Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression

Yuqiao Gong[1], Jingsi Xu[1], Ruitian Gao[1], Jianle Sun[1], Zhangsheng Yu[1,2,3,4,*] and Yue Zhang[1,*]

[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang District, 200240, Shanghai, China
[2]SJTU-Yale Joint Center for Biostatistics and Data Science Organization, Shanghai Jiao Tong University, Shanghai, China
[3]Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[4]Center for Biomedical Data Science, Translational Science Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China
*Corresponding author. yuzhangsheng@sjtu.edu.cn, yue.zhang@sjtu.edu.cn

Abstract

Alzheimer's disease (AD) is a complex and debilitating neurodegenerative disorder that has received widespread attention and research in recent years. The pathogenesis of AD involves complex regulation between genes and intricate changes in different cell types. Searching for a single pathogenic gene or investigating changes in a single cell type has limitations in understanding the disease. Therefore, there is a need for a more comprehensive approach to analyze AD. Here, we propose a single-cell Bayesian biclustering (scBC) framework for the cell-specific detection of network gene biomarkers in scRNA-seq data, utilizing the biclustering method to analyze the perturbations in functional gene modules of complex diseases at single-cell level. By applying our framework to AD scRNA-seq data, we uncover the perturbations of functional gene modules under different cell groups and shed light on gene-cell correlations during AD progression. Our method can handle the challenges of single-cell data such as batch effects and the high proportion of dropout, which are common issues in single-cell data analysis. Moreover, by incorporating existing biological information into our model, we obtain biologically meaningful results. We conducted comparisons on simulated datasets and several highly heterogeneous real-world datasets at different levels, demonstrating that our method is more precise and robust than other state-of-the-art biclustering methods. Our study highlights the great potential of scBC in uncovering the mechanism of polygenic diseases with complex gene co-expression patterns and providing potential treatment options.

Keywords: Functional gene modules, Biclustering, scRNA-seq, scBC, Alzheimer's disease

Introduction

In recent years, the advancement of single-cell sequencing technology has enabled the analysis of single-cell data to reveal meaningful biological information at the cellular level. Specifically, single-cell RNA sequencing (scRNA-seq) enables the sequencing of cells that are hard-to-retrieve or challenging to isolate[1]. This unprecedented resolution into cell states provides us with new insights into the function and dysfunction of cells[2], which is particularly necessary for complex diseases like Alzheimer's disease (AD), as changes in gene expression are related to cell type[3, 4]. Recently, there has been a surge of single-cell studies aimed at understanding the mechanism of AD based on transcriptional profiles[3, 5, 6], which have provided valuable insights into cellular diversity. However, these studies often lack an integrative analysis of functional gene modules, which can reveal how genes work together to regulate biological processes. A recent study used a network-based approach to identify functional gene modules involved in the selective vulnerability of neurons in Alzheimer's disease, demonstrating the importance of analyzing functional gene modules (FGMs) to gain insights into the underlying mechanisms of complex diseases like AD[7]. FGMs are groups of genes that work together to perform a specific biological function and can exhibit complex co-expression or co-regulation patterns, rather than solely comprising differentially expressed genes[8, 9]. Moreover, these local patterns are often cell-specific and may change with disease progression[10-12]. Therefore, it is crucial to identify functional gene modules and their corresponding functional cell groups simultaneously in studies of complex diseases. In this study, we focus on FGMs, the network gene biomarkers, to investigate their potential role in AD.

Unlike clustering methods, which can only conduct clustering in either cell space or gene space, biclustering can identify functional gene modules and their corresponding functional cell groups simultaneously. A cells-genes pair is called a bicluster, the genes in a specific bicluster can be deemed as a FGM shared across related cells. Meanwhile, these cells behave similarly with respect to these relative genes. Therefore, through biclustering, we can easily identify functional gene modules and find the cell populations in which they are active at the same time. It is worth noting that in the complex cell machinery, multiple FGMs are active in a cell group, and different cell groups may share a common FGM (Fig. S1). Fortunately, biclustering with overlapped biclusters can easily capture such complex features[13]. Through biclustering, cell population-specific network gene biomarker and potential gene-cell connections can be identified in a single pass.

Although biclustering is an exquisite tool, it encounters some intractable problems when applied to scRNA-seq data. First, batch effects, due to laboratory conditions, reagent lots, and personnel differences, are widespread and critical to address[14]. If batch effects are not properly accounted for, biclustering algorithms may falsely identify batch-specific co-expression patterns instead of true biological patterns, leading to incorrect conclusions. Second, due to low mRNA content per cell and

molecule losses during the experiment (known as "dropout"), the gene expression matrix has a substantial amount of zero read counts which can cause problems for biclustering algorithms that assume continuous expression values[15]. Biclustering algorithms that are not designed to handle dropout may either ignore the zero read counts, leading to incomplete biclusters, or consider them as low-expressed genes, leading to spurious biclusters. Although algorithms have been designed to address these inherent problems pervasive in scRNA-seq data, they can only improve the performance of the biclustering algorithm in one particular aspect – cell clustering, FGM finding, or the simultaneity of co-clustering[16-18]. However, in complex polygenic diseases, functionally related potential cell groups are finely divided, cell-type conditional gene co-expression patterns are complicated, and the cell-gene correlation changes throughout the progression. Therefore, an algorithm with better performance in functionally related cell group discovery, FGM finding, and cell-gene correlation pattern detection is urgently needed to assist the research of these diseases.

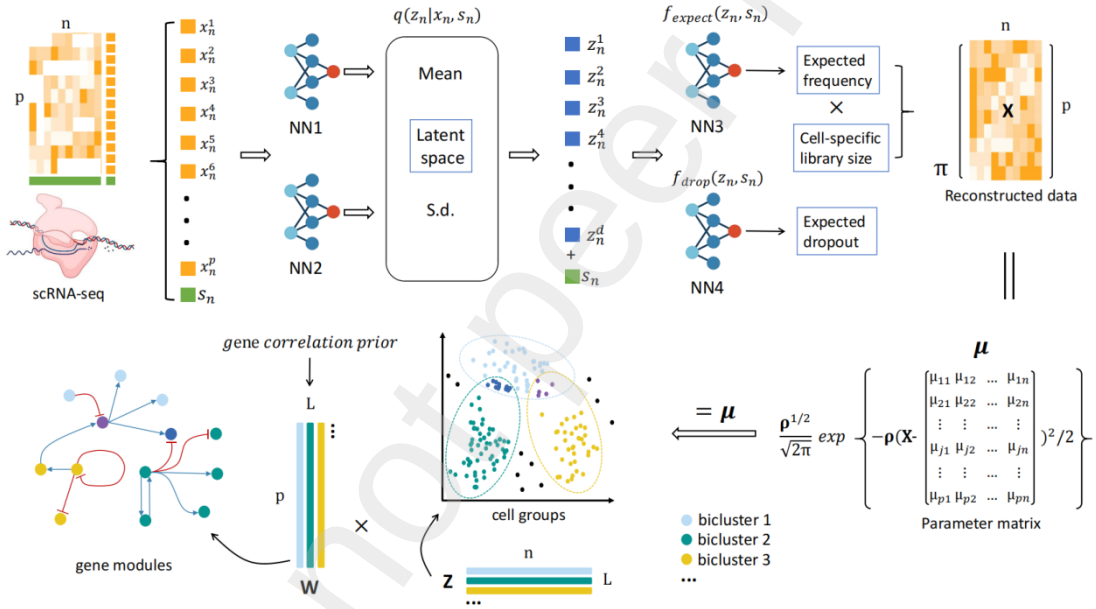


Fig. 1 | flowchart of scBC procedure. The scRNA-seq data with high proportion of dropout and batch annotation (if available) is first fed into the variational autoencoder (VAE). We use x_n^g to denote the g_{th} gene in cell n. s_n is an extra dimension added for each cell to denote the batch annotation. Through training process, we can get a low dimensional approximate posterior distribution $q(z_n \mid x_n, s_n)$ conditional on s_n. At inference stage, the low dimensional representation of each cell z_n is taken to reconstruct the expression data through non-linear mapping. The like likelihood function of gene g from cell n is $\pi(x_{ji} \mid \rho_j, \mu_{ji}) = \frac{\rho_j^{1/2}}{\sqrt{2\pi}} exp\{-\rho_j(x_{ji}-\mu_{ji})^2/2\}$. To reduce randomness, we decompose the parameter matrix $\boldsymbol{\mu}$ of the reconstructed data matrix **X** rather than directly on **X**. Gene correlation prior is used to guide the variable selection at gene level, and results are presented as matrix **W** with each column denoting a module. Matrix **Z** is the result at cell level with each row representing a functionally-related cell group.

Here we propose a novel single-cell biclustering method (scBC) that can handle the problems mentioned above and provide a more reliable result. We utilize a variational

autoencoder (VAE) to model gene expression in single cells, enabling us to gracefully remove batch effects and impute missing data[19]. By estimating the variational posterior distribution, we can obtain a low-dimensional representation of each cell that is conditioned on the batch annotation, enabling us to obtain batch-corrected expression through the generating process. Additionally, we can manually control the procedure of "dropout" during the generating process, leading to the pre-dropout imputed expression. By reconstructing the original data matrix in this way, we can obtain more precise results when conducting biclustering. Furthermore, we incorporate existing biological information (e.g., gene interaction and regulation) into the biclustering procedure through the Bayesian framework, which guides variable selection to more likely capture pathway information and true biological signals[20]. The flowchart of our procedure is depicted in Fig. 1.

Results

scBC outperforms other methods on FGM detecting

To investigate whether scBC can detect biologically meaningful functional gene modules, we analyzed four highly heterogeneous single-cell datasets obtained from different parts and tissues of the human body under different pathological conditions (purified Peripheral Blood Mononuclear Cells dataset, PBMC; cardiac cells with annotation from Heart Cell Atlas, HEART; Single cell RNA sequencing of lung adenocarcinoma, LUAD and single cell RNA sequencing of primary breast cancer, BC). Figure 2a illustrates the preprocessing procedure, and the STAR Methods section provides further details. To verify the feasibility of our method, we compared scBC with six state-of-the-art biclustering algorithms, namely CC[21], xMotifs[22], FABIA[13], Bimax[23], plaid[24] and GBC[20] to see how scBC outperforms alternative methods. For methods that need to set the number of biclusters in advance (eg. Xmotif, CC, Bimax, FABIA, GBC and scBC), we just set the max number of biclusters as the number of cell types (based on cell lable). We then conducted GO enrichment analysis for each FGM detected by each method (see STAR Methods), and used -log10(p) (BH adjusted) as the enrich score. Methods that failed to detect any bicluster were assigned a score of zero (Fig. 2b-d).

We found that the FGMs detected by scBC were consistently more significant than those detected by other algorithms, even in highly heterogeneous settings (Fig. 2b, Table S1). In the PBMC dataset, all methods were able to capture the specific FGM, indicating a relatively simple data structure. Among these methods, scBC performed the best, followed by CC. GBC, Xmotif, and FABIA also performed well, but plaid and Bimax had unsatisfactory results. In the HEART dataset, Xmotif and Bimax failed to detect any biologically meaningful gene modules, and plaid was almost invalid (Fig. 2b), suggesting that biclustering on cardiac tissue data may be more challenging. GBC and CC had passable performance, second only to scBC. Similar results were observed in the LUAD dataset, which is related to the tumor-associated immune microenvironment. In the BC dataset, where sample size was minimal (only up to three

hundred cells after negative sampling) and tumor cells were mixed with normal cells, most methods failed to detect meaningful functional gene modules (Fig. 2b). Even under such challenging conditions, scBC was able to identify more biologically meaningful functional gene modules. Despite both CC and GBC had similar excellent performance in the first few conditions, CC was far inferior to GBC in BC dataset. Overall, scBC demonstrated robust superior performance in FGM detection across highly heterogeneous conditions.

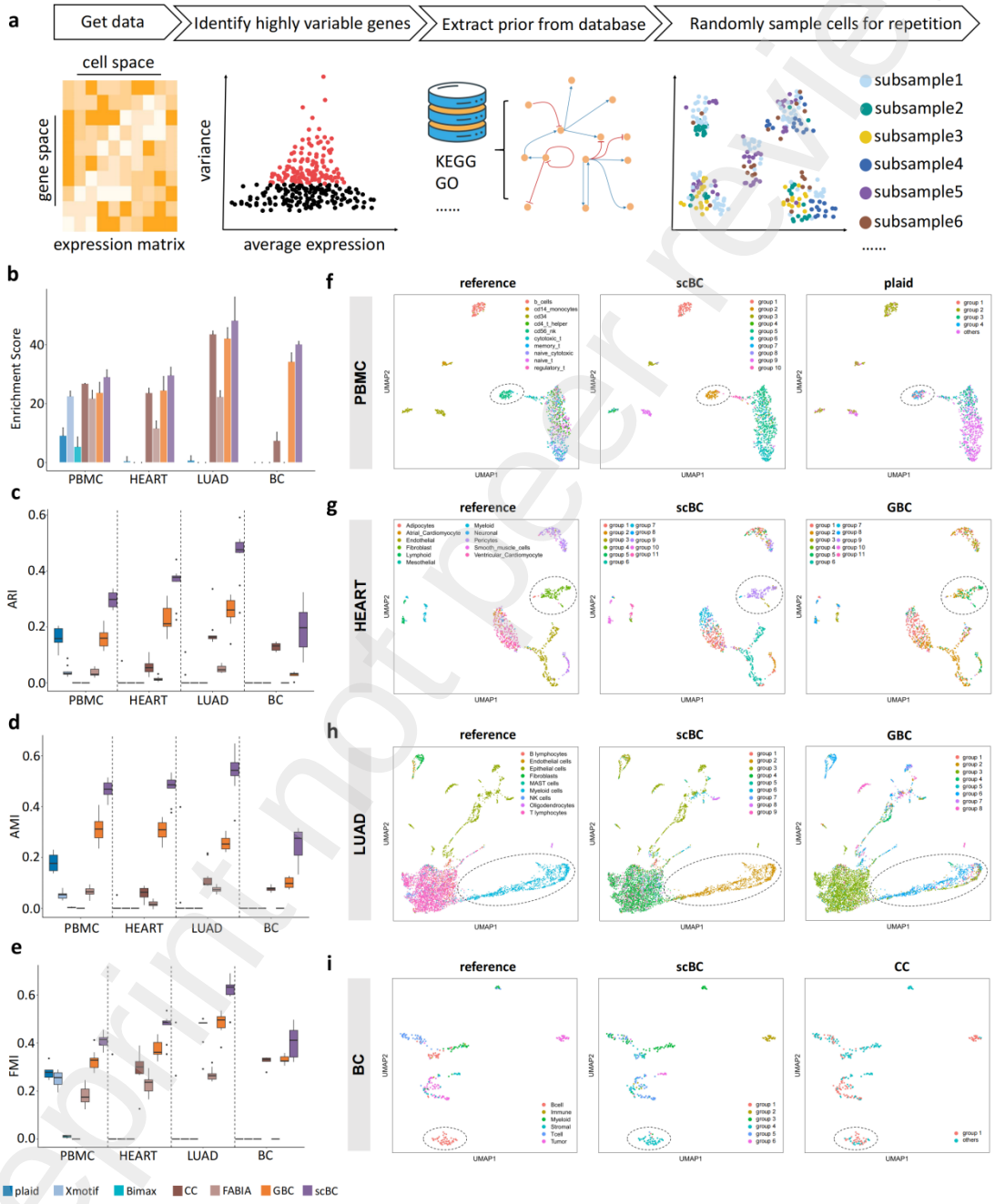


Fig. 2 | scBC outperforms other methods on FGM detecting and cell clustering in 4 real-world datasets. a, The preprocess procedure of the four datasets. In each dataset, highly variable genes are filtered out and used to extract prior co-expression information in database like GO or KEGG. Cells are sampled along with the highly variable genes to generate 10 subsampled datasets for

repetition. **b**, Enrichment score of different methods in four highly heterogenous datasets. The x-axis represents different datasets and y-axis represents the enrichment score (-log10(p), BH adjusted) of different methods. Error bar stands for standard deviation of 10 subsample's results. **c-e,** Benchmarking clustering results at cell level with different criteria in the four datasets. **f-i**, Cell representation of UMAP dimentionality reduction. In addition to the reference labels, shown here are the methods with highest ARI (scBC) and second-highest ARI in the last subsample dataset. The whole comparison of cell clustering can be found in Fig. S2-5. Some methods output too many categories so that we merge some into "others" whose number of cells is less than one percent of the total sample. Highlighted with black dotted lines are cell populations that are correctly identified by scBC but not identified by the second-best method.

Benchmarking clustering results at cell level

Intuitively, cell groups identified by biclustering are more functionally related since each group corresponds to a similar functional gene module. Functionally related cells are also naturally more likely to belong to the same cell type since they have similar functions, although sometimes it is not necessarily correct. In this study, we investigated the clustering performance at the cell level to determine if scBC provides more meaningful clustering results, even when focusing solely on the cell-level clustering results. As mentioned earlier, biclustering results may have overlap between each bicluster, which can result in single cells belonging to different groups. However, the results from scBC and GBC enable us to assign each cell to its most involved groups. For the remaining methods with less well-defined cell-level clustering results, we used the Markov clustering algorithm (MCL)[25] to transform the biclustering results, fully utilizing the information from the biclustering results (see STAR Methods). We used the Adjusted Rand Index (ARI)[26, 27], the Fowlkes Mallows score (FMI)[28], and the Adjusted Mutual Information (AMI)[29] as recommended metrics to quantify the agreement between clusters (see STAR Methods). Their values range from -1 to 1, with higher values indicating better performance. We evaluated the clustering performance at the cell level using the four real-world datasets.

The cell clustering results obtained by scBC are consistently more precise than those of other methods across different heterogeneous conditions (Fig. 2c-e, Table S2-S4). We found that Bimax performed the worst not only in FGM detection, but also cell clustering in all datasets. Although CC performed well in FGM detection in the PBMC dataset, it was unable to perform cell clustering tasks simultaneously. Plaid had the second-best performance on ARI, but was not as good as GBC on AMI and FMI (Fig. 2c-e). In PBMC dataset, multiple T cell subtypes are mixed together and difficult to distinguish, while several other cell types are highly differentiated. This pattern was successfully captured by scBC but not by CC (Fig. 2f). In the HEART dataset, scBC still outperformed other methods, with GBC coming in second. Plaid, Xmotif, and Bimax were invalid in this dataset (Fig. 2c-e). Similar results were found in the LUAD dataset. We observed that scBC captures certain features of these cell populations that GBC does not (Fig. 2h). In the BC dataset, CC ranks second only to scBC on ARI and FMI but is inferior to GBC on AMI (Fig. 2c-e). Plaid, Xmotif, Bimax, and FABIA fail to cluster cells

into functionally related groups as they cannot detect FGMs in this dataset. These results demonstrate that scBC can capture the complex patterns involved in clustering functional cell groups and is more robust and precise than other methods across heterogeneous datasets.

scBC performs best on a bicluster level

Gene co-expression patterns differ across different cell types. These complex gene-cell correlations are of particular interest to us. When we compare the performance of different biclustering methods at the bicluster level, we pay more attention to whether the method can detect cell subgroups with similar FGMs and present these cells and FGMs at the same time. Once such a biosignal is found, we can make guidelines for downstream analysis. In this study, we introduced two evaluation methods, 1-CE and F score, to compare the performance of our scBC with six other state-of-the-art biclustering methods (see STAR Methods for simulation detail). We used simulated datasets with different dropouts under varying sc ales to elucidate how the performance of these methods varies along with the conditions (Fig. 3, Table S5-10).

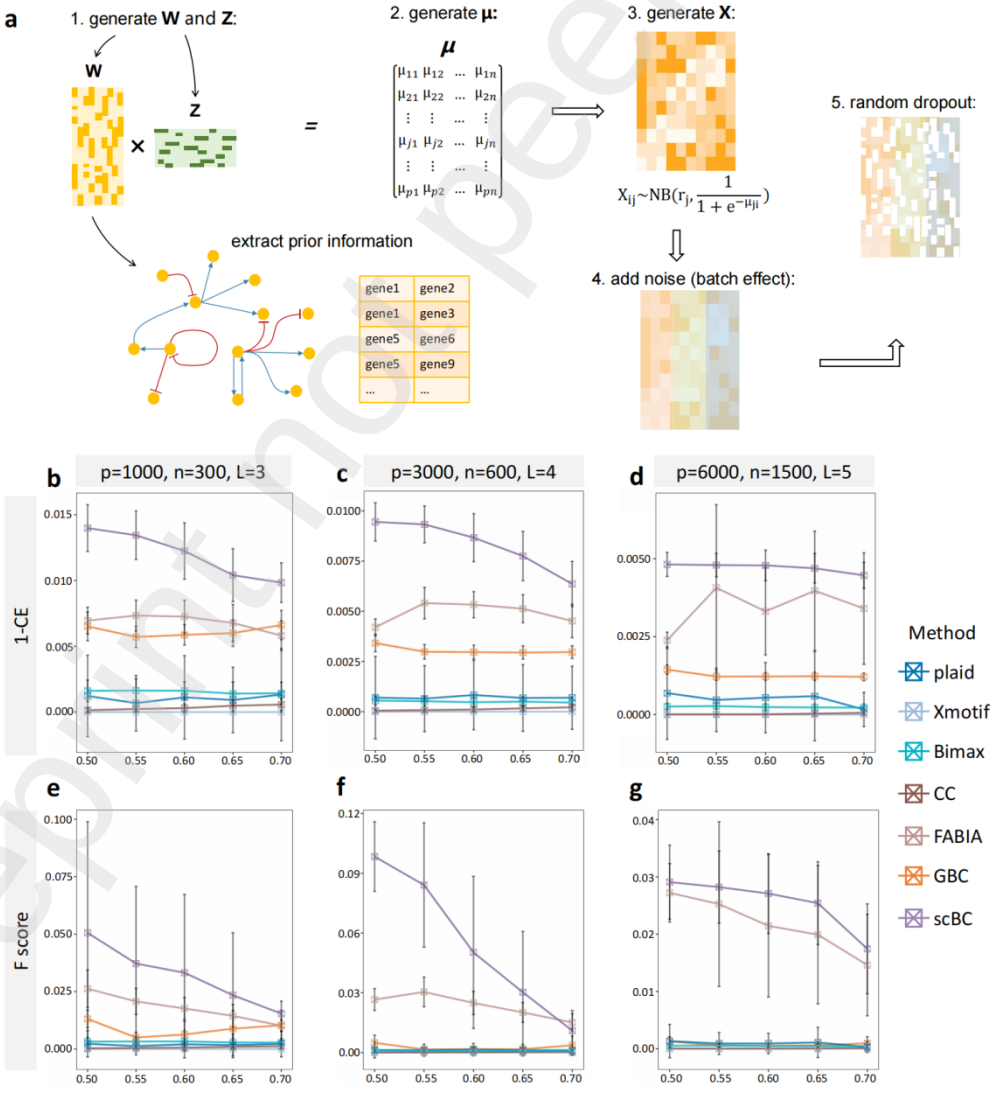


Fig. 3 | scBC performs best on a bicluster level. a, Data simulation process. The parameter **μ** is computed by the multiplicative model **μ** = **WZ**. The prior edge information is generated along with **W**. When generating **X**, each element is generated from $\mathrm{NB}\left(r_j, \frac{1}{1+e^{-\mu_{ji}}}\right)$. To simulate different batches, we divided the dataset into three parts, with different intensities of noise. The implementation of dropout is to perform Bernoulli censoring. See STAR Methods for detail. **b-g,** Performance of different methods under different conditions. For each plot, x axis represents the dropout rate and y axis represents the quantified performance (1-CE or F score). We ran 100 independent simulations for each setting, the data points represent mean value and the error bars represent the standard deviation calculated across repeated simulations.

Since the single-cell data itself is highly sparse, our simulation started directly from dropout=0.5 and explored with a step size of 0.05. It can be observed that scBC was significantly ahead of other methods across different scales with respect to 1-CE in most cases, with FABIA coming in second (Fig. 3b-c), indicating that FABIA may be more concerned about the overall effect of biclustering. However, when the scale of datasets became larger, the performance of FABIA became extremely unstable (Fig. 3d). Methods like CC and Xmotif were invalid mostly (Fig. 3b-d). As expected, the ability of scBC to detect biclusters decreases as dropout rate and data scale increase (Fig. 3b-d). The F score of scBC was quite unstable when the data scale was too small, although it still had the best performance in terms of mean performance (Fig. 3e). When the dropout rate was extremely high, the performance of scBC may be exceeded by FABIA with respect to F score (Fig. 3f). The performance with respect to F score of FABIA also became unstable when the data scale became large, as was observed in 1-CE. Overall, the results at the bicluster level indicate that scBC is more reliable in uncovering complex gene-cell correlation patterns than other methods.

scBC uncover the pathway perturbation in AD progression

Neuropsychiatric disorders involve complex polygenic determinants as well as brain alterations[30]. Biclustering methods can reveal cell population-specific gene co-expression patterns and discover potential gene-cell connections, making them inherently more suitable for the analysis and mining of complex polygenic disorders such as neurodegenerative diseases. At the same time, single-cell-level resolution is critical for neurodegenerative diseases like Alzheimer's disease because changes in gene expression are related to specific cell types[3]. Therefore, scBC is more reliable for analyzing the single-cell data of diseases with complex traits due to its excellent performance.

Alzheimer's disease (AD) is a neurodegenerative disorder associated with aging, characterized by the accumulation of amyloid plaques and neurofibrillary tangles in the brain parenchyma. Recent research, using a single-nucleus RNA sequencing (snRNA-seq) dataset from Alzheimer's disease patients, has shown that AD is a complex disease involving multiple brain cell types, as evidenced by marker gene expression[3]. In this study, we aim to investigate further transcriptomic perturbations

during AD progression using network gene biomarkers identified by our scBC model. This dataset includes 48 post-mortem human brain samples, with or without Alzheimer's disease. The pathology groups are defined based on several pathological traits (Table S11): 'no-pathology' (no amyloid burden, no neurofibrillary tangles, and no cognitive impairment), 'early-pathology' (amyloid burden, but modest neurofibrillary tangles and modest cognitive impairment), and 'late-pathology' (higher amyloid burden, increased neurofibrillary tangles, global pathology, and cognitive impairment) (Fig. 4a). After subsampling, we ensured that cells from different donors were well-blended and not dominated by any one donor or biased by sex (Fig. 4b-d). We also ensured that cells of the same type across individuals were consistent (Fig. 4e-f). To make sure that the results of multiple biclustering analyses corresponded with each other, we matched biclusters in different pathological progression stages and merged some FGM according to the degree of overlap in gene sets (Fig. 4g, STAR Methods). We found that the overlap of each FGM, as expected, is considerable (Fig. 4h).

It is commonly believed that multiple FGMs can be simultaneously active in one cell type, and a single FGM can be shared across different cell types, but the composition percentage in different cells will vary. Our method, single-cell biclustering (scBC), captured this structure perfectly (Fig. 5a-f), indicating that it is very suitable for such analysis. In this study, we focused on the perturbation of FGMs for each cell type during the progression of AD to gain a better understanding of the mechanisms underlying AD and to provide potential recommendations for therapy. To clarify the functional changes represented by specifically altered FGMs, we performed enrichment analysis of specific gene sets before and after a progression stage (see STAR Methods for details) to identify associated pathways that are disrupted during the progression (Fig. 5g-k). The complete enrichment analysis results can be found in Table S12.

For astrocytes and oligodendrocyte precursor cells, FGM perturbation occurs almost only at the early pathology (Fig. 5a, f), indicating transcriptional patterns has largely changed in these two cell types before individual develop severe pathological features, which is consistent with previous research[3]. Inhibitory neurons' change in FGM throughout the disease progression is minimal (Fig. 5c), indicating that this cell type does not have many alterations in transcriptional patterns during AD progression.

Astrocytes are involved in neuronal trophic support, extracellular ion homeostasis and brain fluid balance[31]. Energy metabolism is largely altered in AD astrocyte (Fig. 5g), indicating the inflammatory state of the brain following injury and neurodegeneration since astrocytes are a central driver of energy homeostasis in the brain, which is also mentioned in previous studies[31, 32]. Consistent with previous studies, we found ion transporters are dysregulated in AD astrocytes (Fig. 5g). At the same time, we also found pathways related to myelination and neuron ensheathment are altered with the progression of AD (Fig. 5g).

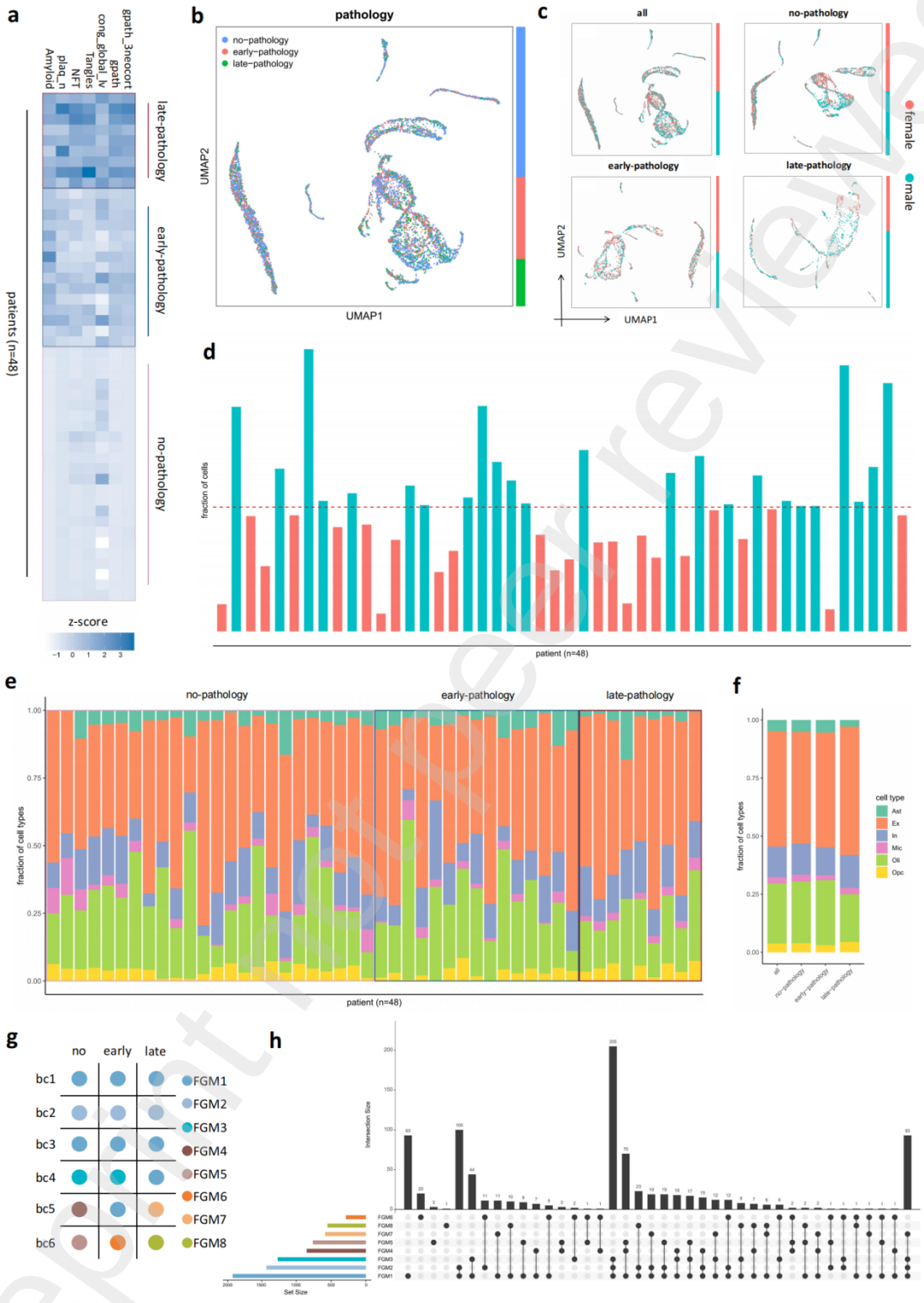


Fig. 4 | overview of the subsampled AD dataset. a, Clinic-pathological variables (columns) of 48 individuals (rows). Amyloid, overall amyloid level; plaq_n, neuritic plaque burden; NFT, neurofibrillary tangle burden; Tangles, neuronal neurofibrillary tangle density; cogn_global_lv, global cognitive function (last valid score). Since the lower the value, the more serious the disease, here we use its opposite number, in order to be consistent with other indicators, so as to more intuitively show the

differences between different pathology groups. gpath, global AD-pathology burden; gpath_3neocort, global measure of neocortical pathology; **b,** UMAP visualization of all cells (n=7063) indicates cells from different donors of different pathological states are well blended. Color bar at the right side represents the fraction composition of cells under different pathology. **c,** UMAP visualization of cells from all sample, no-pathology group, early-pathology group and late-pathology group, but colored by sex. Color bar at the right side represents the fraction composition of cells of different sex. **d,** The proportion of cells provided across individuals (columns). Bars represent the fraction of cells corresponding to each individual. Bar color indicates whether the corresponding value exceeds (blue-green) or does not exceed (rose red) the average value measured across all the donors in the row. Red dashed line indicates the average. **e,** Fraction of cells of each type isolated from each individual (columns; n = 48). **f,** Fraction of cells of each type isolated across all (n = 48), no-pathology (n = 24), early-pathology (n = 15) and late-pathology (n = 9) individuals. **g,** Merge result between different biclusters. Gene sets from different biclusters in different pathology groups labeled with the same color are combined as a new FGM. **h,** The overlap of FGMs**.** The FGM marked with a solid black dot below the bar graph indicates that it is included in the comparison, and the FGM marked with a black transparent dot indicates that it is not included. For example, the first bar chart indicates the number of genes appearing in FGM1 but not appearing in any other FGMs is 93. This result shows the overlap between different FGMs is considerable.

It has been found that gliogenesis and neuron ensheathment-related pathways are largely impaired in AD progression[33]. We found the same conclusion in the progression of AD pathology in excitatory neurons (Fig. 5h). However, the perturbation of FGM composition in excitatory neurons from normal to early pathology is very subtle compared to the change from the early stage to the late stage (Fig. 5b), indicating that the rate at which the cells become abnormal may be slow at first and then fast, at least from the point of view of gene module. Another continuous change in excitatory neurons in AD progression is a general dysregulation in kinase activity (Fig. 5h), which is closely related to neuronal DNA damage, well known to occur in AD neurons[34]. Previous studies mentioned that the immune response is also affected in the progression of AD[31]. Here we found the only gene specific in latter one in the change of AD from early to late pathology is VSIG4 (see Table S12), the encoded protein by which may be a negative regulator of T-cell responses and largely have to do with impaired immune response (Fig. 5h). We also found that cellular cation homeostasis pathway and synapse function are altered in the progression from normal to early pathology (Fig. 5h).

Similar to excitatory neurons, pathways associated with kinase activity are also continuously altered throughout AD progression in microglia (Fig. 5i). Cytokine-mediated signaling pathway is altered in early pathology (Fig. 5i), which may be related to changes in the immune response in AD progression and is also in accordance with previous research[31]. Pathways related to gliogenesis and myelination also altered throughout the disease progression in microglia (Fig. 5i), which is similar to astrocytes and excitatory neurons. Cell migration-related pathways are dysregulated in the early AD microglia (Fig. 5i), which is also consistent with several studies[3, 5, 35, 36] and

largely related to microglial plaque clustering phenotypes, a phenomenon of inappropriate interactions with amyloid. Response to fatty acid becomes odd in early AD microglia (Fig. 5i), which is also an indicator of lipid metabolism dysfunction. We also find cell chemotaxis becomes abnormal in late AD microglia (Fig. 5i), indicating an inflammatory state in AD microglia.

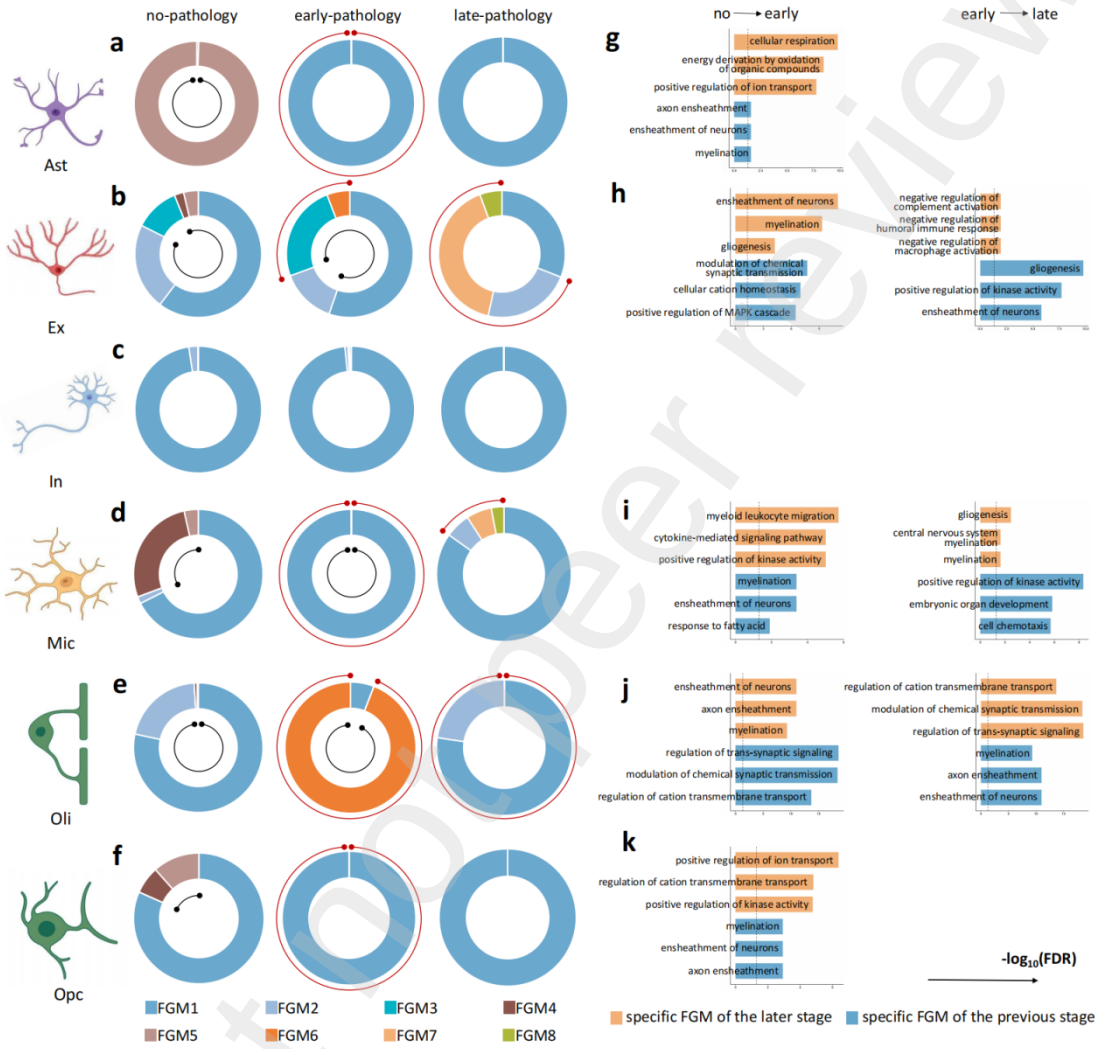


Fig. 5 | scBC uncover the pathway perturbation in AD progression. a-f, perturbation of FGM composition in each cell type during AD progression. Each pie chart quantifies the FGM composition of a cell under a specific progression condition. The outer red circles indicate FGMs whose composition is increasing compared to the previous stage; The inner black circle represents FGMs whose composition is decreasing compared to the latter stage. **g-k**, results of enrichment analysis of FGMs altered in two phases. Orange represents the specific FGM of the later stage, and blue represents the specific FGM of the previous stage, both representing the set of genes that are perturbated during the progression.

We observed that in oligodendrocytes, the main changes in functional gene modules (FGMs) during AD progression occurred in myelination-related and synaptic signaling-related pathways (Fig. 5j). Since memory preservation is thought to require new myelin formation, the impaired capacity of oligodendrocytes to adaptively monitor

neural activity and facilitate myelin remodeling may govern cognitive decline in AD[37]. Moreover, synaptic signaling and axon development are critical for the transmission of excitement in the nervous system, and dysregulation of these processes can result in slower propagation of neural excitation. The changes in FGMs in oligodendrocytes are directly related to the reduction of nervous system excitability. Previous research also suggests that changes in oligodendrocytes may affect the function of other cells in the central nervous system[38-41]. Thus, targeting oligodendrocytes may be a promising strategy for the treatment of AD and other neurological disorders.

Oligodendrocyte precursor cells (OPCs), which are distributed throughout grey and white matter, are thought to dynamically sense and modulate neural activity[40] like oligodendrocytes do. So not surprisingly, pathways related to myelination and the ensheathment of neurons become abnormal along with the progression of AD (Fig. 5k). Pathways related to ion transportation are also dysregulated, providing supports for previous findings that genes related to ion channels are dysregulated in AD OPCs[3, 5]. Besides, pathways related to kinase activity and cellular cation homeostasis are altered at the early stage in AD progression (Fig. 5k).

Except for inhibitory neurons, which did not change significantly throughout disease progression, several other cell types exhibited specific functional gene module (FGM) perturbations, highlighting the importance of single-cell analysis. Notably, we observed that pathways related to myelination and gliogenesis were more or less altered across all these cell types, indicating similar alterations among AD-associated cells and suggesting that AD progression is largely related to dysregulation of this pathway, which is further confirmed in a recent study[42].

Discussion

Molecular biomarkers have been widely used in clinical practice to identify diseases, but they often suffer from low coverage and high false-positive or false-negative rates, limiting their further application[43]. Network biomarkers, also known as module biomarkers, have attracted attention as a more robust form of biomarker than individual molecules for characterizing diseases[44, 45]. This is particularly important for analyzing single-cell data, which is inherently more complex than bulk tissue data due to the heterogeneity of individual cells within a sample. However, network biomarkers are usually cell-specific and may change during the disease progression. To detect cell-specific network biomarkers, we developed scBC, a single-cell Bayesian biclustering method that combines variational autoencoder (VAE) for batch removal and data imputation with matrix factorization-based Bayesian biclustering for utilizing known biological information. Our method outperforms other state-of-the-art methods in finding functional gene modules (FGMs), discovering functionally-related cell groups, and detecting cell-gene correlation patterns in highly heterogeneous single-cell RNA sequencing (scRNA-seq) datasets and simulated data. This makes it well-suited for analyzing diseases with multifactorial etiologies whose functionally-

related potential cell groups are finely divided, cell-type conditioned gene co-expression patterns are complicated, and cell-gene correlation changes throughout the disease progression.

In this study, we applied scBC to a snRNA AD dataset to explore how the transcriptional functional modules of each cell type change as the disease progresses. Our results further confirmed the complex interplay of virtually every major brain cell type in Alzheimer's disease[3, 33]. We found that FGM composition largely changed in astrocytes and oligodendrocyte precursor cells before individuals developed severe pathological features. However, inhibitory neurons showed minimal changes in FGM throughout the disease progression, indicating that this cell type does not have many alterations in transcriptional patterns during AD progression. A consistent FGM perturbation across all other cell types, except inhibitory neurons, was the alteration in pathways related to myelination and gliogenesis, suggesting that this pathway may play a decisive role in the progression of AD.

Specific to each cell type, energy metabolism and ion transporters are dysregulated in AD astrocyte, indicating the inflammatory state of the brain following injury and neurodegeneration. The perturbation of FGM composition in excitatory neurons from normal to early pathology is very subtle compared to the change from the early stage to the late stage, indicating that the rate at which the cells become abnormal may be slow at first and then fast. Another continuous change in excitatory neurons in AD progression is the general dysregulation in kinase activity, which is closely related to neuronal DNA damage. Besides, immune response, cellular cation homeostasis and synapse function are also altered in AD excitatory neurons. Microglia shares a similar alteration in kinase activity with excitatory neurons throughout AD progression. Pathways like immune response-related cytokine-mediated signaling, amyloid interaction-related cell migration and lipid metabolism-related fatty acid response are dysregulated in the early AD microglia. Cell chemotaxis becomes abnormal in late AD microglia, indicating an inflammatory state in AD microglia. Oligodendrocyte is a cell that needs to be more focused on for disease treatment since FGM perturbations in such a cell type are mainly concentrated in myelination-related and synaptic signaling-related pathways, directly related to the reduction of the excitability of the nervous system. At last, pathways related to ion transportation, kinase activity and cellular cation homeostasis are dysregulated in oligodendrocyte precursor cells.

In the context of high-throughput sequencing data, network biomarker-based analytical methods preserve the complex co-expression or co-regulation patterns in the gene module and is more robust to the analysis of complex diseases. We believe that scBC, as a novel technique for cell-specific network biomarkers detection, creates an unprecedented opportunity towards effectively delineating mechanisms of complex diseases at single-cell resolution, providing advice on the treatment of such diseases. However, although the network biomarker may contain complex co-expression or co-regulation patterns, its internal precise and quantitative regulatory

relationship has not been clarified. Future research can focus on the explanation of the regulatory relationship within the cell-specific network structure, so as to have a more accurate inference on the principle of FGM perturbations during disease progression.

Conclusion

Our method, scBC, is a powerful tool for detecting cell-specific network biomarkers in highly heterogeneous single-cell RNA sequencing datasets. It outperforms other state-of-the-art methods in finding functional gene modules (FGMs) and discovering functionally-related cell groups, making it a top contender for analyzing diseases with multifactorial etiologies, where cell-type conditioned gene co-expression patterns are complicated and cell-gene correlation changes throughout disease progression. To demonstrate its potential, we applied scBC to a snRNA AD dataset, revealing FGM perturbations in each cell type as the disease progresses. Our findings support previous studies and offer new insights, such as minimal changes in FGM for inhibitory neurons and a consistent FGM perturbation in pathways related to myelination and gliogenesis across all other cell types. These results suggest that scBC provides unprecedented opportunities for effectively delineating the mechanisms of complex diseases at single-cell resolution on a gene-module-based view, providing valuable insights for the treatment of such diseases.

STAR Methods

Details for Data Reconstruction Using Variational Inference

Taking advantage of recent work by Romain et.al[19], here we also adopt the idea of using variational inference to estimate the posterior distribution for the low-dimensional, latent variables z_n for each cell n which should reflect biological differences among c ells. To remove the nuisance variation due to technique factors such as batch effects, it's reasonable to model the sampling distribution conditioned on the batch annotations s_n[46, 47]. That is, the observed expression x_n^g of each gene g in each cell n is drawn from $\mathrm{p}(x_n^g|z_n,s_n)$. Considering the substantiate amount of dropouts in the scRNA-seq data, here we choose zero-inflated negative binomial(ZINB) distribution as the proper sampling distribution which is recommended by Grün et.al[47].

Now we can present in-detail steps of our pre-defined generating process of how we can get the gene expression of each cell. The generating process can be concluded as follows:

$$z_n \sim N(0, \boldsymbol{I})$$

$$\rho_n \sim f_{expect}(z_n, s_n)$$
$$w_n^g \sim Gamma(\rho_n^g, \theta^g)$$
$$y_n^g \sim Poisson(l_n w_n^g)$$
$$h_n^g \sim Bernoulli(f_{drop}(z_n, s_n))$$

Where z_n is the low-dimensional, latent variable. Here we use a standard multivariate normal prior for z because it can be reparametrized in a differentiable way into any arbitrary multivariate Gaussian random variable, which is extremely helpful in the inference process. f_{expect} is a neuron network which maps the latent space and batch annotations of each cell back to the full dimension of the gene expression: $\mathbb{R}^{d+1} \rightarrow \mathbb{R}^{G}$. At the generating stage, f_{expect} is constrained by a softmax activation function at the last layer so that each element of ρ_n sum up to 1 when we decode the gene expression data. So it is reasonable to deem that ρ_n denotes the mean proportion of transcripts expressed across all genes[19]. This also makes it easy to reconstruct the expression data at an arbitrary scale only to specify the cell-specific library size l_n. $f_{drop}(z_n, s_n)$ is also a neuron network which maps the latent space and batch annotations of each cell to their respective dropout probabilities. w_n^g and y_n^g are two intermediate variable and it can be shown that through this process, h_n^g is a r.v. following ZINB distribution[47] with mean $l_n \rho_n^g$, gene-specific dispersion θ^g and zero-inflation probability $f_{drop}(z_n, s_n)$.

When we conduct data reconstruction to get the batch-removal, imputed gene expression data, we only take advantage of the intermediate variable ρ_n and scale it to our expected library size. That is, multiplying it by a given parameter l_n, which we just use the empirical library size (total number of transcripts per cell) of each cell throughout our experiments. But one should notice we can re-scale it to any expected library size if additional information is given.

Model Training at Learning Stage

Here we introduce a recognition model $q_\varphi(z_n|x_n, s_n)$: an approximation to the intractable true posterior $p_\theta(z_n|x_n, s_n)$. The marginal likelihood can be written as:

$$\log p_\theta(x_n|s_n) = D_{KL}(q_\varphi(z_n|x_n, s_n)||p_\theta(z_n|x_n, s_n)) + L(\boldsymbol{\theta}, \boldsymbol{\varphi}; x_n)$$

Where $L(\boldsymbol{\theta}, \boldsymbol{\varphi}; x_n) = E_{q_\varphi(z_n|x_n,s_n)} [-\log q_\varphi(z_n|x_n,s_n) + \log p_\theta(z_n,x_n|s_n)]$. Since the KL-divergence is always non-negative. We have:

$$\log p_\theta(x_n|s_n) \geq E_{q_\varphi(z_n|x_n,s_n)} [-\log q_\varphi(z_n|x_n,s_n) + \log p_\theta(z_n,x_n|s_n)]$$

The evidence lower bound (ELBO) $L(\boldsymbol{\theta}, \boldsymbol{\varphi}; x_n)$ can also be written as:

$$L(\boldsymbol{\theta}, \boldsymbol{\varphi}; x_n) = E_{q_\varphi(z_n|x_n,s_n)} [\log p_\theta(x_n|z_n,s_n)] - D_{KL}(q_\varphi(z_n|x_n,s_n)||p_\theta(z_n|s_n))$$

Optimizing the ELBO means optimizing both the variational parameters $\boldsymbol{\varphi}$ and generative parameters $\boldsymbol{\theta}$ at the same time. Assuming the true latent variable z_n is batch-free (independent with batch annotation s_n) and the prior follows standard multivariate Gaussian distribution, we can get the closed-form expression of the derivative of $D_{KL}(q_\varphi(z_n|x_n,s_n)||p_\theta(z_n|s_n))$. To get the low-variance Monte Carlo estimation of the gradient of term $E_{q_\varphi(z_n|x_n,s_n)} [\log p_\theta(x_n|z_n,s_n)]$, we use the reparameterization trick in the learning stage[48]:

$$\tilde{E}_{q_\varphi(z_n|x_n,s_n)} [\log p_\theta(x_n|z_n,s_n)] \cong \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x_n|g_\varphi(\epsilon^l,x_n),s_n)$$

Where $g_\varphi(\epsilon^l,x_n)$ is a differentiable transformation to reparameterize the random variable $z_n \sim q_\varphi(z_n|x_n,s_n)$ and $\epsilon \sim p(\epsilon)$ is an auxiliary noise variable.

For a single data point $x_n^{(i)}$ (cell i) we have:

$$\tilde{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}; x_n^{(i)}) = \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x_n^{(i)}|z^{(i,l)},s_n) - D_{KL}(q_\varphi(z_n|x_n^{(i)},s_n)||p_\theta(z_n))$$

Where $z^{(i,l)} = g_\varphi(\epsilon^{(l)},x_n^{(i)})$ and $\epsilon^{(l)} \sim p(\epsilon)$.

At learning stage, we use mini-batch stochastic optimization to optimize the ELBO, suppose our dataset contains N cells and the size of each mini-batch is M, we can get the estimator of marginal likelihood lower bound of the stochastic mini-batch:

$$L(\boldsymbol{\theta}, \boldsymbol{\varphi}; x_n^{(i)}) \cong L^M(\boldsymbol{\theta}, \boldsymbol{\varphi}; x_n^{(i)}) = \frac{1}{M} \sum_{i=1}^{M} \tilde{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}; x_n^{(i)})$$

When M is large enough, Diederik et al[48] found that the number of samples L per datapoint can even be set to 1, hence decrease the time consumption when conduct expectation estimation for $E_{q_\varphi(z_n|x_n,s_n)} [\log p_\theta(x_n|z_n,s_n)]$. Throughout our experiment we set M=128 data points to guarantee the large-sample requirement. We use Adam optimizer with learning rate=0.01. We also use deterministic warm-up and batch normalization during learning to learn an expressive model which is recommended by Sonderby et al[49].

Bayesian Biclustering Incorporate biological information

After reconstructing the original expression matrix, we can conduct biclustering procedure to detect condition-specific FGMs and identify cell subpopulations with distinct functions. Relevant studies have shown that if we can introduce existing biological information (such as the metabolic pathways from the KEGG database) into the process of biclustering, then the accuracy of the biclustering results will be

improved[20, 50-54]. Therefore, we adopt a Bayesian analysis framework, which can introduce prior information to guide variable selection.

Suppose our reconstructed data matrix is $\mathbf{X}$ of size $p \times n$, where p represents the number of genes and n is the number of cells. In order to reduce randomness, here we do not directly decompose the data matrix $\mathbf{X}$, but decompose its parameter matrix. We denote the parameter matrix of $\mathbf{X}$ as $\boldsymbol{\mu}$ (eg. mean) and decompose it: $\boldsymbol{\mu} = \mathbf{m}\mathbf{1}^{\mathbf{T}} + \mathbf{W}\mathbf{Z}$, where $\mathbf{m}$ is a $p \times 1$ bias vector, $\mathbf{1}$ is a $n \times 1$ vector of 1, $\mathbf{W}$ is a $p \times L$ matrix, $\mathbf{Z}$ is a $L \times n$ matrix. Since the observation of gene j from cell i x_{ji} is generated independently, the likelihood function of $\mathbf{X}$ is the product of the likelihood functions of each independent observation. Since we have assumed the generating process of the reconstructed data, we set x_j to be a random variable that follows Gaussian distribution with a likelihood function π_j in the discussion following on.

The likelihood function of an individual observation is:

$$\pi_j(x_{ji}|\mu_{ji},\rho_j) = \frac{\rho_j^{1/2}}{\sqrt{2\pi}} e^{-\rho_j(x_{ji}-\mu_{ji})^2/2}, x_{ji} = 0,1..... \quad (1)$$

Now we discuss how to introduce prior information. To obtain a sparse estimate of W, first use the Laplace prior on the matrix **W**:

$$log\pi(\boldsymbol{W} \mid \boldsymbol{\lambda}) = C + \sum_{j,l} log\,\lambda_{jl} - \sum_{j,l} \lambda_{jl}|w_{jl}|$$

Here the prior parameter λ controls the degree of shrinkage of w, so the prior of the multivariate normal is applied on λ next:

$$log\,\pi(\boldsymbol{\alpha} \mid \boldsymbol{\Omega}) = C_{v_2} + \frac{L}{2} log|\boldsymbol{\Omega}| - \frac{1}{2v_2}\sum_l (\boldsymbol{\alpha}_l - v_1\mathbf{1})\boldsymbol{\Omega}(\boldsymbol{\alpha}_l - v_1\mathbf{1}) \quad (2)$$

Where $\alpha_{jl} = \log\lambda_{jl}$, $\boldsymbol{\alpha}_l = (\alpha_{1l},...,\alpha_{pl})^T$, v_1 and v_2 are hyperparameters. Finally, we apply a suitable prior to the precision matrix $\boldsymbol{\Omega}$ to connect the correlated λ, which will be defined as:

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 + \sum_{j \neq 1} \omega_{1j} & -\omega_{12} & \cdots & -\omega_{1p} \\ -\omega_{21} & 1 + \sum_{j \neq 2} \omega_{2j} & \ddots & -\omega_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ -\omega_{p1} & -\omega_{p2} & \cdots & 1 + \sum_{j \neq p} \omega_{pj} \end{bmatrix}$$

This ensures the symmetry and positive definiteness of the precision matrix, and the prior of the ω is:

$$\pi(\boldsymbol{\omega}) \propto |\boldsymbol{\Omega}|^{-\frac{L}{2}} \prod_{(j,k)\in E} \omega_{jk}^{a_\omega - 1} \exp(-b_\omega \omega_{jk})\, 1(\omega_{jk} > 0) \prod_{(j,k)\notin E} \delta_0(\omega_{jk}) \quad (3)$$

$\delta_0(\cdot)$ is the Dirac function centered at 0, $1(\cdot)$ is an indicative function. Therefore, if x_j and x_k are directly connected in the graph G, then (3) will try to make the

precision matrix components ω_{jk} to be non-zero, and make the contraction term λ_{jl} and λ_{kl} related through (2). Since w_{jl} and w_{kl} are subject to a similar degree of contraction under this condition, they tend to be both zero or non-zero at the same time. In other words, if genes j and k are directly connected in a pathway, they are encouraged to be selected together (or not selected together) in bicluster l. Therefore, a standout feature of this approach is that the selected feature set in each bicluster tends to include functional gene module rather than individual genes, resulting in more biologically meaningful results.

Since the **Z** matrix represents the results on the cell set, there is no special pathway information between the samples, so it is sufficient to perform Laplace sparse prior on it:

$$\log \pi(\boldsymbol{Z} \mid \boldsymbol{\xi}) = C + \sum_{l,j} \log \xi_{li} - \sum_{l,j} \xi_{li} |z_{li}|$$

Where ξ is the contraction factor, on which a conjugate prior is applied, i.e. Gamma prior:

$$\log \pi(\xi) = C_{v_3,v_4} + (v_3 - 1) \sum_{l,i} \log \xi_{il} - \frac{1}{v_4} \sum_{l,i} \xi_{li}$$

This is the general framework of the Bayesian biclustering model we use.

MAP estimation for biclustering result

In the optimization stage, we adopt the Pólya-Gamma latent variable proposed by Polson et al.[55]. We use the identity formula provided in Polson et al.[55]:

$$\frac{e^{\mu_{ji} x_{ji}}}{(1 + e^{\mu_{ji} x_{ji}})^{b_{ji}}} = 2^{-b_{ji}} e^{\kappa_{ji} \mu_{ji}} \int_o^\infty e^{-\rho_{ji} \mu_{ji}^2 / 2} \pi_{ji}(\rho_{ji}) d\rho_{ji}$$

Where $\kappa_{ji} = x_{ji} - b_{ji}/2$, $\pi_{ji}(\rho_{ji})$ is of the Pólya-Gamma class $\mathcal{PG}(b_{ji}, 0)$. So (1) can be written as:

$$\pi_j(\mathbf{x_j} | \mu_j) \propto e^{-\frac{1}{2} \sum_i \rho_{ji} (\mu_{ji} - x_{ji})^2} \pi_j^*(\rho_j)$$

Where $\rho_j \sim \mathcal{G}\left(\frac{\zeta_j + n}{2}, \frac{\zeta_j}{2}\right)$, ζ_j is the prior parameter for variance. After the introduction of latent variable ρ, LASSO can be efficiently solved in the M step of the EM algorithm. Here we use dynamic weighted LASSO algorithm to speed up the calculation[56]. Additionally, we utilize maximum a posteriori estimation (MAP) to estimate the parameters, which is defined as:

$$(\hat{\mathbf{W}}, \hat{\mathbf{Z}}, \hat{\mathbf{m}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\xi}}) = \mathrm{argmax}_{W,Z,m,\alpha,\xi} \int\int \pi(\mathbf{W}, \mathbf{Z}, \mathbf{m}, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\rho}, \boldsymbol{\Omega} | \mathbf{X}) \mathrm{d}\boldsymbol{\rho} \mathrm{d}\boldsymbol{\Omega}$$

This can be efficiently solved using the EM algorithm, and the objective function at t

iterations is:

$$\boldsymbol{Q}_t(\boldsymbol{Z},\boldsymbol{W},\boldsymbol{m},\boldsymbol{\alpha},\boldsymbol{\xi}) = -\frac{1}{2}\sum_{i,j}\rho_j^{(t)}(\mu_{ji}-x_{ji})^2 + \sum_{j,l}\alpha_{jl} - \sum_{j,l}\lambda_{jl}|w_{jl}| + v_3\sum_{l,i}\log\xi_{l,i} -$$

$$\sum_{l,i}\xi_{l.i}\left(|z_{li}| + \frac{1}{4}\right) - \frac{1}{2v_2}\sum_l(\boldsymbol{\alpha}_l - v_1\mathbf{1})^T\boldsymbol{\Omega}^{(t)}(\boldsymbol{\alpha}_l - v_1\mathbf{1})$$

Where $\boldsymbol{\mu} = \boldsymbol{m}^{(t-1)} + \boldsymbol{W}^{(t-1)}\boldsymbol{Z}^{(t-1)}$, $\rho_{ij}^{(t)} = E\left(\rho_{ij} \middle| \boldsymbol{X},\boldsymbol{W}^{(t-1)},\boldsymbol{Z}^{(t-1)},\boldsymbol{m}^{(t-1)},\boldsymbol{\alpha}^{(t-1)},\boldsymbol{\xi}^{(t-1)}\right)$ and $\boldsymbol{\Omega}^{(t)} = E\left(\omega_{ij} \middle| \boldsymbol{X},\boldsymbol{W}^{(t-1)},\boldsymbol{Z}^{(t-1)},\boldsymbol{m}^{(t-1)},\boldsymbol{\alpha}^{(t-1)},\boldsymbol{\xi}^{(t-1)}\right)$.

Strong classification of cell group for different biclustering methods

For scBC and GBC, we can directly observe the contribution of each bicluster to the parameter matrix in each cell from the result of the Z matrix. We assign a cell to the most involved cluster, which is determined by the row with the largest absolute value. For the remaining methods, we aim to achieve optimal cell classification results without losing any information from the biclustering results. Due to the high degree of cell overlap in the biclustering results, we convert the cell-level biclustering results into a graph, where cells in the same bicluster are connected by edges. If the occurrences of a pair of cells increase, the weights of the edges between them also increase accordingly. We then apply the Markov clustering algorithm (MCL)[25] to convert the graph into cell-level clustering results. For each method, we set the number of iterations to a value between 1 and 10 that allows the method to achieve the highest adjusted Rand index (ARI). In fact, we observed that the number of iterations required for the best results usually does not exceed 7. After the transformation, each cell is exclusively assigned to a cluster, and we can evaluate the cell-level clustering results using any clustering evaluation criterion.

Datasets and preprocessing

Here we describe all of the datasets and the preprocessing steps used in the current work as follows. The prior information for all the real-world datasets is extracted by *biomaRt* using the highly variable genes.

heart

This is a combined single cell and single nuclei RNA-Seq data of 485K cardiac cells with annotation from Heart Cell Atlas. Here we use a subsampled version provided at https://github.com/YosefLab/scVI-data/blob/master/hca_subsampled_20k.h5ad, which has been filtered down randomly to 20k cells. In our study, we further filtered 1000 highly variable genes using *scanpy* and generate 10 subsampled datasets with each containing 1000 randomly selected cells.

PBMC

This actually is a purified PBMC dataset from[57]. An organized version can be accessed from https://github.com/YosefLab/scVI-data/raw/master/PurifiedPBMCDataset.h5ad. We also conducted a subsampling procedure here: first screen out 853 highly variable

genes using *scanpy*, then generate 10 subsampled datasets with each containing 1000 random selected cells.

LUAD

Single cell RNA sequencing of lung adenocarcinoma from[58], which can be accessed from the NCBI Expression Omnibus database (accession code GSE131907). This is single cell RNA sequencing (scRNA-seq) for 208,506 cells derived from 58 lung adenocarcinomas from 44 patients, which covers primary tumour, lymph node and brain metastases, and pleural effusion in addition to normal lung tissues and lymph nodes. Here we use *Seurat* to conduct preprocessing: we first randomly selecet 10000 cells to filter 2000 highly variable genes, then generate 10 subsampled datasets with each containing 5000 cells.

BC

Single cell RNA sequencing of primary breast cancer from[59], which can be accessed from the NCBI Expression Omnibus database (accession code GSE75688). This dataset contains 515 cells from 11 patients and most of the cell type is tumor. We first screen out 2000 highly variable genes using *Seurat* then conduct subsampling. Due to the serious category imbalance problem in this dataset (326 cells are labelled as "Tumor"), we only sample 86 cells with the tumor label each time and all cells with other labels are retained so that the results will not be unreliable due to class imbalance during evaluation.

AD

A total of 80660 droplet-based single-nucleus RNA-seq (snRNA-seq) profiles for Alzheimer's disease from[3]. The post-mortem human brain samples came from 48 participants in the Religious Order Study (ROS) or the Rush Memory and Aging Project (MAP), collectively known as ROSMAP with 24 individuals with high levels of β-amyloid and other pathological hallmarks of AD ('AD-pathology'), and 24 individuals with no or very low β-amyloid burden or other pathologies ('no-pathology'). The original study clustered individuals based on nine clinico-pathological traits to further define the pathology groups as 'early-pathology' and 'late-pathology'. And that division is totally adopted in our study. The snRNA-seq data are available on The Rush Alzheimer's Disease Center (RADC) Research Resource Sharing Hub at https://www.radc.rush.edu/docs/omics.html (snRNA-seq PFC) or at Synapse (https://www.synapse.org/#!Synapse:syn18485175) under the doi 10.7303/syn18485175. The data are available under controlled use conditions set by human privacy regulations. To access the data, a data use agreement is needed. Since we are not going to use this data set to conduct benchmarking here, there is no need to repeatedly generate subsamples. When preprocessing the dataset, we first use stratified sampling to draw one out of ten cells, then 2000 highly variable genes are refined by Seurat. This sample is then used to be explored later.

Simulated data

In each simulation setting, we generate 100 simulated datasets. For convenience, we denote p as the number of genes, n as the number of cells. The scale of the FGM increases adaptively with the size of the simulated dataset (actually the size of p). The parameter $\boldsymbol{\mu}$ is computed by the multiplicative model $\boldsymbol{\mu} = \mathbf{WZ}$, where W is a p×L matrix and Z is an L×n matrix. The number of non-zero elements in each column of W is set as p/20, and the number of non-zero elements in each row of Z is randomly drawn from a Poisson distribution with a parameter of 30. The row indices of non-zero elements in W and the column indices of Z with non-zero elements are randomly drawn from 1 to p and 1 to n. The nonzero element values for both W and Z are generated from a normal distribution with mean 1.5 and standard deviation 0.1, and are randomly assigned to be positive or negative. The prior edge is generated along with W.

When generating X, each element is generated from $\mathrm{NB}\left(\mathrm{r_j}, \frac{1}{1+\mathrm{e}^{-\mu_{ji}}}\right)$, and the parameter r_j is randomly drawn from 5 to 20. Finally, in order to simulate different batches, we divided the dataset into three parts, each of n/3 samples, with different intensities of noise. The implementation of dropout is to perform Bernoulli censoring at each data point according to the given dropout rate parameter.

The simulation data generation process is shown in Fig. 3a.

Matching of biclusters when analysing AD dataset

To avoid confusion, we first explain the difference between biclustering and bicluster, two concepts we've been using throughout the paper. A biclustering refers to execute one biclustering algorithm once (eg. scBC). After biclustering is conducted, we can get several columns-rows pairs, each is called a bicluster. In our study, we conduct three independent biclusterings on the three pathologically seperated AD datasets, each with L biclusters. L is the number of biclusters we set beforehand.

When conducting scBC on the AD dataset, genes that are widely present in all biclusters represent the commonality among all cells. We subtract these genes from each bicluster to reduce the homogeneity of different biclusters, since we are not interested in them in this case. Here, we set the number of biclusters in each round of biclustering to 6. However, this is an empirical hyperparameter, and some biclusters may have a high degree of similarity and be more reasonable to merge into a single bicluster. This applies not only to different biclusters in a whole biclustering but also to different biclusters in independent biclusterings. However, the methods for aligning biclusters in a single biclustering and for biclusters in different biclusterings should be different, since biclusters from the latter are somewhat more independent.

Due to their own homogeneity or correlation, more attention should be paid to the exclusivity when merging biclusters from a single biclustering. We denote the number of genes only appear in biclusters i and j as e, which means all the other biclusters

don't have these genes. And genes present in bicluster k is denoted as g_k, the overlap score is defined as:

$$os_{i,j} = \frac{e}{\max\{|g_i|,|g_j|\}}$$

In our study, a pair of biclusters with overlap score > 0.03 are combined as a FGM. The biclusters correspondences in different biclusterings are independent, so more attention should be paid to the degree of overlap. Here we use the "overlap over union"(IoU) criterion to combine different biclusters:

$$IoU_{i,j} = \frac{intersect\{|g_i|,|g_j|\}}{union\{|g_i|,|g_j|\}}$$

each pair of biclusters with IoU > 0.3 are combined as a FGM. The alignment results are in Table S13.

FGM perturbation during AD progression and enrichment analysis

Before merging functional gene modules (FGMs) from different biclusters, we first assign each cell exclusively to a bicluster using the strong classification method as before. When similar FGMs are combined, functionally related cells are also merged as a whole. Next, we obtain the FGM composition contained in each cell type. As observed, multiple FGMs can be simultaneously active in one cell type. To illustrate FGM perturbation for each cell type during AD progression, we first observe the changes in the proportion of each FGM in each cell type. FGMs with an elevated ratio are candidates for "increased activity," while those with a reduced ratio are candidates for "decreased activity." We then examine the differences in the gene set makeup of these two types of FGMs. The overlap between the two represents commonalities exhibited in certain cell types, which are not of interest to us. We focus on the exclusive genes of "increased activity" and "decreased activity," which may uncover pathway perturbations in different pathological states. The exclusive genes in "increased activity" and "decreased activity" are used for functional enrichment using *clusterProfiler*. The results are used to reveal the pathway perturbation during AD progression.

Evaluation

Here we will describe the evaluation metrics used in our study as follows.

Comparison of FGM detection

To quantify the performance of each method in detecting functional gene modules (FGMs), we conduct gene ontology (GO) enrichment using *clusterProfiler* for each gene set of each bicluster and record the most significant p value (BH adjusted). Since the number of biclusters detected by each method differs, we take the most significant p value of all the biclusters detected by a single method and transform it using -log10(p) to denote the performance of this method. Methods that fail to detect any bicluster are labeled as 0. We use 10 subsamples from each dataset for repeated evaluation.

Criterions for clustering performance

There are three metrices we used to benchmark the clustering performance at cell level: ARI, FMI and AMI. Here we will briefly describe how to compute these metrices:

ARI

The Rand Index computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. The raw RI score is then "adjusted for chance" into the ARI score using the following scheme:

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)}$$

To calculate this value, first calculate the contingency table like that:

	Y_1	Y_2	...	Y_s	Sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r1}	$\cdots$	n_{rs}	a_r
Sums	b_1	b_2	$\cdots$	b_s	

each value in the table represents the number of data point located in both cluster (Y) and true class (X), and then calculate the ARI value through this table:

$$\overbrace{\widetilde{ARI}}^{\text{Adjust Index}} = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right]}_{\text{Max Index}} - \underbrace{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}_{\text{Expected Index}}}$$

The adjusted Rand index is thus ensured to have a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clusterings are identical (up to a permutation). The adjusted Rand index is bounded below by -0.5 for especially discordant clusterings.

FMI

The Fowlkes-Mallows index (FMI) is defined as the geometric mean between of the precision and recall:

$$FMI = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}}$$

Where TP is the number of **True Positive** (i.e. the number of pair of points that belongs in the same clusters in both true labels and predicted labels), FP is the number of **False Positive** (i.e. the number of pair of points that belongs in the same clusters in true labels but not in predicted labels) and FN is the number of **False Negative** (i.e the number of pair of points that belongs in the same clusters in predicted labels but not in true labels). The score ranges from 0 to 1. A high value indicates a good similarity between two clusters.

AMI

The Mutual Information is a measure of the similarity between two labels of the same data. Where $|U_i|$ is the number of the samples in cluster U_i and $|V_j|$ is the number of the samples in cluster V_j, the Mutual Information between clusterings U and V is given as:

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for chance. It accounts for the fact that the MI is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared. For two clusterings U and V, the AMI is given as:

$$AMI(U,V) = \frac{MI(U,V) - E(MI(U,V))}{\text{avg}(H(U),H(V)) - E(MI(U,V))}$$

Where $H(*)$ is the information entropy for a label's distribution (eg. $H(U) = \sum_{i=1}^{|U|} P(i)\, log(P(i))$). This metric is independent of the absolute values of the labels: a permutation of the class or cluster label values won't change the score value in any way.

Metrices for biclustering comparion

Suppose $M : \{1...L\} \rightarrow \{1...L\}$ maps the ground true bicluster index to the index of the bicluster detected by an algorithm, T_i denote the i_{th} ground true bicluster and B_i denote

the i_{th} detected bicluster. The Cluster Error (CE) proposed by Anne et.al[60] is defined as:

$$1 - \mathrm{CE}(\mathrm{M}) = \frac{\sum_{i=1}^{L} |\, T_i \cap B_{M(i)}|}{|\bigcup_{i=1}^{L} T_i \cup B_{M(i)}\,|}$$

This is a distance measure of subspace clustering with lower CE indicating better consistency with ground truth. When we evaluate the performance, we choose a M minimizing the CE as the optimal match and is used by other measurements. The corresponding 1-CE is output with the higher the value, the better.

We also use F-score (F) to evaluate the performance. F-score is the harmonic mean of precision (PRE) and recall (REC). Here we use the calculation way proposed by Zhong et.al[18]:

$$\mathrm{PRE}_i = \frac{|T_i \cap B_{M(i)}\,|}{|B_{M(i)}|}$$

$$\mathrm{REC}_i = \frac{|T_i \cap B_{M(i)}\,|}{|T_i|}$$

Where A denote all the elements of the expression data. PRE_i and REC_i are computed for bicluster pair i, we finally output the average for each criterion as PRE and REC, along with their harmonic mean as F-score(F), which is a combination of the two, and we also pay more attention to it. The higher this indicator is, the better.

Supplementary Information

Additional file 1: **Fig. S1.** overlap of biclusters. **Fig. S2.** visualization of cell clustering results on PBMC dataset. **Fig. S3.** visualization of cell clustering results on HEART dataset. **Fig. S4.** visualization of cell clustering results on LUAD dataset. **Fig. S5.** visualization of cell clustering results on BC dataset. **Table S1.** Enrichment score comparison of different methods. **Table S2.** ARI of different methods in different datasets. **Table S3.** FMI of different methods in different datasets. **Table S4.** AMI of different methods in different datasets. **Table S5-S7.** 1-CE of different methods in different datasets. **Table S8-S10**. F score of different methods in different datasets.
Additional file 2. Table S11. Clinico-pathological variables of 48 AD patients.
Additional file 3. Table S12. enrichment results of perturbated gene sets.
Additional file 4. Table S13. matching results of biclusters in different pathological stage.

Declarations

Data and code availability

All data used in this research can be found in **Datasets and preprocessing** in **STAR Methods** section. Our scBC method is available as a Python package on PyPI at

https://pypi.org/project/scBC, free for academic use, and the source code is openly available from our GitHub repository at https://github.com/GYQ-form/scBC.

Competing interests

The authors declare no competing interests.

Funding

The research is supported partly by the National Natural Science Foundation of China (11901387 for Y.Z. and 12171318 for Z.Y.).

Authors' contributions

YG performed the research, analyzed data and wrote the original manuscript, JX participated in data collection, ZY and YZ supervised the research. YG, JX, RG, JS, ZY, and YZ discussed and revised the manuscript.

Acknowledgements

Not applicable.

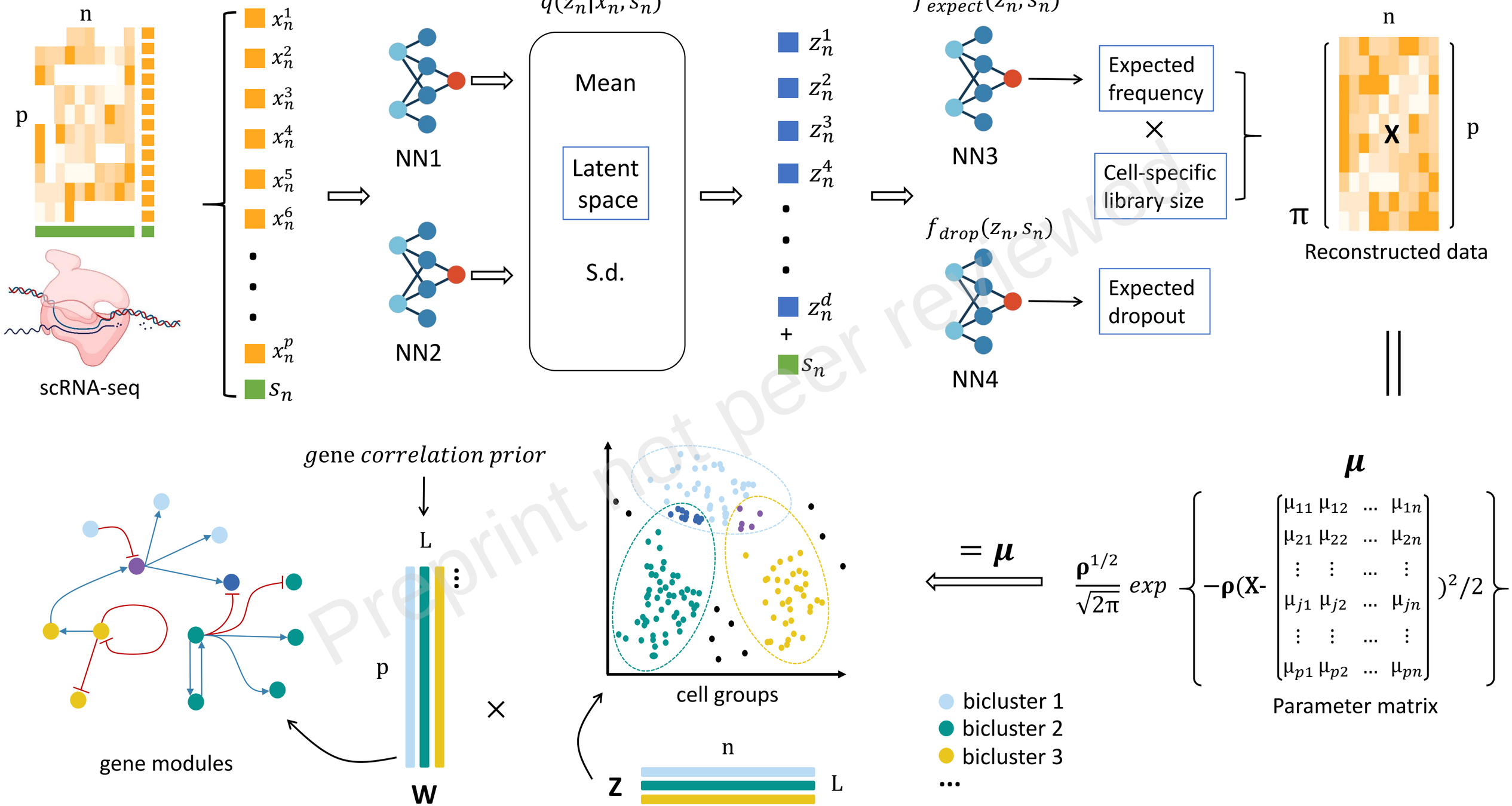
References

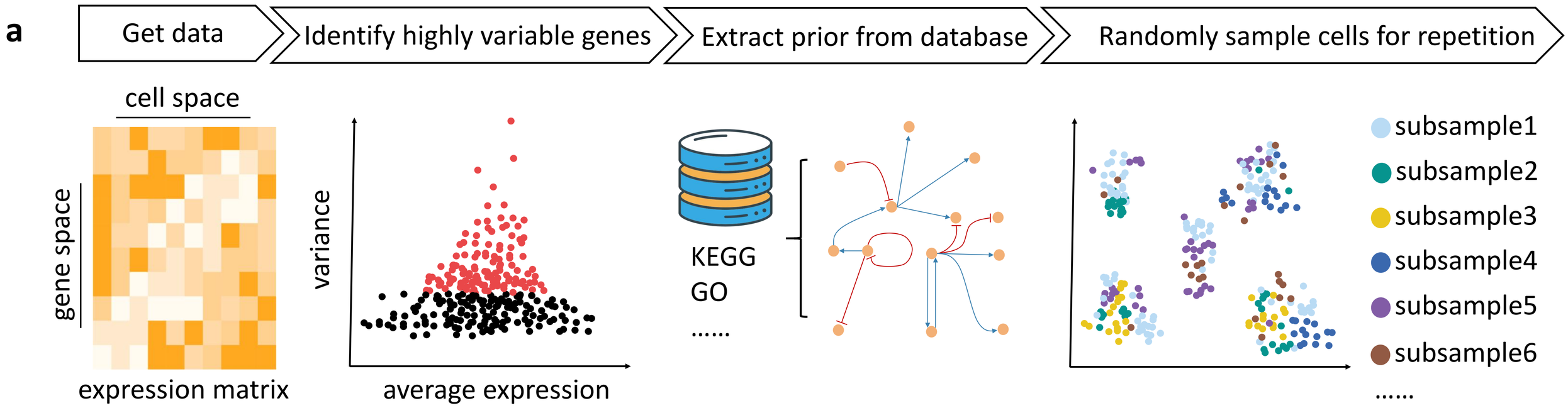
1. Shi, F. and H. Huang, *Identifying Cell Subpopulations and Their Genetic Drivers from Single-Cell RNA-Seq Data Using a Biclustering Approach.* J Comput Biol, 2017. **24**(7): p. 663-674.
2. Eisenberg, E. and E.Y. Levanon, *Human housekeeping genes, revisited.* Trends Genet, 2013. **29**(10): p. 569-74.
3. Mathys, H., et al., *Single-cell transcriptomic analysis of Alzheimer's disease.* Nature, 2019. **570**(7761): p. 332-337.
4. Lake, B.B., et al., *Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain.* Nat Biotechnol, 2018. **36**(1): p. 70-80.
5. Grubman, A., et al., *A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation.* Nat Neurosci, 2019. **22**(12): p. 2087-2097.
6. Habib, N., et al., *Disease-associated astrocytes in Alzheimer's disease and aging.* Nat Neurosci, 2020. **23**(6): p. 701-706.
7. Roussarie, J.P., et al., *Selective Neuronal Vulnerability in Alzheimer's Disease: A Network-Based Analysis.* Neuron, 2020. **107**(5): p. 821-+.
8. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
9. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.
10. Ghazalpour, A., et al., *Comparative analysis of proteome and transcriptome variation in mouse.* PLoS Genet, 2011. **7**(6): p. e1001393.
11. Zhang, B., et al., *Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease.* Cell, 2013. **153**(3): p. 707-20.

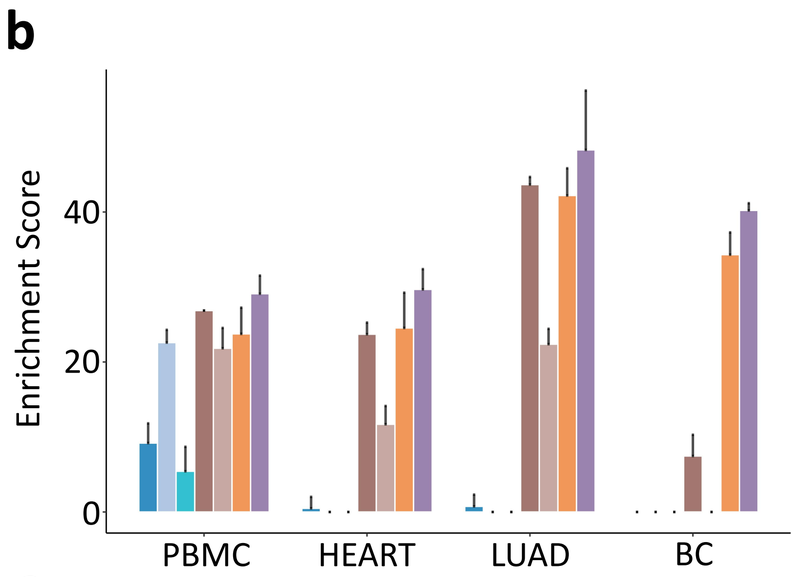
12. Risso, D., et al., *Normalization of RNA-seq data using factor analysis of control genes or samples.* Nat Biotechnol, 2014. **32**(9): p. 896-902.
13. Hochreiter, S., et al., *FABIA: factor analysis for bicluster acquisition.* Bioinformatics, 2010. **26**(12): p. 1520-7.
14. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nat Rev Genet, 2010. **11**(10): p. 733-9.
15. Hu, Z., S. Zu, and J.S. Liu, *SIMPLEs: a single-cell RNA sequencing imputation strategy preserving gene modules and cell clusters variation.* NAR Genom Bioinform, 2020. **2**(4): p. lqaa077.
16. Fang, Q., et al., *An Effective Biclustering-Based Framework for Identifying Cell Subpopulations From scRNA-seq Data.* IEEE/ACM Trans Comput Biol Bioinform, 2021. **18**(6): p. 2249-2260.
17. Xie, J., et al., *QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data.* Bioinformatics, 2020. **36**(4): p. 1143-1149.
18. Zhong, Y. and J.Z. Huang, *Biclustering via structured regularized matrix decomposition.* Statistics and Computing, 2022. **32**(3).
19. Lopez, R., et al., *Deep generative modeling for single-cell transcriptomics.* Nat Methods, 2018. **15**(12): p. 1053-1058.
20. Li, Z., et al., *Bayesian generalized biclustering analysis via adaptive structured shrinkage.* Biostatistics, 2020. **21**(3): p. 610-624.
21. Cheng, Y. and G.M. Church, *Biclustering of expression data.* Proceedings. International Conference on Intelligent Systems for Molecular Biology, 2000. **8**: p. 93-103.
22. Murali, T.M. and S. Kasif, *Extracting conserved gene expression motifs from gene expression data.* Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2003: p. 77-88.
23. Prelic, A., et al., *A systematic comparison and evaluation of biclustering methods for gene expression data.* Bioinformatics, 2006. **22**(9): p. 1122-9.
24. Caldas, J. and S. Kaski, *Bayesian Biclustering with the Plaid Model.* 2008 Ieee Workshop on Machine Learning for Signal Processing, 2008: p. 291-296.
25. Dongen, S.V., *Graph clustering by flow simulation.* PhD thesis, University of Utrecht, 2000.
26. Milligan, G.W. and M.C. Cooper, *A STUDY OF THE COMPARABILITY OF EXTERNAL CRITERIA FOR HIERARCHICAL CLUSTER-ANALYSIS.* Multivariate Behavioral Research, 1986. **21**(4): p. 441-458.
27. Santos, J.M. and M. Embrechts. *On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification*. in *19th International Conference on Artificial Neural Networks (ICANN 2009)*. 2009. Limassol, CYPRUS.
28. Fowlkes, E.B. and C.L. Mallows, *A method for comparing two hierarchical clusterings.* Journal of the American statistical association, 1983. **78**(383): p. 553-569.
29. Strehl, A. and J. Ghosh, *Cluster ensembles- a knowledge reuse framework for combining multiple partitions.* Journal of Machine Learning Research, 2003. **3**(3): p. 583-617.
30. Rahaman, M.A., et al., *Shared sets of correlated polygenic risk scores and voxel-wise grey matter across multiple traits identified via bi-clustering.* Annu Int Conf IEEE Eng Med Biol Soc, 2021. **2021**: p. 2201-2206.
31. Murdock, M.H. and L.H. Tsai, *Insights into Alzheimer's disease from single-cell genomic approaches.* Nat Neurosci, 2023. **26**(2): p. 181-195.

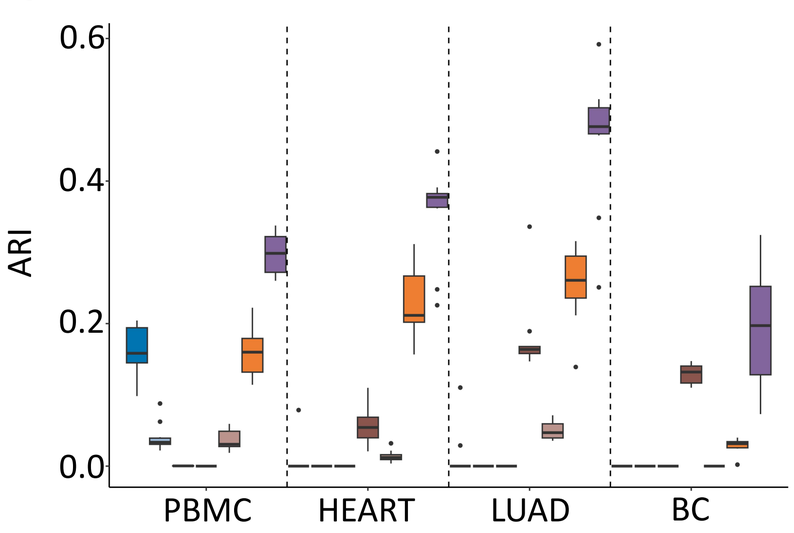
32. Hasel, P., et al., *Neuroinflammatory astrocyte subtypes in the mouse brain.* Nat Neurosci, 2021. **24**(10): p. 1475-1487.
33. Blanchard, J.W., et al., *APOE4 impairs myelination via cholesterol dysregulation in oligodendrocytes.* Nature, 2022. **611**(7937): p. 769-779.
34. Welch, G. and L.H. Tsai, *Mechanisms of DNA damage-mediated neurotoxicity in neurodegenerative disease.* EMBO Rep, 2022. **23**(6): p. e54217.
35. Zhou, Y., et al., *Author Correction: Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease.* Nat Med, 2020. **26**(6): p. 981.
36. Lau, S.F., et al., *Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease.* Proc Natl Acad Sci U S A, 2020. **117**(41): p. 25800-25809.
37. Pan, S., et al., *Preservation of a remote fear memory requires new myelin formation.* Nat Neurosci, 2020. **23**(4): p. 487-499.
38. Fancy, S.P., et al., *Myelin regeneration: a recapitulation of development?* Annu Rev Neurosci, 2011. **34**: p. 21-43.
39. Franklin, R.J. and S.A. Goldman, *Glia Disease and Repair-Remyelination.* Cold Spring Harb Perspect Biol, 2015. **7**(7): p. a020594.
40. Karadottir, R., et al., *Spiking and nonspiking classes of oligodendrocyte precursor glia in CNS white matter.* Nat Neurosci, 2008. **11**(4): p. 450-6.
41. Mitew, S., et al., *Mechanisms regulating the development of oligodendrocytes and central nervous system myelin.* Neuroscience, 2014. **276**: p. 29-47.
42. Depp, C., et al., *Myelin dysfunction drives amyloid-beta deposition in models of Alzheimer's disease.* Nature, 2023. **618**(7964): p. 349-357.
43. Liu, R., et al., *Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers.* Med Res Rev, 2014. **34**(3): p. 455-78.
44. Jin, G.X., et al., *The knowledge-integrated network biomarkers discovery for Major Adverse Cardiac Events.* Journal of Proteome Research, 2008. **7**(9): p. 4013-4021.
45. Ideker, T. and R. Sharan, *Protein networks in disease.* Genome Res, 2008. **18**(4): p. 644-52.
46. Risso, D., et al., *A general and flexible method for signal extraction from single-cell RNA-seq data.* Nat Commun, 2018. **9**(1): p. 284.
47. Grun, D., L. Kester, and A. van Oudenaarden, *Validation of noise models for single-cell transcriptomics.* Nature Methods, 2014. **11**(6): p. 637-+.
48. Kingma, D.P. and M. Welling *Auto-Encoding Variational Bayes*. 2013. arXiv:1312.6114 DOI: 10.48550/arXiv.1312.6114.
49. Sonderby, C.K., et al., *Ladder Variational Autoencoders.* Advances in Neural Information Processing Systems 29 (Nips 2016), 2016. **29**.
50. Li, C. and H. Li, *Network-constrained regularization and variable selection for analysis of genomic data.* Bioinformatics, 2008. **24**(9): p. 1175-82.
51. Zhao, Y., et al., *Hierarchical Feature Selection Incorporating Known and Novel Biological Information: Identifying Genomic Features Related to Prostate Cancer Recurrence.* J Am Stat Assoc, 2016. **111**(516): p. 1427-1439.
52. Li, Z., S.E. Safo, and Q. Long, *Incorporating biological information in sparse principal component analysis with application to genomic data.* BMC Bioinformatics, 2017. **18**(1): p.

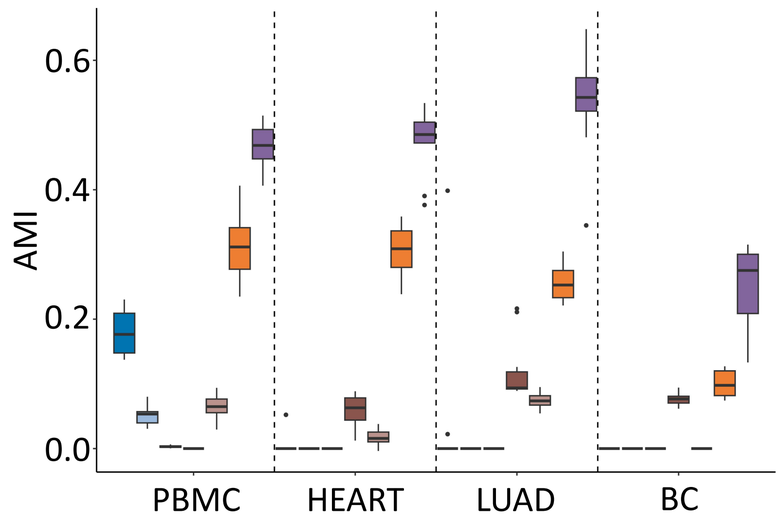
332.
53. Safo, S.E., S. Li, and Q. Long, *Integrative analysis of transcriptomic and metabolomic data via sparse canonical correlation analysis with incorporation of biological information.* Biometrics, 2018. **74**(1): p. 300-312.
54. Chang, C., S. Kundu, and Q. Long, *Scalable Bayesian variable selection for structured high-dimensional data.* Biometrics, 2018. **74**(4): p. 1372-1382.
55. Polson, N.G., J.G. Scott, and J. Windle, *Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables.* Journal of the American Statistical Association, 2013. **108**(504): p. 1339-1349.
56. Chang, C. and R.S. Tsay, *Estimation of covariance matrix via the sparse Cholesky factor with lasso.* Journal of Statistical Planning and Inference, 2010. **140**(12): p. 3858-3873.
57. Zheng, G.X., et al., *Massively parallel digital transcriptional profiling of single cells.* Nat Commun, 2017. **8**: p. 14049.
58. Kim, N., et al., *Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma.* Nat Commun, 2020. **11**(1): p. 2285.
59. Chung, W., et al., *Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer.* Nat Commun, 2017. **8**: p. 15081.
60. Patrikainen, A. and M. Meila, *Comparing subspace clusterings.* Ieee Transactions on Knowledge and Data Engineering, 2006. **18**(7): p. 902-916.

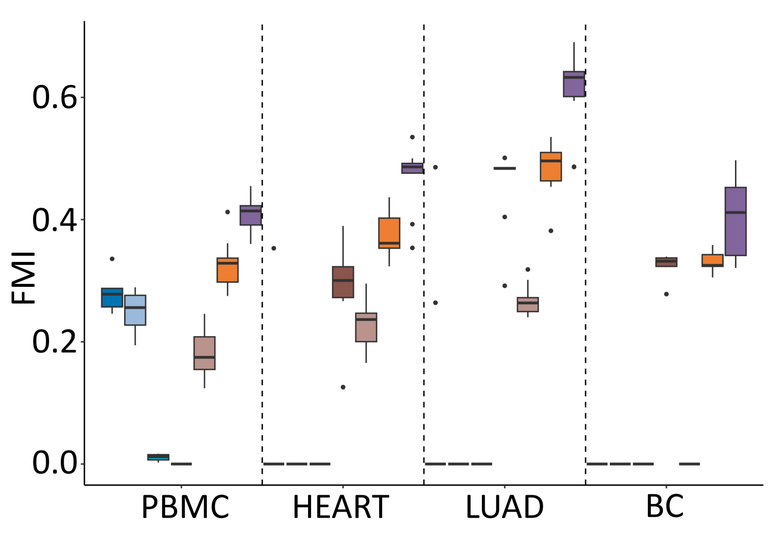


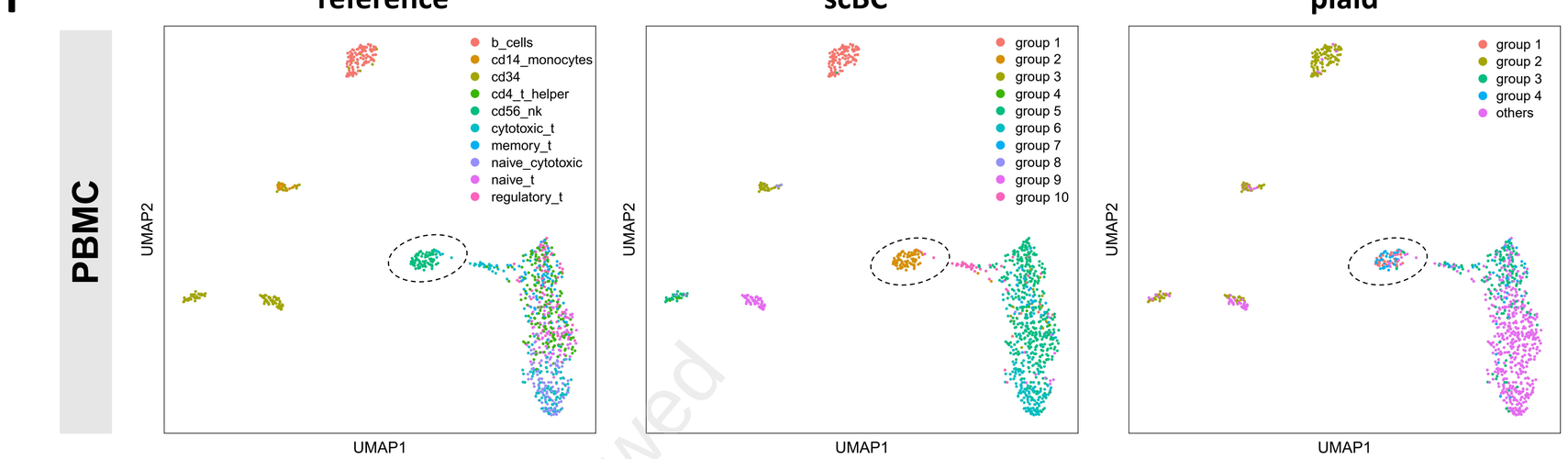


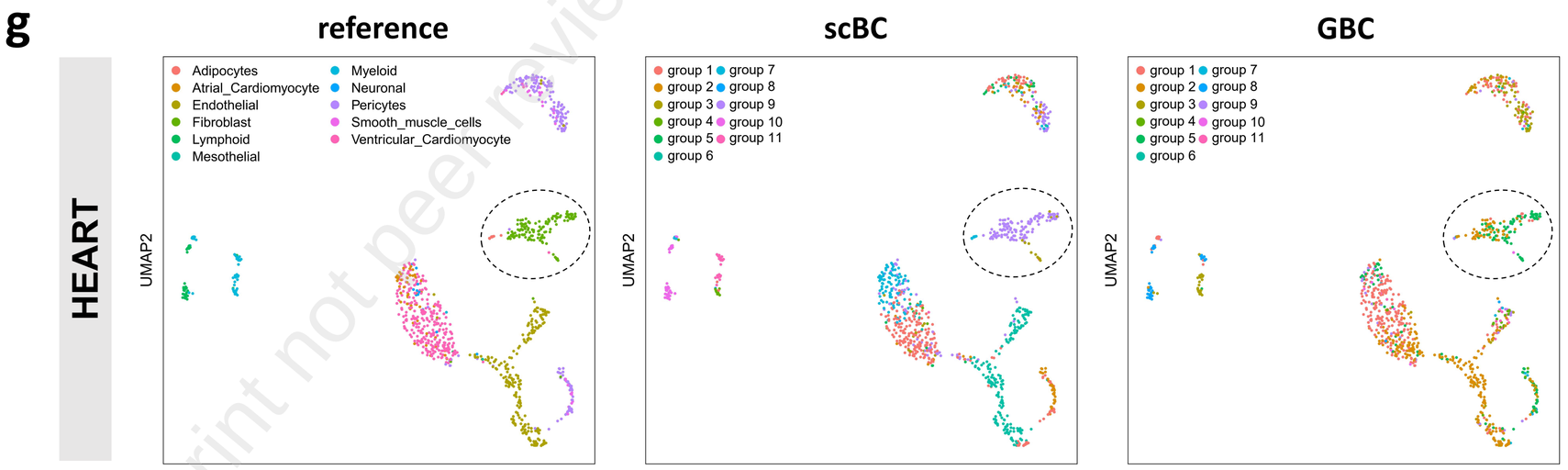


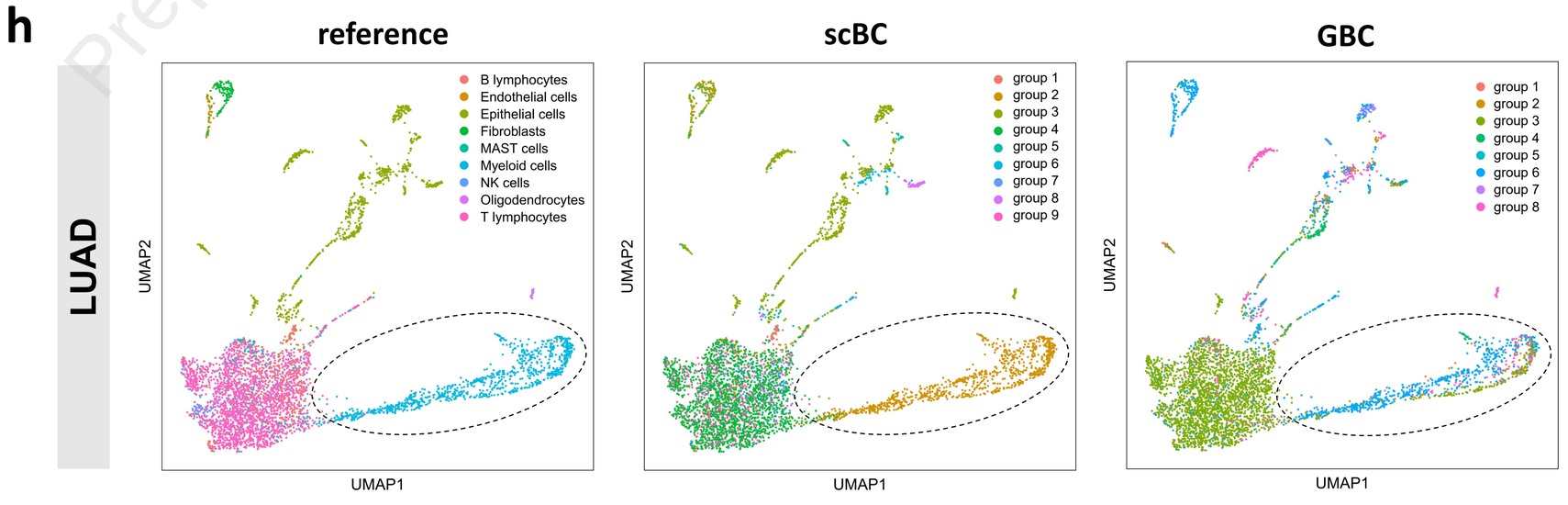


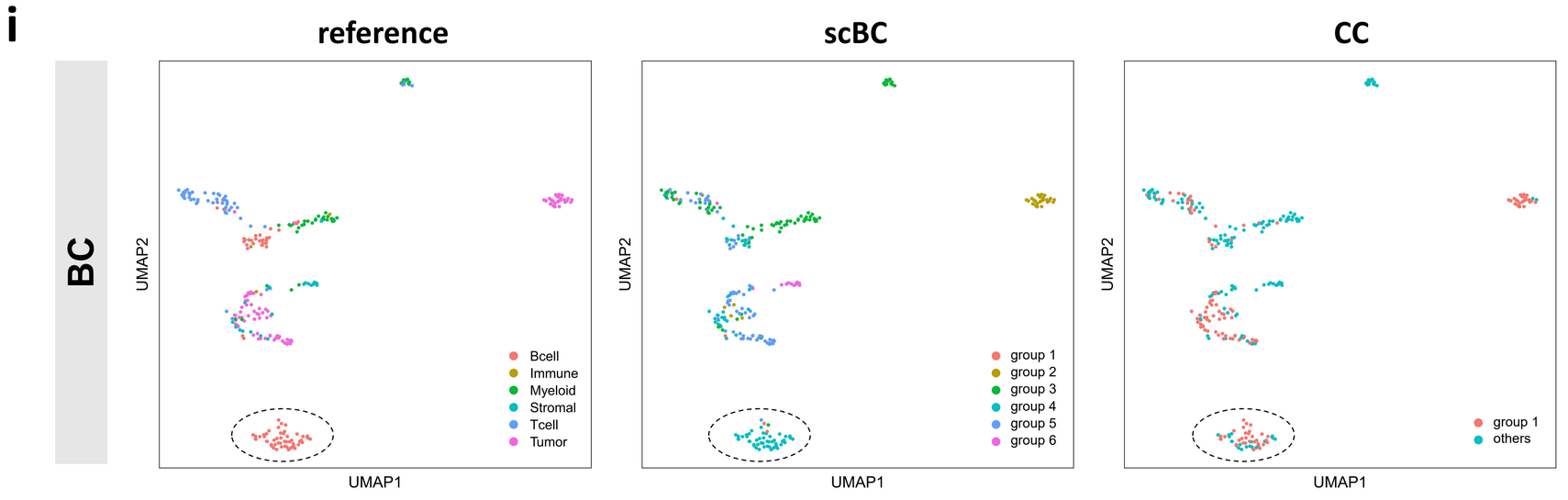












a

1. generate **W** and **Z**:

W × Z =

2. generate **μ**:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2n} \\ \vdots & \vdots & \dots & \vdots \\ \mu_{j1} & \mu_{j2} & \dots & \mu_{jn} \\ \vdots & \vdots & \dots & \vdots \\ \mu_{p1} & \mu_{p2} & \dots & \mu_{pn} \end{pmatrix}$$

3. generate **X**:

$$X_{ij} \sim NB(r_j, \frac{1}{1+e^{-\mu_{ji}}})$$

4. add noise (batch effect):

5. random dropout:

extract prior information

gene1	gene2
gene1	gene3
gene5	gene6
gene5	gene9
…	…

b p=1000, n=300, L=3

c p=3000, n=600, L=4

d p=6000, n=1500, L=5

e

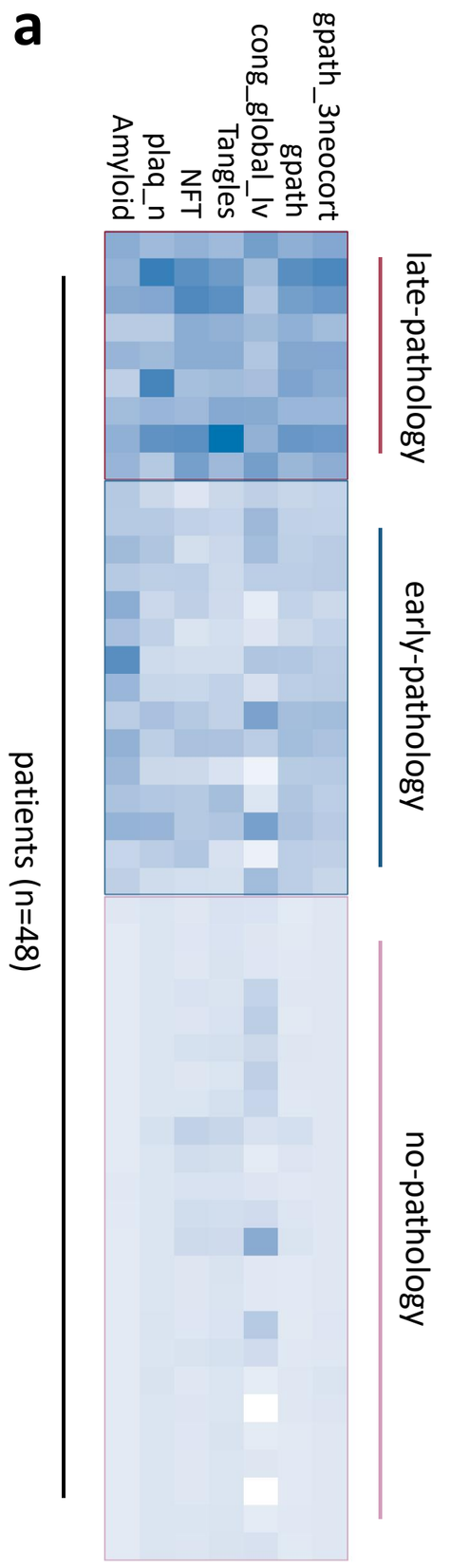
f

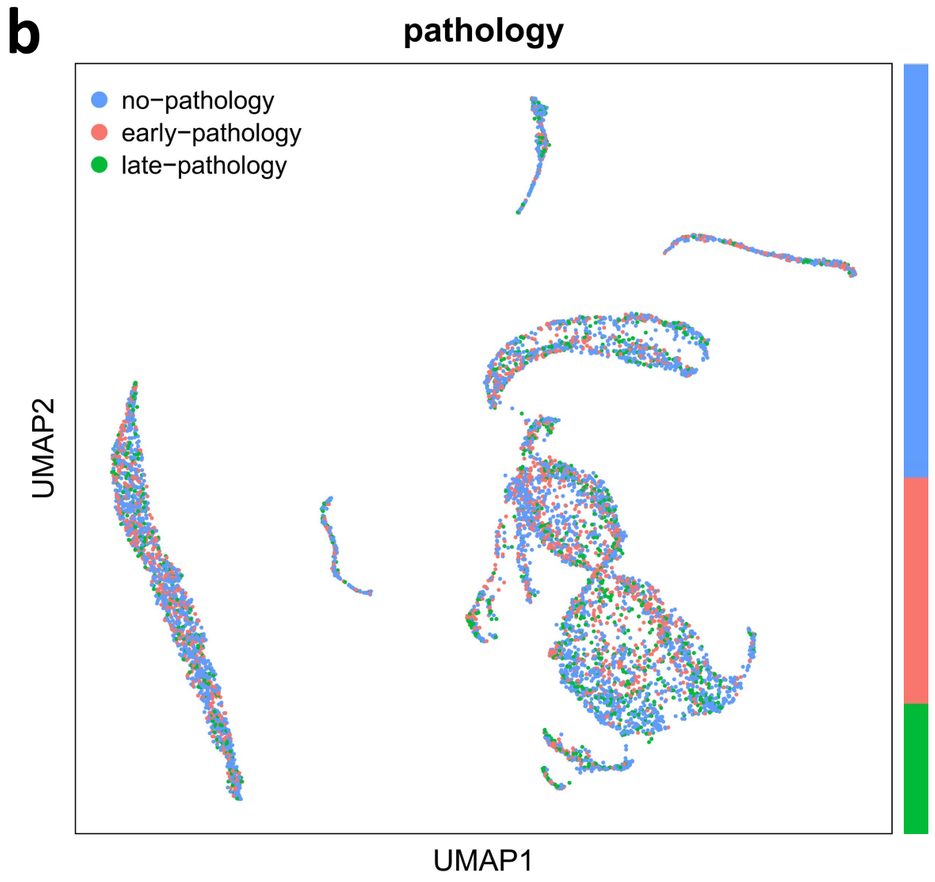
g

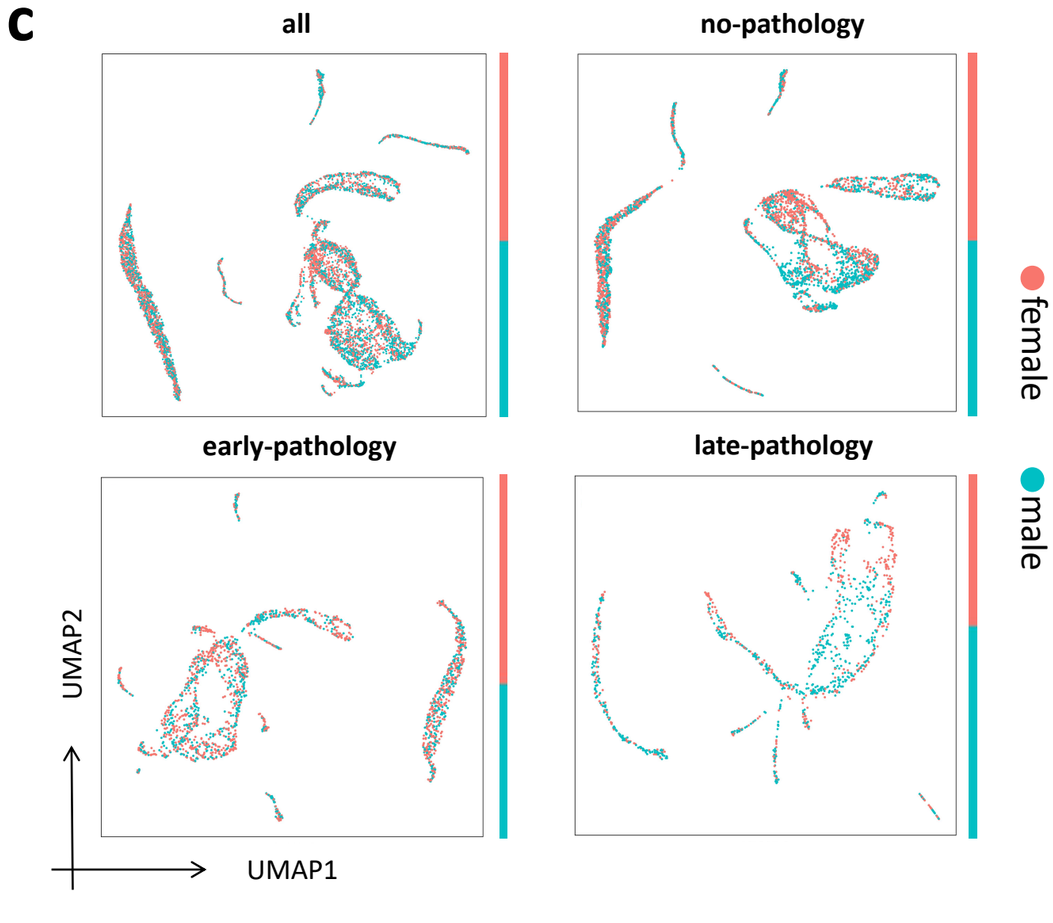
1-CE

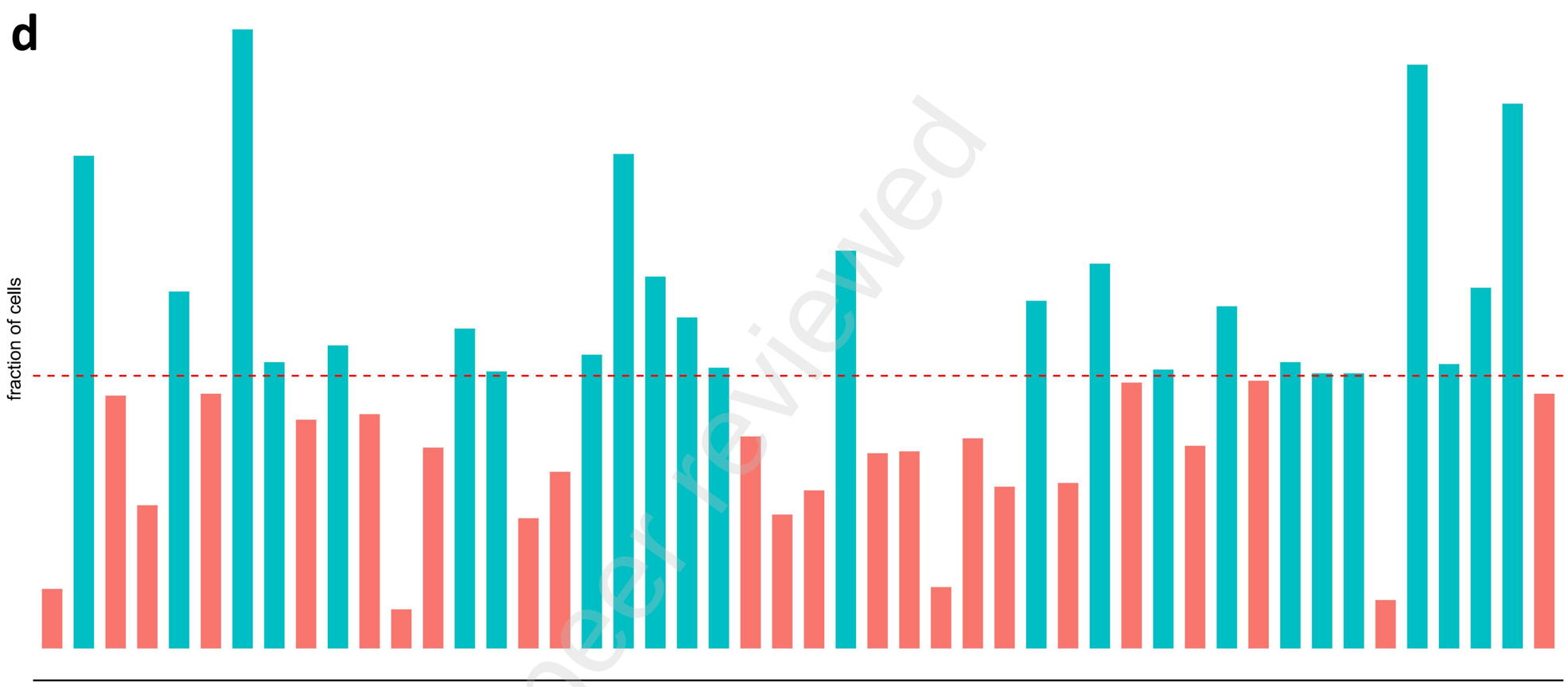
F score

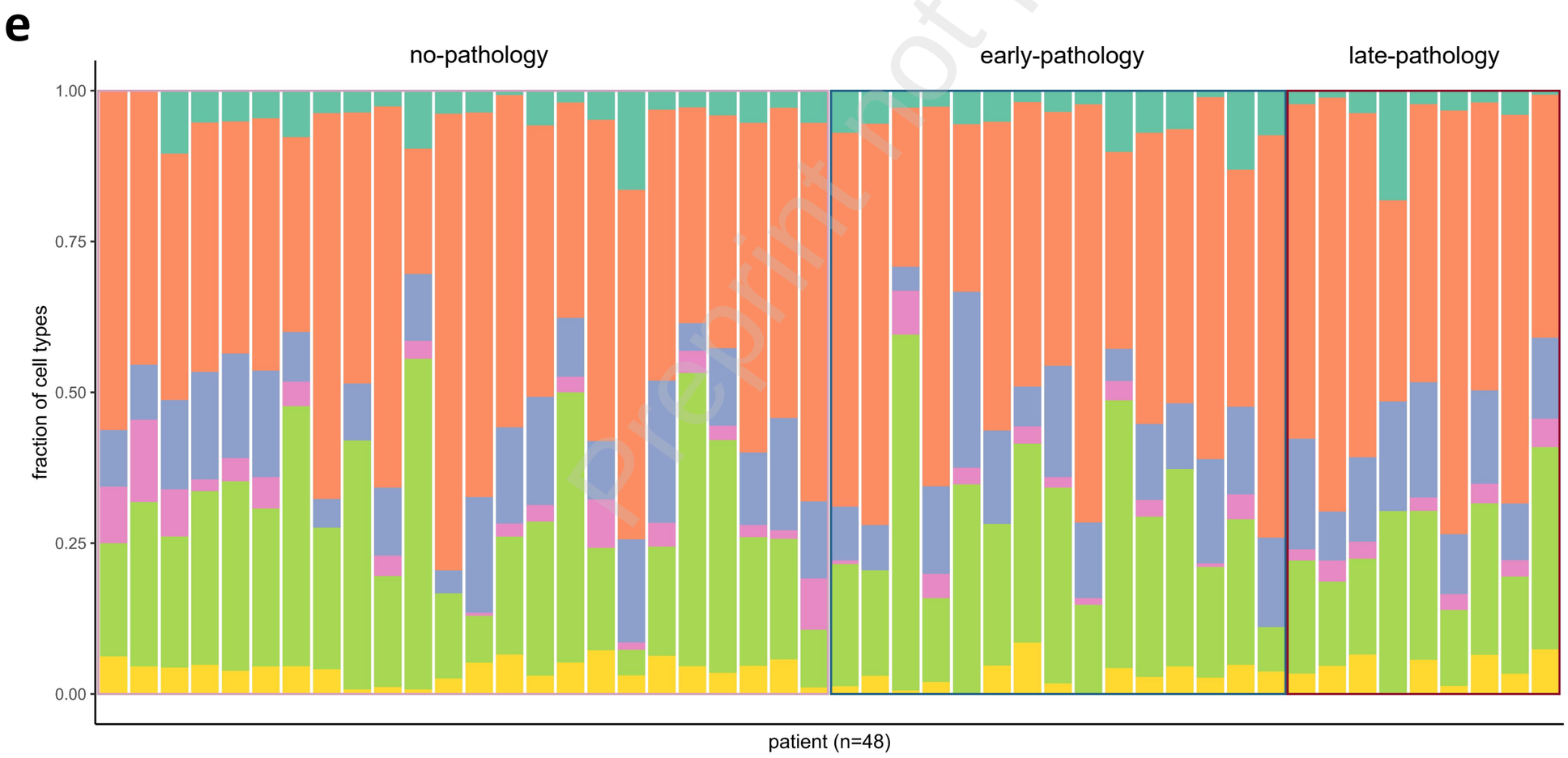
Method: plaid, Xmotif, Bimax, CC, FABIA, GBC, scBC

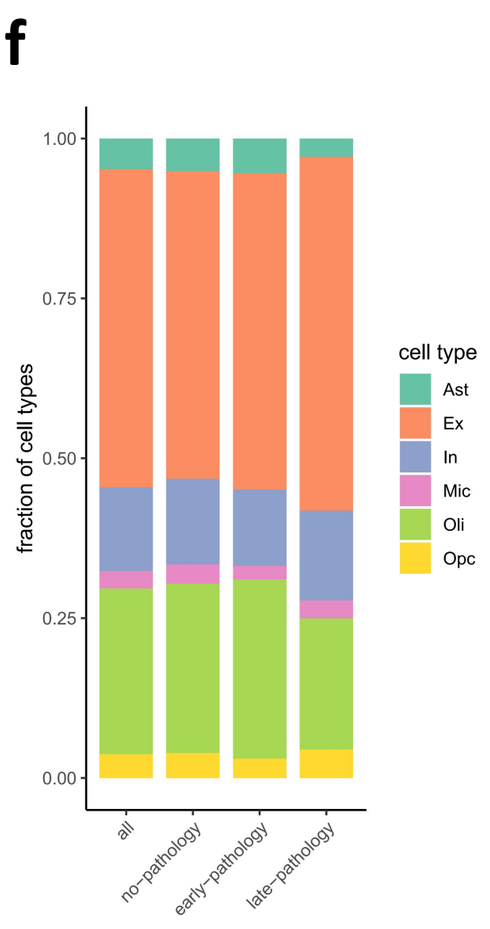


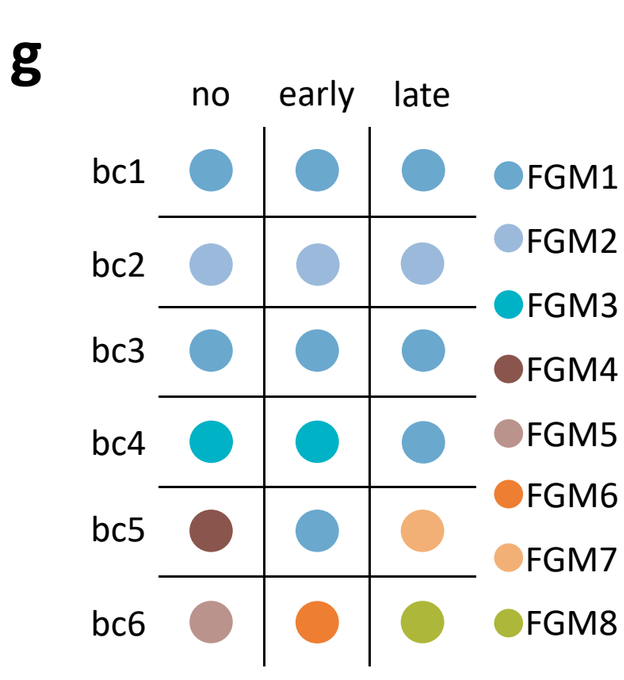


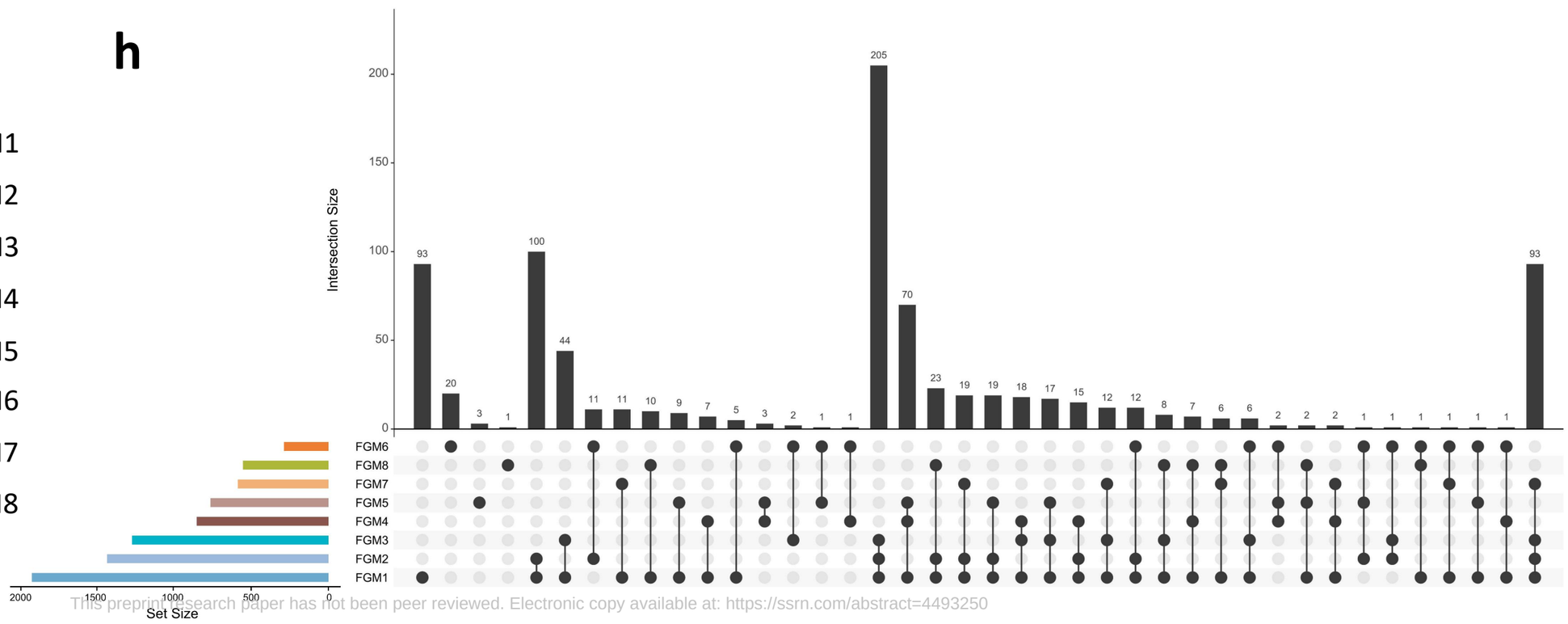


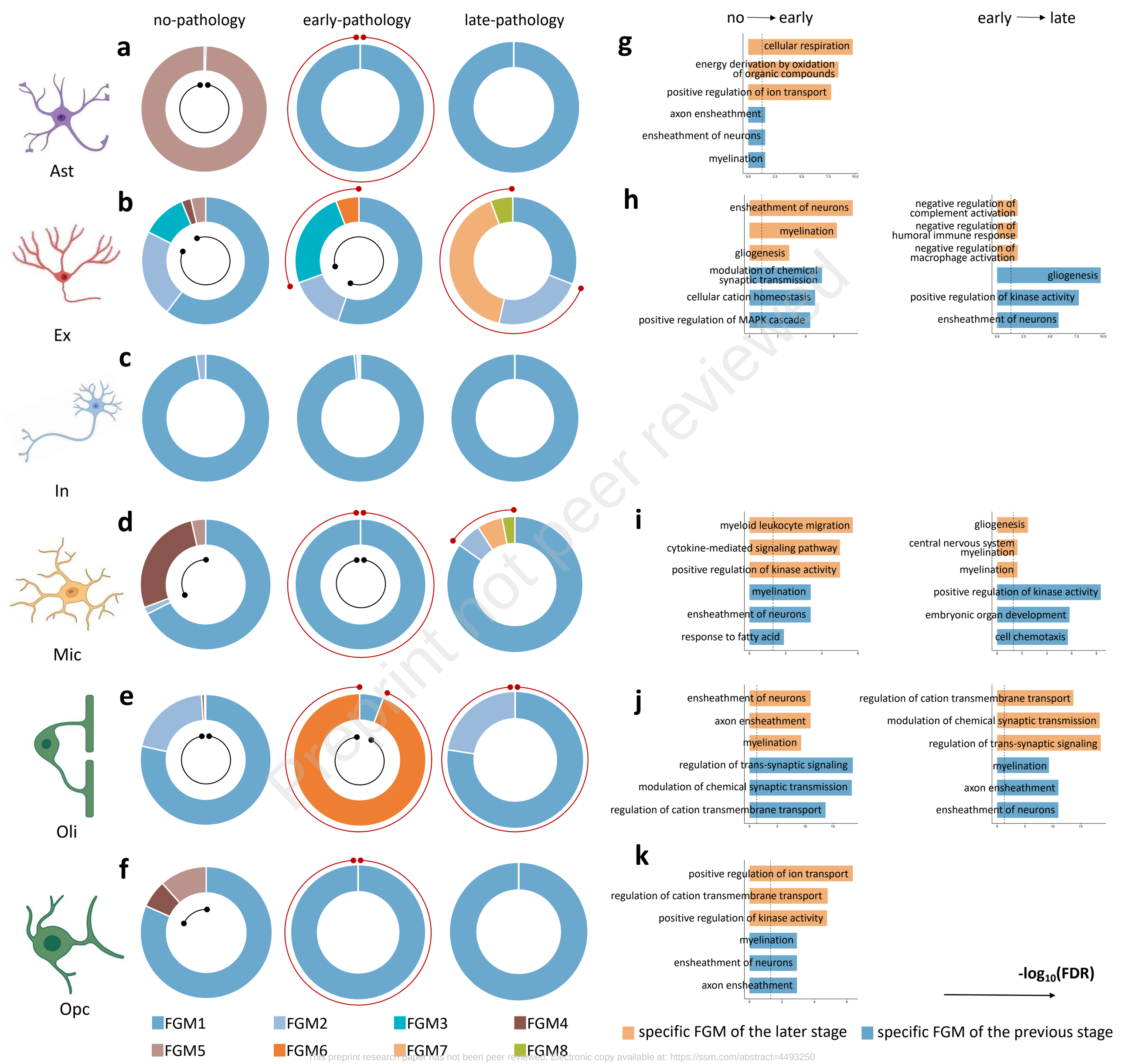












Supplementary Information

Supplementary Figures

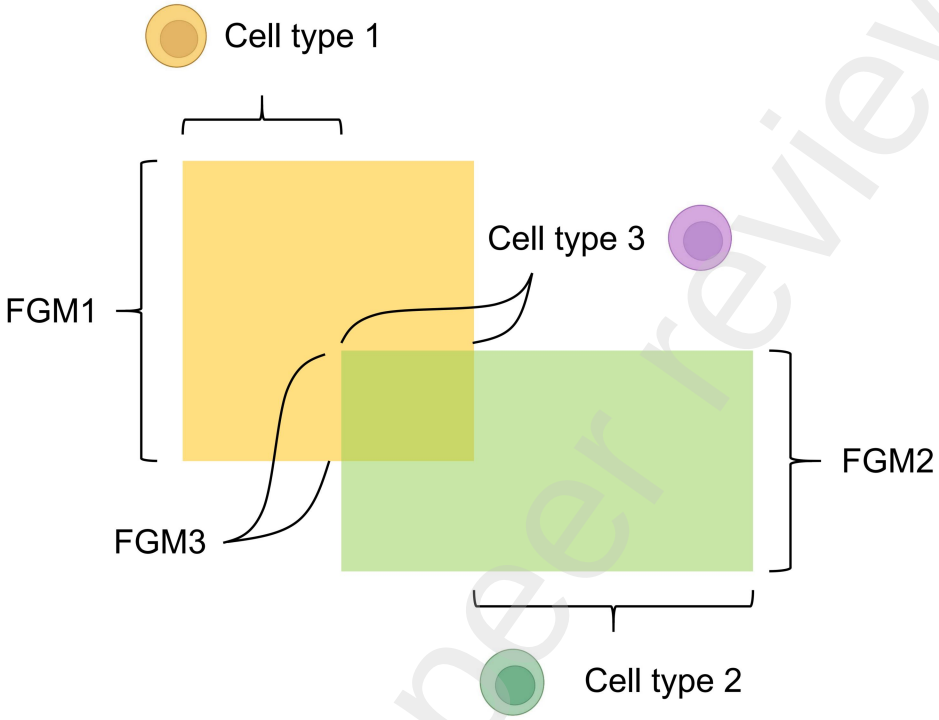


Fig. S1 | overlap of biclusters. This schematic diagram illustrate how biclustering with overlap can uncover the complex patterns in single-cell data. In the complex cell machinery, multiple FGMs are active in a cell group, which is represented by the figure that FGM1 and FGM2 are simultaneously active in cell type 3. Meanwhile, different cell groups may share a common FGM. This can be seen from the fact that cell type 1 and cell type 2 share a common FGM3, since FHM3 is contained in both FGM1 and FGM2.

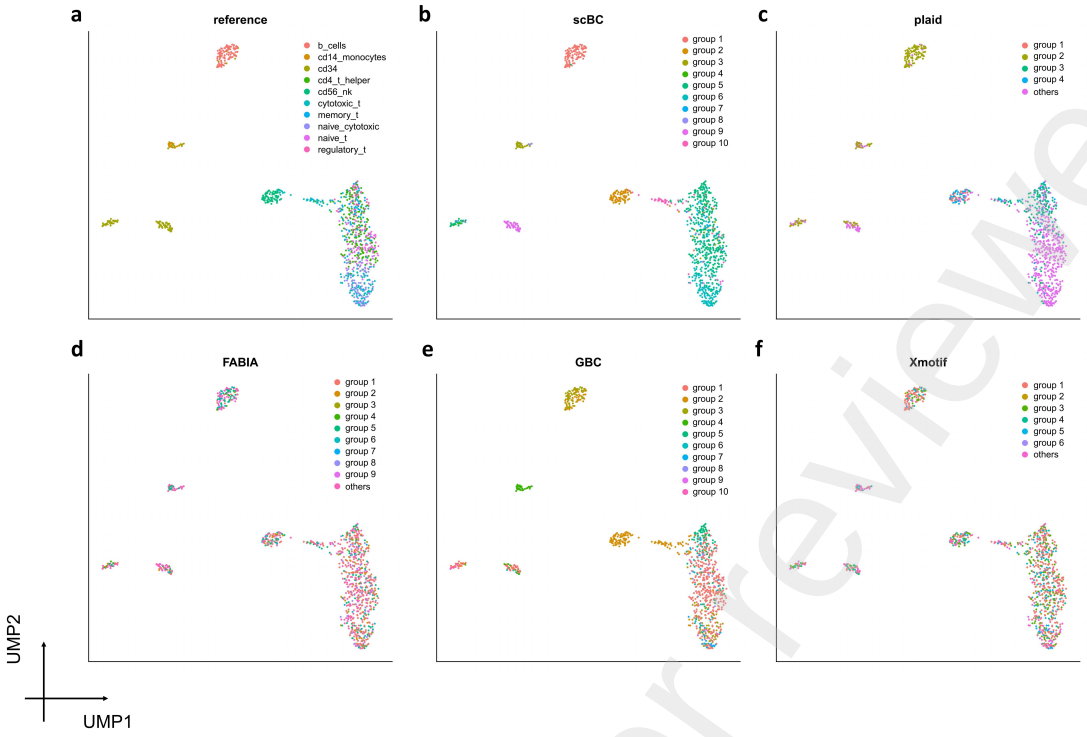


Fig. S2 | visualization of cell clustering results on PBMC dataset. These are the cell representation of the subsample from the last iteration of PBMC dataset. Some methods output too many categories so that we merge some into "others" whose number of cells is less than one percent of the total sample. Methods not shown here fail to detect functional cell groups.

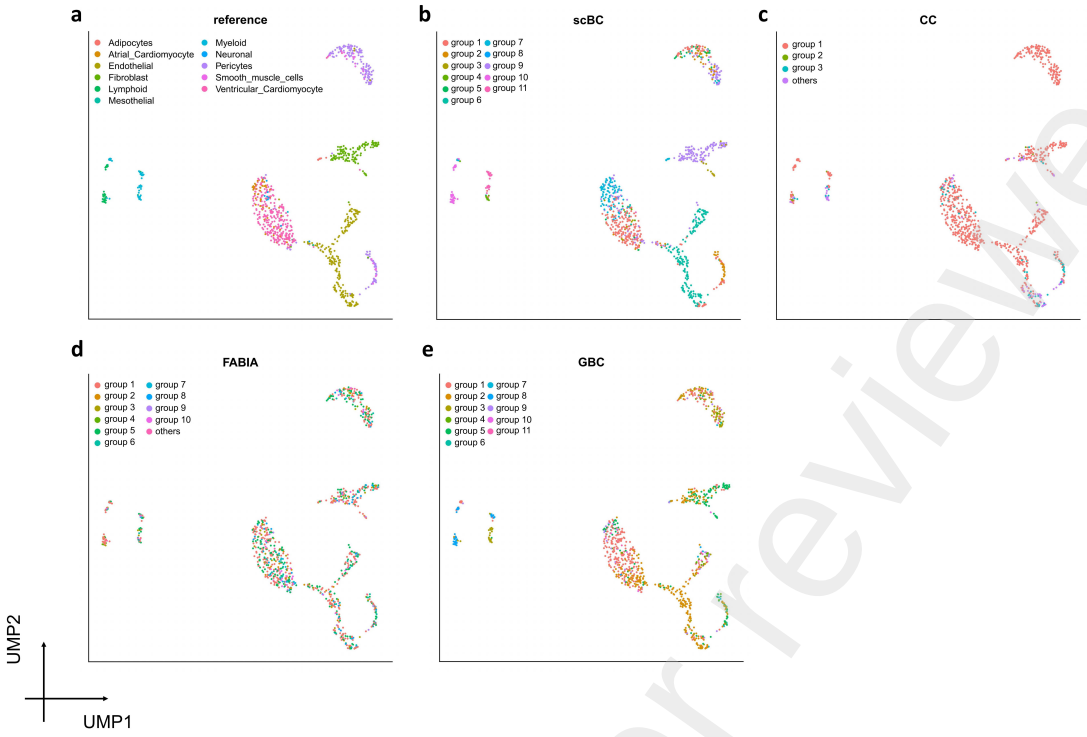


Fig. S3 | visualization of cell clustering results on HEART dataset. These are the cell representation of the subsample from the last iteration of HEART dataset. Some methods output too many categories so that we merge some into "others" whose number of cells is less than one percent of the total sample. Methods not shown here fail to detect functional cell groups.

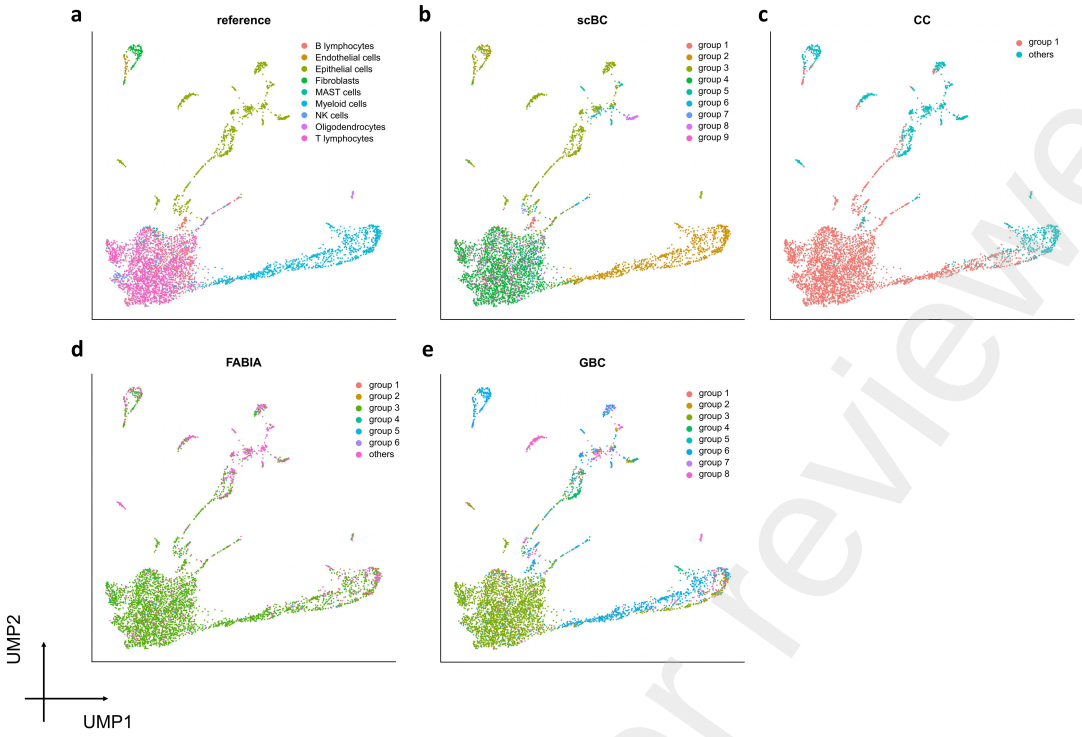


Fig. S4 | visualization of cell clustering results on LUAD dataset. These are the cell representation of the subsample from the last iteration of LUAD dataset. Some methods output too many categories so that we merge some into "others" whose number of cells is less than one percent of the total sample. Methods not shown here fail to detect functional cell groups.

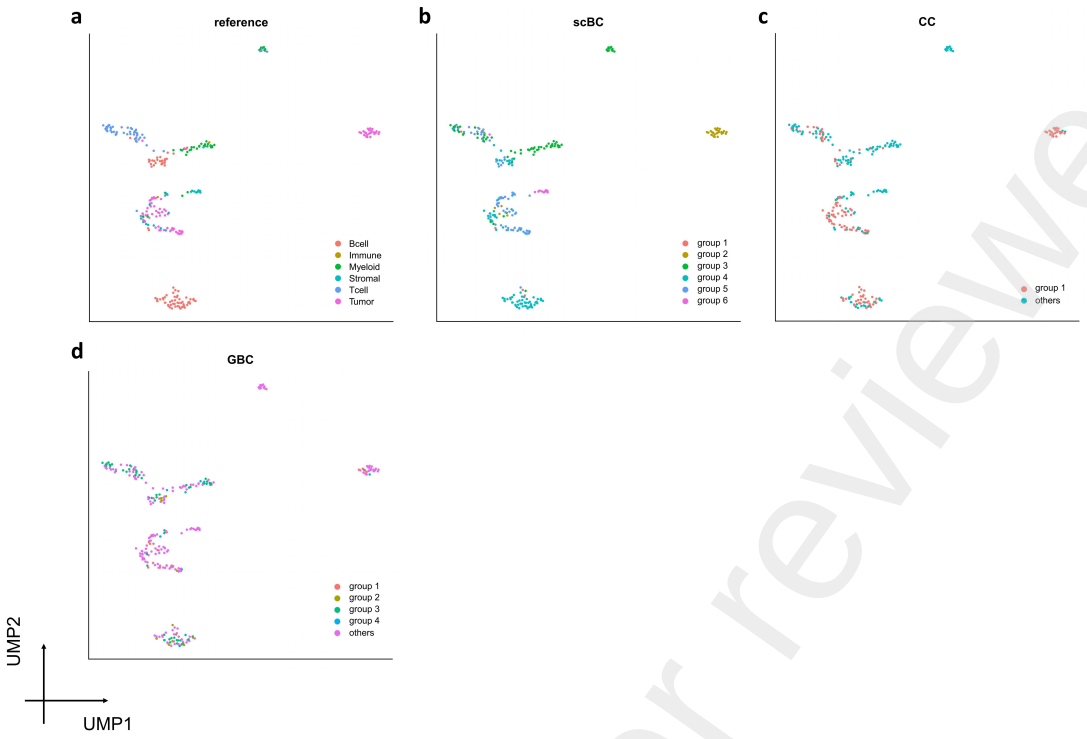


Fig. S5 | visualization of cell clustering results on BC dataset. These are the cell representation of the subsample from the last iteration of BC dataset. Some methods output too many categories so that we merge some into "others" whose number of cells is less than one percent of the total sample. Methods not shown here fail to detect functional cell groups.

Supplementary Tables

Table S1 shows the enrichment scores obtained by comparing different methods through repeated experiments (10 repetitions) on different datasets, corresponding to the performance of different methods in discovering functional gene modules (as shown in Fig. 2b of the main text).

Table S1 Enrichment score comparison of different methods

datasets / methods	PBMC		HEART		LUAD		BC	
	mean	sd	mean	sd	mean	sd	mean	sd
GBC	23.75	3.49	24.54	4.70	42.20	3.63	34.31	2.97
plaid	9.19	2.63	0.48	1.53	0.74	1.57	0.00	0.00
Xmotif	22.59	1.70	0.00	0.00	0.00	0.00	0.00	0.00
CC	26.86	0.00	23.70	1.55	43.65	1.02	7.46	2.84
Bimax	5.43	3.266	0.00	0.00	0.00	0.00	0.00	0.00
FABIA	21.83	2.72	11.70	2.44	22.37	2.05	0.00	0.00
scBC	29.10	2.42	29.68	2.70	48.28	7.92	40.21	0.96

Tables S2 to S4 show the results of ARI, FMI, and AMI obtained by repeating experiments (10 repetitions) of different methods on different datasets, corresponding to the performance of different methods in cell clustering (as shown in Fig. 2c-d of the main text).

Table S2 ARI of different methods in different datasets

datasets / methods	PBMC		HEART		LUAD		BC	
	mean	sd	mean	sd	mean	sd	mean	sd
GBC	0.16	0.03	0.23	0.05	0.26	0.05	0.03	0.01
plaid	0.16	0.04	0.01	0.02	0.01	0.04	0.00	0.00
Xmotif	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00
CC	0.00	0.00	0.06	0.03	0.18	0.06	0.13	0.02
Bimax	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FABIA	0.04	0.01	0.01	0.01	0.05	0.01	0.00	0.00
scBC	0.30	0.03	0.36	0.07	0.46	0.09	0.19	0.09

Table S3 FMI of different methods in different datasets

datasets / methods	PBMC		HEART		LUAD		BC	
	mean	sd	mean	sd	mean	sd	mean	sd
GBC	0.33	0.04	0.37	0.04	0.48	0.05	0.33	0.02
plaid	0.28	0.03	0.04	0.11	0.07	0.17	0.00	0.00
Xmotif	0.25	0.04	0.00	0.00	0.00	0.00	0.00	0.00
CC	0.00	0.00	0.30	0.07	0.46	0.06	0.32	0.02
Bimax	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
FABIA	0.18	0.04	0.23	0.04	0.27	0.02	0.00	0.00
scBC	0.41	0.03	0.47	0.05	0.61	0.07	0.40	0.07

Table S4 AMI of different methods in different datasets

datasets / methods	PBMC		HEART		LUAD		BC	
	mean	sd	mean	sd	mean	sd	mean	sd
GBC	0.31	0.06	0.31	0.04	0.26	0.03	0.10	0.02
plaid	0.18	0.04	0.01	0.02	0.04	0.13	0.00	0.00
Xmotif	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00
CC	0.00	0.00	0.06	0.03	0.12	0.05	0.08	0.01
Bimax	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FABIA	0.06	0.02	0.02	0.01	0.07	0.01	0.00	0.00
scBC	0.47	0.03	0.47	0.05	0.53	0.08	0.25	0.07

Tables S5 to S10 show the results of 1-CE and F scores obtained by repeating experiments (100 repetitions for each setting) of different methods on simulated datasets under different settings. The values in parentheses represent standard deviations. They are corresponding to the comparison of the overall performance of biclustering (as shown in Fig. 3b-g of the main text).

Table S5 1-CE of different methods in different datasets（p=1000, n=300, L=3, $\times 10^{-3}$）

dropout / methods	0.5	0.55	0.6	0.65	0.7
GBC	6.51(1.09)	5.71(0.80)	5.86(0.77)	6.00(1.01)	6.61(1.11)
plaid	1.22(3.09)	0.68(2.11)	1.12(3.17)	0.92(2.49)	1.32(3.51)
Xmotif	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
CC	0.14(0.18)	0.23(0.24)	0.30(0.32)	0.48(0.40)	0.57(0.45)
Bimax	1.62(0.82)	1.63(0.88)	1.62(0.67)	1.40(0.88)	1.43(0.85)
FABIA	6.95(1.02)	7.34(1.17)	7.25(1.25)	6.77(1.40)	5.81(1.13)
scBC	13.98(1.78)	13.44(1.84)	12.24(2.15)	10.41(2.01)	9.84(1.50)

Table S6 1-CE of different methods in different datasets（p=3000, n=600, L=4, $\times 10^{-3}$）

dropout / methods	0.5	0.55	0.6	0.65	0.7
GBC	3.42(0.43)	2.98(0.36)	2.97(0.34)	2.95(0.33)	2.97(0.31)
plaid	0.70(2.05)	0.65(1.65)	0.83(1.73)	0.68(1.66)	0.69(1.57)
Xmotif	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
CC	0.05(0.05)	0.08(0.08)	0.11(0.10)	0.16(0.11)	0.22(0.15)
Bimax	0.55(0.23)	0.52(0.23)	0.47(0.23)	0.50(0.23)	0.45(0.25)
FABIA	4.21(0.42)	5.41(0.78)	5.33(0.65)	5.13(0.70)	4.52(0.83)
scBC	9.45(0.95)	9.33(0.92)	8.67(1.20)	7.75(1.23)	6.36(1.13)

Table S7 1-CE of different methods in different datasets（p=6000, n=1500, L=5, $\times 10^{-3}$）

dropout / methods	0.5	0.55	0.6	0.65	0.7
GBC	1.45(0.15)	1.22(0.12)	1.22(0.12)	1.23(0.13)	1.21(0.12)
plaid	0.69(1.49)	0.47(1.02)	0.54(1.13)	0.59(1.43)	0.16(0.55)
Xmotif	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
CC	0.01(0.02)	0.01(0.02)	0.01(0.01)	0.03(0.03)	0.06(0.04)
Bimax	0.26(0.11)	0.27(0.10)	0.24(0.11)	0.23(0.11)	0.22(0.11)
FABIA	2.39(0.26)	4.07(2.67)	3.31(1.40)	3.97(1.91)	3.40(1.79)
scBC	4.82(0.39)	4.80(0.38)	4.79(0.49)	4.70(0.47)	4.47(0.42)

Table S8 F score of different methods in different datasets（p=1000, n=300, L=3, $\times 10^{-2}$）

dropout / methods	0.5	0.55	0.6	0.65	0.7
GBC	1.31(0.39)	0.50(0.23)	0.63(0.58)	0.88(0.78)	1.03(0.65)
plaid	0.23(0.57)	0.13(0.38)	0.20(0.59)	0.17(0.47)	0.22(0.56)
Xmotif	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
CC	0.03(0.04)	0.05(0.05)	0.06(0.07)	0.10(0.09)	0.12(0.09)
Bimax	0.32(0.18)	0.33(0.18)	0.33(0.14)	0.28(0.18)	0.28(0.17)
FABIA	2.62(0.82)	2.07(0.57)	1.76(0.48)	1.44(0.49)	1.02(0.25)
scBC	5.06(4.83)	3.72(3.35)	3.32(3.41)	2.34(2.72)	1.54(0.54)

Table S9 F score of different methods in different datasets（p=3000, n=600, L=4, $\times 10^{-2}$）

dropout / methods	0.5	0.55	0.6	0.65	0.7
GBC	0.49(0.38)	0.16(0.03)	0.18(0.11)	0.17(0.07)	0.37(0.46)
plaid	0.13(0.40)	0.12(0.30)	0.15(0.31)	0.12(0.29)	0.12(0.26)
Xmotif	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
CC	0.01(0.01)	0.02(0.02)	0.02(0.02)	0.03(0.02)	0.04(0.03)
Bimax	0.11(0.05)	0.10(0.05)	0.09(0.05)	0.10(0.05)	0.09(0.05)
FABIA	2.66(0.55)	3.04(0.74)	2.49(0.58)	2.02(0.50)	1.51(0.44)
scBC	9.84(1.75)	8.41(3.12)	5.03(3.82)	3.03(3.06)	1.10(1.01)

Table S10 F score of different methods in different datasets（p=6000, n=1500, L=5, $\times 10^{-2}$）

dropout / methods	0.5	0.55	0.6	0.65	0.7
GBC	0.13(0.08)	0.05(0.04)	0.05(0.01)	0.06(0.04)	0.09(0.11)
plaid	0.13(0.29)	0.09(0.19)	0.09(0.18)	0.11(0.26)	0.03(0.10)
Xmotif	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
CC	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.01)	0.01(0.01)
Bimax	0.05(0.02)	0.05(0.02)	0.05(0.02)	0.05(0.02)	0.04(0.02)
FABIA	2.72(0.51)	2.53(1.44)	2.15(1.24)	1.99(1.21)	1.46(0.89)
scBC	2.91(0.65)	2.82(0.63)	2.71(0.70)	2.54(0.72)	1.74(0.78)