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], Andrew W. Navia[4-8], Alex K. Shalek[4-7,9,10], Nicolo Fusi[3], Srivatsan Raghavan[5,8,9,11§], Peter S. Winter[5,8,§], Ava P. Amini[3,§,†], and Lorin Crawford[1,3,12,§,†]

1 Center for Computational Molecular Biology, Brown University, Providence, RI, USA
2 Warren Alpert Medical School of Brown University, Providence, RI, USA
3 Microsoft Research, Cambridge, MA, USA
4 Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA
5 Broad Institute of MIT and Harvard, Cambridge, MA, USA
6 Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA
7 Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA
9 Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
10 Harvard Medical School, Boston, MA, USA
11 Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA
11 Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA
12 Department of Biostatistics, Brown University, Providence, RI, USA

§ **Authors Contributed Equally**
† **Corresponding E-mail: ava.amini@microsoft.com; lcrawford@microsoft.com**

Abstract

Clustering is commonly used in single-cell RNA-sequencing (scRNA-seq) pipelines to characterize cellular heterogeneity. However, current methods face two main limitations. First, they require user-specified heuristics which add time and complexity to bioinformatic workflows; second, they rely on post-selective differential expression analyses to identify marker genes driving cluster differences, which has been shown

to be subject to inflated false discovery rates. We address these challenges by introducing nonparametric clustering of single-cell populations (NCLUSION): an infinite mixture model that leverages Bayesian sparse priors to identify marker genes while simultaneously performing clustering on single-cell expression data. NCLUSION uses a scalable variational inference algorithm to perform these analyses on datasets with up to millions of cells. By analyzing publicly available scRNA-seq studies, we demonstrate that NCLUSION (i) matches the performance of other state-of-the-art clustering techniques with significantly reduced runtime and (ii) 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 analysis[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.

However, 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 adds 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 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]. 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 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[20,21]. 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[22–26]. Recently, others have proposed unified frameworks for simultaneous clustering and marker gene selection using hierarchical tree-based algorithms[27], regularized copula models[28], and *post hoc* sensitivity measures[29]. 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 "Nonparametric CLustering of SIngle-cell populatiONs" (NCLUSION): 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 mix-

ture model[30–35], 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 increase 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[20,21]. In this work, we aim to address these challenges using a new approach: NCLUSION.

NCLUSION leverages a Bayesian nonparametric mixture 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 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 always setting the number of possible clusters $K = \infty$, we remove the need for users to iterate through 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 towards zero.

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[36] PIP ≥ 0.5). We use these adjusted inclusion probabilities with the effect size sign (ESS) of each gene to filter for significant genes that are also up-regulated (indicated by positive ESS 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[37–39]. In training, our objective is to minimize the Kullback–Leibler (KL) divergence between the variational posterior and the true posterior[40]. Optimization during model fitting occurs using a coordinate ascent procedure where parameters are sequentially updated based on their gradients (Methods and Supplementary Material). 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 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], k-nearest neighbors followed by the Leiden clustering algorithm (KNN+Leiden)[41], SOUP[42], 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[43] dataset (Figure 1c). To facilitate comparison across all baseline methods, we follow Lopez et al.[14] and limit the analysis to 720 genes, while sub-sampling the number of cells from 500 to 1 million cells. As a practical reference, we use a grey dotted line to highlight the runtime for each method at the median scRNA-seq dataset size as determined in 2020: 31,000 cells[44]. 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 (Supplementary Figure 1). These scalability results show the potential of NCLUSION's utility for emerging large-scale single-cell studies.

NCLUSION achieves competitive clustering performance with less runtime

Next, we assessed the quality of clustering done by NCLUSION as compared to other baseline approaches. Clustering is done as a step to ensure accurate annotation of cell types during the exploratory analysis of new single-cell data sets. Here, we tested each model on scRNA-seq from FACS-purified peripheral blood mononuclear cells (PBMC)[45]. This dataset captures 10 cellular populations including CD14+ monocytes, CD34+ cells, and 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 (Methods), the final dataset contained 94,615 cells and 5,000 genes with the highest standard deviation (post log-normalization)[20,46]. We evaluated the performance of each clustering algorithm 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[45] (Figure 2a-b).

To qualitatively assess clustering performance, we used contingency heat maps to evaluate how well each method captured the unique cell types across clusters (Figure 2c). For a given method, each i-th row of the heat map 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 (i, k)-th element in the heat map indicates the fraction of a i-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 2c). Furthermore, all methods except scCCESS-SIMLR were able to separate the natural killer (NK) cell population with an inferred

cluster occupancy rate of greater than 75% (Supplementary Table 1). 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 sub-populations[47] (Figure 2c).

As a quantitative assessment of each method's clustering performance, we used the FACS-derived experimental annotations as reference labels and calculated three validation metrics, each of which measures how much variation and sample diversity a clustering algorithm can explain from the data (Methods): the Jaccard similarity coefficient, the Gini-Simpson index, and McIntosh evenness, where 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 2d). NCLUSION's performance consistently fell within the top three among all competing approaches. NCLUSION obtained the highest mean Jaccard coefficient of 0.72 ($\pm 8.91 \times 10^{-3}$ standard deviation), with Seurat and scLCA each scoring lower values of 0.71 ($\pm 1.62 \times 10^{-3}$ and $\pm 1.48 \times 10^{-2}$, respectively). When comparing performance using the Gini-Simpson index and McIntosh evenness measure, NCLUSION remained competitive with the other competing methods. Specifically, NCLUSION obtained a mean Gini-Simpson index of 0.84 ($\pm 3.87 \times 10^{-3}$), within range of scLCA ($0.86 \pm 2.20 \times 10^{-2}$). NCLUSION also obtained the third highest McIntosh evenness score of 0.82 ($\pm 1.05 \times 10^{-2}$), with only Seurat and scLCA scoring higher values of 0.88 ($\pm 8.23 \times 10^{-3}$) and 0.96 ($\pm 2.12 \times 10^{-2}$), respectively. Importantly, we found no statistically significant differences when comparing the median performance of NCLUSION to the other approaches (Kruskal–Wallis H-test $P > 0.95$ for all three metrics; Supplementary Table 1). Lastly, NCLUSION recorded the shortest runtime for this analysis without any iterative processes or optimization of hyperparameters. It finished approximately 254 seconds (4.23 minutes) faster than Seurat, more than 12900 seconds (215.00 mins or 3.58 hrs) faster than scLCA, and more than 41331 seconds (688.85 mins or 11.48 hrs) faster than scCCESS-SIMLR.

NCLUSION automatically identifies cell type-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 two criteria: (i) an adjusted posterior inclusion probability, PIP($j;k$), which provides evidence that the j-th gene's mean expression is uniquely nonzero within the k-th cluster and (ii) 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 (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 (Figure 3a-b).

Overall, NCLUSION successfully identified cluster-specific marker genes that are known to be associated with examined cell types (Figure 3c and Supplementary Table 2). For example, in clusters of cells mapping back to the T cell population, we observed that NCLUSION correctly identified marker genes known to play an important role in T cell biology, such as *CD8A* (PIP = 0.538), *CD8B* (PIP = 0.615)[48], and *CD40LG* (PIP = 0.615)[49]. Furthermore, genes associated with cytotoxicity, such as *GZMB* (PIP = 0.692), tended to be selected in clusters largely containing CD8+ T cells and NK cells – two cell types that have been shown to have functionally similar cytotoxic activity[50]. In other clusters, where we observed genes associated with both B cells (e.g., *CD19*, PIP = 0.923; *LINC00926*, PIP = 0.923 *MS4A1*, PIP = 0.846)[51–53] and myeloid lineages (e.g., *MS4A6A*, PIP = 0.846; *S100A8*, PIP 0.846; *LYZ*, PIP = 0.692)[54], NCLUSION distinguished genes known to play an important role in T cell biology as statistically insignificant. 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 3c).

To further evaluate marker gene quality, we computed gene module scores in order to compare the normalized expression for signature genes across clusters (Figure 3d). 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) (Figure 3e and Supplementary Figure 2). In clusters that contained heterogeneous combinations of T cell sub-populations, 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.[45] 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 variable selection approach in the NCLUSION framework results in smaller and more refined transcriptomic signatures for downstream investigation (Figure 3f and Supplementary Figures 3-12). For example, NCLUSION identified 32 cluster-specific marker genes for NK cells, 99% of which were also included in the 1,780 marker genes selected *post hoc* by Seurat. In total, an average of 98% of the marker genes identified by NCLUSION were included in the much larger sets of differentially expressed 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[55–57] (Figure 3g-h and Supplementary Figures 13-24). For example, in the cluster containing predominantly B cells (inferred cluster 1), we observed an up-regulation of B cell receptor signaling (q-value = 7.21×10^{-15}), B cell activation (q-value = 2.62×10^{-11}), and antigen receptor-mediated signaling (q-value = 2.11×10^{-8})[58,59]. 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 lymphocyte chemotaxis (q-value = 2.80×10^{-6}) and positive regulation of natural killer cell chemotaxis (q-value = 4.80×10^{-6})[60–62]. We also observed enrichment of the up-regulation of natural killer cell mediated cytotoxicity (q-value =1.47×10^{-5}), consistent with the known highly cytotoxic behavior of NK cells[60,63]. Each of these gene sets was insignificant when using differentially expressed genes from Seurat (Figure 3g). Lastly, in the monocyte dominated cluster (inferred cluster 3), the NCLUSION-generated module was enriched for both the lipopolysaccharide-mediated signaling pathway (q-value = 2.59×10^{-4}) and tumor necrosis factor production (q-value = 2.89×10^{-4}), two processes known to be associated with monocyte/macrophage-specific inflammatory responses[64,65]. 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 Supplementary Table 2. Together, these results demonstrate that NCLUSION can identify cluster-specific gene signatures that reflect underlying cellular phenotypes.

NCLUSION generalizes to the Tissue Immune Atlas

Next, we assessed the generalizability of NCLUSION by testing it on an additional scRNA-seq study: the tissue immune (IMMUNE) atlas dataset, which contains healthy white blood cells taken from different anatomical locations from multiple donors[66] . After extracting data from a single donor and performing quality control (Methods), we had a total of 88,057 cells and 5,000 genes with the highest standard deviation (post log-normalization) for the analysis.

Once again, we quantitatively evaluated the clustering performance of NCLUSION and the competing baselines using the annotations provided by the original study (Supplementary Figures 25-26). NCLUSION maintains competitive clustering performance as measured by Jaccard similarity, Gini-Simpson index, and McIntosh evenness, with all methods showing a similar median performance (Kruskal–Wallis H-test $P > 0.80$ for all three metrics) (Supplementary Figure 27 and Supplementary Table 3). Next, we analyzed the interpretability of the cluster-specific marker genes inferred by NCLUSION (Supplementary Figures 28-32 and Supplementary Table 4). Examination of the cluster structure between both the annotations provided by the original study and the clusters inferred by NCLUSION showed that NCLUSION primarily groups cells based on their hematopoietic lineage, inferring fewer clusters that broadly capture B, Plasma, cytotoxic T, non-cytotoxic T, and myeloid cell lineages. Here, we continued to observe the general pattern of genes with significant PIPs also being up-regulated in the same inferred cluster; however, this relationship was not exclusive as genes with high PIPs could also be significantly under-expressed (Supplementary Figures 28-29).

By again selecting cluster-specific marker genes to be those with both PIPs above the median probability model threshold and positive ESS, normalized gene module scores revealed increased expression of marker genes in corresponding and related clusters (Supplementary Figure 31). Similar to the PBMC analysis, NCLUSION appeared to use the variance among Macrophage-like and T cell-like expression patterns as a main factor in distinguishing cell types in the IMMUNE dataset. Cluster 1 primarily contained cells labeled as alveolar macrophages. Genes relevant to both macrophage biology, such as *SLC8A1* (PIP = 0.5), and alveolar immune biology, such as *TPSB2* (PIP = 0.749) and *TPSB2* (PIP = 0.750), were selected in this cluster[67,68]. GO analysis in Cluster 1 revealed an enrichment in macrophage-related pathways (e.g., positive regulation of monocyte chemotaxis, q-value = 1.15×10^{-2}; regulation of lipopolysaccharide-mediated signaling pathway, q-value = 1.24×10^{-2}). In addition, we observed enrich-

ment of processes related to the regulation of macrophage-derived foam cell differentiation, consistent with prior reports on alveolar macrophages[67,69]. Conversely, cells with T cell-like expression patterns were found predominantly in clusters 2, 8, and 9 (Supplementary Figure 25). Clusters 2 and 8 showed an up-regulation in salient cytotoxic T and NK cell marker genes such as *KLRC1* (PIP = 0.625), *KLRC2* (PIP = 0.5), and *KLRC3* (PIP = 0.5) and were enriched for multiple T and NK cell cytotoxic behavior-related pathways (Supplementary Table 4). Lastly, cluster 9 contained over-expressed non-cytotoxic T cell markers such as *CD40L* (PIP = 0.5) and *LEF1* (PIP = 0.5)[70,71], and was enriched for GO processes such as T-helper 1 cell differentiation (q-value = 2.17×10^{-2}) and T-helper 17 type immune response (q-value = 2.74×10^{-2}).

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 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 cluster-specific marker genes that significantly drive the clustering. Notably, 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[21,72,73].

The current implementation of the NCLUSION framework offers many directions for future development and applications. First, 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. One possible extension

of NCLUSION would be to incorporate additional genomic information into the sparse prior distributions used for Bayesian variable selection. For example, previous studies have proposed an integrative approach where the importance of a variable also depends on an additional set of covariates[38,74,75]. 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 new prior would assume that biologically related pathways contain shared marker genes, essentially integrating the concept of gene set enrichment analysis into clustering. Second, although it helps NCLUSION scale to large datasets, variational expectation-maximization (EM) algorithms are known to both produce slightly miscalibrated parameter estimates and underestimate the total variation present within a dataset[39,76,77]. 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.[45] (e.g., Figure 2). 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[78]. Lastly, 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[79]. Future efforts could extend NCLUSION to this emerging modality by, for example, reformulating the method as a spatial Dirichlet process mixture model[80].

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.

Online Methods

Overview of NCLUSION

We provide a brief overview of the probabilistic framework underlying the "Nonparametric CLUstering of SIngle-cell populatiONs" (NCLUSION) model. Detailed derivations of the algorithm are provided in

the Supplementary Material. Consider a study with single-cell RNA sequencing (scRNA-seq) expression data for $i = 1, \ldots, N$ cells that each have measurements for $j = 1, \ldots, G$ genes. Let this dataset be represented by the $N \times G$ matrix $\mathbf{X}$ where the row-vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iG})$ denotes the expression profile for the i-th cell. We assume that the log-normalized gene expression for each cell follows a sparse hierarchical Dirichlet process normal mixture model[30–32] of the form

$$x_{ij} \sim \sum_{k=1}^{\infty} \pi_k \, \mathcal{N}(\mu_{jk}, \sigma_{jk}^2) \tag{1}$$

where π_k represents the marginal (unconditional) probability that a cell belongs to the k-th cluster, and μ_{jk} and σ_{jk}^2 are the gene-specific means and variances within the k-th cluster. There are two key features in NCLUSION that are based on the formulation above. First, we assume that the formation of clusters is driven by a few important genes that have mean expression shifted away from zero. To that end, we place a sparsity-inducing spike and slab prior distribution on the mean effect of each gene

$$\mu_{jk} \sim \eta_k \, \mathcal{N}(0, \lambda_j^2) + (1 - \eta_k)\delta_0, \tag{2}$$

where δ_0 is a point mass at zero, λ_j^2 is the variance component of the "slab" distribution, and η_k is the prior probability that any given gene has a nonzero effect when assigning a cell to the k-th cluster. In practice, there are many different ways to estimate η_k. Following previous work[38,39,81–83], one choice would be to assume a uniform prior over $\log \eta_k$ to reflect our lack of knowledge about the correct number of "marker" genes for each cell type that is present in the data. In this work, we set $\eta_k = 1/G$ to represent our prior belief that each cluster is formed with at least one marker gene with a nonzero mean. 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 $\Pr[\rho_{jk} = 1] = \eta_k$. 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. As NCLUSION is trained, the posterior mean for unimportant genes will trend towards zero 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

$$\mathrm{PIP}(j; k) \equiv \Pr[\mu_{jk} \neq 0 \,|\, \mathbf{X}]. \tag{3}$$

The second key feature in Eq. (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_i which indicates that the i-th cell is in the k-th cluster with prior probability π_k. Explicitly, we write this as $\Pr[\psi_i = k] = \pi_k$. Here, we implement the stick-breaking construction of the Dirichlet process[30] where we say

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

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[84]. 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[30,37]. 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 (Supplementary Material). After model training, we use the posterior distribution over the latent categorical indicators $\Pr[\psi_i = k \,|\, \mathbf{X}]$ to determine the cluster assignment for each cell. It is worth noting that, although the prior number of normal components is infinite in Eq. (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[84]. Additionally, the algorithm used to estimate the parameters in the NCLUSION software penalizes empty clusters (see Supplementary Material), 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 Eq. (2) and the resulting PIPs defined in Eq. (3) to find the most salient genes per cluster. Since the model fits to each j-th gene expression for the i-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

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

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[36]. We combine the adjusted posterior inclusion probabilities with the effect size sign of each gene to find the most salient markers per cluster. NCLUSION not only identifies which genes are significant markers for each cluster but also estimates the size and sign of their effects (i.e., whether they are up-regulated or down-regulated depending on the direction of interest). We obtain the effect size sign by taking the sign of Cohen's d[85] 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')

$$\mathrm{ESS}(j;k) = \mathrm{sgn}\left(\frac{\mu_{jk} - \bar{\mu}_{jk'}}{\sqrt{\frac{1}{2}(\sigma^2_{jk} + \bar{\sigma}^2_{jk'})}}\right) \tag{6}$$

where $\bar{\mu}_{jk'} = \sum_{k'} \mu_{jk'}/(K^* - 1)$ and $\bar{\sigma}^2_{jk'} = \sum_{k'} \sigma^2_{jk'}/(K^* - 1)$ are the average mean and variances for the j-th gene in all clusters outside of the k-th. Here, $\mathrm{sgn}(\cdot)$ is the piecewise sign function where $\mathrm{sgn}(u) = +$ (i.e., positive) when $u > 0$, $\mathrm{sgn}(u) = -$ (i.e., negative) when $u < 0$, and $\mathrm{sgn}(u) = 0$ when $u = 0$. In the main text, cluster-specific marker genes are selected as those that have a significant adjusted inclusion probability and are up-regulated in a given cluster such that $\mathrm{PIP}^*(j;k) \geq 0.5$ and $\mathrm{ESS}(j;k) = +$, respectively.

Posterior inference via variational EM algorithm

We combine the likelihood in Eq. (1) and the prior distributions in Eqs. (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[30,31,84], 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" [86]. The algorithm then follows two general steps. In the first step, we iterate through a combination of hyper-parameter 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 $\{\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\pi}\}$. Detailed steps in the variational EM algorithm, explicit coordinate ascent updates for the model parameters, pseudocode, and other derivations are given in the Supplementary Material. 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 150 iterations for the second.

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 [14,87,88]. 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` [89] (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 (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 [32].

BRAIN-LARGE*. This dataset originally contains 1.3 million mouse brain cells from 10x Genomics [43]. 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. [45] 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.

IMMUNE. We obtained filtered scRNA-seq data from approximately 330,000 immune cells from 12 organ donors in Domínguez Conde et al. [66]. To mitigate batch effects, 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: (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` [41] (v.1.2.2) and `scanpy` [89] (v.1.9.1), respectively; (4) a semi-soft clustering algorithm called SOUP [42] (v.0.0.0.9000); and (5) an ensemble method called scCCESS-SIMLR (v.0.2.1), which leverages the spectral clustering approach SIMLR [11,90]. Additional details about each method are provided in the Supplementary Material.

Evaluation

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.

Jaccard similarity. This metric captures the amount of overlap between two finite sets. It ranges between [0, 1] where 1 denotes complete overlap and 0 signifies no overlap. The Jaccard similarity is calculated as follows[90,91]

$$J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A}| + |\mathcal{B}| - |\mathcal{A} \cap \mathcal{B}|}$$

where $\mathcal{A}$ and $\mathcal{B}$ represent the sets of inferred and reference labels, respectively.

Gini-Simpson index. This metric is the complement to Simpson's index[92]. It measures the probability that two cells taken at random from the dataset of interest (with replacement) represent different cell types. The Gini-Simpson index is calculated as the following

$$D^* = 1 - \sum_{k=1}^{K} \frac{N_k(N_k - 1)}{N(N-1)}$$

where N is the total number of cells in the study, K is the number of inferred clusters (or cell types), and N_k represents the number of cells assigned to the k-th cluster. The Gini-Simpson index ranges between $0 < D^* \leq 1$ where larger values indicate that more cellular heterogeneity is being captured by a given clustering algorithm.

McIntosh evenness. This metric can also be seen as a variation of Simpson's index[93,94]. It expresses the cellular heterogeneity captured by a clustering algorithm in geometric terms. Briefly, a group of cells can be represented as a point in a K-dimensional hyper-volume where each inferred cell type represents a dimension. The McIntosh evenness value is calculated as the following

$$E = \sqrt{\frac{\sum_k N_k^2}{(N - K + 1)^2 + K - 1}}$$

where N is the total number of cells in the study, K is the number of inferred clusters (or cell types), and N_k represents the number of cells assigned to the k-th cluster. The McIntosh evenness ranges between $0 < E \leq 1$ where larger values indicate that more cellular heterogeneity is being captured by a given clustering algorithm.

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

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

where n_{ij}, a_i, and b_j are values obtained from a contingency table, and $n = \sum_{ij} n_{ij}$ [90,91,95].

Normalized module expression. Genes with significantly adjusted PIPs in Eq. (5) and positive ESS in Eq. (6) 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 `AddModuleScore` function in Seurat (v.5.0.0) within the `R` programming language (v.4.2.0). We used a min-max scaling procedure to normalize the module scores between 0 and 1 across all clusters. This was to facilitate qualitative evaluation across all generated gene modules. The violin plots were generated using the `VlnPlot` function in Seurat (v.5.0.0).

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[96] (v.1.0.5) in Python (v.3.7.4). 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_2023" [55–57]. 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.

Software availability

An open-source software implementation of NCLUSION is available on GitHub at `https://github.com/microsoft/nclusion`. Guided tutorials and all code needed to reproduce the results and figures in this work can be found at `http://microsoft.github.io/nclusion`.

Data availability

All of the datasets analyzed in this paper are publicly available. 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.[45] can be downloaded directly from `https://www.10xgenomics.com/resources/datasets`. Lastly, the immune cell atlas dataset can be accessed at `https://cellgeni.cog.sanger.ac.uk/pan-immune/CountAdded_PIP_global_object_for_cellxgene.h5ad`.

Acknowledgements

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 LC. CN was a trainee supported under the Brown University Predoctoral Training Program in Biological Data Science (NIH T32 GM128596). SR 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.

Author contributions

CN, APA, and LC conceived the study and developed the methods. CN and MH developed the algorithm and software. CN, MH, and MR led the analyses. SR, PSW, APA, and LC provided resources, supervised the project, and conducted secondary analyses. All authors interpreted the results and wrote and revised the manuscript.

Competing interests

SR holds equity in Amgen and receives research funding from Microsoft. All other authors have declared that no competing interests exist.

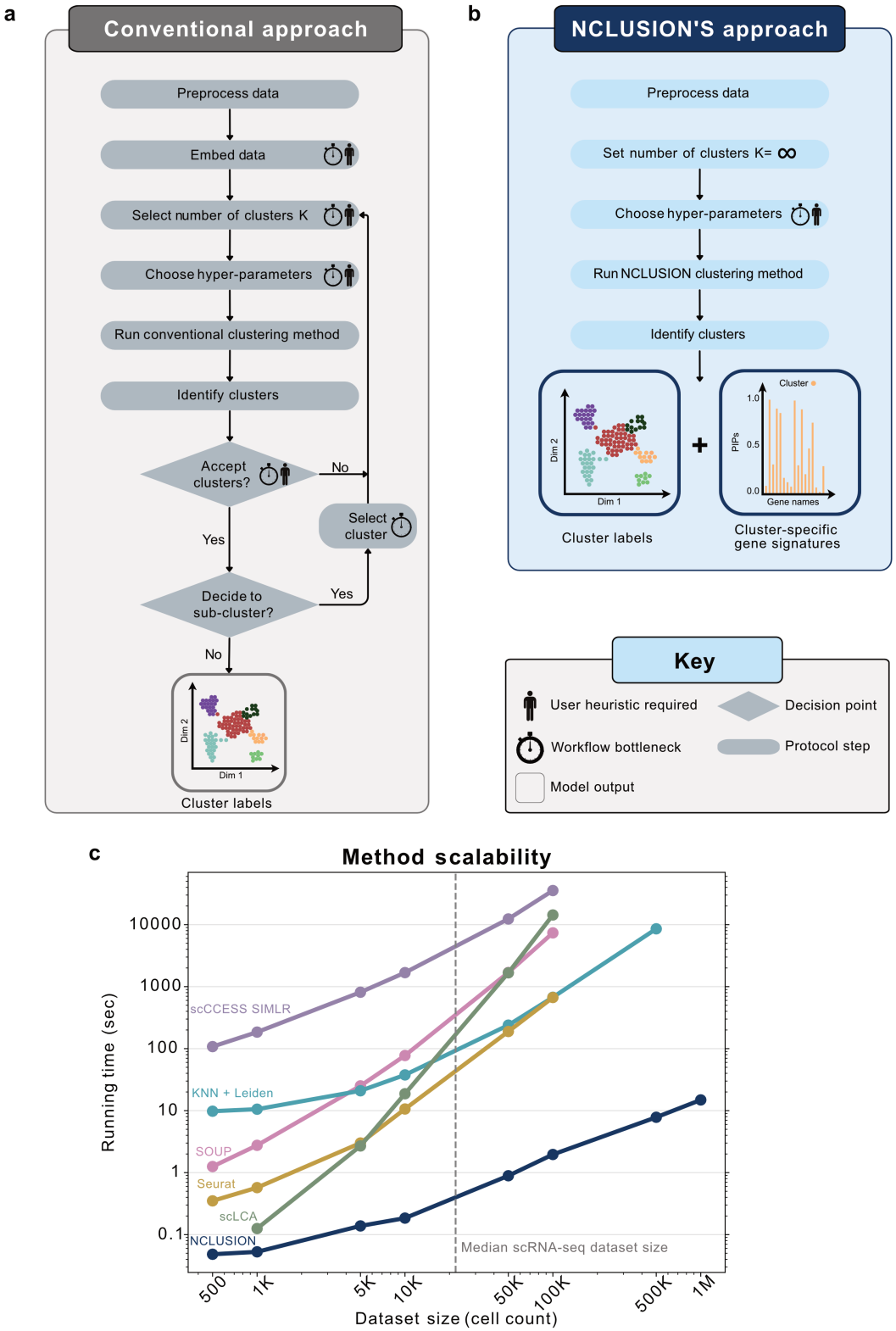

Figure 1. NCLUSION provides a scalable, unified workflow for both clustering and marker gene selection in single-cell analysis. (a) Conventional clustering algorithms require user heuristics and decision making steps that increase wall clock runtime (e.g., selection and human-in-the-loop refinement of 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.

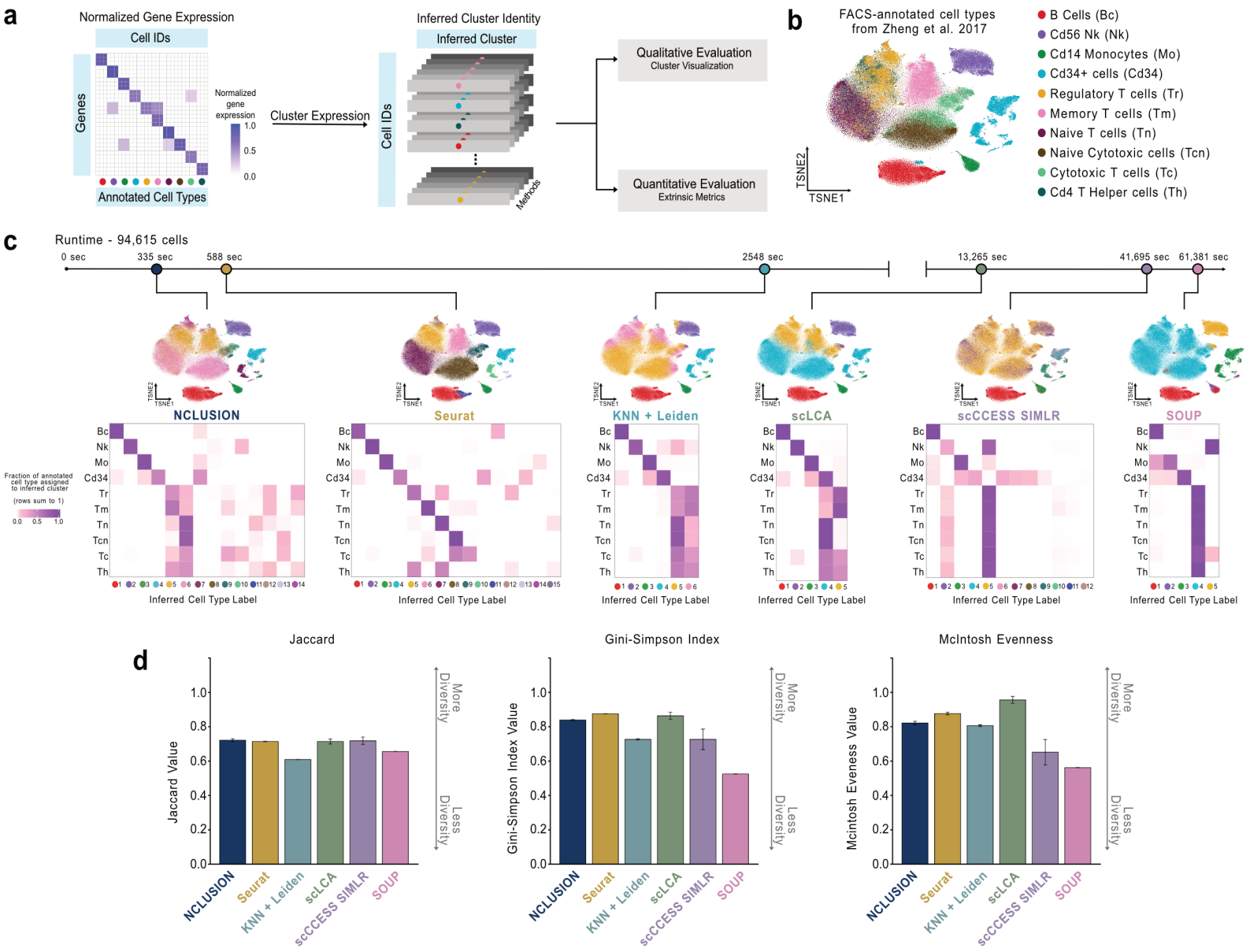

Figure 2. Clustering performance for NCLUSION and other baseline methods on the PBMC scRNA-seq dataset (N = 94,615 cells). (a) The framework used for evaluating the quality of clustering in each method. **(b)** Overview of FACS-based cell type annotations, visualized via t-distributed stochastic neighbor embedding (t-SNE), for the PBMC scRNA-seq dataset. These annotations serve as labels during the evaluation. **(c)** Visualizing the structure of the inferred clusters across all baselines using t-SNEs and a contingency heat map showing the prevalence of each cell type within each cluster. Methods are ordered from fastest (left) to slowest (right) in terms of runtime. Note we reuse the same lower dimensional representation of the data and simply relabel the plots according to the results from each clustering algorithm. **(d)** Assessment of the inferred clusters labels versus the experimental annotations, as quantified by three metrics: Jaccard similarity, Gini-Simpson index, and McIntosh evenness (for each method, we take five random 80% splits of the PBMC dataset; depicted in each bar plot is the mean ± 95% confidence interval).

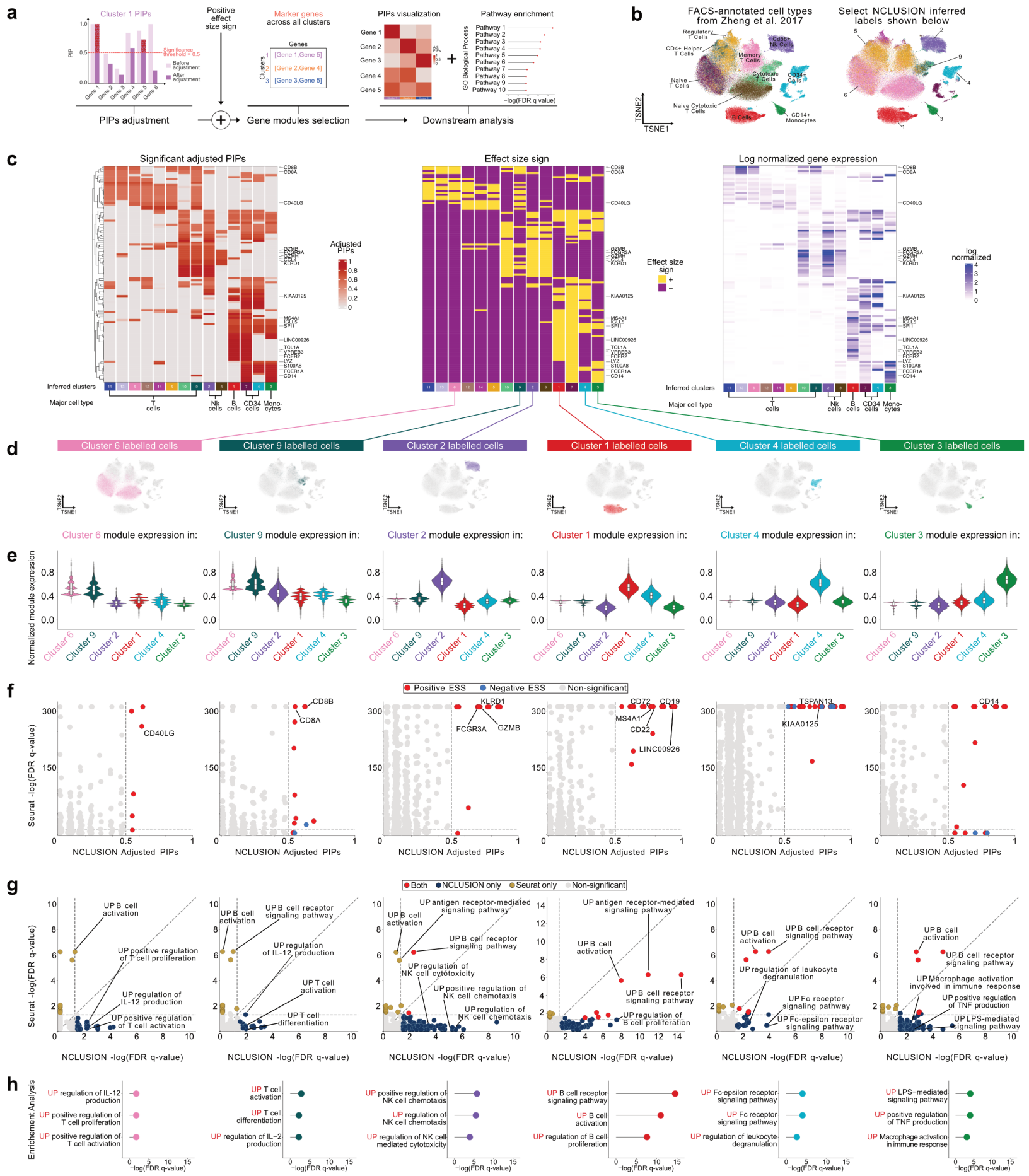
a
Cluster 1 PIPs
Significance threshold = 0.5
Before adjustment
After adjustment
Positive effect size sign
Marker genes across all clusters
Genes
Clusters
[Gene 1,Gene 5]
[Gene 2,Gene 4]
[Gene 3,Gene 5]
PIPs visualization
Pathway enrichment
GO Biological Process
−log(FDR q value)
PIPs adjustment
Gene modules selection
Downstream analysis
b
FACS-annotated cell types from Zheng et al. 2017
Select NCLUSION inferred labels shown below
Regulatory T Cells
CD4+ Helper T Cells
Memory T Cells
CD56+ NK Cells
Naive T Cells
Cytotoxic T Cells
CD34+ Cells
Naive Cytotoxic T Cells
B Cells
CD14+ Monocytes
TSNE2
TSNE1
c
Significant adjusted PIPs
Effect size sign
Log normalized gene expression
Adjusted PIPs
Effect size sign
log normalized
Inferred clusters
Major cell type
T cells
Nk cells
B cells
CD34 cells
Mono-cytes
d
Cluster 6 labelled cells
Cluster 9 labelled cells
Cluster 2 labelled cells
Cluster 1 labelled cells
Cluster 4 labelled cells
Cluster 3 labelled cells
e
Cluster 6 module expression in:
Cluster 9 module expression in:
Cluster 2 module expression in:
Cluster 1 module expression in:
Cluster 4 module expression in:
Cluster 3 module expression in:
Normalized module expression
f
Positive ESS
Negative ESS
Non-significant
Seurat -log(FDR q-value)
NCLUSION Adjusted PIPs
CD40LG
CD8B
CD8A
KLRD1
FCGR3A
GZMB
CD72
CD19
MS4A1
CD22
LINC00926
TSPAN13
KIAA0125
CD14
g
Both
NCLUSION only
Seurat only
Non-significant
Seurat -log(FDR q-value)
NCLUSION -log(FDR q-value)
UP B cell activation
UP positive regulation of T cell proliferation
UP regulation of IL-12 production
UP positive regulation of T cell activation
UP B cell receptor signaling pathway
UP regulation of IL-12 production
UP T cell activation
UP T cell differentiation
UP antigen receptor-mediated signaling pathway
UP regulation of NK cell cytotoxicity
UP positive regulation of NK cell chemotaxis
UP regulation of NK cell chemotaxis
UP regulation of B cell proliferation
UP regulation of leukocyte degranulation
UP Fc receptor signaling pathway
UP Fc-epsilon receptor signaling pathway
UP Macrophage activation involved in immune response
UP positive regulation of TNF production
UP LPS-mediated signaling pathway
h
Enrichment Analysis
UP regulation of IL-12 production
UP positive regulation of T cell proliferation
UP positive regulation of T cell activation
UP T cell activation
UP T cell differentiation
UP regulation of IL-2 production
UP positive regulation of NK cell chemotaxis
UP regulation of NK cell chemotaxis
UP regulation of NK cell mediated cytotoxicity
UP B cell receptor signaling pathway
UP B cell activation
UP regulation of B cell proliferation
UP Fc-epsilon receptor signaling pathway
UP Fc receptor signaling pathway
UP regulation of leukocyte degranulation
UP LPS-mediated signaling pathway
UP positive regulation of TNF production
UP Macrophage activation in immune response
−log(FDR q-value)

Figure 3. 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)** Heat maps of the adjusted posterior inclusion probabilities (PIPs) (left), effect size sign (ESS) (center), and normalized expression (right) of significant genes in each cluster. Cluster-specific marker genes are selected as those that have a significant inclusion probability and are up-regulated in a given cluster such that PIP $\geq$ 0.5 and ESS = +, respectively. **(d)** Highlighted location on t-SNEs of NCLUSION-inferred clusters that contain predominately one cell type. **(e)** Violin plots comparing the normalized expression of cluster-specific marker genes in each of the inferred clusters. **(f)** Scatter plot comparing the marker genes identified using *post hoc* differential expression analysis with Seurat versus the variable selection approach with NCLUSION. Red points have PIP $\geq$ 0.5 and ESS = +, while blue points have PIP $\geq$ 0.5 and ESS = −, respectively. The vertical dashed line marks the median probability criterion[36], and the horizontal dashed line marks the Bonferroni-corrected threshold for significant q-values. Genes in the top right quadrant are identified by both methods. **(g)** Scatter plot comparing gene ontology (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 (orange) and Seurat (blue), 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.

References

1. Zhen Miao, Benjamin D. Humphreys, Andrew P. McMahon, and Junhyong Kim. Multi-omics integration in the age of million single-cell data. Nature Reviews Nephrology, 17(11):710–724, November 2021. ISSN 1759-5061, 1759-507X. doi: 10.1038/s41581-021-00463-x.

2. F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. Genome Biology, 19(1):15, February 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0.

3. Minzhe Guo, Hui Wang, S. Steven Potter, Jeffrey A. Whitsett, and Yan Xu. Sincera: A pipeline for single-cell rna-seq profiling analysis. PLOS Computational Biology, 11(11):e1004575, November 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004575.

4. Aaron T.L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. F1000Research, 5:2122, October 2016. ISSN 2046-1402. doi: 10.12688/f1000research.9501.2.

5. Ashraful Haque, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. Genome Medicine, 9 (1):75, August 2017. ISSN 1756-994X. doi: 10.1186/s13073-017-0467-4.

6. Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics, 16(33):133–145, March 2015. ISSN 1471-0064. doi: 10.1038/nrg3833.

7. Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. Nature Reviews Genetics, 20(55):273–282, May 2019. ISSN 1471-0064. doi: 10.1038/s41576-018-0088-9.

8. Robert A. Amezquita, Aaron T. L. Lun, Etienne Becht, Vince J. Carey, Lindsay N. Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, Levi Waldron, Hervé Pagès, Mike L. Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo, and Stephanie C. Hicks. Orchestrating single-cell analysis with bioconductor. Nature Methods, 17(22): 137–145, February 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0654-x.

9. Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology, 33:495–502, 2015. doi: 10.1038/nbt.3192. URL https://doi.org/10.1038/nbt.3192.

10. Changde Cheng, John Easton, Celeste Rosencrance, Yan Li, Bensheng Ju, Justin Williams, Heather L Mulder, Yakun Pang, Wenan Chen, and Xiang Chen. Latent cellular analysis robustly reveals subtle diversity in large-scale single-cell rna-seq data. Nucleic Acids Research, 47(22):e143, Dec 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz826.

11. Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. Nature Methods, 14 (44):414–416, Apr 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4207.

12. Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(86):2579–2605, 2008. ISSN 1533-7928.

13. Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [cs, stat].

14. Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. Nature Methods, 15(12):1053–1058, Dec 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0229-2.

15. Vladimir Yu Kiselev, Kristina Kirschner, Michael T. Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N. Natarajan, Wolf Reik, Mauricio Barahona, Anthony R. Green, and Martin Hemberg. Sc3: consensus clustering of single-cell rna-seq data. Nature Methods, 14(55): 483–486, May 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4236.

16. Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger rna sequencing reveals rare intestinal cell types. Nature, 525(75687568):251–255, September 2015. ISSN 1476-4687. doi: 10.1038/nature14966.

17. Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T. Gray, Staci A. Sorensen, Tim Dolbeare, Darren Bertagnolli, Jeff Goldy, Nadiya Shapovalova, Sheana Parry, Changkyu Lee, Kimberly Smith, Amy Bernard, Linda Madisen, Susan M. Sunkin, Michael Hawrylycz, Christof Koch, and Hongkui Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature Neuroscience, 19(22):335–346, February 2016. ISSN 1546-1726. doi: 10.1038/nn.4216.

18. Justina žurauskienė and Christopher Yau. pcareduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics, 17(1):140, March 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0984-y.

19. Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. Science, 347(6226):1138–1142, March 2015. doi: 10.1126/science.aaa1934.

20. Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. Molecular Systems Biology, 15(6):e8746, Jun 2019. ISSN 1744-4292. doi: 10.15252/msb.20188746.

21. David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E. Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M. Keizer, Indu Khatri, Szymon M. Kielbasa, Jan O. Korbel, Alexey M. Kozlov, Tzu-Hao Kuo, Boudewijn P.F. Lelieveldt, Ion I. Mandoiu, John C. Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J. Theis, Huan Yang, Alex Zelikovsky, Alice C. McHardy, Benjamin J. Raphael, Sohrab P. Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. Genome Biology, 21(1):31, Feb 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1926-6.

22. Lucy L. Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. Journal of the American Statistical Association, page 2116331, October 2022. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2022.2116331. Web of Science ID: WOS:000866074100001.

23. Anna Neufeld, Lucy L. Gao, Joshua Popp, Alexis Battle, and Daniela Witten. Inference after latent variable estimation for single-cell rna sequencing data. Biostatistics, December 2022. ISSN 1465-4644, 1468-4357. doi: 10.1093/biostatistics/kxac047. Web of Science ID: WOS:000896827000001.

24. Isabella N. Grabski, Kelly Street, and Rafael A. Irizarry. Significance analysis for clustering with single-cell rna-sequencing data. Nature Methods, 20(88):1196–1202, Aug 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01933-9.

25. Alexis Vandenbon and Diego Diez. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. Nature Communications, 11(11):4318, August 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17900-3.

26. Jesse M Zhang, Govinda M Kamath, and N Tse David. Valid post-clustering differential analysis for single-cell rna-seq. Cell Systems, 9(4):383–392, 2019.

27. Jesse M Zhang, Jue Fan, H Christina Fan, David Rosenfeld, and David N Tse. An interpretable framework for clustering single-cell rna-seq datasets. BMC bioinformatics, 19(1):1–12, 2018.

28. Snehalika Lall, Sumanta Ray, and Sanghamitra Bandyopadhyay. Rgcop-a regularized copula based method for gene selection in single-cell rna-seq data. PLOS Computational Biology, 17 (10):e1009464, 2021.

29. F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. Genome Biology, 20:1–16, 2019.

30. David M. Blei and Michael I. Jordan. Variational inference for dirichlet process mixtures. Bayesian Analysis, 1(1):121–143, Mar 2006. ISSN 1936-0975, 1931-6690. doi: 10.1214/06-BA104.

31. Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian Data Analysis, Third Edition. CRC Press, Hoboken, 2013. ISBN 978-1-4398-9820-8.

32. Sandhya Prabhakaran, Elham Azizi, Ambrose Carr, and Dana Pe'er. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. JMLR Workshop and Conference Proceedings, 48:1070–1079, 2016. ISSN 1938-7288.

33. Zhe Sun, Li Chen, Hongyi Xin, Yale Jiang, Qianhui Huang, Anthony R Cillo, Tracy Tabib, Jay K Kolls, Tullia C Bruno, Robert Lafyatis, et al. A bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. Nature communications, 10(1):1649, 2019.

34. Tiehang Duan, José P Pinto, and Xiaohui Xie. Parallel clustering of single cell transcriptomic data with split-merge sampling on dirichlet process mixtures. Bioinformatics, 35(6):953–961, 2019.

35. Zhe Sun, Ting Wang, Ke Deng, Xiao-Feng Wang, Robert Lafyatis, Ying Ding, Ming Hu, and Wei Chen. Dimm-sc: a dirichlet mixture model for clustering droplet-based single cell transcriptomic data. Bioinformatics, 34(1):139–146, 2018.

36. Maria Maddalena Barbieri and James O. Berger. Optimal predictive model selection. The Annals of Statistics, 32(3):870–897, Jun 2004. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053604000000238.

37. Ping Zeng and Xiang Zhou. Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. Nature Communications, 8(1):456, Dec 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-00470-2.

38. Peter Carbonetto and Matthew Stephens. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Analysis, 7:73–108, Mar 2012. ISSN 1936-0975, 1931-6690. doi: 10.1214/12-BA703.

39. Pinar Demetci, Wei Cheng, Gregory Darnell, Xiang Zhou, Sohini Ramachandran, and Lorin Crawford. Multi-scale inference of genetic trait architecture using biologically annotated neural networks. PLOS Genetics, 17(8):e1009754, 2021. ISSN 1553-7404. doi: 10.1371/journal.pgen.1009754.

40. David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518):859–877, April 2017. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1285773. tex.ids: bleiVariationalInferenceReview2017a arXiv: 1601.00670.

41. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

42. Lingxue Zhu, Jing Lei, Lambertus Klei, Bernie Devlin, and Kathryn Roeder. Semisoft clustering of single-cell data. Proceedings of the National Academy of Sciences, 116(2):466–471, Jan 2019. doi: 10.1073/pnas.1817715116.

43. 10x Genomics. Support: single cell gene expression datasets, 2023. URL `https://www.10xgenomics.com/resources/datasets`.

44. Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. A curated database reveals trends in single-cell transcriptomics. Database: The Journal of Biological Databases and Curation, 2020:baaa073, Nov 2020. ISSN 1758-0463. doi: 10.1093/database/baaa073.

45. Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. Nature Communications, 8(11):14049, Jan 2017. ISSN 2041-1723. doi: 10.1038/ncomms14049.

46. Alexander HS Vargo and Anna C Gilbert. A rank-based marker selection method for high throughput scrna-seq data. BMC bioinformatics, 21(1):1–51, 2020.

47. Lieke Michielsen, Marcel J. T. Reinders, and Ahmed Mahfouz. Hierarchical progressive learning of cell identities in single-cell data. Nature Communications, 12(11):2799, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23196-8.

48. Daniel M. Baume, Michael A. Caligiuri, Thomas J. Manley, John F. Daley, and Jerome Ritz. Differential expression of cd8α and cd8β associated with mhc-restricted and non-mhc-restricted

cytolytic effector cells. Cellular Immunology, 131(2):352–365, December 1990. ISSN 0008-8749. doi: 10.1016/0008-8749(90)90260-X.

49. Haeyoun Choi, Hyun-Joo Lee, Hyun-Jung Sohn, and Tai-Gyu Kim. Cd40 ligand stimulation affects the number and memory phenotypes of human peripheral cd8+ t cells. BMC Immunology, 24(1): 15, June 2023. ISSN 1471-2172. doi: 10.1186/s12865-023-00547-2.

50. Suzanne Norris, Derek G. Doherty, Clive Collins, Gerry McEntee, Oscar Traynor, John E. Hegarty, and Cliona O'Farrelly. 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. Human Immunology, 60(1):20–31, Jan 1999. ISSN 0198-8859. doi: 10.1016/S0198-8859(98)00098-6.

51. Simon N. Willis, Julie Tellier, Yang Liao, Stephanie Trezise, Amanda Light, Kristy O'Donnell, Lee Ann Garrett-Sinha, Wei Shi, David M. Tarlinton, and Stephen L. Nutt. Environmental sensing by mature b cells is controlled by the transcription factors pu.1 and spib. Nature Communications, 8(11):1426, November 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-01605-1.

52. Jonathan Zuccolo, Lili Deng, Tammy Unruh, Ratna Sanyal, Jeremy Bau, Jan Storek, Douglas Demetrick, Joanne Luider, Iwona Auer-Grzesiak, Adnan Mansoor, and Julie Deans. Expression of ms4a and tmem176 genes in human b lymphocytes. Frontiers in Immunology, 4, 2013. ISSN 1664-3224. URL `https://www.frontiersin.org/articles/10.3389/fimmu.2013.00195`.

53. Zheng Chen, Mincheng Yu, Jiuliang Yan, Lei Guo, Bo Zhang, Shuang Liu, Jin Lei, Wentao Zhang, Binghai Zhou, Jie Gao, Zhangfu Yang, Xiaoqiang Li, Jian Zhou, Jia Fan, Qinghai Ye, Hui Li, Yongfeng Xu, and Yongsheng Xiao. 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. Frontiers in Immunology, 12, 2021. ISSN 1664-3224. URL `https://www.frontiersin.org/articles/10.3389/fimmu.2021.647209`.

54. Elza Evren, Emma Ringqvist, Kumar Parijat Tripathi, Natalie Sleiers, Inés Có Rives, Arlisa Alisjahbana, Yu Gao, Dhifaf Sarhan, Tor Halle, Chiara Sorini, Rico Lepzien, Nicole Marquardt,

Jakob Michaëlsson, Anna Smed-Sörensen, Johan Botling, Mikael C.I. Karlsson, Eduardo J. Villablanca, and Tim Willinger. Distinct developmental pathways from blood monocytes generate human lung macrophage diversity. Immunity, 54(2):259–275.e7, February 2021. ISSN 10747613. doi: 10.1016/j.immuni.2020.12.003.

55. Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. Bioinformatics, 39(1):btac757, January 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac757.

56. Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, 21(16):3439–3440, August 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti525.

57. Edward Y. Chen, Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. BMC Bioinformatics, 14:128, April 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-128.

58. Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. Nature Genetics, 25(11):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556.

59. Seth Carbon and Chris Mungall. Gene ontology data archive, Jul 2023. URL `https://zenodo.org/record/8200914`.

60. Eric Vivier, Elena Tomasello, Myriam Baratin, Thierry Walzer, and Sophie Ugolini. Functions of natural killer cells. Nature Immunology, 9(55):503–510, May 2008. ISSN 1529-2916. doi: 10.1038/ni1582.

61. Sourav Paul and Girdhari Lal. The molecular mechanism of natural killer cells function and its importance in cancer immunotherapy. Frontiers in Immunology, 8, 2017. ISSN 1664-3224. URL `https://www.frontiersin.org/articles/10.3389/fimmu.2017.01124`.

62. Baptiste N. Jaeger, Jean Donadieu, Céline Cognet, Claire Bernat, Diana Ordoñez-Rueda, Vincent Barlogis, Nizar Mahlaoui, Aurore Fenis, Emilie Narni-Mancinelli, Blandine Beaupain, Christine Bellanné-Chantelot, Marc Bajénoff, Bernard Malissen, Marie Malissen, Eric Vivier, and Sophie Ugolini. Neutrophil depletion impairs natural killer cell maturation, function, and homeostasis. The Journal of Experimental Medicine, 209(3):565–580, March 2012. ISSN 0022-1007. doi: 10.1084/jem.20111908.

63. Kun Jiang, Bin Zhong, Danielle L. Gilvary, Brian C. Corliss, Elizabeth Hong-Geller, Sheng Wei, and Julie Y. Djeu. Pivotal role of phosphoinositide-3 kinase in regulation of cytotoxicity in natural killer cells. Nature Immunology, 1(55):419–425, November 2000. ISSN 1529-2916. doi: 10.1038/80859.

64. Fanying Meng and Clifford A. Lowell. Lipopolysaccharide (lps)-induced macrophage activation and signal transduction in the absence of src-family kinases hck, fgr, and lyn. The Journal of Experimental Medicine, 185(9):1661–1670, May 1997. ISSN 0022-1007.

65. Mark Baer, Allan Dillner, Richard C. Schwartz, Constance Sedon, Sergei Nedospasov, and Peter F. Johnson. Tumor necrosis factor alpha transcription in macrophages is attenuated by an autocrine factor that preferentially induces nf-κb p50. Molecular and Cellular Biology, 18(10):5678–5689, Oct 1998. ISSN 0270-7306.

66. C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. K. Howlett, O. Suchanek, K. Polanski, H. W. King, L. Mamanova, N. Huang, P. A. Szabo, L. Richardson, L. Bolt, E. S. Fasouli, K. T. Mahbubani, M. Prete, L. Tuck, N. Richoz, Z. K. Tuong, L. Campos, H. S. Mousa, E. J. Needham, S. Pritchard, T. Li, R. Elmentaite, J. Park, E. Rahmani, D. Chen, D. K. Menon, O. A. Bayraktar, L. K. James, K. B. Meyer, N. Yosef, M. R. Clatworthy, P. A. Sims, D. L. Farber, K. Saeb-Parsy, J. L. Jones, and S. A. Teichmann. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science, 376(6594):eabl5197, May 2022. doi: 10.1126/science.abl5197.

67. Merav Cohen, Amir Giladi, Anna-Dorothea Gorki, Dikla Gelbard Solodkin, Mor Zada, Anastasiya Hladik, Andras Miklosi, Tomer-Meir Salame, Keren Bahar Halpern, Eyal David, Shalev Itzkovitz, Tibor Harkany, Sylvia Knapp, and Ido Amit. Lung single-cell signaling interaction map reveals ba-

sophil role in macrophage imprinting. Cell, 175(4):1031–1044.e18, November 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.09.009.

68. Patrick Neubert, Arne Homann, David Wendelborn, Anna-Lorena Bär, Luka Krampert, Maximilian Trum, Agnes Schröder, Stefan Ebner, Andrea Weichselbaum, Valentin Schatz, Peter Linz, Roland Veelken, Jonas Schulte-Schrepping, Anna C. Aschenbrenner, Thomas Quast, Christian Kurts, Sabrina Geisberger, Karl Kunzelmann, Karin Hammer, Katrina J. Binger, Jens Titze, Dominik N. Müller, Waldemar Kolanus, Joachim L. Schultze, Stefan Wagner, and Jonathan Jantsch. Ncx1 represents an ionic na+ sensing mechanism in macrophages. PLoS Biology, 18(6):e3000722, June 2020. ISSN 1544-9173. doi: 10.1371/journal.pbio.3000722.

69. D. Gibeon, J. Zhu, A. Sogbesan, W. Banya, C. Rossios, J. Saito, J. P. Rocha, J. H. Hull, A. N. Menzies-Gow, P. K. Bhavsar, and K. F. Chung. Lipid-laden bronchoalveolar macrophages in asthma and chronic cough. Respiratory Medicine, 108(1):71–77, January 2014. ISSN 0954-6111. doi: 10.1016/j.rmed.2013.10.005.

70. A. I. Jaiswal, C. Dubey, S. L. Swain, and M. Croft. Regulation of cd40 ligand expression on naive cd4 t cells: a role for tcr but not co-stimulatory signals. International Immunology, 8(2):275–285, February 1996. ISSN 0953-8178. doi: 10.1093/intimm/8.2.275.

71. Jayati Mookerjee-Basu and Dietmar J. Kappes. New ingredients for brewing cd4+t (cells): Tcf-1 and lef-1. Nature Immunology, 15(77):593–594, July 2014. ISSN 1529-2916. doi: 10.1038/ni.2927.

72. Dimitrios V. Vavoulis, Margherita Francescatto, Peter Heutink, and Julian Gough. Dgeclust: differential expression analysis of clustered count data. Genome Biology, 16(1):39, February 2015. ISSN 1465-6906. doi: 10.1186/s13059-015-0604-6.

73. Charlotte Soneson and Mark D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. Nature Methods, 15(44):255–261, April 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4612.

74. Xiang Zhu and Matthew Stephens. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nature Communications, 9(1): 4361, 2018.

75. Ying Ma, Shiquan Sun, Xuequn Shang, Evan T Keller, Mengjie Chen, and Xiang Zhou. Integrative differential expression and gene set enrichment analysis using summary statistics for scrna-seq studies. Nature Communications, 11(1):1585, 2020.

76. David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518):859–877, 2017.

77. Ryan Giordano, Tamara Broderick, and Michael I Jordan. Covariances, robustness and variational bayes. Journal of Machine Learning Research, 19(51), 2018.

78. Cheng Zhang, Babak Shahbaba, and Hongkai Zhao. Variational hamiltonian monte carlo via score matching. Bayesian Analysis, 13(2):485, 2018.

79. Lukas M Weber, Arkajyoti Saha, Abhirup Datta, Kasper D Hansen, and Stephanie C Hicks. nnsvg for the scalable identification of spatially variable genes using nearest-neighbor gaussian processes. Nature Communications, 14(1):4059, 2023.

80. Brian J Reich and Howard D Bondell. A spatial dirichlet process mixture model for clustering population genetics data. Biometrics, 67(2):381–390, 2011.

81. Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals Applied Statistics, 5(3):1780–1815, 2011. doi: 10.1214/11-AOAS455. URL `https://projecteuclid.org:443/euclid.aoas/1318514285`.

82. Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with Bayesian sparse linear mixed models. PLOS Genetics, 9(2):e1003264, 2013.

83. Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. The Annals Applied Statistics, 11(3):1561–1592, 2017. doi: 10.1214/17-AOAS1046. URL `https://projecteuclid.org:443/euclid.aoas/1507168840`.

84. Michael C Hughes, Dae Il Kim, and Erik B Sudderth. Reliable and scalable variational inference for the hierarchical dirichlet process. Artificial Intelligence and Statistics, page 9, 2015.

85. Jacob Cohen. Statistical Power Analysis for the Behavioral Sciences. Academic press, 2013.

86. Matthew P. Wand, John T. Ormerod, Simone A. Padoan, and Rudolf Frühwirth. Mean field variational bayes for elaborate distributions. Bayesian Analysis, 6(4):847–900, 2011. doi: 10.1214/11-BA631. URL https://doi.org/10.1214/11-BA631.

87. Valentine Svensson. Droplet scrna-seq is not zero-inflated. Nature Biotechnology, 38(2):147–150, Feb 2020. ISSN 1546-1696. doi: 10.1038/s41587-019-0379-5.

88. Valentine Svensson. Reply to: Umi or not umi, that is the question for scrna-seq zero-inflation. Nature Biotechnology, 39(2):160–160, Feb 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-00811-5.

89. F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. Genome Biology, 19(1):15, Feb 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0.

90. Lijia Yu, Yue Cao, Jean Y. H. Yang, and Pengyi Yang. Benchmarking clustering algorithms on estimating the number of cell types from single-cell rna-sequencing data. Genome Biology, 23(1):49, Feb 2022. ISSN 1474-760X. doi: 10.1186/s13059-022-02622-0.

91. Silke Wagner and Dorothea Wagner. Comparing clusterings - an overview, 2007. URL https://publikationen.bibliothek.kit.edu/1000011477.

92. Edward H Simpson. Measurement of diversity. nature, 163(4148):688–688, 1949.

93. Robert P McIntosh. An index of diversity and the relation of certain concepts to diversity. Ecology, 48(3):392–404, 1967.

94. Robert K Peet. The measurement of species diversity. Annual review of ecology and systematics, 5(1):285–307, 1974.

95. Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1073–1080, 2009.

96. Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. Bioinformatics, 39(1):btac757, Jan 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac757.