Evaluating spatially variable gene detection methods for spatial transcriptomics data

Carissa Chen[1,4], Hani Jieun Kim[1,4], Pengyi Yang[1,2,3,*]

[1] Computational Systems Biology Group, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW 2145, Australia

[2] School of Mathematics and Statistics, The University of Sydney, Sydney, NSW 2006, Australia

[3] Charles Perkins Centre, The University of Sydney, Sydney, NSW 2006, Australia

[4] Contributed equally

* Corresponding author (pengyi.yang@sydney.edu.au)

SUMMARY

The identification of genes that vary across spatial domains in tissues and cells is an essential step for spatial transcriptomics data analysis. Given the critical role it serves for downstream data interpretations, various methods for detecting spatially variable genes (SVGs) have been proposed. The availability of multiple methods for detecting SVGs bears questions such as whether different methods select a similar set of SVGs, how reliable is the reported statistics significance from each method, how accurate and robust is each method in terms of SVG detection, and how well the selected SVGs perform in downstream applications such as clustering of spatial domains. Besides these, practical considerations such as computational time and memory usage are also crucial for deciding which method to use. In this study, we address the above questions by systematically evaluating a panel of popular SVG detection methods on a large collection of spatial transcriptomics datasets, covering various tissue types, biotechnologies, and spatial resolutions. Our results shed light on the performance of each method from multiple aspects and highlight the discrepancy among different methods especially on calling statistically significant SVGs across datasets. Taken together, our work provides useful considerations for choosing methods for identifying SVGs and serves as a key reference for the future development of such methods.

INTRODUCTION

Advances in spatial transcriptomics have made it possible to identify genes that vary across spatial domains in tissues and cells (Svensson et al., 2018a). The detection of spatially variable genes (SVGs) is essential for capturing genes that carry biological signals and reducing the high-dimensionality of the spatial transcriptomics data (Svensson et al., 2018a), which is akin to defining highly variable genes (HVGs) (Yip et al., 2019) in single-cell RNA-sequencing (scRNA-seq) data (Kolodziejczyk et al., 2015). These SVGs are therefore useful for various downstream analyses of spatial transcriptomics data. Spatially variable genes are conceptually different from

HVGs found in scRNA-seq data as, by definition, SVGs preserve the spatial relationships of tissues and cells in the biological samples whereas HVGs do not necessarily preserve such relationships.

A fast-growing number of methods for SVG detection have been proposed in the recent literature. Some popular examples include SpatialDE (Svensson et al., 2018a) based on Gaussian process; SPARK (Sun et al., 2020) and SPARK-X (Zhu et al., 2021) based on mixed and non-parametric models, respectively; SOMDE based on self-organising map(Hao et al., 2021a); Giotto based on statistical enrichment of spatial network in neighbouring cells(Dries et al., 2021a); nnSVG based on nearest-neighbour Gaussian processes (Weber et al., 2022); and Moran's I as implemented in the Seurat package(Hao et al., 2021b). While various SVG detection methods have been incorporated into the typical workflows and pipelines for spatial transcriptomics data analysis such as the Giotto and Seurat packages, there is a lack of systematic evaluation and comparison of different methods. In particular, essential questions including the degree of agreement among different methods in terms of the ranking and selection of SVGs, the reproducibility of these methods in terms of SVG detection when the genes included in a given dataset changes, and the accuracy and robustness of SVG detection, and the utility of the selected SVGs to perform in downstream data analysis such as spatial domain clustering remain to be addressed. In addition, practical considerations such as running time and memory usage required by each method have not been systematically benchmarked.

To fill this critical gap, we systematically evaluated a panel of seven popular SVG detection methods on a collection of 20 spatial transcriptomics and simulated datasets. These datasets together capture various sample and tissue types and main spatial biotechnologies with different profiling resolutions, including Visium (10X Genomics), ST (Ji et al., 2020), Slide-seq (Rodriques et al., 2019), MERFISH (Xia et al., 2019), seqFISH+ (Eng et al., 2019), and Stereo-seq (Chen et al., 2022). Our results shed light on the performance of each tested SVG detection method in various aspects and highlight some of the discrepancies among different methods especially on calling statistically significant SVGs across datasets. Taken together, this work provides useful information for considering and choosing methods for identifying SVGs whilst also serving as a key reference for future development of SVG detection methods from spatial transcriptomics data.

RESULTS

Evaluation framework and data summary

We designed an evaluation framework to gain insight into the performance of different SVG detection methods to call SVGs from a collection of real and simulated spatially resolved transcriptomics datasets (Figure 1). These include spatial transcriptomics data generated from a wide range of spatial profiling platforms, species, tissue types, and spatial resolutions (Figure S1a). Specifically, our evaluation framework entailed a wide range of comparative and benchmarking analyses to investigate key questions. First, we compared the concordance between the overall rankings of the SVGs between SVG tools and evaluated their dependence on mean gene expression to assess the variability among methods and their capacity to account for the bias between gene expression and variance. Next, we investigated the capacity

of each SVG method to reproducibly rank SVGs independently of the pool of genes observed in the dataset, to call ground truth spatially variable genes, and to define SVGs required to accurately cluster spatial domains. Finally, using the spatial benchmarking datasets we compared the computational cost in terms of time and memory required for SVGs to be called by each method.

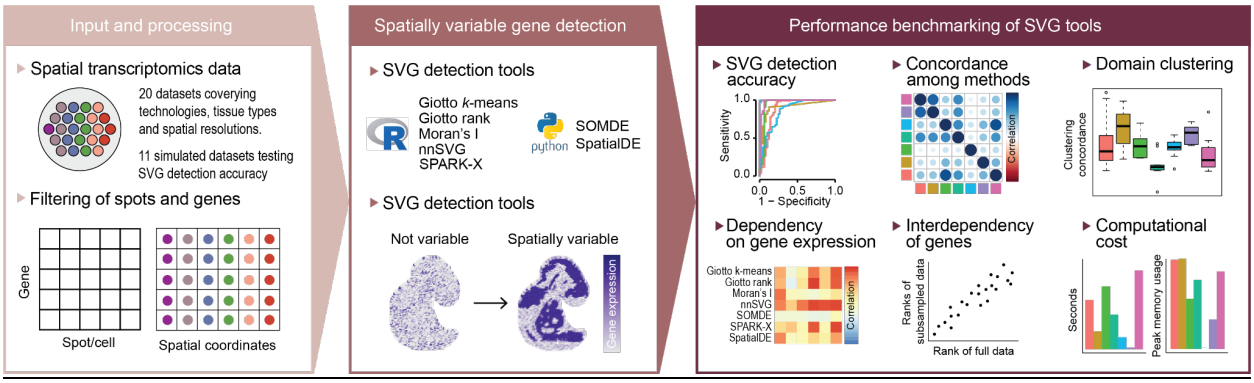

Figure 1. Schematic summary of the evaluation framework used in this study.

Concordance among SVG detection methods

To quantify the degree of agreement amongst the different SVG detection methods, we first obtained the ranking of genes in each dataset ordered from the most to least spatially variable based on the statistics reported by each method and correlated the SVG rankings from each pair of methods. These correlation results were summarised for each SVG detection method with respect to other methods across datasets in Figure 2a. Notably, SpatialDE and Moran's I demonstrated a high correlation with each other. The two spatial network enrichment approaches (i.e. Giotto KM and Giotto rank) implemented in the Giotto package are also highly correlated. Interestingly, SOMDE appears to be generating the least correlated SVG ranking and highest variability across datasets. nnSVG has a comparable and relatively high correlation with all methods except with SOMDE. SPARK-X shows moderate correlation with most methods and relatively higher correlation with nnSVG and lower correlation with SOMDE. Looking at the correlation results on each dataset individually (Figure S2), we found that datasets generated using the Visium platform result in a more consistent SVG ranking among methods except SOMDE in a subset of these datasets. Datasets generated using other spatial transcriptomics profiling platforms generally result in lower consistency in SVG ranking across computational methods. Overall, these results demonstrate that whilst we observed a moderate-to-high correlation between SVG detection tools in terms of SVG ranking, we found considerable variability in reported SVG statistics and their significance across the computational methods, platforms, and datasets.

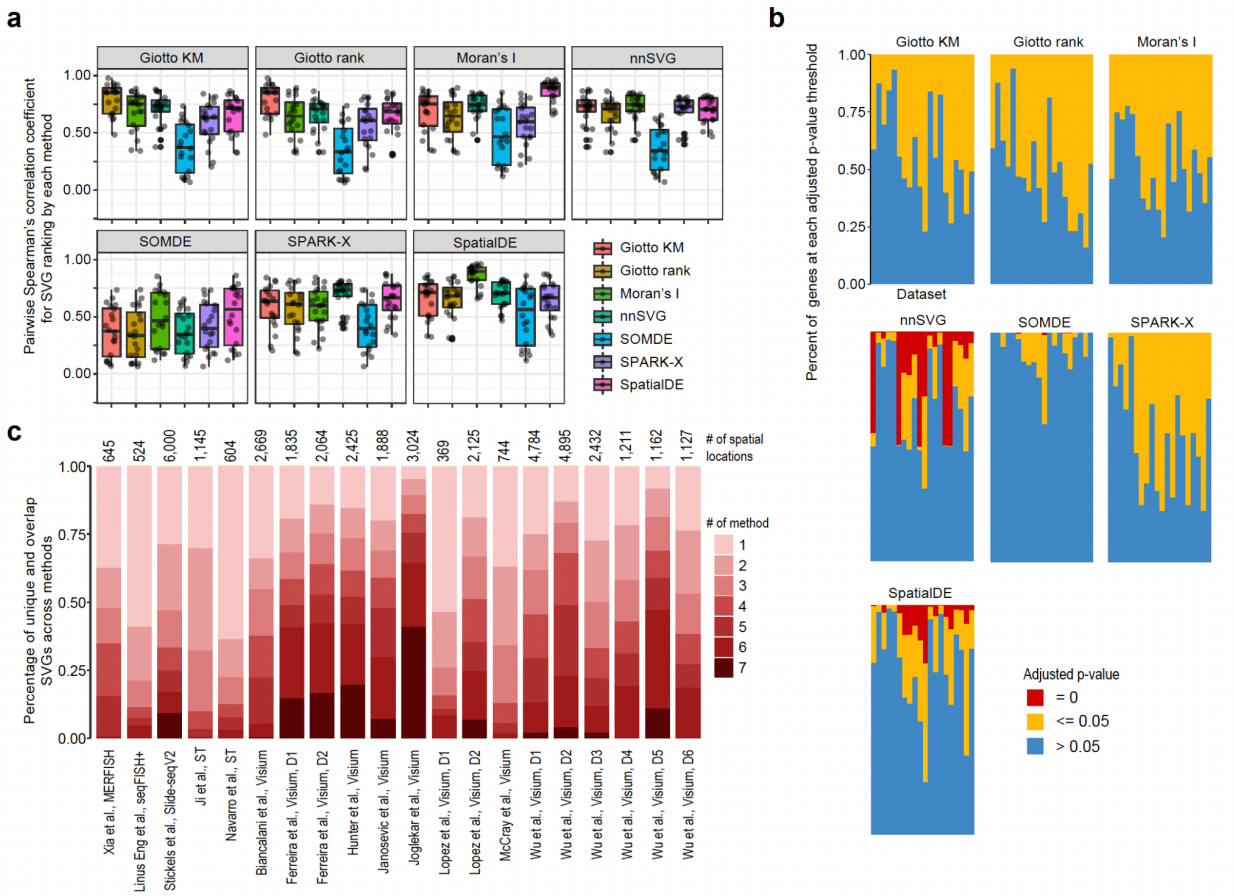

Figure 2. Concordance, statistical significance, and overlap of SVGs detected by different methods. (**a**) Concordance of SVG rankings reported from each of the seven SVG detection methods. Each panel uses one SVG detection as an anchor and the y-axes are pairwise Spearman's correlation coefficient for quantifying concordance in the ranking of each pair of SVG detection methods. Points in each boxplot represent the result from a dataset. Boxplot centre line, median; box limits, upper and lower quartiles; whiskers, 1.5 times the interquartile range. (**b**) Statistical significance of SVGs detected by each method. SVGs are partitioned into three categories based on the adjusted p-values reported by each method (i.e., p = 0; 0 < p <= 0.05; p > 0.05) and presented as a percentage (y-axis). (**c**) A proportional bar plot showing the percentage of unique (# of method = 1) and overlapping (# of method > 1) significant SVGs (adjusted p-value <=0.05) reported by the SVG detection methods for each spatial transcriptomics dataset.

While the ranking of SVGs is useful for selecting the top candidates for subsequent analysis, in practice, statistical significance such as *p*-values is frequently used for selecting SVGs. To this end, we first partitioned the SVGs into three categories (i.e., $p = 0$; $0 < p <= 0.05$; $p > 0.05$) based on the adjusted *p*-value reported from each computational method (Figure 2b). We found that most methods report a large proportion of SVGs at an adjusted *p*-value threshold of 0.05 on a large number of datasets. Among the seven methods, nnSVG, and to a lesser degree SpatialDE, reported a sizable proportion of SVGs with an adjusted *p*-value of 0. Interestingly, SOMDE reported on average the fewest number of significant SVGs with some datasets having

almost no significant SVGs (Figure 2b and Figure S3a). Since the adjusted *p*-value of 0.05 is a commonly used significance cutoff, we next used this adjusted *p*-value threshold reported from each method to find statistically significant SVGs (Figure S3b). Remarkably, the numbers of SVGs found by all methods are strikingly low with many datasets having close to no overlapping SVGs across all seven computational methods (Figure 2c and Figure S4). In addition, a large number of unique genes was found by various individual methods in most datasets (Figure 2c). Together, these findings highlight the discrepancy among methods when an adjusted *p*-value threshold of 0.05 was used for calling statistically significant SVGs.

Dependency of SVG statistics on gene expression levels and across genes

In scRNA-seq data, it is known that variance in gene expression is positively correlated with gene expression level; therefore, most highly variable gene (HVG) detection methods implement procedures to account for this bias (Yip et al., 2019). To test whether methods designed for SVG detection have the tendency to select genes with higher expression levels, we investigated the correlation between mean gene expression and the SVG statistics for each method and dataset pair. We found that indeed the rankings of SVGs from most methods correlated positively with the mean gene expression (Figure 3a, b). In particular, results from nnSVG and SPARK-X showed average correlations around 0.8 across the datasets (Figure 3c), suggesting a high dependency of SVG ranking on gene expression for these two methods. We also correlated the proportion of zeros in gene expression across cells against SVG ranking for each method (Figure S5a, b). Since the proportion of zeros is known to be negatively correlated with their expression levels, the negative correlation observed among each method and dataset pair further confirms the dependency we found between SVG ranking and gene expression among current SVG detection methods.

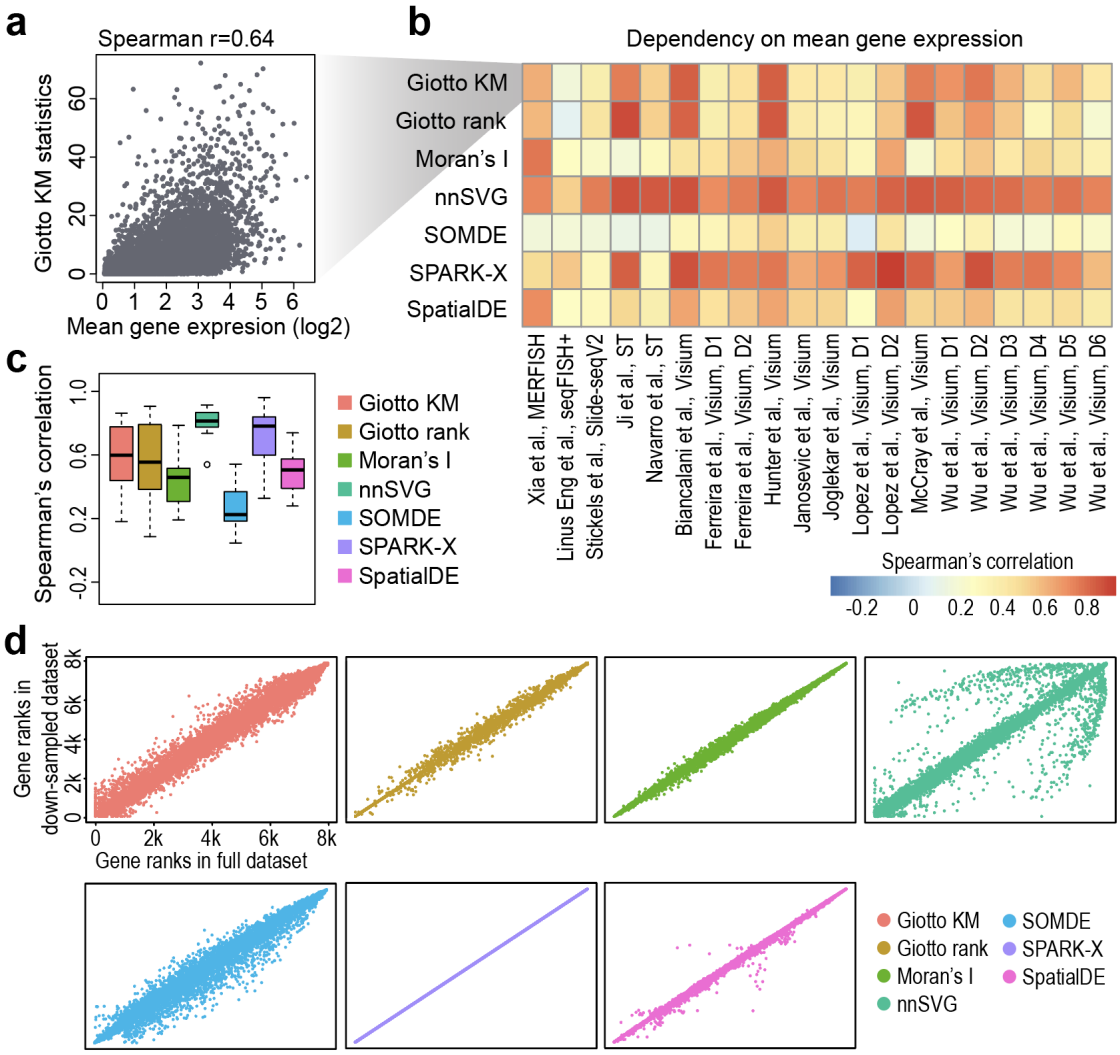

Figure 3. Dependency of SVG statistics on gene expression levels and the interdependency of genes in SVG detection tools. (**a**) An example showing the positive correlation between SVG statistics reported from Giotto KM and mean gene expression across cells in the "Xia et al., MERFISH" dataset (Xia et al., 2019). (**b**) Heatmap summarising the Spearman's correlation of SVG statistics reported by each method and the mean gene expression in each dataset. (**c**) Boxplot of Spearman's correlation of SVG statistics reported by each method and the mean gene expression across the 20 spatial transcriptomics datasets. (**d**) Reproducibility of gene ranks in the down-sampled dataset from "Wu et al. Visium D2" (Wu et al., 2021) and the original full dataset. For the down-sampling, 50% of the genes were down-sampled whilst keeping the number of spots equal.

We next assessed the reproducibility of gene ranks based on the SVG statistics reported from each method when the number of genes included in a dataset changes. To this end, we randomly down-sampled the genes in the dataset Wu et al. Visium D2 (Wu et al., 2021) (Figure S1a) to 50% and re-calculated the ranks of genes from the reported SVG statistics of each method on the down-sampled datasets. Except for SPARK-X, we found that the other six

methods do not completely maintain the exact rank of genes and therefore are not independent of calculating SVG statistics of a gene when the genes included in a dataset changes (Figure 3d). In particular, nnSVG shows the highest level of variability, especially at the top-ranked SVGs. SOMDE and Giotto KM also showed some levels of variability albeit to a lesser degree. The variability in gene ranks from Giotto rank, SpatialDE, and Moran's I implemented in the Seurat package is relatively smaller and may not have a significant impact on downstream analysis. Nevertheless, these analyses reveal that the decisions made on gene filtering, a common step in data preprocessing, may result in a change in SVG statistics and their ranking for most of the SVG detection methods.

Accuracy in detecting SVGs based on simulation data

To test the accuracy of SVG detention methods, we next simulated spatial transcriptomics datasets with known SVGs (Figure S6) and calculated the true positive rate (TPR) and the false discovery rate (FDR) using six adjusted *p*-value thresholds (1e-100, 1e-50, 1e-10, 0.01, 0.05, and 0.1) for each simulated dataset (see Methods for details). At the adjusted *p*-value thresholds of 0.01, 0.05, and 0.1, we found that SPARK-X and SpatialDE performed well with both high TPR and low FDR (Figure 4 and Figure S7). As expected, their TPR dropped when lower adjusted *p*-value thresholds were applied. SOMDE performed well in terms of maintaining a low FDR in its identification but failed to detect many true SVGs and hence demonstrated a lower TPR. Under adjusted *p*-value thresholds of 0.01, 0.05, and 0.1, Giotto rank, Moran's I, and nnSVG all demonstrated a high TPR but suffered from a high level of false positive identification. Compared to other methods, Giotto KM performed relatively poorly in the simulation with lower TPR and higher FDR in most cases (Figure 4a). We further demonstrated this using exaggerated p-value thresholds that the FDR could be improved for methods such as Giotto KM for some of the simulated datasets, confirming the poor FDR control among these methods (Figure 4b). These findings reveal that for methods, except SPARK-X, SpatialDE, and SOMDE, the estimated FDRs (i.e. adjusted *p*-value thresholds) do not accurately represent the true FDRs for SVG detection in these simulated datasets.

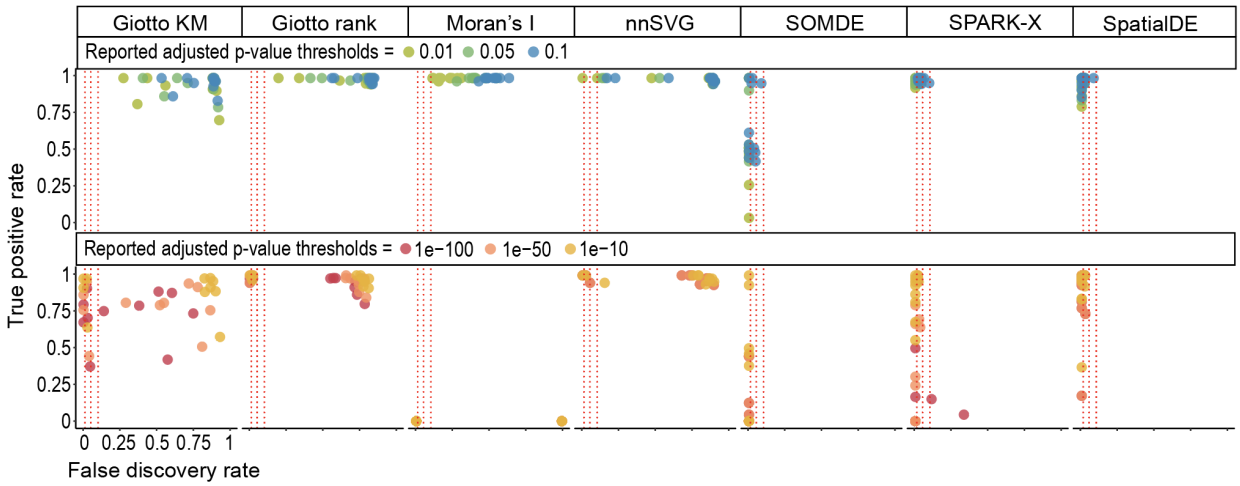

Figure 4. Spatially variable gene detection performance across 11 simulated datasets.

Scatter plot of observed true positive rate (y-axis) and false discovery rate (x-axis) of spatially variable gene detection by the benchmarked tools at six adjusted p-value cutoffs. Each dot is colour-coded by the cutoff used. The three horizontal lines represent the true FDRs of 0.01, 0.05, and 0.1.

Performance on clustering spatial domains

A key task in spatial transcriptomics data analysis is to identify spatial domains that mark distinctive cell and tissue types in a biological sample. One approach to achieve this is to cluster profiled locations into spatial domains using SVGs. To test the utility of SVGs in clustering spatial domains, we took advantage of the spatial transcriptomics data of an E9.5 mouse embryo given the availability of tissue annotations in these samples (Figure 5a) and performed SVG identification using each method. We found that most methods called a large number of significant SVGs (Figure 5b). Using a *k*-means clustering algorithm on all significant SVGs, we found that most SVG methods led to a relatively low concordance with the original tissue annotations (Figure 5c). In light of this, we next selected either the top-500 (Figure 5d) or top-100 (Figure 5e) SVGs ranked by each method and performed the same clustering procedure using these selected SVGs. This led to an overall higher concordance with the tissue annotations for all methods and a better distinction of the different methods in terms of the utilities of the top SVGs, especially when the top-100 SVGs were used. Specifically, we found that SVGs selected by SPARK-X and Giotto rank generally led to a higher clustering concordance whereas nnSVG showed a relatively lower concordance with respect to the tissue annotations based on our four evaluation metrics. These results demonstrate that including all significant SVGs based on the adjusted *p*-value of 0.05 reported by an SVG detection method may not be the most useful approach for certain downstream analyses such as clustering spatial domains in a spatial transcriptomics dataset. Instead, selecting top SVGs based on their ranking may lead to better results for such analyses.

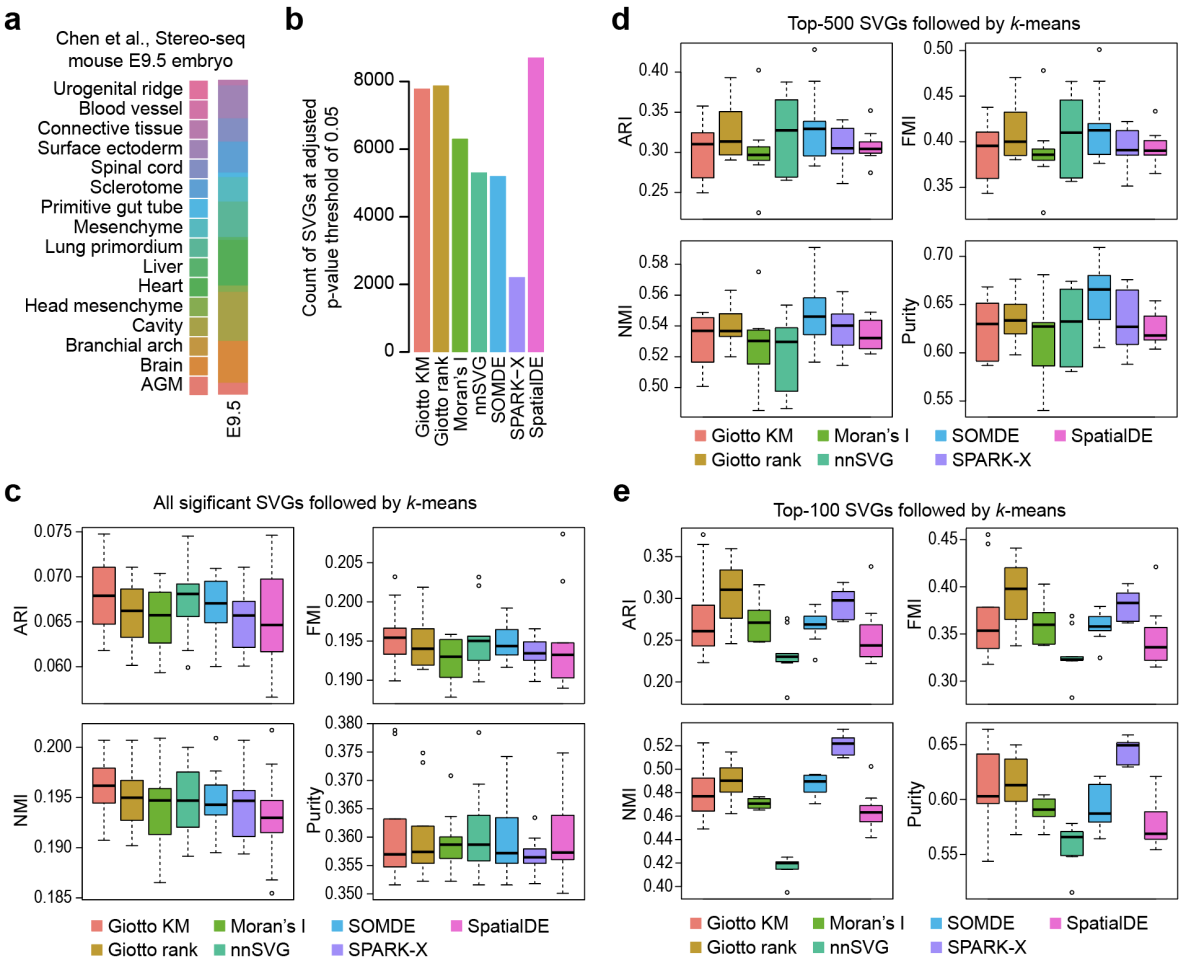

Figure 5. Performance of SVGs selected by each method for clustering spatial domains in mouse embryos. (**a**) Proportion spatial locations annotated to one of sixteen tissue domains in the E9.5 mouse embryo. (**b**) The number of statistically significant SVGs at the adjusted *p*-value threshold of 0.05 reported by each method. Concordance of clustering outputs and pre-defined spatial domains in the mouse embryo using (**c**) all significant SVGs, (**d**) top-500 and (**e**) top-100 SVGs selected by each method. Concordance of clustering outputs and pre-defined spatial domains are quantified by four metrics including ARI, FMI, NMI, and Purity.

Computational time and memory usage

Computational time and memory usage are key considerations in practical applications, especially for large spatial transcriptomics data analyses. In our evaluation, we configured a standard virtual machine, with 16 CPUs and 236 GB of memory and recorded the runtime and the peak memory usage for each SVG detection method on each dataset (Figure 6a, b). As expected, we found the computational time and the peak memory usage are both positively correlated with the number of spatial locations in the datasets. In terms of computational time, comparison across methods revealed that SPARK-X is the fastest method and scales extremely

well with the number of spatial locations. While SOMDE is the second best in most cases, it is significantly slower compared to SPARK-X. In contrast, SpatialDE performed poorer especially on datasets with large numbers of spatial locations. Giotto KM performed poorly in most of the datasets but does scale better than SpatialDE with the number of spatial locations in datasets. Similarly, nnSGV scaled better with the number of spatial locations than SpatialDE but was slower on datasets with a smaller number of spatial locations. In terms of peak memory usage, we found that SOMDE uses the least peak memory across all datasets and SPARK-X ranked the second in most cases although it has a significantly higher peak memory usage. In comparison, the two methods implemented in Giotto and SpatialDE show high peak memory usage, especially in datasets with a large number of spatial locations. While there is a trade-off between speed and memory usage, taken together, these results suggest that SPARK-X and SOMDE are the most efficient methods in terms of speed and memory usage for SVG detection.

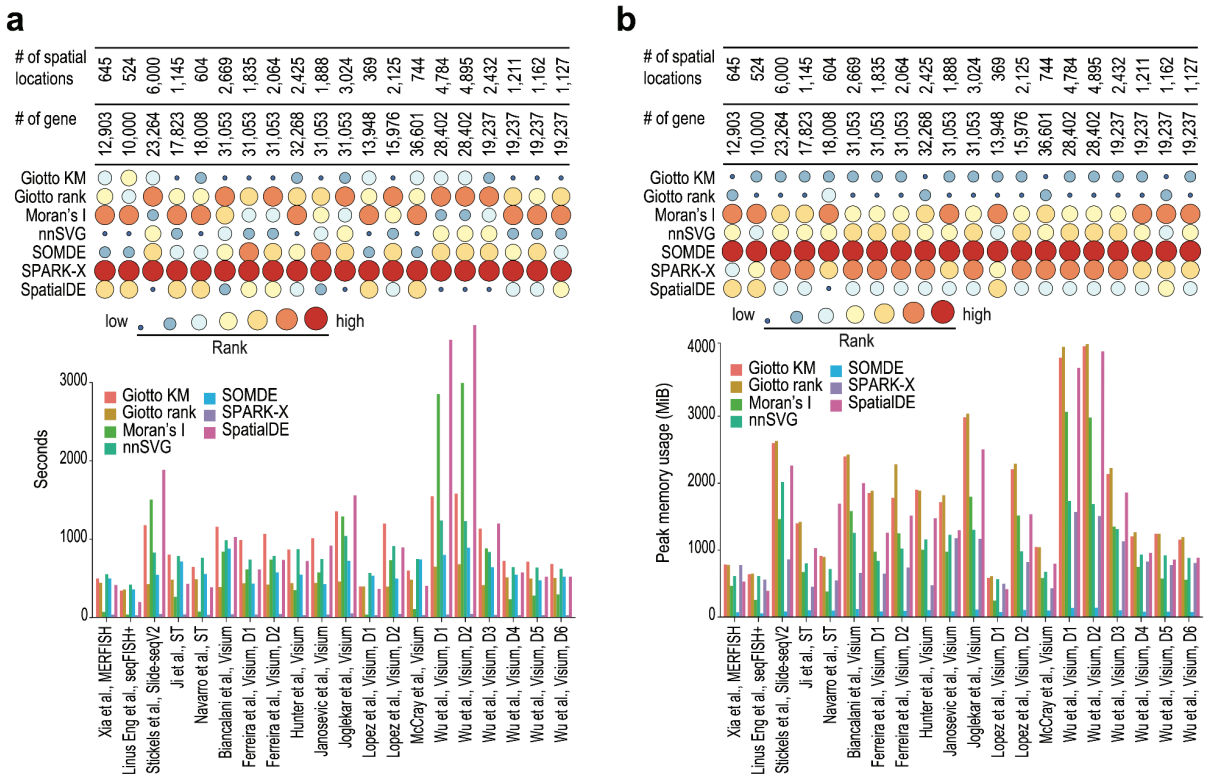

Figure 6. Evaluation of computational time and peak memory usage of SVG detection methods. (**a**) Computation time in units of second and (**b**) peak memory usage in units of MiB across datasets.

DISCUSSION

We found that, for most methods, a significant proportion of genes were detected as SVGs under the adjusted *p*-value of 0.05 in the majority of the tested datasets (Figure 2b and Figure S3). However, the overlaps across the seven methods are relatively small considering the large numbers of SVGs identified from each SVG detection method (Figure 2c), suggesting large

discrepancies among SVG detection methods when a significance cutoff was used to filter for SVGs. Consistent with this, in our simulation study where ground truth SVGs were introduced into simulated spatial transcriptomics data, we found that in most cases the estimated FDRs do not accurately represent the true FDRs (Figure 4). Finally, we found the use of all significant SVGs for spatial domain clustering performed significantly poorly compared to using a small set of the top-ranked SVGs (e.g. 100 and 500). These results highlight that the estimation of statistical significance is difficult and there is much room for improvement. It also cautions the use of and the reliance on such statistical significance from current SVG detection tools for drawing data and biological conclusions.

We also discovered that SVGs identified by most methods show a strong positive correlation with their expression levels (Figure 3). We note that a similar relationship was found between gene variability and expression level in scRNA-seq data and most computational methods designed for HVG detection actively correct for such a 'bias' (Yip et al., 2019). While we could not rule out the possibility that genes that vary spatially are also highly expressed, future work should be performed to investigate the biological basis and plausibility for such a correlation. During practical application, it is important to be aware of the tendency of current SVG detection tools to select genes with high expression levels. Future method development will be required to account for this effect such as to retain relatively lowly expressed genes such as transcription factors in downstream analysis. In addition, we found that for most methods the relative rankings of SVGs change when different pools of genes are included in the datasets (Figure 3d). While taking into account the interdependency among genes may provide useful information for identifying SVGs, it is important to be aware that different SVG detection results may be obtained when different pre-processing steps were used to filter genes prior to SVG analysis.

Lastly, SVG detection can be viewed as a feature selection step in spatial transcriptomics data analysis, where useful features (i.e., SVGs) are selected and/or uninformative ones are removed. In particular, the current SVG detection methods can be considered as unsupervised approaches where no information such as cell types, cell states, or spatial domains are required. A great amount of work has been done in feature selection in single-cell data analysis (Yang et al., 2021), including unsupervised methods and also more advanced methods that perform combinatorial feature selection using supervised learning such as embedded feature selection using random forest and wrapper feature selection using genetic algorithms. We anticipate that future development of SVG detection methods will explore the utility of information such as cell types and states to identify SVGs that not only independently mark the spatial variability but also those that cooperate across multiple genes and together define spatial variability. We believe these developments will introduce additional computational new challenges but will undoubtedly lead to new biological insight from spatial transcriptomics data analyses.

METHODS

SVG detection methods

Datasets were preprocessed as previously described (Li et al., 2022). In addition, each dataset was filtered by first removing cells whose top-50 highly expressed genes contributed to 50% of the total counts and then genes that were expressed in fewer than 30 cells. Log normalisation of raw counts was performed prior to SVG detection as per the recommended default for each method. The same reproducible seed was set prior to running each method.

Giotto KM and Giotto rank. Giotto (Dries et al., 2021a) requires a spatial nearest neighbour network to be built on reduced dimensions to represent the spatial relationships. Then, statistical enrichment using Fisher's exact test of binarized expression in spatial nearest neighbours is performed to determine SVGs. The two methods differ in their binarization method. In *Giotto KM*, expression values for each gene are binarized using k-means clustering (k=2), otherwise simple thresholding on rank is applied in *Giotto rank* (default = 30%). Thus, a gene is considered an SVG if it is highly expressed in neighbouring cells. Normalisation was performed using *normalizeGiotto()* under default parameters. SVG detection was thus performed with two different approaches k-means and rank using *binSpect(bin_method="kmeans")* and *binSpect(bin_method="rank")* respectively, following the author's tutorial. https://rubd.github.io/Giotto_site/articles/mouse_visium_kidney_200916.html

Moran's I. Moran's I ranks genes by the observed spatial autocorrelation (Gittleman and Kot, 1990; Moran, 1950) to measure the dependence of a feature on spatial location. Weights are calculated as 1/distance. Raw counts were first normalised using *SCTransform()*. SVGs were detected using *FindSpatiallyVariableFeatures(selection.method = "moransi")* and statistics for all features were returned.

nnSVG. nnSVG is based on scalable estimation of spatial covariant functions in Gaussian process regression using nearest neighbour Gaussian process (NNGP) models. Assuming an exponential covariance function, a separate model is fitted to the expression values for each gene to obtain maximum likelihood estimates. A likelihood-ratio test is performed to rank genes by estimated LR statistic values. Log normalisation was performed using *scater::LogNormCounts* prior to running *nnSVG()* with default parameters.

SOMDE. A SOM neural network is used to adaptively integrate nearest neighbour data into different nodes, achieving a condensed representation of the spatial transcriptome. Then SVGs are identified on a node-level, using spatial location and gene meta-expression information. A squared exponential Gaussian kernel is applied to generate log-likelihood ratio values wherein a likelihood-ratio test is performed to rank genes by estimated LR statistic values. Log normalisation was performed using *scater::LogNormCounts()*. The procedure was performed as per the recommended tutorial at https://github.com/WhirlFirst/somde using python.

SPARK-X. SPARK-X is a non-parametric method that relies on a robust covariance test framework, including the Hilber-Schmidt independence criteria test and the distance covariance

matrix test. A test statistic is observed by measuring the similarity between two relationship matrices based on gene expression and spatial coordinates respectively. A p-value is computed for each distance covariance matrix constructed and a Cauchy combined p-value is reported. Log normalisation was performed using *scater::LogNormCounts()* and *sparkx()* was run under default parameters.

SpatialDE. SpatialDE fits a linear mixed model for each gene with gaussian kernels and decomposes the gene variation into spatial or non-spatial variation. The non-spatial variation is separately modeled using observed noise, and the spatial variation is explained by an exponential covariance function. For each gaussian kernel, a p-value is calculated from the likelihood test to rank genes by estimated LR statistics. Log normalisation was performed using *scater::LogNormCounts()*. SpatialDE was run under the R implementation using a wrapper package by Corso et al. (Davide Corso, Milan Malfait, Lambda Moses) and the procedure was as follows in the tutorial by the authors.
https://bioconductor.org/packages/release/bioc/vignettes/spatialDE/inst/doc/spatialDE.html

Dependency across genes

To assess the dependency across genes in SVG analysis, we randomly down-sampled 50% of genes in the "Wu et al. Visium D2" dataset (Wu et al., 2021) (Supplementary Figure 1a). We next applied each SVG detection method and calculated SVG statistics of remaining genes in the down-sampled dataset. The relative rank of these genes were compared with their rank in the original full dataset to assess if there is any change of relative ranking when other genes were included in the dataset. Methods that lead to a different ranking of SVGs in the down-sampled dataset when additional genes were included are considered as calculating spatial variability of a gene depending on the presence and absence of other genes.

Simulation of spatial transcriptomics data

To evaluate the capacity of methods to detect SVGs with high sensitivity and specificity, we simulated a set of spatial transcriptomics data, providing us with ground truth spatially variable genes. We generated the simulations across two steps. First, we simulated masks by using five spatial coordinates from an existing spatial transcriptomics dataset as templates (Chen et al., 2022). These masks contain approximately 4300 to 5900 spatial spots. Using the original annotations as a template, we randomly assigned these domains into two distinct spatially variable domains (Supplementary Fig 6a). Next, we simulated the single-cell transcriptomics data using the muscat package (Crowell et al., 2020). We employed the simData function in the muscat package (which is originally designed to simulate single-cell data consisting of multiple samples, conditions, and cell types) by setting parameters 'nk' and 'ns' as 1. To generate spatial transcriptomes of the two distinct spatial domains, we set the log fold-change in differential expression as 10. These spatial transcriptomes were randomly assigned to spots of their respective domains. We simulated a total of 500 genes among which between 40 and 50 were spatially variable.

Benchmarking of simulation studies

To evaluate the performance of the SVG detection methods on the simulated data, we calculated the receiver operating characteristic curve based on the statistics or *p*-values of the genes, indicating the capacity of methods to rank truly spatially variable genes before non-variable ones. We next calculated the true positive rate (TPR) and the false discovery rate to evaluate FDR control at six adjusted *p*-value thresholds (1e-100, 1e-50, 1e-10, 0.01, 0.05, and 0.1) for each simulated dataset. The cutpointr package (Thiele and Hirschfeld, 2021) was used to calculate the TPR and FDR performance metrics.

Clustering and concordance quantification

To quantify the utility of SVGs in spatial domain clustering, we used all significant SVGs or top-500 and top-100 SVGs reported from each method and a *k*-means clustering algorithm to cluster the E9.5 mouse embryo spatial transcriptomics data into 16 tissue domains based on the original annotation (Chen et al., 2022). The concordance of the clustering output from using different sets of SVGs and the original annotation was quantified using adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows index (FMI), and purity.

Time consumption and memory usage

For all methods run in R, the elapsed time to run each method was evaluated using the *system.time()* function. The peak memory usage was monitored using *gc()*. For methods run in python, *perf_counter()* from the *time* package was used to record the elapsed time. To record the peak memory usage, *get_traced_memory()* was used from the *tracemalloc* package.

List of datasets

Human osteosarcoma	(Xia et al., 2019)	https://www.pnas.org/doi/suppl/10.1073/pnas.1912459116/suppl_file/pnas.1912459116.sd12.csv
Mouse primary visual cortex	(Eng et al., 2019)	https://github.com/CaiGroup/seqFISH-PLUS
Mouse hippocampus	(Stickels et al., 2021)	https://singlecell.broadinstitute.org/single_cell/study/SCP815/highly-sensitive-spatial-transcriptomics-at-near-cellular-resolution-with-slide-seqv2#study-download
Human squamous carcinoma	(Ji et al., 2020)	GSM4284322
Mouse hippocampus	(Navarro et al., 2020)	https://data.mendeley.com/datasets/6s959w2zyr/1

Mouse primary motor cortex	(Biancalani et al., 2021)	https://storage.googleapis.com/tommaso-brain-data/tangram_demo/Allen-Visium_Allen1_cell_count.h5ad
Mouse kidney	(Ferreira et al., 2021)	GSE171406
Zebrafish melanoma	(Hunter et al., 2021)	GSE159709
Mouse kidney	(Janosevic et al., 2021)	GSE154107
Mouse pre-frontal cortex	(Joglekar et al., 2021)	GSE158450
Mouse lymph node	(Lopez et al., 2022)	GSE173776
Mouse MCA205 tumour	(Lopez et al., 2022)	GSE173773
Human prostate	(McCray et al., 2021)	GSM4837767
Human breast cancer	(Wu et al., 2021)	https://zenodo.org/record/4739739#.YY6N_pMzaWC
Mouse E9.5 embryo	(Chen et al., 2022)	https://db.cngb.org/stomics/mosta/

List of software

R version 4.1.0	R Development Core Team, 2016	https://www.R-project.org/
Giotto 1.1.1	(Dries et al., 2021b)	https://github.com/RubD/Giotto
Seurat 4.1.1	(Hao et al., 2021b)	https://github.com/satijalab/seurat

nnSVG 1.1.7	(Weber et al., 2022)	https://github.com/lmweber/nnSVG
SOMDE 0.1.8	(Hao et al., 2021a)	https://pypi.org/project/somde/
SPARK 1.1.1	(Zhu et al., 2021)	https://github.com/xzhoulab/SPARK
SpatialDE 1.2.0	(Davide Corso, Milan Malfait, Lambda Moses)	https://github.com/sales-lab/spatialDE
SpatialDE 1.1.3	(Svensson et al., 2018b)	https://pypi.org/project/SpatialDE/
Scater 1.24.0	(McCarthy et al., 2017)	https://github.com/davismcc/scater
muscat 1.10.1	(Crowell et al., 2020)	https://github.com/HelenaLC/muscat
cutpointr 1.1.2	(Thiele and Hirschfeld, 2021)	https://github.com/thie1e/cutpointr
time 3.8.13	(Rossum and Drake, 2010)	https://www.python.org/
tracemalloc 0.9.1	https://pypi.org/	https://github.com/vstinner/pytracemalloc

ACKNOWLEDGEMENTS

We thank our colleagues at the School of Mathematics and Statistics, The University of Sydney, and the Sydney Precision Bioinformatics Alliance for their feedback. This work is supported by a National Health and Medical Research Council (NHMRC) Investigator Grant (1173469) and a Metcalf Prize to P.Y.; and a postgraduate scholarship from Research Training Program and a Children's Medical Research Institute postgraduate scholarship to C.C.

AUTHOR CONTRIBUTIONS

C.C., H.J.K., and P.Y. conceived the study. C.C. and H.J.K. performed data analyses and interpreted the results with input from P.Y. All authors wrote the manuscript and approved the final version of the manuscript.

REFERENCES

Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C.R., Segerstolpe, Å., Zhang, M., et al. (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nat Methods *18*, 1352–1362. https://doi.org/10.1038/s41592-021-01264-7.

Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., et al. (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell *185*, 1777-1792.e21. https://doi.org/10.1016/j.cell.2022.04.003.

Crowell, H.L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., and Robinson, M.D. (2020). muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nat Commun *11*, 6077. https://doi.org/10.1038/s41467-020-19894-4.

Davide Corso, Milan Malfait, Lambda Moses spatialDE (Bioconductor).

Dries, R., Zhu, Q., Dong, R., Eng, C.-H.L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F., et al. (2021a). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biol *22*, 78. https://doi.org/10.1186/s13059-021-02286-2.

Dries, R., Zhu, Q., Dong, R., Eng, C.-H.L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F., et al. (2021b). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biology *22*, 78. https://doi.org/10.1186/s13059-021-02286-2.

Eng, C.-H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature *568*, 235–239. https://doi.org/10.1038/s41586-019-1049-y.

Ferreira, R.M., Sabo, A.R., Winfree, S., Collins, K.S., Janosevic, D., Gulbronson, C.J., Cheng, Y.-H., Casbon, L., Barwinska, D., Ferkowicz, M.J., et al. (2021). Integration of spatial and single-cell transcriptomics localizes epithelial cell–immune cross-talk in kidney injury. JCI Insight *6*, e147703. https://doi.org/10.1172/jci.insight.147703.

Gittleman, J.L., and Kot, M. (1990). Adaptation: Statistics and a Null Model for Estimating Phylogenetic Effects. Systematic Zoology *39*, 227. https://doi.org/10.2307/2992183.

Hao, M., Hua, K., and Zhang, X. (2021a). SOMDE: A scalable method for identifying spatially variable genes with self-organizing map. Bioinformatics btab471. https://doi.org/10.1093/bioinformatics/btab471.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021b). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Hunter, M.V., Moncada, R., Weiss, J.M., Yanai, I., and White, R.M. (2021). Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. Nat Commun *12*, 6278. https://doi.org/10.1038/s41467-021-26614-z.

Janosevic, D., Myslinski, J., McCarthy, T.W., Zollman, A., Syed, F., Xuei, X., Gao, H., Liu, Y.-L., Collins, K.S., Cheng, Y.-H., et al. (2021). The orchestrated cellular and molecular responses of the kidney to endotoxin define a precise sepsis timeline. ELife *10*, e62270. https://doi.org/10.7554/eLife.62270.

Ji, A.L., Rubin, A.J., Thrane, K., Jiang, S., Reynolds, D.L., Meyers, R.M., Guo, M.G., George, B.M., Mollbrink, A., Bergenstråhle, J., et al. (2020). Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell *182*, 497-514.e22. https://doi.org/10.1016/j.cell.2020.05.039.

Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A.K., Marrocco, J., Williams, S.R., Haase, B., Hayes, A., et al. (2021). A spatially resolved brain region- and cell

type-specific isoform atlas of the postnatal mouse brain. Nat Commun *12*, 463. https://doi.org/10.1038/s41467-020-20343-5.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Molecular Cell *58*, 610–620. https://doi.org/10.1016/j.molcel.2015.04.005.

Li, B., Zhang, W., Guo, C., Xu, H., Li, L., Fang, M., Hu, Y., Zhang, X., Yao, X., Tang, M., et al. (2022). Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat Methods *19*, 662–670. https://doi.org/10.1038/s41592-022-01480-9.

Lopez, R., Li, B., Keren-Shaul, H., Boyeau, P., Kedmi, M., Pilzer, D., Jelinski, A., Yofe, I., David, E., Wagner, A., et al. (2022). DestVI identifies continuums of cell types in spatial transcriptomics data. Nat Biotechnol *40*, 1360–1369. https://doi.org/10.1038/s41587-022-01272-8.

McCarthy, D.J., Campbell, K.R., Lun, A.T.L., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics *33*, 1179–1186. https://doi.org/10.1093/bioinformatics/btw777.

McCray, T., Pacheco, J.V., Loitz, C.C., Garcia, J., Baumann, B., Schlicht, M.J., Valyi-Nagy, K., Abern, M.R., and Nonn, L. (2021). Vitamin D sufficiency enhances differentiation of patient-derived prostate epithelial organoids. IScience *24*, 101974. https://doi.org/10.1016/j.isci.2020.101974.

Moran, P.A.P. (1950). NOTES ON CONTINUOUS STOCHASTIC PHENOMENA. Biometrika *37*, 17–23. https://doi.org/10.1093/biomet/37.1-2.17.

Navarro, J.F., Croteau, D.L., Jurek, A., Andrusivova, Z., Yang, B., Wang, Y., Ogedegbe, B., Riaz, T., Støen, M., Desler, C., et al. (2020). Spatial Transcriptomics Reveals Genes Associated with Dysregulated Mitochondrial Functions and Stress Signaling in Alzheimer Disease. IScience *23*, 101556. https://doi.org/10.1016/j.isci.2020.101556.

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467. https://doi.org/10.1126/science.aaw1219.

Rossum, G. van, and Drake, F.L. (2010). The Python language reference (Hampton, NH: Python Software Foundation).

Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., Arlotta, P., Macosko, E.Z., and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat Biotechnol *39*, 313–319. https://doi.org/10.1038/s41587-020-0739-1.

Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods *17*, 193–200. https://doi.org/10.1038/s41592-019-0701-7.

Svensson, V., Teichmann, S.A., and Stegle, O. (2018a). SpatialDE: identification of spatially variable genes. Nat Methods *15*, 343–346. https://doi.org/10.1038/nmeth.4636.

Svensson, V., Teichmann, S.A., and Stegle, O. (2018b). SpatialDE: identification of spatially variable genes. Nat Methods *15*, 343–346. https://doi.org/10.1038/nmeth.4636.

Thiele, C., and Hirschfeld, G. (2021). **cutpointr** : Improved Estimation and Validation of Optimal Cutpoints in *R*. J. Stat. Soft. *98*. https://doi.org/10.18637/jss.v098.i11.

Weber, L.M., Saha, A., Datta, A., Hansen, K.D., and Hicks, S.C. (2022). nnSVG: scalable identification of spatially variable genes using nearest-neighbor Gaussian processes (Genomics).

Wu, S.Z., Al-Eryani, G., Roden, D.L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A.,

Wang, C., Torpy, J.R., Bartonicek, N., et al. (2021). A single-cell and spatially resolved atlas of human breast cancers. Nat Genet *53*, 1334–1347. https://doi.org/10.1038/s41588-021-00911-1.

Xia, C., Fan, J., Emanuel, G., Hao, J., and Zhuang, X. (2019). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proc. Natl. Acad. Sci. U.S.A. *116*, 19490–19499. https://doi.org/10.1073/pnas.1912459116.

Yang, P., Huang, H., and Liu, C. (2021). Feature selection revisited in the single-cell era. Genome Biol *22*, 321. https://doi.org/10.1186/s13059-021-02544-3.

Yip, S.H., Sham, P.C., and Wang, J. (2019). Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. Brief Bioinform *20*, 1583–1589. https://doi.org/10.1093/bib/bby011.

Zhu, J., Sun, S., and Zhou, X. (2021). SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. Genome Biol *22*, 184. https://doi.org/10.1186/s13059-021-02404-0.

Supplementary information

a

Reference and platform	Species	Tissue	# Location	# Gene	Source
Xia et al., MERFISH	Human	Osteosarcoma	645	12,903	Supplementary file
Linus Eng et al., seqFISH+	Mouse	Cortex	524	10,000	Zenodo#2669683
Stickels et al., Slide-seqV2	Mouse	Hippocampus	6,000	23,264	Broad Institute
Ji et al., ST	Human	Squamous carcinoma	1,145	17,823	GSM4284322
Navarro et al., ST	Mouse	Hippocampus	604	18,008	Mendeley#6s959w2zyr.1
Biancalani et al., Visium	Mouse	Primary motor cortex	2,669	31,053	Google Cloud
Ferreira et al., Visium, D1	Mouse	Kidney	1,835	31,053	GSM5524979
Ferreira et al., Visium, D2	Mouse	Kidney	2,064	31,053	GSM5524979
Hunter et al., Visium	Zebrafish	Melanoma	2,425	32,268	GSE159709
Janosevic et al., Visium	Mouse	Kidney	1,888	31,053	GSE154107
Joglekar et al., Visium	Mouse	Prefrontal cortex	3,024	31,053	GSE158450
Lopez et al., Visium, D1	Mouse	Lymph node	369	13,948	GSE173776
Lopez et al., Visium, D2	Mouse	MCA205 tumour	2,125	15,976	GSE173773
McCray et al., Visium	Human	Prostate	744	36,601	GSM4837767
Wu et al., Visium, D1	Mouse	Breast cancer	4,784	28,402	Zenodo#4739739
Wu et al., Visium, D2	Mouse	Breast cancer	4,895	28,402	Zenodo#4739739
Wu et al., Visium, D3	Mouse	Breast cancer	2,432	19,237	Zenodo#4739739
Wu et al., Visium, D4	Mouse	Breast cancer	1,211	19,237	Zenodo#4739739
Wu et al., Visium, D5	Mouse	Breast cancer	1,162	19,237	Zenodo#4739739
Wu et al., Visium, D6	Mouse	Breast cancer	1,127	19,237	Zenodo#4739739

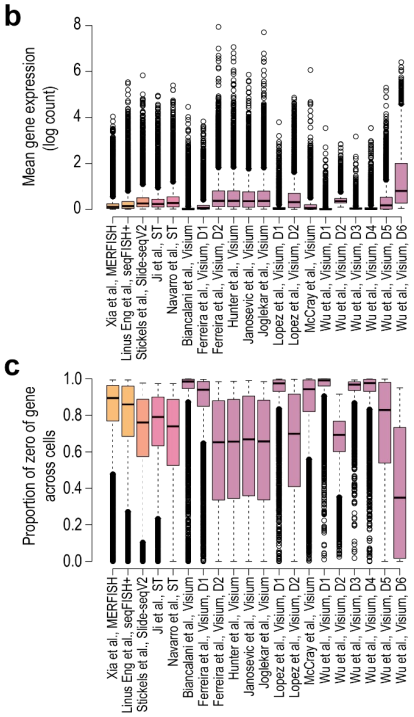

Figure S1. Summary information and statistics of the spatial transcriptomics datasets used for evaluating concordance, statistical significance and reproducibility of SVG detection methods in this study. (**a**) Dataset reference table, summarising basic information and source of each dataset. information. (**b, c**) Boxplot of mean gene expression and proportion of zero of gene across cells in each dataset. Related to Figure 1.

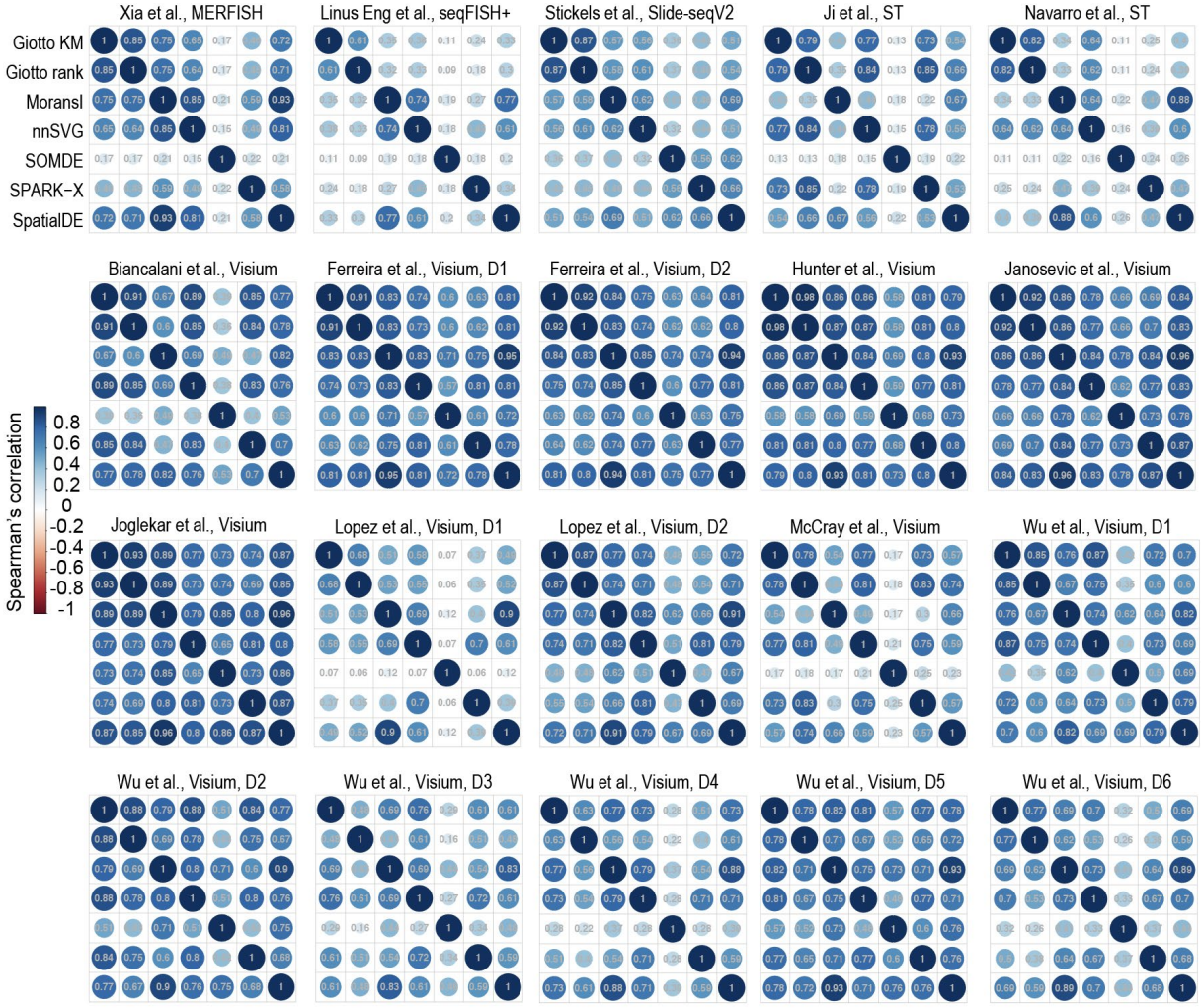

Figure S2. Pairwise correlation of SVG rankings reported by each method for individual spatial transcriptomics datasets. Related to Figure 2.

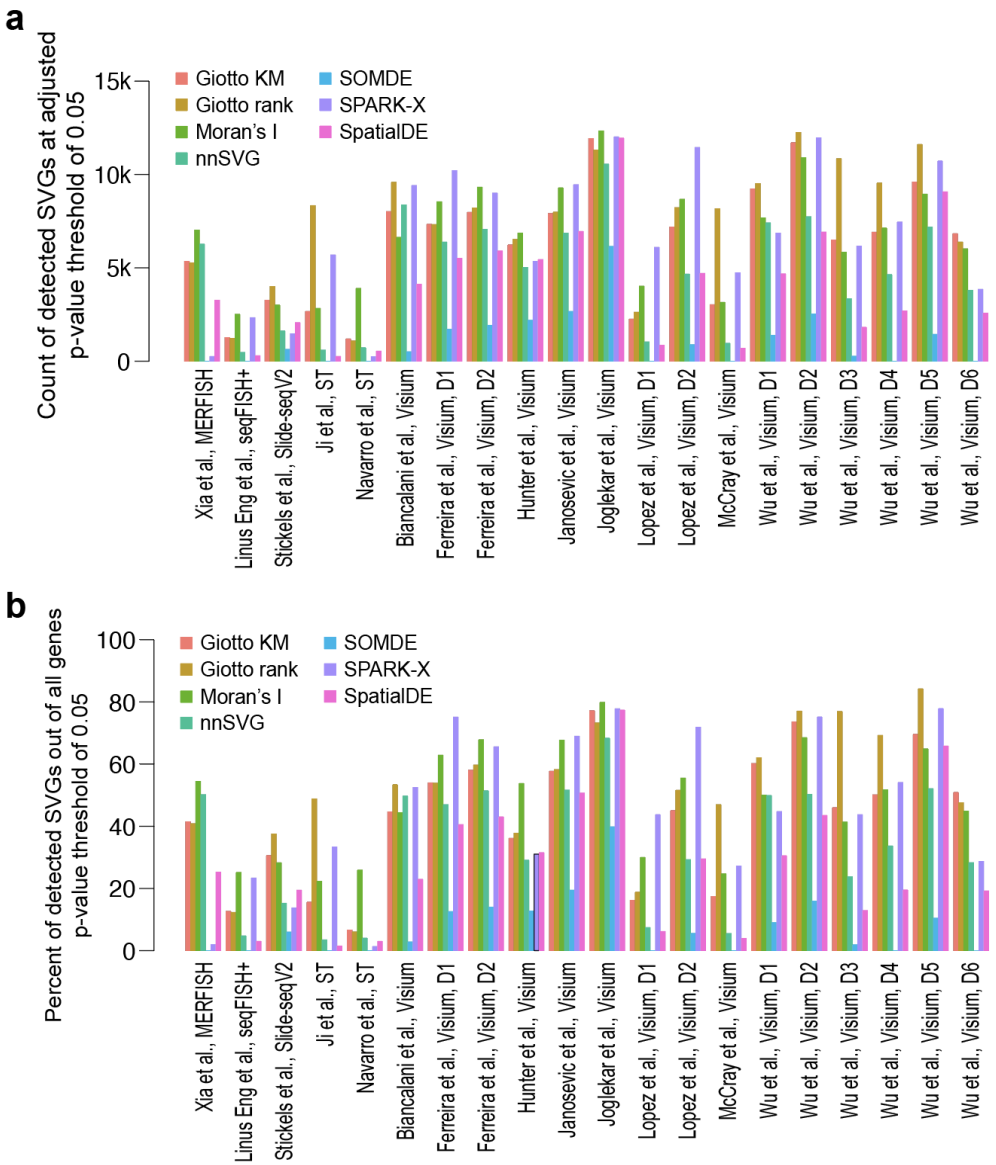

Figure S3. Statistically significant SVGs reported by each method for each spatial transcriptomics dataset. (**a**) Number and (**b**) percentage of statistically significant SVGs under the adjusted p-value threshold of 0.05 reported by each method for each dataset. Related to Figure 2.

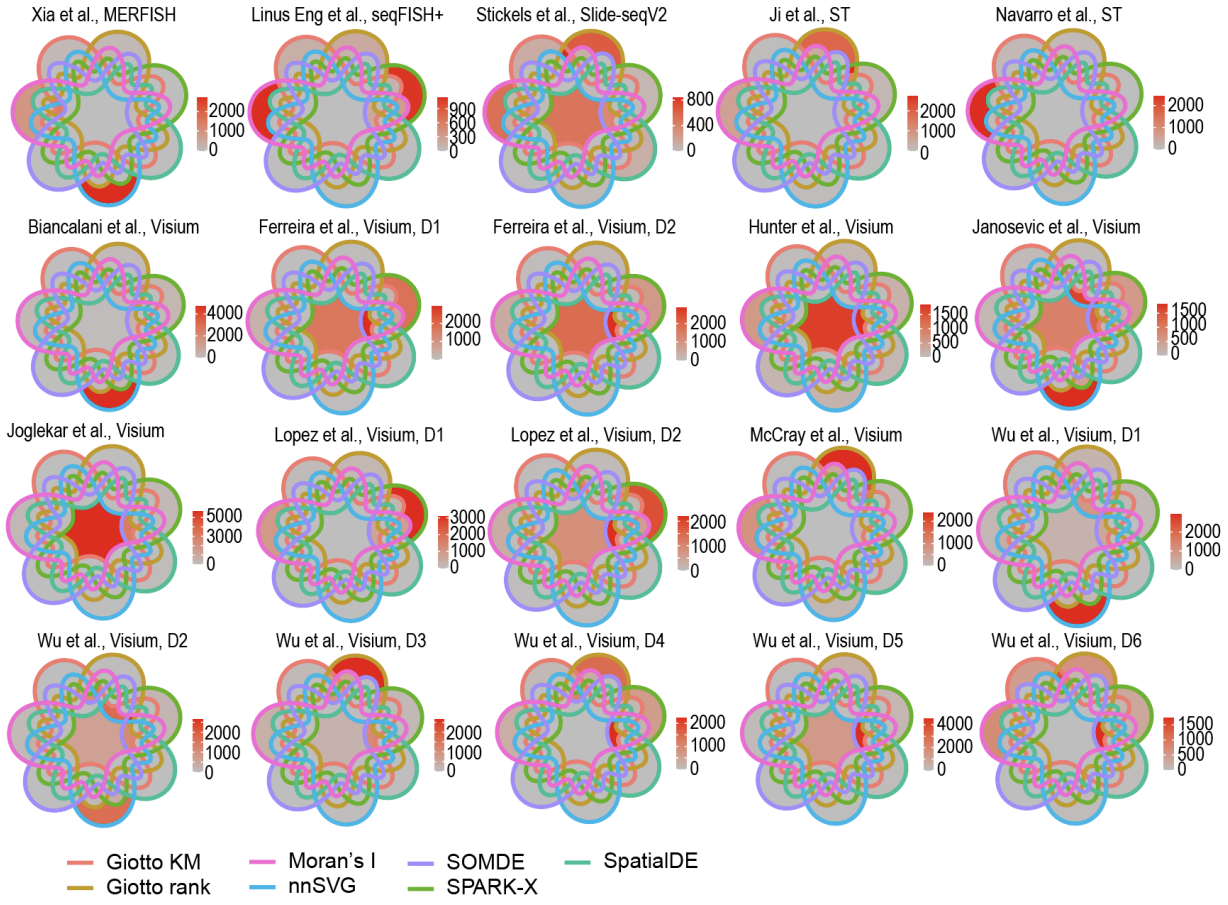

Figure S4. Venn diagram of the overlap of significant SVGs (adjusted p-value <= 0.05) reported by each method for each spatial transcriptomics dataset. Related to Figure 2.

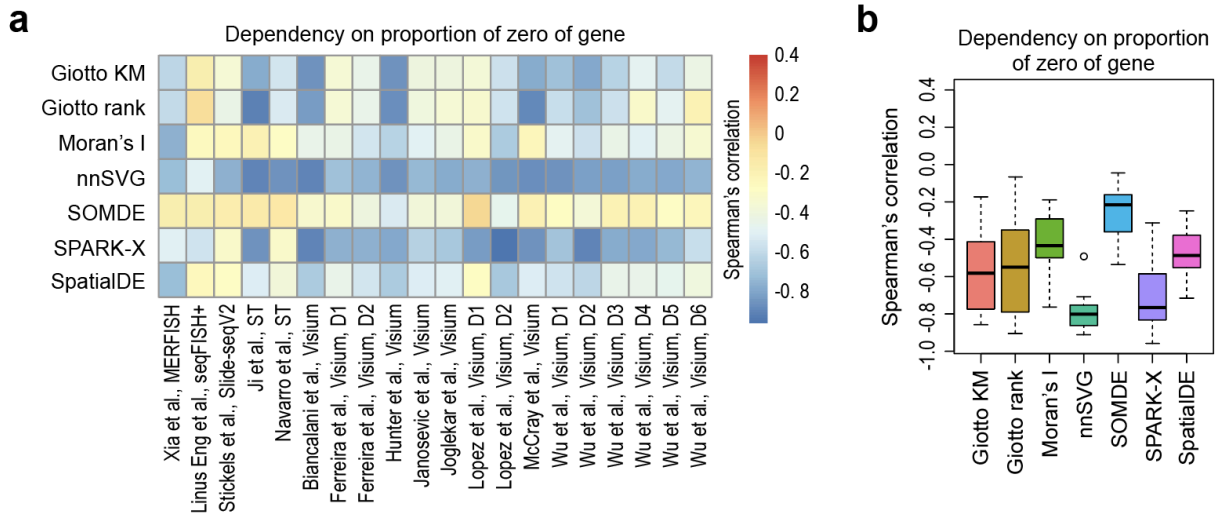

Figure S5. Relationship between SVG statistics and proportion of zero of genes. (**a**) Heatmap summarising the Spearman's correlation of SVG statistics reported by each method and the proportion of zero of genes across cells in each dataset. (**b**) Boxplot of Spearman's correlation of SVG statistics reported by each method and the proportion of zero of genes across the 20 spatial transcriptomics datasets. Related to Figure 3.

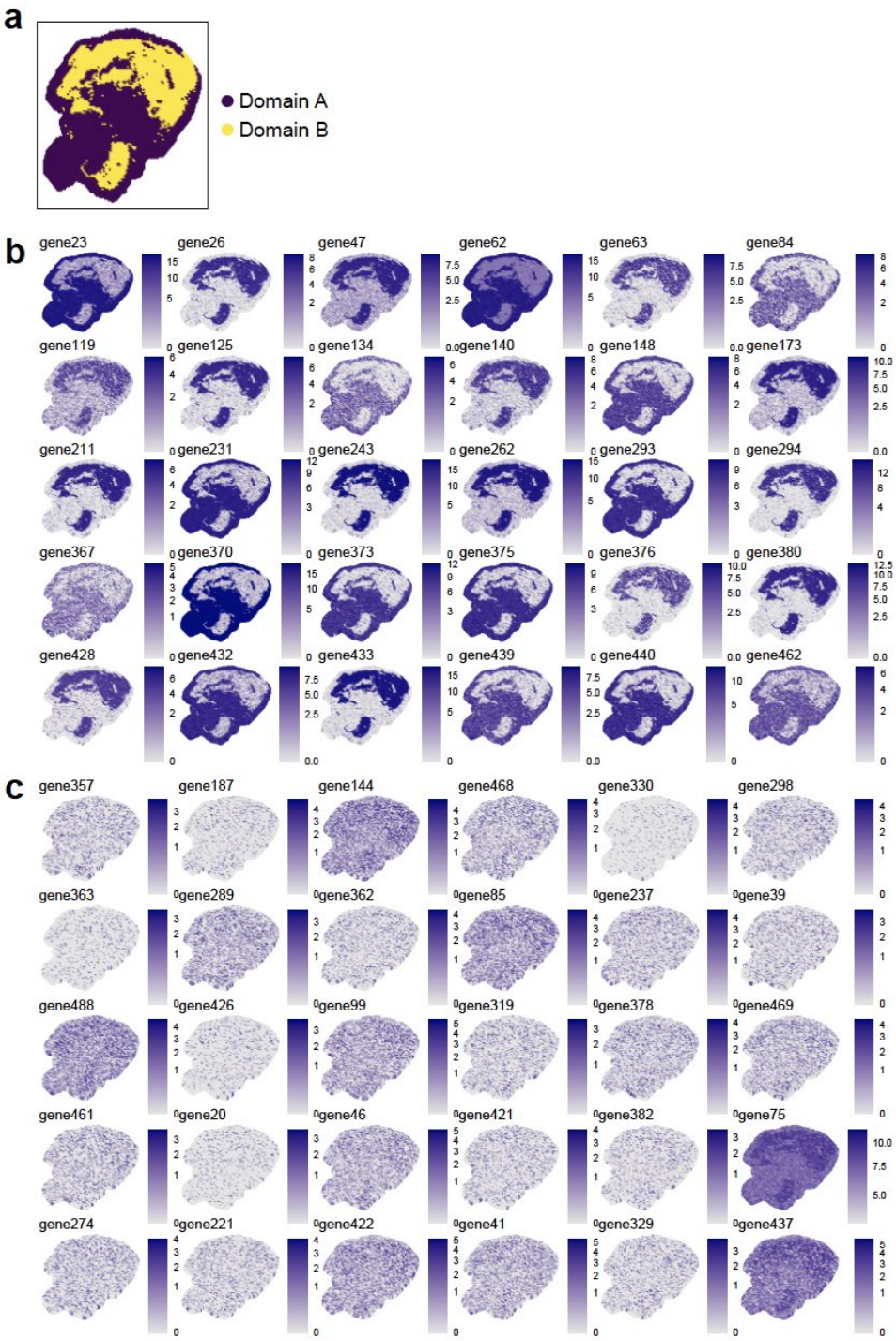

Figure S6. Simulation of spatial transcriptomics data. (a) Spatial domain distribution in an exemplar simulated data. Spatial expression patterns of example **(b)** spatially variable genes and **(c)** non-spatially variable genes. Related to Figure 4.

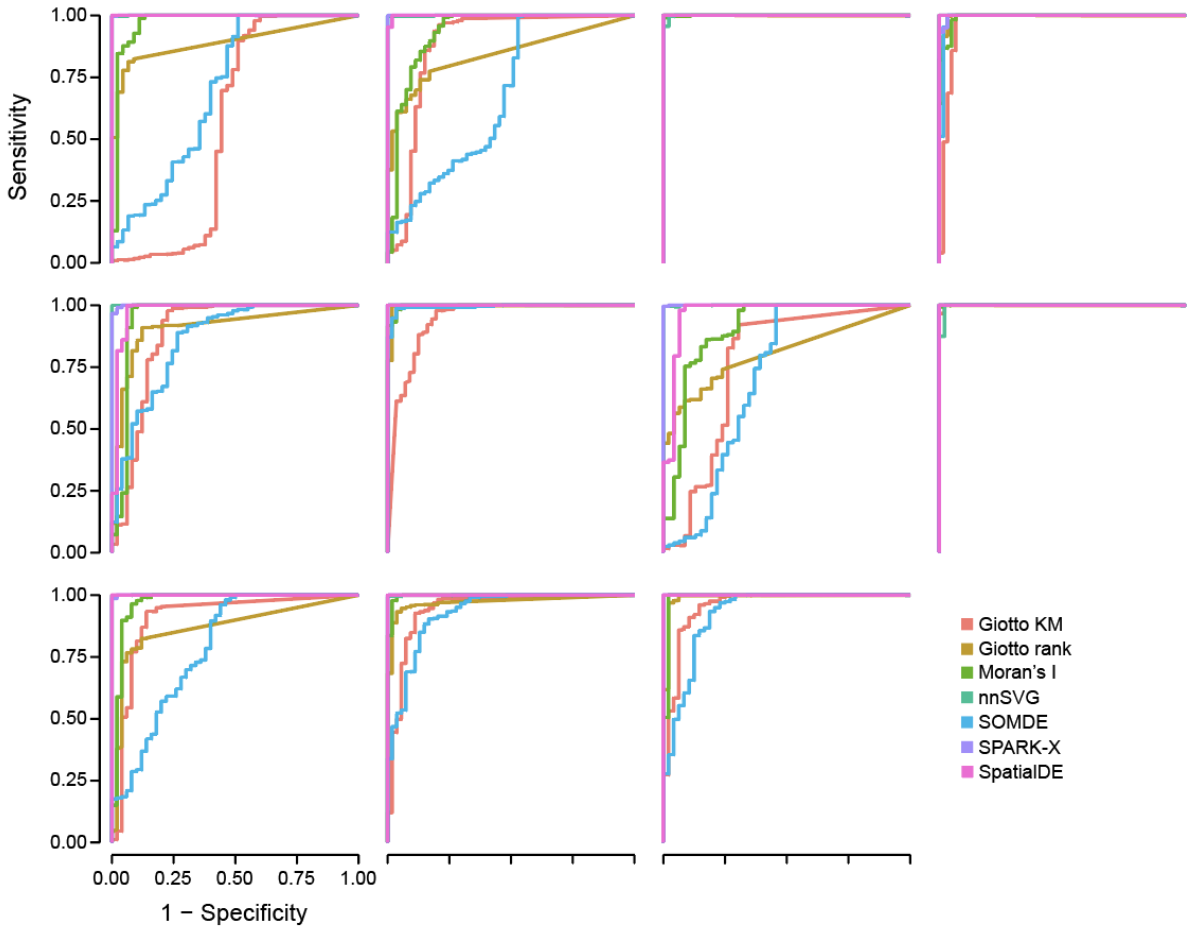

Figure S7. ROC curves of spatially variable gene detection. ROC curves were used to assess the capacity of each method to detect spatially variable genes between two spatial domains. The curves are colour-coded by methods. The 1-specificity and sensitivity of spatially variable gene detection are plotted on the x- and y-axes, respectively. Related to Figure 4.