1   **Integrating gene expression, spatial location and histology to identify spatial**

2   **domains and spatially variable genes by graph convolutional network**

3

4   Jian Hu[1,*], Xiangjie Li[2], Kyle Coleman[1], Amelia Schroeder[1], David J. Irwin[3], Edward B. Lee[4,5], Russell T.

5   Shinohara[1], Mingyao Li[1,*]

6

7   1. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of

8   Pennsylvania, Philadelphia, PA 19104, USA.

9   2. State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular

10   Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100037, China.

11   3. Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

12   19104, USA.

13   4. Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of

14   Pennsylvania, Philadelphia, PA 19104, USA.

15   5. Translational Neuropathology Research Laboratory, Department of Pathology and Laboratory Medicine,

16   Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

17

18   **Correspondence:**

19   Jian Hu, jianhu@pennmedicine.upenn.edu

20   Mingyao Li, mingyao@pennmedicine.upenn.edu

21

22   **Key words:** spatial transcriptomics; histology; spatial domains; spatially variable genes; graph

23   convolutional network.

Abstract

Recent advances in spatial transcriptomics technologies have enabled comprehensive characterization of gene expression patterns in the context of tissue microenvironment. To elucidate spatial gene expression variation, we present SpaGCN, a graph convolutional network approach that integrates gene expression, spatial location and histology in spatial transcriptomics data analysis. Through graph convolution, SpaGCN aggregates gene expression of each spot from its neighboring spots, which enables the identification of spatial domains with coherent expression and histology. The subsequent domain guided differential expression analysis then detects genes with enriched expression patterns in the identified domains. Analyzing five spatially resolved transcriptomics datasets using SpaGCN, we show it can detect genes with much more enriched spatial expression patterns than existing methods. Furthermore, genes detected by SpaGCN are transferrable and can be utilized to study spatial variation of gene expression in other datasets. SpaGCN is computationally fast, making it a desirable tool for spatial transcriptomics studies.

Introduction

Recent advances in spatial transcriptomics technologies have enabled gene expression profiling with spatial information in tissues[1]. Knowledge of the relative locations of different cells in a tissue is critical for understanding disease pathology because spatial information helps in understanding how the gene expression of a cell is influenced by its surrounding environment and how neighboring regions interact at the gene expression level. Experimental methods to generate spatial transcriptomics data can be broadly classified into two categories: 1) single-molecule fluorescence *in situ* hybridization (smFISH) based techniques, such as MERFISH[2] and seqFISH[3], which measure expression level for hundreds of genes with subcellular spatial resolution in a single cell; and 2) spatial barcoding followed by next generation sequencing based techniques, such as SLIDE-seq[4] and 10X Genomics Visium, which measure the expression level for thousands of genes in captured locations, referred to as spots. These different spatial transcriptomics techniques have made it possible to uncover the complex transcriptional architecture of heterogenous tissues and enhanced our understanding of cellular mechanisms in diseases[5,6].

In spatial transcriptomics studies, an important step is identifying spatial domains defined as regions that are spatially coherent in both gene expression and histology. Identifying spatial domains requires methods that can jointly consider gene expression, spatial location, and histology. Traditional clustering methods such as K-means and Louvain's method[7] can only take gene expression data as input, and the resulting clusters may not be contiguous due to the lack of consideration of spatial information and histology. To account for spatial dependency of gene expression, new methods have been developed. For example, stLearn[8] uses features extracted from histology image as well as expression of neighboring spots to spatially smooth gene expression data before clustering; BayesSpace[9] employs a Bayesian approach for clustering analysis by imposing a prior that gives higher weight to spots that are physically close; Zhu *et*

3

61 *al.*[10] uses a Hidden-Markov random field approach to model spatial dependency of gene expression.

62 Although these methods can cluster spots or cells into distinct groups, they do not provide biological

63 interpretations of the identified spatial domains.

64

65 To link spatial domains with biological functions at the gene expression level, it is crucial to identify genes

66 that show enriched expression in the identified domains. Due to spatial variation of cell types in tissue,

67 the difference of gene expression between different domains is mainly driven by cell type composition

68 variation. On the other hand, information on spatial location and the corresponding histology allows the

69 construction of an anatomy-based taxonomy of the tissue, which provides a useful perspective on cell

70 type composition. Although stLearns integrates gene expression, spatial location, and histology

71 information in clustering, the putative correspondence between cell type difference and organizational

72 structure of the tissue remains unclear. As reported in Saiselet *et al.*[11], many spatial regions are highly

73 intermixed in terms of cell types. Without further downstream gene-level analysis, the spatial domains

74 detected by stLearn still suffer from the lack of interpretability. Recently, new methods such as

75 Trendsceek[12], SpatialDE[13], and SPARK[14] have been developed to detect spatially variable genes (SVGs).

76 These methods examine each gene independently and return a p-value to represent the spatial variability

77 of a gene. However, due to the lack of consideration of tissue taxonomy, genes detected by these methods

78 do not have a guaranteed spatial expression pattern, making it difficult to utilize these genes for further

79 biological investigations.

80

81 Rather than considering spatial domain identification and SVG detection as separate problems, we

82 developed SpaGCN, a graph convolutional network-based approach that considers these two problems

83 jointly. Using a graph convolutional network with an added iterative clustering layer, SpaGCN first

84 identifies spatial domains by integrating gene expression, spatial location, and histology together through

85   the construction of an undirected weighted graph that represents the spatial dependency of the data. For

86   each spatial domain, SpaGCN then detects SVGs that are enriched in the domain against its surrounding

87   regions by differential expression analysis guided by domain information. SpaGCN also has the option to

88   detect meta genes that are uniquely expressed in a given domain. The spatial domains and the

89   corresponding SVGs and meta genes detected for these domains provide a comprehensive picture on the

90   spatial gradients in gene expression in tissue.

91

92   **Results**

93

94   **Overview of SpaGCN and evaluation**

95   SpaGCN is applicable to both sequencing-based and smFISH-based data. As shown in Fig. 1a, SpaGCN first

96   builds a graph to represent the relationship of all samples (spots in sequencing-based or cells in smFISH-

97   based data) considering both spatial location and histology information. Next, SpaGCN utilizes a graph

98   convolutional layer to aggregate gene expression information from neighboring samples. Then, SpaGCN

99   uses the aggregated gene expression matrix to cluster samples using an unsupervised iterative clustering

100   algorithm[15]. Each cluster is considered as a spatial domain from which SpaGCN then detects SVGs that are

101   enriched in a domain by differential expression analysis (Fig. 1b). When a single gene cannot mark

102   expression pattern of a spatial domain, SpaGCN will construct a meta gene, formed by the combination

103   of multiple SVGs, to represent gene expression of the domain. Since the expression profile of a spot/cell

104   is heavily influenced by its local microenvironment, SpaGCN also offers the option of subcluster detection

105   within each spatial domain. SVGs can also be detected to help in understanding the function of each sub-

106   spatial domain.

107

108   To showcase the strength and scalability of SpaGCN, we applied it to five publicly available datasets,

5

109   including four datasets generated by sequencing-based techniques and one dataset generated by

110   MERFISH (Supplementary Table 1). The spatial domains identified by SpaGCN agree better with known

111   tissue layer structure than K-means and Louvain's clustering. We also compared SVGs detected by SpaGCN

112   with those detected by SPARK[14] and SpatialDE[13], and found that the SVGs detected by SpaGCN have more

113   coherent expression patterns and better biological interpretability than the other two methods. The

114   specificity of spatial expression patterns revealed by SpaGCN detected SVGs were further confirmed by

115   Moran's *I* statistic[16], a metric that quantifies the spatial autocorrelation of detected genes.

116

117   **Application to mouse olfactory bulb data**

118   To evaluate the performance of SpaGCN, we first analyzed a mouse olfactory bulb (MOB) dataset[17], which

119   consists of 16,218 genes measured in 262 spots. The main olfactory bulb has five layers, ordered from

120   surface to the center as follows: glomerular layer, external plexiform layer, mitral cell layer, internal

121   plexiform layer, and granule cell layer. We compared SpaGCN's clustering results to K-means and Louvain

122   by setting the number of clusters at 5 for all three methods. As shown in Fig. 2a, K-means only identified

123   3 main spatial domains, with only few spots assigned to domains 1 and 3. Louvain's method identified 5

124   main spatial domains. However, since it does not consider spatial and histology information, domains 2,

125   3, and 4 have blurred boundaries and more outliers than SpaGCN. By contrast, the domains detected by

126   SpaGCN agree better with the biologically known 5-layer structure of the MOB.

127

128   To understand the functions of the SpaGCN identified spatial domains, we next detected SVGs for each

129   spatial domain. In total, SpaGCN detected 60 SVGs. Fig. 2b-f shows a randomly selected SVG for each

130   domain, and all genes show strong specificity for the corresponding domain. The *In Situ* Hybridization

131   labelling of these genes from the Allen Brain Institute further confirmed the correspondence of the spatial

132   domains detected by SpaGCN. Additional SVGs detected by SpaGCN are shown in Supplementary Fig. 1.

133

134 As a comparison, we also detected SVGs using SpatialDE and SPARK. SpatialDE identified 67 SVGs, but only

135 12 of them overlapped with SpaGCN results (Supplementary Fig. 2). We further looked into the 55 genes

136 detected exclusively by SpatialDE and found many of the genes are expressed in only a few spots or are

137 highly expressed in most of the spots, leading to false detections of significant spatial patterns

138 (Supplementary Fig. 3). By contrast, SpaGCN avoided this issue by filtering out genes using minimum

139 within group expression fraction and maximum between group expression fraction. SPARK detected 772

140 genes, with 49 overlapping with SapGCN (Supplementary Fig. 2). However, we found that the SPARK

141 results indicate that 274 genes have FDR-adjusted p-values less than 0.00001 with 14 of them having the

142 smallest identical FDR-adjusted p-value of 4.42e-13. As a result, the SPARK p-values are not informative

143 in differentiating the degree of spatial variability between different genes. Of note, none of these 14 genes

144 were detected by SpaGCN. Further examination revealed that some of these genes show spatial variability,

145 but more than half of them are only expressed in a few spots or highly expressed in most of the spots

146 (Supplementary Fig. 4). The FDR-adjusted p-value distribution of SPARK and q-value distribution of

147 SpatialDE are highly skewed toward 0, making it challenging to select informative SVGs based on their p-

148 values or q-values alone (Supplementary Fig. 5).

149

150 To compare SVGs detected by different methods quantitatively, we calculated the Moran's *I* statistic,

151 which measures the spatial autocorrelation for each gene. Fig. 2g shows the distribution of Moran's *I*.

152 Although all SpaGCN detected SVGs have clear spatial patterns, their Moran's *I* values are not significantly

153 higher than the SVGs detected by SPARK and SpatialDE (median of 0.20 for SpaGCN against 0.18 for SPARK

154 and 0.25 for SpatialDE). Further examination revealed that many SVGs detected by SPARK and SpatialDE

155 are expressed in multiple adjacent spatial domains. For example, the gene *PCP4* uniquely detected by

156 SpatialDE is expressed in two adjacent layers (domains 2 and 4 defined by SpaGCN) (Supplementary Fig.

157   6). By contrast, all the SVGs detected by SpaGCN are domain specific, offering interpretation in alignment

158   with our knowledge of layer structure. We note that less informative SVGs with clear, but non-domain

159   specific, spatial patterns, such as *PCP4,* can also be detected by SpaGCN if the user combines domains 2

160   and 4 as the target domain in SVG detection.

161

162   **Application to mouse posterior brain data**

163   Next, we analyzed a dataset generated from mouse posterior cerebrum, cerebellum and brainstem by

164   10X Genomics that includes 3,353 spots and 31,053 genes[18]. We compared the clustering results of

165   SpaGCN with K-means and Louvain's clustering. The number of clusters in K-means and resolution in

166   Louvain were set to generate the same number of clusters as SpaGCN (10 clusters). Fig. 3a shows that

167   Louvain's clustering is similar to SpaGCN, but the spatial domains detected by SpaGCN are more spatially

168   contiguous than Louvain's results. The integrity of SpaGCN's spatial domains stems from the aggregation

169   of gene expression based on spatial information and histology, which ensures that the genes detected by

170   differential expression analysis have clear spatial expression patterns.

171

172   SpaGCN detected 523 SVGs for the 10 spatial domains while SPARK and SpatialDE detected 9,678 and

173   12,676 SVGs, respectively (Supplementary Fig. 7). We hypothesized that the substantially larger number

174   of SVGs detected by SPARK and SpatialDE are due to the lack of spatial expression patterns that exist in

175   the data. To confirm this hypothesis, we calculated the Moran's *I* statistic for all detected SVGs (Fig. 3b).

176   The Moran's *I* values of SpaGCN detected SVGs are much higher than those detected by SPARK and

177   SpatialDE (median of 0.50 for SpaGCN against 0.21 for SPARK and 0.16 for SpatialDE). Closer examination

178   of the SVGs detected by SPARK and SpatialDE revealed that most of the SVGs suffer from one of the two

179   problems observed previously in the MOB dataset: they are (1) only expressed in a few spots or highly

180   expressed in most of the spots, suggesting high false positive rates for SPARK and SpatialDE or (2) spatially

181   variable, but expressed in multiple adjacent spatial domains, making it difficult to interpret. Another

182   limitation of these two methods is that the FDR-adjusted p-value from SPARK and q-value from SpatialDE

183   are not informative. Genes with similar p-values/q-values do not necessarily show similar spatial pattern

184   and a smaller p-value/q-value does not guarantee a better spatial pattern (Supplementary Fig. 8 and

185   Supplementary Fig. 9). The p-value and q-value distributions of SPARK and SpatialDE are highly skewed

186   toward 0 (Supplementary Fig. 10). By contrast, the SVGs detected by SpaGCN were enriched in specific

187   spatial domains (Supplementary Fig. 11) and their expression patterns are transferable to an adjacent

188   tissue slice in the mouse posterior brain (Supplementary Fig. 12). Further, multiple domain adaptive

189   filtering criteria implemented in SpaGCN allow it to eliminate false positive SVGs and ensure all detected

190   SVGs have clear spatial expression patterns.

191

192   To illustrate why appropriate filtering is important, we use domains 1, 5, and 8 as an example. For each of

193   these domains, SpaGCN detected a single SVG enriched in that region. As shown in Fig. 3c, *PVALB* is

194   enriched in domain 1, and *TRM62* is enriched in domain 8. Although domains 1 and 8 are adjacent to each

195   other, these two SVGs can still well mark these domains. *NRGN* is a SVG that SpaGCN detected for domains

196   5 and 7. The high expression of *NRGN* in domains 5 and 7 also indicate that these two domains are

197   neuroanatomically similar – both consisting of cortex and the pyramidal layer of the hippocampus. Both

198   the cortex and hippocampus are regions that are on the curved surface of the brain. This posterior brain

199   tissue section has the top part of the curved surface in domain 5 and the bottom part of the curved surface

200   in domain 7. Domains 5 and 7, which would be contiguous in a complete 3D reconstruction, are

201   artifactually separated due to the way the section was cut. Therefore, it is not surprising that in addition

202   to *NRGN*, SpaGCN also detected many other SVGs, such as *APP*, *ATP6V1G2*, *CALM2*, *CHN1*, *CLSTN1*,

203   *ARPP21*, *CYP46A1*, *DCLK1*, *LINGO1*, and *MARCKS*, that are highly expressed in both domains 5 and 7

204 (Supplementary Fig. 11). The unique and powerful SVG detection procedure in SpaGCN ensures that genes

205 like these are not missed.

206

207 SpaGCN did not identify any SVGs for domain 0. However, we reason that a meta gene, formed by the

208 combination of multiple genes, may better reveal spatial patterns than any single genes. We used domain

209 0 as an example to show how SpaGCN can create informative meta genes to mark a spatial domain (Fig.

210 3d). First, by lowering the filtering thresholds, SpaGCN identified *KLK6* which is highly expressed in the

211 lower part of domain 0. Using *KLK6* as a starting gene, SpaGCN used a novel approach to find a log-linear

212 combination of gene expression of *KLK6*, *MBP* and *ATP1B1*, which accurately marked the spatial domain

213 0. In this meta gene, *KLK6* and *MBP* are considered as positive markers because they are highly expressed

214 in some spots in domain 0, whereas *ATP1B1* is considered a negative marker as it is mainly expressed in

215 regions other than domain 0. Previous studies have shown that *KLK6* and *MBP* expression is restricted to

216 oligodendrocytes, while *ATP1B1* is mainly expressed in neurons and astrocytes[19]. This resonates the fact

217 that domain 0 represents white matter which is dominated by oligodendrocytes and has few neuronal cell

218 bodies. Therefore, the genes that make up this meta gene have meaningful biological interpretation.

219 Using this meta gene detection procedure, we also detected meta genes for domains 2, 7, 8 and 9, and

220 found that these meta genes are transferrable to an adjacent tissue slice (Supplementary Fig. 13).

221

222 The expression profile and biological function of a spot is heavily influenced by its neighboring spots. The

223 surrounding spots can trigger a response pathway or signal the spot to perform certain tasks. Although

224 the spots in one spatial domain detected by SpaGCN are spatially coherent and have similar gene

225 expression patterns, they may still have different functions since their surrounding spots are different. For

226 instance, spots located near the boundary of a spatial domain may have different functions compared to

227 spots located in the inner part of the domain. To learn more about the effect of different neighborhoods

228   on the spots, we performed sub-domain detection. For example, domain 2 is located in the center of the

229   tissue slice and surrounded by multiple other spatial domains. As a result, the neighboring environment

230   for spots in domain 2 varies. As shown in Fig. 3e, domain 2 was separated into 5 sub-domains which are

231   located either in the center or different boundary regions of domain 2, suggesting that differences in the

232   neighborhoods of spots contribute to within-domain heterogeneity. SVGs detected for each sub-domain

233   can help us understand the gene expression variability of spots within each sub-domain.

234

235   **Application to LIBD human dorsolateral prefrontal cortex data**

236   In addition to the datasets described previously, SpaGCN also showed advantage over competing methods

237   when evaluated on the LIBD human dorsolateral prefrontal cortex (DLPFC) data[20]. The LIBD study

238   sequenced 12 slices from DLPFC that spans six neuronal layers plus white matter. We started from

239   analyzing slice 151673, which includes 3,639 spots and 33,538 genes. As the original publication manually

240   annotated the tissue into 7 layers, for fair comparison, the number of clusters was also set at 7 for SpaGCN,

241   K-means, and Louvain. As shown in Fig. 4a, K-means and Louvain failed to separate the tissue into layers

242   with clear boundary. By contrast, SpaGCN successfully identified layer structures with clear boundaries.

243   The Adjusted Rand Indexes (ARIs) for the SpaGCN, K-means, and Louvain identified domains are 0.42, 0.24,

244   and 0.33, respectively, suggesting that the SpaGCN results better agree with the manually curated layer

245   structure reported in the original study.

246

247   To further validate the identified spatial domains, we then detected SVGs. In total, SpaGCN detected 61

248   SVGs, with 53 of them specific to domain 4, which corresponds to the white matter region (Supplementary

249   Fig. 14). Patterns of SVGs for other domains are not very clear. These results indicate that gene expression

250   profiles of spots from white matter are distinct from spots in the neuronal layers, while gene expression

251   differences among the six neuronal layers are much smaller and more difficult to distinguish using

252   individual marker genes. SVGs detected by SPARK and SpatialDE also suffered from the same problem.

253   SPARK detected 3,187 SVGs with 1,131 of them having FDR-adjusted p-values equal to 0, most of which

254   only marked the white matter region. We also found that the SVGs detected by SPARK lack domain

255   specificity (Supplementary Fig. 15). SpatialDE detected 3,654 SVGs with 806 of them having q-values equal

256   to 0, but these genes do not necessarily show better spatial pattern than genes with larger q-values

257   (Supplementary Fig. 16). Although SPARK and SpatialDE detected much larger numbers of SVGs than

258   SpaGCN (Supplementary Fig. 17), the genes detected by these two methods lack ability to distinguish

259   different degrees of spatial variability in expression as their p-value and q-value distributions are highly

260   skewed toward 0 (Supplementary Fig. 18). Fig. 4b shows that the Moran's *I* values for SpaGCN detected

261   SVGs are significantly higher than those detected by SpatialDE and SPARK (median of 0.39 for SpaGCN

262   against 0.09 for SPARK and 0.08 for SpatialDE). For 3 out of the 6 neuronal layers, SpaGCN detected a

263   single SVG to mark that region (Fig. 4c). For example, *NEFM* is enriched in domain 0 (layer 3) and *PCP4* is

264   enriched in domain 1 (layer 4). Although it is difficult to identify single genes to mark the other neuronal

265   layers, SpaGCN was able to find layer-specific meta genes. As shown in Fig. 4c, the meta gene formed by

266   *KRT19*, *MYL9*, *MBP*, *GFAP*, and *SNAP25* for domain 5 is specific to layer 1. Since layer 1 only has few spots,

267   it is difficult to find a highly enriched gene. However, by adding depleted genes like *MBP* and SNAP25, the

268   expression pattern in this region is strengthened. Furthermore, the SVGs and meta genes detected by

269   SpaGCN are transferrable to slice 151676 obtained from the same study (Supplementary Fig. 19 and

270   Supplementary Fig. 20).

271

272   To show the SVGs and meta genes detected by SpaGCN are useful for downstream analysis, we performed

273   K-means clustering on slice 151676 using SVGs and meta genes detected from slice 151673 by SpaGCN.

274   Specifically, we selected 2 SVGs or meta genes detected by SpaGCN for each spatial domain, resulting in

275   14 features (18 unique genes involved in total) used in K-means clustering. Comparing with manually

276   curated layer assignment reported in the original study, this clustering analysis had an ARI of 0.25 (Fig.

277   4d). We performed similar clustering analysis using SVGs detected by SpatialDE and SPARK. When only

278   using their top 18 SVGs, the ARI is only 0.07 for SpatialDE and 0.05 for SPARK. Even when using the 806

279   most significant SpatialDE detected SVGs, the ARI is only 0.14. When using the 1,114 most significant

280   SPARK detected SVGs, the ARI is 0.15 (Fig. 4e). The ARIs of both SpatialDE and SPARK are much lower than

281   SpaGCN, even though both used many more SVGs than SpaGCN, which further confirmed the lack of

282   spatial expression specificity for genes detected by these methods.

283

284   **Application to human primary pancreatic cancer tissue**

285   We also analyzed a human primary pancreatic cancer tissue dataset[5], which includes 224 spots and 16,448

286   genes across 3 manually annotated sections, to show SpaGCN's ability in detecting tumorous regions. The

287   original study identified and annotated the cancer region on the histology image. However, the cancer

288   region detected by their clustering method based on gene expression information alone did not closely

289   match the pathologist annotated cancer region (Fig. 5a). Since the cancer region in the histology image is

290   darker in color than non-cancer regions, it is informative for clustering. To give histology information

291   higher weight, we increased the scaling parameter s in SpaGCN from 1 to 2 when calculating distance

292   between each spot pair. This step ensured that spots in the same dark region in the histology are more

293   likely to be clustered together. Fig. 5a shows that domain 2 detected by SpaGCN has a better

294   correspondence to the cancer region than clusters reported in the original study. In total, SpaGCN

295   detected 12 SVGs, with 3, 8, and 1 SVGs for domains 0, 1, and 2, respectively (Fig. 5b; Supplementary Fig.

296   21). Furthermore, a meta gene using *KRT17*, *MMP11*, and *SERPINA1* marked the cancer region better than

297   the originally identified SVG *KRT17* (Fig. 5c). *KRT17* functions as a tumor promoter and regulates

298   proliferation in pancreatic cancer[21], and *MMP11* has been found to be a prognostic biomarker for

299   pancreatic cancer[22]. Our identification of *KRT17* and *MMP11* as the two positive genes for the cancer

300   region agree well with pancreatic cancer biology. SPARK and SpatialDE detected 203 and 163 SVGs,

301   respectively (Supplementary Fig. 22). However, the Moran's *I* values for their SVGs are much lower than

302   those detected by SpaGCN, suggesting their lack of spatial expression patterns (Fig. 5d).

303

304   **Application to MERFISH mouse hypothalamus data**

305   Next, we show that SpaGCN can also be applied to smFISH-based data. To this end, we analyzed a MERFISH

306   dataset generated from the preoptic region of hypothalamus in mouse brain[2], which includes 5,665 cells

307   and 161 genes. One important difference between MERFISH and sequencing-based spatial

308   transcriptomics data is that the captured tissue area is much smaller and less genes are measured, making

309   it difficult to detect spatial domains since the cells within such a small area are more similar to each other.

310   Thus, when utilizing these types of data, we suggest increasing the contribution of neighboring cells when

311   calculating the weighted gene expression of each cell. Using this approach, SpaGCN detected spatial

312   domains that agreed well with the annotated hypothalamic nuclei (Fig. 6a), with domain 2 corresponding

313   to ACA, domain 3 corresponding to PS, and domain 7 corresponding to MnPo. By contrast, the domains

314   identified from the Hidden Markov Random Field (HMRF) approach showed little overlap with the

315   hypothalamic region annotation. Using SpaGCN, we further detected 19 SVGs including *DGKK*, *ERMN*, and

316   *SLN* that showed enriched expression patterns for domains 2, 3, and 7 (Fig. 6b; Supplementary Fig. 23).

317

318   **Discussion**

319   Identification of spatial domains and detection of SVGs are important steps in spatial transcriptomics data

320   analysis. In this paper, we presented SpaGCN, a graph convolutional network-based approach that

321   integrates gene expression, spatial location, and histology to model spatial dependency of gene

322   expression for clustering analysis of spatial domains and identification of domain enriched SVGs or meta

323   genes. Through the use of a convolutional layer in an undirected weighted graph, SpaGCN aggregates

14

324   gene expression of each spot from its neighboring spots, which enables the identification of spatial

325   domains with coherent gene expression and histology. The subsequent domain guided differential

326   expression analysis also enables the detection of SVGs or meta genes with enriched expression patterns

327   in the identified domains. SpaGCN has been extensively tested on datasets from different species, regions,

328   and tissues generated using both sequencing- and smFISH-based techniques. The results consistently

329   showed that SpaGCN can identify spatial domains with coherent gene expression and histology and detect

330   SVGs and meta genes that have much clearer spatial expression patterns and biological interpretations

331   than genes detected by SPARK and SpatialDE. Additionally, the SpaGCN detected SVGs and meta genes

332   are transferrable and can be utilized for downstream analyses in independent tissue sections.

333

334   The spatial domain detection step in SpaGCN is flexible. For datasets with clear layer structure in histology

335   image, such as the mouse posterior brain data and human primary pancreatic cancer data, higher weight

336   can be given to histology by increasing the scaling parameter s in SpaGCN when calculating distance

337   between each spot pair, which results in spatial domains that are more similar to the anatomy-based

338   taxonomy in the histology image. Another important scaling parameter in SpaGCN is the characteristic

339   length scale l, which controls the relative contribution from other spots when aggregating gene

340   expression. By varying l, users can get spatial domain separations with different patterns in which a higher

341   l will result in spatial domains with higher contiguity.

342

343   The SVG detection procedure in SpaGCN is also flexible. While we mainly demonstrated SVG detection for

344   a single domain, SpaGCN also allows users to combine multiple domains as one target domain or specify

345   which neighboring domains to be included in DE analysis. Additionally, SpaGCN allows the users to

346   customize SVG filtering criteria based on p-value and three additional metrics, i.e., in-fraction, in/out

15

347   fraction ratio, and fold change, to select SVGs. The resulting genes can be ranked by any of these metrics

348   to select SVGs with desired spatial expression patterns.

349

350   SpaGCN is computationally fast and memory efficient. To showcase the computational advantage of

351   SpaGCN, we recorded its run time and memory usage for the mouse posterior brain data and compared

352   with SPARK and SpatialDE. All analyses were conducted on Mac OS 10.13.6 with single Intel® Core(TM) i5-

353   8259U CPU @2.30GHz and 16GB memory. As shown in Supplementary Fig. 24, SpaGCN completed spatial

354   domain and SVG detection in less than one minute, whereas the computing time is ~13 minutes for

355   SpatialDE and more than 18 hours for SPARK. Furthermore, SpaGCN only required 1.3 GB of memory,

356   whereas SpatialDE and SPARK required more than 3.1 GB and 7.2 GB of memory, respectively. With the

357   increasing popularity of spatial transcriptomics in biomedical research, we expect SpaGCN will be an

358   attractive tool for large-scale spatial transcriptomics data analysis. Results from SpaGCN will enable

359   researchers to accurately identify spatial domains and SVGs in their studies.

360

Acknowledgements

365

Author contributions

367   This study was conceived of and led by M.L.. J.H. designed the model and algorithm. J.H. implemented the

368   SpaGCN software and led the data analysis with input from M.L., X.L., K.C., A.S., D.I., E.L., and R.T.S.. J.H.

369   and M.L. wrote the paper with feedback from all other coauthors.

370

371 **Competing financial interests**

372   The authors declare no competing interests.

373   **Figure legends**

374

375   **Figure 1. Workflow of SpaGCN. a**, SpaGCN starts from integrating gene expression, spatial location and

376   histology information using a graph convolutional network (GCN), then separates spots into different

377   spatial domains using unsupervised iterative clustering. The GCN is based on an undirected weighted

378   graph in which the edge weight between every two spots is determined by Euclidean distance between

379   the two spots, defined by the spatial coordinates (x, y) and the 3-rd dimensional coordinate z, obtained

380   from the RGB values in the histology image. **b**, For each detected spatial domain, SpaGCN identifies SVGs

381   or meta genes by domain guided differential expression analysis.

382

383   **Figure 2. Spatial domains and SVGs detected in the mouse olfactory bulb dataset. a**, Histology image of

384   the tissue section and spatial domains detected by SpaGCN, Louvain's method, and K-means clustering.

385   **b-f**, Spatial expression patterns of SVGs detected by SpaGCN for domains 0 (*LCAT*), 1 (*NR2F2*), 2 (*CACNB3*),

386   3 (*SLC17A7*), and 4 (*NECAB2*), and the corresponding *in situ* hybridization of these SVGs obtained from the

387   Allen Brain Atlas. **g**, Boxplot of Moran's *I* values for SVGs detected by SpaGCN, SPARK, and SpatialDE.

388

389   **Figure 3. Spatial domains and SVGs detected in the mouse brain posterior brain dataset. a**, Histology

390   image of the tissue section and spatial domains detected by SpaGCN, Louvain's method, and K-means

391   clustering. **b**, Boxplot of Moran's *I* values for SVGs detected by SpaGCN, SPARK, and SpatialDE. **c**, Spatial

392   expression patterns of SVGs detected by SpaGCN for domain 1 (*PVALB*), 8 (*TRIM62*), and 5 (*NRGN*). **d**,

393   Spatial expression patterns of genes *KLK6*, *MBP*, *ATP1B1*, which form the specific marker meta gene for

394   domain 0 (*KLK6 + MBP - ATP1B1*). **e**, Clustering results for 5 sub-domains detected by SpaGCN for domain

395   2, and the spatial expression patterns of SVGs for sub-domains 0 (*KCNC3*), 1 (*CAMK2A*), and 2 (*NRSN2*).

396

18

397 **Figure 4. Spatial domains and SVGs detected in the LIBD human dorsolateral prefrontal cortex dataset.**

398 **a**, Manually annotated layer structure for slice 151673 from the original study[20], and spatial domains

399 detected by SpaGCN, Louvain's method, and K-means clustering. **b**, Boxplot of Moran's *I* values for SVGs

400 detected by SpaGCN, SPARK, and SpatialDE for slice 151673. **c**, Spatial expression patterns of SVGs for

401 domain 0 (*NEFM*) and domains 1 (*PCP4*), and a meta gene formed by *KRT19*, *MYL9*, *MBP*, *GFAP*, and

402 *SNAP25* for domain 5 (*KRT19 + MYL9 − MBP + GFAP − SNAP25*). **d**, Manually annotated layer structure for

403 slice 151676 from the original study[20], and K-means clustering results for slice 151676 using 18 genes

404 selected by SpaGCN, SPARK, and SpatialDE. For SpaGCN, we selected the following genes, domain 0 (*NEFL*,

405 *NEFM*), domain 1 (*PCP4, TMSB10 + PCP4 − KRT19*), domain 2 (*CCK + KRT17 − MT-ND1, CPLX2 + KRT17 −*

406 *MT-ND2*), domain 3 (*CAMK2N1, ENC1*), domain 4 (*MBP, FTL*), domain 5 (*KRT19 + MYL9 − MBP + GFAP −*

407 *PLP1, KRT8 + MYL9 − MBP + GFAP − PLP1*), and domain 6 (*GFAP, MBP*), resulting in 18 unique genes in

408 total. For SPARK and SpatialDE, the 18 SVGs with the smallest FDR-adjusted p-value or q-value were

409 randomly selected. **e**, ARIs between manually annotated layers and K-means' clustering using SVGs

410 selected by different methods. For SpaGCN, we only used the selected SVGs and meta genes, with 18

411 genes involved in total while for SPARK and SpatialDE, we used top 18, 100, 200, 500 and all SVGs with

412 the identical smallest FDR-adjusted p-value or q-value.

413

414 **Figure 5. Spatial domains and SVGs detected in the human primary pancreatic cancer tissue dataset. a**,

415 Histology image of the tissue section with manually annotated regions from the original study[5], spatial

416 domains detected by SpaGCN, and clustering results from the original study. **b**, Spatial expression pattern

417 of SVGs detected by SpaGCN for domain 0 (*AEBP1*) and domain 1 (*SERPINA1*). **c**, Spatial expression

418 patterns of genes *KRT17, MMP11, SERPINA1*, which form the specific marker meta gene for domain 2

419 (*KRT17 + MMP11 - SERPINA1*). **d**, Boxplot of Moran's *I* values for SVGs detected by SpaGCN, SPARK, and

420 SpatialDE.

19

421

422   **Figure 6. Spatial domains and SVGs detected in the MERFISH mouse brain hypothalamus dataset. a**,

423   Spatial domains detected by SpaGCN and the HMRF method overlayed with annotated hypothalamic

424   nuclei from the original study[2], and cell type distribution from the original study. **d**, Spatial expression

425   patterns of SVGs detected by SpaGCN for domain 2 (*ERMN*), domain 3 (*DGKK*), and domain 7 (*SLN*).

426 **Methods**

427

428 **Data preprocessing**

429 SpaGCN takes spatial gene expression and histology image data (when available) as input. The spatial gene

430 expression data are stored in an $N \times D$ matrix of unique molecular identifier (UMI) counts with N samples

431 and D genes, along with the (x, y) 2-dimensional spatial coordinates of each sample. In sequencing-based

432 data, each sample represents a spot containing multiple cells, whereas in single-molecule fluorescence *in*

433 *situ* hybridization (smFISH)-based data, each sample represents a single cell. For simplicity, we will use

434 'spot' to refer to a sample, as most of the data analyzed in this paper are sequencing based. Genes

435 expressed in less than three spots are eliminated. The gene expression values in each spot are normalized

436 such that the unique molecular identifier (UMI) count for each gene is divided by the total UMI count

437 across all genes in a given spot, multiplied by 10,000, and then transformed to a natural log scale.

438

439 **Conversion of spatial transcriptomics data into graph-structured data**

440 After preprocessing, SpaGCN converts the gene expression and histology image data into a weighted

441 undirected graph, $G(V, E)$. In this graph, each vertex $v \in V$ represents a spot and every two vertices in V

442 are connected via an edge with a specified weight. We focus our analysis on spatial transcriptomics data

443 with histology information, but the method can be easily adapted to analyze smFISH-based data, for which

444 histology information is not available.

445

446 *Calculation of distance between two vertices*

447 The distance between any two vertices u and v in the graph reflects the relative similarity of the two

448 corresponding spots. This distance is determined by two factors: 1) the physical locations of spots u and

449 v in the tissue slice, and 2) the corresponding histology information of these two spots. Although some

21

450   spots are physically close to each other in the tissue, the histology image may reveal that they belong to

451   different tissue layers. Therefore, SpaGCN considers two spots to be close if and only if 1) the two spots

452   are physically close, and 2) they have similar pixel features as shown in the histology image. To define a

453   distance metric considering both aspects, SpaGCN extends the 2-dimensional space in the tissue slice into

454   a 3-dimensional space that incorporates histology information. For spot v, its physical location in the

455   tissue slice is represented by 2-dimensional coordinates (x_v, y_v). To determine the corresponding pixel in

456   the histology image for spot v, SpaGCN maps spot v to the histology image according to its pixel

457   coordinates (x_{pv}, y_{pv}). Instead of using the color of the pixel at (x_{pv}, y_{pv}), SpaGCN draws a square

458   centered on (x_{pv}, y_{pv}) containing 50×50 pixels and calculates the mean color value for the RGB

459   channels, (r_v, g_v, b_v), of all pixels that fall in the square. This step smooths the color value and ensures

460   that the color is not dominated by a single pixel. To derive a single value to represent the histology image

461   features, SpaGCN uses a weighted sum of the RGB values as follows,

462

463
$$z_v = \frac{r_v \times V_r + g_v \times V_g + b_v \times V_b}{V_r + V_g + V_b},$$

464

465   where $V_r = \text{Variance}(r_v)$, $V_g = \text{Variance}(g_v)$, and $V_b = \text{Variance}(b_v)$ for all $v \in V$. In this

466   transformation, higher weight is given to the channel with larger variance so that this combined value z_v

467   captures an accurate representation of the patterns in the histology image.

468

469   Next, SpaGCN rescales z_v as

470

471
$$z_v^* = \frac{z_v - \mu_z}{\sigma_z} \times \max(\sigma_x, \sigma_y) \times s,$$

472

473   where μ_z is the mean of z_v, $\sigma_x, \sigma_y, \sigma_z$ are the standard deviations of x_v, y_v and z_v, respectively, for $v \in$

474   V, and s is a scaling factor. In our analysis, s is usually set at 1 to make sure that z_v^* has the same scale

475   variance as x_v and y_v, and we set s to a value larger than 1 when the goal is to increase the weight of

476   histology. The coordinates of spot v are set to be (x_v, y_v, z_v^*) in the extended 3-dimensional space. Finally,

477   the Euclidean distance between every two spots u and v is calculated as

478

479    $$d(u,v) = \sqrt{(x_u - x_v)^2 + (y_u - y_v)^2 + (z_u^* - z_v^*)^2}.$$

480

481   *Calculation of weight for each edge and construction of graph*

482   The weight of each edge (u, v) measures the degree of relatedness between spots u and v and is

483   negatively associated with their distance. The graph structure G is stored in an $N \times N$ adjacency matrix

484    $\mathbf{A} = [w(u, v)]$, where the edge weight between spot u and spot v and is defined as

485

486    $$w(u,v) = \exp\left(-\frac{d(u,v)^2}{2l^2}\right).$$

487

488   The hyperparameter l, also known as the characteristic length scale, determines how rapidly the weight

489   decays as a function of distance. A similar function has been employed in SpatialDE[13]. Let $\mathbf{I}$ denote the

490   identity matrix. For spot v, the corresponding row sum of $\mathbf{A} - \mathbf{I}$, denoted by a_v, can be interpreted as the

491   relative contribution of other spots to its gene expression. We choose the value of l such that the average

492   of a_v across all spots is equal to a pre-specified value, e.g. 0.5.

493

494   **Graph convolutional layer**

23

495   SpaGCN reduces the dimension of the preprocessed gene expression matrix using principal component

496   analysis (PCA). The top 50 principal components are used as input, which work well for all datasets

497   analyzed in this paper. Next, utilizing the power of a graph convolutional network, SpaGCN concatenates

498   the gene expression information and edge weights in G to cluster the nodes. Following Kipf and Welling[23],

499   the graph convolutional layer can be written as

500

501    $$f(X, A) = \delta(AXB),$$

502

503   where X is the $N \times 50$ embedding matrix obtained from PCA, B is a 50×50 matrix representing filter

504   parameters of the convolutional layer, and $\delta(\cdot)$ is a non-linear activation function such as ReLU. The graph

505   convolutional layer ensures that a corresponding row of parameters in B will control the aggregation of

506   neighborhood information for each feature in X, thus offering the flexibility of feature specific aggregation

507   of information provided by neighboring spots. The filter parameters in B are shared across all vertices in

508   the graph and are automatically updated during an iterative training progress. Through graph convolution,

509   SpaGCN has aggregated the gene expression information according to the edge weights specified in G.

510   The output of this layer is an aggregated matrix that includes information on gene expression, spatial

511   location, and histology. The graph convolutional layer was implemented based on Kipf and Welling[23],

512   where the backpropagation is operated via a localized first-order approximation of spectral graph

513   convolution.

514

515   **Spatial domain identification by clustering**

516   Next, based on the output from the above graph convolutional layer, SpaGCN employs an unsupervised

517   clustering algorithm to iteratively cluster the spots into different spatial domains[15]. Each cluster identified

518   from this analysis is considered to be a spatial domain, which contains spots that are coherent in gene

519   expression and histology. To initialize cluster centroids, we use Louvain's method[7] on the aggregated

520   output matrix from the graph convolutional layer. If the number of domains in the tissue is known, the

521   resolution parameter in Louvain will be set to generate the same number of spatial domains. Otherwise,

522   we vary the resolution parameter from 0.2 to 1.0 and select the resolution that gives the highest

523   Silhouette score[24].

524

525   To update the cluster assignments iteratively, we define a metric to measure the distance from a spot to

526   a cluster centroid using the Student's t-distribution as a kernel. The distance between the embedded

527   point h_i for spot i and centroid μ_j for cluster j

528

529
$$q_{ij} = \frac{\left(1 + \|h_i - \mu_j\|^2\right)^{-1}}{\sum_{j'=1}^{K}\left(1 + \|h_i - \mu_{j'}\|^2\right)^{-1}},$$

530

531   can be interpreted as the probability of assigning cell i to cluster j.

532

533   Next, we iteratively refine the clusters by defining an auxiliary target distribution P based on q_{ij}

534

535
$$p_{ij} = \frac{q_{ij}^2/\sum_{i=1}^{N} q_{ij}}{\sum_{j'=1}^{K}\left(q_{ij'}^2/\sum_{i=1}^{N} q_{ij'}\right)},$$

536

537   which upweights spots assigned with high confidence, and normalizes the contribution of each centroid

538   to the overall loss function to prevent large clusters from distorting the hidden feature space. Now that

539   we have the soft assignment q_{ij} and the auxiliary distribution p_{ij}, we can define the objective function as

540   a Kullback-Leibler (KL) divergence loss,

541

$$L = KL(P||Q) = \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij} log \frac{p_{ij}}{q_{ij}}.$$

543

544 The network parameters and cluster centroids are simultaneously optimized by minimizing L using

545 stochastic gradient descent with momentum. This unsupervised iterative clustering algorithm has been

546 previously utilized for scRNA-seq analysis and showed superior performance over Louvain's method[25,26].

547

548 **Detection of spatially variable genes**

549 We are interested in detecting spatially variable genes (SVGs) that are enriched in each spatial domain.

550 We note that some genes may be expressed in multiple but disconnected domains. Although they are not

551 uniquely expressed in a particular domain, these genes are still useful for understanding spatial variation

552 of gene expression and can be used to form meta genes that are uniquely expressed in a specific domain.

553 Therefore, rather than doing differential expression (DE) analysis using spots from a target domain versus

554 all other spots, we first select spots to form a neighboring set of the target domain. The goal is to detect

555 genes that are highly expressed in the target domain but are not expressed or are expressed at low levels

556 in the neighboring spots. To determine which spots should be considered as neighbors, we draw a circle

557 with a prespecified radius around each spot in the target domain. All spots from non-target domains that

558 reside in the circle are considered its neighbors. The radius is set such that all spots in the target domain

559 have approximately 8 neighbors on average. Next, neighbors of all spots in the target domain are collected

560 and form a neighboring set. For each non-target domain, if more than 50% (default) of its spots are in the

561 neighboring set, this domain is then selected as a neighboring domain. This criterion is set to avoid the

562 situation when a domain is selected as a neighboring domain, but only a small proportion of its spots are

563 adjacent to the target domain.

564

565 After neighboring domains are determined, SpaGCN then performs DE analysis between spots in the

566 target domain and the neighboring domain(s) using Wilcoxon rank-sum test. Genes with a false discovery

567 rate (FDR) adjusted p-value <0.05 are selected as SVGs. To ensure only genes with enriched expression

568 patterns in the target domain are selected, we further require a gene to meet the following three criteria:

569 1) the percentage of spots expressing the gene in the target domain, i.e., in-fraction, is >80%; 2) for each

570 neighboring domain, the ratio of the percentages of spots expressing the gene in the target domain and

571 the neighboring domain(s), i.e., in/out fraction ratio, is >1; and 3) the expression fold change between the

572 target and neighboring domain(s) is >1.5. If a user is interested in finding SVGs for a particular combination

573 of spatial domains, SpaGCN offers the option to do so.

574

575 **Detection of spatially variable meta genes**

576 The spatial domain-specific DE analysis described above typically detects SVGs with enriched expression

577 for the majority of the domains. For domains in which no such SVGs are detected, we aim to identify a set

578 of genes that, when combined to form a meta gene, shows an enriched expression pattern in the given

579 domain. To identify genes to form a meta gene, we employ a multi-step approach. First, we lower the

580 thresholds for SVG filtering, e.g., change the minimum fold change threshold from 1.5 to 1.2, to identify

581 genes showing weaker enriched expression pattern in the target domain. In the presence of multiple such

582 weaker SVGs, we randomly select one of them as the base gene and denote it as $gene_0$. Second, we aim

583 to aggregate expression from other genes to the base gene to enhance the spatial pattern for the target

584 domain. To achieve this goal, we first calculate the mean expression level of $gene_0$ for spots in the target

585 domain as e_0. Then, all spots from non-target domains with $gene_0$'s expression level higher than e_0 are

586 extracted to form a control group. Next, we perform DE analysis using spots from the target domain

587 against spots in the control group using Wilcoxon rank-sum test. The gene with the smallest FDR-adjusted

588 p-value and higher expression in the target domain is selected as $gene_{0+}$. Similarly, we perform DE

27

589   analysis using spots from the control group against those from the target domain and select a gene with

590   the smallest FDR-adjusted p-value and higher expression in the control group as $gene_{0-}$. The meta gene's

591   expression is calculated as

592

593 $$\log(meta\_gene_1) = \log(gene_0) + \log(gene_{0+}) - \log(gene_{0-}) + C_0,$$

594

595   where C_0 is a constant to make $\log(meta\_gene_1)$ non-negative. The log transformation is used to rescale

596   expression and make the expression levels comparable across different genes. We have found that

597   including negative genes can strengthen spatial expression pattern for domains that do not have enriched

598   positive marker genes. This algorithm can be used iteratively to find additional genes to form an updated

599   meta gene with a clearer spatial pattern for the target domain. For the $(t + 1)^{th}$ iteration, the meta gene

600   expression is calculated as

601

602 $$\log(meta\_gene_{t+1}) = \log(meta\_gene_t) + \log(gene_{t+}) - \log(gene_{t-}) + C_t$$

603

604   In the $(t + 1)^{th}$ iteration, after adding $gene_{t+}$ and subtracting $gene_{t-}$, SpaGCN will select the $(t + 1)^{th}$

605   control group based on $meta\_gene_{t+1}$. The size of the new control group, which is the number of spots

606   not in the target domain but have higher expression of $meta\_gene_{t+1}$ than spots in the target domain,

607   should be smaller than the size of the tth control group, to ensure that $meta\_gene_{t+1}$ has a clearer

608   spatial pattern than $meta\_gene_t$. Also, $meta\_gene_{t+1}$ is expected to have a larger difference of mean

609   expression between the target and control groups than $meta\_gene_t$. Therefore, at each iteration,

610   SpaGCN checks whether both criteria are met, and the search of additional genes will stop otherwise. An

611   illustration of this iterative meta gene search is shown in Supplementary Fig. 25.

612

28

613 **Evaluation of spatially variable genes using Moran's *I* statistic**

614 The Moran's *I* statistic[16] is a measure of spatial autocorrelation, which can be used to measure the degree

615 of spatial variability in gene expression[27]. The Moran's *I* value ranges from −1 to 1, where a value close to

616 1 indicates a clear spatial pattern, a value close to 0 indicates random spatial expression, and a value close

617 to −1 indicates a chess board like pattern. To evaluate the spatial variability of a given gene, we calculate

618 the Moran's *I* using the following formula,

619

620
$$I = \frac{N}{W} \frac{\sum_i \sum_j [w_{ij} (x_i - \bar{x})(x_j - \bar{x})]}{\sum_i (x_i - \bar{x})^2},$$

621

622 where x_i and x_j are gene expression of spots i and j, $\bar{x}$ is the mean expression of the gene, N is the total

623 number of spots, w_{ij} is spatial weight between spots i and j calculated using the 2-dimensional spatial

624 coordinates of the spots, and W is the sum of w_{ij}. For each spot, we select the k nearest neighbors using

625 spatial coordinates. Moran's *I* statistic is robust to the choice of k and is set at 4 in our analysis. We assign

626 $w_{ij} = 1$ if spot j is in the nearest neighbors of spot i, and $w_{ij} = 0$ otherwise.

627

628 **Detection of subclusters within a spatial domain**

629 To better characterize heterogeneity within a spatial domain due to the influence of its neighborhood,

630 SpaGCN can further detect sub-domains within each spatial domain by utilizing information from

631 neighboring spots. SpaGCN draws a circle around each spot with a pre-specified radius, and all spots that

632 reside in the circle are considered as neighbors of this spot. The value of the radius is set to ensure that

633 every spot in the target domain have ten neighbors on average. Next, SpaGCN records the number of

634 neighbors from different spatial domains for each spot and stores this information in a $T \times K$ matrix,

635 where T is the number of spots in the target domain and K is the total number of spatial domains

29

636   detected. The value for the ith row and j^{th} column is the number of neighbors of spot i belonging to

637   domain j. Next, this matrix is fed into a *K*-means classifier to detect sub-clusters. Differential expression

638   analysis as described above can be performed to identify subcluster enriched genes.

639

640   **Data availability**

641   We analyzed multiple spatial transcriptomics datasets. Publicly available data were acquired from the

642   following websites or accession numbers: (1) mouse olfactory bulb

643   (https://drive.google.com/drive/folders/1C4l3lBaYl7uuV2AA2o0WDzO_mkc_b0pv?usp=sharing); (2)

644   mouse posterior brain (https://support.10xgenomics.com/spatial-gene-

645   expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Posterior); (3) LIBD human dorsolateral prefrontal

646   cortex Dorsolateral pre-frontal cortex (http://research.libd.org/spatialLIBD/); (4) human primary

647   pancreatic cancer data (GSE111672); (5) MERFISH mouse hypothalamus data

648   (https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248). Details of the datasets analyzed in

649   this paper were described in **Supplementary Table 1.**

650

651   **Software availability**

652   An   open-source   implementation   of   the   SpaGCN   algorithm   can   be   downloaded   from

653   https://github.com/jianhuupenn/SpaGCN

654

655   **Life sciences reporting summary**

656   Further information on experimental design is available in the Life Sciences Reporting Summary.

657

658   **References**

659   1.   Asp, M., Bergenstrahle, J. & Lundeberg, J. Spatially resolved transcriptomes-next generation

660   tools for tissue exploration. *Bioessays* **42**, e1900221 (2020).

661   2.   Moffitt, J.R.*, et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic

662   preoptic region. *Science* **362**, eaau5324 (2018).

663   3.   Eng, C.L.*, et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*

664   **568**, 235-239 (2019).

665   4.   Rodriques, S.G.*, et al.* Slide-seq: A scalable technology for measuring genome-wide expression at

666   high spatial resolution. *Science* **363**, 1463-1467 (2019).

667   5.   Moncada, R.*, et al.* Integrating microarray-based spatial transcriptomics and single-cell RNA-seq

668   reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* **38**, 333-342

669   (2020).

670   6.   Chen, W.T.*, et al.* Spatial transcriptomics and in situ sequencing to study Alzheimer's disease.

671   *Cell* **182**, 976-991 e919 (2020).

672   7.   Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in

673   large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).

674   8.   Pham, D.*, et al.* stLearn: integrating spatial location, tissue morphology and gene expression to

675   find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv*

676   (2020).

677   9.   Zhao, E.*, et al.* BayesSpace enables the robust characterization of spatial gene expression

678   architecture in tissue sections at increased resolution. *bioRxiv* (2020).

679   10.   Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G.C. Identification of spatially associated

680   subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data.

681   *Nat Biotechnol* **36**, 1183-1190 (2018).

682   11.   Saiselet, M.*, et al.* Transcriptional output, cell types densities and normalization in spatial

683   transcriptomics. *J Mol Cell Biol*, mjaa028 (2020).

684   12.   Edsgard, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell

685   gene expression data. *Nat Methods* **15**, 339-342 (2018).

686   13.   Svensson, V., Teichmann, S.A. & Stegle, O. SpatialDE: identification of spatially variable genes.

687   *Nat Methods* **15**, 343-346 (2018).

688   14.   Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved

689   transcriptomic studies. *Nat Methods* **17**, 193-200 (2020).

690   15.   Xie, J., Girshick, R. & Farhadi, A. Unsupervised deep embedding for clustering analysis.

691   *Proceedings of the 33rd International Conference on Machine Learning* **48**(2016).

692   16.   Li, H., Calder, C.A. & Cressie, N. Beyond Moran's I: testing for spatial dependence based on the

693   spatial autoregressive model. *Geographical Analysis* **39**, 357-375 (2007).

694   17.   Stahl, P.L.*, et al.* Visualization and analysis of gene expression in tissue sections by spatial

695   transcriptomics. *Science* **353**, 78-82 (2016).

696   18.   Dataset. https://support.10xgenomics.com/spatial-gene-

697   expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Posterior. (2020).

698   19.   Zhang, Y.*, et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes

699   Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37-53 (2016).

700   20.   Maynard, K.R.*, et al.* Transcriptome-scale spatial gene expression in the human dorsolateral

701   prefrontal cortex. *bioRxiv* (2020).

702   21.   Li, D.*, et al.* KRT17 Functions as a Tumor Promoter and Regulates Proliferation, Migration and

703   Invasion in Pancreatic Cancer via mTOR/S6k1 Pathway. *Cancer Manag Res* **12**, 2087-2095 (2020).

704   22.   Lee, J., Lee, J. & Kim, J.H. Identification of Matrix Metalloproteinase 11 as a Prognostic

705   Biomarker in Pancreatic Cancer. *Anticancer Res* **39**, 5963-5971 (2019).

706   23.   Kipf, T.N. & Welling, M. Semi-supervised classification with graph convolutional networks.

707   *International Conference on Learning Representations* **arXiv:1609.02907**(2017).
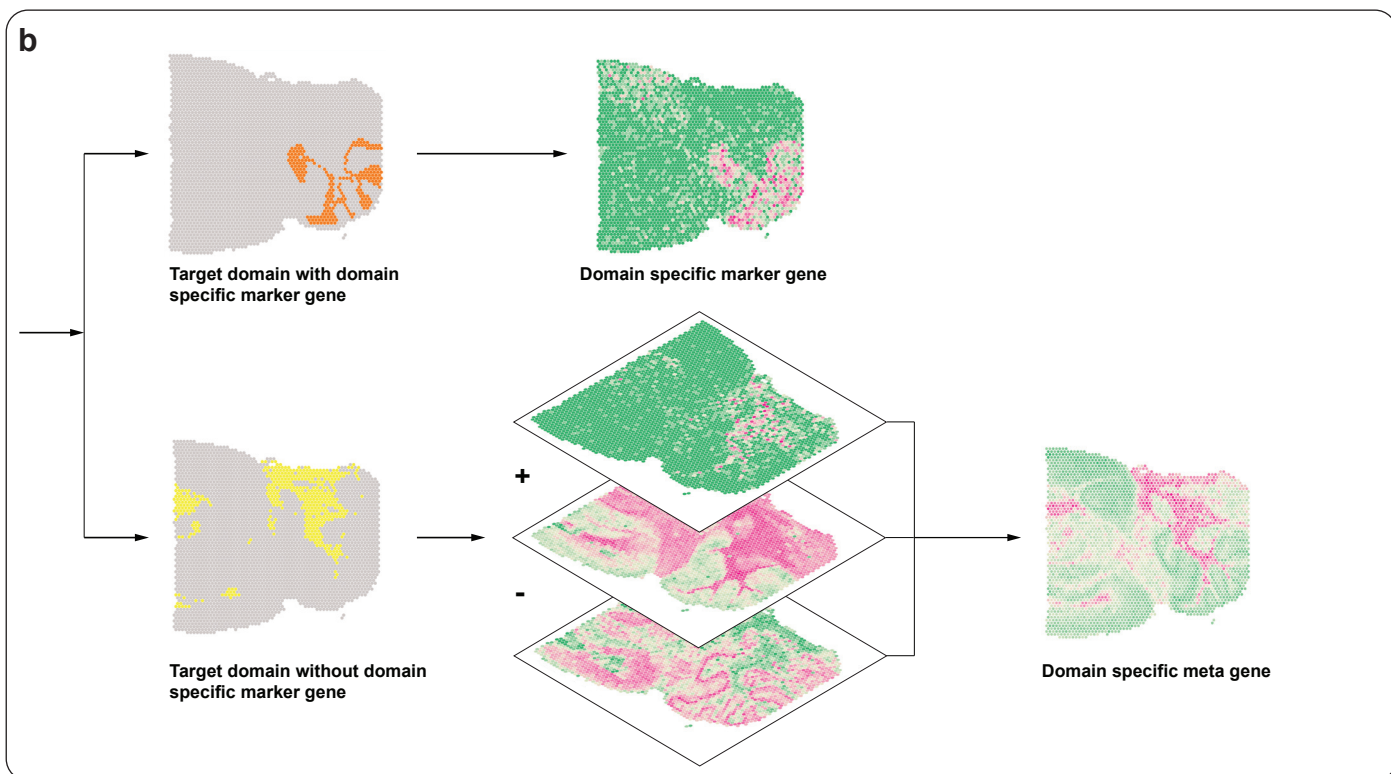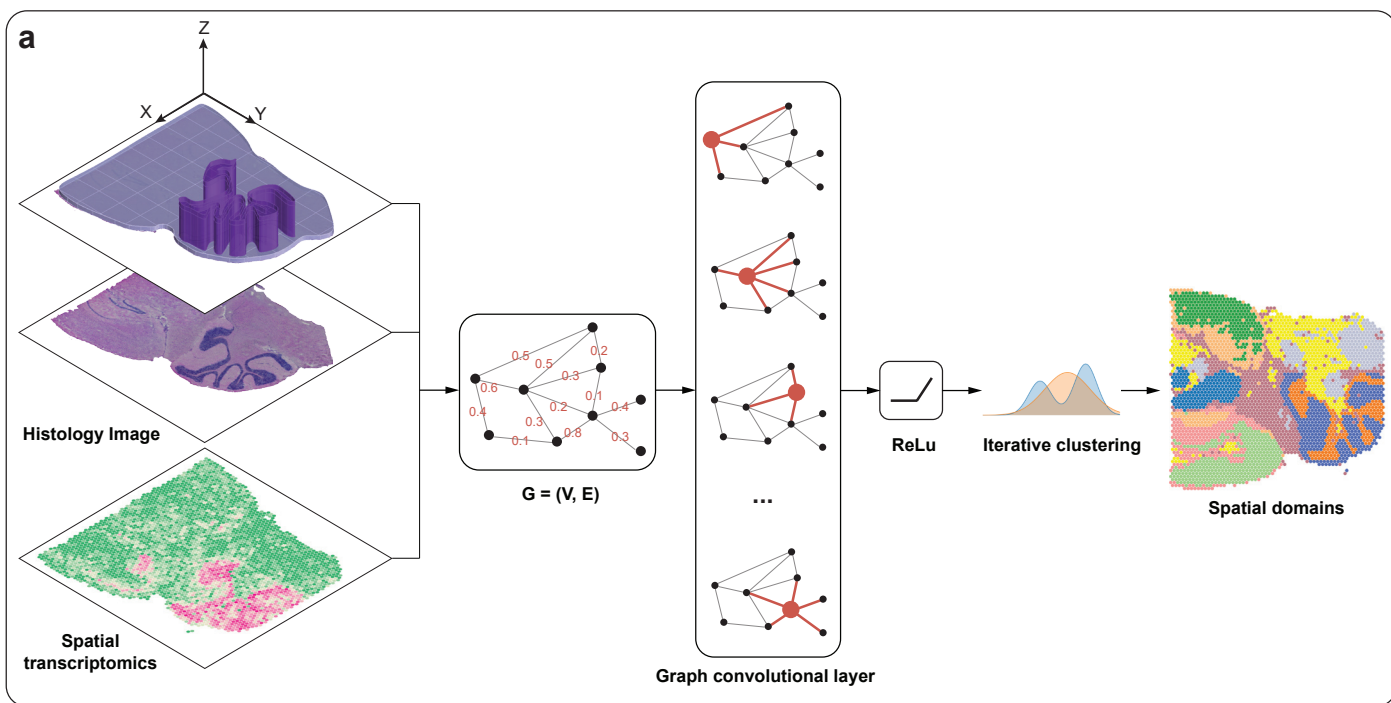
708   24.   Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretaion and validation of cluster

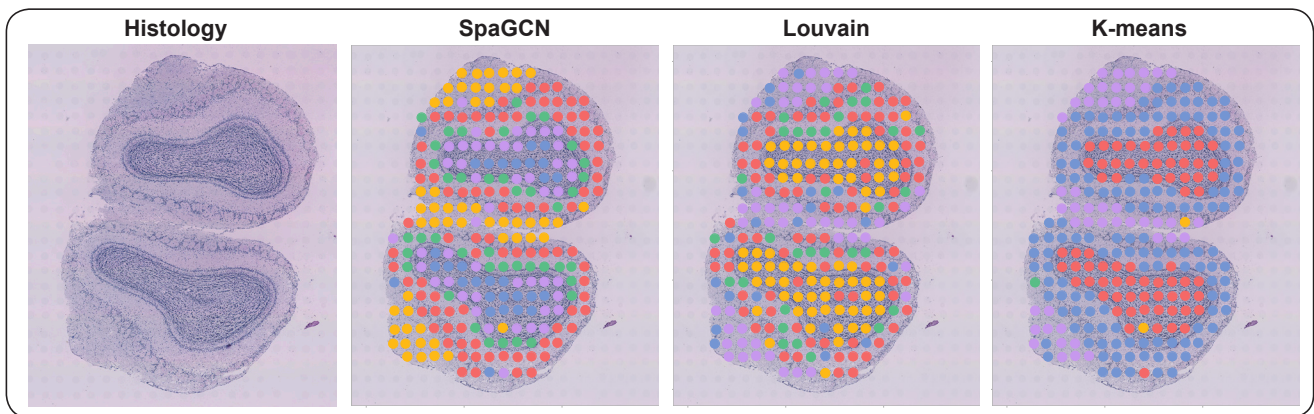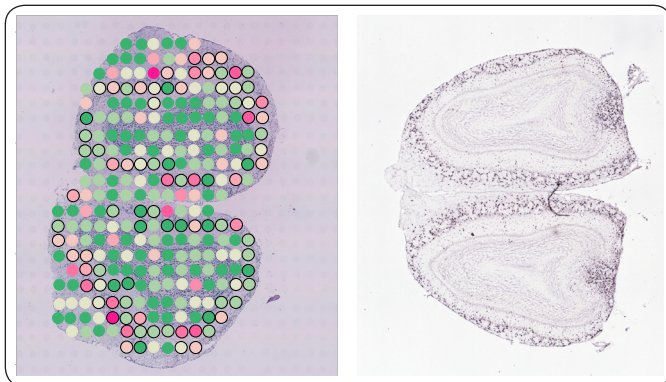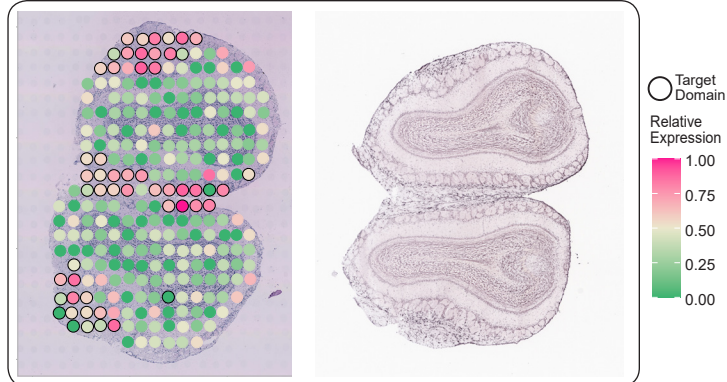709   analysis. *Computational and Applied Mathematics* **20**, 53-65 (1987).
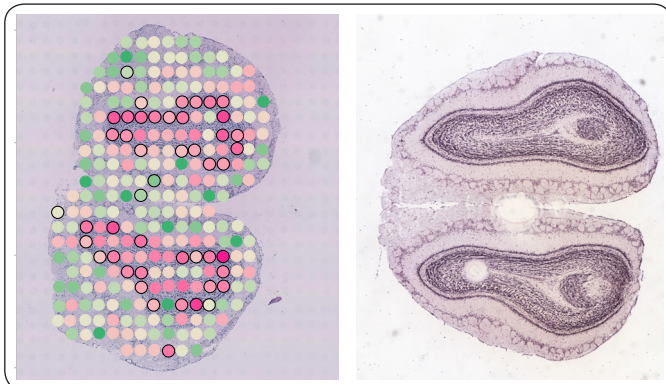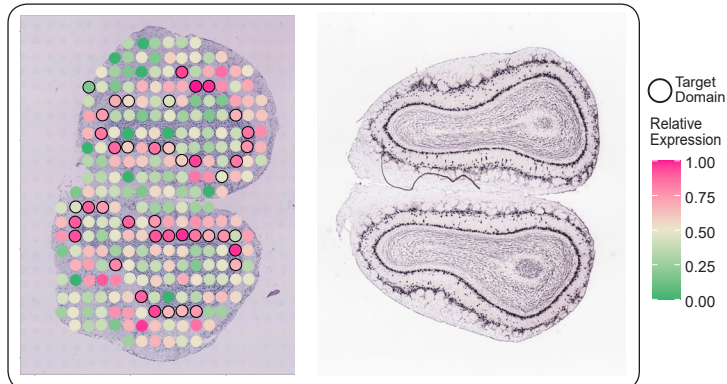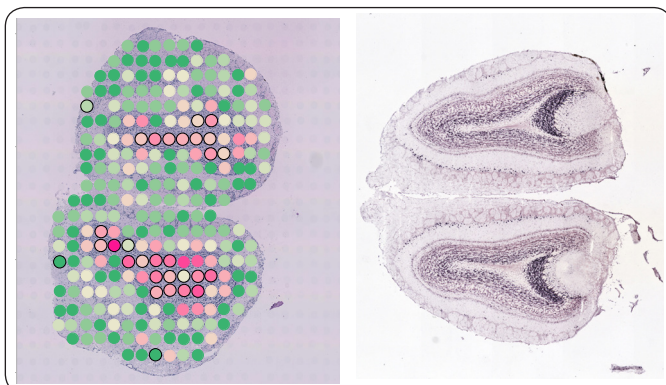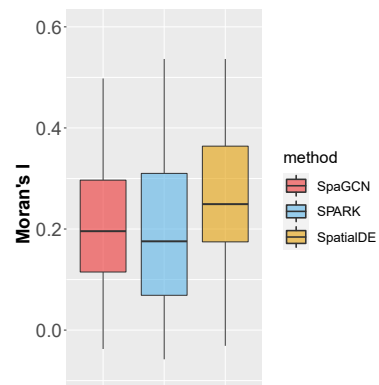
710   25.   Li, X.*, et al.* Deep learning enables accurate clustering with batch effect removal in single-cell

711   RNA-seq analysis. *Nat Commun* **11**, 2338 (2020).

712   26.   Lakkis, J.*, et al.* A joint deep learning model for simultaneous batch effect correction, denoising

713   and clustering in single-cell transcriptomics. *bioRxiv* (2020).

714   27.   Abdelaal, T., Mourragui, S., Mahfouz, A. & Reinders, M.J.T. SpaGE: spatial gene enhancement

715   using scRNA-seq. *Nucleic Acids Res* **48**, e107 (2020).

716

a

Histology Image

Spatial transcriptomics

G = (V, E)

Graph convolutional layer
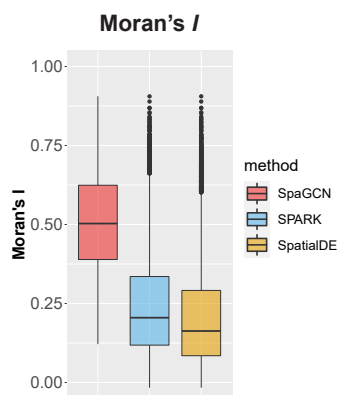
ReLu

Iterative clustering

Spatial domains

b

Target domain with domain specific marker gene

Domain specific marker gene

Target domain without domain specific marker gene

Domain specific meta gene

a

Histology | SpaGCN | Louvain | K-means

Domains
0
1
2
3
4

b Domain 0: *LCAT*

c Domain 1: *NR2F2*

Target Domain
Relative Expression
1.00
0.75
0.50
0.25
0.00

d Domain 2: *CACNB3*

e Domain 3: *SLC17A7*

Target Domain
Relative Expression
1.00
0.75
0.50
0.25
0.00

f Domain 4: *NECAB2*

g Moran's *I*

Moran's I

method
SpaGCN
SPARK
SpatialDE

a

Histology | SpaGCN | Louvain | K-means

Domains
0 1 2 3 4 5 6 7 8 9

b

Moran's *I*

method
SpaGCN
SPARK
SpatialDE

c

Domain 1: *PVALB* | Domain 8: *TRIM62* | Domain 5: *NRGN*

Target Domain

Relative Expression
1.00 0.75 0.50 0.25 0.00

d

KLK6 + *MBP* − *ATP1B1* = Domain 0: *meta gene*

Target Domain

Relative Expression
1.00 0.75 0.50 0.25 0.00

e

Domain 2's subdomains | Subdomain 0: *KCNC3* | Subdomain 1: *CAMK2A* | Subdomain 2: *NRSN2*

Domains
0 1 2 3 4

Target Domain

Relative Expression
1.00 0.75 0.50 0.25 0.00

a

Manual annotation (slice 151673)

Domains
- WM
- Layer1
- Layer2
- Layer3
- Layer4
- Layer5
- Layer6
- na

SpaGCN ARI = 0.419

Louvain ARI = 0.332

K-means ARI = 0.245

Domains
- 0
- 1
- 2
- 3
- 4
- 5
- 6

b

Moran's *I*

Moran's I

method
- SpaGCN
- SPARK
- SpatialDE

c

Domain 0: *NEFM*

Domain 1: *PCP4*

Domain 5: *meta gene*

○ Target Domain

Relative Expression
- 1.00
- 0.75
- 0.50
- 0.25
- 0.00

d

Manual annotation (slice 151676)

Domains
- WM
- Layer1
- Layer2
- Layer3
- Layer4
- Layer5
- Layer6
- na

SpaGCN (18 genes)

SpatialDE (18 genes)

SPARK (18 genes)

Domains
- 0
- 1
- 2
- 3
- 4
- 5
- 6

e

Clustering ARIs

ARI

num_gene
- 18
- 100
- 200
- 500
- 806
- 1131

SpaGCN: 0.25

SpatialDE: 0.07, 0.12, 0.14, 0.15, 0.15

SPARK: 0.05, 0.12, 0.12, 0.14, 0.14

a

Histology | SpaGCN | Original Clustering

Domains
- 0
- 1
- 2

Histological annotation key
- Cancer cells and desmoplasia
- Duct epithelium
- Interstitium

b

Domain 0: *AEBP1* | Domain 1: *SERPINA1*

○ Target Domain

Relative Expression
1.00
0.75
0.50
0.25
0.00

d

Moran's *I*

Moran's I

method
- SpaGCN
- SPARK
- SpatialDE

c

KRT17 + *MMP11* - *SERPINA1* = Domain 2: *meta gene*

○ Target Domain

Relative Expression
1.00
0.75
0.50
0.25
0.00

a

SpaGCN

Illustration of
major
hypothalamic
nucleus

AVPe   PS
BAC   PVA
BNST   StHy
LPO   SHy
MPA   VMPO
MPN   VLPO
MnPO   ACA
PaAP   Fx
Pe   3V

Domains
0
1
2
3
4
5
6
7
8
9
10
11

HMRF

Domains
1
2
3
4
5
6
7
8
9

Cell Type

Cell Type
Ambiguous
Astrocyte
Endothelial 1
Endothelial 2
Endothelial 3
Ependymal
Excitatory
Inhibitory
Microglia
OD Immature 1
OD Immature 2
OD Mature 1
OD Mature 2
OD Mature 3
OD Mature 4
Pericytes

b

Domain 2: *ERMN*

Domain 3: *DGKK*

Domain 7: *SLN*

Target
Domain

Relative
Expression
1.00
0.75
0.50
0.25
0.00