

Published in final edited form as:

*Nat Methods*. 2018 May ; 15(5): 339–342. doi:10.1038/nmeth.4634.

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

**Daniel Edsgård<sup>1,2</sup>, Per Johnsson<sup>1,2</sup>, and Rickard Sandberg<sup>1,2,\*</sup>**

<sup>1</sup>Department of Cell and Molecular Biology, Karolinska Institutet, Sweden

<sup>2</sup>Ludwig Institute for Cancer Research, Stockholm, Sweden

### Abstract

Methods for spatial gene expression analyses at single-cell resolution are becoming available, whereas computational strategies for spatial gene expression analyses are lacking. We present a computational method (*trendsseek*) based on marked point processes that identifies genes with significant spatial expression trends. *Trendsseek* identifies significant genes in spatial transcriptomics and sequential FISH data and also reveal significant gene expression gradients and hotspots in low-dimensional projections of dissociated single-cell RNA-seq data.

---

Analyses of gene expression at single-cell resolution and in spatial contexts will reveal insights into the molecular organizations of tissue and organs. Although transcriptome-wide single-cell gene expression analyses with spatial information is not yet feasible, recent progress in barcoded single-molecule FISH<sup>1,2</sup> have enabled sequential analyses of hundreds of genes in tissues<sup>3</sup>. In parallel, tissue sections lysed and reverse-transcribed on barcoded surfaces represents another approach for high-throughput analyses of gene expression with regional spatial information<sup>4</sup>. Although these approaches enable gene expression analyses in spatial context, spatial gene expression analyses methods are lacking. For example, cellular and regional expression profiles are typically analysed first without the spatial information and only later projected back onto the spatial structure for visual inspection of spatial trends<sup>3,4</sup>.

Here, we introduce a computational method that identifies significant spatial gene expression trends that we named *trendsseek*. In order to identify genes with significant dependencies between the spatial distribution of cells and the gene expression in these cells, we model the data as marked point processes to rank and assess the significance of the spatial expression trends of each individual gene. We first validate the method on simulated data sets and thereafter demonstrate that *trendsseek* can identify genes with significant

---

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:[http://www.nature.com/authors/editorial\\_policies/license.html#terms](http://www.nature.com/authors/editorial_policies/license.html#terms)

\*correspondence to: Rickard.Sandberg@ki.se.

#### 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 Financial Interests

The authors declare no competing financial interests.

spatial patterns in spatial transcriptomics data (mouse olfactory bulb and breast cancer sections<sup>4</sup>) and in sequential FISH (seqFISH) data (hippocampus)<sup>3</sup>. Moreover, *trendsceek* could reveal significant gradients and patterns even for dissociated single-cell RNA-seq data<sup>5</sup> that had been projected to a low-dimensional space (t-distributed stochastic neighbour embedding, t-SNE<sup>6</sup>). Spatial patterns were found within low-dimensional clusters demonstrating the general utility of simultaneously incorporating expression and location information for finding significant spatial trends. Finally, *trendsceek* has been implemented as an R package to allow for broad applications to many types of spatial gene expression data.

To model spatial gene expression we made use of marked point processes, a statistical framework which has previously been applied in the fields of geostatistics, astronomy and material physics<sup>7</sup>. For spatial analyses of gene expression, the points represent the spatial locations of cells (or regions) and the marks of each point constitute expression levels. Importantly, this approach is non-parametric and can identify general non-linear expression patterns without the need to specify a distribution or spatial region of interest. Briefly, our method assesses whether a significant dependency exist 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 points. Summary statistics used for assessing dependencies include conditional mean (E-mark), conditional variance (V-mark), Stoyan's mark correlation (not a true correlation measure) and the mark-variogram (Methods). Notably, if marks and the location 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 where the expression values were permuted, reflecting a null-model with no spatial dependency of the expression. Once a gene with a significant spatial trend has been identified it is also of interest to find the subset of cells in 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 against a null-model derived from permuted expression values.

We first investigated the method on simulated spatial expression data (sampled from empirical seqFISH data; Methods) where cells with higher expression were located in local hotspots, step-gradients or non-radial streaks, or where the expression followed a linear gradient (Figure 1A). Exemplifying the analyses of the hotspot pattern, the mark-correlation and mark-variogram were significant at low  $r$  values (Figure 1B,  $P < 0.05$ ) as determined via 1,000 randomly permuted expression distributions of the same marks (the 5% critical rejection band of these randomizations are shown as grey areas in Figure 1B). Power analysis of the four metrics as a function of the number of cells, expression level, expression level difference and the size of the region with elevated expression are presented in Supplementary Figures 1, 2. The analysis revealed that spatial structures are reliably identified if at least 5% of sampled cells have differing expression levels, in particular when the total number of analysed cells exceed 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). We conclude that the

method has sufficient power to reveal a variety of spatial patterns involving a small number of cells.

Next, we developed an approach to identify the cells belonging to regions with elevated expression patterns. Using wKDE, we identified cells exceeding a 5% significance level (green surface in Figure 1C) by comparison to the upper 5% quantile of a two-dimensional null distribution generated by resampling the mark distribution (blue surface in Figure 1C). Significant cells are shown on top of the density estimate of the expression in Figure 1D. Having developed a method for the identification of genes with spatial expression trends and an approach to pinpoint the cells belonging to regions of interest within the pattern, we next explored recently published gene expression data.

We first analysed spatial transcriptomics data from olfactory bulb tissue<sup>4</sup>, and *trendsseek* identified 35 significant genes (Figures 2A-B, Supplementary Figures 3A, 4A, 5A, 6A-B and Supplementary Tables 1, 2, 3;  $P < 0.05$ , Benjamini-Hochberg adjusted) with expression primarily in non-granular cells. These genes included *Ptn*, *Nr2f2* and *Fabp7* which has been detected as tissue-domain restricted using principal component analyses (PCA)<sup>4</sup> and with clear spatial RNA in situ signatures in the Allen Brain Institute atlas (see Supplementary Figure 9 in Ståhl et al. 2016<sup>4</sup>). We also identified 45 genes with significant expression primarily in granular regions of the bulb (Figures 2C-D, Supplementary Figures 3B, 4B, 5B, 6C-D and Supplementary Tables 1, 4, 5) including known genes such as *Nrgn*, *Camk4* and *Pcp4* but also novel genes such as *Gpsm1* (Figure 2C). A strength of spatial transcriptomics is the ability to profile tumour tissues. Applying *trendsseek* to spatial profiling of human breast cancer tissue (layer 2)<sup>4</sup> identified 14 genes with significant spatial expression (Figure 2E, Supplementary Figures 3C, 4C, 7, 8 and Supplementary Tables 1, 6, 7). Several genes implicated in breast cancer had significant spatial patterns, including the transcription factor *Klf6*, the transmembrane protein *TMEPA19*, and twelve genes related to the extracellular matrix (ECM). We conclude that *trendsseek* can be broadly applied to spatial transcriptomics data to find genes with significant spatial trends.

The vast majority of single-cell RNA-sequencing has been performed on dissociated cells that lack spatial information. The analysis of single-cell gene expression data often include clustering and visualization of cells in low-dimensional spaces (using e.g. PCA or t-SNE). We determined whether *trendsseek* could find spatial patterns within two-dimensional representations of dissociated single-cell data. t-SNE analysis of scRNA-seq data from a mouse gastrulation dataset identified a larger cluster of 481 epiblast cells from E6.5 mice<sup>5</sup>, still *trendsseek* identified 107 genes with significant spatial expression patterns ( $P < 0.05$ , Benjamini-Hochberg adjusted) within that cluster of cells (Supplementary Figure 9). The vast majority among the significant genes were characterized by an expression gradient with higher expression at the narrow part of the cell cluster (Supplementary Figure 10, exemplified by *T* and *Fgf8* in Figure 2F). Moreover, we identified hotspots of male cells that separated from female cells with significant expression of *Eif2s3y* and *Xist*, respectively (Figure 2G and Supplementary Tables 1, 8, 9;  $P < 0.05$ , Benjamini-Hochberg adjusted). Different sets of genes were identified by the mark-correlation and mark-variogram based tests, indicating that the inclusion of multiple summary statistics improves sensitivity (Supplementary Figure 4D,H). We conclude that *trendsseek* can reveal diverse spatial

patterns manifesting as gradients or hotspots within low-dimensional projections of dissociated single-cell data, in a fully unbiased manner without incorporating *a priori* knowledge about candidate genes.

Finally, we applied *trendsceek* to sequential FISH data from 21 mouse hippocampus regions<sup>3</sup> that in total included approximately 3,500 cells and 249 genes. We identified genes with significant spatial patterns in 15 out of 21 regions (median of 54 genes per region) (Supplementary Figures 11, 12 and Supplementary Tables 10, 11), representative spatial patterns for five genes are shown in Figure 2H. 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 single-cell gene expression analyses are being broadly applied in biology and biomedicine several computational analysis strategies have been developed<sup>10</sup>. Analyses of scRNAseq data often include the identification of variable genes<sup>11</sup>, followed by clustering and projections into low-dimensions, e.g. using PCA and t-SNE, to identify discrete groups of cells<sup>10</sup>. Methods have also been developed to assign cells along continuous processes<sup>12–15</sup>, e.g. along a pseudo-time of development<sup>15</sup> or pseudo-space to capture niches<sup>16</sup>. These latter 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 pseudo-time analysis, especially since these methods only take the spatial distribution of cells into account regardless of the presence of a 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 as it performs a gene-level test that jointly incorporates spatial and expression-level information. The spatial analysis complements existing methods that first cluster gene expression profiles, without including spatial information, followed by differential expression tests between clusters. In tissue sections comprised by distinct cell-types defined by multiple genes, a clustering strategy has high power, whereas spatial methods have the additional ability to identify continuous gradients or spatial expression patterns defined by a fewer number of genes which would be hard to identify by cell-cell expression profile correlation based clustering of cells (Supplementary Figure 13). In general, genes with significant spatial expressions were typically also identified as highly variable, although at widely different ranks (Supplementary Figure 14), reflecting that only a subset of highly variable genes have significant spatial expression patterns.

For dissociated scRNAseq 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, i.e. when clustering and low dimensional projections fail to separate cells into further meaningful groups. However, since projections of cells from n-gene to two-dimensional space distorts distances between cells, e.g. tSNE seeks only to preserve local

distances, we recommend users to 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 include extending 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. *Trendsseek* runtimes are provided in Supplementary Figure 15 and could be further sped up by caching the information regarding all pairs of points at each distance. We implemented *trendsseek* as an R package to facilitate the adoption of spatial gene expression analyses.

## Online Methods

### Mark-segregation hypothesis testing

To identify spatial gene expression trends, we made use of the theory for marked point processes, where the spatial distribution of cells are treated as a realization of a two-dimensional point process and the gene expression levels as a mark distribution where 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 at  $\mathbf{x}$  with mark  $m$ . Similarly,  $f_2((\mathbf{x}_1, m_1), (\mathbf{x}_2, m_2))$  quantifies the probability density to find two points at  $\mathbf{x}_1$  and  $\mathbf{x}_2$  with marks  $m_1$  and  $m_2$ . In this study, we use properties of this two-point distribution, and parameterize it by the pair separation  $r = |\mathbf{x}_2 - \mathbf{x}_1|$  and the marks  $m_1, m_2$ . In particular, we are interested in the probability of finding two marks given the separation of two points,

$$M_2(m_1, m_2 | r) = \frac{f_2(m_1, m_2, r)}{f_2(r)}. \text{ Studying the distribution of all pairs at a particular radius, a}$$

mark segregation is said to be present if this distribution is dependent on  $r$ , such that it deviates from what would be expected if the marks would be 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, where the mark distribution was sampled without replacement and expression levels were randomly reassigned to all cells keeping their positions fixed and in effect conditioning on the given spatial locations. Four summary statistics of the pair distribution was 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 was calculated for each radius we implemented a multiple-testing adjustment as to achieve a global  $P$ -value, which adjusts for that all possible radii are tested. This was done by for each permutation, taking the maximum among all null test-statistic across all radii as the null value for that permutation iteration. To account for 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 test-statistic. The implemented tests are 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 that several genes were

tested (here the built-in R-function “*p.adjust*” could be used). Thus, the nominal *P*-value for a gene was calculated 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 permuting 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. 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  $D_Q$  the observed deviations from the expected mean.

Finally, the *P*-value was calculated via the rank,  $k$ , of the observed value compared to the null:

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

$$P(r) = \frac{k(r)}{n+1}, \text{ where } n \text{ is the number of permutations}$$

$$P_{\text{nominal}} = \min_r P(r)$$

### Mark-segregation summary statistics

As mark-segregation summary statistics we used four previously known two-point spatial statistics (as implemented in R-package *spatstat*). All four statistics calculate a summary statistic for all pairs of points, conditioned on their separation distance,  $r$ . The four used statistics were, where  $P$  denotes the set of pairs belonging to pairs of points at distance  $r$  apart:

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 distance<sup>17</sup>:

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

The mean-mark function, which is the arithmetic mean over all points belonging to pairs of points separated by distance  $r$ <sup>18</sup>:

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

The variance-mark function, which is the variance conditioned on the pairs separation<sup>18</sup>:

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

The mark-variogram of a marked point process, which is based on the squared difference of the marks for pairs of points at distance  $r$  apart<sup>19</sup>:

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

As edge correction Ripley's isotropic correction was used<sup>20,21</sup>. Different summary statistics were included in the *trendsceek* test, since 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 power depending on the mark and spatial distribution under consideration (Figure S4).

### Run-time optimization

The computational time complexity of the described approach is relatively high as all possible pair of points ( $O(n_2)$ ), for every permutation ( $O(k)$ ) and gene ( $O(m)$ ), needs to be assessed. To alleviate this, two functionalities were added to *trendsceek*. First, we implemented parallelization with respect to the iteration over genes, using the R Bioconductor package *BiocParallel*. Second, we implemented an early-stopping procedure where the number of permutations are increased in a step-wise fashion one order of magnitude at a time (10, 100, 1000, ...). 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 has been deemed to have an expression distribution that is conditionally dependent on the spatial location of the cells, according to the mark-segregation tests described above, then *trendsceek* provides a test to identify cells located in regions with higher expression level than expected if the marks of that gene would be randomly distributed. To test for this a null distribution is generated by holding the spatial distribution of the cells fixed and marks being randomly resampled without replacement. For each such permutation, a two-dimensional weighted kernel density estimation (wKDE) is performed using the expression levels as weights and a normal distribution with diagonal bandwidth as 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 if the expression from the observed wKDE has higher expression 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 datasets with four possible spatial expression patterns: local hotspot, step-gradient, linear-gradient and non-radial streaks. Sensitivity was calculated as the number of genes with significant P-value (< 0.05)

among 100 independently simulated genes. To assess the robustness of the sensitivity estimate this was repeated three times and the mean sensitivity and bootstrap confidence interval ( $B = 100$ ) of the mean estimate was calculated. The spatial distribution of the cells was generated using 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 then spiked with higher expression values. As the number of cells present within the window varies for each realization of the simulation this can be viewed as including part of the variation introduced due to that the number of cells of a particular cell type that is retrieved when obtaining a tissue sample can vary. The expression values of the cells were bootstrap-sampled from empirical sequential FISH data<sup>3</sup>. 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 were 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, 10 out of 100 cells were spiked in a hotspot region and the fold-change (2, 5, 10) and the expression level (1, 10, 100 among the lowly expressed cells) was varied. 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 from *trendsceek* and a differential expression algorithm that does not include spatial information, one representative *trendsceek*-significant gene for each of the seven spatial patterns that was found in the spatial transcriptomics mouse olfactory bulb, the breast cancer and the scRNA-seq dataset was selected. For each of these genes, the cell-groups identified by *trendsceek*'s wKDE-based cell-detection algorithm were input, as to be contrasted, to the differential expression analysis program SCDE<sup>22</sup>.

### Analysis of mouse scRNA-seq data

Read counts and cell-annotation meta-data from a mouse scRNA-seq gastrulation dataset<sup>5</sup> were downloaded (<http://gastrulation.stemcells.cam.ac.uk/data/counts.gz>) and the subset of cells ( $n=481$ ) belonging to cluster 3 was kept. This cluster of cells was implicated in the paper to contain genes exhibiting spatial patterns within the cluster. Genes were filtered on being expressed in at least 3 cells with a read count of at least 5, leaving 17,625 out of 41,388 Ensembl genes. A gene-variability statistic was calculated that adjusted for the mean-variance relationship present in single-cell RNA-seq data. This was done by assuming that the expression distribution of a gene follow 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 only exhibit technical variability we fitted such a model to the read counts of all genes and a gene-variability statistic was then obtained for each gene by adjusting for the variability present among all genes conditioned on the mean expression level<sup>11</sup>. To stabilize the

estimate, we performed winsorization of the expression distribution of each gene, setting the most extreme value to the expression of the second most extreme cell. Based on this we selected the 500 most variable genes, which were put forward to analysis by *trendsceek*. Read counts were normalized using size-factors<sup>23</sup>, as was done by Scialdone et al<sup>5</sup>, and the 500 most variable genes were subsequently selected. To obtain two-dimensional positions of the cells we used the t-SNE, Student's t distributed stochastic neighbourhood embedding, algorithm<sup>6</sup>, with 100 dimensions from an initial principal component analysis, perplexity = 96, reflecting the number of nearest neighbours to be used, and 400 iterations to reach convergence. As input to *trendsceek*, the position of the 481 cells in the resulting two-dimensional embedding was used as location of the points and the log<sub>10</sub>-normalized expression levels, after adding a pseudo-count of 1 to the size-factor normalized read counts, were used as marks of the points. 10,000 permutations of the marks were used to obtain the null-distribution representing marks being conditionally independent of the spatial location of the cells.

### Analysis of spatial transcriptomics data

Spatial transcriptomic read counts and micro-array spot positions were downloaded (<http://www.spatialtranscriptomicsresearch.org/wp-content/uploads/2016/07>), containing data from 12 arrays with mouse-olfactory bulbs tissue sections from 5 animals and four arrays with sections from a human breast cancer biopsy from a single individual<sup>4</sup>. *Trendsceek* was applied on replicate 3 (n=269 array-spots) and replicate 12 (n=280 array-spots) of mouse-olfactory bulbs tissue sections and layer 2 of human breast cancer biopsy (n=251 array-spots). Genes were filtered on being expressed in at least 3 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 using a negative binomial. This corresponds to that the squared coefficient of variance can be modelled 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, the log<sub>10</sub>-normalized read counts were used, after adding a pseudo-count of 1, and 10,000 permutations of the marks was performed to obtain the spatially independent null-distribution of marks. To group the spatial patterns detected by *trendsceek*, all significant genes (Benjamini-Hochberg adjusted  $P < 0.05$  for at least one of the four statistic tests) were input to *trendsceek*'s wKDE-based cell-detection algorithm. The resulting binary matrix, indicating cells in regions with elevated expression ( $P < 0.05$ ), 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 hippocampus<sup>3</sup> were downloaded from (<https://ars.els-cdn.com/content/image/1-s2.0-S0896627316307024-mmc6.xlsx>). The data set includes a total of 3585 cells and 249 genes from 21 different regions from a single dissected mouse hippocampus. 2050 cells were retained after filtering cells around the edges to eliminate image edge artifacts (keeping cells within 203-822 pixels

of x- and y-axis), similar as was done in the original publication. As input to *trendsseek*, we performed winsorization of each gene setting the four most extreme values to the expression of the fifth most extreme value followed by  $\log_{10}$ -normalization after adding a pseudo-count 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 *trendsseek*, all significant genes (Benjamini-Hochberg adjusted  $P < 0.01$  for at least one of the four statistic tests, two-sided) were input to *trendsseek*'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).

### Data availability statement

The R package *trendsseek* is available at <https://github.com/edsgard/trendsseek>. 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 vignette and the reference-manual of the R-package along with examples of usage.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

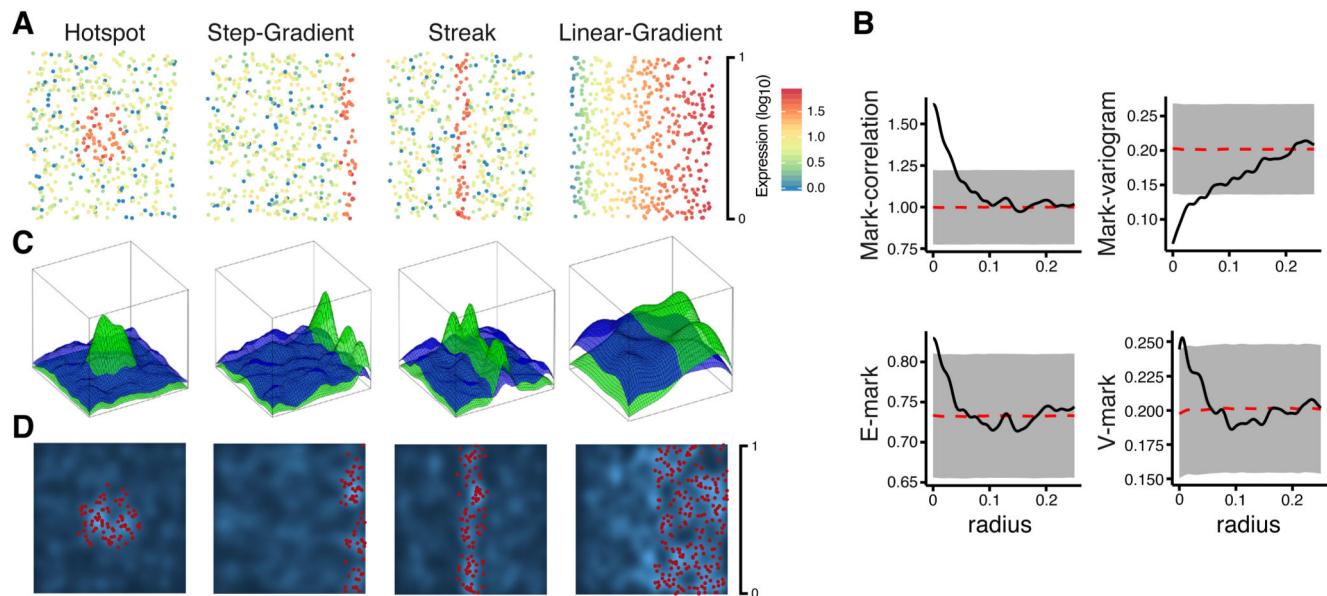
## Acknowledgements

This work was supported by grants from the Swedish Research Council, the European Research Council (CoG 648842) and the Vallee Foundation.

## References

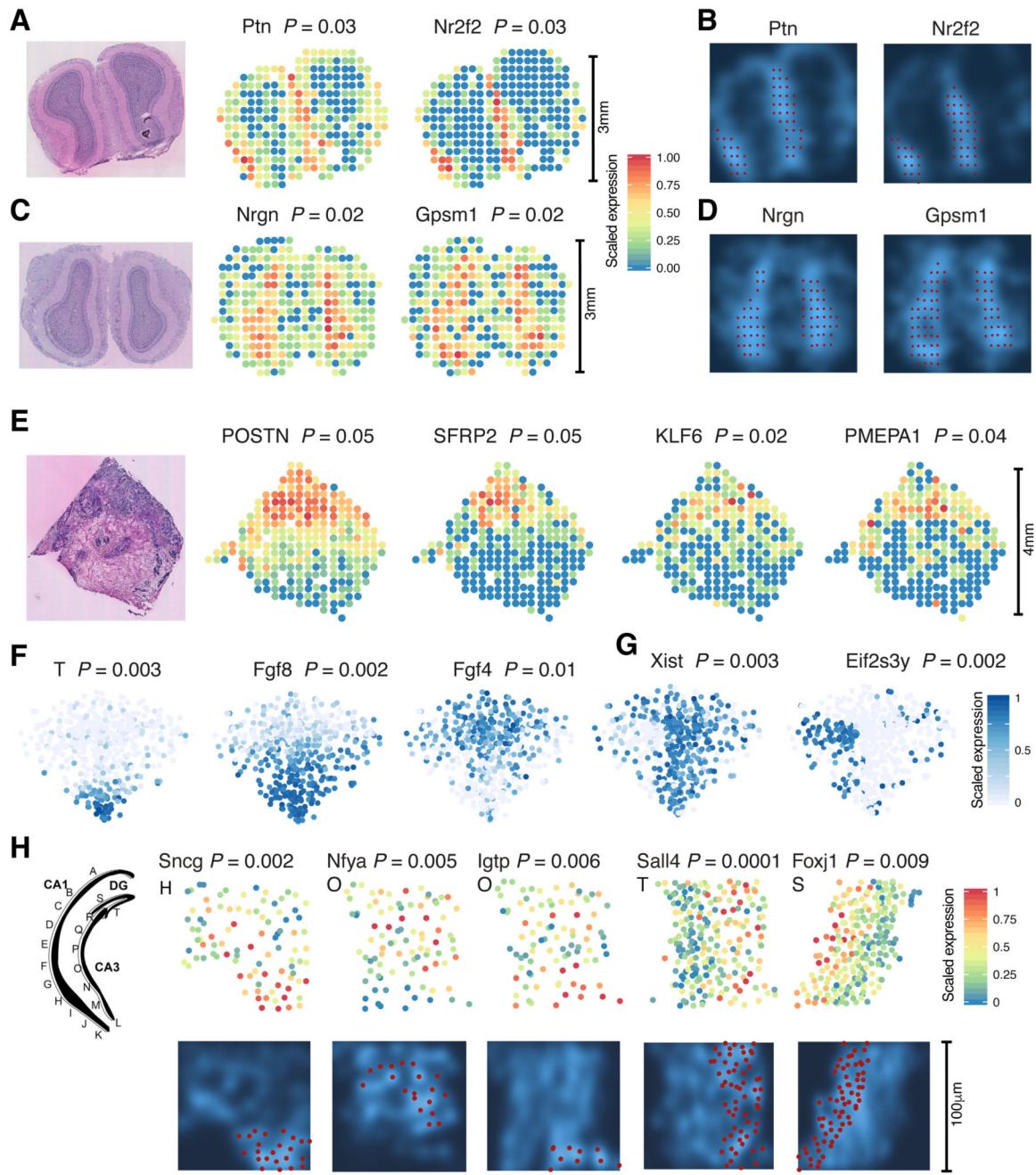
1. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell *in situ* RNA profiling by sequential hybridization. *Nat Methods*. 2014; 11:360–361. [PubMed: 24681720]
2. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*. 2015; 348
3. 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]
4. Ståhl PL, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. 2016; 353:78–82. [PubMed: 27365449]
5. Scialdone A, et al. Resolving early mesoderm diversification through single-cell expression profiling. *Nature*. 2016; 535:289–293. [PubMed: 27383781]
6. van der Maaten L, Hinton G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*. 2008;1–27.
7. Illian J, Penttinen A, Stoyan H, Stoyan D. Modelling and Simulation of Stationary Point ProcessesStatistical Analysis and Modelling of Spatial Point Patterns. John Wiley & Sons, Ltd; 2008. 363–444.
8. Hatami R, et al. KLF6-SV1 drives breast cancer metastasis and is associated with poor survival. *Sci Transl Med*. 2013; 5:169ra12–169ra12.
9. Singha PK, Yeh I-T, Venkatachalam MA, Saikumar P. Transforming growth factor-beta (TGF-beta)-inducible gene TMEPAI converts TGF-beta from a tumor suppressor to a tumor promoter in breast cancer. *Cancer Res*. 2010; 70:6377–6383. [PubMed: 20610632]

10. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol.* 2016; 34:1145–1160. [PubMed: 27824854]
11. Brennecke P, et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods.* 2013; 10:1093–1095. [PubMed: 24056876]
12. Trapnell C, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol.* 2014; 32:381–386. [PubMed: 24658644]
13. Bendall SC, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell.* 2014; 157:714–725. [PubMed: 24766814]
14. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods.* 2016; 13:845–848. [PubMed: 27571553]
15. Petropoulos S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell.* 2016; 165:1012–1026. [PubMed: 27062923]
16. Joost S, et al. Single-Cell Transcriptomics Reveals that Differentiation and Spatial Signatures Shape Epidermal and Hair Follicle Heterogeneity. *Cell Syst.* 2016; 3:221–237.e9. [PubMed: 27641957]
17. Stoyan, D, Stoyan, H. Fractals, random shapes, and point fields. John Wiley & Sons Inc; 1994.
18. Schlather M, Ribeiro PJ, Diggle PJ. Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2004; 66:79–93.
19. Cressie, NAC. Statistics for Spatial Data. John Wiley & Sons; 1991.
20. Ripley, BD. Quantitative Analysis of Mineral and Energy Resources. Springer; Netherlands: 1988. 301–322.
21. Ohser J. On estimators for the reduced second moment measure of point processes. *Series Statistics.* 1983; 14:63–71.
22. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nat Methods.* 2014; 11:740–742. [PubMed: 24836921]
23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014; 15:550. [PubMed: 25516281]



**Figure 1. Illustrating *trendsseek* on simulated data.**

(A) Simulated mark distributions with local hotspot, step gradients, non-radial streaks 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 (cells n=500, spiked cells n=~50, mean expression spiked cells / mean expression background = ~10). (B) Marked point pattern statistics (Methods) for the simulated hotspot shown in (A). The mark correlation and mark variogram clearly indicate a significant spatial pattern as the band, which indicates the 5% significance level of a null distribution based on resampling of the mark distribution, is exceeded at certain radii. (C) 3D-representations of the spatial expression trend, where the green surfaces show weighted kernel density estimation (wKDE) of the simulated datasets. The blue surfaces indicate the upper 5% quantile of a null distribution generated by wKDE of the resampled mark distribution for each dataset. (D) Density plot of the simulated datasets (A) with cells colored red if they exceeded a 5% significance level based on wKDE, indicated by the blue surface in (C). Scale bars in (A) and (D) apply to all items of the figure, including the radius shown in (B).



**Figure 2. Applications of *trendsseek* on spatial and single-cell gene expression data.**

(A) Spatial transcriptomics data from mouse olfactory bulb (replicate 3, n=269 array-spots). Left: hematoxylin and eosin stained tissue-sections (from Ståhl et al4), followed by examples of genes with significant expression trends. Expression was scaled to the range zero to one by unity-based normalization. (B) Density plots of gene expression with cells in regions of significantly elevated expression coloured red. (C) Spatial transcriptomics data from mouse olfactory bulb (replicate 12, n=280 array-spots), as in (A). (D) As in (C) for mouse olfactory bulb, replicate 12. (E) Spatial transcriptomics data from breast cancer

biopsy (histological section “Layer 2”, n=251 array-spots), with examples of genes with significant expression trends. The distance between array-spots is 200 $\mu$ m, each spot covering multiple cells. (F) Examples of distinct spatial expression patterns identified by *trendsceek* within E6.5 mouse epiblast cells (cluster 3 in Scialdone et al.5, n=481 cells). (G) Identification of spatial patterns related to the positions of male and female cells within the cluster, with mutually exclusive expression of *Xist* (expressed in female cells) and *Eif2s3y* (located on the Y-chromosome, only expressed in male cells). (H) Examples of spatial expression patterns identified in mouse hippocampus seqFISH data (cells imaged; H=93, O=89, T=208). Left: Cartoon of hippocampus with the 21 imaged regions labelled according to previous publication3. P-values represent (A-E) mark-correlation (F-G) mark-variogram and (H) Emark (two-sided, Benjamini-Hochberg adjusted).