OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae626

https://doi.org/10.1093/bib/bbae626

Problem Solving Protocol

CosGeneGate selects multi-functional and credible biomarkers for single-cell analysis

Tianyu Liu [1,2,‡], Wenxin Long [1,3,‡,§], Zhiyuan Cao [1,2,4,‡,§], Yuge Wang [1], Chuan Hua He [5], Le Zhang [5,6], Stephen M. Strittmatter [5,6,7], Hongyu Zhao [1,2,*]

[1]Department of Biostatistics, Yale University, New Haven, CT, 06520, United States
[2]Interdepartmental Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT, 06520, United States
[3]Department of Statistics, The Pennsylvania State University, University Park, PA, 16820, United States
[4]Program of Health Informatics, Yale University, New Haven, CT, 06520, United States
[5]Department of Neurology, Yale University School of Medicine, New Haven, CT, 06520, United States
[6]Department of Neuroscience, Yale University School of Medicine, New Haven, CT, 06520, United States
[7]Cellular Neuroscience, Neurodegeneration and Repair Program, Yale University School of Medicine, New Haven, CT, 06520, United States

*Corresponding author. Tel: 203.785.3613; Fax: 203.785.6912; E-mail: hongyu.zhao@yale.edu

‡Tianyu Liu, Wenxin Long and Zhiyuan Cao contributed equally to this work, and the orders can be exchanged.

§Parts of the work were finished during their internship at Yale University.

Availability: https://github.com/VivLon/CosGeneGate.

Abstract

Motivation: Selecting representative genes or marker genes to distinguish cell types is an important task in single-cell sequencing analysis. Although many methods have been proposed to select marker genes, the genes selected may have redundancy and/or do not show cell-type-specific expression patterns to distinguish cell types. Results: Here, we present a novel model, named CosGeneGate, to select marker genes for more effective marker selections. CosGeneGate is inspired by combining the advantages of selecting marker genes based on both cell-type classification accuracy and marker gene specific expression patterns. We demonstrate the better performance of the marker genes selected by CosGeneGate for various downstream analyses than the existing methods with both public datasets and newly sequenced datasets. The non-redundant marker genes identified by CosGeneGate for major cell types and tissues in human can be found at the website as follows: https://github.com/VivLon/CosGeneGate/blob/main/marker gene list.xlsx.

Keywords: marker genes; deep learning; single-cell sequencing; Alzheimer's disease

Introduction

Single-cell sequencing technologies offer high-throughput observations into complex biological systems at the cell level [1, 2], which help elucidate disease mechanisms and improve treatments [3–5]. These technologies enable the characterization of various molecules, such as DNA (scDNA-seq) [6], RNA (scRNA-seq) [7, 8], and proteins [9]. They can also facilitate epigenetic studies through single-cell ATAC sequencing (scATAC-seq) [10, 11] and methylation [12, 13]. There have also been rapid developments of spatial single-cell technologies [14]. These single-cell technologies have been rated as among the most impactful ones in recent years [4, 15].

Since single-cell sequencing allows us to study the properties of different cell types, the inference of cell types has become important [16]. One approach for cell-type annotation is based on marker genes of different cell types [17, 18]. Because scRNA-seq data are high dimensional, sparse and noisy [19], there is significant challenge in cell type assignment using marker genes. In addition, genes with similar biological functions tend to show similar expression patterns, further complicating marker gene selections.

Currently, researchers use three sources of information to select marker genes for cell type annotation. The first source is from experts, who select marker genes based on biological knowledge and prior experiments. However, such selections are subjective and lead to different marker gene sets [20–22]. The second source is through the identifications of differentially expressed genes between groups of cells [23, 24]. By performing the 'one-over-all' statistical test for pre-clustered data (using either Louvain [25] or Leiden [26]), these methods select marker genes based on p-values or fold changes (FCs) or both. However, these methods test the mean difference between the two groups of cells and ignore the effects of gene expression proportions, which may limit their utilities. Moreover, selecting marker genes based on marginal statistics often leads to a number of discovered genes with a high false-positive rate, as shown in Extended Data Fig. 1. The third source is through machine learning models [27, 28], where marker genes are selected based on interpretable machine learning models. A machine learning-based approach

Received: July 12, 2024. **Revised:** October 7, 2024. **Accepted:** November 14, 2024

uses a dataset with annotated cell types as the training set and trains a classifier with feature selection functions or generates correlation on the training dataset to find the features most relevant to the prediction results of different categories [27, 29]. Such features are then treated as marker genes. Machine learning-based models can select marker genes as minimal sets with less redundancy. However, the existing machine learning models either fail to capture patterns of different marker genes or consume much computation resource (see the Results section). Moreover, current methods also cannot identify cells with incorrect cell-type annotations, which means the correctness of marker genes may be confused by incorrect cell types of training datasets.

To overcome the limitations of the existing approaches, we present a model based on interpretable neural networks with stochastic gate design [30] and cosine similarity regularization [29] for marker gene selection, and denote this methods as CosGeneGate. Our model selects marker genes with cell-type-specific expression patterns, which are further discussed in the Results section, by utilizing large-scale annotated scRNA-seq data as the training datasets, and it is also scalable for datasets of different sizes with flexible GPU cores. We offer a list of marker genes with their credible sets for major cell types without redundancy of features. Moreover, we consider several downstream applications with the selected marker genes, e.g. improving the performance of bulk RNA-seq deconvolution tools, reducing the dimensions of single-cell data, identifying spatially-varying marker genes for spatial transcriptomic data, and refining incorrect cell types. We also illustrate how to apply CosGeneGate to uncover disease-specific marker genes for the immune cells from Alzheimer's disease (AD).

Methods

Given one or more scRNA-seq dataset(s) with rows as samples (cells) and columns as features (genes), and known annotated information (including cell types or cell states), our model aims to learn a projection from a high-dimensional space to a lower-dimensional space that preserves biological information. The subset of genes corresponding to the highest prediction accuracy for a given tissue can be treated as the optimal marker-gene candidates for this tissue. To generate the final list of optimal marker genes, we ensure that these genes can not only achieve the highest cell-type classification accuracy, but also follow the cell-type-specific expression patterns for different cell types supported by previous research [27, 29].

Our model has two components. The first component is a stochastic gate (STG) neural network, which selects marker genes based on prediction accuracy. The second component is a post-selector based on the cosine similarity of candidate genes generated by STG, which can focus on candidate genes with similar expression patterns. The final gene list is ranked by scores from the second component and filtered to reduce redundancy. When there are no cell labels, we can assign labels from clustering algorithms [25, 26] or knowledge transformation from another modality [9]. An overview of CosGeneGate is shown in Fig. 1 (a) and (b).

We assess the performance of CosGeneGate based on various datasets and metrics in this manuscript. We also develop a new pipeline for deconvolution analysis to estimate cell-type proportions from bulk RNA-seq datasets. Details of our algorithms are included in Supplementary file 1.

Results

Marker genes selected by CosGeneGate improved the performance of cell-type annotation demonstrated by benchmarking analysis

We first investigated the performance of CosGeneGate with three large-scale scRNA-seq datasets from different tissues (PBMC [31–34], Pancreas [35], and Heart [18]). We illustrate the cell-type similarity across different batches from PBMC in Extended Data Fig. 2, as an example to demonstrate the diversity of our training dataset. We compared the prediction accuracy based on marker genes from different models, including genes from experts (expert design), COSG [29], NSForest [27], scGeneFit [28], STG [30], and CosGeneGate. By using the leave-one-out strategy, we used each batch as the testing dataset and all other batches of the same tissue as the training dataset for selecting markers and training k-nearest neighbor (kNN) classifiers. The information of batches is provided by the sources of these datasets. We evaluated the performance of all models based on metrics including accuracy, weighted F1 score, label Normalized Mutual Information (label NMI), and label Adjusted Rand Index (label ARI). Figure 2 (a) shows the average scores of the above four metrics for each method and each batch. According to Fig. 2 (a), marker genes from CosGeneGate achieved high annotation scores across different datasets. The rank of the annotation scores from CosGeneGate was in the top three for almost all the batches. Based on the average score of all the datasets, CosGeneGate ranked second among all the competitors, as shown in the last column of Fig. 2 (a). Moreover, the variance of annotation scores based on these marker genes is lower than all the other methods except COSG in various metrics, as shown in Extended Data Figs. 3-5. COSG does not have a training process, so it is not affected by random numbers. Using marker genes from CosGeneGate is also significantly better than using marker genes from experts for cell-type annotation.

Validating the marker genes selected by CosGeneGate on the preservation of biological information

Here we investigated the contribution of marker genes for preserving cell-type distinction. By selecting informative marker gene sets (as a method of feature extraction), we can utilize gene expression profiles with a smaller number of features to perform clustering, which is a key step before cell-type annotation. A good marker gene set should preserve the biological distinction induced by cell-type-specific gene expression. Therefore, to evaluate the performance of different gene sets, we computed metrics including PAGA [36] similarity, scaled Average Silhouette Width (scaled ASW), cluster ARI, and cluster NMI based on both highly variable genes (HVGs) and marker genes from different methods and averaged them as the final score. Such scores are relevant for the performance of marker genes for clustering and biological preservation. Details of our evaluations can be found in the Methods section (Supplementary file 1). The average scores are summarized in Fig. 2 (b). Based on this figure, CosGeneGate outperformed other methods in seven out of the 15 batches we tested. By averaging all batches from all three datasets, CosGeneGate ranked second among all competitors, shown in the last column.

Moreover, we computed the Gene Ontology enrichment results (details are in the Methods section (Supplementary file 1)) between each pair of two marker genes and divided the pairs into two groups as genes selected for the same cell type and different cell types. Then, by conducting a t-test between the scores of

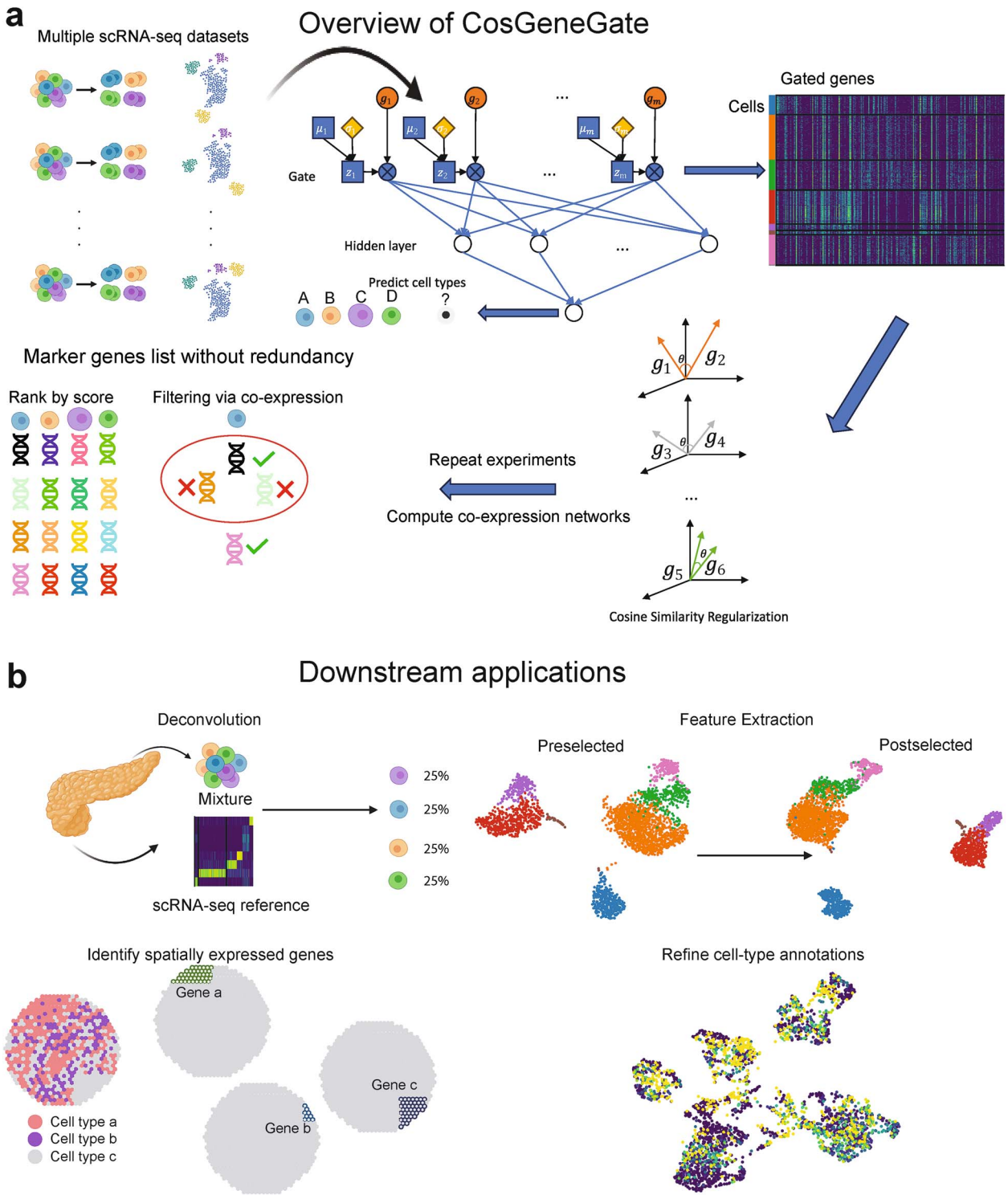

Figure 1. The landscape of CosGeneGate and downstream applications. (a) CosGeneGate utilizes large-scale scRNA-seq as training datasets. Using stochastic gates, we can select candidates of marker genes based on prediction metrics. After collecting all the candidates, we utilize cosine similarity as a method for constraining and filtering target marker genes. Here, $g_{(1\ldots m)}$ represents the input gene expression for m genes. Each gate z is constructed based on a reparameterization trick. The sampling distribution for reparameterization is Gaussian distribution, and μ represents the mean and σ represents the standard deviation. In the cosine-constrained step, θ represents the angle of two gene expression vectors. We run multiple experiments and use correlation to extract equivalent genes, and further filter redundant genes. (b) the downstream applications of marker genes include bulk-seq/spatial transcriptomic data deconvolution, scRNA-seq data feature extraction, spatially expressed marker genes identification, and cell-type annotation correction.

the two groups, the negative logarithm p-value of the t-test can show if there is a statistically significant difference between genes selected for different cell types, thus showing the abilities of biological information preservation. As shown in Extended Data Fig. 6 (a), CosGeneGate performed better than other methods in the PBMC dataset, while ranked second in Pancreas and Heart.

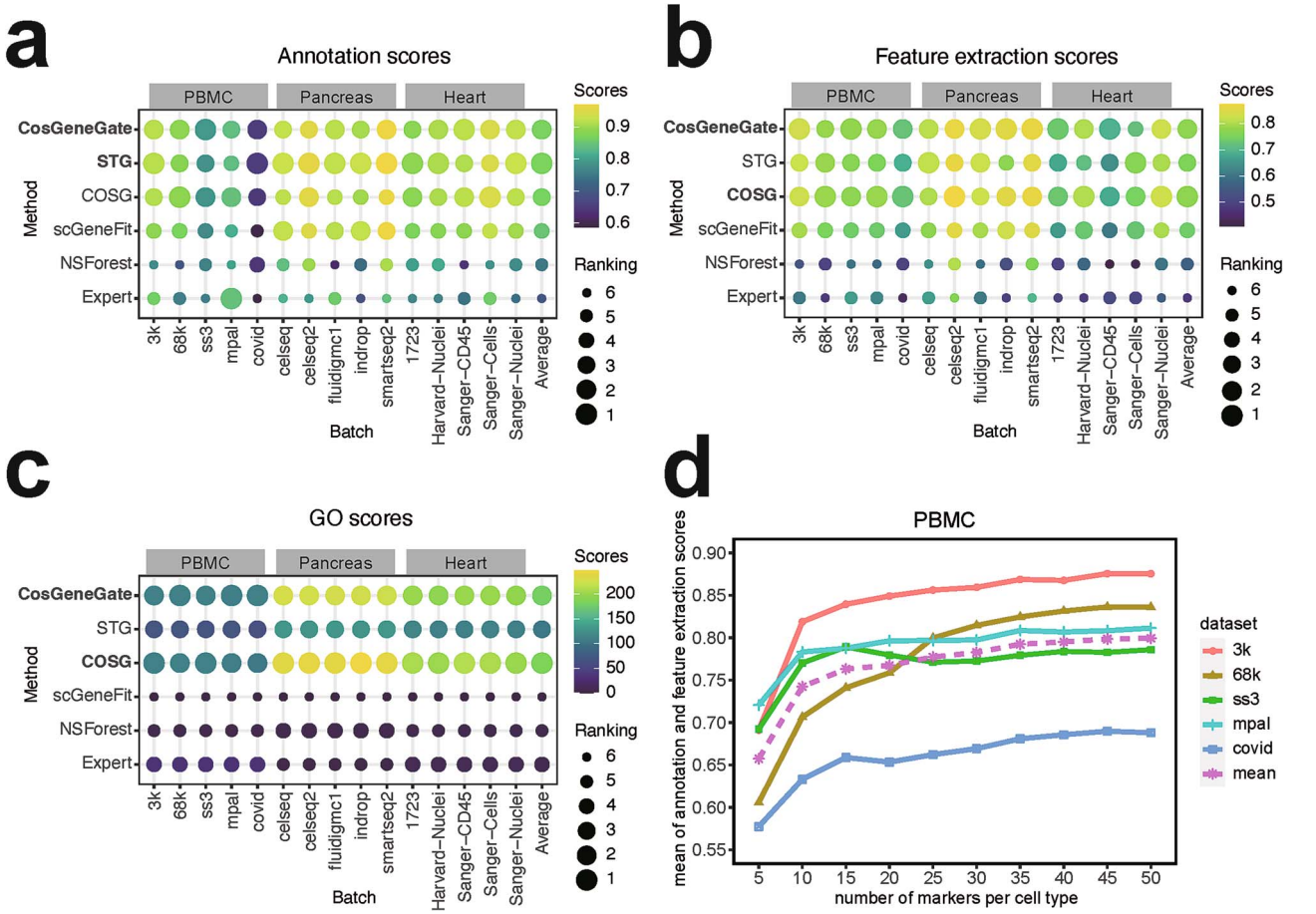

Figure 2. Experimental results for the comparison in the cell-type annotation task. We boldfaced the top 2 methods in panels (a)-(b). (a): Bubble plot for annotation scores (i.e. the average score of accuracy, weighted F1, label ARI, and label NMI) for 12 batches of PBMC, pancreas, and heart. The last column is the averaged annotation scores for each method. (b) Bubble plot for feature extraction scores (i.e. the average of PAGA similarity, scaled ASW, cluster ARI, and cluster NMI) for 12 batches of PBMC, pancreas, and heart. The last column is the averaged feature extraction scores for each method. (c) Line chart for hyperparameter tuning. (d) Boxplot and test result for redundancy removal.

These results demonstrate the ability of CosGeneGate to select functional-specific markers for different cell types. CosGeneGate can also select marker genes with more coherent expressions in the corresponding cell type, and thus these genes have more clear cell-type-specific patterns. Based on Extended Data Figs. 6 (b) and (c), genes selected by STG are usually expressed across several cell types, whereas COSG usually selects genes with ideal expression profiles, however at the cost of reducing prediction accuracy. Overall, our analysis demonstrated the superiority of CosGeneGate in selecting marker genes for biological information preservation.

Regarding the gene number selection, we used the average score of all annotation and feature selection metrics to select the number of marker genes used in the input of CosGeneGate. This criterion is also a standard to select the optimal number of marker genes for each dataset. Figure 2 (c) shows that 50 is a suitable number of markers for each cell type in the PBMC dataset, and more than 50 selected genes will lead to performance drop. Meanwhile, the tuning metrics of the Pancreas dataset are shown in Extended Data Fig. 7. For other methods, we tuned their parameters to reach their best performance in each dataset for fair comparison. It is worth noting that, by averaging all evaluation metrics of cell-type annotation and feature extraction from all datasets, CosGeneGate ranked first among all methods, suggesting that marker genes selected by CosGeneGate can optimize these the cell-type annotation and feature extraction tasks simultaneously.

Removing redundant genes by uncertainty and co-expressions

Here we investigated the performance of CosGeneGate after removing gene redundancy based on changing random seeds and constructing co-expression networks. The workflow of redundancy removal is explained in the Methods section and visualized in Fig. 1 (a). We compared the four metrics of cell-type annotation between the original marker gene set and the credible marker gene set (i.e. the gene set after redundancy removal). For the PBMC dataset, the result of the paired t-test between the two gene sets is shown in Fig. 2 (d), and the other results are shown in Extended Data Fig. 8. According to the boxplot and test result, we improved annotation accuracy for PBMC and Pancreas by generating a credible set of genes, demonstrating that removing redundant genes can lead to more accurate cell-type annotation. We have provided credible sets for each major cell type of different tissues (PBMC, Pancreas, and Heart) in Supplementary file 2.

Removing redundant genes by uncertainty and co-expressions

In this section, we investigate the preference of CosGeneGate for marker gene selection and present an algorithm to choose a

suitable number of marker genes from a statistical perspective. We computed the Wilcoxon rank-sum test score for the marker genes from PBMC and Pancreas based on candidate genes selected by CosGeneGate, and visualized the relation between the number of marker genes and the score in Extended Data Fig. 9 (a). This figure shows that the number of marker genes with the best performances depends on the most difficult cells to classify in the tissue, i.e. the cells with the highest number of genes needed for the score to reach its maximum value. For example, in the PBMC dataset, selecting 50 genes per cell type is the optimal choice based on our metrics. On the other hand, in the Pancreas dataset, selecting 20 genes per cell type is the optimal choice based on our metrics. These two choices correspond to the cell types highlighted by the red block in Extended Data Fig. 9 (a). We also tried to use two different approaches to explain the preference of our marker genes, known as the COSG score (shown in Extended Data Fig. 9 (b)) and the Metamarker score [37] (shown in Extended Data Fig. 9 (c)). However, in the analysis of the COSG score for the PBMC dataset, we did not see a positive relation between the COSG score and the performance of marker genes. The original manuscript for Metamarker score advocates to choose 50–200 genes per cell type for accurate cell-type annotation based on the trade-off between the FC score and the area-under-curve (AUC) score. However, based on our analysis, not all the cell types had such patterns, and it is difficult to balance between the FC score and the AUC score for all cell types in the Pancreas dataset. Therefore, previous methods could not be used to explain the multifunctional marker genes from CosGeneGate, and our analysis is more suitable for model explanation. Details of our algorithm are in the Methods section (Supplementary file 1).

Selecting marker genes improved the performance of deconvolution

We first investigated the impact of marker gene selection on deconvolution. For signature matrix-based deconvolution methods, we can deconvolve the bulk RNA-seq datasets by only keeping selected marker genes rather than using all genes. Since only informative genes are kept, the deconvolution performance is expected to be better. Then, we compared the deconvolution results of all genes, NSForest [27], scGeneFit [28], STG [30], and COSG [29] on different deconvolution models: CIBERSORTx [38], MuSiC [39], and NNLS [40]. To evaluate the performance, we used Root Mean Square Error (RMSE), Pearson Correlation Coefficient, and Coefficient of Variation as metrics. Details of these metrics are included in the Methods section (Supplementary file 1).

Regarding the parameter selection, we used RMSE to select the optimal number of marker genes. For other competitors, we tuned the number of marker genes to reach their best performance accordingly for a fair comparison. Empirical experiments show that the optimal number of marker genes differs for different datasets and different gene selection strategies. Therefore, it should be tuned in practice. To tackle the problem, we designed a deconvolution pipeline that adjusts the number of marker genes to be selected by CosGeneGate. Figure 3 (a) and (b) show the overall workflow of the pipeline for the pseudo bulk mode and the real bulk mode, respectively. In the pseudo bulk mode, the scRNA-seq dataset is split into two parts, one for pseudo bulk data [38] generation, and the other for hyperparameter tuning with CosGeneGate marker genes on these generated pseudo bulk mixtures for this model. In the real bulk mode, the model is directly tuned on the real bulk data [41], and its ground truth cell-type proportion information is provided by the users. Figure 3 (c) shows how the deconvolution performance varies as the number of marker genes for each cell type changes within the pancreas dataset. There is a parallel pattern for RMSE between the validating dataset and the testing dataset, suggesting that the number of marker genes that achieve the best performance in the validation dataset should also achieve almost the best performance in the testing data, guiding the selection of the number of marker genes. Figure 3 (d) shows the performance of deconvolution on a real bulk dataset known as the 3celllines dataset. CosGeneGate performed well on average compared with other gene selection strategies, and outperformed the case of choosing all genes. Figure 3 (e) shows the performance improvement for both the pseudo bulk mode and the real bulk mode. We can observe that in most cases, the pseudo bulk mode improves the performance of deconvolution, which is further improved by the real bulk mode.

Selecting marker genes uncovered spatially-informed patterns in spatial transcriptomic data

We next assessed the performance of CosGeneGate on selecting marker genes from scRNA-seq data for spatial transcriptomics studies. We considered two types of spatial transcriptomic data to investigate the general applicability of our method. In the dataset (visium_fluo) sequenced using 10x Visium [42], every spot within the spatial transcriptomic dataset encompasses a variable number of cells and 10x Visium is a whole-transcriptome-based technique. In contrast, in the dataset (human breast) obtained through Xenium sequencing [43], we can extract a single-cell gene expression profile with hundreds of genes, and each spot corresponds to an individual cell. Using scRNA-seq to select marker genes for the spatial transcriptomic data is meaningful for spatial-level deconvolution and cell-type annotation [44]. We compared the results of 2000 HVGs, COSG, NSForest, scGeneFit, scMAGS [44], STG, and CosGeneGate to evaluate their abilities to identify marker genes for spatial transcriptomic data from scRNA-seq data. We added scMAGS in this section because of its ability to analyze spatial transcriptomic data. Figure 4 (a) shows the evaluation of the preservation of gene–gene correlation selected by different methods for the 10x Visium dataset and the human breast dataset based on different random seeds. According to Fig. 4 (a), CosGeneGate has a high gene–gene correlation across different methods and different datasets. Using marker genes from CosGeneGate is also significantly better than using HVGs. Details of our evaluation methods are summarized in the Methods section (Supplementary file 1). Moreover, Extended Data Fig. 10 and Fig. 4 (b) show the distribution of cortex of the visium_fluo dataset and cell types of the human breast dataset with spatial location respectively. The gene expression profiles of spatially-expressed marker genes shown in Fig. 4 (c) of malignant cells identified by CosGeneGate can illustrate a binary expression profile compared with cells from different cell types of the human breast dataset. These results altogether support the ability of CosGeneGate to select marker genes that uncover cell-type-specific gene expression patterns in spatial transcriptomic data.

Selecting marker genes from CosGeneGate refined uncertain cell types

We assessed CosGeneGate's ability to use selected marker genes to detect incorrect or unknown cell types. According to Fig. 5 (a), the 3-fold intra-dataset hierarchical classification accuracy of all methods is low, leading us to conjecture that there are wrongly labeled sub-cell types in the Zeisel dataset [45]. Figure 5 (b) shows the UMAP for the sub-cell type distribution of the Zeisel dataset. In Fig. 5 (c), we count and then display the number of times a

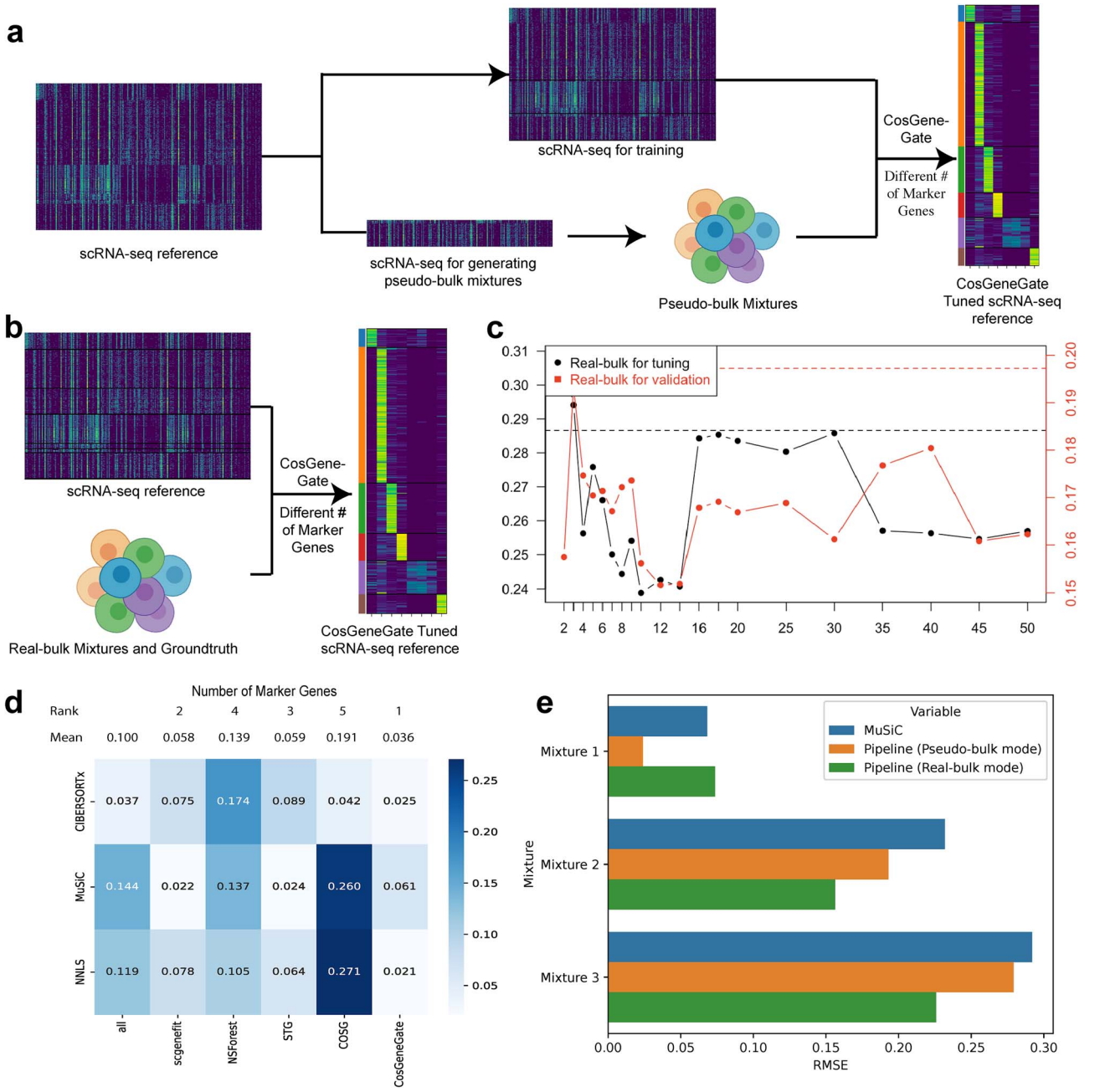

Figure 3. Experimental results for the comparison in the deconvolution task. (a) the workflow for CosGeneGate marker-based deconvolution pipe-line (pseudo-bulk mode) (b) the workflow for CosGeneGate marker-based deconvolution pipeline (real-bulk mode) (c) relationship between the number of markers and deconvolution performance (d) benchmark of performance of deconvolution in the real bulk three-cell dataset (e) the performance of deconvolution pipeline in real bulk pancreas dataset with two different modes.

cell's classification result did not match the label divided by the total number of trials as the uncertainty score. As can be seen in the two UMAPs, tightly clustered cells of the same cell type tend to have uncertainty scores closer to zero and cells between cell-type clusters have higher uncertainty scores, which accords with intuition. We defined unknown or known sub-cell types as wrongly or correctly labeled sub-cell types by markers from CosGeneGate using all ten seeds and all ten n_neighbors parameters of kNN classifiers, i.e. cells' uncertainty scores equal zero for cells which are classified as their annotated types under all the experiments. Then, using the expert markers from Zeisel [45], we generated the dotplots of unknown and known sub-cell types in oligodendrocytes, as shown in Fig. 5 (d). Comparing the upper panel with the bottom panel, we observe that the expert marker genes from unknown cell types such as Oligo 5 and Oligo 6 have unclear patterns. The cells with unknown labels are indeed wrongly labeled based on their expression of the expert markers. We offer the full marker gene list in Extended Data Fig. 11. Finally, we collected the experiment results from the kNN classifier and re-annotated cell types with the maximal cell-type proportion of prediction. These results are summarized in Fig. 5 (e).

Discovery of disease-specific marker genes for AD

CosGeneGate can also discover disease-specific marker genes based on the analysis of sub-cell types across samples with

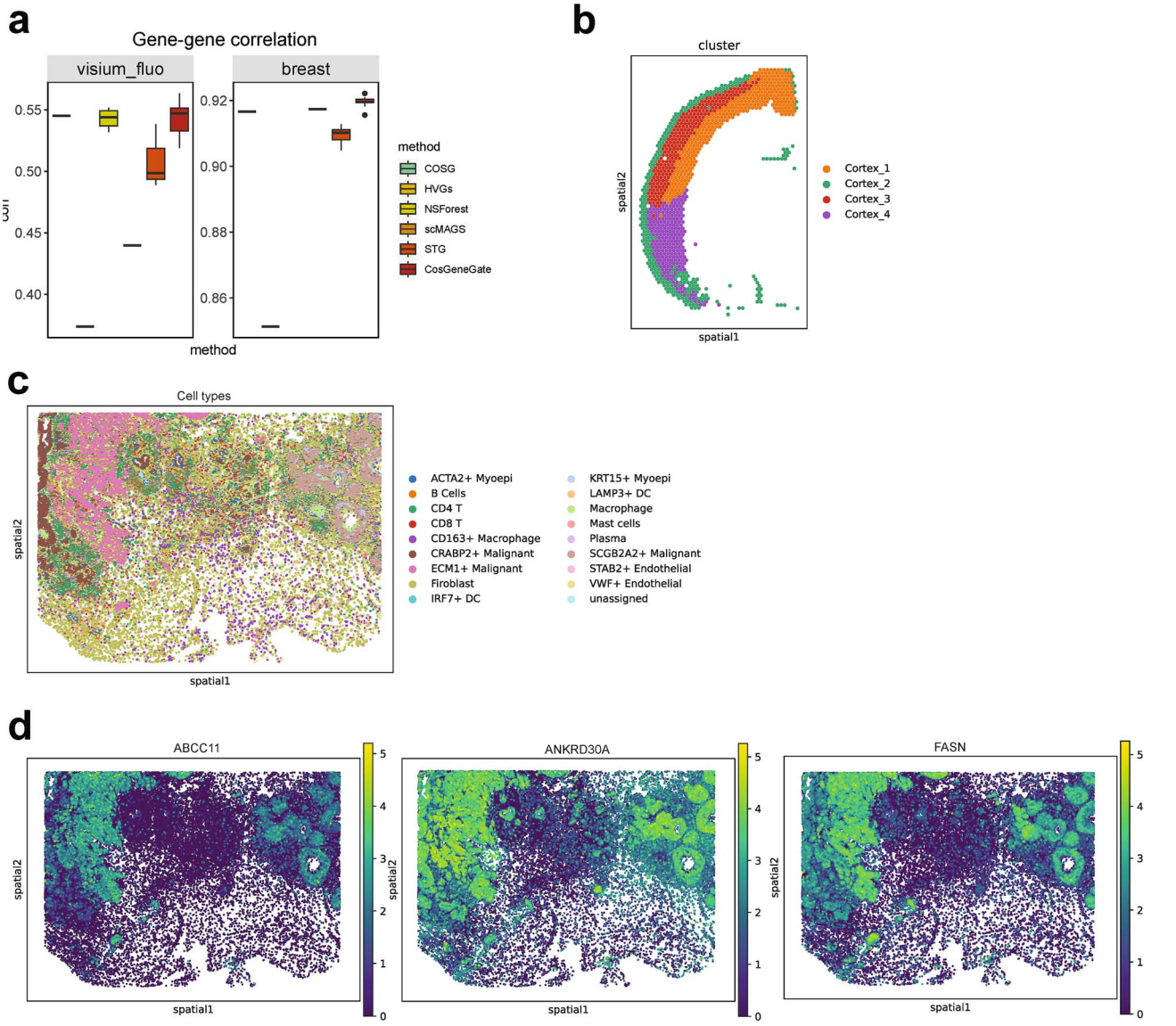

Figure 4. Experimental results for the application of marker genes. (a) the evaluation for the preservation of gene–gene correlation of the genes selected by different methods for the 10x Visium dataset and the human breast dataset. We did not report the results of NS-Forest for the human breast dataset be-cause its running time exceeded our limit. (b) Visualization of the cell-type distribution of the human breast dataset. (c) Spatially-expressed marker genes identified by CosGeneGate in the human breast dataset.

different conditions. Here we focus on AD, the most common age-related neurodegenerative disease. Clinical symptoms of AD are characterized by progressive cognitive decline and dementia. Therefore, the analysis of possible genetic risk factors for AD is important for clinical practice today. Microglia cells are important immune cells in the brain, which are highly correlated with AD [46–49]. To identify disease-specific marker genes, we sequenced a new dataset containing 126,687 cells from seven samples using 10x Multiome sequencing. We show the distribution of diseased conditions in Fig. 6 (a) and the distribution of cell types in Fig. 6 (b). Here we tried three different approaches and used the overlap score (known as precision) between selected disease-specific marker genes and known disease-associated genes as the criterion. Details are included in Extended Data Fig. 12. Finally, we identified the sub-clusters/sub-cell types of microglia cells from our AD-Health Control (HC) datasets and then utilized CosGeneGate to select possible genetic risk factors. Such genes were identified in the AD samples as marker genes of certain sub-cell types rather than in HC samples. Figure 6 (c) shows the expression patterns of selected disease-specific marker genes. The selected AD-specific marker genes showed higher expression levels in AD-associated sub-cell types. The overlapping information between selected marker genes and known AD-associated genes can be found in Supplementary file 3. Those genes which are not previously identified as AD-associated can be treated as novel AD-specific marker genes. Therefore, we can also use CosGeneGate to identify the risk genes for diseases supported by disease-specific expression patterns and the results for mining AD-associated genes from known databases.

Discussion

Marker genes not only work as the safeguards for the cell-type annotation prior to downstream analysis of single-cell data, but can also facilitate downstream applications including bulk RNA-seq data deconvolution, feature extraction, spatial gene

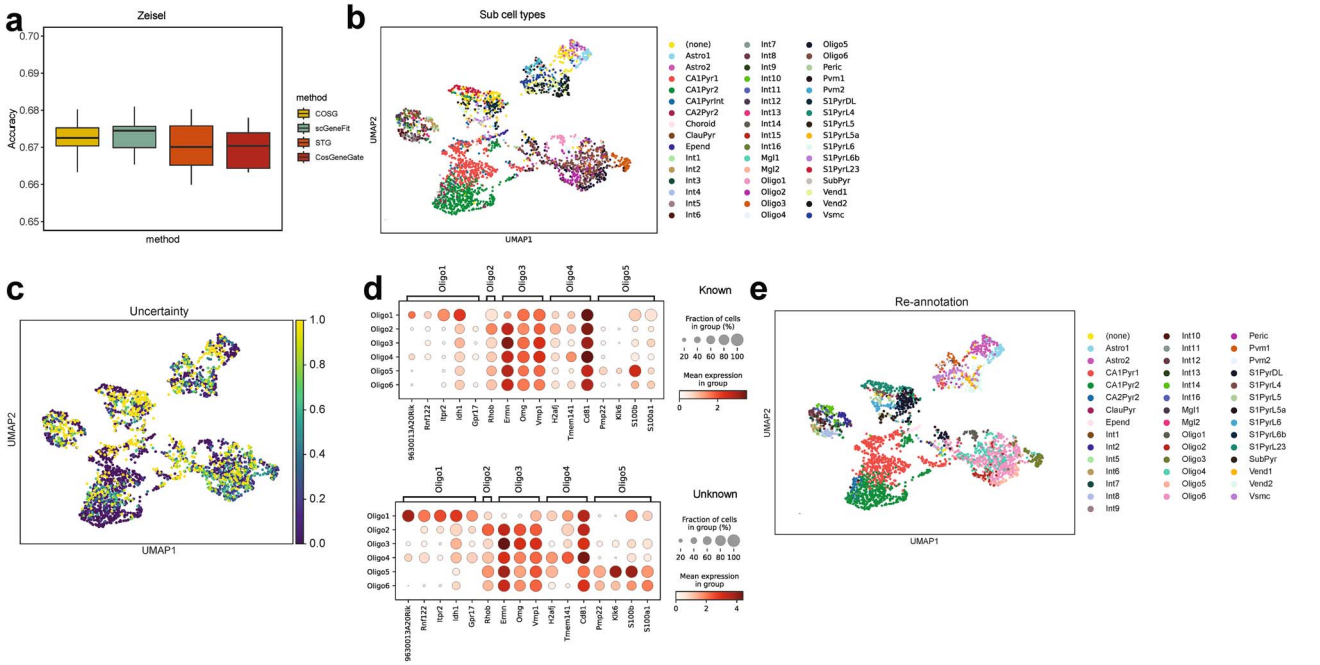

Figure 5. Utilizing marker genes to detect unknown cell types. (a) the accuracy of cell-type annotation based on different methods for the Zeisel dataset. (b) UMAPs for the sub-cell type distribution of the Zeisel dataset. (c) Sub-cell type-specific gene expression plots for cells with incorrect labels (top) and cells with correct labels (bottom). (d) UMAPs for the cells with uncertainty. (e) UMAPs for the re-annotated cells.

expression pattern identification, sub-cell type correction, the discovery of disease-specific genes, and others. Therefore, the multifunctional marker genes selected by CosGeneGate are important for single-cell data pre-processing and downstream applications. CosGeneGate selects marker genes by considering both the contribution of target genes to cell-type annotation, and the cell-type-specific expression patterns of marker genes. By combining these two ideas together, we showed the superiority of marker genes identified by CosGeneGate in the Results section. Furthermore, we designed a framework to reduce the redundancy of our marker gene list and demonstrated its usefulness in the Results section.

We note that in the cell-type annotation task, CosGeneGate had similar performance compared to STG and COSG. However, by taking a deeper look at the gene expression patterns of STG, the marker genes selected by STG tended to express across different cell types, which were false-positive signals discovered by STG shown in Extended Data Figs. 6 (b) and (c). Moreover, the genes selected by COSG are not ideal choices for other downstream applications, including representative feature extraction and bulk RNA-seq deconvolution. It also consumed more memory usage compared to CosGeneGate. Our method integrated the advantages of COSG and STG while avoiding their shortcomings. CosGeneGate also outperformed NSForest and scGeneFit for multiple tasks. The running time of NSForest was too long to select marker genes effectively. Genes selected by scGeneFit are not cell-type-specific and do not have rank information to quantitatively describe their quality, which limits their functionality.

For the deconvolution task, we developed a pipeline for bulk RNA-seq data deconvolution, which allowed users to choose different modes, and it could select the number of marker genes with the best performance automatically, which is also a user-friendly design. For the feature extraction task, we demonstrated that marker genes selected by CosGeneGate could not only preserve the distinction of cell types, but also retain the structure similarity of different cell types before running the selection step. Reducing the dimensionality we need for single-cell analysis is important for efficiency. For the analysis related to spatial transcriptomic data, we demonstrated that marker genes selected by CosGeneGate from scRNA-seq datasets could be used to identify genes with spatial expression patterns based on the spatial transcriptomic data from the same tissue, which builds a bridge for multi-omics data analysis. We also used CosGeneGate to refine the sub-cell types from scRNA-seq datasets thus increasing the data quality. Therefore, the marker genes from CosGeneGate are high-quality across different scenarios.

We also used CosGeneGate for biological discoveries, e.g. identifying new disease-specific marker genes based on sequenced datasets. We identified a number of disease-specific marker genes for the sub-cell types of Microglia cells by analyzing samples with/without AD. In the final marker gene set, some genes are verified as disease-associated markers by analyzing GWAS statistics and/or related biological experiment results. The rest of genes can be treated as newly discovered marker genes.

To determine the number of marker genes identified by CosGeneGate, we ran experiments by adjusting the number of marker genes as a hyper-parameter and selecting the number corresponding to the cell-type annotation and feature extraction task. We also considered the explainability of our methods and discovered the relation between the number of marker genes per cell type and the related score from statistical tests. After redundancy removal and co-expression patterns, we summarized these genes into a table for major tissues, which could be used by users. Moreover, the number of marker genes is also an editable hyper-parameter. Users can also adjust its value if they prefer more/fewer markers.

In summary, CosGeneGate can select high-quality marker genes for multiple downstream applications and has acceptable running time and memory usage. It has its unique feature selection design by combining ideas from the machine learning

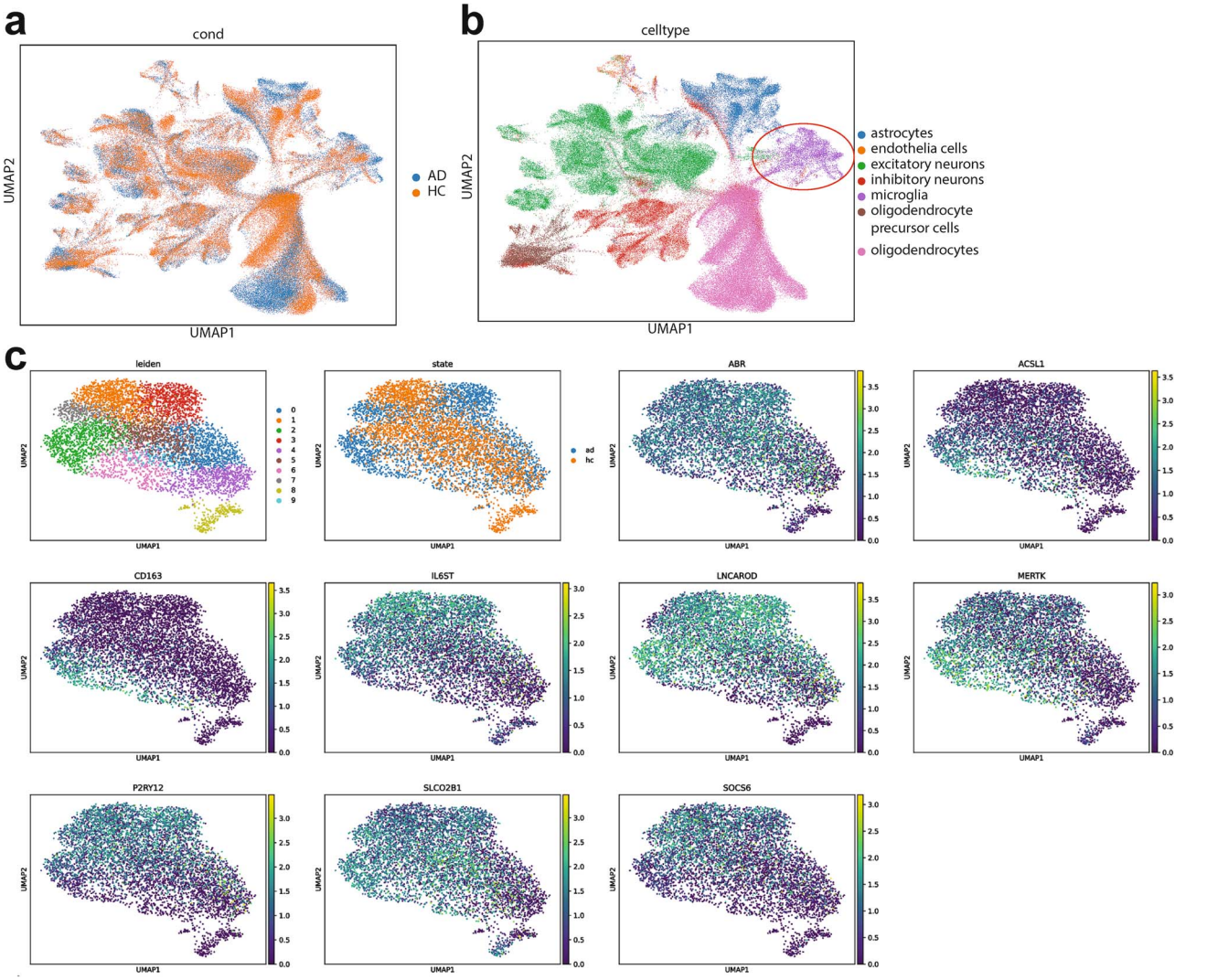

Figure 6. Utilizing marker genes for sub-cell types to explore the biological system. (a) UMAPs of the AD-HC dataset colored by disease condition. (b) UMAPs of the AD-HC dataset colored by cell type. Microglia cells are highlighted by a red circle, located in the right most part of this figure. (c) UMAPs for the AD-specific marker genes of microglia cells.

area and the biology area. The next step of CosGeneGate will shift to multi-omic data analysis and multi-species data analysis.

Key Points

- Selecting representative genes or marker genes to distinguish cell types is an important task in single-cell sequencing analysis and is helpful for downstream analyses.
- we present a novel model to select marker genes for more effective marker selections. Our method inspired by combining the advantages of selecting marker genes based on both cell-type classification accuracy using stochastic gating and marker gene specific expression patterns.
- We demonstrate the outstanding performances of our method in selecting marker genes and generate a statistical framework to explain our results. We also analyze a new ADHC dataset to discover new disease-related marker genes.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Acknowledgements

We thank the suggestions from Jia Zhao, Yingxin Lin and Ning Sun for improving the quality of our manuscript. This work was supported in part by NIH grants R01 GM134005 and R56 AG074015.

Conflict of interests. All authors do not have conflict of interests.

Funding

This work was supported in part by NIH grants R01 GM134005 and R56 AG074015.

Data availability

We summarize the sources and statistics of all the datasets we used in Supplementary file 4. All the datasets can be accessed based on the links in this file.

References

1. Han X, Zhou Z, Fei L *et al.* Construction of a human cell landscape at single-cell level. *Nature* 2020;**581**:303–9. https://doi.org/10.1038/s41586-020-2157-4.
2. Saliba A-E, Westermann AJ, Gorski SA *et al.* Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* 2014;**42**:8845–60. https://doi.org/10.1093/nar/gku555.
3. Mathys H, Davila-Velderrain J, Peng Z *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* 2019;**570**:332–7. https://doi.org/10.1038/s41586-019-1195-2.
4. Stubbington MJT, Rozenblatt-Rosen O, Regev A *et al.* Single-cell transcriptomics to explore the immune system in health and disease. *Science* 2017;**358**:58–63. https://doi.org/10.1126/science.aan6828.
5. Zhang Z, Mathew D, Lim T. *et al.* Signal recovery in single cell batch integration. bioRxiv. 2023. Preprint biorxiv:2023.05.05.539614.
6. Evrony GD, Hinch AG, Luo C. Applications of single-cell DNA sequencing. *Annu Rev Genom Hum Genet* 2021;**22**:171–97. https://doi.org/10.1146/annurev-genom-111320-090436.
7. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 2018;**50**:1–14. https://doi.org/10.1038/s12276-018-0071-8.
8. Zheng GXY, Terry JM, Belgrader P *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**:14049. https://doi.org/10.1038/ncomms14049.
9. Stoeckius M, Hafemeister C, Stephenson W *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;**14**:865–8. https://doi.org/10.1038/nmeth.4380.
10. Cusanovich DA, Daza R, Adey A *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 2015;**348**:910–4. https://doi.org/10.1126/science.aab1601.
11. Chen M, Zhu G, Xu J *et al.* Differentiation of isomeric methylanilines by imidization and gas chromatography/mass spectrometry analysis. *Rapid Comm Mass Spectrometry* 2018;**32**:342–8. https://doi.org/10.1002/rcm.8043.
12. Luo C, Keown CL, Kurihara L *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* 2017;**357**:600–4. https://doi.org/10.1126/science.aan3351.
13. Mulqueen RM, Pokholok D, Norberg SJ *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nat Biotechnol* 2018;**36**:428–31. https://doi.org/10.1038/nbt.4112.
14. Zhang M, Eichhorn SW, Zingg B *et al.* Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* 2021;**598**:137–43. https://doi.org/10.1038/s41586-021-03705-x.
15. Flynn E, Almonte-Loya A, Fragiadakis GK. Single-cell multiomics. *Annu Rev Biomed Data Sci* 2023;**6**:313–37. https://doi.org/10.1146/annurev-biodatasci-020422-050645.
16. Fleck JS, Camp JG, Treutlein B. What is a cell type? *Science* 2023;**381**:733–4. https://doi.org/10.1126/science.adf6162.
17. Yu L, Wu Y, Dunn JF *et al.* In-vivo monitoring of tissue oxygen saturation in deep brain structures using a single fiber optical system. *Biomed Opt Express* 2016;**7**:4685–94. https://doi.org/10.1364/BOE.7.004685.
18. Litviňuková M, Talavera-López C, Maatz H *et al.* Cells of the adult human heart. *Nature* 2020;**588**:466–72. https://doi.org/10.1038/s41586-020-2797-4.
19. Yang P, Huang H, Liu C. Feature selection revisited in the single-cell era. *Genome Biol* 2021;**22**:321. https://doi.org/10.1186/s13059-021-02544-3.
20. Giladi A, Paul F, Herzog Y *et al.* Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat Cell Biol* 2018;**20**:836–46. https://doi.org/10.1038/s41556-018-0121-4.
21. Gómez-Chávez F, Cañedo-Solares I, Ortiz-Alegría LB *et al.* Maternal immune response during pregnancy and vertical transmission in human toxoplasmosis. *Front Immunol* 2019;**10**:285. https://doi.org/10.3389/fimmu.2019.00285.
22. Domínguez Conde C, Xu C, Jarvis LB *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 2022;**376**:eabl5197. https://doi.org/10.1126/science.abl5197.
23. Stuart T, Butler A, Hoffman P *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
24. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15. https://doi.org/10.1186/s13059-017-1382-0.
25. Blondel VD, Guillaume J-L, Lambiotte R *et al.* Fast unfolding of communities in large networks. *J Stat Mech* 2008;**2008**:P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.
26. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**:5233. https://doi.org/10.1038/s41598-019-41695-z.
27. Aevermann B, Zhang Y, Novotny M *et al.* A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing. *Genome Res* 2021;**31**:1767–80. https://doi.org/10.1101/gr.275569.121.
28. Dumitrascu B, Villar S, Mixon DG *et al.* Optimal marker gene selection for cell type discrimination in single cell analyses. *Nat Commun* 2021;**12**:1186. https://doi.org/10.1038/s41467-021-21453-4.
29. Dai M, Pei X, Wang X-J. Accurate and fast cell marker gene identification with COSG. *Brief Bioinform* 2022;**23**:bbab579. https://doi.org/10.1093/bib/bbab579.
30. Yamada Y, Lindenbaum O, Negahban S *et al.* Feature selection using stochastic gates. *Proceedings of the 37th International Conference on Machine Learning* 2020;**119**:10648–59.
31. Pullin JM, McCarthy DJ. A comparison of marker gene selection methods for single-cell RNA sequencing data. *Genome Biol* 2024;**25**:56. https://doi.org/10.1186/s13059-024-03183-0.
32. Wang Y, Liu T, Zhao H. ResPAN: a powerful batch correction model for scRNA-seq data through residual adversarial networks. *Bioinformatics* 2022;**38**:3942–9. https://doi.org/10.1093/bioinformatics/btac427.
33. Granja JM, Klemm S, McGinnis LM *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* 2019;**37**:1458–65. https://doi.org/10.1038/s41587-019-0332-7.
34. Wilk AJ, Lee MJ, Wei B *et al.* Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19. *Journal of Experimental Medicine* 2021;**218**:e20210582. https://doi.org/10.1084/jem.20210582.
35. Tran HTN, Ang KS, Chevrier M *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 2020;**21**:12. https://doi.org/10.1186/s13059-019-1850-9.
36. Wolf FA, Hamey FK, Plass M *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;**20**:59. https://doi.org/10.1186/s13059-019-1663-x.
37. Fischer S, Gillis J. How many markers are needed to robustly determine a cell's type? *iScience* 2021;**24**:21. https://www.cell.com/iscience/fulltext/S2589-0042(21)01261-X.

38. Newman AM, Steen CB, Liu CL *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;**37**:773–82. https://doi.org/10.1038/s41587-019-0114-2.
39. Wang X, Park J, Susztak K *et al.* Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* 2019;**10**:380. https://doi.org/10.1038/s41467-018-08023-x.
40. Danaher P, Kim Y, Nelson B *et al.* Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nat Commun* 2022;**13**:385. https://doi.org/10.1038/s41467-022-28020-5.
41. Dong M, Thennavan A, Urrutia E *et al.* SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief Bioinform* 2021;**22**:416–27. https://doi.org/10.1093/bib/bbz166.
42. Biancalani T, Scalia G, Buffoni L *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nat Methods* 2021;**18**:1352–62. https://doi.org/10.1038/s41592-021-01264-7.
43. Lin Y, Cao Y, Kim HJ *et al.* scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Mol Syst Biol* 2020;**16**:e9389. https://doi.org/10.15252/msb.20199389.
44. Baran Y, Doğan B. scMAGS: marker gene selection from scRNA-seq data for spatial transcriptomics studies. *Comput Biol Med* 2023;**155**:106634. https://doi.org/10.1016/j.compbiomed.2023.106634.
45. Zeisel A, Muñoz-Manchado AB, Codeluppi S *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015;**347**:1138–42. https://doi.org/10.1126/science.aaa1934.
46. Keren-Shaul H, Spinrad A, Weiner A *et al.* A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* 2017;**169**:1276–1290.e17. https://doi.org/10.1016/j.cell.2017.05.018.
47. Hansen DV, Hanson JE, Sheng M. Microglia in Alzheimer's disease. *J Cell Biol* 2018;**217**:459–72. https://doi.org/10.1083/jcb.201709069.
48. Wang C, Zong S, Cui X *et al.* The effects of microglia-associated neuroinflammation on Alzheimer's disease. *Front Immunol* 2023;**14**:1117172. https://doi.org/10.3389/fimmu.2023.1117172.
49. Zhang L, He CH, Coffey S. *et al.* Single-cell transcriptomic atlas of Alzheimer's disease middle temporal gyrus reveals region, cell type and sex specificity of gene expression with novel genetic risk for MERTK in female. medRxiv. 2023. Preprint medrxiv:2023.02.18.23286037.

Briefings in Bioinformatics, 2025, **26(1)**, bbae626
https://doi.org/10.1093/bib/bbae626
Problem Solving Protocol