Briefings in Bioinformatics, 2023, **24(2)**, 1–11
https://doi.org/10.1093/bib/bbad048
Advance access publication date 13 February 2023
Problem Solving Protocol

Identifying spatial domain by adapting transcriptomics with histology through contrastive learning

Yuansong Zeng†, Rui Yin†, Mai Luo, Jianing Chen, Zixiang Pan, Yutong Lu, Weijiang Yu and Yuedong Yang

Corresponding authors: Weijiang Yu. E-mail: weijiangyu8@gmail.com; Yuedong Yang. E-mail: yangyd25@mail.sysu.edu.cn

†YZ and RY contributed equally.

Abstract

Recent advances in spatial transcriptomics have enabled measurements of gene expression at cell/spot resolution meanwhile retaining both the spatial information and the histology images of the tissues. Accurately identifying the spatial domains of spots is a vital step for various downstream tasks in spatial transcriptomics analysis. To remove noises in gene expression, several methods have been developed to combine histopathological images for data analysis of spatial transcriptomics. However, these methods either use the image only for the spatial relations for spots, or individually learn the embeddings of the gene expression and image without fully coupling the information. Here, we propose a novel method ConGI to accurately exploit spatial domains by adapting gene expression with histopathological images through contrastive learning. Specifically, we designed three contrastive loss functions within and between two modalities (the gene expression and image data) to learn the common representations. The learned representations are then used to cluster the spatial domains on both tumor and normal spatial transcriptomics datasets. ConGI was shown to outperform existing methods for the spatial domain identification. In addition, the learned representations have also been shown powerful for various downstream tasks, including trajectory inference, clustering, and visualization.

Keywords: spatial transcriptomics, histopathological image, contrastive learning, spatial clustering, multi-modality

Introduction

Spatial transcriptomics (ST) technologies measure gene expression with spatial information [1], and current technologies are widely divided into two types: (1) high-plex RNA imaging (HPRI) methods employ probes targeting a few specific genes to localize mRNA transcripts. This type of methods include *in situ* sequencing or fluorescence *in situ* hybridization (FISH), such as MERFISH [2], seqFISH [3], CyCIF [4] and STARmap [5]; (2) the spatial tagging approaches use spatial barcodes to capture mRNA transcripts across tissue cross-sections and then perform deep sequencing. Typical methods include ST [6], XYZeq [7] and Visium from 10x Genomics (https://www.10xgenomics.com/). While HPRI methods can obtain single-cell resolution ST with greater depth, the spatial tagging platforms often provide the whole slide of hematoxylin and eosin (H&E)-stained histology images of the tissue, which might be utilized as another modality to remove the noise or error in gene expression [8]. Though these ST technologies have achieved great successes on the developing amyotrophic lateral sclerosis, human heart, Alzheimer's disease and the human squamous cell carcinoma [9, 10], exploiting the spatial domains remains challenging. One way for the spatial domain identification is the clustering of sequenced spots [11–14]. Nevertheless, conventional clustering methods are meeting grand challenges since they do not incorporate the spatial information of spots and are difficult to deal with the sparsity and high-dimensional features of gene expression data [15, 16].

To resolve these challenges, a wide variety of clustering algorithms have been developed for ST analysis [17, 18]. Several early spatial clustering methods account for the spatial dependency of gene expression by considering the similarity between adjacent spots. For example, BayesSpace [19] is a Bayesian statistical model that introduces spatial neighbor structure into the prior so that the nearest spots tend to be assigned into the same cluster. Similarly, Giotto [20] deciphers spatial domains by applying a hidden Markov random field model that integrates the spatial neighbor information. Unfortunately, these methods achieved limited performance since they don't capture the nonlinear characteristics of gene expression. For this reason, SEDR [21], a deep learning-based method, embeds the spatial structure of ST via a variational graph autoencoder network and simultaneously learns latent gene representations through the autoencoder network. Similarly, other methods include CCST [22] and DeepST [23]. These methods assume neighbored spots as the same type, which is not always true. STAGATE [24] refined the relations between spots by reconstructing the gene expression via an attention graph convolutional network [25]. These methods used only the gene expression information and are greatly influenced by noises inherent in the data. In fact, histopathological images have been shown

Yuansong Zeng is a PhD student in the School of Computer Science and Engineering at the Sun Yat-Sen University.
Rui Yin is a master student in the Department of Computer Science at the University of Hong Kong.
Mai Luo is a master student in the School of Computer Science and Engineering at the Sun Yat-Sen University.
Jianing Chen is a master student in the School of Computer Science and Engineering at the Sun Yat-Sen University.
Zixiang Pan is a master student in the School of Computer Science and Engineering at the Sun Yat-Sen University.
Yutong Lu is a professor in the School of Computer Science and Engineering and the National Super Computer Center at Guangzhou, Sun Yat-Sen University.
Weijiang Yu is a PhD student in the School of Computer Science and Engineering at the Sun Yat-Sen University.
Yuedong Yang is a professor in the School of Computer Science and Engineering and the National Super Computer Center at Guangzhou, Sun Yat-Sen University.
Received: September 27, 2022. **Revised:** December 26, 2022. **Accepted:** January 23, 2023

able to predict gene expression [15, 26]. Thus, it is promising to introduce image features as complementary information for gene expression.

To take full use of the information from images, several methods have been developed. For example, stLearn [27] computes the morphological distance between spots by using their corresponding features extracted from the histopathological image and applies these distances and the spatial structure to smoothen gene expression. SpaGCN [28] uses both the histopathological image and spatial coordinates to build relations between spots, which are then fed into a graph convolutional layer to propagate gene expression information between neighboring spots. Though these two methods only used the image to construct relations between spots, they did not include the image in the late training process. Whereas, incorrect spot relations may exist due to noise in the image data. To fully utilize the image information, conST [29] first concatenates gene expressions and the pre-extracted morphology features that are extracted from images via the Masked Auto-encoder into a feature vector [30], which is then fed into a graph convolutional network to learn latent representations. However, conST simply concatenates pairs of the image feature and gene expression into one feature vector, where two modalities can't be used to compensate each other to remove their respective noises.

With the development of deep learning techniques, multiple modalities could be learned by aligning their embeddings in the low-dimensional space using the contrastive learning strategy [31, 32]. The contrastive learning is a discriminative method that aims to pull similar samples closer and push apart different samples in an unsupervised manner [33] and has also been applied to scRNA-seq [34]. However, the technique hasn't been utilized to align multiple modalities in spatial multi-omics, such as gene expressions and histology images.

In this study, we propose a novel method ConGI to accurately decipher spatial domains by integrating gene expression and histopathological images, where the gene expression is adapted to image information through contrastive learning. The natural rationale of our method is to leverage the correspondence between gene expression and cellular phenotypical information at the spot level. To learn the common representations across two modalities while avoiding their respective noise information, we introduce three contrastive loss functions within and between modalities, including gene expression to gene expression, image to image and image to gene expression. The learned representations are then used for identifying the spatial domains through clustering methods. By comprehensive tests on tumor and normal datasets, ConGI was shown to outperform existing methods in terms of spatial domain identification. In addition, the learned representations from our model have also been used efficiently for various downstream tasks, including trajectory inference, clustering and visualization.

Materials and methods

Datasets and preprocessing

To evaluate the performance of our method, we employed seven ST datasets, including the human HER2-positive breast tumor dataset (HER2+) [35], the human dorsolateral prefrontal cortex dataset (spatialLIBD) [36], the human epidermal growth factor receptor (HER) 2-amplified (HER+) invasive ductal carcinoma (IDC) sample, the sections coronal and anterior of the mouse brain tissues, the breast invasive carcinoma (BRCA) and the mouse brain with single-cell resolution [37]. The HER2+ dataset was measured by ST technology [6], which included eight tissue sections with annotations from pathologists. We removed section C in the HER2+ due to the number of spots being less than 200. The other datasets were measured by 10x Visium (https://www.10xgenomics.com/resources/datasets). Each dataset contained histopathological images, gene expression at the spatial spots and their corresponding coordinates. For each histopathological image, we cropped$(\mathbf{W} \times \mathbf{H})$ pixels around each spot, where H and W were the height and width of image patches, respectively. Both W and H were set to 112, matching the diameter of each spot. We provided two strategies to preprocess the gene expression data of each tissue section. For the dataset generated by 10x Visium, we used PCA to reduce the dimension of gene expression to 300 as recommended in ref [28, 29]. For other datasets, we followed reference [15] to select the top 1000 highly variable genes; For a given spot, its counts were divided by the total counts for the given spot and multiplied by the scale factor of 1000 000. This was then natural-log transformed via log (1+x), where x was the normalized count.

The architecture of ConGI

ConGI is a deep learning-based method for deciphering spatial domains by integrating histopathological images and ST via contrastive learning. To achieve this, as illustrated in Figure 1, given pairs of ST (gene expression) and the image patch cropped from histopathological images at each spot, we apply two independent encoders (a convolutional neural network [CNN] $\mathbf{f}^{\mathrm{i}}$ and an multi-layer perceptron [MLP] $\mathbf{f}^{g}$) to learn the low-dimensional representations via contrastive learning. Concretely, we first distort the paired image patch and gene expression data slightly by adding noises to them through their corresponding data augmentation techniques. The augmented data sets are then fed into the encoders for images and gene expressions to learn the low-dimensional representations, separately. These learned representations are then projected into the space where contrastive loss is applied to jointly learn from pairwise data via three contrastive learning losses, including gene expression to gene expression (L_{g2g}), image to image (L_{i2i}), and image to gene expression (L_{i2g}). ConGI pulls the pairwise data within and between modalities together and contrasts the unmatching pairs apart. After training, the low-dimensional representations of gene expression and image patch were combined together for clustering via the mclust [38] package. More importantly, the learned representations have been used efficiently for various downstream tasks, including trajectory inference, clustering and visualization.

Encoders for gene expression and image

Gene expression encoder

The gene expression encoder is used for extracting the low-dimensional representations of gene expression $\mathbf{X} \in \mathbb{R}^{n\times d}$, where n and d are the number of spots and features, respectively. The gene expression encoder $\mathbf{f}^{g}$ is a neural network-based model applying a MLP with fully connected layers and nonlinearities. Specifically, for a given spot, we first generate two augmented gene expressions $\mathbf{x}_{g\_u}$ and $\mathbf{x}_{g\_v}$ by distorting the original gene expression $\mathbf{x}_{\mathrm{g}} \in \mathbf{X}$ of the spot through data augmentation techniques, such as the random mask and random swap. The augmented data sets are then fed into gene expression encoder $\mathbf{f}^{g}$ to obtain the corresponding latent vectors $\mathbf{z}_{g\_u}$ and $\mathbf{z}_{g\_v}$ as follows:

$$\mathbf{z}_{g\_u} = \mathbf{f}^{g}\ \left(\mathbf{x}_{g\_u}\right) \tag{1}$$

$$\mathbf{z}_{g\_v} = \mathbf{f}^{g}\ \left(\mathbf{x}_{g\_v}\right) \tag{2}$$

Image encoder

The image encoder $\mathbf{f}^{\mathrm{i}}$ aims to capture the morphological feature of each spot from the image patch. Here, the backbone of the image encoder $\mathbf{f}^{\mathrm{i}}$ is a classic CNN DenseNet121 [39] with the pretrained ImageNet weights. In our setting, we reserve the pretrained weights of DenseNet121 due to the limited training images. The input of the image encoder is the image patch cropped from the whole histopathological image, which is only matching to the corresponding gene expression. Similarly, we also generate two augmented image patches $\mathbf{x}_{\mathrm{i}\_u}$ and $\mathbf{x}_{\mathrm{i}\_v}$ by distorting the original image patch $\mathbf{x}_{\mathbf{i}}$ through data augmentation techniques, such as RandomGrayscale and RandomHorizontalFlip. The augmented data sets are then fed into the image encoder $\mathbf{f}^{\mathrm{i}}$to obtain the corresponding latent vectors $\mathbf{z}_{\mathrm{i}\_u}$ and $\mathbf{z}_{\mathrm{i}\_v}$ as follows:

$$\mathbf{z}_{\mathrm{i}\_u} = \mathbf{f}^{\mathrm{i}}\ \left(\mathrm{x}_{\mathrm{i}\_u}\right) \tag{3}$$

$$\mathbf{z}_{\mathrm{i}\_v} = \mathbf{f}^{\mathrm{i}}\ \left(\mathrm{x}_{\mathrm{i}\_v}\right) \tag{4}$$

Projection head

For learning the common information from image and gene expression, we apply a small neural network projection head $\mathbf{pg}\,(.)$ layer to take the gene expression and the image representations into a shared space. Here, the $\mathbf{pg}\,(.)$ layer consists of a two-layer neural network, which connect directly with the image and gene expression encoders. The $\mathbf{pg}\,(.)$ projects the latent features $\mathbf{z}_{g\_u}$ and $\mathbf{z}_{g\_v}$ of gene expression and the latent features $\mathbf{z}_{\mathrm{i}\_u}$ and $\mathbf{z}_{\mathrm{i}\_v}$ of the image patch into a shared space as follows:

$$\mathbf{h}_{g\_u} = pg\left(\mathbf{z}_{g\_u}\right) \tag{5}$$

$$\mathbf{h}_{g\_v} = pg\left(\mathbf{z}_{g\_v}\right) \tag{6}$$

$$\mathbf{h}_{\mathrm{i}\_u} = pg\left(\mathbf{z}_{\mathrm{i}\_u}\right) \tag{7}$$

$$\mathbf{h}_{\mathrm{i}\_v} = pg\left(\mathbf{z}_{\mathrm{i}\_v}\right) \tag{8}$$

Contrastive learning loss functions

To learn the common representations between images and gene expressions while avoiding their noise information, we introduce three contrastive learning losses within and between modalities. The contrastive learning between modalities aims to pull paired low-dimensional representations $\mathbf{h}_{i\_u}$ *and* h_{g_u} of image patch and gene expression together while contrasting those unmatching pairs apart. To better learn the characteristics of each modality, we conduct contrastive learning within the modality for image and gene expression, respectively. For achieving these goals, we design the contrastive learning loss function following SimCLR [33], which is a simple framework for contrastive learning of visual representations. SimCLR achieves SOTA performance on many datasets. Concretely, we randomly build a minibatch of N spots and define the contrastive prediction task on paired spots derived from the minibatch, resulting in 2 N paired spots. We do not build negative spots explicitly. By contrast, similar to reference [40], given a positive pair, we take the other 2(N-1) spots within a minibatch as negative spots. Thus, the within and between contrastive loss functions for positive pairs (h_g^u, h_g^v), $(h_\mathrm{i}^u, h_\mathrm{i}^v)$, and (h_i^u, h_g^v) can be defined as follows:

$$\mathbf{L}^{1}_{h_g^u,h_g^v} = -\log\frac{\exp\left(\frac{sim\left(h_g^u,h_g^v\right)}{\tau}\right)}{\sum_{k=1}^{2N} 1_{[k\neq u]}\exp\left(\frac{sim\left(h_g^u,h_g^k\right)}{\tau}\right)} \tag{9}$$

$$\mathbf{L}^{2}_{h_\mathrm{i}^u,h_\mathrm{i}^v} = -\log\frac{\exp\left(\frac{sim\left(h_\mathrm{i}^u,h_\mathrm{i}^v\right)}{\tau}\right)}{\sum_{k=1}^{2N} 1_{[k\neq u]}\exp\left(\frac{sim\left(h_\mathrm{i}^u,h_\mathrm{i}^k\right)}{\tau}\right)} \tag{10}$$

$$\mathbf{L}^{3}_{h_\mathrm{i}^u,h_g^v} = -\log\frac{\exp\left(\frac{sim\left(h_\mathrm{i}^u,h_g^v\right)}{\tau}\right)}{\sum_{k=1}^{2N} 1_{[k\neq u]}\exp\left(\frac{sim\left(h_\mathrm{i}^u,h_g^k\right)}{\tau}\right)} \tag{11}$$

where $1_{[k\neq u]} \in \{0,1\}$ is an indicator function evaluating to 1 if $\mathbf{k} \neq u$, and τ means the temperature parameter. The term $\mathbf{s}im\left(h_\mathrm{i}^u, h_g^v\right) = \left(h_\mathrm{i}^u\right)^T h_g^v / \left\|h_\mathrm{i}^u\right\|\ \left\|h_g^v\right\|$represents the dot product between h_i^u *and* h_g^v. Each contrastive loss is calculated across all positive pairs, i.e. both $\left(h_\mathrm{i}^u, h_g^v\right)$ and $\left(h_g^u, h_\mathrm{i}^v\right)$ in a mini-batch. Thus, three contrastive losses can be formulated as follows:

$$L_{g2g} = \frac{1}{2N}\sum_{k=1}^{N}\left[L^1\left(2k-1,2k\right) + L^1\left(2k,2k-1\right)\right] \tag{12}$$

$$L_{\mathrm{i}2\mathrm{i}} = \frac{1}{2N}\sum_{k=1}^{N}\left[L^2\left(2k-1,2k\right) + L^2\left(2k,2k-1\right)\right] \tag{13}$$

$$L_{g2\mathrm{i}} = \frac{1}{2N}\sum_{k=1}^{N}\left[L^3\left(2k-1,2k\right) + L^3\left(2k,2k-1\right)\right] \tag{14}$$

Finally, the total loss of our model can be summed as follows:

$$\mathbf{L} = \mathbf{L}_{g2\mathrm{i}} + \lambda_1\mathbf{L}_{g2g} + \lambda_2\mathbf{L}_{\mathrm{i}2\mathrm{i}} \tag{15}$$

where λ_1 *and* λ_2 are hyper-parameters used for controlling the contribution of losses $\mathbf{L}_{g2g}$*and* $\mathbf{L}_{\mathrm{i}2\mathrm{i}}$for the final loss. For all datasets, $\lambda_1 = \lambda_2 = 0.1$.

After the model is trained, the final representation of each spot can be defined as follows:

$$\mathbf{z} = \mathbf{z}_g + \alpha\mathrm{z}_\mathrm{i} \tag{16}$$

where α is a hyper-parameter used for controlling the contribution for the final representation of each spot. For all datasets, $\alpha = 0.1$.

Clustering

We take different strategies to identify spatial domains using learned representations from ConGI. When the number of spatial domains is specific, we follow reference [24] to apply the mclust [38] clustering algorithm to decipher spatial domains. For datasets without the number of clusters, we identify spatial domains through the Louvain algorithm implemented in the popular package scanpy [41]. After clustering, we follow SpaGCN to provide an optional refinement step for the clustering results. In this step, for a given spot, we reassign its label to the same spatial domain as the primary label of its neighboring spots if more than half of its neighboring spots are assigned to a different domain. We perform cluster refinement for all datasets.

Hyper-parameters setting

The ConGI was implemented in python and PyTorch. For the gene expression encoder, the dimensions of hidden layers were set to [128,128]. The image encoder used the DenseNet121 with default pretrained weights from torchvision.models. The dimensions of the projection head were set to [128,128]. Our models were optimized via the AdamW optimizer with a learning rate of 0.003. The training batch size was set to 32 when the total number of spots was <1000. In other situations, the training batch size was set to 64. In the processing of the refinement step, the number of neighboring spots was set to 24 when the number of total spots was >1000. Otherwise, the number of neighboring spots was set to 4. All results reported in this paper were conducted on Ubuntu 18.04.7 LTS with Intel® Core (TM) i7-8700K CPU @ 3.70 GHz and 256 GB memory.

Evaluation criteria

Clustering performance

The clustering results are evaluated through three commonly used clustering metrics, including the Normalized Mutual Information (NMI) [42], Adjusted Rand Index (ARI) [43] and Clustering Accuracy (CA) [44].

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}} \qquad (17)$$

where $\mathbf{a}_i$ and $\mathbf{b}_j$ are the number of samples appearing in the *i-th* predicted cluster and the *j-th* true cluster, respectively. $\mathbf{n}_{ij}$ means the number of overlaps between the *i-th* predicted cluster and the *j-th* true cluster.

$$NMI\,(Y, C) = \frac{2 \times [H(Y) - H\,(Y|C)]}{[\mathrm{H}\,(\mathrm{Y}) + \mathrm{H}\,(\mathrm{C})]} \qquad (18)$$

where C and Y are the predicted clusters and the real clusters, respectively. The function H ( ) is used for calculating the entropy.

$$\mathrm{CA} = \max_{m} \frac{\sum_{i=1}^{n} 1\left\{l_i = m\,(c_i)\right\}}{n} \qquad (19)$$

where n is the entire number of samples, and m ranges over all probable one-to-one mapping between clustering assignment c_i and true label l_i. ARI measures the similarity for the clustered results (the true and predicted labels), which takes into account the number of clusters, the number of elements in each cluster and the number of elements that are assigned to the same cluster. In comparison, CA measures the best-matching between labels. NMI measures the mutual information between labels, which does not consider the number of clusters.

Benchmark methods

To evaluate the performance of our method, we compared ConGI with other tools including STAGATE, scanpy, conST, SpaGCN, SEDR, stLearn, BayesSpace, Giotto and Seurat [45]. For all competing methods, we used the default hyper-parameters and preprocessing for datasets recommended in the original paper to test all datasets. For methods scanpy and Seurat, we used the default value of parameter 'resolution' to determine the number of clusters. For other methods, we fed them with the number of true clusters for clustering.

Results

Application to spatially resolved tumor sample transcriptomics

To demonstrate the performance of our method on the spatially resolved tumor sample transcriptomics, we analyzed the human HER2-positive breast tumor dataset (HER2+) and the HER2-amplified (HER+) IDC. We first compared the performance of our method with competing methods on the HER2+ data. As shown in Figure 2A, ConGI outperformed all competing methods in terms of the average ARI. Specifically, the average ARI of our method was 11.2% higher than the 2nd-ranked method Giotto. stLearn obtained a decent performance with an average ARI value of 0.22. SpaGCN and conST achieved similar performance. BayesSpace and STAGATE achieved similar clustering results, both of them performed better than SEDR. Interestingly, methods (i.e. scanpy, SEDR and STAGATE) only using the gene expression information performed worse than methods (i.e. ConGI, stLearn, conST and SpaGCN) considering image information. These results demonstrated that image features were beneficial for clustering spatial domains. To further confirm the superior results, we compared the manually annotated regions for section E1. Figure 2B and Supplementary Figure S1 showed that our method revealed spatial domains that agreed better with the manually annotated tissue regions than competing methods. Although stLearn and conST also utilized histopathological images, they still assigned the wrong spatial domains for most spots. When the HER2+ dataset was tested by the other metrics (NMI and CA), a similar trend could be found (Supplementary Figure S2). The average ARI values of all methods were <0.5, as also observed in the reference [26]. This is likely due to the low number of spots and the high missing rate of the data. As a result, all sections of the HER2+ dataset contain <700 spots with the average missing rate of 85.3%. A similar trend was observed on the breast cancer dataset BRCA measured by another sequencing platform 10x Genomics (Supplementary Figure S3), where our method was 2% higher than the 2nd-best method stLearn in terms of the ARI values. We didn't compare with BayesSpace because it didn't run correctly on this dataset.

We further evaluated the performance of our model on the IDC dataset with the nearly single cell super-resolution. Pathologists identified regions of benign hyperplasia, predominantly invasive carcinoma (IC) and carcinoma *in situ*, which were used as ground truth labels to evaluate the CA. As shown in Figure 2C and Supplementary Figure S4, our model achieved the best CA with an ARI of 0.448, which was 9.8% higher than the 2nd method BayesSpace. Methods STAGATE, SpaGCN, conST and Seurat achieved similar results with an ARI value of around 0.330. Both SEDR and stLearn achieved similar performance with an ARI value of about 0.286. The scanpy method achieved the lowest performance with the ARI value of 0.231. Our method could accurately predict the largest non-tumor region while competing methods mixed the non-tumor region with other tumor regions (Figure 2C). Though SpaGCN and conST also combined the histology images, they didn't exactly predict the non-tumor region. The success of our method should be attributed to the contrastive learning module because its removal (ConGI w/o L_g2i) will also mix the non-tumor region with other tumor regions (Supplementary Figure S4). All

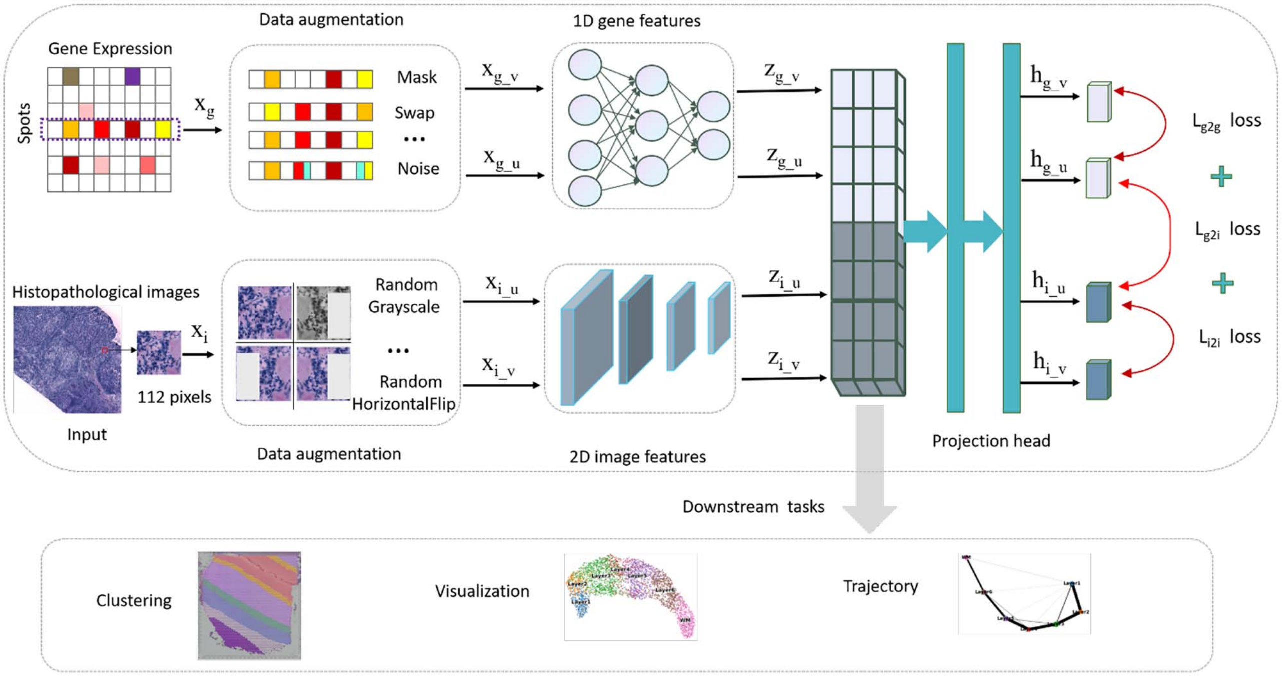

Figure 1. The schematic overview of the ConGI for identifying spatial domains by integrating image and gene expression. The input of ConGI is the paired image patch and gene expression of a spot that is firstly slightly distorted by adding noises through their corresponding data augmentation techniques. The augmented data sets are then fed into encoders to learn the low-dimensional embeddings, respectively. In parallel, these learned representations are pulled together via three contrastive learning loss functions, within gene expression (L_g2g), image (L_i2i) and between the two modules (L_g2i). After the model is trained, the raw images and gene expressions are input into the trained model to generate low-dimensional embeddings for downstream tasks such as clustering, visualization and trajectory inference.

methods are difficult to distinguish *in situ* and benign hyperplasia domains. This is likely because these two domains have similar expressed patterns of genes or identical image features, as also shown in the uniform manifold approximation and projection (UMAP) visualization of the image and gene expression in the 2D space (Supplementary Figure S5).

To confirm that ConGI could investigate the biological relevance, we detected the spatially variable genes (SVGs) for the IDC dataset via the same detection strategies used in SpaGCN. The results showed that FAM234B enriched in domain 0. Similarly, MUC1 was enriched in domain 3 in the histopathological image. The results demonstrated that ConGI showed similar biological tissue patterns to manual annotations.

Application to spatially resolved normal sample transcriptomics

To evaluate the performance of our method on the spatially resolved normal sample transcriptomics, we analyzed a human dorsolateral prefrontal cortex (spatialLIBD) dataset and sections coronal and anterior of the mouse brain tissue. We first analyzed the spatialLIBD consisting of 12 tissue slices from the human dorsolateral prefrontal cortex in three human brains, which spanned six neuronal layers and white matter. As shown in Figure 3B, the average ARI of our method was 4.82% higher than the second-ranked method STAGATE. Though SEDR achieved decent performance and ranked third, the average ARI value was much lower (13.42%) than that of our method. A similar trend could be found when measuring our method by NMI and CA (Supplementary Figure S6). In addition, we compared the manually annotated layer structure for section 151509 in Figure 3C and Supplementary Figure S7. Our method revealed spatial domains that better agreed with the manually annotated tissue layers (Figure 3A) than competing methods. For instance, most spots were mixed wrongly in the spatial domains predicted by scanpy and Giotto. For further investigation of the reasons, we showed UMAP visualizations of image data, gene expressions and concatenated data (simply combined gene expression and image data) and aligned embeddings (embeddings aligned by our method) in the 2D space (Supplementary Figure S8). The UMAP visualization of gene expression and image data showed that most layers were mixed, such as layer 2 and layer 3 due to noises in each modality. The concatenated data also showed similar results since the simple concatenation operation is difficult to couple the image data and gene expression and avoid their respective noise. Nevertheless, our method can separate most spatial domains in the 2D space since noises were removed by learning common embeddings through contrastive learning. We further showed SVGs for section 151509, where HPCAL1 and MBP were enriched in domains 4 and 7, respectively (Figure 3C). On this biggest dataset (section 151509) containing about 5000 spots, our method needed 7Gb memory and 1.75 hours. Since our method was trained through mini-batches, the method can be extended to large datasets. The long-running time was caused by the DesnsNet121 with about 7.98 million parameters that needed to be updated in the training phase. The computationally intensive problem can be further alleviated by pretraining DesnsNet121 using histological images without updating during the training phase.

Our method achieved the highest ARI value of 0.401 on the complex mouse brain anterior tissue (Figure 4 and Supplementary Figure S9). We found that the performance of our method was slightly higher than the second-ranked method Giotto. This is likely because the information from the image included less useful information for deciphering the domains. Actually, we

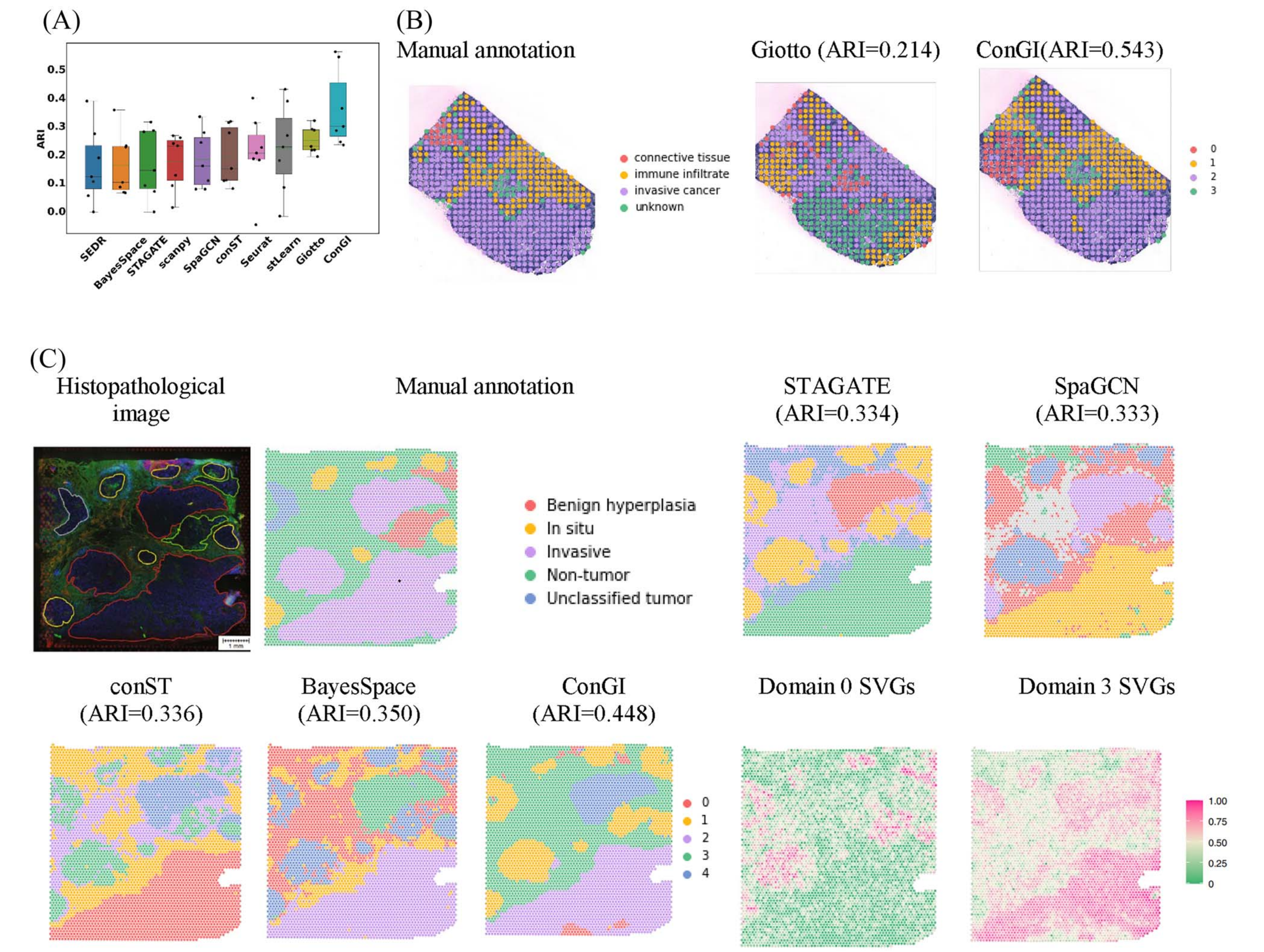

Figure 2. Spatial domains and SVGs detected in the human HER2-positive breast tumor (HER2+) dataset and the IDC dataset. (**A**) Boxplot of CA in all sections of the HER2+ dataset in terms of ARI values for all methods. (**B**) Manually annotated regions of section E1 of the HER2+ dataset, spatial domains detected by Giotto and ConGI. (**C**) The histopathological image and manually annotated regions for the IDC data, spatial domains detected by STAGATE, SpaGCN, conST, BayesSpace and ConGI, and the spatial expression patterns of SVGs for ConGI predicted spatial domains 0 (FAM234B) and 3 (MUC1).

did not find obvious boundaries among spatial domains in the histopathological image of the mouse brain anterior tissue. The other image-based method conST wrongly predicted all spatial domains into one domain. All methods achieved low performance with ARI values of less than 0.5. This might be ascribed to many sub-regions of spots in the mouse brain anterior dataset such as AOB-GI, AOB-GR and AOB-MI and be solved by adding marker genes for each sub-region in the future. For the section coronal of the mouse brain tissue (Supplementary Figure S10), our method achieved 2.5% higher ARI than SEDR, the 2nd-best method. We didn't compare with BayesSpace because it didn't run correctly on this dataset. In addition, we didn't compare our method with scanpy since the ground truth labels of the coronal of the mouse brain were annotated by it [46]. These results showed that our method could still achieve a decent performance when there were not any clear boundaries between spatial domains in the histopathological image.

We also tested our method on the single-cell resolution ST data that don't have the H&E images (by setting L_g2i and L_i2i as zero). On the mouse brain ST dataset generated by MERFISH technology [37], ConGI achieved 1.8% higher ARI than Seurat, the second-ranked method (Supplementary Figure S11). This is as expected because contrastive learning was shown to have superior performance in the method CLEAR [34] for single-cell data analysis.

ConGI learns effective latent representations from image and gene expression

To show that the representations learned by our model could reveal the distance between spatial domains and depict the trajectory, we visualized the low-dimensional representations through the UMAP [47] in the 2D space and plotted the trajectory inference via the PAGA [48] algorithm employed in the package scanpy. We took tissue section E1 from the HER2+ dataset and tissue section 151509 from the spatialLIBD dataset as examples. We first evaluated our method in section E1. As shown in Figure 5A and Supplementary Figure S12, STAGATE and BayesSpace completely mixed spots in regions of connective tissue, invasive cancer and immune infiltrate. A similar trend could be found in the UMAP of other competing methods, such as Giotto and SpaGCN. In contrast, our method separated explicitly the domains invasive cancer and immune infiltrate. Though ConGI didn't separate the immune infiltrate and the connective tissue, other methods were also difficult to distance them. This is likely because they were

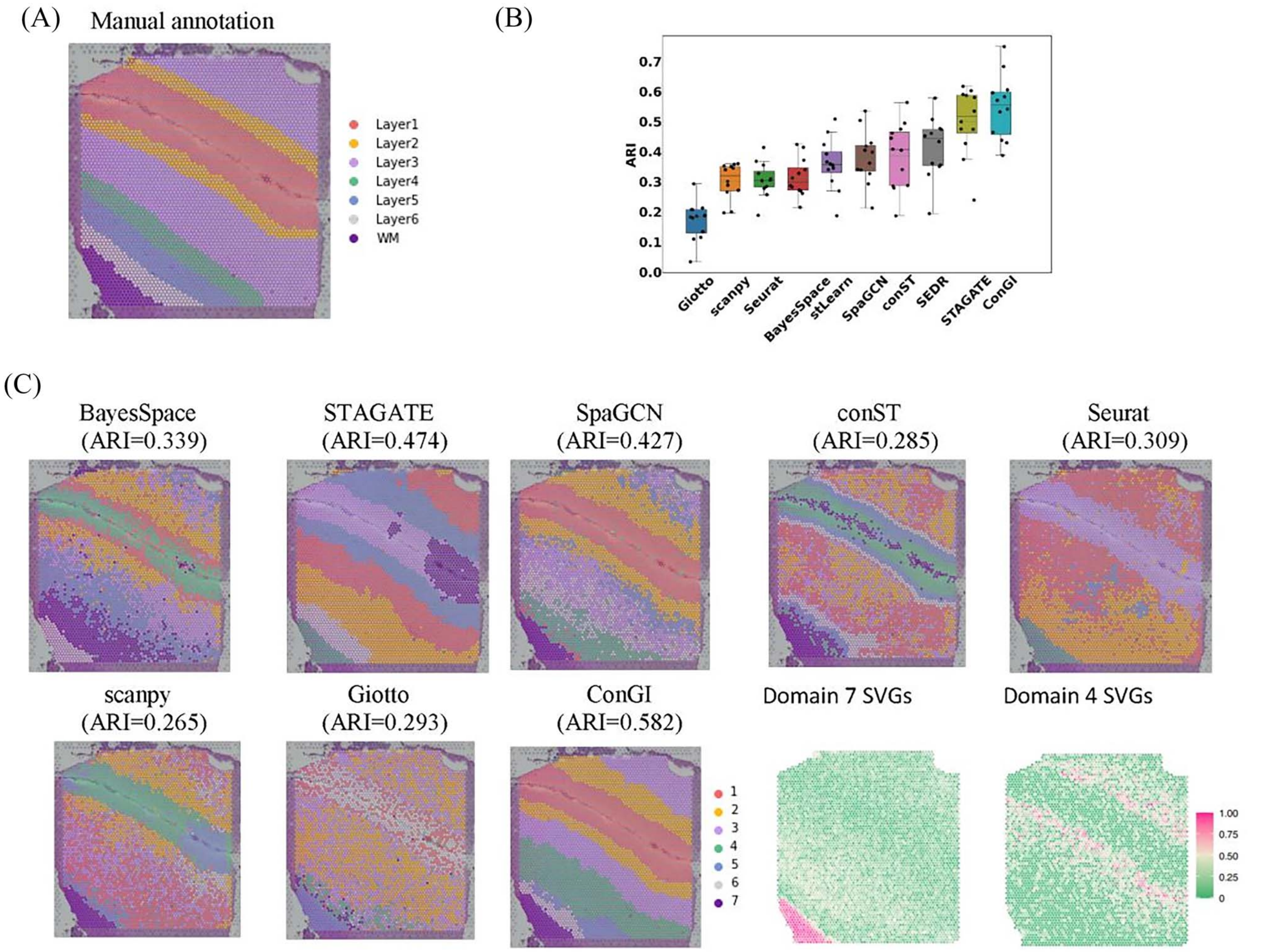

Figure 3. Spatial domains and SVGs detected in the human dorsolateral prefrontal cortex dataset (spatialLIBD). (**A**) Histopathological image of the tissue section 151509 with manually annotated layer structure. (**B**) Boxplot of CA in all sections of the spatialLIBD dataset in terms of ARI values for all methods. (**C**) Spatial domains were detected by BayesSpace, STAGATE, SpaGCN, conST, Seurat, scanpy, Giotto and ConGI in section 151509, respectively, and the spatial expression patterns of SVGs for ConGI predicted domain 4 (HPCAL1) and domain 7 (MBP).

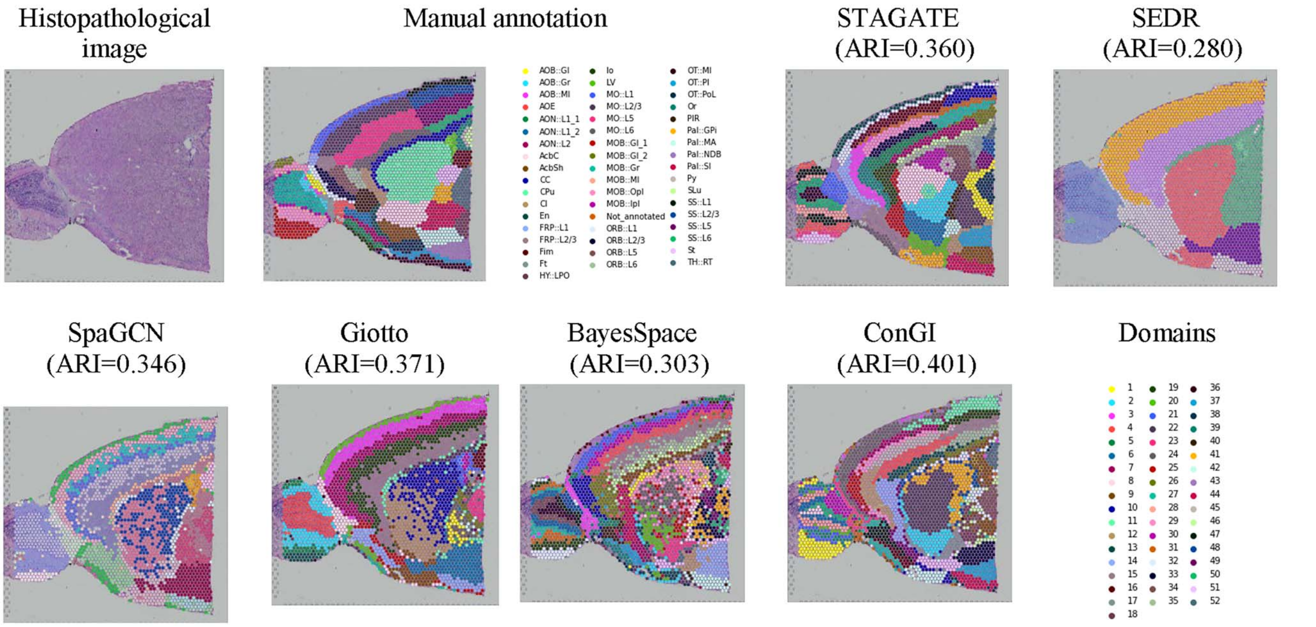

Figure 4. The histopathological image and manually annotated regions for the mouse brain anterior sample, spatial domains detected by SEDR, STAGATE, SpaGCN, Giotto, BayesSpace and ConGI.

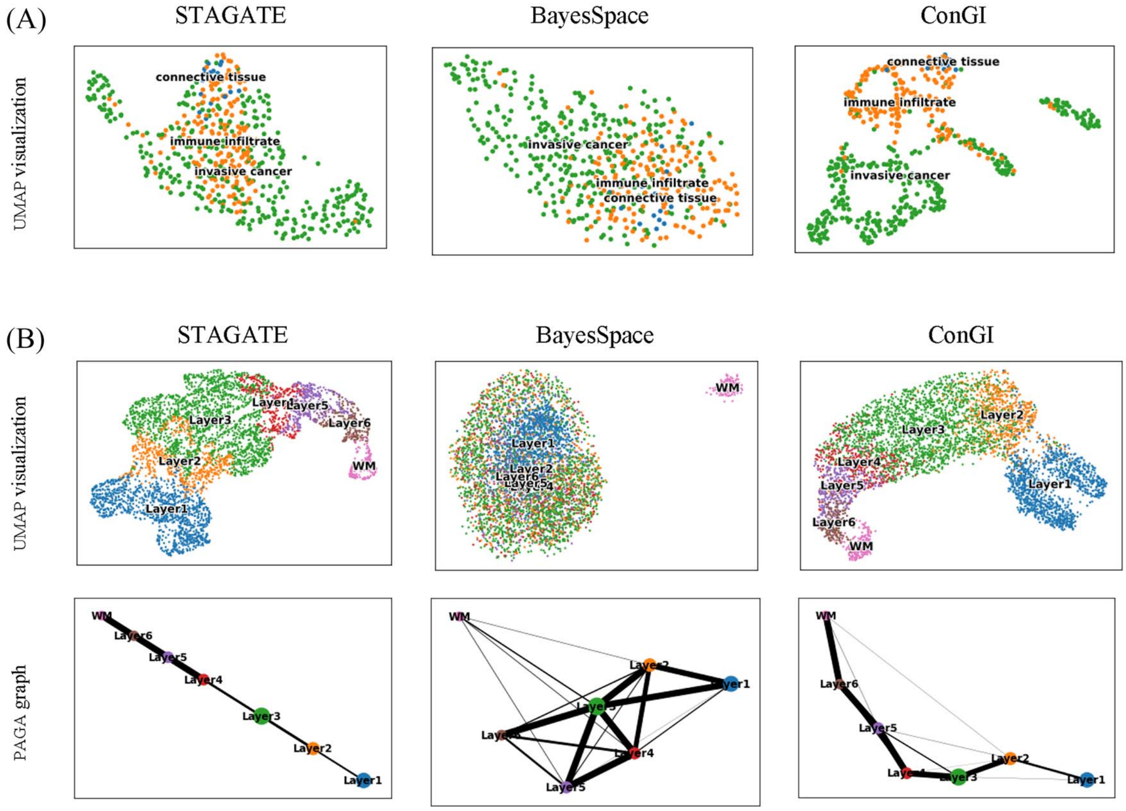

Figure 5. UMAP visualizations and PAGA graphs using representations generated by STAGATE, BayesSpace and ConGI, respectively (**A**) in the section E1 of the dataset HER2+ and (**B**) the section 151509 of the dataset spatialLIBD.

difficult to distinguish. A similar trend could be found when visualizing the tissue section 151509 of the spatialLIBD dataset. As shown in Figure 5B and Supplementary Figure S13, In the UMAP plots of BayesSpace, scanpy and Seurat, most spots of different layers were mixed, such as spots from layers 2 to 4. In the UMAP plots of conST, spots from layer 1 and spots from layer 6 were mixed. However, our method and STAGATE could well separate most spots in different layers. Since section 151509 contained explicitly the layer structure of the human dorsolateral prefrontal cortex, we further validate the inferred trajectory based on the representations generated by all methods. The PAGA graphs of both ConGI and STAGATE representations showed a nearly linear development trajectory from layer 1 to layer 6 as well as the similarity between adjacent layers. Nevertheless, the PAGA results of BayesSpace showed that most layers were mixed. Layer 6 and layer 4 were mixed in the PAGA results of SpaGCN. The results of UMAP and trajectory inference demonstrated that our method could learn effective representations for downstream tasks.

Ablation study

To investigate the contributions of components for our method ConGI, we conducted ablation studies on the HER2+ dataset. As shown in Figure 6, we first tested whether the performance of our model benefited by learning information adaptively from histopathological images. We found that the performance of our method decreased by 3% if we didn't update the weights of the DenseNet121 through the histopathological images (Adaption) in the training stage. These changes demonstrated that the performance of our method was improved by learning information adaptively from histopathological images. The removal of the contrastive learning loss within gene expressions ($\mathbf{L}_{g2g}$) and images ($\mathbf{L}_{i2i}$) caused decreases of 1.5% and 2.4%, respectively. The removal of the contrastive learning loss between gene expression and the image ($\mathbf{L}_{g2i}$) caused a significant decrease of 11%, which indicated that our model was able to effectively learn the common embeddings across image and gene expression and avoid their respective noise information via contrastive learning. We have investigated how weights of losses influence the performance of our method by changing the weights of L_g2i, L_g2g and L_i2i. As shown in Supplementary Figure S14, ConGI changed the performance with different weights of L_g2i, L_g2g and L_i2i but the changes were generally in a reasonable range. As shown in Supplementary Figure S15, for the IDC, the complete removals of L_g2i, L_g2g, L_i2i and adaption caused decreases of 9%, 0.8%, 0.3%, 1% on the ARI values, respectively. A similar trend could be found in the mouse brain dataset. For the spatialLIBD dataset, the complete removals of L_g2i, L_g2g, L_i2i and adaption caused decreases of 13%, 5.1%, 6.3% and 17% in terms of the average ARI values, respectively.

Discussion

Accurate identification of spatial domains is essential for researchers to understand tissue organization and biological functions. In this study, we proposed a novel method ConGI to accurately decipher spatial domains by integrating gene expression and the histopathological image, where the gene expression is adapted to image information through contrastive learning. The learned representations are then used for deciphering the spatial domains through a clustering method. By comprehensive tests on cancer and normal ST datasets, ConGI was shown to outperform existing methods in terms of spatial domain identification. More

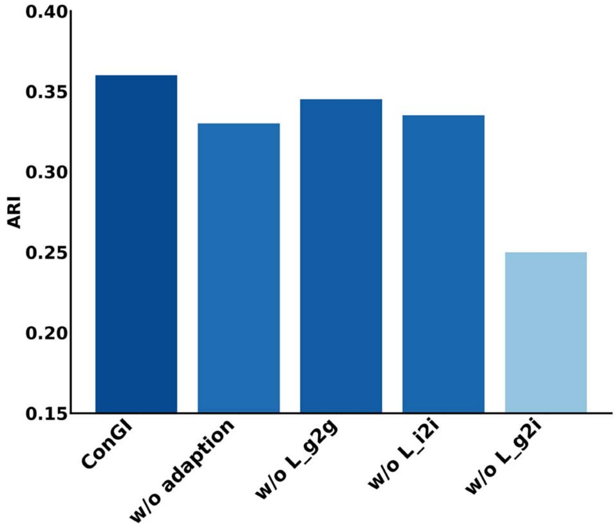

Figure 6. The average ARI values of ConGI on the HER2+ dataset by excluding contrastive losses L_g2i, L_i2i, L_g2g or adaption. The y-axis represents the average ARI value of all sections in the HER2+ dataset.

importantly, the learned representations from our model have also been used efficiently for various downstream tasks, including trajectory inference, clustering and visualization.

While a few methods, such as stLearn and SpaGCN, have been developed for deciphering spatial domains by combining the gene expression and histopathological images to conquer noises in the gene expression, they only use the histopathological images to construct spot relations without updating in the training stage. This may lead to poor performance if the image features construct the wrong spot relationships, since the spot relations are not updated in the training phase. Though the other method conST also integrates the image features, it simply concatenates the information from gene expression and image into one feature vector as the input to train the model. Such structure cannot accommodate the different processing needs of each modality and cannot effectively reduce their respective noises. In parallel, several methods use the pre-extracted image features or simply concatenate image features and gene expressions, but they haven't fully utilized different modalities to remove their inherent noises. In contrast, ConGI adapts gene expression with histology images to reduce their respective noises by aligning embeddings of these two modalities in the latent space through contrastive learning. The aligned embeddings are beneficial for various downstream such as clustering, visualization, and trajectory inference. Concretely, our model applies CNN-based and MLP-based models to learn low-dimensional representations from image and gene expression separately, where the gene expression is adapted to image information through contrastive learning. Especially, the image feature in our framework could be learned adaptively through contrastive loss, which is beneficial to avoid the noise information of images. ConGI can efficiently learn the common embeddings across two modalities while avoiding their respective noise. In addition, our method can be applied to ST datasets with different resolutions. By comprehensive tests on cancer and normal datasets, ConGI was shown to outperform existing methods in terms of spatial domain identification.

Despite the superior performance, ConGI can be improved in several aspects. Firstly, as a deep learning method, our model suffers from poor interpretability. This can be addressed in the future by using interpretation techniques [49], or through downstream analysis such as spatial variable genes identification that can alleviate the problems and bring insights into the cluster labels. Secondly, we use the DenseNet121 with pretrained ImageNet weights for extracting the features from histopathological images while having not fully used the big models trained using the histopathological images in the field of biological medicine. With the relatively easy acquirement of histopathological images, it is promising to pre-train a specific big model on histopathological images and use it in our extraction. In conclusion, this study provided a novel method to decipher the spatial domains by learning efficient representation from gene expressions and images via contrastive learning. The learned representations have been used efficiently for various downstream tasks, including trajectory inference, clustering and visualization. This method will be particularly useful with the rapidly increasing ST datasets.

Key Points

- Existing methods for deciphering spatial domains only use histopathological images to construct spot relations without updating in the training stage, or simply concatenate the information from gene expression and image into one feature vector.
- Here, we propose a novel method ConGI to accurately decipher spatial domains by integrating gene expression and histopathological images, where the gene expression is adapted to image information through contrastive learning. We introduce three contrastive loss functions within and between modalities to learn the common representations across two modalities while avoiding their respective noise.
- By comprehensive tests on tumor and normal spatial transcriptomics datasets, ConGI was shown to outperform existing methods in terms of spatial domain identification. More importantly, the learned representations were powerful for various downstream tasks, including trajectory inference, clustering, and visualization.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

National Key R&D Program of China (2020YFB0204803); National Natural Science Foundation of China (12126610 and 62201150); Guangdong Key Field R&D Plan (2019B020228001, 2018B010109006); Introducing Innovative and Entrepreneurial Teams (2016ZT06D211).

Code availability

All source codes used in our experiments have been deposited at https://github.com/biomed-AI/ConGI.

Data availability

The spatial transcriptomics datasets that support the findings of this study are available here: (1) human HER2-positive breast tumor ST data https://github.com/almaan/HER2st/. (2) The

LIBD human dorsolateral prefrontal cortex (DLPFC) data were acquired with 10x Visium composed of spatial transcriptomics data acquired from twelve tissue slices (http://research.libd.org/spatialLIBD/). (3) The coronal and anterior sections of the mouse brain dataset were obtained from 10x Visium (https://www.10xgenomics.com/resources/datasets). (4) the human epidermal growth factor receptor (HER) 2-amplified (HER+) invasive ductal carcinoma (IDC) (https://support.10xgenomics.com/spatial-gene-expression/datasets). (5) The breast invasive carcinoma (BRCA) dataset was obtained from 10x Visium (https://www.10xgenomics.com/resources/datasets). (6) The mouse brain dataset was obtained from the website: https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248/.

References

1. Moses L, Pachter L. Museum of spatial transcriptomics. *Nat Methods* 2022;**19**(5):534–46.
2. Chen KH, Boettiger AN, Moffitt JR, *et al.* Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015; **348**(6233):aaa6090.
3. Shah S, Takei Y, Zhou W, *et al.* Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell* 2018;**174**(2):363–376.e16.
4. Lin J-R, Izar B, Wang S, *et al.* Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *Elife* 2018;**7**:e31657.
5. Wang X, Allen WE, Wright MA, *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018;**361**(6400):eaat5691.
6. Ståhl PL, Salmén F, Vickovic S, *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;**353**(6294):78–82.
7. Lee Y, Bogdanoff D, Wang Y, *et al.* XYZeq: spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. *Sci Adv* 2021;**7**(17):eabg4755.
8. Song Q, Su J. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief Bioinform* 2021;**22**(5):bbaa414.
9. Chen W-T, Lu A, Craessaerts K, *et al.* Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* 2020;**182**(4):976–991.e19.
10. Ji AL, Rubin AJ, Thrane K, *et al.* Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* 2020;**182**(2):497–514.e22.
11. Blondel VD, Guillaume J-L, Lambiotte R, *et al.* Fast unfolding of communities in large networks. *J Stat Mech* 2008;**2008**(10):P10008.
12. Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat* 1979;**28**(1):100–8.
13. Zeng Y, Wei Z, Zhong F, *et al.* A parameter-free deep embedded clustering method for single-cell RNA-seq data. *Briefings in Bioinformatics* 2022;**23**(5):bbac172.
14. Zeng Y, Zhou X, Rao J, *et al.* Accurately clustering single-cell RNA-seq data by capturing structural relations between cells through graph convolutional network. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Piscataway, NJ: IEEE, 2020, pp. 519–22.
15. Zeng Y, Wei Z, Yu W, *et al.* Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Brief Bioinform* 2022;**23**(5):bbac297.
16. Rao J, Zhou X, Lu Y, *et al.* Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural networks. *Iscience* 2021;**24**(5):102393.
17. Singhal V, Nigel C, Lee J, Liu J, BANKSY: a spatial omics algorithm that unifies cell type clustering and tissue domain segmentation. bioRxiv. 2022.
18. Shan X, Chen J, Dong K, *et al.* Deciphering the spatial modular patterns of tissues by integrating spatial and single-cell transcriptomic data. *J Comput Biol* 2022;**29**(7):650–63.
19. Zhao E, Stone MR, Ren X, *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* 2021;**39**(11): 1375–84.
20. Dries R, Zhu Q, Dong R, *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* 2021;**22**(1):1–31.
21. Fu H, Xu H, Chong K, *et al.* Unsupervised spatially embedded deep representation of spatial transcriptomics. *bioRxiv* 2021.
22. Li J, Chen S, Pan X, *et al.* Cell clustering for spatial transcriptomics data with graph neural networks. *Nat Comput Sci* 2022;**2**(6): 399–408.
23. Xu C, Jin X, Wei S, *et al.* DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res* 2022;**50**:e131.
24. Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention autoencoder. *Nat Commun* 2022;**13**(1):1–12.
25. Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks. *stat* 2017;**1050**(20):10.48550.
26. Pang M, Su K, Li M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv* 2021.
27. Pham DT, Tan X, Xu J, Grice LF, *et al.* stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. bioRxiv. 2020.
28. Hu J, Li X, Coleman K, *et al.* SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* 2021;**18**(11):1342–51.
29. Zong Y, Yu T, Wang X, *et al.* conST: an interpretable multimodal contrastive learning framework for spatial transcriptomics. bioRxiv. 2022.
30. He K, Chen X, Xie S, *et al.* Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2022, pp. 16000–9.
31. Yuan X, Lin Z, Kuen J, *et al.* Multimodal contrastive training for visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2021, pp. 6995–7004.
32. Le-Khac PH, Healy G, Smeaton AF. Contrastive representation learning: a framework and review. *IEEE Access* 2020;**8**:193907–34.
33. Chen T, Kornblith S, Norouzi M, *et al.* A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*. New York, NY: ACM, 2020: PMLR, pp. 1597–607.
34. Han W, Cheng Y, Chen J, *et al.* Self-supervised contrastive learning for integrative single cell RNA-seq data analysis. *Brief Bioinform* 2022;**23**(5):bbac377.
35. Andersson A, Larsson L, Stenbeck L, *et al.* Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat Commun* 2021;**12**(1):1–14.

36. Maynard KR, Collado-Torres L, Weber LM, *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 2021;**24**(3):425–36.
37. Moffitt JR, Bambah-Mukku D, Eichhorn SW, *et al.* Molecular, spa,tial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* 2018;**362**(6416):eaau5324.
38. Fraley C, Raftery AE, Murphy TB, *et al.* mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report* 2012;**597**:1.
39. Huang G, Liu Z, Van Der Maaten L, *et al.* Densely connected convolutional networks In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2017, pp. 4700–8.
40. Chen T, Sun Y, Shi Y, *et al.* On sampling strategies for neural network-based collaborative filtering. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM, 2017, pp. 767–76.
41. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**(1):1–5.
42. Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* 2002;**3**: 583–617.
43. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 1971;**66**(336):846–50.
44. Kuhn HW. The Hungarian method for the assignment problem. *Naval Res Logistics Quarterly* 1955;**2**(1–2):83–97.
45. Hao Y, Hao S, Andersen-Nissen E, *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**(13):3573–3587.e29.
46. Palla G, Spitzer H, Klein M, *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat Methods* 2022;**19**(2):171–8.
47. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. 2018.
48. Wolf FA, Hamey FK, Plass M, *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;**20**(1):1–9.
49. Rao J, Zheng S, Lu Y, *et al.* Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns* 2022;**3**(12):100628.