A Bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies

Zhe Sun, Li Chen, Hongyi Xin, Yale Jiang, Qianhui Huang, Anthony R. Cillo, Tracy Tabib, Jay K. Kolls, Tullia C. Bruno, Robert Lafyatis, Dario A. Vignali, Kong Chen, Ying Ding, Ming Hu & Wei Chen.

The recently developed droplet-based single-cell transcriptome sequencing (scRNA-seq) technology makes it feasible to perform a population-scale scRNA-seq study, in which the transcriptome is measured for tens of thousands of single cells from multiple individuals. Despite the advances of many clustering methods, there are few tailored methods for population-scale scRNA-seq studies. Here, we develop a Bayesian mixture model for single-cell sequencing (BAMM-SC) method to cluster scRNA-seq data from multiple individuals simultaneously. BAMM-SC takes raw count data as input and accounts for data heterogeneity and batch effect among multiple individuals in a unified Bayesian hierarchical model framework. Results from extensive simulation studies and applications of BAMM-SC to in-house experimental scRNA-seq datasets using blood, lung and skin cells from humans or mice demonstrate that BAMM-SC outperformed existing clustering methods with considerable improved clustering accuracy, particularly in the presence of heterogeneity among individuals.
Single-cell RNA sequencing (scRNA-seq) technologies have been widely used to measure gene expression for each individual cell, facilitating a deeper understanding of cell heterogeneity and better characterization of rare cell types. Compared to early generation scRNA-seq technologies, the recently developed droplet-based technology, largely represented by the 10x Genomics Chromium system, has quickly gained popularity because of its high throughput (tens of thousands of single cells per run), high efficiency (a couple of days), and relatively lower cost (<$1 per cell). It is now feasible to conduct population-scale single-cell transcriptomic profiling studies, where several to tens or even hundreds of individuals are sequenced.

A major task of analyzing droplet-based scRNA-seq data is to identify clusters of single cells with similar transcriptomic profiles. To achieve this goal, classic unsupervised clustering methods such as K-means clustering, hierarchical clustering, and density-based clustering approaches can be applied after some normalization steps. Recently, scRNA-seq tailored unsupervised methods, such as SIMLR, CellTree, SC3, TSCAN, and DIMM-SC, have been designed and proposed for clustering scRNA-seq data. Supervised methods, such as MetaNeighbor, have been proposed to assess how well cell-type-specific transcriptional profiles replicate across different datasets. However, none of these methods explicitly considers the heterogeneity among multiple individuals from population studies. In a typical analysis of population-scale scRNA-seq data, reads from each individual are processed separately and then merged together for the downstream analysis. For example, in the 10x Genomics Cell Ranger pipeline, to aggregate multiple libraries, reads from different libraries are downsampled such that all libraries have the same sequencing depth, leading to substantial information loss for individuals with higher sequencing depth. Alternatively, reads can be naively merged across all individuals without any library adjustment, leading to batch effects and unreliable clustering results.

Similar to the analysis of other omics data, several computational approaches have been proposed to correct batch effects for scRNA-seq data. For example, Spitzer et al. adapted the concept of force-directed graph to visualize complex cellular samples via Scaffold (single-cell analysis by fixed force- and landmark-directed) maps, which can overlay data from multiple samples onto a reference sample(s). Recently, two new methods: mutual nearest neighbors (MNN) (implemented in scran) and canonical correlation analysis (CCA) (implemented in Seurat) were published for batch correction of scRNA-seq data. All these methods require the raw counts to be transformed to continuous values under different assumptions, which may alter the data structure in some cell types and lead to difficulty of biological interpretation.

We first conducted an exploratory data analysis to demonstrate the existence of batch effect in multiple individuals using both publicly available and three in-house synthetic droplet-based scRNA-seq datasets, including human peripheral blood mononuclear cells (PBMC), mouse lung and human skin tissues. Detailed sample information was summarized in Fig. 1a and Supplementary Table 1. We use human PBMC as an example. We isolated from whole blood obtained from 4 healthy donors and used the 10x Chromium system to generate scRNA-seq data. We also included one additional healthy donor from a published PBMC scRNA-seq data to mimic the scenario where we combine these 5 donors and observed a clear batch effect: samples from the same batch tend to cluster together.

This illustrative example demonstrates the importance and urgent need for well characterizing different sources of variability and correcting potential batch effects among droplet-based scRNA-seq datasets collected from multiple individuals. In addition, due to the computational burden, many methods cannot be scaled up to analyze population-scale droplet-based scRNA-seq data with tens of thousands of cells collected from many individuals under various conditions. In this study, we propose a BAyesian Mixture Model for Single Cell sequencing (BAMM-SC) to simultaneously cluster large-scale droplet-based scRNA-seq data from multiple individuals. BAMM-SC directly works on the raw counts without any data transformation and models the heterogeneity from multiple sources by learning the distributions of signature genes in a Bayesian hierarchical model framework. In the following sections, we will describe our method, benchmark its performance against existing clustering methods in simulation studies, and evaluate our method for its accuracy, stability, and efficiency in three in-house synthetic
specifically, we simulated droplet-based scRNA-seq data collected from multiple individuals from the posited Bayesian hierarchical Dirichlet multinomial mixture model (see Methods and Supplementary Methods). We considered different experimental designs, including different heterogeneities among multiple individuals and different numbers of individuals (Fig. 2). In our posited hierarchical model, the log normal prior distribution \( \ln(\mu_{ik}, \sigma_{ik}^2) \) measures the heterogeneity of gene \( i \) in cell type \( k \) among multiple individuals, where \( \mu_{ik} \) and \( \sigma_{ik} \) are related to the mean and variation of gene expression. Without loss of generality, we used the mean of \( \sigma_{ik}^2 \) across all genes and all cell types to quantify the overall individual level heterogeneity. We applied BAMM-SC to each synthetic dataset, and compared the inferred cell type label of each single cell with the ground truth, measured by adjusted Rand index (ARI) \(^{18}\). We compared BAMM-SC with other competing clustering methods (K-means, TSCAN, SC3, and Seurat), which are either methods from different clustering categories or recommended by recent reviews on clustering methods for single-cell data\(^{19,20}\). Since none of methods model batch effects and therefore each needs to be combined with a batch correction method as a preprocessing step in data analysis. We applied two recently published and prevalent methods scran MNN\(^ {15}\) and Seurat CCA\(^ {17}\) prior to these clustering methods so that each combination can be a fair comparison with BAMM-SC, which does not need a separate batch correction step.

Specifically, we compared BAMM-SC with the other nine competing methods (MNN+K-means, MNN+TSCAN, MNN+SC3, MNN+Seurat, CCA+K-means, CCA+TSCAN, CCA+SC3, CCA+Seurat, and DIMM-SC) in the simulation studies. Noticeably, DIMM-SC, our previously developed method for clustering scRNA-seq data from a single individual, also takes the raw UMI count matrix as the input without any batch effect correction or data transformation. We pooled single cells from different individuals together while ignoring each individual label, and then applied DIMM-SC to the pooled data. We simulated 100 datasets and summarized the corresponding ARIs for each method.

As shown in Fig. 2a, BAMM-SC consistently outperformed the other nine competing methods across a variety of individual level heterogeneities by achieving higher average ARI and lower variation of ARI among 100 simulations. As expected, the
performance of all ten clustering approaches decreases as the among individual heterogeneity increases, measured by the mean $\sigma^2$ values. In Fig. 2b, with the increase of number of individuals, BAMM-SC achieved higher ARI, while ARIs of other methods either remained stable or decreased.

Furthermore, we performed comprehensive simulation studies by generating simulated scRNA-seq datasets from different number of cell type clusters (Supplementary Fig. 2a), different overall sequencing depths (Supplementary Fig. 2b), and different cell-type-specific heterogeneities (i.e., the mean difference of gene expression profiles between two distinct cell types) (Supplementary Fig. 2c). BAMM-SC consistently outperformed other methods in terms of accuracy and robustness in all these scenarios. Taken together, our comprehensive simulation studies have demonstrated that, when data are generated from the true model, BAMM-SC is able to appropriately borrow information across multiple individuals, account for unbalanced sequencing depths, and provide more accurate and robust clustering results than other competing methods.

To evaluate the robustness of BAMM-SC when data generation model is mis-specified, we simulated additional datasets using R package Splatter21, a commonly used tool for scRNA-seq data simulation using a completely different model. To make our simulated data a good approximation to the real data, we first downloaded the raw UMI count matrix of a purified B-cell scRNA-seq dataset from the 10x Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/b_cells), and used the function splatEstimate to estimate the parameters related to mean of gene, library size, expression outlier, dispersion across genes, and dropout rate. We assumed cell types are shared across multiple individuals, where each individual is treated as one batch with the same number of cells and genes. We further specified batch parameters and differential expression parameters to generate scenarios with different amount of group effect (i.e., cell type differences) and batch effect. As shown in Fig. 3, BAMM-SC still outperformed most other competing methods in terms of clustering accuracy in all scenarios, although the improvement is less substantial than our own model simulations, which is expected.

**Real data analysis on human PBMC dataset.** For aforementioned human PBMC samples, we first pooled cells from five donors together, filtered lowly expressed genes that were expressed in less than 1% cells. We then extracted the top 1000 highly variable genes based on their standard deviations. As shown in Supplementary Fig. 3, we identified seven types of PBMGs based on the biological knowledge of cell-type-specific gene markers (Supplementary Table 2). Using these gene markers, >70% single cells can be assigned to a specific cell type. Since there is no gold standard for clustering analysis in this real dataset, we used the labels of these cells as the approximated ground truth to benchmark the clustering performance for different clustering methods. Cells with uncertain cell types were removed when calculating ARIs.

Similar to the simulation studies, we applied ten clustering methods on these samples and repeated each method ten times to evaluate the stability of its performance (Table 1). The total number of clusters was set as seven based on the biological knowledge from cell-type-specific gene markers. As shown in Table 1, BAMM-SC achieved the highest ARI for human PBMC samples compared to all other competing methods. Both TSCAN and Seurat are deterministic clustering methods and therefore they generate identical results for ten analyses.

We further generated t-SNE plots with each cell colored by their cell-type classification based on specific gene markers (i.e., the approximated truth) (Fig. 4a (left figure)) and cluster labels inferred by BAMM-SC (Fig. 4a (middle figure)), respectively. Despite some dendritic cells were wrongly identified as CD16+Monocytes, these two plots are similar to each other (ARI = 0.532), suggesting that BAMM-SC performed well in human PBMC samples compared with other clustering methods.

Moreover, we calculated the averaged cell proportions of each cell type inferred from BAMM-SC among ten runs for five PBMC samples, compared with cell proportions calculated from the approximated truth based on gene markers. Figure 4a (right figure) shows that the proportions inferred from BAMM-SC are close to the truth, suggesting that BAMM-SC can adequately account for batch effect when clustering cells from multiple individuals. We also generated t-SNE projection plots colored by
Table 1 Performance of clustering across ten times analyses for three real datasets

| Method          | Mean_P | SD_P      | Range_P | Mean_L | SD_L      | Range_L | Mean_S | SD_S      | Range_S |
|-----------------|--------|-----------|---------|--------|-----------|---------|--------|-----------|---------|
| MNN + K-means   | 0.379  | 0.083     | (0.283-0.485) | 0.662  | 0.066     | (0.596-0.815) | 0.597  | 0.075     | (0.461-0.676) |
| MNN + TSCAN     | 0.373  | NA        | NA      | 0.720  | NA        | NA      | 0.553  | NA        | NA      |
| MNN + Sc3       | 0.348  | 0.084     | (0.266-0.511) | 0.640  | 0.061     | (0.556-0.687) | 0.517  | 0.034     | (0.436-0.557) |
| CCA + K-means   | 0.414  | 0.056     | (0.307-0.464) | 0.695  | 0.114     | (0.505-0.883) | 0.619  | 0.129     | (0.424-0.737) |
| CCA + TSCAN     | 0.210  | NA        | NA      | 0.611  | NA        | NA      | 0.398  | NA        | NA      |
| CCA + Sc3       | 0.145  | 0.052     | (0.051-0.215) | 0.510  | 0.068     | (0.431-0.708) | 0.319  | 0.071     | (0.227-0.488) |
| DIMM-SC         | 0.333  | 0.071     | (0.302-0.529) | 0.809  | 0.030     | (0.742-0.868) | 0.275  | 0.045     | (0.671-0.779) |
| BAMM-SC         | 0.487  | 0.056     | (0.362-0.532) | 0.882  | 0.042     | (0.764-0.910) | 0.762  | 0.032     | (0.717-0.843) |

Columns Mean_P, SD_P, and Range_P were calculated from human PBMC dataset. Columns Mean_L, SD_L, and Range_L were calculated from mouse lung dataset. Columns Mean_S, SD_S, and Range_S were calculated from human skin dataset.

Fig. 4 The performance of BAMM-SC clustering for three in-house scRNA-seq datasets. The t-SNE projection of cells (colored by the approximated truth and BAMM-SC clustering results) and bar plots of proportions of cell types among all individuals for a human PBMC, b mouse lung, and c human skin tissues, separately. BAMM-SC clustering labels are from the result with the highest ARI among ten times analysis.
cluster labels inferred by other methods: MNN+K-means clustering (Supplementary Fig. 4a), MNN+TSCAN (Supplementary Fig. 4b), MNN+SC3 (Supplementary Fig. 4c), MNN+Seurat (Supplementary Fig. 4d), CCA+K-means (Supplementary Fig. 4e), CCA+TSCAN (Supplementary Fig. 4f), CCA+SC3 (Supplementary Fig. 4g), CCA+Seurat (Supplementary Fig. 4h), and DIMM-SC (Supplementary Fig. 4i).

Real data analysis on mouse lung dataset. We collected lung mononuclear cells from four mouse samples under two conditions: Streptococcus pneumonia (SP) infected (sample 1 and 2) and naive (sample 3 and 4). Supplementary Figure 5 shows the t-SNE plot of lung mononuclear cells from four mouse samples. Similar to the analysis of PBMC samples, after filtering lowly expressed genes, we pooled cells from 4 mice together and extracted the top 1000 highly variable genes. As shown in Supplementary Fig. 6, we identified six types of cells based on the biological knowledge of cell-type specific gene markers (Supplementary Table 3). Taken together, >66% of single cells can be assigned to a specific cell type. Therefore, we used the labels of these cells as the approximated truth and removed cells with uncertain cell types from the downstream analysis.

Figure 4b (left figure) and Fig. 4b (middle figure) show the t-SNE plots with each cell colored by their cluster label based on cell-type specific gene markers and cluster labels inferred by BAMM-SC, respectively. These two are highly similar (ARI = 0.910), indicating the outstanding performance of BAMM-SC. Table 1 shows that BAMM-SC considerably outperformed other nine clustering methods in terms of ARI. We also generated t-SNE plots colored by cluster labels inferred by other competing clustering methods (Supplementary Fig. 7). As shown in Supplementary Fig. 8, the proportions of neutrophils in SP infected samples (sample 1 and sample 2) are much higher than the proportions in naive samples (sample 3 and sample 4). This is consistent with the fact that infections by bacteria and viruses may increase the number of neutrophils, which is a necessary cell type in infected samples (sample 1 and sample 2) but not in naive samples. The proportions in naive samples (sample 3 is different from others, which may due to unsatisfactory sample quality or unexpected bacterial infections.

Real data analysis on human skin dataset. To evaluate the clustering performance of BAMM-SC in solid human tissues, we collected skin samples from five healthy donors that are part of a systemic sclerosis study. Figure 1a and Supplementary Table 1 list the detailed sample information and Supplementary Fig. 9 shows the t-SNE plot of cells from five human skin samples after the data processing similar to previous analyses. As shown in Supplementary Fig. 10, we identified eight major types of cells based on the biological knowledge of cell-type specific gene markers (Supplementary Table 4). Taken together, >67% of single cells can be assigned to a specific cell type. Similar to the other two real data analyses, we used the labels of these cells as the approximated truth and removed cells with uncertain cell types from the downstream analysis.

As shown in Fig. 3c, BAMM-SC performed well in human skin samples, since the t-SNE plot with each cell colored by their cell-type label based on gene markers is highly similar to the plot generated from the clustering result of BAMM-SC (ARI = 0.843). Also, BAMM-SC achieved higher ARI compared with all the other clustering methods (Table 1). As comparisons, we generated t-SNE plots colored by cluster labels inferred by different clustering (Supplementary Fig. 11).

Other evaluation criteria. To further demonstrate the validity of BAMM-SC, we calculated the confusion matrix for three real datasets and reported the clustering accuracy (defined as the proportion of cells being classified into the correct cell-type cluster) (Supplementary Table 5, Supplementary Methods). Our method outperformed other competing methods in all three datasets. In addition, we performed a flow cytometry experiment, a gold standard method for quantifying cell population through cell surface markers, on the sample 3 from the human PBMC dataset, which has an additional aliquot from the same pool of cells. We used FlowJo software to gate each cell population through specific antibodies and calculated the percentage of each cell type. Then, we compared the proportions of different cell types from flow cytometry and the clustering result of BAMM-SC from scRNA-seq. Supplementary Figure 12 shows that the proportion of cells in each cell type classified by BAMM-SC is consistent with that being estimated by flow cytometry. We also calculated the Pearson’s correlation coefficient of cell proportions for each clustering method (Supplementary Table 6). Despite the different technology, the high correlation (Pearson correlation coefficient is 0.98) suggests that BAMM-SC is able to adequately account for heterogeneity among multiple individuals and provide reliable clustering results. To be noted, unlike some deterministic clustering methods we considered, Seurat cannot directly pre-specify the number of clusters K. Rather it needs to set a resolution parameter that indirectly controls the cluster number. In all three real data sets, after an extensive grid search, we found the resolution parameter that yields the same number of clusters as the one based on the biological knowledge. Therefore, for the two Seurat clustering methods, instead of using the clustering assignments that produced the highest ARI among ten times analysis, we computed the confusion matrix and the proportions of different cell types based on this specific resolution parameter.

It is challenging to evaluate clustering algorithms in experimental data since the ground truth of cell type label is generally unknown. Other than using ARI based on cell-type specific gene markers as approximated ground truth, we also used cluster stability and tightness to evaluate the clustering performance. Specifically, we calculated the average proportion of nonoverlap (APN)25 clustered cells and silhouette width26 in three real datasets, respectively. APN is a cluster stability measurement which evaluates the stability of a clustering result by comparing it with the clusters obtained by removing one feature (i.e., one gene in our study) at a time. It measures the average proportion of observations not placed in the same cluster under both cases. To make computation affordable in our real data analysis, after extracting the top 1000 highly variable genes, we compared the clustering results based on the full data (1000 genes) to the clustering results based on a subset of data with 100 genes randomly removed. We repeated this step ten times to calculate the APN. For cluster tightness, the silhouette width ranges from -1 to 1, where a higher value indicates that the observation is better matched to its own cluster and worse matched to other clusters. For both measurements, BAMM-SC achieved high cluster stability and high cluster tightness in most scenarios, compared with all other competing methods (Supplementary Table 7, Supplementary Table 8).

Uncertainty assessment. Different from other deterministic methods, BAMM-SC has the ability to assess clustering uncertainty through the posterior probability for each cell to belong to each cell-type cluster. As shown in Supplementary Fig. 13, we highlighted vague cells in the t-SNE projection plot, where vague cells are defined as cells with the largest posterior cluster-specific probability <0.95. In the human PBMC samples, most of the vague cells (colored in red) are located at the boundary of different clusters, which reassuring the validity of the clustering
results. In real data analysis, users can decide to remove vague cells under a user-specified criterion (based on the posterior probability) for the downstream analysis such as differential gene expression analysis within each cell type.

Discussion
In summary, we have developed a novel Bayesian framework for clustering population-scale scRNA-seq data. BAMM-SC retains the raw data information by directly modeling UMI counts without data transformation or normalization, facilitating straightforward biological interpretation. The Bayesian hierarchical model enables the joint characterization of multiple sources of uncertainty, including single-cell level heterogeneity and individual level heterogeneity. Furthermore, BAMM-SC can borrow information across different individuals through its mixture hierarchical model structure and Bayesian computational techniques, leading to improved clustering accuracy. BAMM-SC is based on probabilistic models, thus providing the quantification of clustering uncertainty for each single cell.

Our model coupled with a computationally efficient MCMC algorithm is able to cluster large-scale droplet-based scRNA-seq data with feasible computational cost. For example, using 1000 highly variable genes, it takes about 1.5, 2.5, and 4.5 h when analyzing the 3 real datasets (human PBMC, mouse lung and human skin), respectively. For the simulated dataset consists of 10 individuals with 4000 cells each, the computational time for clustering is about 30 min. Supplementary Figure 14 demonstrates that the computational time of BAMM-SC increases approximately linearly with the increase of the number of cells in each individual, the number of individuals and the number of clusters, respectively. To further improve the computational efficiency, we provided a supervised clustering option in BAMM-SC for very large-scale datasets. Specifically, users can first apply BAMM-SC on a small subset of single cells in each individual, and save predicted cluster labels as well as other informative parameters such as $\alpha$. Then for the remaining single cells, users can perform supervised classification via BAMM-SC instead of unsupervised clustering (see Methods). By clustering a small number of single cells, this procedure will substantially reduce the computational cost. We used the simulated dataset of ten individuals with 4000 cells each, and the computational time for clustering is about 30 min. Supplementary Figure 14 demonstrates that the computational time of BAMM-SC increases approximately linearly with the increase of the number of cells in each individual, the number of individuals and the number of clusters, respectively. To further improve the computational efficiency, we provided a supervised clustering option in BAMM-SC for very large-scale datasets. Specifically, users can first apply BAMM-SC on a small subset of single cells in each individual, and save predicted cluster labels as well as other informative parameters such as $\alpha$. Then for the remaining single cells, users can perform supervised classification via BAMM-SC instead of unsupervised clustering (see Methods). By clustering a small number of single cells, this procedure will substantially reduce the computational cost. We used the simulated dataset of ten individuals with 4000 cells each, and the computational time for clustering is about 30 min. Supplementary Figure 14 demonstrates that the computational time of BAMM-SC increases approximately linearly with the increase of the number of cells in each individual, the number of individuals and the number of clusters, respectively. To further improve the computational efficiency, we provided a supervised clustering option in BAMM-SC for very large-scale datasets. Specifically, users can first apply BAMM-SC on a small subset of single cells in each individual, and save predicted cluster labels as well as other informative parameters such as $\alpha$. Then for the remaining single cells, users can perform supervised classification via BAMM-SC instead of unsupervised clustering (see Methods). By clustering a small number of single cells, this procedure will substantially reduce the computational cost. We used the simulated dataset of ten individuals with 4000 cells each, and the computational time for clustering is about 30 min. Supplementary Figure 14 demonstrates that the computational time of BAMM-SC increases approximately linearly with the increase of the number of cells in each individual, the number of individuals and the number of clusters, respectively. To further improve the computational efficiency, we provided a supervised clustering option in BAMM-SC for very large-scale datasets.

In addition, we can predefine the number of clusters based on prior knowledge on the tissue or determine it using some standard model checking criterion such as Akaike’s Information Criteria (AIC) or Bayesian Information Criteria (BIC). As shown in Supplementary Fig. 15, AIC and BIC work as expected in the analysis of simulated datasets and provide a reliable range of cluster numbers to guide real data analysis based on prior knowledge. However, in a biological study, the number of clusters is often considered as a continuum because of the nature of cell growth, so we recommend trying a range of cluster numbers in practice. BAMM-SC is shown to be robust against model misspecification. In our simulation studies, we applied Splatter to simulate scRNA-seq data in which the data generation mechanism is different from our proposed BAMM-SC model. BAMM-SC still achieved higher clustering accuracy than other competing methods. In addition, we compared BAMM-SC with other clustering methods when the number of clusters is different from the true number of cell types. Supplementary Fig. 16 shows that BAMM-SC still achieved the highest ARI in most scenarios.

Other than MNN and CCA, several other approaches have been proposed to correct batch effect across multiple individuals. One straightforward approach is taking one individual as the reference, producing a low-dimensional embedding of it and then projecting the other individuals onto that embedding. To perform low-dimensional embedding, diffusion map is known as a tool for non-linear dimension reduction and has recently been adapted for the visualization of single-cell gene-expression data. In addition, single-cell variational inference (scVI) is a scalable framework for batch correction based on variational inference and stochastic optimization of deep neural networks. The performance of diffusion map and scVI combined with other clustering method was examined, which is worse than MNN and CCA in the three synthetic datasets (possibly due to unmet model underlying assumptions). We will explore more emerging methods in our future work.

There are several limitations of BAMM-SC. First, we filtered out genes with excessive zeros from the analysis under the assumption that lowly-expressed genes do not contribute much to clustering. This may be problematic for rare cell type
identification. Second, we do not explicitly model a zero-inflation pattern, which may or may not affect clustering accuracy. A refined model that can handle inflated zeros can be further developed with a balance between computational complexity and model flexibility. Third, in our model, we assume that each cell belongs to one distinct cluster. The posterior probability measures the clustering uncertainty, which cannot be directly interpreted as a quantification of cell cycle or developmental stage. Finally, although our supervised strategy is proven to work for large datasets efficiently, it may potentially miss some rare clusters.

Our method has the potential to be extended to perform trajectory analysis \(^29,30\), and accounts for both individual and batch level heterogeneity (e.g., two individuals spread evenly across two 10x chips in a properly blocked design) by adding another level of structure. In addition, the model parameters can be used for downstream differential gene expression analysis or construct cell-type specific biomarker panels. These interesting directions are beyond the scope of this paper and will be studied in future papers. Additionally, unlike the traditional way of analyzing scRNA-seq data, BAMM-SC can be also used with batch effect correction. As shown in Supplementary Fig. 17, we ran BAMM-SC on the mouse lung dataset first and extracted cells in cluster 4. Then we applied CCA (implemented in Seurat) on this specific cluster of cells and replotted the t-SNE plot. From Supplementary Fig. 17e, cells from different samples are superimposed on each other, suggesting that most batch effect has been removed. In practice, we recommend using BAMM-SC for clustering raw count data and then use other methods, such as MNM and CCA, to remove batch effect for each individual cell type if needed.

We have applied BAMM-SC to simulated datasets and three in-house synthetic datasets to showcase its performance on different tissue types and species. With the increased popularity of population-based scRNA-seq studies, BAMM-SC will become a powerful tool for elucidating single cell level transcriptional heterogeneity from population-based studies and a complementary approach to existing clustering methods.

**Methods**

**Statistical model.** We propose a Bayesian hierarchical Dirichlet multinomial mixture model to explicitly characterize different sources of variability in population scale scRNA-seq data. Specifically, let \( x_{jl} \) represent the number of unique UMIs for gene \( i \) in cell \( j \) from individual \( l \) (1 \( \leq l \leq L \), 1 \( \leq j \leq G_l \), 1 \( \leq l \leq L \)). Here, \( G_l \), \( C_l \), and \( L \) denote the total number of genes, cells (in individual \( l \)), and individuals, respectively. Our goal is to perform simultaneous clustering for cells from all \( L \) individuals. We assume that within each individual, all single cells consist of \( K \) distinct cell types. Cell type clusters are discrete, and each cell belongs to one cell type exclusively. Here, \( K \) is predefined according to prior biological knowledge, or will be estimated from the data, and \( K \) is the same among all \( L \) individuals.

Assume that \( x_{jl} = (x_{j1l}, x_{j2l}, \ldots, x_{jG_l}) \), the gene expression for cell \( j \) in individual \( l \), follows a multinomial distribution \( \left( T_j, P_j, P_{jG_l} \right) \). Here, \( T_j = \sum_{i=1}^{G_l} x_{jil} \) is the total number of UMIs, \( P_{jG_l} = (p_{j1}, p_{j2}, \ldots, p_{jG_l}) \) is the probability vector for gene expression with \( \sum_{j=1}^{G_l} p_{jG_l} = 1 \), (where larger \( p_{jG_l} \) is associated with more UMI counts \( x_{jil} \)). In addition, let \( x_{jl} \in \{1, 2, \ldots, K\} \) indicate the cell type label for cell \( j \) in individual \( l \), where \( x_{jl} = k \) indicates that cell \( j \) in individual \( l \) belongs to cell type \( k \). Cells of the same cell type share a similar gene-expression pattern. If cell \( j \) in individual \( l \) belongs to cell type \( k \), \( x_{jl} = k \), we assume that \( p_{jk} \) follows a cell-type specific Dirichlet prior \( D(k) \), where \( \alpha_{ik} = (\alpha_{i1k}, \alpha_{i2k}, \ldots, \alpha_{iG_l}) \) is the Dirichlet prior parameter for cell type \( k \) in individual \( l \). Then, the joint distribution of \( x_{jl} \) and \( z_{jl} \) is

\[
P(x_{jl}, z_{jl} | \alpha_k) = \frac{G_l}{\Gamma(z_{jl}+\alpha_{jk})} \prod_{i=1}^{G_l} \frac{\Gamma(x_{jil}+\alpha_{iik})}{\Gamma(\alpha_{iik})} (\Gamma(\alpha_{jk})/\Gamma(z_{jl}+\alpha_{jk}))
\]

(3)

where \( |\alpha_k| = G_l \sum_{i=1}^{G_l} \alpha_{iik} \). The joint distribution of \( x_{jl} \) and \( z_{jl} \) is

\[
P(\mathbf{x}_{jl}, \mathbf{z}_{jl} | \mathbf{\alpha}_k) = \frac{G_l!}{\prod_{i=1}^{G_l} x_{jil}!} \frac{1}{\prod_{i=1}^{G_l} \Gamma(\alpha_{iik})} \prod_{i=1}^{G_l} \frac{\Gamma(x_{jil}+\alpha_{iik})}{\Gamma(\alpha_{iik})} \frac{\Gamma(\alpha_{jk})}{\Gamma(z_{jl}+\alpha_{jk})}
\]

(4)

where \( |\mathbf{\alpha}_k| = \sum_{i=1}^{G_l} \alpha_{iik} \). For a given individual \( l \), the posterior probability distribution of \( x_{jl} \) and \( z_{jl} \) is

\[
P(\mathbf{x}_{jl}, \mathbf{z}_{jl} | \mathbf{\alpha}_k) = \frac{1}{\prod_{i=1}^{G_l} \Gamma(\alpha_{iik})} \prod_{i=1}^{G_l} \frac{\Gamma(x_{jil}+\alpha_{iik})}{\Gamma(\alpha_{iik})} \frac{\Gamma(\alpha_{jk})}{\Gamma(z_{jl}+\alpha_{jk})}
\]

(5)

In this model, the two sets of parameters of interest are \( \mathbf{z}_{jl} \) and \( \mathbf{\alpha}_k \). Specifically, the joint posterior distribution for \( \mathbf{z}_{jl} \) and \( \mathbf{\alpha}_k \) is

\[
P(\mathbf{z}_{jl}, \mathbf{\alpha}_k | \mathbf{x}_{jl}) = \prod_{l=1}^{L} \prod_{j=1}^{G_l} P(\mathbf{x}_{jl}, \mathbf{z}_{jl} | \mathbf{\alpha}_k)
\]

(6)

Prior \( \mathbf{\alpha}_k \) is estimated by the mean of \( \mathbf{\alpha}_k = \sum_{l=1}^{L} \mathbf{\alpha}_k / L \). Estimation of \( \alpha_{jk} \) can be challenging due to limited number of individuals. We can assume all \( \alpha_{jk} \)'s follow a hyper-prior: Gamma distribution \( \mathbf{\alpha}_k \sim \mathbf{\alpha}_k / \mathbf{\alpha}_k \), and use information across all genes to estimate variance. In addition, we assume a noninformative prior for \( \mathbf{\alpha}_k \). Taken all together, we have the full posterior distribution as follows:

\[
P(\mathbf{z}_{jl}, \mathbf{\alpha}_k | \mathbf{x}_{jl}) = \prod_{l=1}^{L} \prod_{j=1}^{G_l} P(\mathbf{z}_{jl}, \mathbf{\alpha}_k | \mathbf{x}_{jl})
\]

(7)

We use Gibbs sample to iteratively update \( \mathbf{\alpha}_k \) and \( \mathbf{\alpha}_k \). Details can be found in Supplementary Methods.

**Classification and computational acceleration.** To further improve the computational efficiency, we provide a supervised option in BAMM-SC. Specifically, for very large-scale dataset, we use BAMM-SC to train a prediction model using a subset of cells from each individual and predict the clustering labels for the rest of cells. First, we randomly select a subset of cells from each individual and apply BAMM-SC on these selected cells. The estimate of \( \mathbf{\alpha}_k \) is computed as the average after deletion of the first 100 (default) iterations as burn-in. Then we predict the cell type labels for other cells with realization of parameters: \( \mathbf{\theta} = (\mathbf{a}_1, \ldots, \mathbf{a}_L, \mathbf{\alpha}_1, \ldots, \mathbf{\alpha}_L) \).

\[
P(z_{jl} = k | \mathbf{x}_{jl}, \mathbf{\theta}) = \frac{\prod_{i=1}^{G_l} \frac{\Gamma(x_{jil}+\alpha_{iik})}{\Gamma(\alpha_{iik})} \frac{\Gamma(\alpha_{jk})}{\Gamma(z_{jl}+\alpha_{jk})}}{\sum_{k'=1}^{K} \prod_{i=1}^{G_l} \frac{\Gamma(x_{jil}+\alpha_{iik})}{\Gamma(\alpha_{iik})} \frac{\Gamma(\alpha_{jk'})}{\Gamma(z_{jl}+\alpha_{jk'})}}
\]

(9)

This approach can substantially reduce the computational cost for very large-scale datasets while maintaining the accuracy as shown in Supplementary Fig. 14.

**Single-cell sequencing library construction.** 10x Genomics Chromium system, which is a microfluidics platform based on Gel bead in EMulsion (GEM) technology, was used for generating real test datasets. Cells mixed with reverse transcription reagents were loaded into the Chromium instrument. This instrument separated cells into minireaction partitions formed by oil microdroplets, each containing a gel bead and a cell, known as GEMs. GEMs contain a gel bead, scaffold for an oligonucleotide that is composed of an oligo-dT section for priming reverse transcription, and barcodes for each cell and each transcript as described. GEM generation takes place in a multiple-channel microfluidic chip that encapsulates single-gel beads. Reverse transcription takes place inside each droplet. Approximately, 1000-fold excess of partitions compared to cells ensured low capture of duplicate cells. The reaction mixture/emulsion was removed from the
Chromium instrument, and reverse transcription was performed. The emulsion was then broken using a recovery agent, and following Dynabead and SPRI clean up of cDNA were amplified by PCR (Clontech, Bio-Rad), cDNAs were sheared (Covaris) into ~200 bp length. DNA fragment ends were repaired, A-tailed and adapters ligated. The library was quantified using KAPA Universal Library Quantification Kit KK4824 and further characterized for cDNA length on a Bioanalyzer using High Sensitivity DNA kit. All sequencing experiments were conducted using Illumina NextSeq 500 in the Genomics Sequencing Core at the University of Pittsburgh.

Data description. Human PBMC dataset: Under a protocol approved by the University of Pittsburgh Institutional Review Board, peripheral blood was obtained from healthy donors by venipuncture. Each subject gave written informed consent. PBMCs were isolated from whole blood by density gradient centrifugation using Ficoll–Hypaque. PBMCs were then counted and resuspended in phosphate buffered saline with 0.04% bovine serum albumin, and were processed through the Chromium 10x Controller according to the manufacturers’ instructions, targeting a recovery of ~2000 cells. The following steps were all performed under the aforementioned protocol developed by 10x Genomics.

Human skin dataset: Skin samples were obtained by performing 3 mm punch biopsies from the dorsal midforearm of healthy control subjects after informed consent under a protocol approved by the University of Pittsburgh Institutional Review Board. Skin for scRNA-seq was dissected enzymatically (Miltenyi Biotec Whole Skin Dissociation Kit, human) for 2 h and further dispersed using the Miltenyi gentleMACS Octo Dissociator. The resulting cell suspension was filtered through 70 micron cell strainers twice and re-suspended in phosphate-buffered saline containing 0.04% bovine serum albumin. Cells from biopsies were mixed with reverse transcription reagents then loaded into the Chromium instrument (10x Genomics). Totally, ~2600–4300 cells were loaded into the instrument to obtain data on ~1100–1800 cells, anticipating a multiplet rate of ~1.2% of partitions. The following steps were all performed under the aforementioned protocol developed by 10x Genomics.

Mouse lung dataset: Lung single cell suspension from naive and infected C57BL/6 mice were subject to scRNA-seq library preparation protocol. Briefly, left lobes of both naive and infected mice were removed and digested with Collagenase/DNase to obtain single-cell suspension. Mononuclear cells after filtration with a 40 µm cell strainer were separated into minireaction partitions or GEMS formed by oil microdroplets, each containing a gel bead and a cell, by the Chromium instrument (10x Genomics). The reaction mixture/emulsion with captured and barcoded mRNAs were removed from the Chromium instrument followed by reverse transcription. The cDNA samples were fragmented and amplified using the Nextera XT kit (Illumina). The following steps were all performed under aforementioned the protocol developed by 10x Genomics. We have compiled with all relevant ethical regulations for animal research. The animal protocol was approved by the University of Pittsburgh Institutional Animal Care and Use Committee.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. The study uses various publicly available scRNA-seq datasets. Both human PBMC (sample 5) and purified CD19+B cell scRNA-seq data that support the findings of this study are available at https://support.10xgenomics.com/single-cell-gene-expression/datasets. The raw fastq files and preprocessed experimental test datasets (human PBMCs, mouse lung and human skin tissues) have been deposited in the gene expression omnibus (GEO) database under accession number GSE128066. All other relevant data are available upon request.

Code availability. BAMM-SC, including all source and example code, is freely available as an R package with a detailed tutorial at https://github.com/CHPGenetics/BAMMSSC.

Received: 6 August 2018 Accepted: 15 March 2019

References
1. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. Nat. Rev. Genet. 17, 175–188 (2016).
2. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382 (2009).
3. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214 (2015).
4. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. Nat. Commun. 8, 14049 (2017).
5. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343, 776–779 (2014).
6. Pollen, A. A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol. 32, 1053–1058 (2014).
7. van der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies cell-type-specific cis-eQTLs and co-expression QTLs. Nat. Genet. 50, 493–497 (2018).
8. Rodriguez, A. & Liao, A. Machine learning. Clustering by fast search and find of density peaks. Science 344, 1492–1496 (2014).
9. Wang, B. et al. SIMLR: a tool for large-scale genomic analyses by multikernel learning. Proteomics 18, 1702322 (2018).
10. duVerle, D. A., Yotsukura, S., Nomura, S., Abaratian, H. & Tsuda, K. CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. BMC Bioinformatics 17, 363 (2016).
11. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. Nat. Methods 14, 483–486 (2017).
12. Li, Z. & Ji, H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 44, e117 (2016).
13. Sun, Z. et al. DIMM-SC: a Dirichlet mixture model for clustering droplet-based single-cell transcriptomic data. Bioinformatics 34, 139–146 (2018).
14. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-seq data using MetaNeighbor. Nat. Commun. 9, 884 (2018).
15. Spitzer, M. H. et al. IMMUNOLOGY. An interactive reference framework for modeling a dynamic immune system. Science 349, 1259425 (2015).
16. Haghighi, V., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. 36, 421–427 (2018).
17. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. 33, 495–502 (2015).
18. Rand, W. M. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66, 846–850 (1971).
19. Freytag, S., Tian, L., Lonstedt, I., Ng, M. & Bahlo, M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. F1000Res. 7, 1297 (2018).
20. Duo, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Res. 7, 1141 (2018).
21. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. Genome. Biol. 18, 174 (2017).
22. Chen, K. & Kolls, J. K. T cell-mediated host immune defenses in the lung. Annu. Rev. Immunol. 31, 605–633 (2013).
23. Weiser, J. N. The pneumococcus: why a commensal misbehaves. J. Mol. Med. 88, 97–102 (2010).
24. Tabib, T., Morse, C., Wang, T., Chen, W. & Lafyatis, R. SFRP2/DPP4 and FMO1/LSPIF define major fibroblast populations in human skin. J. Invest. Dermatol. 138, 802–810 (2018).
25. Datta, S. & Datta, S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19, 459–466 (2003).
26. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65 (1987).
27. Coifman, R. R. et al. Geometric diffusions as a tool for harmonic analysis and data mining in Euclidean and invariant settings. J. Comput. Appl. Math. 215, 1–35 (2008).
28. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. Nat. Methods 15, 1053 (2018).
29. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386 (2014).
30. Trapnell, C. Defining cell types and states with single-cell genomics. Genome Res. 25, 1491–1498 (2015).

Acknowledgements
This work is supported by National Institute of Health grants R56HL137709 (K.C.), P50 CA097190 and P30 CA047904 (D.A.A.V.), P50 AR060780 (R.L. and W.C.), R35HL139930 (J.K.) and Children’s Hospital of Pittsburgh (W.C. and Z.S.).

Author contributions
M.H. and W.C. conceived the study; Z.S. led the statistical modeling and data analysis; L. C helped with developing the R package; A.R.C., T.C.B., and D.A.A.V. performed the experiments for the human PBMC data; T.J. and R.L. performed the experiments for the human skin data; K.C. and J.K. designed the experiments for the mouse data; H.X., Y. J., and Q.H. helped with the preprocessing and analyzing the data; Y.D., W.C., and M.H. supervised the research; and Z.S., W.C., Y.D., and M.H. led the writing of the paper with input from all the other authors.
Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09639-3.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019