CCST: Cell clustering for spatial transcriptomics data with graph neural network

Jiachen Li
Shanghai Jiao Tong University https://orcid.org/0000-0001-9137-6262
Siheng Chen
Shanghai Jiao Tong University
Xiaoyong Pan
Shanghai Jiao Tong University https://orcid.org/0000-0001-5010-464X
Ye Yuan (yuanye_auto@sjtu.edu.cn)
Shanghai Jiao Tong University
Hong-bin Shen
Shanghai Jiao Tong University

Article

Keywords:

Posted Date: December 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-990495/v1

Abstract

Spatial transcriptomics data can provide high-throughput gene expression profiling and spatial structure of tissues simultaneously. An essential question of its initial analysis is cell clustering. However, most existing studies rely on only gene expression information and cannot utilize spatial information efficiently. Taking advantages of two recent technical development, spatial transcriptomics and graph neural network, we thus introduce CCST, **C**ell **C**lustering for **S**patial **T**ranscriptomics data with graph neural network, an unsupervised cell clustering method based on graph convolutional network to improve *ab initio* cell clustering and discovering of novel sub cell types based on curated cell category annotation. CCST is a general framework for dealing with various kinds of spatially resolved transcriptomics. With application to five *in vitro* and *in vivo* spatial datasets, we show that CCST outperforms other spatial cluster approaches on spatial transcriptomics datasets, and can clearly identify all four cell cycle phases from MERFISH data of cultured cells, and find novel functional sub cell types with different micro-environments from seqFISH+ data of brain, which are all validated experimentally, inspiring novel biological hypotheses about the underlying interactions among cell state, cell type and micro-environment.

Introduction

A number of spatial transcriptomics technologies have been developed to achieve high-throughput gene expression profiling and spatial structure of tissues simultaneously. Most of them are based on fluorescence in situ hybridization (FISH) approaches, such as osmFISH (1), MERFISH (2–5), seqFISH (6, 7), seqFISH+ (5), and STARmap (8), which can quantify RNA transcripts of genes and their locations in the sample. Integrated with image analysis, FISH enables single cell resolution high-throughput gene expression quantification and spatial location recording. FISH methods have been applied to different species and tissues, such as lung (9), brain (1, 6, 8, 10), kidney (11), etc. These studies have provided new biological insights on single cell location, neighborhood and interaction with *in vivo* tissue context. Alternative approaches include RNA-seq based technologies, like spatial transcriptomics (ST) (12), Slide-Seq (13), LCM-Seq (14), and etc. While these methods lead to whole transcriptomics profiling, most cannot provide single cell resolution.

An essential question of singe cell gene expression data is cell state or type identification, which is always one of the key steps in any processing pipeline of the data, including lineage (15), cell cycle (5) and cell-cell interaction analysis (16, 17), etc. Now there have been several clustering approaches developed for single cell RNA-seq data, which are mainly based on clustering of low dimension representation of gene expression of single cells (18–21). Most spatial data studies also rely on such strategies. For the MERFISH dataset of cultured U-2 OS cells (5), graph-based Louvain community detection (22, 23) is applied to top principal components of gene expression of single cells (24, 25). Integration of scRNA-seq is also adopted. For example, In the seqFISH study (26), a multiclass support vector machine (SVM) classifier is trained by cell type information from scRNA-seq data, and then applied to map seqFISH cells to corresponding cell types.

For spatial data, these expression-based methods cannot make full use of spatial location information, which is often coupled with cell identities. *In vitro* cultured cells in the same cell cycle phase are more likely to resident together (5), and certain cell type of *in vivo* tissue is known to be spatially proximal to itself or to specific cell types (27). Spatial structure thus can be used as an informative feature to improve cell clustering. Giotto (28) is a package designed for processing spatial gene expression data as well. Recently, stLearn (29) has been developed. It firstly utilizes the standard Louvain clustering procedure as used in scRNA-seq analysis to get a k-nearest neighbor (kNN) graph. Next the initial cluster is split into sub-clusters if its spots are spatially separated. smfishHmrf (30) is another spatial clustering method that starts by the SVM classifier trained using scRNA-seq data as mentioned above. It then updates cell clustering according to the principle that neighbor cells of the same identity have higher score. BayesSpace (31) is a fully Bayesian statistical clustering method designed for only spatial transcriptomics (ST) data which encourages neighboring spots to belong to the same cluster. SpaGCN (32) utilizes a vanilla graph convolutional network (GCN) to integrate gene expression with spatial location and histology in ST. SEDR (33) uses a deep autoencoder to map the gene latent representation to a low-dimensional space. Most of spatial clustering approaches simply assume that the same cell group is spatially close to each other and do not take into consideration the whole complex global cell interactions across the tissue sample. Much work still needs to be investigated on this promising spatial representation.

Here we develop a cell clustering method, **C**ell **C**lustering for **S**patial **T**ranscriptomics data (CCST), based on graph convolutional networks (GCNs), which can simultaneously joint both gene expression and complex global spatial information of single cells from spatial gene expression data. A few years ago, GCN (34) was introduced to handle non-Euclidean relationship data, maintaining the power of convolutional neural network (CNN) (35, 36). The relationship data is encoded as graph with adjacent matrix representing relationship among variables and node feature matrix representing variable observations. GCN layer is designed to integrate graph (spatial structure in our case) and node feature (gene expression). For the cell clustering of spatial data, we first convert the data as graph, where node represents cell with gene expression profile as attributes and edge represents neighborhood relationship between cells. Next a series of GCN layers is used to transfer graph and gene expression information as cell node embedding vectors, meanwhile the graph is corrupted to generate negative embeddings. By learning the discrimination task, the neural network (NN) model is trained to encode cell embedding from spatial gene expression data, which is used for cell clustering.

CCST is tested on both FISH-based single cell transcriptomics and spot-based ST. CCST is also tested on both *in vitro* and *in vivo* spatial datasets, with tasks of *ab initio* cell clustering and sub cell type discovering based on manually curated cell category annotation. Our experimental results suggest CCST can greatly improve *ab initio* cell clustering upon prior methods in MERFISH dataset (5), by clearly recognizing cell groups of all four cell cycle phases of cultured cells of the same cell type. CCST can also be used to find novel sub cell types and their interactions with biological insights from seqFISH+ datasets of mouse olfactory bulb (OB) and cortex tissues (10). In addition, to show superior to recently developed methods, CCST is evaluated on two ST datasets and achieves better clustering results. All above results

indicate that CCST can provide informative clues for better understanding cell identity, interaction, spatial organization in tissues and organs.

The Ccst Framework

We extended the unsupervised node embedding method Deep Graph Infomax (DGI) (37), and developed CCST to discover novel cell subpopulation from spatial single cell expression data. As shown in **Fig. 1**, with both single cell location and gene expression information as inputs, CCST firstly encodes the spatial data into two matrices. One is hybrid adjacent matrix based on cell neighborhood where a hyperparameter λ (set as 0.8 by default on FISH and 0.2 on ST) is used for intracellular (gene) and extracellular (spatial) information balance (**Methods**), and the other one is gene expression profile matrix of single cell. Both matrices are fed into the DGI network to calculate embedding vector for each cell. DGI employs a series of GCN layers that enables it to integrate both graph (cell location) and node attribute (gene expression) as node (single cell) embedding vectors. The edges in the graph are also permuted to generate negative node embedding vectors that do not have any spatial structure information. By being trained how to discriminate the two embedding types, CCST learns to encode cell node embedding that contains both spatial structural information and gene expression. After dimension reduction by Principal Component Analysis (PCA), k-means++ (38) was used for node clustering to find novel cell groups or cell subpopulations.

Applying CCST to spatial gene expression data

While a number of spatial gene expression data have been created, here we focus on three FISH-based data that both contain thousands of genes with single cell resolution. The first one is MERFISH data (5) from *in vitro* cultured U-2 OS cells that provides 10,050 genes in 1,368 cells in three batches. This data only includes one cell type with different cell cycle phases. As the authors of the MERFISH paper mentioned, they have discovered obvious spatial structures of cell cycle phase within this cell population, so it would be an ideal spatial dataset to test clustered cell groups since cell cycle can be used as ground truth here. The second one is seqFISH+ data (10), consisting of 10,000 genes from 2,050 (913) cells in separated fields of view from mouse OB (cortex). Unlike the MERFISH data of only one cultured cell type, seqFISH+ data include several *in vivo* cell types and so it can be used to explore potential cell subpopulations with complex biological molecular and spatial features. See ***Methods*** for dataset and preprocessing details.

Although CCST is designed to find novel sub cell type and single cell interactions, CCST is also applied to two more ST datasets here to test the generalization ability and extend potential application scope. These two ST datasets are human dorsolateral prefrontal cortex (DLPFC) and 10x Visium spatial transcriptomics data of human breast cancer. The Adjust Rand Index (ARI) and local inverse Simpson's index (LISI) (39) are used for evaluating the performance of CCST and other approaches.

CCST identifies spatial heterogeneity from MERFISH dataset

We assess CCST's ability to cluster cells on the cultured U-2 OS MERFISH dataset, which includes all four cell cycle phases within only one single cell type. CCST is firstly trained with normalized gene expression matrix and hybrid adjacent matrix from spatial structure to generate the embedding vector with size of 256. To further reduce the feature dimension, PCA is performed, and the top 30 principal components are selected for k-means clustering with k of 5, as suggested in the MERFISH paper(5). Given the fact that C2 only has two cells, the following analysis focuses on the other four groups.

Figs. 2a-2c shows the spatial distribution of grouped cells by CCST on all three replicates. Cells of C0 to C4 are points in different colors respectively, and the position is the center of each cell. To make fully use of the dataset, we encode all the three replicates with just one adjacent matrix where cells are only connected within each batch such that the matrix is a block diagonal matrix composed of three sub adjacent matrices (**Methods** and **Fig. S1**). To further investigate the neighborhood spatial structure of cells assigned to different group, the neighbor enrichment ratios for C0, C1, C3 and C4 are shown in **Figs. 2e-h**. For all cells in a certain group, we first collect their neighbor cells according to the initial adjacent matrix, next we count how many of the neighbors are assigned to each group, and calculate their proportions as the neighbor enrichment ratios. The ratios clearly show that cells tend to be spatially neighbored to those in the same group, which is similar to the conclusion in the MERFISH literature. As discussed in the next section, GO term analysis suggests that each cluster corresponds to one cell cycle phase exclusively (C0: M, C1: S, C3: G2, C4: G1). It is also noticed that C0 (M) is spatially proximal to C3 (G2), so is C1 (S) to C3 (G2) and C4 (G1) to C0 (M), which indicates that cells of adjacent phases co-locate with each other as well. This could be explained by the fact that spatially proximal cells may be divided from the same mother cell.

CCST clearly identifies all four cell cycle phases

We next perform differential expression (DE) analysis to verify different biological functions of each clustered cell group. Here Mann-Whitney U Test is used to find highly expressed DE genes in each cell group compared with all other groups. Then Gene Ontology (GO) term enrichment analysis is done using top 200 significantly DE genes with the whole MERFISH gene list as background gene set.

The top 10 significantly enriched GO terms for each cell group sorted by False Discovery Rate (FDR) value are shown in the **Fig. 3**. These results indicate that CCST can clearly identify all four cell cycle phases. The significantly highly expressed genes in C1 are mostly related with GO terms of DNA replication (GO:0006260), DNA-dependent DNA replication (GO:0006261), cell cycle (GO:0007049) and cell cycle DNA replication (GO:0044786). This means that C1 refers to the cells in S phase, the stage when DNA is replicated. The significant DE genes in C3 are mostly related with GO terms of ribosome biogenesis (GO:0042254), ribonucleoprotein complex biogenesis (GO:0022613), rRNA processing (GO:0006364), ncRNA processing (GO:0034470), rRNA metabolic process (GO:0016072), which indicates that cells in C3 are mainly in phase G2 when macromolecules for multiplication and cell growth are produced, preparing for the next M stage. The top GO terms of C0 are mitotic cell cycle process (GO:1903047), mitotic cell cycle (GO:0000278), cell cycle process (GO:0022402), cell cycle (GO:0007049) and mitotic spindle

organization (GO:0007052), which indicates C0 refers to cells in M (Mitosis) phase when cells give birth to new progeny cell. C4 are enriched with GO terms of negative regulation of various processes, including negative regulation of cellular process (GO:0048523), negative regulation of biological process (GO:0048519), negative regulation of signal transduction (GO:0009968), negative regulation of signaling (GO:0023057), etc, indicating the cells are in G1 phase with a rest for preparation of next cell cycle.

Despite that G1 phase is very complicated, including a variety of biological processes (40), the top differential expression genes can further confirm CCST's prediction (**Tab. S5**). MALAT1 is the most differentially highly expressed gene in C4 with p value of 2.74e-40. It has been proved that MALAT1 control the gene expression and cell cycle progression in G1/S phase when cell makes decision to enter either S or G0 (41, 42). The second significant gene is ABI2 with p value of 4.77e-20. ABI2 is also found to play a promotive role in promoting G1-to-S phase transition as well (43). Additionally, ABI2 phosphatase is a negative regulator of ABA signaling (44) that can prevent DNA replication, keeping the cells in the G1 stage (45). All the above analyses suggest that cells in C4 are in G1 phase.

In addition, CDT1 and CDC6 are essential for the initiation of DNA replication and are well-known gene markers for cell cycle stage. It was shown (46, 47) that expression of CDT1 increases from a very low level in G1 and starts to decrease after entering S stage, which is consistent with the mean trend of CDT1 of our predicted cell cycle stage (**Fig. 3e**). The STD trend form CCST's prediction shows that the expression in C4 (G1) varies most, which is supported by the recent study as well (46). We also find similar results for CDC6, which further validate our predictions. As a result, CCST can identify all four cell cycle phases and C1, C3, C0 and C4 belong to phase S, G2, M and G1 respectively.

For comparison, we first evaluate clustering result with only gene expression. We do the same analysis based on the cell grouping performed in the MERFISH paper using PCA and graph-based Louvain clustering (5). We download that top DE genes of all five clustered groups, and carry out GO term enrichment analysis. As can be seen in **Fig. S3**, these enriched GO terms are much less significant and more overlapped, which results in difficulties to distinguish different cell cycle phases.

We then further compare with five recently developed spatial clustering methods, Giotto (28) (**Fig. S3**), stLearn (29) (**Fig. S4**), SEDR (33) (**Fig. S7**), BayesSpace (31) (**Fig. S8**) and SpaGCN (32) (**Fig. S9**), and two single cell expression analyzing approaches without taking spatial information into consideration, which are the methods used in MERFISH study (5) (**Fig. S5**) and Seurat (21) (**Fig. S6**). In Giotto's results, C0 and C2 are mainly related with mitotic cell cycle, while C3 has much less significant GO terms, thus it is hard to interpret from the perspective of cell cycle. In stLearn's results, C0 is related with DNA replication, both C1 and C3 are highly related with mitotic cell cycle, while no GO term is discovered on C2. In SEDR's results, only 2 clusters are associated with GO terms and None of them are about cell cycle. In SpaGCN results, only C1 and C3 are relevant to cell cycles and the Go terms are mixed. The clustering results still cannot be used to discover the full cell cycle phases. In BayesSpace's results, there are four clustered cell groups associated with corresponding cell cycle GO terms, however, the result has less spatial neighborhood structure and the GO terms are less significant, compared to CCST. We also quantify the

comparison of CCST with these methods by calculating the overlap ratio of GO terms associated with each cluster discovered by methods (**Fig. 3g**). As can be seen, CCST has the lowest ratio with a median value of 0, further indicating the outperformance of CCST.

CCST outperforms prior methods on ST datasets

The first ST data we used is the LIBD human dorsolateral prefrontal cortex (DLPFC) including the 10x Genomics Visium spatial transcriptomics and manually annotated layers. There are 12 samples in DLFPC, each of which consists of up to six cortical layers and the white matter. For Seurat, Giotto, stLearn, SpaGCN, BayesSpace and SEDR, recommended default parameters are adopted. To measure the consistency between the clustering labels and reference labels, ARI is employed to compare the performance of different clustering algorithms, as shown in **Fig. 4a**. Spots in the same biological layer in brain should be spatially close to each other while separated between different layers. To quantify such property, local Inverse Simpson's Index (LISI) (39) is introduced (33). LISI is a metric for accessing the local diversity of cells. A lower LISI indicates that clusters are better spatially separated. See the last two sections of supplementary materials for details in the implementation of ARI and LISI. The LISI is shown in **Fig. 4b**, and the annotation and cluster results of each method on slice 151674 of DLPFC is shown in **Fig. 4d**. As can be seen, CCST is the closest to annotated layer segmentation numerically, and its cluster boundary is significantly smoother than other approaches visually.

CCST is also tested on more one ST data, the 10x Visium spatial transcriptomic data of human breast cancer. We utilize the manual annotation provided in SEDR (33). The tissue is segmented into 20 regions and grouped into 4 main morphotypes: Ductal Carcinoma in Situ/Lobular Carcinoma in Situ (DCIS/LCIS), healthy tissue (Healthy), Invasive Ductal Carcinoma (IDC), and tumor surrounding regions with low features of malignancy (Tumor edge). Here we only compare ARI rather than LISI in **Fig. 4c**, because tumor tissues are highly heterogeneous. The annotation and cluster result of each method is shown in **Fig. 4e**. Again, CCST cluster has smoother boundary, while clusters obtained by other methods are more fragmented with spot-level noise.

CCST finds novel sub cell types from seqFISH+ mouse OB dataset

In addition to *ab initio* discovering cell groups, next we show that CCST can also be used to find novel sub cell type and interactions from manually curated cell type annotation based on prior biological knowledge. For this we firstly select the seqFISH+ dataset from mouse OB. We apply CCST to all 11 cell types to discover novel subpopulation within each annotated cell type. With the same hyperparameter settings as for MERFISH dataset, the embedding vector generated by GCN is fed into PCA, and the top 30 principal components are utilized to divide each annotated cell type group into two clusters to discover potential sub cell types.

We first analyze the sub cell type result of interneuron cells in **Fig. 5**. Based on the spatial embedding, the annotated interneuron cells can be clearly divided into two subgroups in the reduced two-dimension UMAP space (**Fig. 5a**). Bar plots of neighbor enrichment ratios for the two subgroups indicate that the

two subset cells have very different micro-environments (**Fig. 5b**). Specifically, cells of C1 tend to be spatially proximal to Mitral/Tufted cells, endothelial and Olfactory ensheathing cells. We then did GO term analysis based on the top 200 differentially highly expressed genes in C0 (**Fig. 5c**) and C1 (**Fig. 5d**). For C0, the top enriched GO terms are relevant to neural functions, like anterograde trans-synaptic signaling (GO:0098916), secretion by cell (GO:0032940), neurotransmitter transport (GO:0006836), export from cell (GO:0140352) and signal release from synapse (GO:0099643). Such GO results indicate that interneuron cells of C0 are functional mature neural cells that can communicate with other neural cells. In contrast, the top GO terms for C1 are not quite related to neural functions, instead, they include regulation of multicellular organismal process (GO:0051239), nervous system process (GO:0050877), regulation of localization (GO:0032879) and regulation of cell migration (GO:0030334). In addition, the most significantly high expressed gene in C1 is NRSN1 (p=1.67e-35), which may be important for neural organelle transport, nerve growth and neurite extension (48). Such results indicate that interneuron cell can be divided into two subgroups: one is functional mature neural cell group, and the other one is group of cells still in development, including localization and migration, which are interacting with its neighbor cells, like Mitral/Tufted or endothelial cells. Interestingly, such sub cell type discovery and its interaction are validated by a recent study that, a subclass of interneuron, GABAergic interneuron migration can be regulated via embryonic forebrain endothelial cells (49), and partial loss of GABA release from endothelial cells can still impair long-distance migration and localization of interneurons during embryogenesis.

Distinct microenvironment settings of two cell subgroups within one annotated cell type are also found for astrocyte, neuroblast, and endothelial cells (**Fig. 5**). The C0 subgroup of astrocyte cells are more spatially proximal to interneuton, neuroblast, etc (**Fig. 6a**), and its DE genes are mainly enriched in GO terms related to visual and learning functions, while C1 subgroup of astrocytes has GO terms related to cell interaction (**Figs. 6b** and **6c**). For neuroblast cells, we find similar pattern to that of interneuron cells. The C1 of neuroblast are more spatially close to Mitral/Tufted or endothelial cells compared to C0 (**Fig. 6d**). The top GO terms of C0 are relevant to neural functions, like regulation of trans-synaptic signaling (GO:0099177) , regulation of synapse structure or activity (GO:0050803) and modulation of chemical synaptic transmission (GO:0050804), etc, indicating that C0 are functional mature neural cell, while C1 is unmatured and related to cell adhesion (GO:0007155) according to its GO terms (**Fig. 6e** and **6f**). Moreover, the most significantly high expressed gene in C1 is EOMES (p= 5.57e-39), which is essential for the central nervous system in vertebrates (50). The sub neuroblast discovery and its interaction are also supported by a very recent study (51). It was shown that there are direct contacts between endothelial cells and neuroblasts, and the authors further argue that endothelial cytonemes only contact part of neuroblasts, allowing a scattered pattern of cell-cycle-arrested neuroblasts between other cells with proliferative capacities. As a result, our clustering results provide more detailed complex communication mechanism that involves interneuron, neuroblast and endothelial cells together, which is validated by recent studies. For endothelial cells, we also find two sub cell types with different micro-environments. Specifically, we find GO terms associated with synaptic signaling and cell-cell signaling for C0 group (**Fig. 6g** and **h**). In addition, we do neighbor enrichment analysis of all the four cell types, and find more complex cell interactions with sub cell type resolution (**Fig. S10**).

For comparison, we perform CCST with $\lambda = 1$ where no spatial structure information is taken into consideration. However, no significant GO term is found for all sub cell type groups, which indicates that spatial information is essential to find novel sub cell types.

CCST finds novel sub cell types from seqFISH+ mouse cortex dataset

We also perform the analysis for another seqFISH+ dataset from 913 cells in mouse cortex which are assigned to 10 cell types. Sub cell types of four annotated cell categories can be found for astrocyte, excitatory neuron cells, endothelial cells and neural Stem cells respectively (**Fig. S11**). For astrocyte cells, cells of C0 tend to be spatially proximal to excitatory neuron, while cells of C1 do not. The GO terms of C0 are relevant to cell-cell interaction, like synaptic signaling (GO:0099536), neurotransmitter transport (GO:0006836), secretion (GO:0046903), export from cell (GO:0140352), signal release (GO:0023061) and signal release from synapse (GO:0099643). Such GO results indicate that C0 are more likely to be active to communicate with other cells, like excitatory neuron. In contrast, much fewer DE genes are found in C1, which are not enough to get any significant GO term. The interaction between excitatory and astrocyte cells has been studied by recent studies. Excitatory neurons release the neurotransmitter glutamate, which is the main excitatory neurotransmitter (52). To maintain the metabolism of glutamate, the *de novo* synthesis of glutamine in astrocytes is essential (53, 54). Our results indicate that only part of astrocyte cells is responsible for glutamatergic neurotransmission in brain. We also perform CCST with $\lambda = 1$ where no spatial structure information is taken into consideration. As is shown in **Fig. S12**, subpopulation can only be found for Astrocyte cells, rather than Excitatory neuron, Endothelial and Neural Stem cells. The less significant GO term results of top DE genes in the subgroups further indicate that spatial information is essential to find novel sub cell types.

Discussions

Cell state or type identification is a key biological question, and the analysis of it is always one key step for high-throughput single cell omics data. The recently developed spatial transcriptomics data can provide both gene expression profiling and spatial location of single cells, opening the door how to identify cells using both molecular information in cell and spatial information out of cell.

It has been shown that spatial location is deeply coupled with biological insights including cell state, type and their interactions in micro-environment. For example, cultured cells of same cell cycle phase would be more likely to resident together because neighbor cells may come from one parent cell, and several cell types are known to interact and co-locate with each other. However, most existing studies of spatial transcriptomics data rely on only the gene expression information, using expression level of scRNA-seq data or tools developed for scRNA-seq, and cannot utilize spatial information efficiently. Several recently developed spatial clustering methods simply assume that the same cell group is spatially close to each other and did not take into consideration the complex global cell group interactions across the tissue sample. To make full use of spatial information and gene expression level, here we introduce CCST, which

uses unsupervised graph convolution network to learn cell embedding representation based on graph extracted from spatial transcriptomics data.

Results of different spatial transcriptomics datasets show CCST's power. CCST can greatly improve *ab initio* cell clustering over existing methods from *in vitro* MERFISH dataset of cultured cell. Firstly, CCST's prediction has more spatial structure. secondly, CCST can clearly recognize cell groups of all four cell cycle phases. Thirdly, it can find the spatial proximity between adjacent phases. In addition, for *in vivo* dataset, CCST can also be used to find novel sub cell types with different biological functions and their interactions with biological insights from seqFISH+ datasets of mouse cortex and OB tissues, which is supported by DE gene analysis, GO term enrichment and other literatures. In the quantitative comparison, CCST obtained the highest ARI and lowest LISI on both ST datasets compared to all prior clustering approaches. So CCST can help understand cell identity, interaction and spatial organization from spatial data.

CCST is implemented in Python. All the source code and spatial data can be downloaded from the supporting website, https://github.com/xiaoyeye/CCST.

Methods

1, Dataset

Recently, with the cutting-edge technology in imaging the transcriptome in situ with high accuracy, multiple high-throughput spatial expression datasets are available for analyzing cells based on both gene expression and spatial distribution. We take experiments on two benchmark datasets. The first one is obtained by multiplexed error-robust fluorescence in situ hybridization (MERFISH) (5). consisting of the expression of 10,050 genes in 1,368 human osteosarcoma cells from 3 batches (replicates). The second is obtained by sequential fluorescence in situ hybridization (seqFISH+) (10). The seqFISH+ dataset from mouse cortex contains the expression of 10,000 genes in 913 cells assigned to 10 cell types, and seqFISH+ dataset from MOB contains the expression of 10,000 genes in 2,050 cells assigned to 11 cell types. Additionally, two ST datasets are utilized in our experiment, which are human dorsolateral prefrontal cortex (DLPFC) and 10x Visium spatial transcriptomics data of human breast cancer. There are 12 samples in DLFPC, each of which consists of up to six cortical layers and the white matter. In the annotation of human breast cancer provided by SEDR (33), the tissue is segmented to 20 areas.

2, Graph construction

A graph can be described by two matrices, an adjacent matrix for representing the graph structure and a feature matrix for representing node attributes. To represent the spatial information among cells, an undirected graph is constructed for each field of view, where cell is represented as node and edge connects pair of cells spatially close to each other. For

this purpose, we firstly calculate the distance over each cell pair d_{ij}, where i and j are the indices of two cells. Utilizing a proper threshold d_{thres}, we can obtain the adjacent matrix $A_0 \in R^{N\times N}$, where N is the number of cells, $a_{ij} = 1$ if $d_{ij} < d_{thres}$ and $a_{ij} = 0$ otherwise. The constructed graph on MERFISH is illustrated in **Fig. 2(d)**. To balance the weight between spatial information and the gene expression of an individual cell, we introduce a hyperparameter λ to generate the hybrid adjacent matrix.

$$A = \lambda \times I + (1-\lambda)A_0 \qquad (1)$$

where $I \in R^{N\times N}$ is an identity matrix. We conduct experiments with various λ to better explore the influence of spatial information.

Similar to (5, 10), a series of preprocessing steps are introduced on the raw gene count data to extract nodes features, including filtering out lowly expressed genes and lowly variable genes, normalizing the counts per cell, and batch correction if necessary.

Here, we mainly follow the preprocessing strategy given by (5). Since the MERFISH dataset is collected from 3 replicates, batch correction is required. After removing the lowly expressed genes whose expression is fewer than 1 count per cell on average, we employ Scanorama (55) to correct batch effect. Then we normalize the corrected expressions following equation 2. Finally, the lowly variable genes whose variance of normalized expression is lower than 1 are dropped.

$$expression_{ij} = \frac{count_{ij}}{\sum_j count_{ij}} \times 10000 \qquad (2)$$

where i represents cell i and j represents gene j.

On the seqFISH+ and two ST datasets, we adopt similar preprocessing steps without the batch correction that is not needed.

3, Node embedding and clustering

With the recent progress in graph convolutional network (GCN), several approaches to learn node representations from graph-structured data have been proposed. Here, we utilize an unsupervised graph embedding method, Deep Graph Infomax (DGI) (37). Different from previous approaches based on random walk, DGI relies on maximizing mutual information between local representations and global summaries of graphs. In GCN, nodes are embedded by repeatedly aggregating the features of neighbor nodes. The extracted local feature contains the information of a subgraph centered on each individual node. To better explore the high-level feature of the whole graph, DGI is designed to learn an encoder by maximizing mutual information over patches. This feature contains not only local features, but also global features.

The input to DGI is the hybrid adjacent matrix $A \in R^{N\times N}$ and a set of node features, $X = \{x_1, x_2, ..., x_N\}$, where N is the number of nodes, $x_i \in R^F$ represents the features of node i and F is the number of node features. In the vanilla DGI and majority applications of GCN, the

adjacent matrix A is assumed to be filled with binary numbers, i.e., $A_{ij} = 1$ if there exists an edge between node i and j in the graph and $A_{ij} = 0$ otherwise. Here we further apply DGI to the weighted Graph constructed with hybrid adjacent matrix.

The objective of DGI is learning an encoder E that maps input feature and adjacent matrix to embedding space: $E(X, A) = H = \{h_1, h_2, ..., h_N\}$, where H represents high-level representations, $h_i \in R^M$ for each node i and M is the number of embedding features. The encoder is composed of four graph convolutional layers for passing massage over neighbored nodes with a Parametric Rectified Linear Unit (PReLU) as the activation function. The $l\text{-}th$ graph convolutional layer is:

$$H^{(l+1)} = GCN^{(l)}\left(H^{(l)}, A\right) = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \quad (3)$$

where H^l and H^{l+1} are the input and output of the $l\text{-}th$ graph convolutional layer, $W^{(l)}$ is the weight matrix used for feature transformation. $\tilde{A}$ is the adjacent matrix after being added by self-loops,

$$\tilde{A} = A + I, I \in R^{N*N} \quad (4)$$

$$\tilde{D}_{ii} = \sum_i \tilde{A}_{ij} \quad (5)$$

The PReLU function is:

$$PReLU(x) = \begin{cases} x, & if\ x \geq 0 \\ ax, & otherwise \end{cases} \quad (6)$$

where a is a learnable parameter.

The global representation s is obtained by mapping from the local representations with a readout function S: $s = S(\mathrm{E}(X, A))$ and $S: R^{N\times M} \rightarrow R^M$. With the local and global features, a discriminator $D: R^M \times R^M \rightarrow R$ is introduced to evaluate how much graph level information is contained by a local patch. The higher $D(h_i, s)$ indicates, the patches are more likely to be contained within the summary. For training the discriminator, we generate negative samples by a corruption function C: $\bar{A} = C(A)$, where the edges in the graph are reconstructed randomly. Then we obtain the local representations $\overline{h_i}$ for negative samples as well. The full objective is:

$$L = \sum_{i=1}^{N} E_{X,A}[\log D(h_i, s)] + E_{X,\bar{A}}\left[\log(1 - D(\overline{h_i}, s))\right] \quad (7)$$

By maximizing the approximate representation of mutual information between h_i and s, DGI outputs the node embedding that contains structural information of the graph.

PCA is performed on the obtained embedding vector for dimension reduction. Clustering algorithms UMAP is employed on top principal components to discover novel cell groups or cell subpopulations.

4, Differential gene expression analysis

To verify different biological functions of each clustered cell subpopulation, we find differentially expressed (DE) genes that are expressed highly in each subpopulation by Mann-Whitney U Test for all cell types. With the top 200 DE genes and the whole gene list in the corresponding dataset as the background, we carried out Gene Ontology (GO) term enrichment analysis for each subpopulation to construct a functional enrichment profile. We also take GO term analysis on the differential expressed genes of the five clusters discovered by (5). The results are shown in supplementary materials. the FDR values of those GO terms obtained in their approaches are not as low as those got by ours. In addition, the significantly related GO terms are mixed up in there 5 clusters.

Declarations

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61725302, 62073219, 61972251).

References

1. Codeluppi S*, et al.* (2018) Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature methods* 15(11):932-935.

2. Moffitt JR & Zhuang X (2016) RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH). *Methods in enzymology* 572:1-49.

3. Moffitt JR*, et al.* (2016) High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences* 113(39):11046-11051.

4. Chen KH, Boettiger AN, Moffitt JR, Wang S, & Zhuang X (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348(6233).

5. Xia C, Fan J, Emanuel G, Hao J, & Zhuang X (2019) Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences* 116(39):19490-19499.

6. Eng C-HL, Shah S, Thomassie J, & Cai L (2017) Profiling the transcriptome with RNA SPOTs. *Nature methods* 14(12):1153-1155.

7. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, & Cai L (2014) Single-cell in situ RNA profiling by sequential hybridization. *Nature methods* 11(4):360-361.

8. Wang X, *et al.* (2018) Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361(6400).

9. Schiller HB, *et al.* (2019) The human lung cell atlas: a high-resolution reference map of the human lung in health and disease. *American journal of respiratory cell and molecular biology* 61(1):31-41.

10. Eng C-HL, *et al.* (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* 568(7751):235-239.

11. Park J, Liu CL, Kim J, & Susztak K (2019) Understanding the kidney one cell at a time. *Kidney international* 96(4):862-870.

12. Ståhl PL, *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353(6294):78-82.

13. Rodriques SG, *et al.* (2019) Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363(6434):1463-1467.

14. Nichterwitz S, *et al.* (2016) Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nature communications* 7(1):1-11.

15. Pal B, *et al.* (2017) Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nature communications* 8(1):1-14.

16. Yuan Y & Bar-Joseph Z (2019) GCNG: Graph convolutional networks for inferring cell-cell interactions. *bioRxiv*.

17. Arnol D, Schapiro D, Bodenmiller B, Saez-Rodriguez J, & Stegle O (2019) Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell reports* 29(1):202-211. e206.

18. Stuart T, *et al.* (2019) Comprehensive integration of single-cell data. *Cell* 177(7):1888-1902. e1821.

19. Abdelaal T, *et al.* (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome biology* 20(1):1-19.

20. Wolf FA, Angerer P, & Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* 19(1):1-5.

21. Hao Y, *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*.

22. Blondel VD, Guillaume J-L, Lambiotte R, & Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008(10):P10008.

23. Traag VA, Waltman L, & Van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* 9(1):1-12.

24. Shekhar K*, et al.* (2016) Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 166(5):1308-1323. e1330.

25. Pandey S, Shekhar K, Regev A, & Schier AF (2018) Comprehensive identification and spatial mapping of habenular neuronal types using single-cell RNA-seq. *Current Biology* 28(7):1052-1065. e1057.

26. Zhu Q, Shah S, Dries R, Cai L, & Yuan G-C (2018) Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nature biotechnology* 36(12):1183-1190.

27. Stoltzfus CR*, et al.* (2020) CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. *Cell reports* 31(3):107523.

28. Dries R*, et al.* (2021) Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology* 22(1):1-31.

29. Pham D*, et al.* (2020) stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv*.

30. Teng H, Yuan Y, & Bar-Joseph Z (2021) Cell Type Assignments for Spatial Transcriptomics Data. *bioRxiv*.

31. Zhao E*, et al.* (2021) Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*:1-10.

32. Hu J*, et al.* (2021) SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*:1-10.

33. Fu H, Hang X, & Chen J (2021) Unsupervised Spatial Embedded Deep Representation of Spatial Transcriptomics. *bioRxiv*.

34. Kipf TN & Welling M (2016) Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

35. Krizhevsky A, Sutskever I, & Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25:1097-1105.

36. LeCun Y, Bottou L, Bengio Y, & Haffner P (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278-2324.

37. Veličković P*, et al.* (2018) Deep graph infomax. *arXiv preprint arXiv:1809.10341*.

38. Arthur D & Vassilvitskii S (2006) k-means++: The advantages of careful seeding. (Stanford).

39. Korsunsky I*, et al.* (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* 16(12):1289-1296.

40. Donjerkovic D & Scott DW (2000) Regulation of the G1 phase of the mammalian cell cycle. *Cell research* 10(1):1-16.

41. Tripathi V*, et al.* (2013) Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. *PLoS genetics* 9(3):e1003368.

42. Wang J*, et al.* (2014) MALAT1 promotes cell proliferation in gastric cancer by recruiting SF2/ASF. *Biomedicine & Pharmacotherapy* 68(5):557-564.

43. Lu H*, et al.* (2020) miR-25 expression is upregulated in pancreatic ductal adenocarcinoma and promotes cell proliferation by targeting ABI2. *Experimental and therapeutic medicine* 19(5):3384-3390.

44. Merlot S, Gosti F, Guerrier D, Vavasseur A, & Giraudat J (2001) The ABI1 and ABI2 protein phosphatases 2C act in a negative feedback regulatory loop of the abscisic acid signalling pathway. *The Plant Journal* 25(3):295-303.

45. Swiatek A, Lenjou M, Van Bockstaele D, Inzé D, & Van Onckelen H (2002) Differential effect of jasmonic acid and abscisic acid on cell cycle progression in tobacco BY-2 cells. *Plant physiology* 128(1):201-211.

46. Mahdessian D*, et al.* (2021) Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature* 590(7847):649-654.

47. Sakaue-Sawano A*, et al.* (2008) Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* 132(3):487-498.

48. Cheng C*, et al.* (2002) Cloning, expression and characterization of a novel human VMP gene. *Molecular biology reports* 29(3):281-286.

49. Li S*, et al.* (2018) Endothelial cell-derived GABA signaling modulates neuronal migration and postnatal behavior. *Cell research* 28(2):221-248.

50. Russ AP*, et al.* (2000) Eomesodermin is required for mouse trophoblast development and mesoderm formation. *Nature* 404(6773):95-99.

51. Taberner L, Bañón A, & Alsina B (2020) Sensory neuroblast quiescence depends on vascular cytoneme contacts and sensory neuronal differentiation requires initiation of blood flow. *Cell Reports* 32(2):107903.

52. Bekkers JM (2011) Pyramidal neurons. *Current biology* 21(24):R975.

53. Parpura V*, et al.* (2017) Glutamate and ATP at the interface between signaling and metabolism in astroglia: examples from pathology. *Neurochemical research* 42(1):19-34.

54. Schousboe A, Scafidi S, Bak LK, Waagepetersen HS, & McKenna MC (2014) Glutamate metabolism in the brain focusing on astrocytes. *Glutamate and ATP at the Interface of Metabolism and Signaling in the Brain*, (Springer), pp 13-30.

55. Hie B, Bryson B, & Berger B (2019) Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature biotechnology* 37(6):685-691.

Figures

Figure 1

The framework of our subpopulation discovering algorithm. Steps in our approach: (a) Graph construction. From distance matrix to adjacent matrix. 1.2 Filter out low expression genes and low variance genes. 1.3 Batch correction (optional). 1.4 Gene expression normalization. 1.5 Constructing hybrid adjacent matrix with various lambda. (b) Deep Graph Infomax (DGI) for node embedding with spatial information and principal component analysis (PCA) for further dimension reduction. (c) Node clustering for discovering novel cell subpopulations. (d) Differential gene (DE) expression analysis with Mann-Whitney U Test and GO analysis for differential gene expression.

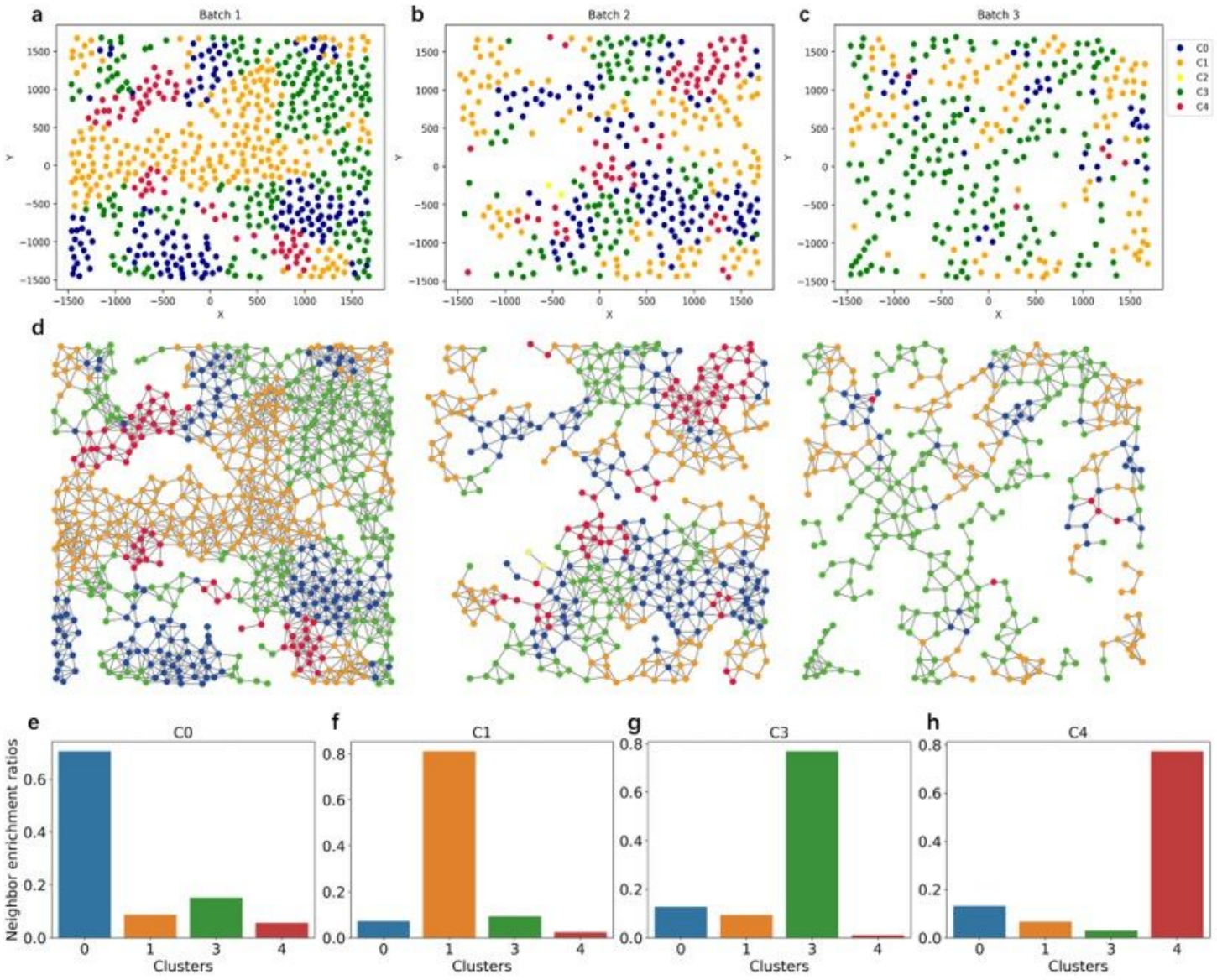

Figure 2

The spatial distribution of cells in different clustered groups on MERFISH dataset. (a-c) the spatial distributions of cells on the three batches (replicates), where cells in C0 to C5 are represented by points in different colors. (d) The constructed graph. Those nodes without neighbors are not shown on the graph. (e-g) The bar plots of neighbor enrichment ratios for C0, C1, C3 and C4 respectively.

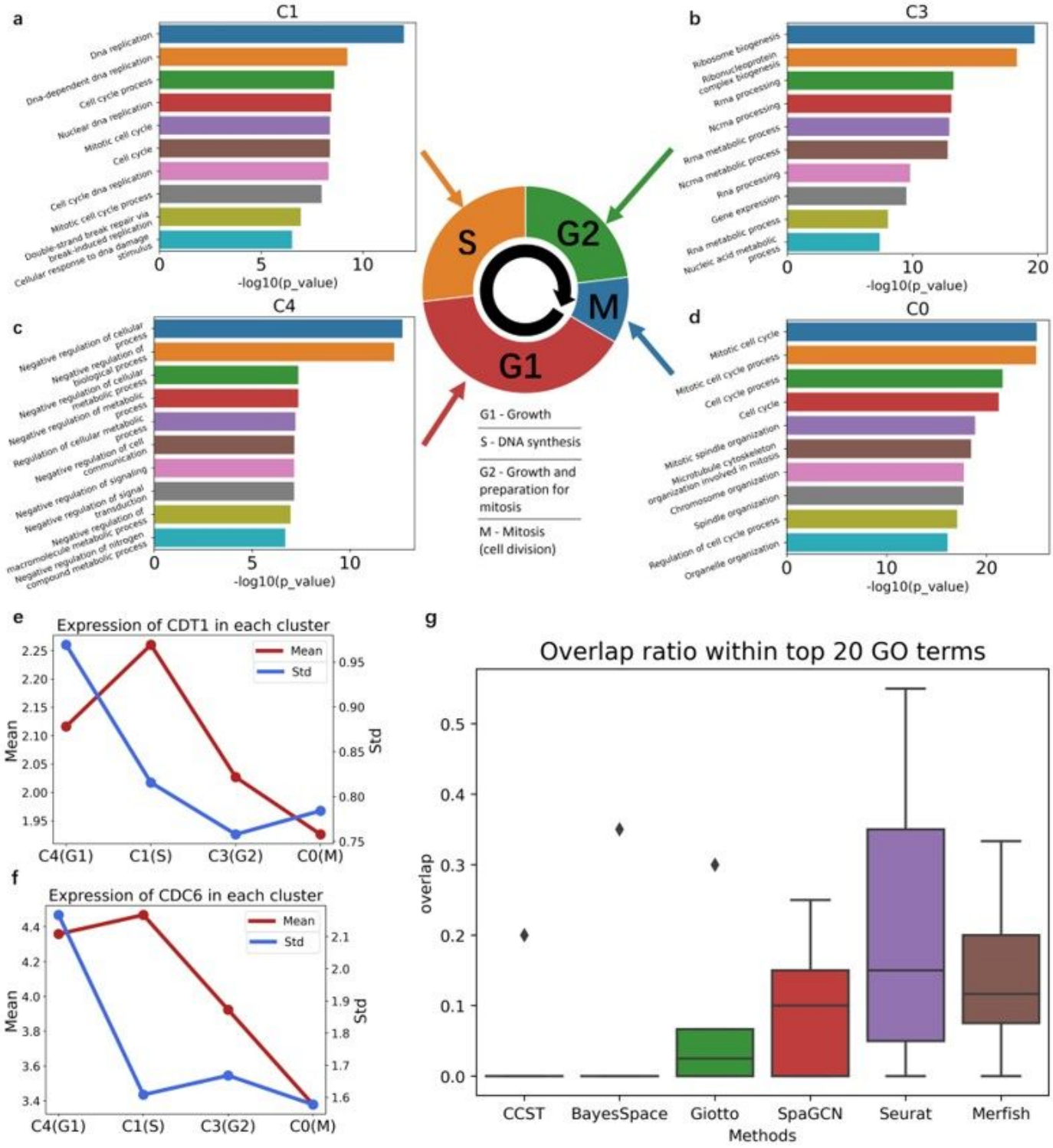

Figure 3

CCST can identify four cell cycle phases clearly. (a-d) Top GO terms of clustered cell groups of C1, C3, C0 and C4, corresponding to S, G2, M and G1 phase respectively. (e, f) The mean and standard deviation (std) of CDT1 and CDC6. (g) A GO result comparison of our CCST with prior methods, including BayesSpace, Gitto, SpaGCN, Seurat and MERFISH pipeline. Stlearn and SEDR are not illustrated in the figure, because only part of clusters given by them are associated with GO terms.

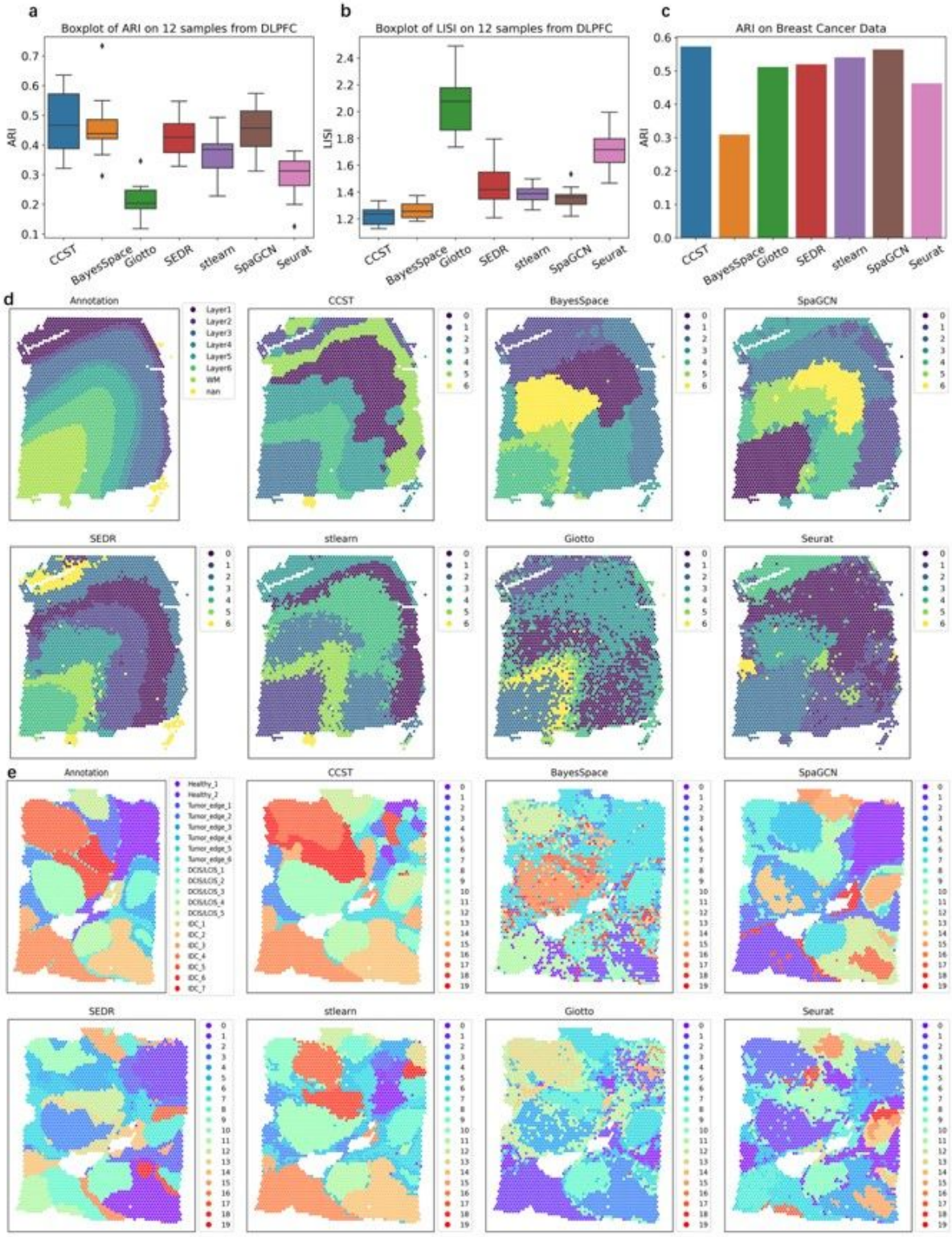

Figure 4

CCST outperforms on two annotated datasets. (a) Barplot of adjusted rand index (ARI) on 12 samples from DLPFC (higher the better). (b) Barplot of local inverse Simpson's index (LISI) on 12 samples from DLPFC (lower the better). (c) Adjusted rand index (ARI) on 10x Visium spatial transcriptomics data of human breast cancer. (d) Annotation and cluster labels obtained by CCST and prior methods, including BayesSpace, SpaGCN, SEDR, stLearn, Giotto and Seurat, on sample 151674 of DLPFC. (e) Annotation

and cluster labels obtained by CCST and prior methods on 10x Visium spatial transcriptomics data of human breast cancer.

Figure 5

Cell subgroup results on interneuron cells of seqFISH+ MOB dataset. (a) The two-dimension UMAP clustering result with Silhouette coefficient. (b) Bar plot of neighbor enrichment ratios for two subgroups. Here, the cells of interneurons, are excluded in the histogram for better demonstration of distribution difference. (c, d) The significant GO terms based on the top 200 differentially highly expressed genes for C0 and C1 respectively.

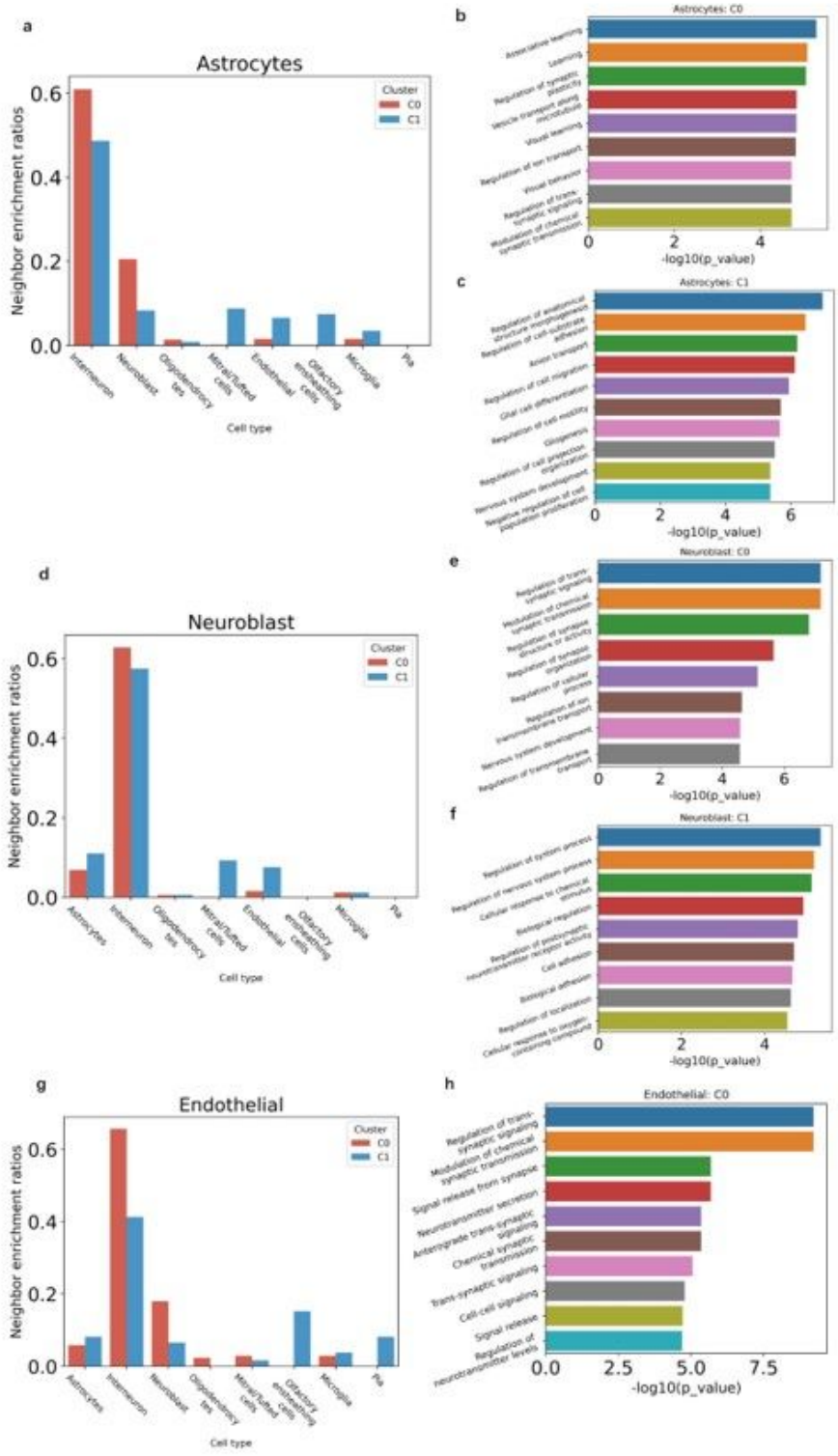

Figure 6

Neighbor enrichment ratios and GO term analysis on each sub cell type of astrocytes, endothelial cells and neural stem cells of seqFISH+ MOB dataset. (a-c) Results on astrocytes. (d-f) Results on endothelial cells. (g, h) Results on neural stem cells where there is no significant GO term for C1.

Supplementary Files

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

- supplementaryCCST.docx
- YYuanCSflat.pdf
- YYuanEPCflat.pdf