

Computer-assisted interpretation, in-depth exploration and single cell type annotation of RNA sequence data using k-means clustering algorithm

Pranshu Saxena, Amit Sinha & Sanjay Kumar Singh


To cite this article: Pranshu Saxena, Amit Sinha & Sanjay Kumar Singh (18 Jan 2024): Computer-assisted interpretation, in-depth exploration and single cell type annotation of RNA sequence data using k-means clustering algorithm, Computer Methods in Biomechanics and Biomedical Engineering, DOI: [10.1080/10255842.2023.2300685](https://doi.org/10.1080/10255842.2023.2300685)

To link to this article: <https://doi.org/10.1080/10255842.2023.2300685>



Published online: 18 Jan 2024.



Submit your article to this journal 



View related articles 



View Crossmark data 



Computer-assisted interpretation, in-depth exploration and single cell type annotation of RNA sequence data using k-means clustering algorithm

Pranshu Saxena^a, Amit Sinha^a and Sanjay Kumar Singh^b

^aDepartment of Information Technology, ABES Engineering College, Ghaziabad, India; ^bUniversity School of Automation and Robotics, Guru Gobind Singh Indraprastha University, Surajmal Vihar, Delhi, India

ABSTRACT

At now, the majority of approaches rely on manual techniques for annotating cell types subsequent to clustering the data obtained from single-cell RNA sequencing (scRNA-seq). These approaches require a significant amount of physical exertion and depend substantially on the user's skill, perhaps resulting in uneven outcomes and inconsistency in treatment. In this paper, we provide a computer-assisted interpretation of every single cell of a tissue sample, along with an in-depth exploration of an individual cell's molecular, phenotypic and functional attributes. The paper will also perform k-means clustering followed by silhouette validation based on similar phenotype and functional attributes, and also, cell type annotation is performed, where we match a cell's gene profile against some known database by applying certain statistical conditions. Finally, all the genes are mapped spatially on the tissue sample. This paper is an aid to medicine to know which cells are expressed/not expressed in a tissue sample and their spatial location on the tissue sample.

ARTICLE HISTORY

Received 21 September 2023
Accepted 24 December 2023

KEYWORDS

RNA sequencing; k-means; cluster-validation; cell; transcriptomics

1. Introduction

The genomic method of RNA sequencing, also known as RNA-seq, is beneficial for analyzing cellular responses. This method enables the detection and quantitative analysis of messenger RNA molecules present in a biological sample. RNA-seq has become a major driver of medical innovation and discovery in recent years. Because of the practical considerations involved, the method is often applied to samples that contain thousands to millions of cells. This has prevented a direct evaluation of the cell, the basic building block of biology. Since the first one was published in 2009, numerous single-cell RNA-sequencing (scRNA-seq) investigations have been conducted, the majority by specialized labs with expertise in single-cell genomics, bioinformatics, and computation.

Single-cell Spatial Transcriptomics [single-cell RNA sequencing] is a compound of three keywords Spatial¹, Transcript², and omics³ (Ståhl et al. 2016; Maynard et al. 2021; Williams et al. 2022).

First, 'omics' refers to the collective characterization and quantification of pools of biological molecules. These pools of molecules collectively translate into an organism's structure, functions, and dynamics.

The second is ²Transcript, meaning an RNA strand is produced when a gene is transcribed, i.e. duplicating a gene's DNA sequence to synthesize an RNA molecule (Cao et al. 2020). These transcriptions follow a series of procedure (Lacar et al. 2016; Svensson et al. 2018).

Step 1: The **initiation** phase of transcription entails using RNA polymerase, the principal enzyme responsible for synthesizing a complementary RNA strand from a single-stranded DNA template. The RNA polymerase interacts with a specific region of the DNA molecule that is close to the initiation site of a gene. In bacteria, it is seen that each gene possesses its own promoter, which is accompanied by a collection of co-transcribed genes. Following the process of binding, RNA polymerase effectively separates the DNA strands, resulting in the formation of individual single strands. This action establishes the necessary template for the subsequent transcription process.

Step 2: The process of **elongation** in RNA polymerase involves the utilization of a single DNA strand, referred to as the template strand, for the purpose of serving as a template. The RNA transcript employs the nucleotide uracil (U) instead of thymine

(T) yet includes the same information as the non-template (coding) strand of DNA.

Step 3: Terminator sequences indicate the completion of the RNA transcript. They induce the release of the transcript from the RNA polymerase after its transcription.

Not all genes exhibit continuous transcription. In contrast, the regulation of gene transcription occurs separately for each gene or, in the case of bacteria, for small clusters of related genes. Cells have precise control over transcription, selectively activating only those genes whose resulting products are necessary during a particular moment.

Moreover, **finally**, ³Spatial means assigning RNA to a location onto a histopathological sample. Identifying the location is desirable to know where some stuff is happening, and doctors can devise the desired therapy only for specific cells, not for entire regions (Chen et al. 2020; Chen et al., 2015; Lee et al. 2021; Petukhov 2020). Conclusively, the analysis of all of the RNA molecules present inside a cell is referred to as the transcriptome.

Cell describes the molecular, phenotypic, or functional attributes of an individual. Cell type annotation is the process of assigning or identifying the specific cell types or identities present in a biological sample based on gene expression patterns (Armingol et al. 2021).

DNA sequencing is the same in almost all cells of a given type of organism. These sequences generally consist of different cells within that organism. The DNA sequence of an organism's genome contains the instructions necessary for the development, functioning and maintenance of that organism. Every cell in a mouse or human body contains the same set of genes with the same DNA sequence. However, a cell may have specific (combination of) genes that are either expressed (turn-on) or not expressed (turn-off), making them perform their specific function expression or gene regulation (Payne et al. 2021). Different cell type has specific gene expression profile. Cellular diversity and cell-specific function are best assessed not at the DNA level but at the protein level/gene level (Tang et al. 2009; Sasagawa et al. 2013). So, these DNA are transcribed into RNA for further gene expression analysis and to identify/locate the region where genes are undesirably triggered (Chen et al., 2015).

Many studies of the transcriptome only concentrate on messenger (m)RNA molecules, which are responsible for reflecting the genes that are being actively expressed (as protein structure) in a cell or tissue at a

specific time or in a specific environment (Tang et al. 2009). RNA can provide insight into whether or not genes are expressed in a given cell. RNA molecules are synthesized from a DNA template. Analyzing RNA in a cell can provide information on which genes are actively being transcribed and, hence, expressed to perform the intended task. RNA sequencing allows researchers to determine the identity and abundance of different RNA molecules, including messenger-RNA (mRNA) transcript. By comparing the RNA-Seq data across different samples or conditions, researchers can identify which genes are unregulated (turned on) or downregulated (turned off) under specific circumstances (Femino et al. 1998; Tang et al. 2009; Ke et al. 2013; Haque et al. 2017; Lopez et al. 2018; Xu et al. 2023).

A thorough comprehension of how particular cells utilize their mRNA and proteins in various tissues of the human body can yield novel approaches for preventing or treating infections, malignancies, neurological or metabolic diseases, and several other ailments.

This paper uses visium data by 10× genomics, which allows for the simultaneous profiling of gene expression and spatial information within intact tissue sections (Merritt et al. 2020). By preserving the spatial context of gene expression, Visium enables studying the spatial organization of cells and molecular interactions within complex tissues. This paper calculates what the various genes are expressed in a specific cell. These identified genes help us to know the cell type and interpret the genes, followed by visualization of a gene at which part of the sample it is activated. This paper also counts the optimal number of clusters based on elbow methods and the validity of the number of clusters calculated by silhouette clustering. A cluster consisting of similar kinds of cells, these cells are annotated with the help of two publicly available datasets. This cell annotation procedure is done based on the top 20 genes identified in a cell. Later, these annotated cells are transferred spatially to tissue samples. Based on cell type, spatial projection of the disease can be predicted.

2. Proposed methodology

The proposed methodology shown in Figure 1 follows a series of procedures starting from reading the visium tissue data at 10× resolution with spots coloured by UMI count. This data contains 2987 spots/cells; each cell has 31053 genes (Kleshchevnikov et al. 2022). In the next step, the highest expressed genes

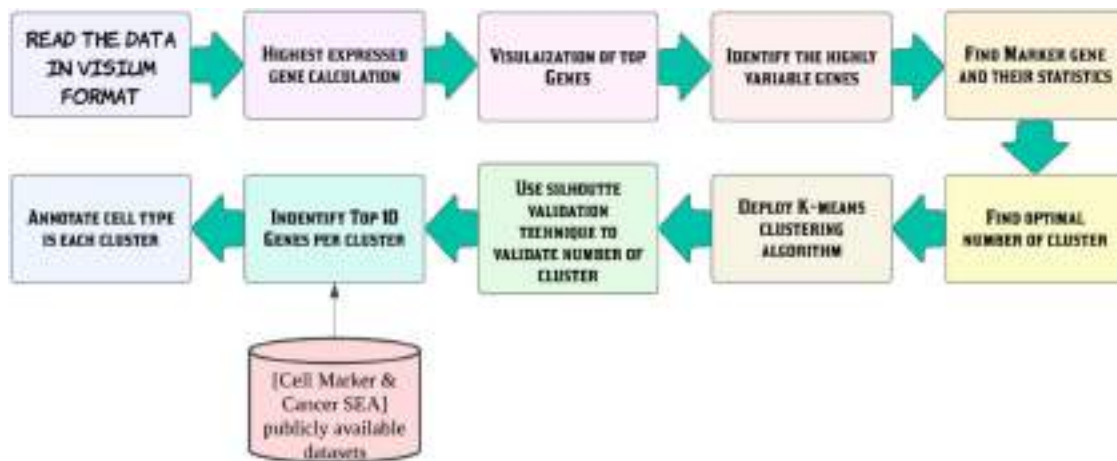


Figure 1. Graphical abstract of the proposed system.

are calculated in each cell. These highest expressed genes are visualized on top of the tissue data by generating a spatial plot of cells using the gene expression levels of the most abundant gene. In the third step, highly variable genes and markers are identified, and their statistics (log-fold-changes, z-scores, *p*value) are calculated. Based on these statistical values, the optimal number of clusters is identified in the fourth step. The k-means clustering is used to cluster similar kinds of cells into specific clusters. The Silhouette validation technique is used to validate the number of clusters. In the next step, the top 10 genes are identified per cluster. These top 10 genes are matched with two public data sets (cell marker, & cancer SEA) (Cao et al. 2020) to annotate cells in the cluster, and finally, these annotated cells are mapped based on spatial location on the tissue sample.

2.1. Data collection

This paper uses a 10× single nucleus RNA-Sequencing (scRNA-Seq) and visium spatial transcriptomic data generated from adjacent mouse brain tissue sections (Kleshchevnikov et al. 2022). These public available data sets [ST8059048], [ST8059049], [ST8059050], [ST8059051], and [ST8059052] freely accessible at https://cell2location.cog.sanger.ac.uk/tutorial/mouse_brain_visium_wo_cloupe_data.zip.

This visium data from genome 10× follows a series of procedures. In the *first* step, a visium slide containing a matrix of capture areas is prepared. The slide has a unique barcode pattern assigned to each capture area. Then, in the *second* step, tissue of interest, such as a section from a biopsy or a tissue slide, is placed onto the visium slide, aligning it with the capture area. In the *third* step, the tissue on the visium slide is permeabilized to allow the capture of RNA

molecules. RNA transcript from the tissue binds to capture areas while other cellular components are washed away. Later in the *fourth* step, reverse transcription is performed within each capture area to convert the captured RNA molecules into complementary DNA (cDNA). This step allows for amplification and preservation of the original RNA information. Library generation is the *fifth* step. The cDNA molecules are amplified and prepared into sequencing libraries in this step. These libraries can be subjected to high-throughput sequencing generation data that can be analyzed to determine the spatial gene expression patterns within the tissue sample. *Finally*, these libraries are subjected to high-throughput sequencing using next-generation sequencing. The Data (2987 × 31053) contains 2987 cells/spots, and each spot having 31053 number of genes.

2.2. Highest expressed gene calculation

This process calculates the percentage of counts attributed to each gene in a cell for each gene. The boxplots represent the *n*_{top} genes with the highest mean fraction across all cells. Figure 2 shows mitochondrial genes (mt-Co3, mt-Co1, mt-Atp6, mt-Co2), Protine coding gene like Fth1, and Ttr gene, which responsible for transthyretin protine instruction are what we anticipate to see. A few spike-in transcripts might also exist in this region, but if all of the spike-ins are among the top 25, it might have added too much spike-in RNA. A large number of anticipated or pseudo-genes may point to alignment issues.

All the genes can be visualized on the tissue sample. For example, in the next section, we have shown 3 genes and their spatial location on the tissue sample.

2.3. Spatial visualization of top genes

Among the top 25 genes identified in the previous step, three chosen genes (*mt-Co3*, *Ttr*, and *Fth1*) are spatially located in Figure 3. As the result shows, *mt-Co3* is the most prominent gene. This gene is active in most of the cells. While *Fth1* and *Ttr* genes are triggered in some of the specific cells and spatially located in Figure 3. This spatial location can help neurologists access and correlate the situation with their wisdom.

Identifying the variability among genes is also essential to exhibit diverse expression.

2.4. Identify the highly variable gene

An essential stage in processing single-cell RNA sequencing (scRNA-seq) data is the identification of highly variable genes shown in Figure 4. It aids in identifying genes that show notable expression variation between cells, which may be a sign of biological

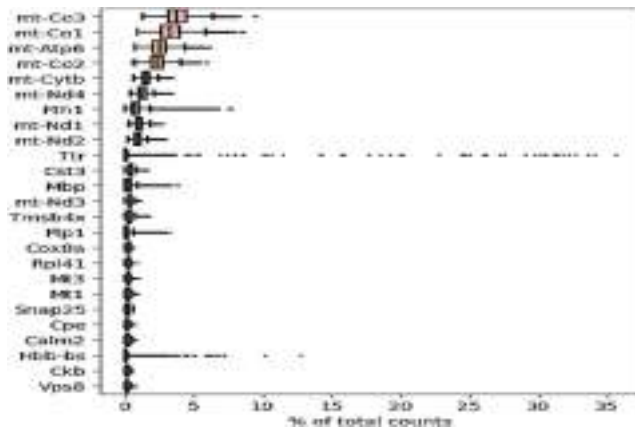


Figure 2. Top-25 expressed genes.

heterogeneity, a marker for a particular cell type, or a gene involved in important regulatory functions.

Identification of the cell type: Highly variable genes frequently exhibit diverse expression patterns in various cell types. We can discover gene signatures that distinguish particular cell types and utilize them as markers for cell type identification by identifying these genes.

Although scRNA-seq datasets can contain thousands of genes, not all significantly contribute to the underlying biological variance. By locating highly variable genes, we can narrow the attention to the most useful genes, decreasing the dataset's complexity and enhancing subsequent analysis like grouping and visualization.

Differential expression analysis shows that highly variable genes are more likely to show notable variations in expression under various settings or in different cell states. We can prioritize these genes for additional research and identify the genes responsible for biological variation. Highly variable genes are frequently linked to important biological functions, such as signalling networks, cell cycle regulation, and important transcription factors. Finding these genes can illuminate the biological underpinnings and regulatory systems underlying cellular heterogeneity.

3. Genes data and their statistics

Later, all the genes are quantified based on quantitative metrics (*Log₂ fold changes*, *z-score*, and *P-values*). Table 1 shows a sub-sample of 10 genes among the 31053 genes. This quantification can help access the different cell types and their nature in the tissue sample.

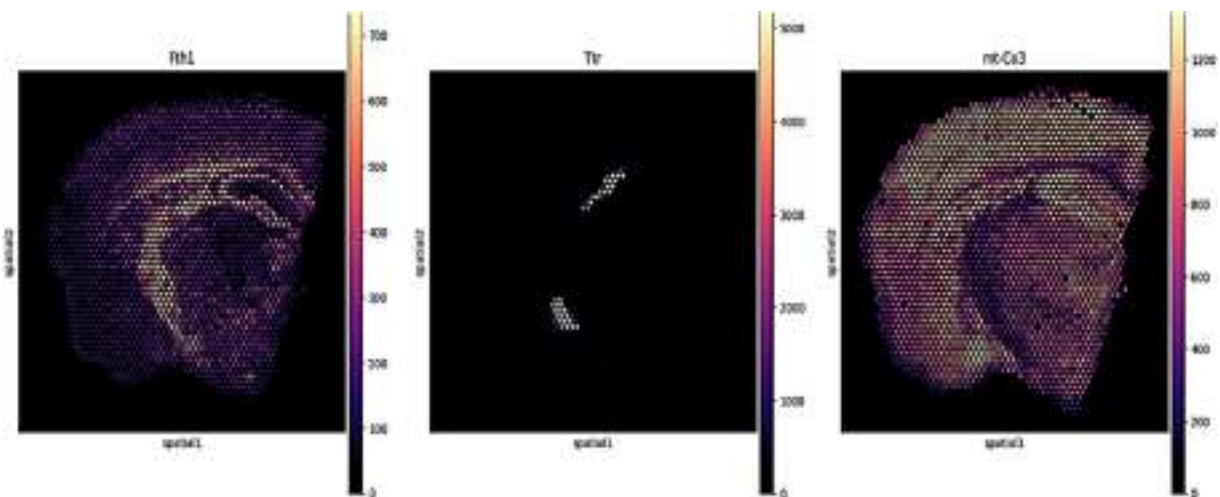


Figure 3. Spatial distribution 3-genes from top 25 expressed genes over sample tissue.

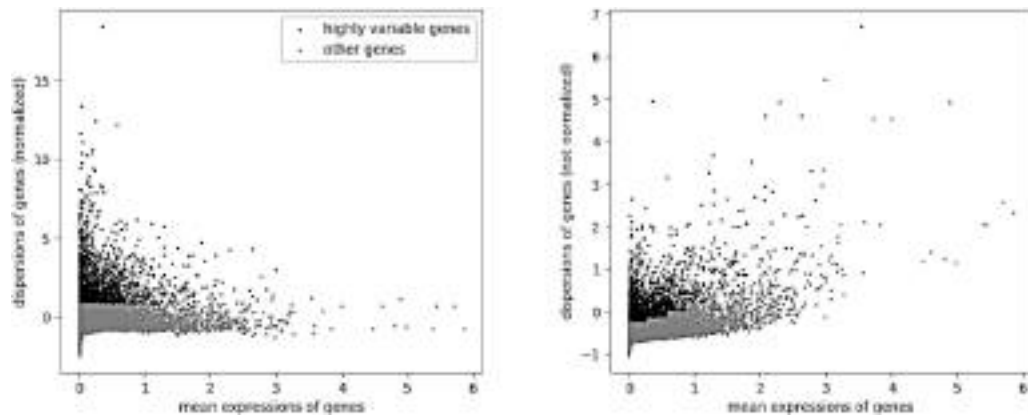


Figure 4. Distribution of highly variable genes.

Table 1. A small sample of 10 genes among 31053 genes: Statistical data.

	Name of Gene	Log fold changes	z - score	P values
0	1110008P14Rik	1.8983871	41.56318	6.24E-193
1	3110035E14Rik	2.8234143	41.05083	4.46E-156
2	Diras2	1.9230663	34.813873	9.21E-139
3	Slc17a7	1.8118644	33.838364	3.99E-179
4	Nrgn	2.111198	33.586037	1.69E-184
5	Ttc9b	2.008768	33.03382	1.05E-127
6	Egr1	2.0530055	33.009403	2.76E-147
7	Hs3st2	3.3050497	31.111254	1.72E-100
8	Prkcb	1.4074348	30.980652	1.44E-147
9	Ccl27a	1.7982706	30.780811	7.96E-115
10	Ier5	2.2711916	28.888874	1.30E-102

On this data preprocessing, differential analysis is done. Moreover, using PCA, k-means clustering techniques, with silhouette coefficient analysis applied to the reduced dataset.

3.1. Find an optimal number of clusters

The Elbow method is a frequent heuristic in mathematical optimization used to find a point at which the decreasing returns are no longer worth the additional cost. This point is determined by selecting a cutoff point (Na et al. 2010; Jin and Han 2011).

Based on the data gathered, the first step is finding an optimal number of clusters to represent data into different cell types. That will eventually help to map cell type in cluster. Figure 5 shows a graph between the with-in-cluster sum of squares vs. the number of clusters. With this graph, it can be inferred that eight can be an optimal number, considering that k-means clustering is discussed in the next section.

3.2. Deploy k-mean clustering algorithm

Based on the number of clusters identified in the previous step, it is used to perform k-means clustering

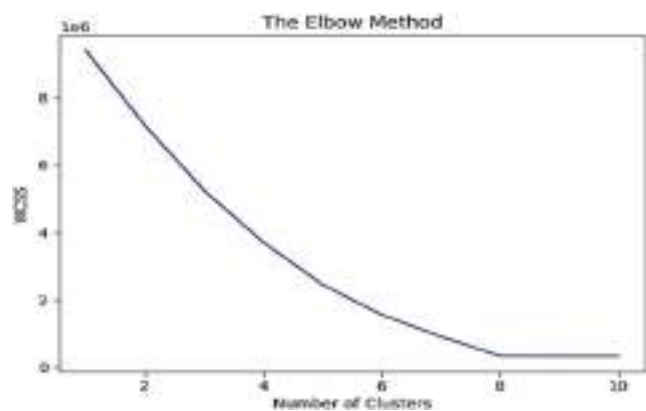


Figure 5. The process of identifying the optimal number of clusters.

on the data. k-means divide the data into eight different clusters with their silhouette coefficient value. Clustering is done based on similar kinds of gene scores. That means each cluster has similar cells that have the same gene type (Na et al. 2010; Jin and Han 2011; Yu et al. 2021).

When the clustering is performed on the basic 2 to 10 clusters, the silhouette coefficient varies from 0.3729 to 0.7042. Silhouette score is maximized when the number of cluster values is 8 (silhouette coefficient = 0.7717) (Wang and Xu 2019) and decreases when the number of clusters exceeds 8. The visualization of cluster data and silhouette plots for various Clusters is displayed in Figure 6.

3.3. Silhouette validation technique to validate the number of clusters

The silhouette value is a metric that quantifies the similarity of an object to other clusters, specifically in terms of its resemblance to its own cluster relative to

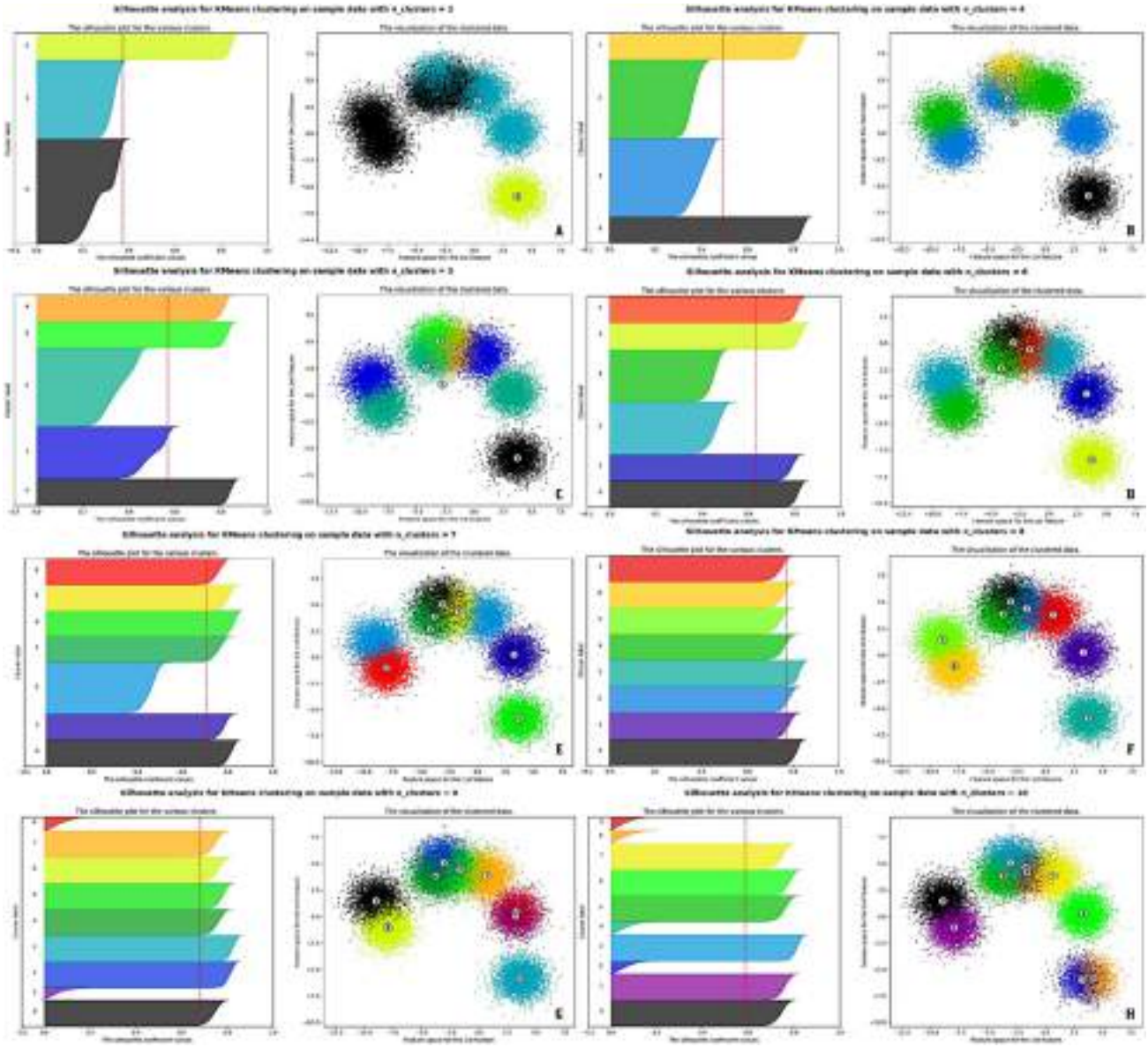


Figure 6. Experimental results considering different numbers of clusters along with the silhouette coefficient values.

its dissimilarity from other clusters (Wang and Xu 2019; Shahapure and Nicholas 2020).

The silhouette coefficient ranges from -1 to $+1$, where a higher number suggests a strong correspondence between the object and its assigned cluster. At the same time, a lower value indicates a weak correspondence with nearby clusters. The selected clustering structure is deemed appropriate when a significant proportion of the objects has a considerable value. If a considerable proportion of data points exhibit negative or low values, the clustering arrangement may suffer from an excessive or inadequate number of clusters (Wang and Xu 2019).

This paper is validating from a number of Cluster (K) 2 to 10 using k-means clustering, and for every K

cluster, the silhouette score is calculated and displayed in Figure 7.

These values are calculated by considering D_i is the data point belonging to cluster C_l . Then, the similarity coefficient for the data point $D_i \in C_l$ is expressed as Avg_{wcss} (within-cluster sum of square distance)

$$Avg_{wcss}(D_i) = \frac{1}{|C_l| - 1} \sum_{j \in C_l, i \neq j} dist(D_i, D_j) \quad (1)$$

Let $Avg_{wcss}(D_i)$ be the distance between the data point D_i with all other data points in the same Cluster, Euclidian distance is the measured distance between data points (D_i, D_j) , and $|C_l|$ will keep track

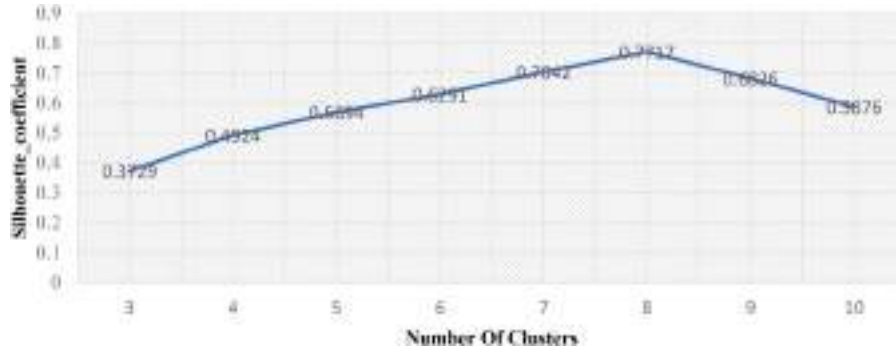


Figure 7. Number of clusters along with silhouette coefficient score.

Table 2. Top-10 genes annotated with cell types and represented cluster.

Cluster 1	Top-10 Genes	Cluster 2	Top-10 Genes	Cluster 3	Top-10 Genes	Cluster 4	Top-10 Genes
Cell Name: Brush cell (Tuft cell)	Prkcd Adarb1 Tcf7l2 Cccl336 Cnih2 Zbtb18 Wipf3 Hpca Psd Chn1	Cell Name: Cardiomyocyte	Camk2n1 Nrgn Atp1a1 Egr1 Mef2c Ppp3ca Atp2b2 Camk2a Lingo1 Lamp5	Cell Name: Ciliated cell	1110008P14Rik Mobp Fth1 Cryab Cldn11 Ly6h Nptxr Hpcal4 Nnat Id3	Cell Name: Mesenchymal stem cell	Tcf7l2 Slc17a6 Tmsb10 Cbln1 Cbln4 Gap43 Nap1l5 Sparc Nxph1 Zcchc12
Cluster 5	Top-10 Genes	Cluster 6	Top-10 Genes	Cluster 7	Top-10 Genes	Cluster 8	Top-10 Genes
Cell Name: Neuroendocrine cell	Sparc Nap1l5 Slc6a11 Peg3 Resp18 Nptxr Syn2 Snca Nov Ddn	Cell Name: Stem cell	Cartpt Gpx3 Scg2 Nrsn2 Sparc Nap1l5 Foxb1 Podxl2 Ndn Peg3	Cell Name: T cell	Ppp1r1b Penk Tmem158 Gpr88 Meis2 Gad2 Adcy5 Arpp21 Gng7 Slc32a1	Cell Name: Type I spiral ganglion neuron	Pvalb Gad1 Gad2 A1593442 Kif5a Slc32a1 Ramp3 Cplx1 Six3 Ubash3b

of the number of points in the cluster C_I . Eq. (1) provides how well D_i is clustered.

If we talk about how the data point D_i is separated from other cluster then dissimilarity of data point D_i to some other cluster C_j as the mean of the distance from D_i to all other points in C_j .

$$Avg_{dissimilarity} = \min_{j \neq I} \cdot \frac{1}{|C_j|} \sum_{j \in C_j} dist(D_i, D_j) \quad (2)$$

For each data point $D_i \in C_I$, $Avg_{dissimilarity}$ to be the smallest mean distance of I to all points in any other cluster.

With the help of these formulations, the silhouette coefficient can be estimated for the data point D_i

$$Sil(D_i) = \frac{Avg_{wcss}(D_i) - Avg_{dissimilarity}}{\max\{Avg_{wcss}(D_i) - Avg_{dissimilarity}\}}, \text{ if } |C_I| > 1$$

$$Sil(D_i) = 0, \text{ if } |C_I| = 1 \quad (3)$$

After formulating all 8 clusters, the next step is to annotate them with cell type to identify the spatial

location of cells and their corresponding top-10 genes.

3.4. Identify the top genes per cluster and annotate cell type in each cluster

Based on gene scores, eight different clusters are identified later, and these clusters are annotated as cell types. Table 2 shows cluster numbers along with their Cell name, and Top-10 represented genes. The names of 8 cells mapped with top-10 genes are Brush cell, Cardiomyocyte, Ciliated cell, Mesenchymal cell, Neuroendocrine cell, Stem cell, T cell, and Type I spiral ganglion neuron. These cells are further mapped with their spatial location in Figure 8. This spatial location helps the clinical expert to verify the relationship among expressed cells/unexpressed cells based on spatial location and can propose a future course of treatment and planning.

Table 3 shows the number of cells per cluster. This table provides a glimpse into the cluster along with the number of cells per cluster with cell type.

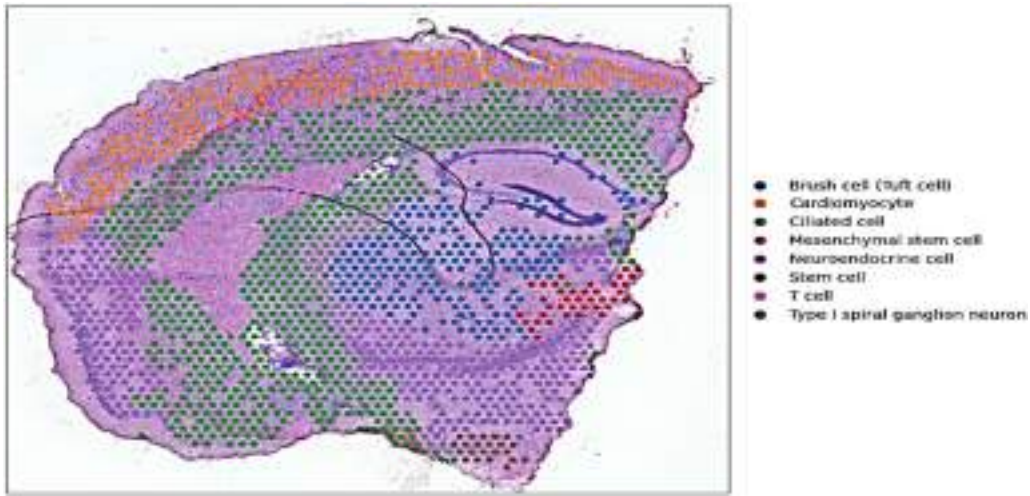


Figure 8. The spatial location of each cell type over the tissue sample.

Table 3. Cluster-wise number of cells.

Cluster	Number of Cells	Cell Type
Cluster_1	3840	Brush Cell (Tuft cell)
Cluster_2	1960	Cardiomyocyte
Cluster_3	7834	Ciliated Cell
Cluster_4	40	Mesenchymal stem cell
Cluster_5	5890	Neuroendocrine cell
Cluster_6	22	Stem cell
Cluster_7	1260	T cell
Cluster_8	4989	Type I spiral ganglion neuron

Figure 8 shows the spatial mapping of each identified cell on the tissue sample. In assigning to the cell type of each cluster, for example, we have eight different clusters, each of which comes with a gene profile. These profiles match against two publically available datasets (Cell Marker (Zhang et al. 2019), Cancer SEA (Yuan et al. 2019)) to call this type of cell. List of more publically available data set is projected in Table 4.

For cell type annotation, we match the pattern against the database by applying certain statistical conditions.

In order to benchmark the cell annotation process against the other cell-type annotation methods, we utilized five scRNA-seq datasets from public domain and re-analysed these using automated system. Five datasets, Mouse Lung (GSE63269) Mouse Brain (ST8059048), (ST8059049), (ST8059050) and (ST8059052). To make the compare unbiased, we annotate cell type with the help of public data set available in the Table 4 and find that annotation performed with different dataset provide similar cells types for ST8059052 dataset of mouse brain.

3.5. Gene set Enrichment analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a computational approach utilised to assess the statistical

significance of concordant differences between two biological states in relation to a pre-defined set of genes. Steps involved in this method to overcome the problem of Over Representation Analysis (ORA) methods. Actually, the big difference (ORA and GSEA) is that in GSEA the input is not a list of genes but a ranked gene.

Ranking basically means that the genes are ranked by some score so common way of ranking genes by level of differential expression like p value and \log_2 -fold. The p -value records about the significance of changes while \log_2 fold changes talk about the direction and the strength of the change basically if the genes are operation up regulated or down regulated. So, combination of both give ranked of list of genes which orders them both not only by significance but also direction of genes. At the top of the list, we have most up regulated and significant genes and at the bottom of the list most down regulated significance genes. Ranking of genes is calculate by the Eq. (4);

$$Gene_{ranking} = Sign (Fold_Changes)^* - \log_{10}(P_{value}) \quad (4)$$

3.6. Pathway analysis ST8059052 mouse brain GeneSet mouse MSigBD collection

The Mouse Molecular Signatures Database (MSigDB) has a total of 16,090 gene sets, which are categorized into six primary collections along with other subcollections. M8 collection: cell type signature gene sets are mapped with ST8059052 dataset.

The gene sets encompassing cluster marker genes for cell types that have been found in single-cell sequencing investigations of mouse tissue. The purpose of these gene sets is to aid in the categorization

Table 4. Publicly available dataset for annotating cell types.

References	Public Dataset	Dataset Type	Species	Location/parts
(Regev et al. 2017)	HCA	scRNAseq	Human	33 multi-organs data
(Han et al. 2018)	MCA	scRNAseq	mice	98 multi-organs dataset
(Schaum et al. 2018)	Tabula Muris	scRNAseq	mice	20 multi-organs dataset and tissue sample
(Hodge et al. 2019)	Allen Brain Atlas	scRNAseq	Both human and mice	69 neuronal cell types
(Zhang et al. 2019)	CellMaker	Genes marker	Both human and mice	467 (human), 389 (mice)
(Franzén et al. 2019)	PanglaoDB	Genes marker	Human	155 cell types
(Yuan et al. 2019)	CancerSEA	Genes marker	Human	14 cancer functional states

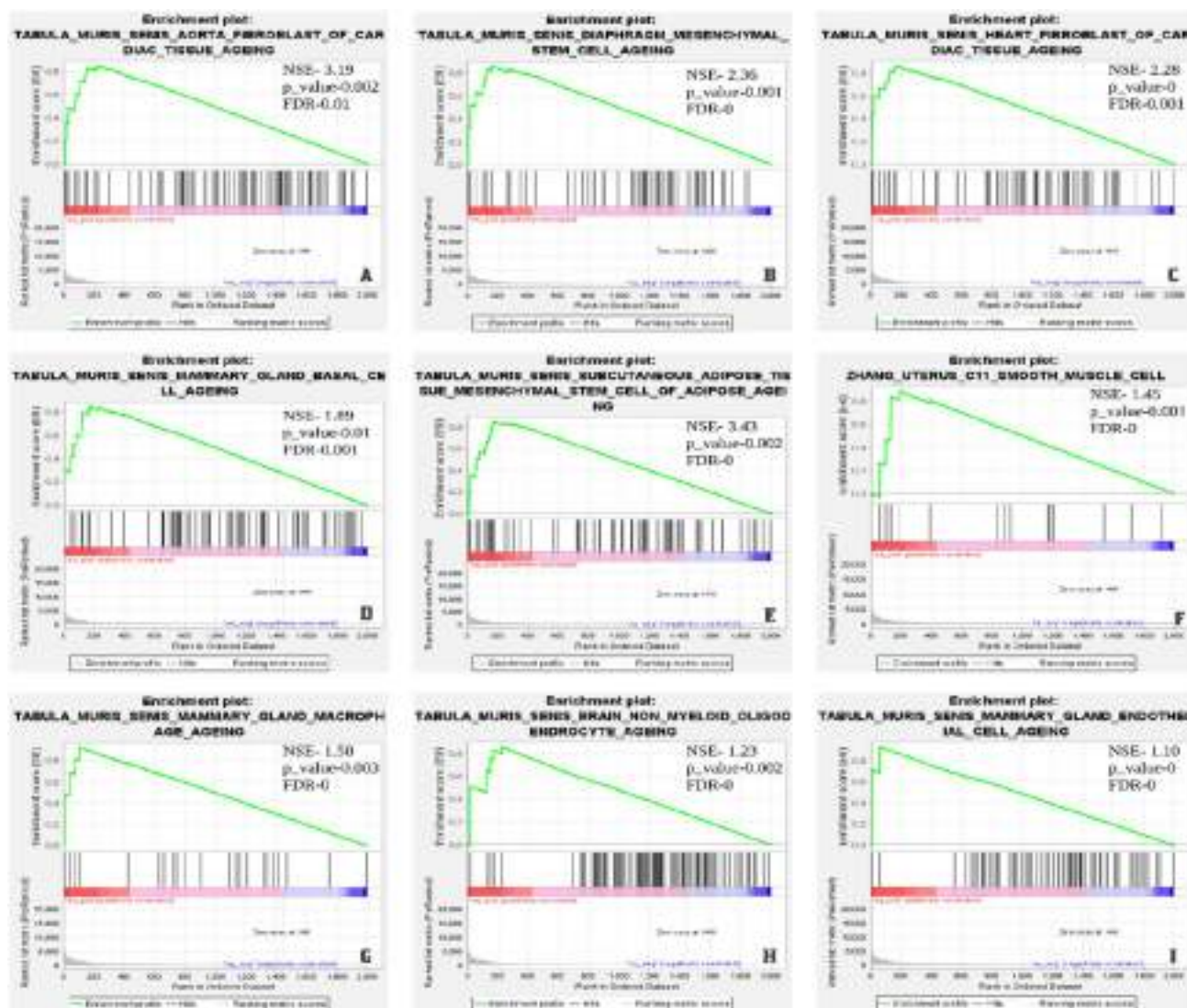


Figure 9. Gene set enrichment analysis of mouse brain sample correlated with GeneSet mouse MSigBD collection. Each image from [A–I] shows gene set enriched pathways of mouse brain sample associated gene with normalized enrichment score (NES), probabilistic values (p -value), and false discovery rate (FDR) inside each diagram.

of cell types within datasets, particularly those derived from investigations involving the development of organoid models.

As figure shows mouse brains sample (ST8059052) shows a strong relation with Subcutaneous_adipose_tissue_mesenchymal (Figure 9E) beside others Aorta_fibroblast_of_cardiac (Figure 9A), and Diaphragm_mesenchymal (Figure 9B). On the

other hand, we have also observed a slightly dip positive correlation with Heart_fibroblast_of_cardiac (Figure 9C), Mammary_gland_basal (Figure 9D), and Zhang_Uterus_C11_Smooth_Muscle (Figure 9F). Rest Mammary_gland_macrophage (Figure 9G), Sinus_brain_non_myeloid (Figure 9H), and mammary_gland_endothelial (Figure 9I), are relatively shows lower side of correlation.

Single-cell Specificity Analysis (SSCA) is a library that allows for accurate cell type annotation by comparing scRNA-Seq data to reference cell type profiles. SSCA calculates a specificity score for each cell type, measuring the likelihood of a cell belonging to a specific type based on its gene expression profile. SSCA library includes a pre-built reference database for various organisms, enabling cell-type annotation in different biological contexts.

4. Discussion and conclusion

Currently, annotating cell types inside cell clusters resulting from unsupervised clustering for single-cell RNA sequencing (scRNA-seq) data is predominantly carried out manually. Due to the inherent constraints associated with the manual methodology, it is unfeasible to provide annotation results that exhibit high quality, reproducibility, and standardization across the expanding array of single-cell RNA sequencing datasets. Identifying location is crucial for comprehending the ongoing events, as healthcare practitioners are limited to administering medication exclusively to specific cells rather than an entire region.

Therefore, this automation helps medical professionals measure the perceptual problems they encounter while diagnosing the intricate architecture of tissue samples. This analysis, based on k-means clustering, assists medicine in its interpretation and in-depth examination of the many cells in tissue samples. A comprehensive understanding of how individual cells in the many tissues of the human body use the mRNA and proteins they produce could lead to the development of new strategies that can be used to prevent or treat a wide variety of conditions, such as infections, malignancies, neurological or metabolic abnormalities, and a source of other diseases. These conditions include infections, malignancies, and neurological or metabolic abnormalities.

The main objective of this study is to discover the pathways and processes that exhibit a substantial association with the activity of the regulatory factor. The approach employed in this study involves the utilisation of a hypergeometric test to establish associations between genes and their corresponding regulatory regions. This enables the inference of proximal gene regulatory domains.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

- Armingol E, Officer A, Harismendy O, Lewis NE. 2021. Deciphering cell-cell interactions and communication from gene expression. *Nat Rev Genet.* 22(2):71–88. doi: [10.1038/s41576-020-00292-x](https://doi.org/10.1038/s41576-020-00292-x).
- Cao Y, Wang X, Peng G. 2020. SCSA: a cell type annotation tool for single-cell RNA-seq data. *Front Genet.* 11: 490. doi: [10.3389/fgene.2020.00490](https://doi.org/10.3389/fgene.2020.00490).
- Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. 2015. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science.* 348(6233):aaa6090. doi: [10.1126/science.aaa6090](https://doi.org/10.1126/science.aaa6090).
- Chen W-T, Lu A, Craessaerts K, Pavie B, Sala Frigerio C, Corthout N, Qian X, Laláková J, Kühnemund M, Voytyuk I, et al. 2020. Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell.* 182(4): 976–991 e19. doi: [10.1016/j.cell.2020.06.038](https://doi.org/10.1016/j.cell.2020.06.038).
- Femino AM, Fay FS, Fogarty K, Singer RH. 1998. Visualization of single RNA transcripts in situ. *Science.* 280(5363):585–590. doi: [10.1126/science.280.5363.585](https://doi.org/10.1126/science.280.5363.585).
- Franzén O, Gan LM, Björkegren JLM. 2019. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database.* 2019:1–9. doi: [10.1093/database/baz046](https://doi.org/10.1093/database/baz046).
- Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F, et al. 2018. Mapping the mouse cell atlas by microwell-seq. *Cell.* 172(5):1107.e17. doi: [10.1016/j.cell.2018.02.001](https://doi.org/10.1016/j.cell.2018.02.001).
- Haque A, Engel J, Teichmann SA, Lönnberg T. 2017. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* 9(1):75. doi: [10.1186/s13073-017-0467-4](https://doi.org/10.1186/s13073-017-0467-4).
- Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, Close JL, Long B, Johansen N, Penn O, et al. 2019. Conserved cell types with divergent features in human versus mouse cortex. *Nature.* 573(7772):61–68. doi: [10.1038/s41586-019-1506-7](https://doi.org/10.1038/s41586-019-1506-7).
- Jin X, Han J. 2011. K-means clustering. In: Sammut C, Webb GI, editors. *Encyclopedia of machine learning*. Boston (MA): Springer. doi: [10.1007/978-0-387-30164-8_425](https://doi.org/10.1007/978-0-387-30164-8_425).
- Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, Nilsson M. 2013. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods.* 10(9):857–860. doi: [10.1038/nmeth.2563](https://doi.org/10.1038/nmeth.2563).
- Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, Elmentaite R, Lomakin A, Kedlian V, Gayoso A, et al. 2022. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol.* 40(5):661–671. doi: [10.1038/s41587-021-01139-4](https://doi.org/10.1038/s41587-021-01139-4).
- Lacar B, Linker SB, Jaeger BN, Krishnaswami SR, Barron JJ, Kelder MJE, Parylak SL, Paquola ACM, Venepally P, Novotny M, et al. 2016. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat Commun.* 7(1):11022. doi: [10.1038/ncomms11022](https://doi.org/10.1038/ncomms11022).

- Lee Y, Bogdanoff D, Wang Y, Hartoularos GC, Woo JM, Mowery CT, Nisonoff HM, Lee DS, Sun Y, Lee J, et al. 2021. XYSeq: spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. *Sci Adv.* 7(17):eabg4755. doi: [10.1126/sciadv.abg4755](https://doi.org/10.1126/sciadv.abg4755).
- Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nat Methods.* 15(12):1053–1058. doi: [10.1038/s41592-018-0229-2](https://doi.org/10.1038/s41592-018-0229-2).
- Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, Tran MN, Besich Z, Tippani M, et al. 2021. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci.* 24(3):425–436. doi: [10.1038/s41593-020-00787-0](https://doi.org/10.1038/s41593-020-00787-0).
- Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, Hoang M, Jung J, Liang Y, McKay-Fleisch J, et al. 2020. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat Biotechnol.* 38(5):586–599. doi: [10.1038/s41587-020-0472-9](https://doi.org/10.1038/s41587-020-0472-9).
- Na S, Xumin L, Yong G. 2010. Research on k-means clustering algorithm: An improved k-means clustering algorithm. 2010 Third international symposium on intelligent information technology and security informatics, Jian, China, pp. 63–67. doi: [10.1109/IITSI.2010.74](https://doi.org/10.1109/IITSI.2010.74).
- Payne AC, Chiang ZD, Reginato PL, Mangiameli SM, Murray EM, Yao C-C, Markoulaki S, Earl AS, Labade AS, Jaenisch R, et al. 2021. In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science.* 371(6532):eaay3446. doi: [10.1126/science.aay3446](https://doi.org/10.1126/science.aay3446).
- Petukhov V. 2020. Bayesian segmentation of spatially resolved transcriptomics data. *bioRxiv.* 2020. doi: [10.1101/2020.10.05.326777](https://doi.org/10.1101/2020.10.05.326777).
- Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al. 2017. Science forum: the human cell atlas. *Elife.* 6:1–30. pp. doi: [10.7554/eLife.27041](https://doi.org/10.7554/eLife.27041).
- Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, Ueda HR. 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* 14(4):R31. doi: [10.1186/gb-2013-14-4-r31](https://doi.org/10.1186/gb-2013-14-4-r31).
- Schaum N, Karkanas J, Neff NF, May AP, Quake SR, Wyss-Coray T, et al. 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature.* 562:367–372. doi: [10.1038/s41586-018-0590-4](https://doi.org/10.1038/s41586-018-0590-4).
- Shahapure KR, Nicholas C. 2020. Cluster quality analysis using silhouette score. 2020 IEEE 7th international conference on data science and advanced analytics (DSAA), Sydney, NSW, Australia, pp. 747–748. doi: [10.1109/DSAA49011.2020.00096](https://doi.org/10.1109/DSAA49011.2020.00096).
- Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, et al. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science.* 353(6294):78–82. doi: [10.1126/science.aaf2403](https://doi.org/10.1126/science.aaf2403).
- Svensson V, Vento-Tormo R, Teichmann SA. 2018. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc.* 13(4):599–604. doi: [10.1038/nprot.2017.149](https://doi.org/10.1038/nprot.2017.149).
- Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods.* 6(5):377–382. doi: [10.1038/nmeth.1315](https://doi.org/10.1038/nmeth.1315).
- Wang X, Xu Y. 2019. An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conf Ser: mater Sci Eng.* 569(5):052024. doi: [10.1088/1757-899X/569/5/052024](https://doi.org/10.1088/1757-899X/569/5/052024).
- Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. 2022. An introduction to spatial transcriptomics for biomedical research. *Genome Med.* 14(1):68. doi: [10.1186/s13073-022-01075-1](https://doi.org/10.1186/s13073-022-01075-1).
- Xu J, Xu J, Meng Y, Lu C, Cai L, Zeng X, Nussinov R, Cheng F. 2023. Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. *Cell Report Methods.* 3(1):100382. doi: [10.1016/j.crmeth.2022.100382](https://doi.org/10.1016/j.crmeth.2022.100382). PMID:36814845;PMCID:PMC9939381.
- Yu W, Mahfouz A, Reinders MJT. 2021. CBA: cluster-guided batch alignment for single cell RNA-seq. *Front Genet.* 12:644211. doi: [10.3389/fgene.2021.644211](https://doi.org/10.3389/fgene.2021.644211).
- Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, Xu L, Luo T, Yan H, Long Z, et al. 2019. CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.* 47(D1):D900–D908. doi: [10.1093/nar/gky939](https://doi.org/10.1093/nar/gky939).
- Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M, et al. 2019. Cell Marker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* 47(D1):D721–D728. doi: [10.1093/nar/gky900](https://doi.org/10.1093/nar/gky900).