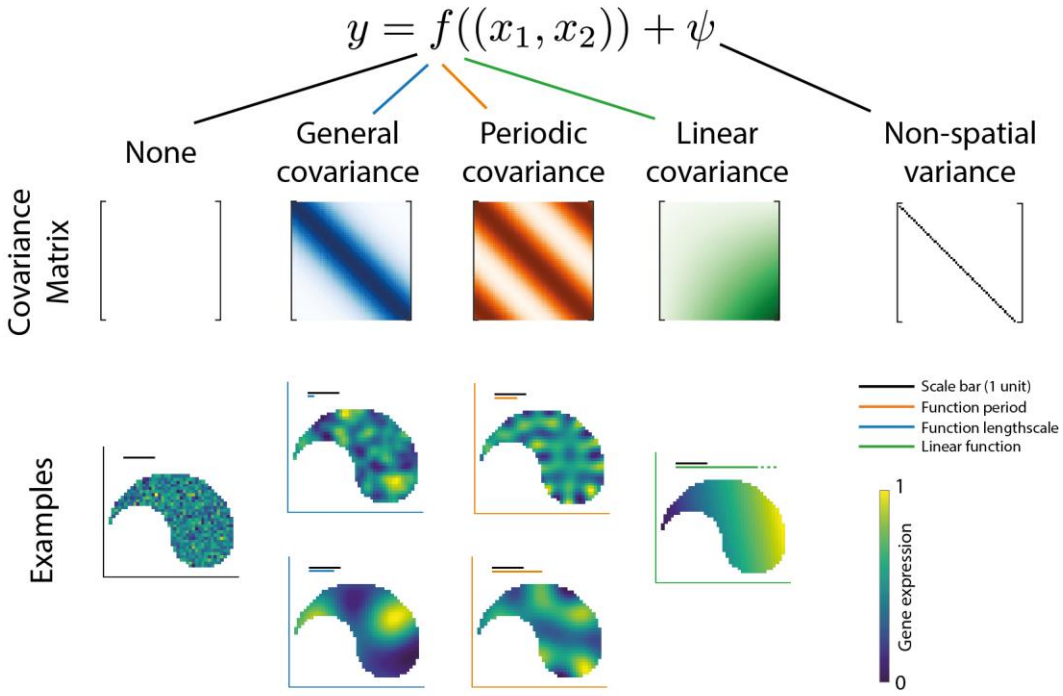
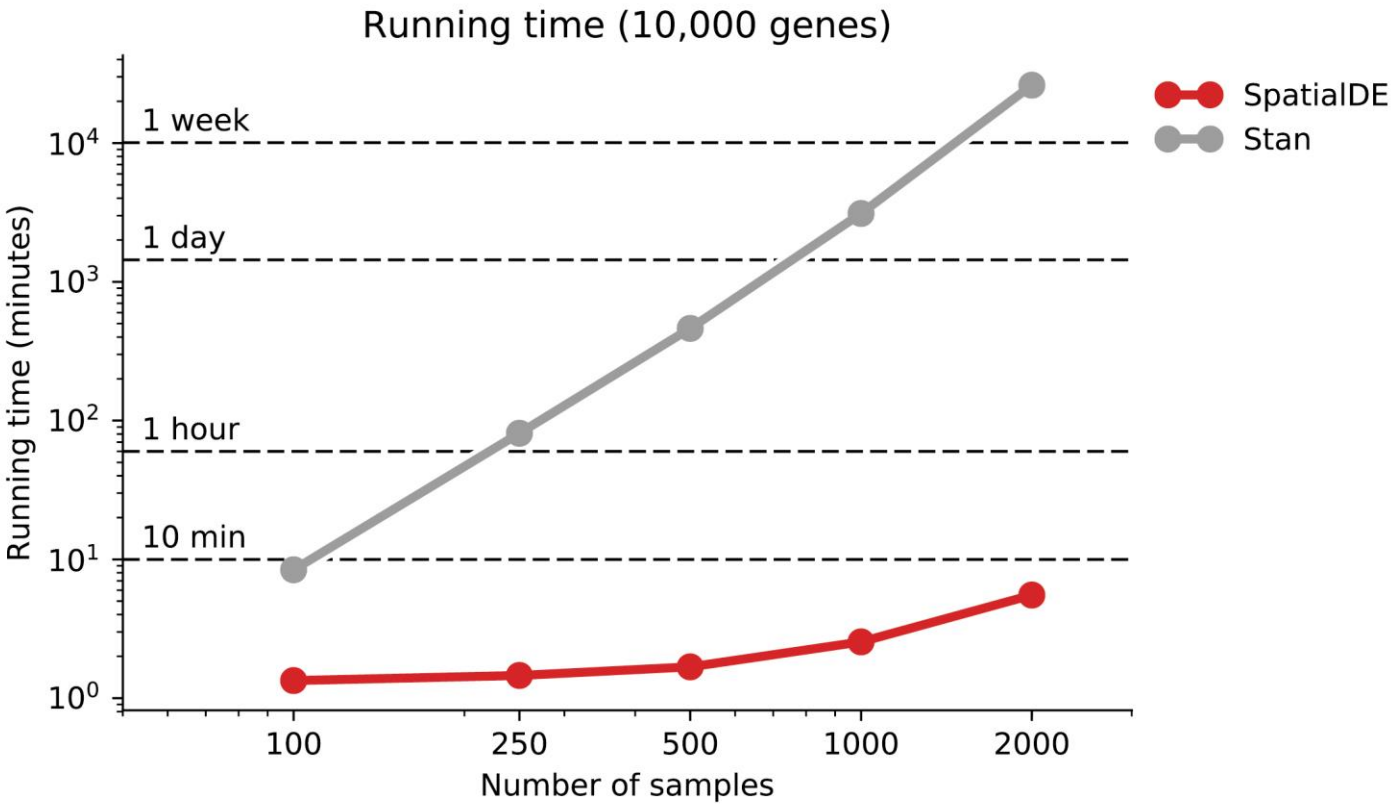


Supplementary Figure 1

Model selection of different covariance functions.

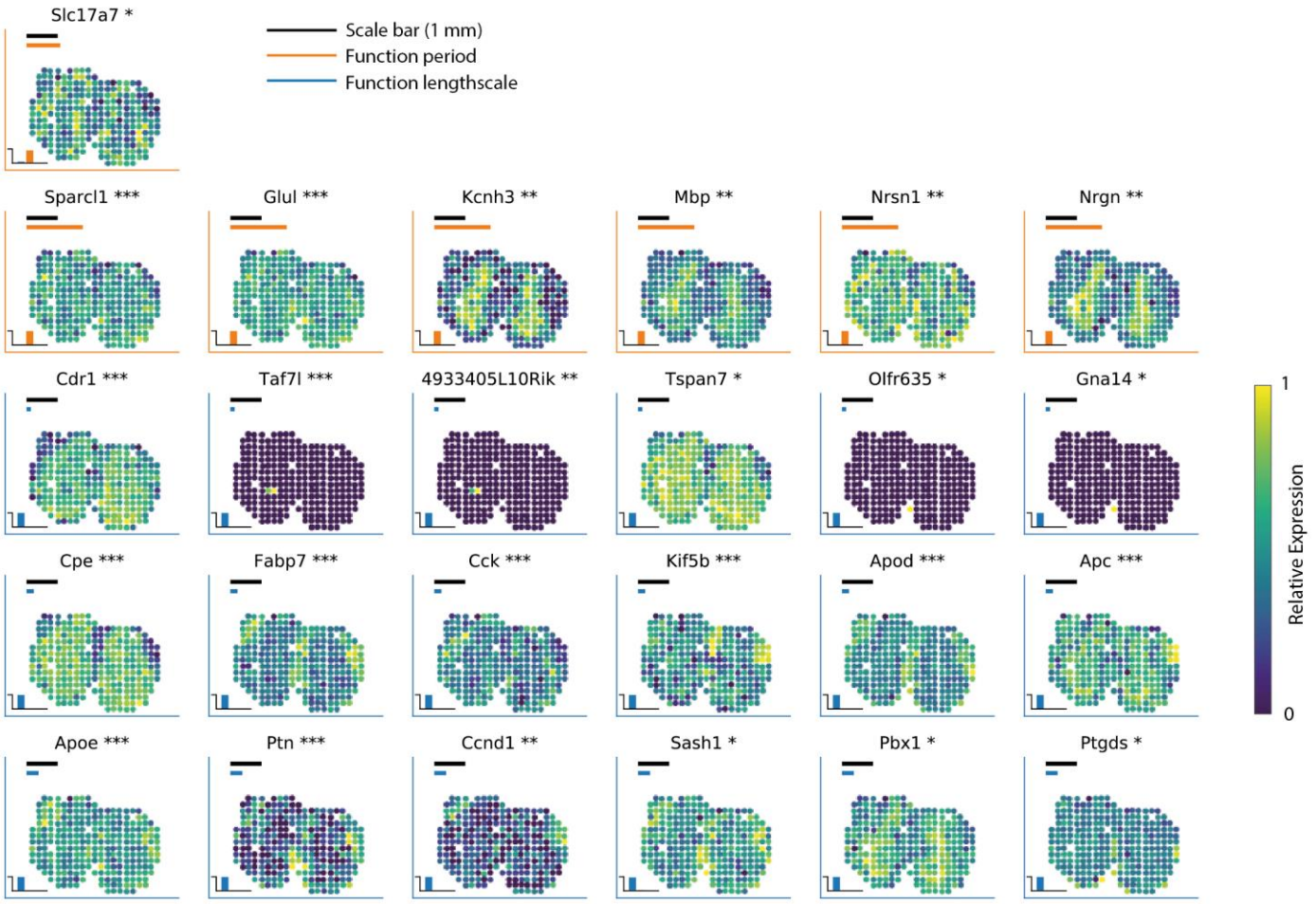
In addition to the hypothesis test of spatial vs non-spatial expression variance, SpatialDE can be used to classify spatially variable genes into genes with periodic, linear patterns, or general spatial patterns. Illustrative examples of simulated functional dependencies are shown below the corresponding covariance matrices.



Supplementary Figure 2

Computational efficiency of SpatialDE.

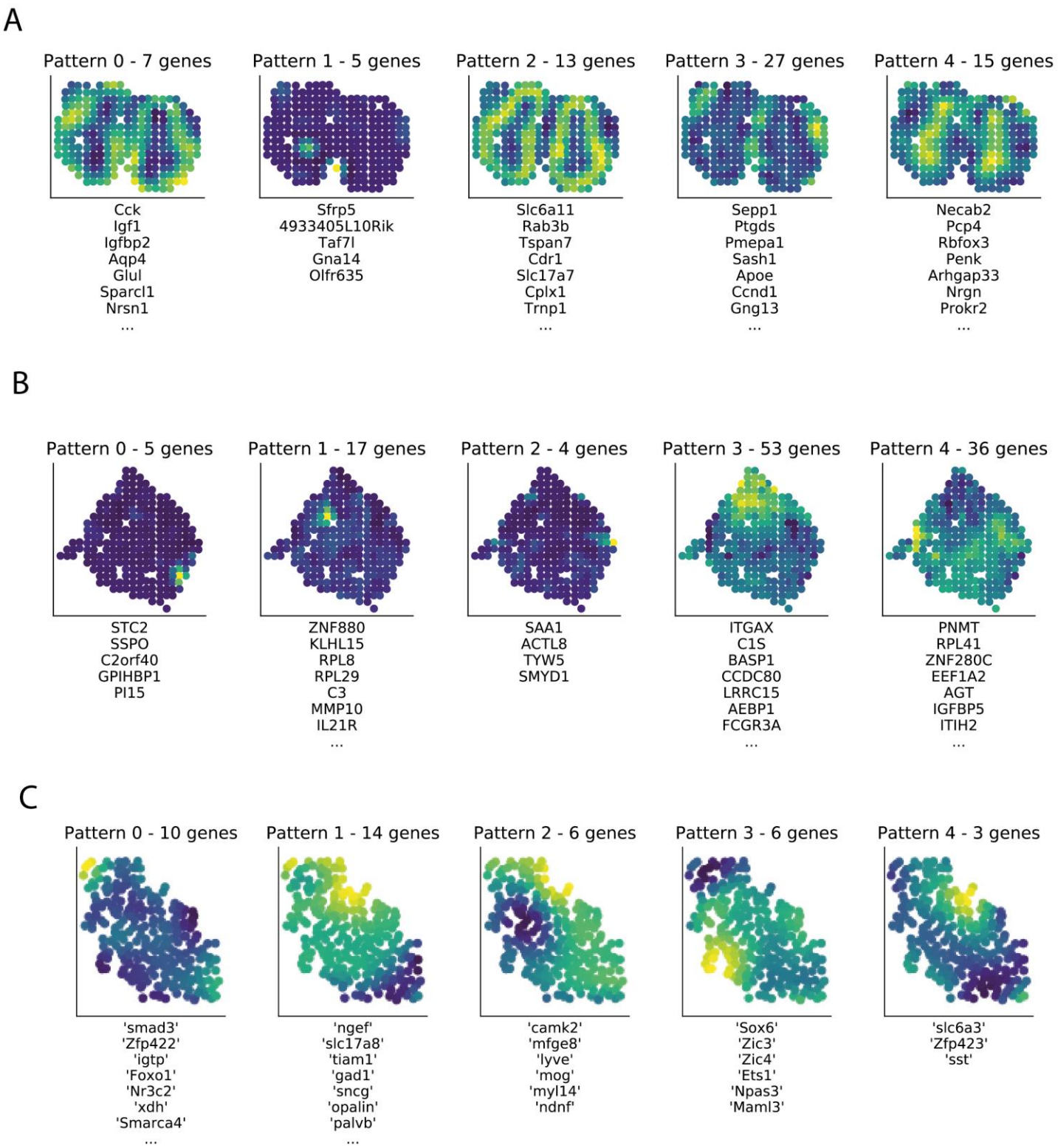
Compared is the SpatialDE native implementation versus a Stan implementation of the same model. Caching operations and linear algebra speedups are used where possible, enabling tractable genome-wide analyses with thousands of samples or cells. Shown is the empirical runtime for the SpatialDE test applied to 10,000 genes, using a late 2013 iMac with 3.2 GHz Intel Core i5 processor.



Supplementary Figure 3

Expanded example of spatially variable genes identified in the mouse olfactory bulb data.

Spatial expression patterns for 25 additional SV genes (out of 67, FDR<0.05, Q-value adjusted), selected to represent expression patterns with different function periods and length scales.

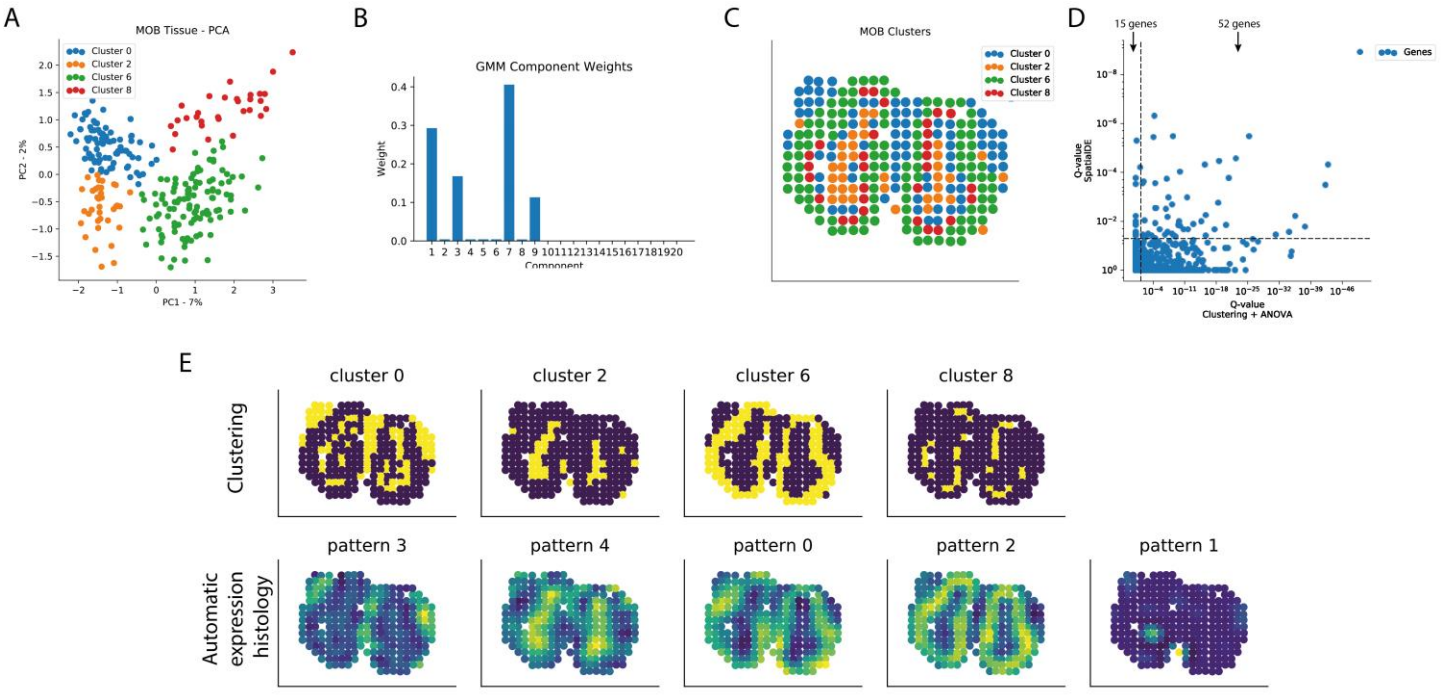


Supplementary Figure 4

Results from automatic expression histology.

(A) AEH applied to spatial transcriptomics data of mouse olfactory bulb, for K=5 spatial expression patterns. Color denotes the

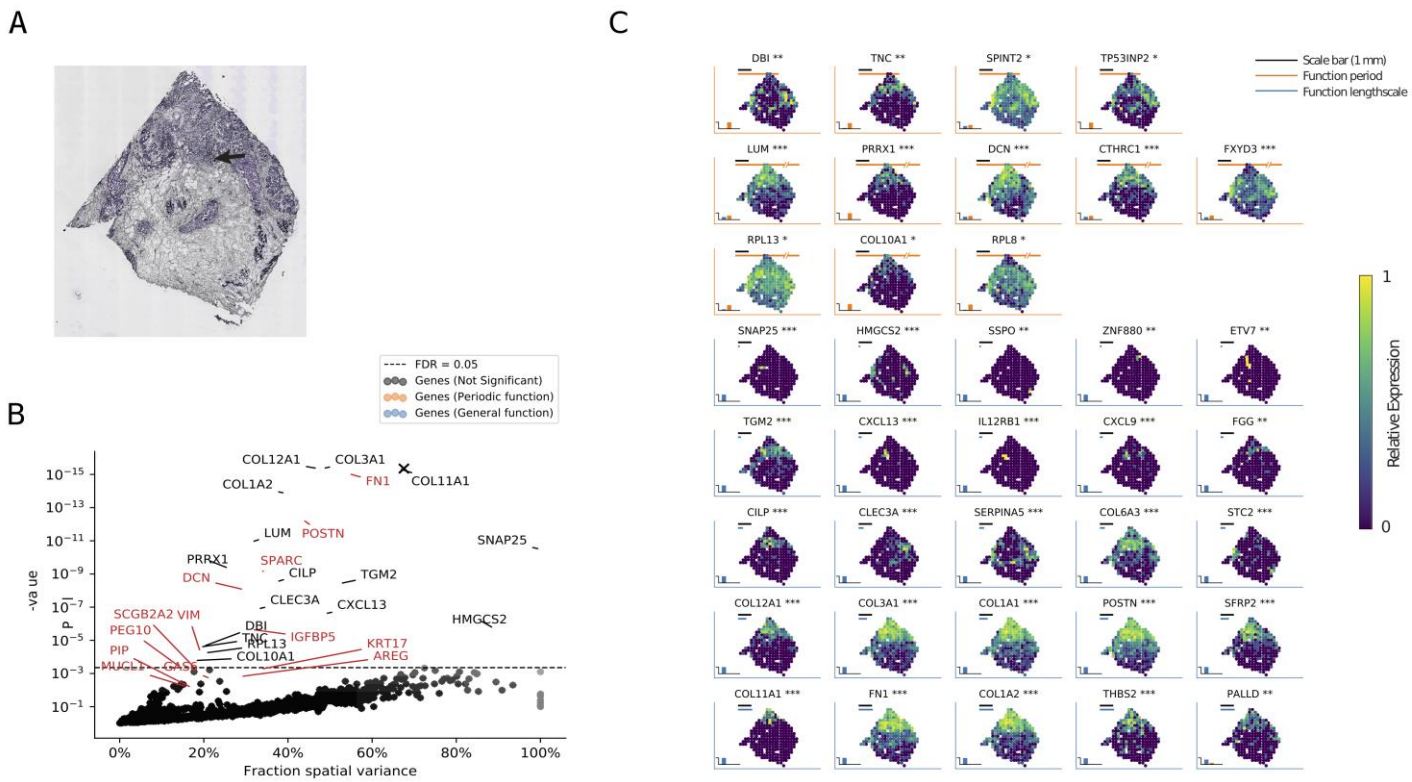
expression level of the inferred spatial pattern. The number of genes assigned to each pattern is indicated in the title of each panel, with representative gene assignments listed below. **(B)** Same as **A** but considering AEH applied to spatial transcriptomics data from breast cancer biopsy. **(C)** As in **A** but for SeqFISH data from mouse hippocampus.



Supplementary Figure 5

Comparison to clustering analysis for the mouse olfactory bulb data.

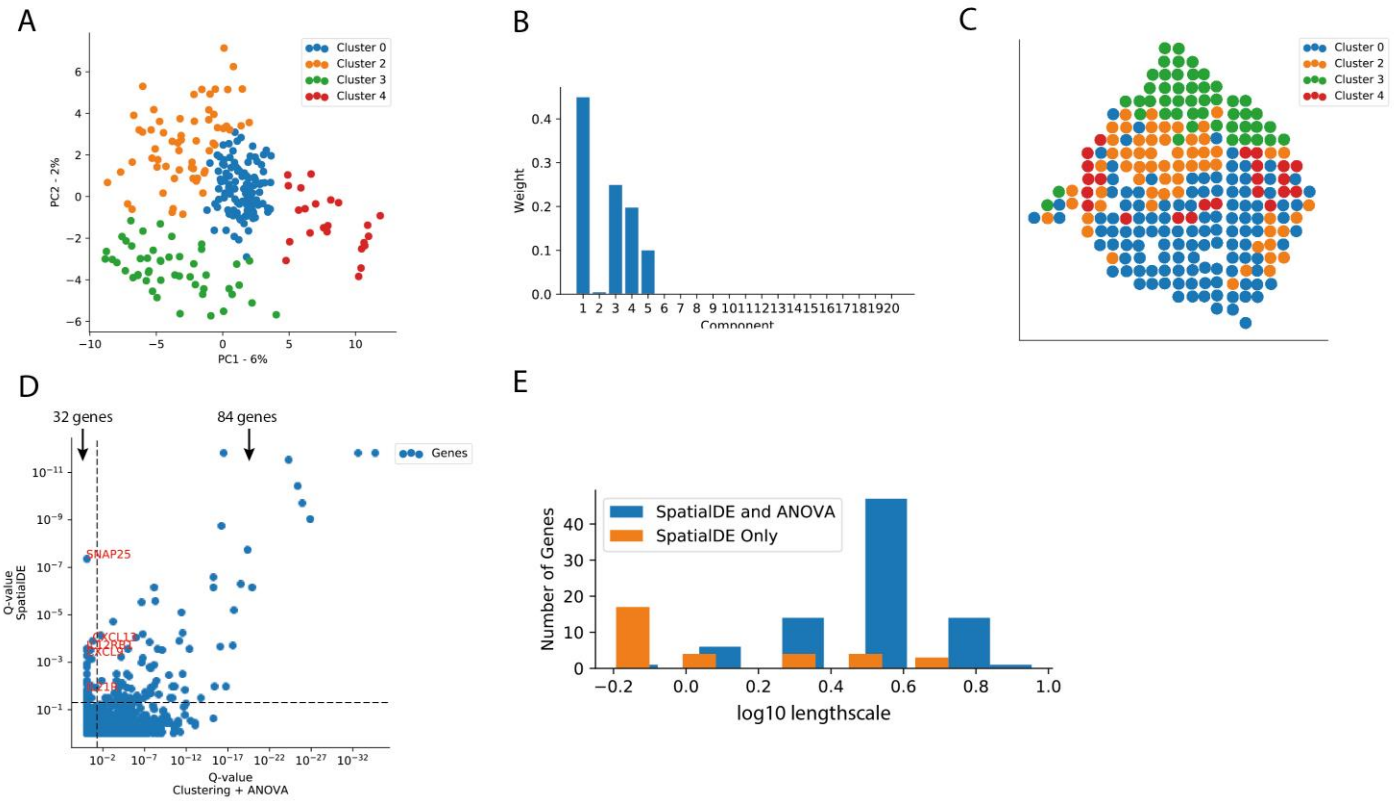
(A) Principal component analysis using genome-wide expression profiles of individual “spots” for the mouse olfactory bulb data. The spots are color coded by the cluster assignment from conventional Bayesian Gaussian Mixture Modelling (K=4 clusters). **(B)** Bayesian Gaussian Mixture Model cluster assignment probabilities, discretizing the 260 spatial “spots” into four clusters (analysis ignores spatial structure). **(C)** Visualization of cluster membership in spatial context of the tissue. **(D)** Scatter plot of negative log P-values from an ANOVA test between clusters (x-axis) versus negative log P-values from the SpatialDE test. 52 genes were identified by both methods, while 15 genes were unique to SpatialDE. **(E)** Comparison of clusters of spots versus AEH patterns.



Supplementary Figure 6

SpatialDE applied to breast cancer biopsy tissue.

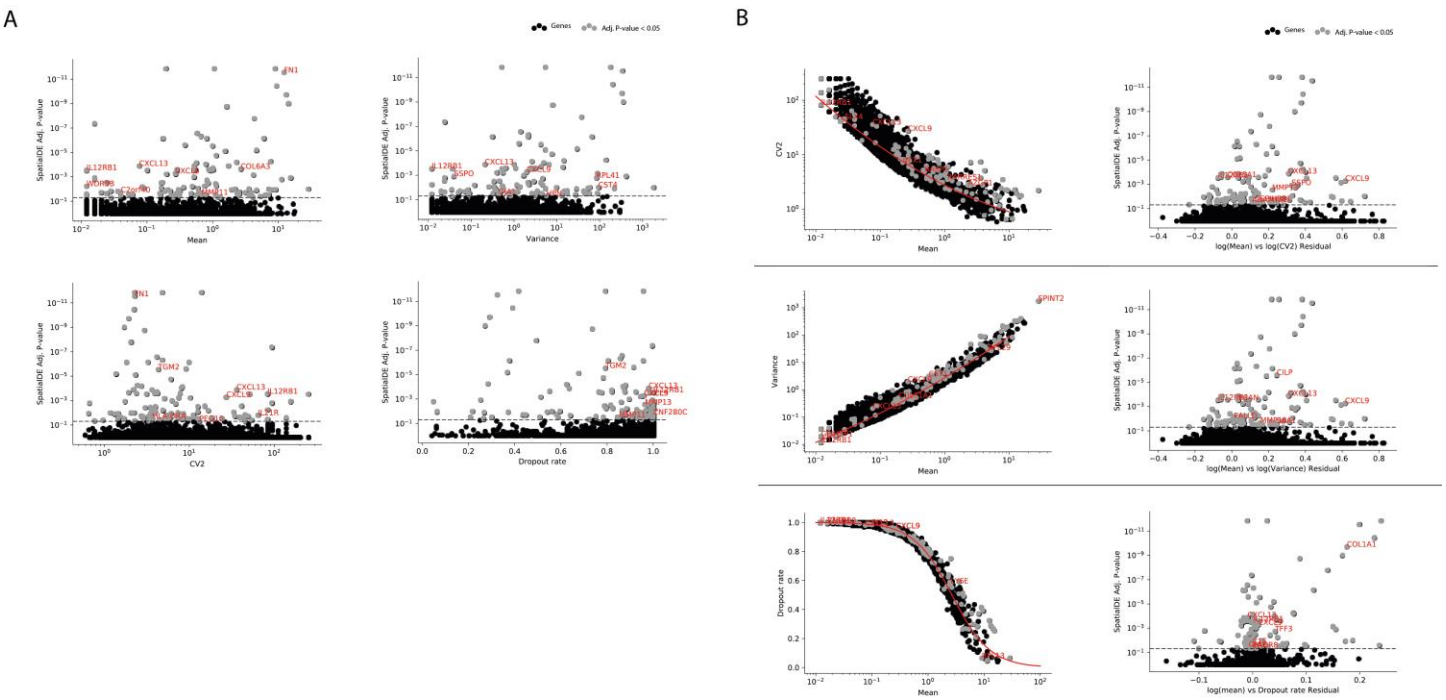
(**A**) Corresponding HE image of breast cancer tissue from spatial transcriptomics. (**B**) Fraction of variance explained by spatial variation (x-axis) versus SpatialDE negative log P-value (y-axis) for all genes. Dashed line corresponds to the FDR=0.05 significance threshold (N=115 genes, Q-value adjusted). Genes classified as periodically variable are shown in orange (N=22), genes with a general spatial dependency in blue (N=93). Disease-implicated genes annotated based on prior knowledge (Stahl et al.[6]) are highlighted with red labels. Other representative genes are annotated with black labels. The X symbol shows the result of applying SpatialDE to the estimated total RNA content per spot. (**C**) Visualization of 37 selected spatially variable genes with different periods and length scales. The black scale bar corresponds to 1 mm. Colors and labels as in (**Fig. 2**). Stars next to gene names denote significance levels (* FDR < 0.05, ** FDR < 0.01, *** FDR < 0.001, Q-value adjusted) of spatial variation.



Supplementary Figure 7

Comparison to differential expression analysis using unsupervised clustering of spots.

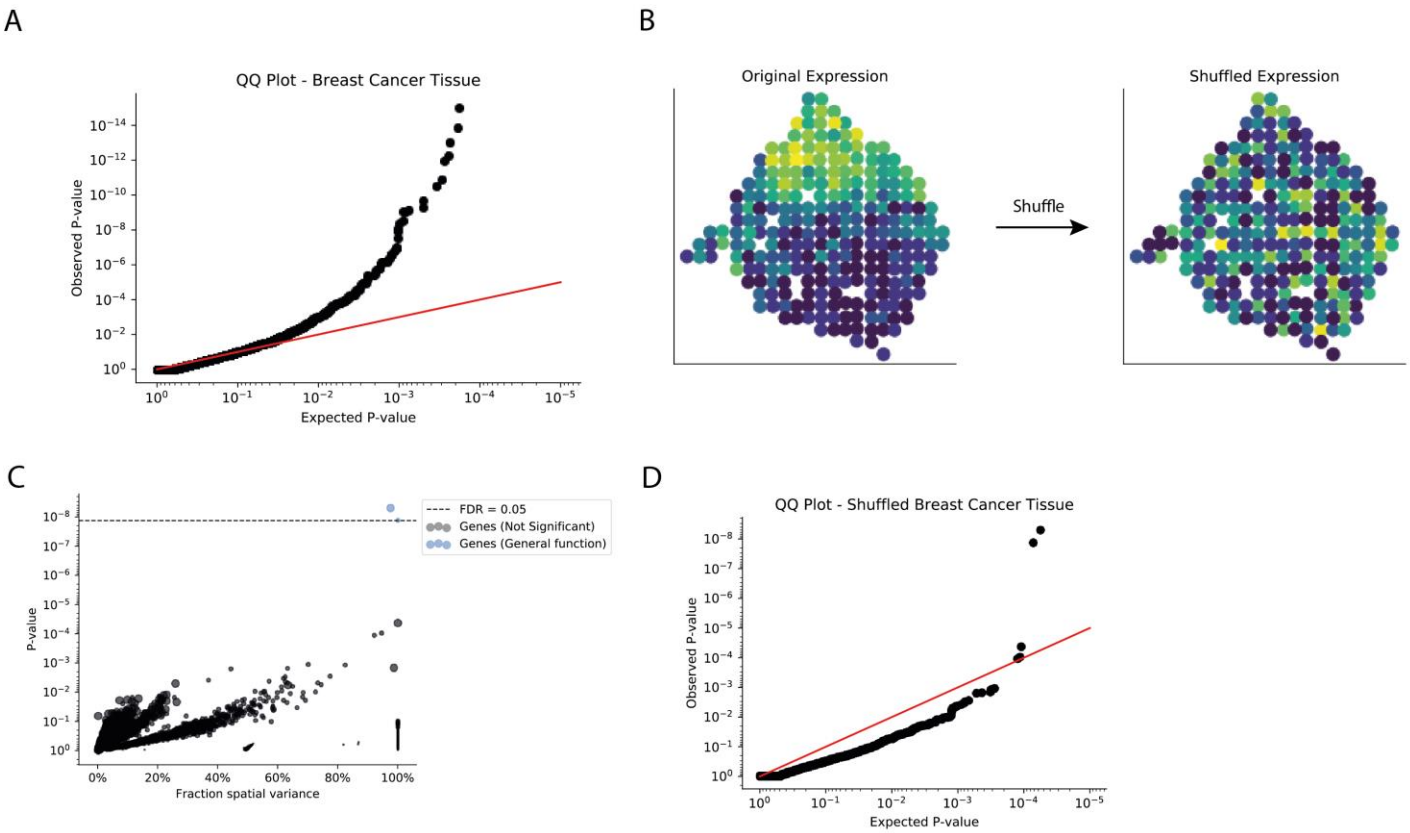
(A) Principal component analysis using genome-wide expression profiles of individual "spots" from the spatial transcriptomics breast cancer data, color coded by cluster membership for K=4 clusters (using Bayesian Gaussian Mixture Modelling). **(B)** Bayesian Gaussian Mixture Model cluster probabilities, discretizing the 250 spatial breast cancer "spots" into four clusters (analysis ignores spatial structure). (**C**) Visualization of cluster membership in the original tissue context. (**D**) Scatter plot of negative log P-values from an ANOVA test between clusters (x-axis) versus negative log P-values of spatial variation from SpatialDE (y-axis). 83 genes were identified as significantly variable by both approaches (FDR<0.05, Q-value adjusted); 32 genes are significant only in the SpatialDE test, among them immune genes. (**E)** Histogram of the fitted length scales for SV genes detected by both approaches (blue) and SV genes exclusively detected by SpatialDE (orange). Genes that were only detected by SpatialDE were associated with smaller length scales, indicating localized expression patterns.



Supplementary Figure 8

Comparison of SpatialDE to alternative measures of expression heterogeneity in the breast cancer tissue.

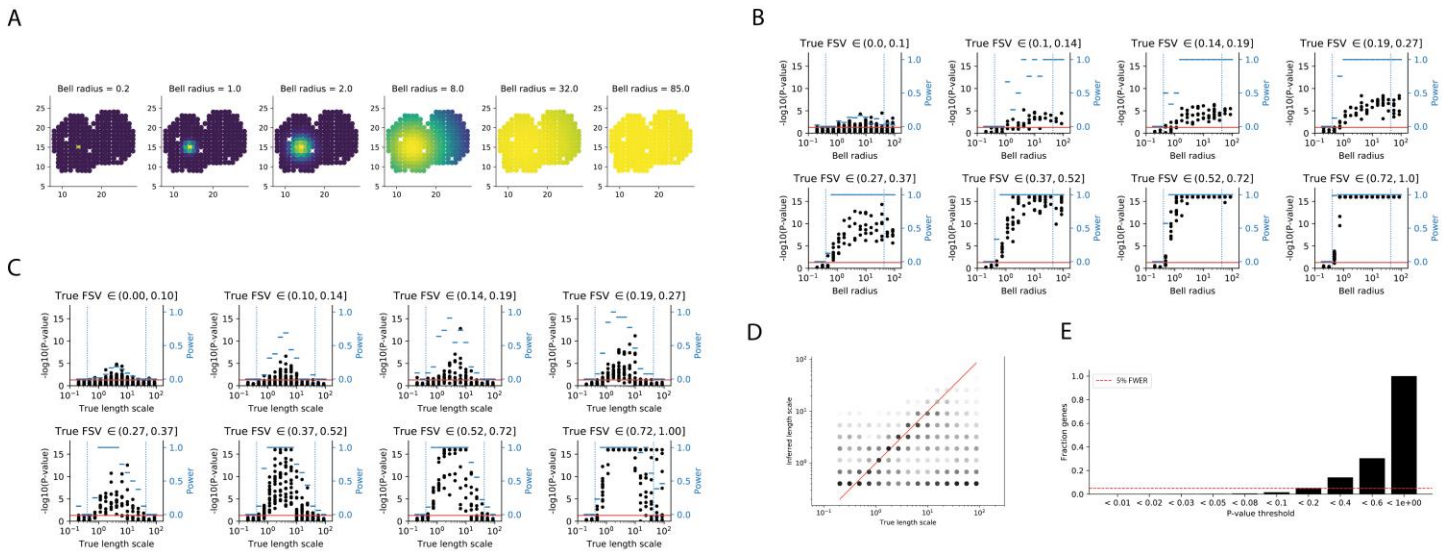
(**A**) Comparison of adjusted negative log P-values from the SpatialDE test (y-axis) versus commonly used statistics for expression heterogeneity (x-axis) - Upper left: Mean, Upper right: Variance, Lower left: CV2 (squared coefficient of variation), Lower right: Dropout rate (fraction of cells/samples a gene is not detected in). Random selection of significant SV genes highlighted in red for context. No dependence between SpatialDE significance levels and expression level (mean) or variance was observed. Statistics calculated for 12,856 genes using 250 “spots”. (**B**) Comparison of significance of SV for genes identified by SpatialDE versus commonly used strategies for defining highly variable genes based on regression models between summary statistics: Relation with CV2 (Upper) or Variance (Middle), or with dropout fraction (Bottom). The rightmost column of plots show residuals compared with the SpatialDE significance; polynomial regression for CV2 and Variance, logistic regression for dropout rate. Significant SV genes as identified by SpatialDE (FDR<0.05, Q-value adjusted) are shown in grey. Other, non-significant genes are shown in solid black. The SV genes significance are orthogonal to HVG measures, indicating that spatial variation is different from general variability. Statistics calculated for 12,856 genes using 250 “spots”.



Supplementary Figure 9

Assessment of statistical calibration of SpatialDE through data randomization.

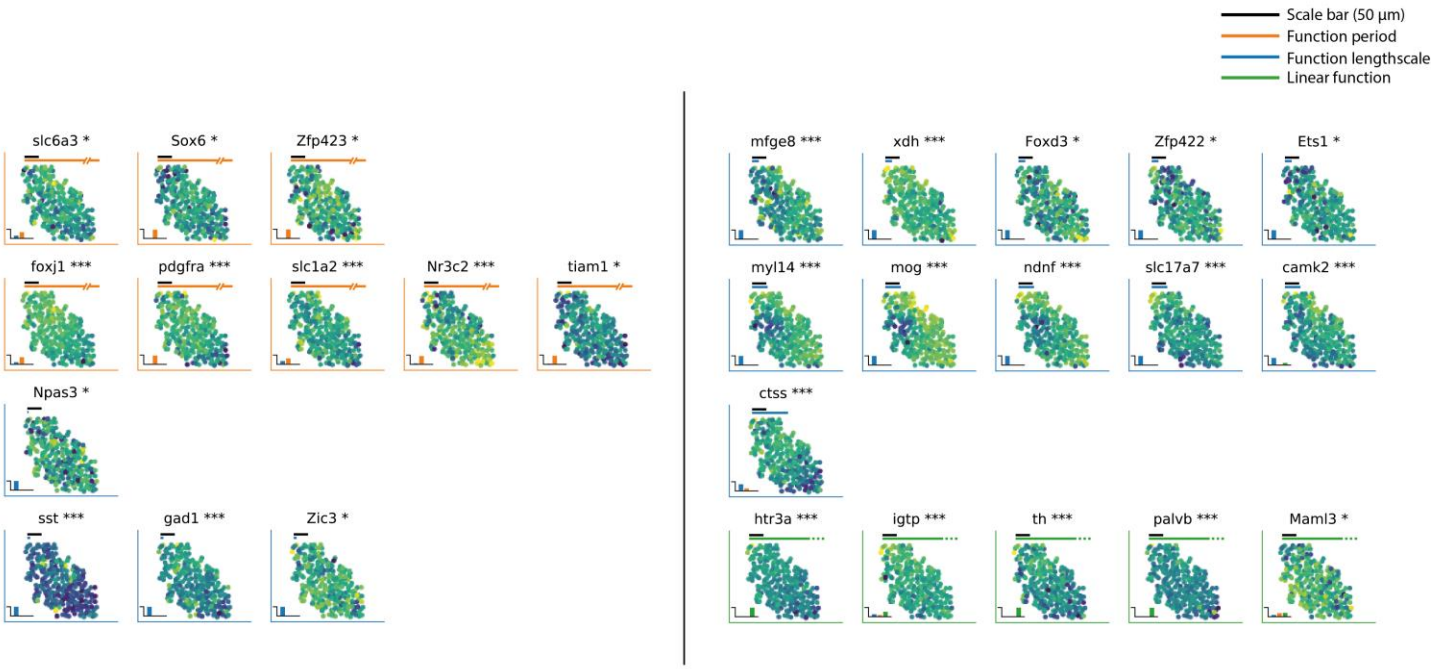
(**A**) QQ-plot of expected P-values (Chi2 distribution with 1 degree of freedom) versus observed P-values from the SpatialDE tests on the breast cancer data. (**B**) To simulate data from an empirical null, without spatial structure, expression values were shuffled among the sampled coordinates. Shown is the resulting expression pattern for *COL3A1* expression, as representative example. (**C**) SpatialDE negative log P-values for genes on shuffled data, with the number of detected SV genes (FD<0.05) being consistent with the selected false discovery rate. (**D**) Analogous QQ-plot as in **A** on shuffled expression values. SpatialDE P-values follow the null distribution, indicating that the model is calibrated.



Supplementary Figure 10

Assessing statistical calibration of SpatialDE through simulations.

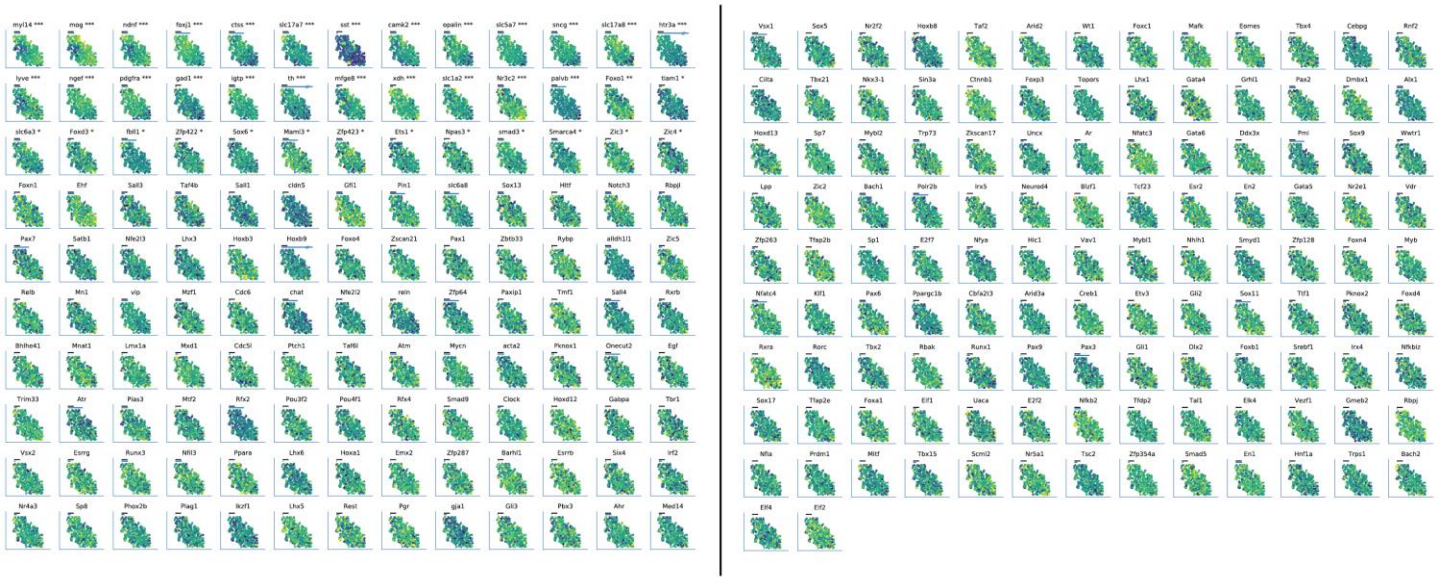
(**A-B**) Simulation of bell curve shaped data on the mouse olfactory bulb coordinates. (**A**) Example of six bell curves with different radii. (**B**) Results from applying SpatialDE to data from 3,000 bell curves with different radii and stratified over different levels of simulated noise (fraction of spatial variance, FSV). Red line denotes P=0.05 significance level, vertical blue dotted lines indicate smallest and largest pairwise distances observed in mouse olfactory bulb data respectively. Black dots denote negative log P-values (left axis), while blue dashes indicating statistical power for detecting true simulated SV genes (fraction true positives) for each bell radius (right axis). (**C**) Analogous results as in **B**, however when simulating data from the generative model underlying SpatialDE, considering different values for the fraction of spatial variance (FSV) and length scales, considering 3,000 simulations. (**D**) Scatter plot of inferred length scales (y-axis) versus simulated length scale (x-axis) for the data shown in **C**. (**E**) Simulation of 3,000 genes from the null model with no spatial covariance. Bars denote the fraction of (false positive) SV genes for different SpatialDE P-value thresholds (x-axis). The proportion of false positive genes was lower than the controlled family-wise error rate (FWER), indicating that the test is conservative.



Supplementary Figure 11

Expanded examples of spatially variable genes for the mouse hippocampus data set.

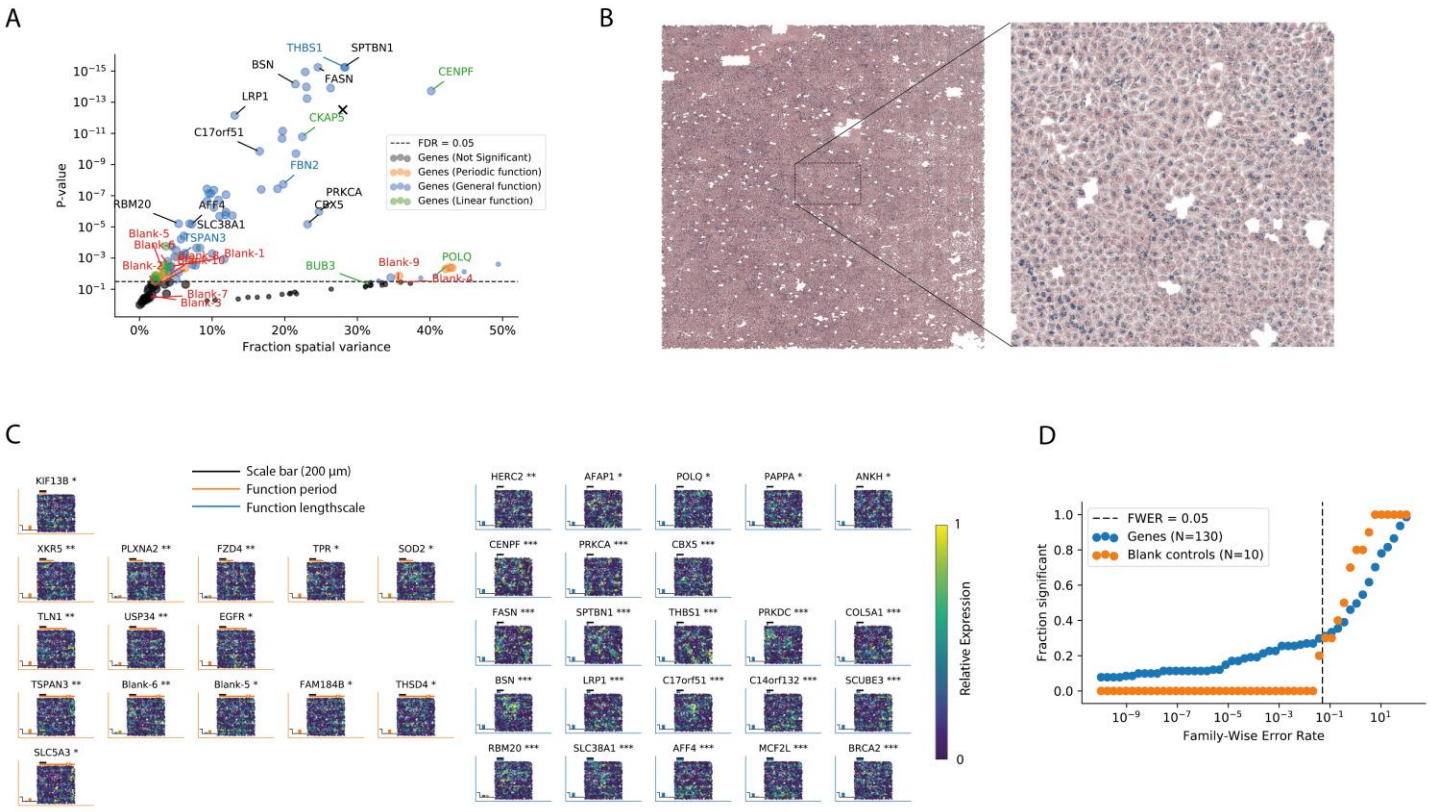
Visualization of 28 SV genes (out of 249, FDR<0.05, Q-value adjusted) from the mouse hippocampus SeqFISH data, displaying selected genes with periodic, linear, and general spatial dependencies with different estimated length scales. Black scale bar correspond to 50 µm. Colors and labels as in **Fig. 2**, stars next to gene names denote significance levels (* FDR < 0.05 , ** FDR < 0.01, *** FDR < 0.001, Q-value adjusted).



Supplementary Figure 12

Visual inspection of genes from seqFISH data.

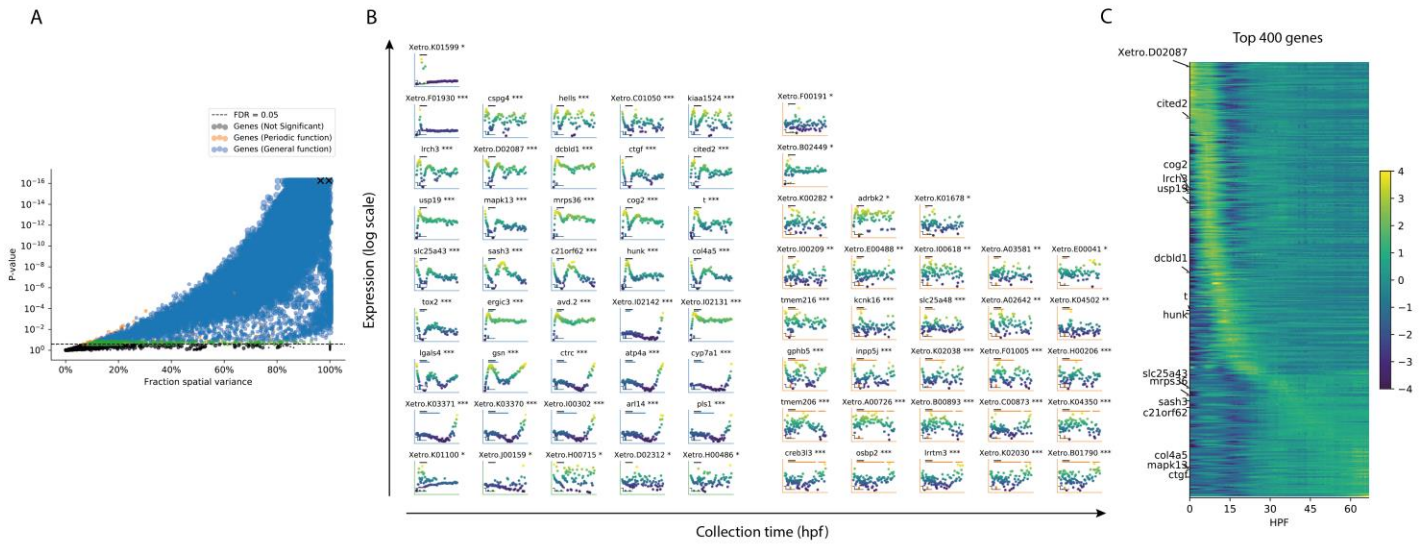
All 249 genes measured in the SeqFISH data. Plots are ordered left to right then top to down in order of decreasing significance of spatial variation using the SpatialDE test (in two columns, left column first). Stars next to gene names denote significance levels after correcting for multiple testing (* FDR < 0.05 , ** FDR < 0.01, *** FDR < 0.001, Q-value adjusted).



Supplementary Figure 13

Application to MERFISH data.

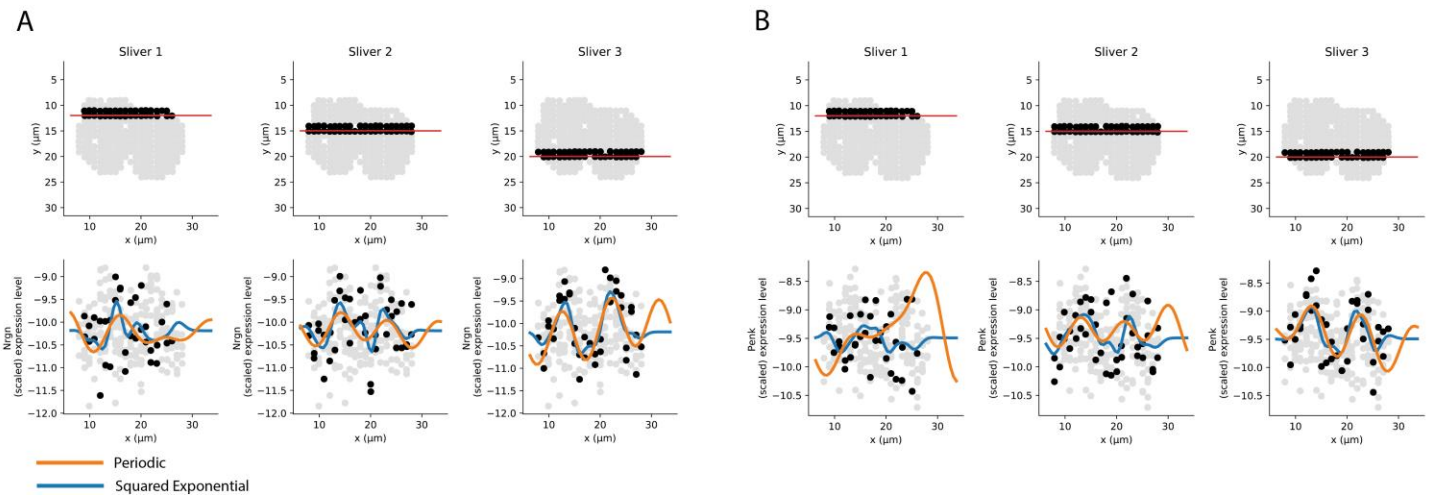
(**A**) In a MERFISH study of an osteosarcoma cell culture of 139 probes from Moffitt *et al*, the SpatialDE test identified the majority of the probes as spatially variable (66%, FDR<0.05, dashed line, Q-value adjusted). 21 of 92 significant SV genes were assigned to a periodic function by the model, and nine genes had linear functions. Red labels indicate negative control probes. Genes indicated as enriched in proliferating cells in the original study are marked in green, and depleted genes in blue. (**B**) Visualization of the MERFISH data by plotting general RNA probes in pink and MALAT1 probes in blue on two 512 x 512 virtual pixel grids at different scales. The original imaged region was 5.2 mm wide and 8.2 mm high totaling 38,594 cells (upper). We analyzed a region of 1 mm x 1 mm in the middle of the cell culture with 1,056 cells (lower). (**C**) Expression levels in the cell culture region visualized for selected SV genes with various fitted periods and length scales (significance levels and colors as in **Fig. 2**). Black scale bar correspond to 200 µm. Stars next to gene names denote significance levels (* FDR < 0.05 , ** FDR < 0.01, *** FDR < 0.001, Q-value adjusted). (**D**) Fraction of gene probes and control probes detected as significant SV genes as a function of the family-wise error rate (FWER). The number of significant control probes was in line with the FWER.



Supplementary Figure 14

Application to a gene expression time-course data set.

(**A**) SpatialDE applied to a developmental time course (89 time points from Owens *et al*), identifying the majority of genes as spatially variable (21,009 out of 22,256 genes, FDR < 0.05, Q-value adjusted). Of these, 241 were assigned to periodic patterns, and 269 were detected with linear trends. Colors and point sizes as in **Fig. 2**. The X marks indicates result of running SpatialDE on RNA spike-in content and the number of detected genes, proxies for the RNA content in the embryos. (**B**) Examples of temporally variable genes of various periods and length scales. Black scale bar corresponds to 12 hours in the time-course, periods and length scales of functions are indicated relative to this. Collection time in units of hours post fertilization (hpf). Stars next to gene names denote significance levels (* FDR < 0.05 , ** FDR < 0.01, *** FDR < 0.001, Q-value adjusted). (**C**) The expression patterns of the top 400 significantly SV genes are visualized, ordered by the time they reach their highest expression value. Example genes from **B** are annotated.



Supplementary Figure 15

Subtle noise can affect spatial function classification.

(**A**) *Nrgn* which is assigned as periodic with posterior probability 1. Red line in top row indicates where in spatial coordinates the Gaussian processes are predicted in the bottom row, with a sliver of close by points colored in black. Bottom row shows expression level of all “spots”, with “spots” close to the predicted curves in black for spatial context. (**B**) Same as (**A**) but for *Penk* which have posterior probability 0.12 of being periodic. Noise at the right side of “Sliver 1” cause the periodic covariance to not fit the data as well as squared exponential covariance.

SPATIALDE SUPPLEMENTARY NOTE 1

VALENTINE SVENSSON, SARAH A TEICHMANN, OLIVER STEGLE

Contents

Date: February 22, 2018.

1. SpatialDE model

SpatialDE builds on the Gaussian process framework, assessing the evidence that the gene expression patterns of individual genes are explained by functions with different spatio-temporal dependencies.

In the following we assume that $\mathbf{y} = (y_1, \ldots, y_N)$ corresponds to a vector of expression values of a given gene at N spatial locations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$. The coordinates of the spatial locations are typically two-dimensional, i.e. $\mathbf{x}_i = (x_{i_1}, x_{i_2})$, however the model is general and can also be applied to data wither other dimensionality, such as three-dimensional or uni-dimensional (e.g. time-series) data.

1.1. Gaussian Processes regression.

A Gaussian Process (GP) is a probability distribution over functions $y = f(\mathbf{x})$,

$$f \sim \mathcal{GP}(k\left(\mathbf{x}, \mathbf{x}' \,|\, \boldsymbol{\theta}_k\right)), \tag{1}$$

which is defined by a covariance function $k(\mathbf{x}, \mathbf{x}' \,|\, \boldsymbol{\theta}_k)$ that parameterizes the dependency between any pair of function values f and f' based on their inputs $\mathbf{x}$ and $\mathbf{x}'$; and $\boldsymbol{\theta}_k$ denotes a vector of additional hyperparameters of the covariance function (see below).

While a GP defines a distribution on functions with infinite dimension, a finite representation of a GP for observed data can be obtained by marginalising over all unobserved function values. This results in a realisation of joint Gaussian distributions for a vector of function values $\mathbf{f}$:

$$p(\mathbf{f} \,|\, \mathcal{H}_{\text{GP}}, \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{f} \,\middle|\, \mathbf{0}, \sigma_s^2 \cdot \boldsymbol{\Sigma}_{k\left(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta}_k\right)}\right), \tag{2}$$

where covariance matrix $\boldsymbol{\Sigma}_{k\left(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta}_k\right)}$ is derived by evaluating the covariance function for all pairs of observed datums $\boldsymbol{\Sigma}_{k\left(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta}_k\right)_{i,j}} = k(\mathbf{x}_i, \mathbf{x}_j \,|\, \boldsymbol{\theta}_k)$, which can depends on hyperparameters $\boldsymbol{\theta}_k$ of the covariance function. We have factored out the scaling of the covariance σ_s^2, which determines the magnitude of total variance. Later, we will introduce a parametrisation of $\boldsymbol{\Sigma}_{k\left(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta}_k\right)}$ such that σ_s^2 can be interpreted as the variance explained by spatial effects.

Assuming Gaussian distributed noise, i.e. $p(y_n \,|\, f_n) = \mathcal{N}\left(y_n \,\middle|\, \mu, f_n, \sigma_e^2\right)$, and marginalising over the function values f_n results in multivariate normally distributed marginal likelihood of the observed expression values,

$$p(\mathbf{y} \,|\, \mathcal{H}_{\text{GP}}, \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \mu \cdot \mathbf{1}, \sigma_s^2 \cdot \left(\boldsymbol{\Sigma}_{k\left(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta}_k\right)} + \delta \cdot \mathbf{I}\right)\right), \tag{3}$$

where we have defined $\delta = \frac{\sigma_e^2}{\sigma_s^2}$ for reasons that will become clear later. The fixed effect $\mu \cdot \mathbf{1}$ models an offset or mean expression expression level for a given gene.

The resulting total covariance in (3) decomposes into two components, where $\sigma_2^2 \cdot \boldsymbol{\Sigma}_{k\left(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta}_k\right)}$ can be thought of as explaining spatial variation in the data, whereas $\sigma_s^2 \cdot \delta \cdot \mathbf{I}$ represent the non-spatial residual variation. The full set of model parameters, $\boldsymbol{\theta} = (\boldsymbol{\theta}_k, \sigma_s^2, \delta, \mu)$ can be determined using maximum likelihood (see Section 1.4):

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, P(\mathbf{y} \,|\, \mathcal{H}_{\text{GP}}, \mathbf{X}, \boldsymbol{\theta}). \tag{4}$$

1.2. **Covariance functions.** To find significant spatial variation and to compare between alternative hypothesis of spatial variation for expression patterns, we assess GP model with different covariance functions. An overview of many covariance functions can be found in [Williams and Rasmussen, 2006]. In this work we make use of the following:

- Null model: no spatial variance component
  $k_{\text{null}}(\mathbf{x}, \mathbf{x}') \propto 0$

- General spatial pattern (known as the *squared exponential* (*Gaussian*) covariance function)
  $k_{\text{spatial}}(\mathbf{x}, \mathbf{x}' \,|\, l) \propto e^{-\frac{1}{2l^2}|\mathbf{x}-\mathbf{x}'|^2}$

- Linear covariance function
  $k_{\text{lin}}(\mathbf{x}, \mathbf{x}' \,|\, p) \propto \mathbf{x}{\mathbf{x}'}^{\mathrm{T}}$

- Periodic covariance function (*cosine* covariance function)
  $k_{\text{periodic}}(\mathbf{x}, \mathbf{x}' \,|\, p) \propto \cos(\frac{\pi}{p}|\mathbf{x} - \mathbf{x}'|)$

Interpretation of model parameters. In order to estimate the proportion of variance explained by spatial covariance, we use Gower's transformation to adjust for the sample variance explained by $\mathbf{\Sigma}_{k(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta}_k)}$ [Kostem and Eskin, 2013]:

$$g = \frac{\mathrm{Tr}(P\mathbf{\Sigma}_{k(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta}_k)}P)}{n-1},$$

where

$$P = I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}.$$

This allows for defining the Fraction of Spatial Variance, $\text{FSV} = \frac{\sigma_s^2 \cdot g}{\sigma_s^2 \cdot g + \sigma_s^2 \cdot \delta} = \frac{g}{g+\delta}$, which corresponds to the fraction of the total variance explained by spatial variance.

1.3. **Statistical significance and classification of spatially variable genes.**

P-values from hypothesis testing. The significance of the spatial variance component is assessed via model comparison:

$$p(\mathbf{y} \,|\, \mathcal{H}_1, \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \mu \cdot \mathbf{1}, \sigma_s^2 \cdot \left(\mathbf{\Sigma}_{k(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta}_k)} + \delta \cdot \mathbf{I}\right)\right),$$
$$p(\mathbf{y} \,|\, \mathcal{H}_0, \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \mu \cdot \mathbf{1}, \sigma^2 \cdot \mathbf{I}\right).$$

Here, $\mathcal{H}_1$ denotes the alternative model, which includes both a spatial and non-spatial component, and $\mathcal{H}_0$ denotes the null model without a spatial variance component. These two models are nested, allowing estimation of the significance of the spatial variance component using a log-likelihood ratio (LLR) test,

$$LLR = 2 \cdot \log \frac{p(\mathbf{y} \,|\, \mathcal{H}_1, \mathbf{X}, \hat{\boldsymbol{\theta}_1})}{p(\mathbf{y} \,|\, \mathcal{H}_0, \mathbf{X}, \hat{\boldsymbol{\theta}_0})}. \tag{5}$$

P-values can be estimated in closed form, assuming that the log-likelihood ratios under the null model are χ^2 distributed with one degree of freedom.

We employ the Q-values procedure by [Storey and Tibshirani, 2003] to adjust for multiple testing. Unless stated otherwise, we report genes at FDR < 0.05 as significantly spatially variable.

Calibration of the P-values was confirmed using permutation experiments (**Supp. Fig. 9**) and data simulated from the null model (**Supp. Fig. 10E**). As additional assessment of the empirical calibration of the method, we considered negative control probes reported in the MERFISH data, confirming that the fraction of significant negative control probes is in line with the expectation at a given family-wise error rate (**Supp. Fig. 13D**).

Classification of spatial expression using model comparison. In order to characterise the functional form of spatial trends, SpatialDE allows for model comparisons with GP models that make stronger assumptions about the spatial dependency. Specifically, SpatialDE supports model comparisons between GP models with alternative prior covariance functions: the general spatial model using a squared exponential covariance function; a GP prior with periodic covariance function, using the cosine kernel (See Section 1.2); and finally a GP prior with linear covariance function.

As these models differ in their number of parameters, we employ the Bayesian Information Criterion (BIC), which has been shown to be effective for model comparisons of alternative GP models [Lloyd et al., 2014]. Notably, a GP is typically defined by a small number of hyperparameters, because the latent function values f_n are marginalised out, which means that the log marginal likelihood is a good approximation for the model evidence. BIC penalises the maximum log-likelihood by the number of parameters that are optimised, thereby accounting for differences in model complexity:

$$BIC = \log(N) \cdot M - 2 \cdot \hat{LL}.$$

Here, $\hat{LL}$ denotes the log marginal likelihood (Eq. 4), M corresponds to the number of observations and N denotes the number of hyperparameters of a given model. Each gene is then classified into different spatial trends by selecting the GP model that minimises the BIC.

We also use the BIC to estimate posterior probabilities of specific models. Briefly, the BIC is an estimate of $-\log p(\mathbf{y} \mid \mathcal{H}_i, \mathbf{X})$, which allows for deriving an approximate form of the marginal likelihood of a given model $\mathcal{H}_i$,

$$\begin{aligned} p(\mathcal{H}_i \mid \mathbf{y}, \mathbf{X}) &= \frac{1}{Z} \cdot p(\mathbf{y} \mid \mathcal{H}_i, \mathbf{X}) \cdot p(\mathcal{H}_i) \\ &= \frac{1}{Z} \cdot \int_{\boldsymbol{\theta}} p(\mathbf{y} \mid \mathcal{H}_i, \mathbf{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\theta \; p(\mathcal{H}_i) \\ &\approx -\frac{1}{Z} \cdot BIC_i, \end{aligned} \tag{6}$$

where

$$Z = \sum_i p(\mathbf{y} | \mathcal{H}_i, \mathbf{X}) \cdot p(\mathcal{H}_i) \approx \sum_i -BIC_i,$$

assuming a uniform prior over models ($p(\mathcal{H}_i)$). In the analysis, we consider the models $\{\mathcal{H}_{\text{spatial}}, \mathcal{H}_{\text{linear}}, \mathcal{H}_{\text{periodic}}\}$ (see Section 1.2), and estimate marginal posterior probabilities for all models using the BIC approximation.

Interpreting characteristic length scales and periods. The parameter l of the squared exponential covariance function corresponds to the prior *characteristic length scale* of the underlying function. In the Gaussian process regression models, this parameter determines the distance at which observations stop covarying. As a concrete example, if we put $l = 1.0$ then observations 0.66 units apart will have covariance

0.8, observations 1.2 units apart will have covariance 0.5, and observations 2.0 units apart will have covariance 0.13. It is important to realise that for any given length scale, the covariance will only be strong within a region of distance, which is a fraction of the selected length scale. For example, for observations to have spatial covariance of at least 0.9, they can only be $0.46 \cdot l$ units of distance apart.

A Gaussian process with a squared exponential covariance functions can capture any smooth function (infinitely differentiable) with constant characteristic length scale, including linear and periodic functions. As such, Gaussian processes with linear or periodic covariance functions are special cases with stronger assumptions on the functional forms.

In the periodic covariance function, observations very close to each other covary, observations at medium distances have negative covariance, and further apart observations start positively covarying again. For example, if we put the period $p = 1.0$, then observations 1.0 units apart will have covariance -1.0, while observations 2.0 units apart will have covariance 1.0. For observations to (locally) have covariance at least 0.9 they can only within $0.14 \cdot p$ units apart. But this will also enforce a repeating pattern globally. A periodic Guassian process can describe two different things, depending on the scale of the period: either a regularly repeating pattern, or curves which decrease (or increase) throughout the domain in a sigmoidal fashion. The latter functional form will take effect when the period is about twice the length of the data. See **Supp. Note Fig. 1A-C** for examples of functions sampled from different length scales and curves.

Cases where data fit the sigmoidal patterns suggested by long-perioded periodic covariances happen in practice, as for example with the gene Foxj1 in the SeqFISH dataset shown in **Figure 2G**. **Supp. Note Fig. 1D** shows the predicted mean effect of the periodic and linear Gaussian processes along three projections of the 2D, showing a periodic-sigmoidal functions fits the data better than a linear plane.

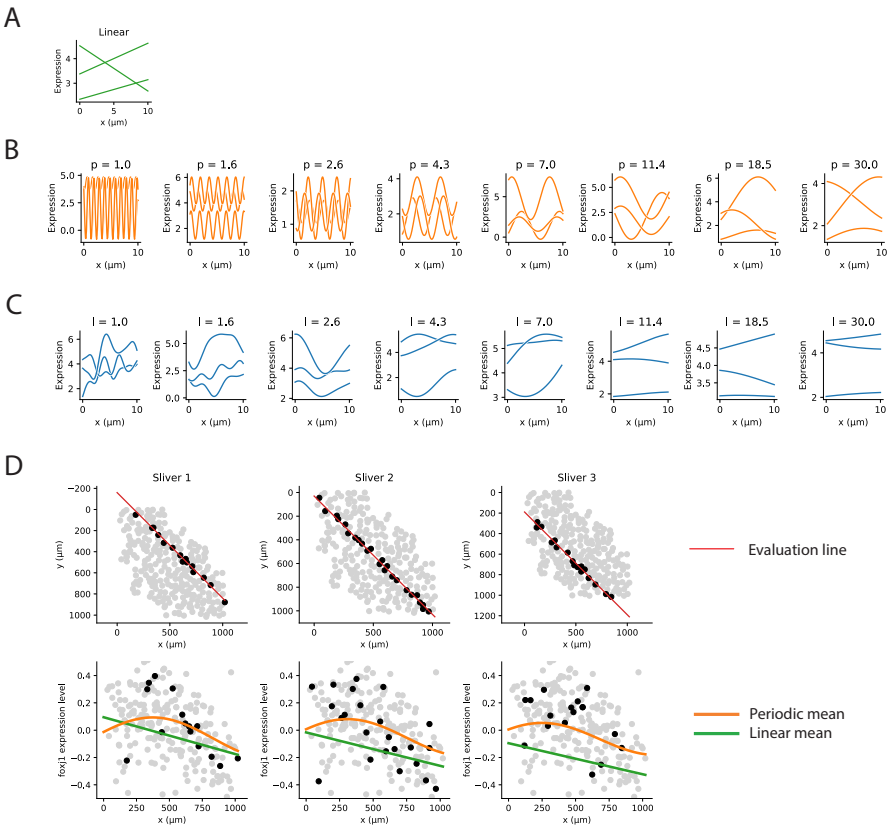
1.4. **Parameter inference.** Maximum likelihood inference (Eq. 4), requires determining $\boldsymbol{\theta} = (\mu, \sigma_s^2, \delta, \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$ denotes additional covariance-specific parameters (e.g. the length-scale l, see Section 1.2).

The log likelihood is

$$LL(\mathbf{y} \mid \mathcal{H}_{\text{GP}}, \mathbf{X}, \boldsymbol{\theta}) = \log p(\mathbf{y} \mid \mathcal{H}_{\text{GP}}, \mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2} N \cdot \log(2 \cdot \pi)$$
$$-\frac{1}{2} \log \left(\left| \sigma_s^2 \left[\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta}_k)} + \delta \cdot \mathbf{I} \right] \right| \right) - \frac{1}{2} (\mathbf{y} - \mu \cdot \mathbf{1})^{\mathrm{T}} \left(\sigma_s^2 \left[\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta}_k)} + \delta \cdot \mathbf{I} \right] \right)^{-1} (\mathbf{y} - \mu \cdot \mathbf{1})$$

Evaluation of the likelihood requires inverting the covariance matrix $\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta}_k)}$, which depend on the parameter $\boldsymbol{\theta}_k$, rendering gradient-based optimisation of $\boldsymbol{\theta}_k$ a key bottleneck in inference ($O(N^3)$). We comment on this later, but for now, assume $\boldsymbol{\theta}_k$ is known.

In statistical genetics, models that use covariance based models to capture population structure have been sped up through spectral decomposition of the covariance, to circumvent inverting the covariance matrix for each evaluation on of the model [Lippert et al., 2011]. Here we adapt this approach for SpatialDE to avoid repeated matrix inversions of $\sigma_s^2 \cdot \left[\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta}_k)} + \delta \cdot \mathbf{I} \right]$, by factoring the covariance matrix with spectral decomposition $\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta}_k)} = \mathbf{U}\mathbf{S}\mathbf{U}^{\mathrm{T}}$. This allows



SUPP. NOTE FIG. 1. **Interpretation of periods and length scales. (A-C)** Simulated noise free 1D data with different covariance functions. **(A)** Three simulations with linear covariance. **(B)** Simulations from periodic covariance functions with different period values. **(C)** Simulations from squared exponential covariance function with different characteristic length scales. **(D)** Visualization of a gene with very long inferred period length. Top row shows the region where the Gaussian processes are predicted. Bottom row shows expression level on the y-axis instead of the spatial y-coordinate. Curves show expression level prediction from the Gaussian process along the red line indicated in the top panel. Cells in grey or black are spatially closer to the red line.

us to take advantage of the fact that eigenvectors are orthogonal, $\mathbf{U}\mathbf{U}^T = \mathbf{I}$:

$$\sigma_s^2 \cdot (\mathbf{\Sigma}_{k(\mathbf{x},\mathbf{x}' \mid \boldsymbol{\theta}_k)} + \delta \cdot \mathbf{I}) = \sigma_s^2 \cdot (\mathbf{U}\mathbf{S}\mathbf{U}^T + \delta \cdot \mathbf{I}) = \sigma_s^2 \cdot \mathbf{U}(\mathbf{S} + \delta \cdot \mathbf{I})\mathbf{U}^{\mathrm{T}}.$$

Nature Methods: doi:10.1038/nmeth.4636

Now if we write the log likelihood as a function of δ, σ_s^2 and μ, we obtain

$$
\begin{aligned}
LL(\delta, \sigma_s^2, \mu) &= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \\
&\cdot \log(|\mathbf{\Sigma}_{k(\mathbf{x},\mathbf{x}' \mid \boldsymbol{\theta}_k)} + \delta \cdot \mathbf{I}|) - \frac{1}{2 \cdot \sigma_s^2} \cdot (\mathbf{y} - \mu \cdot \mathbf{1})^T (\mathbf{\Sigma}_{k(\mathbf{x},\mathbf{x}' \mid \boldsymbol{\theta}_k)} + \delta \cdot \mathbf{I})^{-1} (\mathbf{y} - \mu \cdot \mathbf{1}) \\
&= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot \log(|\mathbf{U}(\mathbf{S} + \delta \cdot \mathbf{I})\mathbf{U}^T|) - \frac{1}{2 \cdot \sigma_s^2} \cdot (\mathbf{y} - \mu \cdot \mathbf{1})^T (\mathbf{U}(\mathbf{S} + \delta \cdot \mathbf{I})\mathbf{U}^T)^{-1} (\mathbf{y} - \mu \cdot \mathbf{1}) \\
&= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot \log(|\mathbf{U}| \cdot |\mathbf{S} + \delta \cdot \mathbf{I}| \cdot |\mathbf{U}^T|) - \frac{1}{2 \cdot \sigma_s^2} \cdot (\mathbf{y} - \mu \cdot \mathbf{1})^T \mathbf{U} (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} \mathbf{U}^T (\mathbf{y} - \mu \cdot \mathbf{1}) \\
&= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot \log(|\mathbf{S} + \delta \cdot \mathbf{I}|) - \frac{1}{2 \cdot \sigma_s^2} \cdot ((\mathbf{U}^T \mathbf{y}) - (\mathbf{U}^T \mathbf{1}) \cdot \mu)^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} ((\mathbf{U}^T \mathbf{y}) - (\mathbf{U}^T \mathbf{1}) \cdot \mu) \\
&= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot \sum_{i=1}^{N} \cdot \log(\mathbf{S}_{i,i} + \delta) - \frac{1}{2 \cdot \sigma_s^2} \cdot \sum_{i=1}^{N} \frac{([\mathbf{U}^T \mathbf{y}]_i - [\mathbf{U}^T \mathbf{1}]_i \cdot \mu)^2}{\mathbf{S}_{i,i} + \delta})
\end{aligned}
$$

The key features used are that $|\mathbf{U}| = |\mathbf{U}^T| = 1$, and $\mathbf{S} + \delta \cdot \mathbf{I}$ is a diagonal matrix, so both its determinant and inverse are trivial to compute (in $O(N)$). The expression $\mathbf{U}^T \mathbf{1}$ only depends on the coordinates $\mathbf{X}$ and can be pre-computed and reused for every gene. The expression $\mathbf{U}^{\mathrm{T}} \mathbf{y}$ will need to be re-computed for each gene, however, it can be re-used for function evaluations during parameter inference.

We make use of the constraint that for the optimal $\mu = \hat{\mu}$ we must have

$$
\frac{\partial LL(\delta, \sigma_s^2, \mu)}{\partial \mu} = 0,
$$

and so

$$
\begin{aligned}
&\frac{1}{\sigma_s^2}((\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{y}) - (\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{1}) \cdot \hat{\mu}) = 0 \\
&\quad \Rightarrow (\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{1}) \cdot \hat{\mu} \\
&\quad = (\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{y}) \\
&\quad \Rightarrow \hat{\mu} \\
&\quad = ((\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{1}))^{-1} (\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{y}) \\
&\quad = \left(\sum_{i=1}^{n} \frac{1}{\mathbf{S}_{i,i} + \delta} [\mathbf{U}^T \mathbf{1}]_i^T [\mathbf{U}^T \mathbf{y}]_i \right) / \left(\sum_{i=1}^{n} \frac{1}{\mathbf{S}_{i,i} + \delta} [\mathbf{U}^T \mathbf{1}]_i^T [\mathbf{U}^T \mathbf{1}]_i \right).
\end{aligned}
$$

When data is given, this expression only depends on δ and we write this as $\hat{\mu}(\delta)$.

The same procedure for σ_s^2 results in

$$
\hat{\sigma}_s^2(\delta) = \frac{1}{N} \sum_{i=1}^{N} \frac{([\mathbf{U}^T \mathbf{y}]_i - [\mathbf{U}^T \mathbf{1}]_i \cdot \hat{\mu}(\delta))^2}{\mathbf{S}_{i,i} + \delta},
$$

which also depend only on δ. So, the entire expression for the log likelihood can be written as

$$LL(\delta) = -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot S_1(\delta) - \frac{N}{2} - \frac{N}{2} \cdot \log(\frac{1}{N} S_2(\delta))),$$
$$S_1(\delta) = \sum_{i=1}^{N} \cdot \log(\mathbf{S}_{i,i} + \delta),$$
$$S_2(\delta) = \sum_{i=1}^{N} \frac{([\mathbf{U}^T\mathbf{y}]_i - [\mathbf{U}^T\mathbf{1}]_i \cdot \hat{\mu})^2}{\mathbf{S}_{i,i} + \delta}.$$

To optimise $LL(\delta)$ with respect to δ we use gradient based optimisation with the bounded limited memory Broyden Fletcher Goldfarb Shanno (L-BFGS-B) algorithm and a numerically approximated gradient.

To avoid gradient based optimization of the length scale ℓ, we pre-compute a grid of covariance matrices $\mathbf{\Sigma}_{k(\mathbf{x},\mathbf{x}' \mid \boldsymbol{\theta}_k)}$ and factorise them. The number of grid points can be specified by the user, but our default settings is to put 10 grid points, logarithmically spaced. The boundaries of the grid are by default set as the shortest observed distance, divided by 2, and the longest observed distance multiplied by 2. We have found these grids to give sufficient sensitivity, with very minor improvement from additional grid points. After factoring the $\mathbf{\Sigma}_{k(\mathbf{x},\mathbf{x}' \mid \boldsymbol{\theta}_k)}$ matrices the $\mathbf{U}$ and $\mathbf{S}$ matrices can be pre-computed and reused for each gene. We only need to do as many $O(N^3)$ matrix decompositions as we have grid points. Each gene under investigation will have a $O(N^2)$ step for each grid point to calculate the $\mathbf{U}^T\mathbf{y}$ factor. All other calculations, including each optimisation iteration, will be $O(N)$. Since our aim to investigate data where $G >> 10$, this greatly reduces computational burden, as illustrated in the comparison of the inference in SpatialDE and an implementation in Stan (**Supp. Fig. 2**).

Estimation of standard errors. The FSV defined earlier is the main value of interest from the fitted model and can be expressed as a function of δ,

$$\mathrm{FSV}(\delta) = \frac{\hat{\sigma}_s^2(\delta) \cdot g}{\hat{\sigma}_s^2(\delta) \cdot g + \delta \cdot \hat{\sigma}_s^2(\delta)} = \frac{g}{g + \delta},$$

where g is the Gower factor for covariance matrix $\mathbf{\Sigma}_{k(\mathbf{x},\mathbf{x}' \mid \boldsymbol{\theta}_k)}$ for a given grid point. To estimate the uncertainty of the FSV, we use the rules of Gaussian error propagation. The uncertainty of the maximum likelihood estimate $\hat{\delta}$ of the parameter δ is the inverse of $\frac{\partial^2 LL(\delta)}{\partial \delta^2}$ evaluated at $\hat{\delta}$. So, the standard error of FSV is

$$s^2_{\mathrm{FSV}} = \left(\frac{\partial \mathrm{FSV}(\delta)}{\partial \delta}\Big|_{\delta=\hat{\delta}} \right)^2 \cdot s^2_{\delta},$$

where

$$s^2_{\delta} = {}^1\Big/\left(\frac{\partial^2 LL(\delta)}{\partial \delta^2}\Big|_{\delta=\hat{\delta}} \right)^2.$$

To evaluate the two derivatives, finite difference approximation is used on the LL and FSV functions.

2. Automatic expression histology model

Expression histology can be defined as patterns in tissues determined by co-expressed spatially variable genes. Intuitively, genes activated by the same pattern can be said follow the same spatial function, but perturbed in different ways from the underlying patterns. Let $\{\boldsymbol{\mu}_k\}$ be a collection of K histological patterns with activation strengths at each spatial coordinate. Individual genes can then be linked to the patterns through a binary matrix $\mathbf{Z}$, where gene g is explained by pattern k if $z_{g,k} = 1$. With these variables and the expression matrix $\mathbf{Y}$, which consist of the expression vectors $(\mathbf{y}_1, \ldots, \mathbf{y}_G)$ for the G spatially variable genes, the expression histology model can be formulated as a Gaussian process mixture model:

$$P(Y \mid \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) = \prod_{k=1}^{K} \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{y}_g \,\middle|\, \boldsymbol{\mu}_k, \sigma_e^2 \cdot \mathbf{I}\right)^{z_{g,k}} \tag{7}$$

with informative priors

$$p(\{\boldsymbol{\mu}_k\}) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k \,\middle|\, \mathbf{0}, \boldsymbol{\Sigma}_{k\left(\mathbf{x},\mathbf{x}' \mid \boldsymbol{\theta}_k\right)}\right), \tag{8}$$

$$P(\mathbf{Z}) = \prod_{k=1}^{K} \prod_{g=1}^{G} (\pi_k)^{z_{g,k}}. \tag{9}$$

The parameters for the prior on cluster membership, π_k, are constrained such that $\sum_{k=1}^{K} \pi_k = 1$, and σ_e^2 is the noise level of the model.

A user need to make choices for two parameters: the number of patterns K and the prior covariance $\boldsymbol{\Sigma}_{k\left(\mathbf{x},\mathbf{x}' \mid \boldsymbol{\theta}_k\right)}$. In practice, the use of the squared exponential kernel with lengthscale set to an average inferred lengthscale of significantly spatially variable genes is recommended.

The key quantify of interest is the posterior distribution $P(\mathbf{Z}, \{\boldsymbol{\mu}_k\} \mid \mathbf{Y}, \sigma_e^2)$, which defines the assignment of genes to clusters and the activation strength of the inferred patterns.

2.1. Variational Bayesian inference.

To infer the posteriors of $\mathbf{Z}$ and $\{\boldsymbol{\mu}_k\}$ we make use of variational inference [Bishop, 2006], a deterministic inference scheme with the goal to fit a parametric approximation to the true posterior. This is achieved by making factorizing assumptions of the parameters of interest. In the AEH model we define this factorization as

$$Q(\mathbf{Z}, \{\boldsymbol{\mu}_k\}) = \prod_{g=1}^{G} Q(\mathbf{z}_g) \cdot \prod_{k=1}^{K} Q(\boldsymbol{\mu}_k).$$

The variational distribution $Q(\boldsymbol{\mu}_k)$ is a multivariate normal distribution with variational parameters $\mathbf{m}_k$ and $\mathbf{W}_k$,

$$Q(\boldsymbol{\mu}_k \mid \mathbf{m}_k, \mathbf{W}_k) = \mathcal{N}(\boldsymbol{\mu}_k \mid \mathbf{m}_k, \mathbf{W}_k).$$

The variational parameters are identified through variational updates, taking the expectation of a single parameter conditional on the current estimate of all other parameters (for brevity in these derivations, we set $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{k\left(\mathbf{x},\mathbf{x}' \mid \boldsymbol{\theta}_k\right)}$, and C_*'s denote independent constants factored out in the derivation).

$$
\begin{aligned}
\ln Q(\boldsymbol{\mu}_k) &\propto \left\langle \ln P(Y \mid \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) \cdot P(\{\boldsymbol{\mu}\}) \cdot P(\mathbf{Z}) \right\rangle_Q \\
&= \left\langle \ln P(Y \mid \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) + \ln P(\boldsymbol{\mu}_k) + C_0 \right\rangle_Q \\
&= \ln P(\boldsymbol{\mu}_k) + \left\langle P(Y \mid \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) \right\rangle_Q + C_1 \\
&= \ln P(\boldsymbol{\mu}_k) + \sum_{g=1}^{G} \left\langle \ln P(\mathbf{y}_g \mid \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) \right\rangle_Q + C_1 \\
&= \ln \mathcal{N}\left(\boldsymbol{\mu}_k \mid \mathbf{0}, \boldsymbol{\Sigma}\right) + \sum_{g=1}^{G} \langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \ln \mathcal{N}\left(\boldsymbol{\mu}_k \,\middle|\, \mathbf{y}_g, \sigma_e^2\right) + C_1 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \sum_{g=1}^{G} \Big(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \\
&\left(-\frac{1}{2} \cdot (\boldsymbol{\mu}_k - \mathbf{y}_g)^{\mathrm{T}} \left(\sigma_e^2 \cdot I\right)^{-1} (\boldsymbol{\mu}_k - \mathbf{y}_g) \right) \Big) + C_2 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \sum_{g=1}^{G} \Big(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \\
&(\boldsymbol{\mu}_k - \mathbf{y}_g)^{\mathrm{T}} \left(\frac{1}{\sigma_e^2} \cdot I \right) (\boldsymbol{\mu}_k - \mathbf{y}_g) \Big) + C_2 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \sum_{g=1}^{G} \Big(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \\
&\left(\boldsymbol{\mu}_k^{\mathrm{T}} \left(\frac{1}{\sigma_e^2} \cdot I \right) \boldsymbol{\mu}_k - 2 \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \left(\frac{1}{\sigma_e^2} \cdot I \right) \mathbf{y}_g + \mathbf{y}_g^{\mathrm{T}} \left(\frac{1}{\sigma_e^2} \cdot I \right) \mathbf{y}_g \right) \Big) + C_2 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \sum_{g=1}^{G} \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \left(\frac{1}{\sigma_e^2} \cdot I \right) \boldsymbol{\mu}_k \right) + \\
&\sum_{g=1}^{G} \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \left(\frac{1}{\sigma_e^2} \cdot I \right) \mathbf{y}_g \right) + C_3 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \left(\frac{\sum_{g=1}^{G} \langle z_{g,k} \rangle_{Q(z_{g,k})}}{\sigma_e^2} \cdot I \right) \boldsymbol{\mu}_k + \\
&\boldsymbol{\mu}_k^{\mathrm{T}} \left(\frac{1}{\sigma_e^2} \cdot I \right) \sum_{g=1}^{G} \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \mathbf{y}_g \right) + C_3
\end{aligned}
$$

For clarity, we define $G_k = \sum_{g=1}^{G} \langle z_{g,k} \rangle_{Q(z_{g,k})}$ and $\hat{\mathbf{y}}_k = \sum_{g=1}^{G} \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \mathbf{y}_g \right)$.

$$
= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \left(\frac{G_k}{\sigma_e^2} \cdot I \right) \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}} \left(\frac{1}{\sigma_e^2} \cdot I \right) \hat{\mathbf{y}}_k + C_3
$$

We can also summarise the matrix $\left(\frac{G_k}{\sigma_e^2} \cdot I\right)$ as D_k

$$
\begin{aligned}
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} D_k \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}} D_k \hat{\mathbf{y}}_k + C_3 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \left(\boldsymbol{\Sigma}_\theta^{-1} + D_k\right) \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}} \hat{\mathbf{y}}_k \cdot \frac{1}{\sigma_e^2} + C_3 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \left(\boldsymbol{\Sigma}_\theta^{-1} + D_k\right) \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}} \left(\boldsymbol{\Sigma}_\theta^{-1} + D_k\right) \left(\boldsymbol{\Sigma}_\theta^{-1} + D_k\right)^{-1} \hat{\mathbf{y}}_k \cdot \frac{1}{\sigma_e^2} + C_3
\end{aligned}
$$

This takes the same form as

$$
\begin{aligned}
\ln Q(\boldsymbol{\mu}_k \,|\, \mathbf{m}_k, \mathbf{W}_k) &= -\frac{1}{2} \cdot (\boldsymbol{\mu}_k - \mathbf{m}_k)^{\mathrm{T}} \mathbf{W}^{-1} (\boldsymbol{\mu}_k - \mathbf{m}_k) + C_0 \\
&= -\frac{1}{2} \cdot \left(\boldsymbol{\mu}_k^{\mathrm{T}} \mathbf{W}^{-1} \boldsymbol{\mu}_k - 2 \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \mathbf{W}^{-1} \mathbf{m}_k + \mathbf{m}_k^{\mathrm{T}} \mathbf{W}^{-1} \mathbf{m}_k\right) + C_0 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^{\mathrm{T}} \mathbf{W}^{-1} \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}} \mathbf{W}^{-1} \mathbf{m}_k + C_1.
\end{aligned}
$$

with final update equations:

$$
\begin{aligned}
\mathbf{W}_k^{-1} &= \boldsymbol{\Sigma}_\theta^{-1} + D_k, \\
\mathbf{m}_k &= \mathbf{W}_k \hat{\mathbf{y}}_k \cdot \frac{1}{\sigma_e^2}.
\end{aligned}
$$

The variational distribution $Q(\mathbf{z}_g)$ takes the form of a multinomial distribution with parameters r,

$$
Q(\mathbf{z}_g \,|\, \mathbf{r}_g) = \prod_{k=1}^{K} r_{g,k}^{z_{g,k}}.
$$

The variational updates for the parameters r follows as

$$\begin{aligned}
\ln Q(\mathbf{z}_g) &\propto \left\langle \ln P(Y \mid \boldsymbol{\mu}, Z, \sigma_e^2) \cdot P(\boldsymbol{\mu}) \cdot P(Z) \right\rangle_Q \\
&= \left\langle \ln P(Y \mid \boldsymbol{\mu}, Z, \sigma_e^2) + \ln P(Z) + C_0 \right\rangle_Q \\
&= \ln P(\mathbf{z}_g) + \sum_{k=1}^{K} \left\langle \ln P(\mathbf{y}_g \mid \boldsymbol{\mu}_k, Z, \sigma_e^2 \right\rangle_Q + C_1 \\
&= \sum_{k=1}^{K} \mathbf{z}_{g,k} \cdot (\ln \pi_k) + \sum_{k=1}^{K} \mathbf{z}_{g,k} \cdot \ln \left\langle \mathcal{N}\left(\boldsymbol{\mu}_k \,\middle|\, \mathbf{y}_g, \sigma_e^2 \cdot I\right) \right\rangle_Q + C_1 \\
&= \sum_{k=1}^{K} \mathbf{z}_{g,k} \cdot (\ln \pi_k - \\
&\frac{1}{2} \cdot \left\langle \ln|\sigma_e^2 \cdot I| + (\boldsymbol{\mu}_k - \mathbf{y}_g)^{\mathrm{T}} (\frac{1}{\sigma_e^2} \cdot I)(\boldsymbol{\mu}_k - \mathbf{y}_g) + N \cdot \ln(2 \cdot \pi) \right\rangle_Q \Bigg) + C_1 \\
&= \sum_{k=1}^{K} \mathbf{z}_{g,k} \cdot \Bigg(\ln \pi_k - \frac{1}{2} \cdot \big(\ln|\sigma_e^2 \cdot I| + \\
&\left\langle (\boldsymbol{\mu}_k - \mathbf{y}_g)^{\mathrm{T}} (\frac{1}{\sigma_e^2} \cdot I)(\boldsymbol{\mu}_k - \mathbf{y}_g) \right\rangle_Q + N \cdot \ln(2 \cdot \pi) \Big) \Bigg) + C_1 \\
&= \sum_{k=1}^{K} \mathbf{z}_{g,k} \cdot \ln \rho_{g,k} + C_1.
\end{aligned}$$

This implies

$$Q(\mathbf{z}_g) \propto \prod_{k=1}^{K} \rho_{g,k}^{\mathbf{z}_{g,k}}.$$

By scaling the ρ-terms we arrive at the final updates of $r_{g,k}$:

$$r_{g,k} = \frac{\rho_{g,k}}{\sum_{k=1}^{K} \rho_{g,k}}. \tag{10}$$

Remaining parameters are determined by maximising the evidence lower bound (ELBO). The ELBO for the AEH model is

$$\begin{aligned}
\mathcal{L} &= \left\langle \ln P(Y, \mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \right\rangle - \left\langle \ln Q(\mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \right\rangle \\
&= \left\langle \ln P(Y \mid \mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \right\rangle + \langle \ln P(\mathbf{Z}) \rangle + \langle \ln P(\boldsymbol{\mu}) \rangle - \langle \ln Q(\mathbf{Z}) \rangle - \langle \ln Q(\boldsymbol{\mu}) \rangle
\end{aligned}$$

where the individual terms evaluate to

$$\begin{aligned}
\left\langle \ln P(Y \mid \mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \right\rangle &= \sum_{k=1}^{K} \sum_{g=1}^{G} r_{g,k} \cdot \ln(\rho_{g,k}), \\
\left\langle \ln P(Y \mid \mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \right\rangle &= \sum_{k=1}^{K} \sum_{g=1}^{G} r_{g,k} \cdot \left(\ln \pi_k - \frac{1}{2} \cdot \right. \\
&\qquad \left. \left(\ln|\sigma_e^2 \cdot I| + (\mathbf{m}_k - \mathbf{y}_g)^{\mathrm{T}} (\frac{1}{\sigma_e^2} \cdot I)(\mathbf{m}_k - \mathbf{y}_g) + N \cdot \ln(2 \cdot \pi) \right) \right), \\
\langle \ln P(\mathbf{Z}) \rangle &= \sum_{k=1}^{K} \sum_{g=1}^{G} r_{g,k} \cdot \ln(\pi_k), \\
\langle \ln P(\boldsymbol{\mu}) \rangle &= -\frac{1}{2} \cdot \sum_{k=1}^{K} \left(\ln|\boldsymbol{\Sigma}_\theta| + \mathbf{m}_k^{\mathrm{T}} \boldsymbol{\Sigma}_\theta^{-1} \mathbf{m}_k + N \cdot \ln(2 \cdot \pi) \right), \\
\langle \ln Q(\mathbf{Z}) \rangle &= \sum_{k=1}^{K} \sum_{g=1}^{G} r_{g,k} \cdot \ln(r_{g,k}), \\
\langle \ln Q(\boldsymbol{\mu}) \rangle &= -\frac{1}{2} \cdot \sum_{k=1}^{K} \left(\ln|\mathbf{W}_k| + N \cdot \ln(2 \cdot \pi) \right).
\end{aligned}$$

Interleaved with the variational updates, the ELBO is maximised with respect to σ_e^2, and the π parameters are updated by

$$\pi_k = \frac{G_k}{G},$$

which further increases the ELBO and allows the model to discard superfluous histological patterns during inference.

3. Data normalisation

The presented Gaussian process models is based on the assumption of normally distributed residual noise and independent observations across cells. To meet these requirements, we have identified two necessary normalisation steps.

First, both spatial transcriptomics and in-situ hybridisation data produce counts of transcripts. Spatial Transcriptomics uses unique molecular identifiers (UMIs) to count amplified transcript tags from next generation sequencing reads, while smFISH counts fluorescent probes inside cell boundaries. By investigating mean-variance relationships in multiple data sets from spatial technologies, we empirically observed that negative binomial (NB) noise is able to capture these dependencies. To stabilise the variance, we use the approximate Anscombe's transform for NB data on the observed counts $\hat{\mathbf{y}}$ and compute $\mathbf{y} = \log(\hat{\mathbf{y}} + \frac{1}{\phi})$, where ϕ is the overdispersion parameter [Anscombe, 1948]. This parameter ϕ is estimated by fitting the quadratic $\mathrm{Var}(\mathbf{y}) = \mathbb{E}(\mathbf{y}) + \phi \cdot \mathbb{E}(\mathbf{y})^2$ across all genes in a study.

Second, we note that in the datasets we investigated, every gene's expression correlates with the total transcript count in the cells. In particular, for MERFISH data the area of cells is provided and we note that the total count correlates strongly with the cytoplasmic area. This relation has previously been described

by [Padovan-Merhar et al., 2015], who showed that cells compensate mRNA content in response to the cytoplasmic volume. The total count is thus a proxy of cell size.

While there are many instances where variation in cell size is self of biological interest, cell size assays are easier than gene expression assays, and here we aim to study the regulation of gene expression independent of cell size. In particular, if the distribution of relative cell sizes show spatial dependencies, *every gene* will be considered spatially variable.

Consequently, we consider expression levels that are adjusted for variation in cell size, using linear regression to account for this dependence by regressing out the log total count from the Anscombe transformed expression values before fitting the spatial models.

For comparison, we also considered the spatial variation test on the total count in each data set. In all data sets the variation is significant, with between 30% and 80% FSV (results marked as X's in **Figure 2** and supplementary figures). In the frog development RNA-seq data, proxies for cell size (ERCC expression and number of genes detected) are over 95% spatially variable.

4. Unsupervised clustering analysis and ANOVA

To compare our results to analysis not considering spatial information we applied principal component analysis to reduce the dimensionality followed by clustering using a Bayesian Gaussian mixture model applied to the first two principal components. The Bayesian Gaussian mixture model was initialized with 20 clusters. The implementation in scikit-learn was used (`BayesianGaussianMixture`), with the `max_iter` parameter set to 10000. After inference, in both datasets 4 clusters were retained (**Supp. Fig. 5A-C** and **Supp. Fig. 7A-C**). Genes varying between clusters were identified by an ANOVA test expression levels processed analogously as for SpatialDE.

5. Assessing statistical calibration through data simulation

To investigate the limits and power of SpatialDE we used simulated data with known ground truth. These data allowed us to assess the power for detecting spatial variation as a function of the simulated FSV level, and for alternative simulated settings. To ensure the settings we considered were in line with real data, we used the spatial X-coordinates from the mouse olfactory bulb data set.

5.1. **Simulation of bell curve shaped data.** A two dimensional bell curve is calculated as $k_{\text{spatial}}(\mathbf{x}, \mathbf{x}' \mid l)$ in Section 1.2, where l is the radius of the bell, and $\mathbf{x}'$ is fixed as the center of the bell (which we chose as $(14.0, 15.0)^{\mathrm{T}}$ here) (**Supp. Fig. 10A**). Bell shaped expression profiles were simulated for 15 radii evenly distributed on a log scale from $\frac{1}{4}$ of the smallest pairwise observed distance (0.2) and 4 times the largest pairwise observed distance (85.9). For non-spatial variation, 200 σ_n^2 values were sampled uniformly on a logarithmic scale from the interval $[10^{-3}, 10]$, and noise sampled from $\mathcal{N}(0, \sigma_n^2)$ was added to the bell curves. Ground truth FSV values were calculated using the variance of the bell curve and the variance parameter.

After applying SpatialDE to the 3,000 simulated bell curve genes, we considered P-values from the tests stratified over FSV ranges and bell radii. Statistical power was assessed as the fraction of genes with $P < 0.05$ (**Supp. Fig 10B**).

5.2. **Simulation from the SpatialDE model.** True spatial covariances $\mathbf{\Sigma}$ were generated from $k_{\text{spatial}}(\mathbf{x}, \mathbf{x}' \,|\, l)$ for 15 values of l as described above. True FSV values were sampled 200 times uniformly on a log scale from the interval $[10^{-2}, 1]$, and for each FSV, $\delta = \frac{1}{\text{FSV}} - 1$ was calculated. Using these parameters, gene expression levels on the spatial locations were simulated by randomly drawing from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma} + \delta \cdot \mathbf{I})$ for the $15 \cdot 200 = 3,000$ parameter combinations.

After applying SpatialDE to the simulated data, P-values were stratified by FSV and the true simulated characteristic length scale. Statistical power was assessed as the fraction of genes with $P < 0.05$ (**Supp. Fig 10C**).

5.3. **Simulation from null model.** Data with no true spatial covariance were simulated by 3,000 draws from the normal distribution $\mathcal{N}(0, 1)$.

After applying SpatialDE to simulated data, the fraction of genes passing increasingly stringent significance thresholds was calculated, indicating the significance test is conservative. (**Supp. Fig. 10E**)

6. Computational Performance Benchmark

Data for 10,000 genes were simulated according to the SpatialDE model with various effect magnitudes for multiple sample sizes. For SpatialDE, the test was run on these data and timed according to a wall clock. For the Stan implementation, 100 random genes were sampled for each sample size, and timing was extrapolated by multiplying the time by 100 (**Supp. Fig. 2**). It should be noted that this problem is trivially parallelizable across genes, and the considered implementations are comparable and that they do not make use this. Benchmarks were performed on a Late 2013 iMac with a 3.2 GHz Intel Core i5 processor and 32 GB of DDR3 RAM, a typical consumer level PC.

7. Software availability

The primary implementation of SpatialDE is a Python 3 package, which can be installed from PyPI using pip. Development is public on Github[1]. A Stan implementation is also provided in the same repository, as well as code for all analysis presented in this paper, and additional tutorials and notebooks illustrating how to use the package. All data used in our analysis is also available in preprocessed form from the Github repository using git-LFS.

References

[Anscombe, 1948] Anscombe, F. J. (1948). THE TRANSFORMATION OF POISSON, BINOMIAL AND NEGATIVE-BINOMIAL DATA. *Biometrika*, 35(3-4):246–254.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

[Kostem and Eskin, 2013] Kostem, E. and Eskin, E. (2013). Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *American journal of human genetics*, 92(4):558–564.

[Lippert et al., 2011] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835.

[Lloyd et al., 2014] Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2014). Automatic construction and Natural-Language description of nonparametric regression models. *Proc. 28th AAAI Conference on Artificial Intelligence*.

[1] `https://github.com/Teichlab/SpatialDE`

[Padovan-Merhar et al., 2015] Padovan-Merhar, O., Nair, G. P., Biaesch, A. G., Mayer, A., Scarfone, S., Foley, S. W., Wu, A. R., Churchman, L. S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Molecular cell*, 58(2):339–352.

[Storey and Tibshirani, 2003] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–9445.

[Williams and Rasmussen, 2006] Williams, C. K. I. and Rasmussen, C. E. (2006). Gaussian processes for machine learning. *the MIT Press*, 2(3):4.