Identification of spatial co-expression patterns and intra-tissue heterogeneity in spatially resolved transcriptomics by region-specific denoising

Linhua Wang
Baylor College of Medicine https://orcid.org/0000-0002-6717-860X
Zhandong Liu (zhandong.liu@bcm.edu)
Baylor College of Medicine https://orcid.org/0000-0002-7608-0831

Article

Keywords: Spatial transcriptomics, cell spatial organization, gene-expression

Posted Date: November 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-647777/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on November 14th, 2022. See the published version at https://doi.org/10.1038/s41467-022-34567-0.

Identification of spatial co-expression patterns and intra-tissue heterogeneity in spatially resolved transcriptomics by region-specific denoising

Linhua Wang[1], Zhandong Liu [2, 3*]

1. Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, USA
2. Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, USA
3. Department of Pediatrics, Baylor College of Medicine, Houston, USA

* Corresponding author: zhandonl@bcm.edu

Abstract

We are pleased to introduce a first-of-its-kind tool that combines *in-silico* region detection and missing value estimation for spatially resolved transcriptomics. Spatial transcriptomics by 10X Visium (ST) is a new technology used to dissect gene and cell spatial organization. Analyzing this new type of data has two main challenges: automatically annotating the major tissue regions and excessive zero values of gene expression due to high dropout rates. We developed a computational tool—MIST—that addresses both challenges by automatically identifying tissue regions and estimating missing gene-expression values for each detected region. We validated MIST detected regions across multiple datasets using manual annotation on the histological staining images as references. We also demonstrated that MIST can accurately recover ST's missing values through hold-out experiments. Furthermore, we showed that MIST could identify intra-tissue heterogeneity and recover spatial gene-gene co-expression signals. We therefore strongly encourage using MIST before downstream ST analysis because it provides unbiased region annotations and enables accurately denoised spatial gene-expression profiles.

Introduction

To understand the biological mechanisms underlying diseases, it is essential to delineate cell and gene spatial organizations; the scientific community has therefore invested significant time and effort in detecting positional gene-expression[1,2]. Of all the spatial gene-expression profiling techniques that have been developed, 10X Visium Spatial Transcriptomics (ST) gained its most popularity due to its whole-genome scalability and cost-efficiency[2,3]. So far, ST has been used to study many tissues' spatial gene-expression, including breast cancer, prostate tumors, melanoma tumors, human and mouse brains, and more[3–8].

While the location of every sequenced tissue domain (spot) in ST is assigned, regions of the tissue are not directly provided. Routinely, researchers circled out anatomical regions by looking at the histological staining image that is aligned with ST. However, in many cases, pathological regions are not detectable through human eyes. Moreover, assigning each spot to a region is labor-intensive and might be erroneous. Therefore, it is of great significance to automatically and unbiasedly extract regions within ST tissues.

Another challenge researchers faced when analyzing ST data is the sparsity of the gene expression profile caused by technical dropouts, which leads to excessive zero values in the gene expression data[3]. As a consequence, it drastically reduces ST's signal-to-noise ratio and

prevents accurate co-expression calculation, cluster detection, and other downstream analyses. Therefore, accurately estimating dropped-out expression values in ST data is vital to rescuing such signals and facilitating more accurate downstream analyses.

While we lack computational methods specifically designed to estimate missing values in ST, plenty of such methods have been developed for single-cell RNA sequencing (scRNA-seq) data[9–14]. Many of these methods— like MAGIC[9], PRIME[10] and knnSmoothing[11]—smooth the missing gene expression values by taking the weighted average of the gene expression values from cells with similar molecular profiles. None of these methods, however, interpret ST's spatial connectivity information to determine the functionally and physically similar neighbors to learn from. Other kinds of methods—like McImpute[12] and ALRA[14], estimate the missing values by approximating a low rank of the gene expression matrix, assuming a relative number of cell types within the tissue. However, they are performed at the entire tissue and lack regional specificity that could allow us to rescue regional heterogeneity.

We realized that these two challenges: (1) difficulty automatically annotating tissue regions and (2) excessive zero gene expression values due to technical dropouts are not independent of each other. On the contrary, solving the first challenge could facilitate region-specific denoising.

An intuitive way of denoising ST using spatial information is to use spatially adjacent neighbors to estimate missing values. However, using adjacent neighbors introduces errors regarding spots on functionally different tissue regions' boundaries. In such cases, detecting the boundaries is vital to avoiding spatial-information misusage when denoising ST data and allows ST to be grouped into functionally similar regions. Denoising within such functional and spatial regions, therefore, avoids errors at tissue boundaries.

We hypothesize that the number of cell types within a functional region in the tissue is limited, leading us to believe that a more accurate denoising result will be achieved by region-specific denoising through low-rank matrix completion. Therefore, we introduce our new tool – Missing-value Imputation for Spatially resolved Transcriptomics (MIST) that addresses the *in-silico* detection and region-specific denoising tasks in a two-step pipeline.

Results

The MIST Algorithm

MIST provides a two-step tool to address the challenges in ST's analyses including automatically detecting ST's functional regions and denoising ST data at a region-specific schema (Fig. 1).

In the first step, MIST automatically detects functional regions in the sample of interest. MIST embeds ST as a 2D graph where every spot will be presented as a node. Every pair of adjacent nodes will have an edge connecting each other with a weight defined by the molecular similarity between the pair of nodes.

To simulate the region boundaries, MIST filtered out low-weight edges with a threshold and extract the connected components within the remaining graph. To avoid bias in selecting the filtering threshold, MIST searches for the optimal value by assessing different thresholds' accuracy in recovering the hold-out values of a set of highly expressed genes.

In the second step, MIST estimates the missing values in each detected region by averaging the outcomes from multiple runs of a low-rank approximation algorithm[12]. Under the assumption that each detected region should have a small number of cell types, we expect the denoised region-specific expression matrix to have a low rank. To achieve this goal, we used a low-rank matrix completion approach that estimates the missing values by minimizing the singular values of the denoised matrix.

To increase the reliability of the estimated values and avoid an extremely singular gene-expression matrix, before denoising, MIST adds contrast to each region by augmenting the region with random spots from other regions. To remove the bias from randomly selected spots and stabilize the estimated values, MIST runs multiple repeats of low-rank approximation on the augmented regions with different sets of augmented spots and takes the average of the outcomes as the estimated values for region-specific spots.

MIST detected functional tissue regions that agree with manual annotations

To prove that MIST could faithfully detect functional regions within tissues, we tested MIST on two examples including a melanoma sample and a mouse brain sample (Fig. 2a, b). We showed that under different filtering thresholds, different shapes of regions were detected. However, regions using the threshold automatically selected by MIST mostly resembles the manual annotation provided by human experts.

While higher filtering thresholds on the edges will find strongly correlated spots within regions, they also lead to a significant number of isolated spots with no connected edges (Fig. 2c). But, while small thresholds will include most spots, they will lose the specificity that allows us to accurately detect tissue regions (Fig. 2c). Balancing the proportion of non-isolated spots and region-detection accuracy is, therefore, crucial to successful region detection and region-specific denoising.

To demonstrate that MIST strikes this balance automatically and accurately, we used Adjusted Rand Index (ARI) to evaluate the consistency between MIST-detected regions and the human experts' annotations. An ARI of zero shows the concordance of two random clustering results while a score of one indicates a perfect match of two clustering results. Using the mouse wild-type brain sample, we validated that the threshold that MIST automatically selected found a balance point that maximizes the ARI while minimizing the proportion of isolated spots (Fig. 2c). Using the selected threshold, the mouse brain regions detected by MIST agreed with the anatomical regions (Fig. 2b) with an ARI value of 0.64.

To confirm the biological relevance of MIST-detected regions in the melanoma sample, we looked at the gene profiles in these regions. Specifically, we identified significantly activated genes in the tumor region relative to the lymphoid region and vice versa using the Wilcoxon rank-sum test. By selecting genes with a log2 fold-change greater than 0.58 and an adjusted P-value less than 10^{-5}, we found 143 and 49 marker genes for the tumor and lymphoid regions, respectively (Fig. 2d). We observed that some well-known melanoma marker genes, including *MLANA*[15], MCAM[16], *SPP1*[17], and *HSP90AA1*[18], topped the list of the melanoma-activated genes that we identified, which demonstrates our detected region's accuracy. We then performed gene set enrichment analysis on the significantly activated genes and observed that the immune response gene ontology term was significantly enriched (adjusted P-value = 9×10^{-8}, Fig. 2e) for the detected lymphoid region. Taken together, the region automatically detected from the Melanoma ST reflected the hidden functional regions within the tissue.

MIST accurately recovers hold-out values across multiple data sets

To assess MIST's accuracy when estimating missing values, we performed 5-fold random hold-out experiments in which we withheld a random set of the observed non-zero values and used these as ground truth to evaluate the models' performances. By withholding some of the

observed values, we simulated cases in which non-zero expression values have dropped out. To consider the heterogeneity of datasets that might lead to biased performance, we tested four datasets from samples including a wild-type mouse brain sample, an AD mouse brain sample, a melanoma tumor sample, and a prostate tumor sample that vary in the number of genes, number of spots, and sparsity level (Table 1). For each dataset, we selected genes that are expressed across at least half of all spots in the sample to generate hold-out test data sets. For each gene, we partitioned the non-zero expression values into five non-overlapping sets. Then, we iteratively held out one-fold of the values and assessed MIST and other methods' accuracy in recovering the held-out gene expression values.

To demonstrate MIST's supremacy in missing value recovery, we compared MIST with state-of-the-art scRNA-seq methods[19], including MAGIC[9], knn-smoothing[11], McImpute[12] and SAVER[13], and a baseline k-nearest neighbor method we constructed (spKNN) that estimates missing values by averaging spatially adjacent neighbors.

To evaluate the accuracy of missing value estimation in hold-out experiments, we used two metrics including Rooted Mean Square Error (RMSE) and Pearson Correlation Coefficient (PCC) where RMSE represents the error and PCC shows the agreement between the ground truth and estimated values. Better imputation methods are expected to have lower RMSE and higher PCC scores.

We found that MIST consistently outperformed other methods across all datasets with higher PCC and lower RMSE scores during hold-out value evaluation (Fig. 3a, b). For the spots that are assigned to functional regions, MIST had an average RMSE improvement (decreasing values) of 12% (P-value=3×10^{-4} compared with McImpute, and 14% (P-value=10^{-14}) and 35% (P-value=7×10^{-10}) improvement compared with the baseline spKNN algorithm. Moreover, MIST's mean PCC is also 8% (P-value=9×10^{-6}), 4% (P-value=2×10^{-4}), and 16% (P-value=6×10^{-11}) higher than McImpute, MAGIC and spKNN respectively. Knn-smoothing and SAVER consistently performed substantially worse than the other four methods (Ext. Fig. 3.1, 3.2).

To investigate the impact of genes' sparsity on denoising accuracy, we further stratified the performance assessment at a per-gene level grouped by the sparsity level (zero-value proportion). While MAGIC and spKNN's estimation error monotonically increased with sparsity

level, MIST's performance was not vulnerable to gene sparsity (Fig. 3c). While McImpute's performance was also not influenced by gene sparsity, MIST has superior performance than McImpute at every gene sparsity level (Fig. 3c).

To demonstrate that MIST recovers the gene-expression patterns, we then visualized and showed that MIST can faithfully recover the gene-expression spatial pattern for gene *GAPDH* after denoising using the Melanoma tissue sample (Fig. 3d). With the hold-out input, MIST accurately estimated the original expression values by increasing Spearman's correlation coefficient from 0.65 to 0.96 (Fig. 3e). When evaluating all genes across the four tested data sets, after denoising, the median correlation had significantly improved from 0.64 to 0.88 (P-value $\approx$ 0, Fig 3f, Ext. Fig. 3.3, 3.4).

MIST discovered intra-cortex heterogeneity within an Alzheimer's Disease (AD) mouse brain

To demonstrate that MIST could improve the clustering results of ST data, we used Uniform Manifold Approximation and Projection (UMAP)[20] to reduce the dimension of Mouse brain ST samples' gene expression data and visualize the clustering structures. First, we performed UMAP on the original and denoised wild-type mouse brain sample. After denoising, MIST enhanced the heterogeneity with most of the spots within the cortex, hippocampus, and thalamus, forming individual clusters (Fig. 4a, Ext. Fig. 4.1).

While similar enhanced clustering patterns were also identified in the Mouse AD brain sample after denoising, in addition, we observed that the cortex region was separated into two individual clusters, something we did not detect using the WT mouse brain nor in the original Mouse AD data (Fig. 4b, Ext. Fig. 4.2). Further cortex analysis revealed a clear separation of the cortex region into two spatially separable parts (Fig. 4c, Ext. Fig. 4.3). The first cluster consisted of the cortical subplate, olfactory, entorhinal, ectorhinal, temporal association, and perirhinal areas. The second cluster contained the auditory, primary somatosensory, posterior parietal association, and retrosplenial areas. When these two clusters were mapped to the anatomical reference, cluster 1 occupied the upper quadrant while cluster 2 occupied the lower quadrant (Fig. 4d-f). Interestingly, such heterogeneity was only detected in the MIST-denoised AD cortex but not the WT cortex.

To understand the biological difference of these two clusters in AD progression, we further performed differential gene analysis to extract AD activated genes for these two clusters, respectively. By selecting upregulated genes in the AD sample with a fold change greater than 50% and adjusted P-value < 0.01, we identified 55 markers for cluster 1 and 41 markers for cluster 2 (Ext. Fig. 4.4, 4.5). We found only 21 AD-activated genes, such as *Clu*, are shared for these two clusters (Fig. 4d). 34 genes, such as *Hap1*, are AD-upregulated only in cluster 1 (Fig. 4e) and 20 genes, such as *Sez6*, that are only AD-upregulated in cluster 2 (Fig. 4f). We showed that the spatial clusters identified using MIST-denoised cortex data play different roles in AD progression by activating different sets of genes, demonstrating that MIST allows the discovery of biological heterogeneity from ST data.

MIST recovers spatial gene-gene co-expression patterns

Spatial gene-gene co-expression plays an important role in understanding gene interactions across 2D space. The dropouts within the ST data undermine the correlation analysis's power and cause inaccurate estimation of gene-gene spatial correlation, which is the fundamental element in many analyses such as weighted correlation network analysis (WGCNA)[21].

To demonstrate that MIST recovers the spatial co-expression patterns, we examined two pairs of genes: *Cldn11-Arhgef10* and *Gfap-Aqp4*.

The first pair of genes *Cldn11-Arhgef10* showed a high spatial correlation score of 0.97 based on the reference Allen Brain Atlas[22] (Fig. 5a). However, in the original ST data, the correlation score between *Cldn11* and *Arhgef10* is only 0.15 (P-value = 2×10^{-3}, Fig. 5b). After denoising by MIST, the correlation score was improved to 0.5 (P-value = 3.5×10^{-29}, Fig. 5c). To visualize the gene expression patterns, we plotted the heatmap of log-scale expression values and showed that *Cldn11* and *Arhgef10* have similar gene expression patterns only after MIST denoising (Fig. 5d-g).

To show that MIST can recover co-expressed gene pairs that are not significantly correlated with the original data, we examined the second pair of genes: *Gfap-Aqp4,* which have a moderately good spatial correlation with a score of 0.74 in the reference Allen Brain Atlas database (Fig. 5h). Before denoising, we observed an insignificance correlation with a score of 0.08 (P-value = 0.56, Fig. 5i). After denoising, we could recover a significant spatial correlation with a score of 0.35 (P-value = 9×10^{-12}, Fig. 5j). Similar to the first pair of genes, the visual

pattern improvement of *Gfap-Aqp4* could also be observed only after denoising by MIST (Fig. 5k-n).

By showing these two cases, we demonstrated that MIST could rescue the spatial correlation of gene pairs whose co-expression patterns are either lessened or lost in the original ST data. Given co-expression estimation is vital in many downstream analyses such as identifying gene modules[21], we see the significance and importance of using MIST to denoise ST data before carrying out such analyses.

Conclusions

In this study, we tackled two major problems encountered in ST data analyses – (1) in-silico region detection and (2) missing value estimation. We developed a computational tool, MIST, that solved both problems.

We solved the first challenge by combining the molecular similarity and spatial connectivity between spots and enabled detecting tissue regions automatically. Before our work, researchers align the histological images with ST coordinates to manually assign regions to every spot. This procedure is laborious and might be inaccurate in cases where pathological changes are not visually detectable through human eyes. With MIST, we could bypass manual annotation and allows unbiased region assignment to every spot. We proved that MIST could accurately detect regions by comparing MIST-detected regions with manually assigned regions at the per-spot level. Addressing this challenge allows us to analyze ST data at a region-specific level and improve the specificity for denoising.

We solved the second challenge by approximating a low rank of region-specific gene expression matrix through singular value decomposition. This is based on a simple yet interpretable assumption that the number of cell types for every region is small, which has been adopted by many other single-cell RNA-sequencing denoising methods such as McImpute[12] and ALRA[14]. Compared with McImpute and ALRA, MIST utilized spatial information to define regions and improved the denoising specificity.

To demonstrate that MIST's accuracy in estimating the missing values, we developed hold-out experiments to simulate dropout events and assessed MIST's performance in the hold-out trials.

As a result, MIST could accurately recover the hold-out values and outperformed the state-of-the-art single-cell denoising methods and a baseline spatial-information-based approach.

While addressing these two major challenges, MIST allows researchers to drastically recover biological signals that are missing in the original ST data. Such biological signals include spatial co-expression patterns, intra-region heterogeneity, and more.

We demonstrated that MIST could improve or recover spatial patterns of co-expressed genes that are highly correlated in reference atlas but either poorly or insignificantly correlated in the original ST. Since many downstream analyses such as WGCNA[21] are based on co-expression estimation, denoising ST by MIST is rather essential before such analyses because it avoids false conclusions due to inaccurate co-expression estimation.

It is also of great significance to understand the heterogeneous responses of different tissue regions to pathology, which will shed light on the treatment of diseases. Therefore, it is critical to recovering such heterogeneity that was lost in the original ST data.

We showed that MIST enables recovering the spatial heterogeneity within the mouse cortex during AD progression. By extracting the differentially expressed genes between AD and wild-type mouse brain samples, we identified genes that are AD-upregulated only in specific regions of the mouse cortex.

We therefore unequivocally recommend using MIST for ST analyses. Our *sui generis* *in-silico* region detection enables analyzing the ST at a brand-new level that combines anatomical connectivity and molecular similarity. It will be useful in many areas, including identifying local subregions within tumors whose heterogeneity is hard to see through pathological staining images. Second, MIST allows researchers to recover important biological signals in downstream analyses, such as when identifying spatial gene-gene co-expression patterns. While the original ST data provided by 10X Visium might hinder ST analyses' accuracy, MIST will accurately rescue the missing values and drastically increase the signal-to-noise ratio.

Methods

Data collection and preprocessing

The ST data sets we used in this study include a 12-month wild-type mouse brain sample, a 12-month Alzheimer's Disease (AD) mouse brain sample, a Melanoma tumor sample, and a prostate tumor sample[5,6,8] (Table 1). Every dataset has a raw mRNA count matrix form where rows indicate spots and columns indicate genes.

To filter out low-quality genes that might otherwise introduce noise to the pipeline, we kept genes with raw counts >= 3 in more than 2 spots within each sample.

To account for the different library sizes of every spot due to variance in sequencing depth and the number of cells, we normalized the raw mRNA count matrix using count per million ($CPM = \frac{Raw\ count * 10^6}{Library\ size}$).

In-silico region detection

Graph embedding

Suppose the ST expression matrix has M spots and N genes, the spatial gene-expression profile can be defined as $Y \in R_*^{M \times N}$, where Y is the observed gene expression matrix, and R_* denotes non-negative real matrices with M rows and N columns. The M spots in an ST slide can form a lattice graph, $G = <V, E>$, where V is the node-set and E is the edge-set. Every pair of adjacent (u, v) spots are connected with edge $E(u, v)$.

Weight calculation

To infer the weights for every connected edge $E(u, v)$, we calculated the Pearson correlation coefficient between the gene expression profile of spot u and v. To remove the noise in high-dimensional gene expression data while keeping the major signals, we used Principal Component Analysis (PCA)[23] to reduce the dimension of the gene expression matrix and kept the first p principal components that could explain 80% of the variance. The weight for edge $E(u, v)$ is then inferred using the correlation score between the first p principal components of spot u and v.

Edge removal and parameter selection

To simulate the boundaries between functionally dissimilar regions, we removed edges whose weights are less than a threshold ε.

To automatically select the threshold and avoid bias, we performed a hold-out experiment on a set of genes that are expressed in at least 80% of all spots. We randomly held out 25% of the gene expression values for every such gene and used them as ground truth to test MIST's denoising error defined as $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_i')^2}{n}}$, where X is the hold-out non-zero values, X' represents the MIST denoised gene expression matrix using certain ε and n denotes the number of non-zero hold-out values. To determine the optimal threshold, we did a grid search with values ranged from 0.1 to 0.9 and selected the one with the minimal RMSE.

Region detection by extracting connected components

To simulate the functional regions within tissues, we used a depth-first search algorithm[24] to identify all the connected components in graph G after edge removal. Each connected component with greater than 5 spots is predicted as an independent tissue region.

Region-specific denoising through multiple runs of low-rank matrix completions

Low-rank matrix completion

Suppose the observed gene expression matrix for region C is Y_C where Y_C is a sparse matrix with m rows (spots) and n columns (genes). The task is to estimate X_C, which represents the denoised gene expression matrix for region C. We adapted the low-rank-matrix completion algorithm through singular value decomposition used by McImpute[12].

Given the assumption that the number of cell types within a functional region is small, we expect X_C to have a low rank. To achieve this goal, the task is turned into a low-rank matrix completion problem by solving the following objective function:

$$\min_{X_C} \left|\left|Y_C - A(X_C)\right|\right|^2 + \lambda * rank(X_C) \qquad (1)$$

The first term in equation (1) minimizes the error between the non-missing gene expression values of X_C and Y_C with a projection function A that returns values in X_C at the indices of non-missing values in Y_C. The second term in equation (1) minimizes the rank of the denoised gene expression matrix. The objective function is a linear combination of these two terms regularized a non-zero tuning parameter λ. Theoretically, a larger λ will give us a lower-ranked denoised gene expression matrix whose values on the non-missing indices might deviate from the ground

truth. On the other hand, a small λ will result in a relatively high-rank denoised matrix with a lower error on the non-missing indices.

However, minimizing the rank of a matrix is non-convex. To transform it to a convex problem with a globally optimal solution, we relaxed equation 1 as:

$$\min_{X_C} \left|\left|Y_C - A(X_C)\right|\right|^2 + \lambda * \|X_C\|_{nuc} \quad (2),$$

, where we transformed the term 2 in equation 1 as a nuclear norm of X_C, which can be calculated by summing up the singular values obtained through singular value decomposition. Specifically, equation (2) can be further transformed as

$$\min_{X_C} \left|\left|B - X_C\right|\right|^2 + \lambda * \|X_C\|_{nuc} \quad (3),$$

where $B_{k+1} = X_{k,C} + \frac{1}{\alpha} A^T(Y_C - A(X_{k,C}))$. Using the inequality $\|B - X\|_2 \geq \|s_B - s_X\|$, where s_X denotes the singular value vector for matrix X, equation (3) can be rewritten as

$$\min_{X_C} \left|\left|s_B - s_{X_C}\right|\right|^2 + \lambda * \left\|s_{X_C}\right\|_1 \quad (4).$$

Therefore, by taking the derivative of (4), the solution of s_{X_C} is achieved by soft thresholding the singular values of s_B with a threshold equals to $\lambda/2$.

To tune the parameter λ that strikes the balance between low matrix singularity and low error on non-missing indices, we find the maximal λ that achieves a fixed low error (10^{-12}) calculated by the sum of the absolute difference between denoised and observed values on non-missing indices.

Compared with the original solution provided by McImpute, MIST also forces the observed values to be unchanged. By doing so, MIST guaranteed that it will only touch the missing values and keep the observed non-zero expression values unharmed. Moreover, MIST's matrix completion was implemented in a free and open-source language—Python, as compared to MATLAB, which was used by McImpute.

Region augmentation

To avoid an extremely singular denoised gene expression matrix that might otherwise introduce instability, we added contrast to every detected region before low-rank matrix completion. Specifically, we augmented the region C with r randomly selected spots from other regions

where $r = \min(\frac{size(Other\ regions)}{10},\ size(C))$. The low-rank matrix completion algorithm will be performed on the augmented region C^a with input X^a.

To include more diversity and reduce the randomness in region augmentation, we repeated the previous step T times. For every spot $s \in C$, we take the average of denoised gene expression values $Y^a(s)$ from outcomes of T repeats. T is set to be 10 as a default to avoid high computational load while including as much signals as possible.

To demonstrate the essentiality of region augmentation, we tested MIST's performance in the hold-out experiments using detected regions and augmented regions, respectively. We found that region augmentation significantly improved MIST's denoising accuracy (RMSE P-value = 1.6×10^{-7}, PCC P-value = 4.4×10^{-7}, Ext. Fig. 6).

Quantifying region detection accuracy

Adjusted rand index is used to quantify the agreement between the manual spot-level annotation and MIST-predicted regions. A score of one means a perfect agreement between two grouping results while zero score means the predicted regions are similar to the results from random guessing. Adjusted rand index scores are calculated using the Python *scikit-learn* package[26].

Differential gene expression analysis

Wilcoxon rank-sum test provided by the Python package *Scipy*[25] was used to infer the significant level of differentially expressed genes. Fold change of genes between condition and control was calculated based on the difference of average gene expression within groups. To get the significantly upregulated genes in the MIST-predicted melanoma region relatively to lymphoid region, we selected genes with a fold change greater than 50% and adjusted P-value less than 10^{-5}. To get the AD upregulated genes in the mouse cortex regions, we changed the adjusted P-value threshold to 10^{-2} because of smaller sample sizes.

Gene set enrichment analysis

The R package *clusterProfiler*[27] was used to perform gene ontology enrichment analysis using the list of differentially expressed genes in the melanoma sample.

Data generation for hold-out experiments

To improve the diversity in the hold-out experiments, we tested samples including a mouse wild-type brain sample, a mouse AD brain sample, a melanoma tumor sample, and a prostate tumor sample. To extract good-quality genes to simulate the dropout events, we removed genes that were expressed in less than 50% of the spots.

To generate the hold-out data, we used a 5-fold cross-validation schema for the non-zero values. Specially, we first randomly partitioned every gene's nonzero expression values into 5 groups. In each hold-out, we created missing values by setting one group to zero and performed imputation based on the remaining values. The held-out values served as ground truth for evaluating the imputation algorithms' accuracy.

Evaluating hold-out experiments' performance

To quantifying the accuracy in recovering the held-out values, we reported two metrics including RMSE and PCC, where RMSE measures the estimation error and PCC measures the linear correlation between the true expression values and the estimated values. RMSE is defined as $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{n}}$ and PCC is defined as $PCC = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$, where Y is the hold-out non-zero values, X represents the MIST estimated values and n denotes the number of hold-out values.

To quantify the recovery of gene expression patterns after denoising the hold-out data, we used Spearman's rank correlation test implemented by *Scipy*[25] to assess the correlation and the corresponding significance level between the original ST and denoised gene expression values.

Co-expression analysis using Allen Brain Atlas as a reference

We obtained mouse brain (coronal section) regional expression values for gene *Cldn11* (experiment: *RP_070116_01_G04*), *Arhgef10* (experiment: RP_070116_01_B05), *Gfap* (experiment: RP_Baylor_253913) and *Aqp4* (experiment: RP_040324_01_F07) from Allen Brain Atlas[22] as references. Gene expression values provided by Allen Brain Atlas are at the log2 scale.

Correlation scores between gene pairs in both reference and ST data are represented by Spearman's correlation coefficient calculated using the Python package *Scipy*[25].

Table 1. Data summary for hold-out experiments

	MouseWT	*MouseAD*	*Melanoma*	*Prostate*
#Genes (preprocessed)	10178	11241	4498	8073
#Genes (<50% sparsity)	3024	4081	952	1718
#Spots	447	488	293	406
Non-zero values (%)	22	26	15	21

Source Code Availability

The MIST algorithm is implemented in Python and is available at https://github.com/linhuawang/MIST.git.

All the source code and raw data for this manuscript can be downloaded from https://github.com/LiuzLab/MIST_manuscript.git.

References

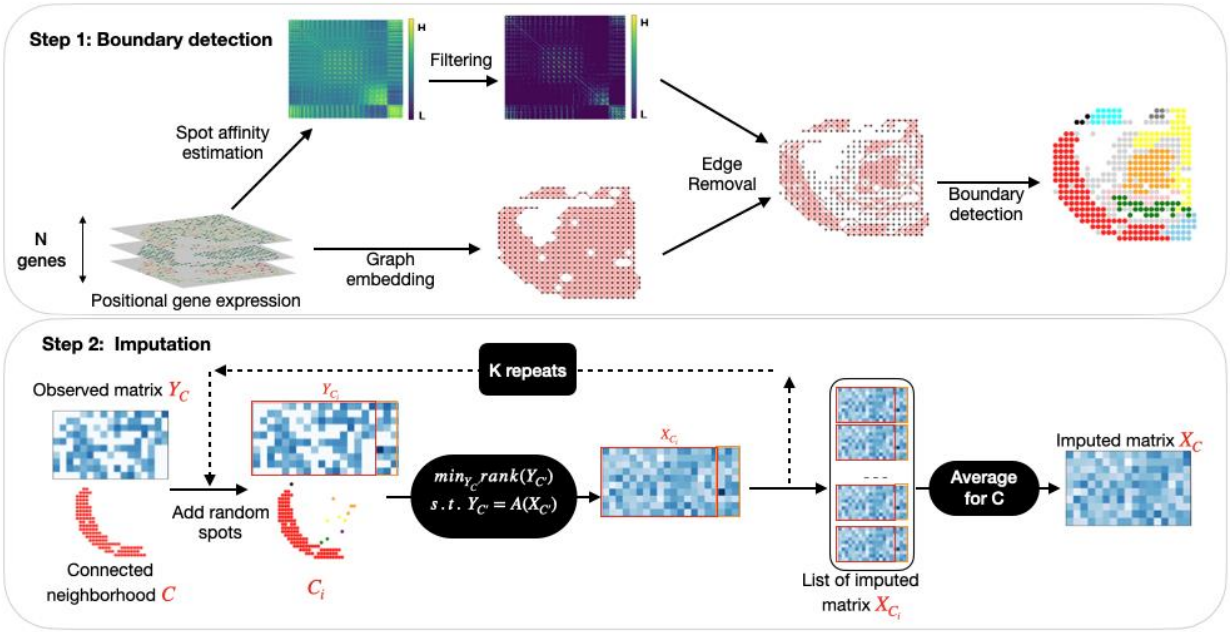
1. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).
2. Asp, M., Bergenstråhle, J. & Lundeberg, J. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays* **42**, 1900221 (2020).
3. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (80-.).* **353**, 78–82 (2016).
4. He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* 1–8 (2020).
5. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9**, 1–13 (2018).
6. Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. & Lundeberg, J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.* **78**, 5970–5979 (2018).
7. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* 1–12 (2021).
8. Chen, W.-T. *et al.* Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* **182**, 976–991 (2020).
9. Van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion.

Cell **174**, 716–729 (2018).
10. Jeong, H. & Liu, Z. PRIME: a probabilistic imputation method to reduce dropout effects in single-cell RNA sequencing. *Bioinformatics* **36**, 4021–4029 (2020).
11. Wagner, F., Yan, Y. & Yanai, I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *BioRxiv* 217737 (2017).
12. Mongia, A., Sengupta, D. & Majumdar, A. McImpute: matrix completion based imputation for single cell RNA-seq data. *Front. Genet.* **10**, 9 (2019).
13. Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539–542 (2018).
14. Linderman, G. C., Zhao, J. & Kluger, Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv* 397588 (2018).
15. Chen, Y.-T. *et al.* Serological analysis of Melan-A (MART-1), a melanocyte-specific protein homogeneously expressed in human melanomas. *Proc. Natl. Acad. Sci.* **93**, 5915–5919 (1996).
16. Xie, S. *et al.* Expression of MCAM/MUC18 by human melanoma cells leads to increased tumor growth and metastasis. *Cancer Res.* **57**, 2295–2303 (1997).
17. Zhou, Y. *et al.* Osteopontin expression correlates with melanoma invasion. *J. Invest. Dermatol.* **124**, 1044–1052 (2005).
18. Shomali, N. *et al.* Heat shock proteins regulating toll-like receptors and the immune system could be a novel therapeutic target for melanoma. *Curr. Mol. Med.* **21**, 15–24 (2021).
19. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21**, 1–30 (2020).
20. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv Prepr. arXiv1802.03426* (2018).
21. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 1–13 (2008).
22. Jones, A. R., Overly, C. C. & Sunkin, S. M. The Allen brain atlas: 5 years and beyond. *Nat. Rev. Neurosci.* **10**, 821–828 (2009).
23. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **2**, 37–52 (1987).
24. Tarjan, R. Depth-first search and linear graph algorithms. *SIAM J. Comput.* **1**, 146–160 (1972).
25. Virtanen, P. *et al.* {SciPy} 1.0: Fundamental Algorithms for Scientific Computing in

Python. *Nat. Methods* **17**, 261–272 (2020).

26. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

27. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omi. a J. Integr. Biol.* **16**, 284–287 (2012).



568 **Figure 1. | The MIST pipeline.** In step 1, region boundaries are detected by extracting locally
569 connected components through graph embedding and edge filtering. In step 2, missing values
570 for every detected region are estimated using the average of multiple runs of low-rank matrix
571 completion algorithm.
572

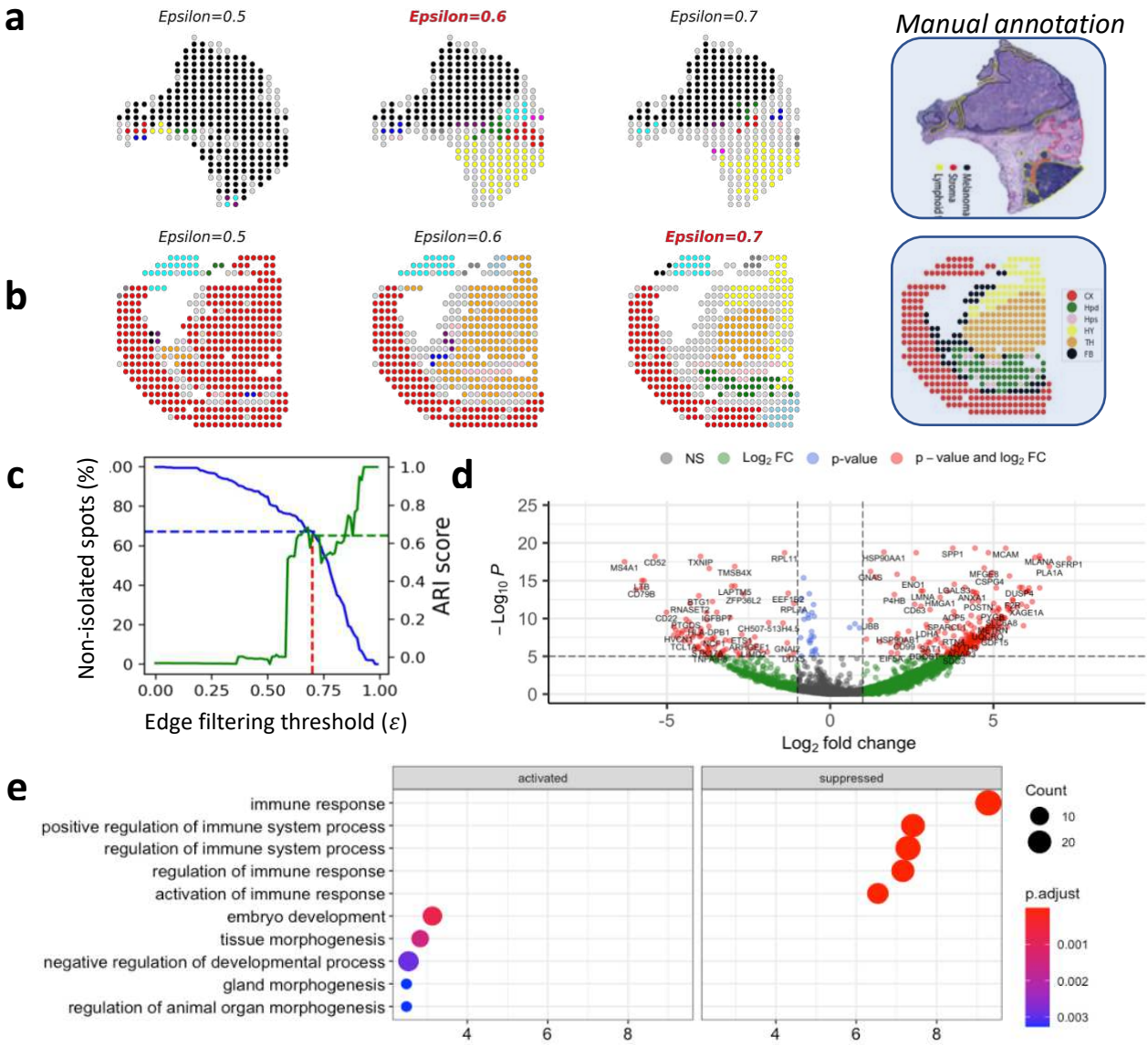


Figure 2. | MIST detected regions agree with manual annotations. a, Melanoma tissue regions detected using different epsilon values. Column 2 with epsilon of 0.6 was selected by MIST. Column 4 is the pathological annotation on the aligned histological staining image. **b,** Mouse brain tissue regions detected using different epsilon values. Column 3 with epsilon of 0.7 was selected by MIST. Column 4 is the spot-level manual annotation from the original study. **c,** Percentage of non-isolated spots (left y-axis, blue curve) and adjusted rand index (ARI, right y-axis, green curve) as functions of edge-filtering parameter epsilon. **d**, Volcano plot of differential expressed genes in tumor region contrasted to lymphoid region of the Melanoma sample. **e**, Gene ontologies activated and suppressed for MIST-detected melanoma tumor region compared relatively to lymphoid region.

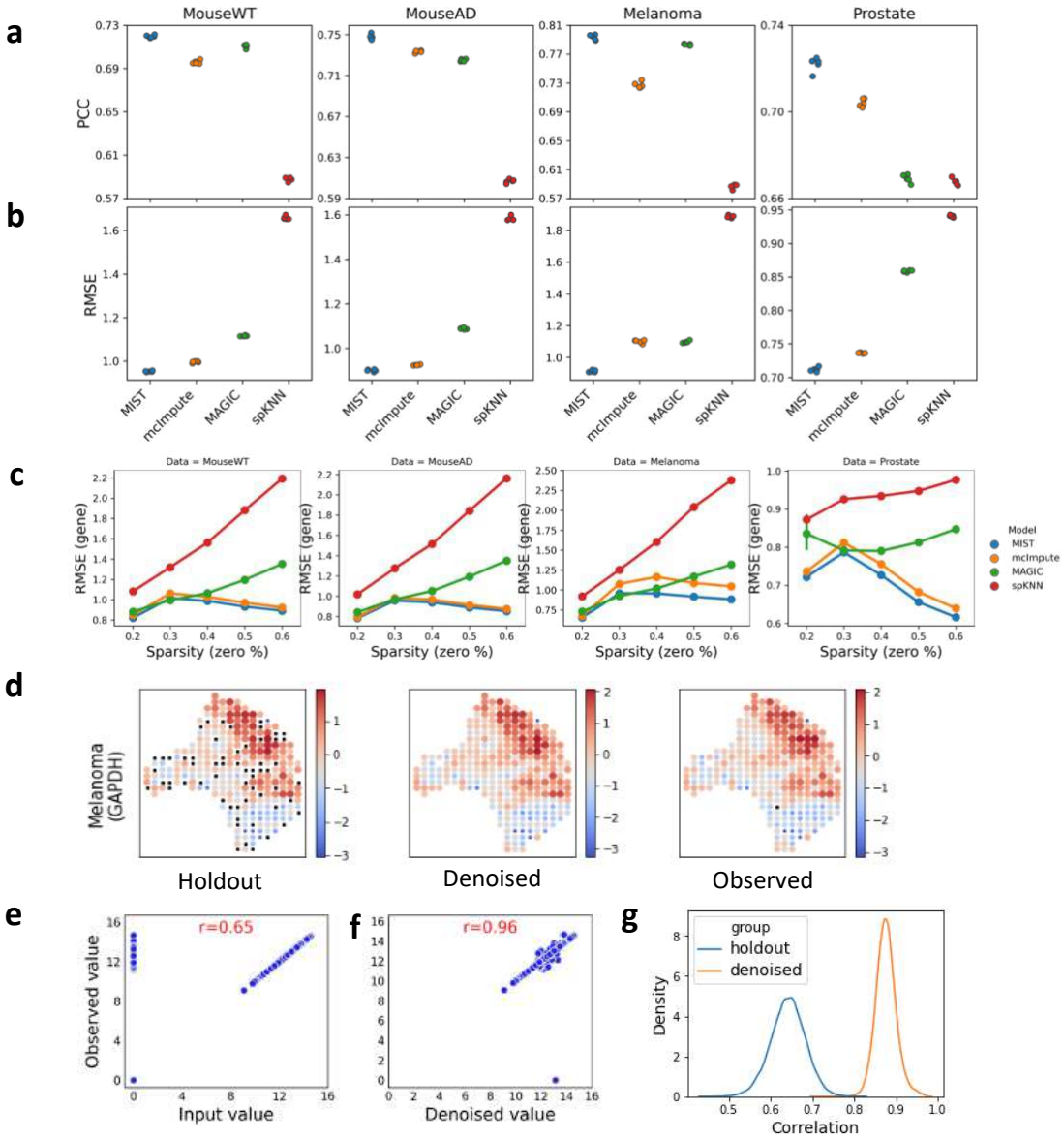


Figure 3. | MIST's outperforms other imputation methods in holdout experiments. a-b, Holdout experiment performance across multiple data sets using metrics **a**, PCC and **b**, RMSE. Each column is an individual data set. Points in the same color indicate the 5 non-overlap fold of tests for each model. **c**, Gene-level performance of each model represented by the RMSE (y-axis) as a function of zero-value percentage of the gene expression values (sparsity, x-axis). **d**, Expression patterns recovered for gene GAPDH in the Melanoma sample. From left to right shows the spatial pattern of the input to MIST with holdouts, the denoised gene expression pattern and the original observed gene expression pattern. Color and size indicate relative gene expression abundance. **e**, the correlation between holdout input and observed GAPDH expression values. **f,** the correlation between MIST denoised and observed GAPDH expression values. **g**, distribution of the gene-level correlation between original non-zero gene expression values and the holdout-input (blue), and denoised values (orange), respectively.

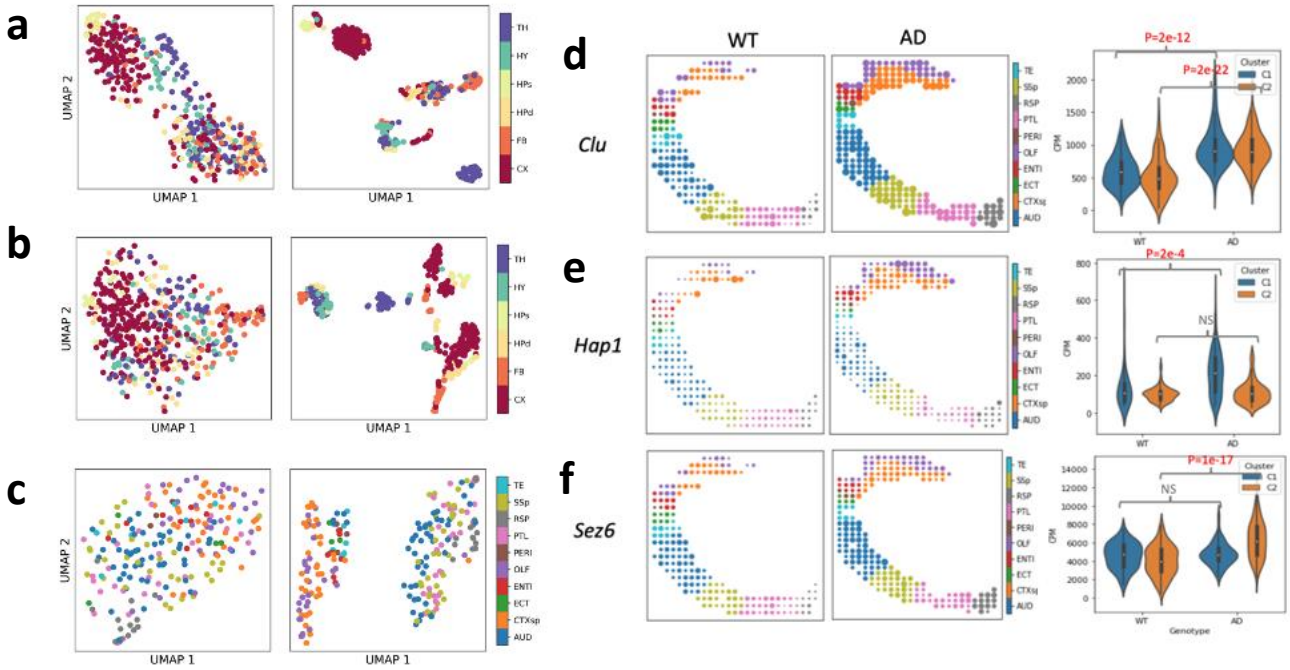


Figure 4. | MIST identified intra-cortex heterogeneity within Alzheimer's Disease (AD) mouse brain. a, UMAP of Mouse WT brain using original observed ST data (left) and MIST-denoised ST data (right). **b,** UMAP of Mouse AD brain using original observed ST data (left) and MIST-denoised ST data (right). **c,** UMAP of Mouse AD cortex region using original observed ST data (left) and MIST-denoised ST data (right). **d-f,** Examples of AD up-regulated genes that are activated in both clusters (**d**), only significantly upregulated in Cluster 1 (**e**, CTXsp: cortical subplate, OLF: olfactory, ENTI: entorhinal, TE: temporal association, ECT: ectorhinal, and PERI: perirhinal areas) and Cluster 2 (**f**, AUD: auditory, PTL: posterior parietal association area, RSP: retrosplenial area, SSp: primary somatosensory). Left, spatial expression pattern in WT mouse brain cortex; middle, spatial expression pattern in AD mouse brain cortex; right, boxplot of CPM grouped by genotype and spatial clusters.

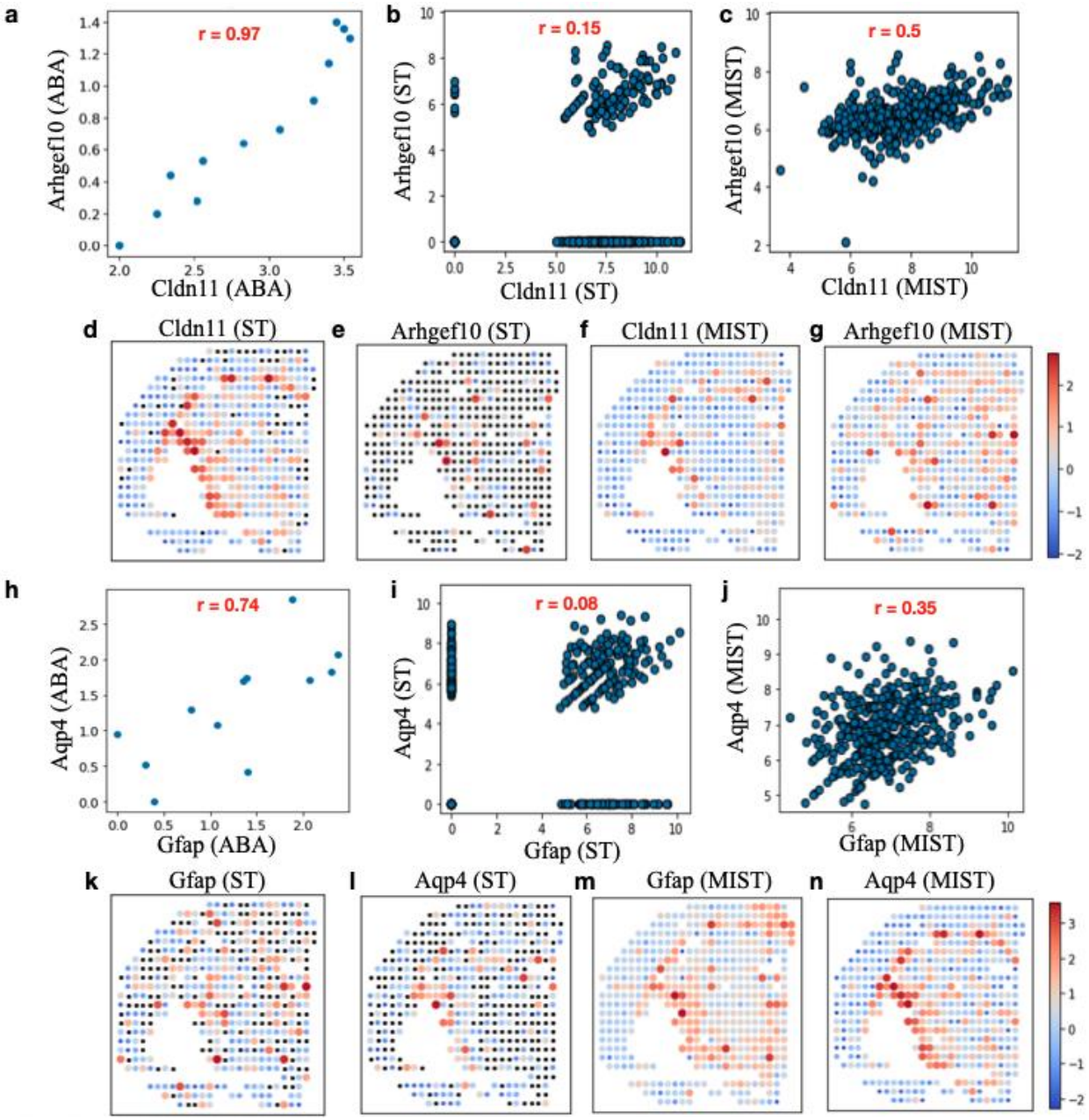


Figure 5. | MIST recovers spatially co-expressed gene pairs. a-c, spatial correlation of *Cldn11-Arhgef10* in the (**a**) Allen Brain Atlas reference, (**b**) original ST data and (**c**) MIST-denoised ST data. **d,** expression patterns of gene *Cldn11* using the original ST data. **e,** expression patterns of gene *Arhgef10* using the original ST data. **f,** expression patterns of gene *Cldn11* using the MIST-denoised data. **g,** expression patterns of gene *Arhgef10* using the MIST-denoised data. **h-j,** spatial correlation of *Gfap-Aqp4* in the (**h**) Allen Brain Atlas reference, (**i**) original ST data and (**j**) MIST-denoised ST data. **k,** expression patterns of gene *Gfap* using the original ST data. **l,** expression patterns of gene *Apq4* using the original ST data. **m,** expression patterns of gene *Gfap* using the MIST-denoised data. **n,** expression patterns of gene *Aqp4* using the MIST-denoised data.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- extendedfigures.docx