


Identification of spatial expression trends in single-cell gene expression data

Daniel Edsgård^{1,2}, Per Johnsson^{1,2}
& Rickard Sandberg^{1,2} 

As methods for measuring spatial gene expression at single-cell resolution become available, there is a need for computational analysis strategies. We present trendsceek, a method based on marked point processes that identifies genes with statistically significant spatial expression trends. trendsceek finds these genes in spatial transcriptomic and sequential fluorescence *in situ* hybridization data, and also reveals significant gene expression gradients and hot spots in low-dimensional projections of dissociated single-cell RNA-seq data.

Analyses of gene expression at single-cell resolution and in spatial contexts can reveal insights into the molecular organization of tissues and organs. Although it is not yet feasible to measure single-cell gene expression with spatial information at the scale of the full transcriptome, recent progress in sequential barcoded fluorescence *in situ* hybridization (seqFISH)^{1,2} has allowed the analysis of hundreds of genes in individual tissue sections³. Lysis and reverse transcription of tissue sections on barcoded substrates is another approach for high-throughput analyses that can provide spatial information⁴. These and other approaches to measure spatial gene expression are being actively developed, yet analysis methods that take full advantage of the resulting data are lacking. Currently, cellular and regional expression profiles are typically first analyzed without spatial information, and are only later projected back onto the spatial structure for visual inspection of spatial trends^{3,4}.

Here we introduce trendsceek, a computational method that identifies statistically significant spatial gene expression trends (<https://github.com/edsgard/trendsceek> and **Supplementary Software**). To identify genes for which dependencies exist between the spatial distribution of cells and gene expression in those cells, we modeled data as marked point processes, which we used to rank and assess the significance of the spatial expression trends of each gene. We first validated the method on simulated data and then demonstrated its ability to identify genes with significant

spatial patterns in spatial transcriptomics data from mouse olfactory bulb and breast cancer sections⁴ and in seqFISH data from hippocampus³. Moreover, we show that trendsceek revealed significant gradients and patterns in dissociated single-cell RNA-seq (scRNA-seq) data⁵ projected onto a low-dimensional space (e.g., via *t*-distributed stochastic neighbor embedding (t-SNE))⁶. We found spatial patterns within low-dimensional t-SNE clusters, thus demonstrating the general utility of simultaneously incorporating expression and location information to find significant spatial trends. We have made trendsceek available as an R package to allow its application to a broad variety of spatial gene-expression data types.

Marked point processes constitute a statistical framework that has previously been applied in the fields of geostatistics, astronomy and material physics⁷. For spatial analyses of gene expression, we assigned points to represent the spatial locations of cells (or regions) and marks on each point to represent expression levels. This approach is nonparametric and can identify general non-linear expression patterns without the need to specify a distribution or spatial region of interest. Our method tests for significant dependency between the spatial distributions of points and their associated marks (expression levels) through pairwise analyses of points as a function of the distance *r* (radius) between them. Summary statistics used for dependency assessment include conditional mean (E-mark), conditional variance (V-mark), Stoyan's mark correlation and the mark-variogram (Online Methods). Notably, if marks and the locations of points are independent, the scores obtained should be constant across the different distances *r*. To assess the significance of a gene's spatial expression pattern, we implemented a resampling procedure in which the expression values are permuted, reflecting a null model with no spatial dependency of expression. Once a gene with a significant spatial trend has been identified, it is also useful to determine the subset of cells that occupy spatial regions of interest. To identify cells located in regions with higher expression than expected by chance, we implemented a method based on weighted kernel density estimation (wKDE) and compared it to a null model derived from permuted expression values.

We first used trendsceek on simulated spatial expression data sampled from empirical seqFISH data (Online Methods). Cells with higher expression were located in local hot spots, step gradients or nonradial streaks, or expression followed a linear gradient (**Fig. 1a**). In the trendsceek analysis of the hot-spot pattern, mark-correlation and mark-variogram were significant at low *r* values, as determined via 1,000 randomly permuted expression distributions of the same marks (**Fig. 1b**; *P* < 0.05, permutation test). We conducted a power analysis of the four metrics as a function of the number of cells, expression level, expression level

¹Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. ²Ludwig Institute for Cancer Research, Stockholm, Sweden. Correspondence should be addressed to R.S. (Rickard.Sandberg@ki.se).

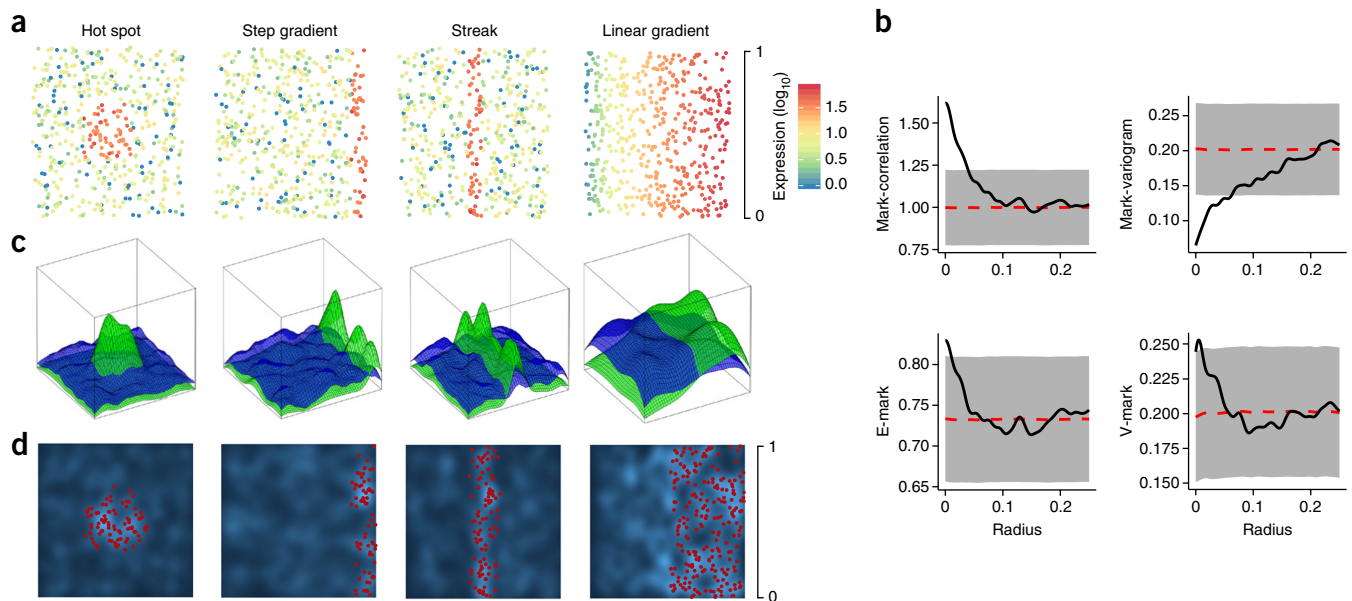


Figure 1 | trendsceek detects spatial gene expression trends in simulated data. **(a)** Simulated mark distributions showing local hot spot, step gradient, non-radial streak and linear gradient patterns. Expression values were sampled from empirical seqFISH data³, and cells in certain regions were spiked by sampling from the upper quantile of the expression distribution ($n = 500$ cells and $n \sim 50$ spiked cells per simulation; mean expression of spiked cells/mean expression background = ~ 10). **(b)** Marked-point pattern statistics (Online Methods) for the simulated hot spot shown in **a**. The red dashed line indicates the sample mean of a null distribution based on resampling of the mark distribution, and the gray band indicates the 2.5% and 97.5% quantiles of the null distribution (two-sided 5% significance level). The mark-correlation and mark-variogram metrics clearly indicate the existence of a significant spatial pattern, represented by values outside of the gray zone at certain radii. **(c)** 3D representations of the spatial expression trend. Green surfaces indicate the wKDE of the simulated data sets; blue surfaces indicate the upper 5% quantile of a null distribution generated by wKDE of the resampled mark distribution for each data set. **(d)** trendsceek identifies cells in regions of spatially significant gene expression. Shown are density plots of the simulated data sets in **a**; red cells exceed a 5% significance level based on wKDE, indicated by the blue surfaces in **c**. Scale bars denoting an arbitrary distance in **a** and **d** apply to all figure panels, including the radius shown in **b**.

difference and the size of the region with elevated expression, for all four simulated spatial patterns (**Supplementary Figs. 1 and 2**). The analysis showed that spatial structures were reliably identified if at least 5% of sampled cells had differing expression levels, in particular when the total number of analyzed cells exceeded 500. For these patterns, the mark-variogram- and mark-correlation-based tests had the highest detection power, but E-mark and V-mark can have higher power in other cases (see results from real data below). Our simulations demonstrated that trendsceek has sufficient power to reveal a variety of spatial patterns involving a small number of cells.

Next, we developed an approach to identify cells in regions with elevated expression levels. Using wKDE, we identified cells with differences in expression exceeding a 5% significance level, on the basis of comparison with the upper 5% quantile of a two-dimensional null distribution generated by resampling of the mark distribution (**Fig. 1c**). The pattern of cells identified as significant recapitulated the pattern of spiked cells for each spatial pattern (**Fig. 1d**). Having developed a method to identify genes with spatial expression trends and an approach to pinpoint the cells belonging to regions of interest within the pattern, we decided to apply it to recently published gene expression data.

We first used trendsceek to analyze spatial transcriptomics data from mouse olfactory bulb tissue⁴, and identified 35 significant genes (**Fig. 2a,b**, **Supplementary Figs. 3a, 4a, 5a and 6a,b** and **Supplementary Tables 1–3**; $P < 0.05$, Benjamini–Hochberg adjusted) with expression primarily in nongranular cells. These

genes included *Ptn*, *Nr2f2* and *Fabp7*, which are known to be restricted to specific tissue domains on the basis of principal component analysis (PCA)⁴ and have clear spatial RNA *in situ* signatures in the Allen Brain Institute atlas (see **Supplementary Fig. 9** in ref. 4). We also identified 45 genes with significant expression primarily in granular regions of the bulb (**Fig. 2c,d**, **Supplementary Figs. 3b, 4b, 5b and 6c,d** and **Supplementary Tables 1, 4 and 5**), including known genes such as *Nrgn*, *Camk4* and *Pcp4*, as well as novel genes such as *Gpsm1* (**Fig. 2c**). A strength of spatial transcriptomics is its ability to profile tumor tissues. When we applied trendsceek to spatial profiling of human breast cancer tissue (layer 2)⁴, we identified 14 genes with significant spatial expression (**Fig. 2e**, **Supplementary Figs. 3c, 4c, 7 and 8** and **Supplementary Tables 1, 6 and 7**). Several genes implicated in breast cancer had significant spatial patterns, including the transcription factor *KLF6* (ref. 8), the transmembrane protein *PMEPA1* (also known as *TMEPAI*; ref. 9), and 12 genes related to the extracellular matrix. We conclude that trendsceek can be broadly applied to spatial transcriptomics data to find genes with significant spatial trends.

The vast majority of scRNA-seq has been done on dissociated cells that lacked spatial information. Analysis of single-cell gene expression data often includes clustering and visualization of cells in low-dimensional spaces (using, for example, PCA or t-SNE). We reasoned that biologically meaningful gene expression patterns might still be present in the resulting clusters, and we therefore investigated whether trendsceek could find spatial patterns in

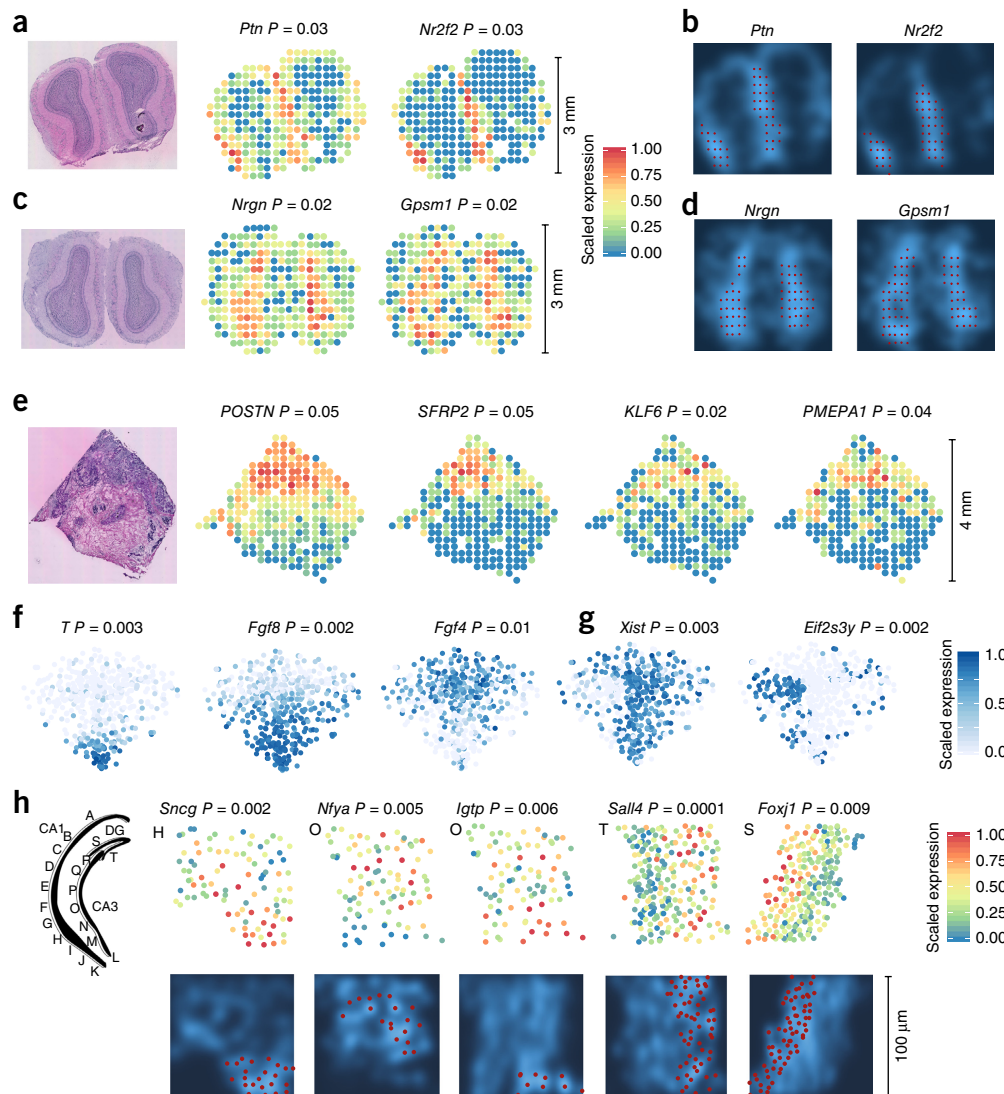


Figure 2 | Application of trendsceek to spatial and single-cell gene expression data. (a) Spatial transcriptomics data from mouse olfactory bulb (tissue section “replicate 3”; $n = 269$ array spots). Left, hematoxylin-and-eosin-stained tissue section (from ref. 4). Right, examples of genes with significant expression trends in the tissue sample. Expression was scaled to the range of 0–1 by unity-based normalization. (b) Density plots of gene expression in the sample shown in a, with cells in regions of significantly elevated expression colored red. (c,d) Spatial transcriptomics data (c) and density plots (d) from mouse olfactory bulb (tissue section “replicate 12”; $n = 280$ array spots) from ref. 4, as in a and b, respectively. (e) Spatial transcriptomics data from breast cancer biopsy (histological section “Layer 2,” $n = 251$ array spots), with examples of genes with significant expression trends within the tissue sample. (a,c,e) The distance between array spots is 200 μm , and each spot covers multiple cells. (f) Examples of distinct spatial gene expression patterns identified by trendsceek in E6.5 mouse epiblast cells (cluster 3 in ref. 5; $n = 481$ cells). (g) Identification of spatial patterns related to the positions of male and female cells within the cluster shown in f, with mutually exclusive expression of *Xist* (expressed in female cells) and *Eif2s3y* (located on the Y-chromosome and expressed only in male cells). (h) Examples of spatial expression patterns identified in mouse hippocampus seqFISH data (cells imaged in each highlighted region: H, 93; O, 89; T, 208; S, 214). Left, an illustration of hippocampus with labels indicating the 21 regions imaged in a previous study³. CA, cornu ammonis; DG, dentate gyrus. P values represent mark-correlation (a–e), mark-variogram (f,g) and E-mark (h) (two-sided, Benjamini–Hochberg adjusted). Panels a,c,e, reproduced with permission from ref. 4. Panel h reproduced with permission from ref. 3.

two-dimensional representations of dissociated single-cell data. t-SNE analysis of scRNA-seq data from a mouse gastrulation data set identified a large cluster of 481 epiblast cells from mice at age embryonic day 6.5 (E6.5)⁵. Within this cluster, trendsceek identified 107 genes with significant spatial expression patterns ($P < 0.05$, Benjamini–Hochberg adjusted; **Supplementary Fig. 9**). The vast majority of significant genes were characterized by an expression gradient with higher expression at the narrow part of the cell cluster (**Supplementary Fig. 10**; exemplified by *T* and

Fgf8 in **Fig. 2f**). Moreover, we identified hot spots of male cells distinct from female cells with significant expression of *Eif2s3y* and *Xist*, respectively (**Fig. 2g** and **Supplementary Tables 1, 8** and **9**; $P < 0.05$, Benjamini–Hochberg adjusted). These two types of expression patterns identified in these cells suggest that there can be functional meaning to spatial patterns identified in this way. Different sets of genes were identified by the mark-correlation- and mark-variogram-based tests, which indicates that the inclusion of multiple summary statistics improves sensitivity

(Supplementary Fig. 4d,h). All of these diverse spatial patterns within low-dimensional projections of dissociated single-cell data were identified by trendsceek in a fully unbiased manner, without incorporation of a priori knowledge about candidate genes.

Finally, we applied trendsceek to seqFISH data from 21 mouse hippocampus regions³ that in total included approximately 2,000 cells and 249 genes. We identified genes with significant spatial patterns in 15 out of 21 regions, with a median of 54 genes per region (Fig. 2h, Supplementary Figs. 11 and 12 and Supplementary Tables 10 and 11). Regions A, C–F and Q contained no significant genes, which may indicate greater homogeneity in these regions. This analysis demonstrated that trendsceek can identify a variety of spatial gene expression patterns in multiplexed FISH data.

As the application of single-cell gene expression analyses in biology and biomedicine has increased, several computational analysis strategies have been developed¹⁰. Analyses of scRNA-seq data often include the identification of variable genes¹¹, followed by clustering and projections into low dimensions, for example, using PCA and t-SNE, to identify discrete groups of cells¹⁰. Methods have also been developed to assign cells along continuous processes^{12–15}, for example, along a pseudotime of development¹⁵ or pseudospace to capture niches¹⁶. Such methods map the positions of the cells onto a one-dimensional axis, which allows for regression against gene expression for the purpose of univariate gene selection. However, there can exist spatial expression trends that are not captured by a pseudotime analysis, especially because these methods take only the spatial distribution of cells into account, regardless of the presence of spatial segregation with respect to the expression of individual genes. Similarly, a gene with high expression variability among cells may be unrelated to the spatial distribution of the cells.

Trendsceek differs from these methods in that it performs a gene-level test that incorporates both spatial and expression-level information. The spatial analysis complements existing methods that first cluster gene expression profiles, without including spatial information, and then carry out differential expression tests between clusters. In tissue sections that comprise distinct cell types defined by multiple genes, a clustering strategy has high power to resolve cell types. In contrast, spatial methods have the ability to identify continuous gradients or spatial expression patterns defined by fewer genes that would be hard to identify through clustering of pairwise cellular expression profile correlations (Supplementary Fig. 13). In general, genes with significant spatial expression were typically also identified as highly variable, although at widely different ranks (Supplementary Fig. 14), reflecting the fact that only a subset of highly variable genes have significant spatial expression patterns.

For dissociated scRNA-seq data, trendsceek is not intended to replace existing clustering approaches in high-dimensional space. Instead, trendsceek can reveal interesting patterns within a cluster of cells, for example, when clustering and low-dimensional projections are not able to separate cells into further meaningful groups. However, projections of cells from n -gene to two-dimensional

space distort distances between cells. For example, t-SNE seeks only to preserve local distances; thus, we recommend that users restrict such analyses to within clusters, and assess their results using several dimensionality-reduction techniques.

In this study, we explored various types of two-dimensional gene expression data, but future work could extend the methodology to 3D data sets. The spatial metrics are currently based on first and second moments of the expression distribution, but higher-order moments may be needed to identify certain types of spatial structures. Caching the information for all pairs of points at each distance could improve trendsceek runtimes (Supplementary Fig. 15). We implemented trendsceek as an R package to facilitate the adoption of spatial gene expression analyses.

METHODS

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

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

ACKNOWLEDGMENTS

This work was supported by the Swedish Research Council (grant 2017-01062 to R.S.), the European Research Council (grant 648842 to R.S.) and the Bert L. and N. Kuggie Vallee Foundation (R.S.).

AUTHOR CONTRIBUTIONS

D.E. conceived the idea, developed the method, performed the analyses and wrote the manuscript. P.J. performed seqFISH and clustering analyses. R.S. supervised the project and wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

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

1. Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M. & Cai, L. *Nat. Methods* **11**, 360–361 (2014).
2. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. *Science* **348**, aaa6090 (2015).
3. Shah, S., Lubeck, E., Zhou, W. & Cai, L. *Neuron* **92**, 342–357 (2016).
4. Ståhl, P.L. *et al. Science* **353**, 78–82 (2016).
5. Scialdone, A. *et al. Nature* **535**, 289–293 (2016).
6. van der Maaten, L. & Hinton, G. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
7. Illian, J., Penttinen, A., Stoyan, H. & Stoyan, D. *Modelling and Simulation of Stationary Point Processes. Statistical Analysis and Modelling of Spatial Point Patterns* 363–444 (John Wiley & Sons, Ltd, 2008).
8. Hatami, R. *et al. Sci. Transl. Med.* **5**, 169ra12 (2013).
9. Singha, P.K., Yeh, I.-T., Venkatachalam, M.A. & Saikumar, P. *Cancer Res.* **70**, 6377–6383 (2010).
10. Wagner, A., Regev, A. & Yosef, N. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
11. Brennecke, P. *et al. Nat. Methods* **10**, 1093–1095 (2013).
12. Trapnell, C. *et al. Nat. Biotechnol.* **32**, 381–386 (2014).
13. Bendall, S.C. *et al. Cell* **157**, 714–725 (2014).
14. Haghverdi, L., Büttner, M., Wolf, F.A., Büttner, F. & Theis, F.J. *Nat. Methods* **13**, 845–848 (2016).
15. Petropoulos, S. *et al. Cell* **165**, 1012–1026 (2016).
16. Joost, S. *et al. Cell Syst.* **3**, 221–237 (2016).

ONLINE METHODS

Mark-segregation hypothesis testing. To identify spatial gene expression trends, we made use of the theory for marked point processes, which treats the spatial distribution of cells as a realization of a two-dimensional point process and the gene expression levels as a mark distribution in which each point has a scalar-valued mark attached to it, corresponding to the expression of a gene for that cell. The marked point process can be described by a joint probability density $f_1(\mathbf{x}, m)$, denoting the probability of finding a point \mathbf{x} with mark m . Similarly, $f_2((\mathbf{x}_1, m_1), (\mathbf{x}_2, m_2))$ quantifies the probability density to find two points \mathbf{x}_1 and \mathbf{x}_2 with marks m_1 and m_2 . In this study, we used properties of this two-point distribution and parameterized it by the pair separation $r = |\mathbf{x}_2 - \mathbf{x}_1|$ and marks m_1 and m_2 . In particular, we were interested in the probability of finding two marks given the separation of two points,

$$M_2(m_1, m_2 | r) = \frac{f_2(m_1, m_2, r)}{f_2(r)}$$

For the distribution of all pairs at a particular radius, a mark segregation is said to be present if the distribution is dependent on r such that it deviates from what would be expected if the marks were randomly distributed over the spatial locations of the points, that is,

$$M_2(m_1, m_2 | r) \neq M_1(m_1)M_1(m_2)$$

To test for the presence of mark segregation, we implemented permutation tests in which we sampled the mark distribution without replacement and randomly reassigned expression levels to all cells, keeping their positions fixed and in effect conditioning on the given spatial locations. Four summary statistics of the pair distribution were calculated for each radius and compared to the null distribution of the summary statistic derived from the permuted expression labels. As one null distribution and corresponding P value were calculated for each radius, we implemented a multiple-testing adjustment to obtain a global P value, which adjusted for the fact that all possible radii were tested. We did this for each permutation, taking the maximum among all null test statistics across all radii as the null value for that permutation iteration. To account for the fact that the expectation value of the summary statistic can differ between different radii, we used the deviation from the summary statistics sample mean of the null distribution as a test statistic. The implemented tests were two-sided, as the absolute value from the null sample mean was used. The resulting P value for each gene was again multiple-testing adjusted to account for the fact that several genes were tested (here the built-in R function “p.adjust” could be used). Thus, we calculated the nominal P value for a gene by deriving the following entities: a null distribution $D_0 = \max_r |O(r) - E[O](r)|$, where $O(r)$ is a null distribution of the summary statistic for each radius r after permutation of the expression labels, and D_0 is a null distribution with n values, corresponding to the number of permutations, containing the maximum deviation from the mean for each permutation; and an observed deviation for every r , $D_Q(r) = |Q(r) - E[O](r)|$, where Q is the value of the summary statistic for the observed (not permuted) data and D_Q is the observed deviation from the expected mean.

Finally, the P value was calculated via the rank k of the observed value compared with the null:

$$k(r) = \sum_{j=1}^n I(D_{0j} \geq D_Q(r)) \text{ where } I = \begin{cases} 1 & \text{if } D_{0j} \geq D_Q \\ 0 & \text{otherwise} \end{cases}$$

$$P(r) = \frac{k(r)}{n+1}$$

where n is the number of permutations $P_{\text{nominal}} = \min_r P(r)$.

Mark-segregation summary statistics. As mark-segregation summary statistics, we used four previously known two-point spatial statistics (as implemented in the R package spatstat). All four statistics calculate a summary statistic for all pairs of points, conditioned on the distance r between them. The four used statistics are described below; for each, P denotes the set of pairs belonging to pairs of points at a distance r .

(1) Stoyan's mark-correlation function, which uses the squared geometric mean of the marks for all pairs at a distance r , normalized with the squared mean over all points, regardless of distance¹⁷:

$$\rho(r) = \frac{E(m_1 m_2)_P(r)}{\bar{m}^2}$$

(2) The mean-mark function, which is the arithmetic mean over all points belonging to pairs of points separated by distance r (ref. 18):

$$E_{\text{mark}}(r) = \frac{E(m_1 + m_2)_P(r)}{2}$$

(3) The variance-mark function, which is the variance conditioned on the pair separation¹⁸:

$$V_{\text{mark}}(r) = E\left[\left(m_1 - E(m_1)_P(r)\right)^2\right]_P(r)$$

(4) The mark-variogram of a marked point process, which is based on the squared difference of the marks for pairs of points at a distance r from each other¹⁹:

$$\gamma(r) = E\left[\frac{1}{2}(m_1 - m_2)^2\right]_P(r)$$

As edge correction, Ripley's isotropic correction was used^{20,21}. Different summary statistics were included in the trendsceek test, because even if they are not all independent of each other, they capture different aspects of the first and second moment of a distribution and thereby have different powers depending on the mark and spatial distribution under consideration (**Supplementary Fig. 4**).

Run-time optimization. The computational complexity of the described approach is relatively high, as all possible pairs of points ($O(n_2)$) for every permutation ($O(k)$) and gene ($O(m)$) need to be assessed, and thus computational times are often long. To alleviate this, we added two functionalities to trendsceek. First, we implemented parallelization with respect to the iteration over genes, using the R Bioconductor package BiocParallel. Second, we implemented an early-stop procedure in which the number

of permutations was increased in a step-wise fashion one order of magnitude at a time (10, 100, 1,000, ...). If the nominal P value exceeds a given threshold (default = 0.2), then no further permutations are done for that gene.

Identification of cells located in regions of high expression.

If a gene is found to have an expression distribution that is conditionally dependent on the spatial location of cells, according to the mark-segregation tests described above, then trendsseek provides a test to identify cells located in regions with higher gene expression levels than would be expected if the marks of that gene were randomly distributed. To test for this, we generated a null distribution by holding the spatial distribution of the cells fixed and randomly resampling marks without replacement. For each such permutation, a two-dimensional wKDE is performed, using the expression levels as weights and a normal distribution with diagonal bandwidth as the kernel. This generates a two-dimensional smoothed null distribution of expression values against which the wKDE of the observed expression values is compared. A one-sided test is performed to assess whether the expression from the observed wKDE is higher than that corresponding to an upper significance level, and cells located in regions passing the test are extracted.

Power analysis with synthetic data. To assess the power of the mark-segregation tests, we simulated data sets with four possible spatial expression patterns: local hot spot, step gradient, linear gradient and non-radial streak. Sensitivity was calculated as the number of genes with significant P values (≤ 0.05) among 100 independently simulated genes. To assess the robustness of the sensitivity estimate, we repeated this three times and calculated the mean sensitivity and bootstrap confidence interval ($B = 100$) of the mean estimate. The spatial distribution of the cells was generated via a random-point-pattern Poisson process. For the linear gradient, the expression of cells was linearly increased along one dimension. For the other three patterns, cells in a window shaped as one of the three evaluated spatial patterns were spiked with higher expression values. As the number of cells present within the window varied for each realization of the simulation, this can be viewed as including part of the variation introduced, as the number of cells of a particular cell type can vary among samples from a given tissue. The expression values of the cells were bootstrap-sampled from empirical seqFISH data³. Three parameters, apart from the shape of the spike window, were varied in the power analysis: (1) the number of cells; (2) the size of the spike window, corresponding to the number of cells that were to be spiked; and (3) the expression values of the spiked cells, sampled from the upper quantile of the expression distribution where the quantile cutoff was set according to how the fold change between the mean value of the upper quantile and the global mean was varied. For the linear gradient, two parameters were varied: the number of cells, and the fold change between the maximum and minimum expression. To assess the effect of the expression level on sensitivity, we spiked 10 out of 100 cells in a hot-spot region and varied the fold change (2, 5, 10) and the expression level (1, 10, 100 among the low-expression cells). All combinations of these fold changes and expression levels were assessed.

Comparison to spatially unaware differential expression algorithm. To assess the difference between significant genes identified by trendsseek and a differential expression algorithm that does not include spatial information, we selected one representative trendsseek-significant gene for each of the seven spatial patterns found in the spatial transcriptomics mouse olfactory bulb, breast cancer and scRNA-seq data sets. For each of these genes, the cell groups identified by trendsseek's wKDE-based cell-detection algorithm were input, so as to be contrasted, to the differential-expression-analysis program SCDE²².

Analysis of mouse scRNA-seq data. Read counts and cell-annotation metadata from a mouse scRNA-seq gastrulation data set⁵ were downloaded (<http://gastrulation.stemcells.cam.ac.uk/data/counts.gz>) and the subset of cells ($n = 481$) belonging to cluster 3 was kept. The original study⁵ suggested that this cluster of cells contained genes with spatial expression patterns within the cluster. Genes were filtered on expression in at least three cells with a read count of at least 5, which left 17,625 out of 41,388 Ensembl genes. A gene-variability statistic was calculated that adjusted for the mean-variance relationship present in scRNA-seq data. For this we assumed that the expression distribution of a gene follows a negative binomial for which the variance v depends on the mean m : $v = m + m^2/r$, where r is the overdispersion, implying that the coefficient of variance $cv^2 = v/m^2 = 1/m + 1/r$. Assuming that the majority of genes exhibit only technical variability, we fitted such a model to the read counts of all genes and obtained a gene-variability statistic for each gene by adjusting for the variability present among all genes conditioned on the mean expression level¹¹. To stabilize the estimate, we carried out winsorization of the expression distribution of each gene, setting the most extreme value to the expression of the second most extreme cell. On this basis we selected the 500 most variable genes, which we moved forward for analysis by trendsseek. Read counts were normalized using size factors²³, as was done by Scialdone *et al.*⁵, and the 500 most variable genes were subsequently selected. To obtain two-dimensional positions of the cells, we used the t-SNE algorithm⁶ with 100 dimensions from an initial PCA, perplexity of 96, reflecting the number of nearest neighbors to be used, and 400 iterations to reach convergence. As input to trendsseek, the position of the 481 cells in the resulting two-dimensional embedding was used as location of the points and the \log_{10} -normalized expression levels, after addition of a pseudocount of 1 to the size-factor-normalized read counts, were used as marks of the points. We used 10,000 permutations of the marks to obtain the null distribution representing marks being conditionally independent of the spatial location of the cells.

Analysis of spatial transcriptomics data. We downloaded spatial transcriptomic read counts and micro-array spot positions (<http://www.spatialtranscriptomicsresearch.org/datasets/doi-10.1126science-aaf2403/>) containing data from 12 arrays with mouse olfactory bulb tissue sections from five animals and four arrays with sections from a human breast cancer biopsy from a single individual⁴. Genes were filtered on expression in at least three array spots with a read count of at least 5. The most variable genes for each array were derived in a similar manner as described above for the scRNA-seq data, but with the difference

that the expression distribution was assumed to follow a Poisson distribution without any overdispersion, as no good fit was found when a negative binomial was used. This implies that the squared coefficient of variance can be modeled with linear regression with respect to the inverse of the mean expression level ($v = m \Rightarrow cv^2 = 1/m$). The 500 most variable genes from each of the 16 arrays were used as input to trendsceek, along with the spot positions as spatial locations for the point pattern. As expression level marks, we used the \log_{10} -normalized read counts, after adding a pseudocount of 1, and we carried out 10,000 permutations of the marks to obtain the spatially independent null distribution of marks. To group the spatial patterns detected by trendsceek, we input all significant genes (Benjamini–Hochberg-adjusted $P \leq 0.05$ for at least one of the four statistic tests) into trendsceek’s wKDE-based cell-detection algorithm. The resulting binary matrix, indicating cells in regions with elevated gene expression ($P \leq 0.05$), was then clustered by hierarchical agglomerative clustering (Euclidean distance, Ward’s criterion).

Analysis of mouse hippocampus seqFISH data. Raw expression data, cell positioning and vector fields from mouse hippocampus³ were downloaded from <https://ars.els-cdn.com/content/image/1-s2.0-S0896627316307024-mm6.xlsx>. The data set includes a total of 3,585 cells and 249 genes from 21 different regions from a single dissected mouse hippocampus. We retained 2,050 cells after filtering cells around the edges to eliminate image edge artifacts (keeping cells within 203–822 pixels of x - and y -axes, respectively), similar to what was done in the original study³. As input to trendsceek, we carried out winsorization of each gene, setting the four most extreme values to the expression of the fifth most extreme value, and then applied \log_{10} -normalization after adding a pseudocount of 1. The positionings of the cells were specified by the x - and y -coordinates within the 21 individual regions. To

group the spatial patterns detected by trendsceek, we input all significant genes (Benjamini–Hochberg-adjusted $P < 0.01$ for at least one of the four statistic tests, two-sided) into trendsceek’s wKDE-based cell-detection algorithm. The resulting binary matrix, indicating cells in regions with elevated expression ($P < 0.05$, one-sided), was then clustered by hierarchical agglomerative clustering (Euclidean distance, Ward’s criterion).

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

Data availability. The trendsceek R package is available at <https://github.com/edsgard/trendsceek>. Apart from the core functions calculating the trend statistics, a number of additional functions to facilitate the spatial analysis, such as plotting and selection functions, are provided. These are documented in the “vignettes” folder at the above website and in the reference manual of the R package, along with examples of usage. Source data for **Figure 2** are available online.

17. Stoyan, D. & Stoyan, H. *Fractals, Random Shapes, and Point Fields* (John Wiley & Sons Inc, 1994).
18. Schlather, M., Ribeiro, P.J. & Diggle, P.J. *J. R. Stat. Soc. Series B Stat. Methodol.* **66**, 79–93 (2004).
19. Cressie, N.A.C. *Statistics for Spatial Data* (John Wiley & Sons, 1991).
20. Ripley, B.D. Point processes for the earth sciences. in *Quantitative Analysis of Mineral and Energy Resources* (eds. Chung, C.F., Fabbri, A.G. & Sinding-Larsen, R.) 301–322 (Springer, 1988).
21. Ohser, J. *Statistics* **14**, 63–71 (1983).
22. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. *Nat. Methods* **11**, 740–742 (2014).
23. Love, M.I., Huber, W. & Anders, S. *Genome Biol.* **15**, 550 (2014).

Life Sciences Reporting Summary

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

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

► Experimental design

1. Sample size

Describe how sample size was determined.

We evaluated the power to detect significant spatial patterns as a function of the number of cells sampled through simulations. This analysis revealed that spatial structures were reliably identified if at least 5% of sequenced cells have differing expression levels, in particular when the total number of analysed cells exceed 500. Analyses of published spatial gene expression data made use of all available information.

2. Data exclusions

Describe any data exclusions.

For all analyzed data sets (Spatial transcriptomics and seqFISH) we report the number of replicate slides within which we identified significant spatial patterns. For seqFISH we excluded cells around the edges of each image, based on a list provided by the authors of that publication, to remove image edge artifacts. They removed cells where the centroid of cells were within 200-250 pixel of the edge.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Simulated patterns were reidentified by the algorithm to the degree presented in the power analysis, including a 95% bootstrap confidence interval. The algorithm works on a per-gene basis and in the real data the same spatial patterns were identified for multiple genes. We analyzed all replicates/regions provided in the respective data set analyzed.

4. Randomization

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

No randomization performed.

5. Blinding

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

No blinding was performed.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

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

n/a Confirmed

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

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

► Software

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

7. Software

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

The computational method introduced in this study has been made available as an R/Bioconductor package. This package was used for all analyses in the manuscript.

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

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

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

No restrictions.

9. Antibodies

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

No antibodies were used.

10. Eukaryotic cell lines

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

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

N/A

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

N/A

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

N/A

► Animals and human research participants

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

11. Description of research animals

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

No animals were used.

12. Description of human research participants

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

The study did not involve human research participants.