SpaGE

Spatial Gene Enhancement using scRNA-seq

Abdelaal, Tamim; Mourragui, Soufiane; Mahfouz, Ahmed; Reinders, Marcel J.T.

DOI
10.1093/nar/gkaa740

Publication date
2020

Document Version
Final published version

Published in
Nucleic Acids Research

Citation (APA)
Abdelaal, T., Mourragui, S., Mahfouz, A., & Reinders, M. J. T. (2020). SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Research*, *48*(18), Article e107. https://doi.org/10.1093/nar/gkaa740

Published online 21 September 2020

Nucleic Acids Research, 2020, Vol. 48, No. 18 ***e107***
doi: 10.1093/nar/gkaa740

SpaGE: Spatial Gene Enhancement using scRNA-seq

Tamim Abdelaal[1,2], Soufiane Mourragui[1,3], Ahmed Mahfouz[1,2,4,†] and Marcel J.T. Reinders[1,2,4,*,†]

[1]Delft Bioinformatics Lab, Delft University of Technology, Delft 2628XE, The Netherlands, [2]Leiden Computational Biology Center, Leiden University Medical Center, Leiden 2333ZC, The Netherlands, [3]Computational Cancer Biology, Division of Molecular Carcinogenesis, Oncode Institute, the Netherlands Cancer Institute, Amsterdam 1066 CX, The Netherlands and [4]Department of Human Genetics, Leiden University Medical Center, Leiden 2333ZC, The Netherlands

Received May 16, 2020; Revised July 30, 2020; Editorial Decision August 24, 2020; Accepted August 25, 2020

ABSTRACT

Single-cell technologies are emerging fast due to their ability to unravel the heterogeneity of biological systems. While scRNA-seq is a powerful tool that measures whole-transcriptome expression of single cells, it lacks their spatial localization. Novel spatial transcriptomics methods do retain cells spatial information but some methods can only measure tens to hundreds of transcripts. To resolve this discrepancy, we developed SpaGE, a method that integrates spatial and scRNA-seq datasets to predict whole-transcriptome expressions in their spatial configuration. Using five dataset-pairs, SpaGE outperformed previously published methods and showed scalability to large datasets. Moreover, SpaGE predicted new spatial gene patterns that are confirmed independently using in situ hybridization data from the Allen Mouse Brain Atlas.

INTRODUCTION

Single cell technologies rapidly developed over the last decade and have become valuable tools for enhancing our understanding of biological systems. Single-cell RNA-sequencing (scRNA-seq) allows unbiased measurement of the entire gene expression profile of each individual cell and has become the *de facto* technology used to characterize the cellular composition of complex tissues (1,2). However, single cells often have to be dissociated before performing scRNA-seq and results in losing the spatial context and hence limits our understanding of cell identities and relationships. Recently, spatial transcriptomics technologies have advanced and provide localizations of gene expressions and cellular structure at the cellular level (3,4). Many current protocols can be divided in two categories: (i) imaging-based methods (e.g. osmFISH, MERFISH and seqFISH+) (5–7), and (ii) sequencing-based methods (e.g. STARmap and Slide-seq) (8,9). Imaging-based protocols have a high gene detection sensitivity; capturing high proportion of the mRNA molecules with relatively small dropout rate. While seqFISH+ and the latest generation of MERFISH can measure up to ∼10 000 genes (7,10), many different imaging-based protocols are often limited in the number of genes that can be measured simultaneously. On the other hand, sequencing-based protocols like STARmap can scale up to thousands of genes, it has a relatively lower gene detection sensitivity. Slide-seq is not limited in the number of measured genes and can be used to measure the whole transcriptome. However, similar to STARmap, Slide-seq suffers from a low gene detection sensitivity. In addition, osmFISH, MERFISH and STARmap can capture genes at the single-molecule resolution, which can be averaged or aggregated to the single-cell level. While Slide-seq has a resolution of 10μm, which is comparable to the average cell size, but does not always represent a single-cell.

Given the complementary information provided by both scRNA-seq and spatial transcriptomics data, integrating both types would provide a more complete overview of cell identities and interactions within complex tissues. This integration can be performed in two different ways (11): (i) dissociated single-cells measured with scRNA-seq can be mapped to their physical locations in the tissue (12–14), or (ii) missing gene expression measurements in the spatial data can be predicted from scRNA-seq. In this study, we focus on the second challenge in which measured gene expressions of spatial cells can be enhanced by predicting the expression of unmeasured genes based on scRNA-seq data of a matching tissue. Several methods have addressed this problem using various data integration approaches to account for the differences between the two data types (15–18). All these methods rely on joint dimensionality reduction methods to embed both spatial and scRNA-seq data into a common latent space. For example, Seurat uses canonical correlation analysis (CCA), Liger uses non-negative matrix factorization (NMF), and Harmony uses

*To whom correspondence should be addressed. Tel: +31 15 2786424; Fax: +31 15 2786424; Email: m.j.t.reinders@tudelft.nl
†The authors wish it to be known that, in their opinion, the last two authors contributed equally.

principal component analysis (PCA). While Seurat, Liger and Harmony rely on linear methods to embed the data, gimVI uses a non-linear deep generative model. Despite recent benchmarking efforts (19), a comprehensive evaluation of these methods for the task of spatial gene prediction from dissociated cells is currently lacking. For example, Seurat, Liger and gimVI, have only been tested using relatively small datasets (<2,000 cells) (15,16,18). It is thus not clear whether a complex model, like gimVI, is really necessary. Moreover, Seurat, Harmony and gimVI lack interpretability of the integration procedure, so that it does not become clear which genes contribute in the prediction task.

Here, we present SpaGE (Spatial Gene Enhancement), a robust, scalable and interpretable machine-learning method to predict unmeasured genes of each cell in spatial transcriptomic data through integration with scRNA-seq data from the same tissue. SpaGE relies on domain adaptation using PRECISE (20) to correct for differences in sensitivity of transcript detection between both single-cell technologies, followed by a *k*-nearest-neighbor (kNN) prediction of new spatial gene expression. We demonstrate that SpaGE outperforms state-of-the-art methods by accurately predicting unmeasured gene expression profiles across a variety of spatial and scRNA-seq dataset pairs of different regions in the mouse brain. These datasets include a large spatial data with >60,000 cells, used to illustrate the scalability and computational efficiency of SpaGE compared to other methods.

MATERIALS AND METHODS

SpaGE algorithm

The SpaGE algorithm takes as input two gene expression matrices corresponding to the scRNA-seq data (reference) and the spatial transcriptomics data (query). Based on the set of shared genes between the two datasets, SpaGE enriches the spatial transcriptomics data using the scRNA-seq data, by predicting the expression of spatially unmeasured genes. The SpaGE algorithm can be divided in two major steps: (i) alignment of the two datasets using the domain adaptation algorithm PRECISE (20), and (ii) gene expression prediction using *k*-nearest-neighbor regression.

First, PRECISE was used to project both datasets into a common latent space. Let $R_{(n \times g)}$ be the gene expression matrix of the reference dataset having n cells and g genes, and let $Q_{(m \times h)}$ be the gene expression matrix of the query dataset having m cells and h genes. Using the set of shared genes $p = g \cap h$, PRECISE applies independent Principal Component Analysis (PCA) for each dataset to define two independent sets of Principal Components (*PCs*), such that:

$$R_{(n \times p)} = R'_{(n \times d)} PC_{r(d \times p)} \text{ with } PC_r\, PC_r^T = I_d \quad (1)$$

and

$$Q_{(m \times p)} = Q'_{(m \times d)} PC_{q(d \times p)} \text{ with } PC_q\, PC_q^T = I_d \quad (2)$$

where d is the number of desired *PCs*, PC_r and PC_q represents the principal components of the reference and the query datasets, respectively. We choose $d = 50$ for the **STARmap_AllenVISp, MERFISH_Moffit and seqFISH_AllenVISp** dataset pairs, and $d = 30$ for all the **osmFISH** dataset pairs. Next, PRECISE compares these independent *PCs* by computing the cosine similarity matrix and decomposing it by SVD (21):

$$PC_r\ PC_q{}^T = U \Sigma V^T \quad (3)$$

where U and V represent orthogonal (of size d) transformations on the reference and query *PCs*, respectively, and Σ is a diagonal matrix. U and V are then used to align the *PCs*, yielding the so-called Principal Vectors (*PVs*), such that:

$$PV_r = U^T\ PC_r \quad (4)$$

and

$$PV_q = V^T\ PC_q \quad (5)$$

PV_r and PV_q are the principal vectors of the reference and the query datasets, respectively, retaining the same information as the principal components. However, these *PVs* have now a one-to-one correspondence as their cosine similarity matrix is diagonal (the matrix Σ). *PVs* are pairs of vectors $(PV_r^1, PV_q^1), \ldots, (PV_r^d, PV_q^d)$ sorted in decreasing order based on similarity. To remove noisy components, we choose a limited number of *PVs*, d', for further analysis, where the cosine similarity is higher than a certain threshold (0.3). The reference *PVs*, PV_r, are then used to project and align both the scRNA-seq (reference) and the spatial transcriptomics (query) datasets:

$$R_{aligned\ (n \times d')} = R_{(n \times p)} PV_{r\ (p \times d')}^T \quad (6)$$

and

$$Q_{aligned\ (m \times d')} = Q_{(m \times p)} PV_{r\ (p \times d')}^T \quad (7)$$

After aligning the datasets, SpaGE predicts the expression of the spatially unmeasured genes, $l = g - p$, from the scRNA-seq dataset. For each spatial cell $i \in m$, we define the *k*-nearest-neighbors ($k = 50$) from the n dissociated scRNA-seq cells, using the cosine distance. Next, we calculate an array of weights w_{ij} between spatial cell i and its nearest neighbors $j \in NN(i)$. Out of the 50 neighbors, we only keep neighbors with positive cosine similarity with cell i (i.e. cosine distance < 1), such that:

$$\forall\ j \in NN(i)\ and\ dist(i, j) < 1$$

$$w_{ij} = 1 - \frac{dist(i, j)}{\sum_j dist(i, j)} \quad (8)$$

$$w_{ij} = \frac{w_{ij}}{length\left(w_{ij}\right) - 1} \quad (9)$$

The predicted expression Y_{il} of the set of spatially unmeasured genes l for cell i is calculated as a weighted average of the nearest neighbors dissociated cells:

$$Y_{il} = \sum_{\substack{j \in NN(i) \\ dist(i, j) < 1}} w_{ij} * R_{jl} \quad (10)$$

Gene contribution to the integration

To evaluate the contribution of each gene in forming this common latent space PV_r, we calculated the gene contribu-

tion C_g of gene $_g$ as follows:

$$C_g = \sum_{i=1}^{d'} \beta_{gi}^2 \quad (11)$$

where β_{gi} is the loading of gene g to the i-th principal vector in PV_r, and d' is the final number of PVs in PV_r. To obtain the top contributing genes, the C_g values are sorted in descending order across all genes. We used the same criteria to calculate the contribution of each gene for dataset-specific PCs or PVs.

Datasets

We used six dataset pairs (Table 1) composed of four scRNA-seq datasets (**AllenVISp** (22), **AllenSSp** (23), **Zeisel** (24) and **Moffit** (4)) and four spatial transcriptomics datasets (**STARmap** (8), **osmFISH** (5), **MERFISH** (4) and **seqFISH+** (7)). The **AllenVISp** (GSE115746) and the **AllenSSp** datasets were downloaded from https://portal.brain-map.org/atlases-and-data/rnaseq. The **AllenVISp** is obtained from the 'Cell Diversity in the Mouse Cortex – 2018' release. The **AllenSSp** is obtained from the 'Cell Diversity in the Mouse Cortex and Hippocampus' release of October 2019. We downloaded the whole dataset and used the metadata to only select cells from the SSp region. The **Zeisel** dataset (GSE60361) was downloaded from http://linnarssonlab.org/cortex/, while the **Moffit** 10X dataset (GSE113576) was downloaded from GEO.

The **STARmap** dataset was downloaded from the STARmap resources website (https://www.starmapresources.com/data). We obtained the gene count matrix and the cell position information for the largest 1020-gene replicate. Cell locations and morphologies were identified using Python code provided by the original study (https://github.com/weallen/STARmap). The **osmFISH** dataset was downloaded as loom file from http://linnarssonlab.org/osmFISH/, we obtained the gene count matrix and the metadata using the loompy Python package. The **MERFISH** dataset was downloaded from Dryad repository (https://doi.org/10.5061/dryad.8t8s248), we used the first naïve female mouse (Animal_ID = 1). The **seqFISH+** dataset was obtained from the seqFISH-PLUS GitHub repository (https://github.com/CaiGroup/seqFISH-PLUS), we used the gene count matrix of the mouse cortex dataset.

Data preprocessing

For all the scRNA-seq datasets, we filtered out genes expressed in less than 10 cells. No filtration was applied on the cells, except for the **AllenVISp** dataset for which we filtered low quality cells provided from the metadata ('Low Quality' and 'No Class' cells). For the **Zeisel** dataset, we only used the somatosensory cortex cells excluding the hippocampus cells. Next, scRNA-seq datasets were normalized by dividing the counts within each cell by the total number of transcripts within that cell, scaling by 10^6 and log1p transformed. Further, we scaled the data by making each gene centered and scaled (zero mean and unit variance) using the SciPy Python package (25).

For spatial transcriptomics datasets all gene were used, except for the **MERFISH** dataset for which we removed the blanks genes and the *Fos* gene (non-numerical values). Additionally, we filtered out cells labeled as 'Ambiguous' from the **MERFISH** dataset. Similar to the **Zeisel** dataset, we only kept cells from cortical regions for the **osmFISH** dataset ('Layer 2–3 lateral', 'Layer 2–3 medial', 'Layer 3–4', 'Layer 4', 'Layer 5', 'Layer 6' and 'Pia Layer 1'). For the **seqFISH+** dataset, we only used the cells from the 'Cortex' region. No cells were filtered from the **STARmap** dataset. Further, each dataset was normalized by dividing the counts within each cell by the total number of transcripts within that cell, scaling by the median number of transcripts per cell, and log1p transformed. Similar to the scRNA-seq data, we scaled the spatial data using the SciPy Python package (25).

It is important to note that in all experiments, the scaled datasets are used as input for the alignment part, while the prediction is applied using the normalized version of the scRNA-seq dataset (Equation 10).

Cross validation

We evaluated the prediction performance of all methods using a leave-one-gene-out cross validation. For a set of N shared genes between the spatial and the scRNA-seq datasets, one gene is left out and the remaining N-1 genes are used for integration and prediction of the left-out gene. The prediction is then evaluated by comparing the measured and predicted spatial profiles of the left-out-gene.

For the **STARmap_AllenVISp** dataset pair, we applied a more challenging cross validation setup. Similar to the leave-one-gene-out setup, for a set of N shared genes, one gene is left out to be predicted. From the remaining $N-1$ genes, we excluded the 100 genes that are most correlated (absolute Pearson correlation) with the left-out gene. The remaining $N-101$ genes are then used for the integration and prediction of the left-out gene.

Benchmarked methods

We compared the performance of SpaGE versus three state-of-the-art methods for data integration: Seurat, Liger, and gimVI. Seurat and Liger are available as R packages, while gimVI is available through the scVI Python package (26). We were not able to include Harmony in the comparison, as the code to predict unmeasured gene expression is not available. During the benchmark, all methods were applied using their default settings, or the settings provided in the accompanying examples or vignettes. Data normalization and scaling were performed using the built-in functions in each package, *NormalizeData* and *ScaleData* functions in Seurat, *normalize* and *scaleNotCenter* functions in Liger, while gimVI implicitly preprocess the data while computing.

Moran's I statistic

The Moran's I statistic (27) is a measure of spatial autocorrelation calculated for each spatial gene. The Moran's I values can range from −1 to 1, where a value close to 1 indicates a clear spatial pattern, and a value close to 0 indicates

Table 1. Summary of the dataset pairs used in this study

	Spatial data			scRNA-seq data		
Spatial_scRNA-seq dataset pair	# of cells	# of genes	Tissue	# of cells	# of genes	Tissue
STARmap_AllenVISp (8,22)	1,549	1,020	VISc	14,249	34,617	VISc
osmFISH_Zeisel (5,24)	3,405	33	SMSc	1,691	15,075	SMSc
osmFISH_AllenSSp (5,23)	3,405	33	SMSc	5,577	30,527	SMSc
osmFISH_AllenVISp (5,22)	3,405	33	SMSc	14,249	34,617	VISc
MERFISH_Moffit (4)	64,373	155	POR	31,299	18,646	POR
seqFISH_AllenVISp (7,22)	524	10,000	Cortex	14,249	34,617	VISc

VISc: Visual cortex; SMSc: Somatosensory cortex; POR: Pre-optic region

random spatial expression, while a value close to −1 indicated a chess board like pattern. We calculated the Moran's I using the following equation:

$$I = \frac{N}{W}\frac{\sum_i \sum_j w_{ij}\ (x_i - \bar{x})\,(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \qquad (12)$$

where x is the gene expression array, $\bar{x}$ is the mean expression of gene x, N is the total number of spatial cells, w_{ij} is a matrix containing spatial weights with zeros on the diagonal, and W is the sum of w_{ij}. We calculated the spatial weights w_{ij} using the XY coordinates of the spatial cells, for each cell we calculated the kNN using the spatial coordinates ($k = 4$). We assigned $w_{ij} = 1$ if j is in the nearest neighbors of i, otherwise $w_{ij} = 0$.

Down-sampling

For the 994 shared genes in the **STARmap_AllenVISp** dataset pair, we first selected the top 50 spatial genes with high Moran's I statistic values to be used as test set. For the remaining 944 genes, we calculated the pairwise Pearson correlation using the scRNA-seq dataset. If the absolute value of the correlation of two genes is larger than 0.7, we removed the gene with the lower variance. After removing highly correlated genes, we sorted the remaining genes according to their expression variance in the scRNA-seq dataset. We selected the top 10, 30, 50, 100, 200 and 500 genes with high variance, these genes were used for alignment of the two datasets and prediction of the expression of the test genes. The prediction performance of these gene sets was compared with using all 944 genes.

We applied the same down-sampling criteria on the 9,751 shared genes in the **seqFISH_AllenVISp** dataset pair, except for two differences: (i) the 50 spatial genes used as test set were selected as the top predicted genes in the leave-one-gene-out cross validation experiment, (ii) after removing correlated genes, we selected sets of the top 10, 30, 50, 100, 200, 500, 1000, 2000, 5000 and 7000 most variable genes, as well as all 9,701 genes.

Cell-type marker genes

To evaluate the performance of SpaGE per cell type, we defined sets of marker genes for four major brain cell types: Inhibitory neurons, Excitatory neurons, Astrocytes and Oligodendrocytes. The marker genes of the **osmFISH** dataset were directly obtained from the original paper (5). For the **STARmap** and **MERFISH** datasets, we used the *FindMarkers* function from the Seurat R package to define the top 20 differentially expressed genes per cell type, comparing one cell type vs the rest using a two-sided Wilcoxon rank sum test and the Bonferroni method for multiple test correction, with min.pct = 0.25 and logfc.threshold = 0.25.

A model to predict trustworthiness of the SpaGE prediction

To determine whether we can trust a predicted spatial pattern by SpaGE, we trained a logistic regression model that predicts the trustworthiness of the predicted signal from four characteristics of the data: (i) the Moran's I statistic of the predicted spatial gene expression (pMI_i), (ii) the mean μ_i and (iii) variance σ_i of the expression of that gene in the scRNA-seq data and (iv) the percentage of cells expressing that gene in the scRNA-seq data (e_i). The trustworthiness, Y_i, used to train the model, is determined from the Spearman correlation between the SpaGE-predicted spatial pattern and the measured spatial pattern, i.e. correlations above the median correlation are considered to be trustworthy. This gives the following logistic regression model:

$$Y_i \sim pMI_i + \mu_i + \sigma_i + e_i$$

Note that the inputs to the model can be determined without the need to have access to the measured spatial expression of the gene, and consequently the model can be used to evaluate whether the predicted spatial pattern of expression of an unmeasured spatial gene is to be trusted or not.

RESULTS

SpaGE overview

We developed SpaGE, a platform that enhances the spatial transcriptomics data by predicting the expression of unmeasured genes from a dissociated scRNA-seq data from the same tissue (Figure 1). Based on the set of shared genes, we align both datasets using the domain adaptation method PRECISE (20), to account for technical differences as well as gene detection sensitivity differences. PRECISE geometrically aligns linear latent factors computed on each dataset and finds gene combinations expressed in both datasets. These gene combinations thus define a common latent space and can be used to jointly project both datasets. Next, in this common latent space, we use the kNN algorithm to define the neighborhood of each cell in the spatial data from the scRNA-seq cells. These neighboring scRNA-seq cells are then used to predict the expression of spatially unmeasured

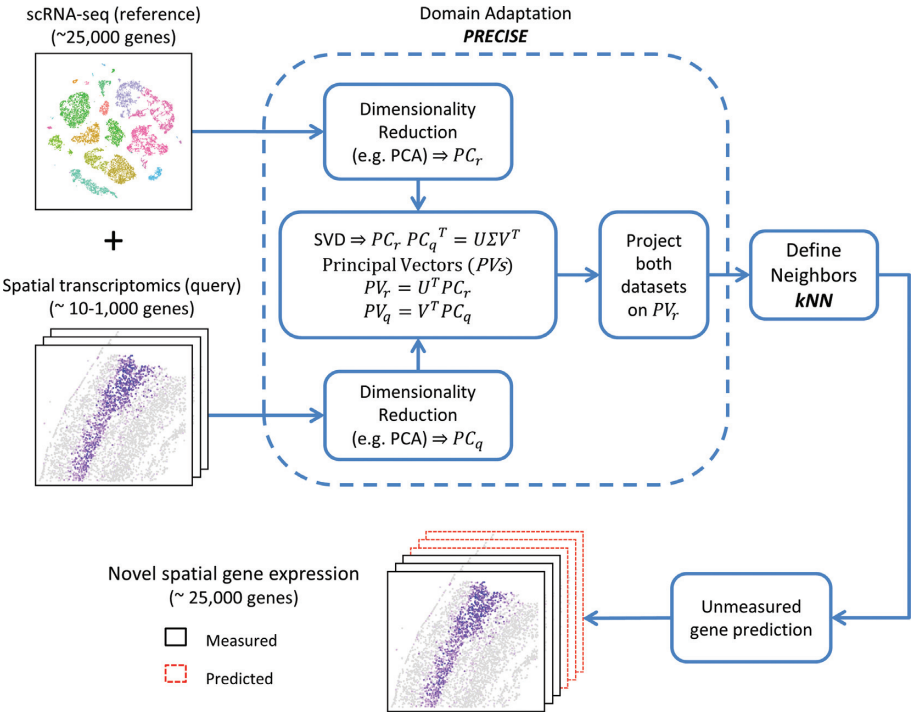

Figure 1. SpaGE pipeline. SpaGE takes as input two datasets, a scRNA-seq dataset and a spatial transcriptomics dataset measured from the same tissue. SpaGE uses gene combinations of equal significance in both datasets to predict spatial locations of unmeasured genes. Using PRECISE, SpaGE finds directions that are important for both datasets, by making use of a geometrical alignment of the independent *PCs* to produce the *PVs*. SpaGE aligns both datasets by projecting on the *PVs* of the reference dataset. Using the aligned datasets, SpaGE applies kNN prediction to define new gene expression patterns for spatially unmeasured genes, predicted from the dissociated scRNA-seq data. Each spatial cell can be enhanced by having the expression of the whole transcriptome.

genes. Finally, we end up with the full gene expression profile of each cell in the spatial data.

The alignment step is the most crucial step in the pipeline of SpaGE. For this purpose, we use PRECISE, a domain adaptation method previously proposed to predict the drug response of human tumors based on pre-clinical models such as cell lines and mouse models. We adapted PRECISE to the task of integrating the spatial data with the scRNA-seq data by defining the common aligned subspace between both datasets (Figure 1). PRECISE takes as input the expression matrix of both datasets, having the same set of (overlapping) genes but measured differently and within different cells. As we are aiming to fit each spatial cell to the most similar scRNA-seq cells, we may refer to the spatial dataset as the 'query' and the scRNA-seq dataset as the 'reference'. First, PRECISE obtains a lower dimensional space for each dataset separately using a linear dimensionality reduction method, such as Principal Component Analysis (PCA). Next, the two independent sets of principal components (*PCs*) are aligned by applying a singular value decomposition. We align the two sets of principal components using the singular vectors to obtain the aligned components, named principal vectors (*PVs*). These *PVs* are sorted in decreasing order based on their similarity between the reference and the query datasets. This allows us to filter out dissimilar or noisy signals, by discarding *PVs* with relatively low similarity, thus keeping only the common latent space (Methods). The principal vectors of the reference dataset (PV_r) are considered as the aligned latent space. We project both datasets on PV_r to obtain the new aligned versions used for the kNN prediction.

We performed SpaGE on six dataset pairs from different regions in the mouse brain, varying in the number of cells and the number of spatially measured genes, summarized in Table 1. To show the alignment performance, we calculated the cosine similarity between the *PCs* and the *PVs*, i.e. before and after the alignment. Across all six dataset pairs, we observed that indeed the relation between the *PCs* is not one-to-one, as these *PCs* are obtained from two different datasets (Supplementary Figure S1 and S2). However, after alignment using PRECISE, the diagonal cosine similarity between the *PVs* is maximized showing a one-to-one relationship between the *PVs* of both datasets. Supplementary Figure S1A shows the diagonal cosine similarity before and after PRECISE (i.e. between *PCs* and *PVs*) across all dataset pairs, showing a relatively large increase in similarity after the alignment using PRECISE. As we used only the informative *PVs*, the final number of *PVs* varied across datasets (Supplementary Table S1) and, as a result, the amount of explained variance for each dataset varied, from ∼6% for the **seqFISH+** dataset to ∼94% for the **osmFISH** dataset.

Another interesting feature of SpaGE is the ability to interpret the most contributing genes defining the latent integration space (Methods). In general, these genes are highly variable and in most cases are related to cell type differences. A good example is the integration of the **osmFISH_Zeisel** dataset pair, in which the top six contributing genes are

Tmem2, *Mrc1*, *Kcnip2*, *Foxj1*, *Apln* and *Syt6*. These genes are related to six different cell categories previously defined in the **osmFISH** paper (5): Oligodendrocytes, Immune cells, Inhibitory neurons, Ventricle, Vasculature and Excitatory neurons, respectively.

We further illustrate the quality of the alignment by examining the overlap in the top contributing genes for the *PCs* (before PRECISE) and the *PVs* (after PRECISE). Using the **STARmap_AllenVISp** dataset pair, we obtained the top 50 contributing genes for the *PCs* of the **STARmap** data and the *PCs* of the **AllenVISp** data. These two sets shared only 2 genes out of 50. After alignment, the shared genes, between the top 50 contributing genes for the *PVs* of the **STARmap** data and the *PVs* of the **AllenVISp** data, increased to 12 genes. Also, we applied GO enrichment on these top contributing gene sets in each case using PANTHER (http://pantherdb.org/, Fisher exact test with Bonferroni multiple test correction). The **STARmap** *PCs* and the **AllenVISp** *PCs* had 9 enriched biological processes each, sharing 3 processes in common (Supplementary Table S2). While the **STARmap** *PVs* and the **AllenVISp** *PVs* had 27 and 41 enriched biological processes, respectively, sharing 12 processes in common. Interestingly many of them related are to regulation processes, such as regulation of biological process, cell population proliferation, metabolic processes, cell motility, locomotion and cellular component movement.

SpaGE outperforms state-of-the-art methods on the STARmap dataset

Using the first dataset pair **STARmap_AllenVISp**, we applied SpaGE to integrate both datasets and predict unmeasured spatial gene expression patterns. In order to evaluate the prediction, we performed a leave-one-gene-out cross validation (Methods). The **STARmap_AllenVISp** dataset pair shares 994 genes. In each cross-validation fold, one gene is left out and the remaining 993 genes are used as input for SpaGE to predict the spatial expression pattern of the left-out gene. We evaluated the prediction performance by calculating the Spearman correlation between the original measured spatially distributed values and the predicted values of the left-out gene. We performed the same leave-one-gene-out cross validation using Seurat, Liger and gimVI, to benchmark the performance of SpaGE. Results show a significant improvement in performance for SpaGE compared to all three methods (P-value < 0.05, two-sided paired Wilcoxon rank sum test), with a median Spearman correlation of 0.125 compared to 0.083, 0.067 and 0.035 for Seurat, Liger and gimVI, respectively (Figure 2A).

Further, we compared the Spearman correlation of SpaGE versus the state-of-the-art methods per gene, to obtain a detailed evaluation. Results show better performance of SpaGE across the majority of genes, but not all (Figure 2B–D). Next, we visually compared a few genes that had high correlations for each method. For the top three predicted genes of SpaGE (*Pcsk2*, *Pgm2l1* and *Egr1*), Seurat obtained a good prediction as well, as these three genes are in the top 10 predicted genes of Seurat. Liger failed to predict *Egr1*, while gimVI failed to predict *Pgm2l1* and *Egr1* (Supplementary Figure S3A). We further looked for examples where other methods obtained higher correlations than SpaGE, excluding the top 10 predicted genes by SpaGE. Compared to Seurat, SpaGE similarly predicted the expression of *Arpp19*, but predicted relatively higher contrast patterns for *Pcp4* and *Arc* (Supplementary Figure S3B). Compared to Liger, SpaGE similarly predicted the expression of *Mobp*, higher contrast pattern for *Hpcal4*, and better predicted the spatial pattern of *Tsnax* (Supplementary Figure S3C). Compared to gimVI, SpaGE predicted a lower contrast pattern for *Arx*, a higher contrast pattern for *Snurf*, but failed to reproduce the measured spatial pattern for *Bcl6* (Supplementary Figure S3D). Remarkably, the predicted spatial patterns of SpaGE, for all three genes, are more in agreement with the data from the Allen Brain Atlas, suggesting that these genes were not accurately measured in the **STARmap** dataset.

Although the correlation values are in general low, SpaGE is capable of accurately reconstructing genes with clear spatial pattern in the brain. Figure 2E shows a set of genes known to have spatial patterns (previously reported by Seurat, Liger and gimVI). In this set of genes, Seurat and Liger are performing well, except that Liger produced a lower contrast expression pattern in some cases (e.g. *Lamp5* and *Bsg*). gimVI produced good prediction for *Lamp5*, however, gimVI was not able to predict the correct gene patterns for the other genes.

To obtain a better understanding and interpretation of these correlation values, we evaluated the effect of the kNN algorithm on the prediction performance. To do so, we divided the **AllenVISp** dataset into two stratified folds ensuring an equal composition of cell types. We used one-fold to predict genes in the other fold using the shared genes. Note that this does not require an alignment (PRECISE), so we can test the influence of the kNN regression. We applied a leave-one-gene-out cross validation using the same set of 994 shared genes of the **STARmap_AllenVISp** dataset pair, which resulted in a median Spearman correlation of 0.551 (Supplementary Figure S4A). While the performance is clearly better compared to that of SpaGE using the **STARmap_AllenVISp** dataset pair (median Spearman correlation = 0.125), it shows that the kNN regression is partially responsible for reduced correlation values.

To investigate the influence of the correlation metric, we tested also the Pearson and Kendall correlation measures, which showed that the highest correlation values are obtained when using the Spearman correlation (Supplementary Figure S4B). Next, we were interested how well SpaGE could predict when there was no difference between measurement modalities (here, spatial and scRNA-seq). Therefore, we used SpaGE to integrate the **Zeisel** and **AllenSSp** datasets, representing two scRNA-seq measured datasets from the same brain region. Using the leave-one-gene-out cross validation and the same shared genes of the **STARmap_AllenVISp** dataset pair, we obtained a median Spearman correlation of 0.303 (query: **Zeisel**, reference: **AllenSSp**) and 0.331 (query: **AllenSSp**, reference: **Zeisel**) (Supplementary Figure S4B). These correlation values suggest that the observed correlation values obtained when applying SpaGE on spatial and scRNA-seq datasets are not as low as they appear.

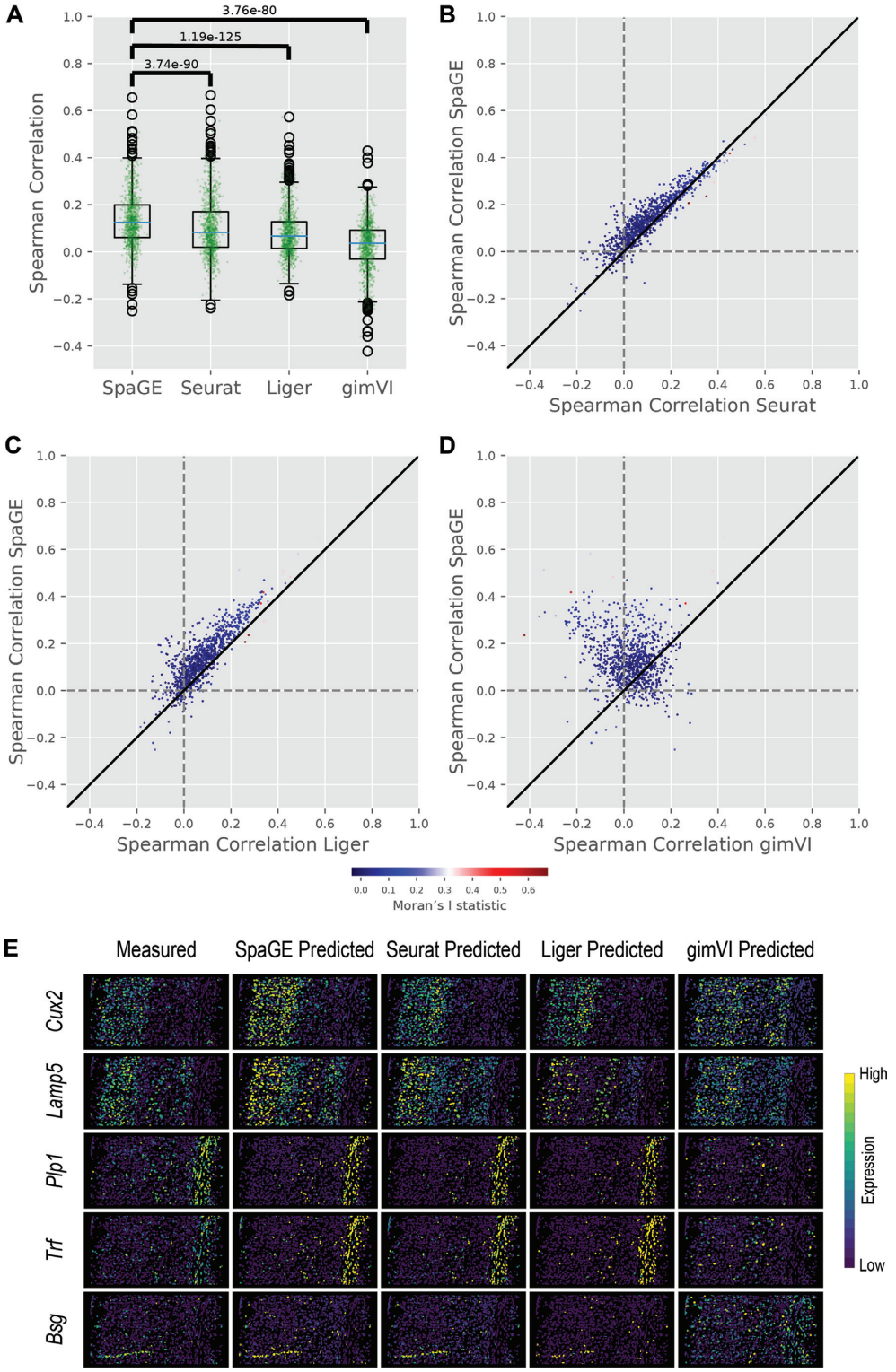

Figure 2. Prediction performance comparison for the STARmap_AllenVISp dataset pair. (**A**) Boxplots showing the Spearman correlations for the leave-one-gene-out cross validation experiment for each method. The blue lines show the median correlation across all genes with a better performance for SpaGE. The green dots show the correlation values for individual genes. The *P*-values show the significant difference between all correlation values of SpaGE and each other method, using a paired Wilcoxon rank-sum test. (**B–D**) Detailed performance comparison between SpaGE and (**B**) Seurat, (**C**) Liger, (**D**) gimVI. These scatter plots show the correlation value of each gene across two methods. The solid black line is the $y = x$ line, the dashed lines show the zero correlation. Points are colored according to the Moran's I statistic of each gene. All scatter plots show that the majority of the genes are skewed above the $y = x$ line, showing an overall better performance of SpaGE over other methods. (**E**) Predicted expression of known spatially patterned genes in the **STARmap** dataset. Each row corresponds to a single gene having a clear spatial pattern. First column from the left shows the measured spatial gene expression in the **STARmap** dataset, while other columns show the corresponding predicted expression pattern by SpaGE, Seurat, Liger and gimVI, using the leave-one-gene-out cross validation experiment. Prediction is performed using the **AllenVISp** dataset.

Additionally, although it is important to accurately predict the expression of all genes, genes with distinct spatial patterns are more important to accurately predict compared to non- or randomly expressed genes. To quantify the existence of spatial patterns, we calculate the Moran's I statistic of each gene using the original **STARmap** spatial data (Methods). We compared the prediction performance of each gene with the corresponding Moran's I value. For SpaGE, Seurat and Liger, we observed a positive relationship between the prediction performance and the Moran's I values, i.e. genes with spatial patterns are better predicted (Supplementary Figure S5A–C). On the other hand, gimVI performed worse on genes with high Moran's I statistic (Supplementary Figure S5D).

Further, we evaluated the prediction performance of all methods using a more challenging cross validation setup. Compared to the (traditional) leave-one-gene-out setup, the left-out gene is predicted using less shared genes in this set up, i.e. we removed the (100) most correlated genes with the left-out gene from the training set (Methods). This more challenging evaluation did result in comparable prediction performance to the leave-one-gene-out setup, with roughly the same differences and ranking across all methods (Supplementary Figure S6A). In addition, we evaluated how well a gene can be predicted when using less shared genes in general. First, we selected a fixed test set of 50 genes, next we down-sampled the remaining set of 944 shared genes in a guided manner (Methods). For down-sampled shared genes sets of 10, 30, 50, 100, 200, 500 and all 944 genes, SpaGE performance always increases with the number of shared genes as expected (Supplementary Figure S6B).

SpaGE predicts unmeasured spatial gene patterns that are independently validated

After validating SpaGE to accurately predict the spatially measured genes, we applied SpaGE to predict new unmeasured genes for the spatial data, with the aim to define novel spatial gene patterns. We illustrate SpaGE's capability of such task using the **STARmap_AllenVISp** dataset pair. First, during the leave-one-gene-out cross validation, SpaGE was able to produce the correct spatial pattern for *Rorb*, *Syt6* and *Tbr1* (Figure 3). These three genes were originally under-expressed, possibly due technical noise or low gene detection sensitivity in the **STARmap** dataset. Our predictions using SpaGE are in agreement with the highly sensitive cyclic smFISH dataset (**osmFISH** (5)) measured from the mouse somatosensory cortex, a similar brain region in terms of layering structure to the visual cortex measured by the **STARmap** dataset. Further, using SpaGE, we were able to obtain novel spatial gene patterns for five genes not originally measured by the **STARmap** dataset, showing clear patterns through the cortical layers (Figure 4). These predicted patterns are supported by the Allen Brain Atlas *in-situ* hybridization (ISH).

To quantitatively evaluate the predicted spatial patterns for non-measured genes, we trained a logistic regression model to estimate whether a predicted spatial gene expression can be trusted or not (Methods). We used three statistical features from the scRNA-seq data, in addition to the Moran's I statistic of the *predicted* spatial pattern. When

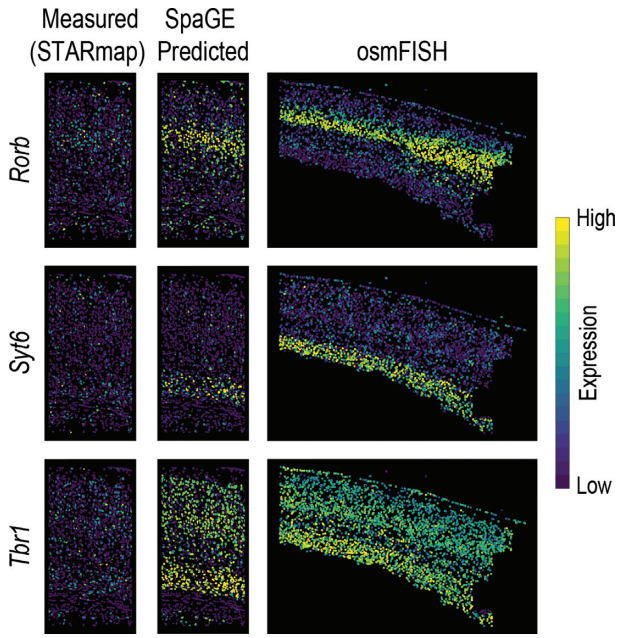

Figure 3. SpaGE accurately predicted the expression of *Rorb*, *Syt6* and *Tbr1* in agreement with the osmFISH data. These three genes (shown in rows) were wrongly measured in the original **STARmap** data (shown in the left column). Using the **STARmap_AllenVISp** dataset pair, SpaGE was able to reconstruct the correct spatial gene expression patterns (middle column). These predicted patterns are in agreement with the measured gene expression patterns measure by the **osmFISH** dataset (right column), a highly sensitive single-molecule technology.

training this model, we used the Spearman correlation between the SpaGE-predicted spatial pattern and the measured spatial pattern to determine whether a gene can be trusted or not, i.e. we assumed that correlations above the median correlation are trustworthy. Using the 994 shared genes of the **STARmap_AllenVISp** dataset pair, we obtained an average accuracy of 0.71 for a stratified 2-fold cross validation. Next, we trained the model using all genes and applied it to the estimated gene patterns in Figures 3 and 4. This model judged the predicted patterns of *Rorb*, *Tbr1*, *Tesc*, *Pvlr3* and *Rora*, trustworthy, and the patterns for *Syt6*, *Chst8* and *Cdh24* were not. Interestingly, when inspecting the model's coefficients we found that the Moran's I statistic of the predicted spatial pattern had the largest contribution.

SpaGE predictions improve with deeply sequenced reference dataset

We wanted to test the effect of changing the reference scRNA-seq data on the spatial gene expression prediction. Here, we used the **osmFISH** dataset which represents a different challenge compared to the **STARmap** dataset. On one hand, the **osmFISH** dataset has a relatively higher gene detection sensitivity, but on the other hand, the **osmFISH** dataset includes only 33 genes. First, we evaluated the **osmFISH_Zeisel** dataset pair, in which we integrated the **osmFISH** dataset with a reference scRNA-seq dataset from the same lab (24). We performed leave-one-gene-out cross validation similar to the **STARmap** dataset. Compared to other methods, SpaGE has significantly better performance (*P*-

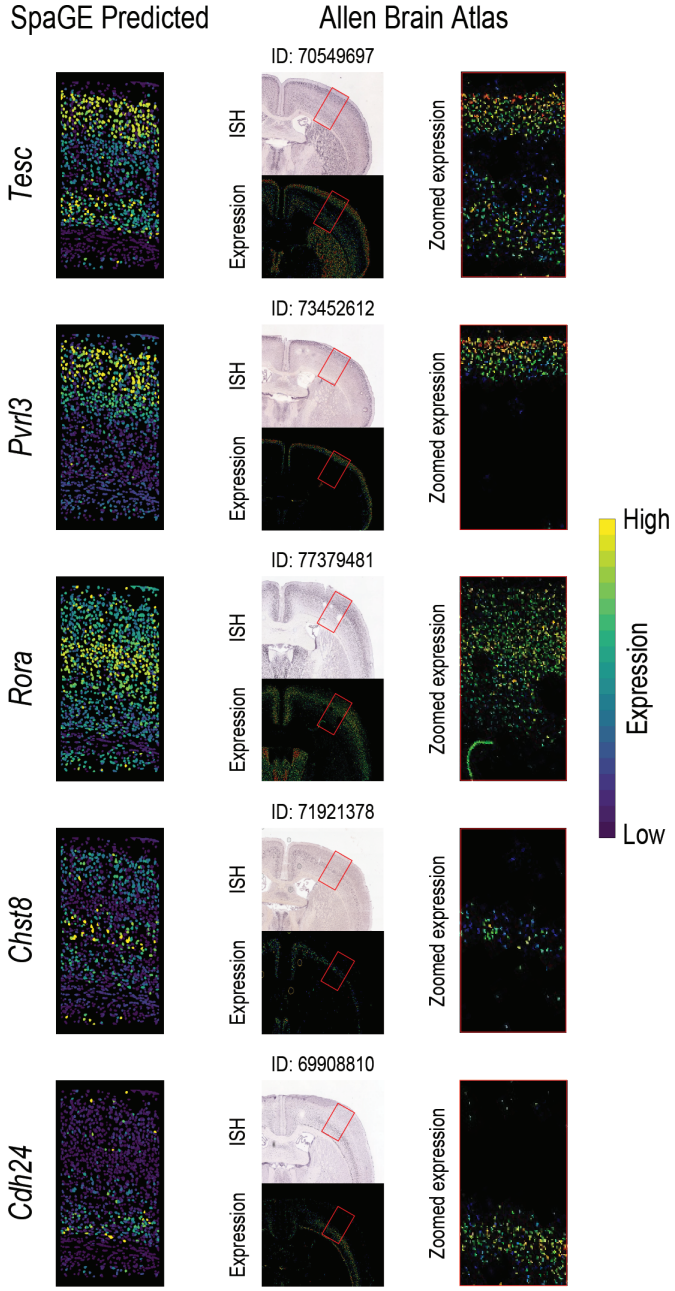

Figure 4. Novel gene expression patterns for five genes not originally measured by the STARmap dataset, validated using the Allen Brain Atlas *in-situ* hybridization ISH. The left column shows the predicted spatial patterns using SpaGE for these five genes (shown in rows). The middle column shows the Allen Brain Atlas ISH data for each gene, stating the image ID on top of each tissue section. The red rectangle highlights the corresponding brain region measured by the **STARmap** dataset. The right column shows a zoomed-in view of the region highlighted using this red rectangle, showing an agreement with the expression patterns predicted by SpaGE.

value <0.05, two-sided paired Wilcoxon rank sum test), with a median Spearman correlation of 0.203 compared to 0.007, 0.090 and 0.133 for Seurat, Liger and gimVI, respectively (Figure 5A, Supplementary Figure S7A). For a more detailed comparison per gene: SpaGE is performing better on the majority of genes compared to Liger and gimVI, while compared to Seurat, SpaGE has better performance across all genes (Supplementary Figure S7B–D). We further investigated the relation between the prediction performance and the Moran's I statistic of the originally measured genes. Similar to the **STARmap** data, for SpaGE and Seurat, we found a positive relationship, i.e. the performance is higher for genes with distinct spatial patterns. However, Liger and gimVI have a negative relationship (Supplementary Figure S8).

Next, we tested the performance of all methods using the **AllenVISp** dataset as reference for the **osmFISH** dataset, similar to the **STARmap** dataset. For the **osmFISH_AllenVISp** dataset pair, we observed similar conclusions where SpaGE has significantly better performance compared to other methods, with a median Spearman correlation of 0.203 compared to 0.014, 0.082 and 0.162 for Seurat, Liger and gimVI, respectively (Figure 5A, Supplementary Figure S9A). SpaGE has better performance across all genes compared to Seurat and Liger, while gimVI is performing better on a few genes (Supplementary Figure S9B–D). All four methods have a positive relationship between their prediction performance and the Moran's I statistic of the measured genes (Supplementary Figure S10). These results show how the reference dataset can affect the prediction. Compared to the **Zeisel** dataset, the **AllenVISp** is more deeply sequenced data, with the average number of detected transcripts per cell being ∼140× more than the **Zeisel** dataset (Supplementary Figure S11A, B). However, not all methods benefit from this, as for Seurat and Liger, the prediction performance using the **AllenVISp** or the **Zeisel** datasets is quite similar (Figure 5A). On the other hand, SpaGE and gimVI get an increase in performance across all genes, although the median correlation for SpaGE remains the same. Similar to the **STARmap** dataset, we tested the performance of the kNN regression within the **AllenVISp** dataset only (excluding the alignment procedure), when using only the 33 genes of the **osmFISH** dataset. In this case, we obtained a median correlation of 0.289 (Supplementary Figure S4A), when predicting the expression of genes in the scRNA-seq data from one-fold to the other, which is slightly higher than SpaGE (0.203) predicting **osmFISH** patterns. This result shows that the alignment of the spatial and scRNA-seq data using SpaGE is performing well, as the overall performance is comparable with predictions within the same dataset.

While the **AllenVISp** is a deeply sequenced reference dataset, it has been measured from a different brain region than the **osmFISH** dataset (Table 1). Therefore, we decided to use a third reference dataset, **AllenSSp**, which has roughly the same sequencing depth as the **AllenVISp** (Supplementary Figure S11B, C) but is measured from the somatosensory cortex, similar to the **osmFISH** dataset. We evaluated the prediction performance of all four tools for the new dataset pair **osmFISH_AllenSSp**. SpaGE obtained a better performance with a median Spearman correlation of 0.199 compared to 0.006 and 0.077 for Seurat and Liger, respectively, while gimVI has similar performance to SpaGE with a median Spearman correlation of 0.199 (Figure 5A, Supplementary Figure S12A). SpaGE has a better performance across almost all genes compared to Seurat and Liger, while gimVI performed better than SpaGE

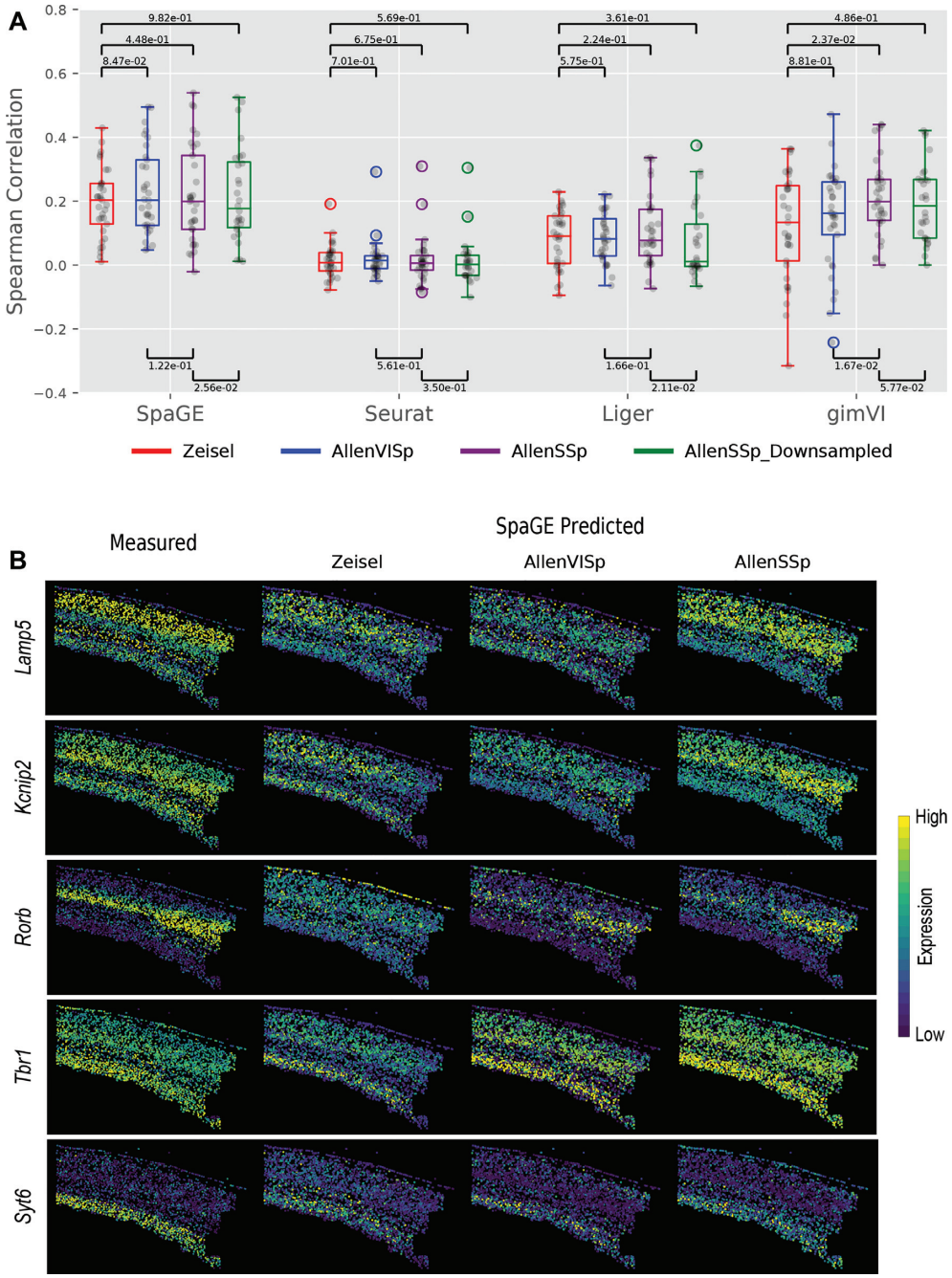

Figure 5. Prediction performance comparison for the osmFISH dataset using different reference scRNA-seq datasets. (**A**) Boxplots showing the Spearman correlations for the leave-one-gene-out cross validation experiment for each method using four different scRNA-seq datasets, **Zeisel**, **AllenVISp**, **AllenSSp** and **AllenSSp_Downsampled**. The median correlations shows a better performance for SpaGE in all dataset pairs. The black dots show the correlation values for individual genes. The *P*-values are obtained using a paired Wilcoxon rank-sum test. SpaGE showed a performance improvement when using the **AllenVISp** over the **Zeisel** data. Although the median correlation is the same, the overall correlation range did improve. Also, gimVI clearly benefits from using the **AllenVISp** and the **AllenSSp** datasets over the **Zeisel** dataset. All methods have decreased performance when using the **AllenSSp_Downsampled** data compared to the original **AllenSSp** data. (**B**) Predicted expression of known spatially patterned genes in the **osmFISH** dataset using different reference scRNA-seq datasets. Each row corresponds to a single gene having a clear spatial pattern. First column from the left shows the measured spatial gene expression in the **osmFISH** dataset, while the second, third and fourth columns show the corresponding predicted expression pattern by SpaGE using **Zeisel**, **AllenVISp** and **AllenSSp** datasets, respectively. Changing from **Zeisel** to **AllenVISp** (deeply sequenced data) improved the prediction, while matching the brain region using the **AllenSSp** improved the prediction further.

for nearly half the genes (Supplementary Figure S12B–D). SpaGE, Liger and gimVI have positive relationship between the prediction performance and Moran's I statistic. However, Seurat has a negative relationship (Supplementary Figure S13).

Several sources of variation do exist between the **Allen** datasets and the **Zeisel** dataset; besides the sequencing depth, these datasets are, for example, generated in different labs and using different sequencing protocols. To separately test the effect of the sequencing depth of the reference scRNA-seq data on the prediction performance, we downsampled the **AllenSSp** dataset to a comparable number of transcripts per cell as the **Zeisel** dataset, using the scuttle R package. Compared to the original **AllenSSp** dataset, we obtained lower prediction performance across all methods when using the downsampled dataset (Figure 5A), clearly showing that a deeply sequenced reference dataset produces a better prediction. Interestingly, compared to the **Zeisel** dataset, the median performance using the downsampled **AllenSSp** dataset was lower for SpaGE, Seurat and Liger, but higher for gimVI.

Changing the brain region did not affect the overall performance of SpaGE (Figure 5A), however, the prediction of genes with known patterns did improve (Figure 5B). When we visually inspect these genes, we can clearly observe that the predicted spatial pattern improved when the reference dataset had a higher sequencing depth, or was obtained from a similar tissue. *Rorb* and *Tbr1* are clear examples, where the prediction using **Zeisel** was almost missing the correct pattern, this became clearer using the **AllenVISp** having a greater sequencing depth. Changing to a matching tissue adds further improves the predicted patterns of these genes (**AllenSSp**). Eventually, all five genes (*Lamp5*, *Kcnip2*, *Rorb*, *Tbr1* and *Syt6*) are more accurately predicted using the **AllenSSp** dataset. Moreover, we used the **AllenSSp** reference dataset to predict the spatial expression of 10 genes not originally measured by the **osmFISH** dataset, with clear patterns through the cortical layers (Figure 6). These predicted patterns are in agreement with the Allen Brain Atlas *in-situ* hybridization (ISH).

SpaGE is scalable to large spatial datasets

So far, SpaGE showed good prediction performance in the leave-one-gene-out predictions, and was also able to predict correct spatial patterns of unmeasured genes within the spatial transcriptomic datasets. All these results were, however, obtained using a relatively small spatial datasets including only a few thousand cells (**STARmap** and **osmFISH**). This opens the question of how does SpaGE scale to large spatial datasets, comparable to the datasets measured nowadays. To assess the scalability of SpaGE, we used a large **MERFISH** dataset with >60,000 cells measured from the mouse brain pre-optic region, and integrated it with the corresponding scRNA-seq dataset published in the same study by **Moffit** *et al.* (4). The **MERFISH_Moffit** dataset pair shares 153 genes on which we applied the same leave-one-gene-out cross validation using all four methods. Similar to the previous results, SpaGE significantly outperformed all other methods (P-value < 0.05, two-sided paired Wilcoxon rank sum test) with a median Spearman correla-

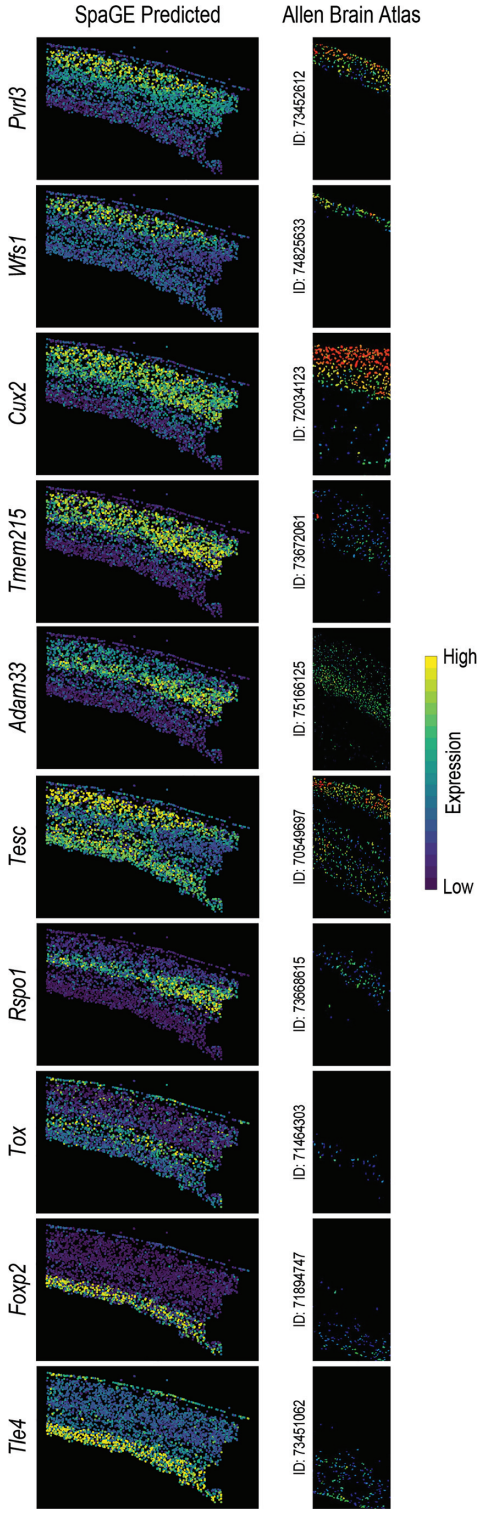

Figure 6. Novel gene expression patterns for 10 genes not originally measured by the osmFISH dataset, validated using the Allen Brain Atlas *in-situ* hybridization ISH. The left column shows the predicted spatial patterns using SpaGE for these 10 genes (shown in rows). The right column shows the Allen Brain Atlas ISH expression for each gene, stating the image ID next to the tissue section, showing an agreement with the expression patterns predicted by SpaGE. These genes show clear expression to specific cortical layers (*Pvrl3* and *Wfs1*: layer 2/3; *Cux2*, *Tmem215* and *Adam33*: layer 2/3 and layer 4; *Rspo1*: layer 4; *Tesc*: layer 2/3 and layer 6; *Tox*: layer 5; *Foxp2* and *Tle4*: layer 6).

tion of 0.275 compared to 0.258, 0.027 and 0.140 for Seurat, Liger and gimVI, respectively (Figure 7A). Per gene comparisons shows a clear advantage of SpaGE versus Liger and gimVI, but more comparable performance with Seurat (Figure 7B–D). The reported *P*-values are quite significant, however, it is important to note that the *P*-values are inflated due to the large sample size, which is also the case for the **STARmap** dataset.

Next to the overall performance across all genes, we evaluated the performance of SpaGE to predict marker genes of four major brain cell types: inhibitory neurons, excitatory neurons, astrocytes and oligodendrocytes (Methods). We observed that SpaGE had higher prediction performance for cell type marker genes compared to the overall performance across all genes (Figure 7E). Similar conclusion can be observed for the **STARmap** dataset (Supplementary Figure S14A), however, this is not the case for the **osmFISH** dataset because almost all 33 genes were cell type marker genes (Supplementary Figure S14B). Additionally, the ranking of the prediction performance across cell types is related to the cell type proportions observed in the data. For instance, the **MERFISH** dataset has approximately 38% inhibitory neurons, 18% excitatory neurons, 15% oligodendrocytes and 13% astrocytes, for which the median correlation per cell type is 0.587, 0.551, 0.402 and 0.398, respectively (Figure 7E). Compared to the pre-optic region, the cortex contains more excitatory neurons than inhibitory. This is directly reflected in the prediction performance of inhibitory and excitatory marker genes, where the latter have higher performance for the cortical datasets **STARmap** and **osmFISH** (Supplementary Figure S14).

Further, we compared the computation times of all four methods across all five dataset pairs. All experiments were run on a Linux HPC server but limited to a single CPU core, with 256GB of memory, to be able to compare runtimes. For all methods, the calculated computation time includes the integration and the prediction time. Overall SpaGE has the lowest average computation time per gene, across all five dataset pairs (Figure 7F). For the large **MERFISH** dataset, SpaGE has a clear advantage compared to the other methods as the average computation time of SpaGE is ∼30×, 63× and 45× faster than Seurat, Liger and gimVI, respectively. In terms of memory, SpaGE has the lowest memory usage across all five dataset pairs, while Seurat and Liger consumed memory the most (Figure 7F). Combined, these results show an overall advantage of SpaGE over other methods for larger datasets with higher prediction performance, lower computation time and less memory requirement.

Increasing the number of shared genes does not always improve the prediction

To investigate whether the performance improves when having many more spatially measured genes, we tested SpaGE when applying it to the **seqFISH+** spatial dataset that measures up to 10,000 genes simultaneously. Using the **seqFISH_AllenVISp** dataset pair, we applied SpaGE using the leave-one-gene-out cross validation setup to predict the spatial expression of 9,751 shared genes. SpaGE produced a median Spearman correlation of 0.154, a minimum correlation of -0.170 and a maximum correlation of 0.716. This result is comparable to the other tested dataset pairs, showing robust performance of SpaGE.

However, with ∼10,000 spatial genes, we expected a better performance as there are many more shared genes with which matching cells can be found in the scRNA-seq data. To further substantiate this, we compared the prediction performance of 494 overlapping genes between the **seqFISH_AllenVISp** and the **STARmap_AllenVISp** dataset pairs, both having the same scRNA-seq reference data. The performance when using the **seqFISH+** data, having ∼10x more shared genes, was significantly higher than when using the **STARmap** data (P-value < 0.05, two-sided paired Wilcoxon rank sum test) (Figure 8A). A detailed comparison per gene shows that the majority of the genes are indeed better predicted in the **seqFISH+** dataset (Figure 8B). However, when comparing the 21 overlapping genes between the **seqFISH_AllenVISp** and the **osmFISH_AllenVISp** dataset pairs, we obtained a contradicting result. The performance when using the **osmFISH** data (only 33 shared genes) was higher than when we used the **seqFISH+** data, for almost all 21 genes for which we could make this comparison (Figure 8C, D).

This opens the question whether having more measured spatial genes (and thus shared genes) is always beneficial to predict the spatial patterns of non-measured genes. To answer that, we performed a downsampling experiment similar to what we did with the **STARmap** data (Methods). We fixed 50 genes as test set and downsampled the remaining genes to sets of the top 10, 30, 50, 100, 200, 500, 1000, 2000, 5000, 7000 and 9,701 (all) highly varying genes as shared genes. The best prediction performance of SpaGE was obtained using 5000 genes, after which the performance decreased (Figure 8E). Apparently, having more genes includes more and more lowly varying, and thus noisy, genes into the matching process, which turns out to confuse the matching process and consequently lower the prediction performance.

DISCUSSION

We demonstrated the ability of SpaGE to enhance spatial transcriptomics data by predicting the expression of unmeasured genes based on scRNA-seq data collected from the same tissue. The ability of SpaGE to produce accurate gene expression prediction highly depends on the alignment part performed using PRECISE, which rotates the principal components of each dataset to produce principal vectors with high one-to-one similarity. Projecting the datasets to the latent space spanned by these principal vectors produces a proper alignment, making a simple kNN prediction sufficient to achieve accurate gene expression estimation.

During the alignment, SpaGE ignores principal vectors with low similarity which excludes uncommon and/or noisy signals. Despite the clear differences in the amount of explained variance for each dataset pair by the final set of principal vectors, SpaGE was able to capture the common sources of variation and produce good predictions of the spatial gene expressions across all dataset pairs. SpaGE captured ∼6% of the variance for the **seqFISH+** dataset that measures ∼10,000 genes spatially, but the majority of which

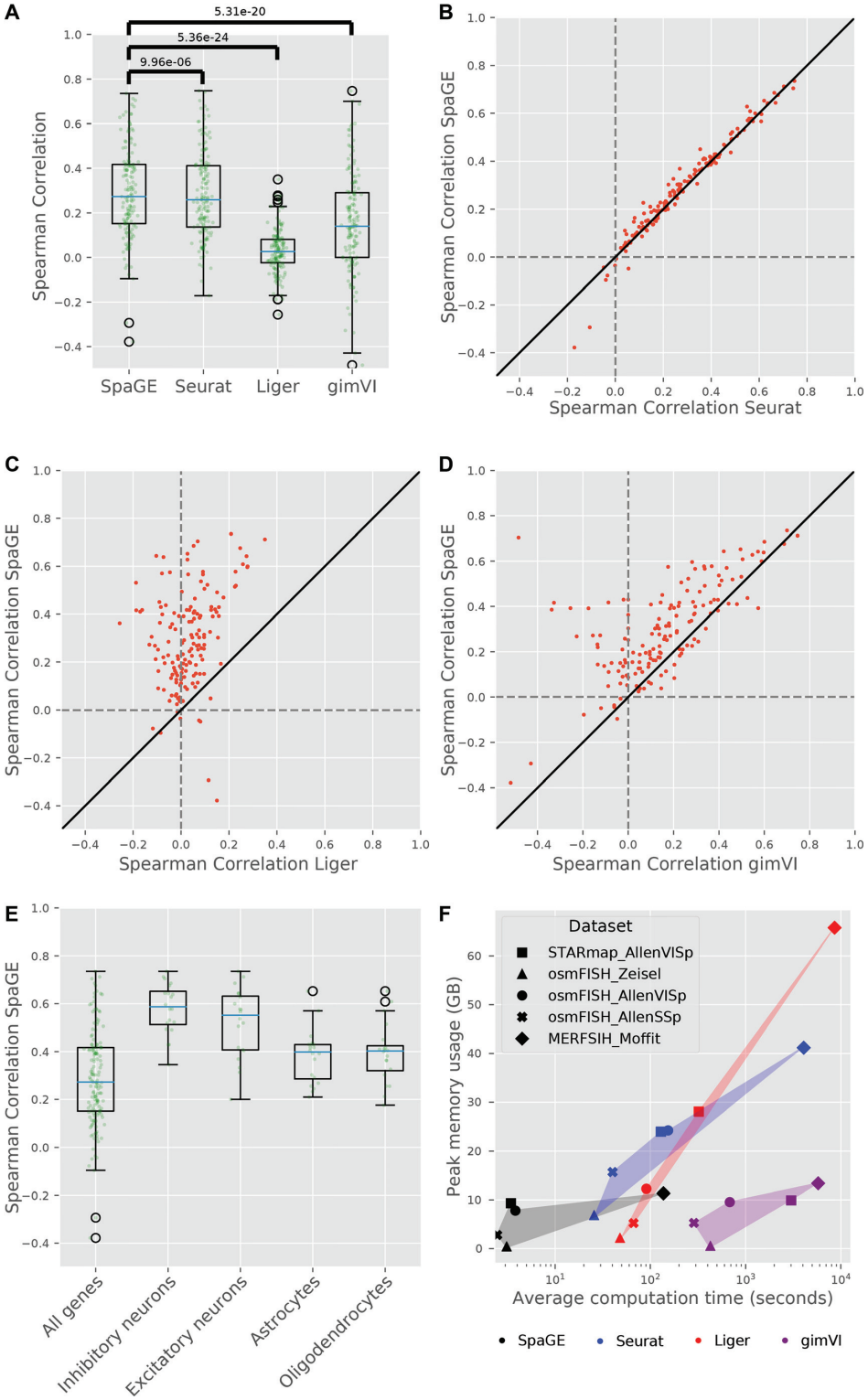

Figure 7. Prediction performance comparison for the MERFISH_Moffit dataset pair. (**A**) Boxplots showing the Spearman correlations for the leave-one-gene-out cross validation experiment for each method. The blue lines show the median correlation across all genes with a better performance for SpaGE. The green dots show the correlation values for individual genes. The *P*-values show the significant difference between all correlation values of SpaGE and each other method, using a paired Wilcoxon rank-sum test. (**B–D**) Detailed performance comparison between SpaGE and (**B**) Seurat, (**C**) Liger, (**D**) gimVI. These scatter plots show the correlation value of each gene across two methods. The solid black line is the $y = x$ line, the dashed lines show the zero correlation. All scatter plots show that the majority of the genes are skewed above the $y = x$ line, showing an overall better performance of SpaGE over other methods. (**E**) Boxplots showing the prediction performance of SpaGE for cell type marker genes compared to the overall performance across all genes. (**F**) scatter plot showing the average computation time (log-scaled) per gene versus the peak memory usage. Methods are represented with different colors and dataset pairs are represented with different symbols. Points of the same method are highlighted for clarity.

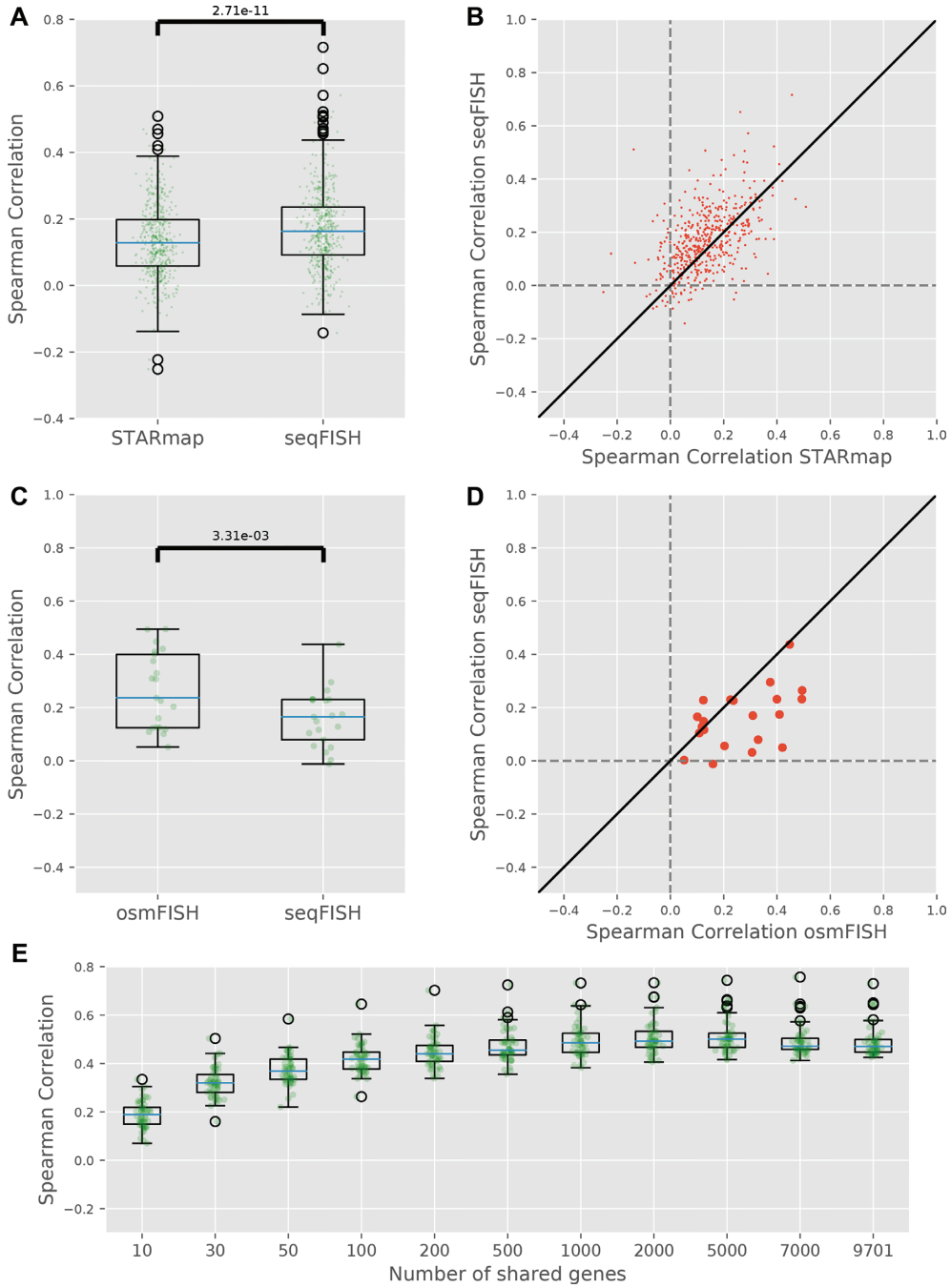

Figure 8. Prediction performance of SpaGE for the seqFISH_AllenVISp dataset pair. (**A,C**) Boxplots comparing the prediction performance of SpaGE for the shared genes between the **seqFISH** and the (**A**) **STARmap**, (**C**) **osmFISH** datasets, using the same **AllenVISp** dataset as reference during prediction. The blue lines show the median correlation across all genes. The green dots show the correlation values for individual genes. The *P*-value is obtained using a paired Wilcoxon rank-sum test. (**B, D**) Detailed performance comparison between **seqFISH** and (**B**) **STARmap**, (**D**) **osmFISH**. These scatter plots show the correlation value of each gene across two datasets. The solid black line is the $y = x$ line, the dashed lines show the zero correlation. (**E**) Boxplots showing the prediction performance of a test set of 50 genes, in terms of Spearman Rank correlations, using downsampled sets of 10, 30, 50, 100, 200, 500, 1000, 2000, 5000 and 7000 shared genes compared to using all 9,701 genes in the **seqFISH_AllenVISp** dataset pair.

are lowly variable in the mouse cortex, thus not contain enough information to contribute to the integration. On the contrary, SpaGE captured ∼94% of the variance for the **osmFISH** dataset, which contains 33 known marker genes for various cell types in the mouse somatosensory cortex. Almost all these genes are highly variable and contain useful information for the integration.

We benchmarked SpaGE against three state-of-the-art methods for multi-omics data integration, using five different dataset pairs. These dataset pairs represent different challenges to the integration and prediction task, as they differ in gene detection sensitivity level and the number of spatially measured genes, which are the basis for the alignment. Increasing the number of shared genes should, in principle, eases the integration task and produces better predictions of spatial patterns of unmeasured genes. However, this is not always the case, as shown by the **seqFISH+** data, where adding more genes eventually also adds genes that have a relatively low variance, and thus are more probably noisy genes. This turns out to negatively influence the matching process and consequently decrease the prediction performance. Apparently, there is an optimum on the number of genes that need to be spatially measured when we want to predict spatial patterns of unmeasured genes. On the other hand, when measuring the spatial patterns measured for ∼10,000 genes, it might not be necessary to predict spatial patterns of unmeasured genes as the initially spatially measured genes already cover most of the transcriptome of interest. Further, imaging-based spatial transcriptomic methods, with high gene detection sensitivity, may also improve the integration and prediction, as they are able to capture the majority of the genes even the ones with relatively low expression. On the other hand, integrating this high sensitivity data with scRNA-seq, which has lower sensitivity, can be more challenging. That is because the differences in gene expression are higher compared to integrating a sequencing-based spatial data with scRNA-seq data, both having comparable sensitivity.

Across all tested dataset pairs, SpaGE outperformed all methods producing better predictions for the majority of the genes. However, for few genes, SpaGE had lower prediction performance than other methods. Seurat produced good gene predictions for the **STARmap** and the **MERFISH** datasets, with similar predictions to SpaGE. However, Seurat had overall the lowest performance for the **osmFISH** dataset, with correlation close to 0, which shows that the performance of Seurat heavily decreased when there are very few shared genes, such as in the **osmFISH** dataset (33 genes). This problem is even more pronounced for Liger, as it performed relatively well for the **STARmap** dataset producing good gene predictions, but has a decreased performance for both the **osmFISH** (33 genes) and the **MERFISH** (155 genes) datasets. On the other hand, gimVI performed relatively well for the **osmFISH** and the **MERFISH** datasets. However, gimVI had overall the lowest performance for the **STARmap** dataset, with inaccurate predictions for genes with spatial patterns such as *Cux2* and *Plp1*. This suggests that gimVI works well with imaging-based technologies having high gene detection sensitivity, but not with the sequencing-based technologies.

Next to the overall best performance, SpaGE is an interpretable algorithm as it allows to find the genes driving the correspondence between the datasets. The principal vectors, used to align the datasets to a latent space, show the contribution of each gene in defining this new latent space. Further, SpaGE is scalable to large spatial data with significantly lower computation time and memory requirement compared to the other methods, as shown on the **MERFISH** dataset having more than 60,000 cells measured spatially. Moreover, SpaGE is a flexible pipeline. Here we used PCA as the initial independent dimensionality reduction algorithm. However, this step can be replaced by any linear dimensionality reduction method.

SpaGE showed high prediction performance for cell type marker genes compared to the overall performance across all genes. These marker genes are often highly variable genes with clear spatial expression patterns. For example, *Cux2* and *Lamp5* represent two excitatory neurons marker genes with clear spatial patterns in the mouse cortex, which were well predicted by SpaGE. We also showed that the cell type proportions directly affect the prediction of the corresponding marker genes. However, the prediction of a marker gene is, in the first place, directly related to the existence of the corresponding cell type across both spatial and scRNA-seq datasets. For example, it is not possible to correctly predict the spatial expression of an astrocyte marker gene, if one or both datasets do not contain any astrocytes. In other words, it is better to measure both spatial and scRNA-seq datasets from the same sample, as we have seen in the **MERFISH_Moffit** dataset pair. However, datasets emerging from different samples but from matching tissue can still produce good spatial gene expression predictions if their cell type compositions are preserved.

We used the Spearman Rank correlation to quantitatively evaluate the predicted gene expressions. The overall evaluation showed relatively low correlations across all methods and all dataset pairs. These low correlations express the difficulty of the problem, as the predicted gene expressions are obtained from a different type of data. Given the low observed correlations, we developed a model that expresses whether we can trust a SpaGE-predicted spatial expression or not, which helps a user of SpaGE to interpret the correlations, improving the practicality of SpaGE. However, the Spearman correlation is not the optimal evaluation metric, as it does not always reflect the spatially predicted patterns, i.e. visual inspection showed good predictions for genes with known spatial pattern in the mouse cortex, while the correlation values were less than 0.2.

CONCLUSION

SpaGE presents a robust, scalable, interpretable and flexible method for predicting spatial gene expression patterns. SpaGE uses domain adaptation to align the spatial transcriptomics and the scRNA-seq datasets to a common space, in which unmeasured spatial gene expressions can be predicted. SpaGE is less complex and much faster when compared to other approaches and generalizes better across datasets and technologies.

DATA AVAILABILITY

The implementation code of SpaGE, as well as the benchmarking code, is available in the GitHub repository, at https://github.com/tabdelaal/SpaGE. The code is released under MIT license. All datasets used are publicly available data, for convenience datasets can be downloaded from Zenodo (https://doi.org/10.5281/zenodo.3967291).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

European Commission of a H2020 MSCA award [675743] (ISPIC); European Union's H2020 research and innovation programme under the MSCA grant agreement [861190 (PAVE)]; NWO TTW project 3DOMICS [NWO: 17126]; ZonMw TOP grant COMPUTE CANCER [40-00812-98-16012]; the collaboration project TIMID [LSHM18057-SGF] financed by the PPP allowance made available by Top Sector Life Sciences & Health (LSH) to Samenwekende Gezondheidsfondsen (SGF) to stimulate public-private partnerships and co-financing by health foundations that are part of the SGF. Partial funding for open access charge: TU Delft Open Access Fund.
Conflict of interest statement. None declared.

REFERENCES

1. Han,X., Wang,R., Zhou,Y., Fei,L., Sun,H., Lai,S., Saadatpour,A., Zhou,Z., Chen,H., Ye,F. *et al.* (2018) Mapping the mouse cell atlas by Microwell-Seq. *Cell*, **172**, 1091–1107.
2. Cao,J., Spielmann,M., Qiu,X., Huang,X., Ibrahim,D.M., Hill,A.J., Zhang,F., Mundlos,S., Christiansen,L., Steemers,F.J. *et al.* (2019) The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, **566**, 496–502.
3. Zeisel,A., Hochgerner,H., Lönnerberg,P., Johnsson,A., Memic,F., van der Zwan,J., Häring,M., Braun,E., Borm,L.E., La Manno,G. *et al.* (2018) Molecular architecture of the mouse nervous system. *Cell*, **174**, 999–1014.
4. Moffitt,J.R., Bambah-Mukku,D., Eichhorn,S.W., Vaughn,E., Shekhar,K., Perez,J.D., Rubinstein,N.D., Hao,J., Regev,A., Dulac,C. *et al.* (2018) Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, **362**, eaau5324.
5. Codeluppi,S., Borm,L.E., Zeisel,A., La Manno,G., van Lunteren,J.A., Svensson,C.I. and Linnarsson,S. (2018) Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods*, **15**, 932–935.
6. Chen,K.H., Boettiger,A.N., Moffitt,J.R., Wang,S. and Zhuang,X. (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, **348**, aaa6090.
7. 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.
8. Wang,X., Allen,W.E., Wright,M.A., Sylwestrak,E.L., Samusik,N., Vesuna,S., Evans,K., Liu,C., Ramakrishnan,C., Liu,J. *et al.* (2018) Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, **361**, eaat5691.
9. 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.
10. 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.
11. Stuart,T. and Satija,R. (2019) Integrative single-cell analysis. *Nat. Rev. Genet.*, **20**, 257–272.
12. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
13. Achim,K., Pettit,J.B., Saraiva,L.R., Gavriouchkina,D., Larsson,T., Arendt,D. and Marioni,J.C. (2015) High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.*, **33**, 503–509.
14. Nitzan,M., Karaiskos,N., Friedman,N. and Rajewsky,N. (2019) Gene expression cartography. *Nature*, **576**, 132–137.
15. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of Single-Cell data. *Cell*, **177**, 1888–1902.
16. Welch,J.D., Kozareva,V., Ferreira,A., Vanderburg,C., Martin,C. and Macosko,E.Z. (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.
17. Korsunsky,I., Millard,N., Fan,J., Slowikowski,K., Zhang,F., Wei,K., Baglaenko,Y., Brenner,M., Loh,P. and Raychaudhuri,S. (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods*, **16**, 1289–1296.
18. Lopez,R., Nazaret,A., Langevin,M., Samaran,J., Regier,J., Jordan,M.I. and Yosef,N. (2019) A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. arXiv doi: https://arxiv.org/abs/1905.02269, 06 May 2019, preprint: not peer reviewed.
19. Tran,H.T.N., Ang,K.S., Chevrier,M., Zhang,X., Lee,N.Y.S., Goh,M. and Chen,J. (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, **21**, 12.
20. Mourragui,S., Loog,M., Van De Wiel,M.A., Reinders,M.J.T. and Wessels,L.F.A. (2019) PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics*. **35**, i510–i519.
21. Gene,H.G and Van Loan,C.F. (2013) In: *Matrix Computations*. 4th edn, Johns Hopkins University Press.
22. Tasic,B., Yao,Z., Graybuck,L.T., Smith,K.A., Nguyen,T.N., Bertagnolli,D., Goldy,J., Garren,E., Economo,M.N., Viswanathan,S. *et al.* (2018) Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, **563**, 72–78.
23. Chatterjee,S., Sullivan,H.A., MacLennan,B.J., Xu,R., Hou,Y.Y., Lavin,T.K., Lea,N.E., Michalski,J.E., Babcock,K.R., Dietrich,S. *et al.* (2018) Nontoxic, double-deletion-mutant rabies viral vectors for retrograde targeting of projection neurons. *Nat. Neurosci.*, **21**, 638–646.
24. Zeisel,A., Môz-Manchado,A.B., Codeluppi,S., Lönnerberg,P., Manno,G.La, Juréus,A., Marques,S., Munguba,H., He,L., Betsholtz,C. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.
25. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
26. Lopez,R., Regier,J., Cole,M.B., Jordan,M.I. and Yosef,N. (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, **15**, 1053–1058.
27. Li,H., Calder,C.A. and Cressie,N. (2007) Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. *Geogr. Anal.*, **39**, 357–375.