SMURF Reconstructs Single-Cells from Visium HD Data to Reveal Zonation of Transcriptional Programs in the Intestine

Juanru Guo[1,2], Simona Sarafinovska[1,5], Ryan A. Hagenson[2,3], Mark C. Valentine[2,4], Joseph D. Dougherty[1,5,6*], Robi D. Mitra[1,2*] and Brian D. Muegge[2,7,8*]

[1]Department of Genetics, Washington University in St. Louis School of Medicine, Saint Louis, 63110, MO, USA.
[2]Edison Family Center for Genome Sciences and Systems Biology, Washington University in St. Louis School of Medicine, Saint Louis, 63110, MO, USA.
[3]Division of Biology and Biomedical Sciences, Washington University in St. Louis School of Medicine, Saint Louis, 63110, MO, USA.
[4]Department of Obstetrics and Gynecology, Washington University in St. Louis School of Medicine, Saint Louis, 63110, MO, USA.
[5]Department of Psychiatry, Washington University in St. Louis School of Medicine, Saint Louis, 63110, MO, USA.
[6]Intellectual and Developmental Disabilities Research Center, Washington University School of Medicine, Saint Louis, 63108, MO, USA.
[7]Department of Medicine,VA Medical Center, Saint Louis, 63106, MO, USA.
[8]Department of Medicine,Washington University in St. Louis School of Medicine, Saint Louis, 63110, MO, USA.

*Corresponding author(s). E-mail(s): jdougherty@wustl.edu; rmitra@wustl.edu; mueggeb@wustl.edu;
Contributing authors: g.juanru@wustl.edu;

Abstract

Spatial transcriptomics (ST) promises to reveal the principles of tissue organization and regional variation. With the advent of high-resolution capture-based ST methods such as Visium HD, new approaches are needed that faithfully assign transcripts to individual cells and enable the analysis of cells in complex tissue substructures. Here we introduce the Segmentation and Manifold UnRolling Framework (SMURF), a computational tool for the analysis of high-resolution ST data. SMURF introduces a novel "soft-segmentation" approach that accurately aggregates and assigns the mRNAs from capture spots to nearby cells. SMURF

is also able to "unroll" cells contained in complex tissue structures and place them on standard Cartesian coordinates, thus empowering studies of cellular organization and regional specification. We demonstrate the power of SMURF in the mouse small intestine. We identified the zonation of metabolic programs along the maturing intestinal villus and the transcription factors that regulate them. We used SMURF to investigate an important outstanding question – how are functionally important proximal-to-distal gene expression gradients in the intestine specified. We discovered that regionally specific gene programs along the proximal-distal axis accumulate in the villus tip and that the tip is reprogrammed by environmental signals in the lumen, suggesting these inputs are a major determinant of regional transcriptional identity. SMURF promises to accelerate the effective use of high-resolution ST data by enabling more accurate segmentation and assignment of mRNAs to cells and by providing a unique toolbox for the analysis of regional transcriptional programs along tissue manifolds.

Introduction

Biology is poised to address a fundamental challenge: describing the molecular features of all cell types in the body at single-cell resolution. However, a full understanding of how cells function can't be achieved without considering how they exist within their microenvironment. Spatial transcriptomics (ST) is transforming biological research by allowing researchers to measure cellular gene expression while preserving spatial context. Previously intractable questions — such as how the tissue microenvironment influences cellular identity, how gradients of cell types or gene expression are established and maintained — can now be addressed [1, 2, 3]. Capture-based ST has further advanced the field by measuring the entire cellular transcriptome using barcoded oligonucleotides that are arrayed in spots on the surface of a slide. Simultaneous collection of hematoxylin and eosin (H&E) staining or DAPI/Immunofluorescence (IF) images provides additional histological context to the gene expression data [4, 5]. Early capture methods only achieved resolutions of 10-100 μm, so that mRNAs from multiple cells were typically captured in a single spot (e.g. Visium, Slide-Seq [6]). But the recently developed Visium HD platform [7] offers a spatial resolution of 2 μm and contiguous coverage of the entire tissue section.

These advances bring new analytical challenges. First, there is no native relationship between the tiled oligonucleotides and overlying cells. Many of the spots aren't aligned with cells and thus contain no or very few transcripts. Second, transcripts from a single cell may be spread across multiple adjacent spots, or conversely, a single spot may contain transcripts from multiple adjacent cells. Existing high-resolution segmentation algorithms typically assume that each capture point contains mRNA from only a single cell [8, 9]. Furthermore, while capturing entire tissue sections should allow researchers to analyze

gene expression in tissue substructures, we lack tools to automatically identify and align these substructures in 2D tissue sections. Finally, substructures often deform during processing and sample preparation, obscuring their native spatial arrangement. Thus, there is a need for new computational approaches that can accurately segment and deconvolve mRNA in capture spots to single cells to enable the analysis of regional transcriptional programs along tissue substructures.

The brain and the small intestine are two examples of tissues with remarkable spatial and functional organization. The central nervous system (CNS) is structurally segregated into regions that control all aspects of cognition, sensation, and motor control. Despite well-described anatomic landmarks, we still lack basic understanding of the molecular gradients that establish these regions. Robust computational pipelines for analyzing spatial transcriptomic data are critical for accurately mapping these molecular gradients and understanding how gene expression patterns in their spatial context relate to brain function and dysfunction. Advanced computational tools that accurately map single-nucleus transcriptional data to individual cells, while preserving spatial relationships, could unlock the full potential of spatial transcriptomics in studies of brain organization, development, and related disorders [10].

The small intestine has a very different but no less complex spatial organization. The intestine is a long tube made up of millions of repeating crypt-villus units [11]. Each crypt-villus unit is a highly organized interface between epithelial, immune, stromal, neural, and muscle cells that collectively regulate nutrient absorption, motility, and host defense. Stem and proliferative Transit Amplifying (TA) cells in the epithelial crypt give rise to differentiated villus cells, including absorptive enterocytes, mucus-secreting goblet cells, and hormone-producing enteroendocrine cells (EEC). Small molecules from stromal cells and the luminal environment provide critical instructions that regulate the balance between proliferation and differentiation. Along the longitudinal axis, epithelial cells in the crypt-villus units are functionally specialized to carry out region-specific digestive, immunological, and hormonal actions. Despite intense research, defining the factors that shape the spatial organization of the intestine remains a challenging problem due to the complexity of the organ. Important studies using Laser Capture Microdissection (LCM) or early generations of ST technologies have revealed some of the dynamics along the crypt-to-villus axes, but these methods are still limited because they lack single-cell resolution [12, 13, 14, 15, 16]. Similarly, applications of ST to study the longitudinal specialization of the small intestine have been limited to broad multicellular aggregates [17]. A recent study used single-cell RNA-seq to study proximal-distal specialization in small intestinal enterocytes and stem cells, but couldn't directly capture spatial relationships along the crypt-villus axis [18].

New ST approaches have the potential to improve our understanding of how tissues are organized in development and disease, but we need new methods to accurately map spatial interactions between cells at single-cell resolution. To address these challenges, we developed the Segmentation and Manifold UnRolling Framework (SMURF), a method that extracts Visium HD gene

expression at single-cell resolution without relying on a reference dataset. We introduce the concept of "soft-segmentation", where mRNA counts within a capture spot may be assigned to a single cell or partitioned and assigned to multiple nearby cells in a fashion that optimizes the cosine similarity of each cell to the gene expression cluster to which it belongs. SMURF excels in maximizing the mRNA counts assigned to each cell, achieves precise cell clustering, and accurately determines cellular boundaries. We also introduce a framework based on manifold theory to unfold complex tissue structures in a two-dimensional histologic section onto "unrolled" projections representing native tissue locations. This enables the analysis of gene expression as a function of spatial location along two different axes on the manifold. Using data from the mouse brain cortex, we verified that SMURF accurately identifies major cell populations and stratifies gene expression to cortical layers. We next used SMURF to analyze the spatial regulation of the mouse small intestine. We used manifold analysis to identify crypt-to-villus zonation in multiple ileal cell types, and the transcription factor (TF) networks that regulate enterocytes at the base and tip of the villus. Finally, we used SMURF to investigate a fundamental question in the development of the small intestine: are proximal-to-distal gradients in gene expression predetermined, or are they established in response to signals in the luminal environment? We found that proximal-to-distal gradients in gene expression are almost exclusively present in cells of the epithelial villus tip and that tip cell gene expression is reprogrammed when a segment of the distal small intestine is surgically moved to a proximal location in an adult mouse, suggesting that signals in the lumen shape proximal-to-distal gradients in gene expression.

Results

SMURF introduces soft-segmentation for spatial transcriptomics

Visium HD collects ST data by capturing cellular mRNA released from a tissue section using barcoded 2 μm capture "spots" printed on a glass slide. Achieving single-cell resolution with this platform remains challenging because the spots are not aligned with cell borders - many spots will capture mRNA from two or more cells, and the mRNA content from a single cell can be distributed across multiple capture spots. This makes it difficult to reliably assign the mRNA transcripts recovered in a capture spot to the nucleus or nuclei from which they originated. SMURF addresses this limitation by using a novel "soft-segmentation" approach to assign transcripts to nuclei in three steps.

In the first step, SMURF starts from nuclei identified by the H&E or DAPI/IF image object using well-validated image segmentation methods [19, 20, 21, 22]. It then creates a single-nuclei gene expression matrix by assigning all the mRNA counts from capture spots that directly overlie each nucleus. Nuclei are clustered by expression similarity using Scanpy [23, 24, 25]. Simultaneously, SMURF creates a nearest neighbor network based on the spatial

locations of the nuclei, which in later steps facilitates the assignment of spots under cell bodies to nearby nuclei (Figure 1a).

Because nuclei represent only a fraction of total cellular volume, the single-nuclei gene expression matrix can be sparsely populated with low UMI counts[26]. To capture the full expression profile of each cell, the second step of SMURF soft-segmentation expands the nucleus-anchored gene expression by assigning additional mRNA transcripts to each nucleus via incremental concentric expansion. Cells are irregularly shaped and spaced from their neighbors; cytoplasmic spots near one nucleus might not come from the same cell. This raises the question: how to determine if spots adjacent to a nucleus arise from the same cell? Since we had already established the gene expression profile of the nucleus and its assignment to a cluster of similar cells, we reasoned that the counts in adjacent spots arising from the same cell should increase the similarity of that nucleus to its cluster identity. Conversely, adding counts from a different cell would decrease the similarity. We used this strategy to expand each nucleus to a "cell", where the cell is made up of the nucleus and counts from nearby spots that increased the cosine similarity of the cell with its gene expression cluster. We repeated the cell expansion process for every nucleus multiple times, until clustering results from the expanded cells' gene expression stabilized by normalized mutual information (NMI) score [27]. Because each spot can capture transcripts from multiple cells, SMURF does not restrict the number of times a given spot may be assigned to a nucleus. Consequently, a spot can be allocated entirely to a single nucleus, several nuclei, or none at all (Figure 1b).

The third and final step of SMURF's soft-segmentation is count deconvolution for spots shared by multiple cells. Many methods have been described for cell-type deconvolution of bulk or spot-based ST transcriptomics data to estimate the proportions of constituent cells in a mixture [28]. In SMURF, we extend this concept to separate individual mRNA counts in those capture spots assigned to multiple cells. We use a mixture of quadratic programming and deep learning methods to deconvolve mRNA counts between cells in a manner that maximizes the overall cosine similarity between the cells and the cluster to which they are assigned [29, 30] (Figure 1c).

SMURF improves single-cell reconstruction from Visium HD data

We sought to benchmark SMURF's ability to accurately soft-segment ST data from the mouse brain and distal small intestine (also known as ileum)[31, 32]. To compare SMURF's accuracy in assigning transcripts to cells against other commonly used methods, we examined how various algorithms assign capture spots to bins. Representative sections of the H&E staining with capture spot bins for the 8 μm method, the Bin2cell method [9], nuclei segmentation, and SMURF for the mouse brain and small intestine are shown (Figure 2a&b). The 8 μm method is the default assay output by 10x Genomics. Bins of 16 2 μm spots are aggregated to form 8×8 μm squares, increasing the number of UMIs per bin to increase statistical power for clustering and differential expression

analysis. This method defines bins without accounting for nuclei' locations. Thus, this method often misassigns transcripts to "cells", leading to bins that may lack nuclei, partially cover a nucleus, or encompass multiple nuclei (Figure 2a&b, 2nd column). Bin2cell is a recently described segmentation method designed to achieve single-cell resolution [33]. It performs significantly better than 8 μm bins in the mouse brain, but it frequently aggregates multiple cells into a single object or fails to assign nuclei to an object in the crowded small intestine (Figure 2a&b, 3rd column). The nuclei segmentation method achieves single-cell resolution by only considering capture spots that substantially overlap nuclei (Figure 2a&b, 4th column). However, this strategy ignores transcripts in capture spots under cell bodies, resulting in a sparser cell-by-transcript matrix. In contrast, SMURF reliably achieves single-cell resolution while still assigning transcripts from spots under cell bodies to cells. This demonstrates the superior performance of SMURF, especially in dense cell regions, compared to other methods (Figure 2a&b, last column).

We calculated the percentages of cells without an associated nucleus and the percentage containing multiple nuclei for each method (Figure 2c). SMURF achieved the best single-cell resolution of all the methods in brain and small intestinal data. Only 0.31% of SMURF cells in the mouse brain lacked a nucleus, 24-fold and 175-fold better than the Bin2Cell or 8 μm method, respectively. Similarly, only 1.44% of SMURF cells in the mouse brain contained multiple nuclei, 7.7-fold and 2.6-fold higher than the Bin2Cell or the 8 μm method, respectively. In the densely crowded small intestine data, SMURF similarly achieved superior rates of cells associated with a nucleus and cells associated with a single nucleus. These results demonstrate that SMURF more accurately assigns transcripts to cells at single-cell resolution than existing methods. We also found that cell-level normalization corrects the severe count imbalance observed in transcript-sparse regions, producing gene-expression profiles that align with *in situ* hybridization (ISH) references and therefore outperform the 8 µm binning approach (see Supplementary Material, Extended Data Figure 1a).

Since Bin2Cell and the 8 μm method do not reliably achieve single-cell resolution with Visium HD data, we next focused on comparing SMURF directly with the nuclei segmentation method. The total number of cells identified by SMURF was very similar to the number of nuclei detected in image segmentation. In contrast, nuclei segmentation and transcript binning without expansion, as implemented in a 10x workflow, retained many fewer features (Figure 2d). In the intestine, the 10x nuclei segmentation method retained only 54.37% of the nuclei identified in image segmentation. Variable nuclear shapes might explain part of this difference; spindle-shaped smooth muscle cell nuclei are less than the 2 μm width of capture spots and were not retained in the 10x nuclei segmentation. Another possibility is mRNA polarization of some transcripts in the cytoplasm, which has been described in intestinal cells [34, 35]. Together, these results demonstrate that SMURF can accurately assign Visium HD counts to single cells, with substantially more counts per cell than nuclei segmentation alone.

SMURF improves cell type identification and capture of rare cell types

We hypothesized that SMURF's higher count assignment to cells might result in more accurate cell-type identification and better estimates of their relative proportions. To test this, we used the cell types identified in a 10x Xenium study of mouse brain [36], an orthogonal probe-based ST method, as a reference. Through measuring Kullback–Leibler (KL) divergence, we quantified how well the cell-type proportions found in the Xenium data are represented by the proportions estimated by SMURF or nuclei segmentation without expansion [37]. SMURF has a lower KL divergence than nuclei segmentation (0.1 versus 0.738), demonstrating that SMURF more accurately measures cell type proportions compared to the reference data (Extended Data Figure 1b). Similar ground truth reference data for small intestine cell-type abundance is limited, so we clustered and assigned cell types for each method and compared those results. Both methods readily detected markers of enterocytes, the most abundant epithelial cell type in the small intestine (Figure 2e). However, SMURF clusters readily identified rarer epithelial cell types, including tuft cells and the heterogeneous EEC, along with smooth muscle cells from the muscle layer that encircles the intestine. In contrast, nuclei segmentation clusters did not capture gene expression for these cell types (Figure 2e). We suspect this is due to irregular nuclear shapes, low nuclear transcript abundance, or mRNA polarization of these cell types, but further study is needed.

Overall, our results demonstrate that SMURF is able to assign transcripts to cells at single-cell resolution, including transcripts from both the nucleus and cytoplasm of cells. SMURF identifies known cell types in Visium HD data with greater transcript counts for each cell than nuclei segmentation alone. Increased counts per cell increase statistical power for downstream analyses and the detection of rare or irregular cell types.

SMURF applies manifold detection to unroll ST data from complex tissue structures

After assigning transcripts to single cells, we next wanted to place the cells on a shared set of coordinates reflecting their *in vivo* locations. To do this, we developed a novel manifold detection approach to convert ST gene expression into a shared coordinate system to align cells. Manifold detection is a mathematical concept that assigns local Euclidean coordinates to small areas of a complex shape. In biology, signaling gradients or cellular structures often define anatomic tissue organization. Examples include cortical layers of neurons in the CNS and the muscularis mucosa that encircles the intestine. SMURF uses cell-resolved transcriptomes to identify the tissue structures and define a manifold. SMURF then "unrolls" the manifold, projecting the cells onto standard Cartesian axes. This functionality empowers studies of cellular organization and regional specification in the coordinate system most natural for that tissue.

We evaluated SMURF's ability to detect and unroll manifolds by analyzing the mouse brain cortex. We selected cells in the cluster corresponding

to the cortical basement membrane [38]. We used a locally linear embedding algorithm to identify a manifold path representing the one-dimensional projection of those cells (Figure 3a) [39]. Every cell can be assigned an "x-axis" coordinate representing its position on this projection. Using this curve as a reference, we expanded the manifold to two dimensions to capture all cells in the cortex. Each cell in the cortex was assigned an x and y axis coordinate based on its position relative to the basement membrane (Figure 3b&c). This procedure projected the cortex onto a standard rectilinear coordinate system, thus "unrolling" it.

To demonstrate how this algorithm can be used to analyze regional patterning of cell types and gene expression, we divided the unrolled y-axis of the cortex into 5 bins from the cortical base to the cortical surface. We plotted gene expression of known cortical layer markers across these bins as a heatmap (Figure 3d) [40]. Layer-specific marker genes aligned precisely with the SMURF-assigned location. Known markers of layer 2-4 neurons, including *Cux1* and *Pou3f2*, were highly expressed at the cortical surface. Similarly, markers of layer 5 and 6 neurons, including *Bcl11b* and *Foxp2*, were expressed at the base of the cortex (Figure 3d). Next, we assigned every neuronal cell in the ST cortex data a cell-type identity using the Allen Brain Cell Atlas as a reference [41, 42]. We unrolled the cortical neurons onto a plane using the calculated x and y coordinates and colored each cell by its assigned cortical layer. As expected, the cells clearly stratified along the calculated y-axis, with Layer 2 and 3 neurons at the cortical surface and Layer 6 neurons at the base (Figure 3e). We also observed increasing cortical thickness moving laterally on the x-axis, consistent with previous reports [43, 44]. We confirmed that non-neuronal cell types in the cortex were also projected according to their known anatomic distributions (Extended Data Figure 2) [45, 46]. Taken together, these results demonstrate that SMURF accurately embeds cells into a manifold and then "unrolls" this manifold onto rectilinear coordinates while preserving patterns of gene expression and cellular organization.

SMURF accurately unrolls complex and discontinuous tissue

Having demonstrated that SMURF can identify and unroll tissue structures in the mouse brain, we next turned to the mouse small intestine. Gene expression, cell type distributions, and function vary within the intestine along both the proximal-distal (longitudinal) and crypt-to-villus (radial) axes. We reasoned that SMURF's methods to identify gene expression in individual intestinal cells and then map those cells to a common proximal-to-distal and crypt-to-villus coordinate system could provide new insight into the molecular logic that controls longitudinal and radial development in the gut, along with the potential cell-cell interactions shaping those gradients.

We utilized Visium HD data from the healthy mouse ileum, the distal-most region of the small intestine [32]. The tissue was prepared in a "Swiss roll" that wraps the longitudinal tissue into a compact plane (Figure 3f). Swiss

roll identification and unrolling is a well-developed problem in manifold detection, and so we thought SMURF would be well suited to identify the original proximal-to-distal axis of the ileum as the manifold of interest. We first applied SMURF soft-segmentation to map gene expression in over forty thousand cells. UMAP clustering confirmed expected epithelial and mesenchymal cell populations (Extended Data Figure 3a). We then used SMURF's generalized unrolling tool to identify the manifold path of the smooth muscle layer at the intestinal base. We effectively unrolled the ileum for direct spatial analysis (Figures 3g). We confirmed that the SMURF x-axis projections accurately placed known representatives of proximal and distal regions of the intestine, including the proximally enriched lactose-degrading enzyme lactase (*Lct*) and the distally enriched Ileal Bile Acid Transporter (*Ibat*) (Figure 3h) [18]. Pseudobulked gene expression profiles of the epithelial ST data along the SMURF x-axis correlated with a recently described atlas of scRNA-seq along the length of the mouse small intestinal epithelium (Extended Data Figure 3b) [18]. Taken together, these results demonstrate that SMURF effectively unrolled the complex Swiss roll geometry onto its original longitudinal position.

We next sought to assign all cells coordinates along the crypt-to-villus axis. There are more than 1000 villi in the dataset which each contains multiple cell types arranged from the muscularis base through the stalk-like villus (Figure 3i). Several challenges emerged during analysis (Figure 3j). First, villi in the tissue image are not necessarily perpendicular to the smooth muscle layer, so that villus position is not simply an orthogonal projection from the base. In extreme cases, the tip of the villus is bent closer to the base than the rest of the villus. Another problem is that villi can be detached or severed during tissue sectioning. To address these challenges, we used an iterative approach that borrows concepts from dynamic programming (see Methods). We identified progressive layers of cells moving up from the base to the tip, and used the path of cells through these layers to identify a corrected y-axis for each villus (Extended Data Figure 3d, Extended Data Figure 4). The corrected y-axis reflects the orthogonal distance of each cell in the villus to the base and accounts for various tissue section artifacts, including bent or disconnected villi. Finally, the height of each villus was normalized to the tallest villi in the data and each cell was given a normalized corrected y-axis position (Figure 3k). Gene markers of the cyrpt base (*Olfm4*), top of the crypt (*Krt19*), villus (*Fabp6*), and villus tip (*Ada*) were distributed appropriately on the normalized y-axis (Extended Data Figure 3e).

We next compared the distribution of cells on our y-axis to known spatial variation of cell types along the crypt-to-villus axis (Figure 3l). As expected, neuronal and muscle cells are at the base of the axis in the muscularis propria. Epithelial stem and Paneth cells are thin, closely approximated cells in the crypt. They form a combined cluster in the ST data. The remaining epithelial cell types followed the expected distribution on the y-axis, with the proliferative TA cells emerging from the Stem cell zone and giving rise to the differentiated cell types of the villus. Interestingly, some EEC and tuft cells are detected at the base of the TA zone, suggesting that these cells either differentiate rapidly after leaving the stem cell compartment or escape the normal

conveyor movement out of the TA zone. Some studies have identified EEC that migrate towards the crypt, but this has not been addressed for tuft cells [47, 48].

Together, these results demonstrate that SMURF can accurately define manifolds along complex tissue structures and project cells onto rectilinear coordinates even in instances where the tissue has complex geometry. Our methods can be applied to subregions that are torn, bent, and/or missing to preserve the *in vivo* spatial arrangements of different cell types. These findings validate our approach and lay a solid foundation for more in-depth analyses.

SMURF identifies the spatial trajectory of intestinal development

The intestinal epithelium is an intriguing model in developmental biology, maintained through a precise balance of cell proliferation and differentiation along the crypt-villus axis (Figure 4a). At the base of the crypt, multipotent stem cells are supported by neighboring mesenchymal cells, which provide essential signals and growth factors to sustain self-renewal. Slowly dividing stem cells differentiate into TA cells, a rapidly dividing progenitor cell type that expands the cell pool necessary for continuous tissue renewal. TA cells exit the cell cycle as they migrate up the villus and differentiate into specialized epithelial cell types, ensuring efficient nutrient absorption and a robust barrier function. Notably, the entire intestinal epithelium is renewed approximately every 3 to 5 days [49].

We hypothesized that we could exploit SMURF's ability to project complex tissue structures onto rectilinear coordinates to quantitatively examine the relationship between the developmental state of intestinal epithelial cells and their spatial positioning along the crypt-villus axis. We focused on key epithelial subsets, including Paneth cells, stem cells, TA cells, and enterocytes. We determine the development trajectory with diffusion map, and visualize the cells by Partition-based graph abstraction (PAGA) and forced-directed graph drawing (Figure 4b) [50, 51, 52, 53, 54, 55, 56]. Each cell in the graph is colored according to its pseudotime progression along the inferred trajectory, with the most immature cells in the bottom-right corner and differentiated cells in the top-left corner. Next, we colored cells by their type and found that the pseudo-time projection mapped the known progression of Stem and Paneth cells to epithelial tips (Figure 4b, middle panel). Remarkably, the corrected spatial position of the y axis of the cells also followed the pseudo-time projection, exhibiting a strong correlation ($corr = 0.79$, $p < 10^{-10}$; Figure 4c). This demonstrates that SMURF can infer the patterns of intestinal epithelial development from the spatial position of cells in the crypt-villus axis.

Taken together, these findings quantitatively reinforce classical models of intestinal epithelial renewal, highlighting the strong spatial-temporal coupling of cell differentiation along the crypt-villus axis. The SMURF framework could be extended to study other tissues with similar continuous renewal processes.

SMURF defines novel ileal zonation gene programs

Having validated the developmental cascade along the crypt-villus axis, we next sought to determine the functional changes of specific cell populations during their maturation. Prior work studying villus zonation used LCM or Visium platforms. These methods inherently aggregate clumps of cells, including non-epithelial cell types, which might confound the results [12, 13, 14, 15]. We hypothesized that SMURF could determine villus zonation separately for each major epithelial cell type.

We subdivided epithelial cells from all fully intact villi into six zones evenly distributed along the crypt-villus axis (Figure 4d). Paneth and stem cells were predominantly located in the first zone at the base of the crypt, and TA cells in the second zone (see Extended Data Figure 5a). Enterocytes are the most abundant cell type in each villus zone, which also contain differentiated secretory cells (Extended Data Figure 5b). The single-cell resolution of SMURF allowed us to focus on enterocyte zonation, so we pseudobulked cells in the stem/Paneth, TA, and enterocyte clusters and analyzed each gene's expression across the zones, normalized to its own maximum expression. Gene expression was remarkably zonated (Figure 4e).

We reconstructed zonated marker profiles for all 17,563 expressed genes across six zones—encompassing stem/Paneth, TA, and enterocyte cells. In total, 1,631 genes were identified as zonated (see Methods, Supplementary table). We then defined zone-specific significant genes using manually selected cutoffs to account for differences in total counts between zones, excluding genes highly expressed in other zones or cell types (see Methods, Extended Data Figure 6).

KEGG pathway enrichment analysis revealed significant functional specialization in each zone (Figure 4f). Zone 1 was enriched for pathways involving known regulators of intestinal stem cells, including Wnt and TGF-β signaling [57, 58]. TA cells in Zone 2 were enriched in proliferative activity, including cell cycle, mismatch and nucleotide repair, and ribosomal pathways. In contrast to the crypt, much less is known about ileal villus zonation because it has been challenging to measure gene expression in single cells with prior methods. Our analysis demonstrated that enterocytes at the villus base in Zones 3 and 4 express programs for carbohydrate metabolism (including galactose and sucrose) and PPAR signaling. As enterocytes moved up the villus, they acquired functional enrichment for protein and vitamin metabolism in Zone 5, followed by lipid metabolism in tip cells of Zone 6 (Figure 4f). This progression is similar to prior work studying zonation in the jejunum, but the particular genes in each region are highly region-specific [12]. For instance, the brush border gene *Vil1* and the lipid transport gene *Apoa1* are broadly expressed throughout the villus. In contrast, *Anpep*, a gene that is involved in peptide digestion, and *Krt20*, a gene responsible for keratinization, are increasingly expressed from mid to upper villus regions (Figure 4g).

SMURF defines novel secretory cell zonation programs

Leveraging the power of SMURF to resolve spatial data to single cells, we next turned our attention to crypt-to-villus zonation in less abundant secretory epithelial cell types. Goblet cells, tuft cells, and the heterogeneous EEC are individually scattered throughout the villus and can't be isolated by LCM or conventional spatial analysis with large bin sizes. Using the same six zones, we pseudobulked the genes detected in each secretory cell type separately and identified genes with significant zonation along the crypt-to-villus axis.

Goblet cell-associated genes followed three distinct expression patterns along the crypt-to-villus axis (Figure 4h). *Spdef* and *Arg2*, both involved in early goblet cell differentiation, are mainly expressed in the crypt. In contrast, *Klf4* — a factor important for goblet cell maturation and function — shows increasing expression toward the villus tip. Lastly, a group of genes was evenly detected across all zones, including the mature goblet cell markers *Zg16* and *Tiff3*.

Tuft cell gene expression was also highly zonated (Figure 4i). The KIT receptor tyrosine kinase was recently identified as a cell surface marker for tuft cells [59]. *Kit* expression is much higher in Zone 2 than in other zones, suggesting its main function may be regulation of early tuft commitment. The essential tuft transcription factor *Pou2f3* was broadly expressed across all zones. Tuft cells are important chemosensors in the intestine, and we found that the sensory machinery was highly zonated across the crypt-villus axis. *Il17rb*, a tuft cell product that coordinates cytokine-mediated signaling with immune cells in the gut niche, was expressed most highly at the base of the villus [60]. The succinate receptor *Sucnr1* can induce small intestinal remodeling via tuft activation, and was sharply increased in the mid villus (zone 4) [61]. The taste receptor *Gnat3*, detected on tuft cells in multiple organs, is highly enriched at the villus tip [62].

Enteroendocrine cells collectively represent multiple subtypes, each expressing one or several hormones. Hormone distribution was highly zonated. The genes for *Tac1* (substance P) and *Gcg*, which give rise to gut hormones GLP-1 and GLP-1, are detected exclusively in the crypt regions. In contrast, cells expressing *Tph1*, the enzyme responsible for serotonin production, and *Nts* and *Pyy* are detected only in villus zones (Figure 4j). These distributions have been observed in prior studies using immunohistochemistry [63, 64]. Morphogen gradients in the villus, including bone morphogenetic proteins (BMPs), have been implicated in the emergence of *Nts* and *Pyy*. However, the factors that limit *Tac1* and *Gcg* expression in the crypt remain unknown. Spatial approaches to identify potential niche signals for these cells will be an important extension of this work.

Zonated transcription factor regulatory networks along the villus

Intestinal specification during embryonic development depends on many well-described TFs [65, 66]. In contrast, the transcriptional networks that control epithelial zonation in the mature villus are less understood. We used SCENIC

to identify gene regulatory networks of TFs and their targets in each zone, using expression from stem/Paneth, TA, and enterocyte cells [67]. Similar to the zonation of enterocytic functions from crypt to villus (Figure 4e), many TF regulons were zonated. We identified 65 TFs with a gene regulon score ≥ 0.05 in at least one of the six zones (Supplemental Table 3). Many TFs that are essential for intestinal function are zonated (Figure 5a). For instance, the *Sox9* regulon is highly specific to the crypt, consistent with that gene's role in stem and Paneth cell specification under the direction of Wnt signaling [68, 69]. Prior studies of villus zonation identified tip-enrichent of the immediate early transcription factors *Jun* and *Fos* in the jejunum, which we confirmed in the ileum [68, 12]. This analysis also suggested an unappreciated zonation of important TF activities. For instance, many well-described factors that regulate intestinal identity, including the Vitamin D receptor *Vdr* and the bile acid receptor *Nr1h4* (*Fxr*), and the peroxisome proliferator-activated receptor delta *Ppard* have the highest regulon scores at the base of the villus. This is intriguing as the luminal factors that activate these receptors would be expected to have higher concentrations near the villus tip.

Because SMURF places all cells in a shared coordinate system, we next asked if the TFs themselves follow the same zonation patterns as their gene targets. The majority of TFs were expressed most highly in the zones where their regulons were most active. For instance, the Vitamin D receptor (*Vdr*) and Peroxisome Proliferator-Activated Receptor Delta (*Ppard*) genes are highly expressed at the villus base, and their gene regulons are sharply enriched in Zones 2 and 3 (Figure 5a&b). There are examples of TFs whose expression distribution does not match their maximal regulon activity. The distal intestinal regulator *Gata6* is required for ileal identity [70]. The TF is expressed highly in crypt and TA zones (Zone 1 and 2), but its regulon scores were highest in the villus and tip (Zone 5 and 6).

Cooperative binding of TFs at regulated genes is a well-recognized mechanism to enhance expression of target regulons. Thus, we sought to identify TF networks that cooperatively regulate zonal identity in key regions. To get an overall view of interactions within each zone, we ploted the TF interaction network for all differential regulons (Extended Data Figure 7).

While many TF's are active in Zone 2, a network of 7 TF's was strikingly interconnected (Extended Data Figure 7, Figure 5c). This network is anchored by interactions with the TF *Elf3*, a well-recognized epithelial TF which mediates Wnt-independent B-catenin transactivation in colon cancer [71, 72]. The network also includes *Klf5*, an essential regulator of intestinal barrier integrity and stemness [73, 74]. Other members of this network, including *Ehf*, *Nono*, *Bclaf1*, and *Vezf1*, have not been as well studied in the intestine. Interestingly, while the TF's together have thousands of potential transcriptional targets, genes with enriched expression in Zone 2 were significantly over-represented in the collection of genes targeted by more than 1 TF (p-value $= 8 \times 10^{-7}$, see Methods, Figure 5c). Functional analysis of the genes targeted by two or more TFs demonstrated that they were involved in RNA biogenesis and Mitochondrial metabolism, core functions of TA cells. Thus, these results identified a

novel cooperative TF network that controls the proliferative functions of TA cells.

We also identified a TF network at the villus tip. Consistent with prior studies, the villus tip network included the AP-1 transcription factor complex members *Jun* and *Fos* whose expression and regulon activities were enriched in Zone 6 [14, 12, 75, 76]. We also identified several novel TFs enriched at the tip, including *Irf7*, *Mxd1*, and *Klf4*. *Irf7* is a virally induced factor that controls the expression of Type 1 and 2 interferons. *Klf4* is known to be a tumor suppressor and regulator of reserve stem cells in the intestine, but it has not been previously linked to a specific role in the villus tip. *Mxd1* antagonizes Myc-driven programs to suppress proliferation and has not been well studied in the intestine. Similar to Zone 2, genes that are significantly zonated in the tip were over-represented in the collection of genes targeted by more than one TF in this network (p-value $< 10^{-10}$, see Methods, Figure 5d). These genes are involved in amino acid, lipid, and vitamin metabolism, as well as host defense. This suggests cooperative actions of the broad AP-1 complex and environmental sensors like *Irf7* to amplify the tip-specific functional program.

Regionally specific gene programs accumulate in the epithelial villus tip

The intestine is functionally specialized along the proximal-to-distal axis [18, 77, 78]. Developmental gradients establish the main axis of proximal-to-distal specialization, but less is understood about the mechanisms that maintain regionality in the adult intestine. To identify the regulators of positional identity on the proximal-to-distal axis, we separated villi in the proximal and distal ileum based on their x-axis coordinate, and performed differential expression analysis in pseudobulked counts from each zone of those villi. Surprisingly, the number of differentially expressed genes was much higher in the upper zones of each villus compared to the base (Figure 6a). For instance, the apolipoprotein genes *Apoa4* and *Apoa3* are significantly enriched in the tips of proximal villi (Figure 6b). While the primary site of fat absorption is the jejunum, the proximal ileum plays an important role in absorbing overflow fats, as demonstrated ultrastructurally after high fat diet [79]. In the distal ileum, the beta-alanine carrier *Slc6a14* is highly expressed in Zones 4, 5, and 6 (Figure 6c) [80]. Functional analysis of the regionally enriched genes showed that proximal villi tips express high levels of programs for metabolism, including cholesterol and pyruvate, while distal tips are enriched in genes for bile secretion and steroid biosynthesis (Figure 6d).

By contrast, less than 0.1% of genes in Zone 1 and less than 1% of genes in Zone 2 were differentially expressed in proximal and distal crypt-villus units (Figure 6d). This is surprising, as the intestinal crypt harbors the stem cells that give rise to all differentiated cell types. Intestinal stem cells intrinsically encode some aspects of regional identity, as demonstrated by organoid cultures grown in the absence of regionally specific niche factors [81, 82]. However, our result is consistent with a recent report using scRNA-seq across multiple intestinal segments that identified 5 metabolic zones of enterocyte identity but

only 3 distinct stem cell patterns [18]. The ST data is limited to approximately 9 cm of the distal intestine, and we wanted to see if the same results extended across greater regions of the intestine. We plotted markers of villus bottom and top zonation identified in the mouse jejunal crypt-villus unit in the ileal ST data (Extended Figure 8a&b) [12]. Interestingly, the genes identified as bottom markers in the jejunum were also expressed predominantly in Zones 2 and 3 of the ileum. In contrast, the jejunal top markers were detected across all zones of the ileum (Extended Data Figure 8)

This novel finding suggests that proximal-to-distal differences in the intestine are not broadly distributed through the crypt-villus axis but instead segregated at the villus tip. Intestinal stem cells at the crypt base give rise to all the differentiated epithelial cells of the intestine. Since there are few transcriptional differences in the stem cell zone in the proximal and distal ileum, these results suggest that intestinal regional identity along the proximal-to-distal axis can't simply result from differences encoded in regional stem cells. If the genetic landscape of stem cell programs can't fully explain the specialization of enterocytes, what else might govern regional identity? We hypothesized that proximal-to-distal luminal gradients of environmental signals, including dietary components, xenobiotics, and microbes, could interact with the stem cell encoded genetic potential to pattern regional identity.

To test our hypothesis, we sought data where proximal luminal contents are presented to distal intestinal cells. Our hypothesis predicts that luminal contents would induce proximal programs segregated at the tips of distal villi. We turned to a recent report of scRNA-sequencing from an animal model of small bowel resection (SBR), in which 50% of the small intestine is surgically removed. The remnant proximal intestine is connected to the remnant ileal tissue, and thus proximal lumenal contents are presented directly to distal intestinal regions (Figure 6e) [83]. The authors performed scRNA-seq on the remnant ileal tissues seven days after resection, and compared gene expression with ileal tissue from sham-operated animals. We identified the genes differentially expressed between mice subjected to resection and controls, and mapped the significantly enriched genes to the corresponding zones in which it is predominantly expressed in the healthy ST ileal data. Consistent with the authors' report, the genes increased after surgery were enriched in lipid processing functions more typical of proximal intestinal regions. The majority of the induced genes are concentrated in the villus region (empirical chi-square test, $p = 9.99 \times 10^{-6}$), often at or near the villus tip (Figure 6f). Notably, none of these genes showed expression in either stem or TA cells, indicating that the observed changes are downstream of regional reprogramming of the stem cells. This result supports our hypothesis that the final specialization of regional identity in the small intestine depends on environmental signals and not just stem cell-encoded potential. The surgical resection model induces a complex adaptive response, and we can't completely isolate other factors, including increased epithelial proliferation and altered hormonal signaling after surgery, that might be important in this model.

Discussion

The recent introduction of Visium HD, a capture-based spatial transcriptomic method with subcellular resolution and high capture efficiency, has sparked new urgency to develop algorithms that accurately assign transcripts to individual cells to fully uncover novel biological insights. Here, we report the development of SMURF, which utilizes a novel "soft-segmentation" algorithm to accurately assign mRNAs from each Visium HD capture spot to individual nuclei-anchored cells. SMURF also "unrolls" tissue substructures, placing cells on Cartesian coordinates representing *in vivo* spatial arrangements. We benchmarked SMURF and demonstrated that it outperforms other analysis methods in achieving single-cell resolution and accurately identifying known cell types. SMURF accurately unrolls tissue substructures even when they are damaged or partially missing.

Prior work to describe spatial organization in the intestine has revealed many important principles, but has been limited by large capture spots or labor-intensive methods that prevent single-cell resolution [12, 14, 17, 13, 16]. We utilized SMURF to analyze the mouse small intestine and identified cell-type specific zonated gene expression programs and novel transcription factor regulatory networks along the crypt-villus axis, many of which were not previously identified. Finally, we used SMURF to investigate a fundamental question about small intestine development: are proximal-to-distal gradients in gene expression predetermined or established in response to environmental signals? We found that proximal-distal variation in gene expression is present almost exclusively in tip cells and that this gene expression can be reprogrammed by moving a segment of the distal small intestine to a proximal location, demonstrating that proximal-to-distal gene expression variation is crucially dependent on environmental signals in adult tissue.

Although effective, SMURF also has some limitations. First, SMURF uses H&E or DAPI/IF stained images to detect nuclei, so it cannot be used to analyze datasets for which tissue images are not available. Second, for capturing spots near cell boundaries, SMURF partitions captured mRNA between nearby cells based on relative nuclei distance and the cells' expression similarity with their respective clusters. While SMURF outperforms previous methods in this regard, it can't always resolve cells that share a high degree of expression similarity and are closely adjacent, such as the stem and Paneth cells in the intestinal crypt. Finally, ST datasets, including associated histology images, can be extremely large. Processing these using SMURF is computationally demanding. SMURF's full implementation of employs a deep-learning algorithm built on PyTorch for deconvolution, which requires GPUs to run efficiently [84]. To make SMURF more widely accessible, we have also developed a "lite" version of SMURF which can run on most workstations. SMURF-lite is benchmarked and described in detail in the Supplementary Methods.

As spatial transcriptomics continues to evolve, new platforms will utilize even higher resolution capture spots and profile larger numbers of cells. The full potential of these improved technologies will only be achieved with approaches like SMURF's soft segmentation algorithm to accurately assign

capture spots to the appropriate cell. SMURF's framework is inherently extensible to emerging technologies for three-dimensional (3D) tissue ST. Finally, SMURF's unrolling algorithm could be extended to other spatial technologies, such as protein-based assays or image-based tools, providing researchers with a more comprehensive view of tissue architecture.

Our study demonstrates the value of SMURF to reveal principles of tissue organization in normal development. An important future extension will be to deploy this approach to understand disease models at single-cell resolution. Many intestinal diseases, including adaptation to massive SBR and epithelial injury in inflammatory bowel disease, result in dysregulation of the normal crypt-villus architecture. SMURF will enable researchers to study how multiple cell types mechanistically interact in these studies, including "4D" studies of evolving responses over time. We demonstrated SMURF's utility in multiple other tissues (see Supplementary) and anticipate that SMURF will provide insight into other spatially complicated tissues such as the kidney, liver, and lung.

We anticipate that SMURF will substantively enhance the analysis of spatial transcriptomics datasets, allowing better identification and characterization of rare cell types, the identification of novel cell types and gene expression program zonations, and provide new insights into tissue homeostasis. Its innovative soft-segmentation and unrolling strategies set the stage for integrative, multi-dimensional analyses to better understand tissue organization, developmental processes, and disease mechanisms.

Methods

Soft Segmentation Analysis

Introduction

In this section, we provide a detailed description of the soft-segmentation method implemented in SMURF. Cells are denoted by the index i, genes by j, and spots by k. Since each cell is associated with exactly one nucleus, we denote both a cell and its corresponding nucleus by i, emphasizing their one-to-one correspondence. Starting from scratch, the inputs consist of: (1) the $2\,\mu$m spot gene expression matrix S; (2) the H&E or DAPI/IF tissue image Img^o; and (3) the "spot to pixel" data frame.

We define the $2\,\mu$m spot gene expression matrix S to have dimensions $K_{\text{spots}} \times J_{\text{genes}}$, where K_{spots} represents the number of spots and J_{genes} the number of genes, with each element $S_{kj} \geq 0$. Each pixel in the original image Img^o represents a grayscale value (0–255) for DAPI/IF images or a 1×3 RGB array for H&E stained images. The matrix S is linked to the processed image Img via the "spot to pixel" data frame. Note that some pixels in Img^o are not covered by spots, and each spot encompasses multiple pixels; hence, we focus on the area covered by spots, defined as Img. We denote by Img_k the set of pixels corresponding to spot k.

SMURF requires users to generate their own nuclei segmentation results. To facilitate this, we provide tutorials for DAPI/IF images (using Cellpose

[19, 85]) and H&E stained images (using StarDist [20, 21, 22]); users may also choose their preferred segmentation tools. The final nuclei segmentation output, denoted as $Nuclei_img$, is a two-dimensional matrix with the same dimensions as Img. In $Nuclei_img$, regions without nuclei are assigned a value of 0, whereas regions containing nuclei receive unique numerical identifiers.

Given the substantial volume of data and resulting outputs, SMURF defines a new Python object, the Spatial Object (SO), initially constructed from the three inputs. The SO records the nuclei segmentation results, detailed information for each cell, and visualization outputs. For clarity, we refer to the components of SO by their object names rather than describing each individual element of SO here.

The soft-segmentation in SMURF involves three main steps: nuclei analysis, cell expansion, and count deconvolution for shared spots. For the first step, for each nucleus, a nucleus-by-gene matrix is constructed and clustered to provide initial cell type annotations, and a nearest neighbor network is constructed to estimate approximate cell size; second, cells are expanded by iteratively assigning adjacent spots based on their similarity to the cell's cluster identity, with periodic reclustering to refine assignments; finally, overlapping spots that are shared by multiple cells are resolved through optimization methods (quadratic programming and deep learning), ensuring accurate gene expression profiles for each cell. Finally, SMURF outputs a spatial single-cell `AnnData` object containing the gene expression matrix G ($I_{\text{cells}} \times J_{\text{genes}}$). Each cell is annotated with its gene expression profile, spatial coordinates, preliminary cluster assignment, and the cosine similarity to its assigned cluster.

Step one: Nuclei Analysis

The nuclei analysis includes mainly three parts: building the nuclei gene expression array, clustering nuclei, and estimating the potential sizes of cells based on location. It takes in the 2μm spot gene expression matrix S and nuclei segmentation results $Nuclei_img$ from the image.

Before the analysis, we first filter the $2\,\mu$m spot gene expression matrix S based on the criterion

$$\sum_{k=1}^{n_{\text{spots}}} S_{kj} \geq \text{mincounts},$$

with the default threshold set at mincounts $= 1000$. We denote the resulting filtered matrix as S'.

The `Nuclei_img` contains pixel-level data, and each $2\,\mu$m spot corresponds to approximately 50 pixels. Consequently, many nuclei span multiple spatial spots and many spots are shared by multiple nuclei. We define $pct_{ik} =$ (number of pixels in nucleus i of spot k)$/Img_k$, with $0 \leq pct_{ik} \leq 1$, to denote the percentage of pixels in spot k that overlap the nucleus of cell i. For each nucleus i, we record all spot IDs for which the fraction of overlapping pixels is at least a specified minimum (the default fraction is $1/6$). The transcripts captured in all spots that meet this criterion for nucleus i are assigned to that nucleus. Using this initial assignment of transcripts to cells, we create the gene

count matrix $X^{(0)}$, defined as:

$$X_{ij}^{(0)} = \sum_{k=1}^{n_{\text{spots}}} \mathbf{1}(pct_{ik} \geq 1/6) \times S'_{kj}$$

where $X^{(0)}$ is an $I_{\text{cells}} \times J_{\text{genes_filtered}}$ matrix and $X_{ij}^{(0)}$ indicates the count expression of gene j in nucleus i.

In the next part, we filter the nuclei according to the total count of nuclei, since extremely low counts of nuclei could be due to errors like waterdrops on the capture surface and influence the accuracy of further analysis. We then analyze matrix $X^{(0)}$ following the Scanpy [23, 24] tutorial, including filtration of genes by a minimum of cells expressed, normalization, dimensionality reduction via principal component analysis (PCA), and clustering by the Leiden algorithm. Through this approach, we obtain initial clustering results $C^{(0)}$ based on the $X^{(0)}$, where $C^{(0)}$ is a $1 * n_{cells}$ array that assigns each nucleus into one of $n^{(0)}$ clusters based on gene expression profile.

Finally, we estimate the potential number of spots belonging to a cell based on its cell density. We calculate the center of each nucleus by averaging the locations of the identified spots associated with that nucleus. We employ the k-nearest neighbors (KNN) algorithm [86], linking each nucleus to its three closest neighbors. For each nucleus i, we record $size_i$ as the number of spots occupied by the nucleus and the distances d_{io} (for $o = 1, 2, 3$) from the center of nucleus i to the centers of its three nearest neighbors. Assuming that the nuclei are approximately circular, we calculate d'_i as:

$$d'_i = \frac{1}{3} \sum_{o=1}^{3} d_{io} - 2\sqrt{\frac{size_i}{\pi}},$$

which represents the average distance between nearby cells minus twice the estimated radius (calculated from the cell size), thus approximating the maximum extent of the spots.

Step two: Cell Expansion

During the nuclei expansion phase, we use an alternating optimization strategy to expand the nuclei, thereby generating robust cell gene profiles and reliable clustering results. In essence, we iteratively assign adjacent spots to cells based on their similarity to the cell's cluster identity, with periodic reclustering to refine the assignments.

Here, we use the iteration variable r (starting from $r = 1$) to record each iteration and define the initial nuclei results as the starting point ($r = 0$). For example, as previously introduced, $X^{(0)}$ represents the nuclei-by-gene expression matrix, $C^{(0)}$ denotes the nuclei clustering results.

In the beginning of each iteration r, we calculate the average expression of each cluster, denoted by $w^{(r-1)}$, which has dimensions $n^{(r-1)} \times n_{\text{genes_filtered}}$.

For each cluster $c = 1, 2, \ldots, n_c^{(r-1)}$ and for each gene j, the average expression is defined as

$$w_{cj}^{(r-1)} = \frac{1}{\|C_c^{(r-1)}\|} \sum_{i \in C_c^{(r-1)}} X_{ij}^{(r-1)},$$

where $C_c^{(r-1)}$ denotes the set of cells belonging to cluster c in $C^{(r-1)}$ and $\|C_c^{(r-1)}\|$ represents the number of cells in this set. Then, we normalize the gene expression by

$$\hat{w}_c^{(r-1)} = \frac{w_c^{(r-1)}}{\|w_c^{(r-1)}\|}.$$

Additionally, we calculate

$$\bar{d}_c^{(r-1)} = \frac{1}{\|C_c^{(r-1)}\|} \sum_{i \in C_c^{(r-1)}} d'_i,$$

to decide the maximum expansion of spots for each nuclei. By default, we use

$$\tilde{d}_c^{(r-1)} = \max\Big(1, \min\Big(2, \lceil \bar{d}_c^{(r-1)} \rceil + 1\Big)\Big).$$

In the second stage, our goal is to derive the matrix $X^{(r)}$ $(n_{\text{cells}} \times n_{\text{genes_filtered}})$ so that cells within the same cluster exhibit high cosine similarity. We denote the gene expression vector for cell i in the current iteration as $X_i^{(r)}$ and let $cell_i^r$ represent the spots assigned to cell i that were not previously associated with nucleus i in iteration r . The objective function is defined as follows:

$$\begin{aligned}\Phi_1(X^{(r)}) &= \sum_{c=1}^{n_c^{(r-1)}} \sum_{i \in C_c^{(r-1)}} \cos_{\text{sim}}(X_i^{(r)}, \hat{w}_c^{(r-1)}) \\ &= \sum_{c=1}^{n_c^{(r-1)}} \sum_{i \in C_c^{(r-1)}} \cos_{\text{sim}} \left(X_i^{(0)} + \sum_{k \in \text{cell}_i^r} S'_{kj}, \hat{w}_c^{(r-1)} \right) \end{aligned} \tag{1}$$

Here, our objective is to determine $cell_i^r$ for $i = 1, 2, \ldots, n_{cells}$. For each spot k included in cell_i^r, the distance from the spot to the nucleus must be less than $\tilde{d}_c^{(r-1)}$, and the spots within a cell must form a connected region.

To achieve this within limited memory resources, we employ a greedy algorithm that expands the spots one-by-one in a specified order. A spot is added if its inclusion improves the objective function (1), and expansion in a particular direction ceases if the previously added spot is excluded.

In the next phase of iteration r, we analyze $X^{(r)}$ following the Scanpy tutorial as described earlier, thereby obtaining the clustering information $C^{(r)}$. To compare the clustering outcomes between iterations r and $r-1$, we calculate the normalized mutual information score (NMI) using the formula:

$$\mathrm{NMI}(C^{(r)}, C^{(r-1)}) = \frac{2 \times I(C^{(r)};\ C^{(r-1)})}{H(C^{(r)}) + H(C^{(r-1)})}$$

where $I(C^{(r)};\ C^{(r-1)})$ denotes the mutual information between the clustering iterations, defined as:

$$I(C^{(r)};\ C^{(r-1)}) = \sum_{u \in C^{(r)}} \sum_{v \in C^{(r-1)}} P(u,v) \log \left(\frac{P(u,v)}{P(u)P(v)} \right),$$

$H(C^{(r)})$ and $H(C^{(r-1)})$ represent the entropies of the respective clustering iterations, defined as:

$$H(C^{(r)}) = -\sum_{u \in C^{(r)}} P(u) \log P(u).$$

This calculation quantifies the consistency of clustering results between iterations r and $r-1$, with higher NMI values indicating greater similarity. Typically, the NMI increases as the iterations progress, and by default, the process terminates when there is no improvement in the NMI between two consecutive iterations.

Step three: Count Deconvolution for Shared Spots

The previous step yields cell profiles for each spot, though a spot may be shared by multiple cells. We denote these final results as $r = *$. At this stage, we isolate spots that are exclusively associated with a single cell and sum their gene counts to obtain X^{final_0}. Next, we focus on the ambiguous spots, which can be categorized into two types: those shared by cells from different clusters, and those shared by cells within the same cluster. We first address the former by transforming situations where multiple clusters share spots into scenarios resembling the latter case. Finally, we apply deep learning optimization to the objective function (Equation 1) to determine the proportion of each cell cluster present within the shared spots.

For the first scenario, we simplify the problem to a case where cells from different clusters, denoted by the set u_k, share a single spot k, with each cluster's contribution represented by the weight matrix $\hat{w}_c^*$. Due to the resolution of Visium HD, u_k typically contains 2 or 3 clusters. We then employ sequential least-squares programming and define the equation as:

$$\sum_{c \in u_k} \mathrm{pct}_{ck} \cdot \hat{w}_c^* = S'_k / (\sum_{j}^{n_{genes}} S'_{kj}),$$

where $\sum_{c \in u_k} \mathrm{pct}_{ck} = 1$ ensures that the proportions sum up to unity and S'_k represents the gene expression array for spot k. After obtaining the values of pct_{ck}, we employ a binomial model to determine the precise count via random number generation. We then define X^{final_1} as $\sum_{k \in cell_i} \bigcap_{(\text{i is only cell of c in k })} \bigcap_{(c \in u_k)} pct_{ck} * S'_k$.

Here, each specific cell i contains spots with predetermined area percentages. However, there remains a subset of spots v_k that are shared by at least two cells within the same cluster. We refine the objective function (2) to incorporate both the already assigned spots and those yet to be assigned, as follows:

$$\Phi_2(X) = \sum_{c}^{n_c^*} \sum_{i \in C_c^*} \cos_{\text{sim}} \left(X_i^{final_0} + X_i^{final_1} + \sum_{k \in v_k} \text{pct}_{ik} \cdot S'_k, \hat{w}_c^* \right) \tag{2}$$

pct_{ik} satisfies the condition:

$$\sum_{i \in (\text{cells partially in spot k})} \text{pct}_{ik} = 1.$$

The maximization of Equation (2) is achieved through gradient-based optimization using the PyTorch library. Subsequently, we employ a binomial model to precisely determine counts via random number generation. Moreover, since each cell comprises a spot and its associated proportion, we revert to the unfiltered S to recover the gene expression count matrix for all genes.

Unrolling Analysis

Visium HD offers high-resolution ST data, and SMURF overcomes the limitation of single-cell resolution by employing a soft-segmentation approach. Many tissues possess intricate biological structures that are challenging to study, and during pathological processing, some tissues can become deformed on slides, obscuring their native spatial organization. With SMURF, we can now identify distinct clusters of surface cells. Viewing these clusters as a manifold, we propose a framework to unroll and elucidate the underlying biological architecture based on the location of cells.

We first select a specific cluster from the soft-segmentation output to serve as the layer to unroll. Cells that are isolated or located at the periphery are filtered out if their average distance to the nearest neighboring cells exceeds a specified threshold. The filtering is performed after augmenting the set of cell coordinates by adding additional points—specifically, by inserting the midpoints between existing cells. Next, we apply Locally Linear Embedding (LLE) to preserve local neighborhood structures, yielding a new x-axis represented by the matrix L (of dimensions $n_l \times 2$), where n_l denotes the number of selected cells. The LLE algorithm comprises several key steps: neighborhood selection, weight matrix construction, and low-dimensional embedding calculation. For each data point l_i in L, its local neighborhood is identified using the k-nearest neighbors method:

$$N(i) = \{l_{j_1}, l_{j_2}, \ldots, l_{j_k}\}$$

where $N(i)$ denotes the neighborhood of the data point l_i. For each data point l_i, we compute weights p_{ij} that best approximate l_i using its neighbors $N(i)$,

with the objective to minimize the reconstruction error:

$$E(P) = \sum_{i=1}^{n} \left\| l_i - \sum_{j \in N(i)} p_{ij} l_j \right\|^2$$

subject to the constraint $\sum_{j \in N(i)} p_{ij} = 1$, which guarantees that the weights sum to one and maintain local linearity. Once the weights p_{ij} are determined, the algorithm seeks a low-dimensional embedding q_i for each l_i that preserves the local geometry defined by these weights. The embedding is then optimized by minimizing the following cost function:

$$\Phi'(Q) = \sum_{i=1}^{n} \left\| q_i - \sum_{j \in N(i)} p_{ij} q_j \right\|^2$$

After obtaining Q, we reorder L according to the magnitude of q_i for each element in L. To present a smooth result with consistent distances between units on the x-axis, we first smooth the points L' $((n_l - n_{avg} + 1) \times 2)$ by averaging the closest n_{avg} points:

$$l'_i = \frac{\sum_{j=i}^{i+n_{avg}-1} l_i}{n_{avg}}, \quad i = 1, \ldots, (n_l - n_{avg} + 1)$$

Next, we use interpolation functions to create an x-axis such that the actual distance between each unit point is consistent. We first calculate the Euclidean distances d_i between two nearby units in L'. The cumulative distance D_i is the total distance from the first point to the i-th point. Specifically, $D_0 = 0$, and for $i = 1, 2, \ldots, (n_l - n_{avg})$, it is given by:

$$D_i = D_{i-1} + d_{i-1}$$

The total distance $D_{n_l - n_{avg}}$ represents the distance from the starting point to the last point. A cubic interpolation function f is established for the coordinates of L'. We define a new sequence of distances containing $n_l - n_{avg}$ points evenly distributed points D'_j, where:

$$D'_j = j \cdot \frac{D_{n_l - n_{avg}}}{n_l - n_{avg}}, \quad \text{for} \quad j = 0, 1, 2, \ldots, n_l - n_{avg}$$

The new coordinates $f(D'_j)$ are computed using the interpolation functions for each D'_j, forming new points ℓ''_j. Thus, we create evenly and smoothly distributed points for the x-axis.

Depending on the tissue type, the y-axis is constructed either based on absolute distances or on cell connectivity. In the first case, for each cell, we identify its nearest unit on the x-axis and calculate the Euclidean distance

from the cell to that unit, assigning this distance as the y-coordinate. In the second case, cells that are in close proximity are considered connected. For each cell, we determine the shortest connected path to the x-axis, define the y-layer as the number of cells along that path, and compute the y-coordinate by summing the distances along this path.

However, given the limited computational resources and the vast number of cells, calculating the shortest path for each cell individually is time-consuming. Moreover, the available paths are normally restricted to a limited set of directions. To overcome these challenges, we progressively generate y-layers by iteration. We begin by constructing the first layer using an iterative approach. For each unit on the x-axis, we identify the 18 closest cells within a specified distance and label them as Layer 1. Then, for each cell in Layer 1, we find the nearest neighboring cells among the remaining cells within the limited distance. To ensure that the cells are captured in the correct direction, we use a central reference point and require that the direction from the center to the original cell is consistent with the direction from the center to the Layer 1 cell. If a cell is selected multiple times by different Layer 1 cells, we retain the path with the shortest total length and record that path as the shortest one, assigning the corresponding x-axis coordinate to that cell. This process continues by iterations: in each iteration, the next layer is generated following the same procedure until no qualified cells are left.

Mouse Brain data Analysis from 10x Genomics[31]

Nuclei Segmentation

SMURF requires nuclei segmentation results as input. In this study, we utilized Squidpy (v1.4.1) [87] with Stardist [22] (v0.8.5) to perform nuclei segmentation on mouse brain H&E data. We first used SMURF to get the exact spot-occupied area of the H&E plot. Since a small portion of the H&E staining in this data was damaged and exhibited abnormally dark red regions, we converted these areas to a pink hue similar to that of the blank regions. Subsequently, we divided the entire image into 20 blocks with a small overlapping area between adjacent sections. For each block, the "2D_versatile_he" model from Stardist was applied, and the segmentation results were then combined. To account for split cells at the block boundaries, we ensured that the overlapping regions were preserved and only retained cells that were completely contained within a block. These refined nuclei segmentation results were used as input for the SMURF analysis of the mouse brain data.

SMURF Analysis

We input the nuclei segmentation data along with the original output from 10x Genomics and followed the SMURF protocol to generate the final gene expression matrix along with primary clustering results. For the unwrapping of the mouse cortex, we first selected the cluster corresponding to the basement membrane of the cortex, then filtered out isolated cells within this cluster and used the filtered cells to build the x-axis. We used the Euclidean distance to

define the y-axis and selected a center point to determine the orientation. Next, we constructed both the x-axis and y-axis based on the spatial coordinates of the cells. The y-axis was evenly divided into five sections, and we subsequently plotted a heatmap of gene expression markers that delineated the distinct layers of the cortex [40, 43].

Cell type annotation

We uploaded the complete gene expression dataset directly from the SMURF output to MapMyCells (RRID: SCR_024672) to align cells with the 10X Whole Mouse Brain (CCN20230722) reference [42]. We then selected layer-specific neurons with subclass bootstrapping probabilities greater than 0.90 and plotted these cells along the new axes, with colors indicating layer information. Similarly, we filtered for specific non-neuronal cell types with subclass bootstrapping probabilities exceeding 0.95 and visualized them on the same coordinate system.

Mouse Small Intestine Data Analysis from 10x Genomics [32]

Nuclei Segmentation

SMURF requires nuclei segmentation results as input. In this study, we utilized Squidpy (v1.4.1) [87] with Stardist (v0.8.5) [22] to perform nuclei segmentation on mouse small intestine H&E data. We first used SMURF to determine the exact spot-occupied areas of the H&E image. The entire image was then divided into 36 blocks with a small overlapping area between adjacent sections. For each block, the "2D_versatile_he" model from Stardist was applied, and the segmentation results were subsequently combined. To account for split cells at block boundaries, we preserved the overlapping regions and retained only those cells that were completely contained within a block. These refined nuclei segmentation results were used as input for the SMURF analysis of the small intestine data.

SMURF analysis

We followed the SMURF protocol to obtain the final gene expression matrix along with primary clustering results. For the unrolling analysis, we manually removed regions where the small intestine was inadvertently cut or damaged at both the beginning and the end. We then selected the clusters corresponding to the muscularis externa and myenteric plexus to define the x-axis, and filtered out isolated cells within these clusters. Finally, we adhered to the SMURF pipeline to define the y-axis layer by layer.

Single-Cell Analysis

We followed the Scanpy (v. 1.10.0) [23] pipeline to analyze the raw counts generated by SMURF. Initial filtering removed low-count cells based on SMURF's cluster assignments: for cells in Peyer's patches we required total counts to be

at least the smaller of 50 and 5 % of that cluster's total counts, and for all other clusters we required total counts to be at least the smaller of 50 and 10 % of that cluster's total counts. We then excluded cells with 2000 or more detected genes, 2000 or more total counts, a mitochondrial gene fraction of 25 % or higher, or a cell size of 45 or greater. These quality-control thresholds ensured the removal of low-quality or abnormal cells, enabling robust downstream analysis.

We then normalized the cells by scaling each cell's total counts to 1000 and performed log-normalization to stabilize variance across cells. Next, we retained the top 4000 variable genes and two key genes (*Pyy* and *Sct*). The data were scaled such that the maximum expression value for each gene was set to 10. We performed PCA to reduce the dimensionality to 40 components, followed by the construction of a nearest-neighbor graph using 25 neighbors. For visualization, we applied UMAP, and cell clustering was conducted using the Leiden algorithm with a resolution of 3 [88, 25]. Differential expression analysis was carried out using the default method of scanpy (scanpy.tl.rank_genes_groups, t-test) [23]. We annotate the cell types by gene markers according to reference data and knowledge with the help of location information [18, 89, 90].

To detect the crypt–villus boundary, we first performed a rough division based on a normalized distance threshold of 9, temporarily assigning cells with lower values to the crypt group and those with higher values to the villus group. We then conducted differential gene expression analysis between these two groups, selecting genes with an adjusted p-value below 1×10^{-5}. Using these genes, we built a random forest classifier by XGBoost (v 2.1.1) [91] to distinguish between crypt and villus cells based on their expression profiles. Recognizing that the crypt–villus continuum is a continuous structure, we further examined the local neighborhood of each cell. We assumed that the distribution of normalized y-axis positions for the crypt and villus groups follows a mixed Gaussian distribution and used the resulting boundary to precisely define the crypt–villus interface at each location.

Pseudotime Analysis

We used Scanpy (v1.10.0) [23] to perform pseudotime analysis on a subset of cells, including Paneth and stem cells, TA cells, and enterocytes. First, we computed diffusion maps to capture the underlying cellular trajectories. One cell was randomly selected from the Paneth and stem cells group as the putative starting point, and pseudotime was computed using the diffusion pseudotime function (sc.tl.dpt) [51]. To better visualize the structure of this cell subset, we employed PAGA (sc.tl.paga) [54] and force-directed graph drawing (sc.tl.draw_graph) [92, 55, 56] to generate a layout on which the pseudotime values were mapped.

We then selected cells that were not part of the highly concentrated Peyer's patches and normalized both the pseudotime values and the normalized corrected y-axis to a scale from 0 to 1. Using pseudotime (0-1) as the independent variable and the normalized corrected y-axis (0-1) as the dependent variable,

we performed linear regression analysis and computed the Pearson correlation coefficient to assess the strength and significance of their relationship.

Zonation Analysis

During the experiment, it was very challenging to maintain intact villi; many were truncated, which posed significant difficulties for analyzing villus structure. To address this, we selected only those villi that exhibited a full structure based on the proportion of tip enterocytes. The intestine was divided into small sections, and we applied a Kolmogorov–Smirnov test to evaluate the distribution of tip enterocytes in each section, following a uniform distribution. To ensure comprehensive coverage, multiple candidate sections from each area that passed the test were retained, and the final selection was refined through manual curation.

Following the exclusion of other cell types, we retained Paneth and stem cells, TA cells, and enterocytes. These cells were then subdivided into six distinct zones according to their normalized, corrected y-axis coordinates, ensuring that each zone contained an equivalent number of cells. Genes with counts greater than 10 were retained (12,482 genes in total), and a heatmap of the average gene expression in each zone was plotted with values normalized to a maximum of 1.

We then analyzed the differentially expressed genes of each zone using the default differential analysis for scanpy (scanpy.tl.rank_genes_groups, t-test) [23] and compared genes for each zone with all the rest of the zones. We generated the donated gene list with the adjusted p-value below 1×10^{-8} and a log2-fold-change (LFC) of 1.5. We also identified the specific significant zonated genes list for each zone by the same method, with constraints (i) an LFC above the zone-specific threshold and (ii) an adjusted p-value below the corresponding zone-specific cutoff; (iii) the candidate list was purged of any gene significantly up-regulated in another zone or muscle/immune/Fibroblast clusters. Here are the specific rules for each zone: (1) Zone 1, marker genes were required to have LFC > 2.1 and adjusted p-value $< 1 \times 10^{-23}$, candidates enriched in zones 2 or 3, or in muscle/immune datasets (adjusted p-value $< 10^{-50}$, LFC > 0) were removed; (2) Zone 2, LFC > 2.0, adjusted p-value $< 1 \times 10^{-40}$; genes enriched in muscle, fibroblast, and immune datasets (adjusted p-value $< 10^{-100}$, LFC > 0) were excluded; (3) Zone 3, LFC > 1.5, adjusted p-value $< 1 \times 10^{-30}$; genes up-regulated in zones 5 or 6, muscle, and immune datasets (adjusted p-value $< 10^{-100}$, LFC > 0) were removed. (4) Zone 4, LFC > 1.5, adjusted p-value $< 1 \times 10^{-10}$; genes up-regulated in zones 3 or 6, muscle, and immune datasets (adjusted p-value $< 10^{-100}$, LFC > 0) were excluded; (5) Zone 5, LFC > 1.5, adjusted p-value $< 1 \times 10^{-10}$, genes up-regulated in zones 3 or 6, muscle, and immune datasets (adjusted p-value $< 10^{-100}$, LFC > 0) were excluded; (6) Zone 6, LFC > 2.3, adjusted p-value $< 1 \times 10^{-20}$; genes up-regulated in zones 3, 4, or 5 (adjusted p-value $< 10^{-50}$, LFC > 0), muscle, and immune datasets (adjusted p-value $< 10^{-100}$, LFC > 0) were removed.

Additionally, we used gprofiler-official (v1.0.0) [93] to analyze KEGG pathway enrichment for each zonation. KEGG pathways were retained if any zone

exhibited an adjusted p-value below 1×10^{-8} and an LFC of 1.5. A bubble plot was then generated for the pathways across all zones, where the bubble size indicates the fraction of genes intersected relative to the largest count in the zone for that specific pathway, and the bubble color represents the corresponding p-value. Finally, for other cell types, we utilized the boundaries of these zones to assign the remaining cells to specific zones. For each cell type—specifically enterocytes, EEC, tuft cells, and Goblet cells—we selected the subset of cells corresponding to that type and plotted the average expression of selected marker genes within each zone.

We then sought to capture proximal–distal differential genes across different zones. Cells were first divided along the x-axis, with the first half designated as the proximal group and the second half as the distal group. Only the subset of Paneth and stem cells, TA cells, and enterocytes was used, with each zone containing an approximately equal number of cells. For each zone, mitochondrial genes were filtered out, and only genes expressed in at least 30 cells were retained. Differential expression analysis was performed using an LFC threshold of 1.5 and an adjusted p-value cutoff of 0.001. The number of differential genes was then normalized by dividing by the total number of genes remaining after filtration. Finally, we generated plots of the differential gene count and the fraction of differential genes versus zones. KEGG pathways of the proximal and distal zone 6 genes were calculated if they exhibited an adjusted p-value below 0.001 and an LFC of 1.5.

Transcription Factor Analysis

We selected Paneth cells, stem cells, TA cells, and enterocytes — each annotated with villus-zone information — from the previous section, and then followed the pySCENIC protocol (https://github.com/aertslab/SCENICprotocol) by applying `pyscenic` to infer TF–target gene regulatory networks and to identify and score regulons significantly enriched in our dataset by comparison with established reference databases [67].

A total of 192 regulons have been called by the results. For each regulon, we computed its mean activity score across the six spatial zones. We then visualized the normalized activity, scaled such that the maximum score for every regulon across the six zones equals 1, for 19 representative regulons in a heatmap. We generated violin plots showing the distribution of the corrected y-axis positions (μm) of the constituent genes for each regulon.

We selected zone-specific regulons (LFC > 0.3 and adjusted p-value $< 1 \times 10^{-10}$) and plotted the TF–target network for all six zones. For Zone 2, we built a TF–target network for the seven regulons whose activities differed significantly across zones by differential t-test from scanpy. In the graph, large blue nodes denote the TF genes, small green nodes mark non-zonated targets, red nodes mark zonated targets, and nodes in additional colours highlight zonated targets enriched for specific functional categories. To probe zonation enrichment, we applied Fisher's exact tests to compare genes targeted by at

least two TFs with all remaining targets, and to compare each TF — comparing its unique targets with the pool of all other target genes separately. We completed the same analysis for Zone 6 with five differential regulons.

Single-cell Analysis of Seiler et al [83]

We downloaded the dataset from GEO (GSE130113), which contains raw counts from three sham and two SBR replicates. We first filtered cells by requiring a minimum of 200 genes per cell, and filtered genes by retaining only those expressed in at least 3 cells. The data were then normalized to a total of 10,000 counts per cell and log-normalized. We selected highly variable genes, scaled the data, performed PCA, and applied Harmony for batch correction. UMAP was used for dimensionality reduction, and clustering was performed using the Leiden algorithm [88, 25]. Differential expression analysis was carried out using a t-test. Although the differential results were not exactly identical to those reported in the original paper, they were extremely similar—the top five genes were exactly the same, and all the top 10 genes reported in the paper appeared as the top 11 in our analysis. Finally, we selected differential genes using a threshold of adjusted p-value less than 1×10^{-50} and an absolute LFC larger than 1, resulting in a total of 77 genes significantly differentially expressed in the SBR group.

We then mapped the differential genes to the spatial dataset, and 72 of them were expressed in the spatial data. Next, we sought to determine the specific zone in which each gene was predominantly expressed by defining a gene as expressed in a particular zone if its LFC in that zone was greater than 1.5 and its adjusted p-value was less than 1×10^{-8}. Finally, we plotted the zones in which all SBR differential genes were expressed in the spatial dataset.

Comparison Analysis for Results from Different Methods

We compared the results of 8 μm spot data, Bin2cell results, nuclei segmentation results, and SMURF results. For each dataset, we used the direct output of 4×4 spots (corresponding to an 8 μm × 8 μm spot size) as the 8 μm spot data. For Bin2cell, we followed the instructions provided in the Bin2cell GitHub repository [33]. Nuclei segmentation results were obtained using Squidpy (v1.4.1) [87] with Stardist (v0.8.5) [22].

In the analysis of the percentage of cells covered by nuclei, we encountered the issue that a single spot might encompass several nuclei, leading to multiple nuclei being assigned to a cell, an outcome that is not accurate. To address this, we defined a cell as covered by a nucleus only if at least 30% of the pixels in the cell region are occupied by that nucleus.

For the definition of nuclei data in the Mouse Brain dataset, we considered a spot to belong to a nucleus if at least 90% of the spot's counts were within that nucleus. We then uploaded the raw counts matrix of nuclei, SMURF, and Xenium results from 10x [36] to MapMyCells[41, 42] by selecting the "10x Whole Mouse Brain (CCN20230722)" as the reference dataset. Using the Hierarchical Mapping strategy available in MapMyCells, we obtained the cell type results for the three datasets. We calculated the KL divergence according

to the "class_name" from the MapMyCells results. In the mouse small intestine dataset, we applied the same segmentation and clustering pipeline as provided by 10x Genomics[94].

Data Availability

The public mouse brain Visium HD and mouse small intestine datasets are available on the 10x Genomics website (mouse brain Visium HD, mouse intestine). Public mouse brain Xenium data can be accessed on the 10x Genomics website (fresh-frozen mouse brain replicates). The public SBR experiment dataset from Seiler et al (2019). is available from GEO under accession number GSE130113. The ISH staining for gene *Slc1a2* was obtained from the Allen Brain Atlas (https://mouse.brain-map.org/gene/show/20273).

Code Availability

The SMURF package is available at https://github.com/The-Mitra-Lab/SMURF. The code to reproduce the results presented in this study can be accessed at https://github.com/The-Mitra-Lab/SMURF_figures. The documentation for SMURF is available at https://the-mitra-lab.github.io/SMURF/.

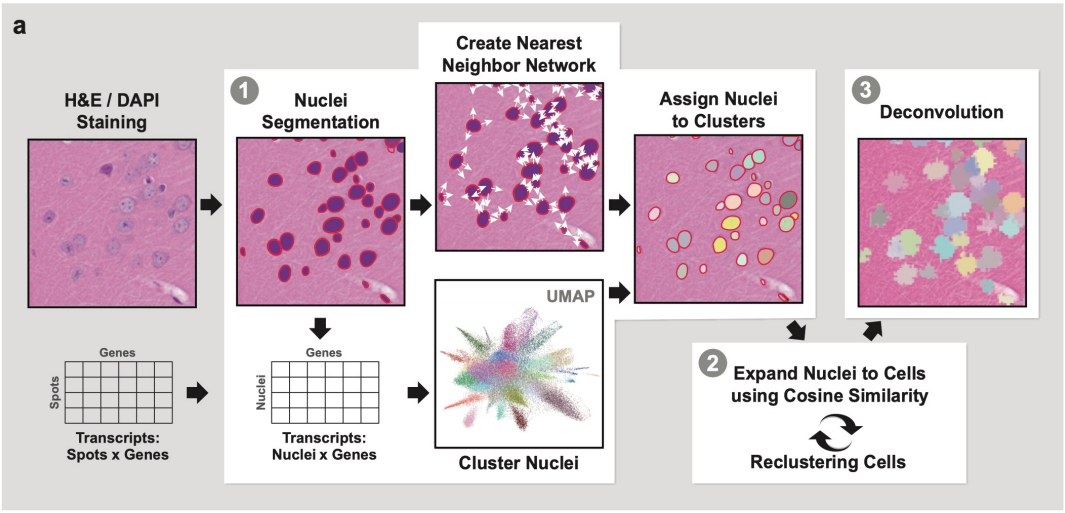
Funding

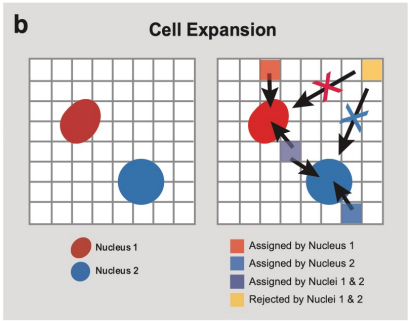
This work has been funded by grants from National Institute of Mental Health (RF1MH126723, R01DA056829, and RM1MH138313 to J.D.D. and R.D.M.; R01MH124808 to J.D.D.; R01DE032865 to R.D.M.). S.S. was supported in part by the Autism Science Foundation (22-007). B.D.M. was supported by the Department of Veterans Affairs, Veterans Health Administration, Biomedical Laboratory Research and Development (1IK2BX004909) and the Washington University Digestive Disease Research Core Center (P30DK052574).

Acknowledgements

We thank Hao Fan for discussions that helped shape this project, and Jennifer Hayden for scientific illustrations. The contents of this manuscript do not represent the views of the Department of Veterans Affairs or the United States Government.

Figures


a
H&E / DAPI Staining
1
Nuclei Segmentation
Create Nearest Neighbor Network
Assign Nuclei to Clusters
3
Deconvolution
Genes
Spots
Transcripts: Spots x Genes
Genes
Nuclei
Transcripts: Nuclei x Genes
UMAP
Cluster Nuclei
2
Expand Nuclei to Cells using Cosine Similarity
Reclustering Cells


b
Cell Expansion
Nucleus 1
Nucleus 2
Assigned by Nucleus 1
Assigned by Nucleus 2
Assigned by Nuclei 1 & 2
Rejected by Nuclei 1 & 2

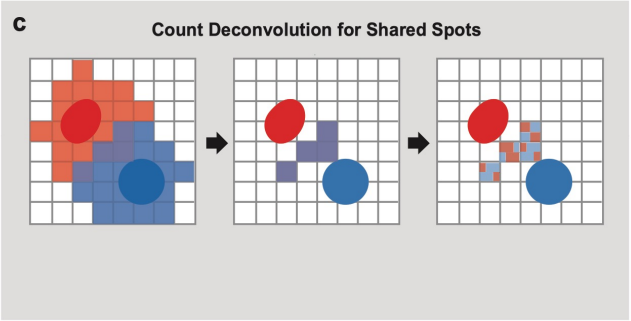

c
Count Deconvolution for Shared Spots

Fig. 1: **SMURF uses soft-segmentation to reconstruct single-cell transcriptomes.** (a) Schematic of SMURF soft-segmentation algorithms. SMURF begins by segmenting nuclei from the image data and creating nuclear gene expression clusters. The algorithm iteratively expands cell boundaries to include additional capture spots. Finally, SMURF leverages deep learning to deconvolve spots assigned to multiple nuclei. (b) Example of capture spots assignment to cells. Capture spots can be assigned to zero, one, or multiple nuclei during cell expansion, depending on their proximity and expression similarity. (c) Count deconvolution for shared spots. After the cell expansion step, capture spots may be shared between multiple cells (purple squares in the first two panels). Transcripts from shared capture spots are divided among cells using a quadratic programming or deep learning algorithm (see Methods).

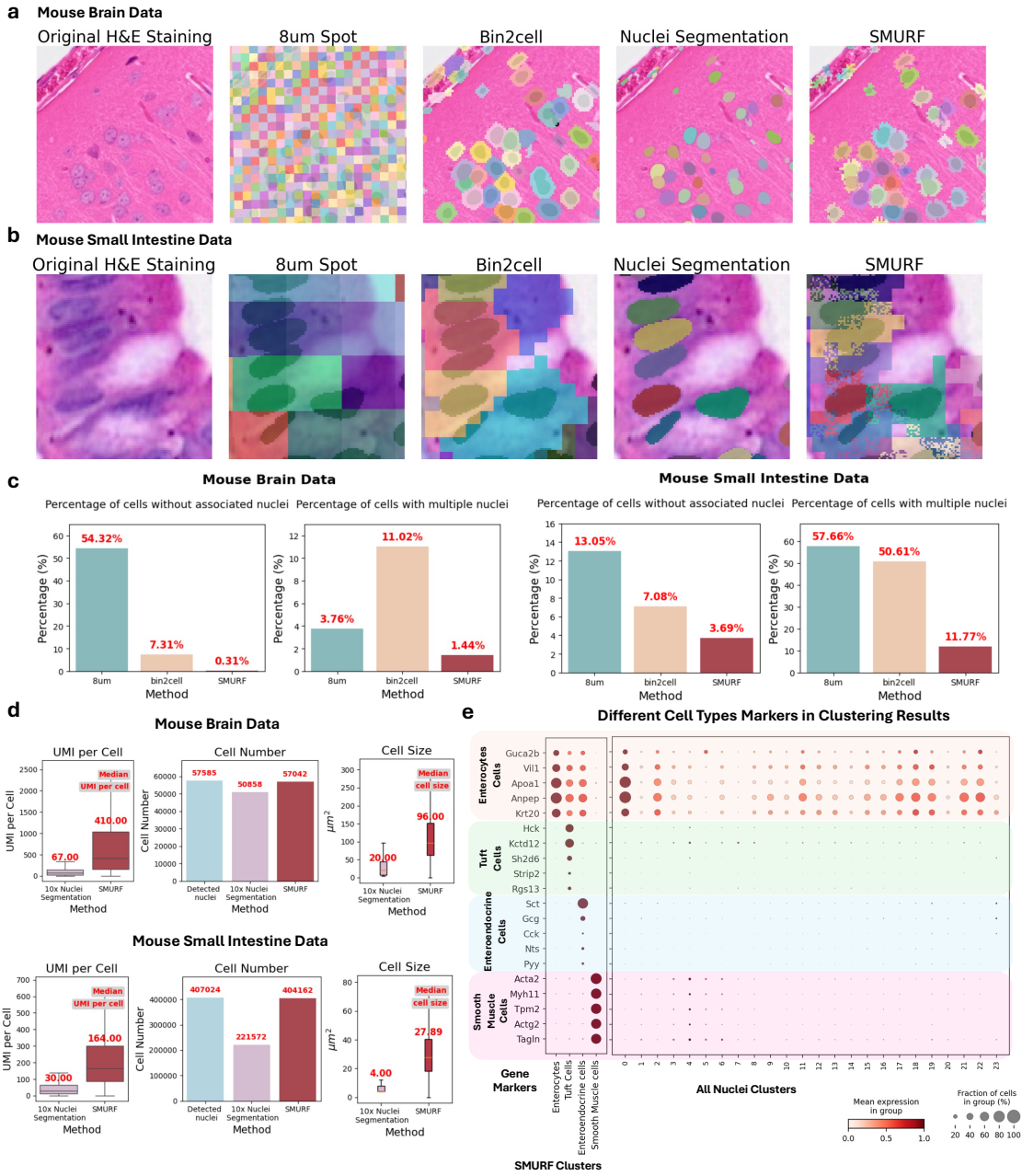

a
Mouse Brain Data
Original H&E Staining
8um Spot
Bin2cell
Nuclei Segmentation
SMURF
b
Mouse Small Intestine Data
Original H&E Staining
8um Spot
Bin2cell
Nuclei Segmentation
SMURF
c
Mouse Brain Data
Percentage of cells without associated nuclei
Percentage of cells with multiple nuclei
54.32%
7.31%
0.31%
3.76%
11.02%
1.44%
Mouse Small Intestine Data
Percentage of cells without associated nuclei
Percentage of cells with multiple nuclei
13.05%
7.08%
3.69%
57.66%
50.61%
11.77%
Percentage (%)
8um
bin2cell
SMURF
Method
d
Mouse Brain Data
UMI per Cell
Median UMI per cell
67.00
410.00
10x Nuclei Segmentation
SMURF
Cell Number
57585
50858
57042
Detected nuclei
10x Nuclei Segmentation
SMURF
Cell Size
Median cell size
20.00
96.00
10x Nuclei Segmentation
SMURF
Mouse Small Intestine Data
UMI per Cell
Median UMI per cell
30.00
164.00
10x Nuclei Segmentation
SMURF
Cell Number
407024
221572
404162
Detected nuclei
10x Nuclei Segmentation
SMURF
Cell Size
Median cell size
4.00
27.89
10x Nuclei Segmentation
SMURF
Method
e
Different Cell Types Markers in Clustering Results
Enterocytes Cells
Guca2b
Vil1
Apoa1
Anpep
Krt20
Tuft Cells
Hck
Kctd12
Sh2d6
Strip2
Rgs13
Enteroendocrine Cells
Sct
Gcg
Cck
Nts
Pyy
Smooth Muscle Cells
Acta2
Myh11
Tpm2
Actg2
Tagln
Gene Markers
Enterocytes
Tuft Cells
Enteroendocrine cells
Smooth Muscle cells
SMURF Clusters
All Nuclei Clusters
Mean expression in group
0.0
0.5
1.0
Fraction of cells in group (%)
20
40
60
80
100

Fig. 2: **SMURF accurately assigns transcripts to cells in mouse brain and mouse small intestine data.** (a,b) Representative images illustrating how SMURF and other ST segmentation methods aggregate Visium HD capture spots to define cells in the mouse brain (top row) and mouse small intestine (bottom row). Capture spots are shaded according to their assigned cell. (c) Percentage of cells without an associated nucleus or containing multiple nuclei in the mouse brain (left) and mouse small intestine using different ST segmentation strategies. (d) Cell properties and UMI per cell using SMURF or nuclei-based segmentation. (e) Cell type annotation of clusters in mouse small intestine using SMURF or nuclei-segmentation. Bubble plot of marker gene expression for four intestinal cell types across clusters. Abbreviations: UMI, Unique Molecular Identifier.

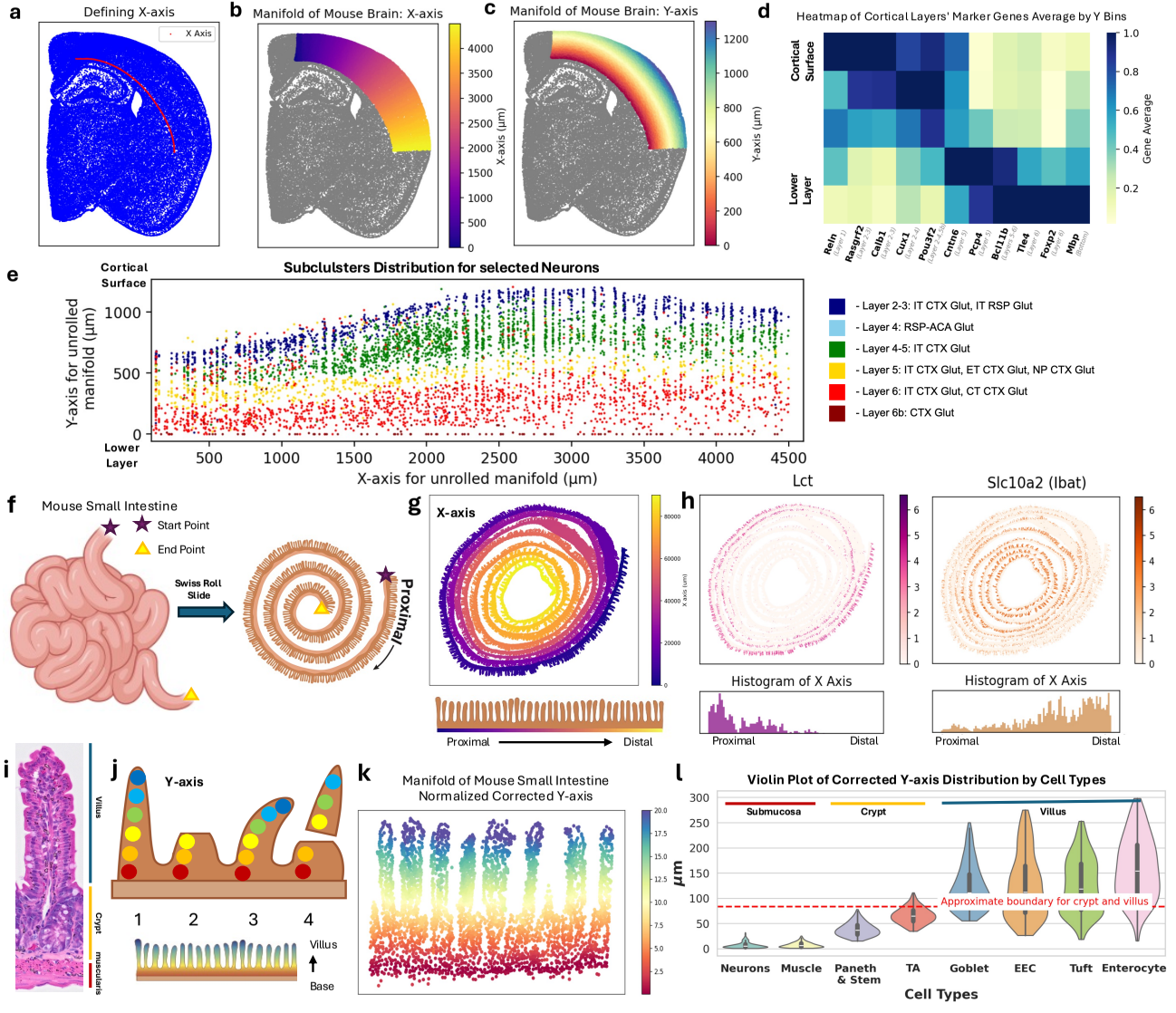

a
Defining X-axis
X Axis
b
Manifold of Mouse Brain: X-axis
X-axis (µm)
c
Manifold of Mouse Brain: Y-axis
Y-axis (µm)
d
Heatmap of Cortical Layers' Marker Genes Average by Y Bins
Cortical Surface
Lower Layer
Gene Average
Reln
Rasgrf2
Calb1
Cux1
Pou3f2
Cntn6
Pcp4
Bcl11b
Tle4
Foxp2
Mbp
e
Subclulsters Distribution for selected Neurons
Cortical Surface
Lower Layer
Y-axis for unrolled manifold (µm)
X-axis for unrolled manifold (µm)
- Layer 2-3: IT CTX Glut, IT RSP Glut
- Layer 4: RSP-ACA Glut
- Layer 4-5: IT CTX Glut
- Layer 5: IT CTX Glut, ET CTX Glut, NP CTX Glut
- Layer 6: IT CTX Glut, CT CTX Glut
- Layer 6b: CTX Glut
f
Mouse Small Intestine
Start Point
End Point
Swiss Roll Slide
Proximal
g
X-axis
Proximal
Distal
h
Lct
Slc10a2 (Ibat)
Histogram of X Axis
Proximal
Distal
i
Villus
Crypt
muscularis
j
Y-axis
1
2
3
4
Villus
Base
k
Manifold of Mouse Small Intestine
Normalized Corrected Y-axis
l
Violin Plot of Corrected Y-axis Distribution by Cell Types
Submucosa
Crypt
Villus
µm
Approximate boundary for crypt and villus
Neurons
Muscle
Paneth & Stem
TA
Goblet
EEC
Tuft
Enterocyte
Cell Types

Fig. 3: **Manifold unwrapping of Visium HD data in cortex and small intestine.** (a) Selection of the X-axis contour in the mouse cortex based on positions of cells in the basement membrane cluster. (b) Overlay of cortical manifold onto mouse brain section, colored by X-axis coordinates (µm). (c) Overlay of cortical manifold onto mouse brain section, colored by Y-axis coordinates (µm). (d) Heatmap of marker genes for cortical layers across different Y-bins of the mouse brain manifold. (e) Distribution of cortical layer-specific neurons in the unrolled rectilinear coordinate system. (f) Schematic illustration of the mouse small intestine and its "Swiss roll" histological orientation. (g) Overlay of mouse small intestine data, colored by unrolled X-axis coordinates based on cells in the muscle layer (µm). (h) Expression of selected proximal and distal enriched genes in the original ST image and histogram plot of those genes on the unrolled X-axis of the mouse small intestine. (*Lct*, *Slc10a2(Ibat)*). (i) Illustration of the crypt-villus axis in the mouse small intestine. (j) Schematic of four common villus structures encountered during construction of the Y-axis for mouse small intestine data. (k) Overlay of intestinal manifold in selected villi with cells colored by normalized, corrected Y-axis. (l) Cell type distribution along the corrected Y-axis. Red dashed line indicates the approximate boundary between crypt and villus regions.

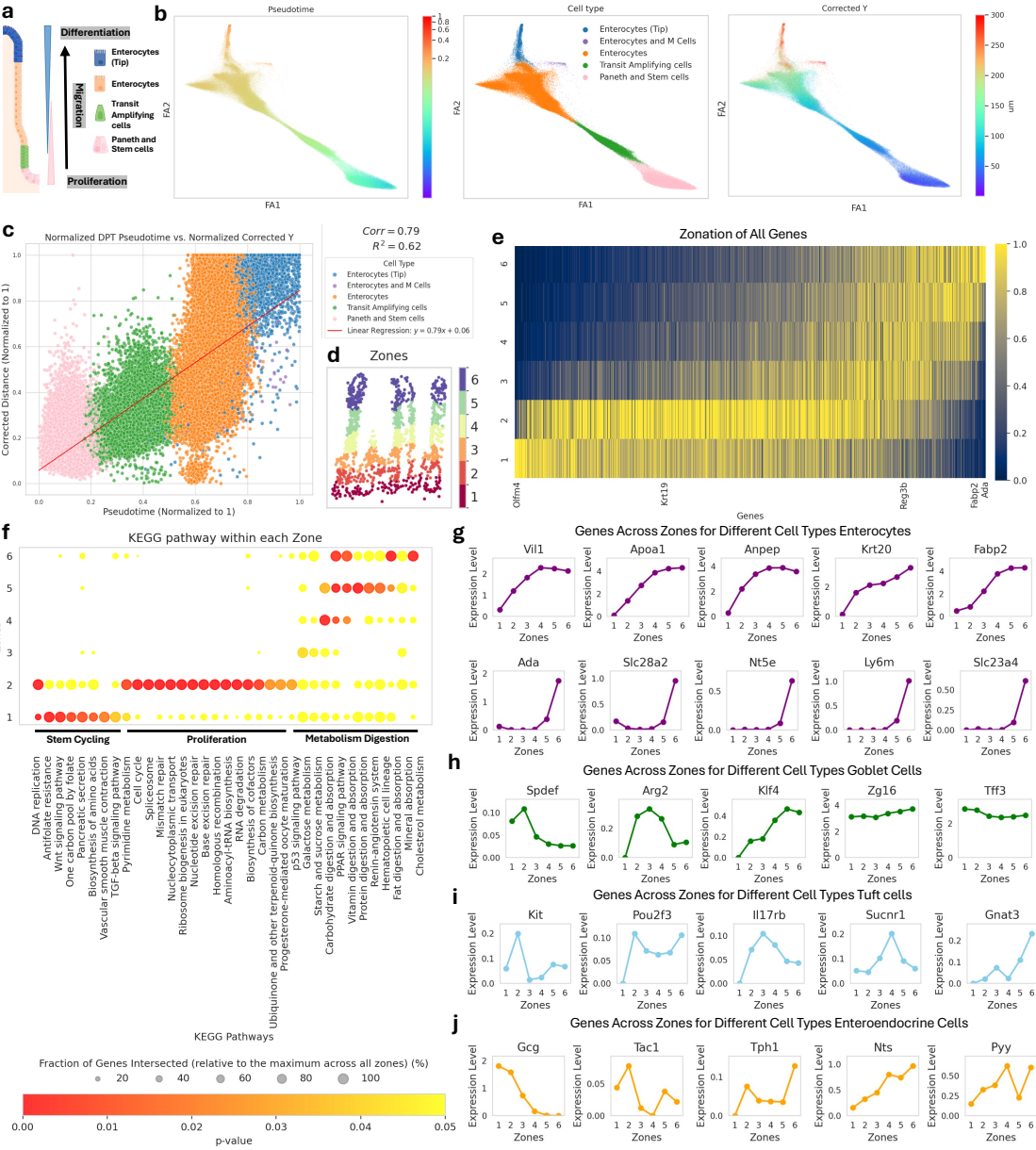

a
Differentiation
Enterocytes (Tip)
Enterocytes
Transit Amplifying cells
Paneth and Stem cells
Migration
Proliferation
b
Pseudotime
Cell type
Corrected Y
FA1
FA2
Enterocytes (Tip)
Enterocytes and M Cells
Enterocytes
Transit Amplifying cells
Paneth and Stem cells
um
c
Normalized DPT Pseudotime vs. Normalized Corrected Y
Corr = 0.79
R² = 0.62
Cell Type
Linear Regression: y = 0.79x + 0.06
Corrected Distance (Normalized to 1)
Pseudotime (Normalized to 1)
d
Zones
e
Zonation of All Genes
Olfm4
Krt19
Reg3b
Fabp2
Ada
Genes
f
KEGG pathway within each Zone
Zones
Stem Cycling
Proliferation
Metabolism Digestion
DNA replication
Antifolate resistance
Wnt signaling pathway
One carbon pool by folate
Pancreatic secretion
Biosynthesis of amino acids
Vascular smooth muscle contraction
TGF-beta signaling pathway
Pyrimidine metabolism
Cell cycle
Spliceosome
Mismatch repair
Nucleocytoplasmic transport
Ribosome biogenesis in eukaryotes
Nucleotide excision repair
Base excision repair
Homologous recombination
Aminoacyl-tRNA biosynthesis
RNA degradation
Biosynthesis of cofactors
Carbon metabolism
Ubiquinone and other terpenoid-quinone biosynthesis
Progesterone-mediated oocyte maturation
p53 signaling pathway
Galactose metabolism
Starch and sucrose metabolism
Carbohydrate digestion and absorption
PPAR signaling pathway
Vitamin digestion and absorption
Protein digestion and absorption
Renin-angiotensin system
Hematopoietic cell lineage
Fat digestion and absorption
Mineral absorption
Cholesterol metabolism
KEGG Pathways
Fraction of Genes Intersected (relative to the maximum across all zones) (%)
p-value
g
Genes Across Zones for Different Cell Types Enterocytes
Vil1
Apoa1
Anpep
Krt20
Fabp2
Ada
Slc28a2
Nt5e
Ly6m
Slc23a4
Expression Level
Zones
h
Genes Across Zones for Different Cell Types Goblet Cells
Spdef
Arg2
Klf4
Zg16
Tff3
i
Genes Across Zones for Different Cell Types Tuft cells
Kit
Pou2f3
Il17rb
Sucnr1
Gnat3
j
Genes Across Zones for Different Cell Types Enteroendocrine Cells
Gcg
Tac1
Tph1
Nts
Pyy

Fig. 4: **Temporal and functional zonation of transcriptional programs on the ileal crypt-villus axis.** (a) Schematic illustrating the epithelial differentiation trajectory from proliferative stem cells (bottom) to differentiated enterocytes (top). (b) Concordance of pseudotime trajectory and corrected y-axis position in enterocyte maturation. PAGA embedding of intestinal stem and Paneth cells, TA, and enterocytes, colored by pseudotime position, cell type, and corrected y-axis position. (c) Scatter plot comparing the normalized pseudotime progression score and the normalized corrected Y position. Each point represents a single cell. The dashed red line indicates a linear fit (Pearson correlation = 0.79). (d) Example of discrete binning of cells into 6 zones along the y-axis. (e) Heatmap of normalized gene expression across zones. Gene expression in all stem and Paneth cells, TA, and enterocytes is included. The heatmap displays normalized expression of each gene (columns) in each zone (rows). (f) Bubble plot displaying KEGG pathway enrichment of significantly zonated genes. (g-j) Zonation profiles of selected genes in enterocytes (g), goblet cells (h), tuft cells (i), and EEC (j).

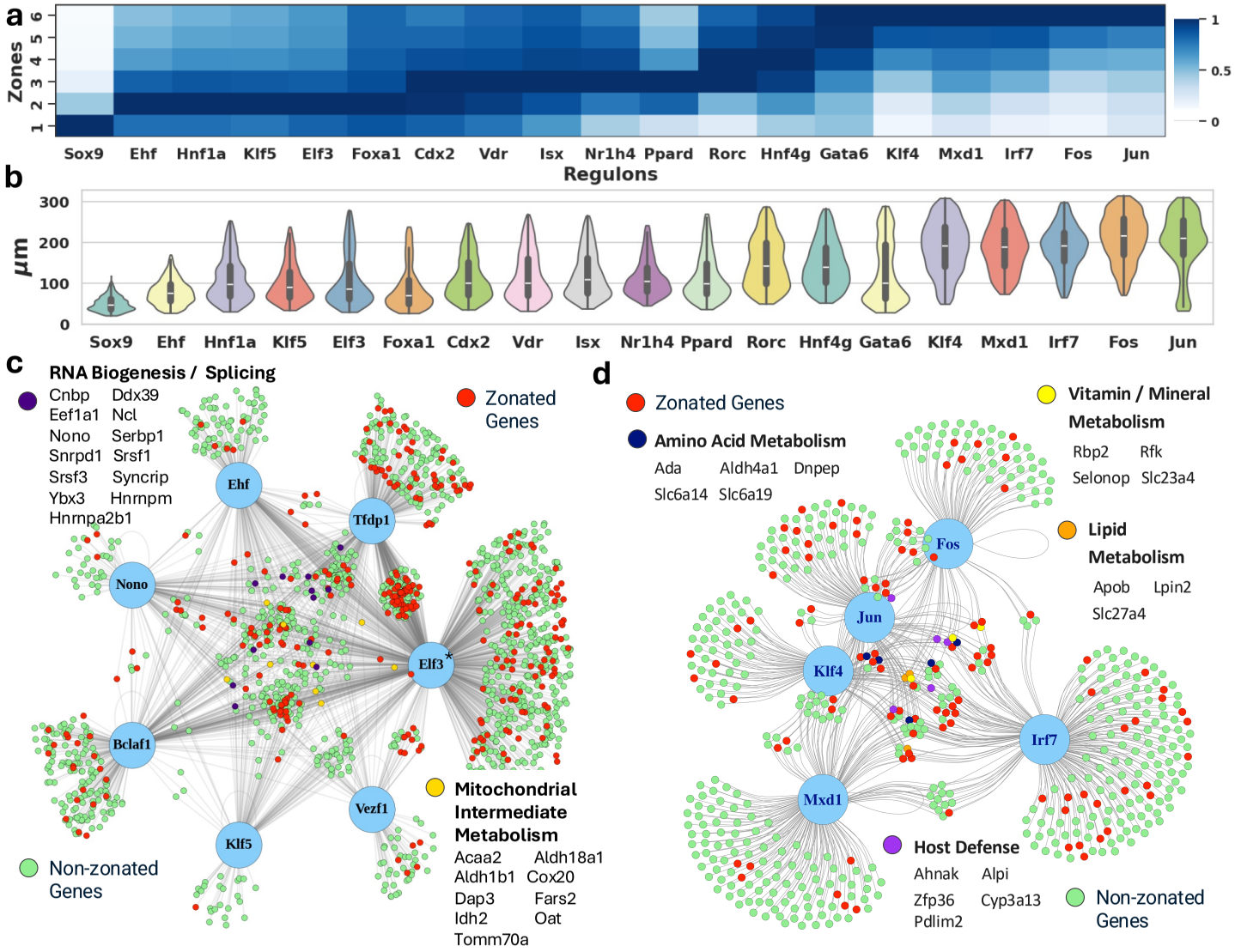

a
Zones
1 2 3 4 5 6
Sox9 Ehf Hnf1a Klf5 Elf3 Foxa1 Cdx2 Vdr Isx Nr1h4 Ppard Rorc Hnf4g Gata6 Klf4 Mxd1 Irf7 Fos Jun
1
0.5
0
b
Regulons
µm
300
200
100
0
Sox9 Ehf Hnf1a Klf5 Elf3 Foxa1 Cdx2 Vdr Isx Nr1h4 Ppard Rorc Hnf4g Gata6 Klf4 Mxd1 Irf7 Fos Jun
c
RNA Biogenesis / Splicing
Cnbp Ddx39
Eef1a1 Ncl
Nono Serbp1
Snrpd1 Srsf1
Srsf3 Syncrip
Ybx3 Hnrnpm
Hnrnpa2b1
Zonated Genes
Ehf
Tfdp1
Nono
Elf3*
Bclaf1
Vezf1
Klf5
Mitochondrial Intermediate Metabolism
Acaa2 Aldh18a1
Aldh1b1 Cox20
Dap3 Fars2
Idh2 Oat
Tomm70a
Non-zonated Genes
d
Zonated Genes
Amino Acid Metabolism
Ada Aldh4a1 Dnpep
Slc6a14 Slc6a19
Vitamin / Mineral Metabolism
Rbp2 Rfk
Selonop Slc23a4
Fos
Lipid Metabolism
Apob Lpin2
Slc27a4
Jun
Klf4
Irf7
Mxd1
Host Defense
Ahnak Alpi
Zfp36 Cyp3a13
Pdlim2
Non-zonated Genes

Fig. 5: **Spatially resolved transcription-factor programs along the crypt-villus axis.** (a) Heatmap of normalized regulon activity scores for 20 key transcription factors (TFs) across the six villus zones. Regulon scores calculated with Scenic.
(b) Violin plots of cell number expressing the same TFs, plotted against the unrolled corrected Y-axis coordinate (μm). Black bars indicate the median and inter-quartile range; violin width reflects local cell density.
(c) Integrated TF target networks for Zone 2. TFs are depicted in blue circles. Lines connect TFs to their gene targets. Selected target genes are color-coded by their predominant functional category. *: 500 of the 1362 *ELF3* targets are shown for legibility. (d) Integrated TF target networks for Zone 6. Coloring as in panel c.

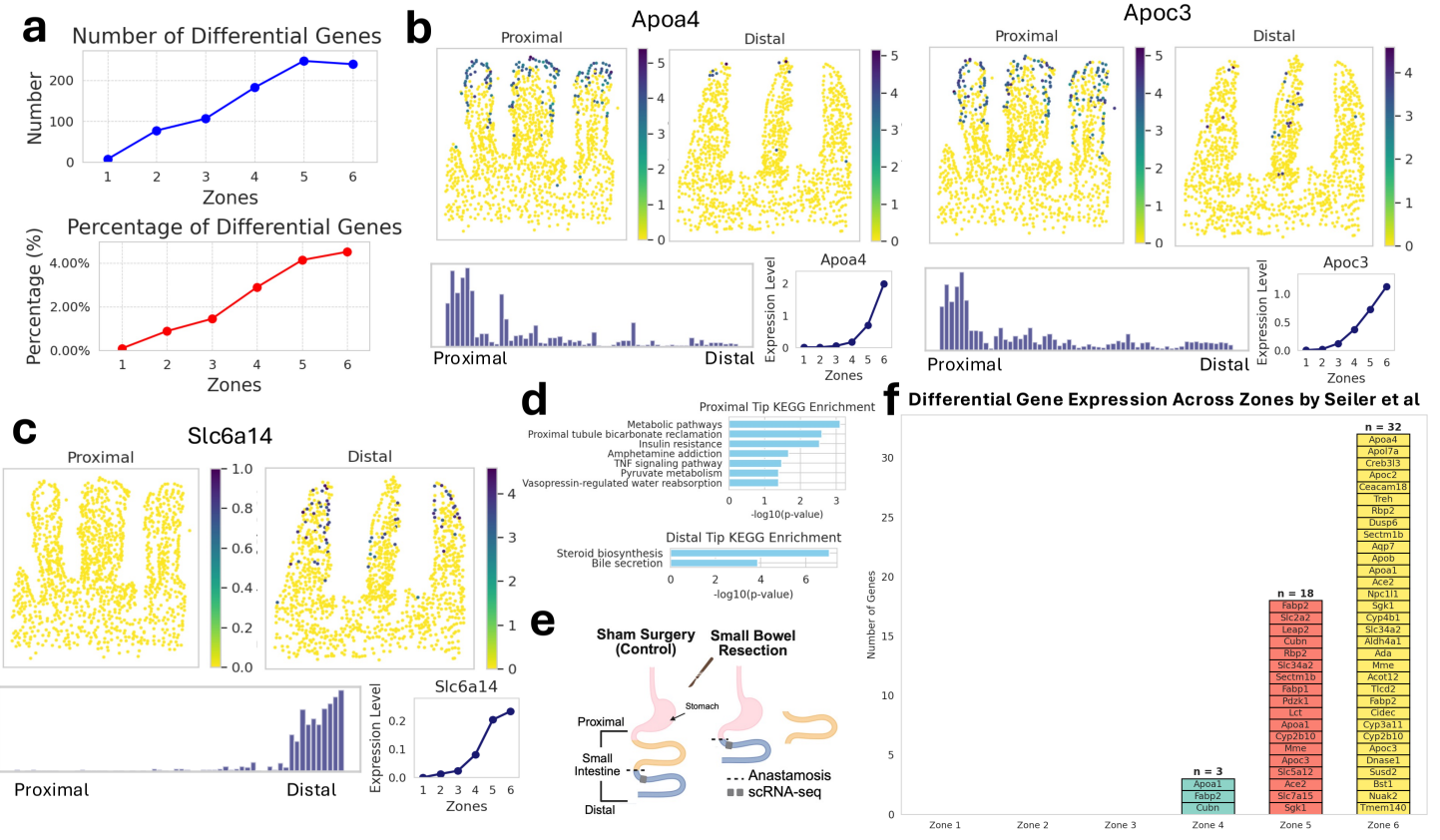
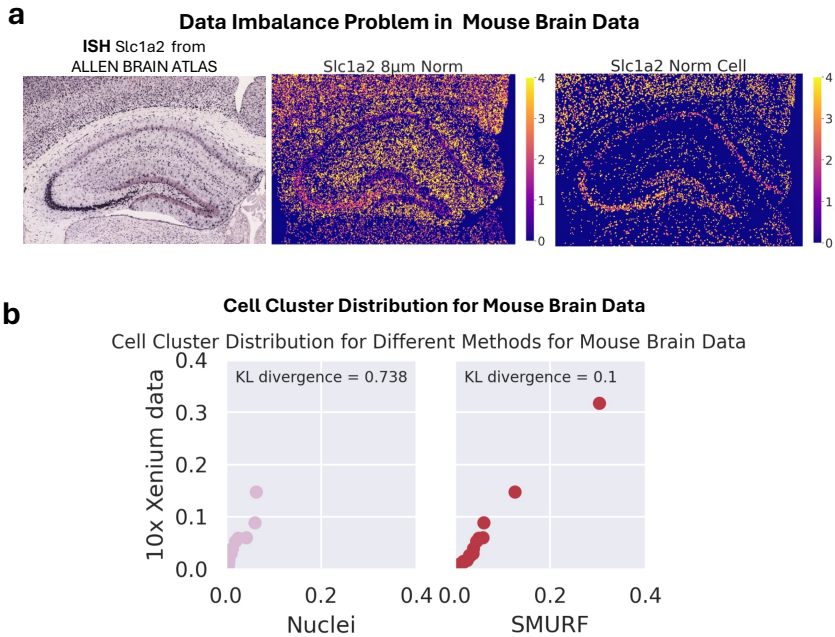
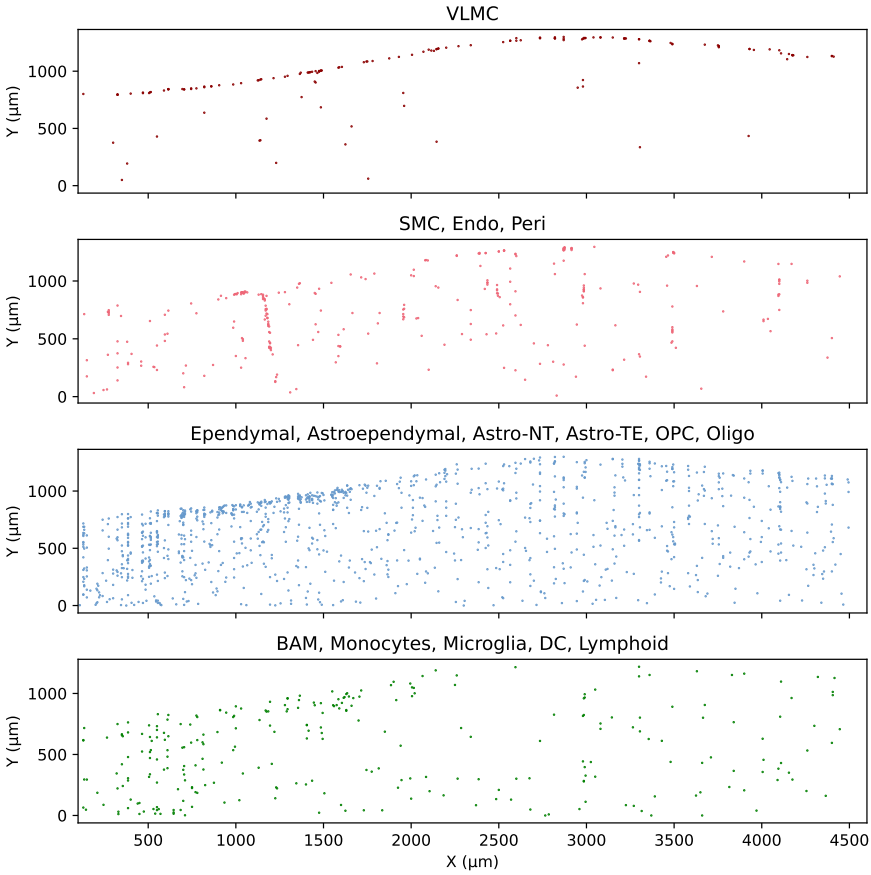

a
Number of Differential Genes
Number
Zones
Percentage of Differential Genes
Percentage (%)
Zones
b
Apoa4
Proximal
Distal
Apoa4
Expression Level
Zones
Apoc3
Proximal
Distal
Apoc3
Expression Level
Zones
c
Slc6a14
Proximal
Distal
Slc6a14
Expression Level
Zones
d
Proximal Tip KEGG Enrichment
Metabolic pathways
Proximal tubule bicarbonate reclamation
Insulin resistance
Amphetamine addiction
TNF signaling pathway
Pyruvate metabolism
Vasopressin-regulated water reabsorption
-log10(p-value)
Distal Tip KEGG Enrichment
Steroid biosynthesis
Bile secretion
-log10(p-value)
e
Sham Surgery (Control)
Small Bowel Resection
Stomach
Proximal
Small Intestine
Distal
Anastamosis
scRNA-seq
f
Differential Gene Expression Across Zones by Seiler et al
Number of Genes
n = 3
Apoa1
Fabp2
Cubn
n = 18
Fabp2
Slc2a2
Leap2
Cubn
Rbp2
Slc34a2
Sectm1b
Fabp1
Pdzk1
Lct
Apoa1
Cyp2b10
Mme
Apoc3
Slc5a12
Ace2
Slc7a15
Sgk1
n = 32
Apoa4
Apol7a
Creb3l3
Apoc2
Ceacam18
Treh
Rbp2
Dusp6
Sectm1b
Aqp7
Apob
Apoa1
Ace2
Npc1l1
Sgk1
Cyp4b1
Slc34a2
Aldh4a1
Ada
Mme
Acot12
Tlcd2
Fabp2
Cidec
Cyp3a11
Cyp2b10
Apoc3
Dnase1
Susd2
Bst1
Nuak2
Tmem140
Zone 1
Zone 2
Zone 3
Zone 4
Zone 5
Zone 6

Fig. 6: **Regionally specific gene programs along the proximal-distal intestinal axis accumulate in the villus tip.** (a) Number and fraction of genes that are differentially expressed in each zone, comparing proximal and distal regions. Only genes expressed in more than 30 cells are considered. (b) Expression of proximal-tip enriched genes (*Apoa4* and *Apoc3*). Data shows expression in the original ST data, as well as expression along the unwrapped x and y axes. (c) Expression of distal-tip enriched gene (*Slc6a14*. Data shows expression in the original ST data, as well as expression along the unwrapped x and y axes. (d) KEGG pathway enrichment for proximal and distal tip-associated genes. (e) Schematic of small bowel resection and scRNA-seq dataset (Seiler et al [83]). (f) Tip cells, but not cells at the base of crypts, are reprogrammed one week after small bowel resection. For each zone, the number and identity of differentially expressed genes from the single-cell data are shown.

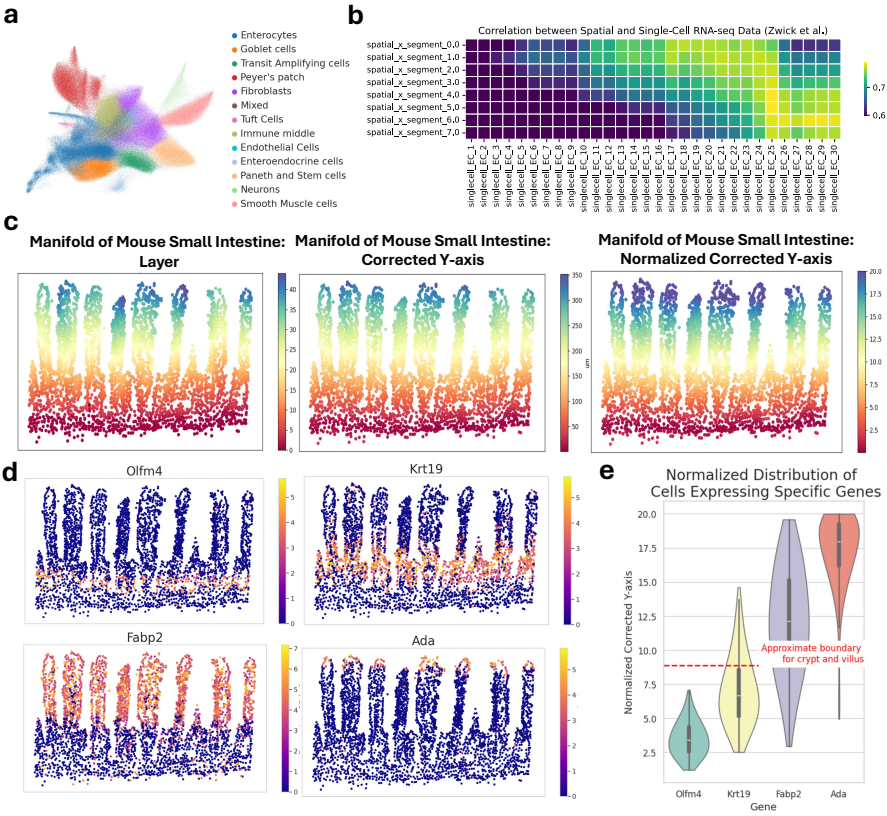
Extended Data Figures



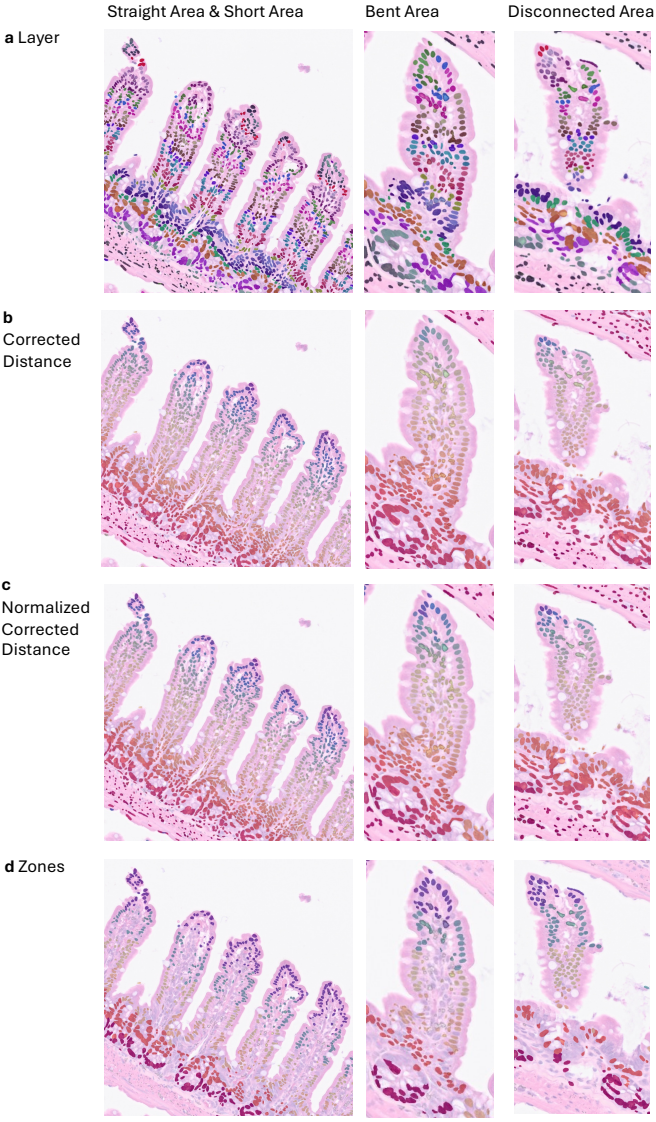
Extended Data Figure 1: **Soft segmentation with SMURF** (a) Illustration of the data imbalance issue when the standard normalized 8 μm Spot method is used. The spatial expression pattern of *Slc1a2* is discordant between Visium HD mouse brain data analyzed by the 8 μm Spot method and ISH data from the Allen Brain Atlas. In contrast, when analyzed by SMURF, there is concordance. (b) Cell type proportions annotated by MapMyCells in 10x Xenium mouse brain data, comparing nuclei results and SMURF results. KL divergence values quantify differences in cell type distributions. Lower KL divergence indicates better concordance.



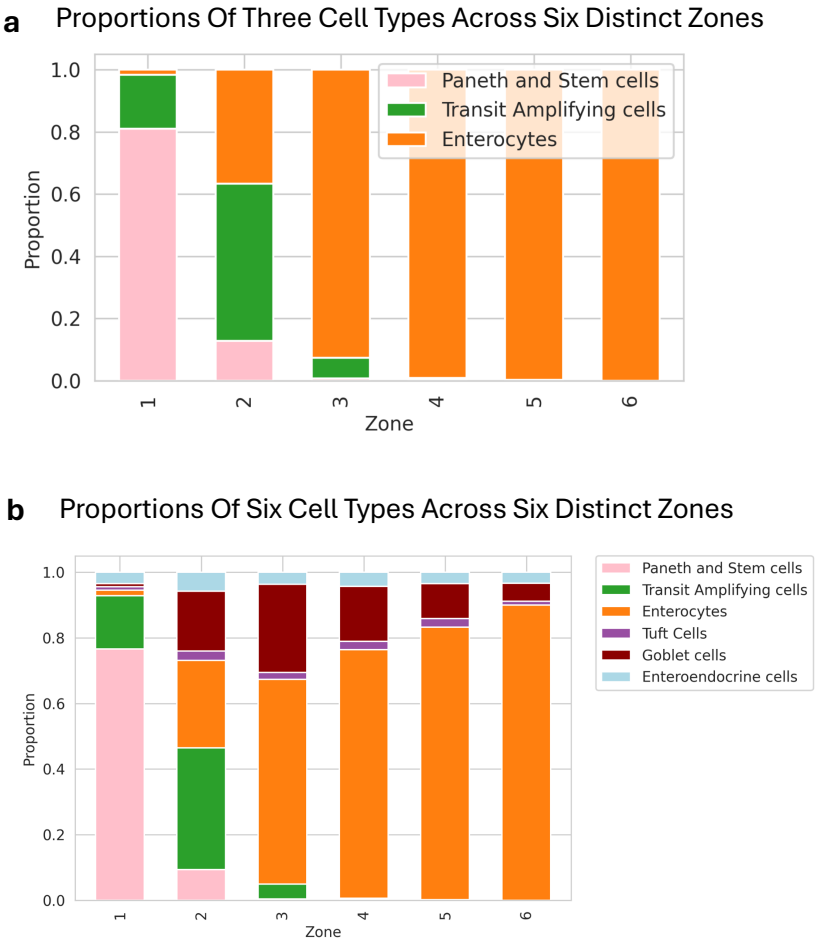
Extended Data Figure 2: Distribution of cortical layer-specific non-neuronal cells in the unrolled rectilinear coordinate system



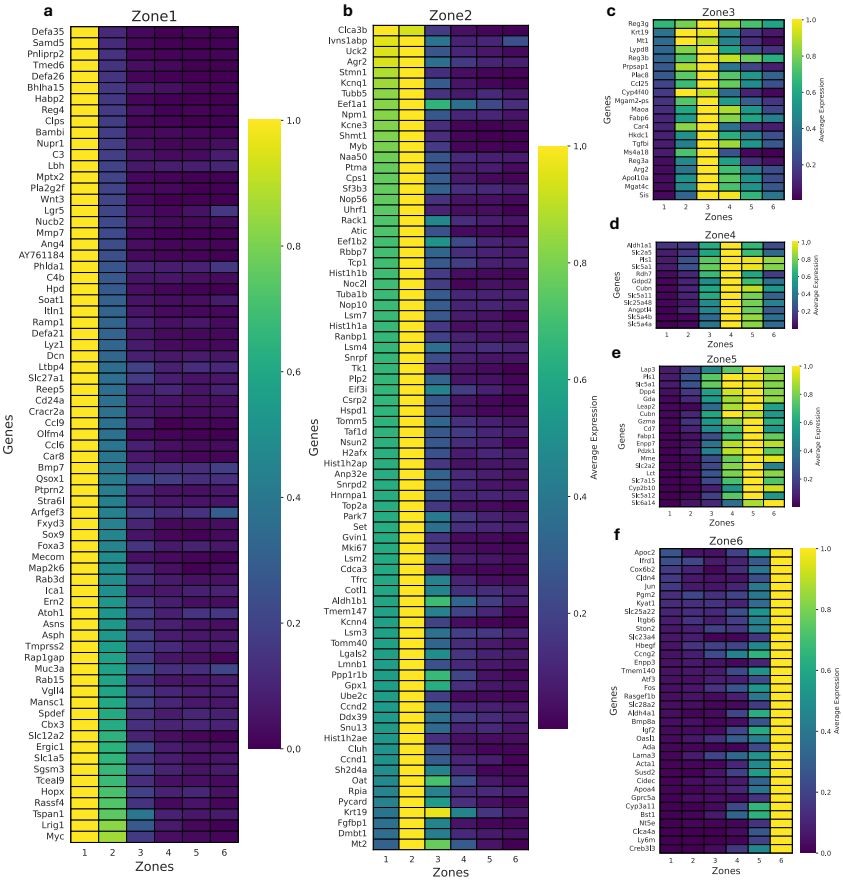
Extended Data Figure 3: **Mouse small intestine analysis.** (a) UMAP embedding of intestinal cells, colored by annotated cell types. (b) Correlation between single-cell transcriptomes from Zwick et al. [18] and spatial transcriptomic profiles across intestinal sections. (c) Overlay of selected mouse small intestine data with cells colored by coordinate value along the Y-axis at different stages of "unrolling": raw layer positioning, corrected Y (µm), and normalized corrected Y. (d) Marker gene expression visualized in the selected unrolled intestinal tissue (*Olfm4*, *Krt19*, *Fabp2*, and *Ada*). (e) Normalized corrected Y-axis distribution of cells expressing selected genes with red dashed lines indicating the approximate boundary between crypt and villus regions.



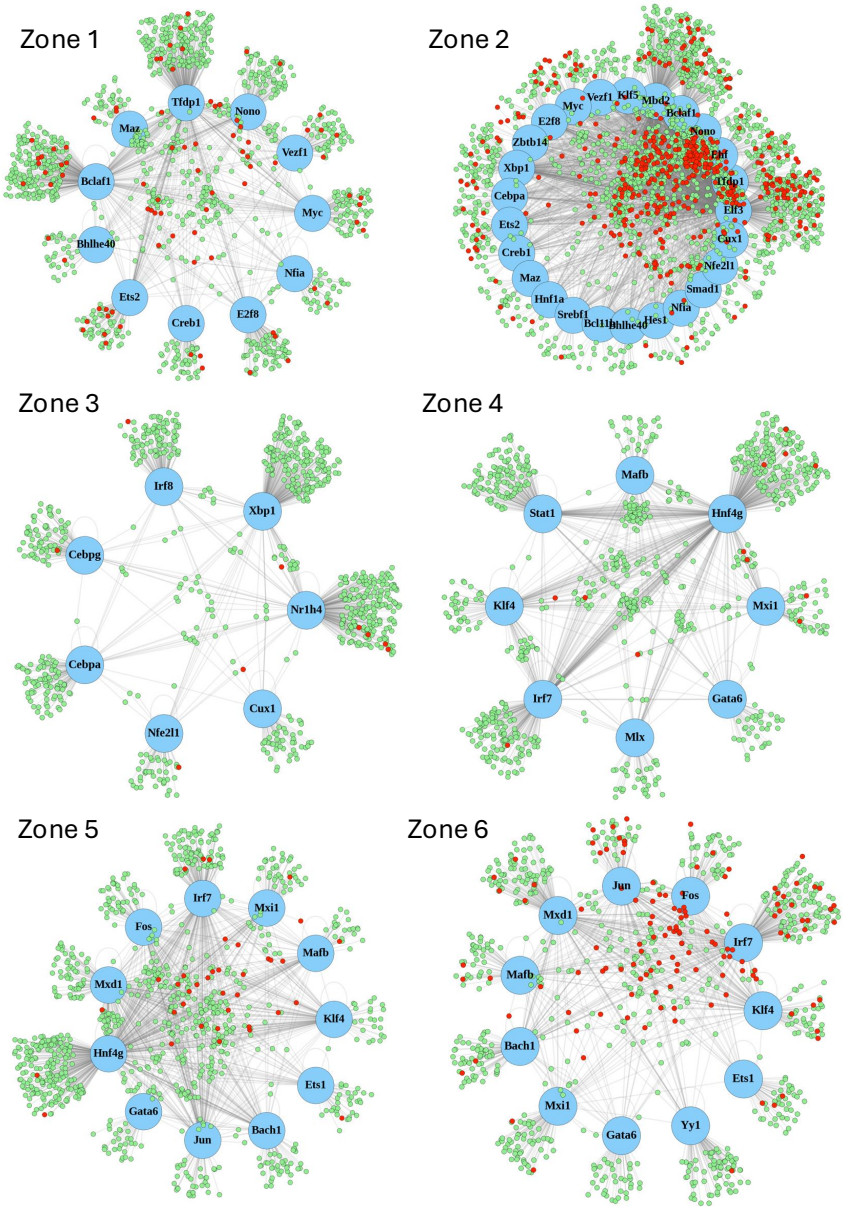
Extended Data Figure 4: Different Distances Examples Mapped on Real H&E Plot. (a) Layer. (b) Corrected Distance. (c) Normalized Corrected Distance. (d) Zone.



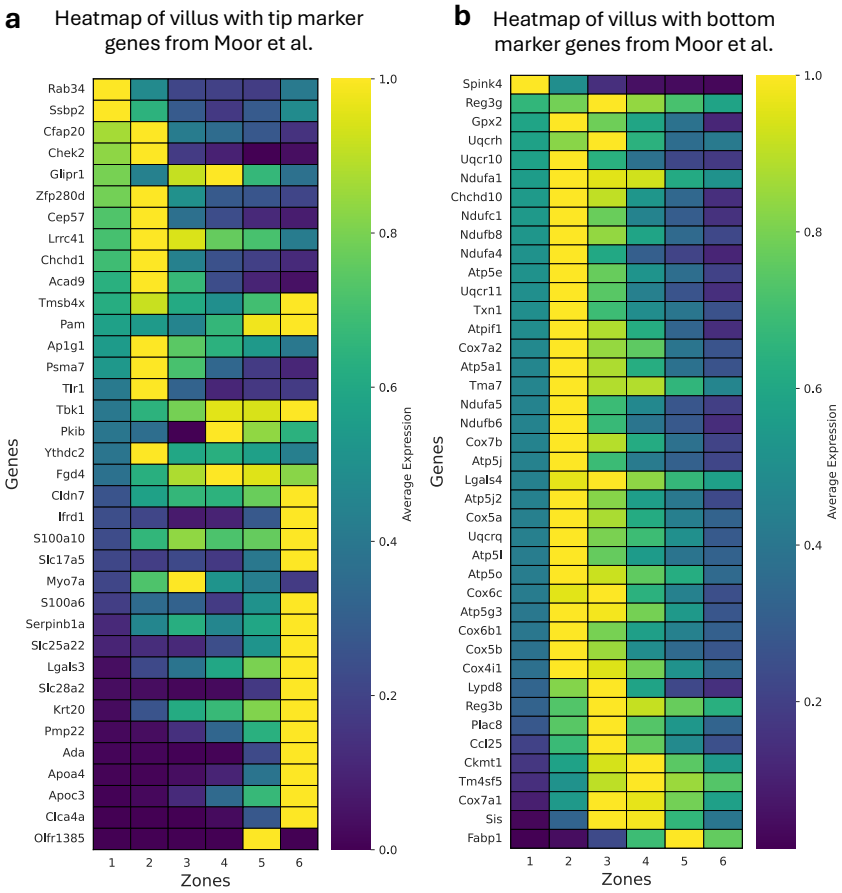
Extended Data Figure 5: **Cell type proportion in different zones.** (a) Cell type proportion of Paneth and Stem cells, TA cells, and Enterocytes. Each zone contains a similar number of cells in total. (b) Cell type proportion of six different cell types. Use the same zone definition of normalized y-axis cutoff in (a), add tuft cells, goblet cells, and EEC according to the normalized y-axis cutoff.



Extended Data Figure 6: **Highly zonated enterocyte genes.** Statistically enriched genes were identified in each zone. These lists were manually inspected to select the most illustrative examples of zonated genes. Heatmaps depict normalized gene expression for Zone 1 (a), Zone 2 (b), Zone 3 (c), Zone 4 (d), Zone 5 (e), and Zone 6 (f).



Extended Data Figure 7: **Transcription-factor (TF) interaction plots in each zone.** Regulons are shown only when the $\log_2$ fold change exceeds 0.3 and the adjusted p-value is below 1×10^{-10}.



Extended Data Figure 8: **Heatmap of villus tip and bottom marker genes from Moor et al. [12] in ileal SMURF dataset.** (a) Heatmap of villus with tip marker genes from Moor et al. (b) Heatmap of villus with bottom marker genes from Moor et al.

References

[1] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.

[2] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C Devlin, Maayan Baron, Cristina H Hajdu, Diane M Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature biotechnology*, 38(3):333–342, 2020.

[3] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[4] Sophia Maniatis, John Petrescu, Richard Rodriguez-Mias, et al. Visual spatial transcriptomics in tissue sections: Enhancing cellular resolution by coupling he or dapi imaging. *Nature Methods*, 16(10):1023–1032, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[6] 10x Genomics. Visium spatial gene expression, n.d. https://www.10xgenomics.com/products/spatial-gene-expression.

[7] 10x Genomics. Visium hd spatial gene expression, 2024. https://www.10xgenomics.com/products/spatial-gene-expression.

[8] Hao Chen, Dongshunyi Li, and Ziv Bar-Joseph. Scs: cell segmentation for high-resolution spatial transcriptomics. *Nature methods*, 20(8):1237–1243, 2023.

[9] Krzysztof Polański, Raquel Bartolomé-Casado, Ioannis Sarropoulos, Chuan Xu, Nick England, Frode L Jahnsen, Sarah A Teichmann, and Nadav Yayon. Bin2cell reconstructs cells from high resolution visium hd data. *Bioinformatics*, 40(9):btae546, 2024.

[10] Joseph D Dougherty, Simona Sarafinovska, Sneha M Chaturvedi, Travis E Law, Titilope M Akinwe, and Harrison W Gabel. Single-cell technology grows up: Leveraging high-resolution omics approaches to understand neurodevelopmental disorders. *Current Opinion in Neurobiology*, 92:102990, 2025.

[11] R.F. Hagemann, C.P. Sigdestad, and S. Lesher. A quantitative description of the intestinal epithelium of the mouse. *American Journal of Anatomy*, 129(1):41–52, 1970.

[12] Andreas E Moor, Yotam Harnik, Shani Ben-Moshe, Efi E Massasa, Milena Rozenberg, Raya Eilam, Keren Bahar Halpern, and Shalev Itzkovitz. Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. *Cell*, 175(4):1156–1167, 2018.

[13] Rita Manco, Inna Averbukh, Ziv Porat, Keren Bahar Halpern, Ido Amit, and Shalev Itzkovitz. Clump sequencing exposes the spatial expression programs of intestinal secretory cells. *Nature communications*, 12(1):3074, 2021.

[14] Yotam Harnik, Lisa Buchauer, Shani Ben-Moshe, Inna Averbukh, Yishai Levin, Alon Savidor, Raya Eilam, Andreas E Moor, and Shalev Itzkovitz. Spatial discordances between mrnas and proteins in the intestinal epithelium. *Nature metabolism*, 3(12):1680–1693, 2021.

[15] Toufic Mayassi, Chenhao Li, Åsa Segerstolpe, Eric M Brown, Rebecca Weisberg, Toru Nakata, Hiroshi Yano, Paula Herbst, David Artis, Daniel B Graham, et al. Spatially restricted immune and microbiota-driven adaptation of the gut. *Nature*, pages 1–10, 2024.

[16] Altay Yuzeir, David Alejandro Bejarano, Stephan Grein, Jan Hasenauer, Andreas Schlitzer, and Jiangyan Yu. Intestline: a shiny-based application to map the rolled intestinal tissue onto a line. *Bioinformatics*, 39(4):btad140, 2023.

[17] Sara M Parigi, Ludvig Larsson, Srustidhar Das, Ricardo O Ramirez Flores, Annika Frede, Kumar P Tripathi, Oscar E Diaz, Katja Selin, Rodrigo A Morales, Xinxin Luo, et al. The spatial transcriptomic landscape of the healing mouse intestine following damage. *Nature communications*, 13(1):828, 2022.

[18] Rachel K Zwick, Petr Kasparek, Brisa Palikuqi, Sara Viragova, Laura Weichselbaum, Christopher S McGinnis, Kara L McKinley, Asoka Rathnayake, Dedeepya Vaka, Vinh Nguyen, et al. Epithelial zonation along the mouse and human small intestine defines five discrete metabolic domains. *Nature Cell Biology*, 26(2):250–262, 2024.

[19] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

[20] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pages 265–273, 2018.

[21] Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, and Gene Myers. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[22] Martin Weigert and Uwe Schmidt. Nuclei instance segmentation and classification in histopathology images with stardist. In *The IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, 2022.

[23] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[24] Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023.

[25] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*,

9(1):1–12, 2019.

[26] John T Chamberlin, Younghee Lee, Gabor T Marth, and Aaron R Quinlan. Differences in molecular sampling and data processing explain variation among single-cell and single-nucleus rna-seq experiments. *Genome Research*, 34(2):179–188, 2024.

[27] Aaron F McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011.

[28] Lucie C Gaspard-Boulinc, Luca Gortana, Thomas Walter, Emmanuel Barillot, and Florence MG Cavalli. Cell-type deconvolution methods for spatial transcriptomics. *Nature Reviews Genetics*, pages 1–19, 2025.

[29] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[30] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022.

[31] 10x Genomics. Visium hd cytassist gene expression libraries of mouse brain he. https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-brain-he, 2024. Accessed: 2025-01-01.

[32] 10x Genomics. Visium hd cytassist gene expression libraries of mouse intestine. https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-intestine, 2024. Accessed: 2025-01-01.

[33] Teichlab. bin2cell. https://github.com/Teichlab/bin2cell, n.d. Accessed: September 14, 2024.

[34] Vicki M Houle, Wei Li, Robert K Montgomery, and Richard J Grand. mrna localization in polarized intestinal epithelial cells. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 284(4):G722–G727, 2003.

[35] Andreas E Moor, Matan Golan, Efi E Massasa, Doron Lemze, Tomer Weizman, Rom Shenhav, Shaked Baydatch, Orel Mizrahi, Roni Winkler, Ofra Golani, et al. Global mrna polarization regulates translation efficiency in the intestinal epithelium. *Science*, 357(6357):1299–1303, 2017.

[36] 10x Genomics. Fresh frozen mouse brain replicates: In situ gene expression dataset (3 replicates). https://www.10xgenomics.com/datasets/fresh-frozen-mouse-brain-replicates-1-standard, 2023. Analyzed using Xenium Onboard Analysis 1.0.2; dataset comprises 3 replicates; licensed under CC BY 4.0. Accessed: 16 February 2025.

[37] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[38] Lingling Xu, Abhijit Nirwane, and Yao Yao. Basement membrane and blood–brain barrier. *Stroke and vascular neurology*, 4(2), 2019.

[39] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[40] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience*, 19(2):335–346, 2016.

[41] Allen Institute for Brain Science. Mapmycells (rrid:scr_024672). https://portal.brain-map.org/atlases-and-data/bkp/mapmycells. Accessed: 16 February 2025.

[42] Zizhen Yao, Cindy TJ van Velthoven, Michael Kunst, Meng Zhang, Delissa McMillen, Changkyu Lee, Won Jung, Jeff Goldy, Aliya Abdelhak, Matthew Aitken, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature*, 624(7991):317–332, 2023.

[43] Pasko Rakic. Evolution of the neocortex: a perspective from developmental biology. *Nature Reviews Neuroscience*, 10(10):724–735, 2009.

[44] Maria Antonietta Tosches and Gilles Laurent. Evolution of neuronal identity in the cerebral cortex. *Current opinion in neurobiology*, 56:199–208, 2019.

[45] Antonio Pinto-Duarte, Terrence J Sejnowski, and Shyam Srinivasan. A conserved principle of glial organization in the paleocortex and neocortex. 2025.

[46] Riikka Pietilä, Francesca Del Gaudio, Liqun He, Elisa Vázquez-Liébanas, Michael Vanlandewijck, Lars Muhl, Giuseppe Mocci, Katrine D Bjørnholm, Caroline Lindblad, Alexander Fletcher-Sandersjöö, et al. Molecular anatomy of adult mouse leptomeninges. *Neuron*, 111(23):3745–3764, 2023.

[47] Matthew Bjerknes and Hazel Cheng. The stem-cell zone of the small intestinal epithelium. iii. evidence from columnar, enteroendocrine, and mucous cells in the adult mouse. *American Journal of Anatomy*, 160(1):77–91, 1981.

[48] Yoshitatsu Sei, Xinping Lu, Alice Liou, Xilin Zhao, and Stephen A Wank. A stem cell marker-expressing subset of enteroendocrine cells resides at the crypt base in the small intestine. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 300(2):G345–G356, 2011.

[49] Nicholas Barker, Johanna H van Es, Johannes Kuipers, Päivi Kujala, Marijke van den Born, M Cozijnsen, A Haegebarth, Joost Korving, H Begthel, Peter J Peters, et al. Identification of stem cells in the small intestine and colon by marker gene lgr5. *Nature*, 449(7165):1003–1007, 2007.

[50] Laleh Haghverdi, Florian Buettner, and Fabian J Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, 2015.

[51] Philipp Angerer, Laleh Haghverdi, Maren Büttner, Fabian J Theis, Carsten Marr, and Florian Buettner. destiny: diffusion maps for large-scale single-cell data in r. *Bioinformatics*, 32(8):1241–1243, 2016.

[52] Mireya Plass, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391):eaaq1723, 2018.

[53] Louise Deconinck, Robrecht Cannoodt, Wouter Saelens, Bart Deplancke, and Yvan Saeys. Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology*, 27:100344, 2021.

[54] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20:1–9, 2019.

[55] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, 21(7):1160–1167, 2011.

[56] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6):e98679, 2014.

[57] Daniel Pinto, Alex Gregorieff, Harry Begthel, and Hans Clevers. Canonical wnt signals are essential for homeostasis of the intestinal epithelium. *Genes & development*, 17(14):1709–1713, 2003.

[58] John A Barnard, R Daniel Beauchamp, Robert J Coffey, and Harold L Moses. Regulation of intestinal epithelial cell growth by transforming growth factor type beta. *Proceedings of the National Academy of Sciences*, 86(5):1578–1582, 1989.

[59] Lulu Huang, Jochem H Bernink, Amir Giladi, Daniel Krueger, Gijs JF van Son, Maarten H Geurts, Georg Busslinger, Lin Lin, Harry Begthel, Maurice Zandvliet, et al. Tuft cells act as regenerative stem cells in the human intestine. *Nature*, 634(8035):929–935, 2024.

[60] Michael R Howitt, Sydney Lavoie, Monia Michaud, Arthur M Blum, Sara V Tran, Joel V Weinstock, Carey Ann Gallini, Kevin Redding, Robert F Margolskee, Lisa C Osborne, et al. Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science*, 351(6279):1329–1333, 2016.

[61] Christoph Schneider, Claire E O'Leary, Jakob von Moltke, Hong-Erh Liang, Qi Yan Ang, Peter J Turnbaugh, Sridhar Radhakrishnan, Michael Pellizzon, Averil Ma, and Richard M Locksley. A metabolite-triggered tuft cell-ilc2 circuit drives small intestinal remodeling. *Cell*, 174(2):271–284, 2018.

[62] Julie Bas, Philippe Jay, and François Gerbe. Intestinal tuft cells: Sentinels, what else? In *Seminars in cell & developmental biology*, volume 150, pages 35–42. Elsevier, 2023.

[63] Keven A Roth and Jeffrey I Gordon. Spatial differentiation of the intestinal epithelium: analysis of enteroendocrine cells containing immunoreactive serotonin, secretin, and substance p in normal and transgenic mice. *Proceedings of the National Academy of Sciences*, 87(16):6408–6412, 1990.

[64] Joep Beumer, Benedetta Artegiani, Yorick Post, Frank Reimann, Fiona Gribble, Thuc Nghi Nguyen, Hongkui Zeng, Maaike Van den Born, Johan H Van Es, and Hans Clevers. Enteroendocrine cells switch hormone

expression along the crypt-to-villus bmp signalling gradient. *Nature cell biology*, 20(8):909–916, 2018.

[65] Yannick D Benoit, Fréderic Paré, Caroline Francoeur, Dominique Jean, Eric Tremblay, François Boudreau, Fabrice Escaffit, and Jean-François Beaulieu. Cooperation between hnf-1α, cdx2, and gata-4 in initiating an enterocytic differentiation program in a normal human intestinal epithelial progenitor cell line. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 298(4):G504–G517, 2010.

[66] Adrianna K San Roman, Boaz E Aronson, Stephen D Krasinski, Ramesh A Shivdasani, and Michael P Verzi. Transcription factors gata4 and hnf4a control distinct aspects of intestinal homeostasis in conjunction with transcription factor cdx2. *Journal of Biological Chemistry*, 290(3):1850–1860, 2015.

[67] Sara Aibar, Carlos B. González-Blas, Thomas Moerman, Vân A. Huynh-Thu, Hana Imrichova, Geert Hulselmans, Florian Rambow, José C. Marine, Pierre Geurts, and Stein Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, 2017.

[68] Philippe Blache, Marc Van De Wetering, Isabelle Duluc, Claire Domon, Philippe Berta, Jean-Noël Freund, Hans Clevers, and Philippe Jay. 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, 2004.

[69] Eric J Formeister, Ayn L Sionas, David K Lorance, Carey L Barkley, Ginny H Lee, and Scott T Magness. Distinct sox9 levels differentially mark stem/progenitor populations and enteroendocrine cells of the small intestine epithelium. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 296(5):G1108–G1118, 2009.

[70] Emily M Walker, Cayla A Thompson, and Michele A Battle. Gata4 and gata6 regulate intestinal epithelial cytodifferentiation during development. *Developmental biology*, 392(2):283–294, 2014.

[71] Ian Y Luk, Camilla M Reehorst, and John M Mariadason. Elf3, elf5, ehf and spdef transcription factors in tissue homeostasis and cancer. *Molecules*, 23(9):2191, 2018.

[72] JL Wang, ZF Chen, HM Chen, MY Wang, X Kong, YC Wang, TT Sun, J Hong, W Zou, J Xu, et al. Elf3 drives β-catenin transactivation and associates with poor prognosis in colorectal cancer. *Cell death & disease*, 5(5):e1263–e1263, 2014.

[73] Kristin N Bell and Noah F Shroyer. Krüpple-like factor 5 is required for proper maintenance of adult intestinal crypt cellular proliferation. *Digestive diseases and sciences*, 60:86–100, 2015.

[74] Chang-Kyung Kim, Madhurima Saxena, Kasmika Maharjan, Jane J Song, Kenneth R Shroyer, Agnieszka B Bialkowska, Ramesh A Shivdasani, and Vincent W Yang. Krüppel-like factor 5 regulates stemness, lineage specification, and regeneration of intestinal epithelial stem cells. *Cellular and Molecular Gastroenterology and Hepatology*, 9(4):587–609, 2020.

[75] Keren Bahar Halpern, Hassan Massalha, Rachel K Zwick, Andreas E Moor, David Castillo-Azofeifa, Milena Rozenberg, Lydia Farack, Adi Egozi, Dan R Miller, Inna Averbukh, et al. Lgr5+ telocytes are a signaling source at the intestinal villus tip. *Nature communications*, 11(1):1936, 2020.

[76] Joep Beumer, Jens Puschhof, Fjodor Yousef Yengej, Lianzheng Zhao, Adriana Martinez-Silgado, Marloes Blotenburg, Harry Begthel, Charelle Boot, Alexander van Oudenaarden, Ye-Guang Chen, et al. Bmp gradient along the intestinal villus axis controls zonated enterocyte and goblet cell states. *Cell Reports*, 38(9), 2022.

[77] Adrianna K San Roman and Ramesh A Shivdasani. Boundaries, junctions and transitions in the gastrointestinal tract. *Experimental cell research*, 317(19):2711–2718, 2011.

[78] Cayla A Thompson, Ann DeLaForest, and Michele A Battle. Patterning the gastrointestinal epithelium to confer regional-specific functions. *Developmental biology*, 435(2):97–108, 2018.

[79] Alyssa S Zembroski, Theresa D'Aquila, and Kimberly K Buhman. Characterization of cytoplasmic lipid droplets in each region of the small intestine of lean and diet-induced obese mice in response to dietary fat. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 321(1):G75–G86, 2021.

[80] Catriona MH Anderson, Vadivel Ganapathy, and David T Thwaites. Human solute carrier slc6a14 is the β-alanine carrier. *The Journal of physiology*, 586(17):4061–4067, 2008.

[81] Sabine Middendorp, Kerstin Schneeberger, Caroline L Wiegerinck, Michal Mokry, Ronald DL Akkerman, Simone Wijngaarden, Hans Clevers, and Edward ES Nieuwenhuis. Adult stem cells in the small intestine are intrinsically programmed with their location-specific function. *Stem cells*, 32(5):1083–1091, 2014.

[82] Ozge Kayisoglu, Franziska Weiss, Carolin Niklas, Isabella Pierotti, Malvika Pompaiah, Nina Wallaschek, Christoph-Thomas Germer, Armin Wiegering, and Sina Bartfeld. Location-specific cell identity rather than exposure to gi microbiota defines many innate immune signalling cascades in the gut epithelium. *Gut*, 70(4):687–697, 2021.

[83] Kristen M Seiler, Sarah E Waye, Wenjun Kong, Kenji Kamimoto, Adam Bajinting, William H Goo, Emily J Onufer, Cathleen Courtney, Jun Guo, Brad W Warner, et al. Single-cell analysis reveals regional reprogramming during adaptation to massive small bowel resection in mice. *Cellular and molecular gastroenterology and hepatology*, 8(3):407–426, 2019.

[84] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[85] Carsen Stringer and Marius Pachitariu. Cellpose3: one-click image restoration for improved cellular segmentation. *Nature Methods*, pages 1–8, 2025.

[86] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[87] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, Isaac Virshup, et al. Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, 19(2):171–178, 2022.

[88] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

[89] Cynthia Arroyo Portilla, Romain Fenouil, Camille Wagner, Cécilia Luciani, Margaux Lagier, Clément Da Silva, Fanny Hidalgo-Villeda, Lionel Spinelli, Mathieu Fallet, Julie Tomas, et al. Peyer's patch phagocytes acquire specific transcriptional programs that influence their maturation and activation profiles. *Mucosal Immunology*, 16(4):527–547, 2023.

[90] Eugene Drokhlyansky, Christopher S Smillie, Nicholas Van Wittenberghe, Maria Ericsson, Gabriel K Griffin, Gokcen Eraslan, Danielle Dionne, Michael S Cuoco, Max N Goder-Reiser, Tatyana Sharova, et al. The human and mouse enteric nervous system at single-cell resolution. *Cell*, 182(6):1606–1622, 2020.

[91] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[92] Bhargav Chippada. Forceatlas2 for python. https://github.com/bhargavchippada/forceatlas2, 2018. GitHub repository.

[93] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198, 2019.

[94] 10x Genomics. Segmentation for visium hd. https://www.10xgenomics.com/analysis-guides/segmentation-visium-hd. Analysis Guide. Accessed: 16 February 2025.