PROBABILISTIC GENE EXPRESSION SIGNATURES IDENTIFY CELL-TYPES FROM SINGLE CELL RNA-SEQ DATA

A PREPRINT

Isabella N. Grabski
Department of Biostatistics, Harvard University
isabellagrabski@g.harvard.edu

Rafael A. Irizarry
Department of Data Sciences, Dana-Farber Cancer Institute
Department of Biostatistics, Harvard University
rafael_irizarry@dfci.harvard.edu

January 22, 2020

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) quantifies the gene expression of individual cells in a sample, which allows distinct cell-type populations to be identified and characterized. An important step in many scRNA-seq analysis pipelines is the classification of cells into known cell-types. While this can be achieved using experimental techniques, such as fluorescence-activated cell sorting, these approaches are impractical for large numbers of cells. This motivates the development of data-driven cell-type identification methods. We find limitations with current approaches due to the reliance on known marker genes and sensitivity to the quality of reference samples. Here we present a computationally light statistical approach, based on Naive Bayes, that leverages public datasets to combine information across thousands of genes and probabilistically assign cell-type identity. Using datasets ranging across species and tissue types, we demonstrate that our approach is robust to low-quality reference data and produces more accurate cell-type identification than current methods.

***K*eywords** Single cell RNA-Seq, Empirical Bayes Models

1 Introduction

Single-cell RNA sequencing (scRNA-seq) quantifies gene expression at the level of individual cells, rather than measuring the aggregated gene expression in a biological sample containing millions of cells, as is done with bulk RNA-sequencing. This improved granularity permits the identification or discovery of distinct populations of cell-types within the tissues under study. To effectively accomplish this, it is important to classify cells reliably into known cell-types, in particular cells that are present in many tissues such as immune system cells. Fluorescence-activated cell sorting (FACS) can be used prior to the sequencing step to physically sort cells from a mixed sample into their cell-type populations. While generally regarded as highly accurate, FACS-sorting has limited throughput, and thus is impractical when sequencing large numbers of cells. As a result, there is a need for data-driven approaches to identify cell-types.

Current methods fall into one of two categories, which we will refer to as *clustering-based* and *reference-based*. In clustering-based methods, the target cells are first grouped using an unsupervised clustering algorithm (for example, [1, 2, 3, 4]). Next, differential expression analysis is used to identify genes that are uniquely expressed in each group and compared to known cell-specific marker genes to annotate the group as a particular cell-type. Although any clustering algorithm can be used in this type of method, we focus primarily on Seurat's clustering algorithm [4], since it is popularly used in this way for cell-type identification.

Reference-based methods (for example, [5, 6, 7, 8, 9]) use supervised learning approaches in which the target cells are compared to reliably annotated, such as FACS-sorted, reference data for each cell-type of interest, and each target cell is annotated using the closest match. Approaches to defining *closest match* vary. Many of these supervised methods are based on complex machine learning algorithms, such as Xgboost [7] and deep neural networks [10], which are prone to over-fitting. In addition, some of the most popular methods rely on marker genes to guide the determination of the closest match [6, 5].

We find that methods using marker genes are limited by the sparsity of scRNA-seq data. Specifically, in a typical scRNA-seq experiment using unique molecular identifiers (UMIs), between roughly 50% and 100% of genes report 0 counts in an individual cell. This has been shown to be a result of technological limitations rather than biologically driven [11]. Here, we demonstrate that the probability of observing a non-zero count for a gene that is actually expressed can be substantially smaller than 1; for instance, in CD4 cells, the markers had non-zero counts in anywhere from 0% to 38% of the cells. As a result, many marker genes that have been established as reliable indicators of cell-type on the bulk level are no longer adequate on the single cell level. This implies that methods relying on them can be highly sensitive to the choice of marker genes. Furthermore, for less-studied or rare cell-types, there are often no well-established marker genes suitable for the single cell resolution.

In this work, to minimize over-fitting and avoid reliance on marker genes, we apply a version of one of the simplest machine learning algorithms, Naive Bayes [12], that defines a conditional distribution for each target cell by estimating a cell-type-specific expression rate for each gene. To overcome the problem where the estimated rates are zero for hundreds of genes due to the sparsity of the data, we use a hierarchical model that defines a cell-type-specific distribution for each gene, with the hierarchical aspect providing statistical power in the presence of said sparsity. We describe the datasets used to build and assess our method, show the limitations of existing approaches, provide a detailed description of our approach, and finally demonstrate its advantages.

2 Results

2.1 Datasets

To benchmark our approach against existing methods, we used three datasets. We chose these datasets by identifying publicly available data satisfying four requirements: (1) the data consisted of UMI counts, (2) the cell-types were FACS-sorted, (3) the donors were healthy and untreated, and (4) we were able to find at least one other UMI count, FACS-sorted, healthy-donor dataset for the same cell-types from a different study that could be used as a test dataset. These requirements greatly limited the number of datasets available but it permitted us to assess the algorithms in a real-world setting and explicitly check for over-training. Specifically, we constructed two test sets for each dataset: one constructed in the traditional way of withholding a subset of the original dataset, and the other by using a dataset generated in a completely different study. We refer to these two as the *withheld* test set and the *external* test set, respectively. Note that over-training, a concern for supervised methods such as the reference-based ones, will result in better performance in the withheld test set compared to the external test set.

The first dataset was constructed from four different FACS-sorted human peripheral blood mononuclear cell (PBMC) datasets from 10X Genomics. In particular, we combined 11,113 CD4 cells, 10,109 CD8 cells, 2,512 CD14 cells, and 8,285 NK cells [13]. We refer to this dataset as the *PBMC* dataset. For the withheld test set, we withheld 100 cells of each of the four cell-types. To construct the external test set, we combined 425 FACS-sorted CD14 cells each from two different donors [14].

The second dataset was created from three FACS-sorted cell-types from different human tissues: 32,841 prostate cells [15], 9,900 brain microglia cells [16], and 2,512 CD14 PBMCs [13]. These are three clearly distinct tissues, and so, in principle, these cell-types should be easy to distinguish. We refer to this dataset as the *Tissues* dataset. To form the withheld test set, we withheld 100 prostate and CD14 cells, and 34 brain cells. The external test was formed from independent FACS-sorted data from other studies. Specifically, we used 425 cells each of prostate primary epithelial cells [17], brain microglial cells [18], and CD14 cells [14].

The third dataset was created from FACS-sorted mouse CD4 Treg and Tmem cells from four different tissues each: 3,883 Treg spleen cells, 4,496 Tmem spleen cells, 4,805 Treg brachial lymph node cells, 3,108 Tmem brachial lymph node cells, 2,982 Treg colon cells, 3,737 Tmem colon cells, 156 Treg skin cells, and 166 Tmem skin cells [19]. This dataset, which we call the *Treg/Tmem* dataset, represents a challenging case of many similar cell-types. The withheld dataset was formed from 100 withheld cells of each cell-type, except for skin cells due to their low number of cells. The external test set combined two independent experiments that sequenced FACS-sorted Treg spleen cells; one experiment is from [20], and the other is unpublished data made available by Dr. Zikai Zhou's laboratory.

2.2 Clustering-based approaches artificially identify more cell-types with increasing dataset size

We used the PBMC dataset to investigate the relationship between the number of clusters and the dataset size when clustering with Seurat [4]. In particular, we applied the clustering algorithm implemented in Seurat v3, which uses a graph-based approach with the Louvain algorithm [21], to successively subsetted versions of this dataset, ranging from a total of 80 cells (20 per cell-type) to a total of 2,000 cells (500 per cell-type). Although there are four cell-types represented in this dataset, the number of clusters found ranges from three to six, with a general increasing trend with

dataset size (Figure 1a). Only a narrow window of dataset size (roughly between 500 and 1,000 total cells) corresponds to the correct number of cell-types.

We further examined the six clusters found in the case with the largest dataset size, and the extra clusters found do not appear reflective of biological ground truth. In particular, there does not seem to be strong separation among the clusters identified (Figure 1b). Furthermore, the clusters are hierarchically related in ways that do not correlate with the true cell-type identities of the cells within those clusters (Figure 1c). For example, clusters 2 and 5 represent a division from a larger group, but in fact, nearly every cell in both of those clusters is a CD4 cell. Note that clusters 0, 3, and 4 are all closely related, but the cells in cluster 0 are denoted as nearly half CD8 and half CD14 by FACS-sorting, with almost all in the cells in cluster 3 corresponding to CD8 cells and almost all the cells in cluster 4 corresponding to CD14 cells. Hence, these extra divisions cannot be simply interpreted as the result of identifying sub-types within larger categories of cell-types. Only cluster 1, which consists almost entirely of NK cells, corresponds to a single cell-type with specificity. These findings suggest that the identity assigned to a given cell depends on the size of the dataset, in a way that is not representative of the actual cell-types in the data.

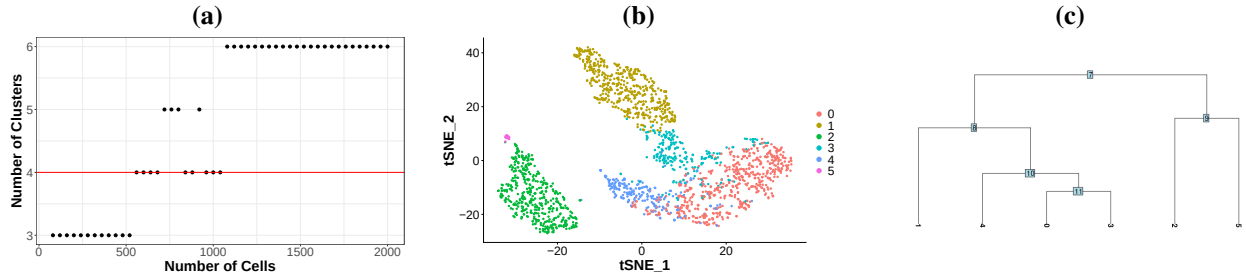

Figure 1: Applying Seurat clustering to successive subsets containing equal amounts of each of the same four cell types (CD4 T-cells, CD14 T-cells, CD8 T-cells, and NK cells) from the PBMC reference data incorrectly identifies more cell-types with increasing dataset size. (a) Number of clusters identified by Seurat plotted against the number of cells included in the analysis. The red line denotes the true number of cell-types. (b) tSNE plot of the six clusters, denoted with different colors, identified in the largest dataset size, which consisted of 2,000 cells. (c) Dendrogram of the six clusters identified in the largest dataset size, constructed using the "average" cell of each cluster. The leaf labels denote the cluster numbers.

2.3 Most marker genes are unlikely to be observed

To demonstrate the unreliability of marker genes in scRNA-Seq data, we use a subset of the reference PBMC dataset as an example and look at the counts for externally validated marker genes for each cell type. Specifically, we selected marker genes for CD4 (IL7R, CD4, CTLA4, FOXP3, IL2RA, PTPRC), CD14 (CD14, LYZ, FCGR3A, CD68, S100A12), CD8 (CD8A, CRTAM, NCR3, CD3D), and NK cells (GNLY, NKG7, PRR5L, S1PR5, NCAM1) that have been curated and well-established in the literature as discriminating among cell types [22, 23, 4]. This list of marker genes is representative of what is likely to be chosen by a user applying a marker-based method to a PBMC dataset.

Among these genes, GNLY and NKG7 are highly sensitive and highly specific to NK cells, as is LYZ for CD14 cells (Figure 2). However, we observe 0 cells with non-zero counts even for marker genes that are known to be expressed (Figure 2). Specifically, both CD14 and CD4 cells had markers with non-zero counts in 0% of cells, and the remaining two cell-types, NK and CD8, both had markers with non-zero counts in as low as 4% of cells. Among CD4 cells in particular, the highest proportion of non-zero counts for any of its markers was only 38% (Figure 2).

2.4 Our approach yields model-based probabilistic classifications

We present a statistical approach, based on Naive Bayes, that defines a cell-type-specific distribution for every gene, by fitting a hierarchical mixture model for each cell-type using FACS-sorted reference datasets for each cell-type of interest. Specifically, if Z_i is an indicator variable that can represent any of the cell-types k, we compute the conditional probability $\Pr(Z_i = k \mid \boldsymbol{Y}_i)$ of each cell-type k given the data $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^\top$, with Y_{ij} the count for gene j in cell i. To keep the approach simple, we assume independence,

$$\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid Z_i = k) = \prod_{j=1}^{J} \Pr(Y_{ij} = y_{ij} \mid Z_i = k),$$

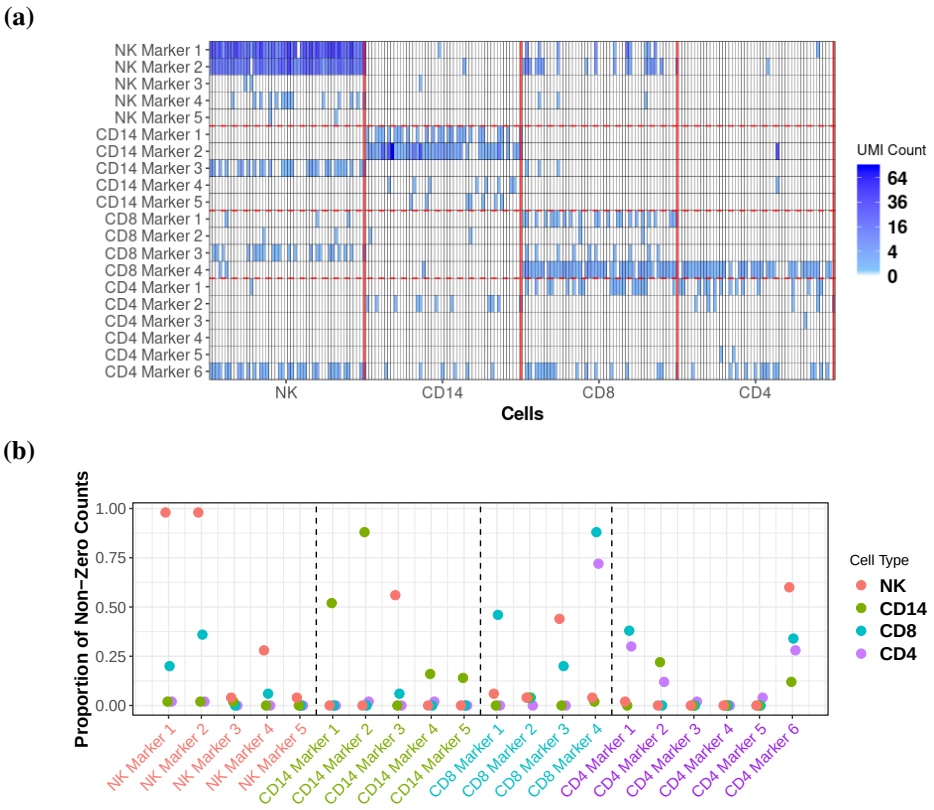

Figure 2: PBMCs data from 20 canonical marker genes show that they are unlikely to be observed. (a) UMI counts, shown in shades of blue, for marker genes in 50 randomly selected cells from each pertinent cell-type. (b) Proportion of non-zero UMI counts for the same canonical marker genes and each cell-type, which is denoted with color.

and Poisson probability model

$$Y_{ij} \mid Z_i = k \sim \text{Poisson}\left(N_i \lambda_{jk}\right)$$

with λ_{jk} the rate for gene j in cell-type k and N_i the total UMI counts in cell i. We explicitly include the total number of observed transcripts N_i to account for the varying coverage across cells [24]. Note that the vector $(\lambda_{1k}, \ldots, \lambda_{Jk})^\top$ can be considered the expression profile that defines cell-type k.

Assuming that the prior probability of a cell belonging to any particular cell-type is uniform, the Naive Bayes procedure is defined by picking the class k that maximizes the posterior probability:

$$\Pr(Z_i = k | \boldsymbol{Y}_i) = \frac{\Pr(\boldsymbol{Y}_i \mid Z_k = k)}{\sum_{k=1}^{K} \Pr(\boldsymbol{Y}_i | Z_i = k)}$$

if there are K total cell-types. Note that this approach allows the certainty of the classification to be directly quantified. For example, if a target cell has a relatively low probability of belonging to any of the reference cell-types, this might indicate contamination or potentially a novel cell-type population.

To implement this model in practice, we need to estimate the λ_{jk} in a training set. The data sparsity presented a challenge because standard estimates, such as maximum likelihood estimtes (MLE), can produce rates of 0 for thousands of expressed genes (Figure 2), which in turn made the Naive Bayes approach too sensitive to expressed genes with low expression. We therefore used a hierarchical model with prior distributions on λ that shrunk estimates away from 0. As done by [25] in the bulk setting, we modeled the λ_{jk} parameters as coming from a mixture of cell-type-specific distributions, corresponding to unexpressed (off) genes and expressed (on) genes.

To motivate the model choice for the on and off distribution in scRNA-seq, we used the microarray barcode [25] to obtain independent calls of which genes were off and on in each tissue type. The empirical distributions from these two groups suggest that two distributions are needed to describe the off genes (Figure 3a). This result motivated modeling the rates λ_{jk} as a mixture of a zero-inflated exponential and log-normal distribution for genes that are not expressed, and a log-normal distribution for expressed genes. Section 4 provides details.

Additionally, we found that off genes may still have non-zero counts, and on genes may have low rates. Our model fits the data (Figure 3b) and provides a statistical explanation for this observation. Furthermore, by hierarchically modeling the gene- and cell-type-specific rate parameters as coming from a mixture of distributions, we are able to increase power despite the high sparsity observed in UMI count data. For example, a key advantage of our approach is that we are able to recover non-zero rate estimates for genes with 0 counts in the reference sample, which would not be possible in a simplified approach that only uses, e.g., MLE from a non-hierarchical Poisson model.

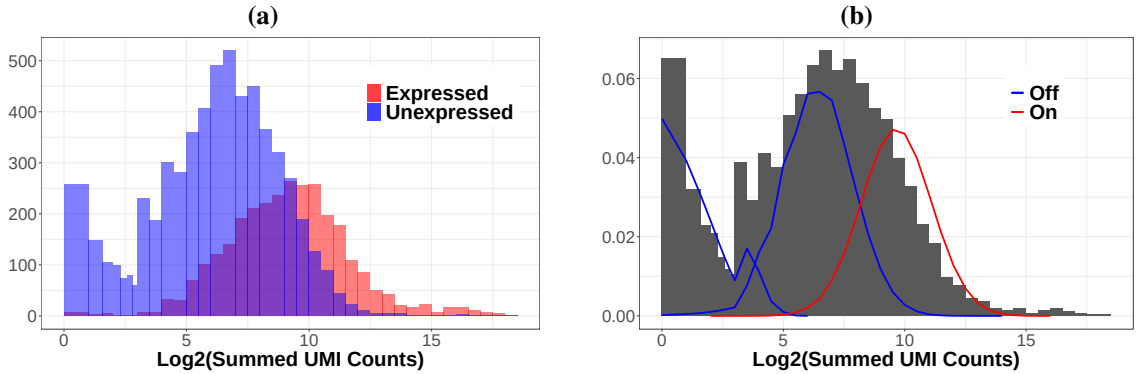

Figure 3: Histograms of log-transformed CD4 cell UMI counts from the PBMC reference data, summed for each gene across all cells in the dataset, excluding genes whose total summed counts are 0. (a) The red histogram shows genes that the microarray barcode reports as expressed in at least 75% of samples, and the blue histogram shows genes that are not expressed in at least 75% of samples. (b) The histogram is overlaid with densities from simulating under the fitted on and off distributions, similarly conditional on observing non-zero total counts. These densities use the posterior mean parameters, and are scaled appropriately for the total number of transcripts across all cells.

After fitting these distributions for the reference datasets for each cell-type, we can use the posterior mean of each λ_{jk}, and compute the posterior probability of observing the target cell under the distributions for each cell-type.

A key advantage of our approach is that these probabilistic classifications are made on the basis of the thousands of genes available, rather than relying largely or exclusively on a subset of marker genes. For example, in the PBMC dataset, a large number of genes can be seen to have very different posterior mean rates between CD4 cells and the other three cell-types (Figure 4). Furthermore, the six canonical CD4 marker genes discussed in subsection 2.3 do not appear useful on their own in distinguishing these cell-types, which handicaps methods that are reliant on such marker genes. By using all genes available in the reference sample, our approach circumvents these marker gene limitations.

2.5 Our approach improves current reference-based methods

We used the datasets described in section 2.1 to compare our approach to six current reference-based methods. In each of the three sets (PBMC, Tissues, and Treg/Tmem), we use two types of test sets, which we refer to as withheld and external respectively. The withheld test set consists of cells from the same experiment as the reference sample, and the external test set consists of cells from an entirely different study.

The current methods we evaluated were scVI [10], CHETAH [9], scmap [8], CaSTLe [7], Garnett [6], and CellAssign [5]. For each of these methods, we processed the reference and test datasets as recommended in their respective documentations, and for the methods requiring markers (Garnett and CellAssign), we identified five markers for each cell-type using `scran` [26] on the reference dataset, as recommended in the CellAssign documentation. This was done out of consistency, since not every cell-type under consideration here has readily available, independently verified marker genes.

Our approach compares favorably to current methods across all datasets (Table 1). In particular, our approach outperformed other methods in the external test sets. In the PBMC dataset, all methods performed almost perfectly in both withheld and external test sets. In the Tissues datasets, our method greatly outperformed all other methods in the external test set. scVI and CaSTLe outperformed our method in the withheld dataset, but their performance

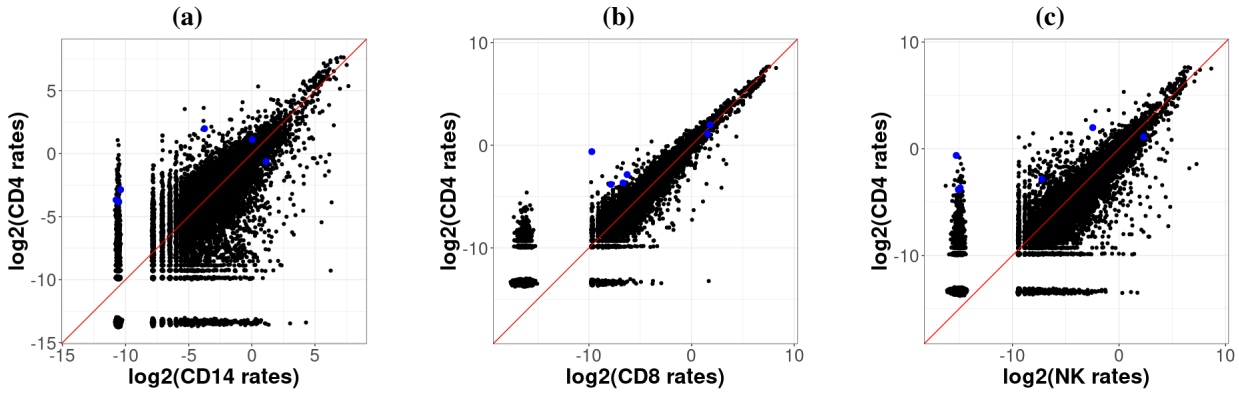

Figure 4: Log-transformed posterior mean rates (in counts per 10,000) for each gene, compared between CD4 cells and each of the other three cell-types in the PBMCs reference dataset (CD14, CD8, and NK cells). The red line indicates the identity line, and the blue points are the six canonical CD4 marker genes (IL7R, CD4, CTLA4, FOXP3, IL2RA, PTPRC).

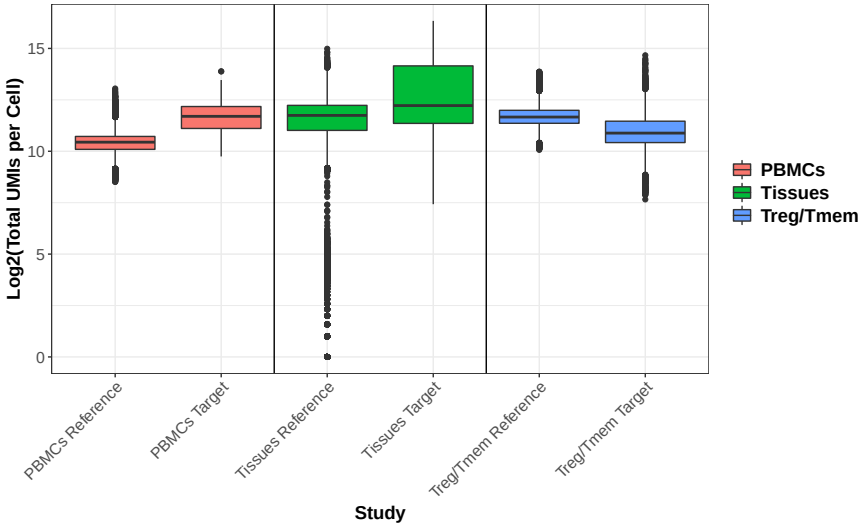

Figure 5: Total UMI counts per cell in each reference sample and corresponding external test dataset.

in the external set was substantially worse. In the Treg/Tmem datasets, our method again greatly outperformed all other methods in the external test set, except for CellAssign. However, our method performed substantially better than CellAssign on the withheld set, which had a larger variety of cell-types. scVI and CaSTLe again outperformed our method in the withheld dataset with much worsened performance on the external set, which is consistent with over-training. A possible explanation for the over-training is the fact that many of the cells in the external set had higher coverage than the reference samples (Figure 5). It should be noted that for the external Treg/Tmem test set, which consisted of two separate experiments, CaSTLe had similar performance as our approach on one of the experiments.

3 Discussion

Currently, cell-type identification in scRNA-seq datasets is typically done with either clustering-based or reference-based methods. We showed with the example of Seurat, one of the most popular clustering-based methods, that such methods can artificially identify more cell-types as the dataset size increases, without reflecting biological ground truth. Moreover, clustering-based methods can depend on arbitrary user decisions. For example, many cell-types do not have well-established and widely-accepted marker genes, so different users examining the same clusters might draw two different conclusions about the cell-type identities. This method of direct annotation by a user also implies that probabilistic classification is not possible; a given target cell is assigned to some particular cell-type without any

	PBMC		Tissues		Treg/Tmem	
	Withheld	**External**	**Withheld**	**External**	**Withheld**	**External**
Our approach	0.98	0.98	0.96	0.80	0.77	0.54
scVI	1.00	0.98	1.00	0.66	0.92	0.31
CHETAH	0.95	0.97	0.17	0.36	0.66	0.30
scmap	0.71	0.75	0.85	0.61	0.15	0.01
CaSTLe	0.99	0.99	1.00	0.36	0.79	0.31
Garnett	0.86	1.00	0.01	0.67	0.30	0.22
CellAssign	0.74	0.98	0.85	0.67	0.31	0.67

Table 1: Accuracy of cell-type assignment for each test dataset, in our approach as well as six current reference-based methods.

information about the certainty of this assignment. Furthermore, there is no way to specify the desired granularity of the cell-types. Since clusters are identified in an unsupervised manner, there is no differentiation between, for instance, a use case where it is enough to identify cells as "T-cells" and a use case where identification at the level of T-cell subtypes is needed.

Reference-based methods can avoid many of the pitfalls of cluster-based methods. The supervised approach means that the desired granularity can be controlled with the choice of reference data, and the number of distinct cell-types identified should not grow with increasing dataset size. Nevertheless, we showed that the currently available algorithms are susceptible to over-training. Many continue to rely on marker genes for their classifications [6, 5], which we demonstrated can be problematic even for cell-types with well-defined marker genes, let alone for rare or poorly-characterized cell-types. In addition, while most quantify the certainty of their match with some metric, these often do not have direct interpretations as probabilities, so they still do not address the need for probabilistic classifications.

We also showed that existing methods are not robust to reference and test data with varying coverage, or to datasets with very similar cell-types. While most methods performed well on the PBMCs dataset, we considered the Tissues and Treg/Tmem datasets to be the most challenging. In the Tissues reference dataset, the brain microglia data have an especially high rate of zero counts compared to the other two cell-types, which makes generalizing to the external test set difficult. In the Treg/Tmem dataset, the cell-types are very similar to each other, which also makes classification challenging.

We showed that our approach compared very favorably to existing methods on these external test sets, which suggests that our approach is more robust both to low-quality reference data and to very similar cell-types. Overall, our approach was more successful at generalizing to external test sets than the other methods. We attribute the poor performance of CellAssign and Garnett on some datasets to their reliance on marker genes. While scVI and CaSTLe both perform strongly on withheld test sets, they are substantially weaker on the external test sets when the reference and test sets have different coverage. This is consistent with overtraining, and may also be attributed to the fact that both methods downsample the number of genes. We note that cell-type identification is a use case but not the primary purpose of scmap and scVI, which may explain their suboptimal performances.

The strength of our approach lies in leveraging thousands of genes to make the identifications. We explicitly model the cell-type-specific probability distributions of these genes with a hierarchical approach that increases power in the presence of sparsity. We directly account for coverage in our model, and we employ the comparatively simple Naive Bayes approach to identify each cell. This yields probabilistic classifications for each cell in a way that is robust to varying quality between the reference and test data.

Note that we consider our approach a first attempt at developing a robust stand-alone classification method. To avoid over-training, we kept the model simple by assuming conditional independence between genes and a Poisson model with no over-dispersion. Although these assumptions result in a method that outperforms current more complicated methods, we conjecture that improvements can be achieved by relaxing these assumptions.

4 Methods

4.1 Model

Our model defines a cell-type-specific distribution for every gene using a hierarchical mixture model. For each cell-type k, we model the UMI counts Y_{ijk} for gene j in cell i as

$$Y_{ijk} \sim \text{Poisson}(\mu_{ijk}),$$

with $\mu_{ijk} = N_i \lambda_{jk}$, where N_i is the total number of UMI counts observed in cell i and λ_{jk} defines the gene expression profile for cell-type k and gene j. We assume λ_{jk} follows a mixture distribution with components representing genes that are unexpressed (off) and expressed (on). Specifically, we assume that

$$\lambda_{jk} \sim p_0 f_{0,k} + (1 - p_0) f_{1,k}$$

with $f_{0,k}$ and $f_{1,k}$ representing the off and on distribution for cell-type k, respectively defined as

$$f_{0,k} = p_{0,0}\mathbb{I}_0 + p_{0,1} \cdot \text{Exp}(\alpha_k) + (1 - p_{0,0} - p_{0,1}) \cdot \text{LogNormal}(\mu_{0,k}, \sigma^2_{0,k})$$

where $\mathbb{I}_0$ represents a point mass at 0, and

$$f_{1,k} = \text{LogNormal}(\mu_{1,k}, \sigma^2_{1,k}).$$

Hence, the gene-specific rate parameters for UMI counts come from a mixture of three cell-type-specific distributions and a zero-inflation component.

4.2 Fitting the Model

To avoid arriving at nonsensical solutions, we impose certain context-driven constraints. First, we force $\mu_{1,k} > \mu_{0,k}$. Second, we constrain zero-valued UMI counts to come from either the zero-inflation or the Exponential component. Finally, we permit the non-zero-valued UMI counts to come from any component except the zero-inflation.

We use Gibbs sampling to fit the model using the following prior distributions to allow estimation:

$$\begin{aligned}
\alpha_k &\sim \text{Gamma}(1, 1) \\
\mu_{0,k} &\sim \text{Normal}(0, 1) \\
\sigma^{-2}_{0,k} &\sim \text{Beta}(1, 1) \\
\mu_{1,k} &\sim \text{Truncated Normal}_{(\mu_{0,k}, \infty)}(0, 1) \\
\sigma^{-2}_{1,k} &\sim \text{Beta}(1, 1)
\end{aligned}$$

The conjugate priors for α_k and $\mu_{0,k}$ were used for computational efficiency. The other priors were chosen to enforce the remaining constraints we place on our model. Namely, the prior distribution for $\mu_{1,k}$ is truncated from below so that the mean of the on Log-Normal distribution is larger than the mean of the off Log-Normal distribution. Further, the Beta priors on $\sigma^{-2}_{0,k}$ and $\sigma^{-2}_{1,k}$ ensure that the variances of the Log-Normal distributions are greater than 1.

To fit this model for each cell-type k, we take the portion of the reference dataset with cells corresponding to that cell-type and sum up their UMI counts for each gene. Estimation is done using the R package `rjags` [27]. We run the sampler for 7,000 iterations and treated the first 5,000 as burn-in, unless lack of convergence necessitates more burn-in iterations. Furthermore, we fit the model without marginalizing component assignments, i.e. we infer discrete assignments for each gene to each mixture component at every iteration. To facilitate this, we initialize every non-zero gene above the 75th percentile of counts as belonging to the on distribution, every non-zero gene below the 25th percentile of counts as belonging to the off Exponential distribution, and the remaining non-zero genes as belonging to the off Log-Normal distribution. The genes with all zero counts are randomly initialized as either belonging to the zero point mass or the off Exponential distribution.

4.3 Cell-Type Assignments using Naive Bayes

Once we fit the model for each cell-type k, to identify the cell-types of each target cell, we set the rate parameter $\hat{\lambda}_{jk}$ for gene j in cell-type k as the posterior mean over the 2,000 post-burn-in iterations. For each test cell, we can compute the probability of observing its UMI counts under cell-type identity k using Bayes rule. Specifically, if we define Z_i as an indicator variable that can be any of the cell-types k, then for each cell profile Y_{ij}, the conditional distribution for this vector is defined by

$$Y_{ij} \mid Z_i = k \sim \text{Poisson}\left(N_i \hat{\lambda}_{jk}\right),$$

where N_i is the total UMI counts observed in cell i.

We then use this to compute the conditional probability of each class k given the data Y_{ij}. Specifically, if we define $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^\top$, with J the total number of genes, we can use the above result to compute the conditional probability of the observed data $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})^\top$:

$$\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid Z_i = k) = \prod_{j=1}^{J} \Pr(Y_{ij} = y_{ij} \mid Z_i = k).$$

Assuming that the prior probability of a cell belonging to any particular cell-type is uniform, the Naive Bayes procedure is defined by picking the class k that maximizes the posterior probability:

$$\Pr(Z_i = k | \boldsymbol{Y}_i) = \frac{\Pr(\boldsymbol{Y}_i \mid Z_k = k)}{\sum_{k=1}^{K} \Pr(\boldsymbol{Y}_i | Z_i = k)},$$

if there are K total cell-types.

References

[1] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483, 2017.

[2] Peijie Lin, Michael Troup, and Joshua WK Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology*, 18(1):59, 2017.

[3] Vasilis Ntranos, Govinda M Kamath, Jesse M Zhang, Lior Pachter, and N Tse David. Fast and accurate single-cell rna-seq analysis by clustering of transcript-compatibility counts. *Genome biology*, 17(1):112, 2016.

[4] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single cell data. *bioRxiv*, 2018.

[5] Allen W Zhang, Ciara O'Flanagan, Elizabeth A Chavez, Jamie LP Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, et al. Probabilistic cell-type assignment of single-cell rna-seq for tumor microenvironment profiling. *Nature methods*, pages 1–9, 2019.

[6] Hannah A Pliner, Jay Shendure, and Cole Trapnell. Supervised classification enables rapid annotation of cell atlases. *BioRxiv*, page 538652, 2019.

[7] Yuval Lieberman, Lior Rokach, and Tal Shay. Castle–classification of single cells by transfer learning: Harnessing the power of publicly available single cell rna sequencing experiments to annotate new experiments. *PloS one*, 13(10):e0205499, 2018.

[8] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359, 2018.

[9] Jurrian Kornelis de Kanter, Philip Lijnzaad, Tito Candelli, Thanasis Margaritis, and Frank Holstege. Chetah: a selective, hierarchical cell type identification method for single-cell rna sequencing. *bioRxiv*, page 558908, 2019.

[10] 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, 2018.

[11] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2017.

[12] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[13] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049, 2017.

[14] Christel Goudot, Alice Coillard, Alexandra-Chloé Villani, Paul Gueguen, Adeline Cros, Siranush Sarkizova, Tsing-Lee Tang-Huau, Mylène Bohec, Sylvain Baulande, Nir Hacohen, et al. Aryl hydrocarbon receptor controls monocyte differentiation into dendritic cells versus macrophages. *Immunity*, 47(3):582–596, 2017.

[15] Gervaise H Henry, Alicia Malewska, Diya B Joseph, Venkat S Malladi, Jeon Lee, Jose Torrealba, Ryan J Mauck, Jeffrey C Gahan, Ganesh V Raj, Claus G Roehrborn, et al. A cellular anatomy of the normal adult human prostate and prostatic urethra. *Cell reports*, 25(12):3530–3542, 2018.

[16] Emma Gerrits, Yang Heng, Erik W. G. M. Boddeke, and Bart J. L. Eggen. Transcriptional profiling of microglia; current state of the art and future perspectives. *Glia*, 2020.

[17] Tara McCray, Daniel Moline, Bethany Baumann, Donald J Vander Griend, and Larisa Nonn. Single-cell rna-seq analysis identifies a putative epithelial stem cell population in human primary prostate cells in monolayer and organoid culture conditions. *American journal of clinical and experimental urology*, 7(3):123, 2019.

[18] Jonathan Hasselmann, Morgan A Coburn, Whitney England, Dario X Figueroa Velez, Sepideh Kiani Shabestari, Christina H Tu, Amanda McQuade, Mahshad Kolahdouzan, Karla Echeverria, Christel Claes, et al. Development of a chimeric model to study and manipulate human microglia in vivo. *Neuron*, 103(6):1016–1033, 2019.

[19] Ricardo J Miragaia, Tomás Gomes, Agnieszka Chomka, Laura Jardine, Angela Riedel, Ahmed N Hegazy, Natasha Whibley, Andrea Tucci, Xi Chen, Ida Lindeman, et al. Single-cell transcriptomics of regulatory t cells reveals trajectories of tissue adaptation. *Immunity*, 50(2):493–504, 2019.

[20] Xiying Fan, Bruno Moltedo, Alejandra Mendoza, Alexey N Davydov, Mehlika B Faire, Linas Mazutis, Roshan Sharma, Dana Pe'er, Dmitriy M Chudakov, and Alexander Y Rudensky. Cd49b defines functionally mature treg cells that survey skin and vascular tissues. *Journal of Experimental Medicine*, 215(11):2796–2814, 2018.

[21] Xinyu Que, Fabio Checconi, Fabrizio Petrini, and John A Gunnels. Scalable community detection with the louvain algorithm. In *2015 IEEE International Parallel and Distributed Processing Symposium*, pages 28–37. IEEE, 2015.

[22] Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, et al. Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research*, 47(D1):D721–D728, 2018.

[23] Sergii Domanskyi, Anthony Szedlak, Nathaniel T Hawkins, Jiayin Wang, Giovanni Paternostro, and Carlo Piermarocchi. Polled digital cell sorter (p-dcs): Automatic identification of hematological cell types from single cell rna-sequencing clusters. *bioRxiv*, page 539833, 2019.

[24] 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. *bioRxiv*, page 574574, 2019.

[25] Matthew N McCall, Karan Uppal, Harris A Jaffee, Michael J Zilliox, and Rafael A Irizarry. The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic acids research*, 39(suppl_1):D1011–D1015, 2010.

[26] Aaron Lun, Karsten Bach, Jong Kyoung Kim, Antonio Scialdone, and Laleh Haghverdi. Package 'scran'. 2017.

[27] Martyn Plummer, Alexey Stukalov, Matt Denwood, and Maintainer Martyn Plummer. Package 'rjags'. *update*, 16:1, 2019.