, E. T., Dillon, D., Montero-Llopis, P., Barroso-Sousa, R., Forman, J., Shukla, S. A., Tolaney, S. M., Mittendorf, E. A., von Andrian, U. H., Wucherpfennig, K. W., Hemberg, M., and Agudo, J. (2022). Quiescent cancer cells resist T cell attack by forming an immunosuppressive niche. *Cell*.

Banerjee, A., Herring, C. A., Chen, B., Kim, H., Simmons, A. J., Southard-Smith, A. N., Allaman, M. M., White, J. R., Macedonia, M. C., Mckinley, E. T., Ramirez-Solano, M. A., Scoville, E. A., Liu, Q., Wilson, K. T., Coffey, R. J., Washington, M. K., Goettel, J. A., and Lau, K. S. (2020). Succinate Produced by Intestinal Microbes Promotes Specification of Tuft Cells to Suppress Ileal Inflammation. *Gastroenterology*, 159(6):2101–2115.e5.

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M., and Hamilton, P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):16878.

Bao, S., Chiron, S., Tang, Y., Heiser, C. N., Southard-Smith, A. N., Lee, H. H., Ramirez, M. A., Huo, Y., Washington, M. K., Scoville, E. A., Roland, J. T., Liu, Q., Lau, K. S., Wilson, K. T., Coburn, L. A., and Landman, B. A. (2021). A cross-platform informatics system for the Gut Cell Atlas: integrating clinical, anatomical and histological data. In Park, B. J. and Deserno, T. M., editors, *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*, volume 11601, page 5. SPIE.

Barkley, D., Moncada, R., Pour, M., Liberman, D. A., Dryg, I., Werba, G., Wang, W., Baron, M., Rao, A., Xia, B., França, G. S., Weil, A., Delair, D. F., Hajdu, C., Lund, A. W., Osman, I., and Yanai, I. (2022). Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nature Genetics*, 54(8):1192–1201.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*.

Becker, W. R., Nevins, S. A., Chen, D. C., Chiu, R., Horning, A., Laquindanum, R., Mills, M., Chaib, H., Ladabaum, U., Longacre, T., Shen, J., Esplin, E. D., Kundaje, A., Ford, J. M., Curtis, C., Snyder, M. P., and Greenleaf, W. J. (2021). Single-cell analyses reveal a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *bioRxiv*, page 2021.03.24.436532.

Becker, W. R., Nevins, S. A., Chen, D. C., Chiu, R., Horning, A. M., Guha, T. K., Laquindanum, R., Mills, M., Chaib, H., Ladabaum, U., Longacre, T., Shen, J., Esplin, E. D., Kundaje, A., Ford, J. M., Curtis, C., Snyder, M. P., and Greenleaf, W. J. (2022). Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *Nature Genetics*, 54(7):985–995.

Beger, R. D., Dunn, W., Schmidt, M. A., Gross, S. S., Kirwan, J. A., Cascante, M., Brennan, L., Wishart, D. S., Oresic, M., Hankemeier, T., Broadhurst, D. I., Lane, A. N., Suhre, K., Kastenmï¿½ller, G., Sumner, S. J., Thiele, I., Fiehn, O., Kaddurah-Daouk, R., and for "Precision Medicine (2016). Metabolomics enables precision medicine: "A White Paper, Community Perspective". *Metabolomics*, 12(10).

Belkina, A. C., Ciccolella, C. O., Anno, R., Spidlen, J., Halpert, R., and Snyder-Cappione, J. (2018). Automated optimal parameters for T-distributed stochastic neighbor embedding improve visualization and allow analysis of large datasets. *bioRxiv*, page 451690.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Bergenstråhle, L., He, B., Bergenstråhle, J., Andersson, A., Lundeberg, J., Zou, J., and Maaskola, J. (2020). Super-resolved spatial transcriptomics by deep data fusion.

Blache, P., Van De Wetering, M., Duluc, I., Domon, C., Berta, P., Freund, J. N., Clevers, H., and Jay, P. (2004). SOX9 is an intestine crypt transcription factor, is regulated by the Wnt pathway, and represses the CDX2 and MUC2 genes. *Journal of Cell Biology*, 166(1):37–47.

Black, S., Phillips, D., Hickey, J. W., Kennedy-Darling, J., Venkataraaman, V. G., Samusik, N., Goltsev, Y., Schürch, C. M., and Nolan, G. P. (2021). CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nature Protocols*, 16(8):3802–3835.

Bortolomeazzi, M., Keddar, M. R., Montorsi, L., Acha-Sagredo, A., Benedetti, L., Temelkovski, D., Choi, S., Petrov, N., Todd, K., Wai, P., Kohl, J., Denner, T., Nye, E., Goldstone, R., Ward, S., Wilson, G. A., Al Bakir, M., Swanton, C., John, S., Miles, J., Larijani, B., Kunene, V., Fontana, E., Arkenau, H. T., Parker, P. J., Rodriguez-Justo, M., Shiu, K. K., Spencer, J., and Ciccarelli, F. D. (2021). Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts. *Gastroenterology*, 161(4):1179–1193.

Borza, C. M., Bolas, G., Bock, F., Zhang, X., Akabogu, F. C., Zhang, M. Z., de Caestecker, M., Yang, M., Yang, H., Lee, E., Gewin, L., Fogo, A. B., McDonald, W. H., Zent, R., and Pozzi, A. (2022). DDR1 contributes to kidney inflammation and fibrosis by promoting the phosphorylation of BCR and STAT3. *JCI Insight*, 7(3).

Bourdais, R., Rousseau, B., Pujals, A., Boussion, H., Joly, C., Guillemin, A., Baumgaertner, I., Neuzillet, C., and Tournigand, C. (2017). Polymerase proofreading domain mutations: New opportunities for immunotherapy in hypermutated colorectal cancer beyond MMR deficiency. *Critical Reviews in Oncology/Hematology*, 113:242–248.

Brahimi-Horn, M. C., Chiche, J., and Pouysségur, J. (2007). Hypoxia and cancer. *Journal of Molecular Medicine*, 85(12):1301–1307.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Bronder, D., Tighe, A., Wangsa, D., Zong, D., Meyer, T. J., Wardenaar, R., Minshall, P., Hirsch, D., Heselmeyer-Haddad, K., Nelson, L., Spierings, D., McGrail, J. C., Cam, M., Nussenzweig, A., Foijer, F., Ried, T., and Taylor, S. S. (2021). TP53 loss initiates chromosomal instability in fallopian tube epithelial cells. *DMM Disease Models and Mechanisms*, 14(11).

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420.

Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., and Irizarry, R. A. (2020). Robust decomposition of cell type mixtures in spatial transcriptomics. *bioRxiv*, page 2020.05.07.082750.

Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., and Irizarry, R. A. (2021). Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology 2021*, pages 1–10.

Calon, A., Lonardo, E., Berenguer-Llergo, A., Espinet, E., Hernando-Momblona, X., Iglesias, M., Sevillano, M., Palomo-Ponce, S., Tauriello, D. V., Byrom, D., Cortina, C., Morral, C., Barceló, C., Tosi, S., Riera, A., Attolini, C. S. O., Rossell, D., Sancho, E., and Batlle, E. (2015). Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nature Genetics*, 47(4):320–329.

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., and Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (New York, N.Y.)*, 361(6409):1380–1385.

Cardoso, J., Molenaar, L., De Menezes, R. X., Van Leerdam, M., Rosenberg, C., Möslein, G., Sampson, J., Morreau, H., Boer, J. M., and Fodde, R. (2006). Chromosomal Instability in MYH- and APC-Mutant Adenomatous Polyps. *Cancer Research*, 66(5):2514–2519.

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., Hackl, H., and Trajanoski, Z. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Reports*, 18(1):248–262.

Chen, B., Ramirez-Solano, M. A., Heiser, C. N., Liu, Q., and Lau, K. S. (2021a). Processing single-cell RNA-seq data for dimension reduction-based analyses using open-source tools. *STAR Protocols*, 2(2):100450.

Chen, B., Scurrah, C. R., McKinley, E. T., Simmons, A. J., Ramirez-Solano, M. A., Zhu, X., Markham, N. O., Heiser, C. N., Vega, P. N., Rolong, A., Kim, H., Sheng, Q., Drewes, J. L., Zhou, Y., Southard-Smith, A. N., Xu, Y., Ro, J., Jones, A. L., Revetta, F., Berry, L. D., Niitsu, H., Islam, M., Pelka, K., Hofree, M., Chen, J. H., Sarkizova, S., Ng, K., Giannakis, M., Boland, G. M., Aguirre, A. J., Anderson, A. C., Rozenblatt-Rosen, O., Regev, A., Hacohen, N., Kawasaki, K., Sato, T., Goettel, J. A., Grady, W. M., Zheng, W., Washington, M. K., Cai, Q., Sears, C. L., Goldenring, J. R., Franklin, J. L., Su, T., Huh, W. J., Vandekar, S., Roland, J. T., Liu, Q., Coffey, R. J., Shrubsole, M. J., and Lau, K. S. (2021b). Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. *Cell*, 0(0).

Chen, Y., Zhao, H., Feng, Y., Ye, Q., Hu, J., Guo, Y., and Feng, Y. (2021c). Pan-Cancer Analysis of the Associations of TGFBI Expression With Prognosis and Immune Characteristics. *Frontiers in Molecular Biosciences*, 8.

Chen, Z., Soifer, I., Hilton, H., Keren, L., and Jojic, V. (2020). Modeling Multiplexed Images with Spatial-LDA Reveals Novel Tissue Microenvironments. *Journal of Computational Biology*, 27(8):1204–1218.

Cheng, S., Li, Z., Gao, R., Xing, B., Gao, Y., Yang, Y., Qin, S., Zhang, L., Ouyang, H., Du, P., Jiang, L., Zhang, B., Yang, Y., Wang, X., Ren, X., Bei, J. X., Hu, X., Bu, Z., Ji, J., and Zhang, Z. (2021). A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell*, 184(3):792–809.e23.

Clarke, R. T. and Greenacre, M. J. (1985). Theory and applications of correspondence analysis. *Journal of Animal Ecology*, 54:1031.

Colom, B., Herms, A., Hall, M. W., Dentro, S. C., King, C., Sood, R. K., Alcolea, M. P., Piedrafita, G., Fernandez-Antoran, D., Ong, S. H., Fowler, J. C., Mahbubani, K. T., Saeb-Parsy, K., Gerstung, M., Hall, B. A., and Jones, P. H. (2021). Mutant clones in normal epithelium outcompete and eliminate emerging tumours. *Nature*, 598(7881):510–514.

Combes, A. J., Samad, B., Tsui, J., Chew, N. W., Yan, P., Reeder, G. C., Kushnoor, D., Shen, A., Davidson, B., Barczak, A. J., Adkisson, M., Edwards, A., Naser, M., Barry, K. C., Courau, T., Hammoudi, T., Argüello, R. J., Rao, A. A., Olshen, A. B., Cai, C., Zhan, J., Davis, K. C., Kelley, R. K., Chapman, J. S., Atreya, C. E., Patel, A., Daud, A. I., Ha, P., Diaz, A. A., Kratz, J. R., Collisson, E. A., Fragiadakis, G. K., Erle, D. J., Boissonnas, A., Asthana, S., Chan, V., Krummel, M. F., Spitzer, M., Fong, L., Nelson, A., Kumar, R., Lee, J., Burra, A., Hsu, J., Hackett, C., Tolentino, K., Sjarif, J., Johnson, P., Shao, E., Abrau, D., Lupin, L., Shaw, C., Collins, Z., Lea, T., Corvera, C., Nakakura, E., Carnevale, J., Alvarado, M., Loo, K., Chen, L., Chow, M., Grandis, J., Ryan, W., El-Sayed, I., Jablons, D., Woodard, G., Meng, M. W., Porten, S. P., Okada, H., Tempero, M., Ko, A., Kirkwood, K., Vandenberg, S., Guevarra, D., Oropeza, E., Cyr, C., Glenn, P., Bolen, J., Morton, A., and Eckalbar, W. (2022). Discovering dominant tumor immune archetypes in a pan-cancer census. *Cell*, 185(1):184–203.e19.

Conteduca, V., Sansonno, D., Russi, S., and Dammaco, F. (2013). Precancerous colorectal lesions. *International Journal of Oncology*, 43(4):973–984.

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Dalton, W. B., Yu, B., and Yang, V. W. (2010). p53 suppresses structural chromosome instability after mitotic arrest in human cells. *Oncogene*, 29(13):1929–1940.

De Nooij-van Dalen, A. G., Morolli, B., Van Der Keur, M., Van Der Marel, A., Lohman, P. H., and Giphart-Gassler, M. (2001). Intrinsic genetic instability of normal human lymphocytes and its implication for loss of heterozygosity. *Genes, Chromosomes and Cancer*, 30(4):323–335.

Ding, J., Condon, A., and Shah, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications*, 9(1):2002.

Drews, R. M., Hernando, B., Tarabichi, M., Haase, K., Lesluyes, T., Smith, P. S., Morrill Gavarró, L., Couturier, D. L., Liu, L., Schneider, M., Brenton, J. D., Van Loo, P., Macintyre, G., and Markowetz, F. (2022). A pan-cancer compendium of chromosomal instability. *Nature*, 606(7916):976–983.

Duan, X., Xu, X., Zhang, Y., Gao, Y., Zhou, J., and Li, J. (2022). DDR1 functions as an immune negative factor in colorectal cancer by regulating tumor-infiltrating T cells through IL-18. *Cancer Science*.

Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., and Heyn, H. (2021). SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9):e50–e50.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390.

Erickson, A., He, M., Berglund, E., Marklund, M., Mirzazadeh, R., Schultz, N., Kvastad, L., Andersson, A., Bergenstråhle, L., Bergenstråhle, J., Larsson, L., Alonso Galicia, L., Shamikh, A., Basmaci, E., Díaz De Ståhl, T., Rajakumar, T., Doultsinos, D., Thrane, K., Ji, A. L., Khavari, P. A., Tarish, F., Tanoglidi, A., Maaskola, J., Colling, R., Mirtti, T., Hamdy, F. C., Woodcock, D. J., Helleday, T., Mills, I. G., Lamb, A. D., and Lundeberg, J. (2022). Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature*, 608(7922):360–367.

Fearon, E. R. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767.

Feig, C., Jones, J. O., Kraman, M., Wells, R. J., Deonarine, A., Chan, D. S., Connell, C. M., Roberts, E. W., Zhao, Q., Caballero, O. L., Teichmann, S. A., Janowitz, T., Jodrell, D. I., Tuveson, D. A., and Fearon, D. T. (2013). Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 110(50):20212–20217.

Femia, A. P., Dolara, P., Giannini, A., Salvadori, M., Biggeri, A., and Caderni, G. (2007). Frequent mutation of Apc gene in rat colon tumors and mucin-depleted foci, preneoplastic lesions in experimental colon carcinogenesis. *Cancer research*, 67(2):445–449.

Flamary, R. and Courty, N. (2017). POT Python Optimal Transport library.

Fleming, S. J., Marioni, J. C., and Babadi, M. (2019). CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. *bioRxiv*, page 791699.

Foijer, F., Xie, S. Z., Simon, J. E., Bakker, P. L., Conte, N., Davis, S. H., Kregel, E., Jonkers, J., Bradley, A., and Sorger, P. K. (2014). Chromosome instability induced by Mps1 and p53 mutation generates aggressive lymphomas exhibiting aneuploidy-induced stress. *Proceedings of the National Academy of Sciences of the United States of America*, 111(37):13427–13432.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Galeano Niño, J. L., Wu, H., LaCourse, K. D., Kempchinsky, A. G., Baryiames, A., Barber, B., Futran, N., Houlton, J., Sather, C., Sicinska, E., Taylor, A., Minot, S. S., Johnston, C. D., and Bullman, S. (2022). Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature*, 611(7937):810–817.

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., Zinzindohoué, F., Bruneval, P., Cugnenc, P. H., Trajanoski, Z., Fridman, W. H., and Pagès, F. (2006). Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795):1960–1964.

Gerdes, M. J., Sevinsky, C. J., Sood, A., Adak, S., Bello, M. O., Bordwell, A., Can, A., Corwin, A., Dinn, S., Filkins, R. J., Hollman, D., Kamath, V., Kaanumalle, S., Kenny, K., Larsen, M., Lazare, M., Li, Q., Lowes, C., McCulloch, C. C., McDonough, E., Montalto, M. C., Pang, Z., Rittscher, J., Santamaria-Pang, A., Sarachan, B. D., Seel, M. L., Seppo, A., Shaikh, K., Sui, Y., Zhang, J., and Ginty, F. (2013). Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences*, 110(29):11982–11987.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, 366(10):883–892.

Germann, M., Zangger, N., Sauvain, M.-O., Sempoux, C., Bowler, A. D., Wirapati, P., Kandalaft, L. E., Delorenzi, M., Tejpar, S., Coukos, G., and Radtke, F. (2020). Neutrophils suppress tumor-infiltrating T cells in colon cancer via matrix metalloproteinase-mediated activation of TGFβ. *EMBO Molecular Medicine*, 12(1):e10681.

Gil Vasquez, E., Nasreddin, N., Valbuena, G. N., Tejpar, S., Sansom, O. J., Correspondence, S. J. L., Vasquez, G., Mulholland, E. J., Belnoue-Davis, H. L., Eggington, H. R., Schenck, R. O., Rie, V., Wouters, M., Wirapati, P., Gilroy, K., Lannagan, T. R. M., Flanagan, D. J., Najumudeen, A. K., Omwenga, S., Mccorry, A. M. B., Easton, A., Koelzer, V. H., East, J. E., Morton, D., Trusolino, L., Maughan, T., Campbell, A. D., Loughrey, M. B., Dunne, P. D., Tsantoulis, P., Huels, D. J., and Leedham, S. J. (2022). Dynamic and adaptive cancer stem cell population admixture in colorectal neoplasia. *Cell Stem Cell*, 29(8):1213–1228.e8.

Gould, S. J. and Eldredge, N. (1977). Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology*, 3(2):115–151.

Graf, J., Cho, S., Mcdonough, E., Corwin, A., Sood, A., Lindner, A., Salvucci, M., Stachtea, X., Van Schaeybroeck, S., Dunne, P. D., Laurent-Puig, P., Longley, D., Prehn, J. H., and Ginty, F. (2021). FLINO-A new method for immunofluorescence bioimage normalization. *Bioinformatics (Oxford, England)*, 38(2):btab686–btab686.

Greenwald, N. F., Miller, G., Moen, E., Kong, A., Kagel, A., Fullaway, C. C., McIntosh, B. J., Leow, K., Schwartz, M. S., Dougherty, T., Pavelchek, C., Cui, S., Camplisson, I., Bar-Tal, O., Singh, J., Fong, M., Chaudhry, G., Abraham, Z., Moseley, J., Warshawsky, S., Soon, E., Greenbaum, S., Risom, T., Hollmann, T., Keren, L., Graf, W., Angelo, M., and Valen, D. V. (2021). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *bioRxiv*, page 2021.03.01.431313.

Grist, S. A., McCarron, M., Kutlaca, A., Turner, D. R., and Morley, A. A. (1992). In vivo human somatic mutation: frequency and spectrum with age. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 266(2):189–196.

Guinney, J., Dienstmann, R., Wang, X., De Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster, S., Fessler, E., De Sousa .E Melo, F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G. C., Broom, B., Boige, V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J. W., Hanahan, D., Tabernero, J., Bernards, R., Friend, S. H., Laurent-Puig, P., Medema, J. P., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L., and Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 21(11):1350–1356.

Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., Zabala, M., Scheeren, F. A., Lobo, N. A., Qian, D., Yu, F. B., Dirbas, F. M., Clarke, M. F., and Newman, A. M. (2020). Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, 367(6476):405–411.

Gunnarsson, U., Strigård, K., Edin, S., Gkekas, I., Mustonen, H., Kaprio, T., Böckelman, C., Hagström, J., Palmqvist, R., and Haglund, C. (2020). Association between local immune cell infiltration, mismatch repair status and systemic inflammatory response in colorectal cancer. *Journal of Translational Medicine*, 18(1):178.

Gutierrez, D. B., Gant-Branum, R. L., Romer, C. E., Farrow, M. A., Allen, J. L., Dahal, N., Nei, Y.-W., Codreanu, S. G., Jordan, A. T., Palmer, L. D., Sherrod, S. D., Mclean, J. A., Skaar, E. P., Norris, J. L., and Caprioli, R. M. (2018). An Integrated, High-Throughput Strategy for Multiomic Systems Level Analysis. *Journal of Proteome Research*.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Halekoh, U., Højsgaard, S., and Yan, J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, 15(2):1–11.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G. C., Chen, M., and Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 172(5):1091–1107.e17.

Harris, C. R., McKinley, E. T., Roland, J. T., Liu, Q., Shrubsole, M. J., Lau, K. S., Coffey, R. J., Wrobel, J., and Vandekar, S. N. (2022). Quantifying and correcting slide-to-slide variation in multiplexed immunofluorescence images. *Bioinformatics*.

Heiser, C. N. and Lau, K. S. (2020). A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques. *Cell Reports*, 31(5):107576.

Heiser, C. N., Simmons, A. J., Revetta, F., McKinley, E. T., Ramirez-Solano, M. A., Wang, J., Shao, J., Ayers, G. D., Wang, Y., Glass, S. E., Kaur, H., Rolong, A., Chen, B., Vega, P. N., Drewes, J. L., Saleh, N., Vandekar, S., Jones, A. L., Washington, M. K., Roland, J. T., Sears, C. L., Liu, Q., Shrubsole, M. J., Coffey, R. J., and Lau, K. S. (2023). Molecular cartography uncovers evolutionary and microenvironmental dynamics in sporadic colorectal tumors. *bioRxiv*, page 2023.03.09.530832.

Heiser, C. N., Wang, V. M., Chen, B., Hughey, J. J., and Lau, K. S. (2021). Automated quality control and cell identification of droplet-based single-cell data using dropkick. *Genome Research*, 31(10):1742–1752.

Herring, C. A., Banerjee, A., McKinley, E. T., Simmons, A. J., Ping, J., Roland, J. T., Franklin, J. L., Liu, Q., Gerdes, M. J., Coffey, R. J., and Lau, K. S. (2018). Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut. *Cell Systems*, 6(1):37–51.e9.

Hickey, J. W., Becker, W. R., Nevins, S. A., Horning, A., Perez, A. E., Chiu, R., Chen, D. C., Cotter, D., Esplin, E. D., Weimer, A. K., Caraccio, C., Venkataraaman, V., Schürch, C. M., Black, S., Brbić, M., Cao, K., Leskovec, J., Zhang, Z., Lin, S., Longacre, T., Plevitis, S. K., Lin, Y., Nolan, G. P., Greenleaf, W. J., and Snyder, M. (2021). High Resolution Single Cell Maps Reveals Distinct Cell Organization and Function Across Different Regions of the Human Intestine. *bioRxiv*, page 2021.11.25.469203.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55.

Househam, J., Heide, T., Cresswell, G. D., Spiteri, I., Kimberley, C., Zapata, L., Lynn, C., James, C., Mossner, M., Fernandez-Mateos, J., Vinceti, A., Baker, A.-M., Gabbutt, C., Berner, A., Schmidt, M., Chen, B., Lakatos, E., Gunasri, V., Nichol, D., Costa, H., Mitchinson, M., Ramazzotti, D., Werner, B., Iorio, F., Jansen, M., Caravagna, G., Barnes, C. P., Shibata, D., Bridgewater, J., Rodriguez-Justo, M., Magnani, L., Sottoriva, A., and Graham, T. A. (2022). Phenotypic plasticity and genetic control in colorectal cancer evolution. *Nature*, pages 1–10.

Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., Lee, E. B., Shinohara, R. T., and Li, M. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351.

Huang, S. (2012). The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology? *BioEssays*, 34(2):149–157.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3):99–104.

Jackson, A. L. and Loeb, L. A. (1998). The Mutation Rate and Cancer. *Genetics*, 148(4):1483–1490.

Jackson, H. W., Fischer, J. R., Zanotelli, V. R., Ali, H. R., Mechera, R., Soysal, S. D., Moch, H., Muenst, S., Varga, Z., Weber, W. P., and Bodenmiller, B. (2020). The single-cell pathology landscape of breast cancer. *Nature*, 578(7796):615–620.

Javed, S., Mahmood, A., Fraz, M. M., Koohbanani, N. A., Benes, K., Tsang, Y. W., Hewitt, K., Epstein, D., Snead, D., and Rajpoot, N. (2020). Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical Image Analysis*, 63:101696.

Ji, A. L., Rubin, A. J., Thrane, K., Jiang, S., Reynolds, D. L., Meyers, R. M., Guo, M. G., George, B. M., Mollbrink, A., Bergenstråhle, J., Larsson, L., Bai, Y., Zhu, B., Bhaduri, A., Meyers, J. M., Rovira-Clavé, X., Hollmig, S. T., Aasi, S. Z., Nolan, G. P., Lundeberg, J., and Khavari, P. A. (2020). Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell*, 182(2):497–514.e22.

Joanito, I., Wirapati, P., Zhao, N., Nawaz, Z., Yeo, G., Lee, F., Eng, C. L. P., Macalinao, D. C., Kahraman, M., Srinivasan, H., Lakshmanan, V., Verbandt, S., Tsantoulis, P., Gunn, N., Venkatesh, P. N., Poh, Z. W., Nahar, R., Oh, H. L. J., Loo, J. M., Chia, S., Cheow, L. F., Cheruba, E., Wong, M. T., Kua, L., Chua, C., Nguyen, A., Golovan, J., Gan, A., Lim, W.-J., Guo, Y. A., Yap, C. K., Tay, B., Hong, Y., Chong, D. Q., Chok, A.-Y., Park, W.-Y., Han, S., Chang, M. H., Seow-En, I., Fu, C., Mathew, R., Toh, E.-L., Hong, L. Z., Skanderup, A. J., DasGupta, R., Ong, C.-A. J., Lim, K. H., Tan, E. K. W., Koo, S.-L., Leow, W. Q., Tejpar, S., Prabhakar, S., and Tan, I. B. (2022). Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nature Genetics*, 18:1–13.

Karlsson, N. G., Johansson, M. E., Asker, N., Karlsson, H., Gendler, S. J., Carlstedt, I., and Hansson, G. C. (1996). Molecular characterization of the large heavily glycosylated domain glycopeptide from the rat small intestinal Muc2 mucin. *Glycoconjugate journal*, 13(5):823–831.

Kather, J. N., Suarez-Carmona, M., Charoentong, P., Weis, C.-A., Hirsch, D., Bankhead, P., Horning, M., Ferber, D., Kel, I., Herpel, E., Schott, S., Zörnig, I., Utikal, J., Marx, A., Gaiser, T., Brenner, H., Chang-Claude, J., Hoffmeister, M., Jäger, D., and Halama, N. (2018). Topography of cancer-associated immune cells in human solid tumors. *eLife*, 7.

Kaur, H., Heiser, C. N., McKinley, E. T., Antunes, L. V., Harris, C. R., Roland, J. T., Shrubsole, M. J., Coffey, R. J., Lau, K., and Vandekar, S. N. (2023). Consensus tissue domain detection in spatial -omics data using MILWRM. *bioRxiv*, page 2023.02.02.526900.

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., Yang, S. R., Kurian, A., Van Valen, D., West, R., Bendall, S. C., and Angelo, M. (2018). A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell*, 174(6):1373–1387.e19.

Kim, J., Rustam, S., Mosquera, J. M., Randell, S. H., Shaykhiev, R., Rendeiro, A. F., and Elemento, O. (2022). Unsupervised discovery of tissue architecture in multiplexed imaging. *Nature Methods*, 19(12):1653–1661.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201.

Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., Jain, M. S., Park, J. S., Ramona, L., Tuck, E., Arutyunyan, A., Vento-Tormo, R., Gerstung, M., James, L., Stegle, O., and Bayraktar, O. A. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, pages 1–11.

Kobak, D. and Linderman, G. C. (2019). UMAP does not preserve global structure any better than t-SNE when using the same initialization. *bioRxiv*, page 2019.12.19.877522.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari.*, 4(1):83–91.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., ru Loh, P., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296.

Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., and Sabeti, P. C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife*, 8.

Krishnamurty, A. T., Shyer, J. A., Thai, M., Gandham, V., Buechler, M. B., Yang, Y. A., Pradhan, R. N., Wang, A. W., Sanchez, P. L., Qu, Y., Breart, B., Chalouni, C., Dunlap, D., Ziai, J., Elstrott, J., Zacharias, N., Mao, W., Rowntree, R. K., Sadowsky, J., Lewis, G. D., Pillow, T. H., Nabet, B. Y., Banchereau, R., Tam, L., Caothien, R., Bacarro, N., Roose-Girma, M., Modrusan, Z., Mariathasan, S., Müller, S., and Turley, S. J. (2022). LRRC15+ myofibroblasts dictate the stromal setpoint to suppress tumour immunity. *Nature*, 611(7934):148–154.

Lang, J., Leal, A. D., Marín-Jiménez, J. A., Hartman, S. J., Shulman, J., Navarro, N. M., Lewis, M. S., Capasso, A., Bagby, S. M., Yacob, B. W., MacBeth, M., Freed, B. M., Eckhardt, S. G., Jordan, K., Blatchford, P. J., Pelanda, R., Lieu, C. H., Messersmith, W. A., and Pitts, T. M. (2022). Cabozantinib sensitizes microsatellite stable colorectal cancer to immune checkpoint blockade by immune modulation in human immune system mouse models. *Frontiers in Oncology*, 12.

Larsson, E., Tremaroli, V., Lee, Y. S., Koren, O., Nookaew, I., Fricker, A., Nielsen, J., Ley, R. E., and Bäckhed, F. (2012). Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. *Gut*, 61(8):1124–31.

Lazarus, J., Maj, T., Smith, J. J., Perusina Lanfranca, M., Rao, A., D'Angelica, M. I., Delrosario, L., Girgis, A., Schukow, C., Shia, J., Kryczek, I., Shi, J., Wasserman, I., Crawford, H., Nathan, H., Pasca Di Magliano, M., Zou, W., and Frankel, T. L. (2018). Spatial and phenotypic immune profiling of metastatic colon cancer. *JCI insight*, 3(22).

Lecker, L. S., Berlato, C., Maniati, E., Delaine-Smith, R., Pearce, O. M., Heath, O., Nichols, S. J., Trevisan, C., Novak, M., McDermott, J., Brenton, J. D., Cutillas, P. R., Rajeeve, V., Hennino, A., Drapkin, R., Loessner, D., and Balkwill, F. R. (2021). Tgfbi production by macrophages contributes to an immunosuppressive microenvironment in ovarian cancer. *Cancer Research*, 81(22):5706–5719.

Leduc, P. R., Messner, W. C., and Wikswo, J. P. (2011). How Do Control-Based Approaches Enter into Biology? *Annual Reviews Biomedical Engineering*.

Lee, H.-O., Hong, Y., Etlioglu, H. E., Cho, Y. B., Pomella, V., Van den Bosch, B., Vanhecke, J., Verbandt, S., Hong, H., Min, J.-W., Kim, N., Eum, H. H., Qian, J., Boeckx, B., Lambrechts, D., Tsantoulis, P., De Hertogh, G., Chung, W., Lee, T., An, M., Shin, H.-T., Joung, J.-G., Jung, M.-H., Ko, G., Wirapati, P., Kim, S. H., Kim, H. C., Yun, S. H., Tan, I. B. H., Ranjan, B., Lee, W. Y., Kim, T.-Y., Choi, J. K., Kim, Y.-J., Prabhakar, S., Tejpar, S., and Park, W.-Y. (2020). Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nature Genetics*, pages 1–10.

Lee, Y. C., Kurtova, A. V., Xiao, J., Nikolos, F., Hayashi, K., Tramel, Z., Jain, A., Chen, F., Chokshi, M., Lee, C., Bao, G., Zhang, X., Shen, J., Mo, Q., Jung, S. Y., Rowley, D., and Chan, K. S. (2019). Collagen-rich airway smooth muscle cells are a metastatic niche for tumor colonization in the lung. *Nature Communications*, 10(1):1–16.

Leow, C. C., Romero, M. S., Ross, S., Polakis, P., and Gao, W. Q. (2004). Hath1, Down-Regulated in Colon Adenocarcinomas, Inhibits Proliferation and Tumorigenesis of Colon Cancer Cells. *Cancer Research*, 64(17):6050–6057.

Lepourcelet, M., Tou, L., Cai, L., Sawada, J.-i., Lazar, A. J. F., Glickman, J. N., Williamson, J. A., Everett, A. D., Redston, M., Fox, E. A., Nakatani, Y., and Shivdasani, R. A. (2005). Insights into developmental mechanisms and cancers in the mammalian intestine derived from serial analysis of gene expression and study of the hepatoma-derived growth factor (HDGF). *Development*, 121(10):3163–3174.

Levina, E. and Bickel, P. (2001). The Earth Mover's distance is the Mallows distance: some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 251–256. IEEE Comput. Soc.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., Finck, R., Gedman, A. L., Radtke, I., Downing, J. R., Pe'er, D., and Nolan, G. P. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–97.

Lewis, S. M., Asselin-Labat, M. L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V. C., Merino, D., Rogers, K. L., and Naik, S. H. (2021). Spatial omics and multiplexed imaging to explore cancer biology. *Nature Methods*, 18(9):997–1012.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992.

Lin, J.-R., Wang, S., Coy, S., Lau, K. S., Santagata, S., Sorger Correspondence, P. K., Chen, Y.-A., Yapp, C., Tyler, M., Nariya, M. K., Heiser, C. N., and Sorger, P. K. (2023). Multiplexed 3D atlas of state transitions and immune interaction in colorectal cancer. *Cell*, 186(2):363–381.e19.

Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., and Kluger, Y. (2017). Efficient Algorithms for t-distributed Stochastic Neighborhood Embedding. *arXiv*.

Liu, C. C., Greenwald, N. F., Kong, A., McCaffrey, E. F., Leow, K. X., Mrdjen, D., and Angelo, M. (2022). Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering. *bioRxiv*, page 2022.08.16.504171.

Liu, M., Chen, J., Wang, X., Wang, C., Zhang, X., Xie, Y., Zuo, Z., Ren, J., and Zhao, Q. (2021). MesKit: a tool kit for dissecting cancer evolution of multi-region tumor biopsies through somatic alterations. *GigaScience*, 10(5):1–12.

Liu, Q., Herring, C. A., Sheng, Q., Ping, J., Simmons, A. J., Chen, B., Banerjee, A., Li, W., Gu, G., Coffey, R. J., Shyr, Y., and Lau, K. S. (2018). Quantitative assessment of cell population diversity in single-cell landscapes. *PLOS Biology*, 16(10):e2006687.

Llosa, N. J., Cruise, M., Tam, A., Wicks, E. C., Hechenbleikner, E. M., Taube, J. M., Blosser, R. L., Fan, H., Wang, H., Luber, B. S., Zhang, M., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., Sears, C. L., Anders, R. A., Pardoll, D. M., and Housseau, F. (2015). The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discovery*, 5(1):43–51.

Lomakin, A., Svedlund, J., Strell, C., Gataric, M., Shmatko, A., Rukhovich, G., Park, J. S., Ju, Y. S., Dentro, S., Kleshchevnikov, V., Vaskivskyi, V., Li, T., Bayraktar, O. A., Pinder, S., Richardson, A. L., Santagata, S., Campbell, P. J., Russnes, H., Gerstung, M., Nilsson, M., and Yates, L. R. (2022). Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature*, pages 1–9.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058.

Luca, B. A., Steen, C. B., Matusiak, M., Azizi, A., Varma, S., Zhu, C., Przybyl, J., Espín-Pérez, A., Diehn, M., Alizadeh, A. A., van de Rijn, M., Gentles, A. J., and Newman, A. M. (2021). Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell*.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6).

Luke, J. J., Bao, R., Sweis, R. F., Spranger, S., and Gajewski, T. F. (2019). WNT/b-catenin pathway activation correlates with immune exclusion across human cancers. *Clinical Cancer Research*, 25(10):3074–3083.

Lukowski, S. W., Lo, C. Y., Sharov, A. A., Nguyen, Q., Fang, L., Hung, S. S., Zhu, L., Zhang, T., Grünert, U., Nguyen, T., Senabouth, A., Jabbari, J. S., Welby, E., Sowden, J. C., Waugh, H. S., Mackey, A., Pollock, G., Lamb, T. D., Wang, P., Hewitt, A. W., Gillies, M. C., Powell, J. E., and Wong, R. C. (2019). A single-cell transcriptome atlas of the adult human retina. *The EMBO Journal*.

Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., and Marioni, J. C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*, 20(1):63.

Luoma, A. M., Suo, S., Williams, H. L., Sharova, T., Sullivan, K., Manos, M., Bowling, P., Hodi, F. S., Rahma, O., Sullivan, R. J., Boland, G. M., Nowak, J. A., Dougan, S. K., Dougan, M., Yuan, G. C., and Wucherpfennig, K. W. (2020). Molecular Pathways of Colon Inflammation Induced by Cancer Immunotherapy. *Cell*, 182(3):655–671.e22.

Macosko, E., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., Trombetta, J., Weitz, D., Sanes, J., Shalek, A., Regev, A., and McCarroll, S. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214.

Marabita, F., James, T., Karhu, A., Virtanen, H., Kettunen, K., Stenlund, H., Boulund, F., Hellström, C., Neiman, M., Mills, R., Perheentupa, T., Laivuori, H., Helkkula, P., Byrne, M., Jokinen, I., Honko, H., Kallonen, A., Ermes, M., Similä, H., Lindholm, M., Widén, E., Ripatti, S., Perälä-Heape, M., Engstrand, L., Nilsson, P., Moritz, T., Miettinen, T., Sallinen, R., and Kallioniemi, O. (2022). Multiomics and digital monitoring during lifestyle changes reveal independent dimensions of human biology and health. *Cell Systems*, 13(3):241–255.e7.

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, 28(11):1747–1756.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*.

Mckenna, M. T., Weis, J. A., Brock, A., Quaranta, V., and Yankeelov, T. E. (2018). Precision Medicine with Imprecise Therapy: Computational Modeling for Chemotherapy in Breast Cancer 1. *Translational Oncology*, 11:732–742.

McKinley, E. T., Shao, J., Ellis, S. T., Heiser, C. N., Roland, J. T., Macedonia, M. C., Vega, P. N., Shin, S., Coffey, R. J., and Lau, K. S. (2022). MIRIAM: A machine and deep learning single-cell segmentation and quantification pipeline for multi-dimensional tissue images. *Cytometry Part A*.

McKinley, E. T., Sui, Y., Al-Kofahi, Y., Millis, B. A., Tyska, M. J., Roland, J. T., Santamaria-Pang, A., Ohland, C. L., Jobin, C., Franklin, J. L., Lau, K. S., Gerdes, M. J., and Coffey, R. J. (2017). Optimized multiplex immunofluorescence single-cell analysis reveals tuft cell heterogeneity. *JCI Insight*, 2(11).

Mckinney, W. (2010). Data Structures for Statistical Computing in Python.

Milo, I., Bedora-Faure, M., Garcia, Z., Thibaut, R., Périé, L., Shakhar, G., Deriano, L., and Bousso, P. (2018). The immune system profoundly restricts intratumor genetic heterogeneity. *Science Immunology*, 3(29).

Moehlin, J., Mollet, B., Colombo, B. M., and Mendoza-Parra, M. A. (2021). Inferring biologically relevant molecular tissue substructures by agglomerative clustering of digitized spatial transcriptomes with multilayer. *Cell Systems*, 0(0).

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., and Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*.

Moor, A. E. and Itzkovitz, S. (2017). Spatial transcriptomics: paving the way for tissue-level systems biology. *Current Opinion in Biotechnology*, 46:126–133.

Moses, L. and Pachter, L. (2022). Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546.

Motta, R., Cabezas-Camarero, S., Torres-Mattos, C., Riquelme, A., Calle, A., Figueroa, A., and Sotelo, M. J. (2021). Immunotherapy in microsatellite instability metastatic colorectal cancer: Current status and future perspectives. *Journal of Clinical and Translational Research*, 7(4):511.

Nagle, M. P., Tam, G. S., Maltz, E., Hemminger, Z., and Wollman, R. (2021). Bridging scales: From cell biology to physiology using in situ single-cell technologies. *Cell Systems*, 12(5):388–400.

Nikolos, F., Hayashi, K., Hoi, X. P., Alonzo, M. E., Mo, Q., Kasabyan, A., Furuya, H., Trepel, J., Di Vizio, D., Guarnerio, J., Theodorescu, D., Rosser, C., Apolo, A., Galsky, M., and Chan, K. S. (2022). Cell death-induced immunogenicity enhances chemoimmunotherapeutic response by converting immune-excluded into T-cell inflamed bladder tumors. *Nature Communications*, 13(1):1–16.

Nirmal, A. J., Regan, T., Shih, B. B., Hume, D. A., Sims, A. H., and Freeman, T. C. (2018). Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors. *Cancer Immunology Research*, 6(11):1388–1400.

Nouri Nojadeh, J., Behrouz Sharif, S., and Sakhinia, E. (2018). Microsatellite instability in colorectal cancer. *EXCLI Journal*, 17:159–168.

Nowak, M. A., Komarova, N. L., Sengupta, A., Jallepalli, P. V., Shih, L. M., Vogelstein, B., and Lengauer, C. (2002). The role of chromosomal instability in tumor initiation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(25):16226–16231.

Obuch, J. C., Pigott, C. M., and Ahnen, D. J. (2015). Sessile Serrated Polyps: Detection, Eradication, and Prevention of the Evil Twin. *Current Treatment Options in Gastroenterology*, 13(1):156–170.

Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, 9(3):10–20.

Ortiz, C., Navarro, J. F., Jurek, A., Märtin, A., Lundeberg, J., and Meletis, K. (2020). Molecular atlas of the adult mouse brain. *Science Advances*, 6(26).

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Palla, G., Fischer, D. S., Regev, A., and Theis, F. J. (2022a). Spatial components of molecular tissue biology. *Nature Biotechnology*, pages 1–11.

Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., Rybakov, S., Ibarra, I. L., Holmberg, O., Virshup, I., Lotfollahi, M., Richter, S., and Theis, F. J. (2022b). Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19(2):171–178.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., and Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Pelka, K., Hofree, M., Chen, J. H., Sarkizova, S., Pirl, J. D., Jorgji, V., Bejnood, A., Dionne, D., Ge, W. H., Xu, K. H., Chao, S. X., Zollinger, D. R., Lieb, D. J., Reeves, J. W., Fuhrman, C. A., Hoang, M. L., Delorey, T., Nguyen, L. T., Waldman, J., Klapholz, M., Wakiro, I., Cohen, O., Albers, J., Smillie, C. S., Cuoco, M. S., Wu, J., ju Su, M., Yeung, J., Vijaykumar, B., Magnuson, A. M., Asinovski, N., Moll, T., Goder-Reiser, M. N., Applebaum, A. S., Brais, L. K., DelloStritto, L. K., Denning, S. L., Phillips, S. T., Hill, E. K., Meehan, J. K., Frederick, D. T., Sharova, T., Kanodia, A., Todres, E. Z., Jané-Valbuena, J., Biton, M., Izar, B., Lambden, C. D., Clancy, T. E., Bleday, R., Melnitchouk, N., Irani, J., Kunitake, H., Berger, D. L., Srivastava, A., Hornick, J. L., Ogino, S., Rotem, A., Vigneau, S., Johnson, B. E., Corcoran, R. B., Sharpe, A. H., Kuchroo, V. K., Ng, K., Giannakis, M., Nieman, L. T., Boland, G. M., Aguirre, A. J., Anderson, A. C., Rozenblatt-Rosen, O., Regev, A., and Hacohen, N. (2021). Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, 184(18):4734–4752.e20.

Piccinini, F., Balassa, T., Szkalisity, A., Molnar, C., Paavolainen, L., Kujala, K., Buzas, K., Sarazova, M., Pietiainen, V., Kutay, U., Smith, K., and Horvath, P. (2017). Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data. *Cell Systems*, 4(6):651–655.e5.

Pierson, E. and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241.

Pietro Femia, A., Dolara, P., and Caderni, G. (2004). Mucin-depleted foci (MDF) in the colon of rats treated with azoxymethane (AOM) are useful biomarkers for colon carcinogenesis. *Carcinogenesis*, 25(2):277–281.

Pino, M. S. and Chung, D. C. (2010). The Chromosomal Instability Pathway in Colon Cancer. *Gastroenterology*, 138(6):2059–2072.

Pinyol, R., Sia, D., and Llovet, J. M. (2019). Immune Exclusion-Wnt/CTNNB1 Class Predicts Resistance to Immunotherapies in HCC. *Clinical Cancer Research : an Official Journal of the American Association for Cancer Research*, 25(7):2021–2023.

Pretlow, T. P. and Pretlow, T. G. (2005). Mutant KRAS in aberrant crypt foci (ACF): Initiation of colorectal cancer? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1756(2):83–96.

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C. L., Mroz, E. A., Emerick, K. S., Deschler, D. G., Varvares, M. A., Mylvaganam, R., Rozenblatt-Rosen, O., Rocco, J. W., Faquin, W. C., Lin, D. T., Regev, A., and Bernstein, B. E. (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*, 171(7):1611–1624.e24.

Qi, J., Sun, H., Zhang, Y., Wang, Z., Xun, Z., Li, Z., Ding, X., Bao, R., Hong, L., Jia, W., Fang, F., Liu, H., Chen, L., Zhong, J., Zou, D., Liu, L., Han, L., Ginhoux, F., Liu, Y., Ye, Y., and Su, B. (2022). Single-cell and spatial analysis reveal interaction of FAP+ fibroblasts and SPP1+ macrophages in colorectal cancer. *Nature Communications*, 13(1):1–20.

Rana, A., Yauney, G., Lowe, A., and Shah, P. (2018). Computational Histological Staining and Destaining of Prostate Core Biopsy RGB Images with Generative Adversarial Neural Networks. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 828–834. IEEE.

Rao, A., Barkley, D., França, G. S., and Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., and Yosef, N. (2017). The Human Cell Atlas. *eLife*, 6.

Rheaume, B. A., Jereen, A., Bolisetty, M., Sajid, M. S., Yang, Y., Renna, K., Sun, L., Robson, P., and Trakhtenberg, E. F. (2018). Single cell transcriptome profiling of retinal ganglion cells identifies cellular subtypes. *Nature Communications*, 9(1):2759.

Rhee, Y. Y., Kim, K. J., and Kang, G. H. (2017). CpG Island Methylator Phenotype-High Colorectal Cancers and Their Prognostic Implications and Relationships with the Serrated Neoplasia Pathway. *Gut and Liver*, 11(1):38–46.

Risom, T., Glass, D. R., Averbukh, I., Liu, C. C., Baranski, A., Kagel, A., McCaffrey, E. F., Greenwald, N. F., Rivero-Gutiérrez, B., Strand, S. H., Varma, S., Kong, A., Keren, L., Srivastava, S., Zhu, C., Khair, Z., Veis, D. J., Deschryver, K., Vennam, S., Maley, C., Hwang, E. S., Marks, J. R., Bendall, S. C., Colditz, G. A., West, R. B., and Angelo, M. (2022). Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell*, 185(2):299–310.e18.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):284.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, pages 1463–1467.

Roh, W., Chen, P. L., Reuben, A., Spencer, C. N., Prieto, P. A., Miller, J. P., Gopalakrishnan, V., Wang, F., Cooper, Z. A., Reddy, S. M., Gumbs, C., Little, L., Chang, Q., Chen, W. S., Wani, K., De Macedo, M. P., Chen, E., Austin-Breneman, J. L., Jiang, H., Roszik, J., Tetzlaff, M. T., Davies, M. A., Gershenwald, J. E., Tawbi, H., Lazar, A. J., Hwu, P., Hwu, W. J., Diab, A., Glitza, I. C., Patel, S. P., Woodman, S. E., Amaria, R. N., Prieto, V. G., Hu, J., Sharma, P., Allison, J. P., Chin, L., Zhang, J., Wargo, J. A., and Futreal, P. A. (2017). Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Science Translational Medicine*, 9(379).

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*, 160(1-2):48–61.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J. E., Ashenberg, O., Cerami, E., Coffey, R. J., Demir, E., Ding, L., Esplin, E. D., Ford, J. M., Goecks, J., Ghosh, S., Gray, J. W., Guinney, J., Hanlon, S. E., Hughes, S. K., Hwang, E. S., Iacobuzio-Donahue, C. A., Jané-Valbuena, J., Johnson, B. E., Lau, K. S., Lively, T., Mazzilli, S. A., Pe'er, D., Santagata, S., Shalek, A. K., Schapiro, D., Snyder, M. P., Sorger, P. K., Spira, A. E., Srivastava, S., Tan, K., West, R. B., Williams, E. H., Aberle, D., Achilefu, S. I., Ademuyiwa, F. O., Adey, A. C., Aft, R. L., Agarwal, R., Aguilar, R. A., Alikarami, F., Allaj, V., Amos, C., Anders, R. A., Angelo, M. R., Anton, K., Ashenberg, O., Aster, J. C., Babur, O., Bahmani, A., Balsubramani, A., Barrett, D., Beane, J., Bender, D. E., Bernt, K., Berry, L., Betts, C. B., Bletz, J., Blise, K., Boire, A., Boland, G., Borowsky, A., Bosse, K., Bott, M., Boyden, E., Brooks, J., Bueno, R., Burlingame, E. A., Cai, Q., Campbell, J., Caravan, W., Cerami, E., Chaib, H., Chan, J. M., Chang, Y. H., Chatterjee, D., Chaudhary, O., Chen, A. A., Chen, B., Chen, C., Chen, C.-h., Chen, F., Chen, Y.-A., Chheda, M. G., Chin, K., Chiu, R., Chu, S.-K., Chuaqui, R., Chun, J., Cisneros, L., Coffey, R. J., Colditz, G. A., Cole, K., Collins, N., Contrepois, K., Coussens, L. M., Creason, A. L., Crichton, D., Curtis, C., Davidsen, T., Davies, S. R., de Bruijn, I., Dellostritto, L., De Marzo, A., Demir, E., DeNardo, D. G., Diep, D., Ding, L., Diskin, S., Doan, X., Drewes, J., Dubinett, S., Dyer, M., Egger, J., Eng, J., Engelhardt, B., Erwin, G., Esplin, E. D., Esserman, L., Felmeister, A., Feiler, H. S., Fields, R. C., Fisher, S., Flaherty, K., Flournoy, J., Ford, J. M., Fortunato, A., Frangieh, A., Frye, J. L., Fulton, R. S., Galipeau, D., Gan, S., Gao, J., Gao, L., Gao, P., Gao, V. R., Geiger, T., George, A., Getz, G., Ghosh, S., Giannakis, M., Gibbs, D. L., Gillanders, W. E., Goecks, J., Goedegebuure, S. P., Gould, A., Gowers, K., Gray, J. W., Greenleaf, W., Gresham, J., Guerriero, J. L., Guha, T. K., Guimaraes, A. R., Guinney, J., Gutman, D., Hacohen, N., Hanlon, S., Hansen, C. R., Harismendy, O., Harris, K. A., Hata, A., Hayashi, A., Heiser, C., Helvie, K., Herndon, J. M., Hirst, G., Hodi, F., Hollmann, T., Horning, A., Hsieh, J. J., Hughes, S., Huh, W. J., Hunger, S., Hwang, S. E., Iacobuzio-Donahue, C. A., Ijaz, H., Izar, B., Jacobson, C. A., Janes, S., Jané-Valbuena, J., Jayasinghe, R. G., Jiang, L., Johnson, B. E., Johnson, B., Ju, T., Kadara, H., Kaestner, K., Kagan, J., Kalinke, L., Keith, R., Khan, A., Kibbe, W., Kim, A. H., Kim, E., Kim, J., Kolodzie, A., Kopytra, M., Kotler, E., Krueger, R., Krysan, K., Kundaje, A., Ladabaum, U., Lake, B. B., Lam, H., Laquindanum, R., Lau, K. S., Laughney, A. M., Lee, H., Lenburg, M., Leonard, C., Leshchiner, I., Levy, R., Li, J., Lian, C. G., Lim, K.-H., Lin, J.-R., Lin, Y., Liu, Q., Liu, R., Lively, T., Longabaugh, W. J., Longacre, T., Ma, C. X., Macedonia, M. C., Madison, T., Maher, C. A., Maitra, A., Makinen, N., Makowski, D., Maley, C., Maliga, Z., Mallo, D., Maris, J., Markham, N., Marks, J., Martinez, D., Mashl, R. J., Masilionais, I., Mason, J., Massagué, J., Massion, P., Mattar, M., Mazurchuk, R., Mazutis, L., Mazzilli, S. A., McKinley, E. T., McMichael, J. F., Merrick, D., Meyerson, M., Miessner, J. R., Mills, G. B., Mills, M., Mondal, S. B., Mori, M., Mori, Y., Moses, E., Mosse, Y., Muhlich, J. L., Murphy, G. F., Navin, N. E., Nawy, T., Nederlof, M., Ness, R., Nevins, S., Nikolov, M., Nirmal, A. J., Nolan, G., Novikov, E., Oberdoerffer, P., O'Connell, B., Offin, M., Oh, S. T., Olson, A., Ooms, A., Ossandon, M., Owzar, K., Parmar, S., Patel, T., Patti, G. J., Pe'er, D., Pe'er, I., Peng, T., Persson, D., Petty, M., Pfister, H., Polyak, K., Pourfarhangi, K., Puram, S. V., Qiu, Q., Quintanal-Villalonga, Á., Raj, A., Ramirez-Solano, M., Rashid, R., Reeb, A. N., Regev, A., Reid, M., Resnick, A., Reynolds, S. M., Riesterer, J. L., Rodig, S., Roland, J. T., Rosenfield, S., Rotem, A., Roy, S., Rozenblatt-Rosen, O., Rudin, C. M., Ryser, M. D., Santagata, S., Santi-Vicini, M., Sato, K., Schapiro, D., Schrag, D., Schultz, N., Sears, C. L., Sears, R. C., Sen, S., Sen, T., Shalek, A., Sheng, J., Sheng, Q., Shoghi, K. I., Shrubsole, M. J., Shyr, Y., Sibley, A. B., Siex, K., Simmons, A. J., Singer, D. S., Sivagnanam, S., Slyper, M., Snyder, M. P., Sokolov, A., Song, S.-K., Sorger, P. K., Southard-Smith, A., Spira, A., Srivastava, S., Stein, J., Storm, P., Stover, E., Strand, S. H., Su, T., Sudar, D., Sullivan, R., Surrey, L., Suvà, M., Tan, K., Terekhanova, N. V., Ternes, L., Thammavong, L., Thibault, G., Thomas, G. V., Thorsson, V., Todres, E., Tran, L., Tyler, M., Uzun, Y., Vachani, A., Van Allen, E., Vandekar, S., Veis, D. J., Vigneau, S., Vossough, A., Waanders, A., Wagle, N., Wang, L.-B., Wendl, M. C., West, R., Williams, E. H., Wu, C.-y., Wu, H., Wu, H.-Y., Wyczalkowski, M. A., Xie, Y., Yang, X., Yapp, C., Yu, W., Yuan, Y., Zhang, D., Zhang, K., Zhang, M., Zhang, N., Zhang, Y., Zhao, Y., Zhou, D. C., Zhou, Z., Zhu, H., Zhu, Q., Zhu, X., Zhu, Y., and Zhuang, X. (2020). The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell*, 181(2):236–249.

Rubner, Y., Tomasi, C., and Guibas, L. (1998). A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66. Narosa Publishing House.

Ruddle, N. H. (2016). High endothelial venules and lymphatic vessels in tertiary lymphoid organs: Characteristics, functions, and regulation. *Frontiers in Immunology*, 7(NOV):491.

Rusan, N. M. and Peifer, M. (2008). Original CIN: reviewing roles for APC in chromosome instability. *Journal of Cell Biology*, 181(5):719–726.

Ryser, M. D., Mallo, D., Hall, A., Hardman, T., King, L. M., Tatishchev, S., Sorribes, I. C., Maley, C. C., Marks, J. R., Hwang, E. S., and Shibata, D. (2020). Minimal barriers to invasion during human colorectal tumor growth. *Nature Communications*, 11(1):1–10.

Ryser, M. D., Min, B. H., Siegmund, K. D., and Shibata, D. (2018). Spatial mutation patterns as markers of early colorectal tumor cell mobility. *Proceedings of the National Academy of Sciences of the United States of America*, 115(22):5774–5779.

Sakamoto, N., Feng, Y., Stolfi, C., Kurosu, Y., Green, M., Lin, J., Green, M. E., Sentani, K., Yasui, W., McMahon, M., Hardiman, K. M., Spence, J. R., Horita, N., Greenson, J. K., Kuick, R., Cho, K. R., and

Fearon, E. R. (2017). BRAFV600E cooperates with CDX2 inactivation to promote serrated colorectal tumorigenesis. *eLife*, 6.

Satas, G., Zaccaria, S., El-Kebir, M., and Raphael, B. J. (2021). DeCiFering the elusive cancer cell fraction in tumor heterogeneity and evolution. *Cell Systems*, 12(10):1004–1018.e10.

Schapiro, D., Yapp, C., Sokolov, A., Reynolds, S. M., Chen, Y. A., Sudar, D., Xie, Y., Muhlich, J., Arias-Camison, R., Arena, S., Taylor, A. J., Nikolov, M., Tyler, M., Lin, J. R., Burlingame, E. A., Abravanel, D. L., Achilefu, S., Ademuyiwa, F. O., Adey, A. C., Aft, R., Ahn, K. J., Alikarami, F., Alon, S., Ashenberg, O., Baker, E., Baker, G. J., Bandyopadhyay, S., Bayguinov, P., Beane, J., Becker, W., Bernt, K., Betts, C. B., Bletz, J., Blosser, T., Boire, A., Boland, G. M., Boyden, E. S., Bucher, E., Bueno, R., Cai, Q., Cambuli, F., Campbell, J., Cao, S., Caravan, W., Chaligné, R., Chan, J. M., Chasnoff, S., Chatterjee, D., Chen, A. A., Chen, C., hui Chen, C., Chen, B., Chen, F., Chen, S., Chheda, M. G., Chin, K., Cho, H., Chun, J., Cisneros, L., Coffey, R. J., Cohen, O., Colditz, G. A., Cole, K. A., Collins, N., Cotter, D., Coussens, L. M., Coy, S., Creason, A. L., Cui, Y., Zhou, D. C., Curtis, C., Davies, S. R., Bruijn, I., Delorey, T. M., Demir, E., Denardo, D., Diep, D., Ding, L., DiPersio, J., Dubinett, S. M., Eberlein, T. J., Eddy, J. A., Esplin, E. D., Factor, R. E., Fatahalian, K., Feiler, H. S., Fernandez, J., Fields, A., Fields, R. C., Fitzpatrick, J. A., Ford, J. M., Franklin, J., Fulton, B., Gaglia, G., Galdieri, L., Ganesh, K., Gao, J., Gaudio, B. L., Getz, G., Gibbs, D. L., Gillanders, W. E., Goecks, J., Goodwin, D., Gray, J. W., Greenleaf, W., Grimm, L. J., Gu, Q., Guerriero, J. L., Guha, T., Guimaraes, A. R., Gutierrez, B., Hacohen, N., Hanson, C. R., Harris, C. R., Hawkins, W. G., Heiser, C. N., Hoffer, J., Hollmann, T. J., Hsieh, J. J., Huang, J., Hunger, S. P., Hwang, E. S., Iacobuzio-Donahue, C., Iglesia, M. D., Islam, M., Izar, B., Jacobson, C. A., Janes, S., Jayasinghe, R. G., Jeudi, T., Johnson, B. E., Johnson, B. E., Ju, T., Kadara, H., Karnoub, E. R., Karpova, A., Khan, A., Kibbe, W., Kim, A. H., King, L. M., Kozlowski, E., Krishnamoorthy, P., Krueger, R., Kundaje, A., Ladabaum, U., Laquindanum, R., Lau, C., Lau, K. S. K., LeBoeuf, N. R., Lee, H., Lenburg, M., Leshchiner, I., Levy, R., Li, Y., Lian, C. G., Liang, W. W., Lim, K. H., Lin, Y., Liu, D., Liu, Q., Liu, R., Lo, J., Lo, P., Longabaugh, W. J., Longacre, T., Luckett, K., Ma, C., Maher, C., Maier, A., Makowski, D., Maley, C., Maliga, Z., Manoj, P., Maris, J. M., Markham, N., Marks, J. R., Martinez, D., Mashl, J., Masilionis, I., Massague, J., Mazurowski, M. A., McKinley, E. T., McMichael, J., Meyerson, M., Mills, G. B., Mitri, Z. I., Moorman, A., Mudd, J., Murphy, G. F., Deen, N. N. A., Navin, N. E., Nawy, T., Ness, R. M., Nevins, S., Nirmal, A. J., Novikov, E., Oh, S. T., Oldridge, D. A., Owzar, K., Pant, S. M., Park, W., Patti, G. J., Paul, K., Pelletier, R., Persson, D., Petty, C., Pfister, H., Polyak, K., Puram, S. V., Qiu, Q., Villalonga, Á. Q., Ramirez, M. A., Rashid, R., Reeb, A. N., Reid, M. E., Remsik, J., Riesterer, J. L., Risom, T., Ritch, C. C., Rolong, A., Rudin, C. M., Ryser, M. D., Sato, K., Sears, C. L., Semenov, Y. R., Shen, J., Shoghi, K. I., Shrubsole, M. J., Shyr, Y., Sibley, A. B., Simmons, A. J., Sinha, A., Sivagnanam, S., Song, S. K., Southar-Smith, A., Spira, A. E., Cyr, J. S., Stefankiewicz, S., Storrs, E. P., Stover, E. H., Strand, S. H., Straub, C., Street, C., Su, T., Surrey, L. F., Suver, C., Tan, K., Terekhanova, N. V., Ternes, L., Thadi, A., Thomas, G., Tibshirani, R., Umeda, S., Uzun, Y., Vallius, T., Van Allen, E. R., Vandekar, S., Vega, P. N., Veis, D. J., Vennam, S., Verma, A., Vigneau, S., Wagle, N., Wahl, R., Walle, T., Wang, L. B., Warchol, S., Washington, M. K., Watson, C., Weimer, A. K., Wendl, M. C., West, R. B., White, S., Windon, A. L., Wu, H., Wu, C. Y., Wu, Y., Wyczalkowski, M. A., Xu, J., Yao, L., Yu, W., Zhang, K., Zhu, X., Chang, Y. H., Farhi, S. L., Thorsson, V., Venkatamohan, N., Drewes, J. L., Pe'er, D., Gutman, D. A., Herrmann, M. D., Gehlenborg, N., Bankhead, P., Roland, J. T., Herndon, J. M., Snyder, M. P., Angelo, M., Nolan, G., Swedlow, J. R., Schultz, N., Merrick, D. T., Mazzili, S. A., Cerami, E., Rodig, S. J., Santagata, S., and Sorger, P. K. (2022). MITI minimum information guidelines for highly multiplexed tissue images. *Nature Methods*, 19(3):262–267.

Schürch, C. M., Bhate, S. S., Barlow, G. L., Phillips, D. J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D. R., Samusik, N., Goltsev, Y., and Nolan, G. P. (2020). Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell*, 182(5):1341–1359.e19.

Sheffer, M., Bacolod, M. D., Zuk, O., Giardina, S. F., Pincas, H., Barany, F., Paty, P. B., Gerald, W. L., Notterman, D. A., and Domany, E. (2009). Association of survival and disease progression with chromosomal instability: A genomic exploration of colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 106(17):7131–7136.

Shi, R., Zhang, Z., Zhu, A., Xiong, X., Zhang, J., Xu, J., Sy, M. S., and Li, C. (2022). Targeting type I collagen for cancer treatment. *International Journal of Cancer*, 151(5):665–683.

Shibata, D. (2020). Visualizing Human Colorectal Cancer Intratumor Heterogeneity with Phylogeography. *iScience*, 23(7).

Shoshkes-Carmel, M., Wang, Y. J., Wangensteen, K. J., Tóth, B., Kondo, A., Massassa, E. E., Itzkovitz, S., and Kaestner, K. H. (2018). Subepithelial telocytes are an important source of Wnts that supports intestinal crypts. *Nature*, 557(7704):242–246.

Sieber, O. M., Heinimann, K., Gorman, P., Lamlum, H., Crabtre, M., Simpson, C. A., Davies, D., Neale, K., Hodgson, S. V., Roylance, R. R., Phillips, R. K., Bodmer, W. F., and Tomlinson, I. P. (2002). Analysis of chromosomal instability in human colorectal adenomas with two mutational hits at APC. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):16910–16915.

Sigurdsson, S., Bödvarsdottir, S. K., Anamthawat-Jonsson, K., Steinarsdottir, M., Jonasson, J. G., Ögmundsdottir, H. M., and Eyfjörd, J. E. (2000). p53 Abnormality and Chromosomal Instability in the Same Breast Tumor Cells. *Cancer Genetics and Cytogenetics*, 121:150–155.

Simmons, A. J. and Lau, K. S. (2017). Deciphering tumor heterogeneity from FFPE tissues: Its promise and challenges. *Molecular & Cellular Oncology*, 4(1):e1260191.

Simmons, A. J. and Lau, K. S. (2022). Dissociation and inDrops microfluidic encapsulation of human gut tissues for single-cell atlasing studies. *STAR Protocols*, 3(3):101570.

Sipos, F. and Muzes, G. (2011). Isolated lymphoid follicles in colon: switch points between inflammation and colorectal cancer? *World journal of gastroenterology*, 17(13):1666–1673.

Sofroniew, N., Lambert, T., Evans, K., Nunez-Iglesias, J., Bokota, G., Winston, P., Peña-Castellanos, G., Yamauchi, K., Bussonnier, M., Doncila Pop, D., Can Solak, A., Liu, Z., Wadhwa, P., Burt, A., Buckley, G., Sweet, A., Migas, L., Hilsenstein, V., Gaifas, L., Bragantini, J., Rodríguez-Guerra, J., Muñoz, H., Freeman, J., Boone, P., Lowe, A., Gohlke, C., Royer, L., PIERRÉ, A., Har-Gil, H., and McGovern, A. (2022). napari: a multi-dimensional image viewer for Python.

Sorzano, C. O. S., Vargas, J., and Montano, A. P. (2014). A survey of dimensionality reduction techniques. *arXiv*.

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., and Curtis, C. (2015). A big bang model of human colorectal tumor growth. *Nature Genetics*, 47(3):209–216.

Southard-Smith, A. N., Simmons, A. J., Chen, B., Jones, A. L., Ramirez Solano, M. A., Vega, P. N., Scurrah, C. R., Zhao, Y., Brenan, M. J., Xuan, J., Shrubsole, M. J., Porter, E. B., Chen, X., Brenan, C. J., Liu, Q., Quigley, L. N., and Lau, K. S. (2020). Dual indexed library design enables compatibility of in-Drop single-cell RNA-sequencing with exAMP chemistry sequencing platforms. *BMC Genomics*, 21(1):1–15.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., and Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.

Steitz, A. M., Steffes, A., Finkernagel, F., Unger, A., Sommerfeld, L., Jansen, J. M., Wagner, U., Graumann, J., Müller, R., and Reinartz, S. (2020). Tumor-associated macrophages promote ovarian cancer cell migration by secreting transforming growth factor beta induced (TGFBI) and tenascin C. *Cell Death & Disease*, 11(4):1–15.

Stoler, D. L., Chen, N., Basik, M., Kahlenberg, M. S., Rodriguez-Bigas, M. A., Petrelli, N. J., and Anderson, G. R. (1999). The onset and extent of genomic instability in sporadic colorectal tumor progression. *Proceedings of the National Academy of Sciences of the United States of America*, 96(26):15121–15126.

Stunnenberg, H., Atlasy, N., Bujko, A., Brazda, P. B., Janssen-Megens, E., Bakkenvold, E. S., Jahnsen, J., and Jahsen, F. (2019). Single cell transcriptome atlas of immune cells in human small intestine and in celiac disease. *bioRxiv*, page 721258.

Su, H., Yang, F., Fu, R., Trinh, B., Sun, N., Liu, J., Kumar, A., Baglieri, J., Siruno, J., Le, M., Li, Y., Dozier, S., Nair, A., Filliol, A., Sinchai, N., Rosenthal, S. B., Santini, J., Metallo, C. M., Molina, A., Schwabe, R. F., Lowy, A. M., Brenner, D., Sun, B., and Karin, M. (2022). Collagenolysis-dependent DDR1 signalling dictates pancreatic cancer outcome. *Nature*, pages 1–7.

Sun, R., Hu, Z., Sottoriva, A., Graham, T. A., Harpak, A., Ma, Z., Fischer, J. M., Shibata, D., and Curtis, C. (2017). Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nature Genetics*, 49(7):1015–1024.

Sun, X., Wu, B., Chiang, H.-C., Deng, H., Zhang, X., Xiong, W., Liu, J., Rozeboom, A. M., Harris, B. T., Blommaert, E., Gomez, A., Espin Garcia, R., Zhou, Y., Mitra, P., Prevost, M., Zhang, D., Banik, D., Isaacs, C., Berry, D., Lai, C., Chaldekas, K., Latham, P. S., Brantner, C. A., Popratiloff, A., Jin, V. X., Zhang, N., Hu, Y., Angel Pujana, M., Curiel, T. J., An, Z., and Li, R. (2021). Tumour DDR1 promotes collagen fibre alignment to instigate immune exclusion. *Nature*, 599:673.

Sunkin, S. M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T. L., Thompson, C. L., Hawrylycz, M., and Dang, C. (2013). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Research*, 41(D1):D996–D1008.

Suvà, M. L. and Tirosh, I. (2019). Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges.

Tait, S. W. and Green, D. R. (2010). Mitochondria and cell death: Outer membrane permeabilization and beyond.

Takai, K., Drain, A. P., Lawson, D. A., Littlepage, L. E., Karpuj, M., Kessenbrock, K., Le, A., Inoue, K., Weaver, V. M., and Werb, Z. (2018). Discoidin domain receptor 1 (DDR1) ablation promotes tissue fibrosis and hypoxia to induce aggressive basal-like breast cancers. *Genes & Development*, 32(3-4):244–257.

Talevich, E., Shain, A. H., Botton, T., and Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Computational Biology*, 12(4):e1004873.

Tamura, M., Tanaka, S., Fujii, T., Aoki, A., Komiyama, H., Ezawa, K., Sumiyama, K., Sagai, T., and Shiroishi, T. (2007). Members of a novel gene family, Gsdm, are expressed exclusively in the epithelium of the skin and gastrointestinal tract in a highly tissue-specific manner. *Genomics*, 89(5):618–629.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tighe, A., Johnson, V. L., Albertella, M., and Taylor, S. S. (2001). Aneuploid colon cancer cells have a robust spindle checkpoint. *EMBO reports*, 2(7):609–614.

Townes, F. W. and Engelhardt, B. E. (2021). Nonnegative spatial factorization. *arXiv*.

Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):295.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1).

Tsuyuzaki, K., Sato, H., Sato, K., and Nikaido, I. (2020). Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biology*, 21(1):1–17.

Tu, M. M., Lee, F. Y., Jones, R. T., Kimball, A. K., Saravia, E., Graziano, R. F., Coleman, B., Menard, K., Yan, J., Michaud, E., Chang, H., Abdel-Hafiz, H. A., Rozhok, A. I., Duex, J. E., Agarwal, N., Chauca-Diaz, A., Johnson, L. K., Ng, T. L., Cambier, J. C., Clambey, E. T., Costello, J. C., Korman, A. J., and Theodorescu, D. (2019). Targeting DDR2 enhances tumor response to anti–PD-1 immunotherapy. *Science Advances*, 5(2).

Tumbarello, D. A., Temple, J., and Brenton, J. D. (2012). ß3 integrin modulates transforming growth factor beta induced (TGFBI) function and paclitaxel response in ovarian cancer cells. *Molecular cancer*, 11(1):1–15.

Turajlic, S., Sottoriva, A., Graham, T., and Swanton, C. (2019). Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics*, 20(7):404–416.

Tusi, B. K., Wolock, S. L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J., Klein, A. M., and Socolovsky, M. (2018). Emergence of the erythroid lineage from multipotent hematopoiesis. *bioRxiv*, page 261941.

Tytgat, K. M., Büller, H. A., Opdam, F. J., Kim, Y. S., Einerhand, A. W., and Dekker, J. (1994). Biosynthesis of human colonic mucin: Muc2 is the prominent secretory mucin. *Gastroenterology*, 107(5):1352–1363.

Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., and Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, 11(1):1–13.

Van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Vandekar, S., Tao, R., and Blume, J. (2020). A Robust Effect Size Index. *Psychometrika*, 85(1):232–246.

Vega, P. N., Nilsson, A., Kumar, M. P., Niitsu, H., Simmons, A. J., Ro, J., Wang, J., Chen, Z., Joughin, B. A., Li, W., McKinley, E. T., Liu, Q., Roland, J. T., Washington, M. K., Coffey, R. J., Lauffenburger, D. A., and Lau, K. S. (2022). Cancer-Associated Fibroblasts and Squamous Epithelial Cells Constitute a Unique Microenvironment in a Mouse Model of Inflammation-Induced Colon Cancer. *Frontiers in Oncology*, 12:1888.

Venkatesan, S. and Swanton, C. (2016). Tumor evolutionary principles: How intratumor heterogeneity influences cancer treatment and outcome. *American Society of Clinical Oncology Educational Book*, 36:e141–e149. PMID: 27249716.

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J. F., Gould, J., Griffin, G. K., Borg, Å., Ronaghi, M., Frisén, J., Lundeberg, J., Regev, A., and Ståhl, P. L. (2019). High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*.

Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G., and Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987.

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods*, 14(4):414–416.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164.

Wang, M. (2022). DPEP1 mediates neutrophil and monocyte influx. *Nature Reviews Nephrology*, 18(4):199–199.

Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G. P., Bava, F.-A., and Deisseroth, K. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691.

Warchol, S., Krueger, R., Nirmal, A. J., Gaglia, G., Jessup, J., Ritch, C. C., Hoffer, J., Muhlich, J., Burger, M. L., Jacks, T., Santagata, S., Sorger, P. K., and Pfister, H. (2022). Visinity: Visual Spatial Neighborhood Analysis for Multiplexed Tissue Imaging Data. *IEEE Transactions on Visualization and Computer Graphics*, PP.

Waskom, M., Botvinnik, O., Hobson, P., Cole, J. B., Halchenko, Y., Hoyer, S., Miles, A., Augspurger, T., Yarkoni, T., Megies, T., Coelho, L. P., Wehner, D., Cynddl, Ziegler, E., Diego0020, Zaytsev, Y. V., Hoppe, T., Seabold, S., Cloud, P., Koskinen, M., Meyer, K., Qalieh, A., and Allan, D. (2014). seaborn: v0.5.0 (November 2014).

Weber, L. M. and Robinson, M. D. (2016). Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, 89(12):1084–1096.

Ween, M. P., Oehler, M. K., and Ricciardelli, C. (2012). Transforming Growth Factor-Beta-Induced Protein (TGFBI)/(βig-H3): A Matrix Protein with Dual Functions in Ovarian Cancer. *International Journal of Molecular Sciences*, 13(8):10461–10477.

Werman, M., Peleg, S., and Rosenfeld, A. (1985). A distance metric for multidimensional histograms. *Computer Vision, Graphics, and Image Processing*, 32(3):328–336.

West, J., Schenck, R. O., Gatenbee, C., Robertson-Tessi, M., and Anderson, A. R. A. (2021). Normal tissue architecture determines the evolutionary course of cancer. *Nature Communications*, 12(1):1–9.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15.

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59.

Woodford-Richens, K. L., Rowan, A. J., Gorman, P., Halford, S., Bicknell, D. C., Wasan, H. S., Roylance, R. R., Bodmer, W. F., and Tomlinson, I. P. (2001). SMAD4 mutations in colorectal cancer probably occur before chromosomal instability, but after divergence of the microsatellite instability pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):9719–9723.

Wu, Y., Yang, S., Ma, J., Chen, Z., Song, G., Rao, D., Cheng, Y., Huang, S., Liu, Y., Jiang, S., Liu, J., Huang, X., Wang, X., Qiu, S., Xu, J., Xi, R., Bai, F., Zhou, J., Fan, J., Zhang, X., and Gao, Q. (2022a). Spatiotemporal Immune Landscape of Colorectal Cancer Liver Metastasis at Single-Cell Level. *Cancer Discovery*, 12(1):134–153.

Wu, Z., Trevino, A. E., Wu, E., Swanson, K., Kim, H. J., D'Angio, H. B., Preska, R., Charville, G. W., Dalerba, P. D., Egloff, A. M., Uppaluri, R., Duvvuri, U., Mayer, A. T., and Zou, J. (2022b). Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering*, 6(12):1435–1448.

Yang, K., Popova, N. V., Wan, C. Y., Lozonschi, I., Tadesse, S., Kent, S., Bancroft, L., Matise, I., Cormier, R. T., Scherer, S. J., Edelmann, W., Lipkin, M., Augenlicht, L., and Velcich, A. (2008). Interaction of Muc2 and Apc on Wnt Signaling and in Intestinal Tumorigenesis: Potential Role of Chronic Inflammation. *Cancer Research*, 68(18):7313–7322.

Yang, S., Corbett, S. E., Koga, Y., Wang, Z., Johnson, W. E., Yajima, M., and Campbell, J. D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology*, 21(1):57.

Yoosuf, N., Navarro, J. F., Salmén, F., Ståhl, P. L., and Daub, C. O. (2020). Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. *Breast Cancer Research*, 22(1):6.

Young, M. D. and Behjati, S. (2018). SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *bioRxiv*, page 303727.

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., and Linnarsson, S. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (New York, N.Y.)*, 347(6226):1138–42.

Zeng, C., Qi, G., Shen, Y., Li, W., Zhu, Q., Yang, C., Deng, J., Lu, W., Liu, Q., and Jin, J. (2022). DPEP1 promotes drug resistance in colon cancer cells by forming a positive feedback loop with ASCL2. *Cancer Medicine*, 00:1–13.

Zhang, Q., Jeppesen, D. K., Higginbotham, J. N., Graves-Deal, R., Trinh, V. Q., Ramirez, M. A., Sohn, Y., Neininger, A. C., Taneja, N., McKinley, E. T., Niitsu, H., Cao, Z., Evans, R., Glass, S. E., Ray, K. C., Fissell, W. H., Hill, S., Rose, K. L., Huh, W. J., Washington, M. K., Ayers, G. D., Burnette, D. T., Sharma, S., Rome, L. H., Franklin, J. L., Lee, Y. A., Liu, Q., and Coffey, R. J. (2021). Supermeres are functional extracellular nanoparticles replete with disease biomarkers and therapeutic targets. *Nature Cell Biology*, 23(12):1240–1254.

Zhang, Y., Zheng, L., Zhang, L., Hu, X., Ren, X., and Zhang, Z. (2019). Deep single-cell RNA sequencing data of individual T cells from treatment-naïve colorectal cancer patients. *Scientific Data*.

Zhao, E., Stone, M. R., Ren, X., Guenthoer, J., Smythe, K. S., Pulliam, T., Williams, S. R., Uytingco, C. R., Taylor, S. E., Nghiem, P., Bielas, J. H., and Gottardo, R. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39(11):1375–1384.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):1–12.

Zhou, R., Zhang, J., Zeng, D., Sun, H., Rong, X., Shi, M., Bin, J., Liao, Y., and Liao, W. (2019). Immune cell infiltration as a biomarker for the diagnosis and prognosis of stage I–III colon cancer. *Cancer Immunology, Immunotherapy*, 68(3):433–442.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Appendices

A2 Appendix for Chapter 2: A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques

A2.1 Methods

A2.1.1 Resource Availability

The code generated during this study is available at https://github.com/KenLauLab/DR-structure-preservation.

A2.1.2 Cell Filtering

Raw counts expression matrices downloaded from GEO (accession IDs GSM1626793, GSM2743164) were filtered for high-quality cells prior to downstream analysis. The cumulative sum of total UMI counts for each cell was plotted along with the slope of the secant line to the curve as a function of rank-ordered cell. The distance between these two curves was used as a metric for determining the rate of diminishing cell quality. The cell number at which this distance was 50 % of its maximum was chosen as a cutoff, with cells contributing less UMI counts were removed. Next, a 100-component PCA and UMAP with n_neighbors value of 0.5 % of the total cells in the dataset were used to visualize cell populations and manually gate out clusters containing high mitochondrial counts, indicating dead cells. Analysis performed using `scanpy` (Wolf et al., 2018).

A2.1.3 Clustering

PhenoGraph (Levine et al., 2015) was used to perform Louvain clustering on both datasets in Python. To create coarse, ground-truth clusters, the algorithm was run on 100 principal components of all genes in each dataset. For the retina data, 100 PCs of 20,478 genes explained 33.5 % of the variance in the dataset. For the colon data, 100 PCs of 25,505 genes explained 54.0 % of the variance. k values of 50 and 100 for generating the Knn graph to seed the Louvain algorithm for the retina and colon datasets, respectively, were chosen to provide coarse clustering of major cell types. Nine resulting clusters for the retina dataset and six resulting clusters in the colon dataset were analyzed by Seurat's FindAllMarkers and DoHeatmap functions (Butler et al., 2018) to obtain visualizations of up- and down-regulated genes in each cluster (Figure 2.3A,D).

A2.1.4 Dimensionality Reduction

All dimensionality reduction was performed on feature-selected data containing the most variable genes in each dataset. Genes were rank-ordered by variance using the Pandas (version 0.22.0) (Mckinney, 2010) DataFrame.var function in Python, and the top 500 were chosen. Each dimensionality reduction technique was run "out-of-the-box" with default parameters on the feature-selected data. DCA, scvis, scVI, ZINB-WaVE and GLM-PCA take raw, unnormalized counts as input. Developers of ZIFA recommend a log2 transformation of counts, which we first normalized to the maximum UMI count within each cell. Arcsinh-transformed counts normalized to the maximum UMI count in each cell were used for all other methods (t-SNE, FIt-SNE, UMAP, SIMLR, PCA).

A2.1.5 Visualization

Cumulative cell distance distributions were plotted from the upper triangle of symmetrical cell distance matrices (using `triu_indices` function from the `numpy` Python package (version 1.16.3) (Oliphant, 2007)). The histogram and cumsum functions from numpy were used to plot cumulative distribution functions using n/100 bins, where n is the length of the flattened distance vector. Unique distance correlation was visualized using the JointGrid and kdeplot functions from the seaborn package (version 0.9.0) (Waskom et al., 2014), as well as the `pyplot.hist2d` function from the `matplotlib` package (version 3.0.3) (Hunter, 2007). Cluster topology graphs were plotted using the network function `draw_networkx`.

A2.1.6 `Splatter` Simulation

Simulated single-cell datasets were generated using the `Splatter` package (1.8.0) (Zappia et al., 2017). Continuous dataset was generated with 500 features (`nGenes`) and 3060 observations (`batchCells`), with a `lib.loc` value of 10 and `lib.scale` value of 0.05 to generate data close to observed counts distributions from scRNA-seq. The simulation defined three paths with equal `group.prob` values (0.3333333) originating at the same state (`path.from = c(0,0,0)`). Each path had `path.nSteps` value of 1000 indicating as many possible continuous expression states emanating from the common origin state. These step values are used as pseudotime (PT) measures in our analysis. Discrete simulation data was generated by simply excluding all cells with PT values less than 400, eliminating the common central state. The resulting dataset had 1873 observations for the 500 features of the continuous simulation. When evaluating embeddings of these simulated data, native cell-cell distance distributions were replaced with cell-cell PT sums normalized in the same fashion. These analyses were only performed pairwise between cells in each of the three developmental paths.

A2.1.7 Theoretical basis for difference in dimension reduction performance across single-cell modalities

Based on prior evidence and common practice in the field, we aim to validate our metrics and address the challenges our results pose to current conceptions about popular dimensionality reduction tools. UMAP was benchmarked against t-SNE, FIt-SNE and scvis on three datasets (Becht et al., 2018): two CyTOF – Samusik and Wong – and one scRNA-seq – Han mouse cell atlas. We applied our structural preservation framework to the Samusik and Han datasets (Weber and Robinson, 2016; Han et al., 2018), making interesting observations that both substantiate our metrics and emphasize a major takeaway from this study.

Figure S5A compares t-SNE to UMAP on the hematopoietic subset of the Han scRNA-seq dataset. Interestingly, the two methods perform very similarly, with UMAP slightly outperforming t-SNE as described in Becht et al., 2018 Figure S5B-D reflect a similar analysis of the Samusik CyTOF dataset, where there is a clear improvement in unique distance preservation by UMAP over t-SNE, indicated by a strong increase in cell-cell distance correlation. Again, this result agrees with Becht and coworkers, who also correlated cell distances to show vast improvement over t-SNE and other methods. Nonetheless, our framework identifies a marginal increase in EMD for UMAP over t-SNE, indicating a higher degree of global structural distortion, likely due to more compact clustering by UMAP (Figure S5B). We propose that improved performance of UMAP applied to CyTOF data is due to the input cell distribution.

Because mass cytometry measurements have a larger dynamic range and lower dropout rate than scRNA-seq, the overall variance of cell distance distributions from CyTOF is greater (Figure S4D). This allows for better discrimination between ”large” (global) and ”small” (local) distances in the native space. With this in mind, we can explore the mathematical basis for global distance preservation in UMAP versus t-SNE to help explain why these advantages may not always be clearly observed when applied to scRNA-seq data.

First, t-SNE models the conditional probability that any two points x_i,x_j would be neighbors if neighbors were chosen in proportion to a Gaussian probability density function at x_i (Van der Maaten and Hinton, 2008). We can simplify this probability density function to Equation S1 under the right parameter conditions. Conditional probability for low-dimensional distances between points y_i,y_j, is modeled by the Student t-distribution, simplified in Equation S2.

$$p_{ij} = exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \approx exp\left(-\left\|x_i - x_j\right\|^2\right) \tag{S1}$$

$$q_{ij} = \left(1 + \left\|y_i - y_j\right\|^2\right)^{-1} \tag{S2}$$

UMAP models distance probabilities very similarly. Equations S3 and S5 show simplification of high-dimensional conditional probabilities given the defined symmetrization used by UMAP (Equation S4). Equation S6 shows UMAP’s low-dimensional distance probability model, which is not exactly the Student t-distribution, but approximates to it under the right parameters a and b (McInnes et al., 2018).

$$p_{i|j} = exp\left(-\frac{\left\|x_i - x_j\right\| - \rho_i}{\sigma_i}\right) \approx exp\left(-\left\|x_i - x_j\right\|\right) \tag{S3}$$

$$p_{ij} = p_{i|j} + p_{j|i} - p_{i|j}p_{j|i} \tag{S4}$$

$$p_{ij} \approx 2exp\left(-\left\|x_i - x_j\right\|\right) - exp\left(-\left\|x_i - x_j\right\| - \left\|x_j - x_i\right\|\right) \approx 2exp\left(-\left\|x_i - x_j\right\|\right) \approx exp\left(-\left\|x_i - x_j\right\|^2\right) \tag{S5}$$

$$q_{ij} = \left(1 + a\left\|y_i - y_j\right\|^{2b}\right)^{-1} \approx \left(1 + \left\|y_i - y_j\right\|^2\right)^{-1} \tag{S6}$$

The cost function for optimization of t-SNE coordinates is Kullback-Leibler divergence (D_{KL}), defined in Equation S7 for X representing the set of cell-cell distances in the high-dimensional (native) space, and Y representing the set of corresponding distances in the low-dimensional (latent) space. We additionally notate p_{ij} and q_{ij} as $P(X)$ and $Q(Y)$ for all x_i,$x_j \in X$ and y_i,$y_j \in Y$, respectively. The first term of this equation is close to 0 for both large and small X, so you can approximate D_{KL} by the second term alone and substitute p_{ij} and q_{ij} from Equations S5 and S6 for $P(X)$ and $Q(Y)$ (Equation S8).

$$D_{KL}(X,Y) = P(X)log\left(\frac{P(X)}{Q(Y)}\right) = P(X)logP(X) - P(X)logQ(Y) \tag{S7}$$

$$D_{KL}(X,Y) \approx -P(X)logQ(Y) \approx e^{-X^2}log\left(1 + Y^2\right) \tag{S8}$$

Evaluating the limits of D_{KL} in Equation S8, there is a large penalty at small X and large Y, but for large X the penalty is marginal regardless of Y (Equations S9, S10).

$$\lim_{x\to 0} D_{KL}(X,Y) \approx log\left(1 + Y^2\right) \tag{S9}$$

$$\lim_{x\to\infty} D_{KL}(X,Y) \approx 0 \tag{S10}$$

On the other hand, UMAP uses cross entropy (CE) as its cost function (Equation S11). This function behaves the same as t-SNE for small X, shown in Equations S12 and S13. The difference arises at large X, where the penalty becomes very large for small Y (Equation S14).

$$CE(X,Y) = P(X)log\left(\frac{P(X)}{Q(Y)}\right) + (1 - P(X))\,log\left(\frac{1 - P(X)}{1 - Q(Y)}\right) \tag{S11}$$

$$CE(X,Y) \approx e^{-X^2} log\left(1+Y^2\right) + \left(1-e^{-X^2}\right) log\left(\frac{1+Y^2}{Y^2}\right) \tag{S12}$$

$$\lim_{x\to 0} CE(X,Y) \approx log\left(1+Y^2\right) \tag{S13}$$

$$\lim_{x\to \infty} CE(X,Y) \approx log\left(\frac{1+Y^2}{Y^2}\right) \tag{S14}$$

Undoubtedly, the CE cost function has theoretical advantages over D_{KL}, and strikes a balance between local and global distance preservation for more uniformly distributed samples. So why is performance seemingly equivalent on scRNA-seq data such as our colon and retina datasets, as well as the Han hematopoietic dataset evaluated by Becht et al., 2018? We propose that the negative binomial nature of scRNA-seq data causes the native set of cell distances X to have a small variance (even following normalization and log or arcsinh transformation), resulting in a similar cost function profile in both t-SNE and UMAP. Conversely, CyTOF data with more variant cell distances are predisposed to favorable optimization by the CE cost function. To further test this hypothesis with a simpler example, we generated a synthetic dataset consisting of two 1,000-point Gaussians in three-dimensional space (Figure S5J). The resulting cell distance distribution is bimodal, consisting of local distances between cells in each Gaussian and global distances from one point-cloud to the other. UMAP outperformed t-SNE drastically in correlation and EMD values, with only a slight loss in Knn preservation (to be expected, as t-SNE favors small distances in its optimization) (Figure S5K,L). This result corroborates prior evidence that UMAP distinguishes itself greatly on datasets with clear ”local” and ”global” distance populations, owing to its cost function. Consequently, we assert that behavior of dimensionality reduction methods is predominantly governed by the input data itself, and we encourage evaluation of these techniques on data types, datasets, and normalization and preprocessing approaches specific to an intended application.

A2.1.8 Distance Metric Calculations

Pearson correlation was performed for Euclidean cell-cell distance preservation analysis using the `scipy.stats.pearsonr` function from the `scipy` package (version 1.1.0) (Oliphant, 2007). The `wasserstein_1d` function from the `POT` package (version 0.6.0) (Flamary and Courty, 2017) was used to calculate the Earth Mover’s Distance between vectors containing unique distances between all cells in the dataset (upper triangle of distance matrix), except for local comparisons between clusters, where the entire flattened matrix was used as the cell-cell distance matrices are not symmetrical. K nearest-neighbor graphs were constructed using the `scikit-learn` (version 0.20.0) (Pedregosa et al., 2011) function `sklearn.neighbors.kneighbors_graph`. Knn preservation was calculated as the percentage of elements in the Knn graph matrix that are conserved. Cluster centroid topology graphs and minimum spanning trees were generated using `networkx` (version 2.2) (Hagberg et al., 2008).

A2.2 Supplemental tables and figures

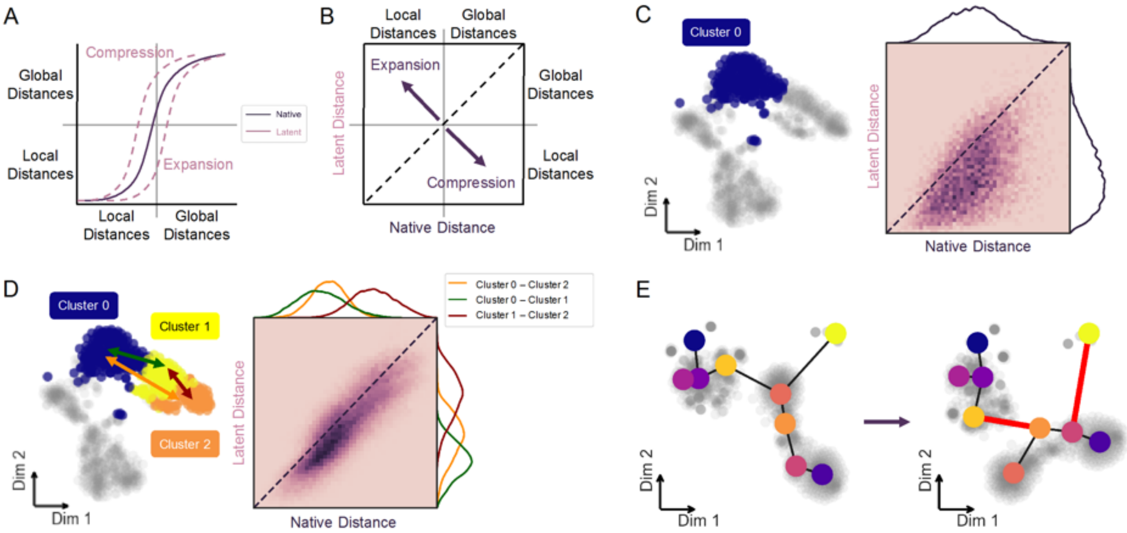


Figure S2: Interpretation of data structure preservation analyses.
(A) Small distances in cumulative distance distribution represent local cell similarity (within cluster), while large distances represent global relationships and arrangement of data (between clusters). A distribution shift left indicates compression of distances from native to latent space, while a shift right results from expansion or exaggeration of native distances.
(B) Correlation of latent to native distances; dispersion below identity line (dashed) indicates compression of distances from native to latent space, while dispersion above identity results from expansion of native distances in low-dimensional space.
(C) Substructure analysis uses same framework as Figure 1 on isolated subset of data to measure intra-cluster distance preservation and determine contribution to global structure.
(D) Distribution of distances from all cells in one cluster to another define relative substructure. Inter-cluster distances are measured pairwise to interrogate cluster arrangement in latent compared to native space.
(E) Evaluation of coarse global cluster topology using minimum spanning tree (MST) graph constructed from cluster centroids and their pairwise distances in native and latent dimensions. Black edges between centroids denote MST. Edges not present in native MST graph are highlighted red, indicating relative rearrangement of clusters following dimension reduction.

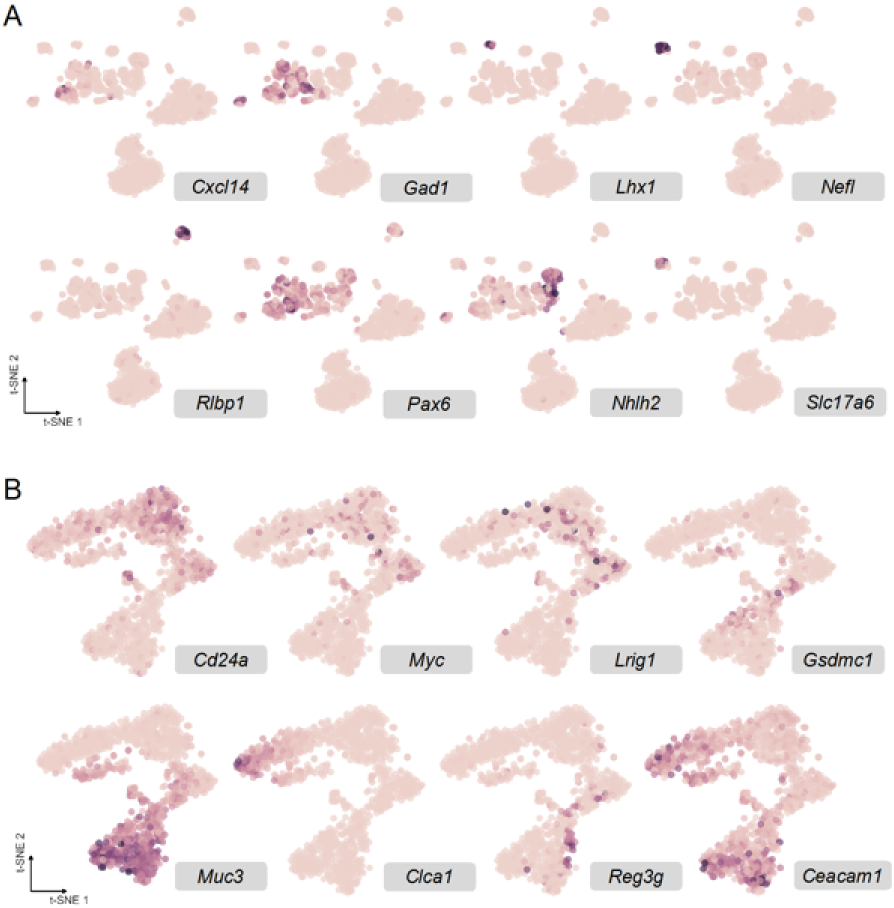


Figure S3: t-SNE visualizations from Figure 2.3 with overlay of arcsinh-normalized expression of marker genes (Macosko et al., 2015; Herring et al., 2018) used to assign cell type to Louvain clusters for retina (A) and colon (B) datasets.

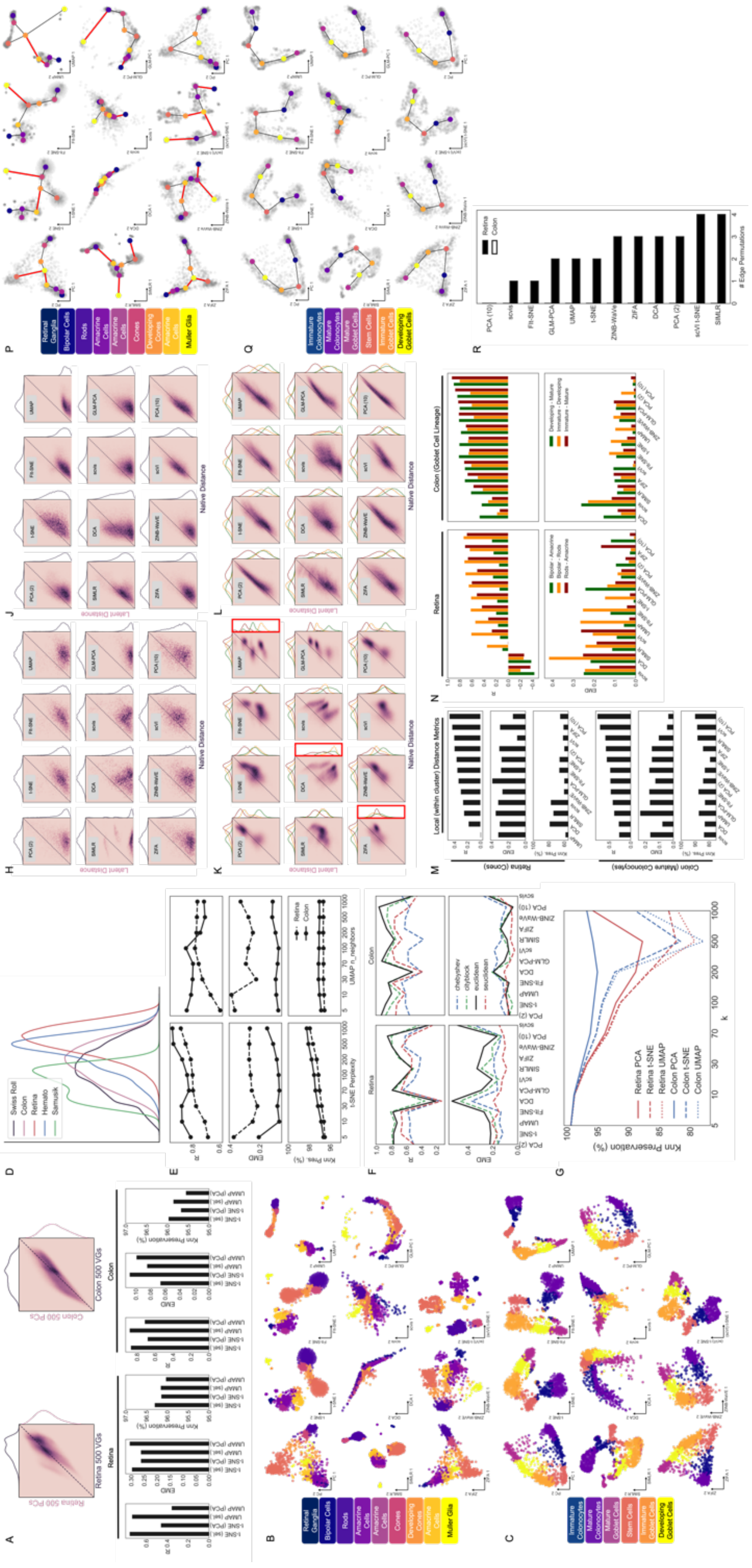
A
Retina 500 PCs
Retina 500 VGs
Colon 500 PCs
Colon 500 VGs
Retina
Colon
R
EMD
Knn Preservation (%)
t-SNE
UMAP
B
Retinal Ganglia
Bipolar Cells
Rods
Amacrine Cells
Cones
Developing Cones
Muller Glia
C
Immature Colonocytes
Mature Colonocytes
Goblet Cells
Stem Cells
Immature Goblet Cells
Developing Goblet Cells
D
Swiss Roll
Colon
Retina
Hemato
Samusik
E
t-SNE Perplexity
UMAP n_neighbors
F
chebyshev
cityblock
euclidean
seuclidean
G
Knn Preservation (%)
k
Retina PCA
Retina t-SNE
Retina UMAP
Colon PCA
Colon t-SNE
Colon UMAP
H
J
K
L
Latent Distance
Native Distance
M
Local (within cluster) Distance Metrics
Retina (Cones)
Colon (Mature Colonocytes)
N
Colon (Goblet Cell Lineage)
Bipolar - Amacrine
Bipolar - Rods
Rods - Amacrine
Developing - Mature
Immature - Developing
Immature - Mature
P
Q
R
# Edge Permutations
PCA (10)
scvis
FIt-SNE
GLM-PCA
UMAP
t-SNE
ZINB-WaVE
ZIFA
DCA
PCA (2)
scVI t-SNE
SIMLR

Figure S4 *(preceding page)*: Related to Figure 2.4.

(A) Comparison of 500 VGs to 500 PCs as latent space for both datasets. B,C) Low-dimensional projections with overlay of consensus Louvain clusters for retina (B) and colon (C) data.

(D) Overlay of normalized probability distributions of native spaces to demonstrate varying structure of different datasets.

(E) Resulting distance metrics from titration of perplexity parameter in t-SNE and UMAP (`n_neighbors`) on retina and colon datasets.

(F) Comparison of alternative distance metrics and their effect on R and EMD results for both datasets and 11 methods.

(G) Titration of k parameter for construction of Knn graphs and calculation of their preservation following dimension reduction by PCA, t-SNE, and UMAP on both retina and colon data.

(H) 2D histograms of cell distance correlations within cone cell cluster of retina dataset.

(J) Same as in H for distances within mature colonocyte cluster of colon dataset.

(K) Same as in H for distances between bipolar, amacrine, and rod cells in retina dataset.

(L) Same as in H for distances between three goblet cell clusters in colon dataset.

(M) Local distance preservation metrics for cone cell cluster and mature colonocyte cluster from retina and colon datasets, respectively.

(N) EMD and distance correlation values for pairwise distance distributions between bipolar cells, rod cells, and amacrine cells in retina dataset, and three clusters along developing goblet cell lineage in colon dataset.

(P) MST graphs constructed from cluster centroids and their pairwise distances in 2D latent projections for retina dataset.

(Q) Same as in P for colon dataset.

(R) Summary of number of edge permutations in 2D latent projections relative to native graph.

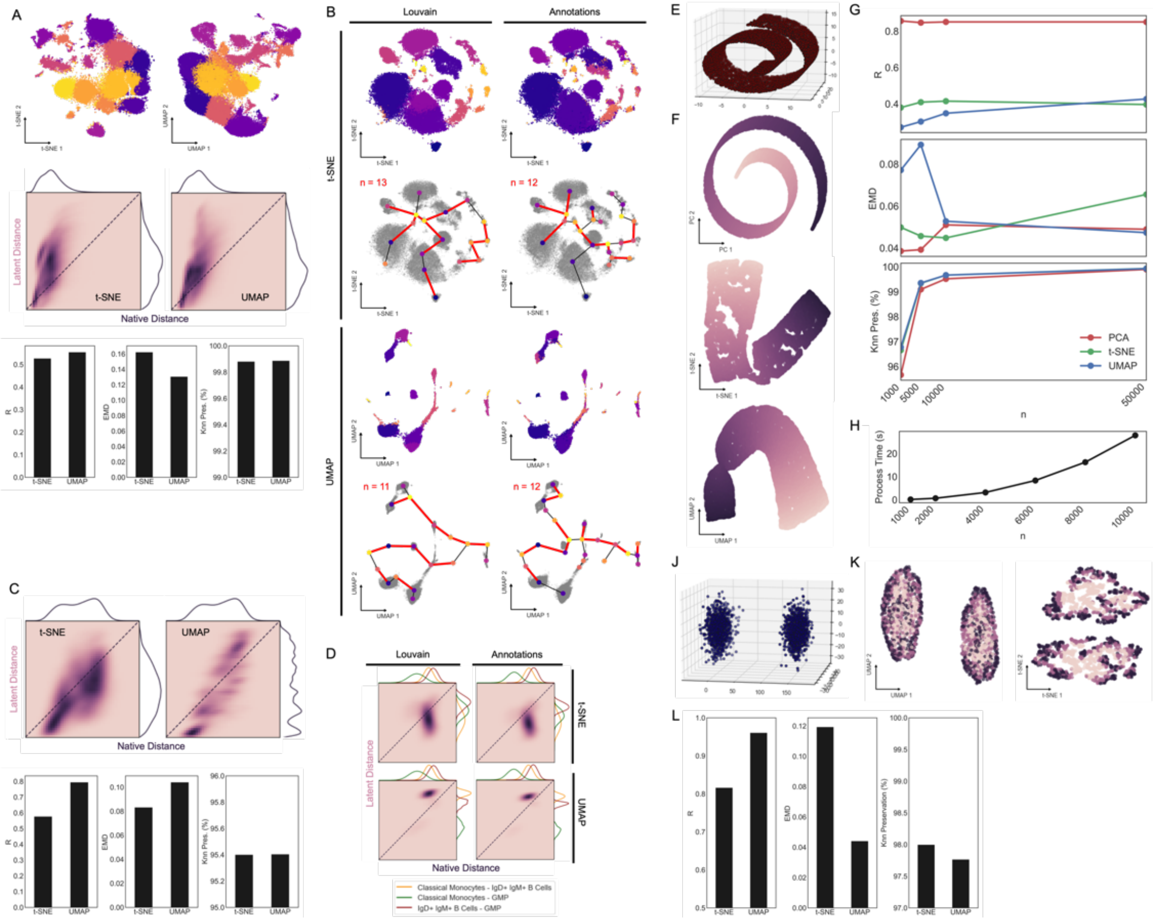

A
t-SNE 2
t-SNE 1
UMAP 2
UMAP 1
Latent Distance
Native Distance
t-SNE
UMAP
R
EMD
Knn Pres. (%)
B
Louvain
Annotations
t-SNE
UMAP
n = 13
n = 12
n = 11
n = 12
C
D
Classical Monocytes - IgD+ IgM+ B Cells
Classical Monocytes - GMP
IgD+ IgM+ B Cells - GMP
E
F
PC 2
PC 1
G
PCA
t-SNE
UMAP
n
H
Process Time (s)
J
K
L
Knn Preservation (%)

Figure S5 *(preceding page)*: Application of framework to datasets from Becht et al., 2018 and 3D synthetic datasets with intuitive structure.

(A) Hematopoietic subset of mouse cell atlas (Han et al., 2018); 14 combined scRNA-seq samples from blood and bone marrow totaling 51,252 cells. Data were normalized and preprocessed with 100-component PCA prior to embedding with t-SNE and UMAP (top) as in Han et al., 2018. 2D histograms of cell distance correlation (middle) and summary metrics (bottom).

(B) Comparison of naïve Louvain clustering and previously published cell annotations via MST topological analysis for Samusik_01 CyTOF data (86,864 cells from mouse bone marrow, 39 features) preprocessed as in Weber and Robinson, 2016.

(C) 2D histograms of cell distance correlation (top) and summary metrics (bottom) for t-SNE and UMAP embeddings of dataset (both shown in A).

(D) Comparison of cluster definition as in B, using neighborhood analysis measuring distance distributions between classical monocytes, IgD+ IgM+ B cells, and granulocyte-monocyte progenitors (GMPs).

(E) Example 3-dimensional swiss roll dataset with 10,000 randomly placed points generated using `sklearn.datasets.make_swiss_roll` function with 0 noise (vertically away from manifold).

(F) Example 2D embeddings of data from A using PCA, t-SNE, and UMAP. Points are colored by their position along the manifold to show expected order.

(G) Correlation, EMD, and Knn preservation metrics (k=30) for swiss roll datasets as in E with increasing number of points.

(H) Processing time for structural preservation framework (calculating R, EMD, and Knn pres. from distance matrices) for up to 10,000 cells.

(J) 3D dataset consisting of two Gaussian point clouds (1,000 points each) generated using `sklearn.datasets.make_gaussian_quantiles`.

(K) UMAP and t-SNE embeddings of data from J with points colored by their distance from center of their respective Gaussian distribution in 3D space.

(L) Summary of structural preservation metrics for t-SNE and UMAP of data from J.

A3 Appendix for Chapter 3: Automated quality control and cell identification of droplet-based single-cell data using dropkick

A3.1 Methods

A3.1.1 inDrop data generation

The human colorectal carcinoma inDrop data – deposited to the Gene Expression Omnibus (GEO) to accompany this manuscript (GSE158636) - were generated according to published protocols (Southard-Smith et al., 2020; Banerjee et al., 2020).

A3.1.2 Quality control and ambient RNA quantification with the dropkick QC module

The dropkick QC module begins by calculating global heuristics per barcode (observation) and gene (variable) using the `scanpy` (Wolf et al., 2018) `pp.calculate_qc_metrics` function. These metrics are used to order barcodes by decreasing total counts (black curve in Figure 3.1A) and order genes by increasing dropout rate (Figure 3.1B). The nth gene ranked by dropout rate determines the cutoff for calling "ambient" genes, with n determined by the `n_ambient` parameter in the `dropkick.qc_summary` function. All genes with dropout rates less than or equal to this threshold are labeled "ambient". In a sample with many (¿ n) genes detected in all barcodes, this ensures that the entire ambient profile is identified. Through observation of samples used in this study, we set the default `n_ambient` = 10. To compile the dropkick QC summary report, the log-total counts versus log-ranked barcodes (Figure 3.1A black curve) are plotted along with total genes detected for each barcode (Figure 3.1A green points), percent counts from "ambient" genes in each barcode (Figure 3.1A blue points), and percent counts from mitochondrial genes in each barcode (Figure 3.1A red points).

A3.1.3 Labeling training set with the dropkick filtering module

The dropkick filtering module also begins by calculating global heuristics per barcode (observation) and gene (variable) using the `scanpy` (Wolf et al., 2018) `pp.calculate_qc_metrics` function. Next, training thresholds are calculated on the histogram of the chosen heuristic(s); arcsinh-transformed `n_genes` by default. dropkick then uses the `scikit-image` function `filters.threshold_multiotsu` to identify two local minima in the `n_genes` histogram that represent the transitions from uninformative barcodes to "empty droplets" and from "empty droplets" to real cells. These locations are also characterized by the two expected drop-offs in the total counts/genes profiles as shown in the dropkick QC report (Figure 3.1, Figure S6). To label barcodes for dropkick model training, barcodes with fewer genes detected than the first multi-Otsu threshold are discarded due to their lack of molecular information. dropkick then labels barcodes below the second threshold as "empty", and remaining barcodes above the second threshold as real cells for initial training. These inputs to the dropkick logistic regression model represent the "noisy" boundary in heuristic space that is to be replaced with a learned cell boundary in gene space.

A3.1.4 Training and optimizing the dropkick filtering model

The dropkick filtering model uses logistic regression with elastic net regularization (Zou and Hastie, 2005), and is fit as described in Friedman et al., 2010. The elastic net combines ridge and lasso (least absolute shrinkage and selection operator) penalties for optimal regularization of model coefficients. The ridge regression penalty pushes all coefficients toward zero while allowing multiple correlated predictors to borrow strength from one another, ideal for a scenario like scRNA-seq with several expected collinearities (Hoerl and Kennard, 1970). The lasso penalty on the other hand, favors model sparsity, driving coefficients to zero and thus selecting informative features (Tibshirani, 1996). The combined elastic net balances feature selection and grouping by preserving or removing correlated features from the model in concert (Zou and Hastie, 2005).

The fraction $\alpha \in [0, 1]$ (alpha) represents the balance between the lasso and ridge penalties for the elastic net model. If $\alpha = 0$, the regularization would be entirely ridge, while if $\alpha = 1$, it would be entirely lasso. By default, dropkick fixes this alpha value at 0.1, but the user may alter this parameter or provide multiple alpha values to optimize through cross-validation (with lambda; explained below) at the expense of slightly longer computational time. All default dropkick results in this manuscript used $\alpha = 0.1$, and we also ran dropkick on all 46 samples with given alpha values [0.1, 0.25, 0.5, 0.75, 0.9]. Only 9 of 46 models chose a value other than $\alpha = 0.1$.

For a desired length of "lambda path," n (default n = 100 for dropkick), the model is fit n + 1 times, where the first pass determines the values of lambda (regularization strength) to test, and subsequent fits determine model performance using cross-validation (CV; default 5-fold for dropkick). Each fit involves selection of highly-variable genes (HVGs; `scanpy pp.highly_variable_genes`; default 2,000 for dropkick) from the training set. For both the first pass and the final model, the training set consists of all available barcodes, while training the model along the lambda path uses only the current training fold as to not bias model fitting with information from the test set. The lambda path is scored using mean deviance from the training labels for all cross-validation folds. The largest value of lambda such that its mean CV deviance is less than or equal to one standard error above the minimum deviance is chosen as the final regularization strength for the model in order to further minimize overfitting. Finally, dropkick fits a logistic regression model using all training labels and the chosen lambda value and assigns cell probability (`dropkick_score`) to all barcodes. By default, the resulting `dropkick_label` is positive (1; real cell) for barcodes with `dropkick_score` $\geq$ 0.5, but the user may define a stricter or more lenient threshold for particular applications.

A3.1.5 Synthetic scRNA-seq data simulation

We used CellBender (Fleming et al., 2019) to build synthetic single-cell datasets. We generated a basic count matrix with 30,000 features (`n_genes`), 12,000 total droplets (including 3,000 `n_cells` and 9,000 `n_empty`), and 6 clusters. The default ratio between the cell size scale factor and the empty droplet size scale factor – `d_cell` at 10,000 and `d_empty` at 200 – created an unrealistic gap between the empty droplets and the real cells but built a foundation on which to produce more realistic simulations. By adjusting these parameters, we simulated two different scenarios with the number of features, total droplets, and clusters held constant. The first scenario modeled a "low background" dataset, with a realistic `n_genes` and total counts profile and relatively low ambient RNA. We set the cell size scale factor (`d_cell`) to 10,000, and the empty droplet size scale factor (`d_empty`) to 1,000. These settings produced a small gap between the real cells and the empty droplets, yet still mimicked a low background droplet profile. We then modeled a "high background" scenario, which had much higher ambient RNA content. For this simulation we set `d_cell` to 10,000, and `d_empty` to 2,000. This simulation mimicked a real scRNA-seq dataset with a high ambient profile, as it had a smaller gap between real cells and empty droplets. Taken together, these simulations recapitulate real-world single-cell data and were tested by dropkick to compare their ground-truth labels to those determined by dropkick filtering.

A3.1.6 High-background PBMC simulation

To imitate empty droplets with high mRNA content over a relatively low-background sample, we used the 10x Genomics 4k human PBMC dataset. Because this encapsulation was derived from suspended blood cells, there was negligible lysis and ambient contamination, and empty droplets are very clearly distinguished from real cells based on their mRNA content alone. Combining reads from the bottom 1,000 genes by dropout rate across all barcodes with less than 100 total UMIs, we normalized this pseudo-bulk as probabilistic weightings for a random generation of count vectors. We drew 2,000 random integers between 10 and 5,000 to determine the total number of counts for each simulated barcode, then drew that number of random integers from a multinomial distribution using the `random.default_rng.multinomial` function from the numpy Python package, with pvals equal to the weightings determined from the true empty droplet pseudo-bulk. We then added these 2,000 count vectors back to the original matrix, labeling them as "simulated" for downstream comparison (Figure 3.4A).

A3.1.7 CellRanger 2, EmptyDrops, CellBender, and manual filtering of real-world scRNA-seq datasets

CellRanger and EmptyDrops filtering algorithms were derived from Lun et al., 2019, with CellRanger 2 described by the function `DefaultDrops` (from the repository github.com/MarioniLab/EmptyDrops2017), and EmptyDrops by the `EmptyDrops` function within the DropletUtils R package (v1.8.0). All 10x datasets were processed as in Lun et al., 2019 (github.com/MarioniLab/EmptyDrops2017). EmptyDrops was run for all inDrop datasets using the "inflection point" from CellRanger 2 analysis as the minimum non-ambient UMI threshold as in Lun et al., 2019 (github.com/MarioniLab/EmptyDrops2017).

Further investigation of user-defined parameters for both methods was performed by titrating the "lower" parameter, which describe the lower proportion of total barcodes to ignore when calculating the inflection point for CellRanger 2, and the maximum total UMI counts under which all barcodes are considered ground-truth empty droplets for EmptyDrops. We compared dropkick scores to the resulting labels (Figure S14), noting that sub-optimal parameter values led to lower concordance.

CellBender remove-background, while primarily an ambient RNA subtraction model, also provides cell labels from raw scRNA-seq counts matrices (Fleming et al., 2019). With the caveat that CellBender likely retains more previously high-background droplets after regressing out ambient reads, CellBender was performed on 10x Genomics samples using the same expected cell number used for EmptyDrops in Lun et al., 2019, and concordance was tested with dropkick labels as before, showing a slightly lower average AUROC of 0.9585 ± 0.0596 for 13 10x Genomics samples (Figure S13G).

Manual filtering was performed for each inDrop sample by initial thresholding beyond the inflection point detected in the first curve of the ranked barcodes profile (as in Figure 3.1A). Then, following standard dimension reduction and high-resolution Leiden clustering, clusters with low quality cells (high mitochondrial/ambient percentage, low total counts/genes) were manually gated out of the final dataset. These manually curated labels were used as an orthogonal "gold standard" for benchmarking automated thresholding methods (Figure S7A) and final AUROC (Figure S10D). Further description of this manual filtering method in Chen et al., 2021b.

Bivariate thresholding was performed for all samples using total UMI counts and percent mitochondrial counts, keeping barcodes that have greater than or equal to the minimum total count threshold and less than 40 % mitochondrial reads.

A3.1.8 sc-UniFrac analysis of shared populations between dropkick, CellRanger 2, and EmptyDrops labels

In order to evaluate the preservation of expected cell clusters between dropkick and alternative labels, we employed sc-UniFrac (Liu et al., 2018) to determine the global and populational differences between the label sets. We used nonnegative matrix factorization (NMF) to analyze the union of barcodes kept by `dropkick_label`, CellRanger_2, and EmptyDrops in order to reduce dimensions into cell identity and activity "metagenes" (Kotliar et al., 2019). We then clustered this low-dimensional space using the Leiden algorithm (Traag et al., 2019) to define consensus cell populations for sc-UniFrac analysis. We then ran sc-UniFrac (v0.9.6) to evaluate statistically significant cluster differences based on both cluster membership and gene expression hierarchies between clusters. The global sc-UniFrac distance quantified the overall similarity of hierarchical trees across barcode label sets.

A3.1.9 Dimension reduction, clustering, projection, and differential expression analysis

We used Consensus Nonnegative Matrix Factorization (cNMF; Kotliar et al., 2019) for initial dimension reduction. The optimal number of factors, k, was determined by maximizing stability and minimizing error across all tested values after 30 iterations of each. We then built a nearest-neighbors graph in `scanpy` (`pp.neighbors` function) from the NMF usage scores for consensus factors in all cells, where we set `n_neighbors` to the square root of the total number of cells in the dataset. We then clustered cells with the Leiden algorithm (`scanpy tl.leiden` function; Traag et al., 2019) applied to this graph. Resulting clusters were used in sc-UniFrac analysis, differential expression, and visualization. We performed differential expression analysis using a Student's t-test with Benjamini-Hochberg p-value correction for multiple testing (`scanpy tl.rank_genes_groups`). To visualize datasets in 2D space, we ran partition-based graph abstraction (PAGA; Wolf et al., 2019; `scanpy tl.paga`) on this nearest-neighbors graph and associated Leiden clustering in order to create a simple representation of cluster similarity. Finally, a UMAP projection (McInnes et al., 2018) seeded with these PAGA positions provided a two-dimensional embedding of all cells in the dataset (`scanpy tl.umap` with `init_pos` = ”paga”).

A3.1.10 Data Access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE158636. All publicly available datasets are listed in the Table S2.

A3.1.11 Software Availability

The dropkick Python package is available for download via “pip” from the Python Package Index (PyPI) at https://pypi.org/project/dropkick/. Source code for the package is also available on GitHub at https://github.com/KenLauLab/dropkick. Scripts for reproducing analyses in this manuscript are hosted on GitHub at https://github.com/codyheiser/dropkick-manuscript.

A3.2 Supplemental tables and figures

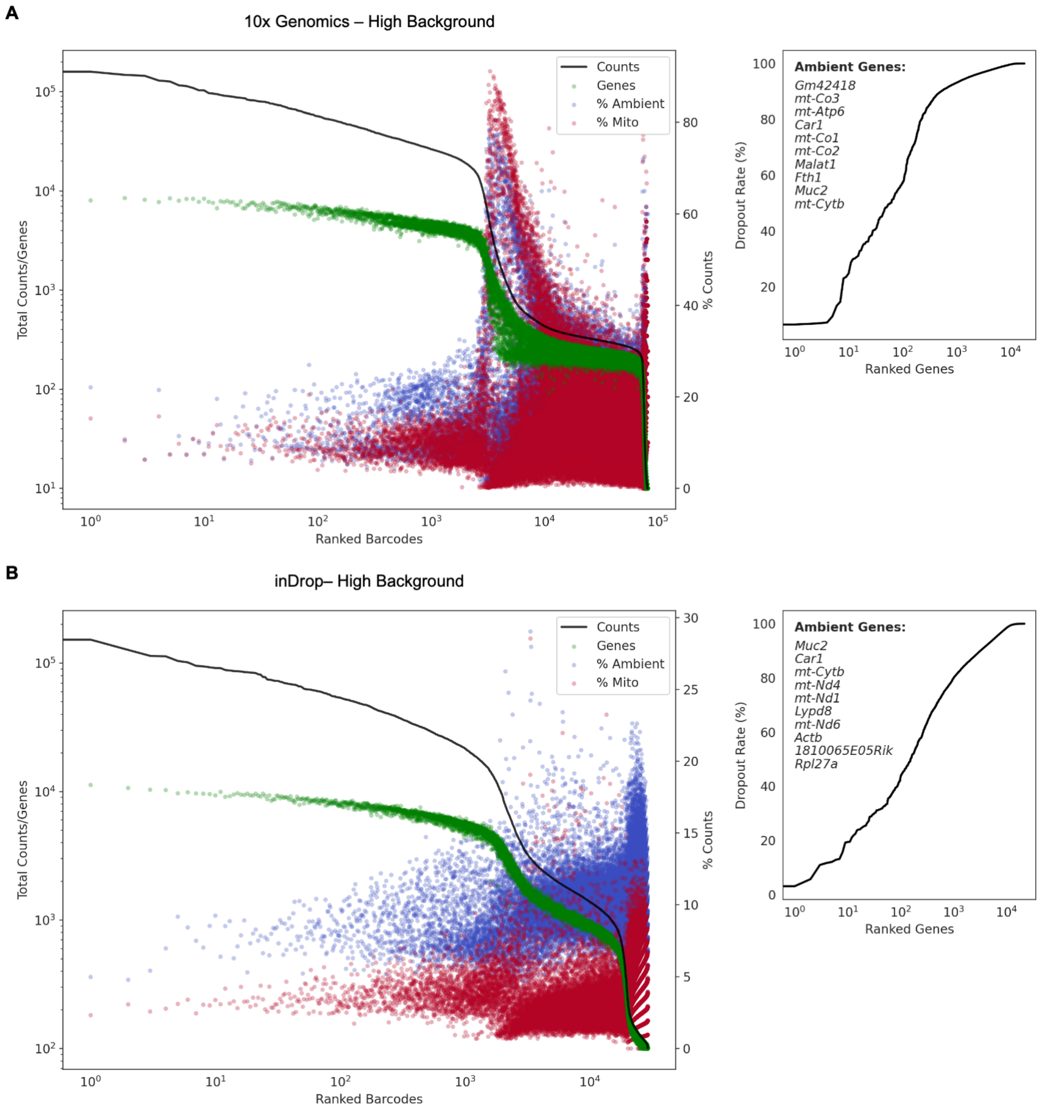


Figure S6: dropkick QC reports for a mouse colonic epithelium sample analyzed by both 10x Genomics (A) and inDrop (B) scRNA-seq. In contrast to Figure 3.1, this is considered a high-background sample due to the height (increased total counts) of the second plateau (empty droplets) and presence of epithelial marker genes (*Car1*, *Muc2*) in the ambient profile.

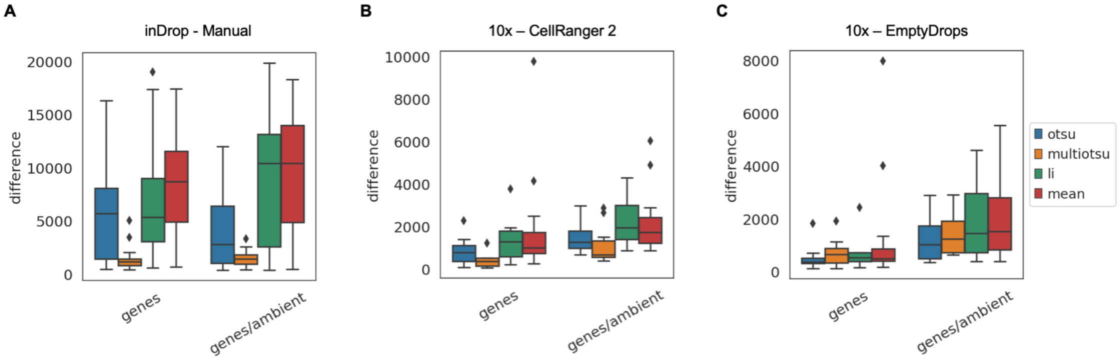


Figure S7: Optimal heuristics and thresholding for determination of dropkick training set.
(A) Barcode set differences between initial dropkick thresholding and manual filtering of 33 inDrop scRNA-seq datasets. Four automated thresholding techniques were used to label cells based on the distribution of arcsinh-transformed genes detected alone (genes), or the combination of genes and percent ambient counts as calculated by the dropkick QC module (genes/ambient).
(B) Same as in A for 13 10x Genomics scRNA-seq datasets, with set differences compared to CellRanger_2.
(C) Same as in B, with set differences compared to EmptyDrops.

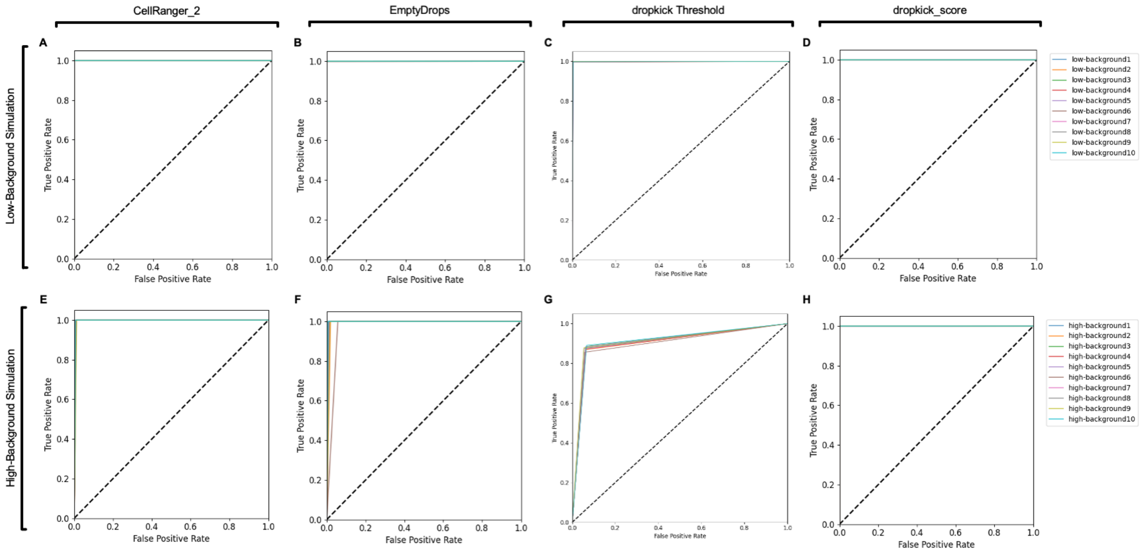


Figure S8: Evaluating dropkick filtering performance with synthetic data.
(A) Receiver operating characteristic (ROC) curves for CellRanger_2 vs. ground truth in ten low-background simulations.
(B) ROC curves for EmptyDrops vs. ground truth in ten low-background simulations.
(C) ROC curves for dropkick training labels (threshold) vs. ground truth in ten low-background simulations.
(D) ROC curves for final dropkick score vs. ground truth in ten low-background simulations.
(E-H) Same as in A-D, for ten high-background simulations.

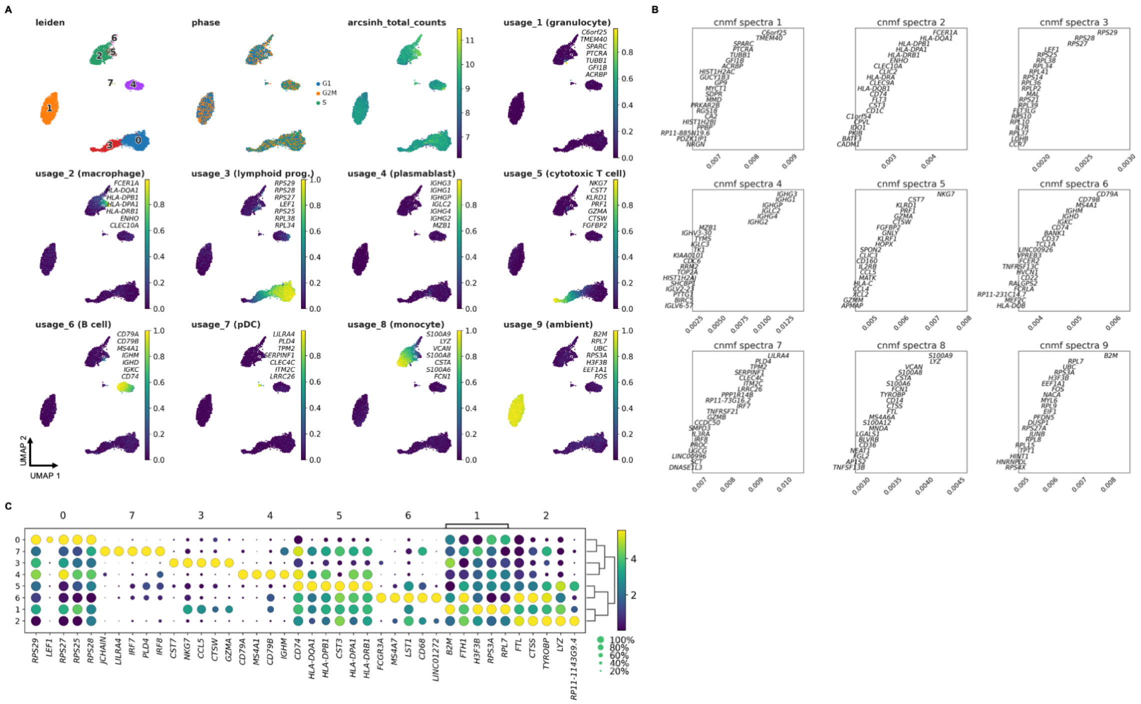


Figure S9: Benchmarking dropkick performance on simulated high-background data.

(A) UMAP embedding of all barcodes kept by dropkick, CellRanger 2, and EmptyDrops. NMF results were used to generate leiden clusters; usage scores shown with a description of each cell type they represent and top 7 gene loadings for each.

(B) Top 20 gene loadings for each NMF metagene.

(C) Top 5 differentially expressed genes in the 8 leiden clusters.

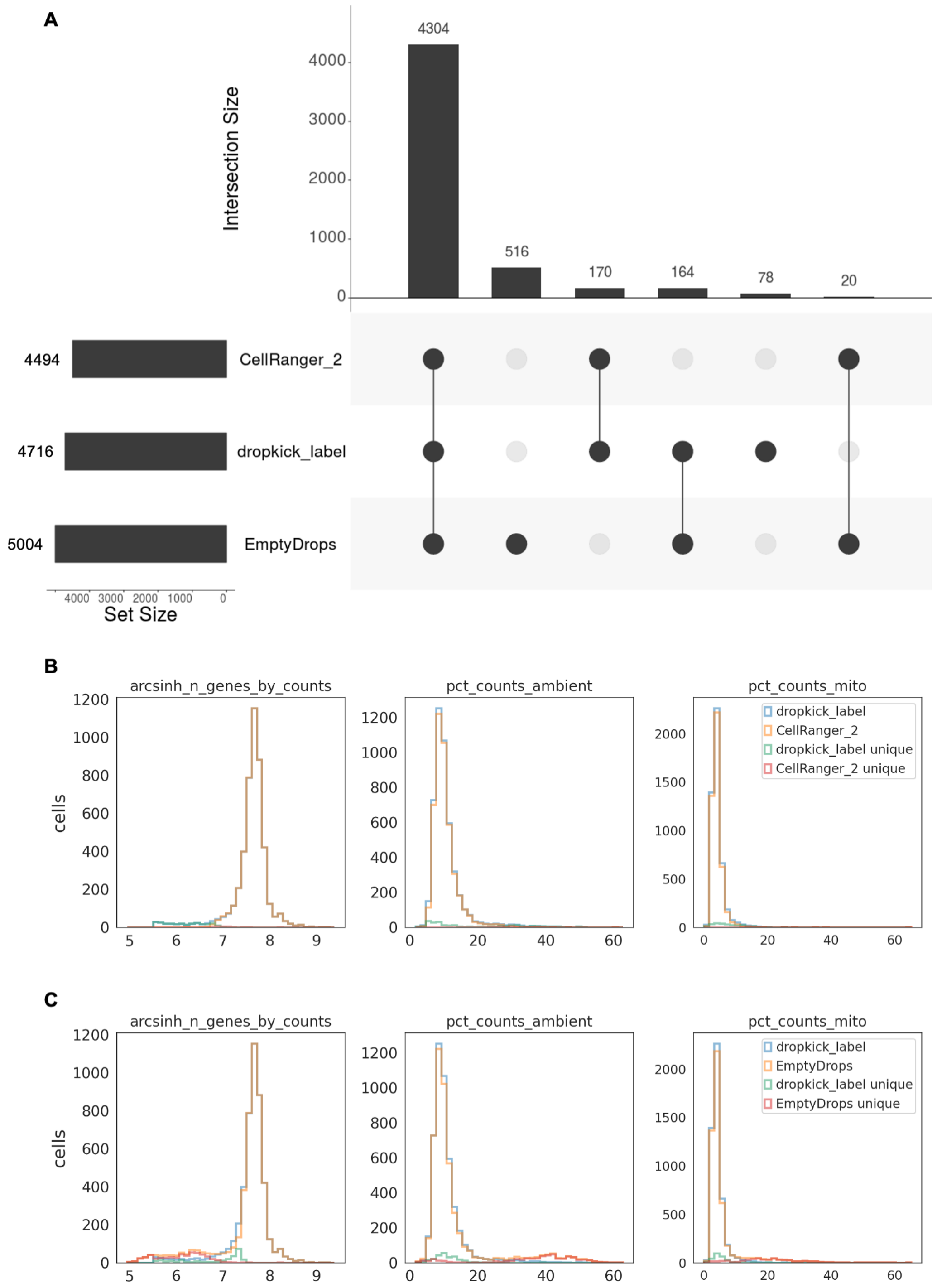


Figure S10: Barcode set differences for 4k pan-T cell dataset.
(A) UpSet plot showing global set differences between `dropkick_label` (dropkick score $\geq$ 0.5), CellRanger_2 and EmptyDrops.
(B) Histograms showing global distribution of heuristics (arcsinh-transformed genes, left, percent ambient counts, middle, and percent mitochondrial counts, right) in barcodes kept by `dropkick_label` and CellRanger_2. Distribution of barcodes unique to each label set also overlaid to show difference.
(C) Same as in B, for `dropkick_label` compared to EmptyDrops.

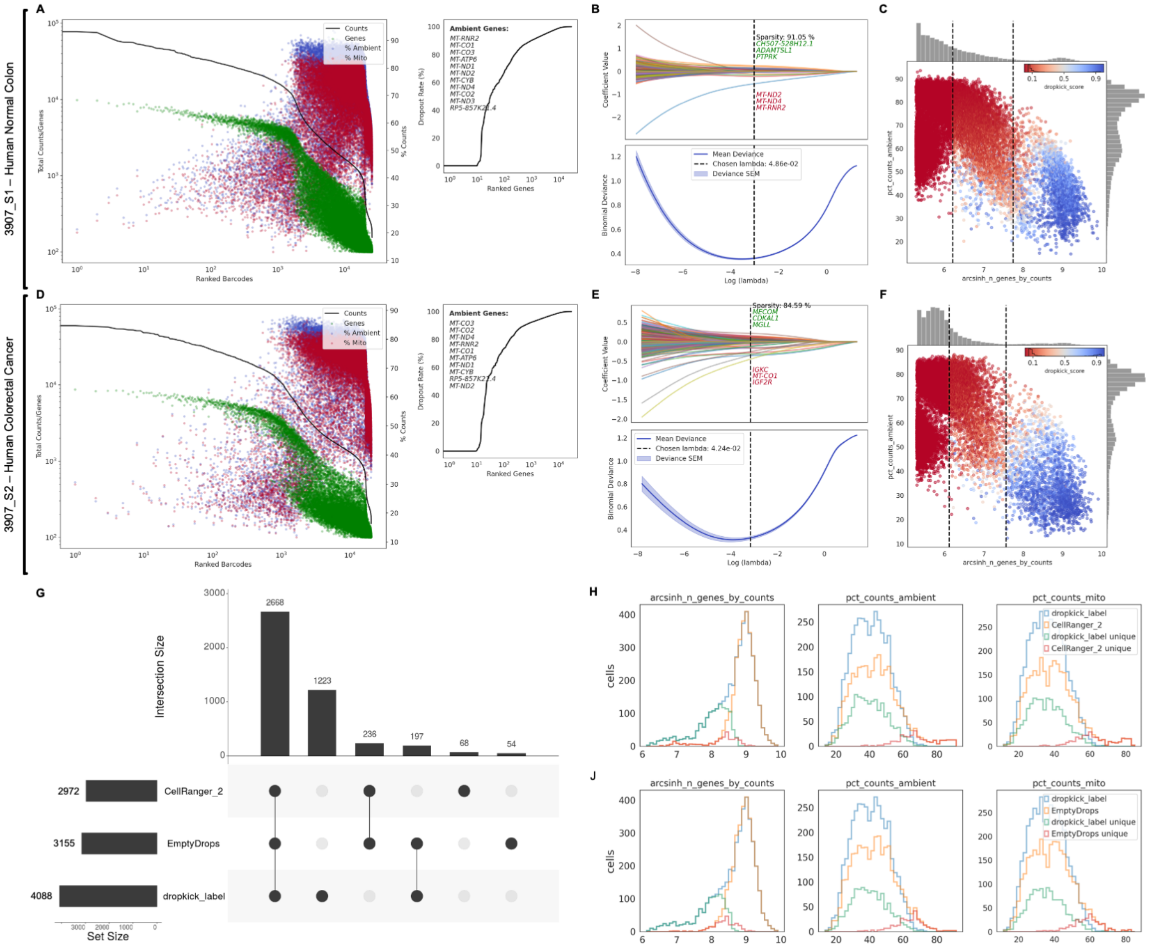


Figure S11: dropkick plots and barcode set differences for human colorectal carcinoma (CRC) inDrop samples.

(A) dropkick QC report for human normal colonic mucosa, 3907_S1 and CRC, 3907_S2.

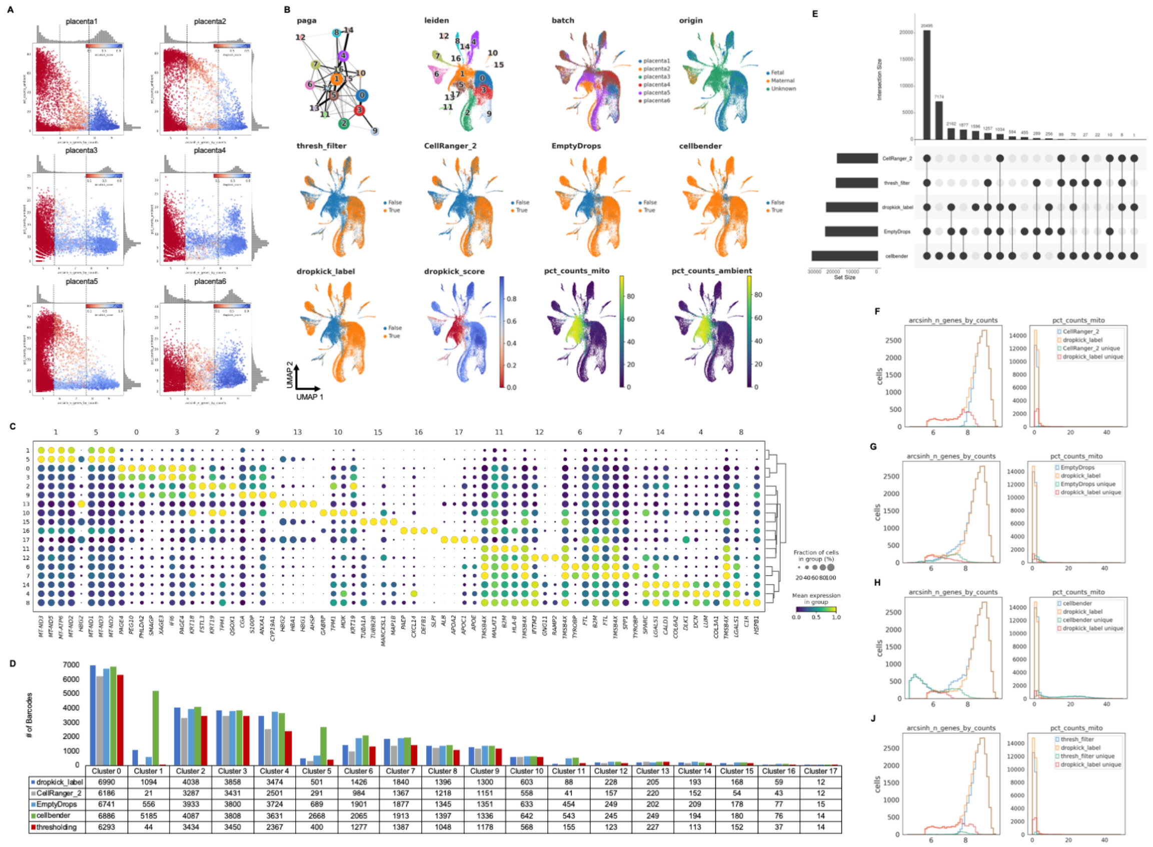
(B) dropkick coefficient plots, showing coefficient values (top) and binomial deviance (bottom) along the tested lambda regularization path. Dashed line indicates chosen lambda value of trained model. Top and bottom three genes by coefficient value and total model sparsity noted in top plot.

(C) dropkick score plot showing scatter of percent counts ambient versus arcsinh-transformed total genes detected per barcode. Dashed lines indicate location of automated dropkick thresholds used for model training. Points colored by final dropkick score. D-F) Same as in A-C, but for adjacent human normal colonic mucosa sample, 3907_S2.

(G) UpSet plot showing global set differences between `dropkick_label` (dropkick score ≥ 0.5), CellRanger_2 and EmptyDrops.

(H) Histograms showing global distribution of heuristics (arcsinh-transformed genes, left, percent ambient counts, middle, and percent mitochondrial counts, right) in barcodes kept by `dropkick_label` and CellRanger_2. Distribution of barcodes unique to each label set also overlaid to show difference.

(J) Same as in H, for `dropkick_label` compared to EmptyDrops.



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12	Cluster 13	Cluster 14	Cluster 15	Cluster 16	Cluster 17
dropkick_label	6990	1094	4038	3858	3474	501	1426	1840	1396	1300	603	88	228	205	193	168	59	12
CellRanger_2	6186	21	3287	3431	2501	291	984	1367	1218	1151	558	41	157	220	152	54	43	12
EmptyDrops	6741	556	3933	3800	3724	689	1901	1877	1345	1351	633	454	249	202	209	178	77	15
cellbender	6886	5185	4087	3808	3631	2668	2065	1913	1397	1336	642	543	245	249	194	180	76	14
thresholding	6293	44	3434	3450	2367	400	1277	1387	1048	1178	568	155	123	227	113	152	37	14

Figure S12: dropkick filters reproducibly across scRNA-seq batches.
(A) dropkick score plots for six placenta replicates.
(B) PAGA graph and UMAP embedding of all barcodes kept by `dropkick_label` (dropkick score $\geq$ 0.5), CellRanger_2, EmptyDrops, CellBender, and bivariate thresholding for the aggregate placenta dataset. Points colored by each of the five filtering labels as well as original batch, tissue of origin, Leiden clusters, dropkick_score (cell probability), percent counts ambient, and percent counts mitochondrial.
(C) Dot plot showing top five differentially expressed genes for each cluster. The size of each dot indicates the percentage of cells in the population with nonzero expression for the given gene, while the color indicates the average expression value in that population.
(D) Table and bar graph enumerating the total number of barcodes detected by each algorithm in all clusters.
(E) UpSet plot showing global set differences between `dropkick_label`, CellRanger_2, EmptyDrops, CellBender, and bivariate thresholding.
(F) Histograms showing global distribution of heuristics (arcsinh-transformed genes, left, and percent mitochondrial counts, right) in barcodes kept by `dropkick_label` and CellRanger_2. Distribution of barcodes unique to each label set also overlaid to show difference.
(G) Same as in F, for `dropkick_label` compared to EmptyDrops.
(H) Same as in F, for `dropkick_label` compared to CellBender.
(J) Same as in F, for `dropkick_label` compared to bivariate thresholding.

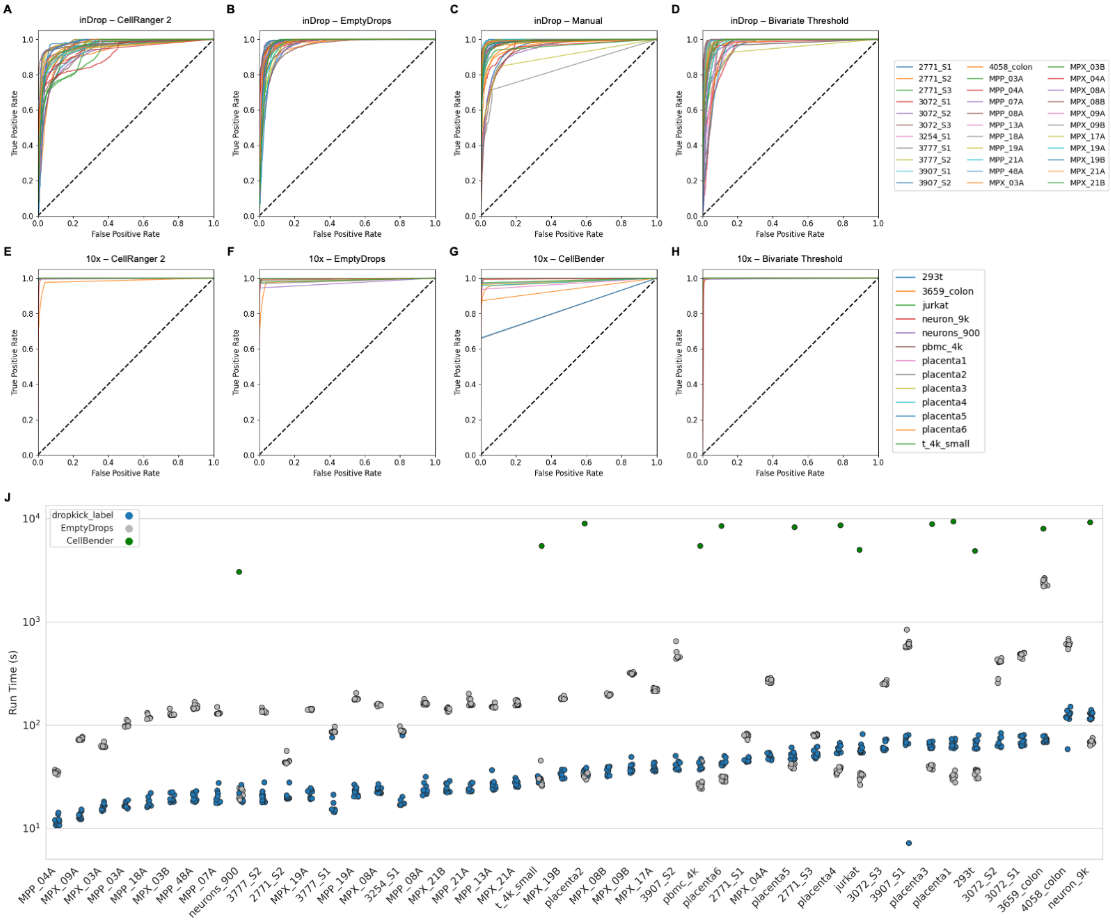


Figure S13: Comparing dropkick probability scores to five alternative cell labels using receiver operating characteristic (ROC) curves. AUC = area under the ROC curve.
(A) ROC curves for 33 inDrop scRNA-seq datasets, using CellRanger_2 as reference.
(B) Same as in A, with EmptyDrops as reference.
(C) Same as in A, with manually curated cell labels as reference.
(D) Same as in A, with bivariate thresholding as a reference.
(E) ROC curves for 13 10x Genomics scRNA-seq datasets, using CellRanger_2 as reference.
(F) Same as in E, with EmptyDrops labels as reference.
(G) Same as in E, with CellBender remove-background labels as a reference.
(H) Same as in E, with bivariate thresholding as a reference.
(J) Total run time, in seconds, for EmptyDrops, CellBender remove-background, and dropkick. EmptyDrops and dropkick were run ten times on all datasets; CellBender was run once on 10x Genomics samples. Points represent single replicates.

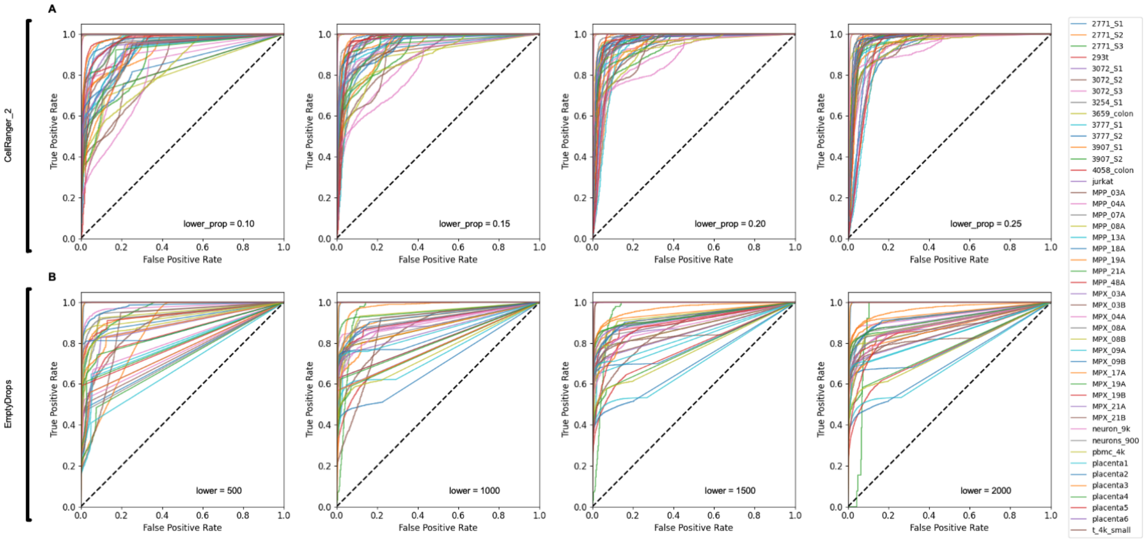


Figure S14: Comparing dropkick probability scores to titrations of CellRanger_2 (A) and EmptyDrops (B) parameters using receiver operating characteristic (ROC) curves. AUC = area under the ROC curve. Results shown for 46 scRNA-seq datasets from 10x Genomics (n = 13) and inDrop (n = 33) encapsulation platforms. The `lower_prop` parameter in CellRanger version 2 (A) is used to exclude a bottom fraction of total barcodes prior to calculating the knee point of the log-rank total counts curve. The ”lower” parameter in EmptyDrops is used to determine the total UMI cutoff below which all droplets are considered empty for model building.

A4 Appendix for Chapter 4: Consensus tissue domain detection in spatial multi-omics data using MILWRM

A4.1 Methods

A4.1.1 Resource Availability

The package is available to be installed directly from pypi website (https://pypi.org/project/MILWRM/). The source code for the package and code for the figures can be found on github (https://github.com/Ken-Lau-Lab/MILWRM).

A4.1.2 MILWRM workflow details

A4.1.2.1 Data preprocessing

Spatial-omics data differ in their acquisition and technological artifacts across modalities, so these preprocessing steps are data type specific. It is important for the user to understand and apply methods that are reasonable for their modality before using MILWRM, otherwise results can be corrupted by batch effects.

A4.1.2.2 Multiplex Immunofluorescence (mIF)

Prior to preprocessing, mIF data were scaled from uint8 (0 to 255) to float (0 to 1) and downsampled by a factor 1/16th resolution to speed computation and normalization process. There is no sacrifice in the quality of the neighborhood identification by downsampling as mIF data have subcellular spatial resolution and MILWRM is designed to identify broad tissue domains. After downsampling we created tissue masks for each image as described in mIF tissue mask generation with MILWRM. Finally, we applied image normalization at slide-level using the formula $y = \log \frac{x}{\mu_x} + 1$, where x is the unnormalized data and μ_x is the mean of non-zero pixels in the image, per marker. This normalization was a modification of an existing method evaluated in segmented mIF data. Here, we implemented the mean of non-zero pixels to accommodate channels with sparse signal intensities Harris et al., 2022. The downsampling performed on images prior to this normalization step also aligns with the unbiased grid-based normalization framework described by Graf et al., 2022 Graf et al., 2021. In order to incorporate broader spatial information within each pixel, after normalization, we applied gaussian smoothing. The radius of blurring can be controlled by adjusting σ parameter in MILWRM for mIF modality. Here, we use σ = 2 for smoothing.

A4.1.2.3 Spatial Transcriptomics (ST)

The above-described steps differ slightly between mIF and ST modality. For ST data the first step is to reduce the dimensionality of the transcriptomics data. For the analysis shown in this paper, we used Principal Component Analysis (PCA) for dimensionally reduction, but other methods can also be used with MILWRM such as Non-Negative Matrix Factorization (NMF) Kotliar et al., 2019. We used Harmony Korsunsky et al., 2019 to correct technical variation between the samples. As in mIF, blurring is applied to the ST slides to preserve spatial information. To perform blurring, each central spot is assigned the average value for the selected reduced components (PCs in this case) across the spots within the neighborhood of the central spot. The spatial neighborhoods are computed using the squidpy Python package Palla et al., 2022b. The neighborhood distance can be controlled by adjusting the n_rings parameter. Here, we use n_rings = 1 for smoothing in ST data.

A4.1.2.4 Identification of tissue domains

The tissue domains in the data are identified across slides by performing unsupervised K-means clustering on the preprocessed data. MILWRM reduced computation time by randomly subsampling pixels for mIF modality. The fraction of pixels (default 0.2, used here) and all the spots are serialized to build the K-means model. If the dimension reduction is performed, then the input data are the PCs, otherwise the input is the batch-adjusted marker channels. Prior to performing K-means, the data are Z-normalized to ensure that the mean and variances are similar across the different channels/PCs of the input data. The k-selection for K-means is done by estimating adjusted inertia metric. Adjusted inertia is inertia weighted by a penalty parameter that controls the number of clusters Clarke and Greenacre, 1985. For MILWRM, the parameter can be adjusted to control the resolution of tissue domains identified.

After performing K-means classification, tissue domains are identified in the full dataset by assigning the tissue domain for the closest cluster centroid from the K-means model. The mean and variance computed for subsampled data is used to Z-normalize the original image data. By performing K-means model estimation in the subsample, MILWRM can reduce computational demand for mIF modality. Kmeans is performed on entire dataset in ST modality.

A4.1.2.5 Quality control and tissue labeling

Once the regions are identified it is useful to label the tissue domains based on their marker expression profile and assess the quality of clustering. The cluster centroids for each tissue domain are plotted in marker or PCA component space to label the tissue cluster based on its expression profile. The centroids can also be plotted in gene space for ST modality or other dimensionally reduced components. The quality of clustering is assessed at the whole slide and pixel levels. To assess the whole slide fit, we compute the variance explained and mean square error within each slide. These metrics allow the user to flag slides for manual review where the overall fit might be bad. We also compute pixel-level confidence score using the formula $y = \frac{dist_x,c2 - dist_x,c}{dist_x,c2}$ where dist is the Euclidean distance between pixel or spot, x, assigned centroid, c, and the second closest centroid `c`. The confidence scores take values between zero and one where higher values indicate smaller distance between the assigned centroid and closest centroid thus, better fit. This metric is a fast simplification of the Silhouette index.

A4.1.2.6 mIF tissue mask generation with MILWRM

MILWRM has a designated function to perform creation of tissue masks through the MILWRM pipeline described above. Each preprocessing step is performed on individual images including log normalization and smoothing with a gaussian filter (σ = 2). Finally, the mask is created using Kmeans clustering with n = 2. The Kmeans cluster centers are then z-normalized and the cluster center with a mean smaller or equal to zero is set as background.

A4.1.3 Imaging data, acquisition, and basic image processing

The mIF data were generated for the Human Tumor Atlas Network (HTAN) consisting of human normal colon and different colonic pre-cancer subtypes (conventional adenomas - AD and serrated polyps - SER) Chen et al., 2021b. These data comprised multichannel fluorescent images from 37 biospecimen consisting of tissues with different morphologies and pathological classification, as confirmed by 2 pathologists (Table S1). Cyclical antibody staining, detection, and dye inactivation was performed as described previously Gerdes et al., 2013. In brief, fluorescent images were acquired at 200x magnification on a GE In Cell Analyzer 2500 using the Cell DIVE platform. Exposure times were determined for each antibody. Dye inactivation was accomplished with an alkaline peroxide solution, and background images were collected after each round of staining to ensure fluorophore inactivation. Staining sequence, conditions, and exposure times are as described in Chen et al., 2021b. Following acquisition, images were processed as described McKinley et al., 2017. Briefly, DAPI images for each round were registered to a common baseline, and autofluorescence in staining rounds was removed by subtracting the previous background image for each position.

A4.1.4 Method evaluation and statistical analysis

In order to assess the sensitivity of MILWRM regions to biological differences between precancer subtypes we computed tissue proportions and connected component statistics for each tissue domain within the tumor region of each image and used generalized estimating equations (GEEs) to model how these variables were associated with precancer subtypes. Connected components were estimated for each image in Python using the label function in `scipy.ndimage.measurements` module. For the tissue proportions, we modeled each tissue proportion separately using a binomial family model assuming that images from the same slide had an exchangeable correlation structure. We modeled the maximum connected component size in order to quantify how the size and connectedness of different tissue domains differed across precancer subtypes. In these analyses, we used log transformation in a gaussian family model with a log transformation on the maximum connected component size and included log of the total tissue volume as a covariate. In all models, we weighted each region by its total image size so that results were not affected by noisy estimates from smaller images. Statistical analyses were performed in R using the `geepack` package Halekoh et al., 2006. We performed plot all results with unadjusted significant p-values and report adjusted p-values using the Benjamini-Hochberg procedure and a robust effect size index Benjamini and Hochberg, 1995; Vandekar et al., 2020 (Tables 4.1 and 4.2).

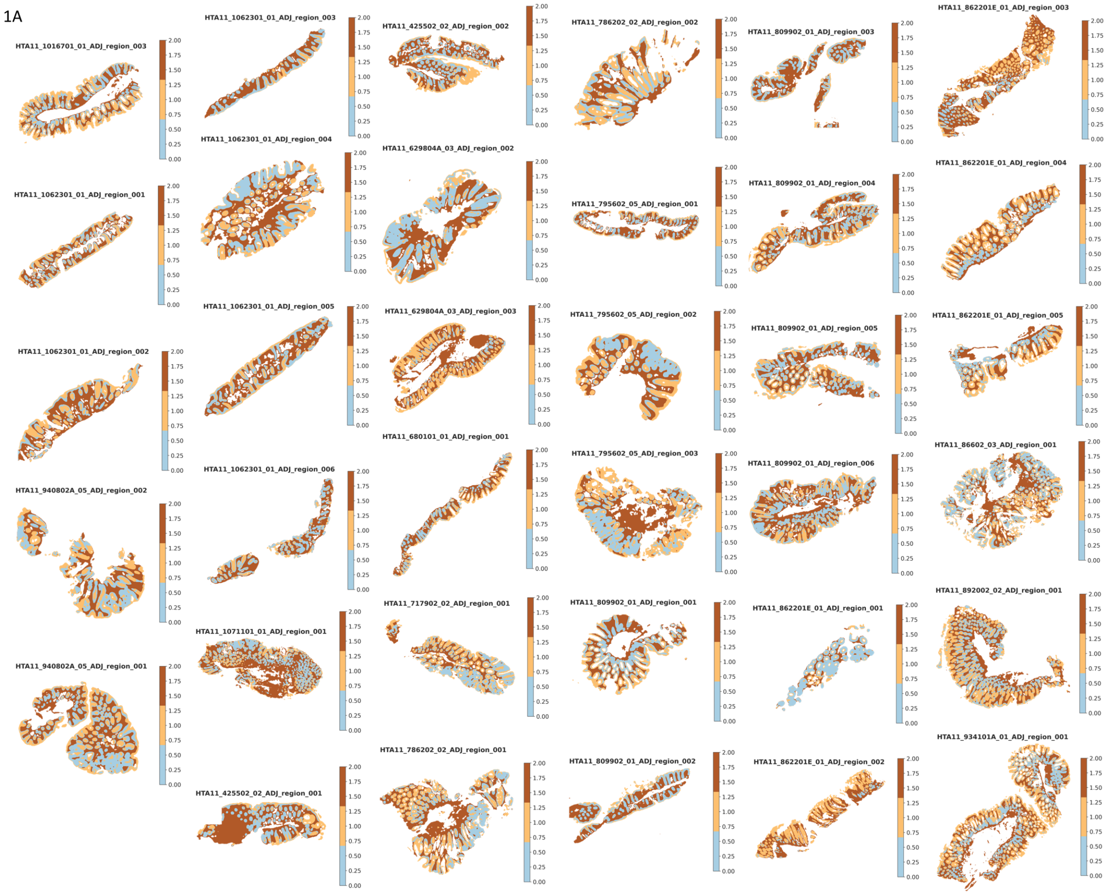
A4.1.5 Tissue domain signature scores for ST data

The manual annotation for tissue domains in ST data were verified by generating signature gene scores specific to each brain region. For this purpose, we extracted differentially expressed genes from Allen brain atlas for all available brain regions and molecular atlas of adult mouse brain for fiber tract and ventricles. MILWRM also identified a set of genes for each tissue domain. We computed a score for both reference signature set and MILWRM gene set using `scanpy` Wolf et al., 2018.

A4.2 Supplemental tables and figures

Table S1: Summary table for human colonic adenoma sample metadata

Batch name	Slide region	Tissue category	Broad precancer type
HTA11_10167_0000_01_01	region_001	normal	SSL
HTA11_10167_0000_01_01	region_002	normal	SSL
HTA11_10167_0000_01_01	region_003	Tumor	SSL
HTA11_10623_0000_01_01	region_001	normal	AD
HTA11_10623_0000_01_01	region_002	normal	AD
HTA11_10623_0000_01_01	region_003	normal	AD
HTA11_10623_0000_01_01	region_004	Tumor	AD
HTA11_10623_0000_01_01	region_005	normal	AD
HTA11_10623_0000_01_01	region_006	normal	AD
HTA11_10711_0000_01_01	region_001	Tumor	AD
HTA11_4255_0000_02_02	region_001	Tumor	SSL
HTA11_4255_0000_02_02	region_002	normal	SSL
HTA11_6298_0000_04A_03	region_001	Tumor	AD
HTA11_6298_0000_04A_03	region_002	Tumor	AD
HTA11_6298_0000_04A_03	region_003	normal	AD
HTA11_6801_0000_01_01	region_001	Tumor	SSL
HTA11_7179_0000_02_02	region_001	Tumor	AD
HTA11_7862_0000_02_02	region_001	Tumor	AD
HTA11_7862_0000_02_02	region_002	normal	AD
HTA11_7956_0000_02_05	region_001	normal	SSL
HTA11_7956_0000_02_05	region_002	Tumor	SSL
HTA11_7956_0000_02_05	region_003	Tumor	SSL
HTA11_8099_0000_02_01	region_001	normal	SSL
HTA11_8099_0000_02_01	region_002	normal	SSL
HTA11_8099_0000_02_01	region_003	normal	SSL
HTA11_8099_0000_02_01	region_004	normal	SSL
HTA11_8099_0000_02_01	region_005	Tumor	SSL
HTA11_8099_0000_02_01	region_006	normal	SSL
HTA11_8622_0000_01E_01	region_001	normal	SSL
HTA11_8622_0000_01E_01	region_002	Tumor	SSL
HTA11_8622_0000_01E_01	region_003	Tumor	SSL
HTA11_8622_0000_01E_01	region_004	Tumor	SSL
HTA11_8622_0000_01E_01	region_005	normal	SSL
HTA11_866_0000_02_03	region_001	Tumor	AD
HTA11_8920_0000_02_02	region_001	Tumor	SSL
HTA11_9341_0000_01A_01	region_001	Tumor	SSL
HTA11_9408_0000_02A_05	region_001	Tumor	AD
HTA11_9408_0000_02A_05	region_002	Tumor	AD



1A
HTA11_1016701_01_ADJ_region_003
HTA11_1062301_01_ADJ_region_001
HTA11_1062301_01_ADJ_region_002
HTA11_940802A_05_ADJ_region_002
HTA11_940802A_05_ADJ_region_001
HTA11_1062301_01_ADJ_region_003
HTA11_1062301_01_ADJ_region_004
HTA11_1062301_01_ADJ_region_005
HTA11_1062301_01_ADJ_region_006
HTA11_1071101_01_ADJ_region_001
HTA11_425502_02_ADJ_region_001
HTA11_425502_02_ADJ_region_002
HTA11_629804A_03_ADJ_region_002
HTA11_629804A_03_ADJ_region_003
HTA11_680101_01_ADJ_region_001
HTA11_717902_02_ADJ_region_001
HTA11_786202_02_ADJ_region_001
HTA11_786202_02_ADJ_region_002
HTA11_795602_05_ADJ_region_001
HTA11_795602_05_ADJ_region_002
HTA11_795602_05_ADJ_region_003
HTA11_809902_01_ADJ_region_001
HTA11_809902_01_ADJ_region_002
HTA11_809902_01_ADJ_region_003
HTA11_809902_01_ADJ_region_004
HTA11_809902_01_ADJ_region_005
HTA11_809902_01_ADJ_region_006
HTA11_862201E_01_ADJ_region_001
HTA11_862201E_01_ADJ_region_002
HTA11_862201E_01_ADJ_region_003
HTA11_862201E_01_ADJ_region_004
HTA11_862201E_01_ADJ_region_005
HTA11_86602_03_ADJ_region_001
HTA11_892002_02_ADJ_region_001
HTA11_934101A_01_ADJ_region_001
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

Figure S15: Tissue domain labels for human colonic-adenoma ($\alpha = 0.05$).

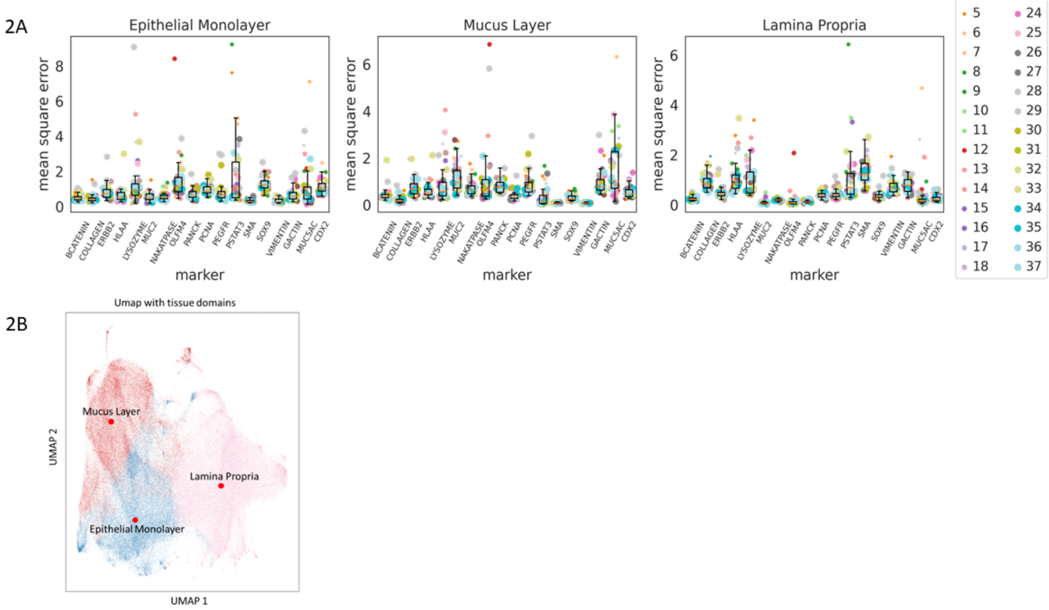
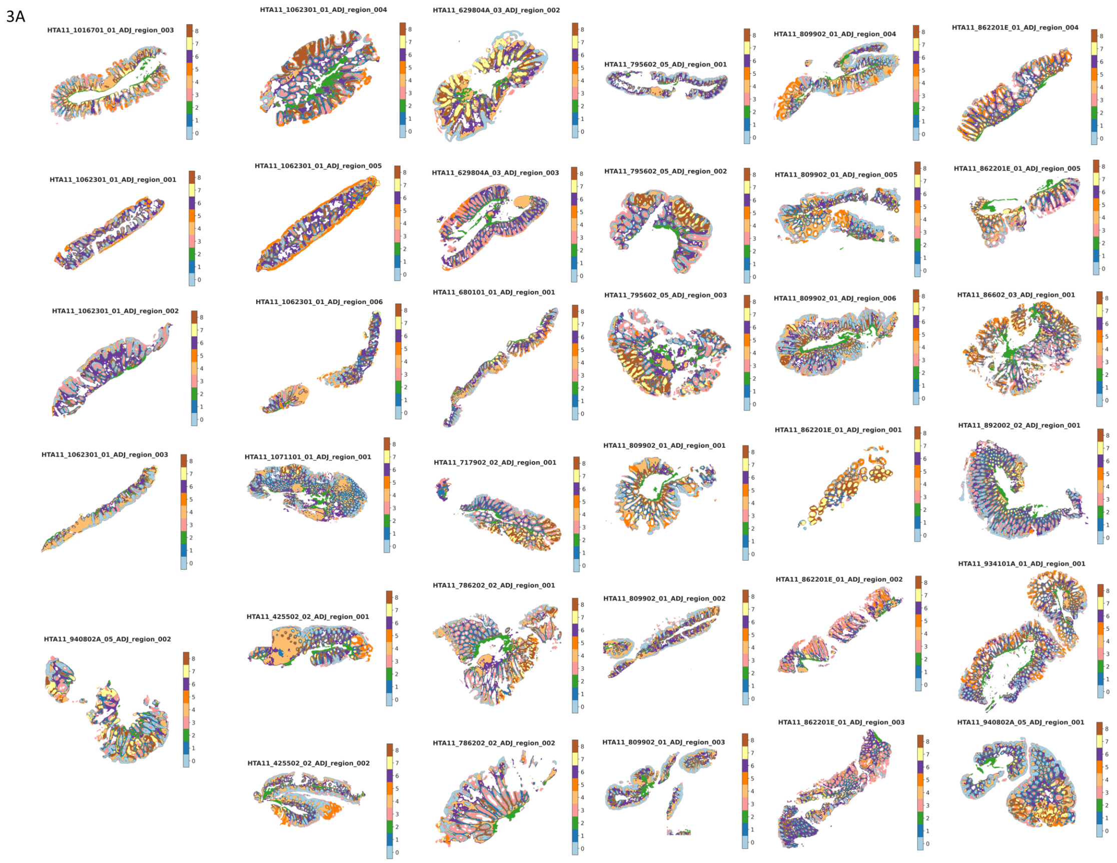


Figure S16: MILWRM QC metrics related to Figure 4.2.
(A) Mean Square error for each tissue domain
(B) UMAP with tissue domains.



HTA11_1016701_01_ADJ_region_003

HTA11_1062301_01_ADJ_region_004

HTA11_629804A_03_ADJ_region_002

HTA11_795602_05_ADJ_region_001

HTA11_809902_01_ADJ_region_004

HTA11_862201E_01_ADJ_region_004

HTA11_1062301_01_ADJ_region_001

HTA11_1062301_01_ADJ_region_005

HTA11_629804A_03_ADJ_region_003

HTA11_795602_05_ADJ_region_002

HTA11_809902_01_ADJ_region_005

HTA11_862201E_01_ADJ_region_005

HTA11_1062301_01_ADJ_region_002

HTA11_1062301_01_ADJ_region_006

HTA11_680101_01_ADJ_region_001

HTA11_795602_05_ADJ_region_003

HTA11_809902_01_ADJ_region_006

HTA11_86602_03_ADJ_region_001

HTA11_1062301_01_ADJ_region_003

HTA11_1071101_01_ADJ_region_001

HTA11_717902_02_ADJ_region_001

HTA11_809902_01_ADJ_region_001

HTA11_862201E_01_ADJ_region_001

HTA11_892002_02_ADJ_region_001

HTA11_940802A_05_ADJ_region_002

HTA11_425502_02_ADJ_region_001

HTA11_786202_02_ADJ_region_001

HTA11_809902_01_ADJ_region_002

HTA11_862201E_01_ADJ_region_002

HTA11_934101A_01_ADJ_region_001

HTA11_425502_02_ADJ_region_002

HTA11_786202_02_ADJ_region_002

HTA11_809902_01_ADJ_region_003

HTA11_862201E_01_ADJ_region_003

HTA11_940802A_05_ADJ_region_001

Figure S17: Tissue domain labels for human colonic-adenoma ($\alpha = 0.02$).

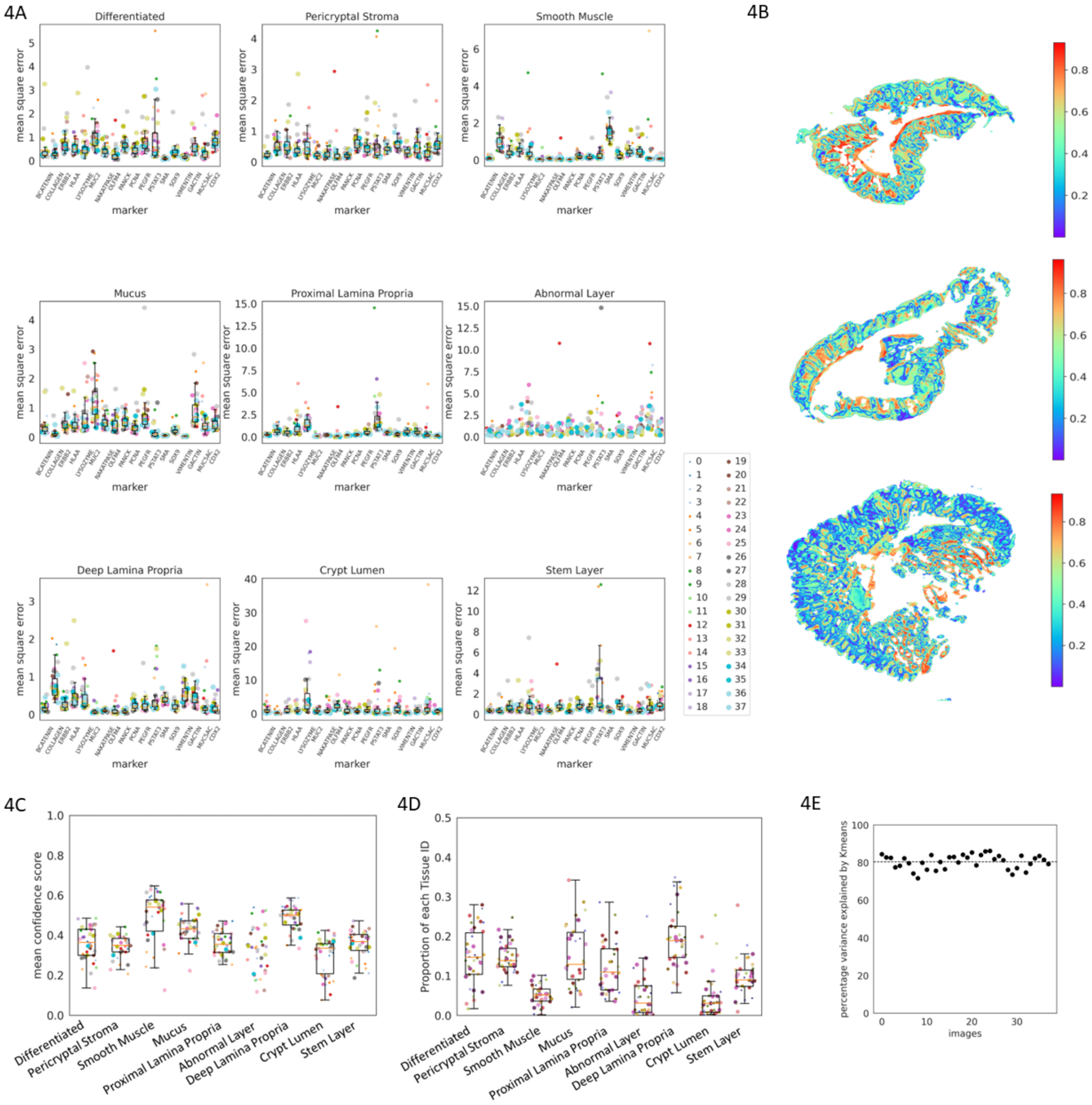


Figure S18: MILWRM QC metrics related to Figure 4.3.
(A) Mean Square error for each tissue domain
(B) Confidence score overlaid on three representative tissues
(C) mean confidence score for each tissue domain for all colonic adenoma samples
(D) Proportion of each tissue domain in each slide
(E) Variance explained by the Kmeans model.

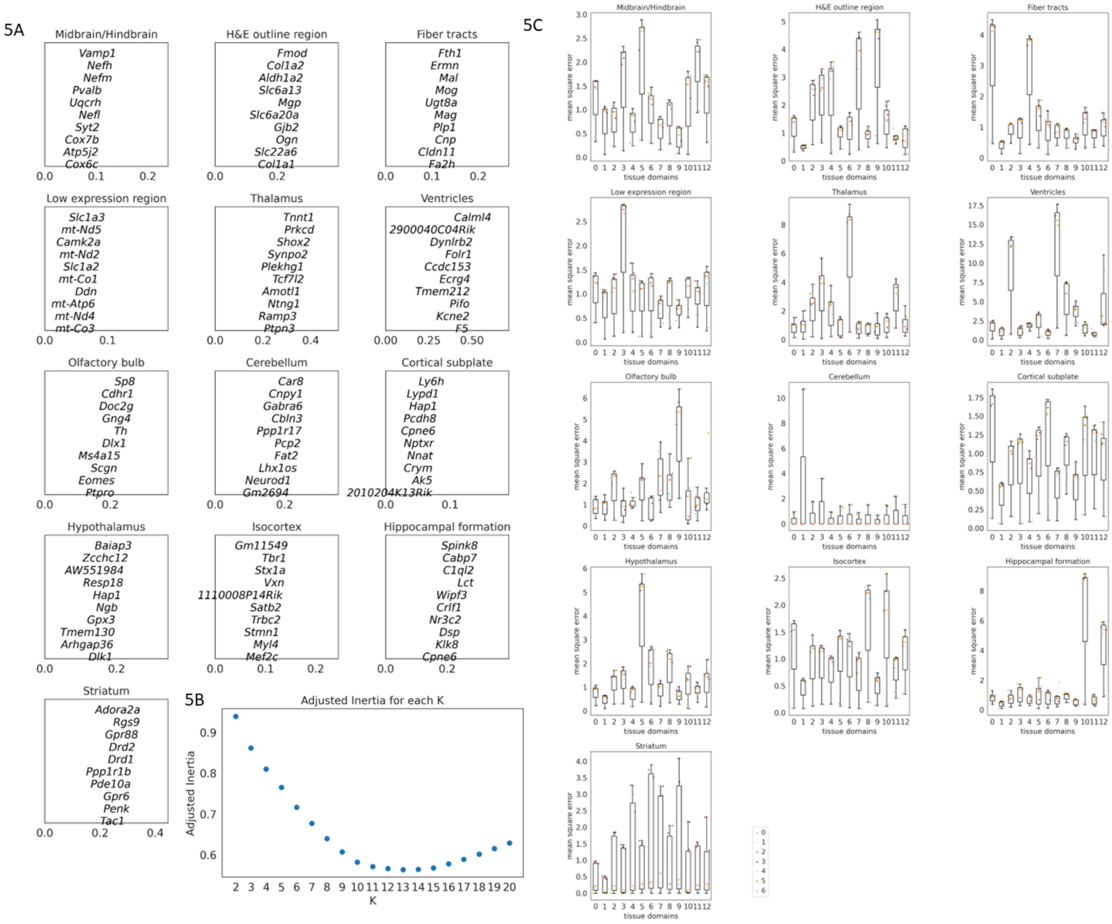


Figure S19: Domain profile for mouse brain tissue domains and MILWRM QC metrics.
(A) Domain profile for tissue domains in mouse brain ST
(B) Estimated number of tissue domains in Adjusted inertia plot
(C) Mean square error for each tissue domain.

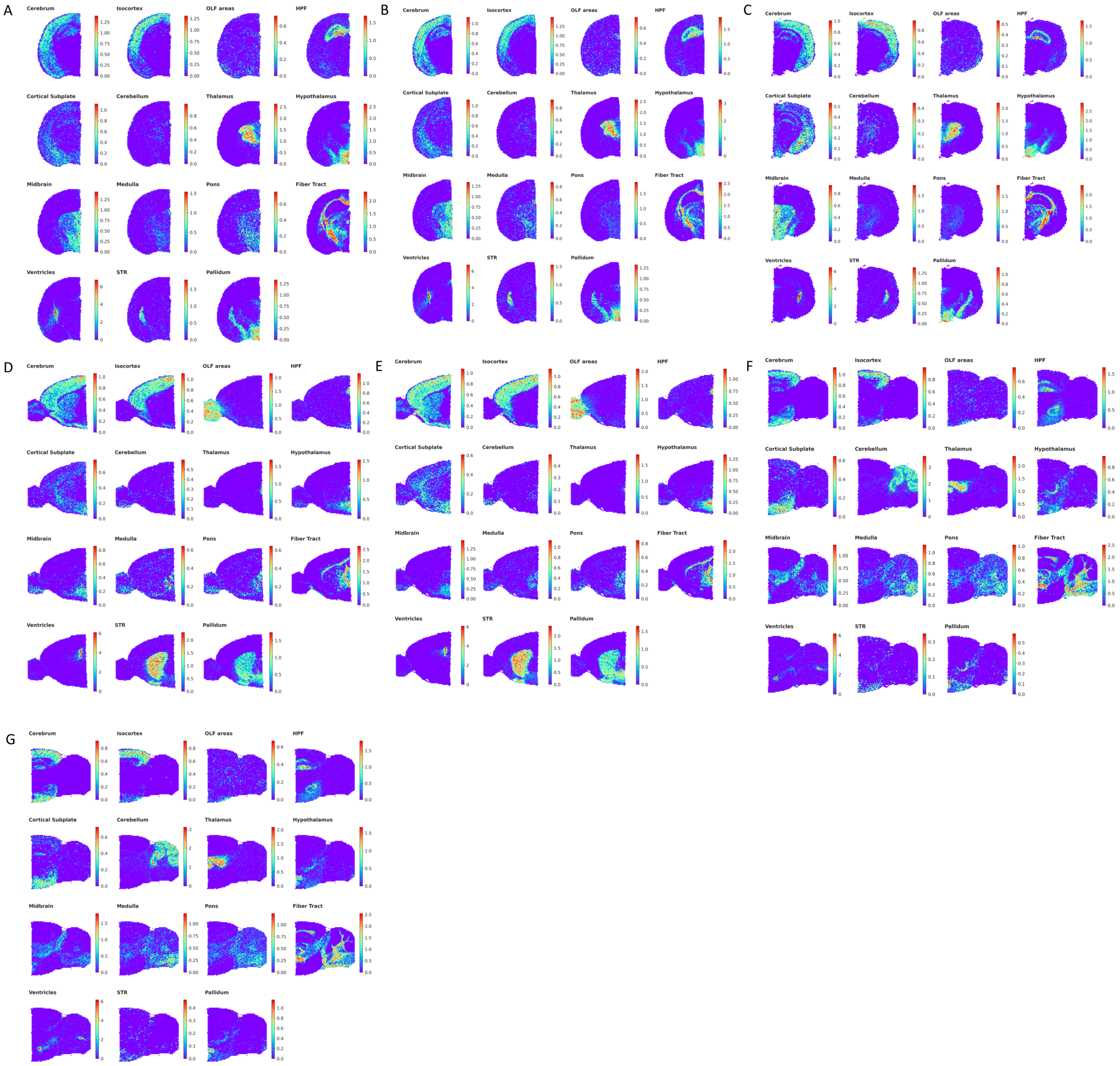


Figure S20: Reference score profiles for all mouse brain samples.
(A) Scores from Allen brain atlas anatomical regions coronal slice
(B) Scores from Allen brain atlas anatomical regions coronal slice replicate
(C) Scores from Allen brain atlas anatomical regions coronal slice 2
(D) Scores from Allen brain atlas anatomical regions sagittal anterior slice
(E) Scores from Allen brain atlas anatomical regions sagittal anterior slice replicate
(F) Scores from Allen brain atlas anatomical regions sagittal posterior slice
(G) Scores from Allen brain atlas anatomical regions sagittal posterior slice replicate.

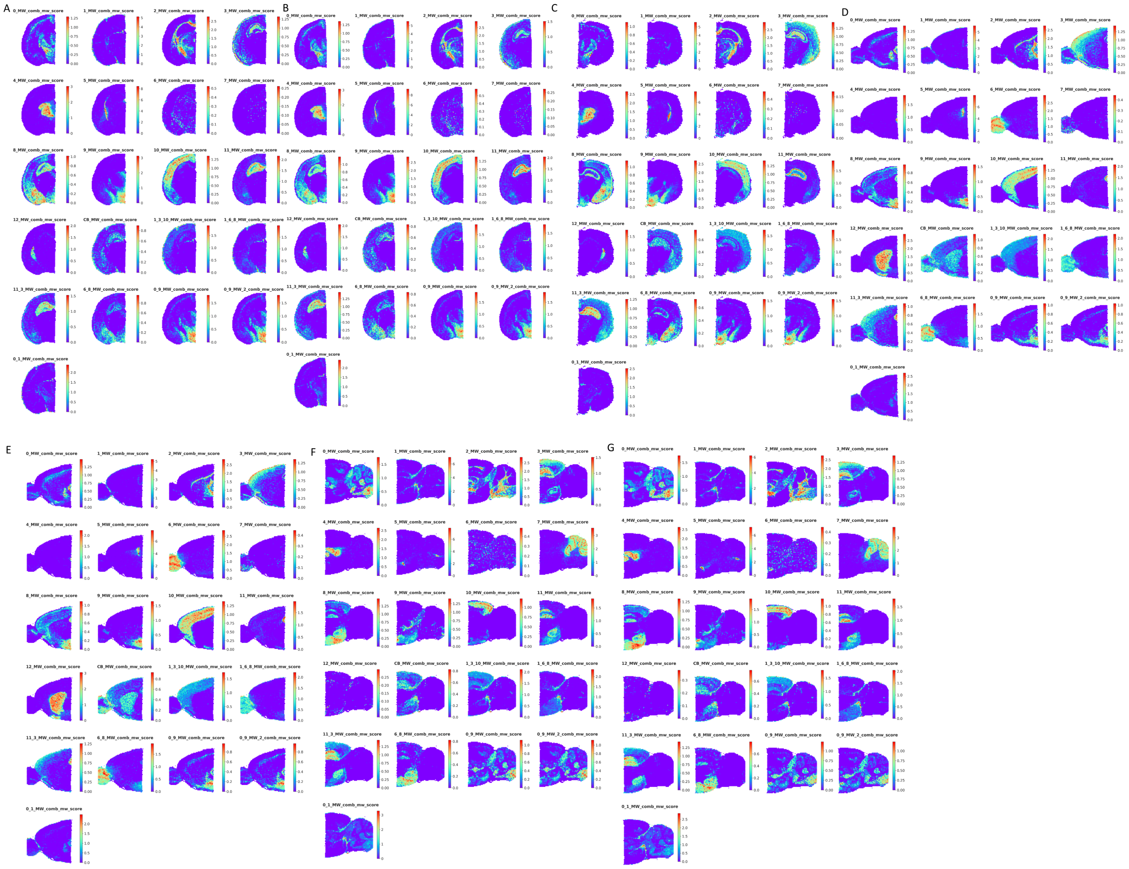


Figure S21: MILWRM domain score profiles for all mouse brain samples.
(A) Scores for MILWRM domains coronal slice
(B) Scores for MILWRM domains coronal slice replicate
(C) Scores for MILWRM domains coronal slice 2
(D) Scores for MILWRM domains sagittal anterior slice
(E) Scores for MILWRM domains sagittal anterior slice replicate
(F) Scores for MILWRM domains sagittal posterior slice
(G) Scores for MILWRM domains sagittal posterior slice replicate.

A5 Appendix for Chapter 5: Molecular cartography uncovers evolutionary and microenvironmental dynamics in sporadic colorectal tumors

A5.1 Methods

A5.1.1 Data and code availability

Data have been deposited to the HTAN Data Coordinating Center Data Portal at the National Cancer Institute: https://data.humantumoratlas.org/ (under the HTAN Vanderbilt Atlas).

A5.1.2 Sample procurement

These specimens were procured through the collaborative human tissue network (CHTN) as formalin-fixed, paraffin-embedded (FFPE) tissue blocks with accompanying pathology reports.

A5.1.3 Visium ST sample handling

Regions of interest (ROIs) for ST were chosen based on histological annotation of FFPE blocks, targeting tumor areas with morphology indicative of various stages of malignancy, and transition points between them. Tissue sections were cut and trimmed (if necessary) into 6.5 mm × 6.5 mm capture areas of 10X Genomics Visium FFPE spatial gene expression slides (Table S2). Serial tissue sections were collected simultaneously for whole-slide MxIF staining and laser capture microdissection.

Visium FFPE spatial gene expression slides were temporarily coverslipped, stained with hematoxylin and eosin (H&E; Table S2), and brightfield imaged at 20X objective prior to tissue permeabilization, probing, and library prep according to the 10X Genomics protocol. Sample libraries were sequenced on an Illumina NovaSeq targeting 125M reads per capture area. Resulting sequencing data were aligned using 10X Genomics Space Ranger software version 1.3.0 (Table S2).

A5.1.4 Visium TMA building

A subset of FFPE blocks was chosen for building tissue microarrays (TMAs), where three to five 1 mm punches were collected and arrayed in a 3 × 3 format for sectioning into 6.5 mm × 6.5 mm capture areas of 10X Genomics Visium FFPE spatial gene expression slides (Table S2). Adjacent 1 mm punches to each TMA region of interest (ROI) were collected in tubes for direct DNA extraction and whole-exome sequencing (WES) library preparation, providing analogous molecular information to LCM-WES data.

A5.1.5 Multiplex immunofluorescence (MxIF) imaging

A cyclic staining, imaging, fluorophore inactivation protocol for multiplexed protein imaging was employed as shown previously Gerdes et al., 2013. A panel of 33 antibodies was used for staining, as detailed in Table S2.

Virtual H&E stains were generated from autofluorescence (AF) images for each block, and are used to orient Visium data to whole-slide tissue morphology following spatial registration.

A5.1.6 Spatial registration

We developed a custom Python plugin for `napari`, a multidimensional viewer for biological images Sofroniew et al., 2022, which allowed us to perform affine transformation and scaling on Visium brightfield images in order to spatially align morphology features of the tissue with whole-slide MxIF. The `napari` plugin exports affine matrices and final image sizes, which can be applied directly to MxIF to align pixels with Visium ST microwells, or applied in reverse to Visium spots to cast them into whole-slide space on top of MxIF images.

Registration of LCM-WES with ST was performed manually by creating masks for each LCM ROI in the 10X Genomics Loupe Browser that were used to subset Visium microwells to LCM ROIs for downstream analysis.

A5.1.7 Gene signature scoring

Gene expression signatures were curated from the literature to interrogate cellular identity and activity in scRNA-seq and ST samples Nirmal et al., 2018; Gulati et al., 2020; Chen et al., 2021b; Combes et al., 2022; Joanito et al., 2022; Barkley et al., 2022; Gil Vasquez et al., 2022. Lists of genes were passed to the `scanpy` function `score_genes`, with default parameters. There were two exceptions:

- ”CytoTRACE” scores were calculated using the CytoTRACE R package version 0.3.3 Gulati et al., 2020
- When scoring ”iCMS2” and ”iCMS3” tumor epithelial signatures from Joanito, *et al.*, genes from the ”Up” list were scored against genes from the ”Down” list from each respective iCMS classification by restricting the `gene_pool` parameter of the `score_genes` function to the combined ”Up” and ”Down” gene lists Joanito et al., 2022

A5.1.8 CNV inference from ST and scRNA-seq

We used the `infercnvpy` Python package to infer somatic CNVs from scRNA-seq and ST gene expression Patel et al., 2014. Analysis was performed separately on scRNA-seq and ST. For scRNA-seq analysis, we used all normal epithelial and stromal cells from each patient, as labeled in Chen et al., 2021b, to provide a normal background for calling CNVs in malignant cells derived from the same patient. In ST, we grouped Visium microwells by patient and used stromal regions and adjacent normal epithelium, manually annotated in the 10X Genomics Loupe Browser, to provide a normal background for calling CNVs in tumor regions. For patient samples lacking sufficient stromal/normal surface area (mostly pre-malignant polyps and TMAs), we grouped multiple patients with the normal human Swiss roll sample (SR00001), which provided a chromosomally-stable reference for inferCNV analysis. Grouped samples included SR00001 (NL), HTA11_01938 (TVA), WD33469 (TMA), WD33473 (TMA), and WD33474 (TMA).

A5.1.9 CNV calling from WES and WGS

FFPE curls were treated with the truXTRAC FFPE total NA kit from Covaris (Table S2) to extract DNA prior to library preparation. Whole-genome (WGS) libraries were prepared using a modified Twist Bioscience Human Genome Panel (Table S2) and sequenced on an Illumina NovaSeq, targeting 50X coverage genome-wide. Whole-exome libraries were prepared using the Twist Bioscience Human Comprehensive Exome Panel (Table S2) and sequenced on an Illumina NovaSeq, targeting 50X coverage exome-wide.

Somatic CNVs were called following `GATK4` Best Practices workflow and annotated with `GATK4` Funcotator data sources v1.6. We used the `CNVkit` Python package to summarize and visualize somatic CNVs in WES and WGS samples Talevich et al., 2016.

A5.1.10 Somatic mutational profiling with LCM-WES

We performed laser capture microdissection (LCM) on FFPE sections serial to Visium ST samples using the Arcturus XT LCM system from Thermo Fisher Scientific. Circular ROIs 1.5 mm - 2.0 mm in diameter were collected from spatially distinct tissue regions, targeting abnormal tumor epithelium with morphology indicative of various stages of malignancy, and transition points between them. For samples lacking adjacent normal colon biopsies or whole-blood samples for bulk WES, we dissected additional ROI(s) from adjacent normal epithelium or stroma present in the FFPE sections to use as germline reference for mutation calling.

We extracted DNA from dissected FFPE ROIs using the Arcturus PicoPure DNA Extraction Kit from Applied Biosystems (Table S2). Alternatively, FFPE cores from Visium TMAs were treated with the truXTRAC FFPE total NA kit from Covaris (Table S2) to extract DNA prior to library preparation. Whole-exome libraries were prepared using the Twist Bioscience Human Comprehensive Exome Panel (Table S2) and sequenced on an Illumina NovaSeq, targeting 50X coverage exome-wide.

FASTQ reads were trimmed to remove adapter sequences using `Cutadapt` v2.10. Quality control on both raw reads and adapter-trimmed reads was performed using `FastQC` v0.11.9. The reads were then aligned to the human reference genome `hg19` using `BWA` v0.7.17. Duplicated reads were removed and the alignments were refined with `GATK4` Mark Duplicates and Base Quality Score Recalibration tools. Somatic variants were called using `GATK4 Mutect2` in “normal-tumor” paired mode and annotated with `ANNOVAR` Wang et al., 2010 (2019/Dec/05 version).

A5.1.11 Phylogenetic tree construction from LCM-WES

Mutation Annotation Format (MAF) output files from WES alignment and mutation calling were processed with the `maftools` R package to generate oncoplots and summarize mutations within and between samples Mayakonda et al., 2018. We then used the `MesKit` R package to generate phylogenetic trees for each patient Liu et al., 2021.

A5.1.12 Global PPT ordering and classification of tumor regions

Following spatial registration of ST and LCM-WES, we divided tumors into areas of overlap between ST CNV clones (non-"E" and non-"S" regions) and LCM ROI masks (referred to as "tumor regions"), in order to provide paired mutational and CNV information for each region. Tumor regions containing less than 30 microwells were disposed for reliable summarization of cell states, genes, and gene signatures. Pre-malignant tumors (TA/TVA and SSL/HP) were simply subset to their major CNV clone region, as their WES data was collected in bulk due to small size of the lesions preventing LCM analysis. The normal colon specimen (SR00001) was also subset to its epithelial area using the two CNV clone regions detected. Additionally, we removed HTA11_01938 from tumor region ordering due to extreme hypermutation likely resulting from compromised DNA excision repair and proofreading machinery (Figure 5.1F). Exclusion of this outlier prevents skewing of cell states and gene expression by a pre-malignant polyp assigned to late HM PPT.

Resulting tumor regions comprising groups of ST microwells sub-divided each macro tumor into smaller, epithelial-only regions which could be summarized based on average expression of genes, gene signatures, and cell states. We then calculated CIN+ PPT, defined as CNV score scaled between 0 and 1 using `sklearn.preprocessing.MinMaxScaler` Pedregosa et al., 2011 across all tumor regions. Likewise, HM PPT was calculated as TMB scaled between 0 and 1 using `sklearn.preprocessing.MinMaxScaler` Pedregosa et al., 2011 across all tumor regions. Next, we created a quantitative "CIN index" for each tumor region, calculated as the difference between CIN+ PPT and HM PPT. In this way, tumor regions with a CIN index greater than zero are more chromosomally unstable than hypermutated, and regions with a negative CIN index are conversely more hypermutated.

We then classified these CNV clone-LCM ROI overlap regions as CIN+ or HM based on their tumor type (MMR status and pathological annotation) coupled with their CIN index. All tumor regions with PPT values less than 0.4 were assigned to tumor classes according to the tumor type they were derived from (NL, TA/TVA and MSS = CIN+; SSL/HP and MSI-H = HM). Seven late-stage (PPT > 0.4) MSI-H tumor regions had CIN indices > 0.0 (2/4 regions from SG0001, 2/3 regions from SG00002, and 3/4 regions from PAT73458), and were thus labeled CIN+ for downstream GAM modeling and analysis (Figure 5.3E,G; Figure S23X-Y).

We assigned CIN+ tumor class to SG00001, SG00002, and PAT73458 for the purposes of patient-level summarization and downstream analyses (Figure 5.4F-G; Figure 5.5C), as the majority of the tumor area of these MSI-H CRCs had transitioned to CIN+. Additionally, one tumor that was excluded from tumor region analysis above due to low number of ST microwells available for each CNV clone-LCM ROI overlap region, PAT15211, was categorized as HM due to high TMB (low CIN index). All other analyses were performed at the tumor region level and employed classifications described above (Figure 5.4H; Figure 5.5A-B; Figure S25A; Figure 5.6A).

Finally, we ordered all tumor regions along PPT according to their tumor class (Figure 5.3E; Figure S23X). These rankings provided a pseudotemporal basis for GAM fitting along two tumor growth trajectories representing the major classes of CRC (CIN+ and HM).

A5.1.13 refNMF cell-state discovery and deconvolution

Cell states from tumors and normal colonic mucosa were discovered in the VUMC scRNA-seq cohort Chen et al., 2021b using the cNMF Python package Kotliar et al., 2019. Normal and abnormal epithelial cells plus stromal and immune cells curated from Chen, *et al.* were combined prior to cNMF analysis, and genes were subset to the union of all genes detected in all 40 ST samples Chen et al., 2021b. The consensus NMF factorization was performed at an optimal k = 30 to yield representative cell-type factors as well as factors that describe cell-state subtypes. We factored the gene loading matrix from scRNA-seq consensus NMF out of the ST expression matrices using the `sklearn.decomposition.NMF` function Pedregosa et al., 2011. This reference NMF (”refNMF”) factorization provided fractional usages of each of the 30 cell-state factors in every ST spot.

A5.1.14 refNMF validation

MxIF stains registered to ST data were averaged within each Visium microwell area for all markers separately. Visium microwells were then blurred using the `squidpy.gr.spatial_neighbors` function with a `radius` value of 1 to capture local spatial neighborhood information of MxIF markers, refNMF cell-state fractions, and gene signature scores Palla et al., 2022b. The blurred values were correlated with one another across all Visium microwells from all tumors in the atlas in order to highlight protein markers and cell identity/activity signature scores that confirm refNMF state characterizations (Figure S24E-G).

A5.1.15 Tissue domain detection with MILWRM

We employed refNMF cell states as predictors for a MILWRM model of macro-level consensus tissue domains across all ST slides Kaur et al., 2023. We only included states from epithelial and stromal compartments (Figure 5.4G), excluding immune states to limit predictors to markers of tissue architecture. We used a radius of 1 Visium ST ring for smoothing predictor values, and an α of 0.02 for regularization of scaled inertia during k optimization.

A5.1.16 Modeling expression dynamics along PPT

We built generalized additive models (GAMs) for genes, gene signatures, and refNMF cell states along global PPT using the `tradeSeq` R package Van den Berge et al., 2020. The pseudotime (`PT`) value given to the model was defined as CNV score for CIN+ tumors and TMB (number of somatic mutations detected per tumor region) for HM tumors. GAMs were built separately for CIN+ and HM PPT, treating each tumor class as a unique trajectory. We employed the `startVsEndTest` to detect genes, gene signatures, and cell states with differential expression between early and late PPT in both tumor classes, as well as the `diffEndTest` for statistically significant differences in late-PPT expression between the two classes. Resulting significant features from the two tests were curated for heatmap plotting (Figure 5.4H; Figure 5.5A; Figure 5.6A).

A5.1.17 Spatial co-occurrence analysis from ST

We employed the `squidpy` Python package for spatial co-occurrence analysis of IES and infiltrating immune cell states Palla et al., 2022b. We first thresholded the signature or cell state of interest to label microwells with ”high” expression. Then, we used major CNV clone regions as a reference label to measure co-occurrence with ”high” values of the signature or cell state in question. We also employed the stromal (”S”) CNV clone region as a negative control (Figure 5.5E,K).

A5.1.18 MxIF immune-exclusion analysis

We segmented MxIF images into single cells using the `MIRIAM` segmentation algorithm McKinley et al., 2022. Thresholds were manually determined for all markers on a slide-by-slide basis in order to call positive and negative pixels for each marker while avoiding slide-specific illumination or staining batch effects. Within each single cell area, markers were further binarized based on 50 % pixel area or greater, reducing markers to present or absent per single cell segment. Immune cells were then identified by co-expression of the following markers (Figure 5.5G,I):

- T helper - CD3D and CD4
- T reg - CD3D, CD4, and FOXP3
- T cytotoxic - CD3D and CD8
- Myeloid - CD11B

- Macrophage - CD11B and CD68
- Macrophage M1 - CD11B, CD68, and Lysozyme

ST data for each corresponding slide were transformed into MxIF space using an affine transformation according to `napari`-based spatial registration between MxIF and ST. Segmented MxIF data, consisting of cell centroids as x-y pixel coordinates and cell-type IDs (subset to immune cells only, identified with markers above), were then counted for each CNV clone region in the transformed ST data based on centroid pixel coordinates and ST CNV clone masks (Figure 5.5L-M; Figure S25C).

A5.1.19 TCGA immune exclusion and survival analysis

Normalized expression data (RSEM TPM) and matched clinical information from TCGA COAD and READ samples were accessed and integrated from cBioPortal. Corresponding MMR status (MSS vs. MSI-H) were extracted from GDC Data Portal. Samples with `NA` values for any of the IES genes were filtered out, yielding 301 total tumors for IES scoring. For stratification by MMR status and survival analysis, these samples were further filtered, removing additional tumors with `NA` values in any of the metadata columns ”Age”, ”Gender”, ”Race”, or ”Disease progressive event and time”. Final sample sizes for survival analysis stratified by MMR status were 244 MSS and 45 MSI-H.

IES scores were calculated as the average TPM expression of *DDR1*, *TGFBI*, *PAK4*, and *DPEP1*. Survival analyses were then performed with the `survival` R package (version 3.4-0) (Figure 5.6B-C; Figure S26A-D).

A5.1.20 Antibody conjugation and immunohistochemical (IHC) imaging

Human CRC TMA slides were cut at 5 µm on positively charged slides. The slides were de-paraffinized in three changes of xylene and hydrated to water in graded alcohol. The slides underwent antigen retrieval using a citrate buffer (Ph 6.0) solution at 105°C in a pressure cooker for 20 minutes followed by a bench cool down for 10 minutes. The slides were washed in distilled water and placed in TBST wash buffer solution for continuation of the staining protocol. Endogenous enzymes were blocked using a 0.03% (H_2O_2) peroxidase block solution for 5 minutes, rinsed in wash buffer and incubated for one hour in the primary antibodies (PAK4 Santa Cruz Biotechnology sc-390507 1:75 primary dilution, TGFBI Abcam ab170874 1:300 primary dilution, DDR1 Cell Signaling #5583 primary dilution 1:2000; Table S2). The slides were gently rinsed and a peroxidase labelled polymer was applied utilizing Dako EnVision + System -HRP labeled polymer for a 30 incubation period. After a gentle rinse, the slides were treated with a DAB+ Substrate-Chromogen for 5 minutes to complete the staining protocol. The slides were washed in distilled water, counterstained in

Mayer's Hematoxylin, blued in running tap water, dehydrated in 3 changes of absolute alcohol, and cleared in xylene prior to cover-slipping.

A5.1.21 IHC immune exclusion and survival analysis

From a CRC TMA representing tumors from 163 patients, 149 patients had at least one IHC marker measurement, and 147 of those had clinical data. These tumors were filtered to the subset where DDR1, DPEP1, or TGFBI staining was detected (*n* = 108, 86 MSS and 22 MSI-H). Each stain was graded as 0, 1, 2, or 3 by a pathologist based on increasing expression intensity, and cores were labeled ”+” for scores of 2 or 3 and ”-” otherwise (Figure 5.6F). Survival analyses were performed with the `survival` R package (version 3.4-0), comparing ”+” to ”-” expression for single markers or combinations of markers (Figure 5.6G; Figure S26E-G).

A5.2 Supplemental tables and figures

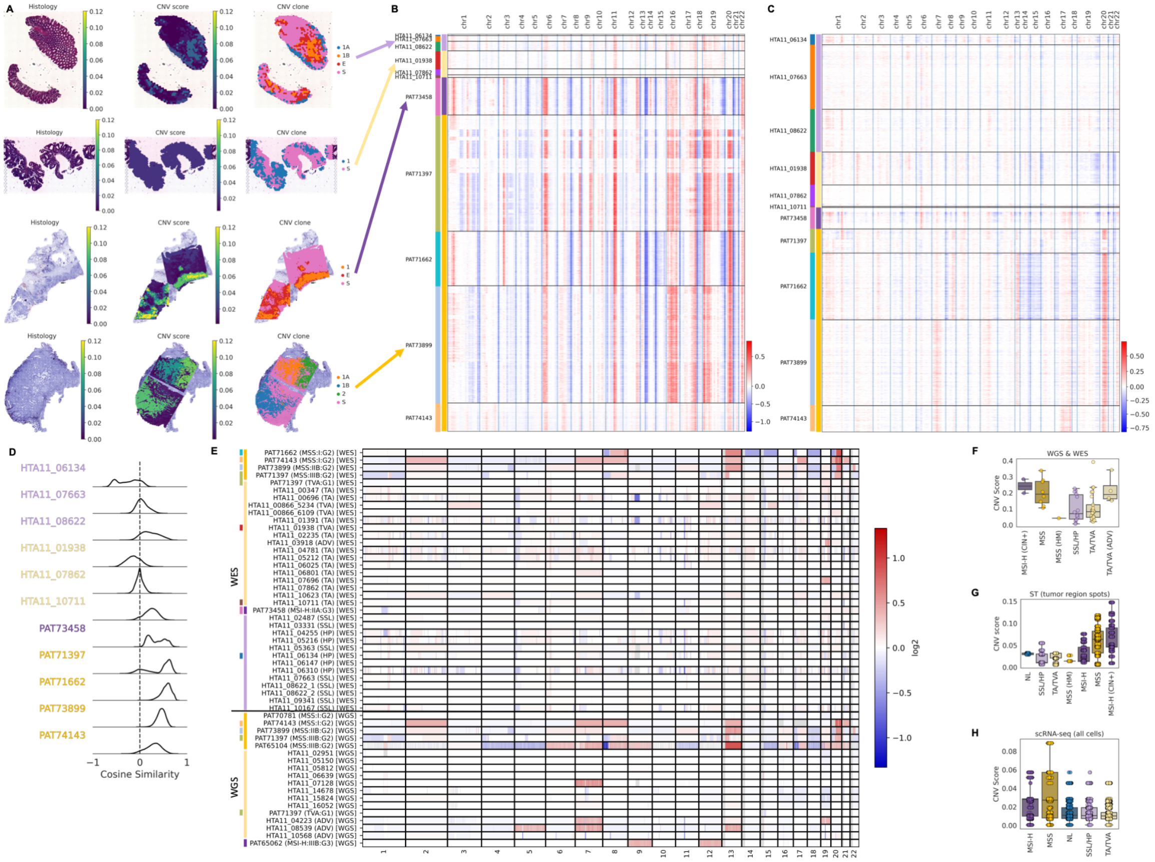


Figure S22: CNV inference establishes spatially resolved tumor clones and their phylogenetic relationships.
(A) Representative SSL, TVA, MSI-H, and MSS tumors from the CNV heatmap in B, showing histology alongside CNV score and CNV clone regions.
(B) Heatmap of inferred CNV profiles in the subset of ST samples with matched scRNA-seq. Samples are grouped by patient (left colorbar) and tumor type (right colorbar).
(C) Same as in B for scRNA-seq.
(D) Distribution of cosine similarity scores between the inferred CNV profiles of ST microwells (B) and single cells (C) derived from each patient. Both modalities were subset to major clone regions/clusters prior to comparison to exclude stroma and adjacent normal epithelium.
(E) Heatmap of CNVs detected in WES and WGS of select patients from ST atlas and Chen, *et al.* cohort of pre-cancerous polyps. Samples are colored by patient (left colorbar) and tumor type (right colorbar).
(F) Boxplots of CNV scores for WES and WGS samples in E, grouped by tumor type.
(G) Boxplots of CNV scores for all ST samples, grouped by tumor type.
(H) Boxplots of CNV scores for all scRNA-seq samples, grouped by tumor type.

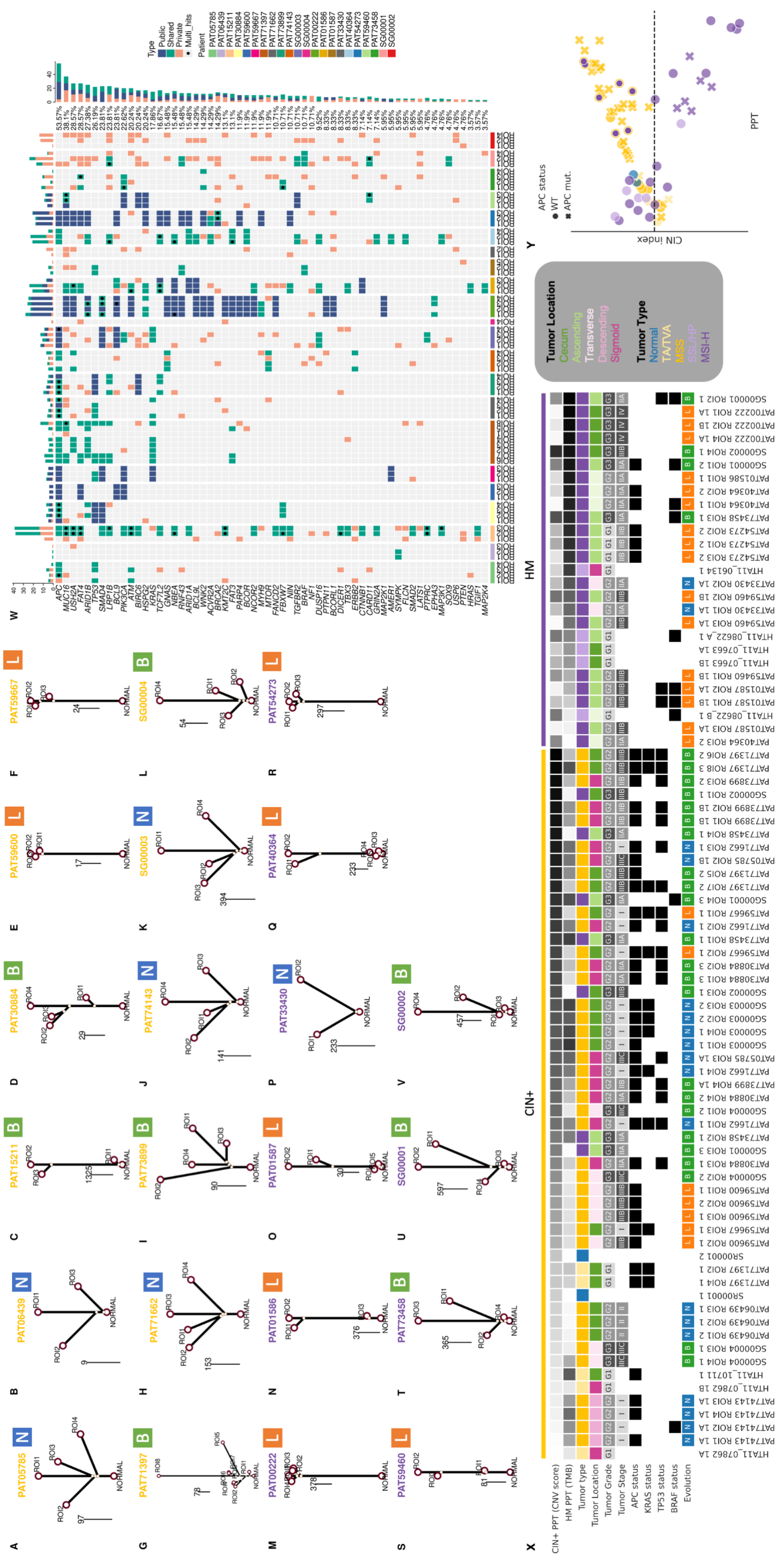

A
PAT05785
N
B
PAT06439
C
PAT15211
B
D
PAT30884
B
E
PAT59600
L
F
PAT59667
L
G
PAT71397
B
H
PAT71662
N
I
PAT73899
B
J
PAT74143
N
K
SG00003
N
L
SG00004
B
M
PAT00222
L
N
PAT01586
L
O
PAT01587
L
P
PAT33430
N
Q
PAT40364
L
R
PAT54273
L
S
PAT59460
L
T
PAT73458
B
U
SG00001
B
V
SG00002
B
NORMAL
W
X
Y
CIN+
HM
Type
Public
Shared
Private
Multi_hits
Patient
Tumor Location
Cecum
Ascending
Transverse
Descending
Sigmoid
Tumor Type
Normal
TA/TVA
MSS
SSL/HP
MSI-H
APC status
WT
APC mut.
CIN index
PPT
CIN+ PPT (CNV score)
HM PPT (TMB)
Tumor type
Tumor Location
Tumor Grade
Tumor Stage
APC status
KRAS status
TP53 status
BRAF status
Evolution

Figure S23 *(preceding page)*: Multiregional somatic mutational profiles provide phylogeographical topology.
(A-V) Phylogenetic trees constructed from LCM-WES of 22 CRCs from atlas. Length of branches are proportional to the number of shared or private somatic mutations detected in each region. Patient ID colors represent tumor type (MMR status). Icon next to patient ID indicates inferred mode of evolution based on tree structure (”L” = linear, ”B” = branching, ”N” = neutral).
(W) Oncoplot of detected driver mutations within the sampled regions of each patient in A-V.
(X) Tumor regions and their clinical and mutational metadata divided by class and ordered left-to-right by corresponding PPT (CNV score for CIN+, TMB for HM).
(Y) CIN index versus PPT for tumor regions from X. Points are colored by tumor class. Dark yellow points with purple centers represent MSI-H/CIN+ tumor regions. Point shape corresponds to regions with detected *APC* mutation.

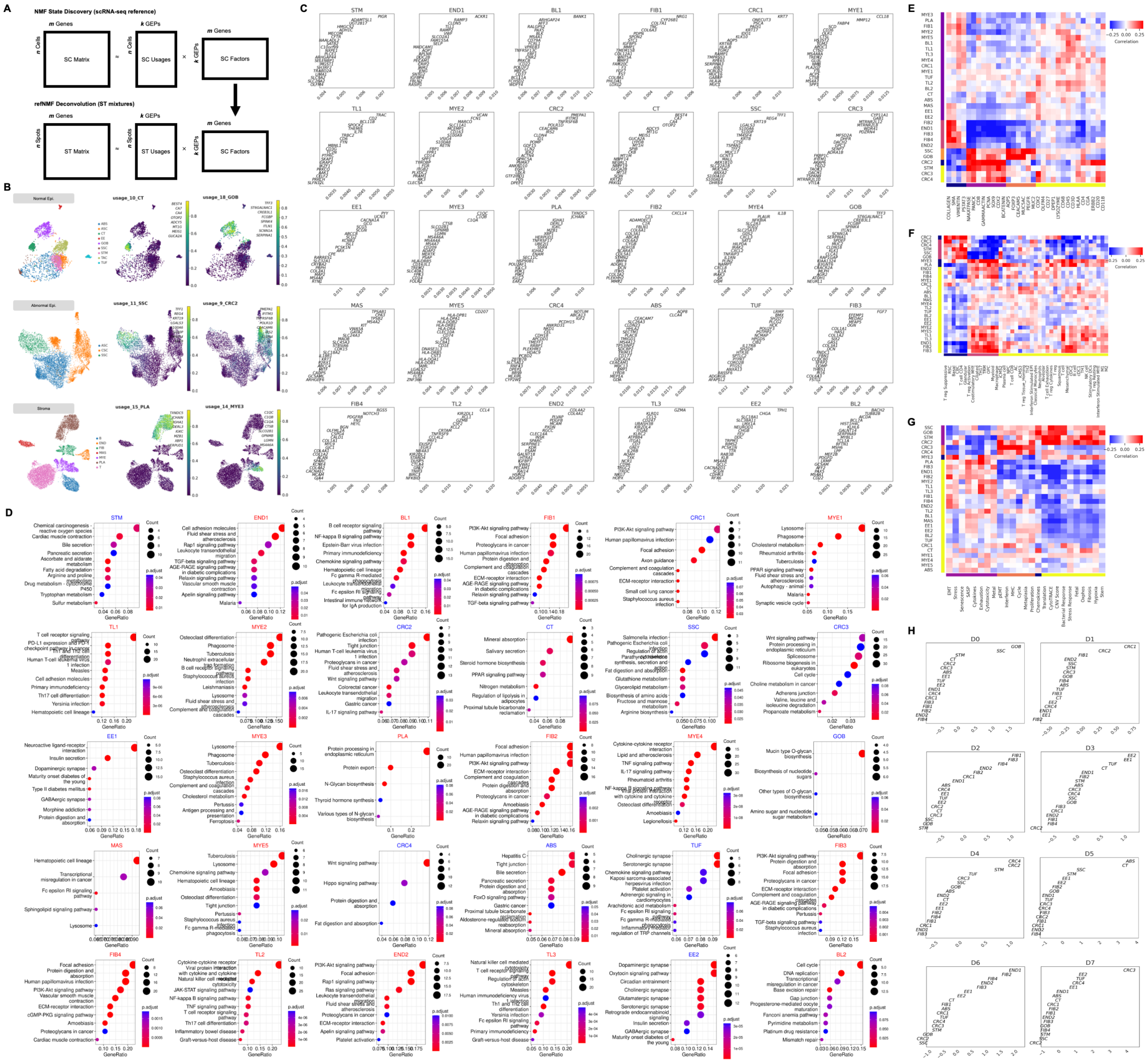


Figure S24: Cell-state deconvolution reveals pseudotemporal tissue dynamics.
(A) Diagram of linear algebraic doconvolution of ST expression mixtures using cell states (”factors”) discovered by NMF of scRNA-seq. These reference states (SC Factors) are factorized out of the ST matrix to assign fractional scores for all cell states to each microwell (ST Usages).
(B) Example ground-truth NMF factors in Chen, *et al.* scRNA-seq cohort profiling cell states in normal epithelial (top), abnormal/dysplastic epithelial (middle), and stromal (bottom) tissue compartments.
(C) refNMF cell-state gene loadings, showing top 25 genes for each of the 30 consensus states.
(D) KEGG term enrichment for the top 200 genes from each cell state in C. Top 2,000 genes were used for CRC3 to yield associated pathway terms.
(E) Spatial correlation matrix between refNMF cell-state usages (left) and average marker intensity from MxIF data spatially registered to ST from a serial tissue section (bottom).
(F) Same as in E, using cell identity gene signatures.
(G) Same as in E, using cell activity gene signatures.
(H) Tissue domain coefficients for MILWRM model built from all ST data in atlas.

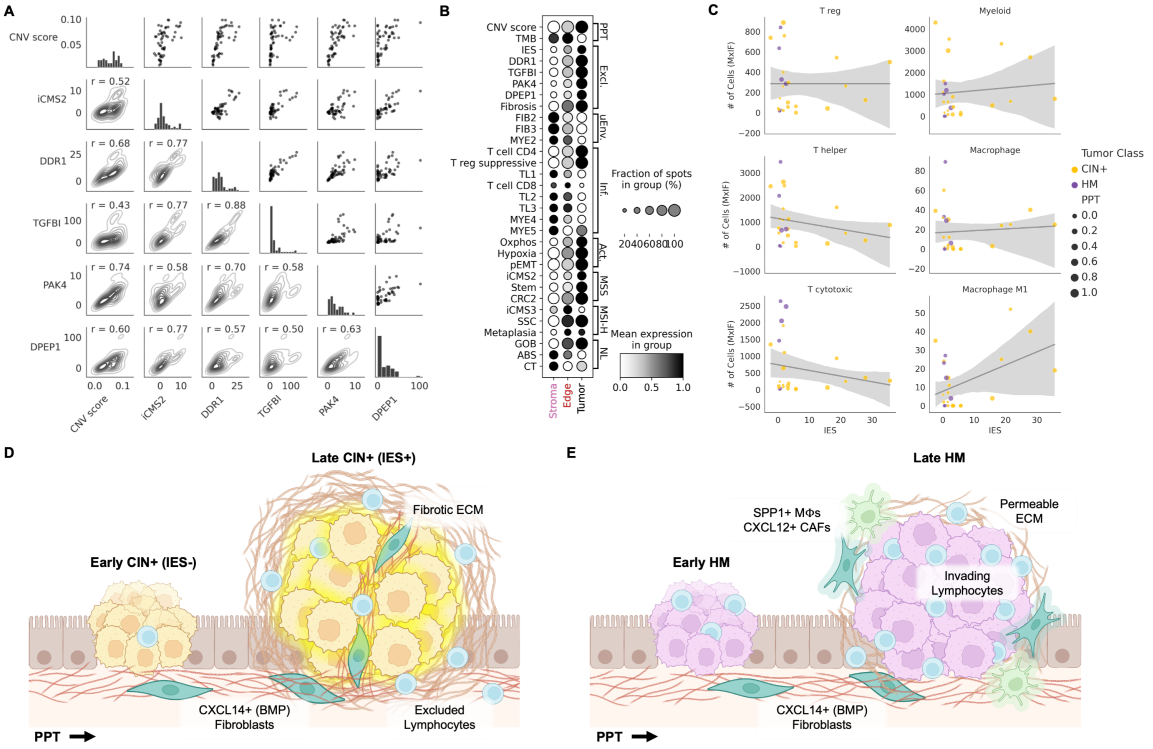


Figure S25: Gene expression features of CIN+ CRCs predict immune exclusion.
(A) Pairwise Pearson correlations between IES genes and progression indicators (CNV score and iCMS2) in all CIN+ tumor regions.
(B) Genes, gene signatures, and refNMF cell states grouped into pseudotime indicators (”PPT”), immune exclusion markers (”Excl.”), microenvironmental cells (”uEnv.”), infiltrating immune cells (”Inf.”), tumor activity (”Act.”), and epithelial-specific markers of MSS, MSI-H, and normal mucosa summarized across CNV tissue domain for all ST samples.
(C) Number of infiltrating immune cells detected in MxIF plotted against average IES score for all CNV clones in atlas. Points are colored by tumor class and sized according to their respective normalized PPT.
(D) Diagram of early-to-late CIN+ CRC development, highlighting proposed IES mechanism.
(E) Diagram of early-to-late HM CRC.

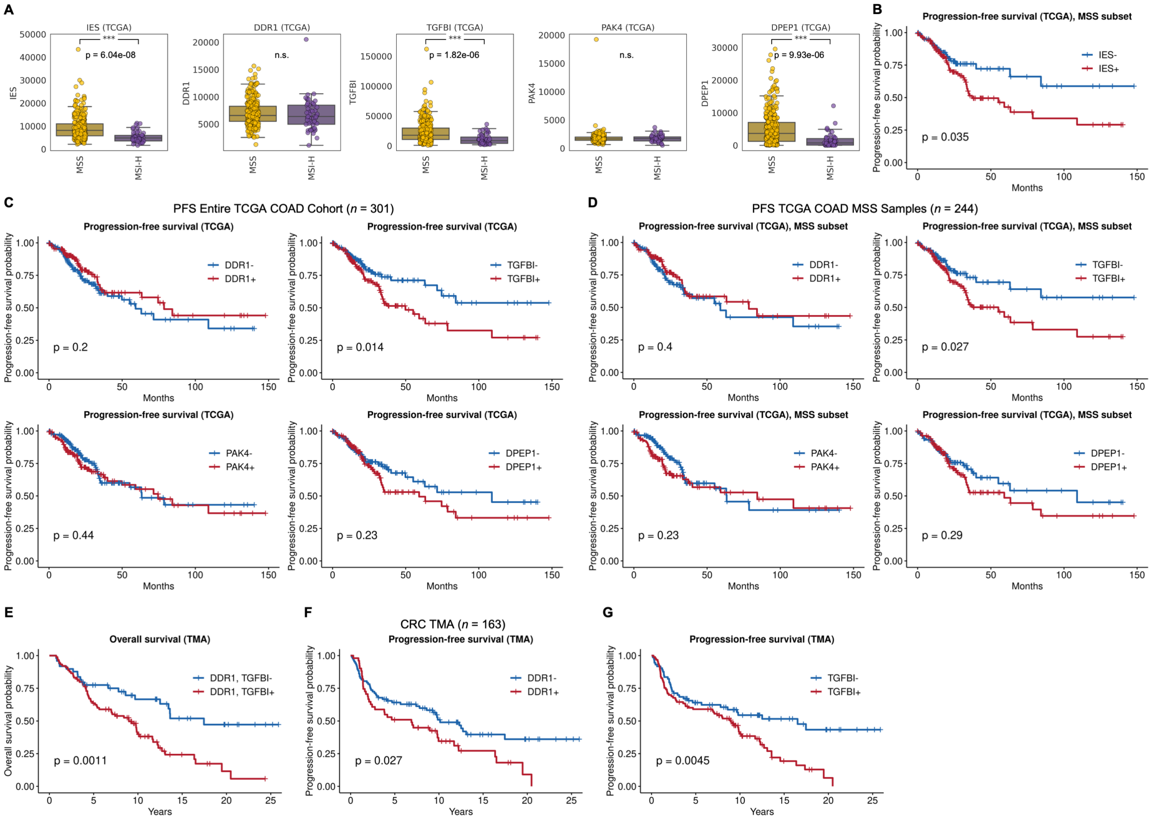


Figure S26: IES trends with tumor progression and predicts poor patient outcomes.
(A) Boxplots of IES scores and expression of individual constituent genes in TCGA COAD and READ samples, stratified by MMR status (MSS *n* = 244; MSI-H *n* = 45). Statistics shown represent Student's T-test with Bonferroni correction.
(B) Kaplan-Meier PFS curves for TCGA COAD and READ samples from A with high (+) and low (-) IES scores, subset to MSS tumors only (*n* = 244).
(C) Kaplan-Meier PFS curves for TCGA COAD and READ samples (*n* = 301) with high (+) and low (-) expression of individual constituent genes from IES (*DDR1*, *TGFBI*, *PAK4*, *DPEP1*).
(D) Same as in C, for MSS tumors only (*n* = 244).
(E) Kaplan-Meier PFS curves for CRC TMA cores with high (+) and low (-) DDR1 IHC staining.
(F) Same as in E, for TGFBI.
(G) Kaplan-Meier OS curves for CRC TMA cores with high (+) and low (-) IHC staining of both DDR1 and TGFBI.

A6 Key resources

Table S2: Key reagents, datasets, and software packages used in these studies

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Laboratory Reagents		
Mayer's Hematoxylin	Sigma	MHS16
Alcoholic Eosin	Sigma	HT110116
Bluing Buffer	Dako	CS702
UltraPure Glycerol	Invitrogen	15514-011
Visium Spatial for FFPE Gene Expression Kit, Human Transcriptome	10X Genomics	1000336
DAPI	Sigma-Aldrich	T9284
MUC2 Antibody (F-2)	Santa Cruz	sc-515032
Collagen Antibody (CHP)	3Helix	R-CHP
SNA Antibody (Lectin)	Vector	CL-1305-1
CD11B Antibody (C67F154)	Abcam	ab133357
CD20 Antibody (D-10)	Santa Cruz	sc-393894
PCNA Antibody (PC-10)	Santa Cruz	sc-56
BCATENIN Antibody (12F751)	Vanderbilt Antibody and Protein Resource	
PSTAT3 Antibody (D3A7)	Cell Signaling	4324S
PEGFR Antibody (EP774Y)	Abcam	ab205828
CGA Antibody (C-12)	Santa Cruz	sc-393941
CD4 Antibody (EPR6855)	Abcam	ab133616
COX2 Antibody (D5H5)	Cell Signaling	13596S
CD3D Antibody (EP4426)	Abcam	ab208514
HLAA Antibody (EP1395Y)	Abcam	ab199837

Continued on next page

Table S2: Key reagents, datasets, and software packages used in these studies (Continued)

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PANCK Antibody (AE1/AE3)	ThermoFisher	53-9003-82
OLFM4 Antibody (D1E4M)	Cell Signaling	Custom (87240BC)
CD8 Antibody (C8/114B)	Biolegend	13983
ACTININ Antibody (EPR2533(2))	Abcam	ab198608
CD68 Antibody (KP1)	Santa Cruz	sc-20060
NAKATPASE Antibody (EP1845Y)	Abcam	ab198367
VIMENTIN Antibody (E-5)	Santa Cruz	sc-373717
SOX9 Antibody (EPR14335)	Abcam	ab202516
FOXP3 Antibody (206D)	Biolegend	320113
LYSOZYME Antibody (E-5)	Santa Cruz	sc-518012
SMA Antibody (1A4)	Santa Cruz	sc-53015
ERBB2 Antibody (MAb414)	Abcam	ab225510
CD45 Antibody (H130)	Biolegend	304020
ACTG1 Antibody (1-17’)	Santa Cruz	sc-65638
MUC5AC Antibody (E309I)	Cell Signaling	Custom (43937BC)
CDX2 Antibody (D11D10)	Cell Signaling	Custom (84638BC)
DDR1 Antibody (D1G6)	Cell Signaling	5583
TGFBI Antibody (EPR12078(B))	Abcam	ab170874
PAK4 Antibody (B-3)	Santa Cruz	sc-390507
Arcturus PicoPure DNA Extraction Kit	Applied Biosystems	KIT0103
truXTRAC FFPE total NA kit	Covaris	520262
Human Comprehensive Exome Panel	Twist Bioscience	102033
Human Genome Panel	Twist Bioscience	custom protocol

Continued on next page

Table S2: Key reagents, datasets, and software packages used in these studies (Continued)

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Mouse retina scRNA-seq	Macosko et al., 2015	GEO: GSM1626793
Mouse colon scRNA-seq	Herring, Banerjee, et al., 2018	GEO: GSM2743164
Mouse bone marrow cyTOF	Weber Robinson, 2016	FlowRepository FR-FCM-ZZPH
Mouse bone marrow/blood scRNA-seq	Han, et al., 2018	Figshare 865e694ad06d5857db4b
293T Cells	Zheng, et al. 2017	support.10xgenomics.com/single-cell-gene-expression/datasets
Jurkat Cells		
9k Neurons	10x Genomics	
900 Neurons		
4k PBMCs		
4k Pan-T Cells		
Placenta		
3907_S1 3907_S2	Heiser, et al., 2021	GSE158636
Human colon ST	Heiser, et al., 2023	humantumoratlas.org/explore
Human colon multiregional WES		
Human colon MxIF		
Human CRC TMA		
Human colonic-adenoma mIF data	HTAN project	humantumoratlas.org/explore
Mouse brain ST data	10X Genomics	10xgenomics.com/resources/datasets
Human colon scRNA-seq	Chen et al., 2021	humantumoratlas.org/explore
Human colon bulk WES		
Human colon MxIF		

Continued on next page

Table S2: Key reagents, datasets, and software packages used in these studies (Continued)

REAGENT or RESOURCE	SOURCE	IDENTIFIER
TCGA COAD READ RNA-seq clinical data	cBioportal	cbioportal.org
TCGA COAD READ MMR status	NCI GDC Data Portal	portal.gdc.cancer.gov
Software and Algorithms		
Space Ranger version 1.3.0	10X Genomics	support.10xgenomics.com/spatial-gene-expression/software/downloads
Loupe Browser version 6.4.0	10X Genomics	10xgenomics.com/products/loupe-browser/downloads
Python version 3.8.10	Python Software Foundation	python.org
R version 4.2.2	The R Foundation	r-project.org
FastQC v0.11.9	Andrews, 2010	bioinformatics.babraham.ac.uk/ projects/fastqc/
Cutadapt v2.10	DOI:10.14806/ej.17.1.200	cutadapt.readthedocs.io
BWA v0.7.17	Li Durbin, 2010	bio-bwa.sourceforge.net
GATK v4.1.8.1	Van der Auwera O’Connor, 2020	gatk.broadinstitute.org
Annovar v2019/Dec/05	Wang, Li, Hakonarson, 2010	annovar.openbioinformatics.org
Software (Python Packages)		
DCA version 0.2.3	Eraslan et al., 2019	github.com/theislab/dca
FIt-SNE	Linderman et al., 2019	github.com/KlugerLab/FIt-SNE
networkx version 2.2	Hagberg, Schult, Swart 2008	networkx.github.io
PhenoGraph 1.5.2	Levine et al., 2015	github.com/jacoblevine/PhenoGraph
POT version 0.6.0	Flamary et al., 2021	github.com/rflamary/POT

Continued on next page

Table S2: Key reagents, datasets, and software packages used in these studies (Continued)

REAGENT or RESOURCE	SOURCE	IDENTIFIER
scVI version 0.5.0	Lopez et al., 2018	github.com/YosefLab/scVI
Scvis version 0.1.0	Ding, Condon and Shah, 2018	github.com/shahcompbio/scvis
umap-learn version 0.3.10	Mcinnes and Healy, 2018	github.com/lmcinnes/umap
ZIFA version 0.1	Pierson and Yau, 2015	github.com/epierson9/ZIFA
dropkick version 1.2.3	Heiser, et al., 2021	pypi.org/project/dropkick/
MILWRM version 1.1.0	Kaur Heiser, et al., 2023	pypi.org/project/MILWRM/
napari version 0.4.16	Sofroniew et al., 2022	napari.org
scanpy version 1.9.1	Wolf, Angerer and Theis, 2018	github.com/theislab/scanpy
matplotlib version 3.0.3	Hunter, 2007	matplotlib.org
numpy version 1.22.4	Oliphant, 2006	numpy.org
pandas version 1.4.2	McKinney et al., 2010	pandas.pydata.org
scipy version 1.8.1	Oliphant, 2007	scipy.org
seaborn version 0.11.2	Waskom, et al., 2014	seaborn.pydata.org
scikit-learn version 1.1.1	Pedregosa et al., 2011	scikit-learn.org
scikit-image version 0.19.2	van der Walt, et al., 2014	scikit-image.org
infercnvpy version 0.4.0	Patel et al., 2014	github.com/icbi-lab/infercnvpy
cnvkit version 0.9.9	Talevich et al., 2016	github.com/etal/cnvkit
squidpy version 1.2.2	Palla et al., 2022	github.com/scverse/squidpy
Software (R Packages)		
Seurat version 3.0.0	Butler et al., 2018	satijalab.org/seurat
SIMLR version 1.8.1	Wang et al., 2017	github.com/BatzoglouLabSU/SIMLR

Continued on next page

Table S2: Key reagents, datasets, and software packages used in these studies (Continued)

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GLM-PCA	Townes et al., 2019	github.com/willtownes/scrna2019
ZINB-WaVE version 1.4.2	Risso et al., 2018	bioconductor.org/packages/zinbwave
splatter version 1.8.0	Zappia, Phipson, Oshlack, 2017	DOI: 10.18129/B9.bioc.splatter
DropletUtils version 3.11	Lun, et al. 2019	10.18129/B9.bioc.DropletUtils
CellBender version 0.2.0	Fleming et al., 2019	github.com/broadinstitute/CellBender
UpSetR version 1.4.0	Lex, et al. 2014	doi:10.1109/TVCG.2014.2346248
sc-UniFrac version 0.9.6	Liu, et al. 2018	github.com/liuqivandy/scUnifrac
geepack version 1.3.9	Halekoh, et al., 2006	jstatsoft.org/article/view/v015i02
CytoTRACE version 0.3.3	Gulati et al., 2020	cytotrace.stanford.edu
maftools version 2.12.0	Mayakonda et al., 2018	bioconductor.org/packages/maftools
MesKit version 1.6.0	Liu et al., 2021	bioconductor.org/packages/MesKit
tradeSeq version 1.8.0	Van den Berge et al., 2020	bioconductor.org/packages/tradeSeq
survival version 3.4-0	Therneau et al., 2009	github.com/therneau/survival