Genome Biology

METHOD **Open Access**

aKNNO: single-cell and spatial transcriptomics clustering with an optimized adaptive k-nearest neighbor graph

Jia Li[1,2], Yu Shyr[1,2*] and Qi Liu[1,2*]

*Correspondence:
yu.shyr@vumc.org;
qi.liu@vumc.org

[1] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, USA
[2] Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN 37203, USA

Abstract
Typical clustering methods for single-cell and spatial transcriptomics struggle to identify rare cell types, while approaches tailored to detect rare cell types gain this ability at the cost of poorer performance for grouping abundant ones. Here, we develop aKNNO to simultaneously identify abundant and rare cell types based on an adaptive *k*-nearest neighbor graph with optimization. Benchmarking on 38 simulated and 20 single-cell and spatial transcriptomics datasets demonstrates that aKNNO identifies both abundant and rare cell types more accurately than general and specialized methods. Using only gene expression aKNNO maps abundant and rare cells more precisely compared to integrative approaches.

Keywords: Single-cell and spatial transcriptomics, Clustering, Rare cells, Adaptive k-nearest neighbors

Background

Single-cell and spatial transcriptomics provide an unprecedented opportunity to navigate cellular landscape in complex tissues [1–4]. To understand cellular heterogeneity, one essential step is to define cell types through unsupervised clustering based on transcriptome similarity [5]. A number of clustering methods have been developed, most of which are generic algorithms adapted for single-cell transcriptomics analysis, such as *k*-means, hierarchical, density-based [6], and community-detection-based clustering [7]. For example, RaceID [8], SIMLR [9], and SC3 [10] refine *k*-means for robust cell clustering. CIDR [11], BackSPIN [12], and pcaReduce [13] extend hierarchical clustering to improve grouping ability on single-cell transcriptomics. Phenograph [14], Seurat [15], and scanpy [16] apply community-detection methods to define cell clusters. Although those typical methods achieved good performance in identifying abundant cell types, they all face the challenge of detecting rare ones.

To address the challenge, several approaches have been specifically designed or tailored to detect rare cell types, such as RaceID [8], GiniClust [17], GiniClust3 [18], FiRE [19], and GapClust [20]. RaceID supplements k-means clustering with outlier detection to identify rare cell types [8]. GiniClust selects genes with high Gini index and then discovers rare cells based on density-based clustering [17]. GiniClust3 extends the method to identify both abundant and rare cell clusters using a cluster-aware, weighted ensemble approach [18]. FiRE uses the Sketching technique to assign a rareness score to each cell [19]. GapClust captures the abrupt local distances change to find rare cell clusters [20]. Those methods either only target rare cells but ignore abundant ones or gain the ability of identifying rare cells at the cost of poorer performance for clustering abundant ones [21].

Here, we propose aKNNO, which builds an optimized adaptive *k*-nearest neighbor graph for community-based detection. Compared to the traditional *k*-nearest neighbor (*k*NN) graph requiring a prespecified and fixed *k* for all cells, aKNNO chooses *k* adaptively for each cell based on its local distance distribution. The adaptive strategy enables the accurate detection of both abundant and rare cell types in a single run. The optimization step gets the most of the adaptive strategy and further improves rare cells identification. Benchmarked on 38 simulated and 17 single-cell transcriptomics datasets, aKNNO outperformed other methods for rare cells identification without scarifying the performance on abundant cells clustering. Applied on three spatial transcriptomics datasets, aKNNO stereotyped fine-grained anatomical patterns using gene expression alone, some of which were even missed by those methods integrating gene expression, spatial locations, and histology image.

Results

Overview of aKNNO

The community-detection methods have become increasingly popular, particularly for analyzing large single-cell transcriptomics datasets [21]. They first construct a *k*NN graph by connecting each cell to its nearest *k* cells measured by transcriptome similarity and then group cells with dense connection. The choice of *k* has great impact on the clustering performance. A large *k* may generate phony connections between rare cells and cells in other clusters, while a small *k* may lead to overclustering of abundant cell types due to dominance of local variances. Cell populations in single-cell data are generally highly imbalanced, including both abundant and rare cells. Therefore, the traditional *k*NN graph using a universal *k* for all cells is unable to capture the inherent cellular structure accurately.

Instead of using a single *k* value for all cells, aKNNO chooses *k* adaptively for each cell based on its local distance. It automatically assigns a small *k* for rare cells to remove spurious long-range connections and connect only true nearest neighbors and a large *k* for abundant cells to balance local and global variances. We used a toy example to illustrate how *k* is chosen adaptively (Fig. 1). In the example, K_{max} is set to 10, meaning that the choice of *k* ranges from 1 to 10. For each cell, aKNNO finds its 10-nearest neighbors and sorts the distance in an ascending order ($d_1 < d_2 < \ldots < d_{10}$). $k = K_{max} = 10$ if $d_{10} < d_{\text{cutoff}}$; otherwise, *k* is chosen if $d_k < d_{\text{cutoff}}$ and $d_{k+1} \geq d_{\text{cutoff}}$. d_{cutoff} is determined by 10-nearest distances of the cell and tuned by a hyperparameter δ

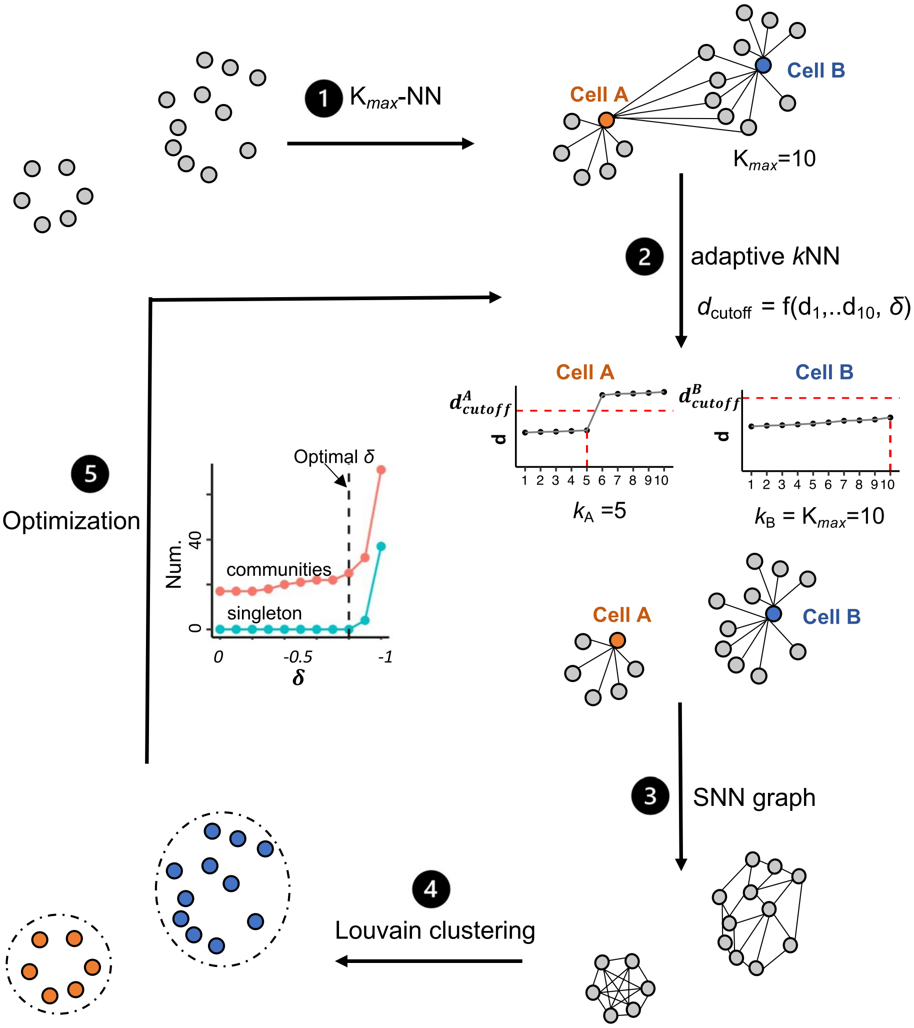


Fig. 1 Overview of aKNNO. aKNNO includes five steps: (1) calculating the K_{max}-nearest neighbors for all cells, (2) choosing k adaptively for each cell based on its local distance distribution, (3) building the shared nearest neighbor graph, (4) clustering based on Louvain community-detection, (5) optimizing the hyperparameter δ by grid search

($d_{\text{cutoff}}=f(d_1,..d_{10},\delta)$). As an example, cell A comes from a rare cluster containing only six cells; thus, its 10-nearest distances increase dramatically from d_5 to d_6, leading to its $d_5 < d^A_{cutoff}$ and $d_6 > d^A_{cutoff}$. Therefore, k of 5 is chosen for the cell A ($k_A = 5$). As another example, cell B belongs to an abundant cell type, where its 10-nearest distances have a slow increase from d_1 to d_{10}, resulting in its $d_{10} < d^B_{cutoff}$. In this case, k of 10 is selected for the cell B ($k_B = 10$). In this way, aKNNO assigns the adaptive and optimal neighbors for each cell. To improve the robustness, the adaptive nearest-neighbor graph is reweighted based on the shared nearest neighbors of pairs of cells (SNN). Finally, Louvain community detection method is applied on the shared nearest neighbor graph to identify clusters (Fig. 1). The hyperparameter δ controls the sensitivity of aKNNO to the local distance change. aKNNO performs a grid search to find the optimal δ that balances the sensitivity and specificity of rare cluster identification (Fig. 1).

Performance of aKNNO in simulated datasets

We first compared the performance of aKNNO to the traditional *k*NN-based method in the Seurat [15] (denoted as KNN) using simulated datasets with ground truth cell-type identity. To make a fair comparison, aKNNO and KNN processed the data exactly the same way except that aKNNO clustered cells based on an adaptive *k*-nearest neighbor graph while KNN grouped cells from a *k*NN graph with a fix *k* for all cells (default $k=20$, details in the "Methods" section). We generated simulated datasets using a public single-cell RNA-seq dataset with 2700 peripheral blood mononuclear cells (PBMC3k) and nine cell types. We simulated two settings, each of which contained two abundant and one rare cell types. In one setting, the two abundant cell types were B cells ($n=200$) and naïve CD4 T ($n=200$), while the rare cell type was nature killer (NK) cell. In the other setting, the two abundant cell types were the same, while the rare cell type was CD14+monocyte. The first setting was more challenging than the second since its rare cells (NK) were similar to one of abundant cell types (naïve CD4 T), indicating that those rare cells were more likely to be hidden by abundant ones. We simulated 19 scenarios for each setting with the number of rare cells ranging from 2 to 20, and we generated 50 datasets for each scenario by random sampling cells from the PBMC3k.

We evaluated the performance by measuring both the accuracy of identifying rare cells (the percentage of rare cells correctly recognized) and the Adjusted Rand Index (ARI) against the true cell clusters. The clustering results were obtained at four different resolutions ($r=0.1$, 0.3, 0.5, and 0.8). aKNNO demonstrated nearly perfect detection of rare cells (accuracy = 1) when the number of rare cells was greater than two in both settings. It achieved an accuracy exceeding 0.9 even when there were only two rare cells (Fig. 2a and b). Furthermore, aKNNO accurately grouped abundant cells and achieved nearly perfect agreement with the ground truth (average ARI > 0.995 in each scenario) (Additional file 1: Fig. S1). In comparison, KNN failed to detect rare cells entirely at every resolution when the number of rare cells was less than six in the first challenging setting (Fig. 2a). For example, with five rare NK cells (as illustrated in Fig. 2c), even at the lowest resolution of 0.1, aKNNO identified three clusters with 100% accuracy. KNN, however, grouped the five NK cells with naïve CD4 T cells, even at the highest resolution (KNN_high in Fig. 2c). As the number of rare cells increased, KNN showed improved performance in identifying them, yet its performance was sensitive to resolution (Additional file 1: Fig. S2). KNN exhibited higher accuracy with increasing resolution, whereas the accuracy of aKNNO remained stable across various resolutions (Additional file 1: Fig. S2). It is important to note that the higher accuracy of KNN at higher resolutions were achieved at the expense of overclustering of abundant cell types (an example was given in the Additional file 1: Fig. S3). In the second setting, KNN failed to detect rare cells when their number was less than five (Fig. 2b). Similarly, its performance was sensitive to resolution (Additional file 1: Fig. S2), and the improved performance at higher resolutions came at the cost of overclustering of abundant cell types. In summary, aKNNO demonstrates accurate identification of both abundant and rare cell types. Subcluster structures identified by aKNNO were supported by high Phiclust scores, derived from random matrix theory, which serves as a measure for identifying non-random substructure within cell clusters [22]. The high Phiclust

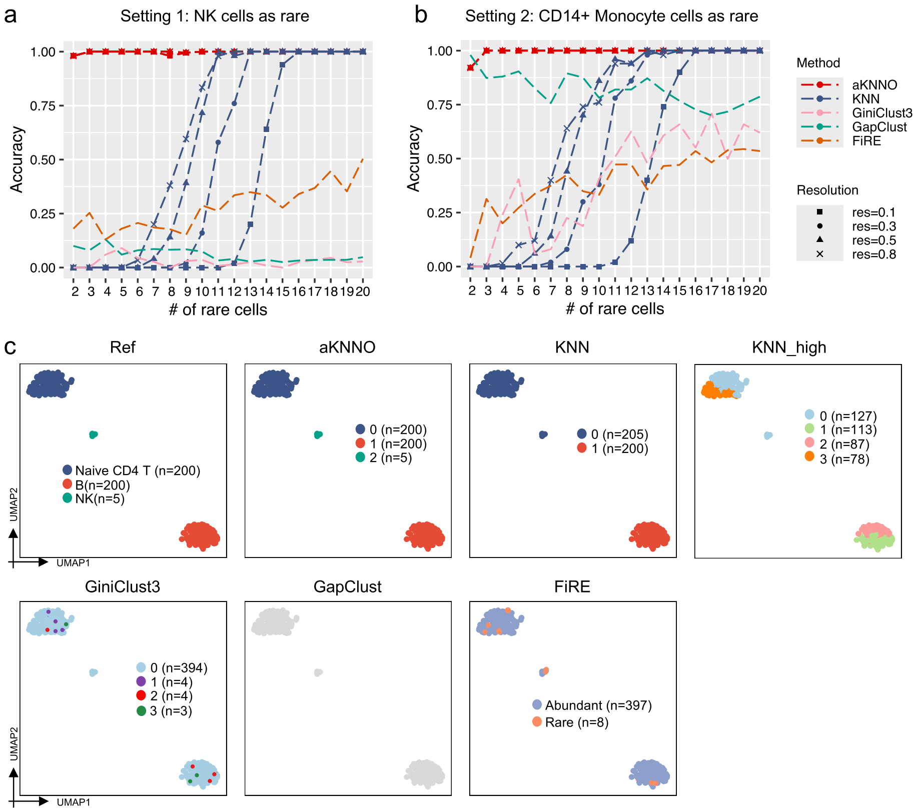


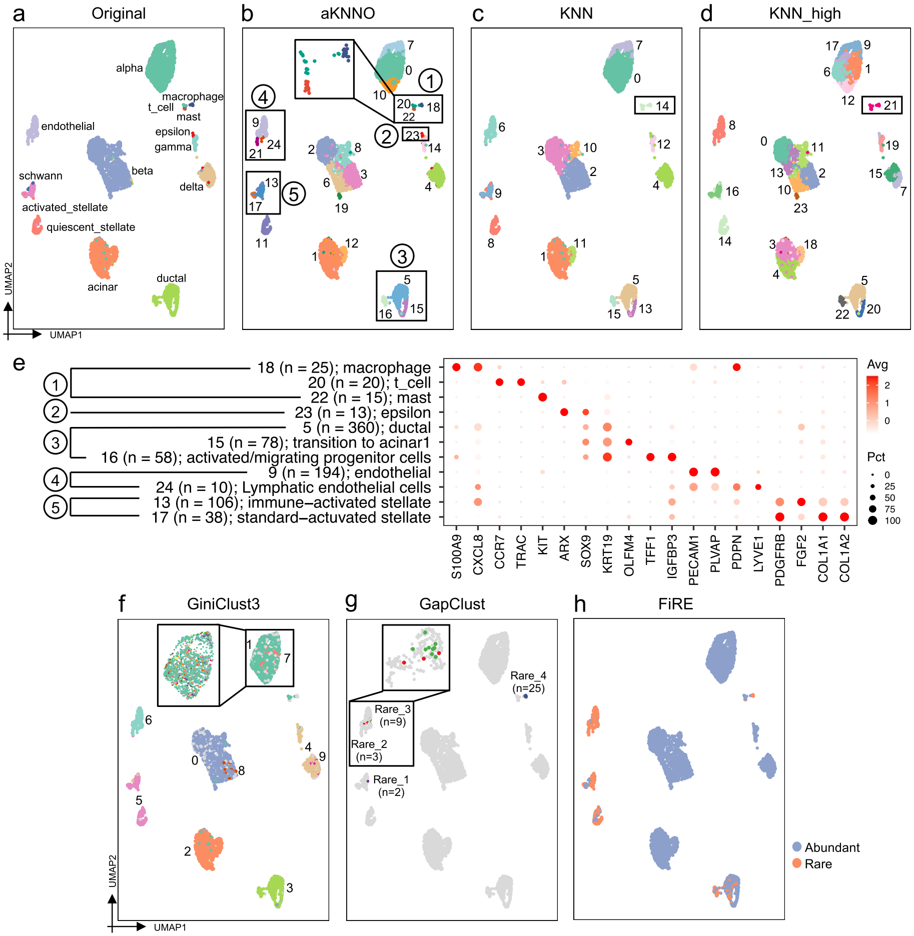
Fig. 2 Application to simulated datasets generated from the PBMC3k dataset. Accuracy of aKNNO and KNN at different resolutions, GiniClust3, GapClust, and FiRE with the number of rare cells ranging from 2 to 20 in the first setting (**a**) and in the second setting (**b**). **c** The UMAP plots labeled by the ground truth (Ref), labeled by aKNNO clustering at a resolution of 0.1 (aKNNO), labeled by KNN clustering at a resolution of 0.1 (KNN), labeled by KNN clustering at a high resolution of 0.8 (KNN_high), labeled by GiniClust3, GapClust, and FiRE

scores demonstrated that aKNNO detected non-random substructures rather than statistical noise (Additional file 1: Fig. S4), that is, aKNNO identifies rare cells without overclustering, whereas KNN either fails to detect rare cells or does so at the expense of overclustering, especially when few rare cells are present.

We further compared aKNNO to three latest methods specifically designed or tailored to identify rare cells, GiniClust3 [18], FiRE [19], and GapClust [20]. GiniClust3 is an extension of GiniClust to identify both abundant and rare cell types, GapClust detects rare clusters, and FiRE quantifies rareness of each cell without clustering. All three methods exhibited poorer performance in identifying rare cells compared to aKNNO in both settings (Fig. 2a and b). In the first setting, FiRE had accuracy < = 0.5, while GapClust and GiniClust achieved accuracy < = 0.1 (Fig. 2a). As an example illustrated in Fig. 2c, FiRE identified two out of five rare NK cells, while both GapClust and GiniClust3 failed to detect any rare cells. Their performance improved notably in the less challenging second setting, particularly GapClust. GapClust performed slightly better than aKNNO when there were only two rare cells, but its performance significantly lagged behind aKNNO when the number of rare cells exceeded two (Fig. 2b).

Application to single-cell transcriptomics data from human pancreas

We applied aKNNO to a single-cell RNAseq dataset from human pancreas by inDrops technology, involving 5542 cells and 14 cell populations manually annotated [23]. The 14 populations include three rare immune cell types, T, macrophage, and mast cells and one rare epsilon cell type (Fig. 3a). aKNNO identified 25 clusters in total (Fig. 3b). aKNNO identified all the manually annotated cell types except Schwann cells, including three rare immune cell types (clusters 18, 20, and 22 in Fig. 3b) and rare epsilon cells



i

Original	aKNNO	KNN	KNN_high	GiniClust3	GapClust	FiRE
macrophage	18 (n=25)	--	--	16 (n=25)	Rare_4 (n=25)	Rare (n=24)
--	19 (n=20); doublets	--	23 (n=18)	--	--	--
t_cell	20 (n=20)	--	--	--	--	--
	21 (n=17); doublets	--	--	--	--	Rare (n=17)
mast	22 (n=15)	--	--	22 (n=15)	--	--
epsilon	23 (n=13)	--	--	--	--	--
--	24 (n=10); LECs	--	--	--	--	Rare (n=10)

Fig. 3 Application to single-cell RNAseq data from human pancreas. The UMAP plot labeled by the manual annotation from the original study (**a**), labeled by aKNNO clustering (**b**), KNN (**c**), KNN_high (**d**). **e** Dot plot of marker genes in clusters detected by aKNNO. **f** The UMAP plot labeled by the GiniClust3 result. Only the largest 10 clusters were shown since there were too many clusters. **g** The UMAP plot labeled by the GapClust result. **h** The UMAP plot labeled by the FiRE result. **i** A summary of rare clusters identified in the original study, aKNNO, KNN, KNN_high, GiniClust3, GapClust, and FiRE

(cluster 23 in Fig. 3b, $n=13$). It also returned refined subclusters for abundant cell types. Compared to one ductal cell type in the manual annotation, aKNNO found three ductal clusters, clusters 5 ($n=360$), 15 ($n=78$), and 16 ($n=58$). The three clusters all expressed known ductal markers *KRT19* and *SOX9*, verifying their ductal identity. Compared to the abundant cluster 5, cluster 15 was specifically positive for *OLFM4*, and cluster 16 was high in *TFF1* and *IGFBP3* (Fig. 3e). These two clusters have been identified by a previous study on multipotent progenitor-like ductal cells, where *OLFM4*+ductal was named as "transition to acinar 1" and *TFF1*+*IGFBP3*+was labeled as "activated/migrating progenitor cells" [24]. Besides, aKNNO identified two endothelial clusters with high expression of *PECAM1* and *PLVAP* (clusters 9 and 24). Compared to cluster 9, the rare cluster 24 ($n=10$) showed highly specific expression of *PDPN* and *LYVE1* (Fig. 3e), which are well-known markers for lymphatic endothelial cells (LEC) [25]. As another example, aKNNO identified two activated stellate clusters with high level of *PDGFRB* (clusters 13 and 17), the abundant cluster 13 ($n=106$) with specific expression of *CXCL8* and *FGF2*, and the minor cluster 17 ($n=38$) with specific expression of *COL1A1* and *COL1A2* (Fig. 3e). This division of activated stellate cells have been reported in the original literature by further analyzing stellate cells [23]. In addition to discovering true rare cell clusters, aKNNO also detected rare doublets, clusters 19 ($n=20$) and 21 ($n=17$). They had relatively higher number of genes and UMIs than other corresponding cell types and showed signatures from two different cell types, indicating they are doublets. Cluster 19 not only showed high expression of beta cell markers like *INS*, but also highly expressed acinar cell markers like *CPA1*. Cluster 21 had both high expression of endothelial (*PECAM1* and *PLVAP*) and stellate markers (*PDGFRB*) (Additional file 1: Fig. S5).

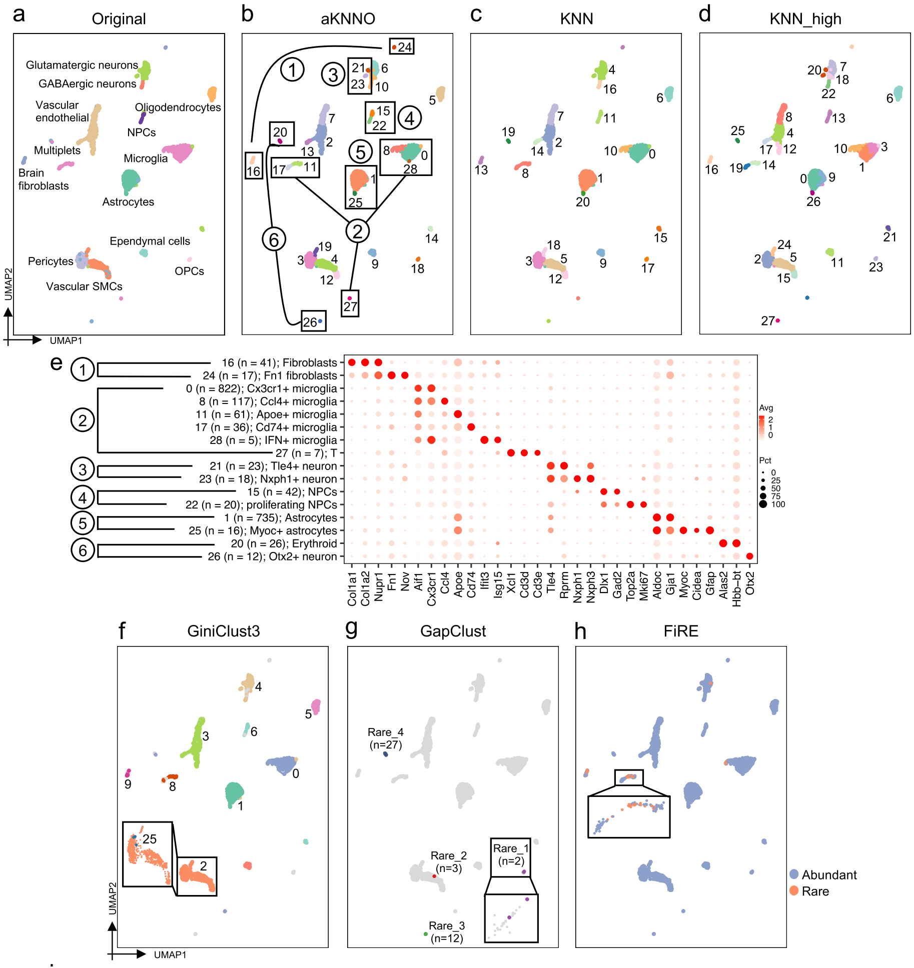
We compared aKNNO to KNN with default ($r=0.8$) and high resolutions ($r=2$, denoted as KNN_high). KNN discovered 16 clusters (Fig. 3c), while KNN_high identified 24 clusters in total (Fig. 3d). Surprisingly, KNN and KNN_high both failed to distinguish three rare immune clusters, T, macrophages, and mast cells (cluster 14 in the Fig. 3c and cluster 21 in the Fig. 3d), although they had very distinct features with high expression of *CCR7* and *TRAC* in T cells, *S100A9* and *CXCL8* in macrophages, and *KIT* in mast cells (Fig. 3e). KNN and KNN_high also missed the epsilon cells. KNN and KNN_high detected the two minor types of ductal cells but were unable to identify lymphatic endothelial cells, the two activated stellate cells, and doublets (Fig. 3c and d). Although KNN_high obtained 24 clusters, it overclustered abundant cell types rather than found rare clusters. With 25 clusters, in contrast, aKNNO found many true rare cell types and doublets without overclustering those abundant ones. There was high agreement between aKNNO and KNN in clustering abundant cells (ARI > 0.9), demonstrating that aKNNO is powerful in rare cell identification without scarifying its performance in abundant cells clustering.

We further compared rare cells identified by aKNNO, GiniClust3, FiRE, and GapClust. aKNNO identified seven rare clusters with less than or equal to 25 cells (clusters 18–24, $n=10\sim25$, Fig. 3i), where cells in each cluster were densely located together in the UMAP embedding. They were either manually annotated in the original literature (T, mast, macrophage, and epsilon cells) or supported by well-known marker genes (LEC, and two types of doublets) (Fig. 3i), demonstrating they were true rare cells or clusters. In comparison, GiniClust3 identified 85 clusters in total (Fig. 3f). Although

GiniClust3 obtained so many clusters, it misclassified even activated and quiescent stellate cells into one group (cluster 5 in the Fig. 3f) and three types of ductal cells into one cluster (cluster 2 in the Fig. 3f), indicating its poor performance in clustering abundant cell types. For the rare cell clusters, GiniClust3 identified mast, and macrophage, but missed epsilon cells, lymphatic endothelial cells, and two types of doublets (Fig. 3f and i). Most rare clusters identified by GiniClust3 scattered and mixed well with other abundant cells in the UMAP embedding and not detected by other methods, suggesting they were highly likely to be false rare cells (Fig. 3f). GapClust discovered four rare clusters, one of which is macrophage ($n = 25$, Rare_4 in the Fig. 3g). The other three rare clusters containing two, three, and nine cells, respectively, were highly likely to be false positives since they scattered and mixed with endothelial and activated stellate cells in the UMAP embedding (Fig. 3g). FiRE quantified 547 cells as being rare (Fig. 3h), which detected macrophage, lymphatic endothelial cells, and one type of doublets correctly but misidentified common endothelial cells and most of stellate cells as being rare ($n > 150$) and most of T cells and mast cells ($n < 20$) and all of epsilon cells ($n = 13$) as being common (Fig. 3h and i). In summary, aKNNO identified more true and less false rare cells than GiniClust3, GapClust, and FiRE.

Application to single-cell transcriptomics from mouse brain

We applied aKNNO to a single-cell RNA-seq dataset from mouse brain by 10 × technology, involving 3985 cells and 12 cell types manually annotated [26]. However, the UMAP embedding showed far more than 12 clearly-separated clusters, suggesting the original annotation is rough and imprecise (Fig. 4a). For example, there are two separated groups annotated as brain fibroblasts and multiple distinct clusters all labeled as microglia (Fig. 4a). aKNNO detected 29 clusters in total, which perfectly identified those separated groups in the UMAP as different clusters (Fig. 4b). For the two distinct groups manually annotated as brain fibroblasts, aKNNO identified one as cluster 16 ($n = 41$), and the other as cluster 24 ($n = 17$) (Fig. 4b). Both clusters expressed known fibroblast markers *Col1a1*, *Col1a2* and *Nupr1* and cluster 24 also specifically expressed *Fn1* and *Nov*, which was known as Fn1 fibroblasts [27] (Fig. 4e). In comparison, KNN and KNN_high failed to distinguish these two types of fibroblasts (cluster 13 in the Fig. 4c and cluster 16 in the Fig. 4d). For the three distinct groups manually annotated as microglia, aKNNO detected six clusters (clusters 0, 8, 11, 17, 27, and 28 in the Fig. 4b). Five clusters had high expression of *Aif1*, a known marker of microglia, while the rare cluster 27 ($n = 7$) had low expression of *Aif1* but specifically expressed *Xcl1*, *Cd3d*, and *Cd3e* (Fig. 4e), suggesting its T cell identity. Each of the five types of microglia had specific gene expression signatures, cluster 0 ($n = 822$) with high *Cx3cr1*, cluster 8 ($n = 117$) with high *Ccl4*, cluster 11 with high *Apoe* ($n = 61$), rare cluster 17 ($n = 36$) with high *Cd74*, and rare cluster 28 ($n = 5$) with high *Ifit3* and *Isg15* (Fig. 4e). The microglia clusters 0, 8, 11, and 17 have different functions reported by previous studies [28, 29], and cluster 28 is highly enriched in IFN-response genes, suggesting it is a novel type of microglia. Among the five types of microglia and T cells, KNN only identified three and KNN_high found five types (Fig. 4c and d). Besides GABAergic and Glutamatergic neurons, aKNNO identified two more types of rare neurons (clusters 21 and 23 in the Fig. 4b, n = 23 and $n = 18$ respectively), where cluster 21 had specific expression of *Tle4* and *Rprm* and cluster 23 was positive



i

Original	aKNNO	KNN	KNN_high	GiniClust3	GapClust	FiRE
--	21 (n=23); Tle4+ neuron	--	--	--	--	--
--	22 (n=20); proliferating NPCs	--	--	--	--	--
--	23 (n=18); Nxph1+ neuron	--	--	--	--	--
--	24 (n=17); Fn1 fibroblast	--	--	16 (n=11)	--	--
--	25 (n=16); Myoc+ astrocytes	20 (n=16)	26 (n=16)	--	--	--
--	26 (n=12); Otx2+ neuron	--	27 (n=12)	15 (n=12)	Rare_3 (n=12)	--
--	27 (n=7); Cd3g+ T	--	--	--	--	--
--	28 (n=5); IFN+ microglia	--	--	--	--	--

Fig. 4 Application to single-cell RNAseq data from mouse brain. The UMAP plot labeled by the manual annotation from the original study (**a**), aKNNO (**b**), KNN (**c**), KNN_high (**d**). **e** The dot plot of marker genes in clusters detected by aKNNO. **f** The UMAP plot labeled by the GiniClust3 result. Only the largest 10 clusters were shown. **g** The UMAP plot labeled by the GapClust result. **h** The UMAP plot labeled by the FiRE result. **i** A summary of rare clusters identified in the original study, aKNNO, KNN, KNN_high, GiniClust3, GapClust, and FiRE

for *Nxph1* and *Nxph3* (Fig. 4e). Those genes all play important roles in neurons [30], suggesting they are novel neuron types. KNN failed to identify both types of neurons, while KNN_high missed one type. In addition, AKNNO discovered two rare and distinct clusters (cluster 20, $n = 26$; cluster 26, $n = 10$ in the Fig. 4b), which were annotated as “multiplets” in the original annotation. Cluster 20 had high expression of *Alas2* and *Hbb-bt*, which are erythroid markers. Cluster 26 had specific expression of *Otx2*, which is known to regulate progenitor identity and neurogenesis in the midbrain [31, 32] (Fig. 4e).

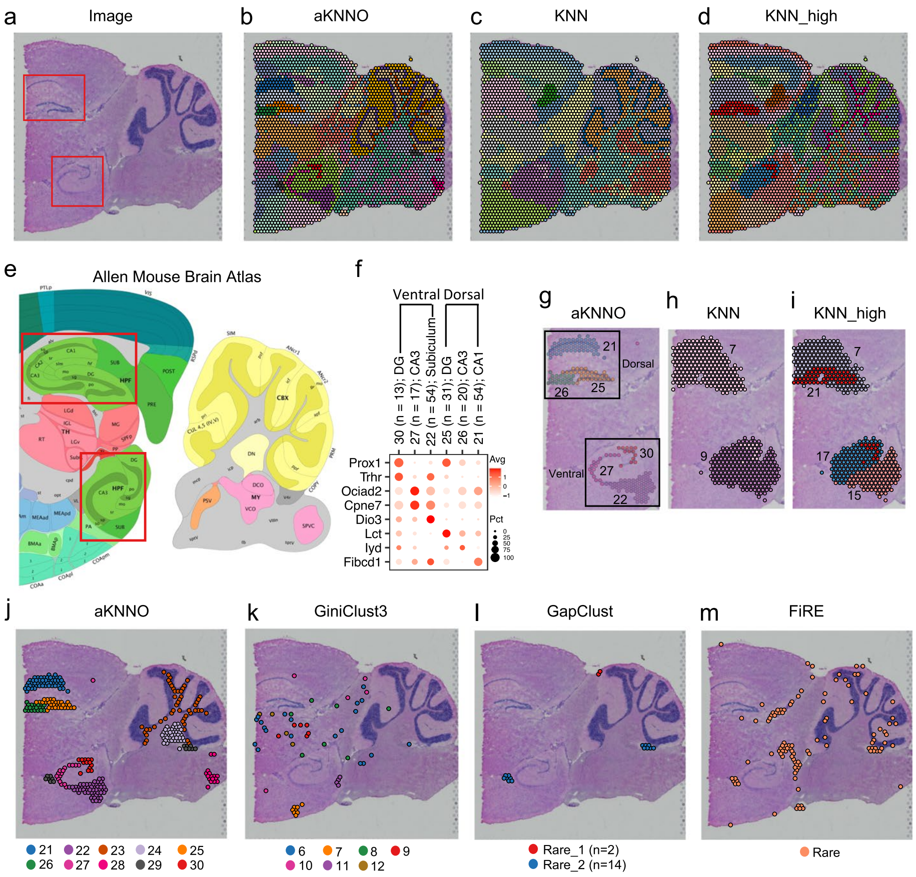
Besides, aKNNO and KNN were in high agreement on their clustering of abundant cells (ARI > 0.92), demonstrating aKNNO does not reduce their power at clustering abundant cells when it achieves great performance in identifying rare ones.

We further compared rare cells identified by aKNNO, GiniClust3, FiRE, and GapClust. aKNNO identified eight rare clusters with less than 25 cells (clusters 21–28, $n = 5 \sim 23$) (Fig. 4b and i). As discussed above, cluster 21 was *Tle4* + neuron, cluster 23 was *Nxph1* + neuron, cluster 24 was *Fn1* fibroblast, cluster 26 was Otx2 + neuron, cluster 27 was T cell, and cluster 28 was IFN + microglia. Cluster 22 was similar to cluster 10 with high expression of *Dlx1* and *Gad2*, which was manually annotated as neural progenitor cell (NPCs). However, cluster 22 expressed high *Top2a* and *Mki67* (Fig. 4e), suggesting it was proliferating NPCs. Cluster 25 and cluster 1 both had high expression of known astrocytes markers, such as *Aldoc* and *Gja1*. Cluster 25 had specific expression of *Myoc*, *Cidea*, and *Gfap* (Fig. 4e), which has been reported as a new type of Astrocytes [33]. In comparison, GiniCluster3 identified 27 clusters in total (Fig. 4f). For abundant cell types, it misidentified pericytes and vascular SMCs as one group (cluster 2 in the Fig. 4f), indicating its inefficiency in large cluster detection. Among the eight rare cell types identified aKNNO, GiniClust3 only identified Fn1 fibroblast and Otx2 + neuron (Fig. 4i). Cells in other rare clusters by GiniClust3 scattered in the UMAP embedding and mixed with other abundant cells, indicating they are not true rare cells (Fig. 4f). GapClust obtained four rare clusters, containing 2, 3, 12, and 27 cells, respectively (Fig. 4g). The cluster with 12 cells (Rare_3 in the Fig. 4g) corresponded to cluster 26 of aKNNO, which is Otx2 + neuron. The cluster with 27 cells (Rare_4 in the Fig. 4g) matched cluster 20 in the aKNNO, which was not that rare compared to the eight rare clusters. The other two rare clusters mixed with abundant cells in the UMAP embedding (Fig. 4g), suggesting they are not real rare. FiRE only found 56 rare cells (Fig. 4h), most of which mixed with microglia and fibroblasts. It failed to detect any of the eight rare clusters in aKNNO (Fig. 4i). Consistent with the results from human pancreas, aKNNO achieved higher sensitivity and specificity than other methods in identifying rare cells.

Application to spatial transcriptomics data from mouse posterior brain

Spatial transcriptomics map out organizational structures of cells along with their transcriptomics profiles, providing powerful tools for understanding spatial and functional arrangement of tissues [34].

We applied aKNNO to a 10 × Visium dataset generated from mouse sagittal posterior brain, which has complicated tissue structures. We used the brain anatomical reference annotations from the Allen Mouse Brain Atlas and H&E image as the ground truth (Fig. 5a and e). aKNNO identified 31 clusters from the 3355 spots, which stereotyped anatomical structures precisely (Fig. 5b). For example, the hippocampal system consists of the dentate gyrus (DG), cornu ammonis (CA) fields that are subdivided into four regions (CA1–CA4), and the subiculum [35]. The brain anatomical reference annotations and the H&E image of the data show dorsal and ventral hippocampus, where the dorsal contains DG, CA3, and CA1, and the ventral includes DG, CA3, and subiculum (highlighted in red boxes in the Fig. 5a and e). Previous transcriptomics studies on neuronal classes of the hippocampus found not only cell-class but also region-specific expression profiles, indicating heterogeneous populations along the dorsal-ventral axis



n

Ref	aKNNO	KNN	KNN_high	GiniClust3	GapClust	FiRE
CA1_dorsal	21 (n=54)	--	--	--	--	--
Sub_ventral	22 (n=54)	--	--	--	--	--
Oligodendrocytes	23 (n=48)	--	26 (n=47)	--	--	--
Dentate Nucleus	24 (n=42)	--	--	--	--	--
DG_dorsal	25 (n=31)	--	--	--	--	--
CA3_dorsal	26 (n=20)	--	--	--	--	--
CA3_ventral	27 (n=17)	--	--	--	--	--
Hox+ hindbrain	28 (n=16)	--	--	--	--	--
lateral ventricle and lateral recess	29 (n=14)	--	--	--	Rare_2 (n=14)	Rare (n=13)
DG_ventral	30 (n=13)	--	--	--	--	--

Fig. 5 Application to 10 × Visium spatial transcriptomics data from mouse posterior brain. **a** H&E image. The spatial plot annotated by aKNNO (**b**), KNN (**c**), and KNN_high (**d**). **e** Allen Brain Institute reference atlas diagram. **f** Dot plot of marker genes in the dorsal and ventral hippocampus structures. The spatial plot focusing on dorsal and ventral hippocampus structures annotated by the six cell clusters identified by aKNNO (**g**), two cell clusters detected by KNN (**h**), three cell clusters by KNN_high (**i**). The spatial plot of rare cell clusters detected by aKNNO (**j**), GiniClust3 (**k**), GapClust (**l**), and FiRE (**m**). **n** A summary of rare clusters identified by aKNNO, KNN, KNN_high, GiniClust3, GapClust, and FiRE

[36]. aKNNO discovered six clusters successfully with known cell- and region-specific expression (clusters 21, 22, 25, 26, 27, and 30 in the Fig. 5g), which mapped to the six anatomical patterns precisely. In the ventral hippocampus, cluster 30 ($n = 13$) aligned exclusively to DG and had high expression of known DG-specific (*Prox1*) and ventral-specific genes (*Trhr*) [36] (Fig. 5f). Cluster 27 ($n = 17$) depicted CA3 with CA3-specific expression of *Ociad2* and *Cpne7* [36], and cluster 22 corresponded to subiculum with known cell-class and region-specific expression of *Dio3* [37] (Fig. 5f). In the dorsal

region, cluster 25 ($n = 31$) lined up with DG with DG-specific gene *Prox1* and dorsal-specific gene *Lct* [36], cluster 26 ($n = 20$) mapped to CA3 with high CA3_dorsal-specific expression of *Iyd* [36], and cluster 21 ($n = 54$) matched CA1 with high expression of *Fibcd1* [36] (Fig. 5f). In comparison, KNN and KNN_high discovered 15 and 28 clusters, respectively, which showed poorer performance in defining anatomical patterns than aKNNO (Fig. 5c and d). For example, KNN and KNN_high failed to distinguish the six hippocampus subpopulations. KNN only found two clusters, where it misclassified DG_ventral and the whole dorsal hippocampus into one cluster, and it miscategorized CA3_ventral and subiculum into one and also masked them into the surrounding region (Fig. 5h). KNN_high detected four clusters, where it misclassified DG_ventral, CA3_dorsal, and DG_dorsal into one cluster, and it failed to distinct CA1_dorsal, CA3_ventral and subiculum from their surrounding regions (Fig. 5i).

The ten rare clusters identified by aKNNO (clusters 21–30 in the Fig. 5j, $n < 55$) all aligned exclusively to a specific brain region with functional meaning, including CA1_dorsal (cluster 21, $n = 54$), subiculum_ventral (cluster 22, $n = 54$), oligodendrocytes in the white matter (cluster 23, $n = 48$), dentate nucleus (cluster 24, $n = 42$), DG_dorsal (cluster 25, $n = 31$), CA3_dorsal (cluster 26, $n = 20$), CA3_ventral (cluster 27, $n = 17$), Hox gene-enriched hindbrain (cluster 28, $n = 16$), lateral ventricle and lateral recess (cluster 29, $n = 14$), and DG_ventral (cluster 30, $n = 13$) (Fig. 5e, j and n). In comparison, GiniClust3 detected only 13 clusters, which agreed poorly with the Allen Brain Institute reference atlas diagram (Additional file 1: Fig. S6). For example, multiple cortical layers were misclassified into one cluster and none of the hippocampus structure were identified correctly (Additional file 1: Fig. S6). The rare clusters (clusters 6–12 in the Fig. 5k with $n < 55$) scattered randomly in the spatial region and did not map to specific anatomical regions, suggesting that they were highly likely to be false. None of the ten rare clusters in aKNNO were detected by GiniCluster3 (Fig. 5n). GapClust only identified two rare clusters (Fig. 5l). One cluster containing 14 cells aligned to lateral ventricle and lateral recess region (cluster 29 in the aKNNO), while the other cluster with two cells was identified to be common by other methods. FiRE identified 105 rare cells (Fig. 5m). Among them, 13 cells were lined up with lateral ventricle and lateral recess region (cluster 29 in the aKNNO). Sixty-six cells were smooth muscle cells with high level of *Ogn* and *Prdm6*, which were also identified by aKNNO (Additional file 1: Fig. S7, $n = 105$) and were not that rare compared to the ten rare clusters. Other cells scattered in the brain region and might not be true rare cells. In summary, the ten rare clusters identified by aKNNO stereotyped subtle anatomical structures, whereas GapClust and FiRE only recovered the lateral ventricle and lateral recess structure (Fig. 5n).

Recently, several integrative approaches combining expression, spatial location, and histology images have been developed to improve the performance of spatial transcriptomics clustering [38–42]. We compared aKNNO with five integrative approaches, stLearn [39], SpaGCN [38], GraphST [40], BayesSpace [41], and DR-SC[42]. We set their number of clusters to 31, matching the cluster number of aKNNO (Fig. 6a–f). Using the fine-grained hippocampus structures as an example, aKNNO stereotyped six anatomical patterns precisely (Fig. 6g). In contrast, all the five integrative approaches failed to resolve the six patterns. Both BayesSpace and DR-SC misclassified CA3_dorsal and CA1_dorsal into one (cluster 28 in Fig. 6h and l) as well as DG_dorsal and DG_ventral

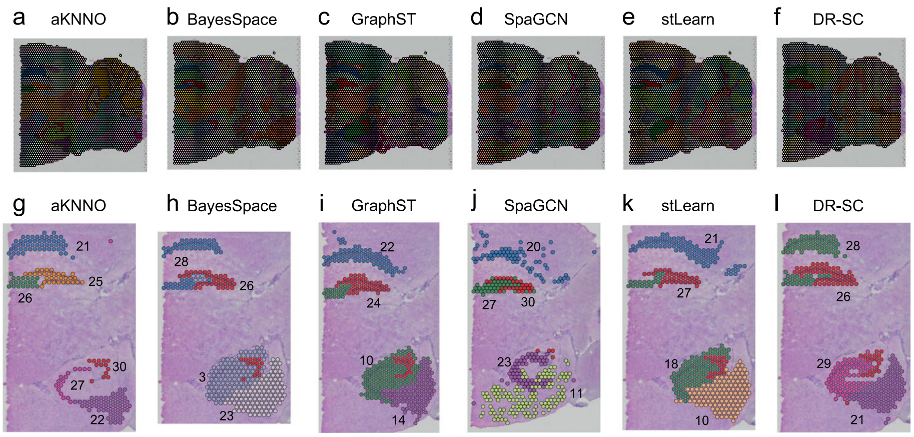


Fig. 6 Comparison between aKNNO, BayesSpace, GraphST, SpaGCN, stLearn, and DR-SC on the 10 × Visium spatial transcriptomics data from mouse posterior brain. The spatial pot annotated by aKNNO (**a**), BayesSpace (**b**), GraphST (**c**), SpaGCN (**d**), stLearn (**e**), and DR-SC (**f**). Detailed view of clustering in hippocampus structures in aKNNO (**g**), BayesSpace (**h**), GraphST (**i**), SpaGCN (**j**), stLearn (**k**), and DR-SC (**l**)

into one (cluster 26 in Fig. 6h and l). However, DR-SC successfully identified CA3_ventral and subiculum (clusters 29 and 21 in Fig. 6l), while BayesSpace failed to distinguish them from their surrounding regions (clusters 3 and 23 in Fig. 6h). GraphST identified CA1_dorsal (cluster 22 in Fig. 6i) but misrecognized DG_dorsal and DG_ventral into one (cluster 24 in Fig. 6i) and CA3_dorsal and CA3_ventral into one (Cluster 10 in Fig. 6i). SpaGCN successfully identified CA3_dorsal (cluster 27 in Fig. 6j) and CA3_ventral (cluster 23 in Fig. 6j) but misclassified DG_dorsal and DG_ventral into one (cluster 30 in Fig. 6j) and also lost the subiculum structure. stLearn, similar to GraphST, detected CA1_dorsal (cluster 21 in Fig. 6k) but failed to distinguish between CA3_dorsal and CA3_ventral (cluster 18 in Fig. 6k) and between DG_dorsal and DG_ventral (cluster 27 in the Fig. 6k). The comparison demonstrated that aKNNO, using gene expression alone, resolved tissue structures more accurately than those integrative approaches.

Discussion

The accurate detection of abundant and rare clusters simultaneously is crucial for characterizing cellular heterogeneity in single-cell and spatial transcriptomics analysis. Here, we presented aKNNO, an adaptive *k*-nearest neighbor graph with optimization for the community-detection-based clustering. Compared to traditional *k*NN specifying a universal *k* for all cells, aKNNO chooses *k* adaptively for each cell based on its local distance distribution. The adaptive strategy assigns a small *k* for rare cells and a large *k* for abundant cells, enabling to capture the inherent cellular structure accurately. aKNNO has been extensively evaluated on 38 simulated scenarios and 20 single-cell and spatial transcriptomics data from different species, tissues, and technologies. Additional analysis on mouse intestine from 10x [43], mouse habenula from inDrops [44], organoids from CEL-seq [8], spatial main olfactory bulb, spatial coronal posterior brain, and 12 pan-cancer datasets [45, 46] were included in the Additional file 1: Sections S1-S6. The results consistently demonstrated that aKNNO outperformed traditional *k*NN-based approach in identifying abundant and rare cell types in a single run. aKNNO identifies rare cells without overclustering abundant cells. In both simulated and real datasets, aKNNO did not overcluster those *k*NN-based clusters with $\emptyset_{clust}=0$. The substructures identified by aKNNO demonstrate both high Phiclust scores and distinct well-known markers, highlighting their biological relevance over mere statistical noise (Additional file 1: Figs. S3, S4, and S8). It is important to note that aKNNO aims to construct a graph that accurately capture inherent cellular structure, rather than pinpointing the optimal clustering resolution. In fact, aKNNO can be clustered at any resolution. Due to Phiclust scores being sensitive to technical and biological noise, high Phiclust scores, coupled with the presence of distinct, meaningful markers, might reliably indicate the existence of true subclusters. aKNNO was also far more superior than those methods specifically designed or tailored for rare cells detection in terms of both sensitivity and specificity.

aKNNO provides an optimization step to tune δ, a hyperparameter controlling the sensitivity to local distance change. High sensitivity to local distance change would result in many localized subclusters and lead to overclustering, while low sensitivity would lose the adaptive ability and reduce to the traditional *k*NN. aKNNO uses grid search to find the optimal δ for each dataset. We compared the performance between the optimal and the default δ (-0.5). We found the optimal δ was different across datasets. aKNNO using the optimal δ was able to find more true rare cell types than the approach using the default δ without optimization (Additional file 1: Figs. S9-S12). That is, the optimization step did help balance between sensitivity and specificity in rare cells identification. aKNNO is less sensitive to clustering resolution than *k*NN-based approaches thanks to its ability to capture the inherent cellular structure more accurately (Additional file 1: Fig. S2). Furthermore, the performance of aKNNO is unaffected by the parameter $K_{\max}$, representing the maximum number of nearest neighbors. We conducted a sensitivity analysis for the parameter $K_{\max}$ in simulation studies, demonstrating that the performance of aKNNO remains consistent across varying values of $K_{\max}$ (ranging from 10 to 50) when the number of rare cells ranges from 2 to 10 (Additional file 1: Fig. S13).

aKNNO builds a nearest neighbor graph, which can be seamlessly incorporated into existing analysis pipelines (see tutorials in the GitHub) and also used for any graph-based clustering approaches, such as Louvain clustering, spectral clustering, Leiden clustering, and Minimal Spanning Tree. aKNNO not only identified true rare clusters but also found doublets and empty droplets. For example, aKNNO found two types of doublets in the human pancreas and one cluster with only two empty droplets in the mouse habenula dataset, that is, aKNNO can be used to remove poor quality cells that are not eliminated completely in the preprocessing step. Although an adaptive strategy is included to build the NN-graph, the computational burden added is neglectable compared to the traditional *k*NN-based approach. aKNNO is comparable to the traditional *k*NN-based approach and significantly more efficient than GapClust, GiniClust3, and FiRE in terms of run time and memory usage (Additional file 1: Fig. S14).

Spatial transcriptomics is an emerging technology that provides a roadmap of transcriptional activity within tissue sections. To better decipher domains or cell types that are spatially coherent in both gene expression and histology, a number of integrative approaches to combine gene expression, spatial location, histology, and H&E image have been developed [38–42]. Integrating multi-modal information are expected to define cell types or domains accurately than using gene expression alone. Surprisingly, aKNNO is able to stereotype anatomical cell types precisely using gene expression alone. For example, the six clusters identified by aKNNO aligned almost perfectly with the six dorsal and ventral hippocampus structures (DG_dorsal, DG_ventral, CA3_dorsal, CA3_ventral, CA1_dorsal, and subiculum_ventral). Each common and rare cluster detected by aKNNO mapped to specific regions in the Allen Brain Institute reference atlas diagram, suggesting they are functionally meaningful. Some clusters or spatial regions were even not identified by those integrative approaches (Fig. 6 and Additional file 1: Sections S4-S5). The results suggest that gene expression alone is enough to define spatial cell types or domains when the appropriate strategy is used.

Conclusions

We proposed aKNNO, a novel clustering method to identify abundant and rare cell types simultaneously for single-cell and spatial transcriptomics data. aKNNO benefits from choosing *k* adaptively for each cell based on its local distance distribution, which captures inherent cellular structure accurately. Without sacrificing performance for clustering abundant cell types, aKNNO discovered known and novel rare cell types that those typical and even specifically tailored methods failed to detect. Notably, aKNNO using transcriptome alone stereotyped fine-grained anatomical structures more precisely than those spatially and histology informed approaches. aKNNO provides a powerful and accurate way in unsupervised clustering of single-cell and spatial transcriptomics data.

Methods

Single-cell and spatial transcriptomics data processing

Single-cell RNAseq datasets were filtered and processed by the Seurat package [15]. Specifically, cells expressing less than 200 genes and genes expressed in less than three cells were excluded. Data were normalized to the 10,000 UMI, and the top 2000 highly variable genes were selected by the vst method. Principal component analysis (PCA) was used to reduce dimension and UMAP embedding was computed for visualization.

Spatial transcriptomics data were processed by the Seurat package [47]. Particularly, data were normalized and the top 3000 highly variable genes were selected using SCTtransform [48]. Genes expressed in less than 10% of spots in the mouse sagittal posterior brain were filtered out, following preprocessing steps in previous studies [38, 49]. Principal component analysis (PCA) was used to reduce dimension and UMAP embedding was computed for visualization.

aKNNO

aKNNO includes five steps: (1) calculate the distances within the K_{max} nearest neighbors for all cells; (2) choose the actual number of the true nearest neighbors *k* adaptively for each cell to construct an adaptive *k*NN graph; (3) build the shared nearest neighbor graph; (4) Louvain clustering on the shared neighbor graph; (5) repeat steps (2)–(4) for optimization (Fig. 1).

1) Calculate the distances within the K_{max} nearest neighbors for all cells; the top PCs (default: 50) were used to calculate the distances between cells. Given a maximum

number of the nearest neighbors K_{max}, RANN package is used to find the K_{max} nearest neighbors of all cells and calculate the distances

2) Construct an adaptive *k*NN graph by choosing *k* adaptively for each cell; for each cell *i*, the distance to its K_{max} nearest neighbors are sorted in an ascending order ($d_{i1} < d_{i2} < \ldots d_{iK\mathrm{max}}$) and *k* is chosen based on the K_{max} distance distribution [50]. Several algorithms have been proposed to find adaptive neighbors [51–54]. We used the algorithm described in Cai et al. [50], which does not require the regularization term and thus be more computational efficient

Assuming there are *n* cells in total (1,2,...*n*), for the cell *i*, the cell *j* can be connected to *i* as a neighbor with probability S_{ij}. Intuitively, a smaller distance should be assigned a larger probability. Cai et al. (Cai et al. [50]) computes the probability by minimizing the following objective function:

$$min \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} \sum\nolimits_{m=0}^{\infty} (d_{ij} s_{ij}^{m}) \tag{1}$$

$$s.t. \forall_i, 0 \leq s_{ij} \leq 1, s_i^T 1 = 1$$

where *m* is the number of power, and d_{ij} is the distance between the cell *i* and *j*. The standard Karush–Kuhn–Tucker conditions are used to solve the Eq. (1). The optimal solution is:

$$s_{ij} = \begin{cases} 1, if\ d_{ij} < (\frac{\sum_{l=1}^{K_{max}} \sqrt{d_{il}}}{K_{max}-1-\delta})^2 \\ 0, otherwise \end{cases}$$

where δ is a hyperparameter to control the sensitivity to the local distance change ($\delta \leq 0$). $s_{ij} = 1$ indicates there is a connection between *i* and *j*, while $s_{ij} = 0$ means no connection. Therefore, for the K_{max} nearest neighbors of the cell *i* ($d_{i1} < d_{i2} < \ldots d_i K_{max}$),

$$k_i = \begin{cases} K_{max}, if\ d_{iK_{max}} < (\frac{\sum_{l=1}^{K_{max}} \sqrt{d_{il}}}{K_{max}-1-\delta})^2 \\ j, if\ d_{ij} < (\frac{\sum_{l=1}^{K_{max}} \sqrt{d_{il}}}{K_{max}-1-\delta})^2\ and\ d_{i(j+1)} \geq \left(\frac{\sum_{l=1}^{K_{max}} \sqrt{d_{il}}}{K_{max}-1-\delta}\right)^2, j = 1,2, \ldots K_{max} - 1 \end{cases}$$

For example, the true nearest neighbor *k* for the cell *i* is set to 3 automatically ($k_i = 3$) if $d_{i3} < (\frac{\sum_{l=1}^{K_{max}} \sqrt{d_{il}}}{K_{max}-1-\delta})^2\ and\ d_{i4} \geq (\frac{\sum_{l=1}^{K_{max}} \sqrt{d_{il}}}{K_{max}-1-\delta})^2$. In this way, aKNNO chooses the *k* adaptively based on the local distance distribution for each cell.

3) Build the shared nearest neighbor graph; The adaptive *k*NN graph is reweighted based on the shared nearest neighbors of each pair of cells, which make it more robust to outliers and noise. The Jaccard similarity is used to build the shared nearest neighbor graph.
4) Louvain clustering on the shared neighbor graph; the Louvain clustering in the Seurat package is used to group cells on the shared nearest neighbor graph.
5) Repeat steps (2)–(4) for optimization. δ is a hyperparameter to control the sensitivity to the local distance change (default = − 0.5). A too negative δ would make aKNNO very sensitive to local distance change and lead to overclustering, while aKNNO

would lose the adaptive ability if δ is not negative enough. To find the optimal δ, aKNNO employs a grid search to find the δ that balances between the sensitivity and specificity. aKNNO decreases δ in each repeat and choose the δ before there is a rapid increase in the number of communities detected, which suggests overclustering (Fig. 1).

Comparison with other methods

We compared aKNNO with the conventional *k*NN-based clustering method in the Seurat package (denoted as KNN). aKNNO identified clusters at the default solution ($r = 0.8$), while KNN found clusters at the default resolution ($r = 0.8$) and a high resolution ($r = 2$, denoted as KNN_high). To make a fair comparison, aKNNO and KNN processed the data exactly the same way, including the highly variable genes selection, the number of PCs to calculate the distance, the method to find the nearest neighbors, and Louvain clustering. The only difference between aKNNO and KNN is the graph used for clustering. aKNNO generates clustering from the adaptive *k*NN graph, while KNN from the traditional *k*NN graph.

We also compared aKNNO with three methods specifically designed or tailored to identify rare cell types, GiniClust3 [18], GapClust [20], and FiRE [19]. GiniClust3 is an extension of GiniClust, which discovers rare cells based on genes with high Gini index. FiRE uses the Sketching technique to assign a rareness score to each cell [19]. GapClust captures the abrupt local distances change to find rare cell clusters [20]. GiniClust3 identifies abundant and rare clusters simultaneously, GapClust finds rare clusters only, and FiRE quantifies rareness of each cell without clustering. Default parameters were used to run GiniClust3, GapClust, and FiRE. GapClust requires a normalized expression matrix as the input; therefore, the data were normalized to 10,000 UMI without log-transformation for GapClust.

In the spatial transcriptomics data, we further compared aKNNO with five integrative approaches, BayesSpace [41], GraphST [40], SpaGCN [38], stLearn [39], and DR-SC[42]. BayesSpace implements a full Bayesian model that uses the information from spatial neighborhoods for resolution enhancement for spatial clustering. stLearn first normalizes gene expression by distance measures on morphological similarity and neighborhood smoothing and then clustering cell types from the normalized expression profiles. SpaGCN uses a graph convolutional network that integrate gene expression, spatial location, and histology images. GraphST is a graph self-supervised contrastive learning method that combines spatial location and expression. DR-SC simultaneously performs dimension reduction and spatial clustering within a unified framework, which encourages spatial smoothness based on a latent hidden Markov random field. Default parameters were used to run BayesSpace, GraphST, SpaGCN, stLearn, and DR-SC. The number of clusters was set to the same as aKNNO.

Simulation studies

We generated simulated datasets by sampling four cell types from the PBMC3k dataset: B cells, naïve CD4 T cells, NK, and CD14+monocytes. We created two settings. In one setting, the rare cell type was similar to one of the abundant cell types, i.e., rare NK but abundant B ($n = 200$) and naïve CD4 T ($n = 200$) cells. In the other setting, the rare cell type was distinct from both abundant cell types, i.e., rare CD14+monocytes but abundant B ($n = 200$) and naïve CD4 T cells ($n = 200$). In each setting, we simulated 19 scenarios with the number of rare cells ranging from 2 to 20. In each scenario, 50 simulated datasets were generated by random sampling the four cell types from the PBMC3k dataset, which was downloaded from https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03339-y.

Additional file 1. Supplementary information, includes supplementary figures and supplementary sections for other applications of aKNNO.

Additional file 2. Review history.

Review history

The review history is available as Additional File 2.

Peer review information

Veronique van den Berghe was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

Q.L. and Y.S. conceived and designed the study. J.L. and Q.L. collected and analyzed the sequencing data. J.L. and Q.L. implemented the algorithm and developed the package. J.L. and Q.L. wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work is supported by National Cancer Institute grants (U2C CA233291, P01CA229123, and U54 CA274367), National Institutes of Health (P01 AI139449), Cancer Center Support Grant (P30CA068485), and The Leona M. and Harry B. Helmsley Charitable Trust [G-1903-03793].

Availability of data and materials

The R package aKNNO, together with tutorials for reproducing the analysis results presented in the study, is freely available on GitHub (https://github.com/liuqivandy/aKNNO) [55] and Zenodo [56]. The source code is released under the GNU General Public License v3.0. The single-cell RNAseq data for human pancreas, mouse brain, mouse intestine, mouse habenula, and intestinal organoids are available at the Gene Expression Omnibus (GEO) under accession numbers GSE84133 [23, 57], GSM3580745 [26, 58], GSM4521364 [59], GSM4411753 [44, 60], and GSE62270 [8, 61–63], respectively. The 10x Visium data for mouse coronal and sagittal posterior brain [64] and main olfactory bulb are downloaded from the 10x Genomics website (https://www.10xgenomics.com/resources/datasets). The single-cell RNAseq dataset of the melanoma is obtained from GEO under accession number GSE72056 [45, 65], and the other 11 cancer datasets were downloaded from the database Tumor Immune Single-cell Hub 2 (TISCH2) [46, 66–73]. Lymphatic endothelial cells (LEC) were identified in 12 datasets [74–79].

Declarations

Ethics approval and consent participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 16 June 2023 Accepted: 16 July 2024
Published online: 01 August 2024

References

1. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, Chen H, Wang J, Tang H, Ge W, et al. Construction of a human cell landscape at single-cell level. Nature. 2020;581:303–9.

2. Rao A, Barkley D, Franca GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature. 2021;596:211–20.
3. Elmentaite R, Dominguez Conde C, Yang L, Teichmann SA. Single-cell atlases: shared and tissue-specific cell types across human organs. Nat Rev Genet. 2022;23:395–410.
4. Fei L, Chen H, Ma L, E W, Wang R, Fang X, Zhou Z, Sun H, Wang J, Jiang M, et al. Systematic identification of cell-fate regulatory programs using a single-cell atlas of mouse development. Nat Genet. 2022;54:1051–61.
5. Yu L, Cao Y, Yang JYH, Yang P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. Genome Biol. 2022;23:49.
6. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. kdd. 1996;96:226–31.
7. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):P10008.
8. Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525:251–5.
9. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods. 2017;14:414–6.
10. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14:483–6.
11. Lin P, Troup M, Ho JW. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18:59.
12. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347:1138–42.
13. Zurauskiene J, Yau C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics. 2016;17:140.
14. Levine JH, Simonds EF, Bendall SC, Davis KL, el Amir AD, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162:184–97.
15. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33:495–502.
16. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:15.
17. Jiang L, Chen H, Pinello L, Yuan GC. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016;17:144.
18. Dong R, Yuan GC. GiniClust3: a fast and memory-efficient tool for rare cell type identification. BMC Bioinformatics. 2020;21:158.
19. Jindal A, Gupta P. Jayadeva, Sengupta D: discovery of rare cells from voluminous single cell expression data. Nat Commun. 2018;9:4719.
20. Fa B, Wei T, Zhou Y, Johnston L, Yuan X, Ma Y, Zhang Y, Yu Z. GapClust is a light-weight approach distinguishing rare cells from voluminous single cell expression profiles. Nat Commun. 2021;12:4197.
21. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20:273–82.
22 Mircea M, Hochane M, Fan X, de Chuva Sousa Lopes SM, Garlaschelli D, Semrau S. Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations. Genome Biol. 2022;23:18.
23. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst. 2016;3:346–360 e344.
24. Qadir MMF, Alvarez-Cubela S, Klein D, van Dijk J, Muniz-Anquela R, Moreno-Hernandez YB, Lanzoni G, Sadiq S, Navarro-Rubio B, Garcia MT, et al. Single-cell resolution analysis of the human pancreatic ductal progenitor cell niche. Proc Natl Acad Sci U S A. 2020;117:10876–87.
25. Fujimoto N, He Y, D'Addio M, Tacconi C, Detmar M, Dieterich LC. Single-cell mapping reveals new markers and functions of lymphatic endothelial cells in lymph nodes. PLoS Biol. 2020;18:e3000704.
26. Heng JS, Rattner A, Stein-O'Brien GL, Winer BL, Jones BW, Vernon HJ, Goff LA, Nathans J. Hypoxia tolerance in the Norrin-deficient retina and the chronically hypoxic brain studied at single-cell resolution. Proc Natl Acad Sci U S A. 2019;116:9103–14.
27. DeSisto J, O'Rourke R, Jones HE, Pawlikowski B, Malek AD, Bonney S, Guimiot F, Jones KL, Siegenthaler JA. Single-cell transcriptomic analyses of the developing meninges reveal meningeal fibroblast diversity and function. Dev Cell. 2020;54:43–59 e44.
28. Hammond TR, Dufort C, Dissing-Olesen L, Giera S, Young A, Wysoker A, Walker AJ, Gergits F, Segel M, Nemesh J, et al. Single-cell RNA sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex cell-state changes. Immunity. 2019;50:253–271 e256.
29. Stratoulias V, Venero JL, Tremblay ME, Joseph B. Microglial subtypes: diversity within the microglial community. EMBO J. 2019;38:e101997.
30. Di Bella DJ, Habibi E, Stickels RR, Scalia G, Brown J, Yadollahpour P, Yang SM, Abbate C, Biancalani T, Macosko EZ, et al. Molecular logic of cellular diversification in the mouse cerebral cortex. Nature. 2021;595:554–9.
31. Zeisel A, Hochgerner H, Lonnerberg P, Johnsson A, Memic F, van der Zwan J, Haring M, Braun E, Borm LE, La Manno G, et al. Molecular architecture of the mouse nervous system. Cell. 2018;174:999–1014 e1022.
32. Vernay B, Koch M, Vaccarino F, Briscoe J, Simeone A, Kageyama R, Ang SL. Otx2 regulates subtype specification and neurogenesis in the midbrain. J Neurosci. 2005;25:4856–67.
33. Wu YE, Pan L, Zuo Y, Li X, Hong W. Detecting activated cell populations using single-cell RNA-Seq. Neuron. 2017;96:313–329 e316.

34. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 2021;22:627–44.
35. Wible CG. Hippocampal physiology, structure and function and the neuroscience of schizophrenia: a unified account of declarative memory deficits, working memory deficits and schizophrenic symptoms. Behav Sci (Basel). 2013;3:298–315.
36. Cembrowski MS, Wang L, Sugino K, Shields BC, Spruston N. Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons. Elife. 2016;5:e14997.
37. Ding SL, Yao Z, Hirokawa KE, Nguyen TN, Graybuck LT, Fong O, Bohn P, Ngo K, Smith KA, Koch C, et al. Distinct transcriptomic cell types and neural circuits of the subiculum and prosubiculum along the dorsal-ventral axis. Cell Rep. 2020;31:107648.
38. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, Lee EB, Shinohara RT, Li M. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods. 2021;18:1342–51.
39. Pham D, Tan X, Balderson B, Xu J, Grice LF, Yoon S, Willis EF, Tran M, Lam PY, Raghubar A, Kalita-de Croft P. Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. Nat Commun. 2023;14(1):7739.
40. Long Y, Ang KS, Li M, Chong KLK, Sethi R, Zhong C, Xu H, Ong Z, Sachaphibulkij K, Chen A, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. Nat Commun. 2023;14:1155.
41. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, Williams SR, Uytingco CR, Taylor SEB, Nghiem P, et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol. 2021;39:1375–84.
42. Liu W, Liao X, Yang Y, Lin H, Yeong J, Zhou X, Shi X, Liu J. Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data. Nucleic Acids Res. 2022;50:e72.
43. Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, Burgin G, Delorey TM, Howitt MR, Katz Y, et al. A single-cell survey of the small intestinal epithelium. Nature. 2017;551:333–9.
44 Wallace ML, Huang KW, Hochbaum D, Hyun M, Radeljic G, Sabatini BL. Anatomical and single-cell transcriptional profiling of the murine habenular complex. Elife. 2020;9:e51271.
45. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352:189–96.
46. Sun D, Wang J, Han Y, Dong X, Ge J, Zheng R, Shi X, Wang B, Li Z, Ren P, et al. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. Nucleic Acids Res. 2021;49:D1420–30.
47. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019;177:1888–1902 e1821.
48. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296.
49. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020;17:193–200.
50. Cai Y, Huang JZ, Yin J. A new method to build the adaptive k-nearest neighbors similarity graph matrix for spectral clustering. Neurocomputing. 2022;493:191–203.
51. Nie F, Wang X, Huang H. Clustering and projected clustering with adaptive neighbors. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014. p. 977–86.
52. Kang Z, Pan H, Hoi SCH, Xu Z. Robust graph learning from noisy data. IEEE Trans Cybern. 2020;50:1833–43.
53. Bian Z, Ishibuchi H, Wang S. Joint learning of spectral clustering structure and fuzzy similarity matrix of data. IEEE Trans Fuzzy Syst. 2019;27:31–44.
54. Yuan M, Zhu Q. Spectral clustering algorithm based on fast search of natural neighbors. IEEE Access. 2020;8:67277–88.
55. Jia Li YS, Qi Liu: aKNNO: single-cell and spatial transcriptomics clustering with an optimized adaptive k-nearest neighbor graph. GitHub. 2024. https://github.com/liuqivandy/aKNNO.
56. Jia Li YS, Qi Liu: aKNNO: single-cell and spatial transcriptomics clustering with an optimized adaptive k-nearest neighbor graph. Zenodo. 2024. https://zenodo.org/records/12585444.
57. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al: A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. GSE84133. Gene Expression Omnibus. 2016. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133.
58. Heng JS, Rattner A, Stein-O'Brien GL, Winer BL, Jones BW, Vernon HJ, Goff LA, Nathans J: Hypoxia tolerance in the Norrin-deficient retina and the chronically hypoxic brain studied at single-cell resolution. GSM3580745. Gene Expression Omnibus. 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3580745.
59. Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, Burgin G, Delorey TM, Howitt MR, Katz Y, et al: A single-cell survey of the small intestinal epithelium. GSM4521364. Gene Expression Omnibus. 2017. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4521364.
60. Wallace ML, Huang KW, Hochbaum D, Hyun M, Radeljic G, Sabatini BL: Anatomical and single-cell transcriptional profiling of the murine habenular complex. GSM4411753. Gene Expression Omnibus. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4411753.
61. Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A: Single-cell messenger RNA sequencing reveals rare intestinal cell types. GSE62270. Gene Expression Omnibus. 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62270.
62. Lopez-Diaz L, Jain RN, Keeley TM, VanDussen KL, Brunkan CS, Gumucio DL, Samuelson LC. Intestinal neurogenin 3 directs differentiation of a bipotential secretory progenitor to endocrine cell rather than goblet cell fate. Dev Biol. 2007;309:298–305.
63. Zhu Y, Liu Q, Zhou Z, Ikeda Y. PDX1, neurogenin-3, and MAFA: critical transcription regulators for beta cell development and regeneration. Stem Cell Res Ther. 2017;8:240.

64. Chevreau R, Ghazale H, Ripoll C, Chalfouh C, Delarue Q, Hemonnot-Girard AL, Mamaeva D, Hirbec H, Rothhut B, Wahane S, et al. RNA profiling of mouse ependymal cells after spinal cord injury identifies the oncostatin pathway as a potential key regulator of spinal cord stem cell fate. Cells. 2021;10(12):3332.
65. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al: Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. GSE72056. Gene Expression Omnibus. 2016. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72056.
66. Zhang M, Yang H, Wan L, Wang Z, Wang H, Ge C, Liu Y, Hao Y, Zhang D, Shi G, et al. Single-cell transcriptomic architecture and intercellular crosstalk of human intrahepatic cholangiocarcinoma. J Hepatol. 2020;73:1118–30.
67. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171:1611–1624 e1624.
68. Pandiani C, Strub T, Nottet N, Cheli Y, Gambi G, Bille K, Husser C, Dalmasso M, Beranger G, Lassalle S, et al. Single-cell RNA sequencing reveals intratumoral heterogeneity in primary uveal melanomas and identifies HES6 as a driver of the metastatic disease. Cell Death Differ. 2021;28:1990–2000.
69. Jeong HY, Ham IH, Lee SH, Ryu D, Son SY, Han SU, Kim TM, Hur H. Spatially distinct reprogramming of the tumor microenvironment based on tumor invasion in diffuse-type gastric cancers. Clin Cancer Res. 2021;27:6529–42.
70. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, Bassez A, Decaluwe H, Pircher A, Van den Eynde K, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med. 2018;24:1277–89.
71. Gaydosik AM, Tabib T, Geskin LJ, Bayan CA, Conway JF, Lafyatis R, Fuschiotti P. Single-cell lymphocyte heterogeneity in advanced cutaneous T-cell lymphoma skin tumors. Clin Cancer Res. 2019;25:4443–54.
72. Gaydosik AM, Queen DS, Trager MH, Akilov OE, Geskin LJ, Fuschiotti P. Genome-wide transcriptome analysis of the STAT6-regulated genes in advanced-stage cutaneous T-cell lymphoma. Blood. 2020;136:1748–59.
73. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, Kumar T, Hu M, Sei E, Davis A, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39:599–608.
74. Jiang X, Wang J, Deng X, Xiong F, Zhang S, Gong Z, Li X, Cao K, Deng H, He Y, et al. The role of microenvironment in tumor angiogenesis. J Exp Clin Cancer Res. 2020;39:204.
75. He M, He Q, Cai X, Chen Z, Lao S, Deng H, Liu X, Zheng Y, Liu X, Liu J, et al. Role of lymphatic endothelial cells in the tumor microenvironment-a narrative review of recent advances. Transl Lung Cancer Res. 2021;10:2252–77.
76. Zhang J, Lu T, Lu S, Ma S, Han D, Zhang K, Xu C, Liu S, Gan L, Wu X, et al. Single-cell analysis of multiple cancer types reveals differences in endothelial cells between tumors and normal tissues. Comput Struct Biotechnol J. 2023;21:665–76.
77. Qian J, Olbrecht S, Boeckx B, Vos H, Laoui D, Etlioglu E, Wauters E, Pomella V, Verbandt S, Busschaert P, et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. Cell Res. 2020;30:745–62.
78. Yang D, Guo P, He T, Powell CA. Role of endothelial cells in tumor microenvironment. Clin Transl Med. 2021;11:e450.
79. Wang W, Wang L, She J, Zhu J. Examining heterogeneity of stromal cells in tumor microenvironment based on pan-cancer single-cell RNA sequencing data. Cancer Biol Med. 2021;19:30–42.

Publisher’s Note