Gene expression

SD²: Spatially resolved transcriptomics deconvolution through integration of dropout and spatial information

Haoyang Li[1,2], Hanmin Li[1,2], Juexiao Zhou[1,2], Xin Gao[1,2,*]

[1]Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

[2]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

* To whom correspondence should be addressed: xin.gao@kaust.edu.sa.

Abstract

Motivation: Unveiling the heterogeneity in the tissues is crucial to explore cell-cell interactions and cellular targets of human diseases. Spatial transcriptomics (ST) supplies spatial gene expression profile which has revolutionized our biological understanding, but variations in cell type proportions of each spot with dozens of cells would confound downstream analysis. Therefore, deconvolution of ST has been an indispensable step and a technical challenge towards the higher-resolution panorama of tissues.

Results: Here, we propose a novel ST deconvolution method called SD² integrating spatial information of ST data and embracing an important characteristic, dropout, which is traditionally considered as an obstruction in single-cell RNA sequencing data (scRNA-seq) analysis. First, we extract the dropout-based genes as informative features from ST and scRNA-seq data by fitting a Michaelis-Menten function. After synthesizing pseudo-ST spots by randomly composing cells from scRNA-seq data, auto-encoder is applied to discover low-dimensional and non-linear representation of the real- and pseudo-ST spots. Next, we create a graph containing embedded profiles as nodes, and edges determined by transcriptional similarity and spatial relationship. Given the graph, a graph convolutional neural network is used to predict the cell-type compositions for real-ST spots. We benchmark the performance of SD² on the simulated seqFISH+ dataset with different resolutions and measurements which show superior performance compared with the state-of-the-art methods. SD² is further validated on three real-world datasets with different ST technologies, and demonstrates the capability to localize cell-type composition accurately with quantitive evidence. Finally, ablation study is conducted to verify the contribution of different modules proposed in SD².

Availability: The SD² is freely available in github (https://github.com/leihouyeung/SD2) and Zenodo (https://doi.org/10.5281/zenodo.7024684).

Contact: xin.gao@kaust.edu.sa

1 Introduction

Understanding the arrangement of cells and tissues, and its impact on biological function is a fundamental pursuit in life science research (Method of the Year 2020: spatially resolved transcriptomics 2021). Spatially resolved transcriptomics (Ståhl Patrik. et al. 2016), which aims to characterize the gene expression profiles while retaining information of spatial tissue context, sheds light on the understanding of structure and function of cells and tissues in recent years (Burgess 2019). This technique renders the panorama for the organization and heterogeneity of complex tissues by equipping multimodal data containing gene expression profiles with spatial information by capturing the mRNA population of molecules in situ and the super-resolution histological staining image integrating morphological features (Andrews and Hemberg 2019). Spatial transcriptomics (ST) technique has been utilized for exploring the biological mechanisms among a variety of diseases, tissues and species, such as human heart (Asp et al. 2019), mouse brain (Cantin et al. 2021), Alzheimer's disease (W.-T. Chen et al. 2020) and so on.

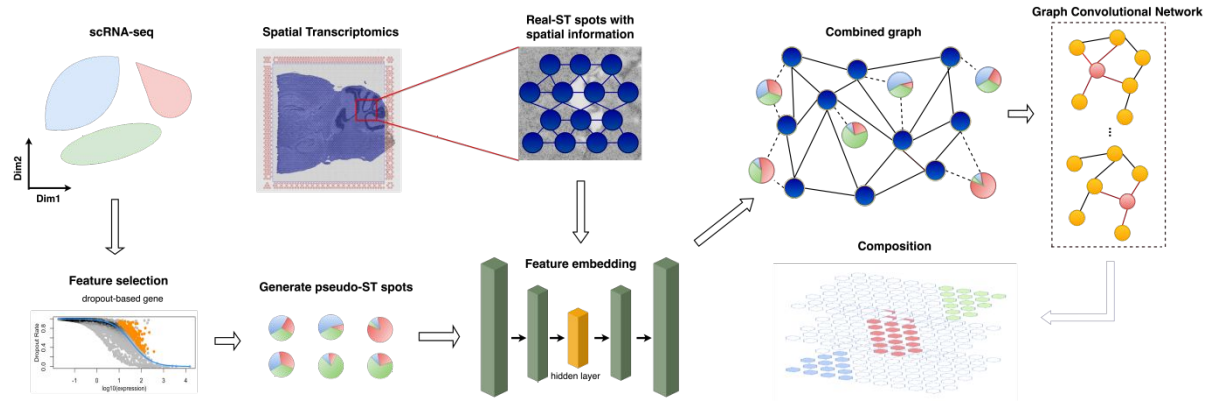

Figure 1 The pipeline of SD². First, we extract the dropout genes by fitting a Michaelis-Menten function and generate the pseudo-ST spots. Then we extract the embedded feature by AE from pseudo-ST and real-ST spots. The pseudo-ST and real-ST spots are constructed as a graph by transcriptional similarity and spatial connection. Finally, graph convolutional neural network is used to output the cell-type composition for real-ST spots.

The cellular composition of biological samples is heterogeneous and varying inherently. Characterizing the variation of cell-type composition across subjects could identify cellular targets of diseases. On the other hand, adjusting for these variations could also clarify the cell-cell interactions and reconstruct the topological and spatial distribution of all cell types (Czerwińska 2018, Avila Cobos et al. 2020, R. Dong and Yuan 2021). The traditional way of bulk sequencing may confound downstream data analysis because less abundant cell types will be masked by that of more abundant ones (M. Dong et al. 2021, Jin and Liu 2021). The ability to measure the cellular heterogeneity under specific conditions is therefore critical. However, in the original few generations of ST techniques, the resolution of ST data is much lower than the single-cell level. For instance, the 10X Visium, a commonly-used ST technique developed by 10X Genomics, utilizes the spots with 50 μm diameter containing 10-20 cells on average. Thus, unveiling the mixture of cells in the ST spots is a key to depict the precise panorama for the tissues and advance better understanding of the precise tissue organization.

To tackle this problem, several methods have been proposed. SPOTLight (Elosua-Bayes et al. 2021) integrates ST and single-cell RNA sequencing (scRNA- seq) data to infer the cell types of spots in the tissue by seeded non-negative matrix factorization (NMF) regression and non-negative least squares (NNLS) to subsequently deconvolute ST spots. The performance of SPOTLight shows that it could return accurate predictions with shallow sequenced references. DSTG (Su and Song 2020) utilizes the graph-based model to accurately deconvolute the ST spots and reveal the spatial architecture of cellular heterogeneity in tissues. DSTG has great quantitative performance in benchmarking scRNA-seq of different protocols and identifying cellular heterogeneity in mouse cortex layer, hippocampus slice and pancreatic tumor tissues. Cell2location (Kleshchevnikov et al. 2020), a principled and versatile Bayesian model, integrates the scRNA-seq and ST data to map cell types in situ in a comprehensive manner. The applications on several data sets demonstrate that cell2location could serve as a versatile first-line analysis tool to map tissue architectures. Despite the technical advances, these state-of-the-art methods did not utilize the spatial information among the spots and the nonlinear relationship behind various genes. In addition, dropout has been considered as non-informative component by these methods, although both ST and scRNA-seq data show the nature of containing extremely high levels of dropout which may contribute to the deconvolution of ST.

Here we propose a novel method called SD², which integrates spatial information and embraces an important characteristic, traditionally considered as an obstruction in scRNA-seq analysis, called dropout information, through graph convolutional networks (GCN). In SD², scRNA-seq data with cell-type annotations are complementary resource to generate pseudo-ST data. SD² explicitly utilizes the dropout-based genes of scRNA-seq and ST data, as well as spatial information of all spots as additional information. Nonlinear relationship among dropout genes is revealed by auto-encoder (AE) to extract low-dimensional representations of the real- and pseudo-spots. Comprehensive benchmark and real-world experiments demonstrate not only the accuracy of our method over other methods, across various resolutions and measurements, but also the capability to localize the cell types accurately on different ST techniques from mouse brain, mouse kidney and human pancreatic tumor tissues. The utility of dropout information is thus guaranteed as an essential role in the deconvolution task. We further conduct ablation studies to evaluate the importance of each proposed component.

2 Methods

2.1 Pre-processing

For scRNA-seq data, dropout has been treated as an obstruction to be tackled. However, inspired by (Qiu 2020), we leverage dropout as an informative pattern to extract dropout-based genes from ST and scRNA-seq data instead of highly-variable genes (HVGs). The input data are ST expression profile $R \in \mathbb{R}^{n_r \times g_r}$ and scRNA-seq expression profile $S \in \mathbb{R}^{n_s \times g_s}$ where n_r (n_s) and g_r (g_s) represent the number of spots and genes respectively. We analyze the distribution of dropout rates in R and S, and find that most of dropout rates are over 80% and some of them are even close to 100% (Supp. Figure 1A, B). We identify the dropout-based genes by M3Drop (Andrews and Hemberg 2019). Through all spots in R and cells in S, a Michaelis-Menten function is fitted to the relationship between mean expression (E) through cells (spots) and dropout-rate ($P_{dropout}$) for each specific gene and the gene-specific parameter K_i is estimated by this function. Through the following equation, the global parameter K_M is optimized by maximum likelihood estimation across all genes.

$$P_{dropout} = 1 - \frac{E}{E + K_M}. \#(1)$$

The hypothesis is that gene-specific K_i is equal to K_M. After estimating the deviation error from K_i to K_M, significance of each K_i is evaluated by *t-test*. Finally, *top-k* significant genes g_r^d and g_s^d are chosen for R and S, and we use their expression profiles as the informative features for each spot (cell). The $g_d = g_r^d \cap g_s^d$ are intersected genes considered as the

SD²

preserved features for scRNA-seq and ST data, thus $R \in \mathbb{R}^{n_r \times g_d}$ and $S \in \mathbb{R}^{n_s \times g_d}$.

dimension g_a as the attributes of each spot after the training process. Thus, $R' = \psi(R) \in \mathbb{R}^{n_r \times g_a}$ and $P' = \psi(P) \in \mathbb{R}^{n_p \times g_a}$, where

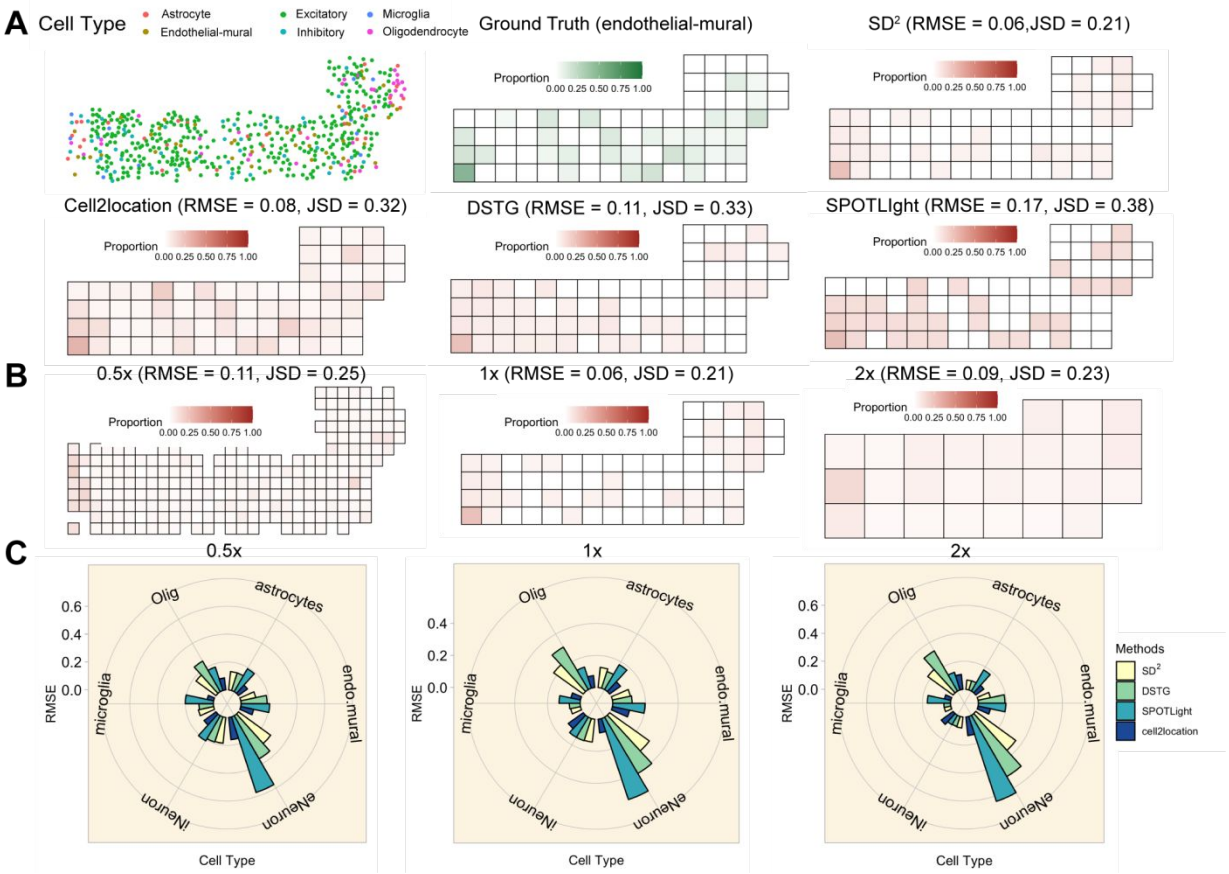

Figure 2 A. The figure at the top left corner shows the spatial and cell-type distribution of the seqFISH+ dataset. The top middle figure shows the ground truth of abundance of endothelial-mural cells in the simulated dataset. The other four figures show the corresponded deconvolution results for endothelial-mural cells from four methods. **B**. The three spatial distributions of endothelial-mural cells deconvolved by SD² are shown through three kinds of resolution (0.5x, 1x and 2x). **C.** The three radial column figures compare the RMSE of deconvolution for each cell type on these four methods through three kinds of resolution.

To mimic the gene expressional distribution of real-ST spots better forconstructing a homogeneous graph, we generate pseudo-ST spots by annotated scRNA-seq data. Inspired by the concept of Markov Chain Monte Carlo sampling (van Ravenzwaaij, Cassey, and Brown 2018), we hope that distribution of more sampled pseudo-ST spots could match the distribution of real-ST spots better. We randomly synthesize m cells as a new pseudo-ST spot whose expression vector would be $\sum_r^m S_r / m$ and the proportion of specific cell-type c is m_c/m where m_c denotes the number of selected cells in c. The selection of m could refer to the resolution of real-ST spots. For now, pseudo-ST profiles would be $P \in \mathbb{R}^{n_p \times g_d}$, where n_p denotes the number of pseudo-ST spots. Then, P and R are divided by the library size through the spots and multiplied by a size factor of 10000 for normalization. Through generating pseudo-ST spots, the density distribution of counts in pseudo-ST is closer to real-ST than scRNA-seq which means that pseudo-ST generation could mimic the pattern of real-ST better than using individual cells directly (Supp. Figure 1C).

To preserve nonlinear relationships in the gene expression profiles of all spots, we utilize the AE as the embedding method of the gene expression vector (Dwivedi et al. 2020, Eraslan et al. 2019). We train a three-layer AE network to reconstruct the concatenation of R and P, and minimize the reconstruction error with the encoder ψ and decoder ϕ following Equation 2. Then, we capture the embedded space with

$$\phi, \psi = argmin_{\phi,\psi} MSE([R, P], (\psi \circ \phi)[R, P]). \#(2)$$

2.2 Graph construction

Now, we integrate all real- and pseudo-ST spots as nodes into an unweighted graph. For the edges of the graph, we define them at the transcriptional and spatial level. At the transcriptional level, we identify the mutual nearest neighbors between the spots from pseudo-ST and real-ST data. If a pseudo-ST spot s_p is among the top-k nearest neighbors of the real-ST spot s_r calculated by the *K-Nearest Neighbor (KNN)* algorithm and *vice versa*, we would define that there is a connected unweighted edge between s_p and s_r. k is used to control the sparsity of the graph. This kind of connection preserves the transcriptional similarity between pseudo-ST and real-ST spots. Under the assumption that the expression level among adjacent spots tend to be similar, pseudo-ST data is useful for us to unveil the composition of cell types in the adjacent real-ST spots. At the spatial level, we define the edges among the real-ST spots relying on their spatial coordinates. We set the spot-to-spot horizontal distance as h and the relative coordinates of one specific spot is (x,y), the four nearest neighbors for that spot would be: $(x - h/2, y - h/2)$, $(x - h/2, y + h/2)$, $(x + h/2, y + h/2)$ and $(x + h/2, y - h/2)$. We link these four nearest neighbors to that spot respectively as four unweighted edges and apply

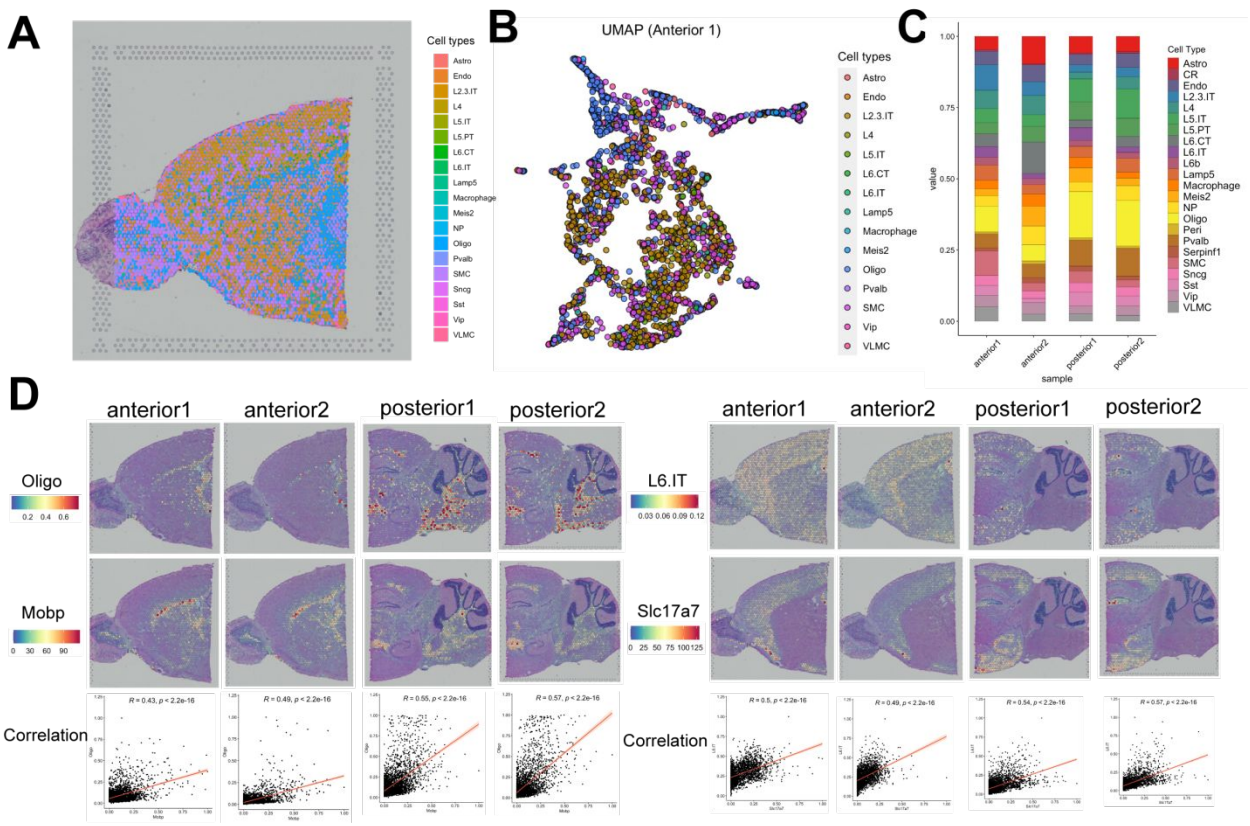

Figure 3 A. The deconvolution results of one spot with all cell types for adult mouse brain. In the figure, each little pie chart indicates a spot with different compositions of cell types noted by different colors. **B.** UMAP figure of all spots with their highest-content cell type for each spot. Each color represents a specific cell type and all spots are mostly separated following the nature of expression profiles. **C.** The abundance of all cell types through two anterior slices and two posterior slices. **D.** Comparison between the abundance of two pairs of cell types and their marker genes: Oligo and Mobp, L6.IT and Slc17a7. The visualized results and Pearson correlation through four slices all show great relationship of them.

this process to all real-ST spots. This spatial connection from real-ST spots could enrich more real-world information for the topological structure of the graph. Finally, the linked graph is

$$X = [P', R'] \in \mathbb{R}^{N \times g_a}, N = n_p + n_r \#(3)$$

which preserves both spatial and transcriptional structure among all spots, and its adjacency matrix $A \in \mathbb{R}^{N \times N}$ would be

$$A_{ij} = \begin{cases} 1, if\ i \sim j, \\ 0, otherwise. \end{cases} \#(4)$$

2.3 SD²

As mentioned above, adjacent spots tend to have similar gene expression patterns. We could unveil the cell-type composition of real-ST spots by utilizing the adjacent pseudo-ST spots with known cell-type composition, which could be formulated as a semi-supervised problem and it comes to GCN as a proper solution. GCN also has the ability for aggregating the graph signals within the node neighborhood which shows capabilities to learn the graph representations and achieves superior performance in a wide range of tasks and applications (Chen et al. 2020). GCN was originally used in the node classification problem (Kipf and Welling 2016). In that task, the output would contain the probabilities of all classes for each node which would be classified as the class with the highest probability. Since probabilities of all classes for each node are added up to 1 after that softmax activation function, here we treat the classes as cell types and the output probabilities as the composition of all cell types in our case (Figure 1).

The input data of SD² are expression matrix X and it's adjacency matrix of the constructed graph A. These input data would be fed into GCN with three convolutional layers. To preserve the information of the nodes themselves and train the network more efficiently, the new adjacency matrix is defined as $\tilde{A} = \hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}$, where $\hat{A} = A + I$ and $\hat{D}$ is the degree matrix of $\hat{A}$. Each layer of GCN could be defined as

$$H^{(l+1)} = f(H^{(l)}, \tilde{A}) = \sigma(\tilde{A}H^{(l)}W^{(l)}) = ReLU(\tilde{A}H^{(l)}W^{(l)}), \#(5)$$

where $H^{(l)}$ is the previous layer, $W^{(l)}$ is the weight of layer l and we use ReLU as the activation function after each graph convolutional layer. At the output of three convolutional layers, we use softmax as the activation function to normalize the output into the range of $[0,1]$. Through the network, the cross-entropy is used as the loss function. The output matrix $C = [C_p, C_r] = (c_{tn}) \in \mathbb{R}^{T \times N}$ represents the cell-type composition matrix of all pseudo-ST and real-ST spots (N spots totally) through T cell types and the composition of cell type t in spot n would be c_{tn}. The sum of vector $\sum_{t=1}^{T} c_{tn}$ would be 1 in each spot. Thus, final output of SD² would be matrix C_r.

3 Experiments

3.1 Experimental setup

During the selection of dropout-based genes, we selected top-2000 significant dropout-based genes and used their expression profiles as the feature for each spot. In the process of generating pseudo-ST spots, we

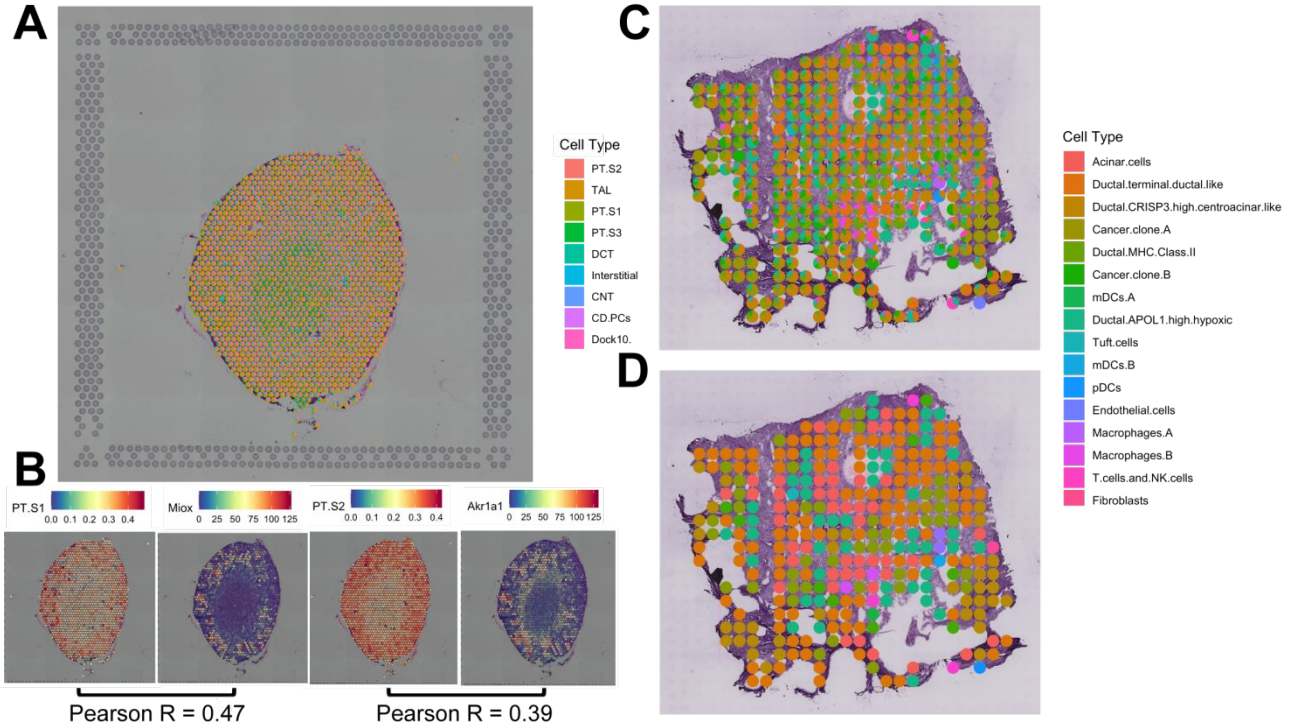

Figure 4 A. The deconvolution results of mouse kidney. **B.** Comparison between deconvolution results of two cell types (PT S1 and PT S2) with their marker genes' (Miox and Akr1a1) expression profiles. The Pearson correlation show great relationship between them. **C.** The deconvolution results of PDAC. **D.** The low-dimensional deconvolution results where each little pie chart indicates a spot with the highest-content cell type assigned by its specific color.

randomly mixed 10 to 20 cells for each pseudo-ST spots. In AE, we set the dimension of the hidden layer as 200, the batch size as 300 and epoch as 5. In the *KNN* algorithm, we set k as 100. We split the pseudo-ST data into 80 % as the training set, 10% as the validation set and 10% as the test set. As SD² is an approach of inductive learning, the training, validation and test set are all utilized in the training process. The validation set is used for preventing overfitting where we set the epoch number of early stopping as 10 and the test set is used for evaluating the performance. Under the grid search technique for hyperparamter optimization, we used Adam optimizer (Kingma and Ba 2014) with the learning rate as 0.005 among 0.001, 0.005 and 0.01, and training epochs as 200 among 100, 200 and 300. All the experiments were implemented on the 2 Quadro M6000 GPUs in Ubuntu 18.04 operating system.

3.2 Benchmark evaluation

For evaluating the performance of SD², we designed quantitive experiments under different resolution of spots and measurements by synthesizing the cells from seqFISH+ dataset (Eng et al. 2019) and MERFISH dataset (Moffitt et al. 2018) to mimic the spots of ST.

seqFISH+ dataset contains 523 cells with corresponding cell-type annotations and spatial coordinates from the cortex of mouse brain at the single-cell resolution. The 6 cell types of the seqFISH+ dataset are clustered and annotated by the cell type reference of the scRNA-seq dataset (Amit et al. 2015) from the same tissue: excitatory neurons (eNeuron), inhibitory neurons (iNeuron), astrocytes, oligodendrocytes (Olig), microglia cells, and endothelial-mural cells (endo_mural) (Figure 2A). To simulate the low-resolution profiles, we devided these cells by multiple squares and considerd one square as a simulated spot with ground truth of cell-type proprotions. We designed three side lengths of squares to simulate different resolution of spots (25.75 μm (0.5x), 51.5 μm (1x) and 103 μm (2x)) whose resolutions are 1.2, 4.5 and 15.7 cells in each spot averagely.

To benchmark the performance of SD², we compared with other three methods: SPOTLight (Elosua-Bayes et al. 2021), DSTG (Su and Song 2020) and cell2location (Kleshchevnikov et al. 2020). We chose to visualize the distribution of proportion of the endothelial-mural cells which has local and distinct pattern of composition among SD² and the other three compared methods. Through the comparsion with ground truth, we observed that SD² outperformed the other three methods in root mean square error (RMSE) and Jensen-Shannon Divergence (JSD). In particular, SD² has RMSE of 0.06 and JSD of 0.21 which shows the most similar pattern compared with ground truth (Figure 2A). We also visualized the proportion of endothelial-mural cells deconvolved by SD² through 0.5x, 1x and 2x resolutions (Figure 2B). The visualization and corresponding RMSE showed the consistent outperformace over SOTA which demonstrated the robustness of SD². To identify the ability of deconvolution for each cell type, we showed three radial column figures to compare the RMSE of deconvolution for each cell type on these four methods through three kinds of resolution (Figure 2C). SD² performed better than other three methods for the most cell types and showed the robustness through three kinds of resolutions. We also testified the computational time through these methods and the histogram through the different numbers of genes in ST data (3000, 6000 and 10000) showed the most efficiency of SD² compared with the other three methods (Supp. Table 1).

We also evaluated the performance of SD² and three compared methods on the MERFISH datasets of mouse brain medial pre-optic area containing 12 samples from posterior to anterior. MERFISH datasets have 135 genes and 59651 cells classified by 6 cell types which show the opposite

Table 1. The ablation study compared the original SD² and the other three conditions: no spatial connection, no dropout-based feature and no AE. The experiments were conducted through three resolutions of spot and measured by RMSE and JSD score. The results showed the contribution and necesscity of these proposed modules for SD².

Methods	0.5x		1x		2x	
	RMSE	JSD	RMSE	JSD	RMSE	JSD
SD²	**0.188**	**0.198**	**0.169**	**0.216**	**0.191**	**0.153**
No spatial connection	0.301	0.489	0.181	0.235	0.233	0.255
No dropout-based feature	0.280	0.431	0.188	0.245	0.228	0.237
No AE	0.315	0.482	0.182	0.230	0.238	0.253

condition of number of genes and spots compared with seqFISH+ dataset, which is helpful to testify the robustness of SD² in these different extreme situations. As the simulation procedure in seqFISH+ dataset, we binned the square of 100×100 cells as one spot in MERFISH datasets and 3067 spots were simulated with their real cell-type proportions. We visualized the deconvolution results of SD² and three compared methods in all 12 samples and calculated the average RMSE and JSD of them where SD² still achieved great visualized and quantitive performance compared with the other three methods (Supp. Figure 4).

3.4 Evaluation on real-world data

To examine the performance of SD² on the real-world data, we collected three data sets: adult mouse brain, mouse kidney and pancreatic ductal adenocarcinoma (PDAC) (Supp. Table 2).

The cortex of mouse brain is partitioned into multiple subcortical and other cortical regions. The isocortex and hippocampal formation in the mammalian brain could greatly affect the function of perception, cognition, emotion, and learning (Van Essen and Glasser 2018, Rakic 2009). To explore the heterogeneity of the adult mouse brain, we collected ST data of two anterior and two posterior brain slices from 10X Genomics generated by 10X Visium technique and the scRNA-seq data set generated by Smart-seq from the Allen Institute (Yao et al. 2021), which consists of around ten thousand cells in adult mouse cortex and hippocampus tissue with 22 cell types. The number of cells could supply various selections for generating the pseudo-ST spots. We also mapped the deconvolution results with spatial coordinates on the original tissue image. After conducting SD², the deconvolution results of four slices from anterior brain and posterior brain were shown (Figure 3A, Supp. Figure 2).

We outputted the cell-type composition of all spots where each little pie chart indicates a spot with different composition of cell types noted by different colors. The cell type whose content in one spot is lower than 5% was deleted from this spot, because the maximum cell number is 20 in a spot by 10X Visium technique, which means that the expected number of such cells is lower than 1. We also visualized the two-dimensional projection of all spots with their highest-abundant cell type by Uniform Manifold Approximation and Projection (UMAP) (Figure 3B). The UMAP results show that the spots from different colors (cell types) were mostly separated and the spots from the same coloirs were aggregated together even with the low-resolution of cell-type composition which means that SD² could deconvolve the spots well following the cell-type-specific gene-expression nature. In order to compare the change of spatial organization of cell type composition quantitively, we examined abundance of all cell types through four adjacent slices and the smoothness of all adundance of all cell types are shown in the Figure 3C which indicated the steady results of SD² through multiuple adjacent slices.

We further assumed that the expression pattern of cell-type specific marker genes could reflect to the spatial distribution of cells in that specific cell type, which could also be used to compare with the distribution of the deconvolved cell types by SD². Based on this assumption, we extracted two pairs of cell types and their marker genes: Oligo and Mobp (Holz and Schwab 1997), and L6 and Slc17a7 (Hodge et al. 2019). Then, we calculated the Peason correlation and corresponding *p-value* to verify the relationship between cell-type distribution and marker gene's expression. The results demonstrated the trustiness of SD² that through two anterior and two posterior slices, the Pearson correlation ranged from 0.4 to 0.6 steadily and significantly (Figure 3D). The visualized results of two pairs of cell type composition and the distribution of their marker genes' expression profiles were also matched closely.

We next conducted the experiments on the adult mouse kidney. The kidney maintains fluid, electrolyte, and metabolite balance of the body and plays an essential role in blood pressure regulation, red blood cell homeostasis and injury response. It is thus important to understand the cell-type heterogeneity of mouse kidney (Miao et al. 2021, Reidy et al. 2014). We collected scRNA-seq data generated by snATAC-seq from adult mouse kidney including 16119 cells with 14 cell types (Miao et al. 2021).

The ST slice of kidney was from 10X Genomics generated by 10X Visium technique whose resolution was 10-20 cells for each spot. We also outputted the cell-type composition of each spot (Figure 4A) which showed that PT S2 and PT S1 are the top two highest abundant cell types. We further compared two pairs of abundance of cell types (PT S1 and PT S2) and their specific marker genes' expression profiles (Miox and Akr1a1) and the results showed great Pearson correlation as 0.47 and 0.39 (Figure 4B). We then visualized the UMAP results for all spots with its highest-content cell type (Supp Figure 3A). From the visualization of UMAP and the spatial deconvolution results, we could observe that the entropy of cell-type distribution in mouse kidney is higher than that in the mouse brain which means that the cells with the same cell type in mouse kindey are not assembled together.

PDAC is a highly devastating and heterogenic disease for human with poor prognosis and rising incidence (Orth et al. 2019). It is the most prevalent neoplastic disease of the pancreas accounting for more than 90% of all pancreatic malignancies whose 5-year overall survival is less than 8% (Siegel, Miller, and Jemal 2018) (Kleeff et al. 2016). The ST data of PDAC were generated by the original spatial transcriptomics method (L. et al. 2016) with the low resolution of nearly 10 to 40 cells for each spot. We collected two ST data sets of PDAC. Paired scRNA-seq data were generated from the pancreatic adenocarcinoma tissue by the InDrop technique (Moncada et al. 2020). PDAC has 1927 cells with 21 cell types. We conducted the experiments on ST data with the corresponding scRNA-seq data and showed the deconvolution results mapping with spatial coordinates and its low-resolution deconvolution results which meant the highest-abundant cell type of each spot was perserved (Figure 4C, 4D). We also visualized the UMAP results for all spots with the highest-abundance cell type to observe the data separation among these cell types. Despite the low-resolution of these spots, we could still find the relative isolate patterns in UMAP results (Supp. Figure 3B). In the results of deconvolution, the cells with the same cell type were aggregated smoothly.

3.5 Ablation study

We conducted ablation study to explore the contribution of different modules in SD² on the previous simulated seqFISH+ dataset. We considered four conditions: the original SD²; no spatial connection in real ST; no dropout genes; no AE. To replace the dropout-based genes, we used the HVGs as the feature for each spot by Analysis of Variance (ANOVA). For no spatial connection, we only conisdered the connection between pseudo-ST and real-ST spots. To replace the AE for embedding features, we used a linear dimensional reduction method called singular value decomposition and extracted the largest eigenvalues as the final features for nodes. To evaluate the contribution of each module comprehensively, the experiments were conducted in three kinds of resolutions (0.5x, 1x and 2x) and two metrics were used in each experiment: JSD score and RMSE (Table 1). Through the performance of the ablation study, RMSE and the JSD score were both obviously increased through the elimination of our proposed modules which means that each proposed module contributed to the success of SD².

To explore the effectiveness of dropout genes, we also designed the experiements on seqFISH+ datasets by three conditions (dropout genes only, HVGs only and both of them) in three numbers of total used genes (3000, 6000 and 10000 genes) (Supp. Table 3). The results with different conditions revealed that dropout genes achieved better performance in RMSE and JSD than HVGs steadily in different number of total genes. The combination of dropout genes and HVGs did not show lower RMSE and JSD in most of number of total genes which meant usage of HVGs may be redundant selected features in our datasets.

4 Discussion and conclusion

In this paper, we proposed a method called SD² for ST deconvolution, which follows the nature of ST to leverage dropout and spatial information in the paired scRNA-seq and ST data. Our comprehensive experiments demonstrate that SD² reaches great quantitive and visible performance under different resolutions, ST techniques and metrics. The comparison with three state-of-the-art methods also shows the superior performance of SD². Based on the transcriptional connection in ST data, we explore the real connection among spots by adding spatial information in ST data under the assumption that adjacent spots tend to have similar gene expression patterns. Under the high dropout rates among ST data, dropout-based shows more utility than highly-variable-based on the selection of informative genes and we aim to demonstrate a new perspective that dropout could also play an essential role instead of an obstruction during the analysis of ST data.

For the success of using dropout genes in ST data, it results from several reasons: (1) mRNA expressions in ST are much sparser than scRNA-seq where dropout genes could play a more essential role than HVGs (Supp. Figure 1). As the figure shows, the dropout rates of scRNA-seq datasets are around 80% to 90%, but over 90% in ST datasets. The ST datasets has too few counts to select HVGs for deconvolution task. (2) The selected HVGs could be sensitive to the preprocessing procedure (such as normalization and imputation) which causes the biased selection of HVGs. On the other hand, some informative genes may not be highly variable in the expression profile. (3) Mathematically, the selection strategy of dropout genes we used assumed that for all genes, dropout rates and mean expression profiles had a non-linear relationship (Michaelis-Menten function) and the outlier genes could be informative to be features. This strategy is only suitable for high dropout rate, such as scRNA-seq and ST data. On the other hand, strategy of HVGs selection could be more suitable for sequencing techniques with high sequence depth, such as RNA-seq. In our study, we validated that dropout genes could be informative features for deconvolution task and they should also be utilized for exploring cell-cell interaction, cell-type clustering or trajectory inference tasks further.

Despite the success of the proposed method, there are still some limitations to be overcome. First, more and more methods begin to use scRNA-seq data to supply more fine-grained transcriptional information. But scRNA-seq and ST data could not be truly gathered from the same tissue sections which means that there must be some inconsistency which affects the deconvolution results for ST data. Second, SD² could be further improved by utilizing the distribution of all scRNA-seq data for assistance. Selecting scRNA-seq data randomly to generate pseudo-ST spots would miss some low-abundance cell types or lose the biological significance. In future work, we would try to solve this issue by narrowing down the scale of selection to generate more realistic spots.

Funding

This work was supported by the Office of Research Administration (ORA) at King Abdullah University of Science and Technology (KAUST) under award number BAS/1/1624-01, FCC/1/1976-23-01, URF/1/4077-01-01, URF/1/4098-01-01, REI/1/4216-01-01, REI/1/4437-01-01, REI/1/4473-01-01, URF/1/4352-01-01, and REI/1/4742-01-01.

Conflict of Interest: none declared.

References

Amit, Zeisel et al. 2015. "Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-Cell RNA-Seq." *Science* 347(6226): 1138–42. https://doi.org/10.1126/science.aaa1934.

Andrews, Tallulah S, and Martin Hemberg. 2019. "M3Drop: Dropout-Based Feature Selection for ScRNASeq." *Bioinformatics* 35(16): 2865–67. https://doi.org/10.1093/bioinformatics/bty1044.

Avila Cobos, Francisco et al. 2020. "Benchmarking of Cell Type Deconvolution Pipelines for Transcriptomics Data." *Nature Communications* 11(1): 5650. https://doi.org/10.1038/s41467-020-19015-1.

Chen, Fenxiao, Yun-Cheng Wang, Bin Wang, and C-C Jay Kuo. 2020. "Graph Representation Learning: A Survey." *APSIPA Transactions on Signal and Information Processing* 9.

Dong, Meichen et al. 2021. "SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References." *Briefings in Bioinformatics* 22(1): 416–27. https://doi.org/10.1093/bib/bbz166.

Dong, Rui, and Guo-Cheng Yuan. 2021. "SpatialDWLS: Accurate Deconvolution of Spatial Transcriptomic Data." *Genome Biology* 22(1): 145. https://doi.org/10.1186/s13059-021-02362-7.

Dwivedi, Sanjiv K, Andreas Tjärnberg, Jesper Tegnér, and Mika Gustafsson. 2020. "Deriving Disease Modules from the Compressed Transcriptional Space Embedded in a Deep Autoencoder." *Nature Communications* 11(1): 856. https://doi.org/10.1038/s41467-020-14666-6.

Elosua-Bayes, Marc et al. 2021. "SPOTlight: Seeded NMF Regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkab043.

Eng, Chee-Huat Linus et al. 2019. "Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA SeqFISH+." *Nature* 568(7751): 235–39. https://doi.org/10.1038/s41586-019-1049-y.

Eraslan, Gökcen et al. 2019. "Single-Cell RNA-Seq Denoising Using a Deep Count Autoencoder." *Nature Communications* 10(1): 390. https://doi.org/10.1038/s41467-018-07931-2.

Van Essen, David C, and Matthew F Glasser. 2018. "Parcellating Cerebral Cortex: How Invasive Animal Studies Inform Noninvasive Mapmaking in Humans." *Neuron* 99(4): 640–63.

Hodge, Rebecca D et al. 2019. "Conserved Cell Types with Divergent Features in Human versus Mouse Cortex." *Nature* 573(7772): 61–68.

Holz, A, and M E Schwab. 1997. "Developmental Expression of the Myelin Gene MOBP in the Rat Nervous System." *Journal of Neurocytology* 26(7): 467–77. https://doi.org/10.1023/A:1018529323734.

Jin, Haijing, and Zhandong Liu. 2021. "A Benchmark for RNA-Seq Deconvolution Analysis under Dynamic Testing Environments." *Genome Biology* 22(1): 102. https://doi.org/10.1186/s13059-021-02290-6.

Kingma, Diederik P, and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." *arXiv preprint arXiv:1412.6980*.

Kipf, Thomas N, and Max Welling. 2016. "Semi-Supervised Classification with Graph Convolutional Networks." *arXiv preprint arXiv:1609.02907*.

Kleeff, Jorg et al. 2016. "Pancreatic Cancer." *Nature reviews Disease primers* 2(1): 1–22.

Kleshchevnikov, Vitalii et al. 2020. "Comprehensive Mapping of Tissue Cell Architecture via Integrated Single Cell and Spatial Transcriptomics." *bioRxiv*: 2020.11.15.378125. http://biorxiv.org/content/early/2020/11/17/2020.11.15.378125.abstract.

L., Ståhl Patrik et al. 2016. "Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics." *Science* 353(6294): 78–82. https://doi.org/10.1126/science.aaf2403.

Miao, Zhen et al. 2021. "Single Cell Regulatory Landscape of the Mouse Kidney Highlights Cellular Differentiation Programs and Disease Targets." *Nature Communications* 12(1): 2277. https://doi.org/10.1038/s41467-021-22266-1.

Moffitt, Jeffrey R et al. 2018. "Molecular, Spatial, and Functional Single-Cell Profiling of the Hypothalamic Preoptic Region." *Science* 362(6416): eaau5324. https://doi.org/10.1126/science.aau5324.

Moncada, Reuben et al. 2020. "Integrating Microarray-Based Spatial Transcriptomics and Single-Cell RNA-Seq Reveals Tissue Architecture in Pancreatic Ductal Adenocarcinomas." *Nature Biotechnology* 38(3): 333–42. https://doi.org/10.1038/s41587-019-0392-8.

Orth, Michael et al. 2019. "Pancreatic Ductal Adenocarcinoma: Biological Hallmarks, Current Status, and Future Perspectives of Combined Modality Treatment Approaches." *Radiation Oncology* 14(1): 141. https://doi.org/10.1186/s13014-019-1345-6.

Qiu, Peng. 2020. "Embracing the Dropouts in Single-Cell RNA-Seq Analysis." *Nature Communications* 11(1): 1169. https://doi.org/10.1038/s41467-020-14976-9.

Rakic, Pasko. 2009. "Evolution of the Neocortex: A Perspective from Developmental Biology." *Nature Reviews Neuroscience* 10(10): 724–35.

van Ravenzwaaij, Don, Pete Cassey, and Scott D Brown. 2018. "A Simple Introduction to Markov Chain Monte–Carlo Sampling." *Psychonomic Bulletin & Review* 25(1): 143–54. https://doi.org/10.3758/s13423-016-1015-8.

Reidy, Kimberly, Hyun Mi Kang, Thomas Hostetter, and Katalin Susztak. 2014. "Molecular Mechanisms of Diabetic Kidney Disease." *The Journal of clinical investigation* 124(6): 2333–40.

Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal. 2018. "Cancer Statistics, 2018." *Ca-a Cancer Journal for Clinicians* 68(1): 7–30.

Su, Jing, and Qianqian Song. 2020. "DSTG: Deconvoluting Spatial Transcriptomics Data through Graph-Based Artificial Intelligence." *bioRxiv*.

Yao, Zizhen et al. 2021. "A Taxonomy of Transcriptomic Cell Types across the Isocortex and Hippocampal Formation." *Cell* 184(12): 3222-3241.e26.