SpatialDE - Identification of spatially variable genes

Valentine Svensson[1,3], Sarah A Teichmann[1,2], Oliver Stegle[3]

1. Wellcome Trust Sanger Institute, Wellcome Genome Campus, CB10 1SD Hinxton, Cambridge, UK
2. Theory of Condensed Matter Group, Cavendish Laboratory, 19 JJ Thomson Avenue, CB3 0HE, Cambridge, U.K
3. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD, Hinxton, Cambridge, UK

Abstract

Technological advances have enabled low-input RNA-sequencing, paving the way for assaying transcriptome variation in spatial contexts, including in tissue systems. While the generation of spatially resolved transcriptome maps is increasingly feasible, computational methods for analysing the resulting data are not established. Existing analysis strategies either ignore the spatial component of gene expression variation, or require discretization.

To address this, we have developed SpatialDE, a computational framework for identifying and characterizing spatially variable genes. Our method generalizes variable gene selection, as used in population- and single-cell studies, to spatial expression profiles. We apply SpatialDE to Spatial Transcriptomics and to data from single cells expression profiles using multiplexed In Situ Hybridisation (SeqFISH and MERFISH), demonstrating its general use. SpatialDE identifies genes with expression patterns that are associated with histology in breast cancer tissue, several of which have known disease implications and are not detected by variable gene selection. Additionally, our model can be used to classify genes with distinct spatial patterns, including periodic expression profiles, linear trends and general spatial variation

Main text

Technological advances have helped to miniaturize and parallelize genomics, thereby enabling high-throughput transcriptome profiling from low quantities of starting material, including in single cells. Increased experimental throughput has also fostered new experimental designs, where in particular the spatial context of gene expression variation can now be directly assayed, which is critical for decoding complex tissues from multicellular organisms. The spatial context of gene expression is crucial in determining functions and phenotypes of cells[1,2]. In many cases a gene's expression is determined by cellular communication, and in other cases cells migrate to specific locations in tissue to perform their functions.

Several experimental methods to measure gene expression levels in a spatial context have been established, which differ in resolution, accuracy and throughput. These include the computational assignment of transcriptome-profiles from dissociated cells to a spatial reference[3,4], parallel profiling of mRNA using barcodes on a grid of known spatial locations[5–7], and methods based on multiplexed *in situ* hybridization[8,9] or sequencing[10–12].

A first critical step in the analysis of the resulting datasets is to identify the genes that exhibit spatial variation across the tissue. However, existing approaches designed to identify highly variable genes[13,14], used in e.g.

single-cell RNA-sequencing (scRNA-seq) studies, ignore the spatial location and hence do not measure *spatial* variability. Alternatively, researchers have applied ANOVA to test for differential expression between groups of cells, either derived using *a priori* defined (discrete) cell annotations, or based on clustering[3,4,7,8,10], with some clustering strategies incorporating spatial information[15]. Importantly, such strategies fall short in detecting variation that is not well captured by discrete groups, including linear and nonlinear trends, periodic expression patterns and other complex patterns of expression variation.

To address this, we here propose a computational approach termed *SpatialDE* for identifying and characterizing *spatially variable* genes (SV genes). Our method builds on Gaussian Process Regression, a class of models that is widely used in geostatistics, also known as Kriging[16]. For each gene, our model decomposes the expression variability into a spatial and non-spatial component (**Figure 1A**). Significant SV genes can then be identified by comparing this full model to a model that assumes no spatial dependency of expression variation (**Figure 1B**, **Methods**).

In addition to identifying spatially variable genes, SpatialDE also allows for classifying the spatial patterns of individual genes, differentiating between linear trends, periodic expression profiles or general spatial dependencies (**Figure 1C**). By interpreting the fitted model parameters it is possible to identify the length scale (the expected number of changes in direction in a unit interval[16]) or the period length of spatial patterns for individual genes (**Figure 1B**). Finally, SpatialDE achieves unprecedented computational efficiency[1] by leveraging computational tricks for efficient inference in linear mixed models[18] and precomputing operations where possible (**Methods**, **Figure 1C**). Taken together,SpatialDE is a widely applicable tool for the initialanalysis of spatial transcriptomics datasets.

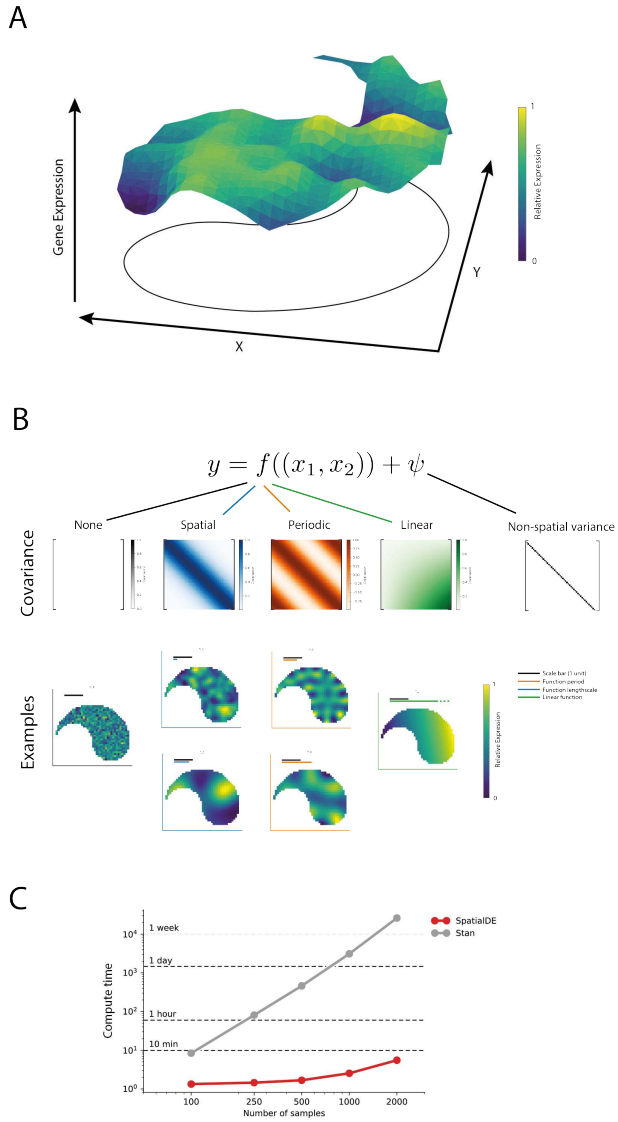


Figure 1 - Overview of SpatialDE for the identification of spatially variable genes. (**A**) In spatial gene expression studies, expression levels vary in ways that depend on spatial coordinates. SpatialDE defines spatial dependence for a given gene using a non-parametric approach, testing whether gene expression levels at different locations covary in a manner that depends on their relative location. (**B**) SpatialDE partitions the expression variation into a spatial component (using functional dependencies f(x,y)), characterized by alternative spatial covariances, and observation noise (Psi). Alternative spatial covariance models considered by SpatialDE: no spatial effect (null model), general spatial, periodic spatial patterns and linear trends. Example expression patterns with the covariances plotted below corresponding matrix. (**C**) Computational efficiency of SpatialDE compared to a Stan[17] implementation of the same model. Caching operations and linear algebra speedups are used where possible, enabling genome-wide analyses with thousands of samples. Benchmarks performed on a late 2013 iMac with 3.2 GHz Intel Core i5 processor.

[1] Comparison made with implementation of the same model in *Stan*[17]

First, we applied our method to Spatial Transcriptomics (ST) data from breast cancer tissue[7]. Briefly, ST gene expression levels are derived from thin tissue sections of frozen material, placed on an array with poly(dT) probes and spatially resolved DNA barcodes in a grid of "spots". Following permeabilization, the mRNA is captured by the probes, and the spatial location can be recovered from sequenced barcodes. The resulting gene expression profiles can be analysed in context with hematoxylin and eosin (HE) stained microscopic images of the tissue (**Figure 2A**).

SpatialDE identified 115 SV genes (FDR < 0.05). Notably, seven highly ranking genes were also included in a set of 14 genes with known roles in the disease that were highlighted in the primary analysis of the data (**Figure 2C,** red text). Significantly SV genes were enriched for collagens, which distinguish tissue substructure[19] (Reactome term "Collagen formation", $P < 5 * 10^{-14}$ using gProfiler[20], **Supp. Table 1**). Additionally, we identified the autophagy related gene, *TP53INP2,* surrounding the fatty tissue (Q = 0.022, **Figure 2B,** extended examples **Supp. Fig. 1**). Interestingly, the set of SV genes also included the cytokines *CXCL9* ($Q = 5.4 * 10^{-4}$) and *CXCL13* ($Q = 1.3 * 10^{-4}$), both of which are expressed in a visually distinct region (**Figure 2A**, black arrow), together with the IL12 receptor subunit gene *IL12RB1* ($Q = 2.8 * 10^{-4}$), indicating a potential tumour related immune response in the tissue. Notably, neither of these genes (and N=29 others), were identified as differentially expressed when applying clustering in conjunction with an ANOVA test between the identified groups of cells (**Supp. Fig. 2**). Nor did they have high rank based on conventional Highly Variable Genes measures (such as the mean - CV2 relation[13] or mean - dropout rate relation[21]), measures that do not take the spatial context into account (**Supp. Fig. 3**). Generally, we observed that variable genes detected by SpatialDE are complementary to existing methods, where in particular spatially variable genes with localized expression patterns, as indicated by small fitted length scales, or periodic patterns, are not detected by methods that ignore spatial contexts (**Supp. Fig. 2E**). Finally, we confirmed the statistical calibration and the robustness of SpatialDE using randomization experiments (**Supp. Fig. 4**).

As a second application, we considered a study of mouse olfactory bulb[7], profiled using the same ST protocol. Again, SpatialDE identified SV genes with clear spatial sub-structure, consistent with the matched HE stained image (**Supp. Fig. 5A-B**). These included canonical marker genes highlighted in Stahl et al, such as *PENK*, *DOC2G*, and *KCTD12*, but also additional genes that define the granule cell layer (GCL) 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 (including *KCNH3*, *NRGN*, or *MBP* with 1.8 mm period length, **Supp. Fig. 5C**). Other genes with periodic patterns, such as the vesicular glutamate transporter *SLC17A7*, were identified with shorter periods (1.1 mm), and inspection revealed regularly dispersed regions, potentially identifying a pattern of regions with higher neuron density[22]. This suggests that periodic expression patterns in tissue contexts are a biological feature of interest to understand tissue biology.

Taken together, these results demonstrate that SpatialDE can be used to characterize clinically relevant features in spatial tissue samples in the absence of *a priori* histological annotation.

SpatialDE is not limited to sequencing technologies, and can be applied to any expression datatype with spatial and/or temporal resolution. To explore this, we applied the method to data generated using multiplexed is single molecule FISH (smFISH), a recent technological development that allows for quantifying gene expression with subcellular resolution for larger numbers of target genes in parallel. Briefly, probes are hybridized to RNA while carrying barcodes of fluorophores, which allows for quantifying gene expression of up to several thousands of probes[23] using high-content imaging.

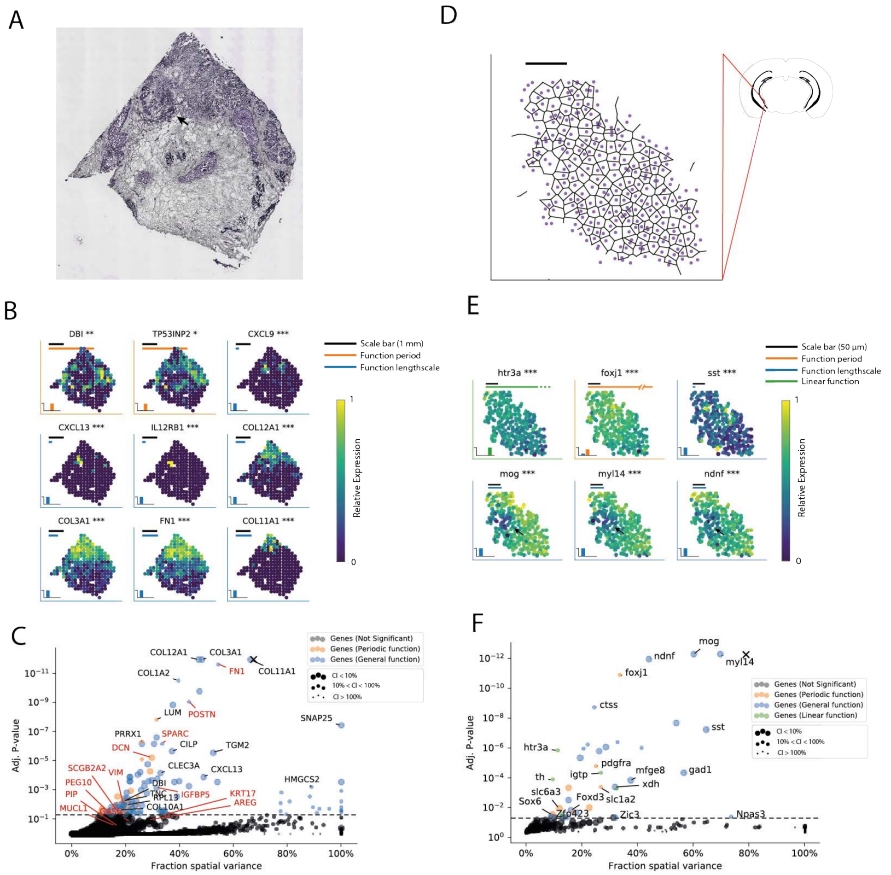


Figure 2 - Applications of SpatialDE to Spatial Transcriptomics and data generated using SeqFISH. (A) Correlated image of breast cancer tissue from Spatial Transcriptomics[7]. **(B)** Visualization of nine selected spatially variable genes (out of 115, FDR<0.05). The black scale bar corresponds to 1 mm. For genes identified with periodic dependencies, the orange bar shows the fitted period length on the same scale. Analogously, the blue bar shows the fitted length scale for genes with general spatial trends. 2D plots show the relative expression level for genes across the tissue section coded in color. Stars next to gene names denote significance levels (* Q < 0.05 , ** Q < 0.01, *** Q < 0.001) of spatial variation. Insets in lower left show the posterior probability of these three function classes for each gene. **(C)** Proportion of variance (x-axis) explained by spatial variation (FSV) versus adj. P-value (y-axis, FDR adjusted[26]) for 12,856 genes. Dashed line corresponds to the FDR=0.05 significance level (N=115 genes). 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.[7]) are indicated with red labels, and are significantly enriched in SpatialDE results ($P=10^{-11}$, Fisher exact test). Other representative genes selected by stratifying over function periods / length scales are annotated with black labels. Size of of points indicate certainty in the estimate of Fraction Spatial Variance (FSV), larger points have smaller standard deviation. The X symbol show the result of running SpatialDE on the estimated total RNA content per spot. **(D)** SeqFISH data from a region of mouse hippocampus from Shah *et al*[8]. Black scale bar correspond to 50 µm, Voronoi tessellation representative of tissue structure. **(E)** Expression patterns of six selected SV genes analogous to panel B (out of 32, FDR < 0.05). Shown are genes with linear (*htr3a*), periodic (*foxj1*), and generally spatial models. Black arrows indicate distinct region of low expression of *Mog*, *Myl14* and *Ndnf*. **(F)** Proportion of variance (x-axis) versus adj. P-value (y-axis, FDR adjusted) for 249 genes, as in (C). Genes with a linear dependency are highlighted in green.

We applied SpatialDE to Multiplexed smFISH data of cells from mouse hippocampus, generated using SeqFISH[8]. This study considered 249 genes that were chosen to investigate the cell type composition along dorsal and ventral axes of the hippocampus (**Figure 2D**). SpatialDE identified 32 SV genes (FDR<0.05), with the three highest ranking genes, *MOG* ($Q = 10^{-14}$), *MYL14*, ($Q = 10^{-14}$) and *NDNF* ($Q = 2 * 10^{-12}$) displaying a distinct region of lower expression (**Figure 2E**, black arrows). Again, SpatialDE identified genes with different types of spatial variation, including linear trends (N=5) and periodic patterns (N=8, **Figure 2F**, extended examples in **Supp. Fig. 6**).

SpatialDE can also be used to test for spatial expression variation in cell culture systems, where spatial variation may not be expected *a priori*. We explored this, and considered data from another recent multiplexed smFISH dataset generated using MERFISH with 140 probes from a human osteosarcoma cell culture[9] (**Supp. Fig. 7A-B**). Interestingly, the model revealed that a substantial proportion of the genes assayed were spatially variable (N=92, 65%, FDR < 0.05). This reconstitutes results from the primary analysis, where the authors noted spatially restricted populations of cells with higher proliferation rates. Indeed, six of the seven genes highlighted as differential between proliferation subpopulations were identified as SV genes (e.g. *THBS1* and *CENPF1*, **Supp Fig. 7C)**. This result is also consistent with previous studies which observed that high confluence in cell culture, promoting cell-to-cell communication and crowding, leads to spatial dependency in gene expression[24]. We also considered negative control probes in the data, which were not detected as spatially variable,thereby confirming the statistical calibration of SpatialDE (**Supp Fig. 7D**).

Herein, we have presented a method for identifying spatially variable genes. The commoditization of high-throughput experiments, 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 model to data from multiple different protocols, from Spatial Transcriptomics to multiplexed single-molecule FISH, considering both tissue systems and cell lines. The extent of spatial variation we observed in cell lines may be surprising, a result that is consistent with recent studies that have reported coordinated expression changes across neighbouring cells[24]. The method is also applicable to temporal data from time-course experiments (**Supp. Fig. 8**), and it can be applied without modification to 3-dimensional data from e.g. *in situ* sequencing when such technologies mature[11,12].

SpatialDE generalizes previous approaches for the detection of highly variable genes, most notably methods designed for conventional scRNA-seq[13]. Our model separates spatial variation from non-spatial effects, which may include biological and technical variability. Underlying this approach is the assumption that technical noise is independent across sampling positions, which circumvents the need to explicitly model technical sources of variation, which enables applications to virtually any protocol.

Future extension of SpatialDE could be tailored towards specific platforms, for example to make use of spike-in standards or unique molecular identifier, thereby explicitly estimating technical variation. Another area of future work are extensions for incorporating information about the tissue makeup or local differences in cell density. Our framework also opens up the possibility for future work to define spatial patterns that are common to groups of genes, using clustering combined with the spatial Gaussian Process framework[25]

Availability of code and data. SpatialDE is implemented in Python 3.5. The open source implementation is available from https://github.com/Teichlab/SpatialDE together with a Stan version, and can be installed from PyPI using the command 'pip install spatialde'. An R-based Bioconductor implementation is in preparation. Tutorials and example vignettes for reproducing the presented analyses can be obtained here. The pre-processed datasets from the public studies we have considered can be obtained from the same repository.

Acknowledgements. The authors wish to thank Damien Arnol and Francesco Paolo Casale for helpful advice on statistics and data normalisation. In addition we are thankful to Martin Hemberg, Daniel Kunz, and Kerstin Meyer for feedback on the manuscript. V.S. was supported by the EMBL International PhD Program, S.A.T. was supported by the Wellcome Trust and ERC Consolidator Grant "ThDEFINE", O.S. received funding from EMBL core funding.

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

Figure Legends

Figure 1

Overview of SpatialDE for the identification of spatially variable genes. (**A**) In spatial gene expression studies, expression levels vary in ways that depend on spatial coordinates. SpatialDE defines spatial dependence for a given gene using a non-parametric approach, testing whether gene expression levels at different locations covary in a manner that depends on their relative location. (**B**) SpatialDE partitions the expression variation into a spatial component (using functional dependencies f(x,y)), characterized by alternative spatial covariances, and observation noise (Psi). Alternative spatial

covariance models considered by SpatialDE: no spatial effect (null model), general spatial, periodic spatial patterns and linear trends. Example expression patterns with the covariances plotted below corresponding matrix. (**C**) Computational efficiency of SpatialDE compared to a Stan[17] implementation of the same model. Caching operations and linear algebra speedups are used where possible, enabling genome-wide analyses with thousands of samples. Benchmarks performed on a late 2013 iMac with 3.2 GHz Intel Core i5 processor.

Figure 2

Applications of SpatialDE to Spatial Transcriptomics and data generated using SeqFISH. (**A**) Correlated image of breast cancer tissue from Spatial Transcriptomics[7]. (**B**) Visualization of nine selected spatially variable genes (out of 115, FDR<0.05). The black scale bar corresponds to 1 mm. For genes identified with periodic dependencies, the orange bar shows the fitted period length on the same scale. Analogously, the blue bar shows the fitted length scale for genes with general spatial trends. 2D plots show the relative expression level for genes across the tissue section coded in color. Stars next to gene names denote significance levels (* Q < 0.05 , ** Q < 0.01, *** Q < 0.001) of spatial variation. Insets in lower left show the posterior probability of these three function classes for each gene. (**C**) Proportion of variance (x-axis) explained by spatial variation (FSV) versus adj. P-value (y-axis, FDR adjusted[26]) for 12,856 genes. Dashed line corresponds to the FDR=0.05 significance level (N=115 genes). 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.[7]) are indicated with red labels, and are significantly enriched in SpatialDE results ($P=10^{-11}$, Fisher exact test). Other representative genes selected by stratifying over function periods / length scales are annotated with black labels. Size of of points indicate certainty in the estimate of Fraction Spatial Variance (FSV), larger points have smaller standard deviation. The X symbol show the result of running SpatialDE on the estimated total RNA content per spot. (**D**) SeqFISH data from a region of mouse hippocampus from Shah *et al*[8]. Black scale bar correspond to 50 µm, Voronoi tessellation representative of tissue structure. (**E**) Expression patterns of six selected SV genes analogous to panel B (out of 32, FDR < 0.05). Shown are genes with linear (*htr3a*), periodic (*foxj1*), and generally spatial models. Black arrows indicate distinct region of low expression of *Mog*, *Myl14* and *Ndnf*. (**F**) Proportion of variance (x-axis) versus adj. P-value (y-axis, FDR adjusted) for 249 genes, as in (C). Genes with a linear dependency are highlighted in green.

Supplementary information

Supp. Methods

Full derivation of the SpatialDE model.

Supp. Fig. 1

Expanded example of Breast Cancer tissue genes. Spatial expression pattern for 37 additional SV genes (out of 115), selected to represent patterns from different function periods and length scales to illustrate different spatial patterns.

Supp. Fig. 2

Comparison to differential expression analysis using clustering. (A) Principal Component Analysis of individual "spots", color coded by cluster membership for N=4 clusters (identified by Bayesian Gaussian Mixture Modelling). **(B)** Bayesian Gaussian Mixture Model cluster probabilities, the 250 spatial breast cancer "spots" can be clustered into four groups when ignoring spatial structure. (**C**) Visualization of cluster membership in the original tissue context. (**D**) Comparison of P-values from an ANOVA test between clusters (x-axis) with significance from SpatialDE (y-axis). 83 genes are identified as significantly variable by both approaches; 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 detected only by SpatialDE (orange). Genes detected only by SpatialDE have smaller length scales, indicating more localized expression patterns.

Supp. Fig. 3

Comparison of SpatialDE to other measures of expression heterogeneity. (**A**) Comparison of P-values from SpatialDE to other commonly used summary statistics - Upper left: Mean, Upper right: Variance, Lower left: CV2 (squared coefficient of variation), Lower right: Dropout rate (fraction of samples a gene is not detected in). Random selection of significant SV genes highlighted in red for context. (**B**) Comparison with common strategies to define highly variable genes, which are based on regression models between summary statistics: Relation with CV2 (Upper) or Variance (Middle), or with dropout fraction (Bottom). Model residuals are compared with the SpatialDE significance to the right of the relation. Polynomial regression for CV2 and Variance, logistic regression for dropout rate. Significant SV genes as identified by SpatialDE are shown in grey. Other, non-significant genes are shown in solid black.

Supp. Fig. 4

Statistical calibration of SpatialDE. (**A**) QQ-plot of expected P-values (Chi2 distribution with 1 degree of freedom) compared to observed P-values derived using the log likelihood ratio test in SpatialDE. (**B**) To simulate data from an empirical null, without spatial structure, expression values were shuffled among the sampled coordinates. Shown is *COL3A1* expression as an example. (**C**) Adj. P-values for genes on shuffled data, which are generally below the FDR = 0.05 threshold. (**D**) Analogous QQ-plot as in A on shuffled expression values. P-values follows the null distribution, indicating that the model is calibrated.

Supp. Fig. 5

Application to Mouse Olfactory Bulb tissue. (**A**) The corresponding image for mouse olfactory bulb data from Stahl et al. (**B**) SpatialDE identified 67 spatially variable genes (SV genes, FDR < 0.05). Of these, 19 were assigned to periodic functions. Genes highlighted in Stahl et al are displayed in red, representative examples of SV genes are annotated with black text (Colors and sizes as in Figure 2). (**C**) Representative examples of SV gene with different periods and length scales (indicated in orange and blue bars, respectively, relative to scale bar). Black scale bar correspond to 1 mm. Colors and significance levels as described in Figure 2.

Supp. Fig. 6

Expanded examples of significant spatially variable genes for the mouse hippocampus dataset. Visualization of 24 SV genes with from the mouse hippocampus SeqFISH data, showing selected genes with periodic, linear, and general spatial dependencies with different estimated length scales. Black scale bar correspond to 50 µm.

Supp. Fig. 7

Application to MERFISH data. (**A**) In MERFISH study of an osteosarcoma cell culture from Moffitt et al[9] the majority of genes are found spatially variable. 21 of 92 significant SV genes were assigned to a periodic function by the model, and 9 genes had linear functions. Negative control probes are indicated with red labels. Genes indicated as enriched in proliferating cells in the original study 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 totalling 38,594 cells (upper). We analysed 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 visualised for selected SV genes with various fitted periods and length scales (Significance levels

and colors as in Figure 2). Black scale bar correspond to 200 µm. (**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.

Supp. Fig. 8

Application to expression time-course data. (**A**) When applying SpatialDE to developmental time course data from Owens et al[27], the majority of genes were found differentially expressed (21,009 out of 22,256 genes, FDR < 0.05). Of these, 241 were assigned to periodic patterns, and 269 were detected with linear trends. Colors and point sizes as in Figure 2. The X marks indicates result of running test on ERCC content and number of detected genes. (**B**) Examples of temporally DE 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) (**C**) The expression patterns of the top 400 significantly SV genes are visualised, ordered by the time they reach their highest expression value. Example genes from **B** are annotated.

Online Methods

The method section is attached as supplementary file (Supplementary methods).

References

1. Ledford, H. The race to map the human body - one cell at a time. *Nature* **542,** 404–405 (2017).
2. Lee, J. H. Quantitative approaches for investigating the spatial context of gene expression. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **9,** (2017).
3. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33,** 503–509 (2015).
4. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).
5. Junker, J. P. *et al.* Genome-wide RNA Tomography in the zebrafish embryo. *Cell* **159,** 662–675 (2014).
6. Chen, J. *et al.* Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat.*

Protoc. **12,** 566–580 (2017).

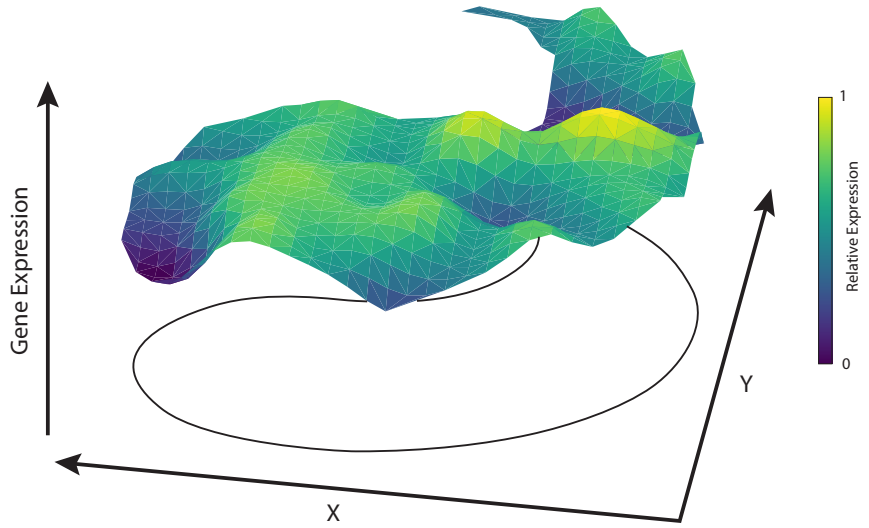
7. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82 (2016).
8. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92,** 342–357 (2016).
9. Moffitt, J. R. *et al.* High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 11046–11051 (2016).
10. Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10,** 857–860 (2013).
11. Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* **343,** 1360–1363 (2014).
12. Lee, J. H. *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10,** 442–458 (2015).
13. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–1095 (2013).
14. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Comput. Biol.* **11,** e1004333 (2015).
15. Pettit, J.-B. *et al.* Identifying cell types from spatially referenced single-cell expression datasets. *PLoS Comput. Biol.* **10,** e1003824 (2014).
16. Rasmussen, C. E. & Williams, C. Gaussian Processes for Machine Learning, Model Selection and Adaptation of Hyperparameters, Chapter 5. (2006).
17. Carpenter, B. *et al.* Stan: A probabilistic programming language. *J. Stat. Softw.* **20,** (2016).
18. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8,** 833–835 (2011).
19. Seewaldt, V. L. Cancer: Destiny from density. *Nature* **490,** 490–491 (2012).
20. Reimand, J. *et al.* g:Profiler-a web server for functional interpretation of gene lists (2016 update).

Nucleic Acids Res. **44,** W83–9 (2016).

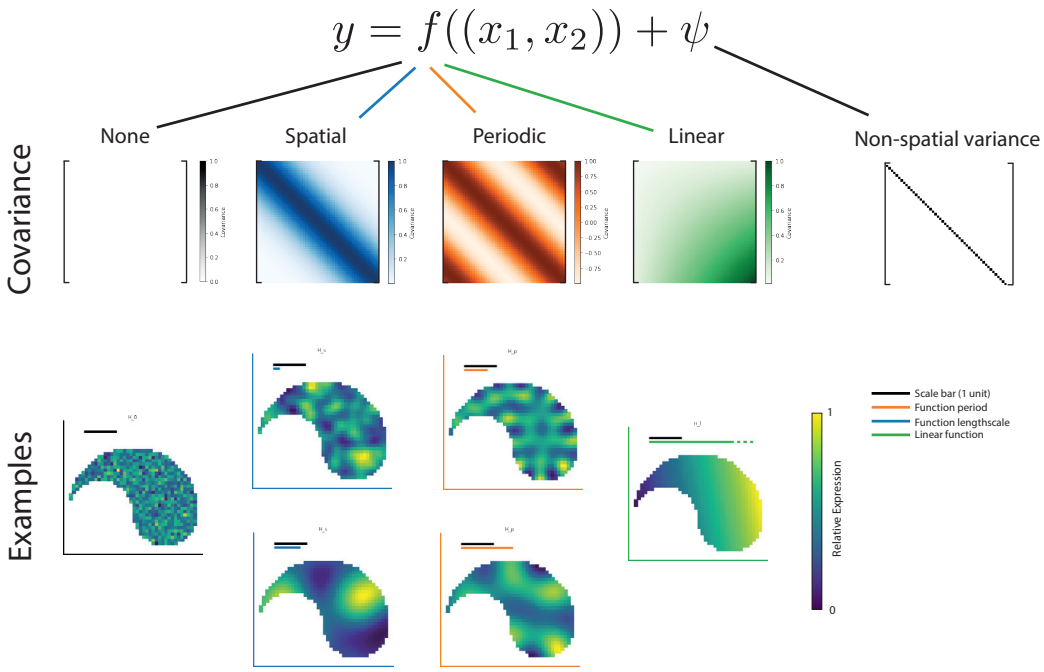
21. Andrews, T. S. & Hemberg, M. Modelling dropouts allows for unbiased identification of marker genes in scRNASeq experiments. *bioRxiv* (2016).
22. Jahn, R., Takamori, S., Rhee, J. S. & Rosenmund, C. 10.1038/35025070. *Nature* **407,** 189–194 (2000).
23. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348,** aaa6090 (2015).
24. Battich, N., Stoeger, T. & Pelkmans, L. Control of Transcript Variability in Single Mammalian Cells. *Cell* **163,** 1596–1610 (2015).
25. Hensman, J., Rattray, M. & Lawrence, N. D. Fast Nonparametric Clustering of Structured Time-Series. *IEEE Trans. Pattern Anal. Mach. Intell.* **37,** 383–393 (2015).
26. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 9440–9445 (2003).
27. Owens, N. D. L. *et al.* Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. *Cell Rep.* **14,** 632–647 (2016).

Figure 1

A


Gene Expression
X
Y
Relative Expression
1
0

B


$y = f((x_1, x_2)) + \psi$
None
Spatial
Periodic
Linear
Non-spatial variance
Covariance
Examples
Scale bar (1 unit)
Function period
Function lengthscale
Linear function
Relative Expression
1
0

C

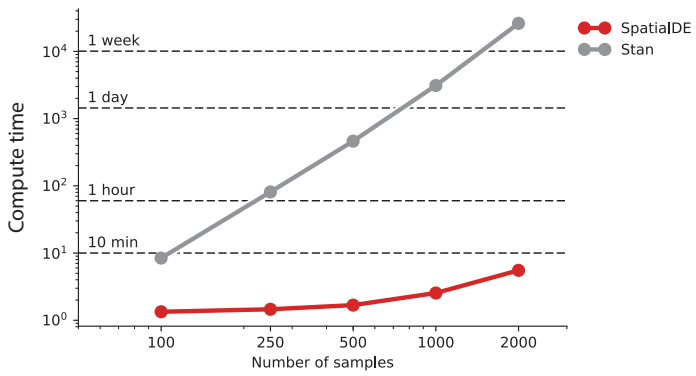
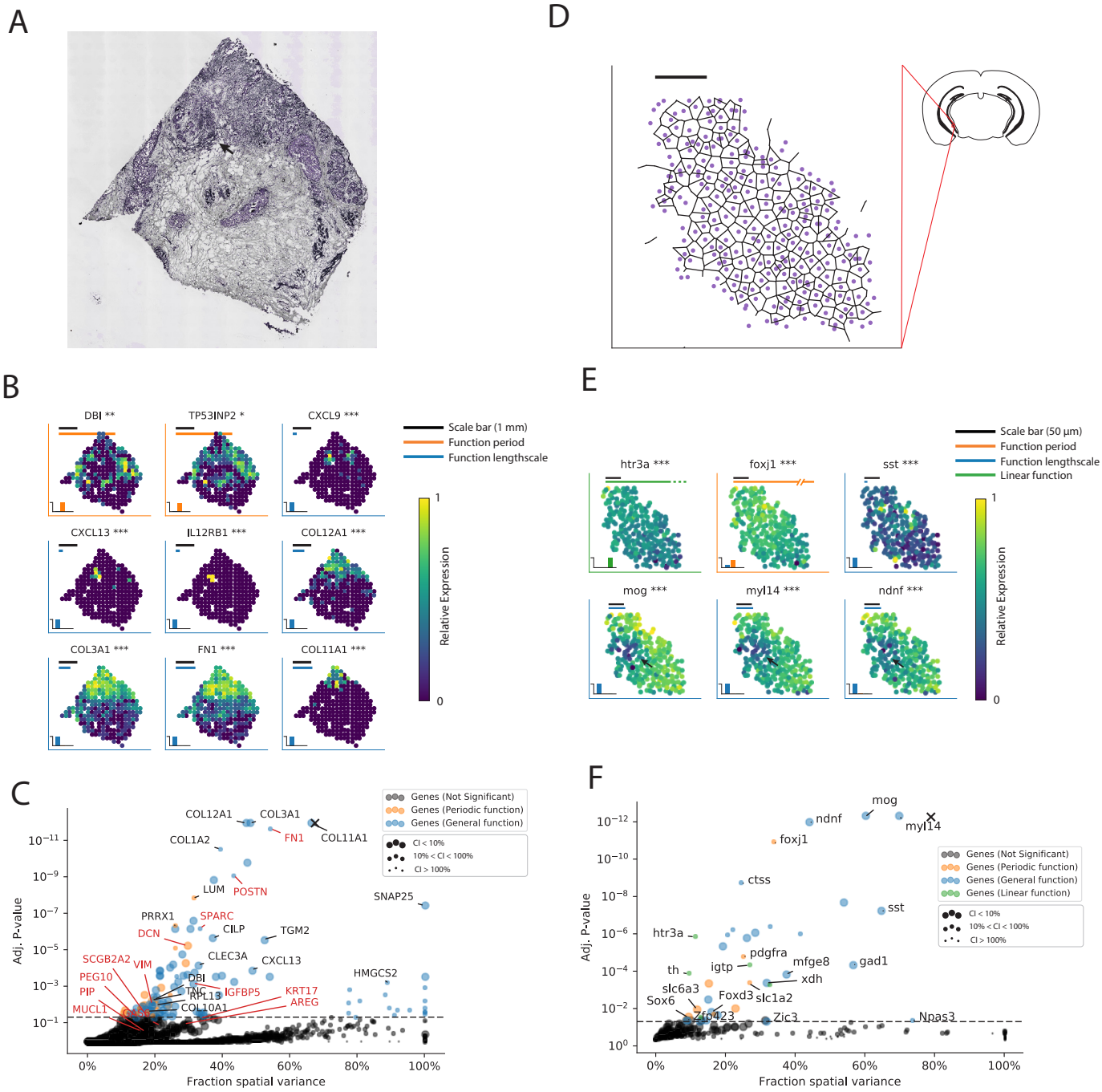
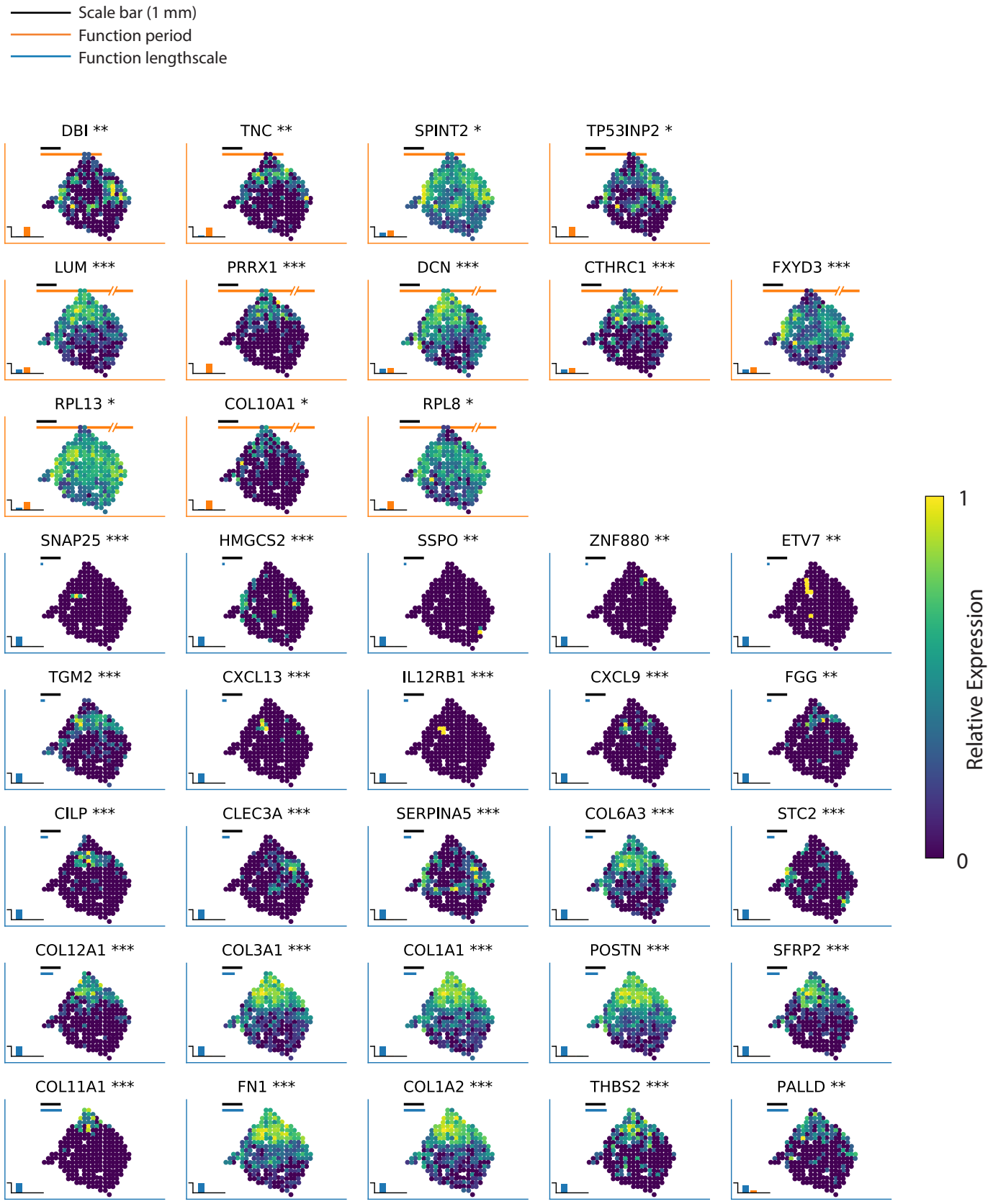

SpatialDE
Stan
1 week
1 day
1 hour
10 min
Compute time
10^4
10^3
10^2
10^1
10^0
100
250
500
1000
2000
Number of samples

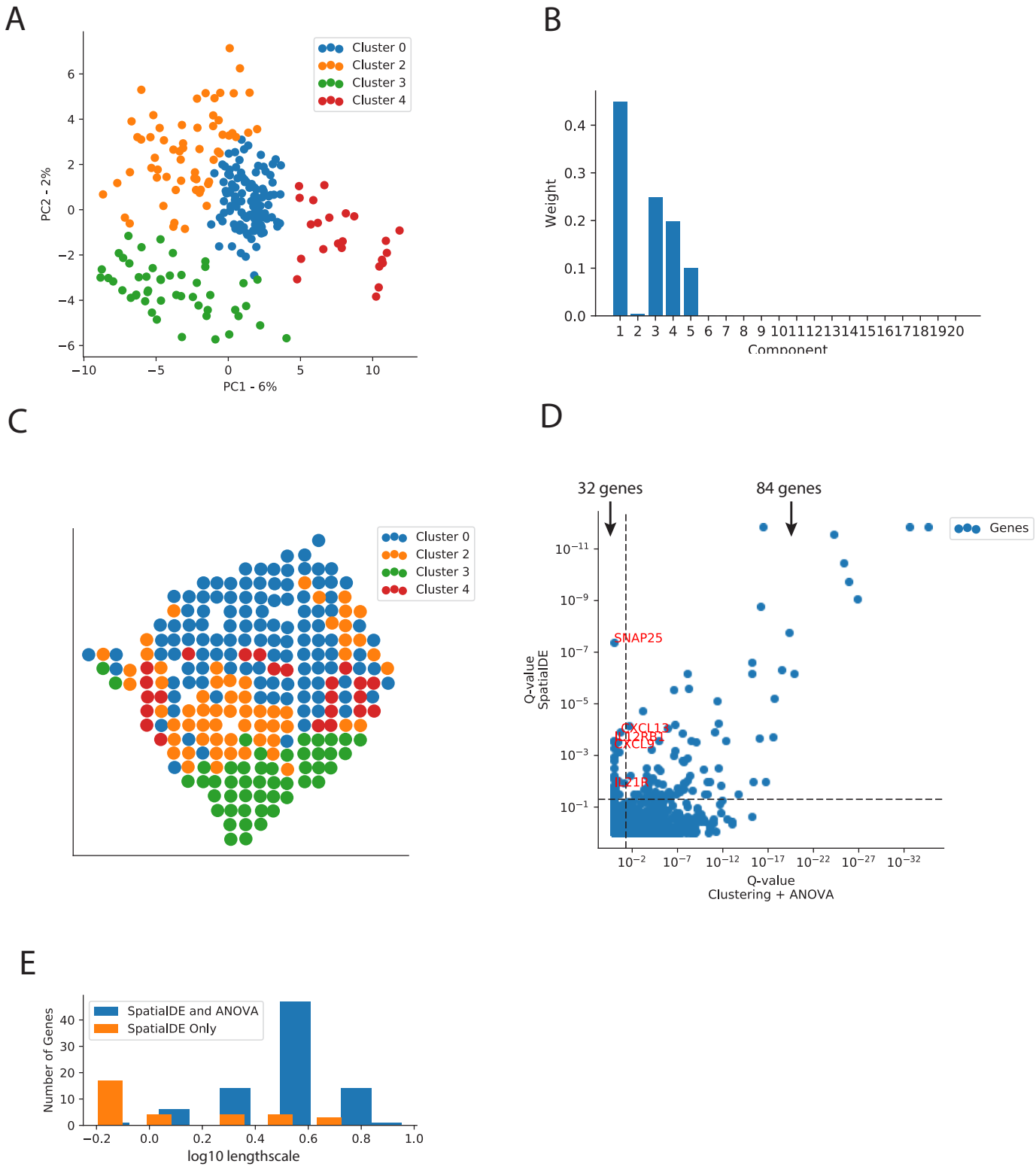
Figure 2


A
D
B
DBI **
TP53INP2 *
CXCL9 ***
CXCL13 ***
IL12RB1 ***
COL12A1 ***
COL3A1 ***
FN1 ***
COL11A1 ***
Scale bar (1 mm)
Function period
Function lengthscale
Relative Expression
1
0
E
htr3a ***
foxj1 ***
sst ***
mog ***
myl14 ***
ndnf ***
Scale bar (50 µm)
Function period
Function lengthscale
Linear function
Relative Expression
1
0
C
Genes (Not Significant)
Genes (Periodic function)
Genes (General function)
CI < 10%
10% < CI < 100%
CI > 100%
COL12A1
COL3A1
FN1
COL11A1
COL1A2
LUM
POSTN
SNAP25
PRRX1
SPARC
DCN
CILP
TGM2
CLEC3A
CXCL13
SCGB2A2
VIM
PEG10
PIP
MUCL1
DBI
TNC
RPL13
IGFBP5
COL10A1
KRT17
AREG
HMGCS2
Adj. P-value
Fraction spatial variance
0%
20%
40%
60%
80%
100%
F
mog
ndnf
myl14
foxj1
ctss
sst
htr3a
pdgfra
mfge8
gad1
th
igtp
xdh
slc6a3
Foxd3
slc1a2
Sox6
Zfp423
Zic3
Npas3
Genes (Not Significant)
Genes (Periodic function)
Genes (General function)
Genes (Linear function)
CI < 10%
10% < CI < 100%
CI > 100%
Adj. P-value
Fraction spatial variance
0%
20%
40%
60%
80%
100%

Supp. Fig. 1

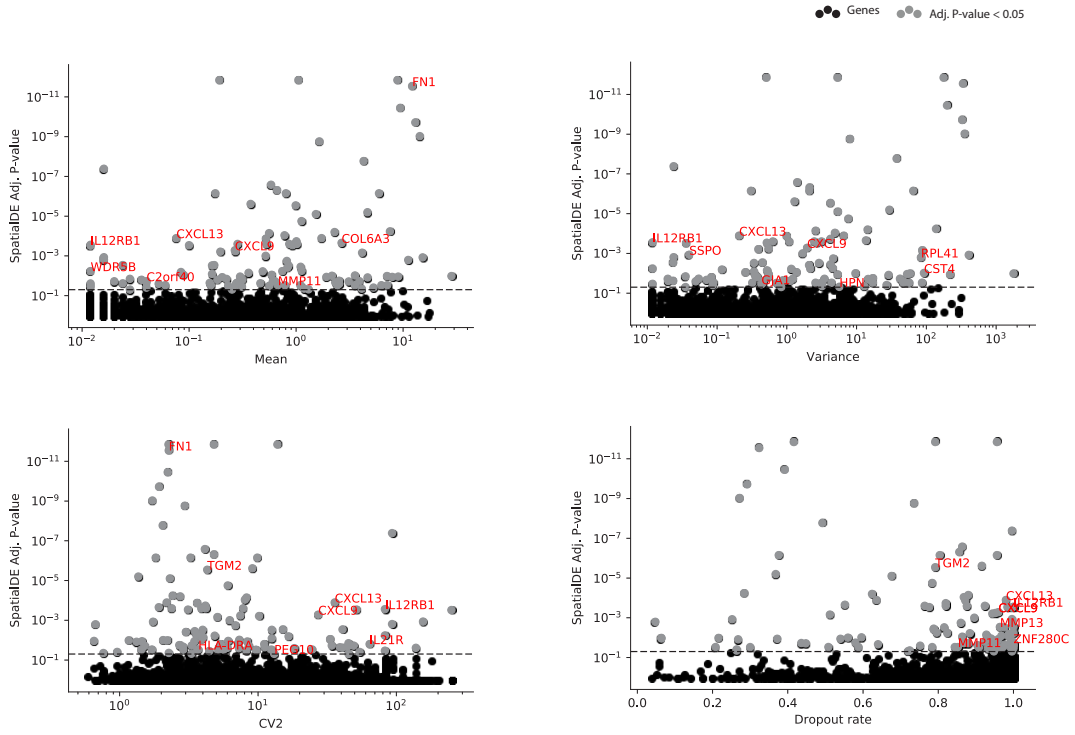

Scale bar (1 mm)
Function period
Function lengthscale
DBI **
TNC **
SPINT2 *
TP53INP2 *
LUM ***
PRRX1 ***
DCN ***
CTHRC1 ***
FXYD3 ***
RPL13 *
COL10A1 *
RPL8 *
SNAP25 ***
HMGCS2 ***
SSPO **
ZNF880 **
ETV7 **
TGM2 ***
CXCL13 ***
IL12RB1 ***
CXCL9 ***
FGG **
CILP ***
CLEC3A ***
SERPINA5 ***
COL6A3 ***
STC2 ***
COL12A1 ***
COL3A1 ***
COL1A1 ***
POSTN ***
SFRP2 ***
COL11A1 ***
FN1 ***
COL1A2 ***
THBS2 ***
PALLD **
1
0
Relative Expression

Supp. Fig. 2

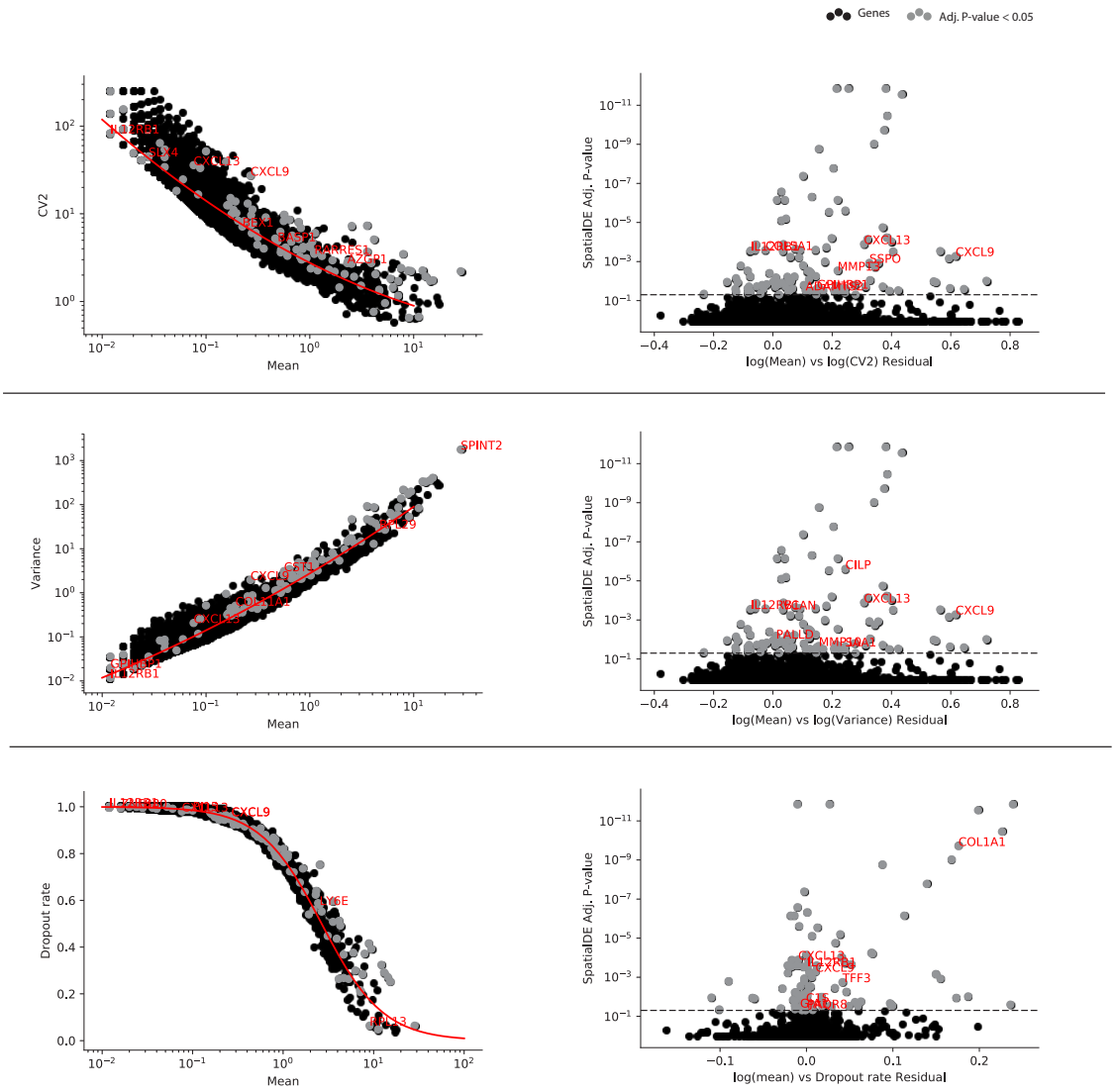

A
Cluster 0
Cluster 2
Cluster 3
Cluster 4
PC2 - 2%
PC1 - 6%
B
Weight
Component
C
Cluster 0
Cluster 2
Cluster 3
Cluster 4
D
32 genes
84 genes
Genes
SNAP25
CXCL13
CD2BP1
CXCL9
IL21R
Q-value
SpatialDE
Q-value
Clustering + ANOVA
E
SpatialDE and ANOVA
SpatialDE Only
Number of Genes
log10 lengthscale

Supp. Fig. 3

A

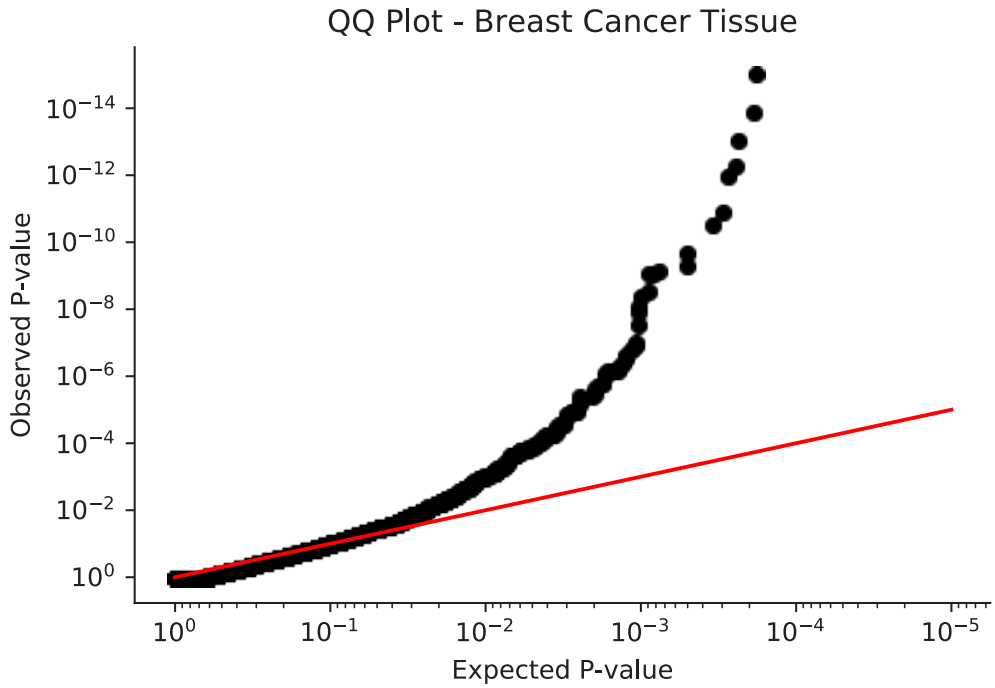

Genes
Adj. P-value < 0.05
SpatialDE Adj. P-value
Mean
Variance
CV2
Dropout rate
FN1
IL12RB1
CXCL13
CXCL9
COL6A3
WDR5B
C2orf40
MMP11
SSPO
RPL41
CST4
HPN
TGM2
IL21R
PEO1
MMP13
ZNF280C

B

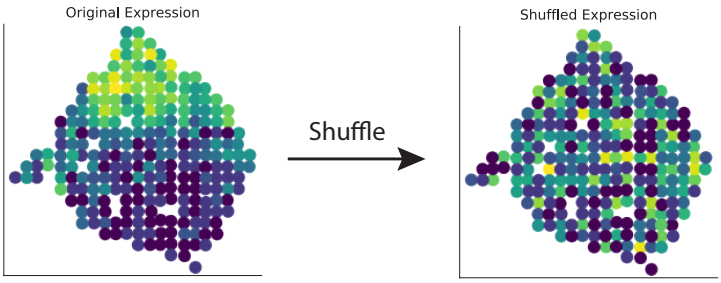

Genes
Adj. P-value < 0.05
CV2
Mean
SpatialDE Adj. P-value
log(Mean) vs log(CV2) Residual
CXCL13
CXCL9
SSPO
MMP13
Variance
SPINT2
log(Mean) vs log(Variance) Residual
CILP
Dropout rate
COL1A1
TFF3
log(mean) vs Dropout rate Residual

Supp. Fig. 4

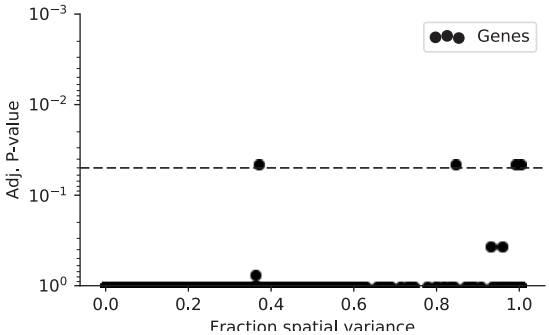
A



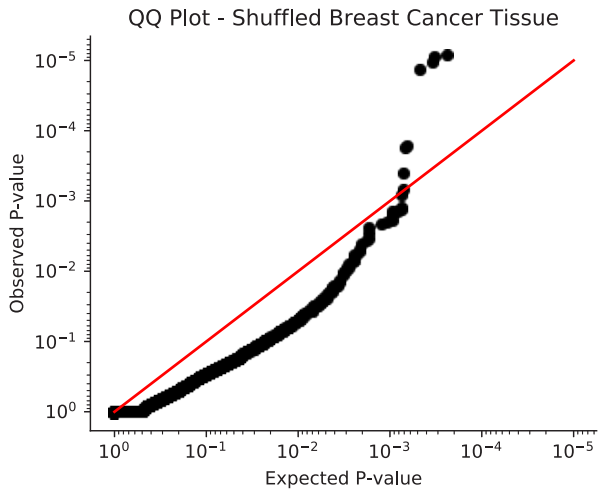
B



C

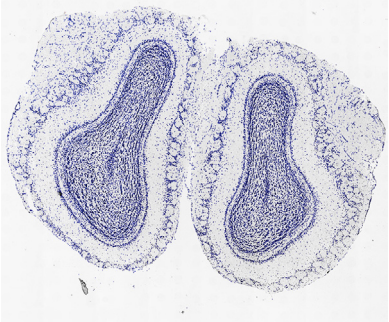


D

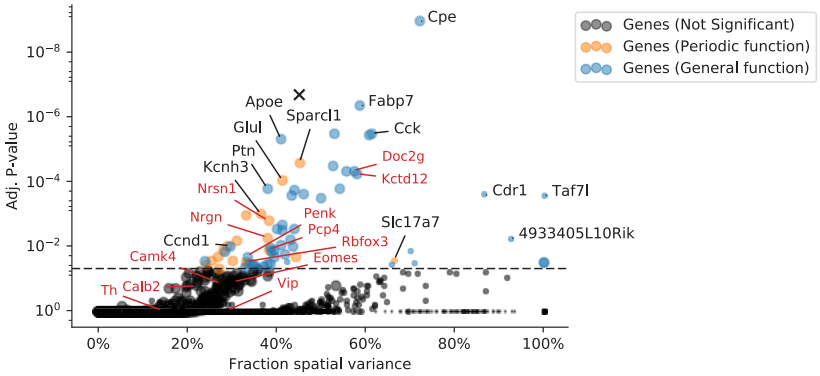


Supp. Fig. 5

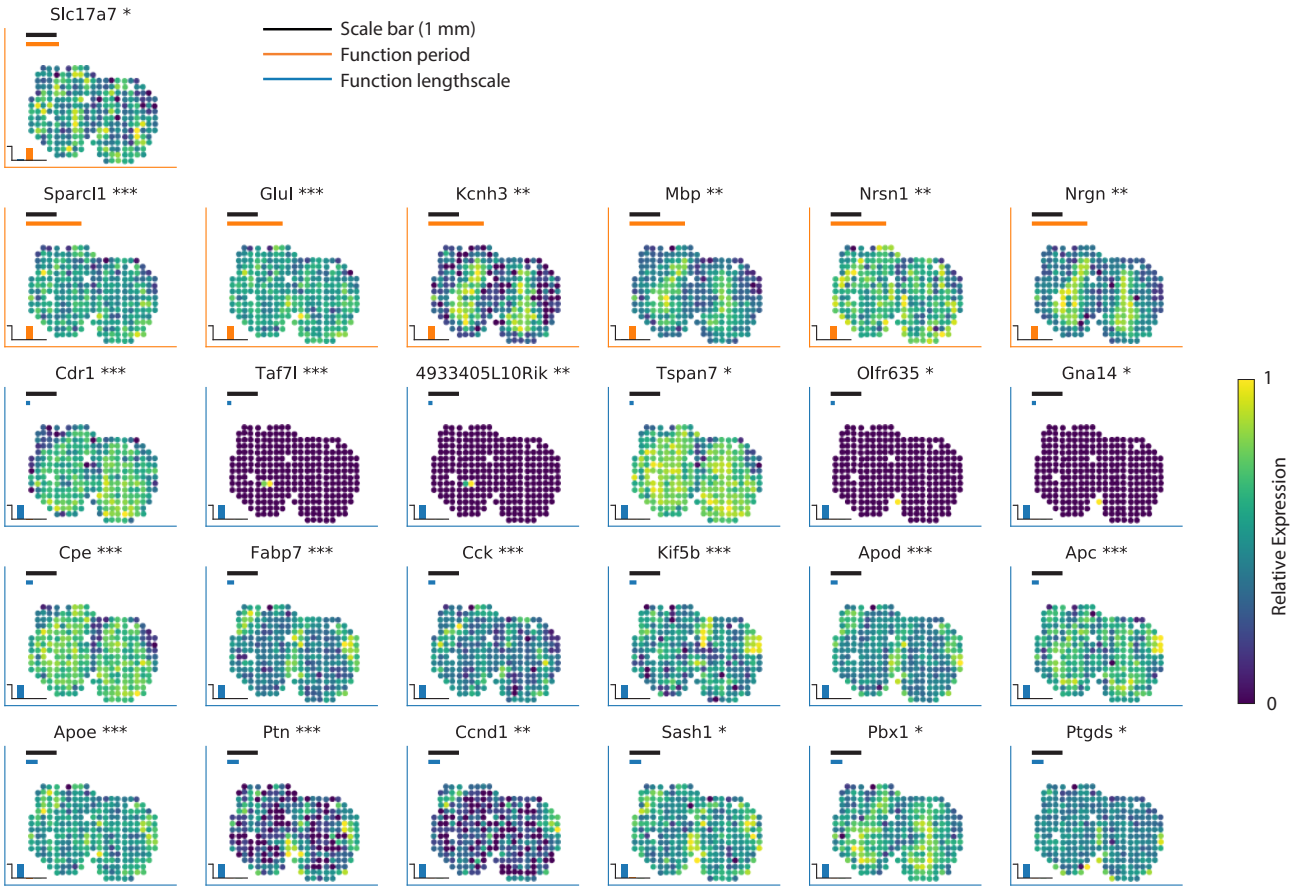
A



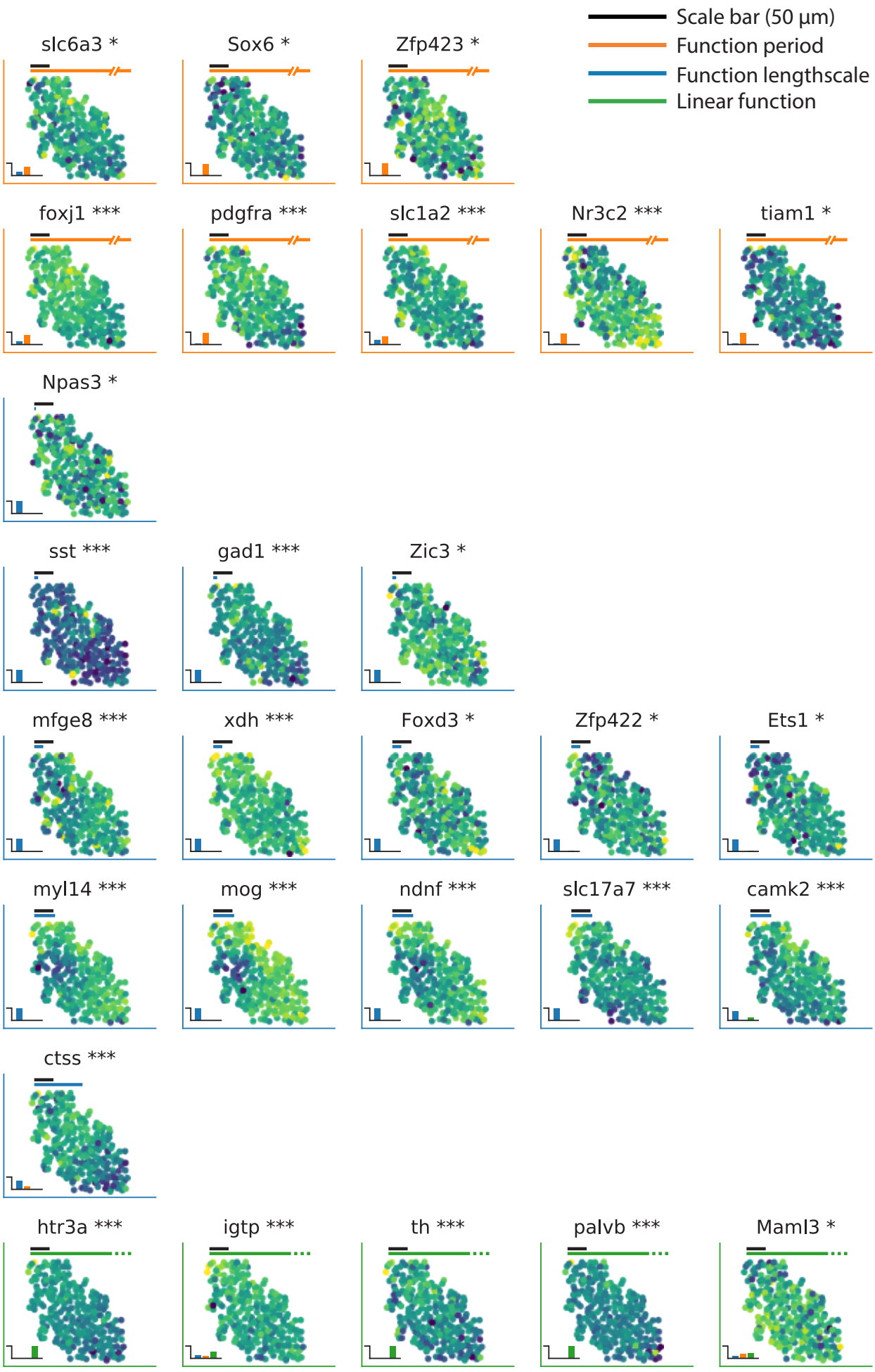
B



C

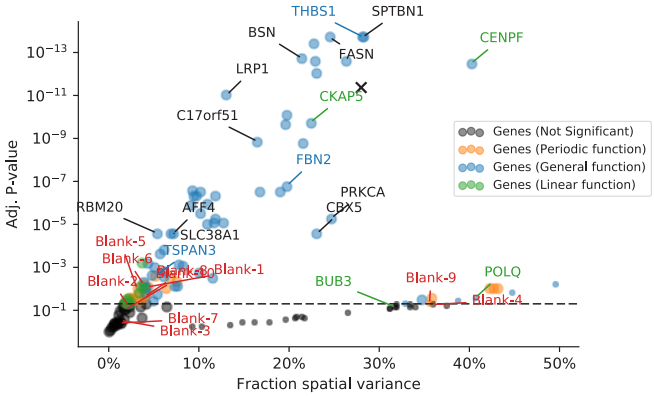


Supp. Fig. 6

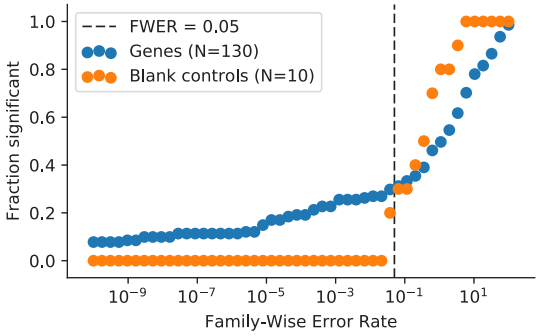

Scale bar (50 µm)
Function period
Function lengthscale
Linear function
slc6a3 *
Sox6 *
Zfp423 *
foxj1 ***
pdgfra ***
slc1a2 ***
Nr3c2 ***
tiam1 *
Npas3 *
sst ***
gad1 ***
Zic3 *
mfge8 ***
xdh ***
Foxd3 *
Zfp422 *
Ets1 *
myl14 ***
mog ***
ndnf ***
slc17a7 ***
camk2 ***
ctss ***
htr3a ***
igtp ***
th ***
palvb ***
Maml3 *

Supp. Fig. 7

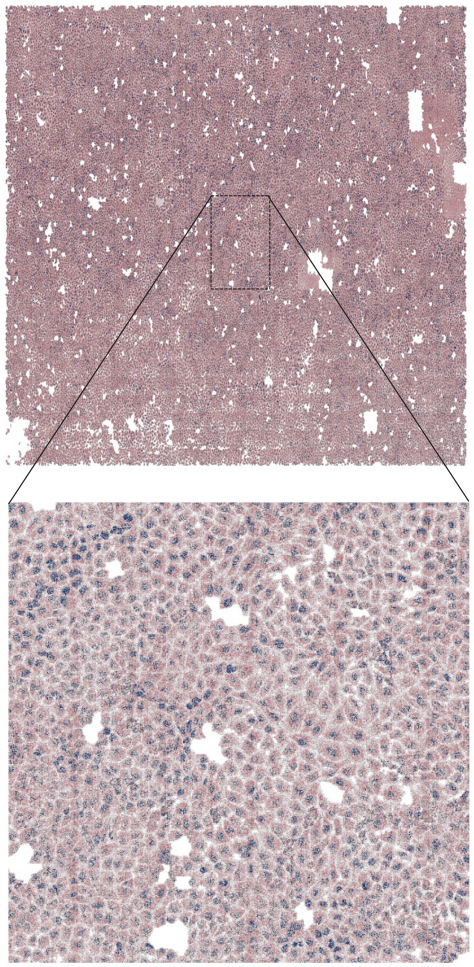
A



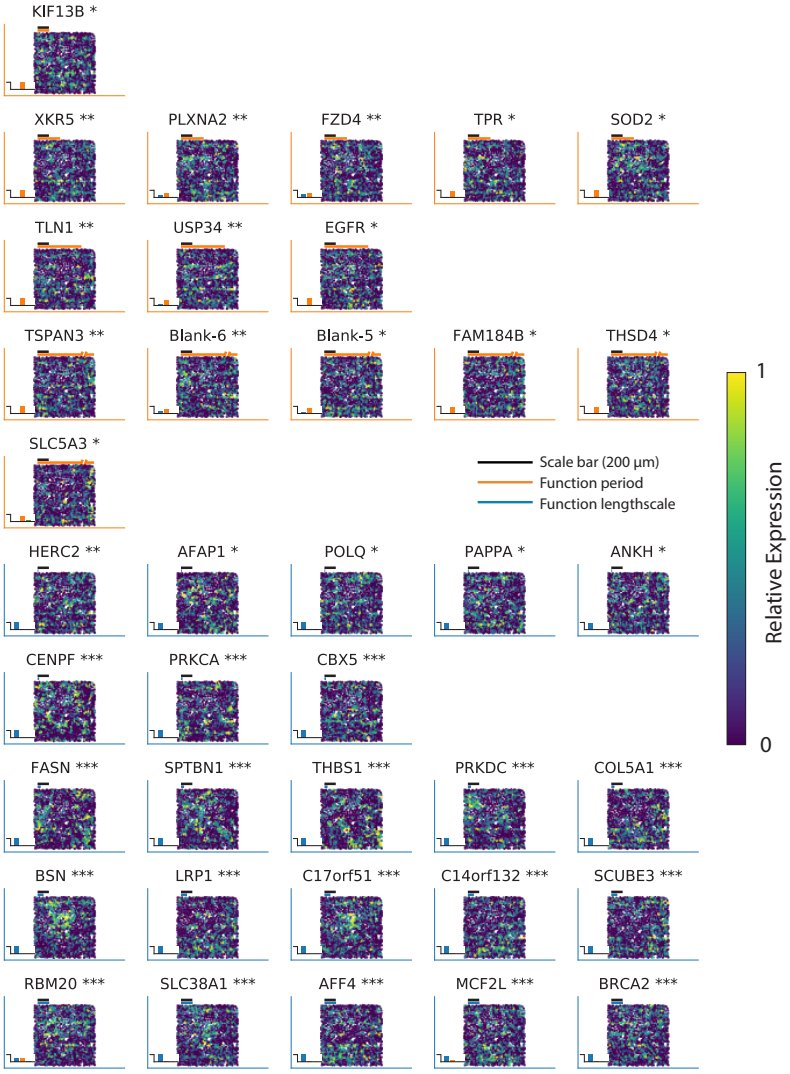
D



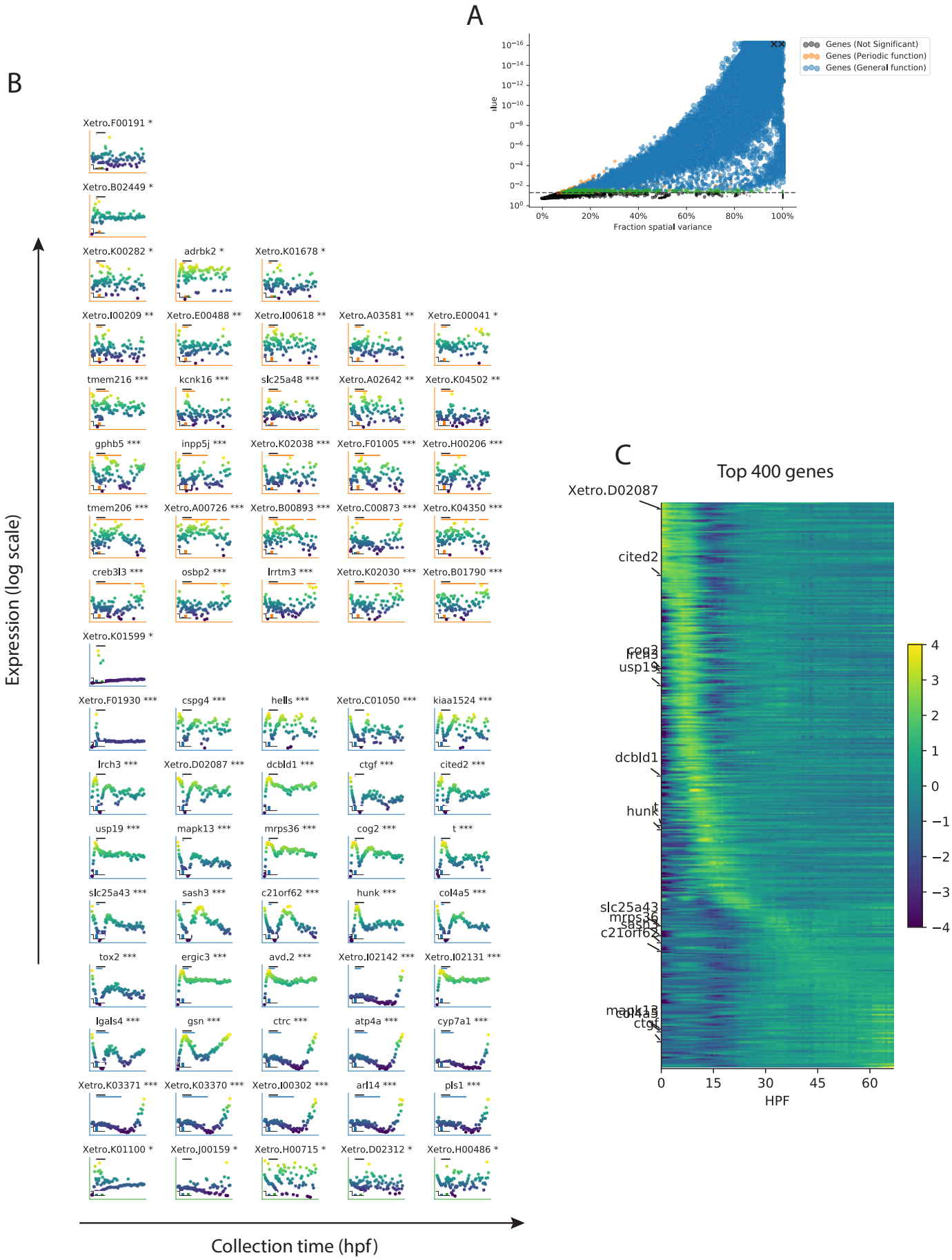
B



C



Supp. Fig. 8


A
Genes (Not Significant)
Genes (Periodic function)
Genes (General function)
Fraction spatial variance
0%
20%
40%
60%
80%
100%
B
Expression (log scale)
Collection time (hpf)
Xetro.F00191 *
Xetro.B02449 *
Xetro.K00282 *
adrbk2 *
Xetro.K01678 *
Xetro.I00209 **
Xetro.E00488 **
Xetro.I00618 **
Xetro.A03581 **
Xetro.E00041 *
tmem216 ***
kcnk16 ***
slc25a48 ***
Xetro.A02642 **
Xetro.K04502 **
gphb5 ***
inpp5j ***
Xetro.K02038 ***
Xetro.F01005 ***
Xetro.H00206 ***
tmem206 ***
Xetro.A00726 ***
Xetro.B00893 ***
Xetro.C00873 ***
Xetro.K04350 ***
creb3l3 ***
osbp2 ***
lrrtm3 ***
Xetro.K02030 ***
Xetro.B01790 ***
Xetro.K01599 *
Xetro.F01930 ***
cspg4 ***
hells ***
Xetro.C01050 ***
kiaa1524 ***
lrch3 ***
Xetro.D02087 ***
dcbld1 ***
ctgf ***
cited2 ***
usp19 ***
mapk13 ***
mrps36 ***
cog2 ***
t ***
slc25a43 ***
sash3 ***
c21orf62 ***
hunk ***
col4a5 ***
tox2 ***
ergic3 ***
avd.2 ***
Xetro.I02142 ***
Xetro.I02131 ***
lgals4 ***
gsn ***
ctrc ***
atp4a ***
cyp7a1 ***
Xetro.K03371 ***
Xetro.K03370 ***
Xetro.I00302 ***
arl14 ***
pls1 ***
Xetro.K01100 *
Xetro.J00159 *
Xetro.H00715 *
Xetro.D02312 *
Xetro.H00486 *
C
Top 400 genes
Xetro.D02087
cited2
cog2
lrch3
usp19
dcbld1
hunk
t
slc25a43
mrps36
sash3
c21orf62
mapk13
col4a5
ctgf
0
15
30
45
60
HPF
4
3
2
1
0
−1
−2
−3
−4

SPATIALDE - METHODS

VALENTINE SVENSSON, SARAH A TEICHMANN, OLIVER STEGLE

1. SpatialDE model

SpatialDE builds on the Gaussian process framework, thereby 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 at N spatial locations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ for a given gene. 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 any 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 $y = f(\mathbf{x})$,

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

A Gaussian process model $\mathcal{H}_{\mathrm{GP}}$ is defined by the covariance function $k(\mathbf{x}, \mathbf{x}' \,|\, \boldsymbol{\theta})$, which parameterizes the dependency between any pair of function values based on their inputs $\mathbf{x}$ and $\mathbf{x}'$; and $\boldsymbol{\theta}$ denotes a vector of additional hyperparameters of the covariance (see below).

Any finite representation of a GP for an observed dataset can be obtained by marginalizing over all unobserved function values, resulting in a finite realisation of joint Gaussian distribution:

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

Here, $\mu\mathbf{1}$ account for mean effects (bias term) and the scaling parameter σ_s^2 determines the proportion of variance explained by the spatial covariance. The term $\sigma_s^2 \delta \mathbf{I}$ explains iid observation noise, i.e. variation in the data does not follow the spatial pattern.

The covariance matrix is derived by evaluating the covariance function for all pairs of observed datums $\boldsymbol{\Sigma}_{k(\mathbf{x},\mathbf{x}' \,|\, \boldsymbol{\theta})_{i,j}} = k(\mathbf{x}_i, \mathbf{x}_j \,|\, \boldsymbol{\theta})$, for which the parameters $\boldsymbol{\theta}$ can be determined using maximum likelihood (see Secion 1.4).

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} LL(\mathcal{H}_{\mathrm{GP}}, \boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \log p(\mathbf{y} \,|\, \mathcal{H}_{\mathrm{GP}}, \boldsymbol{\theta}), \end{aligned} \tag{3}$$

where $LL(\mathcal{H}_{\mathrm{GP}}, \boldsymbol{\theta})$ denotes the log marginal likelihood.

Date: May 28, 2017.

1.2. **Covariance functions.** To test and compare between alternative hypothesis of spatial variation of expression patterns, we asses GP model with different covariance functions.

- Null model
  $k_{\text{null}}(\mathbf{x}, \mathbf{x}') \propto 0$

- General spatial pattern (known as the *RBF* or *Gaussian kernel*)
  $k_{\text{spatial}}(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta}) \propto e^{-\frac{1}{2L^2}|\mathbf{x}-\mathbf{x}'|^2}$

- Linear trend
  $k_{\text{lin}}(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta}) \propto \mathbf{x}\mathbf{x}'^{\mathrm{T}}$

- Periodic pattern (known as the *cosine kernel*)
  $k_{\text{periodic}}(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta}) \propto \cos(\frac{1}{p}|\mathbf{x} - \mathbf{x}'|)$

Interpretation of model parameters. As the scale is parameterized using σ_s^2 in Eq. 2, the proportionality factors does not change the marginal likelihood. However, in order to be able to interpret the parameter σ_s^2 as the proportion of variance explained we use Gower's transformation to correct the σ_s^2 parameter for the structure in the covariance matrix $\boldsymbol{\Sigma}$ [Kostem and Eskin, 2013]:

$$g = \frac{\mathrm{Tr}(P\boldsymbol{\Sigma}P)}{n-1},$$

where

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

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

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

P-values from hypothesis testing. Significant spatial variance component are tested via mode comparison:

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

Here, $\mathcal{H}_1$ denotes the alternative model that includes both a spatial and non-spatial component and $\mathcal{H}_0$ denotes the null model, motting a spatial variance component.

The parameters of both models are optimised using maximum likelihood (see Section 1.4). Significance of the spatial variance component is then assessed using a likelihood ratio (LLR) test between the alternative and the null model. P-values can be estimated in closed form, assuming that the LLR's under the null model are χ^2 distributed with one degree of freedom.

To correct for multiple testing, we use the FDR based strategy by [Storey and Tibshirani, 2003] yielding Q-values. Unless stated otherwise, we report genes at Q-Value < 0.05 as significant spatially variable.

Calibration of the P-values was investigated through negative control probes in the MERFISH experiment. The fraction of significant negative control probes behave as expected with regards to the family-wise error rate (**Supp. Fig. 7D**).

Classification of spatial patterns using model comparison. In order to identify interpretable spatial trends, we can compare the spatial model to alternative models that make stronger assumptions about the spatial dependency. Specifically, for significant spatially variable genes (e.g. Q-value < 0.05), we compare GP models with alternative prior covariances: the general spatial model using an RBF kernel, a GP priors with periodic covariance functions, using the cosine kernel (See Section 1.2), and 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]. The BIC penalises the maximum log-likelihood by the number of effective parameters in the model, 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. 3), 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 specifi models. Briefly, the BIC is an estimate of $-\log p(\mathbf{x}, \mathbf{y}|\mathcal{H}_i)$, which allows for deriving an approximate form of the marginal likelihood of the model $\mathcal{H}_i$,

$$p(\mathcal{H}_i|\mathbf{X}, \mathbf{y}) = \frac{1}{Z} \cdot p(\mathbf{X}, \mathbf{y}|\mathcal{H}_i) \cdot p(\mathcal{H}_i) = \frac{1}{Z} \cdot \int_\theta p(\mathbf{X}, \mathbf{y}|\mathcal{H}_i, \theta) d\theta \approx -\frac{1}{Z} \cdot BIC_i,$$

where

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

We consider the models $\{\mathcal{H}_{\text{spatial}}, \mathcal{H}_{\text{linear}}, \mathcal{H}_{\text{periodic}}\}$ described above (Section 1.2), deriving posterior probabilities of these models given the data.

1.4. **Parameter inference.** Maximum likelihood inference (Eq. 3), requires determining μ, σ_s^2, δ and, depending on the model, additional hyperparameters of the selected covariance function (e.g. the length-scale l, see Section 1.2). The log likelihood is

$$\begin{aligned} LL(\mathbf{y}, \mathbf{X}, \theta) = -\frac{1}{2} \big(& n \log(2\pi) \\ & + \log(|\sigma_s^2 \cdot (\mathbf{\Sigma}_\ell + \delta \cdot \mathbf{I})|) + (\mathbf{y} - \mu)^T (\sigma_s^2 \cdot (\mathbf{\Sigma}_\ell + \delta \cdot \mathbf{I}))^{-1} (\mathbf{y} - \mu) \big) \end{aligned}$$

Evaluation of the likelihood requires inverting the covariance matrix $\mathbf{\Sigma}_\ell$ which depend on the parameter ℓ, this makes gradient based optimisation of ℓ a key bottleneck in inference. We comment on this later, but for now, assumes ℓ is known. To circumvent inverting the entire matrix $\sigma_s^2 \cdot (\mathbf{\Sigma}_\ell + \delta \cdot \mathbf{I})$, we follow [Lippert et al., 2011] and factor the matrix by spectral decomposition:

$$\sigma_s^2 \cdot (\mathbf{\Sigma}_\ell + \delta \cdot \mathbf{I}) = \sigma_s^2 \cdot (USU^T + \delta \cdot \mathbf{I}) = \sigma_s^2 \cdot U(S + \delta \cdot \mathbf{I})U^T$$

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

$$LL(\delta, \sigma_s^2, \mu) = -\frac{1}{2}(n\log(2\pi\sigma_s^2) + \log(|\mathbf{\Sigma}_\ell + \delta\cdot\mathbf{I}|) + \frac{1}{\sigma_s^2}(\mathbf{y}-\mu)^T(\mathbf{\Sigma}_\ell + \delta\cdot\mathbf{I})^{-1}(\mathbf{y}-\mu))$$
$$= -\frac{1}{2}(n\log(2\pi\sigma_s^2) + \log(|U(S+\delta I)U^T|)\frac{1}{\sigma_s^2}(\mathbf{y}-\mu)^T(U(S+\delta\cdot\mathbf{I})U^T)^{-1}(\mathbf{y}-\mu))$$
$$= -\frac{1}{2}(n\log(2\pi\sigma_s^2) + \log(|U||S+\delta\cdot\mathbf{I}||U^T|) + \frac{1}{\sigma_s^2}(\mathbf{y}-\mu)^T U(S+\delta I)^{-1}U^T(\mathbf{y}-\mu))$$
$$= -\frac{1}{2}(n\log(2\pi\sigma_s^2) + \log(|S+\delta\cdot\mathbf{I}|) + \frac{1}{\sigma_s^2}((U^T\mathbf{y}) - (U^T1)\mu)^T(S+\delta\cdot\mathbf{I})^{-1}((U^T\mathbf{y}) - (U^T1)\mu))$$
$$= -\frac{1}{2}(n\log(2\pi\sigma_s^2) + \sum_{i=1}^{n}\log(S_{i,i}+\delta) + \frac{1}{\sigma_s^2}\sum_{i=1}^{n}\frac{([U^T\mathbf{y}]_i - [U^T1]_i\mu)^2}{S_{i,i}+\delta})$$

The key features used is that $|U| = |U^T| = 1$, and $S + \delta\cdot\mathbf{I}$ is diagonal, so both the determinant and inverse are trivial to compute. The expression U^T1 only depends on the coordinates X and can be precomputed for every gene. The expression $U^T\mathbf{y}$ will need to be re-computed for each gene, however, it can be re-used for inference evaluations.

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

$$\begin{aligned}
&\frac{1}{\sigma_s^2}((U^T1)^T(S+\delta\cdot\mathbf{I})^{-1}(U^T\mathbf{y}) - (U^T1)^T(S+\delta\cdot\mathbf{I})^{-1}(U^T1)\hat{\mu}) = 0\\
&\Rightarrow (U^T1)^T(S+\delta\cdot\mathbf{I})^{-1}(U^T1)\hat{\mu}\\
&= (U^T1)^T(S+\delta\cdot\mathbf{I})^{-1}(U^T\mathbf{y})\\
&\Rightarrow \hat{\mu}\\
&= ((U^T1)^T(S+\delta\cdot\mathbf{I})^{-1}(U^T1))^{-1}(U^T1)^T(S+\delta\cdot\mathbf{I})^{-1}(U^T\mathbf{y})\\
&= \left(\sum_{i=1}^{n}\frac{1}{S_{i,i}+\delta}[U^T1]_i^T[U^T\mathbf{y}]_i\right)\Big/\left(\sum_{i=1}^{n}\frac{1}{S_{i,i}+\delta}[U^T1]_i^T[U^T1]_i\right).
\end{aligned}$$

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 gives us

$$\hat{\sigma}_s^2(\delta) = \frac{1}{n}\sum_{i=1}^{n}\frac{([U^T\mathbf{y}]_i - [U^T1]_i\hat{\mu}(\delta))^2}{S_{i,i}+\delta},$$

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

$$\begin{aligned}
LL(\delta) &= -\frac{1}{2}(n\log(2\pi) + S_1(\delta) + n + n\log(\frac{1}{n}S_2(\delta))),\\
S_1(\delta) &= \sum_{i=1}^{n}\log(S_{i,i}+\delta),\\
S_2(\delta) &= \sum_{i=1}^{n}\frac{([U^T\mathbf{y}]_i - [U^T1]_i\hat{\mu})^2}{S_{i,i}+\delta}.
\end{aligned}$$

To optimise $LL(\delta)$ with respect to δ we use gradient based optimisation with l-bfgs-b and numerically approximated gradient. Empirically, we observed that an

analytically calculated gradient would require more floating point operations per iteration step with no gain in performance.

To avoid gradient based optimization of the length scale ℓ, we precalculate a grid of covariance matrices $\mathbf{\Sigma}_\ell$ and factorise them. The number of grid points can be specified by the user, but our default settings put 10 grid points logratihmically spaced between half shortest and twice the longest distance observed in the data. We have found to give sufficient sensitivity. After factoring the $\mathbf{\Sigma}_\ell$'s, the U and S matrices can be reused for each gene. We only need to do as many $O(n^3)$ matrix inversions as we have grid points. Each gene under investigation will have a $O(n^2)$ step for each grid point to calculate the $U^T\mathbf{y}$ 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 Figure 1C of the main text.

Estimation of standard errors. The only optimised parameter in our model is δ, the uncertainty of the maximum likelihood estimate of this parameter is the inverse of $\frac{\partial^2 LL(\delta)}{\partial \delta^2}$ evaluated at $\hat{\delta}$. We use rules of uncertainty propagation to estimate uncertainty of FSV since this can be expressed as a function of δ,

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

where g is the Gower factor for covariance matrix $\mathbf{\Sigma}_\ell$ for a given grid point. 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, we use finite difference approximation on the LL and FSV functions.

2. Data normalisation

The presented Gaussian process model 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 produces counts of transcripts. Spatial Transcriptomics uses Unique Molecular Identifiers (UMI's) to count amplified transcript tags from next generation sequencing reads, while smFISH counts fluorescent probes inside cell boundaries. By investigating the mean-variance relation for all genes in multiple data sets from all spatial technologies we note that the data empirically correspond to negative binomial (NB) noise.

To stabilise the variance, we use the approximate Anscombe's transform for NB data on the observed counts $\hat{\mathbf{y}}_g$, $\mathbf{y}_g = \log(\hat{\mathbf{y}}_g + \frac{1}{\phi})$, where ϕ is the overdispersion parameter, so that $\mathrm{Var}(\mathbf{y}) = \mathbb{E}(\mathbf{y}) + \phi \cdot \mathbb{E}(\mathbf{y})^2$, and ϕ is estimated by curve fitting across all genes in a study [Anscombe, 1948].

Second, we note that in all the data we investigated, every gene's expression correlates with the total 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 Paravan-Medhar et al. [Padovan-Merhar et al., 2015], who showed that cells compensate mRNA content in response to the cytoplasmic volume of a cell. The total count thus correspond to the size of cells.

While there are many cases where cells grow for biologically interesting reasons, cell size assays are easier than gene expression assays, and here we focus on regulation of gene expression. 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, regressing out the log total count from the Anscombe transformed expression values before fitting the spatial models.

For context, we also perform 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 figures and supplementary figures). In the frog development data, proxies for cell size (ERCC expression and number of genes detected) are over 95% spatially variable.

3. Data sets and specific processing steps

The analysis presented in this study is based on a number of publicly available datasets. Some of these data were however not available in typical data repositories owing to their novel nature.

3.1. **Spatial Transcriptomics data.** The count tables from Stahl et al were downloaded from the website `http://www.spatialtranscriptomicsresearch.org/datasets/doi-10-1126science-aaf2403`, linked from the publication. For the breast cancer data, we used the file annotated as "Layer 2" with the corresponding HE image. For the mouse olfactory bulb, we used the file named "Replicate 11" with corresponding HE image. Images included in figures were cropped, down-scaled and converted to grey scale to conserve file sizes.

3.2. **SeqFISH data.** We downloaded the expression table from the supplementary material of Shah et al, and extracted cell counts from the region annotated with number 43 in the 249 gene experiment (Table S8 in the original publication). The shape of the data suggested this corresponded to a region in the lower left part of the corresponding supplementary figure, informing our sketch in Fig. 2D (this was only relevant for illustration, and not used for analysis or results).

3.3. **MERFISH data.** From the website `http://zhuang.harvard.edu/merfish` we downloaded the file "data_for_release.zip" which contain data from Moffitt et al. We used the files in the folder called "Replicate 6", as these had the larges number of cells and highest confluency. Jeffrey Moffitt helped us understand the data format through personal communication.

3.4. **Frog development RNA-seq data.** We downloaded the TPM expression table for Clutch A from GEO accession GSE65785.

4. 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 wall clock. For the Stan implementation, 100 random genes were sampled for each sample size, and timing was extrapolated by multplying the time by 100. It should be noted that this problems is trivially parallelizable over the genes, and neither of the implementations make use of this fact. The 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.

5. 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 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 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.

[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.

[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.

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