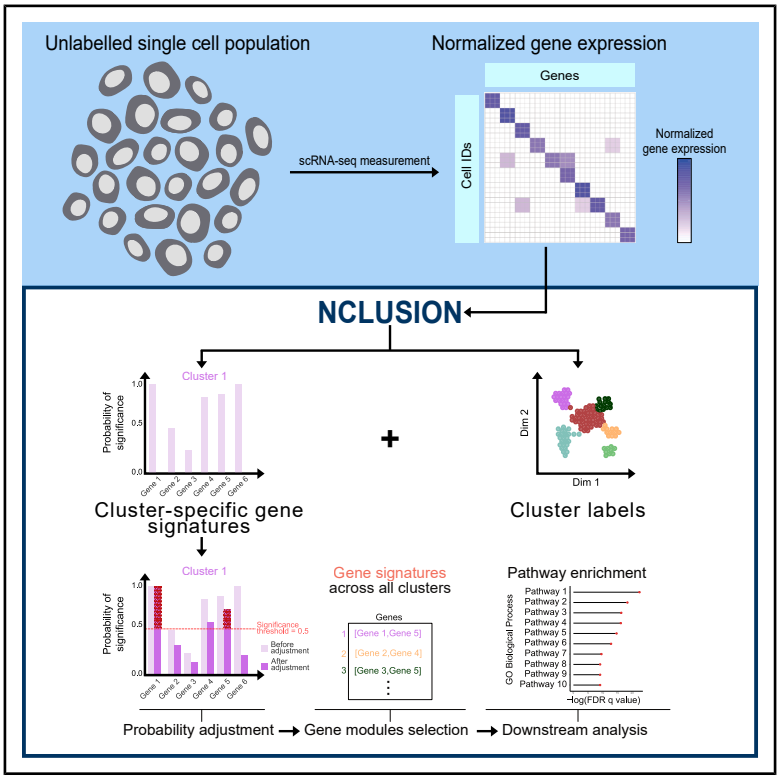
Cell Reports
Methods

Article

Scalable nonparametric clustering with unified marker gene selection for single-cell RNA-seq data

Graphical abstract



Authors

Chibuikem Nwizu, Madeline Hughes, Michelle L. Ramseier, ..., Peter S. Winter, Ava P. Amini, Lorin Crawford

Correspondence

ava.amini@microsoft.com (A.P.A.), lcrawford@microsoft.com (L.C.)

In brief

Nwizu et al. develop a nonparametric statistical model that simultaneously clusters single-cell data and identifies cell type marker genes. They demonstrate the utility of their approach on simulations and five different publicly available single-cell RNA sequencing (scRNA-seq) studies.

Highlights

- NCLUSION is a nonparametric clustering algorithm for scRNA-seq data analysis
- The method does not require users to specify the number of clusters *a priori*
- It jointly learns cluster-defining marker genes while learning cluster memberships
- Variational inference enables the model to scale up to millions of cells

Nwizu et al., 2026, Cell Reports Methods *6*, 101329
March 23, 2026
https://doi.org/10.1016/j.crmeth.2026.101329

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Scalable nonparametric clustering with unified marker gene selection for single-cell RNA-seq data

Chibuikem Nwizu,[1,2] Madeline Hughes,[3] Michelle L. Ramseier,[4,5,6,7,8,12] Andrew W. Navia,[4] Alex K. Shalek,[4,5,6,7,8] Nicolo Fusi,[3] Srivatsan Raghavan,[4,9,10,11,13,14] Peter S. Winter,[4,13,14] Ava P. Amini,[3,13,14,*] and Lorin Crawford[3,13,14,15,*]

[1]Center for Computational Molecular Biology, Brown University, Providence, RI 02906, USA
[2]Warren Alpert Medical School of Brown University, Providence, RI 02906, USA
[3]Microsoft Research, Cambridge, MA 02142, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[5]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[6]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[7]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[8]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA
[9]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[10]Harvard Medical School, Boston, MA 02115, USA
[11]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[12]Present address: Genentech, South San Francisco, CA 94080, USA
[13]These authors contributed equally
[14]Senior author
[15]Lead contact
*Correspondence: ava.amini@microsoft.com (A.P.A.), lcrawford@microsoft.com (L.C.)
https://doi.org/10.1016/j.crmeth.2026.101329

MOTIVATION Common clustering methods for single-cell pipelines require user-specified heuristics and secondary differential expression analyses, which lead to inflated false discovery rates. We, therefore, developed NCLUSION, which performs clustering and identifies marker genes simultaneously. NCLUSION eliminates the need for human-in-the-loop decisions, reducing the runtime and complexity of single-cell analyses.

SUMMARY

Clustering is commonly used in single-cell RNA sequencing (scRNA-seq) to assess cellular heterogeneity, but standard methods often require user-specified heuristics and rely on post-selective differential expression analyses, which often lead to inflated false discovery rates. Here, we present NCLUSION: a nonparametric infinite mixture model that leverages Bayesian sparse priors to identify marker genes and cluster single-cell expression data simultaneously. NCLUSION uses a variational inference algorithm, which enables it to scale up to millions of cells. Through simulations and analyses of publicly available scRNA-seq studies, we demonstrate that NCLUSION (1) matches the performance of other state-of-the-art clustering techniques with significantly reduced runtime and (2) provides statistically robust and biologically relevant transcriptomic signatures for each of the clusters it identifies. Overall, NCLUSION represents a reliable hypothesis-generating tool for understanding patterns of expression variation present in single-cell populations.

INTRODUCTION

Recent advances in sequencing technologies have increased the throughput of genomic studies to millions of single cells, necessitating computational workflows to explore and analyze these data.[1] In single-cell RNA sequencing (scRNA-seq), unsupervised clustering and marker gene selection are integral steps in the exploratory phase of analyses.[2–5] Clustering facilitates the identification of cell types, while marker gene selection enables the annotation of gene modules and cluster-specific biological programs. However, there has yet to be a consensus on the best approach to clustering cells and identifying the transcriptomic signatures that characterize them.[6,7] This has resulted in a multitude of proposed clustering methods for single-cell data, many of which are reliant on various user-defined heuristics that prevent practitioners from performing an unbiased survey of data and limit each method's "out-of-the-box" applicability when analyzing multiple studies.

Many current clustering approaches take a subset of highly variable genes as input, use dimensionality reduction techniques to simplify the representation of single-cell expression for these genes, and then perform clustering on top of this reduced representation. K-nearest neighbor (KNN) algorithms,[8] for example, generate nearest-neighbor (NN) graphs using transcriptomic similarity scores between cells and then perform Louvain clustering on the estimated graphs. Popular methods such as Seurat[9] and scLCA[10] use principal-component analysis (PCA) and singular value decomposition to learn a lower-dimensional representation of cells, respectively. Ensemble approaches such as scCCESS-SIMLR[11] learn over a mixture of kernels to generate a final cell-cell similarity matrix, which is then used in a spectral clustering algorithm.

Notably, selecting an appropriate embedding for single-cell data is not always a straightforward task. Previous studies have shown that if the generated lower-dimensional representation does not accurately capture underlying biological relationships between cells, then both the quality of clustering and the generalizability of findings in downstream analyses can be compromised.[12,13] Factors such as the number of highly variable genes retained during data preprocessing and the number of components used to define the lower-dimensional embedding can affect the ability of a clustering algorithm to identify fine-grained differences between cell types.[7,14] Furthermore, nearly all state-of-the-art clustering methods require users to specify the number of clusters K to be used in the algorithm. Strategies such as consensus finding,[11,15] outlier detection,[16] and iterative cluster merging and splitting[17–19] rely on human-in-the-loop interactive steps within their algorithms to determine an "optimal" choice for K. Generally, requiring users to make these additional decisions can add significant time and complexity when using clustering as a preliminary analysis in bioinformatic workflows.

Perhaps the biggest limitation of current single-cell clustering algorithms is that most do not directly identify top marker genes that are driving the inference of different biologically significant clusters; instead, they use post-selective inference to find genes that are differentially expressed between the inferred cell groups.[7,20] For example, MAST implements a two-component mixture model that accounts for both biological condition and technical covariates, such as cluster identity, to jointly model gene expression levels and the probability of dropout events in single-cell data.[21] In contrast, limma-voom applies ordinary least-squares regression to model the association between cluster identity and a cell's log-counts per million.[22] Both approaches require known specification of cluster identities and exhibit poor computational scalability when applied to large datasets.[20] Moreover, since the point of clustering algorithms is to separate dissimilar data into different groups, it is expected *a priori* that there are differences between the groups and any test statistics computed by comparing the groups (e.g., via a t test or Wilcoxon rank-sum) are likely to be inflated due to "data double dipping. Many studies have shown that performing this post-selective inference uncorrected can lead to inflated type I error rates.[20,23] Though there has been work developed to correct for post-selection, these are still in nascency, and most do not yet scale to high-dimensional settings.[24–29] Recently, others have proposed unified frameworks for simultaneous clustering and marker gene selection using hierarchical tree-based algorithms,[30] regularized copula models,[31] and post hoc sensitivity measures.[32] However, these approaches rely on arbitrary thresholding to find "significant" marker genes and fail to theoretically test a well-defined null hypothesis, making them difficult to biologically interpret.

We present NCLUSION (nonparametric clustering of single-cell populations): a unified Bayesian nonparametric framework that simultaneously performs clustering and marker gene selection. NCLUSION works directly on normalized single-cell count data, bypassing the need to perform dimensionality reduction. By modeling the expression of each gene as a sparse hierarchical Dirichlet process normal mixture model,[33–39] NCLUSION both learns the optimal number of clusters based on the variation observed between cellular expression profiles and uses sparse prior distributions to identify genes that significantly influence cluster definitions. The key to our proposed integrative framework is that clustering and extracting marker genes concurrently is a more efficient approach to the exploratory analysis of scRNA-seq data, as it effectively allows each process to inform the other. Most importantly, our approach eliminates the need for human-in-the-loop decisions, significantly reducing the runtime and complexity of these analyses. Altogether, NCLUSION mitigates the need for heuristic choices (e.g., choosing specific lower-dimensional embeddings), avoids iterative hyper-parameter optimization, bridges the interpretability gap suffered by many unsupervised learning algorithms in single-cell applications, and scales to accommodate the growing sizes of emerging scRNA-seq datasets.

RESULTS

NCLUSION simplifies traditional clustering workflows

Conventional scRNA-seq clustering approaches include numerous steps that require user heuristics or human-in-the-loop decisions, which increases runtime and complexity (Figure 1A). These can range from deciding how to optimally embed high-dimensional expression data into a lower-dimensional space to selecting the number of clusters, K, to identify in the data. Furthermore, current methods require that marker gene selection is performed post-clustering, which can lead to inflated rates of false discovery.[20,23] In this work, we aim to address these challenges using NCLUSION.

NCLUSION leverages a Bayesian nonparametric modeling framework to reduce the number of choices that users need to make while simultaneously performing variable selection to identify top cluster-specific marker genes for downstream analyses (Figure 1B; see STAR Methods for details). There are three key components of our model formulation that distinguish it from traditional bioinformatic workflows. First, NCLUSION is fit directly on single-cell expression matrices and does not require the data to be embedded within a lower-dimensional space. Second, we implicitly assume *a priori* that cells can belong to one of infinitely many different clusters. This is captured by placing a Dirichlet process prior over the cluster assignment probabilities for each cell. By allowing the number of possible clusters $K = \infty$, we remove the need for users to iterate through

CellPress
OPEN ACCESS

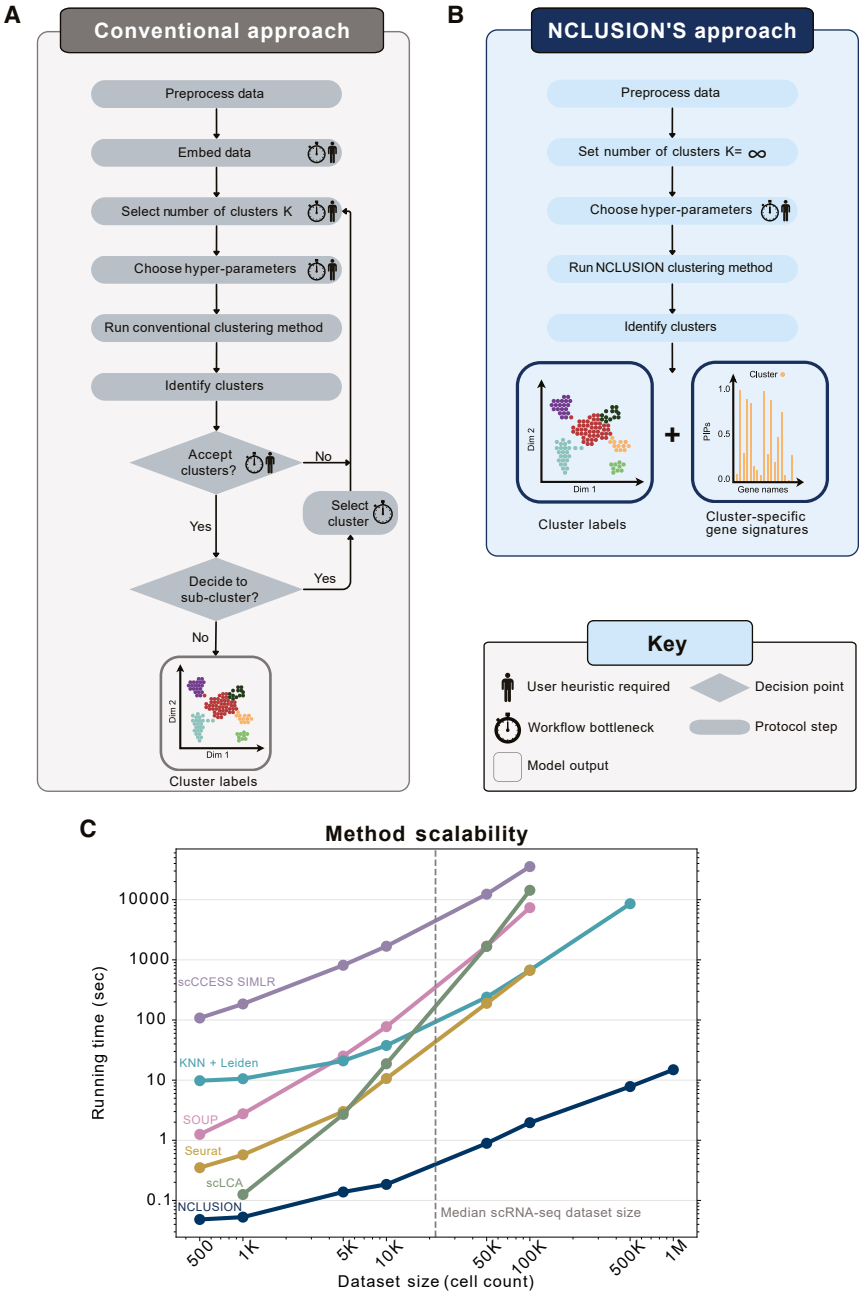


Figure 1. NCLUSION provides a scalable, unified workflow for both clustering and marker gene selection in single-cell analysis

(A) Conventional clustering algorithms require user heuristics and decision-making steps that increase wall clock runtime (e.g., selection and human-in-the-loop refinement of the number of clusters *K*).

(B) The nonparameteric workflow of NCLUSION reduces the number of choices and heuristics that users have to make, while also performing cluster-specific variable selection to identify top marker genes for downstream investigation.

(C) Runtimes of NCLUSION and other baselines on the BRAIN-LARGE dataset with a fixed set of 720 genes and an increasing sample size ranging from $N = 500$ to 1 million cells.

See also Methods S1, Figure S1, and Data S1.

different values until they find the optimal choice. Third, NCLUSION assumes that not all genes are important when assigning cells to a given cluster. To model this, we place a spike and slab prior on the mean expression of each gene within each cluster. This prior shrinks the mean of genes that play an insignificant role in cluster formation toward zero. Note that we use "nonparametric" when describing NCLUSION to match the nomenclature of previous work,[33–39] but indeed this is approximated with the sparse hierarchical Dirichlet process normal mixture model (STAR Methods).

To identify cluster-specific marker genes, we start by estimating posterior inclusion probabilities (PIPs), which represent our confidence that a gene's mean expression within a cluster is nonzero. These PIPs act as a signature that can be used to distinguish clusters. Since NCLUSION fits each gene and cell independently, signatures learned between clusters can share the same subsets of genes. To select for unique cell type markers, we multiply each PIP with a usage weight, which is calculated by performing min-max normalization over the proportion of clusters in which a gene is determined to be statistically significant (i.e., using the median probability model threshold[40] PIP $\geq$0.5). We use these adjusted inclusion probabilities with the effect size sign (ESS) and strictly standardized mean difference (SSMD) of each gene to filter for significant genes that are substantially up-regulated (indicated by positive ESS and large SSMD values). The genes remaining after filtering make up cluster-specific marker gene modules that provide insight into the biological features underlying cluster assignments.

To enable efficient posterior inference that can scale as the size of scRNA-seq datasets continues to grow, we train NCLUSION using variational expectation-maximization (EM). This algorithm leverages a "mean-field" assumption to approximate the true posterior distribution over model parameter estimates with a product of simpler distributions.[41–43] In training, our objective is to minimize the Kullback-Leibler (KL) divergence between the variational posterior and the true posterior.[44] Optimization during model fitting occurs using a coordinate ascent procedure where parameters are sequentially updated based on their gradients (STAR Methods and Methods S1). With this variational approach, NCLUSION is capable of scaling well up to 1 million cells without applying any dimensionality reduction to the input data.

We evaluated the runtime of NCLUSION against a set of state-of-the-art single-cell clustering methods using publicly available datasets. The methods we used for comparison include Seurat,[9] scLCA,[10] KNN followed by the Leiden clustering algorithm (KNN+Leiden),[45] SOUP,[46] and scCCESS-SIMLR.[11] Each of these methods operates by first reducing the dimensionality of the input data and then performing clustering on the reduced representation. To the best of our knowledge, NCLUSION is the only method to date that clusters on the expression matrix directly while jointly identifying cluster-specific salient genes. NCLUSION does not incur additional runtime cost due to this model design choice. In fact, NCLUSION boasts a faster runtime compared to other methods, particularly as the number of cells in a dataset grows. We showcase this scalability on the BRAIN-LARGE[47] dataset (Figure 1C). To facilitate comparison across all baseline methods, we follow Lopez et al.[14] and limit the analysis to 720 genes, while subsampling the number of cells from 500 to 1 million cells. As a practical reference, we use a gray dotted line to highlight the runtime for each method at the median scRNA-seq dataset size as determined in 2020: 31,000 cells.[48] We observed that only NCLUSION and KNN+Leiden were able to scale past 100,000 cells, while NCLUSION was the only method able to run on 1 million cells. Additionally, NCLUSION records competitive runtimes across varying numbers of genes (Figure S1). These scalability results show the potential of NCLUSION's utility for emerging large-scale single-cell studies.

NCLUSION accurately identifies clusters and marker genes in simulations

We used simulations to evaluate the performance of NCLUSION in controlled settings. Here, we used `scDesign3`[49] to generate synthetic datasets consisting of 10,000 cells and 1,000 genes (due to computational constraints of the software) distributed over five clusters. We considered four different scenarios (with 20 replicates per scenario), where we varied cluster size and marker gene composition (STAR Methods). Scenario I was the simplest, where we evenly distributed all cells across the five clusters and each cluster had 50 marker genes. In scenarios II and III, all five clusters had 50 marker genes but one cluster had significantly fewer cells than the other four clusters. Last, in scenario IV, each of the five clusters had the same number of cells, but one cluster had a signature of only 20 marker genes, while the other four clusters had 50 marker genes each. All results are displayed in Figures 2, S2, S3, and Data S1.

We first compared the cluster identification accuracy of NCLUSION with the previously described algorithms: Seurat,[9] scLCA,[10] KNN followed by the Leiden clustering algorithm (KNN+Leiden),[45] SOUP,[46] and scCCESS-SIMLR.[11] The relatively smaller size of these simulated datasets also allowed us to include three additional widely used methods: CIDR,[50] SC3,[15] and scDeepCluster.[51] We should highlight that all competing methods, except for NCLUSION, first perform dimensionality reduction prior to clustering, allowing us to evaluate the impact of dimensionality reduction on cluster recovery.

The normalized mutual information (NMI) and adjusted Rand index (ARI) were calculated to quantitatively evaluate the clustering results given by each algorithm. In all scenarios, scDeepCluster had the best performance, with NCLUSION, scLCA, and scCCESS-SIMLR rounding out the top four. The reason for scDeepCluster's top performance is that it first performs Leiden clustering on principal components from the expression data to obtain the initial cluster assignments. Then, it uses a deep neural network to refine the cluster assignments; therefore, performance is highly dependent on the initialization of the cluster assignments and on the success of the initial clustering. NCLUSION, on the other hand, only performs clustering once to achieve comparable results. Notably, all methods performed relatively worse in scenarios II and III than their performance in scenarios I and IV (Figures 2B and S2; Data S1) due to the class imbalance in how the cells were distributed across clusters.

We also evaluated the performance of NCLUSION on marker gene detection using the same simulated datasets and compared it with three popular differential expression algorithms: DUBStepR,[52] singleCellHaystack,[27] and FESETM.[53] Notably, of these methods, only NCLUSION is able to find cluster-specific marker genes. FESTEM uses an alternative algorithm (via the Scott-Knott test) to assign marker genes to clusters found by an independent clustering method, while both DUBStepR and singleCellHaystack aim to identify differentially expressed genes. To that end, we evaluated the global marker gene detection of each model using true-positive rate (TPR), false discovery rate (FDR), and false-positive rate (FPR; computed as 1−Specificity for each method) (STAR Methods; Table 1). We found no statistically significant differences when comparing the median power of NCLUSION to the other approaches (Kruskal-Wallis H-test $p > 0.99$; Figures 2C and S3; Data S1). However, importantly, NCLUSION and FESTEM were the only two methods to have high power, while maintaining a low FDR and FPR. On the other hand, singleCellHaystack and DUBStepR achieved similar power but only because they incorrectly labeled many genes as being marker genes (resulting in markedly higher FDR and FPR).

NCLUSION achieves competitive clustering performance on PBMC data with less runtime

Next, we assessed the quality of clustering done by NCLUSION as compared to other baseline approaches on real data. Here, we analyzed scRNA-seq from fluorescence-activated cell sorting (FACS)-purified peripheral blood mononuclear cells (PBMCs).[54] This dataset captures 10 cellular populations, including $CD14^+$ monocytes, $CD34^+$ cells, major lymphoid lineages (B and T cells), as well as other phenotypic lineages within the T cell population, including $CD4^+$ helper T cells, $CD8^+$ cytotoxic T cells, $CD4^+$ regulatory T cells, and $CD4^+$ memory T cells. After quality control (STAR Methods), the final dataset contained 94,615 cells and 5,000 genes with the highest standard deviation (post-log-normalization).[20,55] We evaluated the performance of NCLUSION, Seurat, scLCA, the KNN+Leiden algorithm, SOUP, and scCCESS-SIMLR by comparing inferred cluster assignments to the cell type annotations from the original study, which were obtained via a combination of FACS analysis and clustering with Seurat[54] (Figures 3A and 3B).

CellPress
OPEN ACCESS

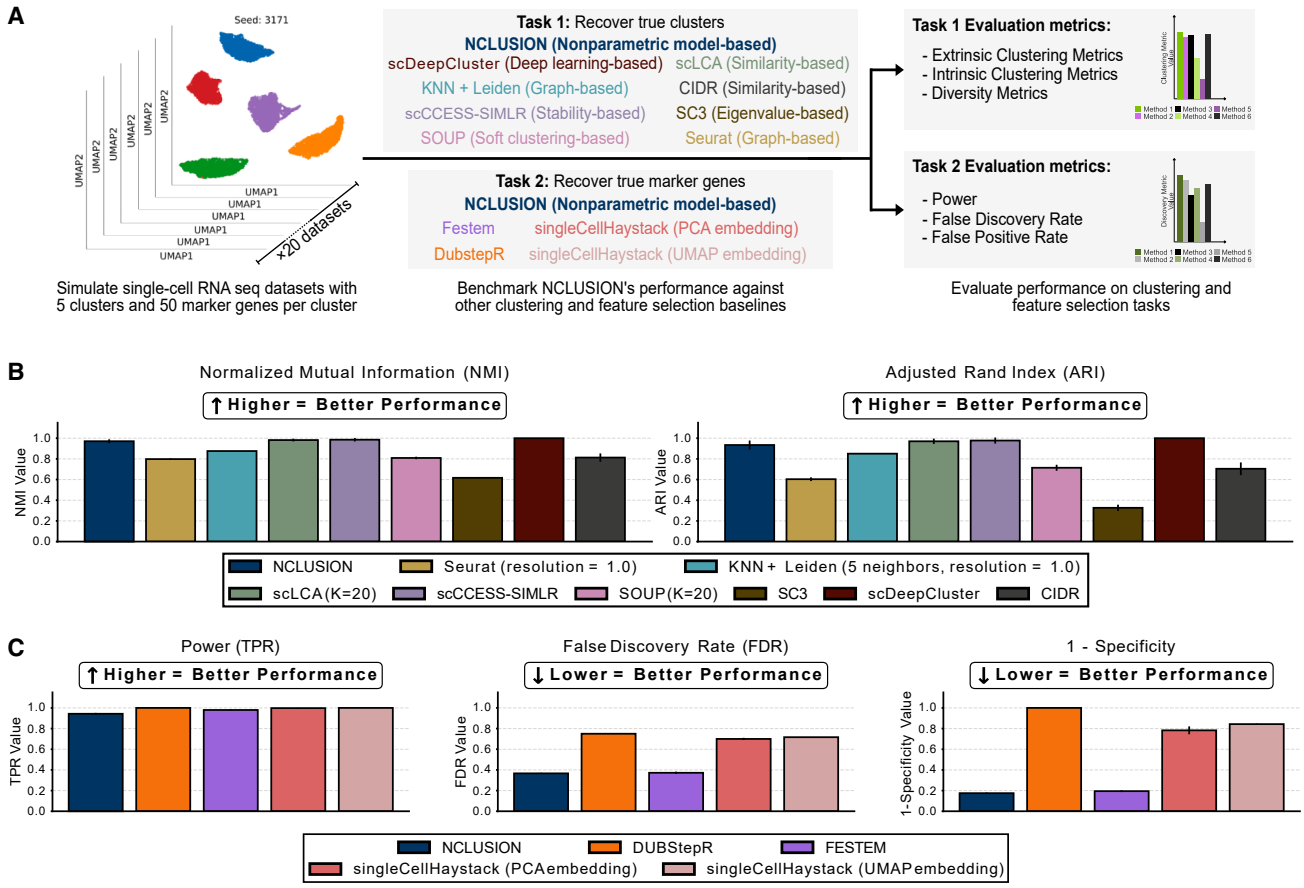


Figure 2. Comparing NCLUSION and competing algorithms on performing clustering and marker gene selection in a simulation study

Depicted are results for scenario I where we evenly distributed all synthetically generated cells across five clusters and each cluster had a unique set of 50 marker genes.

(A) Overview of the simulation framework used for evaluating the quality of clustering and marker gene selection for NCLUSION and each competing method.

(B) Inferred cluster labels were compared to "true" annotations created during the simulation, where performance was measured according to (left) NMI and (right) ARI.

(C) Assessment of marker gene selection was done on the global scale, where methods were evaluated on how well they could detect a "true" causal gene without taking cluster assignment into account. This was due to the limitation of competing methods not being able to identify cluster-specific genes. Evaluations were done by measuring the TPR (or power), FDR, and FPR (computed as 1−specificity) for each approach.

Results for (B) and (C) are based on 20 simulations, with each bar plot representing the mean and the error bars covering a ±95% confidence interval.

See also Figures S2 and S3 and Data S1.

To qualitatively assess clustering performance, we used contingency heat maps to evaluate how well each method captured the unique cell types across clusters (Figure 3C). For a given method, each n-th row of the heatmap represents an annotation from the original study and each k-th column represents an inferred cluster identified by the method. The color saturation of each (n,k)-th element in the heatmap indicates the fraction of an n-th cell annotation that a given method assigned to the k-th inferred cluster. Overall, all baselines were able to distinguish the B cell population from other cells, and each approach uniquely clustered a majority of the $CD14^+$ monocytes together (Figure 3C). Furthermore, all methods except scCCESS-SIMLR and KNN+Leiden were able to separate the natural killer (NK) cell population with an inferred cluster occupancy rate of greater than 90% (Data S1). NCLUSION's performance was most similar to that of Seurat (χ^2 test p = 0.99 when assessing independence between their contingency tables). Both methods were able to identify major PBMC cell lineages and were the only approaches to divide the B cell, cytotoxic $CD8^+$ T cells, and regulatory T cells into subpopulations[56] (Figure 3C).

As a quantitative assessment of each method's clustering performance, we used the FACS-derived experimental annotations as reference labels and computed NMI and ARI, each measuring how well the clustering algorithm's labels matched the reference labels (STAR Methods). For both metrics, values closer to 1 indicate better clustering performance. To assess the robustness and consistency of each method, we ran them on five different randomly subsampled partitions containing 80% of the cells in the dataset. We report the mean metric score for each clustering algorithm across these partitions, along with corresponding 95% confidence intervals (Figure 3D).

Table 1. Confusion matrix showing true and inferred marker gene labels

		(Inferred label)		
		Marker gene	Non-marker gene	Total
(True label)	Marker gene	TP	FN	b_1
	Non-marker gene	FP	TN	b_2
Total		a_1	a_2	n

Here, TP represents the number of correctly identified marker genes (true positives), FN represents the number of incorrectly identified marker genes (false negatives), TN represents the number of correctly identified non-marker genes (true negatives), and FP represents the number of incorrectly identified non-marker genes (false positives). In the table above, we let a_1 and a_2 be the total number of genes inferred as markers and the total number of genes not inferred as markers, respectively. Likewise, we let b_1 and b_2 be the total number of genes that are truly markers and non-markers, respectively. It follows that the total number of genes is defined as $n = a_1+b_1+a_2+b_2$. From here, we can compute the following metrics.

NCLUSION outperformed all competing approaches across both metrics. It obtained the highest mean NMI coefficient of 0.80 ($\pm 1.62 \times 10^{-2}$ standard deviation), with Seurat and scLCA each scoring lower values of 0.77 ($\pm 2.78 \times 10^{-3}$) and 0.62 ($\pm 3.04 \times 10^{-2}$), respectively. These differences were statistically significant, as determined by two-sided t tests ($p = 1.08 \times 10^{-3}$ and $p = 1.90 \times 10^{-6}$, respectively). When comparing performance using the ARI, NCLUSION remained competitive and significantly outperformed other methods (Data S1). Specifically, NCLUSION achieved a higher mean ARI of 0.67 ($\pm 2.02 \times 10^{-2}$) when compared to 0.62 ($\pm 2.48 \times 10^{-3}$; two-sided t tests $p = 1.17 \times 10^{-3}$) for Seurat and 0.50 ($\pm 2.52 \times 10^{-2}$; two-sided t tests $p = 3.10 \times 10^{-6}$) for scLCA.

Last, NCLUSION recorded the shortest runtime for this analysis without any iterative processes or optimization of hyperparameters. It finished approximately 254 s (4.23 min) faster than Seurat, more than 12,900 s (215.00 min or 3.58 h) faster than scLCA, and more than 41,331 s (688.85 min or 11.48 h) faster than scCCESS-SIMLR.

NCLUSION is well-powered to identify PBMC-specific marker genes

The key distinguishing property of NCLUSION is its inherent ability to perform variable selection. NCLUSION thus provides users with cluster labels for each cell as well as unique gene signatures that define each cluster. The statistical model underlying NCLUSION selects cluster-specific marker genes based on three criteria: (1) an adjusted PIP, PIP(j;k), which provides evidence that the j-th gene's mean expression is uniquely nonzero within the k-th cluster; (2) the sign of the j-th gene's effect, ESS(j;k), which is used to determine whether it is uniquely up-regulated or down-regulated within the k-th cluster; and (3) the magnitude of the j-th gene's effect, SSMD(j;k), on the definition of the k-th cluster (STAR Methods). We assessed the marker genes identified by NCLUSION for each of the inferred clusters in the PBMC dataset to determine whether they provide insights into the biology of different cell types (Figures 4A and 4B).

Overall, NCLUSION successfully identified cluster-specific marker genes that are known to be associated with examined cell types (Figure 4C and Data S1). For example, in cluster 8, which has cells mapping back to the cytotoxic and naive cytotoxic T cell population, we observed that NCLUSION correctly identified marker genes known to play an important role in cytotoxic T cell biology, such as *CD8A* (adjusted PIP = 0.90), *CD8B* (adjusted PIP = 0.70),[57] and *CD27* (adjusted PIP = 0.62).[58] Furthermore, genes associated with cytotoxicity, such as *GZMM*, tended to be selected in clusters largely containing $CD8^+$ T cells (cluster 8; adjusted PIP = 0.60) and NK cells (cluster 2; adjusted PIP = 0.61)—two cell types that have been shown to have functionally similar cytotoxic activity.[59] In other clusters, where we observed genes associated with both B cells (e.g., in cluster 1, *CD19*, adjusted PIP = 1.00; LINC00926, adjusted PIP = 1.00; *MS4A1*, PIP = 0.90)[60–62] and myeloid lineages (e.g., in cluster 3, *MS4A6A*, PIP = 1.00; S100A8, adjusted PIP = 0.93; *LYZ*, PIP = 0.90),[63] NCLUSION distinguished genes known to play an important role in T cell biology as statistically significant. This observation suggests that NCLUSION accounted for the variance among T cell and T cell-like expression patterns when distinguishing cell types in the PBMC dataset. We observed that our criteria for cluster-specific marker genes, based on high PIPs and positive ESS scores, strongly agreed with the relative over-expression of each gene in its respective cluster (Figure 4C). Imposing a threshold on SSMD allowed NCLUSION to filter out less-relevant genes, narrowing the list of cluster-specific over-expressed genes to those most salient.

To further evaluate marker gene quality, we computed gene module scores in order to compare the normalized expression for signature genes across clusters (Figure 4D and Data S1). Here, we find that each module exhibits the highest expression within its respective cluster, with the most definitive signatures occurring within the B (inferred cluster 1), NK (inferred cluster 2), monocytic (inferred cluster 3), and $CD34^+$ cells (inferred cluster 4) (Figures 4E and S4). In clusters that contained heterogeneous combinations of T cell subpopulations, we still see an increased relative expression among cluster-specific marker genes, although not as distinct as in the other cell types.

As an additional analysis, we compared the similarity between the marker genes identified by NCLUSION with the list of marker genes that are identified by using a post hoc differential expression analysis with Seurat. Here, we took the FACS-derived experimental annotations from Zheng et al.[54] and found differentially expressed genes by doing a one-versus-all Wilcoxon rank-sum test for each cluster (mirroring the typical procedure in a conventional bioinformatic workflow). As expected, this post-selective inference procedure resulted in Seurat identifying a multitude of candidate marker genes for each cell type, even after Bonferroni correction. A direct comparison between NCLUSION and Seurat showed that the proposed Bayesian

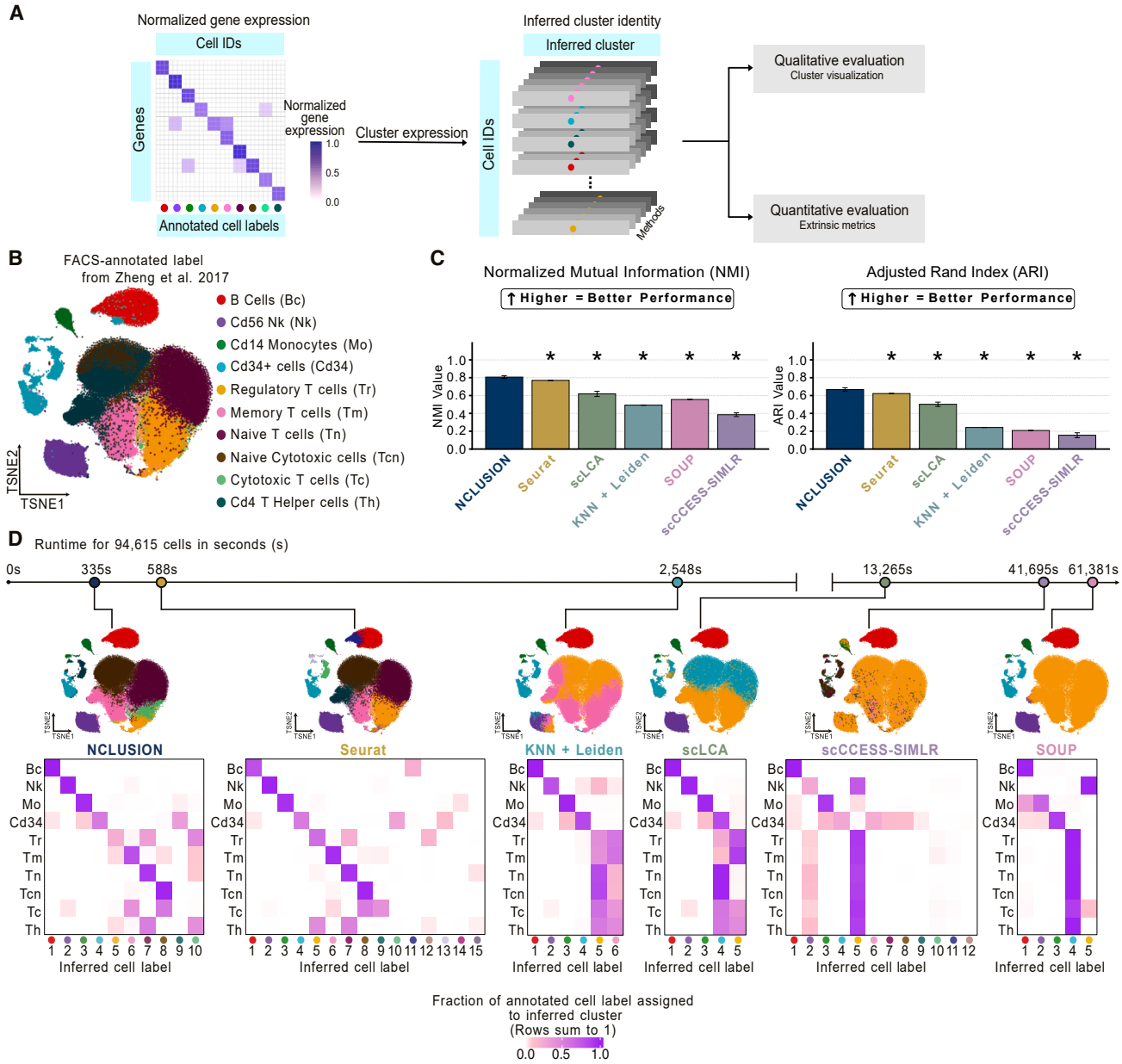


Figure 3. Clustering performance for NCLUSION and other baseline methods on the PBMC scRNA-seq dataset (*N* = 94,615 cells)

(A) The framework used for evaluating the quality of clustering in each method.

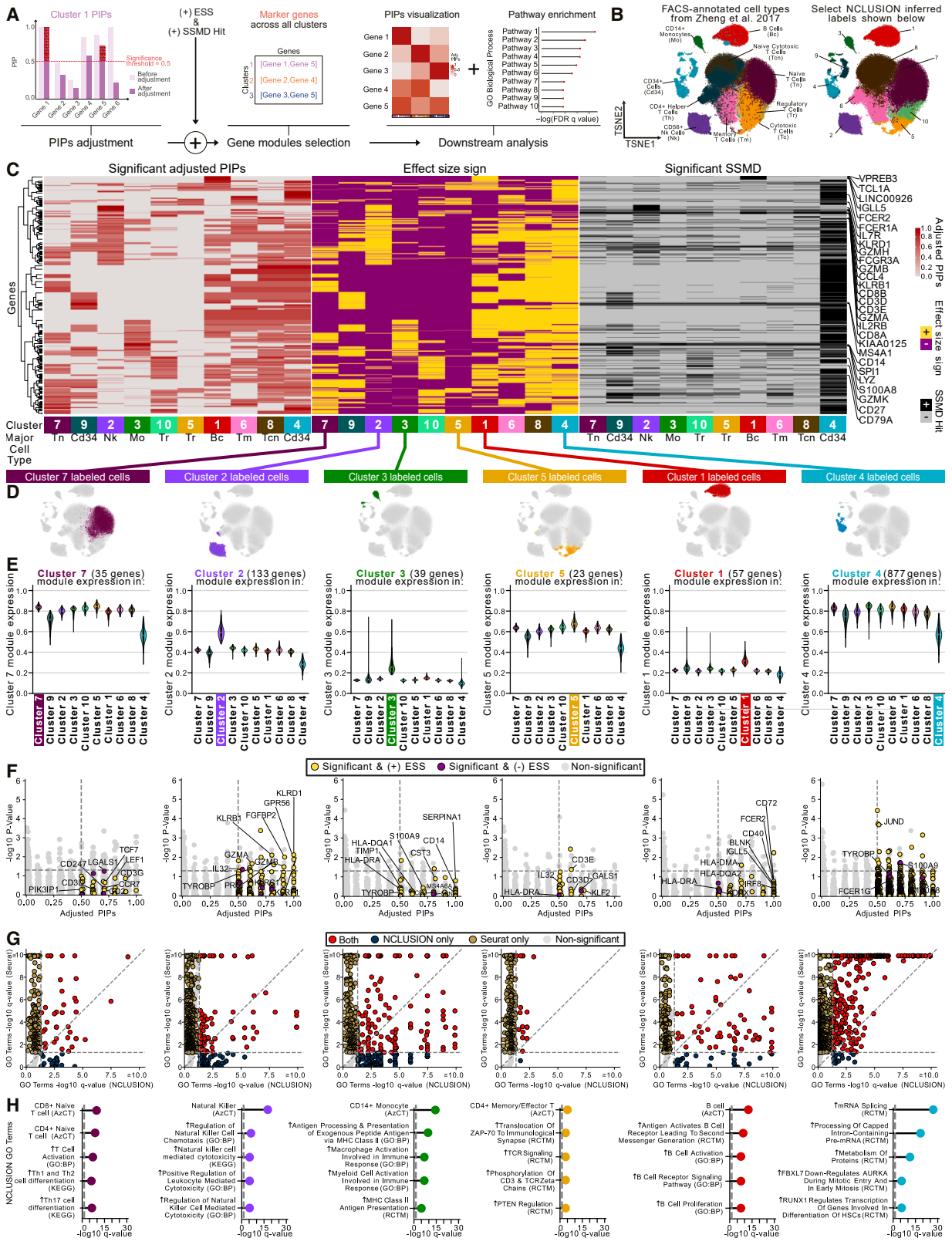
(B) Overview of FACS-based cell type annotations, visualized via t-distributed stochastic neighbor embedding (t-SNE), for the PBMC scRNA-seq dataset. These annotations serve as labels during the evaluation.

(C) Assessment of the inferred cluster labels versus the experimental annotations, as quantified by two metrics: NMI and ARI (for each method, we take five random 80% splits of the PBMC dataset; depicted in each bar plot is the mean ±95% confidence interval). Asterisks indicate that there is a statistically significant difference in performance between NCLUSION and a corresponding method (two-sided *t* test $p < 0.05$).

(D) Visualizing the structure of the inferred clusters across all baselines using t-SNEs and a contingency heatmap showing the prevalence of each cell type within each cluster. Methods are ordered from fastest (left) to slowest (right) in terms of runtime. The same lower-dimensional representation of the data is reused with relabeling of the plots according to the results from each clustering algorithm.

See also Data S1.

variable selection approach in the NCLUSION framework results in smaller and more refined transcriptomic signatures for downstream investigation (Figures 4F, S5, and S6). For example, NCLUSION identified 134 cluster-specific marker genes for NK cells, 97% of which were also included in the 1,780 marker genes selected post hoc by Seurat. In total, an average of 96% of the marker genes identified by NCLUSION were included in the much larger sets of differentially expressed



(legend on next page)

genes selected by Seurat across each of the FACS-annotated cell types.

Over-representation analysis using gene product annotations in Gene Ontology (GO) further confirmed that the selective set of gene modules inferred by NCLUSION reflect known immune cell biology[64–66] (Figures 4G and 4H and Data S1). For example, in the cluster containing predominantly B cells (inferred cluster 1), we observed an up-regulation of B cell receptor signaling (adjusted $p = 6.68 \times 10^{-7}$), B cell activation (adjusted $p = 2.3 \times 10^{-7}$), and B cell proliferation (adjusted $p = 2.51 \times 10^{-7}$).[67,68] Notably, some of the biologically relevant GO terms found when using NCLUSION were not statistically significant when applying the larger sets of marker genes provided post hoc by Seurat. For instance, in the cluster with NK cells (inferred cluster 2), significant gene sets from NCLUSION included the positive regulation of leukocyte chemotaxis (adjusted $p = 1.53 \times 10^{-4}$) and positive regulation of NK cell chemotaxis (adjusted $p = 4.88 \times 10^{-6}$).[69–71] We also observed enrichment of the up-regulation of NK cell-mediated cytotoxicity (adjusted $p = 3.81 \times 10^{-5}$), consistent with the known highly cytotoxic behavior of NK cells.[69,72] Each of these gene sets was insignificant when using differentially expressed genes from Seurat (Figure 4G). Last, in the monocyte-dominated cluster (inferred cluster 3), the NCLUSION-generated module was enriched for macrophage activation involved in immune response (adjusted $p = 2.41 \times 10^{-5}$) and antigen-presenting activity (e.g., MHC class II antigen presentation, adjusted $p = 5.39 \times 10^{-4}$). Complete marker gene and GO analyses for all clusters inferred by NCLUSION and Seurat as a baseline in the PBMC dataset can be found in Data S1. Together, these results demonstrate that NCLUSION can identify cluster-specific gene signatures that reflect underlying cellular phenotypes.

NCLUSION's performance generalizes to the other single-cell datasets

Finally, we assessed the generalizability of NCLUSION by testing it on three additional large scRNA-seq datasets of various sample sizes: a pancreatic ductal adenocarcinoma (PDAC) dataset from Raghavan et al.[73] with $N = 23,042$ cells, an acute myeloid leukemia (AML) dataset from van Galen et al.[74] with $N = 43,690$ cells, and a tissue immune (IMMUNE) atlas dataset from Domínguez Conde et al.[75] with $N = 88,057$ cells. These datasets represent a range of tissue and disease states to assess our method's performance in different use cases. Both the PDAC and AML datasets contain a mixture of malignant and non-malignant cells from different patient biopsies, while the IMMUNE dataset contains healthy white blood cells from different anatomical locations. After performing quality control (STAR Methods), we had a total of 5,000 genes with the highest standard deviation (after log-normalization) for the analysis. For each dataset, we ran each method on five different randomly subsetted partitions that contained 80% of total cells in the respective dataset, and we report the mean and corresponding 95% confidence interval.

We observed similar scalability in the runtime of NCLUSION and competing baselines on all three datasets (Figure 5A). NCLUSION maintains its computational efficiency, now only being slightly outperformed by Seurat on the PDAC and AML datasets due to longer convergence time in its variational EM algorithm. We also found that NCLUSION continued to remain competitive in terms of clustering performance. When quantitatively evaluating the clustering ability of NCLUSION versus the competing baselines using the annotations provided by the original studies, NCLUSION was often statistically significantly better (as determined via a two-sided t test, $p < 0.05$) according to ARI and NMI across all datasets (Figures 5B, 5C, and S7–S19; Data S1).

We then analyzed the interpretability of the cluster-specific marker genes inferred by NCLUSION (Figures 5D–5G and S8–S24; Data S1). For brevity, we highlight just notable results from the PDAC and IMMUNE datasets in the main text. Additional analyses for the AML dataset can be found in the supplemental information (see Figures S12–S17 and Data S1).

To begin, we focused on evaluating the gene modules generated from the NCLUSION-inferred clusters in the PDAC dataset (Figures 5D–5G). As with the PBMC data, we observed higher module expression within the respective clusters (Figure 5F). NCLUSION appeared to use immune cell signatures as the primary axis for distinguishing malignant and non-malignant populations (Figure 5E and Data S1). For example, the inferred cluster 6 predominantly contained cells that were originally

Figure 4. Evaluation of cluster-specific marker genes identified by NCLUSION on the PBMC dataset ($N = 94,615$ cells)

(A) The framework used for assessing cluster-specific marker genes.

(B) Embeddings of the experimental annotations for major cell types from the PBMC dataset compared to the clusters inferred by NCLUSION.

(C) Heatmaps of the adjusted PIPs (left), ESS (center), and SSMD (right) of significant genes in each cluster. Cluster-specific marker genes are selected as those that have a significant inclusion probability, are up-regulated in a given cluster, and have a large effect size magnitude such that PIP$\geq$0.5, ESS = +, and $|\text{SSMD}(j;k)| \geq S^*(j;k)$, respectively. Here $S^*(j;k)$ is a threshold set to preserve a false-positive rate of 0.05.

(D) Highlighted location on t-SNEs of NCLUSION-inferred clusters that contain predominantly one cell type.

(E) Violin plots comparing the normalized expression of cluster-specific marker genes in each of the inferred clusters.

(F) Scatterplot comparing the marker genes identified using post hoc differential expression analysis with Seurat (yellow) versus the variable selection approach with NCLUSION (blue). Yellow points have PIP$\geq$0.5 and ESS = +, while purple points have PIP$\geq$0.5 and ESS = −, respectively. The vertical dashed line marks the median probability criterion,[40] and the horizontal dashed line marks the Bonferroni-corrected threshold for significant q values (i.e., an adjusted p). Genes in the top right quadrant are identified by both methods.

(G) Scatterplot comparing GO pathway enrichment analyses using cluster-specific marker genes from Seurat versus NCLUSION. The horizontal and vertical lines correspond to significant q values being below 0.05. Pathways in the top right quadrant are selected by both approaches (red), while elements in the bottom right and top left quadrants are uniquely identified by NCLUSION (blue) and Seurat (orange), respectively.

(H) Highlight of select top GO pathway enrichment analysis for the marker genes identified by NCLUSION. Plotted on the x axis are the negative log-transformed q values for each GO gene set. Gene sets with a q value below 0.05 are deemed to be significant.

See also Figures S4–S6 and Data S1.

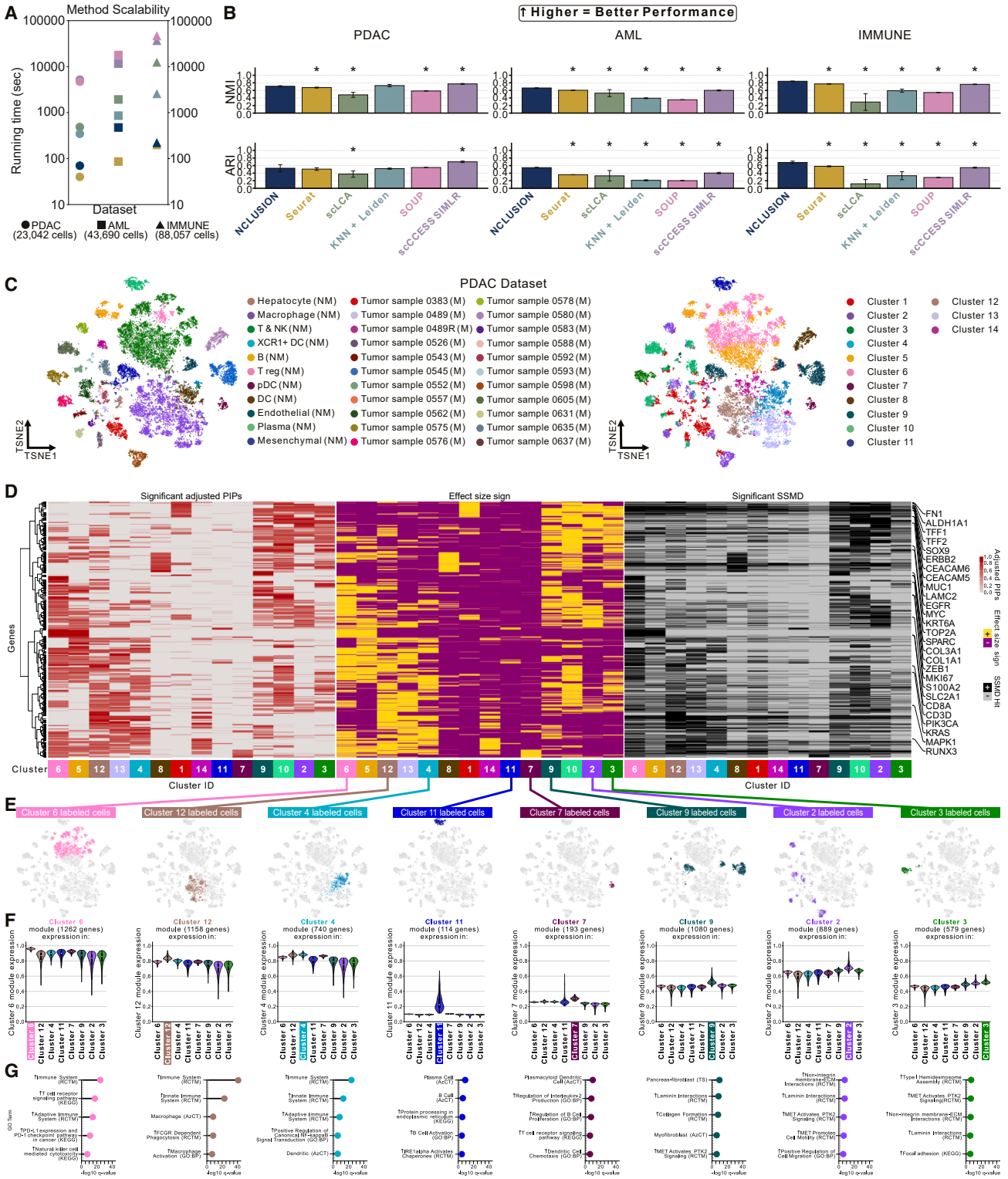


Figure 5. Scalability and generalizability of NCLUSION across diverse datasets

NCLUSION and baselines were applied to the following scRNA-seq datasets: PDAC (N = 23,042 cells),[73] AML (N = 43,690 cells),[74] and IMMUNE (N = 88,057 cells).[75]

(A) Runtimes for all methods when applied to each dataset.

(legend continued on next page)

annotated as NK and T cells by Raghavan et al.[73] This inferred cluster had immune cell type-specific marker genes such as *CD2* (PIP = 0.92), *GZMB* (PIP = 0.85), *IL7R* (PIP = 0.85), and *NCAM1* (PIP = 1.00)[76,77] (Figure 5D). An additional GO analysis of this cluster showed an enrichment of NK cell-mediated cytotoxicity (adjusted $p = 6.60 \times 10^{-9}$) and T cell receptor signaling (adjusted $p = 1.17 \times 10^{-21}$) (Figure 5G).

Notably, the other clusters that primarily contained non-malignant cells (inferred clusters 4, 5, 11, 12, 13, and 14) also directly aligned with cell type labels originally annotated by Raghavan et al.[73] For the clusters that primarily contained malignant and metastatic cells (i.e., inferred clusters 1, 2, 3, 9, and 10), a GO analysis revealed an enrichment of extracellular matrix (ECM) organization and cell migration processes (Figure 5G). Importantly, however, NCLUSION also had the power to divide these cells into more granular subpopulations based on their level of differentiation. For example, the inferred cluster 9 was enriched for both cell migration processes (e.g., MET-activated PTK2 signaling, adjusted $p = 2.97 \times 10^{-5}$; MET-promoted cell motility, adjusted $p = 4.02 \times 10^{-5}$) and fibroblast cell activity (pancreatic fibroblasts, adjusted $p = 2.1 \times 10^{-9}$; collagen formation, adjusted $p = 1.13 \times 10^{-7}$).[78,79]

Finally, we evaluated the granularity of NCLUSION's clustering on the IMMUNE dataset, which contained 33 manually annotated labels from experts.[75] When analyzing this study, NCLUSION was able to delineate between multiple T cell sub-lineages, whereas methods like Seurat, KNN+Leiden, and scCCESS-SIMLR merged these subpopulations into 1 or 2 clusters. For example, the inferred cluster 12 by NCLUSION was enriched (∼95% occupancy rate) for $CD8^+$ effector memory (T_{EM}) and effector memory cells re-expressing CD45RA (T_{EMRA}), while NCLUSION's inferred cluster 8 was enriched (∼90% occupancy rate) for $CD8^+$ tissue-resident memory (T_{RM}). These two populations have been shown to be functionally distinct subpopulations in the $CD8^+$ T cell lineage.[80,81] Likewise NCLUSION's inferred clusters 6 and 10 were enriched (∼77% and ∼72% occupancy rates, respectively) for functionally distinct subpopulations of $CD4^+$ T cells, namely, T follicular helper cells (T_{FH}) and $CD4^+$ effector/effector memory T cells, respectively[80,82,83] (Data S1).

The GO analysis of NCLUSION-generated gene modules also showed an enrichment of T cell phenotypes. The inferred cluster 12's top ontology term was indeed "CD8+ Effector Memory T4" (adjusted $p = 4.18 \times 10^{-9}$), while the inferred cluster 10 had "CD4+ Central Memory T1" as a top enriched term (adjusted $p = 3.53 \times 10^{-3}$). Similar results showing how NCLUSION-generated gene modules granularly distinguish these cellular populations can be found in the supplemental information (Data S1).

DISCUSSION

We present NCLUSION: a scalable Bayesian nonparametric framework designed to serve as an unbiased method for inferring phenotypic clusters and identifying cluster-specific marker genes in scRNA-seq experiments. We show how our approach simplifies traditional single-cell transcriptomic workflows, which often rely on the transformation of the data to a lower-dimensional representation to facilitate clustering and iteratively tune the number of clusters used to obtain optimal results. In contrast, NCLUSION operates on the full normalized gene expression matrix, eliminating the need for transformation to a lower-dimensional space; infers the optimal number of clusters without iterative user refinement; and simultaneously identifies the cluster-specific marker genes that significantly drive the clustering. By leveraging a variational inference algorithm, NCLUSION can scale to scRNA-seq studies with a million cells. Through the analysis of a collection of large-scale publicly available datasets, we show that NCLUSION not only achieves clustering performance comparable to state-of-the-art methods but also provides refined sets of gene candidates for downstream analyses. By unifying clustering and marker gene selection, NCLUSION provides a flexible and unified statistical framework for inferring complex differential gene expression patterns observed in heterogeneous tissue populations.[23,84,85]

The current implementation of the NCLUSION framework offers many directions for future development and applications. First, NCLUSION assumes a normal mixture model for log-normalized gene expression data. We use this assumption both because log-based transformations have been shown to reduce the effects of sparsity in single-cell analyses[20,86] and because the Gaussian-based specification offers computational advantages for scalable posterior inference. By using a sparse prior on the mean, we are also able to mitigate the overabundance of zeroes and small expression values. Still, future extensions of the NCLUSION framework should explore the utility of Poisson- and negative binomial-based likelihoods to deal with the zero-inflated nature of scRNA-seq studies in their raw form.

Second, the current formulation of NCLUSION models the gene expression of each cell independently and does not consider, for example, the correlation between genes with similar functionality or co-expression patterns between genes within the same signaling pathway. This is, in part, due to scalability considerations when fitting the algorithm. One possible extension of NCLUSION would be to incorporate additional

(B) Assessment of the inferred cluster labels from each method versus cell type annotations from the original studies. Evaluation is quantified by NMI and ARI (for each method in each dataset, we take five random 80% splits; depicted in each bar plot is the mean ± 95% confidence interval). Asterisks indicate that there is a statistically significant difference in performance between NCLUSION and a corresponding method (two-sided t test $p < 0.05$).
(C–F) Results from running NCLUSION on the PDAC dataset. (C) Shown is a t-SNE visualization of the PDAC scRNA-seq dataset, annotated by the cell type labels from the PDAC study (top) compared to the clusters inferred by NCLUSION (bottom), where the "NM" labels indicate non-malignant cells and the "M" labels indicate malignant cells. (D) Heatmaps of the adjusted PIPs (left), ESS (center), and SSMD (right) of the significant genes in each cluster. (E) Highlighted location on t-SNEs of NCLUSION-inferred clusters that contain predominantly one cell type. (F) Violin plots comparing the normalized expression of cluster-specific marker genes across clusters.
(G) GO pathway enrichment analysis for the marker genes identified for each cluster. Gene sets with a q value below 0.05 are deemed to be significant.
See also Figures S7–S24 and Data S1.

genomic information into the sparse prior distributions used for Bayesian variable selection, thereby inducing more sophisticated co-expression relationships. For example, previous studies have proposed an integrative approach where the importance of a variable also depends on an additional set of covariates.[42,87,88] In the case of single-cell applications, we could assume that the prior probability of the *j*-th gene being a marker of the *k*-th cluster is also dependent upon its cellular pathway membership. Unlike the current spike-and-slab prior NCLUSION implements, this prior would assume that biologically related pathways contain shared marker genes, essentially integrating the concept of gene set enrichment analysis into clustering. An alternative approach would be to extend NCLUSION to incorporate non-diagonal correlation structures by exploring sparse covariance models, which could provide a balance between the need for maintaining computational efficiency while representing a richer set of gene dependencies.[89,90]

Third, a thrust of recent work in genomics has been to develop methods that identify spatially variable marker genes as a key step during analyses of spatially-resolved transcriptomics data.[91] Future efforts could extend NCLUSION to this emerging modality by, for example, reformulating the method as a spatial Dirichlet process mixture model.[92]

In sum, NCLUSION provides a unified framework for simultaneous clustering and marker gene selection in single-cell transcriptomic data, yielding improvements in computational efficiency, interpretability, and scalability. We envision that NCLUSION will accelerate key analytic steps universal to single-cell analysis across diverse applications.

Limitations of the study

Although it helps NCLUSION scale to large datasets, variational EM algorithms are known to both produce slightly miscalibrated parameter estimates and underestimate the total variation present within a dataset.[43,44,93] While this does not greatly affect the performance of NCLUSION in the evaluations presented in this paper, this can be seen as a limitation depending on the application of interest. For example, in the PBMC dataset, NCLUSION is unable to resolve all the different T cell subtypes that were annotated by Zheng et al.[54] (Figure 3). This is most likely due to variational approximations being well-suited to describe the global variation across cells but at the cost of smoothing over local variation between smaller subpopulations. Considering other (equally scalable) ways to carry out approximate Bayesian inference may be relevant for future work.[94]

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Lorin Crawford (lcrawford@microsoft.com).

Materials availability

This study did not generate new materials.

Data and code availability

- All of the datasets analyzed in this paper are publicly available. The PDAC dataset from Raghavan et al.[73] can be accessed at https://singlecell.broadinstitute.org/single_cell/study/SCP1644/microenvironment-drives-cell-state-plasticity-and-drug-response-in-pancreatic-cancer#/. The AML data from van Galen et al.[74] has been deposited in the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE116256. The BRAIN-LARGE dataset can be accessed at https://www.10xgenomics.com/datasets/1-3-million-brain-cells-from-e-18-mice-2-standard-1-3-0. The individual PBMC data from Zheng et al.[54] can be downloaded directly from https://www.10xgenomics.com/resources/datasets. Last, the immune cell atlas dataset can be accessed at https://cellgeni.cog.sanger.ac.uk/pan-immune/CountAdded_PIP_global_object_for_cellxgene.h5ad.
- An open-source software implementation of NCLUSION is available on GitHub at https://github.com/microsoft/Nclusion.jl and archived at https://doi.org/10.5281/zenodo.18049775.[95] Guided tutorials and all code needed to reproduce the results and figures in this work can be found at https://microsoft.github.io/Nclusion.jl/.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank members of the Crawford, Raghavan, and Shalek Labs for insightful comments on earlier versions of this manuscript. This research was conducted by using a combination of computational resources and services provided by Microsoft Research and the Center for Computation and Visualization at Brown University. This research was also supported in part by an Alfred P. Sloan Research Fellowship and a David & Lucile Packard Fellowship for Science and Engineering awarded to L.C. C.N. was a trainee supported under the Brown University Predoctoral Training Program in Biological Data Science (NIH T32GM128596). S.R. is supported by NCI K08 award 1K08CA260442, the Claudia Adams Barr Program in Innovative Basic Cancer Research, and the Dana-Farber Cancer Institute Hale Family Center for Pancreatic Cancer Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

AUTHOR CONTRIBUTIONS

C.N., A.P.A., and L.C. conceived the study and developed the methods. C.N. and M.H. developed the algorithm and software. C.N., M.H., and M.L.R. led the analyses. A.W.N., A.K.S., N.F., S.R., P.S.W., A.P.A., and L.C. provided secondary analyses. A.K.S., N.F., S.R., P.S.W., A.P.A., and L.C. provided resources. S.R., P.S.W., A.P.A., and L.C. supervised the project. All authors interpreted the results and wrote and revised the manuscript.

DECLARATION OF INTERESTS

S.R. holds equity in Amgen. P.S.W. reports compensation for consulting/speaking from Engine Ventures and AbbVie unrelated to this work. A.K.S. reports compensation for consulting and/or scientific advisory board membership from Honeycomb Biotechnologies, Cellarity, Ochre Bio, Relation Therapeutics, Fog Pharma, Bio-Rad Laboratories, IntrECate Biotherapeutics, Passkey Therapeutics, and Dahlia Biosciences unrelated to this work. S.R. and P.S.W. receive research funding from Microsoft. M.H., N.F., A.P.A., and L.C. are employees of Microsoft and own equity in Microsoft.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Overview of NCLUSION
  - Selection of cluster-specific marker genes
  - Posterior inference via variational EM algorithm
  - Simulation study design
  - Real datasets and preprocessing

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2026.101329.

Received: April 29, 2024
Revised: September 19, 2025
Accepted: January 22, 2026
Published: March 12, 2026

REFERENCES

1. Miao, Z., Humphreys, B.D., McMahon, A.P., and Kim, J. (2021). Multi-omics integration in the age of million single-cell data. Nat. Rev. Nephrol. *17*, 710–724. https://doi.org/10.1038/s41581-021-00463-x.
2. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). Scanpy: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.
3. Guo, M., Wang, H., Potter, S.S., Whitsett, J.A., and Xu, Y. (2015). Sincera: A pipeline for single-cell rna-seq profiling analysis. PLoS Comput. Biol. *11*, e1004575. https://doi.org/10.1371/journal.pcbi.1004575.
4. Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. F1000Res. *5*, 2122. https://doi.org/10.12688/f1000research.9501.2.
5. Haque, A., Engel, J., Teichmann, S.A., and Lönnberg, T. (2017). A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. Genome Med. *9*, 75. https://doi.org/10.1186/s13073-017-0467-4.
6. Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. *16*, 133–145. https://doi.org/10.1038/nrg3833.
7. Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell rna-seq data. Nat. Rev. Genet. *20*, 273–282. https://doi.org/10.1038/s41576-018-0088-9.
8. Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2020). Orchestrating single-cell analysis with bioconductor. Nat. Methods *17*, 137–145. https://doi.org/10.1038/s41592-019-0654-x.
9. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. https://doi.org/10.1038/nbt.3192.
10. Cheng, C., Easton, J., Rosencrance, C., Li, Y., Ju, B., Williams, J., Mulder, H.L., Pang, Y., Chen, W., and Chen, X. (2019). Latent cellular analysis robustly reveals subtle diversity in large-scale single-cell rna-seq data. Nucleic Acids Res. *47*, e143. https://doi.org/10.1093/nar/gkz826.
11. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. Nat. Methods *14*, 414–416. https://doi.org/10.1038/nmeth.4207.
12. Maaten, L.v. d., and Hinton, G. (2008). Visualizing data using t-sne. J. Mach. Learn. Res. *9*, 2579–2605.
13. McInnes, L., Healy, J., and Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. http://arxiv.org/abs/1802.03426.
14. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.
15. Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., and Hemberg, M. (2017). Sc3: consensus clustering of single-cell rna-seq data. Nat. Methods *14*, 483–486. https://doi.org/10.1038/nmeth.4236.
16. Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger rna sequencing reveals rare intestinal cell types. Nature *525*, 251–255. https://doi.org/10.1038/nature14966.
17. Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. *19*, 335–346. https://doi.org/10.1038/nn.4216.
18. žurauskienė, J., and Yau, C. (2016). pcareduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinf. *17*, 140. https://doi.org/10.1186/s12859-016-0984-y.
19. Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. Science *347*, 1138–1142. https://doi.org/10.1126/science.aaa1934.
20. Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746. https://doi.org/10.15252/msb.20188746.
21. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome Biol. *16*, 278. https://doi.org/10.1186/s13059-015-0844-5.
22. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for rna-seq read counts. Genome Biol. *15*, R29. https://doi.org/10.1186/gb-2014-15-2-r29.
23. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. *21*, 31. https://doi.org/10.1186/s13059-020-1926-6.
24. Gao, L.L., Bien, J., and Witten, D. (2024). Selective inference for hierarchical clustering. J. Am. Stat. Assoc. *119*, 332–342. https://doi.org/10.1080/01621459.2022.2116331.
25. Neufeld, A., Gao, L.L., Popp, J., Battle, A., and Witten, D. (2023). Inference after latent variable estimation for single-cell rna sequencing data. Biostatistics *25*, 270–287. https://doi.org/10.1093/biostatistics/kxac047.
26. Grabski, I.N., Street, K., and Irizarry, R.A. (2023). Significance analysis for clustering with single-cell rna-sequencing data. Nat. Methods *20*, 1196–1202. https://doi.org/10.1038/s41592-023-01933-9.
27. Vandenbon, A., and Diez, D. (2020). A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. Nat. Commun. *11*, 4318. https://doi.org/10.1038/s41467-020-17900-3.
28. Zhang, J.M., Kamath, G.M., and Tse, D.N. (2019). Valid post-clustering differential analysis for single-cell rna-seq. Cell Syst. *9*, 383–392.e6. https://doi.org/10.1016/j.cels.2019.07.012.
29. DenAdel, A., Ramseier, M.L., Navia, A.W., Shalek, A.K., Raghavan, S., Winter, P.S., Amini, A.P., and Crawford, L. (2025). Artificial variables help to avoid over-clustering in single-cell rna sequencing. Am. J. Hum. Genet. *112*, 940–951. https://doi.org/10.1016/j.ajhg.2025.02.014.
30. Zhang, J.M., Fan, J., Fan, H.C., Rosenfeld, D., and Tse, D.N. (2018). An interpretable framework for clustering single-cell rna-seq datasets. BMC Bioinf. *19*, 93. https://doi.org/10.1186/s12859-018-2092-7.
31. Lall, S., Ray, S., and Bandyopadhyay, S. (2021). Rgcop-a regularized copula based method for gene selection in single-cell rna-seq data. PLoS Comput. Biol. *17*, e1009464. https://doi.org/10.1371/journal.pcbi.1009464.
32. Townes, F.W., Hicks, S.C., Aryee, M.J., and Irizarry, R.A. (2019). Feature selection and dimension reduction for single-cell rna-seq based on a

multinomial model. Genome Biol. *20*, 295. https://doi.org/10.1186/s13059-019-1861-6.

33. Blei, D.M., and Jordan, M.I. (2006). Variational inference for dirichlet process mixtures. Bayesian Anal. *1*, 121–143. https://doi.org/10.1214/06-BA104.

34. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). Bayesian Data Analysis, Third Edition (CRC Press).

35. Prabhakaran, S., Azizi, E., Carr, A., and Pe'er, D. (2016). Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data. JMLR Workshop Conf. Proc. *48*, 1070–1079.

36. Sun, Z., Chen, L., Xin, H., Jiang, Y., Huang, Q., Cillo, A.R., Tabib, T., Kolls, J.K., Bruno, T.C., Lafyatis, R., et al. (2019). A bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. Nat. Commun. *10*, 1649. https://doi.org/10.1038/s41467-019-09639-3.

37. Duan, T., Pinto, J.P., and Xie, X. (2019). Parallel clustering of single cell transcriptomic data with split-merge sampling on dirichlet process mixtures. Bioinformatics *35*, 953–961. https://doi.org/10.1093/bioinformatics/bty702.

38. Sun, Z., Wang, T., Deng, K., Wang, X.-F., Lafyatis, R., Ding, Y., Hu, M., and Chen, W. (2018). Dimm-sc: a dirichlet mixture model for clustering droplet-based single cell transcriptomic data. Bioinformatics *34*, 139–146. https://doi.org/10.1093/bioinformatics/btx490.

39. Adossa, N.A., Rytkönen, K.T., and Elo, L.L. (2022). Dirichlet process mixture models for single-cell rna-seq clustering. Biol. Open *11*, bio059001. https://doi.org/10.1242/bio.059001.

40. Barbieri, M.M., and Berger, J.O. (2004). Optimal predictive model selection. Ann. Statist. *32*, 870–897. https://doi.org/10.1214/009053604000000238.

41. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. Nat. Commun. *8*, 456. https://doi.org/10.1038/s41467-017-00470-2.

42. Carbonetto, P., and Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. *7*, 73–108. https://doi.org/10.1214/12-BA703.

43. Demetci, P., Cheng, W., Darnell, G., Zhou, X., Ramachandran, S., and Crawford, L. (2021). Multi-scale inference of genetic trait architecture using biologically annotated neural networks. PLoS Genet. *17*, e1009754. https://doi.org/10.1371/journal.pgen.1009754.

44. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. J. Am. Stat. Assoc. *112*, 859–877. https://doi.org/10.1080/01621459.2017.1285773.

45. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

46. Zhu, L., Lei, J., Klei, L., Devlin, B., and Roeder, K. (2019). Semisoft clustering of single-cell data. Proc. Natl. Acad. Sci. USA *116*, 466–471. https://doi.org/10.1073/pnas.1817715116.

47. 10× Genomics. Support: single cell gene expression datasets (2023). https://www.10xgenomics.com/resources/datasets.

48. Svensson, V., da Veiga Beltrame, E., and Pachter, L. (2020). A curated database reveals trends in single-cell transcriptomics. Database *2020*, baaa073. https://doi.org/10.1093/database/baaa073.

49. Song, D., Wang, Q., Yan, G., Liu, T., Sun, T., and Li, J.J. (2024). scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. Nat. Biotechnol. *42*, 247–252. https://doi.org/10.1038/s41587-023-01772-1.

50. Lin, P., Troup, M., and Ho, J.W.K. (2017). Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. Genome Biol. *18*, 59. https://doi.org/10.1186/s13059-017-1188-0.

51. Tian, T., Wan, J., Song, Q., and Wei, Z. (2019). Clustering single-cell rna-seq data with a model-based deep learning approach. Nat. Mach. Intell. *1*, 191–198. https://doi.org/10.1038/s42256-019-0037-0.

52. Ranjan, B., Sun, W., Park, J., Mishra, K., Schmidt, F., Xie, R., Alipour, F., Singhal, V., Joanito, I., Honardoost, M.A., et al. (2021). Dubstepr is a scalable correlation-based feature selection method for accurately clustering single-cell data. Nat. Commun. *12*, 5849. https://doi.org/10.1038/s41467-021-26085-2.

53. Chen, Z., Wang, C., Huang, S., Shi, Y., and Xi, R. (2024). Directly selecting cell-type marker genes for single-cell clustering analyses. Cell Rep. Methods *4*, 100810. https://doi.org/10.1016/j.crmeth.2024.100810.

54. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

55. Vargo, A.H.S., and Gilbert, A.C. (2020). A rank-based marker selection method for high throughput scrna-seq data. BMC Bioinf. *21*, 477. https://doi.org/10.1186/s12859-020-03641-z.

56. Michielsen, L., Reinders, M.J.T., and Mahfouz, A. (2021). Hierarchical progressive learning of cell identities in single-cell data. Nat. Commun. *12*, 2799. https://doi.org/10.1038/s41467-021-23196-8.

57. Baume, D.M., Caligiuri, M.A., Manley, T.J., Daley, J.F., and Ritz, J. (1990). Differential expression of cd8α and cd8β associated with mhc-restricted and non-mhc-restricted cytolytic effector cells. Cell. Immunol. *131*, 352–365. https://doi.org/10.1016/0008-8749(90)90260-X.

58. Hendriks, J., Gravestein, L.A., Tesselaar, K., van Lier, R.A., Schumacher, T.N., and Borst, J. (2000). Cd27 is required for generation and long-term maintenance of t cell immunity. Nat. Immunol. *1*, 433–440. https://doi.org/10.1038/80877.

59. Norris, S., Doherty, D.G., Collins, C., McEntee, G., Traynor, O., Hegarty, J.E., and O'Farrelly, C. (1999). Natural t cells in the human liver: cytotoxic lymphocytes with dual t cell and natural killer cell phenotype and function are phenotypically heterogenous and include vα24-jαq and $\gamma\delta$ t cell receptor bearing cells. Hum. Immunol. *60*, 20–31. https://doi.org/10.1016/S0198-8859(98)00098-6.

60. Willis, S.N., Tellier, J., Liao, Y., Trezise, S., Light, A., O'Donnell, K., Garrett-Sinha, L.A., Shi, W., Tarlinton, D.M., and Nutt, S.L. (2017). Environmental sensing by mature b cells is controlled by the transcription factors pu.1 and spib. Nat. Commun. *8*, 1426. https://doi.org/10.1038/s41467-017-01605-1.

61. Zuccolo, J., Deng, L., Unruh, T.L., Sanyal, R., Bau, J.A., Storek, J., Demetrick, D.J., Luider, J.M., Auer-Grzesiak, I.A., Mansoor, A., and Deans, J.P. (2013). Expression of ms4a and tmem176 genes in human b lymphocytes. Front. Immunol. *4*, 195.

62. Chen, Z., Yu, M., Yan, J., Guo, L., Zhang, B., Liu, S., Lei, J., Zhang, W., Zhou, B., Gao, J., et al. (2021). Pnoc expressed by b cells in cholangiocarcinoma was survival related and lair2 could be a t cell exhaustion biomarker in tumor microenvironment: Characterization of immune microenvironment combining single-cell and bulk sequencing technology. Front. Immunol. *12*, 647209. https://doi.org/10.3389/fimmu.2021.647209.

63. Evren, E., Ringqvist, E., Tripathi, K.P., Sleiers, N., Rives, I.C., Alisjahbana, A., Gao, Y., Sarhan, D., Halle, T., Sorini, C., et al. (2021). Distinct developmental pathways from blood monocytes generate human lung macrophage diversity. Immunity *54*, 259–275.e7. https://doi.org/10.1016/j.immuni.2020.12.003.

64. Fang, Z., Liu, X., and Peltz, G. (2023). Gseapy: a comprehensive package for performing gene set enrichment analysis in python. Bioinformatics *39*, btac757. https://doi.org/10.1093/bioinformatics/btac757.

65. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440. https://doi.org/10.1093/bioinformatics/bti525.

66. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. BMC Bioinf. *14*, 128. https://doi.org/10.1186/1471-2105-14-128.

67. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

68. Carbon, S., and Mungall, C. Gene ontology data archive (2023). Zenodo. https://zenodo.org/record/8200914.

69. Vivier, E., Tomasello, E., Baratin, M., Walzer, T., and Ugolini, S. (2008). Functions of natural killer cells. Nat. Immunol. *9*, 503–510. https://doi.org/10.1038/ni1582.

70. Paul, S., and Lal, G. (2017). The molecular mechanism of natural killer cells function and its importance in cancer immunotherapy. Front. Immunol. *8*, 1124. https://doi.org/10.3389/fimmu.2017.01124.

71. Jaeger, B.N., Donadieu, J., Cognet, C., Bernat, C., Ordoñez-Rueda, D., Barlogis, V., Mahlaoui, N., Fenis, A., Narni-Mancinelli, E., Beaupain, B., et al. (2012). Neutrophil depletion impairs natural killer cell maturation, function, and homeostasis. J. Exp. Med. *209*, 565–580. https://doi.org/10.1084/jem.20111908.

72. Jiang, K., Zhong, B., Gilvary, D.L., Corliss, B.C., Hong-Geller, E., Wei, S., and Djeu, J.Y. (2000). Pivotal role of phosphoinositide-3 kinase in regulation of cytotoxicity in natural killer cells. Nat. Immunol. *1*, 419–425. https://doi.org/10.1038/80859.

73. Raghavan, S., Winter, P.S., Navia, A.W., Williams, H.L., DenAdel, A., Lowder, K.E., Galvez-Reyes, J., Kalekar, R.L., Mulugeta, N., Kapner, K.S., et al. (2021). Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. Cell *184*, 6119–6137.e26. https://doi.org/10.1016/j.cell.2021.11.017.

74. van Galen, P., Hovestadt, V., Wadsworth II, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi Story, J., et al. (2019). Single-cell rna-seq reveals aml hierarchies relevant to disease progression and immunity. Cell *176*, 1265–1281.e24. https://doi.org/10.1016/j.cell.2019.01.031.

75. Domínguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T., Howlett, S.K., Suchanek, O., Polanski, K., King, H.W., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science *376*, eabl5197. https://doi.org/10.1126/science.abl5197.

76. Almehmadi, M., Flanagan, B.F., Khan, N., Alomar, S., and Christmas, S.E. (2014). Increased numbers and functional activity of cd56+ t cells in healthy cytomegalovirus positive subjects. Immunology *142*, 258–268. https://doi.org/10.1111/imm.12250.

77. Kansler, E.R., and Li, M.O. (2019). Innate lymphocytes—lineage, localization and timing of differentiation. Cell. Mol. Immunol. *16*, 627–633. https://doi.org/10.1038/s41423-019-0211-7.

78. Beviglia, L., and Kramer, R.H. (1999). Hgf induces fak activation and integrin-mediated adhesion in mtln3 breast carcinoma cells. Int. J. Cancer *83*, 640–649. https://doi.org/10.1002/(sici)1097-0215(19991126)83:5<640::aid-ijc13>3.0.co;2-d.

79. Silzle, T., Randolph, G.J., Kreutz, M., and Kunz-Schughart, L.A. (2004). The fibroblast: Sentinel cell and local immune modulator in tumor tissue. Int. J. Cancer *108*, 173–180. https://doi.org/10.1002/ijc.11542.

80. Meraviglia, S., Carlo, P.D., Pampinella, D., Guadagnino, G., Presti, E.L., Orlando, V., Marchetti, G., Dieli, F., and Sergi, C. (2019). T-cell subsets (tcm, tem, temra) and poly-functional immune response in patients with human immunodeficiency virus (hiv) infection and different t-cd4 cell response. Ann. Clin. Lab. Sci. *49*, 519–528.

81. Schenkel, J.M., and Masopust, D. (2014). Tissue-resident memory t cells. Immunity *41*, 886–897. https://doi.org/10.1016/j.immuni.2014.12.007.

82. Opata, M.M., Ibitokou, S.A., Carpio, V.H., Marshall, K.M., Dillon, B.E., Carl, J.C., Wilson, K.D., Arcari, C.M., and Stephens, R. (2018). Protection by and maintenance of cd4 effector memory and effector t cell subsets in persistent malaria infection. PLoS Pathog. *14*, e1006960. https://doi.org/10.1371/journal.ppat.1006960.

83. Crotty, S. (2014). T follicular helper cell differentiation, function, and roles in disease. Immunity *41*, 529–542. https://doi.org/10.1016/j.immuni.2014.10.004.

84. Vavoulis, D.V., Francescatto, M., Heutink, P., and Gough, J. (2015). Dgeclust: differential expression analysis of clustered count data. Genome Biol. *16*, 39. https://doi.org/10.1186/s13059-015-0604-6.

85. Soneson, C., and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods *15*, 255–261. https://doi.org/10.1038/nmeth.4612.

86. Ahlmann-Eltze, C., and Huber, W. (2023). Comparison of transformations for single-cell rna-seq data. Nat. Methods *20*, 665–672. https://doi.org/10.1038/s41592-023-01814-1.

87. Zhu, X., and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat. Commun. *9*, 4361.

88. Ma, Y., Sun, S., Shang, X., Keller, E.T., Chen, M., and Zhou, X. (2020). Integrative differential expression and gene set enrichment analysis using summary statistics for scrna-seq studies. Nat. Commun. *11*, 1585. https://doi.org/10.1038/s41467-020-15298-6.

89. Bien, J., and Tibshirani, R.J. (2011). Sparse estimation of a covariance matrix. Biometrika *98*, 807–820. https://doi.org/10.1093/biomet/asr054.

90. Aboutaleb, Y.M., Danaf, M., Xie, Y., and Ben-Akiva, M.E. (2021). Sparse covariance estimation in logit mixture models. Econom. J. *24*, 377–398. https://doi.org/10.1093/ectj/utab008.

91. Weber, L.M., Saha, A., Datta, A., Hansen, K.D., and Hicks, S.C. (2023). nnsvg for the scalable identification of spatially variable genes using nearest-neighbor gaussian processes. Nat. Commun. *14*, 4059. https://doi.org/10.1038/s41467-023-39748-z.

92. Reich, B.J., and Bondell, H.D. (2011). A spatial dirichlet process mixture model for clustering population genetics data. Biometrics *67*, 381–390. https://doi.org/10.1111/j.1541-0420.2010.01484.x.

93. Giordano, R., Broderick, T., and Jordan, M.I. (2018). Covariances, robustness and variational bayes. J. Mach. Learn. Res. *19*, 1–49.

94. Zhang, C., Shahbaba, B., and Zhao, H. (2018). Variational hamiltonian monte carlo via score matching. Bayesian Anal. *13*, 485–506. https://doi.org/10.1214/17-ba1060.

95. Nwizu, C., Hughes, M., and Crawford, L. (2025). microsoft/nclusion.jl. Zenodo. https://doi.org/10.5281/zenodo.18049774.

96. Yu, L., Cao, Y., Yang, J.Y.H., and Yang, P. (2022). Benchmarking clustering algorithms on estimating the number of cell types from single-cell rna-sequencing data. Genome Biol. *23*, 49. https://doi.org/10.1186/s13059-022-02622-0.

97. Guan, Y., and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat. *5*, 1780–1815. https://doi.org/10.1214/11-AOAS455.

98. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264. https://doi.org/10.1371/journal.pgen.1003264.

99. Zhu, X., and Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Ann. Appl. Stat. *11*, 1561–1592. https://doi.org/10.1214/17-AOAS1046.

100. Hughes, M.C., Kim, D.I., and Sudderth, E.B. (2015). Reliable and scalable variational inference for the hierarchical dirichlet process. Artificial Intelligence and Statistics *38*, 370–378.

101. Cohen, J. (2013). Statistical Power Analysis for the Behavioral Sciences (Academic Press).

102. Zhang, X.D. (2007). A pair of new statistical parameters for quality control in rna interference high-throughput screening assays. Genomics *89*, 552–561. https://doi.org/10.1016/j.ygeno.2006.12.014.

103. Zhang, X.D. (2007). A new method with flexible and balanced control of false negatives and false positives for hit selection in rna interference high-throughput screening assays. SLAS Discovery *12*, 645–655. https://doi.org/10.1177/1087057107300645.

104. Zhang, X.D. (2010). Strictly standardized mean difference, standardized mean difference and classical t-test for the comparison of two groups. Stat. Biopharm. Res. *2*, 292–299. https://doi.org/10.1198/sbr.2009.0074.

105. Zhang, X.D. (2011). Optimal High-Throughput Screening: Practical Experimental Design and Data Analysis for Genome-Scale RNAi Research, 1 ed. (Cambridge University Press)978-0-521-51771-3. URL: https://www.cambridge.org/core/product/identifier/9780511973888/type/book

106. Zhang, X.D., Marine, S.D., and Ferrer, M. (2009). Error rates and powers in genome-scale rnai screens. SLAS Discovery *14*, 230–238. https://doi.org/10.1177/1087057109331475.

107. Zhang, X.D., Lacson, R., Yang, R., Marine, S.D., McCampbell, A., Toolan, D.M., Hare, T.R., Kajdas, J., Berger, J.P., Holder, D.J., et al. (2010). The use of ssmd-based false discovery and false nondiscovery rates in genome-scale rnai screens. SLAS Discovery *15*, 1123–1131. https://doi.org/10.1177/1087057110381919.

108. Wand, M.P., Ormerod, J.T., Padoan, S.A., and Frühwirth, R. (2011). Mean field variational bayes for elaborate distributions. Bayesian Anal. *6*, 847–900. https://doi.org/10.1214/11-BA631.

109. Svensson, V. (2020). Droplet scrna-seq is not zero-inflated. Nat. Biotechnol. *38*, 147–150. https://doi.org/10.1038/s41587-019-0379-5.

110. Svensson, V. (2021). Reply to: Umi or not umi, that is the question for scrna-seq zero-inflation. Nat. Biotechnol. *39*, 160. https://doi.org/10.1038/s41587-020-00811-5.

111. Vinh, N.X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In Proceedings of the 26th Annual International Conference on Machine Learning (ICML), pp. 1073–1080.

112. Wagner, S., and Wagner, D. (2007). Comparing Clusterings - an Overview (Karlsruhe).

113. Gene Ontology Consortium; Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., et al. (2023). The gene ontology knowledgebase in 2023. Genetics *224*, iyad031. https://doi.org/10.1093/genetics/iyad031.

114. Tabula Sapiens Consortium*; Jones, R.C., Karkanias, J., Krasnow, M.A., Pisco, A.O., Quake, S.R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., et al. (2022). The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science *376*, eabl4896. https://doi.org/10.1126/science.abl4896.

115. Tabula Sapiens Consortium; and Quake, S.R. (2025). Tabula sapiens reveals transcription factor expression, senescence effects, and sex-specific features in cell types from 28 human organs and tissues. Preprint at bioRxiv. https://www.biorxiv.org/content/10.1101/2024.12.03.626516v1.

116. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

117. Kanehisa, M., and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

118. Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Sci. *28*, 1947–1951. https://doi.org/10.1002/pro.3715.

119. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y., and Ishiguro-Watanabe, M. (2025). Kegg: biological systems database as a model of the real world. Nucleic Acids Res. *53*, D672–D677. https://doi.org/10.1093/nar/gkae909.

120. Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marin-Garcia, P., Ping, P., Wu, G., Stein, L., D'Eustachio, P., and Hermjakob, H. (2018). Reactome graph database: Efficient access to complex pathway data. PLoS Comput. Biol. *14*, e1005968. https://doi.org/10.1371/journal.pcbi.1005968.

121. Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., D'Eustachio, P., Stein, L., and Hermjakob, H. (2017). Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinf. *18*, 142. https://doi.org/10.1186/s12859-017-1559-2.

122. Fabregat, A., Sidiropoulos, K., Viteri, G., Marin-Garcia, P., Ping, P., Stein, L., D'Eustachio, P., and Hermjakob, H. (2018). Reactome diagram viewer: data structures and strategies to boost performance. Bioinformatics *34*, 1208–1214. https://doi.org/10.1093/bioinformatics/btx752.

123. Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A., and Hermjakob, H. (2020). Reactomegsa - efficient multi-omics comparative pathway analysis. Mol. Cell. Proteomics *19*, 2115–2125. https://doi.org/10.1074/mcp.TIR120.002155.

124. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res. *48*, D498–D503. https://doi.org/10.1093/nar/gkz1031.

125. Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., et al. (2024). The reactome pathway knowledgebase 2024. Nucleic Acids Res. *52*, D672–D678. https://doi.org/10.1093/nar/gkad1025.

126. Sidiropoulos, K., Viteri, G., Sevilla, C., Jupe, S., Webber, M., Orlic-Milacic, M., Jassal, B., May, B., Shamovsky, V., Duenas, C., et al. (2017). Reactome enhanced pathway visualization. Bioinformatics *33*, 3461–3467. https://doi.org/10.1093/bioinformatics/btx441.

127. Wu, G., and Haw, R. (2017). Functional interaction network construction and analysis for disease discovery. Methods Mol. Biol. *1558*, 235–253. https://doi.org/10.1007/978-1-4939-6783-4_11.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
PDAC Dataset	Raghavan et al.[73]	https://singlecell.broadinstitute.org/single_cell/study/SCP1644/microenvironment-drives-cell-state-plasticity-and-drug-response-in-pancreatic-cancer#/
AML Dataset	van Galen et al.[74]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116256
BRAIN-LARGE Dataset	10× Genomics[47]	https://www.10xgenomics.com/datasets/1-3-million-brain-cells-from-e-18-mice-2-standard-1-3-0
PBMC Dataset	Zheng et al.[54]	https://www.10xgenomics.com/resources/datasets
Immune Cell Atlas Dataset	Domínguez Conde et al.[75]	https://cellgeni.cog.sanger.ac.uk/pan-immune/CountAdded_PIP_global_object_for_cellxgene.h5ad
Software and algorithms		
NCLUSION	This study	https://github.com/microsoft/Nclusion.jl https://doi.org/10.5281/zenodo.18049775
CIDR	Lin et al.[50]	https://github.com/VCCRI/CIDR
SC3	Kiselev et al.[15]	http://bioconductor.org/packages/SC3/
scCCESS+SIMLR	Yu et al.[96] Wang et al.[11]	https://github.com/PYangLab/scCCESS https://github.com/BatzoglouLabSU/SIMLR
scLCA	Cheng et al.[10]	https://bitbucket.org/scLCA/single_cell_lca
KNN+Leiden	Pedregosa et al.[45] Wolf et al.[2]	https://scikit-learn.org/stable/ https://scanpy.readthedocs.io/en/stable/
Seurat	Satija et al.[9]	https://satijalab.org/seurat/
SOUP	Zhu et al.[46]	https://github.com/lingxuez/SOUPR
scDeepCluster	Tian et al.[51]	https://github.com/ttgump/scDeepCluster
DUBStepR	Ranjan et al.[52]	https://github.com/prabhakarlab/DUBStepR
FESTEM	Chen et al.[53]	https://github.com/Edward-Z-Chen/Festem
singleCellHaystack	Vandenbon and Diez[27]	https://github.com/alexisvdb/singleCellHaystack

METHOD DETAILS

Full derivation of the NCLUSION model, the variational expectation maximization (EM) algorithm, corresponding pseudocode, details on comparisons between clustering algorithms, and description of the computing resources used for this study can be found in Methods S1.

Overview of NCLUSION

We provide a brief overview of the probabilistic framework underlying the "Nonparametric CLUstering of SIngle-cell populatiONs" (NCLUSION) model. Note that we use "nonparametric" when describing NCLUSION to match the nomenclature of previous work.[33–39] Consider a study with single-cell RNA sequencing (scRNA-seq) expression data for $n = 1, \ldots, N$ cells that each have measurements for $j = 1, \ldots, J$ genes. Let this dataset be represented by the $N \times J$ matrix $\mathbf{X}$ where the row-vector $\mathbf{x}_n = (x_{n1}, \ldots, x_{nJ})$ denotes the expression profile for the n-th cell. We assume that the log-normalized gene expression for each cell follows a sparse hierarchical Dirichlet process normal mixture model[33–35] of the form

$$x_{nj} \sim \sum_{k=1}^{\infty} \pi_k \, \mathcal{N}\left(\nu_j + \mu_{jk}, \sigma_j^2\right) \qquad \text{(Equation 1)}$$

where π_k represents the marginal (unconditional) probability that a cell belongs to the k-th cluster, ν_j and σ_j^2 are the global means and variances for the j-th gene across all cells (i.e., not conditioned on cluster identity), and μ_{jk} is the mean shift of expression for the j-th gene within the k-th cluster. There are two key features in the model formulation of NCLUSION specified above. First, we

assume that the formation of clusters is driven by a few important genes that have mean expression shifted away from a baseline gene-specific expression level, ν_j. To that end, we place a sparsity-inducing spike and slab prior distribution on the mean effect of each gene

$$\mu_{jk} \sim \eta\, \mathcal{N}\left(0, \lambda_{jk}\sigma_j^2\right) + (1 - \eta)\delta_0, \quad \text{(Equation 2)}$$

where δ_0 is a point mass at zero, λ_{jk} scales the global variance to form a cluster-specific "slab" distribution for each gene, and η is the prior probability that any given gene has a nonzero effect when assigning a cell to any cluster. In practice, there are many different ways to estimate η. Following previous work,[42,43,97–99] one choice would be to assume a uniform prior over $log\eta$ to reflect our lack of knowledge about the correct number of "marker" genes for each cell type that is present in the data. Instead, in this work, we assume $\eta \sim$Beta(1,1) to represent this uncertainty and learn its value during model inference. To facilitate posterior computation and interpretable inference, we introduce a binary indicator variable $\rho_{jk} \in \{0,1\}$ where we implicitly assume *a priori* that $Pr[\rho_{jk} = 1] = \eta$. Alternatively, we say that ρ_{jk} takes on a value of 1 when the effect of a gene μ_{jk} on cluster assignment is nonzero and deviates from the baseline gene expression level ν_j. As NCLUSION is trained, the posterior mean for unimportant genes will trend toward the global mean (i.e., $\mu_{jk} \to 0$) as the model attempts to identify subsets of marker genes that are relevant for each cluster. We then use posterior inclusion probabilities (PIPs) as general summaries of evidence that the j-th gene is statistically important in determining when a cell is assigned to the k-th cluster where

$$\mathrm{PIP}(j;k) \equiv \Pr\left[\mu_{jk} \neq 0 \mid \mathbf{X}\right]. \quad \text{(Equation 3)}$$

The second key feature in Equation 1 is that we do not assume to know the true number of clusters K. Instead, we take a nonparametric approach and attempt to learn K directly from the data. Once again, to facilitate posterior computation, we introduce a categorical latent variable ψ_n which indicates that the n-th cell is in the k-th cluster with prior probability π_k. Explicitly, we write this as $Pr[\psi_n = k] = \pi_k$. Here, we implement the stick-breaking construction of the Dirichlet process[33] where we say

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha_0\boldsymbol{\beta}), \beta_k \sim \chi_k \prod_{l=1}^{k-1}(1 - \chi_l), \chi_k \sim \mathrm{Beta}(1, \gamma_0) \quad \text{(Equation 4)}$$

with $\boldsymbol{\pi}=(\pi_1, \ldots, \pi_{>K})$ having mean $\boldsymbol{\beta}$ and variance determined by the concentration hyper-parameters α_0 and γ_0.[100] The concentration hyper-parameters α_0 and γ_0 are both non-negative scalars that effectively help to determine the number of clusters used in the model.[33,41] Larger values for these parameters increase the model's sensitivity to variation in the data and encourage the creation of a greater number of smaller clusters. Smaller values for these parameters, on the other hand, decrease the model's sensitivity to variation in the data and encourage the creation of fewer larger clusters. In this work, we encourage the creation of fewer clusters and fix α_0 and γ_0 to be less than or equal to 1 (Methods S1). After model training, we use the posterior distribution over the latent categorical indicators $Pr[\psi_n = k|\mathbf{X}]$ to determine the cluster assignment for each cell. It is worth noting that, although the prior number of normal components is infinite in Equation 1, the posterior number of components after model fitting will be finite. This truncation reflects the fact that not all infinite states are used when conditioning on finite data.[100] Additionally, the algorithm used to estimate the parameters in the NCLUSION software penalizes empty clusters (see Methods S1), and, as a result, the model has the flexibility to automatically adjust its complexity based on the inferred complexity of the data being analyzed. This helps to increase the utility and adaptability of NCLUSION across a wide range of single-cell applications.

Selection of cluster-specific marker genes

NCLUSION jointly performs clustering on single-cell populations while also learning cluster-specific gene signatures. To achieve this, we use the spike and slab prior distribution specified in Equation 2 and the resulting PIPs defined in Equation 3 to find the most salient genes per cluster. Since the model fits to each j-th gene expression for the n-th cell independently, signatures learned between clusters can share subsets of the same genes. Genes that are identified as "important" across many different clusters can effectively be seen as ubiquitous housekeeping variables rather than significant marker genes of unique cell types. Therefore, we down-weight the inclusion probabilities to proportionally penalize genes based on the number of clusters in which they appear

$$\mathrm{PIP}^*(j;k) = w_j \times \mathrm{PIP}(j;k), w_j = \left(1 - \frac{S_j}{K^*}\right) \Big/ \left(1 - \frac{1}{K^*}\right) \quad \text{(Equation 5)}$$

where $K^* \leq K$ is the finite number of occupied clusters learned by the model, and S_j is the number of clusters that the j-th gene is significant in according to a given selection threshold. We set this threshold to be 0.5 which corresponds to the median probability criterion in Bayesian statistics.[40]

While Equations 3 and 5 can be used to identify the genes that are differentially expressed in a given cluster, they do not indicate the direction or magnitude of this shift. Therefore, for each gene, we combine the adjusted posterior inclusion probabilities with effect size sign (ESS) and strictly standardized mean difference (SSMD) measures to find the most salient markers per cluster. Here, we

obtain the effect size sign by taking the sign of Cohen's d[101] between the expression of the j-th gene for cells in the k-th cluster and cells not in the k-th cluster (denoted by k').

$$\mathrm{ESS}(j;k) \simeq \mathrm{sgn}\left(\mu_{jk} - \overline{\mu}_{jk'}\right) \quad \text{(Equation 6)}$$

where, in addition to previous notation, $\overline{\mu}_{jk'} = \sum_{k'} \mu_{jk'} / (K^* - 1)$ is the average mean for the j-th gene in all clusters outside of the k-th. Here, sgn(·) is the piecewise sign function where sgn(u) = + (i.e., positive) when $u > 0$, sgn(u) = - (i.e., negative) when $u < 0$, and sgn(u) = 0 when $u = 0$.

The strictly standardized mean difference (SSMD) is a metric often used in high-throughput screenings to test for the significance of an effect size magnitude.[102–105] It is computed as the following

$$\mathrm{SSMD}(j;k) = \frac{\mu_{jk} - \overline{\mu}_{jk'}}{\sqrt{\sigma_j^2 \left[(N_k - 1)/N_k + (N_{k'} - 1)/N_{k'} \right]}}. \quad \text{(Equation 7)}$$

Asymptotically, the SSMD follows a normal distribution.[102,103,105] To determine a significant value, we follow a previous procedure[105] by calculating a threshold $|\mathrm{SSMD}(j;k)| \geq S^*(j;k)$ which controls for a predetermined false positive rate (FPR). Here, this threshold is given by

$$S^*(j;k) = \mathrm{SSMD}_{min} + \mathbf{\Phi}^{-1}\left(1 - \frac{\mathrm{FPR}}{2}\right) \varsigma_{jk} \quad \text{(Equation 8)}$$

where FPR is set to 0.05, $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal, and SSMD_{min} is the minimum SSMD magnitude that one considers to be significant. In practice, this minimum value is often set between 0 and 0.25 in order to identify weak effect sizes. In the main text, we follow previous work[106–108] and let $\mathrm{SSMD}_{min} = 0.15$. The parameter ς_{jk} is used to denote the asymptotic variance which is given by

$$\varsigma_{jk} = \frac{(N_k - 1)/N_k^2 + (N_{k'} - 1)/N_{k'}^2}{(N_k - 1)/N_k + (N_{k'} - 1)/N_{k'}} + \frac{(N_k - 1)^2/N_k^2 + (N_{k'} - 1)^2/N_{k'}^2}{2\sigma_j^2 \left[(N_k - 1)/N_k + (N_{k'} - 1)/N_{k'} \right]^3} \left(\mu_{jk} - \overline{\mu}_{jk'}\right)^2. \quad \text{(Equation 9)}$$

In the main text, cluster-specific marker genes are selected as those that have a significant adjusted inclusion probability and are notably up-regulated in a given cluster meaning that they satisfy the following criteria: (1) $\mathrm{PIP}^*(j;k) \geq 0.5$, (2) $\mathrm{ESS}(j;k) = +$, and (3) $\mathrm{SSMD}(j;k) \geq S^*(j;k)$, respectively.

Posterior inference via variational EM algorithm

We combine the likelihood in Equation 1 and the prior distributions in Equations 2 and 4 to perform Bayesian inference. In current scRNA-seq datasets, it is less feasible to implement traditional Markov Chain Monte Carlo (MCMC) algorithms due to the large number of cells being studied. For model fitting, we instead use a variational expectation maximization (EM) algorithm,[33,34,100] which allows us to estimate parameters within an optimization framework. The overall goal of variational inference is to approximate the true posterior distribution for model parameters using a set of approximating distributions. The EM algorithm optimizes parameters such that it minimizes the Kullback-Leibler divergence between the exact and approximate posterior distributions. To compute the variational approximations, we make the mean-field assumption that the true posterior can be "fully-factorized".[108] The algorithm then follows two general steps. In the first step, we iterate through a combination of hyperparameter values and compute variational updates for the other parameters using coordinate ascent. In the second step, we empirically compute (approximate) posterior values for the main model parameters $\{\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\psi}\}$. Detailed steps in the variational EM algorithm, explicit coordinate ascent updates for the model parameters, pseudocode, and other derivations are given in Methods S1. Parameters in the variational EM algorithm are initialized by taking random draws from their assumed prior distributions. Iterations in the algorithm are terminated when at least one of two stopping criteria are met: (i) the difference between the lower bound of two consecutive updates is within some small range (specified by argument ϵ), or (ii) a maximum number of iterations is reached. For the analyses run in this paper, we set $\epsilon = 1$ for the first criterion and used a maximum of 1×10^4 iterations for the second.

Simulation study design

Generating simulated datasets

To evaluate the robustness and sensitivity of NCLUSION under controlled conditions, we generated synthetic single-cell RNA-seq datasets using `scDesign3`[49] (v.1.4.0). The reference dataset we used was derived from the FACS-sorted peripheral blood mononuclear cell (PBMC) dataset produced by Zheng et al.[54] Initial preprocessing for this reference dataset included mitochondrial gene content assessment, ribosomal and hemoglobin gene filtering, and quality control to remove both low-quality cells and lowly expressed genes. Highly variable genes (HVGs) were identified using the `modelGeneVar` function in the `scran` `R` package, and the top 1000 HVGs were retained for downstream simulation. We used five immune cell types (B cells, $CD14^+$ monocytes, $CD56^+$ natural killer (NK) cells, cytotoxic T cells, and regulatory T cells) for these analyses. To ensure balanced representation across cell types, we

implemented a stratified subsampling scheme which selected an equal number of cells per type while enforcing non-zero gene expression across all selected cells and genes.

Each simulated dataset comprised of $N = 10,000$ cells across five clusters and 1,000 genes where we preserved realistic transcriptomic correlation structures through Gaussian copula modeling. Simulations were conducted across four different scenarios (with 20 replicates per scenario), each varying in cluster size imbalance and marker gene composition.

- **Scenario I:** Balanced clusters of 2000 cells per cell type, each with 50 marker genes.
- **Scenario II:** Imbalanced cluster design where one small cluster had 200 cells and the other four larger clusters each had 2450 cells. All clusters contained 50 marker genes.
- **Scenario III:** Imbalanced cluster design where one cluster had 20 (rare) cells and the other four larger clusters each had 2495 cells each. All clusters contained 50 marker genes.
- **Scenario IV:** Balanced clusters of 2000 cells per cell type, but one cluster had only 20 marker genes while the other four clusters had 50 marker genes.

More specifically, synthetic datasets were generated using the `construct_data`, `fit_marginal`, `fit_copula`, `extract_para`, and `simu_new` functions within `scDesign3` to create gene expression vectors using a negative binomial distribution that is conditioned on cell type from the reference data. To introduce differentially expressed genes (DEGs), we first ranked genes by their cell type specific mean expression in the reference data and sampled some number of top-ranked genes to be markers. Then in the synthetic data, these DEGs were then artificially upregulated in one cluster while maintaining the baseline expression in others. This was done by apply a log-fold change factor sampled uniformly over the interval [1.5, 2.5]. This ensured that we maintained realistic variance but still had distinct signal between cell types.

Real datasets and preprocessing

Below we briefly describe all of the datasets and the preprocessing steps used in this work. Each of these datasets is relatively large (containing at least 20,000 cells) with unique molecular identifiers (UMI). The latter is important because prior research suggests that UMIs provide enough information to avoid overcounting issues due to amplification and zero-inflation.[14,109,110] We use an asterisk by the BRAIN-LARGE dataset to indicate that it was exclusively to test the scalability of NCLUSION and competing methods; therefore, clustering performance was not recorded. For the other datasets, we use cell type annotations provided by the original study as "true" reference labels during our analyses. Cells were filtered for quality using a custom `scanpy`[2] (v.1.9.1) pipeline script (see Software availability). Unless otherwise stated, all data was preprocessed by taking the logarithm (to the base 2) of the counts, dividing by a scaling factor of 10000, and then adding a pseudo-count of 1.0 for stability. Additionally, unless otherwise stated, all results were produced using the top 5000 highly variable genes (HVG), which were determined by sorting the standard deviation of the transformed counts.[35] Note this number of HVGs was chosen to maximize the biological variability captured in each study.

*BRAIN-LARGE**

This dataset originally contains 1.3 million mouse brain cells from 10× Genomics.[47] During preprocessing, we subset the data to a collection of 720 genes following a procedure outlined by Lopez et al.[14] Next, we further filtered by only keeping cells that had at least one of these genes expressed. This left a total of 64,071 cells. Since the original study did not provide cell labels, we exclusively used this dataset to compare runtime performance. To do so, we up-sampled by randomly selecting groups of 64,071 cells to create a synthetic dataset of 1 million cells. We report the runtime for each method on datasets with 500, 1K, 5K, 10K, 50K, 100K, 500K, and 1M cells.

PBMC

We took scRNA-seq data from fluorescence-activated cell sorted (FACS) populations of peripheral blood mononuclear cells (PBMCs) provided by Zheng et al.[54] and concatenated each population into one dataset. During preprocessing, we filtered out genes that were expressed in fewer than three cells. We also dropped cells with (i) fewer than 200 genes expressed, (ii) greater than 20% mitochondrial reads, and (iii) fewer than 5% ribosomal reads. This resulted in a final dataset with 94,615 high-quality cells representing 10 distinct cell types.

PDAC

We used scRNA-seq data from pancreatic ductal adenocarcinoma (PDAC) tissue obtained from 23 patients according to methods documented in Raghavan et al.[73] This dataset contains 23,042 total cells made up of 15,302 non-malignant cells of 11 distinct cell types and 7,740 malignant cells.

AML

The scRNA-seq data obtained from van Galen et al.[74] contains 43,690 acute myeloid leukemia (AML) and non-malignant donor cells taken from 16 AML patients, 5 healthy donors, and 2 cell lines. It is comprised of 13,489 patient-derived malignant cells, 23,005 non-malignant donor cells, 6,018 cells from the MUTZ-3 AML cell line, and 1,178 cells from the OCI-AML3 cell line. To account for the biological differences between cell lines and donor cells of the same cell type annotation, we appended the cell line name onto cell type labels where applicable, producing 33 distinct cell types overall. To process the data, we filtered out all cells with "unclear" cell state labels, retaining only "malignant" or "non-malignant" cells.

IMMUNE

We obtained filtered scRNA-seq data from approximately 330,000 immune cells from 12 organ donors in Domínguez Conde et al.[75] We isolated 88,057 cells that were taken from a single organ donor (donor D496) with uniform chemistry annotations, containing 44 distinct immune cell types.

Other methods

We selected five additional methods to compare against the performance of NCLUSION in real data and in simulations: (1) a Louvain algorithm implemented using the `FindClusters` function in Seurat[9] (v.4.3.0.1); (2) a spectral clustering method called scLCA[10] (v.0.0.0.9000), which optimizes both intra- and inter-cluster similarity; (3) a combination of a K-Nearest Neighbor (KNN) classifier with the Louvain community detection algorithm to find clusters implemented via `scikit-learn`[45] (v.1.2.2) and `scanpy`[2] (v.1.9.1), respectively; (4) a semi-soft clustering algorithm called SOUP[46] (v.0.0.0.9000); and (5) an ensemble method called scCCESS-SIMLR (v.0.2.1), which leverages the spectral clustering approach SIMLR.[11,96] In the simulation experiments, we also compared the clustering performance of NCLUSION against three additional methods: (6) a consensus clustering method, SC3[15](v.1.34.0); (7) a deep-learning based method, scDeepCluster[51] (v.1.0.0); and (8) an imputation and dimensionality reduction method, CIDR[50] (v.0.1.5). Also in simulations, when assessing the ability of NCLUSION to perform robust marker gene selection, we compare it against: (9) a method that leverage differential correlation patterns in the local structure of a PCA-derived cell neighborhood graph, DUBStepR[52] (v.1.2.0); (10) a feature selection via an EM algorithm, FESTEM[53] (v.1.2.1); and (11) a divergence-based strategy with permutation tests, singleCellHaystack[27] (v.1.0.2). Additional details about each method are provided in Methods S1.

Evaluation metrics

Below we describe the metrics and approaches used to compare performance across all methods. Our clustering evaluation procedure used extrinsic metrics that require reference labels to serve as the ground truth in our calculations.

Normalized mutual information (NMI)

This metric is a normalized variant of mutual information (MI). It is an entropy-based metric that captures the amount of shared information between the inferred label distribution and the reference label distribution. NMI ranges between [0, 1] where 1 represents total information sharing between label sets and 0 represents no information sharing between label sets. NMI is calculated by

$$\mathrm{NMI} = \frac{I(Q;R)}{\sqrt{\mathbb{H}(Q)\mathbb{H}(R)}}$$

where Q and R are the empirical label distributions from the inferred and reference labels, respectively. The function $I(\cdot)$ is the mutual information between the inferred labels distribution and reference labels distributions; $\mathbb{H}(\cdot)$ represents the Shannon entropy of a given label distribution.[96,111]

Adjusted Rand index (ARI)

This metric captures the similarity between labels inferred by a method and the reference labels. It is based on the Rand index (RI) but corrects for the measurement's sensitivity to chance. ARI ranges between [-1, 1] where 1 represents perfect agreement between label sets, 0 represents random agreement, and -1 represents perfect disagreement. ARI is calculated by

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\binom{n}{2}}{1/2\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\binom{n}{2}}$$

where n_{ij}, a_i, and b_j are values obtained from a contingency table, and $n = \sum_{ij} n_{ij}$.[96,111,112]

Metrics to evaluate marker gene selection in simulations

In the simulation studies, we evaluated the accuracy of marker gene detection of NCLUSION and competing methods by treating the task as a classification problem. In order to do so, we defined the confusion matrix defined in Table 1.

- **True positive rate (TPR; also referred to as power)** captures the proportion of correctly identified marker genes using a given method. It is defined as TPR = TP/(TP + FP).
- **False discovery rate (FDR)** details the proportion of all identified marker genes that are actually non-marker genes It is defined as FDR = FP/(TP + FP).
- **False positive rate (FPR)** captures the proportion of non-marker genes that will be incorrectly labeled marker genes. It is defined as FP/(TN + FP).

Note that the false positive rate can also be computed as FPR = 1-Specificity.

Normalized module expression

Genes with significantly adjusted PIPs in Equation 5, positive ESS in Equation 6, and significant SSMD in Equation 7 were used to generate modules (i.e., a collection of marker genes) for each cluster. We calculated a score for each cluster to asses the exclusivity

of expression within each module. This was done using the `score_genes` function in `scanpy` (v.1.10.4). The violin plots were generated using the `violinplot` function in matplotlib (v.3.10.0).

Gene set over-enrichment analysis

We also performed gene set enrichment analysis on each of the learned gene modules across clusters. This was done via an over-enrichment analysis within the `GSEApy` package[64] (v.1.1.5) in Python (v.3.11.0). This method uses a hypergeometric test to calculate the enrichment of genes in a supplied module with respect to the gene sets within an ontology. In this work, we use the ontology labeled `GO_Biological_Process_2025`,[64–66,113] `Tabula_Sapiens`,[114,115] `Azimuth_Cell_Types_2021`,[116] `KEGG_2021_Human`,[117–119] and `Reactome_2022`.[120–127] The gene sets in this particular ontology represent a combination of biological processes, pathways, and phenotypes. In this analysis, we use *q*-values to determine the enrichment of a given gene set with a significance threshold set to 0.05. The *q*-value is the analog of a *p*-value that has been corrected for testing multiple hypotheses (i.e., an adjusted *P*).

QUANTIFICATION AND STATISTICAL ANALYSIS

This study reports the development of a new statistical model with details and evaluation metrics described in the "method details" section.