



# SpatialDE: identification of spatially variable genes

Valentine Svensson<sup>1,2</sup> , Sarah A Teichmann<sup>1,3</sup>  
& Oliver Stegle<sup>2,4</sup> 

**Technological advances have made it possible to measure spatially resolved gene expression at high throughput. However, methods to analyze these data are not established. Here we describe SpatialDE, a statistical test to identify genes with spatial patterns of expression variation from multiplexed imaging or spatial RNA-sequencing data. SpatialDE also implements ‘automatic expression histology’, a spatial gene-clustering approach that enables expression-based tissue histology.**

Miniaturization and parallelization in genomics have enabled high-throughput transcriptome profiling from low quantities of starting material, including single cells. The increased throughput has also fostered new experimental techniques that can directly assay the spatial context of variations in gene expression. Spatial resolution of gene expression is crucial for determining the functions and phenotypes of cells in multicellular organisms<sup>1</sup>. Spatial expression variation can reflect communication between adjacent cells, position-specific states, or cells that migrate to specific tissue locations to perform their functions.

Several experimental methods for measuring gene expression levels in a spatial context have been established, and all differ in resolution, accuracy, and throughput. These include the computational integration of single-cell RNA-seq (scRNA-seq) data with a spatial reference data set<sup>2,3</sup>, careful collection and recording of the spatial location of samples<sup>4</sup>, parallel profiling of mRNA using barcodes on a grid of known spatial locations<sup>4–6</sup>, and methods based on multiplexed *in situ* hybridization<sup>7,8</sup> or sequencing<sup>1</sup>.

A first critical step in the analysis of these data sets is to identify genes with spatial expression variation across the tissue. However, existing approaches for identifying highly variable genes (HVGs)<sup>9</sup>, as used for conventional scRNA-seq data, ignore spatial information and thus do not measure spatial variability (Fig. 1a). As an alternative, researchers have applied analysis of variance (ANOVA) to test for differential expression between groups of cells, either using a priori-defined cell annotations or based on sample clustering<sup>2,3,6,7</sup>, with some methods incorporating spatial

information<sup>10</sup>. Critically, such methods can detect only variations that are captured by differences between discrete groups.

Here we propose SpatialDE, a method for identifying and characterizing spatially variable (SV) genes. Our method builds on Gaussian process regression, a class of models used in geostatistics. Briefly, for each gene, SpatialDE decomposes expression variability into spatial and nonspatial components (Fig. 1a,b), using two random effect terms: a spatial variance term that parametrizes gene expression covariance by pairwise distances of samples, and a noise term that models nonspatial variability. The ratio of the variance explained by these components quantifies the fraction of spatial variance. One can identify significant SV genes by comparing this full model to a model without the spatial variance component (Fig. 1b and Online Methods).

By interpreting the fitted model parameters, one can gain insights into the underlying spatial function, such as the length scale (i.e., the expected number of changes in a unit interval; Fig. 1b). SpatialDE can also be used to classify these functions, thereby identifying genes with linear or periodic expression patterns (Supplementary Fig. 1 and Online Methods). Finally, SpatialDE provides a spatial clustering method within the same Gaussian process framework, which identifies sets of genes that mark distinct spatial expression patterns (Fig. 1c). This provides a means to apply automatic expression histology (AEH), which relates tissue structure and cell-type composition on the basis of the expression patterns of marker genes. Because it leverages efficient inference methods previously developed for linear mixed models<sup>11</sup> and takes advantage of the data structure from massively parallel molecular assays, SpatialDE is computationally efficient (Online Methods and Supplementary Fig. 2).

First, we applied our method to spatial transcriptomics data from mouse olfactory bulb<sup>6</sup>. Briefly, we derived spatial transcriptomics gene expression levels from thin tissue sections placed on an array with poly(dT) probes and spatially resolved DNA barcodes. These formed a grid of circular ‘spots’ with a diameter of 100  $\mu\text{m}$ , where at each spot the mRNA abundance of 10–100 cells is measured using probes with barcodes that encode spatial locations.

The SpatialDE test identified 67 SV genes (false discovery rate (FDR) < 0.05; Supplementary Table 1), with spatial dependencies explaining up to 70% of the gene expression variance (Fig. 2a). This set of genes was also markedly disjointed from genes identified via conventional HVG methods that ignore spatial dependencies (3,497 genes, 40 overlapping; Online Methods). The SV genes identified showed clear spatial substructure, consistent with matched hematoxylin and eosin (H&E)-stained samples of the same tissue (Fig. 2b,c). These included canonical marker genes

<sup>1</sup>Wellcome Trust Sanger Institute, Hinxton, UK. <sup>2</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. <sup>3</sup>Theory of Condensed Matter Group, The Cavendish Laboratory, University of Cambridge, Cambridge, UK. <sup>4</sup>European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. Correspondence should be addressed to V.S. (v.s@ebi.ac.uk) or O.S. (oliver.stegle@ebi.ac.uk).

RECEIVED 14 JUNE 2017; ACCEPTED 22 FEBRUARY 2018; PUBLISHED ONLINE 19 MARCH 2018; DOI:10.1038/NMETH.4636

highlighted in the primary analysis by Ståhl *et al.*<sup>6</sup>, such as *Penk*, *Doc2g*, and *Kctd12*, and additional genes that define the granule cell layer of the bulb. Genes in the latter set were classified as periodically variable, with period lengths corresponding to the distance between the centers of the hemispheres; these genes included *Kcnh3*, *Nrgn*, and *Mbp*, with 1.8-mm period lengths (Fig. 2c; further examples are shown in Supplementary Fig. 3). Other genes with periodic patterns, such as *Slc17a7*, which encodes a vesicular glutamate transporter, were identified with shorter periods (1.1 mm), and inspection revealed regularly dispersed regions potentially identifying a pattern associated with higher neuron density<sup>12</sup>, thus suggesting that periodic expression in tissues is of biological interest.

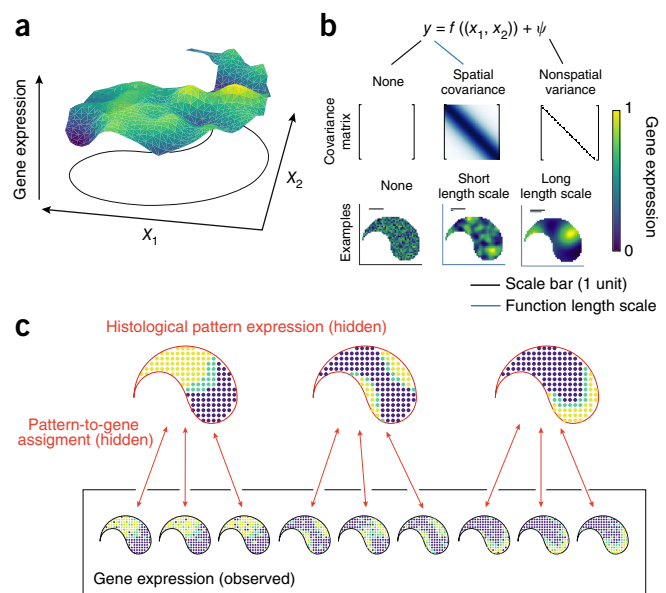
Application of AEH identified five canonical expression patterns, clearly demarcating structures visible in the H&E-stained tissue (Fig. 2d, Supplementary Fig. 4a). For comparison, we also considered conventional clustering based on the expression profile of each spot. However, this approach ignores spatial information and does not establish relationships between genes that define cell types as in AEH (Supplementary Fig. 5).

As a second application, we considered tissue slices from human breast cancer biopsies<sup>6</sup>, profiled with the same spatial transcriptomics protocol (Supplementary Fig. 6). SpatialDE identified 115 SV genes (FDR < 0.05, compared with 3,503 detected by the HVG approach; overlap of 34 genes), including 7 genes with known disease relevance that were highlighted in the primary analysis (Supplementary Fig. 6b,c). Significantly SV genes were enriched for collagens, which distinguish tissue substructure<sup>13</sup> (reactome “Collagen formation,”  $P = 3.38 \times 10^{-14}$ , gProfiler<sup>14</sup>; Supplementary Table 1). Additionally, we identified the autophagy-related gene *TP53INP2* in the region surrounding the structured tissue (Supplementary Figure 6c). The set of SV genes also included the cytokine-expressing genes *CXCL9* and *CXCL13*, which were expressed in a visually distinct region (Supplementary Fig. 6a), together with the IL-12 receptor subunit gene *IL12RB1*, indicating a potential tumor-related local immune response. Notably, these genes (and  $N = 29$  others) were not identified as differentially expressed when we applied unsupervised clustering in conjunction with ANOVA (Supplementary Fig. 7). Furthermore, these genes did not have high rankings on the basis of nonspatial HVG measures (including comparisons of mean versus squared coefficient of variation<sup>9</sup> and mean versus dropout<sup>15</sup>; Supplementary Fig. 8).

AEH of the SV genes in the breast cancer biopsy (Supplementary Fig. 4b) most clearly separated the adipocytic from the denser region of the tissue, but also identified a small region overlapping the tumor feature in the H&E image. Among the 17 genes assigned to this pattern were *CXCL9* and *CXCL13*, expressing cytokines, and *IL12RB1* and *IL21R*, expressing interleukin receptors (Supplementary Table 1).

Overall, we found that variable gene detection by SpatialDE is complementary to existing methods. In particular, SpatialDE identifies genes with localized expression patterns, as indicated by small fitted length scales, which are missed by methods that ignore spatial contexts (Supplementary Fig. 7e). We confirmed the statistical calibration and robustness of SpatialDE by applying the method to randomized data (Supplementary Fig. 9) and simulations (Supplementary Fig. 10).

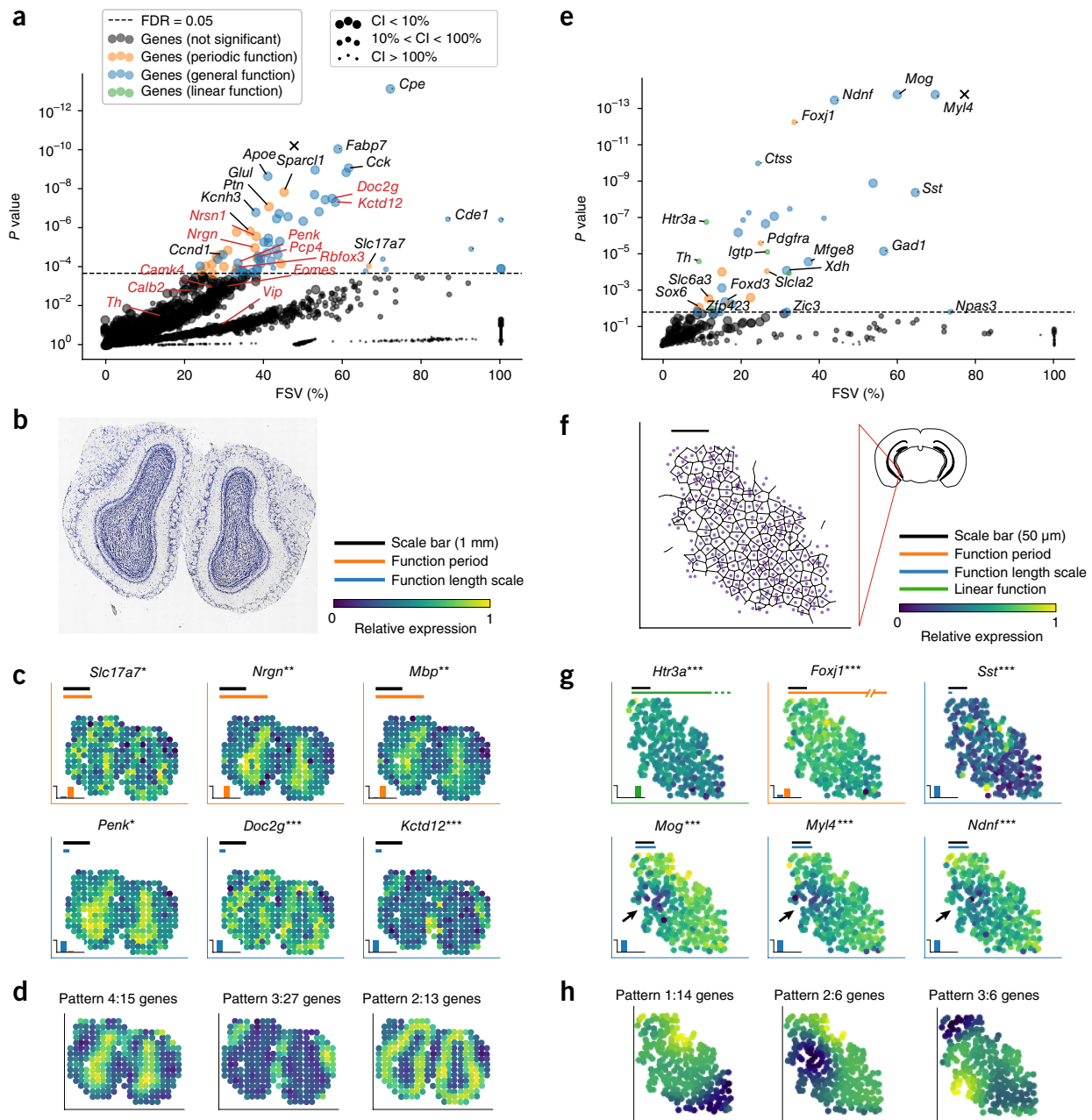
SpatialDE is not limited to sequencing technologies and can be applied to any expression data with spatial and/or temporal



**Figure 1** | Overview of SpatialDE for the identification of spatially variable genes. **(a)** In spatial gene expression studies, expression levels are measured as a function of spatial coordinates of cells or samples. SpatialDE defines spatial dependence for a given gene by using a nonparametric regression model, testing whether gene expression levels at different locations covary in a manner that depends on their relative location, and thus are spatially variable. **(b)** SpatialDE partitions expression variation into a spatial component (using functional dependencies  $f(x_1, x_2)$ ), characterized by spatial covariance, and independent observation noise ( $\psi$ ). Representative simulated expression patterns are plotted below the corresponding covariance matrices for the null model (None) and the alternative model (Spatial covariance) with different length scales. **(c)** Automatic expression histology uses spatial clustering to model the expression levels of spatially variable genes, using a set of unobserved tissue-structure patterns. Both the underlying patterns and the gene-pattern assignments are learned from data.

annotation. To explore this, we applied SpatialDE to data generated by multiplexed single-molecule fluorescence *in situ* hybridization (smFISH), a method that quantifies gene expression with subcellular resolution for a large number of target genes in parallel. Briefly, probes are sequentially hybridized to RNA while carrying different temporal combinations of fluorophores, which act as barcodes and can be used to quantify the expression of thousands of transcripts<sup>16</sup> through imaging.

We applied SpatialDE to multiplexed smFISH data of cells from mouse hippocampus, generated via sequential barcoded fluorescence *in situ* hybridization (seqFISH)<sup>7</sup>. This study considered 249 genes chosen to investigate the cell-type composition along dorsal and ventral axes of the hippocampus (Fig. 2e). SpatialDE identified 32 SV genes (Fig. 2e; FDR < 0.05; 58 genes were detected as HVGs, with an overlap of 5 genes), and again identified genes with different types of spatial variation, including linear ( $N = 5$ ) and periodic patterns ( $N = 8$ ; examples are shown in Supplementary Fig. 11). The three highest-ranking genes—*Mog*, *Myl4*, and *Ndnf*—were associated with a distinct region of lower expression (Fig. 2f,g, black arrows). These genes were grouped into histological expression patterns by the AEH method (Fig. 2h, Supplementary Fig. 2c). Visual inspection of all 249 genes supported the ranking of spatial variation from SpatialDE (Supplementary Fig. 12).



**Figure 2** | Application of SpatialDE to spatial transcriptomics and seqFISH data. **(a)** Fraction of variance explained by spatial variation (FSV) versus significance of spatial variation (SpatialDE  $-\log P$  value) for all genes in the mouse olfactory bulb data. The dashed horizontal line indicates the significance cutoff (FDR = 0.05;  $N = 67$  SV genes;  $Q$ -value adjusted). Genes are classified as periodically variable ( $N = 19$ ) or with a general spatial dependency ( $N = 48$ ). Classical histological marker genes highlighted by Ståhl *et al.*<sup>6</sup> are in red text. Point size denotes the uncertainty of FSV estimates as indicated in the key (CI, confidence interval). The “X” symbol indicates the result of applying SpatialDE to the estimated total RNA content per spot. **(b)** Image of H&E-stained mouse olfactory bulb from ref. 6. **(c)** Visualization of selected SV genes. Orange bars indicate the fitted period length for genes with periodic dependencies; blue bars indicate the fitted length scale for genes with general spatial trends. 2D plots depict expression levels for genes across the tissue section, color-coded according to the key. Insets in the lower left show the posterior probability for gene assignments as general spatial, periodic spatial, or linear trends. \*FDR < 0.05, \*\*FDR < 0.01, \*\*\*FDR < 0.001. **(d)** Example histological expression patterns identified by AEH analysis, with expression levels color-coded according to the key. **(e)** FSV versus significance of spatial variation (SpatialDE  $-\log P$  value) for all 249 genes in the seqFISH data from a region of mouse hippocampus from ref. 7, as in **a**, with genes with linear dependency denoted by green points. **(f)** Voronoi tessellation representative of tissue structure in mouse hippocampus. **(g)** Expression of selected SV genes (out of 32; FDR < 0.05,  $Q$ -value adjusted) with linear (*Htr3a*), periodic (*Foxj1*), and general spatial trends. Black arrows indicate distinct regions of low expression of *Mog*, *Myl4*, and *Ndnf*. Color-coded according to the key. \*\*\*FDR < 0.001. **(h)** Three examples of histological expression patterns identified by AEH. Panels **b** and **f** reproduced with permission from ref. 6.



SpatialDE can also be used to test for spatial expression variation in cell culture systems, where spatial variation is not typically expected. As an example, we considered data from another multiplexed smFISH data set generated with MERFISH with 140 probes in a human osteosarcoma cell line<sup>8</sup> (**Supplementary Fig. 13a,b**). In the primary analysis, surprisingly, Moffitt *et al.*<sup>8</sup> discovered spatially restricted cell populations with higher proliferation rates. Consistent with these findings, our method identified a substantial proportion of the genes assayed as SV ( $N = 91$  (65% of all genes);  $\text{FDR} < 0.05$ ; 29 HVGs, with overlap of 24 genes), including six of the seven genes highlighted as differentially expressed between proliferating and resting subpopulations (for example, *THBS1* and *CENPF*; **Supplementary Fig. 13c**). This indicates that high confluence in cell culture can lead to spatial dependency in gene expression<sup>17</sup>. Negative control probes in these data were not detected as SV, which further confirms the statistical calibration of SpatialDE (**Supplementary Fig. 13d**).

Our results demonstrate that SpatialDE identifies SV genes and allows biologically relevant features to be detected in tissue samples without a priori histological annotation. The increased availability of high-throughput techniques, including spatially resolved RNA-seq, means that there will be a growing need for methods that account for this new dimension of expression variation, such as SpatialDE.

We applied our method to data generated via several different protocols, considering both tissues and cell cultures. SpatialDE can also be applied to temporal data from time-course experiments to identify genes with dynamic expression (**Supplementary Fig. 14**). Some methods already exist for this application<sup>18</sup>, but they are typically computationally more demanding. In principle, SpatialDE can also be applied to three-dimensional data, for example, from aligned serial sections of two-dimensional data, or from *in situ* sequencing<sup>1</sup>.

SpatialDE is related to and generalizes previous approaches for the detection of temporal<sup>19</sup> and periodic gene expression patterns<sup>20</sup> in time series. Although biologically important, the identification of periodic patterns has technical limitations, in particular in edge cases, where noise can mask statistical significance for visually similar patterns (**Supplementary Fig. 15**).

Future extensions of SpatialDE could be tailored toward specific platforms, for example, to more explicitly model technical sources of variation. Other areas of future work might include the incorporation of information about tissue makeup or local differences in cell density. Finally, there exist spatial clustering methods that are focused on clustering cell positions rather than genes<sup>10</sup>, which could be combined with the AEH presented here.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

The authors thank D. Arnol and F.P. Casale for helpful advice on statistics and data normalization. J. Moffitt helped us understand the data format for available MERFISH data. In addition, we thank A. Lun, M. Hemberg, D. Kunz, and K. Meyer for feedback on the manuscript. This work was supported by the EMBL (EMBL International PhD Program support to V.S.; core funding to O.S.), the Wellcome Trust (S.A.T. and O.S.), the ERC (Consolidator Grant “ThDEFINE” to S.A.T.), and the EU (O.S.).

## AUTHOR CONTRIBUTIONS

V.S. and O.S. conceived the method. V.S. implemented the method and generated the results. V.S., S.A.T., and O.S. interpreted the results and wrote the paper.

## COMPETING INTERESTS

The authors declare no competing interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Lee, J.H. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **9**, e1369 (2017).
2. Achim, K. *et al. Nat. Biotechnol.* **33**, 503–509 (2015).
3. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. *Nat. Biotechnol.* **33**, 495–502 (2015).
4. Junker, J.P. *et al. Cell* **159**, 662–675 (2014).
5. Chen, J. *et al. Nat. Protoc.* **12**, 566–580 (2017).
6. Ståhl, P.L. *et al. Science* **353**, 78–82 (2016).
7. Shah, S., Lubeck, E., Zhou, W. & Cai, L. *Neuron* **92**, 342–357 (2016).
8. Moffitt, J.R. *et al. Proc. Natl. Acad. Sci. USA* **113**, 11046–11051 (2016).
9. Brennecke, P. *et al. Nat. Methods* **10**, 1093–1095 (2013).
10. Pettit, J.-B. *et al. PLOS Comput. Biol.* **10**, e1003824 (2014).
11. Lippert, C. *et al. Nat. Methods* **8**, 833–835 (2011).
12. Takamori, S., Rhee, J.S., Rosenmund, C. & Jahn, R. *Nature* **407**, 189–194 (2000).
13. Seewaldt, V.L. *Nature* **490**, 490–491 (2012).
14. Reimand, J. *et al. Nucleic Acids Res.* **44**, W83–W89 (2016).
15. Andrews, T.S. & Hemberg, M. *bioRxiv* Preprint at <https://www.biorxiv.org/content/early/2016/10/20/065094> (2016).
16. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. *Science* **348**, aaa6090 (2015).
17. Battich, N., Stoeger, T. & Pelkmans, L. *Cell* **163**, 1596–1610 (2015).
18. Owens, N.D.L. *et al. Cell Rep.* **14**, 632–647 (2016).
19. Kalaitzis, A.A. & Lawrence, N.D. *BMC Bioinformatics* **12**, 180 (2011).
20. Durrande, N., Hensman, J., Rattray, M. & Lawrence, N.D. *PeerJ Comput. Sci.* **2**, e50 (2016).

## ONLINE METHODS

Full details of the derivation and implementation of SpatialDE are provided in **Supplementary Note 1**.

**SpatialDE model.** SpatialDE models gene expression profiles  $y = (y_1, \dots, y_N)$  for a given gene across spatial coordinates  $X = (x_1, \dots, x_N)$ , using a multivariate normal model of the form

$$P(y | \mu, \sigma_s^2, \delta, \Sigma) = N(y | \mu \cdot 1, \sigma_s^2 \cdot (\Sigma + \delta \cdot I)) \quad (1)$$

The fixed effect  $\mu_g \cdot 1$  accounts for the mean expression level, and  $\Sigma$  denotes a spatial covariance matrix defined on the basis of the input coordinates of pairs of cells. SpatialDE uses the so-called squared exponential covariance function to define  $\Sigma$ :

$$\Sigma_{i,j} = k(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2 \cdot l^2}\right) \quad (2)$$

whereby the covariance between pairs of cells  $i$  and  $j$  is modeled to decay exponentially with the squared distance between them. The hyperparameter  $l$ , also known as the characteristic length scale, determines how rapidly the covariance decays as a function of distance<sup>21</sup>.

The second covariance term  $\delta \cdot I$  accounts for independent non-spatial variation in gene expression, where the ratio  $1/(1 + \delta)$  can be interpreted as the fraction of expression variance attributable to spatial effects. Model parameters are fit by maximizing the marginal log likelihood (LL),

$$\begin{aligned} LL = & -\frac{1}{2} \cdot N \cdot \log(2\pi) - \frac{1}{2} \cdot \log(|\sigma_s^2 \cdot (\Sigma + \delta \cdot I)|) \\ & - \frac{1}{2} \cdot (y - \mu \cdot 1)^T (\sigma_s^2 \cdot (\Sigma + \delta \cdot I))^{-1} (y - \mu \cdot 1) \end{aligned} \quad (3)$$

This optimization problem has closed-form solutions for the parameters  $\mu$  and  $\sigma_s$  for given parameter values  $\delta$ . Gradient-based optimization is used to determine  $\delta$ , and the hyperparameter  $l$  is determined via grid search. Naive methods for evaluating the marginal likelihood in equation (1) scale cubically in the number of cells, thus prohibiting applications to larger data sets. We adapted algebraic reformulations that have been proposed in statistical genetics<sup>11,22</sup>, coupled with efficient precomputations of all terms possible, to improve the scalability of the model (**Supplementary Fig. 2**).

**Statistical significance.** To estimate statistical significance, we compared the model likelihood of the fitted SpatialDE model with the likelihood of a model that corresponds to the null hypothesis of no spatial covariance,

$$P(y | \mu, \sigma^2) = N(\mu \cdot 1, \sigma^2 \cdot I) \quad (4)$$

We then estimated  $P$  values analytically on the basis of the  $\chi^2$  distribution transformation with one degree of freedom. Unless stated otherwise, we used the  $Q$ -value method<sup>23</sup> to adjust for multiple testing, thereby controlling the FDR.

**Model selection.** After significance testing, the spatial covariance patterns identified can be further investigated through comparisons of models with alternative covariance functions.

In addition to the squared exponential covariance (equation (2)), SpatialDE implements covariance functions that assume linear trends and periodic patterns of gene expression variation (**Supplementary Fig. 1**), which are compared using the Bayesian information criterion (BIC):

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

Here  $M$  is the number of hyperparameters of a given model,  $N$  is the number of samples, and LL (equation (3)) is the log marginal likelihood of the data. Guidance on how to interpret these inferences and alternative functional forms is available in **Supplementary Note 1**.

**Automatic expression histology.** To group SV genes with similar spatial expression patterns, SpatialDE implements a clustering model based on the same spatial Gaussian-process-based (GP) prior as used to test for SV genes (equation (1)). Let  $Y = (y_1, \dots, y_G)$  be the expression matrix of  $G$  spatially variable genes in each spatial location (now each  $y_g$  is a vector of  $N$  observations), and  $\mu = \{\mu_1, \dots, \mu_K\}$  be the matrix of  $K$  underlying patterns, so the vector  $\mu_k$  represents pattern  $k$ . Further, let  $Z$  be a binary indicator matrix that assigns gene  $g$  to pattern  $k$  if  $z_{g,k} = 1$ . Then the full model across all genes can be written as

$$P(Y, \mu, Z, \sigma_e^2, \Sigma) = P(Y | \mu, Z, \sigma_e^2) \cdot P(\mu | \Sigma) \cdot P(Z)$$

$$P(Y | \mu, Z, \sigma_e^2) = \prod_{k=1}^K \prod_{g=1}^G N(y_g | \mu_k, \sigma_e^2)^{(z_{g,k})}$$

$$P(\mu | \Sigma) = \prod_{k=1}^K N(\mu_k | 0, \Sigma)$$

$$P(Z) = \prod_{k=1}^K \prod_{g=1}^G \left(\frac{1}{K}\right)^{(z_{g,k})}$$

The parameter  $\sigma_e^2$  is the noise level for the model, and  $\Sigma$  is the spatial covariance matrix defined on the basis of spatial coordinates (see equation (2)). This model can be regarded as an extension of the classical Gaussian mixture model<sup>24</sup>, with the addition of a spatial prior on cluster centroids. Approximate posterior distributions for  $\mu$  and  $Z$  are estimated by variational inference<sup>24</sup>, and the noise level  $\sigma_e^2$  is estimated by maximization of the variational lower bound. The length scale  $l$  for the covariance  $\Sigma$  is specified by the user, as is the number of fitted patterns,  $K$ . The choice of  $l$  can be informed by the fitted length scales in the SpatialDE significance test. **Supplementary Note 1** includes details on inference and derivation of variational updates.

After inference, the posterior expectations  $\bar{\mu}$  and  $\bar{Z}$  of the parameters can be used to visualize any histological pattern through plotting of  $\bar{\mu}_k$  over the  $x$  coordinates. The most likely assignment of genes to an individual pattern is determined by the largest value  $\bar{Z}_{g,k}$  for each pattern  $k$ , which corresponds to the posterior probabilities of a gene belonging to each pattern.

**Highly variable gene selection.** For each data set, HVGs were identified using the ScanPy implementation<sup>25</sup> of the Seurat method of HVG filtering<sup>3</sup> with default parameters.

**Relationship to prior work.** SpatialDE is related to existing GP gene expression models. First applied in geostatistics<sup>26</sup>, GP models have been used to test for differential gene expression over time<sup>27</sup>, including in the analysis of bifurcation events<sup>28</sup>, and to define general tests for temporal variability<sup>19,28–31</sup>.

Here we adapted GP models to spatial transcriptome data, although the model can also be applied to univariate data (**Supplementary Fig. 14**) or higher-dimensional inputs. The main technical innovations presented here are threefold. First, the model presented is faster than existing methods because it leverages computational tricks previously proposed in the context of statistical genetics (**Supplementary Fig. 2** and preceding section). Second, we combined spatial Gaussian processes with model selection using BIC<sup>32</sup>. Third, we propose efficient and versatile spatial clustering within the same statistical framework.

**Data sets and processing.** *Spatial transcriptomics data.* We downloaded the count tables from ref. 6 from the Spatial Transcriptomics Research website (<http://www.spatialtranscriptomicsresearch.org/datasets/doi-10-1126science-aaf2403>), which links to the original publication. For the breast cancer data, we used the file “Breast Cancer Layer 2” with the corresponding H&E image. For the mouse olfactory bulb, we used the file “MOB Replicate 11” with the corresponding H&E image. Images included in figures were cropped, down-scaled, and converted to grayscale to minimize file sizes. For AEH, the number of patterns was set to five for both data sets, and the characteristic length scale was set to 105  $\mu\text{m}$  for the breast cancer data and to 150  $\mu\text{m}$  for the olfactory bulb data.

*SeqFISH data.* We downloaded the expression table from the supplementary material associated with ref. 7 and extracted cell counts from the region annotated as number 43 in the 249-gene experiment (Table S8 in ref. 7). The shape of the data suggested that this corresponded to a region in the lower left part of the corresponding supplementary figure, which informed the schematic shown in **Figure 2f** (used only for the purpose of illustration). In the AEH analysis, the number of patterns was set to five, and the characteristic length scale was set to 50  $\mu\text{m}$ .

*MERFISH data.* From the Zhuang lab website (<http://zhuang.harvard.edu/merfish>), we downloaded a .zip file containing MERFISH data from ref. 8 ([http://zhuang.harvard.edu/MERFISHData/data\\_for\\_release.zip](http://zhuang.harvard.edu/MERFISHData/data_for_release.zip)). We used the files in the folder called “Replicate 6” from this .zip file, as these had the largest number of cells and highest confluency.

*Frog development RNA-seq data.* We downloaded the TPM expression table for clutch A from GEO accession [GSE65785](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65785), which is referenced in ref. 18.

**Expression count normalization.** The SpatialDE model is based on the assumption of normally distributed residual noise and independent observations across cells. To meet these requirements with spatial expression count data, we identified two normalization steps (**Supplementary Note 1**). First, we use a variance-stabilizing transformation for negative-binomial-distributed data to satisfy the first condition, known as Anscombe’s transformation. Second, we noticed that generally the expression level of a given gene correlates with the total count in a cell or spatial location. To ensure that SpatialDE captures the spatial covariance for each gene beyond this effect, we regress log total count values out from the Anscombe-transformed expression values before fitting the spatial models.

**Code availability.** An open source implementation of SpatialDE is available at GitHub (<https://github.com/Teichlab/SpatialDE>). The release includes tutorials and example vignettes for reproducing the presented analyses, as well as all preprocessed data sets considered in this study. The software version used to generate the results presented in this paper is also available as **Supplementary Software**.

**Life Sciences Reporting Summary.** Further information on experimental design is available in the **Life Sciences Reporting Summary**.

**Data availability.** In addition to the data sources mentioned above, all data used for analysis are available at GitHub (<https://github.com/Teichlab/SpatialDE>) using git-lfs. All results from analysis are reported in **Supplementary Table 1**.

21. Rasmussen, C.E. & Williams, C.K.I. *Gaussian Processes for Machine Learning* (MIT Press, 2006).
22. Zhou, X. & Stephens, M. *Nat. Genet.* **44**, 821–824 (2012).
23. Storey, J.D. & Tibshirani, R. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
24. Bishop, C.M. *Pattern Recognition and Machine Learning* (Springer, 2006).
25. Wolf, F.A., Angerer, P. & Theis, F.J. *Genome Biol.* **19**, 15 (2018).
26. Krige, D.G. *J. S. Afr. Inst. Min. Metall.* **52**, 119–139 (1951).
27. Stegle, O. et al. *J. Comput. Biol.* **17**, 355–367 (2010).
28. Lönnberg, T. et al. *Sci. Immunol.* **2**, eaal2192 (2017).
29. Åijö, T. et al. *Bioinformatics* **30**, i113–i120 (2014).
30. Macaulay, I.C. et al. *Cell Rep.* **14**, 966–977 (2016).
31. Eckersley-Maslin, M.A. et al. *Cell Rep.* **17**, 179–192 (2016).
32. Lloyd, J.R., Duvenaud, D., Grosse, R., Tenenbaum, J.B. & Ghahramani, Z. in *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence* 1242–1250 (AAAI Press, 2014).

## Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

### ► Experimental design

#### 1. Sample size

Describe how sample size was determined.

Only previously published data was used.

#### 2. Data exclusions

Describe any data exclusions.

For all data sets genes with fewer than 3 measurements were excluded from analysis as they are guaranteed to get a P-value of 1.

#### 3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

No experiments were performed.

#### 4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No experiments were performed.

#### 5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not considered necessary and not performed.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

#### 6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

- ☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- ☐ ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ A statement indicating how many times each experiment was replicated
- ☐ ☒ The statistical test(s) used and whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- ☐ ☒ Test values indicating whether an effect is present  
*Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
- ☐ ☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- ☐ ☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on [statistics for biologists](#) for further resources and guidance.

## ► Software

Policy information about [availability of computer code](#)

### 7. Software

Describe the software used to analyze the data in this study.

Analyses were performed using our new software, as well as classical methods from the scikit-learn package and SciPy. All analysis scripts as well as our software is available at <http://github.com/Teichlab/SpatialDE>

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

## ► Materials and reagents

Policy information about [availability of materials](#)

### 8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No physical material were used in the study.

### 9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used in the study.

### 10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No cell lines were used.

b. Describe the method of cell line authentication used.

No cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

No cell lines were used.

## ► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

### 11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used in this study.

Policy information about [studies involving human research participants](#)

### 12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

This study did not involve any human participants.