Demystifying Dropouts in 10X UMI Data

January 7, 2020

Abstract

Main

Droplet-based single-cell RNA-sequencing (scRNA-seq) methods have changed the landscape of research in biological systems [15, 23, 12, 24] by producing cell-level data at improved costs. Moreover, recent sequencing technologies have an additional step of barcoding unique molecular identifiers (UMI) to remove amplification bias, and hence, the quality of the data sets has improved as well. Recent literature [3, 18] suggests that barcoding has led to a different error structure in the count data set with much less technical noise. Regardless, many tools assume that the both suffer from excessive technical noise, and they do not acknowledge the difference between the read count data or UMI count data.

Various softwares have attempted at cleaning the scRNA-seq data. Some focus on correcting the effects of sequencing depths or size factors [19, 2]. Some focus on building and fitting for parametric models that are designed to reflect the true data-generating process [8, 5, 21, 22]. Another branch of research has focused on only downstream analysis such as clustering [20, 5] and differential analysis [16, 14]. Some of the softwares present an integrated tool for both pre-processing and downstream analysis tool [2], but even so, the analysis pipeline has been dichotomized into the pre-processing and downstream analysis.

Here, we present extensive analyses of existing UMI data sets that contradict the assumptions of most existing pre-processing tools that they correct the technical noise of UMI data. Adding to the arguments of Chen (2018) and Townes (2019) that the UMI data is much cleaner than the read count data, our analyses demonstrate that the parametric model of the simplest form — Poisson — is sufficient to fully leverage the biological information contained in the UMI data. Moreover, the results suggest that explaining the cell-type heterogeneity should be the foremost step of the scRNA-seq analysis pipeline. Normalizing or cleaning the data set before clustering can lead to adversary consequences in the further analysis results. Therefore, we provide a new perspective on scRNA-seq data analysis by fully integrating the pre-processing and clustering, which was classified as part of the down-stream analysis. Finally, we introduce a novel software HIPPO (Heterogeneity Induced Pre-Processing tOol) that specializes in analyzing the UMI data.

Results

Drop-outs as Indicator of Biological Heterogeneity

Extensive analyses of scRNA-seq UMI data show the following three things. First, when cell heterogeneity is appropriately accounted for, almost all genes' zero proportions — proportion of cells that have zeros — align with the expected zero proportions under Poisson distribution. Second, the excessive zeros, or "drop-outs", are a better indicator of cell-type heterogeneity than gene variance, as once suggested in past literature [1]. Moreover, by selecting important genes based on the inflated zeros, the data can be effectively modeled using Poisson distribution, not Negative Binomial, Zero-Inflated Negative Binomial, or other more complicated models. Third, some immune-related genes are consistently noisier in PBMC data because they inherently have more heterogeneity across the cells. This suggests that it is practically impossible to fully account for the biological heterogeneity. Therefore, we suggest a "blacklist" of genes that are expected to be noisy for PBMC data sets. The following model reflects the new results.

$$p_g = \sum_{k=1}^{K_g} \pi_k e^{-\lambda_{kg}}$$

where g is gene index, K_g is the number of cell-types for given gene g, p_g is the proportion of cells with zero UMI count in gene g, λ_g is the true mean UMI count, and π_k is the proportion of cells for cell-type k. This model is set up assuming the Poisson model under complete homogeneity. However, we do not claim that all UMI counts follow Poisson distribution; we only claim that the zero proportions can be modeled using Poisson. All genes are tested for the hypotheses framework below, and those that are rejected are used for further downstream analyses.

$$H_0 : p_g = e^{-\lambda_g}, \tag{1}$$

$$H_A : p_g > e^{-\lambda_g} \tag{2}$$

Note that we do not fully leverage the mixture Poisson model for hypothesis testing, e.g. $H_A : p_g = \sum_{k=1}^{K_g} \pi_k e^{-\lambda_{kg}}$. Instead, we simply test whether zero proportion is inflated, which is always the case under cell heterogeneity due to Jensen's inequality. This makes the feature selection method robust to various parametric models, and the details are discussed in the Methods section.

Above model is significantly different from existing ones in several ways. First, it allows different number of cell-types for each gene, suggesting that cell-types are not well-defined biologically by a fixed number of genes, which is implicitly assumed by most existing methods by selecting one set of genes to cluster all the cells [2, 4, 11, 20]. For example, there is a collection of genes commonly known as house-keeping genes who are not differentially expressed across different cell-types, and hence for such gene g, $K_g = 1$. Meanwhile, immune-related genes are known to evolve very quickly [9, 10], and they might be more finely differentiated among the sampled cells. By assigning different numbers of grouping for each gene, the model acknowledge the dynamic nature of the cellular processes, and hence reflects underlying data

generating process better. Our proposed method is yet simpler to understand and easier to solve for latent variables than traditional pseudo-time inference, and computationally closer to cell-type clustering method, so it enjoys the advantage of both aspects (computation of clustering and interpretation of trajectory inference) of single-cell analysis.

Second, the model uses the proportion of zeros for each gene instead of using counts directly because the proportion of zeros (p_g) is enough of evidence of a gene for biological heterogeneity. The analysis shows that modeling p_g works just as well as modeling the distribution of all UMI counts such as Negative Binomial or Zero-Inflated Negative Binomial. Andrews (2018) [1] corroborates this claim by analyzing the sampling error. Moreover, this model ignores the genes that are highly expressed anywhere and deems them unimportant. In other words, if a certain gene g has high λ_g across all cells and hence no observed zeros ($p_g = 0$), the model does not reject the null hypothesis. This reflects the prior belief that any truly important biological marker should be expressed in some and not in other cells.

Third, the proposed model suggests that information contained in the single-cell UMI data is mostly biological and rarely technical. From analyses of real single-cell data, we observe that the low capture rate affects every UMI count equally, across different cell-types and even across different data sets. So, many past attempts in the past to "normalize" or "de-noise" the UMI counts ironically introduce noise to the data. Therefore, the proposed model provides a much more optimistic view of the UMI data analysis.

To support the claims above, various example data sets with true and inferred labels were analyzed, including those used in Duo (2018) [4], Freytag (2018) [6], and Zheng (2017) [23]. We also study the data sets from the 10X website with inferred labels. The description of each data set is in the references and also in the Supplementary Materials.

The initial motivation of the analysis is a well-known fact that zeros are induced by the low capture rate, so zero proportions are expected to be high when average counts $\hat{\lambda}$ are low. Hence, we plot the zero proportions twice, once with the entire cell population and once separately for each cell-type, to see whether the behavior or pattern is different in each cell type or in each data set. The black dots are plotted using heterogeneous cell populations, while all others represent the results from one cell-type. Figure 1 shows the result from three data sets created from Duo (2018) [4] based on the data sets in Zheng (2017) [23]. When the zero proportions computed from a heterogeneous cell population are distinctly higher than those computed from homogeneous cell populations. Moreover, in the case of homogeneous cell populations, most genes' zero proportions align with Poisson distribution $e^{-\lambda}$, while the same cannot be concluded with heterogeneous cell populations. This shows that high zero proportions are an indicator of biological heterogeneity. Meanwhile, the same does not apply to gene variance. When the gene variance is plotted against the gene mean, the mean-variance relationship is not very different between the two cases of homogeneous and heterogeneous cell populations. This means that genes with higher variance do not contain the signal for cell-type heterogeneity. We believe they contain more information about gene-specific characteristics, not the characteristics of the cell population.

The same analysis is repeated with 68K PBMC cells where true labels are unavailable, and the same phenomenon occurs where the proportions of zeros align with the expected Poisson curve ($e^{\hat{\lambda}}$)(Supplementary Materials) when plotted after cell grouping according to the inferred labels. However, one cell-type, CD34+, was particularly noisy with very high zero proportions, but when the cells were further separated into 3 sub-types, the zero proportions again aligned much closer to the Poisson exponential curve.

However, the data is still too noisy to conclude that every single gene follows the Poisson distribution. There are still some outliers left with particularly high number of zeros even when gene mean is reasonably high. When further investigated, certain gene types have shown to consistently inflated zero proportions. The last panel in Figure 1 shows the plot of the distribution of the difference between observed and expected zero proportions (under Poisson distribution). Note that there is a natural variance of zero proportions even under Poisson distribution. For example, a vast majority of the observed genes are categorized as "protein-coding genes", and so their zero proportions have a much wider range, but they are still centered at 0. The three gene types — HLA (human leukocyte antigen) genes, IG-C (immunoglobin constant) genes, and TR-C (T-cell receptor constant) genes — have distinctly more zeros than expected. These genes are immune-system related genes and are known to have much higher sequencing diversity, leading to higher cell-type heterogeneity. This result corroborates with the proposed hypothesis that cell heterogeneity is the main driver of zero-inflation. These results suggest that the three gene types be blacklisted for PBMC data sets. Due to lack of available data sets, there is no well supported conclusion for other tissues.

Table 1 shows an intuitive understanding of how cell heterogeneity drives excessive zeros. An outlier gene PPBP originally has a high zero proportion, but this phenomenon disappears after cell clustering. When the gene information was summarized under cell heterogeneity (cell-type CD34+), it had an average mean of 25.89 with a high zero proportion of 0.26. When this gene is tested for the null hypothesis that it follows Poisson distribution (detail in Methods), the z-score is more than 1 million. However, when separated into three subtypes, it is separated into cell groups where this gene is highly expressed with no zeros (subtype 2 and 3) and another cell group with low gene mean and high drop out rate (subtype 1).

All of the presented evidence is based on data sets from 10X. The same analysis does not hold true for other droplet-based methods such as in-Drop. The analyses are in the Supplementary Materials.

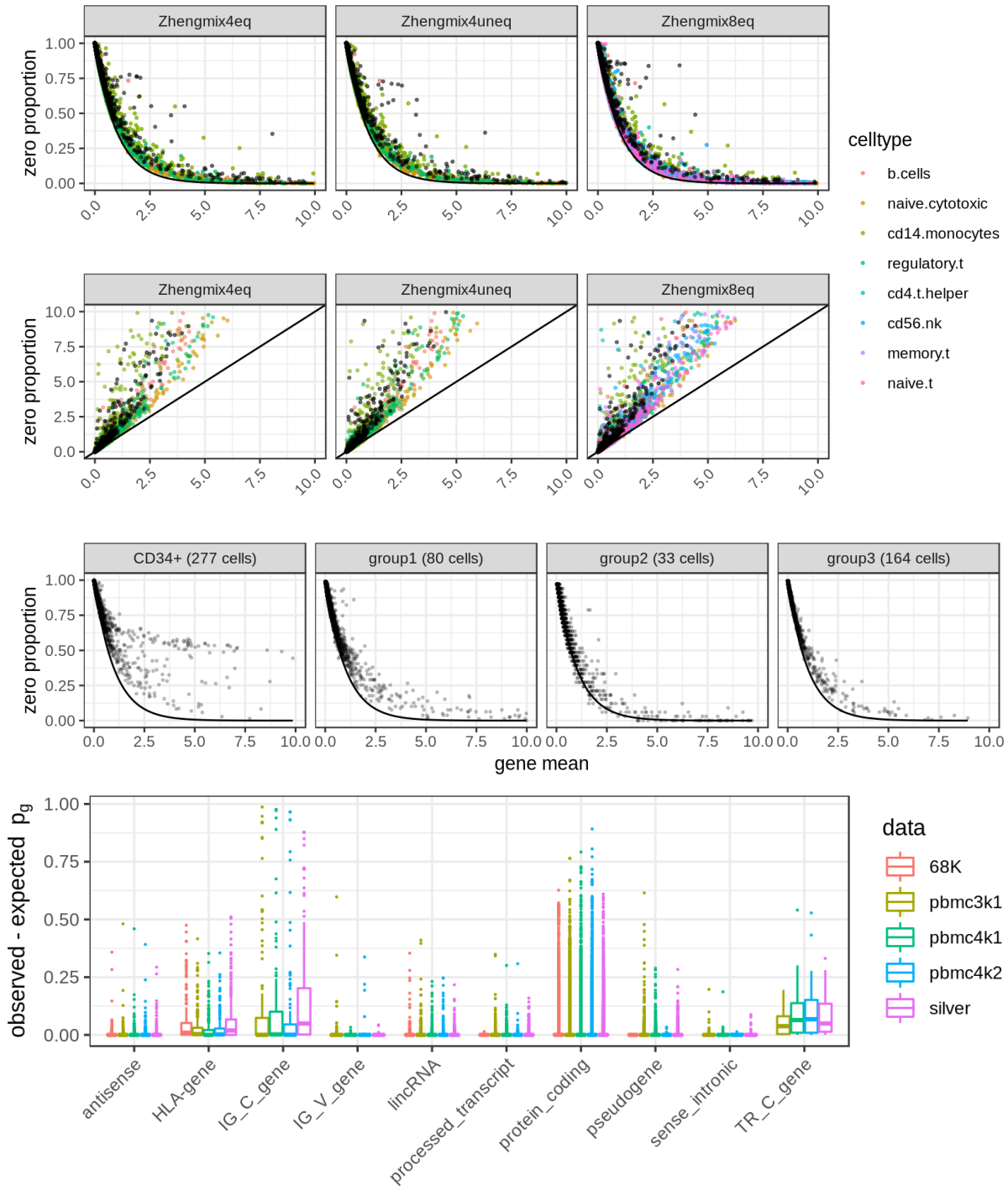

Figure 1: A. The black dots are plotted from a heterogeneous cell population where all cell-types are combined. In the first row, the black line represents the expected zero proportion under Poisson distribution. B. The left plot shows the relationship between zero proportion and gene mean. The next three plots show the same plot after further clustering of the cells. C. Distribution of the difference between expected and observed zero proportions in five different data of PBMC cells. Note that each gene type has a different number of genes so the distributions have different variances, i.e. there are more than 10,000 protein-coding genes for all 5 data sets while there are less than twenty IG C genes. A detailed plot and analysis are in Supplementary Materials.

Cell Population	gene mean	expected p	observed p	z-score
CD34+	25.89	5.69e-12	0.26	1838203
Subtype 1	0.5625	0.57	0.91	6.19
Subtype 2	22.36	1.93e-10	0	0
Subtype 3	38.96	1.25e-17	0	0

Table 1: Test result for gene PPBP before and after clustering within cell-type CD34+.

Limitations of Existing Pre-Processing Methods

Above analysis shows that the biological heterogeneity alone, without any other pre-processing steps, cleans the data so that the zero proportions of UMI data across different cell-types and data sets closely follow the expected zero proportions of Poisson distribution. Such results change the perspective of single-cell UMI analysis, especially in terms of pre-processing. Therefore, the first step of the UMI analysis should be accounting for cell-type heterogeneity. This section presents the negative consequence of three inappropriate pre-processing practices.

First, normalizing by size factor assumes that the sequencing depth for each cell is purely technical. However, sequencing depths are in fact confounded with cell-type, so any method based on the size-factor adjustment can obscure biological information. Additionally, multiplying UMI count by a pre-determined or inferred size factor destroys the Poisson structure. Next, similarly, sctransform regresses out the log of sequencing depth information and therefore inherits the previously mentioned issues. The Pearson residual from sctransform no longer has a clear distribution. Moreover, residuals may introduce nonexistent cell to cell variation, inflating the biological signals that do not exist. In case of "cleaned UMI" from sctransform output, the difference between cell-types is often compromised due to excessive cleaning. Lastly, Deep Count Autoencoder (DCA), a denoising tool using deep neural network with parametric assumptions, also leads to weaker signals of differential expression. When DCA is used after accounting for cell heterogeneity, differential expression tests show a much stronger signal compared to DCA applied before cell-type clustering. Therefore, de-noising the UMI data using DCA without clustering can lose biological information.

Several normalizing tools have been developed for UMI data, acknowledging the fundamental difference of the data-generating process between UMI data and read-count data [19]. The basic assumption for normalizing scRNA-seq data is to assume separate cell-specific effects (scaling-factor, technical) and gene-specific effects (gene mean, biological), so the most basic and popular protocol is dividing each UMI count by estimated scaling factors. However, as seen in Figure 2, forcing each cell to have the same UMI count destroys the Poisson structure. Dividing UMI counts by scaling factors does not change the zero-proportion but changes the gene mean, so the existing pattern of zero proportions following the exponential curve will no longer be true. More importantly, as shown in Figure 2, different cell-types are stratified when scaling-factors are adjusted. This means the original total count number of each cell contains valuable information about the cell-type, and forcefully deleting this information can induce problems in downstream analyses.

Next, sctransform suffers from similar consequences when applied to UMI data. Sctransform is one of the recent popular UMI analysis tools [7]. Its key idea is to assume Negative Binomial distribution for all UMI count, and regress out log(sequencing depth) from each cell. The underlying assumption here is similar to above, that the total count number of each cell is purely technical and must be removed from the data set. Therefore, sctransform destroys the Poisson structure as mentioned above. Moreover, two specific negative consequences of sctransform in downstream analysis are shown in Figure 2. The two markers for B cells are specifically studied [17], and the distributions of their raw UMI count, cleaned UMI count, and the residuals are observed from sctransform in two cell-types (B cells and monocytes) in three different data sets (Zhengmix4eq, Zhengmix4uneq, Zhengmix8eq). The software recommends the use of residuals for the downstream analysis. However, the residuals introduce cell-to-cell heterogeneity that does not exist in the raw data. The marker MS4A1 is not expressed at all in monocytes in two data sets (Zhengmix4eq and Zhengmix8eq), but the residuals suggest that there are some variations of the expression across the monocyte cells. This can easily lead to inflated biological signals when the true label of cell-types does not exist. Meanwhile, in the case of cleaned UMI of sctransform, the difference between the distribution of UMI count between two cell-types drastically decreases. Especially, the UMI counts of B cells are in general deflated, which is a by-product of removing the outliers of sequencing depth. Inappropriate processing through sctransform can therefore poor inference about biological signals.

Another dangerous practice of pre-processing is de-noising before accounting for cell heterogeneity. The existing de-noising tools blur some of this information especially through shrinkage because it regularizes each cell to resemble one another. Specifically, shrinkage can reduce the difference between each cell-type, and this can have a negative impact on differential expression analysis. Differential expression is examined under two scenarios: imputing the heterogeneous cell population and then comparing the expression, and imputing homogeneous cell population separately and then comparing the expression. Naive T cells and regulatory T cells are selected because they are similar to each other so that many clustering algorithms struggle to differentiate. Four known markers are specifically investigated: CD4, CTLA4, FOXP3, and IL2RA [17]. Figure 2 shows that in Zhengmix8eq data, the imputed data (heterogeneous cells) completely misses the signals for CD4, CTLA4, and FOXP3 by showing p values greater than 0.05. The log fold change values are also much greater for imputing cell-types separately rather than imputing before clustering.

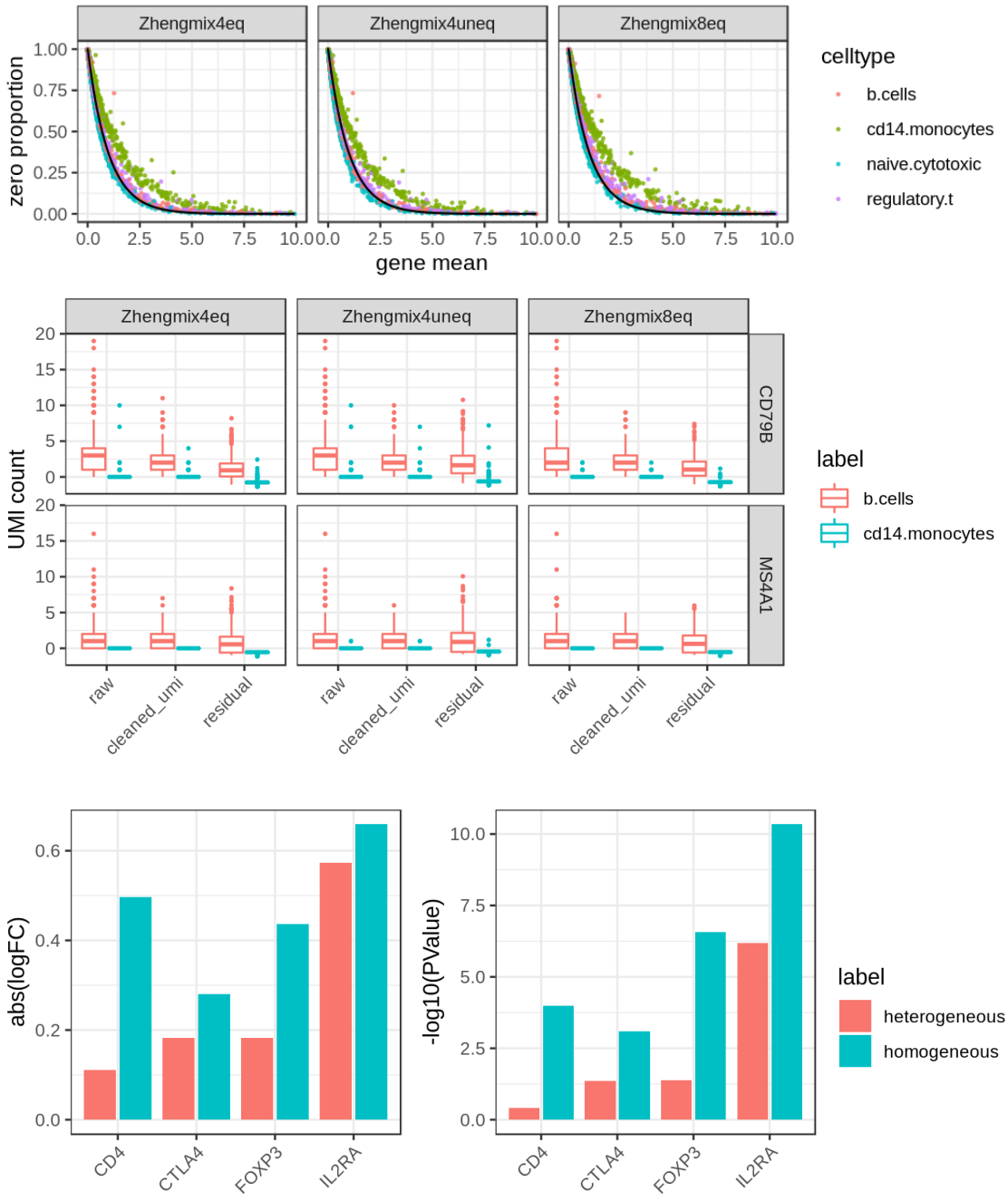

Figure 2: A. For each data set, for each cell, the sequencing depth (sum of UMI counts) was normalized to be the median sequencing depth of all the cells. Then, each gene's zero proportion across cells in a given cell-type is plotted against the mean UMI count, after adjusting for the size factor. The black curve is the expected exponential curve of Poisson distribution. B. Distribution of UMI counts before and after sctransform, both cleaned UMI and the residuals, for two marker genes for B cells, CD79B and MS4A1. C. Effects of DCA on homogeneous cell population and heterogeneous cell population. Log of fold change and log of p-values are plotted for three distinct data sets.

HIPPO: Heterogeneity-Induced Pre-Processing Tool

Based on the above results, we provide a new perspective on the preprocessing of single-cell data: Heterogeneity-Induced Pre-Processing tOol (HIPPO). The first and foremost step is to account for the cell heterogeneity, and we achieve this by simultaneously selecting features at each round of hierarchical clustering. We believe this is the most critical first step to clean the UMI data before downstream analyses or even de-noising. The clustering performance of the proposed method is comparable to Seurat and SCtransform in many data sets and often better with less computational cost.

The feature selection method is a hypothesis test from (2) that can select genes that have strong indications of cell heterogeneity. Under the null hypothesis in (2), we can derive the null distribution of the test statistic of dropout rate p, and subsequently get the z-score using the null distribution. Users can assign either the z-score cut-offs or the number of genes to select.

The selected features are used to cluster the cells into 2 groups using PCA + K-means. Past works have proposed ways to reduce the dimension of Poisson data such as generalized PCA, rather than regular PCA that is optimized for Gaussian data [13], but in practice, the performance of generalized PCA was quite similar to log transformation + regular PCA . Then, the within-group variability is preserved for each cluster, and the group with the most intra-variability is selected, assigned for next round of clustering. The feature selection and clustering steps are iteratively repeated until one of the two ending criteria are met: K round of clustering for pre-determined K, or there are no more significant genes whose z scores exceed the pre-determined threshold.

The software returns the clustering result for each round $k = 2, \cdots, K$. Selecting the final number of groups is a difficult problem without any prior knowledge, so we offer another natural stopping criterion; HIPPO will stop the clustering procedure as soon as the number of zero-inflated genes is less than a certain percentage of the genes.

Figure 3 shows the results of HIPPO for its each step. In Zhengmix4eq data with 4 types of PBMC cells, it successfully classifies Monocytes, B cells, and Regulatory T cells from Naive T cells in the respective order. This order reflects the amount of biological heterogeneity of each cell type — it is commonly observed that clustering monocytes is easy, while classifying different types of T cells (naive T cells and regulatory T cells) are more difficult. Since the proposed method uses different gene set to do each round of clustering, HIPPO leverages the relevant information well without getting noise from irrelevant genes.

HIPPO is computationally very cheap mainly because we use less and less features for each round of clustering, and it is based on a simple Poisson distribution. When the data set runs out of features, the computing time plateaus, as shown in Figure 3. This provides a natural stopping point of K, although it will lean towards higher K than the known cell types. HIPPO will detect genes that differentiate cells even in the same cell-type.

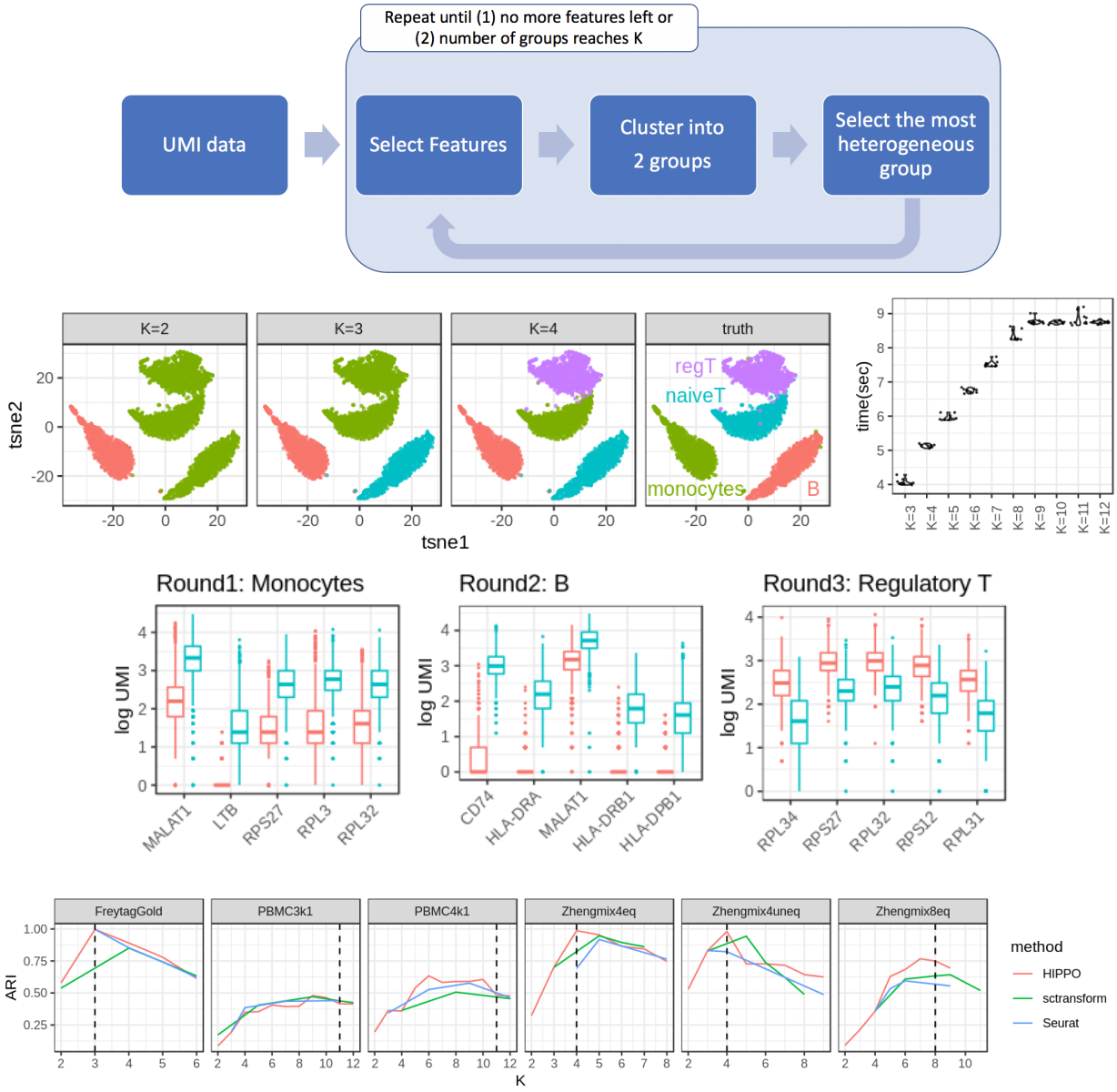

Figure 3: HIPPO framework applied to Zhengmix4eq data. A. Summary of framework B. t-SNE plot for each number of cluster k juxtaposed to the true label result. C. Distribution of computing time of HIPPO, run 10 times for each K. D. Example of differential expression in each round. The blue boxplot is from each cell type in the title, and the red is the rest of the cells. On the third round, red boxes represent the naive T cells, blue regulatory T cells. E. Clustering results compared with sctransform and Seurat using adjusted rand index

Discussion

The presented results provide a new perspective on the analysis of sigle cell UMI data sets. Through tackling the cell-type heterogeneity as the first step, the proposed method leads to a more reliable downstream analyses. Moreover, through a streamlined feature selection method that reflects the dynamic nature of cellular process, the method overall provides a computationally and mathematically simple analysis tool. The method is implemented in the R toolkit HIPPO. The results confirm the claims of recent literature [3] that a different tool must be applied to the UMI data set from the tools for read count data set. UMI data set is free from amplification bias so that the level of technical noise is much lower.

All of the analyses were performed on data sets sequenced from 10X, but there are other droplet-based methods. The conclusions from 10X data do not extend to data sets from in-Drop or Drop-seq [12, 24], and the noise level was too high to assume the zero proportions follow the exponential curve relative to the gene mean.

Methods

Poisson Mixture Model

To understand the behavior of the zeros for each gene, the first step is to reduce the information from each gene to the proportion of zeros across the cells

$$\hat{p}_g = \sum_c \frac{\mathbb{1}_{X_{cg}=0}}{C} \tag{3}$$

which is an estimator for the true zero proportion of gene g: p_g. We study its relationship against the mean expression for the set of cells, because p_g would decrease as the expression level increases. With the test statistic above, we test a one-sided hypothesis for each gene g, whether the zero proportion is higher than the expected rate under the Poisson model. For the alternative hypothesis, we believe that UMI counts follow finite Poisson Mixture. The hypotheses for each gene g are formally specified below.

$$H_0 : p = e^{-\lambda_g}, \qquad H_A : p = \sum_{k=1}^{K_g} \pi_k e^{-\lambda_{kg}}$$

In practice, we re-frame the hypotheses as $H_0 : K_g = 1, H_A : K_g > 1$ when $p = p = \sum_{k=1}^{K_g} \pi_k e^{-\lambda_{kg}}$. In other words, we claim that noisy gene means there is cell heterogeneity across the samples. If the cell population is truly homogeneous, the count data follows Poisson data with expected zero proportion, $e^{-\lambda_g}$.

Chen (2018) demonstrates that most genes in UMI data follow Poisson distribution [3] while other noisy genes follow Negative Binomial or Zero-Inflated Binomial distribution. Such model, although fundamentally different, is closely tied to the Poisson mixture model because Negative Binomial is the limiting distribution of Gamma-Poisson. If

λ_{cg} for each cell is drawn independently from the gamma distribution $\Gamma(r_g, \frac{1-p_g}{p_g})$, then $\sum_{c=1}^{C} X_{cg} \sim \frac{1}{C} Pois(\lambda_{cg}) \Leftrightarrow X_{cg} \sim NB\left(r, \frac{1-p_g}{p_g}\right)$. While Negative Binomial assumes a continuous mixture of Poisson, the proposed model assumes a finite mixture of Poisson, which is simpler and directly explains the source of zero inflation.

In practice, we do not explicitly estimate π_k, but instead simply test if observed $\hat{p}_g$ is larger than expected p with estimated gene mean λ. ($H_A : p_g > e^{-\lambda_g}$). It might seem counterintuitive that this test statistic does not fully leverage the specification of the alternative hypothesis; we never estimate the mixture parameters π_k. Alternatively, for example, one might alternatively suggest that we can conduct a likelihood ratio test of Poisson versus Poisson mixture. The main strength of the proposed reduced test statistic is its robustness to the modeling assumptions. Table 2 shows that the proportion of zeros are always larger than expected under different alternative hypotheses. Under the proposed alternative, mixture of Poisson, the proportion of zeros under the null hypothesis would be $e^{-\lambda}$ where λ is the weighted mean of the gene mean for each cell-type. Due to Jensen's inequality, p under alternative hypothesis is always greater than that under the H_0. The same conclusion does not hold when we look at gene variance in Table **??**, again suggesting that zero proportion is a better feature selection criterion.

Underlying Distribution	p under H_0	p under H_A
Mixture of Poisson	$e^{-\lambda} = e^{-\sum_k \pi_k \lambda_k}$	$\sum_k \pi_k e^{-\lambda_k}$
Negative Binomial	$e^{-\lambda}$	$\left(\frac{r}{r+\lambda}\right)^r$
Zero-inflated Negative Binomial	$e^{-\lambda}$	$\left(\frac{r}{r+\lambda}\right)^r + \pi_0$

Table 2: The alternative hypothesis $H_A : p_g > e^{-\lambda}$ is robust to different model hypotheses. In the first row, the right column is larger than the left column due to Jensen's inequality. For negative binomial, the dispersion parameter r is constructed so that the variance is $\frac{\lambda^2}{r} + \lambda$, so that Poisson is a special case of Negative Binomial with $r = \infty$. The zero inflated negative binomial distribution is parameterized as $\pi_0 \delta_0 + (1 - \pi_0)$ NB(λ, r)

Alternative Hypothesis	variance under H_0	variance under H_A
Mixture of Poisson	λ	$\sum_{k=1}^{K} \pi_k^2 \lambda$
Negative Binomial	λ	$\frac{\lambda^2}{r} + \lambda$
Zero-inflated Negative Binomial	λ	$(1 - \pi_0)^2 \left(\frac{\lambda^2}{r} + \lambda\right)$

Table 3: High gene variance is not a good indicator of cell type heterogeneity under some of the alternative hypotheses. Only under the negative binomial distribution, gene variance is higher than λ with any nonnegative dispersion parameter.

Feature selection and Inference

For gene g with count data for cells $c = 1, \cdots, C$, we define an estimate for the proportion of zeros $\hat{p}_g$. Below is our test statistic and the null hypothesis where $\bar{X}_g$ is the average counts, the maximum likelihood estimator of the true mean λ_g.

$$\hat{p}_g = \frac{\sum_{c=1}^{C} \mathbb{1}_{X_g=0}}{C}, \quad \hat{\lambda} = \bar{X}_g \quad H_0 : p_g = e^{-\lambda_g} \tag{4}$$

We have the following results from the above set-up.

$$\bar{X}_g \sim \mathcal{N}(\lambda_g, \lambda_g/C) \Rightarrow \hat{p}_g = e^{-\bar{X}_g} \sim \log\mathcal{N}(-\lambda_g, \lambda_g/C)$$

$$E(e^{-\bar{X}_g}) = e^{-\lambda_g + \frac{\lambda_g}{2C}} = \frac{2C-1}{2C}p, \quad Var(e^{-\bar{X}_g}) = (e^{-\lambda_g} - 1)e^{-2\lambda_g + \frac{\lambda_g}{C}} = \frac{2C-1}{2C}p(1-p)$$

$$E(\hat{p}_g) = p_g, \quad Var(\hat{p}_g) = \frac{p_g(1-p_g)}{C}$$

We can get the p value from the null distribution below

$$\hat{p}_g - \frac{2C}{2C-1}e^{-\bar{X}} \sim \mathcal{N}\left(0, \frac{p_g(1-p_g)}{C-0.25}\right)$$

including the z-statistic.

Hierarchical Model

Algorithm 1 Cell-Type Hierarchical Clustering

K: upper limit of cluster number
z-threshold: threshold for feature selection
$\ell = 1$
for $k = 2, \cdots, K$ **do**
  if No genes exceed z threshold **then**
    no more important features; terminate algorithm
  else
    update the matrix by selecting new features
    separate cells with label ℓ into two groups, one with label ℓ and another with label k, using PCA + kmeans
    update ℓ = cluster with the highest intra-cluster distance
  end if
end for
return cluster labels for each k

Differential Expression

After the group labeling, we can do a similar but simpler hypothesis test to see if a certain gene is differentially expressed in two groups.

$$X_{cg}|c \in \mathcal{C}_1 \sim \text{Poisson}(\lambda_1), \qquad X_{cg}|c \in \mathcal{C}_2 \sim \text{Poisson}(\lambda_2)$$

$$H_0 : \lambda_1 = \lambda_2$$

We can use a 2-sample t-test and order the genes in the order of significance in the mean difference of two groups.

$$t = \frac{\bar{X}_{\mathcal{C}_1 g} - \bar{X}_{\mathcal{C}_2 g}}{\sqrt{\frac{\bar{X}_{\mathcal{C}_1 g}}{|\mathcal{C}_1|} + \frac{\bar{X}_{\mathcal{C}_2 g}}{|\mathcal{C}_2|}}}$$

References

[1] Tallulah S Andrews and Martin Hemberg. M3drop: Dropout-based feature selection for scrnaseq. *Bioinformatics*, 2018.

[2] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411, 2018.

[3] Wenan Chen, Yan Li, John Easton, David Finkelstein, Gang Wu, and Xiang Chen. Umi-count modeling and differential expression analysis for single-cell rna sequencing. *Genome biology*, 19(1):70, 2018.

[4] Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7, 2018.

[5] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):390, 2019.

[6] Saskia Freytag, Luyi Tian, Ingrid Lönnstedt, Milica Ng, and Melanie Bahlo. Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data. *F1000Research*, 7, 2018.

[7] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *bioRxiv*, page 576827, 2019.

[8] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539, 2018.

[9] Austin L Hughes. Rapid evolution of immunoglobulin superfamily c2 domains expressed in immune system cells. *Molecular biology and evolution*, 14(1):1–5, 1997.

[10] Laurence D Hurst and Nick GC Smith. Do essential genes evolve slowly? *Current biology*, 9(14):747–750, 1999.

[11] 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.

[12] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[13] Yoonkyung Lee. Generalized principal component analysis. *Journal of Educational Psychology*, 24(6):417–441, 2015.

[14] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[15] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[16] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[17] Max Schelker, Sonia Feau, Jinyan Du, Nav Ranu, Edda Klipp, Gavin MacBeath, Birgit Schoeberl, and Andreas Raue. Estimation of immune cell content in tumour tissue using single-cell rna-seq data. *Nature communications*, 8(1):2032, 2017.

[18] 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.

[19] Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods*, 14(6):565, 2017.

[20] 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(4):414, 2017.

[21] Jingshu Wang, Divyansh Agarwal, Mo Huang, Gang Hu, Zilu Zhou, Chengzhong Ye, and Nancy R Zhang. Data denoising with transfer learning in single-cell transcriptomics. *Nature methods*, 16(9):875–878, 2019.

[22] Jingshu Wang, Mo Huang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, John Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Gene expression distribution deconvolution in single-cell rna sequencing. *Proceedings of the National Academy of Sciences*, 115(28):E6437–E6446, 2018.

[23] 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.

[24] Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nature protocols*, 12(1):44, 2017.

Supplementary Materials

- Table of example data sets
- Same curve + plots in many more data sets

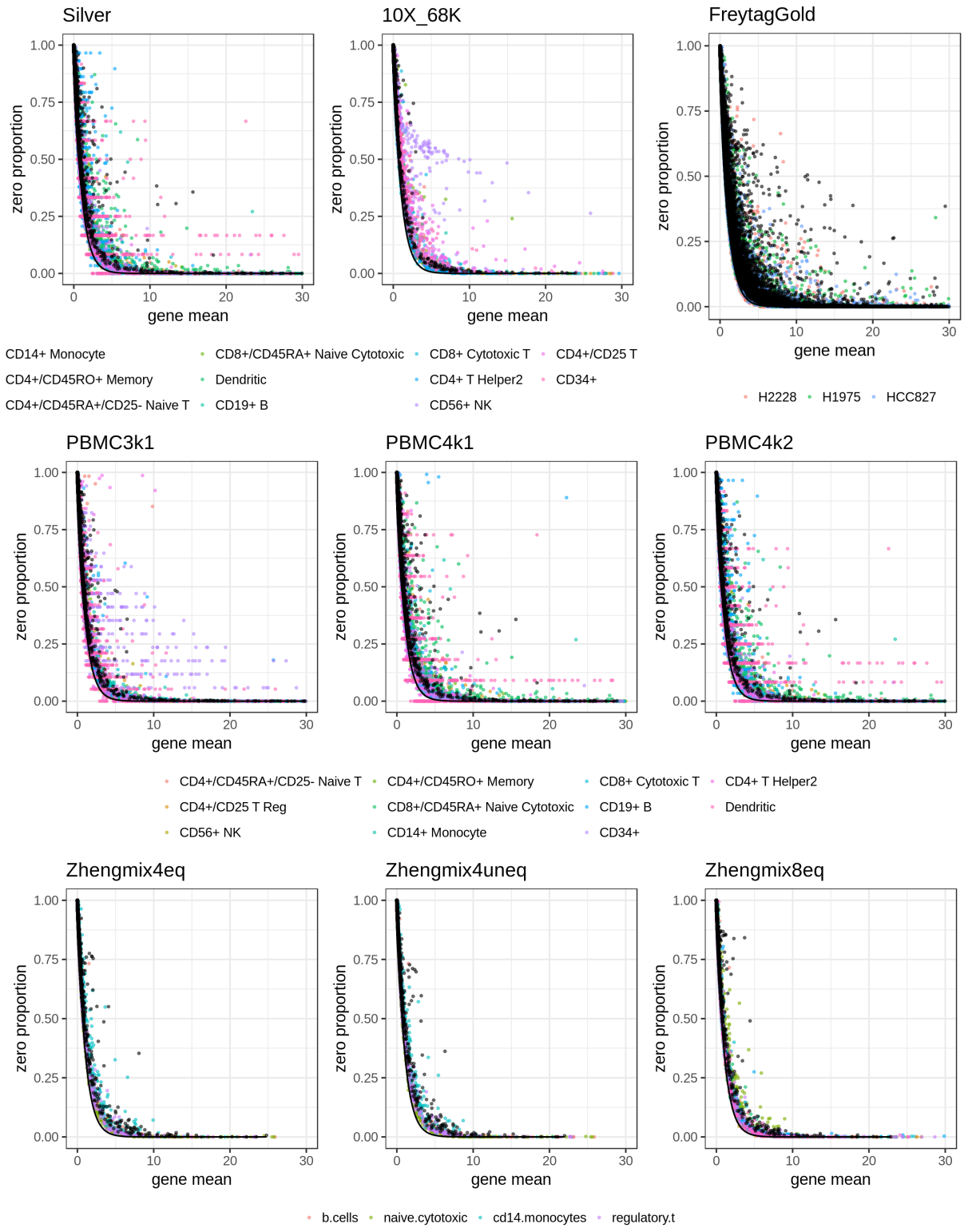

Figure 4: Zero proportion against gene mean for diverse UMI data set.

- Outlier genes: table of sample sizes for each gene type

	68K	pbmc3k1	pbmc4k1	pbmc4k2	silver
antisense	12013	1490	1749	1707	1239
HLA-gene	226	277	267	202	278
IG_C_gene	0	120	109	105	115
IG_C_pseudogene	0	18	20	19	8
IG_J_gene	0	2	0	0	1
IG_V_gene	0	204	189	196	182
IG_V_pseudogene	0	8	6	6	1
lincRNA	10323	1045	1245	1178	900
miRNA	0	2	11	0	1
misc_RNA	16	555	547	0	197
Mt_rRNA	0	22	22	0	22
Mt_tRNA	0	45	88	0	29
polymorphic_pseudogene	0	51	65	47	57
processed_transcript	27	620	666	302	612
protein_coding	131801	106551	112902	113239	103162
pseudogene	9	7830	8662	77	4016
rRNA	0	47	44	0	24
sense_intronic	0	223	237	11	124
sense_overlapping	0	12	16	0	14
snoRNA	14	92	83	0	30
snRNA	0	325	437	0	154
TR_C_gene	0	52	55	55	54
TR_V_gene	0	312	330	336	241
TR_V_pseudogene	0	5	13	16	7

Table 4: Gene counts for each data set and each gene type

- DCA results in other cell types, SAVER results in the same cell type

Figure 5

- Comparison of ARI with other clustering methods (SC3, scater, SIMLR) (+ computation time)

- in-drop & drop-seq result (Macosko and Baron)

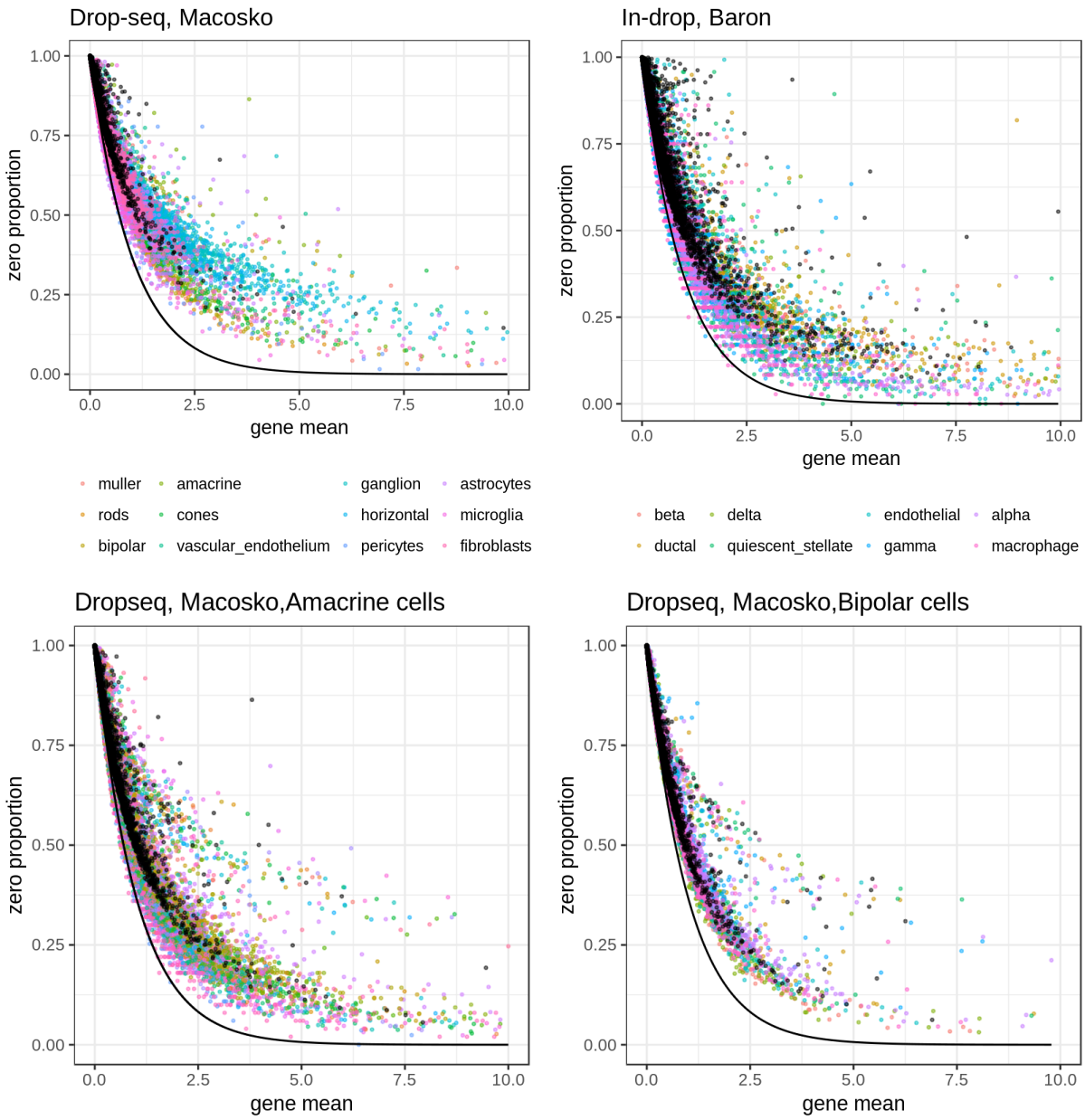

Figure 6: Results from Drop-seq and In-drop.