Identifying spatially variable genes by projecting to morphologically relevant curves

Phillip B. Nicol[1,2], Rong Ma[1,2], Rosalind J. Xu[3,4],
Jeffrey R. Moffitt[3,4,5], and Rafael A. Irizarry[1,2,*]

[1] *Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA*

[2] *Department of Data Science, Dana Farber Cancer Institute, Boston, MA 02215, USA*

[3] *Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston MA 02115, USA*

[4] *Department of Microbiology, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, USA*

[5] *Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA*

[*] Correspondence: rafael_irizarry@dfci.harvard.edu

November 21, 2024

Abstract

Spatial transcriptomics enables high-resolution gene expression measurements while preserving the two-dimensional spatial organization of the sample. A common objective in spatial transcriptomics data analysis is to identify spatially variable genes within predefined cell types or regions within the tissue. However, these regions are often implicitly one-dimensional, making standard two-dimensional coordinate-based methods less effective as they overlook the underlying tissue organization. Here we introduce a methodology grounded in spectral graph theory to elucidate a one-dimensional curve that effectively approximates the spatial coordinates of the examined sample. This curve is then used to establish a new coordinate system that better reflects tissue morphology. We then develop a generalized additive model (GAM) to detect genes with variable expression in the new *morphologically relevant* coordinate system. Our approach directly models gene counts, thereby eliminating the need for normalization or transformations to satisfy normality assumptions. We demonstrate improved performance relative to current methods based on hypothesis testing, while also accurately estimating gene expression patterns and precisely identifying spatial loci where deviations from constant expression are observed. We validate our approach through extensive simulations and by analyzing experimental data from multiple platforms

such as Slide-seq and MERFISH.

1 Introduction

Spatial transcriptomics (ST) technologies permit high-resolution measurement of gene expression while maintaining the spatial coordinates of the samples (Rao et al., 2021; Moses and Pachter, 2022). These technologies have the potential to improve our understanding of the influence of cellular spatial organization on important biological processes and disease. One of the starting points for ST analysis is the identification of spatially variable genes (SVGs) (Adhikari et al., 2024). Since spatial variability can often be explained by differences in cell type (Cable et al., 2022b), it is common to test for SVGs within a predefined cell type or spatial domain (Yu and Luo, 2022).

Current statistical approaches for SVG detection perform a hypothesis test for each gene, quantifying the evidence of spatial variability using a p-value (Svensson et al., 2018; Sun et al., 2020; Hao et al., 2021; Zhu et al., 2021; Weber et al., 2023). However, these methodologies are incapable of distinguishing genes whose spatial expression patterns manifest in fundamentally different ways, such as along distinct anatomical features within the tissue. For example, MERFISH measurements of a healthy mouse colon revealed two dominant patterns of spatial variability, denoted here as *localized* (**Fig 1a, left**) and *radial gradient* (**Fig 1a, right**), respectively. Genes exhibiting localized variation, such as *Ddx58*, are characterized by a distinct patch of expression in one region of the colon, whereas genes exhibiting radial gradient variation, such as *Apob*, are characterized by a gradual change in expression between the outside and inside of the mucosa. Importantly, while both of these examples are illustrations of spatially variable genes, their distinct spatial distributions have important implications for their biological interpretation. *Ddx58* is an interferon-stimulated gene, and the localized expression observed in the mucosa is representative of local patches of interferon activity and interferon-stimulated gene expression described previously (Van Winkle et al., 2022). By contrast, the radial distribution of *Apob*–a marker of the final stages of mature enterocytes–shows the known variation of epithelial cells from the base to the tip of colonic crypts (Moor et al., 2018). Although current leading approaches succesfully identified these genes as spatially variable (**Fig S1, S2**), they lack the capability of distinguishing between these two modes of spatial variation. Additionally, these existing methodologies do not allow for precise pinpointing of the locations where spatially relevant gene expression occurs.

Although numerous statistical techniques exist for the estimation of two-dimensional surfaces (Wood, 2003; Schulz et al., 2018), and some of these used for ST (Cable et al., 2022a), it is noteworthy that in many applications the primary interest is in genes that exhibit variation along one-dimensional paths. For example, the exercise of visually detecting the localized pattern shown in *Ddx58* could be described as searching for

deviation from a baseline expression level along a curve tracing through the mucosa. The radial gradient pattern exhibited by *Apob* could be described as change in expression in the direction perpendicular to the curve. This implies that a curve-based coordinate transform could help separate genes with a localized burst from those with radial gradient, which, in turn, could facilitate new biological findings. In addition to the colon (**Fig 1a**), cell types in the brain also commonly exhibit distinct one-dimensional spatial structure. In this paper we also consider two Slide-seq datasets: granule cells from the mouse cerebellum (Cable et al., 2022b) (**Fig 1b**) and CA3 cells from the mouse hippocampus (Stickels et al., 2021) (**Fig 1c**), although numerous additional examples exist.

In this paper, we introduce a statistical framework that estimates a one-dimensional curve passing through the ST coordinates and then uses that estimated curve to define a *morphologically relevant* coordinate system. Although similar curve-estimation methods have been used for pseudotime analysis in single-cell RNA-seq (Street et al., 2018), we find that our methodology grounded in spectral graph theory yields better results on two-dimensional ST data. Moreover, pseudotime methods do not measure variation *orthogonal* to the curve which is critical to distinguish between localized and radial gradient patterns.

Upon estimating the curve, we employ a generalized additive model (GAM) to model expression as a (possibly non-linear) function of the morphologically relevant coordinates. We refer to our approach as *MorphoGAM*. Unlike previously published hypothesis tests for SVGs, MorphoGAM identifies the exact location and mode of the pertinent expression pattern, thereby resulting in more interpretable findings. An additional advantage of summarizing the two-dimensional coordinates using one-dimensional projections is increased statistical power to detect relevant SVGs due to reduced complexity of model fit to estimate spatial effects.

2 Results

MorphoGAM identifies morphologically relevant coordinates in spatial transcriptomics data. We begin by modeling the spatial location of cell j, $1 \leq j \leq n$, using *morphologically relevant* coordinates t_j and r_j. Specifically, we assume that the standard two-dimensional spatial coordinates $x_j \in \mathbb{R}^2$ lie close to a latent curve:

$$\mathbb{E}(x_j) = f(t_j) \tag{2.1}$$

where $f : [a, b] \to \mathbb{R}^2$ is a smooth one-dimensional parametric curve. We write $f(t) = (f_1(t), f_2(t))$ to denote the two component functions of the curve. The first morphologically relevant coordinate t_j describes the position of cell j along the curve. In the Methods we describe in detail our approach based on spectral

graph theory to estimate t_j and f. Briefly, our approach relates the distance between coordinates $|t_i - t_j|$ to the shortest path in a k-nearest neighbor graph G_k and then shows that t_j can be estimated through an eigendecomposition of a centered shortest path matrix. Upon obtaining an estimate $\hat{t}_j$, the curve f can be estimated by smoothing each dimension separately. We plug in $\hat{t}_j$ to (2.1) to obtain

$$\mathbb{E}(x_{j1}) = f_1(\hat{t}_j) \tag{2.2}$$

$$\mathbb{E}(x_{j2}) = f_2(\hat{t}_j) \tag{2.3}$$

We thus obtain $\hat{f}_1$ and $\hat{f}_2$ by using regression splines as implemented in *mgcv* (Wood, 2017).

Our methodology to estimate t_j in the case of a linear curve (i.e., when $f(a) \neq f(b)$) is motivated by the ISOMAP (Tenenbaum et al., 2000) technique for non-linear dimensionality reduction. We extend this approach to allow our method to address scenarios wherein f constitutes a closed curve (i.e., when $f(a) = f(b)$).

A detailed visual assessment indicates that, when applied to the slice of healthy mouse colon, MorphoGAM excels in estimating $f(t)$ (**Fig 2a**) and the morphologically relevant coordinate t_j (**Fig 2b**).

The second *morphologically relevant* coordinate, denoted here by r_j, is defined by using the distance from the cell's coordinates to the position on the curve $f(t)$. Explicitly, the magnitude of r_j is given by

$$|r_j| = ||x_j - \hat{f}(\hat{x}_j)||_2. \tag{2.4}$$

To determine the sign of r_j, we set

$$\text{sign}(\hat{r}_j) = \text{sign}\left[\langle x_j - \hat{f}(\hat{t}_j), R\hat{f}'(\hat{t}_j)\rangle\right] \tag{2.5}$$

where $R : \mathbb{R}^2 \to \mathbb{R}^2$ is a counter-clockwise 90 degrees rotation: $R(v_1, v_2) = (-v_2, v_1)$. The conceptual framework behind equation (2.5) can be understood by envisioning a traversal along the curve, where the velocity vector at time t is $f'(t)$. Residuals exhibiting a positive sign would be placed on the left-hand side of the curve as one progresses, whereas those with a negative sign would be on the right-hand side. The left-hand side can be ascertained through a counter-clockwise rotation R of the velocity vector $f'(t)$. This coordinate for each cell is also morphologically pertinent, as illustrated in Figure (**Fig 2c**).

After transforming to the morphologically relevant coordinate system, the difference between localized and radial gradient patterns becomes immediately clear. SVGs with localized patterns show variation in the first morphologically relevant coordinate t_j (**Fig 2d**) whereas SVGs exhibiting a radial gradient pattern

show variation in the second morphologically relevant coordinate r_j (**Fig 2e**).

MorphoGAM outperforms existing curve estimation approaches. Hastie and Stuetzle (1989) introduced model (2.1) as a general approach to estimate a curve passing through a set of points (in arbitrary dimension). This method, known as *Principal Curves*, is used by the popular pseudomtime method *Slingshot* (Street et al., 2018). Hastie and Stuetzle (1989) employ an iterative algorithm that alternates between updating f and updating t_j. However, we find that this iterative approach is unsuitable for the highly non-linear structures observed in spatial transcriptomics data. To demonstrate this, we applied *Principal Curves* to granule cells from the mouse cerebellum, measured using Slide-seqV2 (Cable et al., 2022b) (see Figure 1b). We found that this approach did not accurately estimate the curve for a variety of tuning parameter choices (**Fig 3a-c**). In contrast, if we set the number of nearest neighbors k to 5, MorphoGAM accurately identified the path and the first morphologically relevant coordinate (**Fig 3d, Fig S3**). The performance is similar for $k = 10$ (**Fig 3e**) and only begins to degrade once $k = 30$ (**Fig 3f**).

To quantitatively evaluate the robustness of these methods with respect to their tuning parameters, we defined *ground truth* t_j by carefully hand-drawing a path (**Fig S4**) and compared estimates $\hat{t}_j$ obtained with different values of the tuning parameter to the ground truth t_j. We observed that the spearman correlation between the estimates $\hat{t}_j(\text{df})$ and t_j was below 0.6 for Principal Curves (**Fig 3g**). In contrast, with MorphoGAM, the correlation exceeded 0.95 for $k < 10$ and remained above 0.9 for values up to $k = 30$ (**Fig 3h**). We also applied both methods to simulated swiss roll data (**Fig S5**) and the mouse colon dataset (**Fig S6**) and found similar performance gains from using MorphoGAM. An additional advantage of MorphoGAM is that the tuning parameter k (number of nearest neighbors) is more interpretable than degrees of freedom (df) for a smoothing spline, which makes it easier to find a reasonable value in practice.

MorphoGAM allows for interpretable detection of spatially variable genes. Following the estimation of morphologically relevant coordinates $\hat{t}_j$ and $\hat{r}_j$, MorphoGAM identifies spatially variable genes using a generalized generalized additive model (GAM) (Hastie and Tibshirani, 1986). Specifically, we denote the count for gene g in cell j with Y_{gj} and model it with

$$\begin{aligned} Y_{gj} &\sim \text{NegBinom}(n_j \mu_{gj}, \theta_g) \\ \log \mu_{gj} &= \beta_{g0} + h_g(\hat{t}_j) + s_g(\hat{r}_j) \end{aligned} \tag{2.6}$$

where h_g and s_g are unknown smooth functions, β_{g0} is an unknown intercept, and n_j is the total counts for cell j. θ_g is the inverse dispersion parameter because $\text{Var}(Y_{gj}) = n_j \mu_{gj} + (n_j \mu_{gj})^2/\theta_g$. In this model, the gene g is spatially variable if $h_g \neq 0$ or $s_g \neq 0$. Estimating the parameters of model (2.6) is achieved

by writing h_g and s_g as the sum of basis functions and then adding a penalty to encourage smoothness (Methods). Although p-values can be computed by testing the null hypothesis $h_g = 0$ or $s_g = 0$, we recommend inspecting the estimated functions $\hat{h}_g$ and $\hat{s}_g$ along with estimates of their covariance (to measure uncertainty). We specifically use *adaptive shrinkage* (Stephens, 2017) to further regularize the functions with higher uncertainty (see Methods).

Although we can examine the entire function estimate, we also introduce two summaries useful for ranking genes automatically. Specifically, we consider the *peak*

$$\hat{P}_g := \sup_t \hat{h}_g(t) \tag{2.7}$$

which estimates the maximum log-fold change from the baseline log expression $\hat{\beta}_{g0}$. Because this measurement could prioritize large multiplicative changes in small genes we also define the *range*

$$\hat{R}_g := \sup_t \left[n_{\text{med}} \exp(\hat{\beta}_{g0} + \hat{h}_g(t)) \right] - \inf_t \left[n_{\text{med}} \exp(\hat{\beta}_{g0} + \hat{h}_g(t)) \right] \tag{2.8}$$

to account for genes that have large differences on the original scale of the counts. Here n_{med} is defined as the median of the n_j, so that $\hat{R}_g$ can be directly interpreted as a count difference. We note $\hat{s}_g(t)$ could replace $\hat{h}_g(t)$ in both equations (2.7) and (2.8).

We emphasize that the model (2.6) can easily be modified depending on the particular scientific question. For example, if only variation along the curve is of interest then we only need to examine $\hat{h}_g(t)$. Moreover, the model is flexible enough to account for other potential confounders in the linear predictor.

MorphoGAM improves power to detect relevant spatially variable genes. We applied MorphoGAM to the CA3 mouse hippocampus cells (**Fig 1c**) to estimate a one-dimensional curve $\hat{f}$ and morphologically relevant coordinates $\hat{t}_j$ (**Fig 4a**). In the original analysis of this dataset, Cable et al. (2022b) used 2D locally weighted regression to identify genes with a high coefficient of variation (CV). This analysis identified two genes *Rgs14* and *Cpne9* that exhibited variable expression at different ends. We applied the GAM model (2.6) with s_g removed to identify genes varying along the curve $\hat{f}$. Our approach corroborated the finding of *Rgs14* ($\hat{P}_g = 2.77$, $p < 10^{-16}$) and *Cpne9* ($\hat{P}_g = 1.41$, $p < 10^{-16}$) (**Fig 4b**).

We hypothesized that MorphoGAM increases statistical power to detect SVGs by projecting the two-dimensional ST coordinates to a one-dimensional morphologically relevant coordinate. To demonstrate this, we simulated a gene such that $\mu_{gj} = 1 + \kappa \exp\left(-\sigma(\hat{t}_j - 0.5)^2\right)$, where $\hat{t}_j$ is as above. In order to compare to the approaches based on hypothesis testing, we labeled a gene as spatially variable if the p-value was below the transcriptome wide significance level of 0.05/20000. MorphoGAM had consistently higher power than

two state-of-the-art methods for detecting SVGs, nnSVG (Weber et al., 2023) and SPARK-X (Zhu et al., 2021) (**Fig 4c**). To show that the increase in power did not come at the price of an inflated type I error rate, we randomly permuted all spatial locations (to generate a null dataset with no spatially variable genes) and found that our method was conservative at a variety of significance levels (**Fig S7**). When ranking by genes with a large peak or range, our approach identified genes that were not reported in the original analysis of Cable et al. (2022b) such as *Fxyd6* ($\hat{P}_g = 3.28$, $p < 10^{-16}$) and *Hpca* ($\hat{R}_g = 12.67$, $p < 10^{-16}$) (**Fig 4d**).

MorphoGAM identifies spatially variable genes in the mouse colon data. We applied MorphoGAM to the MERFISH measurements of a slice of healthy mouse colon (**Fig 1**) with the goal of separating genes with localized and radial gradient patterns of expression. Because localized genes are characterized by a burst in expression along the curve, we used the peak of estimated functions $\hat{h}_g(\hat{t}_j)$ to rank genes (**Fig 5a**). Radial gradient genes, on the other hand, are characterized by a smooth transition along the second morphologically relevant coordinate, so for this we found genes with a large range in $\hat{s}_g$ (**Fig 5b**). **Figure 5** also lists the ranking of each gene of both nnSVG and SPARK-X, showing that the targeted analysis of MorphoGAM prioritizes genes that could have been missed if only hypothesis-based tests were used for SVGs. In particular, *Ddx58* was found to have a large peak in the first morphologically relevant coordinate ($\hat{P}_g = 2.01$, $p < 10^{-16}$) and *Apob* was found to have a large peak in the direction of the second morphologically relevant coordinate ($\hat{P}_g = 1.24$, $p < 10^{-16}$). We also plot the genes with the largest range in the direction of the first morphologically relevant coordinate and the genes with the largest peak in the direction of the second morphologically relevant coordinate in **Figure S8**.

3 Discussion

We introduced an approach to estimate the curve passing through spatial transcriptomics coordinates and leveraged this curve to define *morphologically relevant* coordinates. A GAM is used to model spatial variation along these morphologically relevant coordinates, which we have shown to be an interpretable and powerful approach to find relevant spatially variable genes. Importantly, we have advocated to directly use summaries of the estimated functions rather than relying on a null hypothesis test, as p-values do not provide information about the mode of spatial variation and are in general sensitive to misspecification in the assumed model (Greenland et al., 2016).

The proposed methodology presents certain limitations. First, the final results depend on the accurate annotation of cell types or spatial domains, and inaccuracies at this stage may propagate to MorphoGAM. Furthermore, the approach is not inherently applicable in scenarios where the tissue structure cannot be adequately represented by a one-dimensional framework. As a result, as part of the software package sup-

porting the implementation of MorphoGAM, we have developed a tool allowing manual curve drawing $f(t)$, as shown in **Figure S4**. This facilitates the application of MorphoGAM, although manually, in instances where variation along a predetermined trajectory is to be identified. More broadly, our GAM methodology, accompanied by summaries of estimated functions (such as range and peak), can be easily extended to the two-dimensional domain by employing thin-plate splines (Wood, 2003).

Morphologically relevant coordinates may offer considerable utility beyond the scope of spatially variable genes. For instance, the alignment of multiple spatial transcriptomics (ST) slices may be enhanced by leveraging these morphologically relevant coordinates instead of conventional two-dimensional coordinates. Future research could profitably explore the application of morphologically pertinent coordinates in conducting multi-sample ST analyses.

References

S. D. Adhikari, J. Yang, J. Wang, and Y. Cui. Recent advances in spatially variable gene detection in spatial transcriptomics. *Computational and Structural Biotechnology Journal*, 2024.

D. M. Cable, E. Murray, V. Shanmugam, S. Zhang, L. S. Zou, M. Diao, H. Chen, E. Z. Macosko, R. A. Irizarry, and F. Chen. Cell type-specific inference of differential expression in spatial transcriptomics. *Nature methods*, 19(9):1076–1087, 2022a.

D. M. Cable, E. Murray, L. S. Zou, A. Goeva, E. Z. Macosko, F. Chen, and R. A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature biotechnology*, 40(4):517–526, 2022b.

P. Cadinu, K. N. Sivanathan, A. Misra, R. J. Xu, D. Mangani, E. Yang, J. M. Rone, K. Tooley, Y.-C. Kye, L. Bod, et al. Charting the cellular biogeography in colitis reveals fibroblast trajectories and coordinated spatial remodeling. *Cell*, 187(8):2010–2028, 2024.

R. Cannoodt. princurve 2.0: Fit a principal curve in arbitrary dimension, Jun 2018. URL https://doi.org/10.5281/zenodo.3351282.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3): 211–218, 1936.

B. Ghojogh, M. Crowley, F. Karray, and A. Ghodsi. Multidimensional scaling, sammon mapping, and isomap. In *Elements of Dimensionality Reduction and Manifold Learning*, pages 185–205. Springer, 2023.

S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350, 2016.

M. Hao, K. Hua, and X. Zhang. Somde: a scalable method for identifying spatially variable genes with self-organizing map. *Bioinformatics*, 37(23):4392–4398, 2021.

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American statistical association*, 84(406):502–516, 1989.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986.

B. Kégl, A. Krzyzak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE transactions on pattern analysis and machine intelligence*, 22(3):281–297, 2000.

A. Lastra. *Parametric Geometry of Curves and Surfaces*. Springer, 2021.

A. E. Moor, Y. Harnik, S. Ben-Moshe, E. E. Massasa, M. Rozenberg, R. Eilam, K. B. Halpern, and S. Itzkovitz. Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. *Cell*, 175(4):1156–1167, 2018.

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

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

D. Righelli, L. M. Weber, H. L. Crowell, B. Pardo, L. Collado-Torres, S. Ghazanfar, A. T. Lun, S. C. Hicks, and D. Risso. Spatialexperiment: infrastructure for spatially-resolved transcriptomics data in r using bioconductor. *Bioinformatics*, 38(11):3128–3131, 2022.

E. Schulz, M. Speekenbrink, and A. Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of mathematical psychology*, 85:1–16, 2018.

M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

R. R. Stickels, E. Murray, P. Kumar, J. Li, J. L. Marshall, D. J. Di Bella, P. Arlotta, E. Z. Macosko, and F. Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319, 2021.

K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19:1–16, 2018.

S. Sun, J. Zhu, and X. Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature methods*, 17(2):193–200, 2020.

V. Svensson, S. A. Teichmann, and O. Stegle. Spatialde: identification of spatially variable genes. *Nature methods*, 15(5):343–346, 2018.

J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

J. A. Van Winkle, S. T. Peterson, E. A. Kennedy, M. J. Wheadon, H. Ingle, C. Desai, R. Rodgers, D. A. Constant, A. P. Wright, L. Li, et al. Homeostatic interferon-lambda response to bacterial microbiota stimulates preemptive antiviral defense within discrete pockets of intestinal epithelium. *elife*, 11:e74072, 2022.

L. M. Weber, A. Saha, A. Datta, K. D. Hansen, and S. C. Hicks. nnsvg for the scalable identification of spatially variable genes using nearest-neighbor gaussian processes. *Nature communications*, 14(1):4059, 2023.

S. N. Wood. mgcv: Gams and generalized ridge regression for r. *R news*, 1(2):20–25, 2001.

S. N. Wood. Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1):95–114, 2003.

S. N. Wood. *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2017.

J. Yu and X. Luo. Identification of cell-type-specific spatially variable genes accounting for excess zeros. *Bioinformatics*, 38(17):4135–4144, 2022.

J. Zhu, S. Sun, and X. Zhou. Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome biology*, 22(1):184, 2021.

4 Methods

Statistical model for latent curve. Let $x_j \in \mathbb{R}^2$ denote the spatial coordinates of cell $1 \leq j \leq n$. We assume that

$$\mathbb{E}(x_j) = f(t_j) \tag{4.1}$$

where $f : [a, b] \to \mathbb{R}^2$ is a smooth parametric curve. We will write $f(t) = (f_1(t), f_2(t))$ to denote the two component functions of the curve. For the moment, we assume that f does not intersect itself, so that $t_i \neq t_j$ implies $f(t_i) \neq f(t_j)$. Note that both f and t_j are unknown and must be estimated. See Section S1.2 for detailed discussions on the identifiability conditions for f and t_j.

The arclength of f can be expressed in terms of the first derivative $f'(t) := (f_1'(t), f_2'(t))$:

$$\int_a^b ||f'(t)||_2 dt \tag{4.2}$$

We will assume that f has a *unit-speed parametrization*, which means that $||f'(t)||_2 = 1$ for all t. Any parametric curve such that $f'(t) \neq 0$ for all t can be reparametrized to satisfy this requirement (Section S1.2). In particular, this means the arc-length between two points on the curve is equal to the difference in their coordinates:

$$t_i - t_j = \int_{t_i}^{t_j} ||f'(t)||_2 dt \quad (t_j < t_i) \tag{4.3}$$

First morphologically relevant coordinate. Our approach to estimate t_j in the case of a non-intersecting curve (i.e., $x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$) leverages the relationship in (4.3). Interestingly, it turns out that the estimated $\hat{t}_j$ is the same as the first component produced by the ISOMAP algorithm for manifold learning (Tenenbaum et al., 2000). Based on equation (4.3), we can estimate the arclength between t_i and t_j using shortest paths in a k-nearest neighbor (KNN) graph G_k. If k is chosen sufficiently small, the shortest path between two vertices (cells) will have a similar shape as f. To make this precise, we define $d_{G_k}(i, j)$ to be the shortest path between x_i and x_j in G_k, so that

$$d_{G_k}(i, j) \approx |t_i - t_j| \tag{4.4}$$

Our estimate $\hat{t}_j$ will be chosen to satisfy the approximation in (4.4). To construct this, we follow the steps of classical multidimensional scaling (cMDS) (Torgerson, 1952). Squaring both sides of (4.4) yields

$$d^2_{G_k}(i, j) \approx t_i^2 + t_j^2 - 2t_i t_j \tag{4.5}$$

Viewing $d^2_{G_k}$ as an $n \times n$ matrix, the operation of double centering (Lemma S1.1) yields $b_{G_k} \in \mathbb{R}^{n \times n}$ such that

$$-\frac{1}{2} b_{G_k}(i, j) \approx t_i t_j \tag{4.6}$$

Given this, we set

$$\hat{t} = \text{argmin}_{t \in \mathbb{R}^n} || - \frac{1}{2} b_{G_k} - t \otimes t||_F^2 \tag{4.7}$$

where $\otimes$ is the outer product defined in Section S1.1. The optimization problem in (4.7) has (under mild conditions) a closed form solution given by the leading eigenvector of $-\frac{1}{2} b_{G_k}$ (scaled by the square root of the leading eigenvalue), see Lemma S1.2. In practice, we standardize the resulting $\hat{t}$ so that it takes values between 0 and 1.

Once $\hat{t}_j$ is obtained, the curve f can be estimated by smoothing each component function separately. We plug in $\hat{t}_j$ to (2.1) to obtain

$$\mathbb{E}(x_{j1}) = f_1(\hat{t}_j) \tag{4.8}$$

$$\mathbb{E}(x_{j2}) = f_2(\hat{t}_j) \tag{4.9}$$

We obtain $\hat{f}_1$ and $\hat{f}_2$ by using regression splines as implemented in *mgcv* (Wood, 2017).

Extending the method to closed curves. For closed curves, we have $f(a) = f(b)$, violating the non-intersecting condition required above. In this case, the approximation in (4.3) no longer holds because there could be a shorter path passing over the endpoint. However, we can still obtain an explicit form for the shortest path between t_i and t_j by applying the law of cosines:

$$\begin{aligned} \theta_i &:= 2\pi \left(\frac{t_i - b}{b - a} \right) \\ d_{G_k}(i,j) &\approx (b-a) \arccos \left(1 - \frac{(\cos(\theta_i) - \cos(\theta_j))^2 - (\sin(\theta_i) - \sin(\theta_j)^2}{2} \right). \end{aligned} \tag{4.10}$$

As $d_{G_k}^2$ appears to have no simple form, we make a second-order Taylor approximation:

$$d_{G_k}^2(i,j) \approx c \left[(\cos(\theta_i) - \cos(\theta_j))^2 - (\sin(\theta_i) - \sin(\theta_j)^2 \right] \tag{4.11}$$

where c is some constant (the entire derivation is in Section S1.4). Applying the double centering operation to (4.11) yields

$$-\frac{1}{2} b_{G_k}(i,j) \approx c \cos(\theta_i) \cos(\theta_j) + c \sin(\theta_i) \sin(\theta_j). \tag{4.12}$$

Because $\cos(\theta), \sin(\theta) \in \mathbb{R}^n$ are expected to be approximately orthogonal, a reasonable approximation θ_j is given by

$$\hat{\theta}_j = \text{atan2}\left(v_2(-b_{G_k}/2)_j, v_1(-b_{G_k}/2)_j \right) \tag{4.13}$$

where $v_k(\cdot)$ denotes the k-th leading eigenvector of a matrix and atan2 is the 2-argument arctangent function. Thus $\hat{\theta}_j$ can be converted back to $\hat{t}_j$ via equation (4.13), although our downstream analysis of SVG detection will be invariant to this scaling.

Second morphologically relevant coordinate. The second morphologically relevant coordinate $\hat{r}_j \in \mathbb{R}$ describes how far from the estimated curve a cell's coordinates are. The magnitude of the coordinate is defined as

$$||\hat{r}_j||_2 := ||x_j - \hat{f}(\hat{t}_j)||_2 \tag{4.14}$$

The sign of the second coordinate is determined by

$$\text{sign}(\hat{r}_j) = \text{sign}\langle x_j - \hat{f}(\hat{t}_j), R\hat{f}'(\hat{t}_j)\rangle \tag{4.15}$$

where $R : \mathbb{R}^2 \to \mathbb{R}^2$ is a counter-clockwise rotation by 90 degrees: $R(v_1, v_2) = (-v_2, v_1)$.

The intuition behind this equation is that cells/spots with a positive sign would be on the left-hand side if one was driving along the curve. The left-hand side is identified by a counter-clockwise rotation R of the velocity vector $f'(t)$. Again, in practice we standardize $\hat{r}_j$ to be in the interval $[0, 1]$ although this could be modified depending on the specific scientific question.

Disconnected graphs. The estimation procedure described above requires G_k to be connected. However, if k is chosen large enough to ensure the graph is fully connected, then d_{G_k} may not capture more subtle morphological features. For this reason, we permit the procedure to be applied separately to disconnected components of G_k and then *stitched* together to create the final curve. Given G_k has C connected components, let $x_1^{(c)}$ and $x_2^{(c)}$ denote endpoints of the curve describing the c-th component. We then identify the connections between $x_i^{(c)}$ and $x_{i'}^{(c')}$, $i, i' \in \{1, 2\}$, $c \neq c'$ of minimum Euclidean distance that produce a single (connected) curve. Note that this may require reversing the direction of a curve fit to one particular component. We identify the optimal connections through a brute force search of the $C! \cdot 2^C$ possibilities. Because this is computationally infeasible for large C, we require that k is at least large enough to ensure that $C \leq 5$.

Generalized additive model to identify spatially variable genes. Let Y_{gj} denote the count for gene g $(1 \leq g \leq G)$ in cell/spot j $(1 \leq j \leq n)$. As noted before, we consider the model

$$\begin{aligned} Y_{gj} &\sim \text{NegBinom}(n_j \mu_{gj}, \theta_g) \\ \log \mu_{gj} &= \beta_{g0} + h_g(\hat{t}_j) + s_g(\hat{r}_j) \end{aligned} \tag{4.16}$$

where h_g and s_g are unknown smooth functions, β_{g0} is an unknown intercept, and $n_j := \sum_{g=1}^{G} Y_{gj}$ is a known offset. Note that we are using the following standard parameterization of the negative binomial distribution:

$$\mathbb{P}(Y_{gj} = y) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)y!}\left(\frac{\theta}{\theta + n_j\mu_{gj}}\right)^{\theta}\left(\frac{n_j\mu_{gj}}{\theta + n_j\mu_{gj}}\right)^{y} \quad y = 0, 1, 2, \ldots \tag{4.17}$$

For identifiability, we also assume that $\sum_{j=1}^{n} h_g(\hat{t}_j) = \sum_{j=1}^{n} s_g(\hat{r}_j) = 0$. We use *mgcv* (Wood, 2017) to estimate the functions in model (2.6); this method writes h_g and s_g as a linear combination of known basis functions

$$\log \mu_{gj} = \beta_{g0} + \sum_{\ell=1}^{L_t} \beta_{g\ell}^{(t)} \phi(\hat{t}_j) + \sum_{\ell=1}^{L_r} \beta_{g\ell}^{(r)} \psi(\hat{r}_j) \tag{4.18}$$

Although there is flexibility in the choice of ϕ and ψ, we use cubic regression splines (cyclic for ϕ when f is a closed curve). Estimation proceeds by maximizing the log-likelihood of the model parameters $\ell(\beta_{g0}, \beta_g^{(t)}, \beta_g^{(r)})$ (here $\beta_g^{(t)} \in \mathbb{R}^{L_\ell}$ and $\beta_g^{(r)} \in \mathbb{R}^{L_r}$ are vectors of coefficients) subject to a smoothness penalty:

$$\text{argmin}_{\beta}\left\{ -\ell(\beta_{g0}, \beta_g^{(t)}, \beta_g^{(r)}) + \lambda_t (\beta_g^{(t)})^\top S_t \beta_g^{(t)} + \lambda_r (\beta_g^{(r)})^\top S_r \beta_g^{(r)} + \lambda \left((\beta_g^{(t)})^\top \beta_g^{(t)} + (\beta_g^{(r)})^\top \beta_g^{(r)} \right) \right\} \tag{4.19}$$

where S_t and S_r are (known) matrices depending on the second derivative of the chosen basis functions (Wood, 2001). *mgcv* performs a procedure to select the best choice of λ_t and λ_s and we set $\lambda = 1$ by default. Upon estimating the coefficients, *mgcv* returns a Bayesian covariance matrix for uncertainty quantification. We use this to obtain the posterior standard deviation $\text{sd}(\hat{\beta}_{g\ell}^{(\cdot)})$ for each coefficient. For further shrinkage towards 0, we apply adaptive shrinkage (ash) (Stephens, 2017) to the estimated coefficients and their standard deviations to obtain the final estimate of $\hat{h}$ and $\hat{s}$.

Data and code availability. The following datasets were used:

- The granule cells in **Figure 3** were obtained from the data provided by Cable et al. (2022b). Cells such that the 5-th nearest neighbor was 2 times greater than the median 5-th nearest neighbor were excluded. This procedure removed outlier cells.

- The CA3 cells were obtained from *STexampleData* (Righelli et al., 2022). Cells such that the 20-th nearest neighbor was 3 times greater than the median 20-th nearest neighbor were excluded. These values were used so that the retained set of cells visually matched Figure 5 of Cable et al. (2022b).

- MERFISH measurements of the adult healthy colon are available upon request from the authors. Briefly, these measurements were performed using standard MERFISH protocols (Cadinu et al., 2024) targeting a custom set of 1,920 genes.

MorphoGAM is available as an R package at https://github.com/phillipnicol/MorphoGAM. The repository also includes scripts to reproduce all results in the paper.

Acknowledgements

PBN is supported by the National Institutes of Health grant T32CA009337.

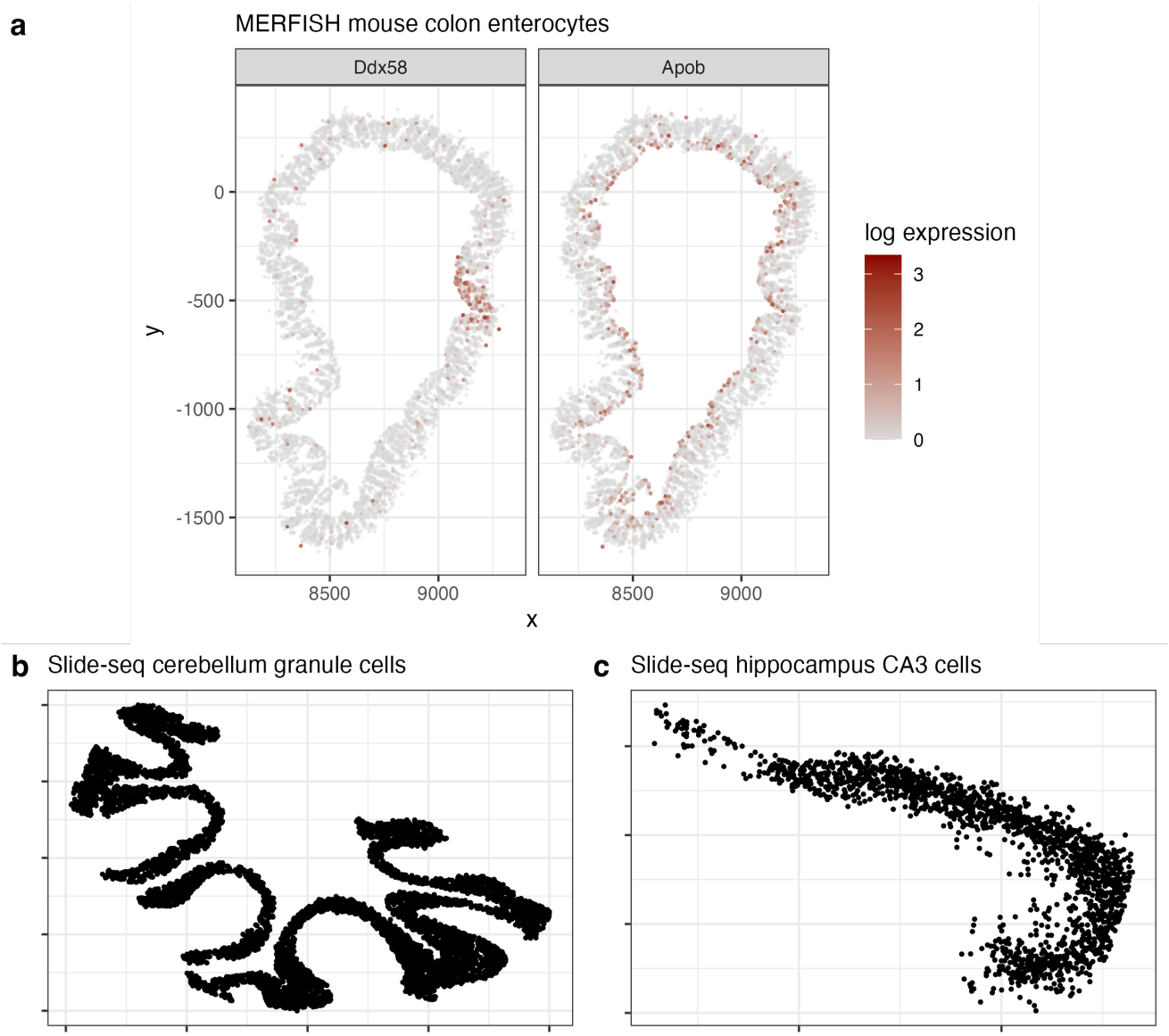

Figure 1: **a.** The spatial location of enterocytes identified in a cross section of the healthy mouse colon as measured with MERFISH. Cells are colored by the log-transformed expression of the two listed genes. The expression of *Ddx58* is called localized whereas the expression pattern of *Apob* is called radial gradient. The spatial locations of the plotted enterocytes lie close to a one-dimensional circular manifold. Additional examples of cell types with coordinates that lie close to a one-dimensional curve can be found in **b.** granule cells from the mouse cerebellum (Cable et al., 2022b) and **c.** CA3 cells in the mouse hippocampus (Stickels et al., 2021)

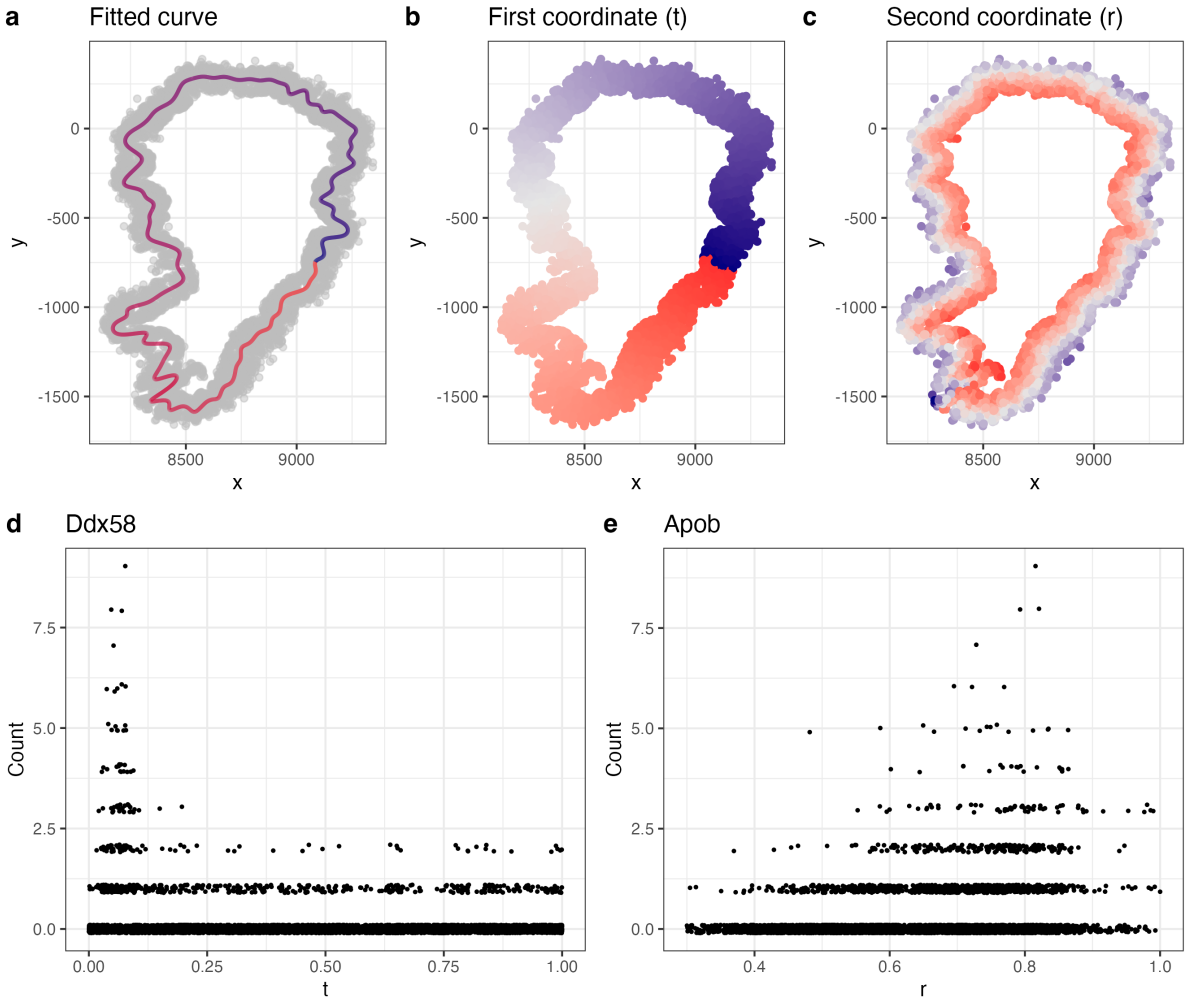

Figure 2: **Overview of MorphoGAM. a.** MorphoGAM begins by estimating a smooth parametric curve passing through the spatial transcriptomic sample coordinates. **b.** The first morphologically relevant coordinate is defined as the position of each cell (or more generally, sample) along the estimated curve from the previous step. **c.** The second morphologically relevant coordinate is defined as the position of the cell in the direction orthogonal to the curve at a given point. **d.** Genes with localized variation such as *Ddx58* show strong expression variation as a function of the first morphologically relevant coordinate. **e.** Genes with a radial gradient pattern such as *Apob* show expression variation as a function of the second morphologically relevant coordinate.

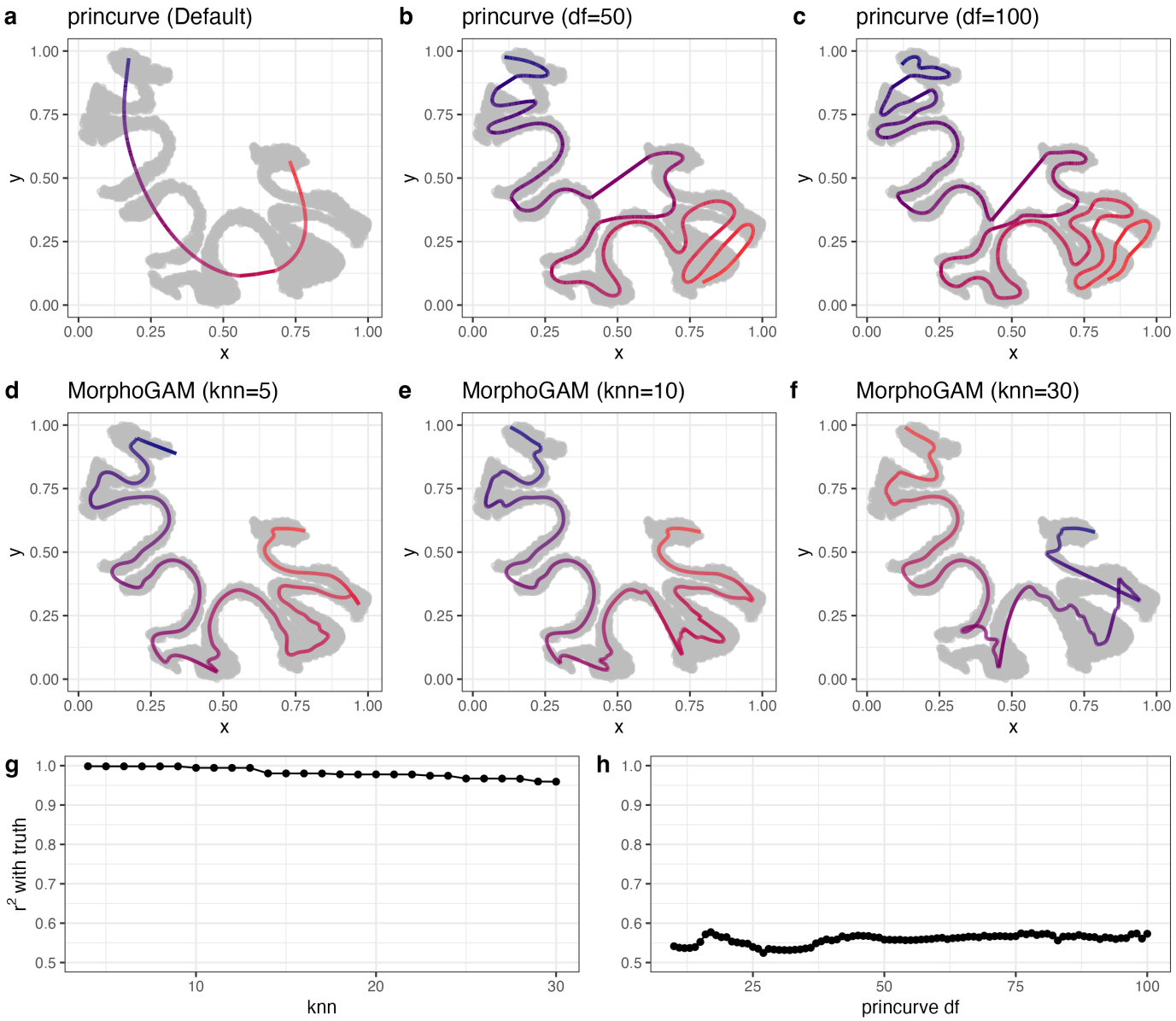

Figure 3: **MorphoGAM outperforms existing curve estimation approaches.** **a.** Applying the principal curves method (Hastie and Stuetzle, 1989) as implemented by the *princurve* (Cannoodt, 2018) R package. **a.**, **b.**, and **c.** show the estimated curve for three different values of the tuning parameter, which is degrees of freedom (df) of the smoothing spline. **d.**, **e.** and **f.** show the estimated curve from MorphoGAM with three different values of its tuning parameter k-nearest neighbor (kNN). **g.**, **h.** Using the estimated coordinate t_j from a hand-drawn ground truth (**Fig S3**) we compute the squared spearman correlation between this and the estimated coordinate from both methods as the tuning parameters vary.

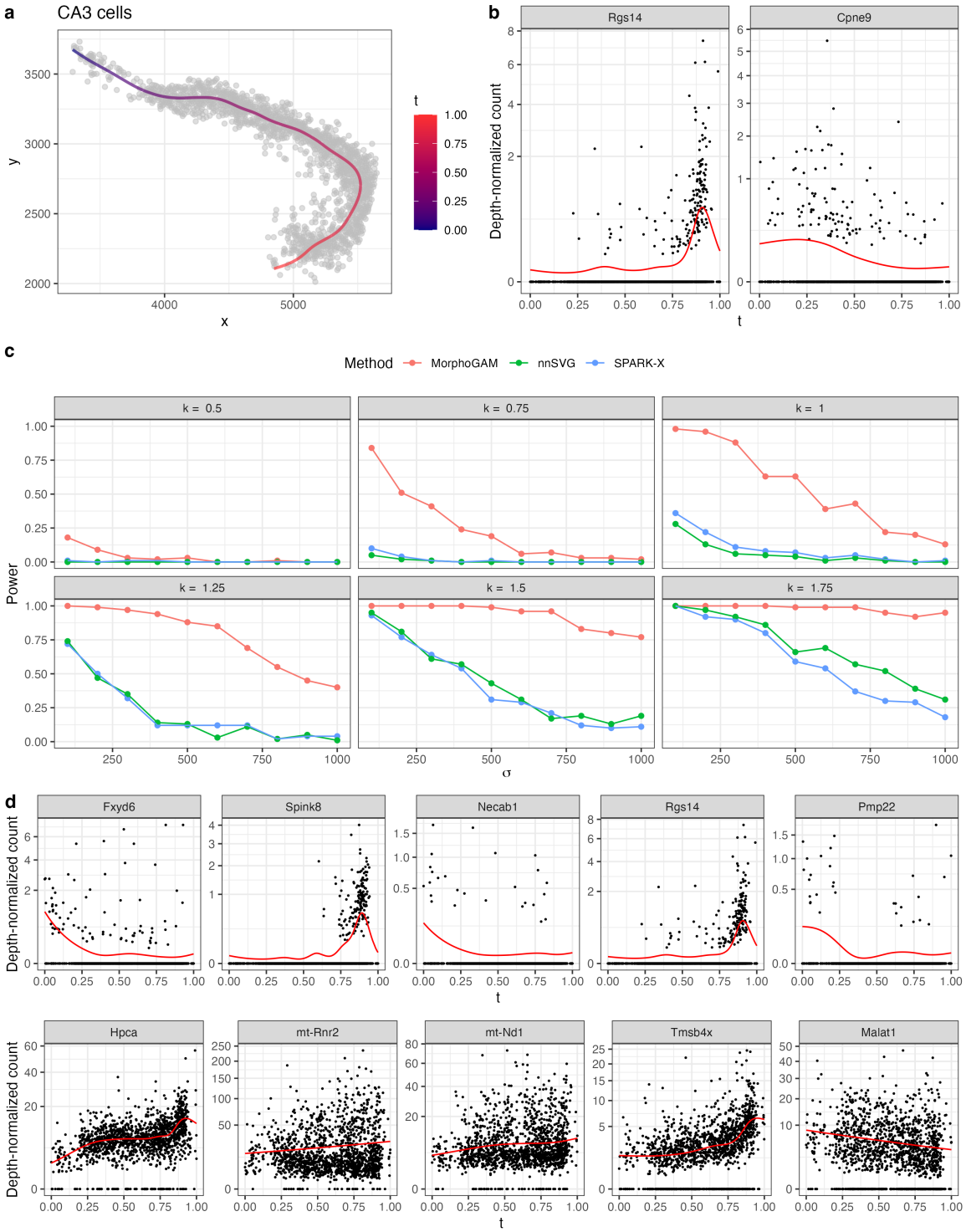

Figure 4: **MorphoGAM increases power to detect relevant spatially variable genes. a.** The estimated curve $\hat{f}$ on the CA3 cells from mouse hippocampus (see **Fig 1d**). **b.** The estimated functions $\hat{h}(\hat{t}_j)$ for two previously reported spatially variable genes *Rgs14* and *Cpne9*. **c.** Comparing hypothesis-based frameworks to detect SVGs; a gene with $\mu_{gj} = 1 + \kappa \exp(-\sigma(\hat{t} - 0.5)^2)$ and $\theta = 5$ was simulated and labeled as SVG is the corresponding p-value was smaller than $0.05/20000$. The power reflects the proportion of 100 trials where the null hypothesis was correctly rejected. **d.** Plotting the genes with the largest peak and range summaries.

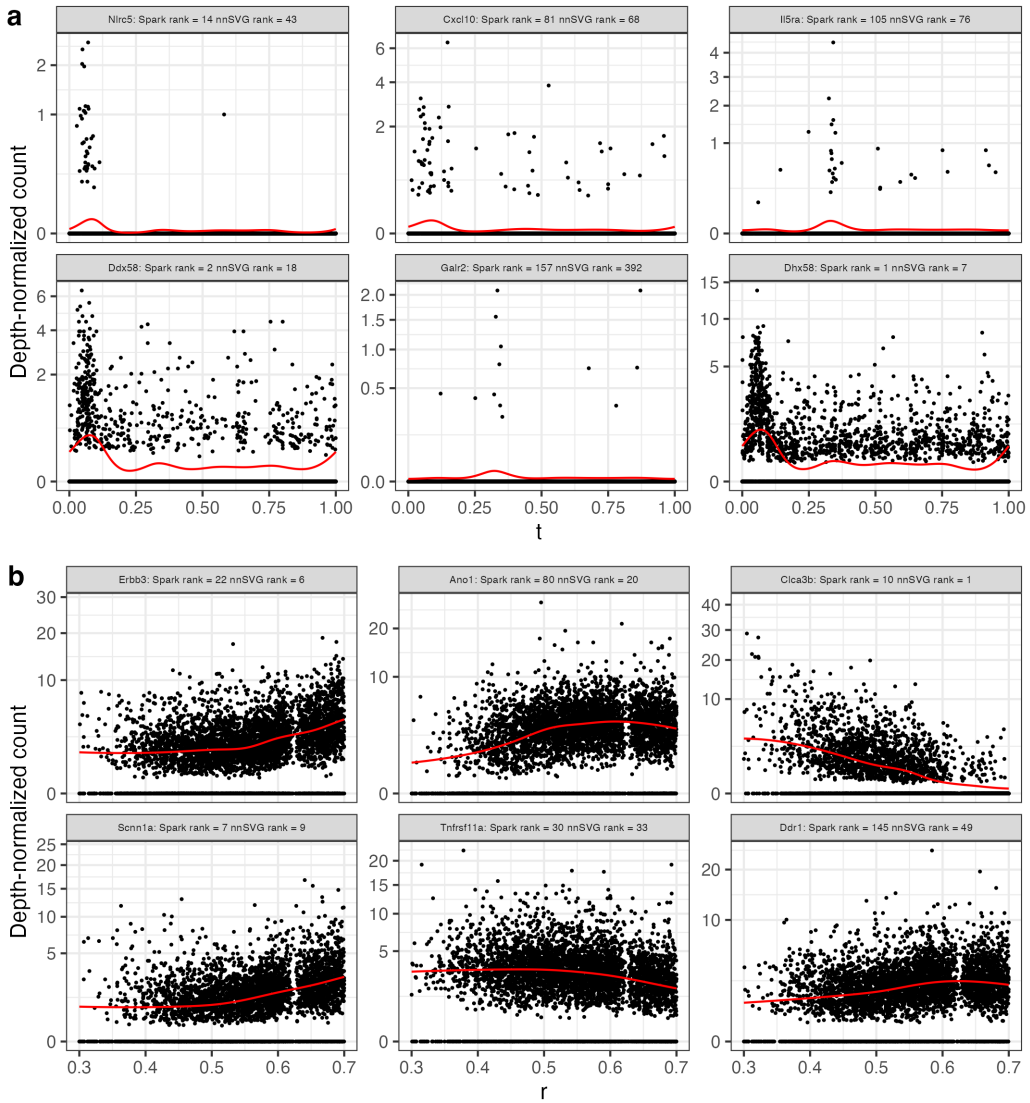

Figure 5: **MorphoGAM identifies additional genes with localized and radial gradient pattern. a.** The top six genes identified when ranking by the peak of h_g. That is, genes with a high log fold-change relative to baseline in the direction of the first morphologically relevant coordinate. **b.** The top six genes identified when ranking by the range of s_g. That is, genes with a large changes on the scale of the counts in the direction of the second morphologically relevant coordinate. Each label shows the ranking of the gene from SPARK-X (Zhu et al., 2021) and nnSVG (Weber et al., 2023).

Supplementary Material for "Identifying spatially variable genes by projecting to morphologically relevant curves"

S1 Mathematical details of curve and coordinate estimation

S1.1 Notation

We use $\otimes$ to denote outer-product: if $u, v \in \mathbb{R}^n$ then $u \otimes v := uv^\top \in \mathbb{R}^{n\times n}$. If $f : [a, b] \to \mathbb{R}^k$ denotes a parametric curve, then f can be written in terms of k *component functions* $f(t) = (f_1(t), \ldots, f_k(t))$ and $f'(t) := (f_1'(t), \ldots, f_k'(t))$. We say f is *smooth* if $f_i'(t)$ exists and is continuous for all t. For a matrix $A \in \mathbb{R}^{m\times n}$, $||A||_F^2 = \sum_{i=1}^n \sum_{j=1}^m A_{ij}^2$ denotes Frobenius norm.

S1.2 Assumptions

We assume the following conditions, which are necessary (but not sufficient) for the identifiability of model (2.1):

1. $\bar{t} := \frac{1}{n}\sum_{j=1}^n t_j = 0$ and $t_1 < 0$.
2. $||f'(t)||_2 = 1$ for all t.

For condition 1, note that for any constant c, $\tilde{f} : [a - c, b - c] \to \mathbb{R}^2$ defined by $\tilde{f}(t) = f(t - c)$ satisfies $f(t_j) = \tilde{f}(t_j + c)$ for every j. Similarly, $\tilde{f} : [-b, -a] \to \mathbb{R}^2$ defined by $\tilde{f}(t) = f(-t_j)$ satisfies $f(t_j) = \tilde{f}(-t_j)$. For condition 2, define $h(t) = \int_a^t ||f'(s)||_2 ds$ and note that $h'(t) = ||f'(t)||_2$. Then we can define the reparameterized curve $\tilde{f} := f(h^{-1}(t))$. h^{-1} exists and is differentiable by the inverse function theorem. In particular,

$$||(f \circ h^{-1})'(t)||_2 = ||f'(h^{-1}(t)) \cdot (h'(h^{-1}(t)))^{-1}||_2 = ||f'(h^{-1}(t)) \cdot (f'(h^{-1}(t))^{-1}||_2 = 1. \tag{S1.1}$$

A full introduction to parametric curves is given by Lastra (2021). We also note that these conditions are not sufficient to ensure the identifiability of model (2.1) as a fully identifiable model would likely need to specify a distribution or a procedure from which the t_j are obtained.

S1.3 Linear curve

We now describe how to estimate the first coordinate in the case of a linear curve (i.e., $x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$). If we assume that the approximation in equation (4.3) is equality, i.e., $d_{G_k}(i,j) = |t_i - t_j|$, then $d^2_{G_k} \in \mathbb{R}^{n\times n}$ can be written in matrix form as

$$d^2_{G_k} = t^2 \otimes 1_n + 1_n \otimes t^2 - 2t \otimes t \tag{S1.2}$$

where t^2 is applied entry-wise to $t := (t_1, \ldots, t_n)$ and $1_n \in \mathbb{R}^n$ is a vector of 1's. Now define the centering matrix $H \in \mathbb{R}^{n\times n}$ as

$$H = I - \frac{1}{n} 1_n \otimes 1_n \tag{S1.3}$$

Applying H on the right has the property of subtracting the row means while appliying H on the left subtracts the column means.

Lemma S1.1. *The "double centered" matrix b_{G_k} satisfies*

$$-\frac{1}{2} b_{G_k} := -\frac{1}{2} H d^2_{G_k} H = t \otimes t \tag{S1.4}$$

Proof. The proof is derived from Ghojogh et al. (2023). Because $H(1_n \otimes t^2) = 0$ and $(t^2 \otimes 1_n)H = 0$, we have

$$\frac{-1}{2} H_{G_k^2} H = \left(-\frac{1}{2} H(t^2 \otimes 1_n) + H(t \otimes t) \right) H \tag{S1.5}$$

$$= H(t \otimes t)H \tag{S1.6}$$

$$= t \otimes t \tag{S1.7}$$

where the last line follows because $\bar{t} = 0$ by assumption. □

The above result shows that $-\frac{1}{2} b_{G_k}$ is a rank 1 matrix with a positive eigenvalue $\langle t, t\rangle > 0$. In practice, however, $-\frac{1}{2} b_{G_k}$ will be expected to have higher rank due to noise. For this reason, we estimate t using the top eigenvector associated with the largest eigenvalue. The following theorem shows that, under some conditions, the top eigenvector of a symmetric matrix leads to the best rank-one approximation with the smallest reconstruction error.

Lemma S1.2. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix, and suppose that $\lambda_{\max}(A) > 0$ and $\lambda_{\max}(A) >$*

$|\lambda_{\min}(A)|$, *where* $\lambda_{\max}$ *and* $\lambda_{\min}$ *denotes the largest and smallest eigenvalues, respectively. Then*

$$argmin_{t\in\mathbb{R}^n}||A - t\otimes t||_F^2 = \sqrt{\lambda_{\max}(A)}u_1 \tag{S1.8}$$

where u_1 *is the unit eigenvector corresponding to* $\lambda_{\max}(A)$.

Proof. As A is symmetric, we may write

$$A = \sum_{i=1}^{n} \lambda_i(u_i \otimes u_i) \tag{S1.9}$$

with $u_1, \ldots, u_n \in \mathbb{R}^n$ orthornormal. Then

$$A = \sum_{i=1}^{n} |\lambda_i|(\text{sign}(\lambda_i)u_i) \otimes u_i \tag{S1.10}$$

is a singular value decomposition (SVD) of A. By Eckart and Young (1936), we have

$$\min_{u,v\in\mathbb{R}^n} ||A - u\otimes v||_F^2 = \sum_{i=2}^{n} |\lambda_i| \tag{S1.11}$$

Moreover,

$$\min_{u,v\in\mathbb{R}^n} ||A - u\otimes v||_F^2 \leq \min_{t\in\mathbb{R}^n} ||A - t\otimes t||_F^2 \tag{S1.12}$$

so $\min_{t\in\mathbb{R}^n} ||A - t\otimes t||_F^2 \geq \sum_{i=2}^{n} |\lambda_i|$ as well. This minimum is achieved by setting $t = \sqrt{\lambda_{\max}(A)}u_1$. □

In practice, it seems to be the case that the condition $\lambda_{\max}(A) > 0$ and $\lambda_{\max}(A) > |\lambda_{\min}(A)|$ always holds.

S1.4 Closed curve

In the case of a closed curve $f(a) = f(b)$ and the approximation in equation (4.10) must be used for d_{G_k}:

$$\begin{aligned} \theta_i &:= 2\pi\left(\frac{t_i - b}{b - a}\right) \\ d_{G_k}(i,j) &\approx (b-a)\arccos\left(1 - \frac{(\cos(\theta_i) - \cos(\theta_j))^2 - (\sin(\theta_i) - \sin(\theta_j)^2)}{2}\right). \end{aligned} \tag{S1.13}$$

Consider $(b-a)^2\arccos^2(1-x^2/2)$ as a function of x. We make a second order Taylor approximation around $x = 0$, which yields

$$\arccos(1) + (b-a)^2x^2 = (b-a)^2x^2 \tag{S1.14}$$

Taking $c = (b-a)^2$ yields the approximation in (4.12). Then by double centering,

$$-\frac{1}{2} b_{G_k}(i,j) \approx c\cos(\theta_i)\cos(\theta_j) + c\sin(\theta_i)\sin(\theta_j). \tag{S1.15}$$

This implies that $-\frac{1}{2} b_{G_k}$ will be approximately rank 2. Moreover, if n is large and θ_i densely populated within $[0, 2\pi]$ then we have

$$\frac{1}{n}\sum_{i=1}^{n} \cos(\theta_i)\sin(\theta_i) \approx \int_0^{2\pi} \cos(\theta)\sin(\theta) d\theta = 0 \tag{S1.16}$$

which shows that (S1.15) is also an (approximate) eigendecomposition of $-\frac{1}{2} b_{G_k}$. In particular, if the two eigenvectors recover $\cos(\theta_i)$ and $\sin(\theta_i)$, respectively, then taking the arctangent function of the ratio should be a reasonable approximation to θ_i. Because both $\cos(\theta)$ and $\sin(\theta)$ are approximate eigenvectors with eigenvalue c, the top two eigenvectors could be invariant to rotation. However, in any case, the top two eigenvectors would still represent the location of each cell on a circle, and the two-argument arctangent function would still recover the angle along that circle.

S2 Supplementary figures

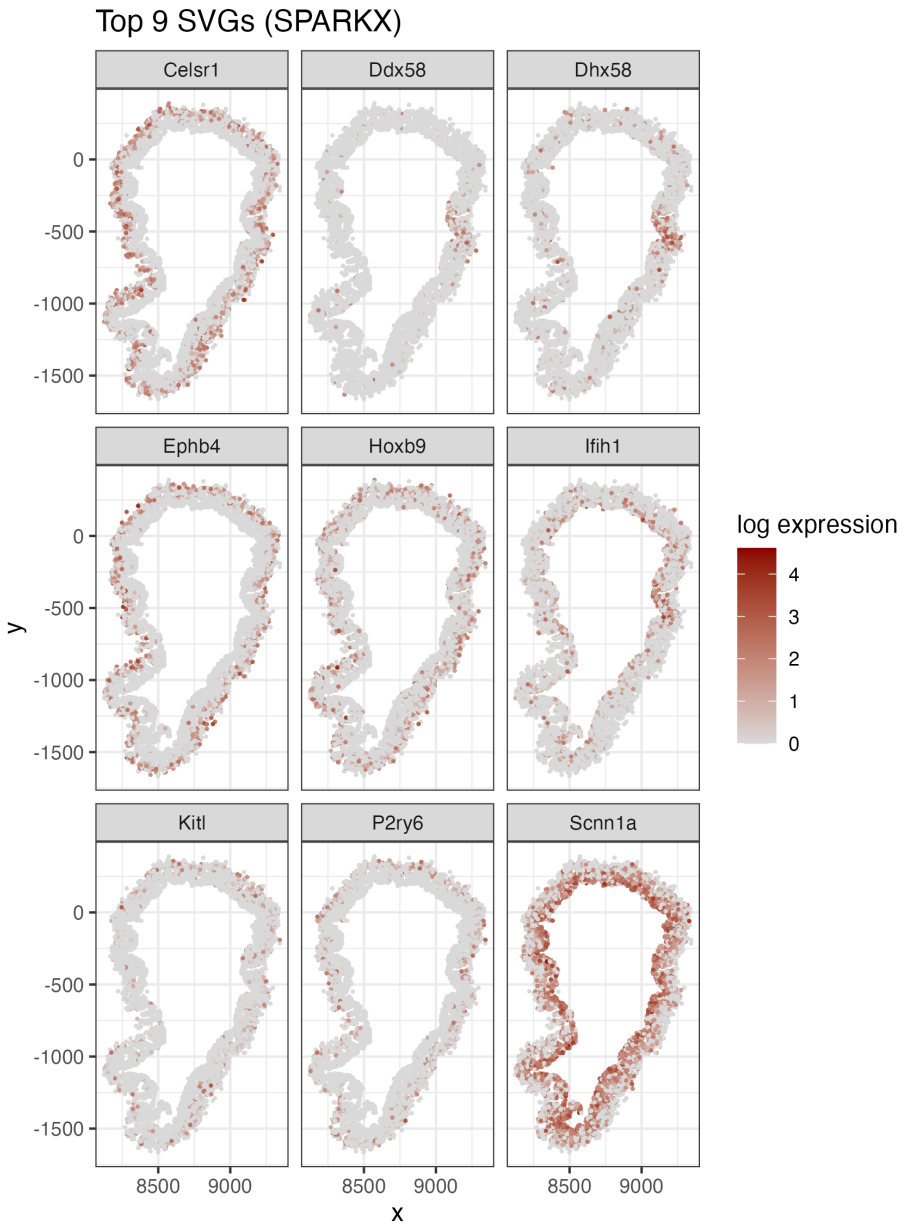

Figure S1: The top 9 SVGs identified by SPRAK-X (Zhu et al., 2021) in a MERFISH measurement of a slice of the healthy mouse colon. Specifically, *Ddx58* had a reported (adjusted) p-value of 6.38×10^{-67} and *Apob* had a reported (adjusted) p-value of 8.04×10^{-9}.

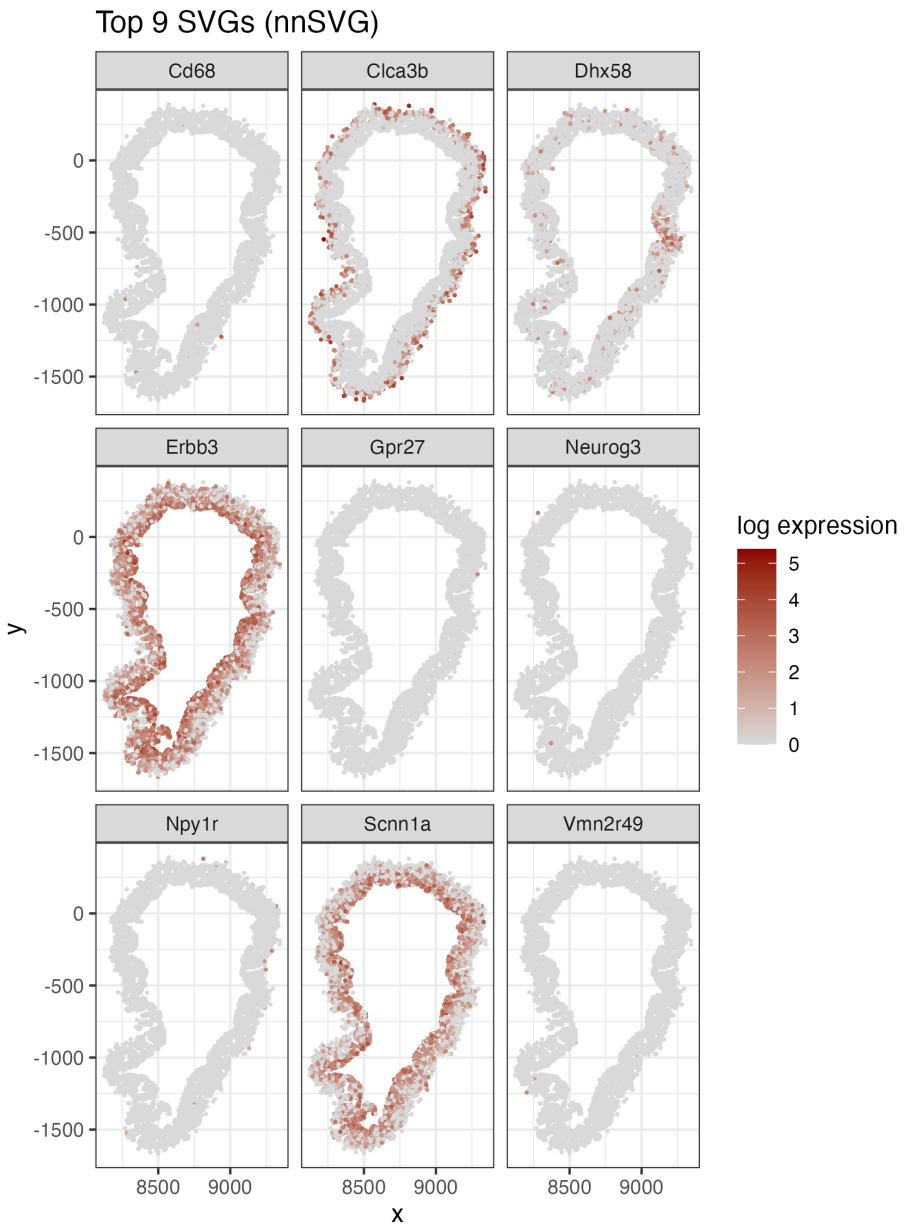

Figure S2: The top 9 SVGs identified by nnSVG (Weber et al., 2023) in a MERFISH measurement of a slice of the healthy mouse colon. Specifically, *Ddx58* and *Apob* both had reported (adjusted) p-values of 0.

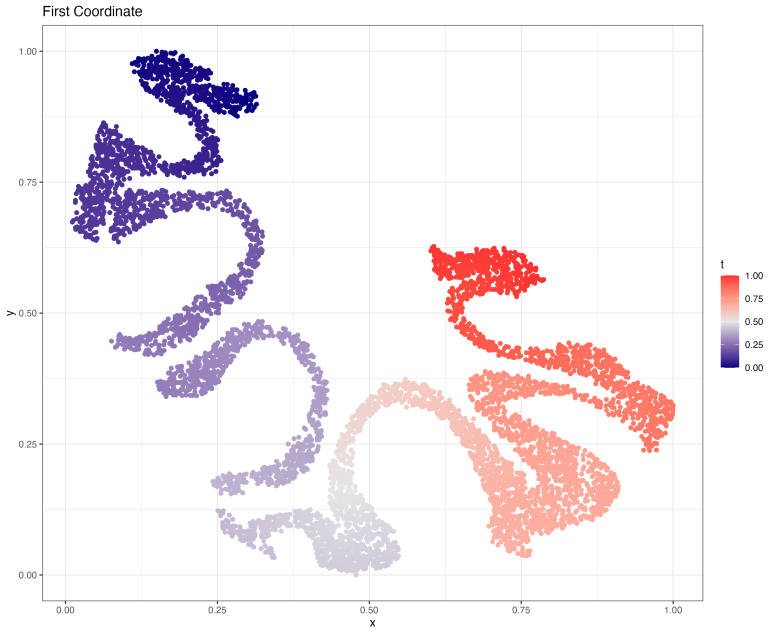

Figure S3: The estimated coordinate from the hand-drawn path on the granule cells.

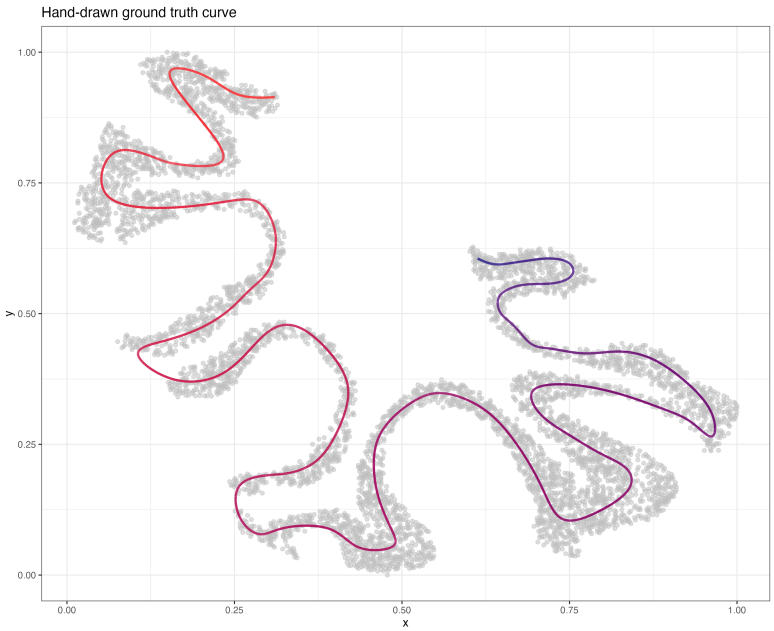

Figure S4: The hand-drawn curve on the granule cells.

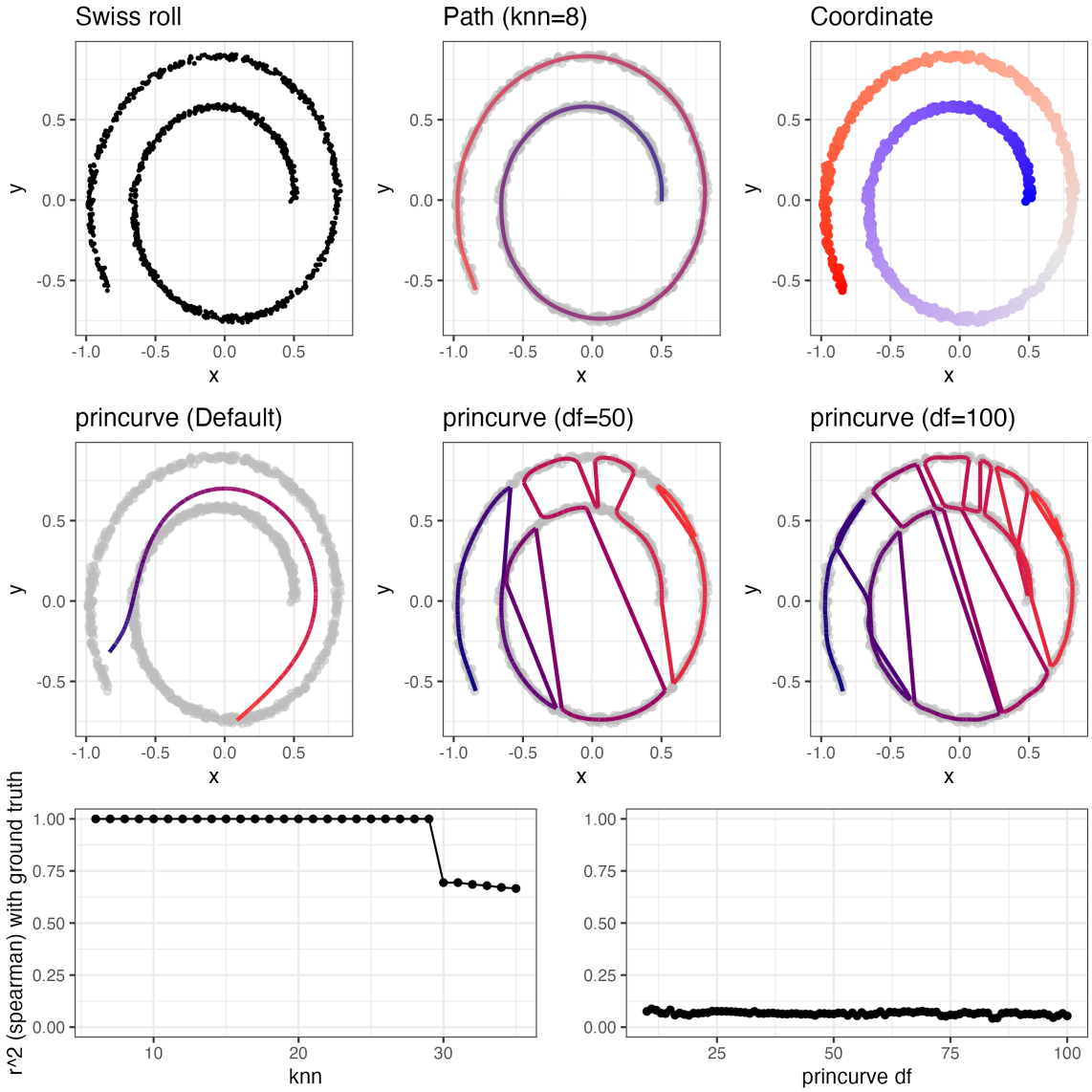

Figure S5: Repeating the analysis in Figure 3 instead using a simulated swiss roll. The inability of the standard principal curves algorithm to accurately reconstruct the swiss roll was discussed in (Kégl et al., 2000).

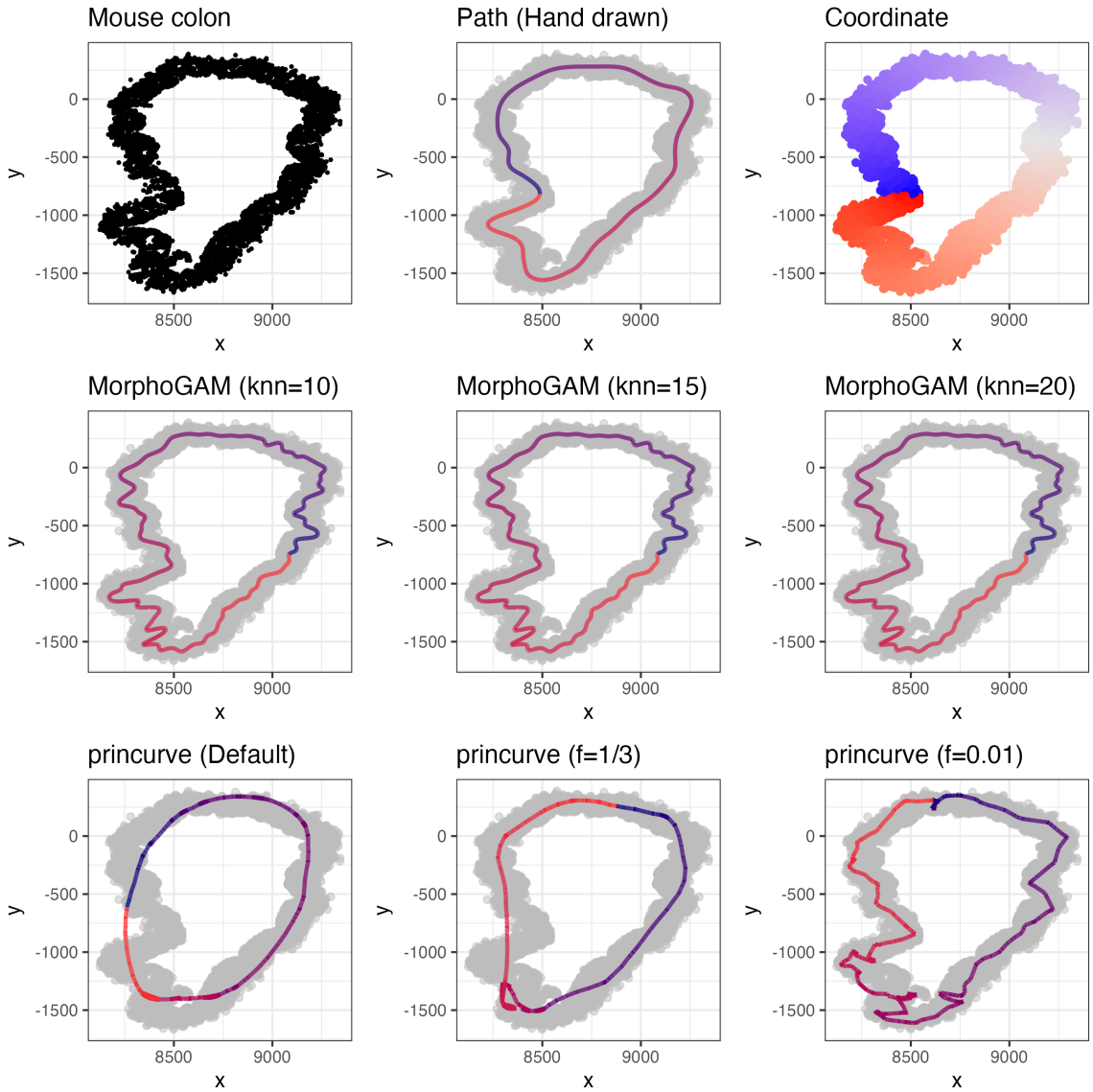

Figure S6: Repeating the analysis in Figure 3 instead using the mouse mucosa data. In this case a periodic smoother was used in princurve.

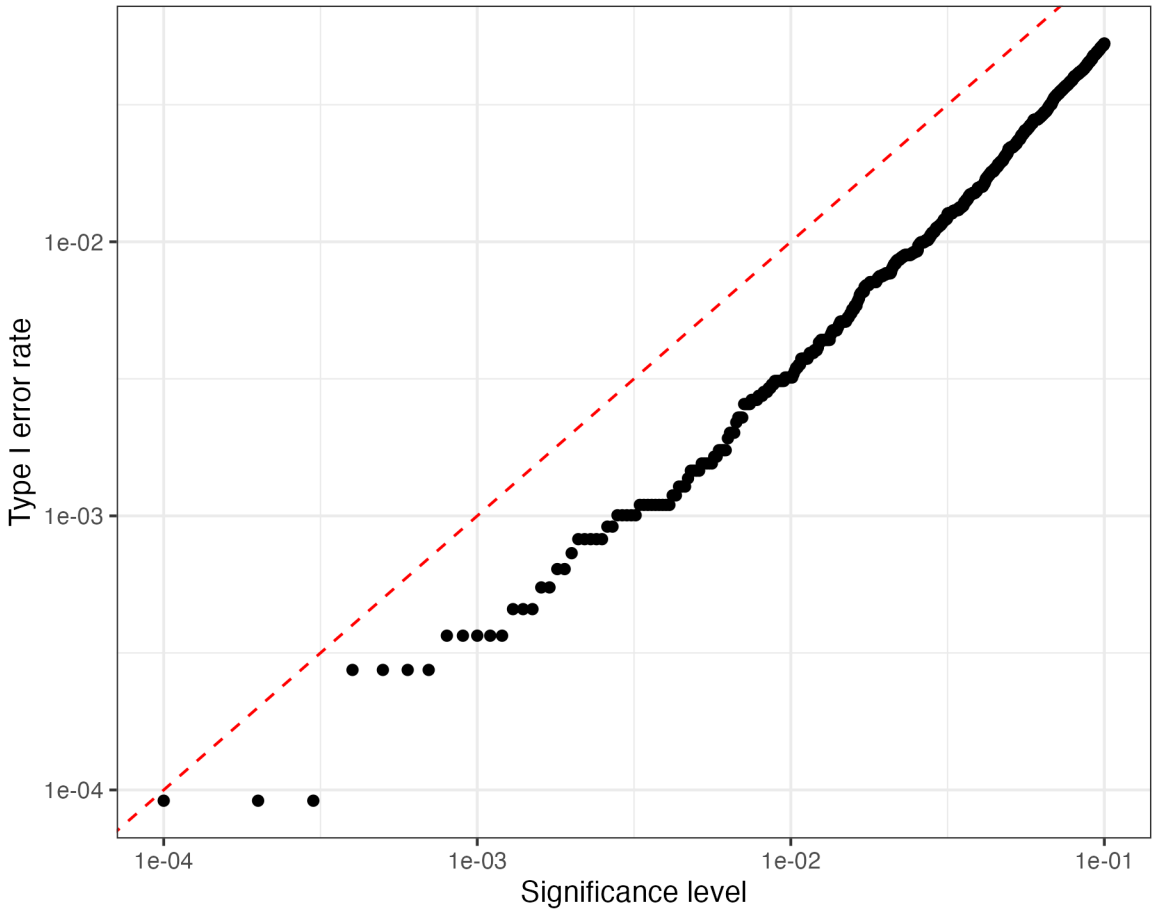

Figure S7: Spatial locations in the CA3 data were randomly permuted to produce a null dataset where there should be no SVGs. The proportion of genes with a p-value smaller each significance level was computed (the Type I error rate). The red-dashed line indicates the nominal type I error rate.

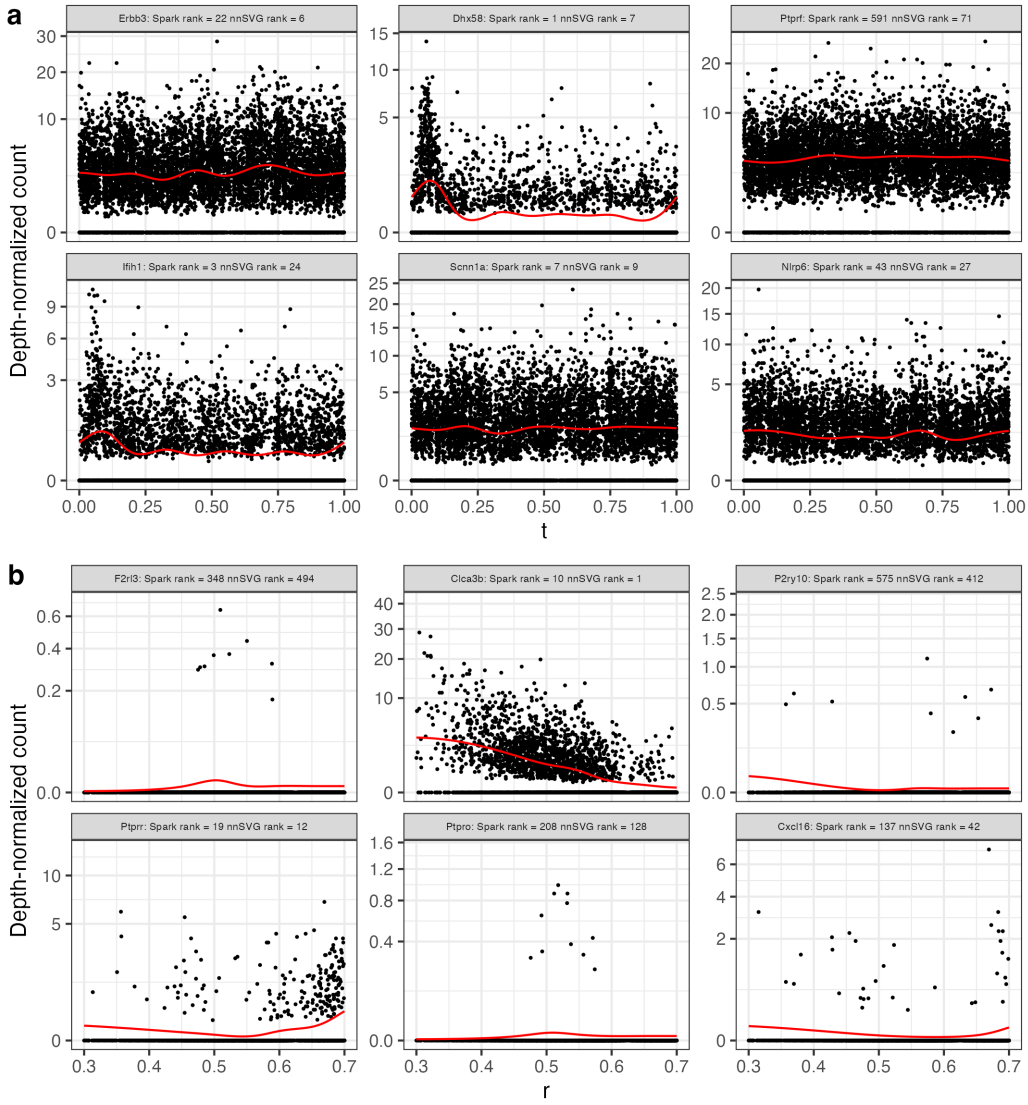

Figure S8: Repeating the analysis of Figure 5 plotting the genes with the largest range in the direction of the first morphologically relevant coordinate t_j and the the genes with the largest peak in the direction of the second morphologically relevant coordinate r_j.