Published in final edited form as:
Nat Methods. 2018 May ; 15(5): 343–346. doi:10.1038/nmeth.4636.

SpatialDE: identification of spatially variable genes

Valentine Svensson[1,2], **Sarah A Teichmann**[1,3], and **Oliver Stegle**[2,4]

[1]Wellcome Trust Sanger Institute, Wellcome Genome Campus, CB10 1SD Hinxton, Cambridge, UK

[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD, Hinxton, Cambridge, UK

[3]Theory of Condensed Matter Group, Cavendish Laboratory, 19 JJ Thomson Avenue, CB3 0HE, Cambridge, U.K

[4]European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany

Abstract

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 develop 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 has enabled high-throughput transcriptome profiling from low quantities of starting material, including in single cells. Increased throughput has also fostered new experimental designs that directly assay the spatial context of gene expression variation. Spatially resolved gene expression is crucial for determining the functions and phenotypes of cells in multicellular organisms1. Spatial expression variation can reflect communication between adjacent cells, position-specific states, or cells that migrate to specific tissue locations to perform their functions.

Corresponding authors: Valentine Svensson (vale@ebi.ac.uk), Oliver Stegle (oliver.stegle@ebi.ac.uk).

Code availability

An open source implementation of SpatialDE is available from https://github.com/Teichlab/SpatialDE. The release includes tutorials and example vignettes for reproducing the presented analyses, as well as all pre-processed datasets considered in this study. The software version used to generate the results as presented in this manuscript is also available as Supplementary Software.

Data availability

In addition to data sources mentioned above, all data used for analysis is available from https://github.com/Teichlab/SpatialDE using git-lfs. All results from analysis are reported in Supplementary Table 1.

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.

Competing financial interests

The authors declare no financial conflicts of interest.

Europe PMC Funders Author Manuscripts

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 integration of single cell RNA-seq (scRNA-seq) data with a spatial reference dataset2,3, careful collection and recording of the spatial location of samples4, parallel profiling of mRNA using barcodes on a grid of known spatial locations4–6, and methods based on multiplexed *in situ* hybridization7,8 or sequencing1.

A first critical step in the analysis of these datasets is to identify genes that exhibit spatial variation across the tissue. However, existing approaches for identifying highly variable genes (HVG)9, as used for conventional scRNA-seq data, ignore spatial information and hence do not measure *spatial* variability (Fig. 1A). Alternatively, 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 clustering2,3,6,7, with some methods incorporating spatial information10. Critically, such methods can only detect variations that are captured by differences between discrete groups.

Here, we propose *SpatialDE*, a method for identifying and characterizing *spatially variable* genes (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 non-spatial 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 non-spatial variability. The ratio of the variance explained by these components quantifies the Fraction of Spatial Variance (FSV). Significant SV genes can be identified by comparing this full model to a model without the spatial variance component (Fig. 1B, Methods).

By interpreting the fitted model parameters, we can gain insights into the underlying spatial function, such as its length scale (Fig. 1B, the expected number of changes in a unit interval). SpatialDE can also be used to classify these functions, thereby identifying genes with linear or periodic expression patterns (Supp. Fig. 1, 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 perform *automatic expression histology* (AEH), which relates tissue structure and cell type composition using the expression patterns of marker genes. Leveraging efficient inference methods previously developed for linear mixed models11, and taking advantage of the data structure from massively parallel molecular assays, SpatialDE is computationally very efficient (Methods, Supp. Fig. 2).

First, we applied our method to spatial transcriptomics data from mouse olfactory bulb6. Briefly, spatial transcriptomics gene expression levels were derived from thin tissue sections placed on an array with poly(dT) probes and spatially resolved DNA barcodes. These form a grid of circular “spots” with a diameter of 100 μm, measuring mRNA abundance of 10-100 cells per spot using probes with barcodes that encode spatial locations.

The SpatialDE test identified 67 SV genes (FDR < 0.05, Supp. Table 1), with spatial dependencies explaining up to 70% of the gene expression variance (Fig. 2A). This set of

genes was also markedly disjoint from genes identified when using conventional HVG methods that ignore spatial dependencies (3,497 genes, 40 overlap, Methods). The SV genes identified exhibit clear spatial substructure, consistent with matched hematoxylin and eosin (HE) stained images of the same tissue (Fig. 2B-C). These included canonical marker genes highlighted in the primary analysis by Ståhl *et al6*, such as *Penk*, *Doc2g*, and *Kctd12*, but also 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, including *Kcnh3*, *Nrgn*, or *Mbp* with 1.8 mm period length (Fig. 2C, further examples in Supp. Fig. 3). 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 associated with higher neuron density12, suggesting that periodic expression in tissues is of biological interest.

Applying automatic expression histology identified five canonical expression patterns, clearly demarcating structures visible in the HE image (Fig. 2D, Supp. 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 defining cell types as in AEH (Supp. Fig. 5).

As a second application, we considered tissue slices from breast cancer biopsies6, profiled using the same spatial transcriptomics protocol (Supp. Fig. 6). SpatialDE identified 115 SV genes (FDR < 0.05, compared to 3,503 detected by HVG; overlap 34 genes), including seven genes with known disease relevance that were highlighted in the primary analysis (Supp. Fig. 6B-C). Significantly SV genes were enriched for collagens, which distinguish tissue substructure13 (Reactome “Collagen formation”, $P = 3.38 * 10^{-14}$, gProfiler14, Supp. Table 1). Additionally, we identified the autophagy related gene *TP53INP2,* surrounding the structured tissue (Supp. Fig. 6C). The set of SV genes also included the cytokines *CXCL9* and *CXCL13*, which are expressed in a visually distinct region (Supp. Fig. 6A, black arrow), together with the IL12 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 applying unsupervised clustering in conjunction with an ANOVA test (Supp. Fig. 7). Furthermore, these genes do not have high rank based on non-spatial HVG measures (including mean-CV^2 relation9 or mean-dropout relation15, Supp. Fig. 8).

Automatic expression histology of the SV genes in the breast cancer biopsy (Supp. Fig. 4B) most clearly separated the adipocytic from the denser region of the tissue, but additionally identified a small region overlapping the tumor feature in the HE image. Among the 17 genes assigned to this pattern were the cytokines and receptors *CXCL9*, *CXCL13*, *IL12RB1*, and *IL21R* (Supp. Table 1).

Overall, we found that variable genes detected by SpatialDE are 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 (Supp. Fig. 7E). We confirmed the statistical calibration and the robustness of SpatialDE using randomized data (Supp. Fig. 9) and simulations (Supp. Fig. 10).

SpatialDE is not limited to sequencing technologies, and can be applied to any expression data with spatial and/or temporal annotation. To explore this, we applied SpatialDE to data generated using multiplexed single molecule FISH (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 quantify the expression of thousands of transcripts16 by imaging.

We applied SpatialDE to multiplexed smFISH data of cells from mouse hippocampus, generated using SeqFISH7. 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 HVG, with an overlap of 5 genes) and again SpatialDE identified genes with different types of spatial variation, including linear (N=5) and periodic patterns (N=8, examples in Supp. Fig. 11). The three highest ranking genes: *Mog*, *Myl14*, and *Ndnf* displayed a distinct region of lower expression (Figure 2F-G, black arrows). These genes were grouped into histological expression patterns by the AEH method (Figure 2H, Supp. Fig. 2C). Visual inspection of all 249 genes supports the ranking of spatial variation from SpatialDE (Supp. Fig. 12).

SpatialDE can also be used to test for spatial expression variation in cell culture systems, where spatial variation is not typically expected *a priori*. As an example, we considered data from another multiplexed smFISH dataset generated using MERFISH with 140 probes on a human osteosarcoma cell line8 (Supp. Fig. 13A-B). In the primary analysis, surprisingly *Moffitt et al.8* discovered spatially restricted cell populations with higher proliferation rates. Consistent with these findings, our method identified a substantial proportion of the genes assayed as spatially variable (N=91, 65% of all genes, FDR<0.05, 29 genes HVG with overlap of 24 genes), including six of the seven genes highlighted as differentially expressed between proliferating and resting subpopulations (e.g. *THBS1* and *CENPF1*, Supp Fig. 13C). This indicates that high confluence in cell culture can lead to spatial dependency in gene expression17. Negative control probes in these data were not detected as spatially variable, further confirming the statistical calibration of SpatialDE (Supp Fig. 13D).

Our results demonstrate that SpatialDE identifies spatially variable genes and allows biologically relevant features to be detected in tissue samples without *a priori* histological annotation. The increased availability 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 method to data from multiple 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 (Supp. Fig. 14). Methods already exist for this application18, but are typically computationally more demanding. In principle, SpatialDE

Europe PMC Funders Author Manuscripts

can also be applied to 3-dimensional data, e.g. from aligned serial sections of 2-dimensional data, or from *in situ* sequencing1.

SpatialDE is related to and generalizes previous approaches for detecting temporal19 and periodic gene expression patterns20 in time series. While 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 (Supp. Fig. 15).

Future extensions of SpatialDE could be tailored towards specific platforms, for example to more explicitly model technical sources of variation. Other areas of future work are the incorporation of information about the tissue makeup or local differences in cell density. Finally, there exist spatial clustering methods that are focused on clustering cell positions rather than genes10, which could be combined with the AEH presented here.

Online methods

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper. Full details of the derivation and implementation of SpatialDE are provided in Supp. Note 1.

SpatialDE model

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

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

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

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

whereby the covariance between pairs of cells i and j is modelled 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 distance21.

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

$$LL = -\frac{1}{2}\cdot N\cdot\log(2\cdot\pi) - \frac{1}{2}\cdot\log(|\sigma_s^2\cdot[\Sigma+\delta\cdot I]|) - \frac{1}{2}\cdot(y-\mu\cdot 1)^T(\sigma_s^2\cdot[\Sigma+\delta\cdot I])^{-1}(y-\mu\cdot 1)\,. \quad (3)$$

This optimization problem with closed form solutions for the parameters μ and σ_s, for given parameters values δ. Gradient-based optimization is used to determine δ, and the hyperparameter I is determined via grid search. Naïve methods for evaluating the marginal likelihood in Eq. (1) scale cubically in the number of cells, thus prohibiting applications to larger datasets. We adapt algebraic reformulations that have been proposed in statistical genetics11,22, coupled with efficient pre-computations of all terms possible, to improve scalability of the model (Supp. Fig. 2).

Statistical significance

To estimate statistical significance, the model likelihood of the fitted SpatialDE model is compared to the likelihood of a model that corresponds to the null hypothesis of no spatial covariance,

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

P-values are then estimated analytically based on the χ^2 distribution transformation with one degree of freedom. Unless stated otherwise, we use the Q-value method23 to adjust for multiple testing, thereby controlling the false discovery rate (FDR).

Model selection

Following significance testing, the spatial covariance patterns identified can be further investigated by comparisons of models with alternative covariance functions. In addition to the squared exponential covariance (Eq. (2)), SpatialDE implements covariance functions that assume linear trends as well as periodic patterns of gene expression variation (Supp. Fig. 1), which are compared using the Bayesian information criterion:

$$BIC = \log(N)\cdot M - 2\cdot LL\,.$$

Here M denotes the number of hyperparameters of a given model, N the number of samples, and LL (Eq. (3)) is the log marginal likelihood of the data. For guidance on how to interpret these inferences and alternative functional forms, see Supp. Note 1.

Automatic expression histology

To group spatially variable genes with similar spatial expression patterns, SpatialDE implements a clustering model based on the same spatial GP prior as used to test for

spatially variable genes (Eq. (1)). Let $Y = (y_1, \ldots, y_G)$ be the expression matrix of G spatially variable genes in each spatial location (now each y_g is a vector of N observations), $\mu = \{\mu_1, \ldots, \mu_K\}$ is the matrix of K underlying patterns, so the vector μ_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 σ_e^2 is the noise level for the model, and Σ is the spatial covariance matrix defined based on spatial coordinates (see Eq. (2)). This model can be regarded as an extension of the classical Gaussian mixture model24, with the addition of a spatial prior on cluster centroids. Approximate posterior distributions for μ and z are estimated using variational inference24, while the noise level σ_e^2 is estimated by maximising the variational lower bound. The length scale l for the covariance Σ 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. See Supp. Note 1 for 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 visualise any histological pattern through plotting $\overline{\mu_k}$ over the x coordinates. The most likely assignment of genes to an individual pattern is determined by the largest value in the vector $\overline{z_g}$, which corresponds to the posterior probabilities of a gene belonging to each pattern.

Highly variable gene selection

For each dataset, highly variable genes were identified using the ScanPy implementation25 of the Seurat method of highly variable gene filtering3 using default parameters.

Relationship to prior work

SpatialDE is related to existing Gaussian-processes based gene expression models. First used in geostatistics26, GP models have been applied to test for differential gene expression over time27, including the analysis of bifurcation events28, and to define general tests for temporal variability28–32.

We have here adapted GP models to spatial transcriptome data, although the model can also be applied to univariate data (Supp. Fig. 14) or higher-dimensional inputs. The main technical innovations presented here are three-fold. First, the model presented is faster than

existing methods by leveraging computational tricks previously proposed in the context of statistical genetics (Supp. Fig. 2, Section above). Second, we combine spatial GPs with model selection using BIC33. Third, we propose an efficient and versatile spatial clustering within the same statistical framework.

Data sets and processing

Spatial Transcriptomics data—The count tables from Stahl et al6 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 grayscale to conserve file sizes. When performing automatic expression histology, the number of patterns was set to 5 for both data sets, the characteristic length scale was set to 105 μm for the breast cancer data, and to 150 μm for the olfactory bulb data.

SeqFISH data—We downloaded the expression table from the supplementary material of Shah *et al17* 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 the schematic shown in Fig. 2F (only used for the purpose of illustration). In the automatic histology analysis, the number of patterns was set to 5, and the characteristic length scale was set to 50 μm.

MERFISH data—From the website http://zhuang.harvard.edu/merfish we downloaded the file "data for release.zip" which contain data from Moffitt et al8 We used the files in the folder called "Replicate 6", 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 which was referenced in the original publication18.

Expression count normalisation

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 have identified two normalisation steps (Supp. 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 / spatial location. To ensure that SpatialDE captures the spatial covariance for each gene beyond this effect, log total count values are regressed out from the Anscombe-transformed expression values before fitting the spatial models.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors wish to thank Damien Arnol and Francesco Paolo Casale for helpful advice on statistics and data normalization. Jeffrey Moffitt helped us understand the data format for available MERFISH data. In addition, we are thankful to Aarun Lun, 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, the Wellcome Trust and the EU.

References

1. Lee JH. Quantitative approaches for investigating the spatial context of gene expression. Wiley Interdiscip Rev Syst Biol Med. 2017; 9
2. Achim K, et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat Biotechnol. 2015; 33:503–509. [PubMed: 25867922]
3. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015; 33:495–502. [PubMed: 25867923]
4. Junker JP, et al. Genome-wide RNA Tomography in the zebrafish embryo. Cell. 2014; 159:662–675. [PubMed: 25417113]
5. Chen J, et al. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. Nat Protoc. 2017; 12:566–580. [PubMed: 28207000]
6. Ståhl PL, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016; 353:78–82. [PubMed: 27365449]
7. 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. 2016; 92:342–357. [PubMed: 27764670]
8. Moffitt JR, et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proc Natl Acad Sci U S A. 2016; 113:11046–11051. [PubMed: 27625426]
9. Brennecke P, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013; 10:1093–1095. [PubMed: 24056876]
10. Pettit J-B, et al. Identifying cell types from spatially referenced single-cell expression datasets. PLoS Comput Biol. 2014; 10:e1003824. [PubMed: 25254363]
11. Lippert C, et al. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011; 8:833–835. [PubMed: 21892150]
12. Jahn R, Takamori S, Rhee JS, Rosenmund C. Nature. 2000; 407:189–194. [PubMed: 11001057]
13. Seewaldt VL. Cancer: Destiny from density. Nature. 2012; 490:490–491. [PubMed: 23099400]
14. Reimand J, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016; 44:W83–9. [PubMed: 27098042]
15. Andrews TS, Hemberg M. Modelling dropouts allows for unbiased identification of marker genes in scRNASeq experiments. bioRxiv. 2016
16. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015; 348
17. Battich N, Stoeger T, Pelkmans L. Control of Transcript Variability in Single Mammalian Cells. Cell. 2015; 163:1596–1610. [PubMed: 26687353]
18. Owens NDL, et al. Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. Cell Rep. 2016; 14:632–647. [PubMed: 26774488]
19. Kalaitzis AA, Lawrence ND. A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. BMC Bioinformatics. 2011; 12:180. [PubMed: 21599902]

20. Durrande N, Hensman J, Rattray M, Lawrence ND. Detecting periodicities with Gaussian processes. 2016; doi: 10.7717/peerj-cs.50
21. Rasmussen, CE, Williams, CKI. Gaussian Processes for Machine Learning. MIT Press; 2006.
22. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44:821–824. [PubMed: 22706312]
23. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003; 100:9440–9445. [PubMed: 12883005]
24. Bishop, CM. Pattern recognition and machine learning. springer; 2006.
25. Alexander Wolf F, Angerer P, Theis FJ. Scanpy for analysis of large-scale single-cell gene expression data. bioRxiv. 2017; doi: 10.1101/174029
26. Krige DG. A statistical approach to some basic mine valuation problems on the Witwatersrand. J South Afr Inst Min Metall. 1951; 52:119–139.
27. Stegle O, et al. A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. J Comput Biol. 2010; 17:355–367. [PubMed: 20377450]
28. Lönnberg T, et al. Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria. Science Immunology. 2017; 2
29. Kalaitzis AA, Lawrence ND. A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. BMC Bioinformatics. 2011; 12:180. [PubMed: 21599902]
30. Äijö T, et al. Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. Bioinformatics. 2014; 30:i113–20. [PubMed: 24931974]
31. Macaulay IC, et al. Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. Cell Rep. 2016; 14:966–977. [PubMed: 26804912]
32. Eckersley-Maslin MA, et al. MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. Cell Rep. 2016; 17:179–192. [PubMed: 27681430]
33. Lloyd JR, Duvenaud D, Grosse R, Tenenbaum JB, Ghahramani Z. Automatic Construction and Natural-Language Description of Nonparametric Regression Models. arXiv [stat.ML]. 2014

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Editor's Summary

SpatialDE identifies genes with significant spatial expression patterns from multiplexed imaging or spatial RNA sequencing data, and can cluster genes with similar spatial patterns as a form of expression-based tissue histology.

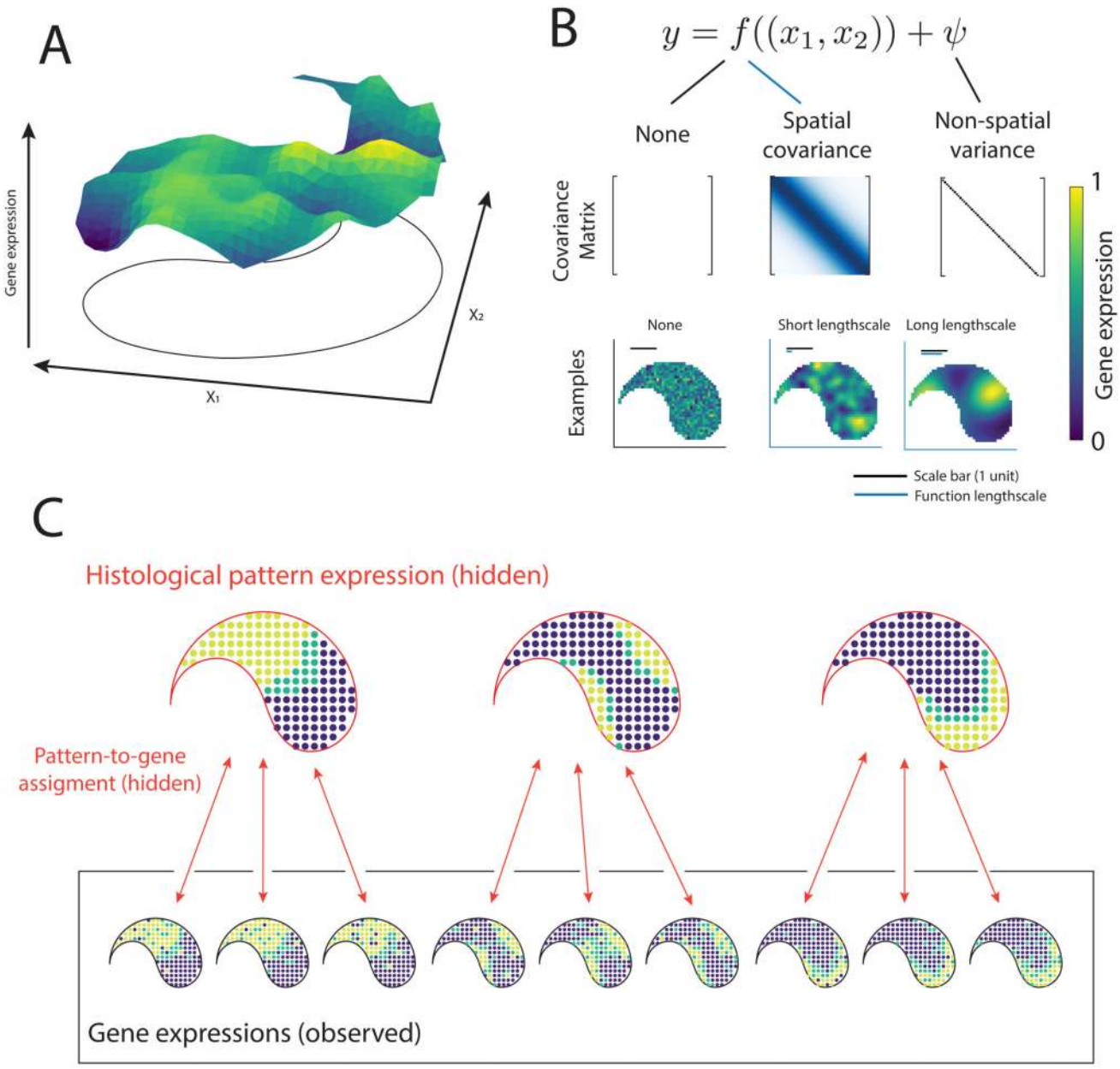

Figure 1. Overview of SpatialDE for the identification of spatially variable genes.

(**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 using a non-parametric regression model, testing whether gene expression levels at different locations co-vary 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 (ψ). Representative simulated expression patterns are plotted below the corresponding covariance matrices for the null model (None) and the alternative model (Spatial covariance) with different lengthscales. (**C**) Automatic expression histology uses

Europe PMC Funders Author Manuscripts

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.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

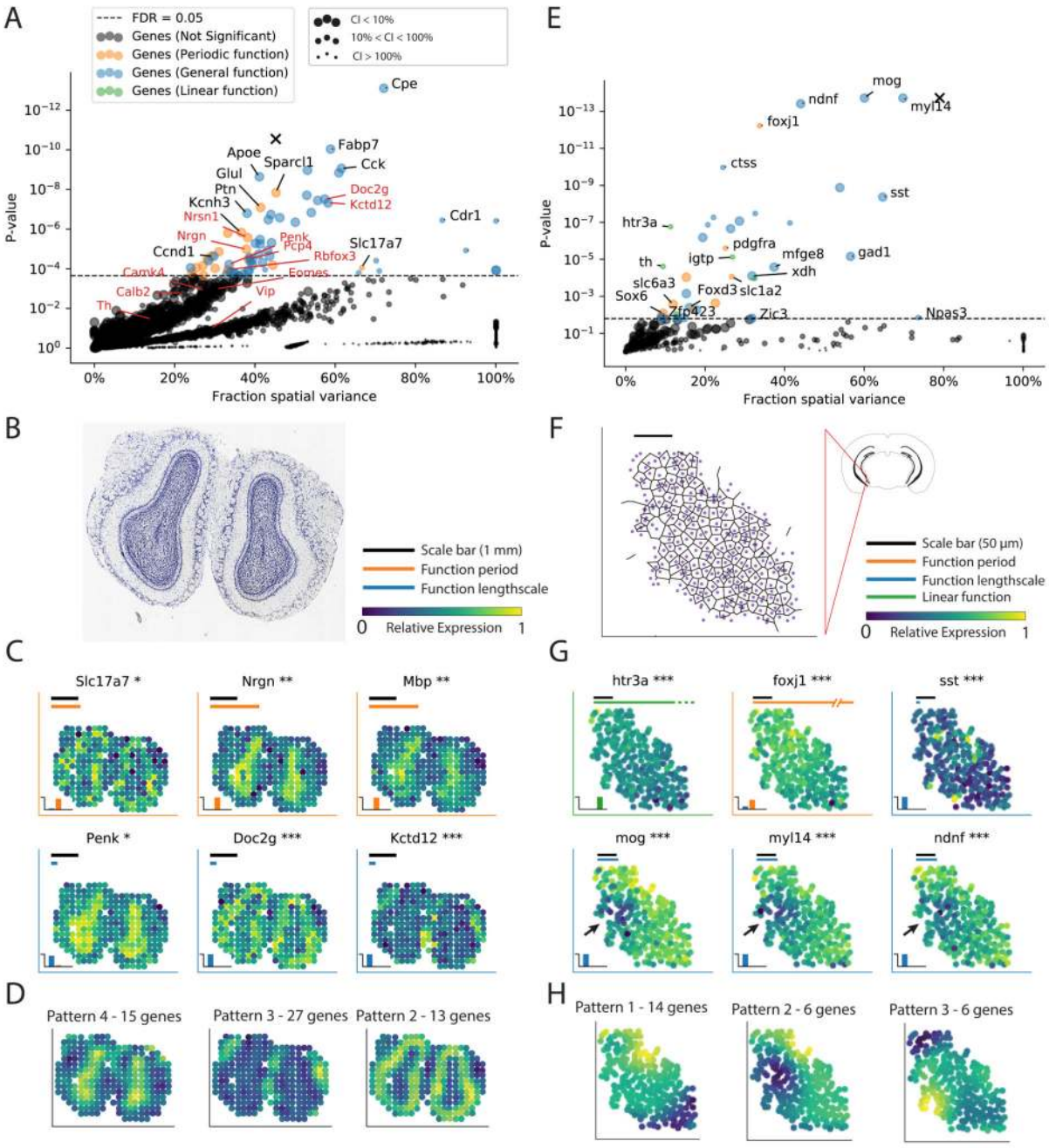

Figure 2. Application of SpatialDE to spatial transcriptomics and SeqFISH data.

(**A**) Fraction of variance explained by spatial variation (FSV) versus significance of spatial variation (SpatialDE negative log P-value) for all genes in the mouse olfactory bulb data. Dashed line corresponds to FDR=0.05 significance level (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 in Stahl et al are in red text. Point size indicates uncertainty of FSV estimates; CI, confidence intervals. The X symbol shows the result of applying SpatialDE to the estimated total RNA content per spot.

(**B**) Hematoxylin and eosin image for mouse olfactory bulb data from Stahl *et al.* (**C**) Visualization of selected SV genes. Orange bar shows fitted period length for genes with periodic dependencies; blue bar shows fitted length scale for genes with general spatial trends. 2D plots depict expression level for genes across the tissue section coded in color. Asterisks denote statistical significance of spatial variation (* FDR < 0.05, ** FDR < 0.01, *** FDR < 0.001). Insets in lower left show the posterior probability for gene assignments as general spatial, periodic spatial, or linear trend. (**D**) Example histological expression patterns identified by automatic expression histology analysis, with expression levels encoded in color. The number of genes assigned to each pattern are noted. (**E**) Proportion of spatial variance versus significance of spatial variation (SpatialDE negative log P-value) for all 249 genes in the SeqFISH data from a region of mouse hippocampus from Shah *et al7*, as in **A**, showing genes with linear dependency in green. (**F**) Voronoi tessellation representative of tissue structure. (**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 region of low expression of *Mog*, *Myl14* and *Ndnf*. (**H**) Three examples of histological expression patterns identified by AEH.

Europe PMC Funders Author Manuscripts