Kim *et al. Genome Biology* (2020) 21:196
https://doi.org/10.1186/s13059-020-02096-y

Genome Biology

RESEARCH Open Access

Demystifying “drop-outs” in single-cell UMI data

Tae Hyun Kim[1], Xiang Zhou[2*] and Mengjie Chen[3*]

*Correspondence:
xzhousph@umich.edu;
mengjiechen@uchicago.edu
[2]Department of Biostatistics, University of Michigan, Ann Arbor, USA
[3]Department of Human Genetics and Department of Medicine, University of Chicago, Chicago, USA
Full list of author information is available at the end of the article

Abstract

Many existing pipelines for scRNA-seq data apply pre-processing steps such as normalization or imputation to account for excessive zeros or “drop-outs." Here, we extensively analyze diverse UMI data sets to show that clustering should be the foremost step of the workflow. We observe that most drop-outs disappear once cell-type heterogeneity is resolved, while imputing or normalizing heterogeneous data can introduce unwanted noise. We propose a novel framework HIPPO (Heterogeneity-Inspired Pre-Processing tOol) that leverages zero proportions to explain cellular heterogeneity and integrates feature selection with iterative clustering. HIPPO leads to downstream analysis with greater flexibility and interpretability compared to alternatives.

Introduction

Droplet-based single cell RNA-sequencing (scRNA-seq) methods have changed the landscape of genomics research in complex biological systems [1–4] by producing single-cell resolution data at affordable costs. In the state-of-the-arts protocols, a step called barcoding unique molecular identifiers (UMI) has been introduced to remove amplification bias and further improve data quality [5]. Some literature [6–8] suggests that barcoding leads to a different data structure from read count data structure but many tools remain to not acknowledge the difference between the count data produced with and without barcoding.

Many pipelines have been built for scRNA-seq UMI data analysis. Despite subtle differences in these pipelines, the general order of a scRNA-seq analysis is as follows: quality control (filtering), cleaning (normalization, imputation, de-noising, batch-correction, etc.), feature selection which often involves dimension reduction, and downstream analysis such as clustering and lineage analysis [9]. In this paper, we do not discuss filtering and focus on the later three steps. First, the challenge of scRNA-seq data cleaning has led to the development of a wide array of tools. Some methods adjust for sequencing depths using size factors [10, 11]. Some impute the reads directly using a zero inflated model, to reduce the noise from drop-outs [12]. Some try to de-noise the entire data set by fitting parametric models, where one example is sctransform that uses the residuals from negative binomial regression [13] and another example is SAVER that uses Poisson LASSO regression [14]. Despite the diversity of proposed methods, the general consensus has been reached to use one of the following distributions to model the counts: Poisson, negative binomial, or zero-inflated negative binomial distribution. Second, methods for feature selection have been less controversial. Most tools use some form of gene variance to mean ratio to identify genes that are highly dispersed, where the dispersion level is interpreted as a signal of biological heterogeneity [6, 10, 13]. Another less recognized approach is to use the zeros in the read or UMI counts; genes with inflated zeros are interpreted as biologically important signals [15]. Lastly, after data cleaning and feature selection, the pre-processed data will then be piped into downstream analysis tools for clustering analysis [16–18], trajectory inference [19, 20], or differential expression analysis [21, 22]. Currently, pre-processing and downstream analysis have been mostly considered as separate and consecutive steps [10, 14, 16, 23].

Here, we present extensive analyses of publicly available UMI data sets that challenge most existing pre-processing tools' assumption, mainly that pre-processing is a necessary step before feature selection and downstream analysis. Our results suggest that clustering, or resolving the cell heterogeneity, should be the foremost step of the scRNA-seq analysis pipeline, not as part of the downstream analysis. Normalizing or imputing the data set before resolving the heterogeneity can lead to adverse consequences in downstream analysis. Adding to the arguments that the UMI data is much cleaner than the read count data [6, 8], our analyses demonstrate that the simple Poisson distribution is sufficient to fully leverage the biological information contained in the UMI data if the cell-type heterogeneity has been appropriately accounted for. As a result, we provide a new perspective on scRNA-seq data analysis by integrating the pre-processing step and clustering, which was classified as part of the down-stream analysis. The proposed procedures have been implemented in software HIPPO (https://github.com/tk382/HIPPO). HIPPO leverages zero proportions to detect different levels of cell-type heterogeneity in each gene and can be particularly useful for low UMI data sets with excessive zeros, such as typical datasets generated from 10X protocols.

Results

Demystifying drop-outs

We started by exploring zero detection rates in three UMI datasets generated by 10X protocols for both homogeneous and heterogeneous cell populations. Taking a subset of data in Zheng [3] as created in Freytag [17] as an example, we computed zero proportions, defined as the proportion of cells with zero counts per gene, across 15,568 genes, in CD19+ B cells, CD4+/CD25 regulatory T cells, and combined. The obtained statistics were plotted against gene-level average count and were compared with expected zero proportions under the Poisson, negative binomial, and zero-inflated negative binomial distribution, respectively (Fig. 1a). For a homogeneous cell population, we observe most genes align well with the expected curve under the Poisson assumption. Few genes can benefit from using the negative binomial model to account for extra dispersion from the Poisson, but our results strongly suggest that to model the drop-outs by introducing an extra zero-inflation component by the zero-inflated negative binomial distribution is

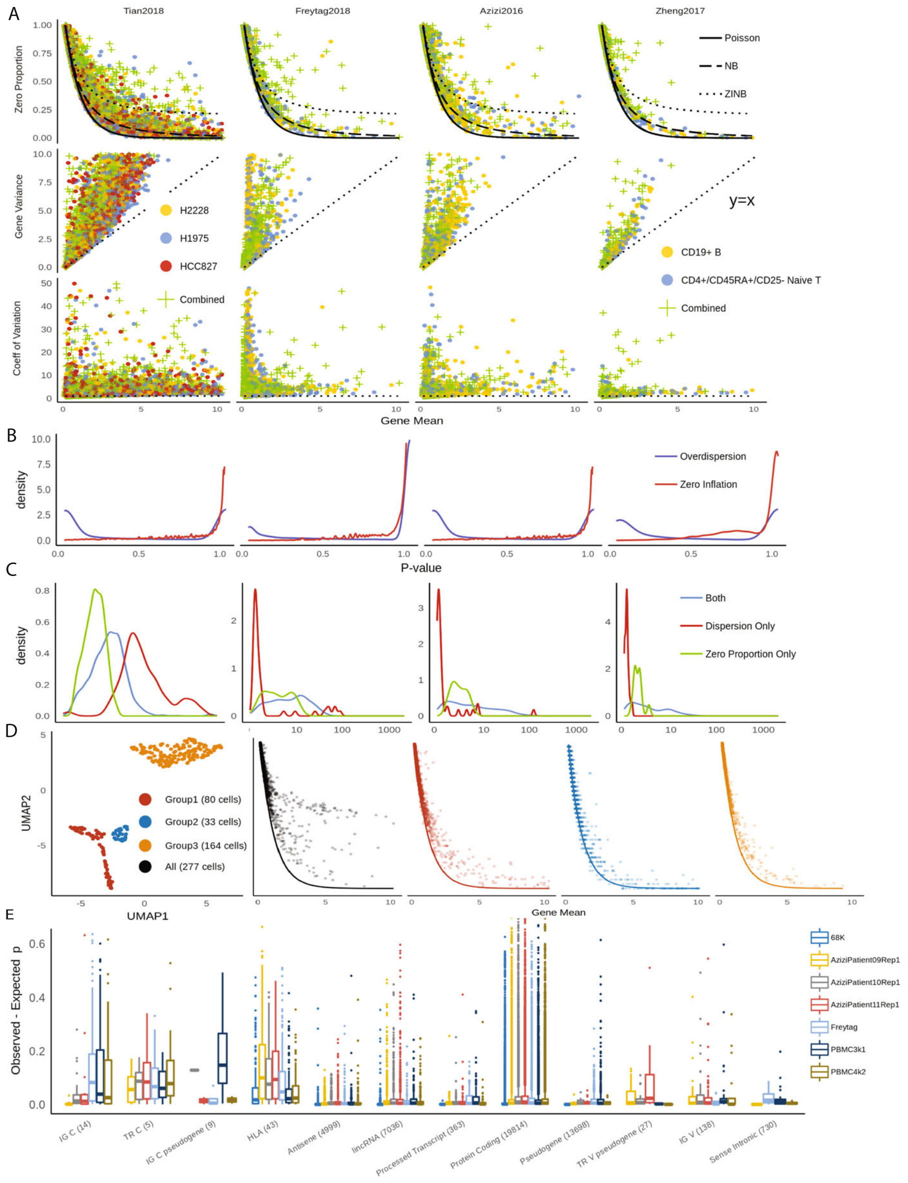

Fig. 1 **a** Comparisons of zero proportion, gene variance, and CV as indicators for cellular heterogeneity in different UMI data sets. **b** Distributions of *p* values from likelihood ratio test for over-dispersion and zero-inflation. **c** Comparisons feature selection using likelihood ratio test and zero proportions. LR test tends to select the genes that have mean UMI count close to 0. **d** t-SNE plots of CD34+ cells in Zheng data, and relationship between zero proportions and gene means before (black) and after (colors) clustering of CD34+ cells. **e** Distributions of zero inflation in different PBMC data sets. The *x*-axis labels represent gene types from GENCODE annotations and the number of genes within each type

unnecessary. For example, in Zheng dataset, 257 genes out of 5,568 genes would benefit from negative binomial modeling (*p* values pass Bonferroni criterion in likelihood ratio test at 0.05 type I error level), but no gene would benefit from extra zero inflation parameter. The *p* values are not calibrated to the uniform distribution because there are many genes that have UMI count of 1 in one cell and 0 in everywhere else, in which case *p* value is close to 1 (Fig. 1b). This result shows that drop-outs are within the range of natural Poisson sampling noise in UMI data for a homogeneous cell population, and they do not introduce excessive zero inflation, which is contradictory to prevalent opinions [11, 16, 24, 25]. Extra zero inflation can be measured by comparing observed zero to expected zero counts under Poisson distribution within a homogeneous cell population (Methods). Through the following analysis, we show zero proportions are as effective measures for cell-type heterogeneity as other widely used alternatives, gene variance, coefficient of variation (CV), or dispersion parameter in negative binomial distribution [15]. It provides simplicity and interpretability in particular for data sets with low UMI counts and reasonable number of zeros, as zero-inflation is meaningless when no zero is observed.

Analysis in multiple UMI data sets shows that zero proportions in most genes can be effectively modeled by the Poisson distribution, as more than 95% of absolute z values (Methods) are below 2. For mixed cell types, zero proportions considerably deviate from expected values under the Poisson model, as only less than 30% of the genes have z values below 2. This shows that the zero inflation test is an effective way to find genes that contribute to cellular heterogeneity. On the contrary, gene variance of mixed cell types does not always surpass those of a single cell type. In Zheng data, 62% of the genes had higher variance in pure naive cytotoxic cells than in mixed PBMC cells. On average, gene variance is similarly distributed for homogeneous and heterogeneous cell populations (Additional file 1: Table S2 and Figure S8). Therefore, the gene variance is rather more of a gene-specific characteristic while being less informative about the characteristics of the entire cell population. CV, on the other hand, suffers from an inherent numerical instability issue when gene mean is close to 0, because when mean is close to 0, CV estimates have high variability. Another popular option is to conduct model selection to assign genes to one of three candidate distributions of Poisson, NB, and ZINB, but measuring overdispersion also suffers from a similar problem in selecting biologically meaningful genes [6, 26]. When we used statistics from likelihood ratio test and select top genes from the resulting statistics, the selected genes were different from those selected when we used zero proportion (Fig. 1c). For three data sets of [3, 17, 27] (median sequencing depth of 4371, 1298, and 2393.5 respectively), the likelihood ratio test selects genes that are overly focused on those with mean close to 0. Intuitively, the dispersion parameter scales with the square of the ratio of gene mean to the gene variance and in nature very similar to CV. These genes with very low mean are likely to have little information about the cells. Still, dispersion parameter can be more useful than zero-inflation statistic when the data set has high UMI counts, so deviance has been implemented in HIPPO as an alternative feature selection method (Additional file 1: Figure S9, S10).

We expand the data to study all 68,579 cells from Zheng dataset [3]. When we aligned the zero proportions with the expected Poisson curve according to the provided cell type labels: CD14+ monocyte, CD19+ B, CD34+, CD4+ T helper2, CD4+/CD25 T Reg, CD45RA+/CD25- naive T, CD4+/CD45RO+ memory, CD56+ NK, CD8+ cytotoxic T, CD8+/CD45RA+ naive cytotoxic, and dendritic cells. Most of these cell types look relatively homogeneous. However, one cell type, CD34+, was particularly noisy with very high zero proportions, indicating cellular heterogeneity (Fig. 1d). Based on the diagnosis from t-SNE plots, we identified three subtypes within the CD34+ cells. The alignment of zero proportions against the Poisson curve was immediately improved according to the inferred subtype labels. This indicates the effectiveness of zero proportions as metrics to evaluate cellular heterogeneity and their potentials to discern cell types.

We further checked how zero proportions could be dispersed from the Poisson distribution for genes with various functional annotations across all PBMC datasets (Fig. 1e). Specifically, we calculated the difference between observed zero proportions and expected proportions (under Poisson) for each functional group using reference data from GENCODE of GRCh38 [28]. The vast majority of genes are categorized as "protein-coding genes." Their zero proportions cover a wide range from 0 to 0.7, but centered at 0, indicating variability in zero proportions but no systematic inflation of zero proportions. In contrast, immune-related genes are consistently zero-inflated, with the interquartile range as high as 10 to 20%. The enrichment analysis (Additional file 1: Table S4) shows that immune-related genes have significantly higher proportion of zero-inflated genes compared to genes that are not related to immune function. The top-ranked annotations for zero inflation include IG C genes, TR C genes, and HLA genes. IG C genes are immunoglobulin genes of the constant (C) region, while TR C genes are T cell receptor genes of the constant region, and hence, both gene types are deeply connected to immune system. HLA genes are genes in human leukocyte antigen system that is responsible for the regulation of the immune system. Genes involved in immune functions are expected to be inherently heterogeneous [29]. For example, HLA genes are highly polymorphic than others, and TR-C genes go through VDJ recombinations that lead to more diverse sequences across cells. This result corroborates with the notion that cellular heterogeneity is the main driver of zero inflation. Higher level of heterogeneity in immune genes explain the past studies' results that even within one cell type, there are zero-inflated genes [30]. The high heterogeneity of cells in certain genes suggests that it is difficult, to say the least, to fully account for the cell type for every gene; as the cells are finely clustered, a point will be reached where the number of cells left in each cluster is not enough to do statistically meaningful analysis. Therefore, we use the following stopping criteria for iterative clustering where the procedure stops when the number of genes with zero-inflated is less than a certain threshold. This criterion is designed to take into account the remaining granular biological heterogeneity in certain genes that cannot be fully resolved through cell-type.

Zero inflation test for cellular heterogeneity

Based on the above observations, we propose a new feature selection strategy that uses detected zero proportion of a given gene as the statistic to test for cellular heterogeneity. Under the null hypothesis, where complete cellular homogeneity is assumed, the proportion of zeros is equal to the expected zero proportion under Poisson distribution. Under the alternative hypothesis, zero proportion is inflated, as if the count data follows mixture of Poisson (Methods). Formally, our framework can be presented as follows:

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

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

where g is gene index and λ_g is the mean UMI count for gene g. The above testing framework is based on an assumption that whether UMI count being 0 follows the Bernoulli distribution. Test statistic z_g follows a standard normal under the null hypothesis (Methods). Genes with rejected null hypotheses will be selected for downstream analysis. For example, the CD34+ cell population within Zheng2017 dataset [3] has 2.7% of the genes

Table 1 Zero inflation test statistics of PPBP gene in CD34+ cells in Zheng data before and after clustering into subtypes

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

with significant zero inflation at 5% type I error level after Bonferroni correction. But after clustering into subtypes, each subtype had 1.3%, 0.3%, and 1.2% of genes with zero inflation respectively. We demonstrate the intuition of this test procedure in Table 1 using gene PPBP as an example. PPBP was identified with a high zero proportion of 26% and an average mean UMI count of 25.89 within CD34+ cells, indicating very high zero inflation with z-score greater than 10^6 when the proposed test is applied. After we separate CD34+ cells into three subtypes, the test within each subtype is no longer statistically significant. We observe PPBP is highly expressed in subtypes 2 and 3 and is almost unexpressed in subtype 1. This shows how cellular heterogeneity can drive excessive zeros and how zero proportions can be used to discern cell types.

The proposed framework significantly differs from existing ones in several ways. First of all, only the proportion of zeros (p_g), but not that of other non-zero count values, is used in the test. We empirically show that this statistic is sufficient for cellular heterogeneity analysis in many data sets with low UMI counts. This allows us not to search for a particular parametric distribution to fit all non-zero values, which can be computationally more burdensome. In terms of clustering, analysis shows that deviance and zero-inflation both lead to feature selection with similar performance (Additional file 1: Figure S11, S12); our software can use deviance test for feature selection when a data set has high UMI counts and zeros alone do not hold enough information. Secondly, this framework allows each gene to have different grouping structure across cells. Most existing methods select one set of genes to cluster all the cells [10, 18, 31, 32], which implicitly assume cell types can be well-defined biologically by a common set of genes. This is not realistic given the fact that each gene's heterogeneity level varies with its function. For example, housekeeping genes are expected to behave similarly in all cells but immune-related genes, known to have more diverse genetic profiles with highly polymorphic nucleotides [33, 34], might be more finely differentiated among the sampled cells. Our approach acknowledges this type of variability. Finally, our approach provides a much more optimistic view of the UMI data analysis. No complicated modeling is needed for resolving the cellular heterogeneity.

We observe that the droplet-based data-generating process in the 10X protocols affect UMI counts in different cell types and even across datasets in a similar fashion (Additional file 1: Figure S1, S2). Regardless of samples or cell types, all cells show the same distribution of zero proportions with respect to mean UMI count. This means the technical noise affects each and every cell fairly, and hence, biological heterogeneity alone can largely explain the zero inflation phenomenon. Once the heterogeneity is accounted for, without any other pre-processing steps, zero proportions of UMI data closely follow the expected curve under a Poisson distribution. These observations urge us to re-evaluate some widely used pre-processing methods under this scope.

Inappropriate pre-processing introduces unwanted noise in the downstream analysis

One of the most popular method for normalization is to divide UMI counts by a cell-specific scaling factor so that total UMI counts are equal across cells [11]. This strategy implicitly assumes sequencing depth effects are purely technical. Total UMI count needs to be carefully corrected in case of data integration. Data sets collected from different protocols and batches have different distribution of UMI counts. However, when there is no batch effect, the total UMI count has valuable biological information. Here, we show sequencing depths are confounded with cell types and size factor-based adjustment can obscure biological information. For a given gene, dividing UMI counts by cell-specific factors does not change its zero proportion across cells but changes its mean. As a consequence, zero proportions across genes no longer follow the expected curve under a Poisson distribution (Fig 2a), and the two curves from each cell type are separated from one another. For example, in 6 PBMC data sets (Azizi and Zheng) [3, 27], monocytes have lower UMI counts than B cells. The median UMI counts for monocytes and B cells, respectively, are 787 and 1180 for Zheng data for 68,000 PBMC cells, 4831 and 5575 for Azizi 2018 data for breast cancer tumor patient 9 (replication 1), 4891 and 5372 for patient 10, and 5093 and 5722 for patient 11 (replication 1). When they are forced to match the median UMI count of all cells, the counts for the monocytes are inflated while those for the B cells deflated. In addition, cell types are stratified on the zero proportion plot after adjustment, indicating that total UMI counts of each cell contain valuable information

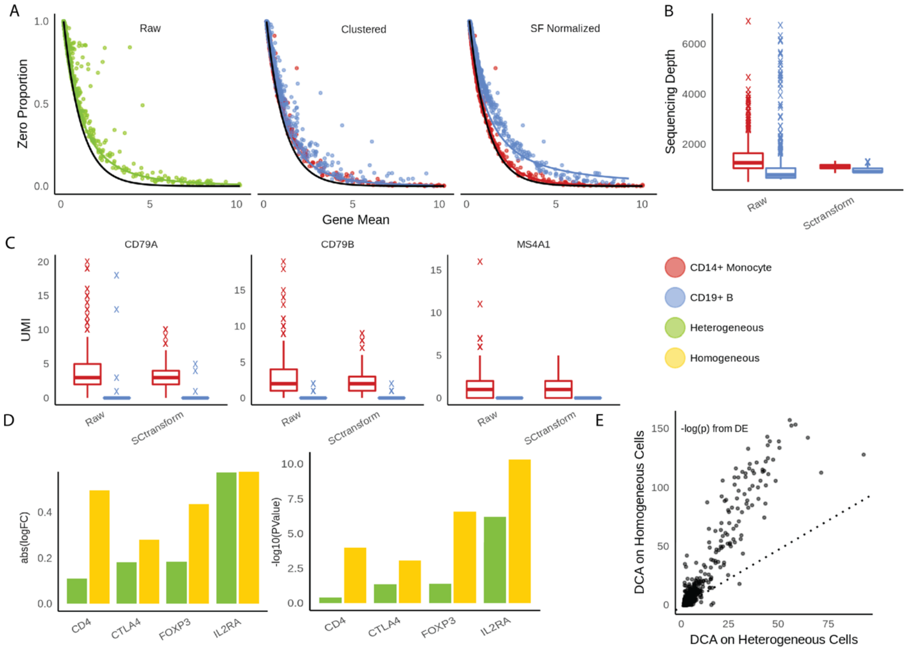

Fig. 2 **a** Scatterplots between gene means and zero proportions across genes calculated from raw UMI data, clustered data, and data after sequencing-depth normalization, respectively. Fitted line is negative binomial curve. **b**, **c** Evaluations of pre-processing in Sctransform. **b** Distributions of sequencing depths across cells in raw UMI data vs. data cleaned by SCtransform. **c** Comparisons of three monocyte markers in raw UMI data vs. data cleaned by SCtransform. **d**, **e** Evaluations of pre-processing in DCA. **d** Log fold changes and log *p* values from differential expression analysis using the same data set but imputed by DCA as heterogeneous and homogeneous cell populations, respectively. **e** The *p* values comparisons between two different imputation strategies show the general deflation of biological signals of DCA when applied to heterogeneous cells

about its cell type (Fig. 2a, Additional file 1: Figure S14). The UMI counts for different cell types do not need to be consistent across different tissues or organisms. For example, in Zhang2019 data, B cells and monocytes have similar total UMI counts (Additional file 1: Figure S14), but fibroblast has more UMI counts than all other types. Forcing fibroblasts to have the same UMI counts as other types would reduce the signal strengths of the markers for fibroblasts.

Sctransform is one recent influential UMI analysis method [13]. The key idea of sctransform pre-processing is to remove sequencing depth effects by introducing log-scale sequencing depth as a covariate and regressing it out from each cell under a negative binomial model. Similarly, this approach destroys the natural Poisson structure for zero proportions. We show in Fig. 2b, c an example of how normalization can further interfere with detection of biological signals. Across all 6 PBMC datasets mentioned above, we observe B cells always have more UMI counts than monocytes before pre-processing. Applying sctransform barely modifies the sequencing depths of monocytes but shrinks the UMI counts of B cells to match those of monocytes. Due to the artificial shrinkage, biological markers for B cells, such as MS4A1, CD79A, and CD79B, lose their power to discern B cells and monocytes [35]. This suggests cell type differences could be potentially compromised due to excessive cleaning from sctransform.

Another popular pre-processing step is to apply deep learning based de-noising tools such as Deep Count Autoencoder (DCA) and SAVER, which de-convolute the technical effects from biological effects and impute zero accounts due to dropouts at the same time. DCA implements deep neural network with flexible parametric options for noise distributions. Similarly, we observe DCA blurs the distinction among cell types, because denoising methods essentially regularize each cell to resemble one another. We illustrate its negative impacts on downstream analysis by comparing differential expression analysis results using two imputation strategies. We selected naive T cells and regulatory T cells from Zhengmix8eq for the experiments, which clustering algorithms often struggle to differentiate because of their similarity. In the first experiment, we imputed naive T cells and regulatory T cells together. In the second, we imputed naive T cells and regulatory T cells separately. Then, we performed DE analysis on imputed data sets using edgeR's likelihood ratio test [22]. We observe much greater log fold change values between naive T and regulatory T cells from data imputed separately than data imputed together. Overall, the signal strength of DE analysis is greatly compromised across all the genes if imputing two cell types together (Fig. 2d–e). Using type I error level of 0.05, 320 genes pass the Bonferroni criterion if clustering is performed first, while only 156 does if imputation is performed first. Known markers including CD4, CTLA4, FOXP3, and IL2RA [35] lost significant amount of biological signals by showing weaker log-fold change (Fig. 2d, Additional file 1: Figure S15). When the cells were first clustered and then imputed, the p values were 1e−04, 8e−04, 2e−07, and 4e−11 respectively for those genes. When the cells were imputed first through DCA, the p values were 3e−01, 4e−02, 4e−02, and 6e−07. Hence, three of the 4 genes lost statistical significance at a very liberal p value threshold of 0.05. This analysis suggests imputing the UMI data without resolving cell heterogeneity can lead to loss of important biological information.

HIPPO: Heterogeneity-Inspired Pre-Processing tOol

The above analyses suggest the first and foremost step in pre-processing is to account for the cellular heterogeneity. Imputation or normalization before resolving the cellular heterogeneity may lead to inevitable loss of biological signals. We implement this new perspective into a computational tool called HIPPO, where we integrate the proposed zero inflation test into a hierarchical clustering framework. Specifically, we first selected genes with strong indication for cellular heterogeneity. We use a cutoff of 2 on z score for selection of genes. The selected features were then used to cluster the cells into 2 groups using PCA + K-means. Then, each cluster was evaluated with their intra-variability using the mean Euclidean distance from the centers of K-mean algorithm. The group with the highest intra-variability was selected and 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 number of clusters K, or the number of zero-inflated genes is less than a certain percentage of the genes. The former one can be difficult to set in real practice without any prior knowledge and the later one offers a more natural stopping criterion. HIPPO is computationally cheap because fewer and fewer features will be left for the next round of clustering, and the Poisson-based test statistic has closed-form expression (Fig. 3a, b). In Fig. 3c, we show the results from each iteration of HIPPO on Zhengmix8eq data. HIPPO successfully identifies monocytes, natural killer cells, B cells, and T cells in the respective order. Then, it further separates naive cytotoxic cells, memory T cells, and naive T cells from a group of regulatory T cells and helper T cells. However, when forced to separate into one more group, instead of clustering the remaining T cells, it created another subgroup of natural killer cells. Meanwhile, Seurat and Sctransform fails to separate the memory T cells, regulatory T cells, and helper T cells, grouping them as one cluster. (Fig. 3d). The adjusted rand index for the three methods show that HIPPO performs the best throughout the different K specification (Fig. 3e). When the selected features' characteristics were studied through CV, gene variance, and zero proportion, Seurat and Sctransform selected more features (2000 and 3000 respectively while HIPPO selected 950), but they are highly concentrated where gene means are near 0. This is because their feature selection focuses on coefficient of variation which becomes numerically unstable as gene mean becomes near zero. HIPPO selects fewer but more relevant genes by using the zero proportion as the selection metric (Fig. 3f). This result is repeated in a different data set from muscular heart tissue in Fig. 4a. Genes selected by both methods are those with non-zero mean UMI counts, but Seurat selects extra number of genes that have mean count very close to 0. These genes are likely to add noise instead of contributing to real biological signals detection.

HIPPO's iterative procedures naturally offer strong interpretability through sequential visualization of the analysis at each round of clustering. We use HIPPO results on an unlabeled 10X UMI data set of 10K E18 mouse heart cells for illustration. Sequential feature selection can be monitored through the visualization of the changing relationships between zero proportions and gene means. As cells are clustered into finer distinct groups, or as more cellular heterogeneity is resolved, regression lines between zero proportions and gene means get more closely aligned with the expected Poisson curve (Fig. 4b). Simultaneously, we can use a heatmap to visualize top features that contribute most at each round of clustering (Fig. 4c). In addition to biomarkers identified based on zero inflation, HIPPO also implements a differential expression test based

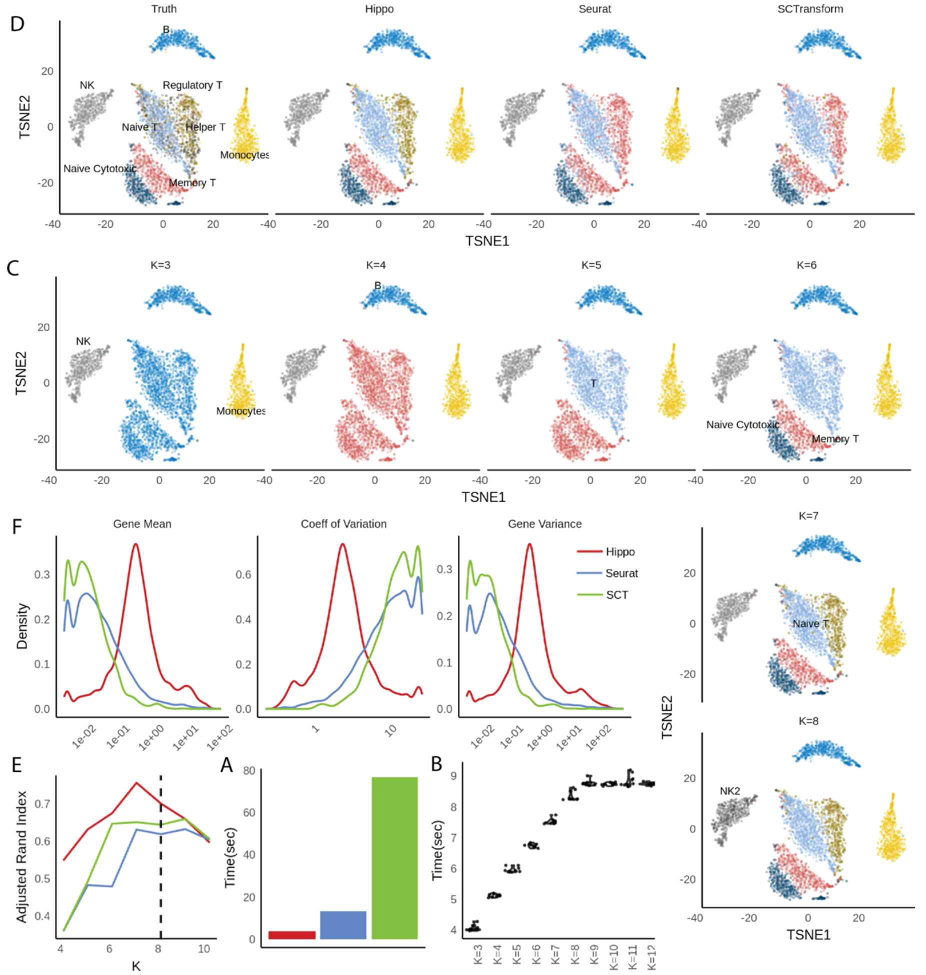

Fig. 3 HIPPO framework applied to Zhengmix8eq data. **a** Computing time for each method using LAMBDA QUAD workstation with Intel Xeon W-2175 processor sequentially (non-parallel). **b** Computing time for HIPPO using different *k*. **c** HIPPO's sequential clustering results for $K = 3, \cdots, 8$. **d** t-SNE plots for clustering results from three methods: HIPPO, Seurat, and SCTransform, compared to true labels. Seurat and SCTransform cannot differentiate helper T/regulatory T and memory T cells. **e** Clustering results comparisons using Adjusted Rand Index. **f** Comparisons of features selected by different methods for their gene mean, CV, and variance. Seurat and SCTransform use CV as the selection criteria, and hence, their features weigh heavily on genes with small mean expression and variance

on all count values to extract more features (Methods, Fig. 4d). The differential analysis can be viewed together with a t-SNE plot constructed with the same color code (Fig. 4e).

Discussion

We have provided a new perspective on the analysis of single-cell UMI data sets of multiple tissues and protocols (Additional file 1: Figures S1, S2, S3, S4, S7). Extensive analyses confirm the claims of recent literature [6] that 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 the level of technical noise is much lower. The results also show that cell-type heterogeneity must be tackled as the first step of analysis for more reliable downstream

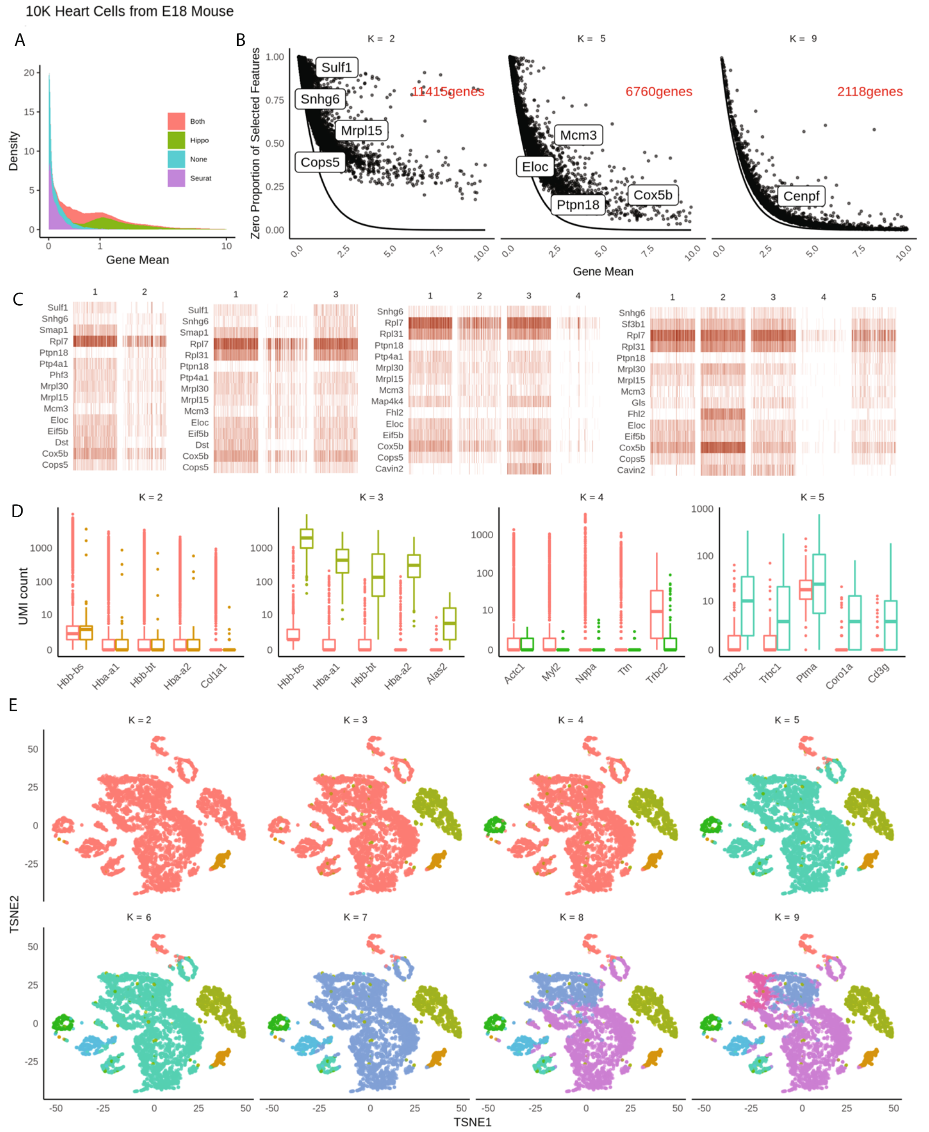

Fig. 4 HIPPO framework applied to 10K E18 mouse heart cells. **a** Distributions of means across gene features selected by Seurat, Hippo, both, or none. **b** Sequential feature selection visualizes how genes gradually align closer to the expected Poisson line as more heterogeneity accounted for. **c** Heatmaps of top features selected at the first 5 rounds of clustering. **d** Top differential expression genes obtained at each round of clustering. **e** Visualization of sequential clustering using t-SNE plots

analyses. Moreover, through a streamlined feature selection method that reflects the dynamic nature of cellular process, the proposed method provides a computationally and mathematically simple analysis tool with great interpretability.

There are remaining challenges that are important in the future development of single cell UMI data analysis. First, lack of labeled data restricts the analyses in certain protocols such as Drop-seq. There is strong evidence for our method in 10X data sets. Supplementary Figures also show that the claims hold in all 10X data, Tung2018 data that uses Hi-Seq 2500 [25] and Baron2016 data that uses in-Drop [36]. In Drop-seq, the noise level was too high to assume the zero proportions follow the exponential curve relative to the gene mean (Additional file 1: Figure S7). It is either that Drop-seq data sets have different noise structure from the 10X data sets, or in particular Macosko data [2] of muscular retina cells have excessively high cellular heterogeneity [1]. Future new Drop-seq data could help resolve the discrepancy between 10X and Drop-seq. Second, although HIPPO is computationally simple compared to existing tools, the current computational bottleneck is the principal component analysis, which could be slow for large cell numbers. In that case, advanced computing techniques such as sub-sampling or more rigorous filtering should be applied. In addition, HIPPO is implemented as evolving modular software. In the current release, we include two different feature selection methods, deviance test and zero-inflation test, and two differential expression detection methods. Alternative dimension reduction or clustering methods can be easily incorporated into the framework.

We focus on the pre-processing with resolving cellular heterogeneity in our analysis tool, but this novel perspective on the noise structure of UMI data can be extended to other steps of analysis pipeline. Batch correction, lineage analysis, or trajectory inference can all benefit from the simpler noise structure not only computationally but also by avoiding unnecessary normalizing steps that can introduce unwanted bias and noise.

Methods and materials

Datasets

Throughout the analysis, we used publicly available single cell UMI sequencing data from various protocols. Most analysis in the main text is focused on SRP073767 which is also available in 10x Genomics, and it sequences 68,000 PBMC cells using Cell Ranger 1.1.0 [3]. We use different subsets of this data sets, namely Zhengmix4eq, Zhengmix4uneq, and Zhengmix8eq as defined in Duo (2018)[31]. Other data sets used in the main text are GSE111108 [37] and GSE115189 [17], and GSE114724 [27]. Supplementary data includes more data sets from 10X including 5k Cells from a combined cortex, hippocampus and subventricular zone of an E18 mouse (v3 chemistry), 1k Brain Cells from an E18 Mouse (v2 chemistry), and 10k Heart Cells from an E18 mouse (v3 chemistry). We also use Tabula Muris data from various mouse tissues [38]. We also use GSE84133 [36] as an example of in-Drop, GSE63473 [2] as an example of Drop-seq, SDY998 [39] as an example of CEL-seq2, and GSE77288 [25] as an example of Hi-Seq. All the data sets were analyzed after their own filtering process (Table 2).

Benchmarked methods

In Fig. 3, we benchmark Seurat 3.0.0 [40] and SCTransform version 0.2.0 that is integrated with Seurat platform. Seurat was implemented following its guided tutorial https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html, and SCTransform through a vignette https://rawgit.com/ChristophH/sctransform/master/inst/doc/seurat.html. All parameters were selected through software's default except resolution parameter for clustering to generate results for various number of clusters. Seurat used in Fig. 4 A was also the same version with the default parameters for feature selection. In Fig. 3a, the t-SNE plots were created using the features selected by the first round of HIPPO because they reflected the division of true cell labels the most accurately.

DCA was installed through Conda and imputation was performed following the tutorial on https://github.com/theislab/dca. In one experiment, we first divide the data set into correct labels, and then impute them separately using DCA (imputing homogeneous

Table 2 List of data sets used in the main text and supplementary materials

ID	Data set	Species	Protocol
10x	5KNeuron	Mouse neuron	10x (v3.1) CR* 3.0.2
10x	10KHeart	Mouse heart	10x (v3) CR 3.0.0
GSE111108 [37]	Tian2018	Human cell lines	10x Chromium
GSE115189[17]	Freytag2018	Human PBMC	10x (v2)
10x	1KNeuron	Mouse neuron	10x (v2) CR 2.1.0
SRP073767[3]	Zhengmix4eq	Human PBMC	10x (v1) CR 1.1.0
SRP073767	Zhengmix4uneq	Human PBMC	10x (v1) CR 1.1.0
SRP073767	Zhengmix8eq	Human PBMC	10x (v1) CR 1.1.0
SRP073767	PBMC3k	Human PBMC	10x (v1) GemCode
SRP073767	PBMC4k	Human PBMC	10x (v1) Chromium
SRP073767	PBMC68k	Human PBMC	10x (v1) CR 1.1.0
GSE84133[36]	Baron2016	Human pancreas	inDrop
GSE114724[27]	AziziPatient09Rep1	Human breast tumor	10x CR 2.1.1
GSE114724	AziziPatient09Rep2	Human breast tumor	10x CR 2.1.1
GSE114724	AziziPatient10Rep1	Human breast tumor	10x CR 2.1.1
GSE114724	AziziPatient11Rep1	Human breast tumor	10x CR 2.1.1
GSE114724	AziziPatient11Rep2	Human breast tumor	10x CR 2.1.1
SDY998[39]	Zhang2019	Human joint synovial	CEL-seq2
GSE63473[2]	Macosko2015	Mouse	Drop-seq
GSE77288 [25]	Tung2017	Human iPSC	HiSeq 2500
Tabula Muris [38]	Tabula Muris	Mouse	10x (v2)

CR cell ranger

cell population). In the other experiment, we impute both cell types together (imputing heterogeneous cell population). One property of DCA is that it automatically removes genes that are 0 in all the cells. Naturally, there are more such genes in homogeneous cell populations. Especially, some of the biomarkers are not expressed at all when cell population is divided into subtype. In that case, we imputed zero to those genes, assuming DCA did not perform any imputation (Fig. 2d, e). SAVER was downloaded from CRAN with version 1.1.1. In Additional file 1: Figure S16 transcriptome-level statistics were compared only using genes that had at least one positive count in each cell type. In both DCA and SAVER, all the parameters the default values as suggested by the software.

Likelihood ratio test in Figs. 1b and 2f was conducted by fitting the distributions using the fitdistr function from *MASS* package [41], and zero-inflated negative binomial distribution was fitted using *pscl* package [42].

Poisson mixture model

Consider a gene by cell matrix if UMI counts X for gene $g = 1, \cdots, G$ and cell $c = 1, \cdots, C$. 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

$$\hat{p}_g = \sum_{c=1}^{C} \frac{\mathbb{1}_{X_{gc}=0}}{C} \tag{3}$$

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 $p = \sum_{k=1}^{K_g} \pi_k e^{-\lambda_{kg}}$. In other words, zero inflation indicates 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) and Sarkar (2020) demonstrates that most genes in UMI data follow Poisson distribution [6, 7] 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 more directly addresses 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; mixture parameters π_k are not estimated. Alternatively, for example, one might 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 3 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 (Additional file 1: Figure S13) (Table 4).

Table 3 The alternative hypothesis $H_A : p_g > e^{-\lambda}$ is robust to different model hypotheses

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	$e^{-\lambda}$	$\left(\frac{r}{r+\lambda}\right)^r + \pi_0$

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)

Table 4 High gene variance is not a good indicator of cell type heterogeneity under the alternative hypothesis of zero-inflated negative binomial, because the variance can be lower under the alternative hypothesis

Alternative hypothesis	Variance under H_0	Variance under H_A
Mixture of Poisson	$\lambda = \sum_k \pi_k \lambda_k$	$\sum_k \pi_k(\lambda_k + \lambda_k^2) - (\sum_k \pi_k \lambda_k)^2$
Negative Binomial	λ	$\frac{\lambda^2}{r} + \lambda$
Zero-inflated negative binomial	λ	$(1-\pi_0)^2 \left(\frac{\lambda^2}{r} + \lambda\right)$

Feature selection and Inference

We provide two ways to select features: zero-inflation test and deviance test. The clustering performance is similar using both methods.

For zero-inflation test, HIPPO defines the observed zero proportion $\hat{p}_g$ and expected zero proportion $e^{-\bar{X}}$. For gene g with count data X_{gc} for cells $c = 1, \cdots, C$, consider an estimate for the proportion of zeros $\hat{p}_g$ as

$$\hat{p}_g = \frac{\sum_{c=1}^{C} \mathbb{1}_{X_{gc}=0}}{C}$$

The gene mean is estimated as the average UMI counts $\bar{X}_g = \frac{1}{C}\sum_{c=1}^{C} X_{gc}$ and is treated as a fixed number. Then,

$$\hat{p}_g = \mathcal{N}\left(e^{-\bar{X}}, \frac{\hat{p}_g(1-\hat{p}_g)}{C}\right)$$

The test statistic z-score for gene g is as below.

$$z_g = \frac{\hat{p}_g - e^{-\bar{X}}}{\frac{\hat{p}_g(1-\hat{p}_g)}{C}}$$

To note, the gene mean $e^{-\bar{X}}$ is also a random variable that follows a log-normal distribution, whose inference is not trivial (further discussion can be found in Supplementary Text 1). However, this feature selection method works well in practice and intuitively interpretable.

Users can also use deviance measure to select the top features [8].

$$d_g = 2 \cdot \sum_{c=1}^{C} \left(X_{cg} log \left(\frac{X_{cg}}{\bar{X}_g} \right) - \left(X_{cg} - \bar{X}_{cg} \right) \right)$$

The deviance threshold is required from the users to select cutoffs to select the features.

Hierarchical clustering

Algorithm 1 outlines the iterative procedure of HIPPO's hierarchical clustering. Several stopping criteria can be determined by the user: the maximum number of clusters K, the feature selection statistic threshold z, and outlier gene proportion o. The algorithm first computes the number of outlier genes to allow, $G \times o$. For example, if there are 30,000 genes in total and o is specified as $1\% = 0.01$, then the algorithm allows 300 features to have zero inflation. During the clustering procedure, HIPPO terminates in either scenarios: there are K identified clusters or if there are less than $G \times o$ genes that exceed the specified z value threshold.

HIPPO takes all the cells and select the features whose zero inflation statistic z exceeds the threshold. Only zero-inflated features are then log-transformed (log(X)+1) and sent

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 Less than the designated number of genes exceed $z$ threshold then
        stopping criterion; terminate algorithm
    else
        update the matrix by selecting new features
        log transformation + centered/scaled PCA + Kmeans
        divide cells into two groups, one with label $\ell$ and another with label $k$
        log transformation + un-centered/un-scaled PCA
        update intra-cluster variation by taking sample variance of un-scaled PCs
        update $\ell$ = cluster with the highest intra-cluster distance
    end if
end for
return cluster labels for each $k$
```

into principal component decomposition with scaling and centering. HIPPO then clusters the cells into two groups using the dimension-reduced cell embeddings through K-means, which is performed multiple times (user-defined, default 10) for stability.

Meanwhile, during the hierarchical clustering, HIPPO keeps track of intra-cluster variation by performing unscaled, uncentered PCA. HIPPO computes the first 10 dimensions (user-defined) of cell embeddings and sum up the sample variance of each component. The PCA for recording the intra-cluster variability is not scaled to avoid potential bias due to the cell population size. Since a subset of cells are considered for clustering at each round, fewer and fewer cells are used for the dimension reduction. If scaled, their dimension-reduced cell embeddings would be artificially more far apart compared to when more cells were considered for clustering. The intra-cluster variance is the criterion for selecting the cell group to be further clustered in the next round.

Differential expression testing

HIPPO provides two methods to identify differentially expressed genes, one by performing a t-test on means, and the other by performing a Poisson likelihood test.

For both methods, the cells are separated into two groups as follows.

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

$$H_0 : \lambda_1 = \lambda_2, \qquad H_A : \lambda_1 \neq \lambda_2$$

The first method conducts the 2-sample t-test by measuring the significance in the mean difference of two groups.

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

which asymptotically follows standard normal distribution.

The second method conducts the likelihood ratio test by measuring the deviance in null and alternative models. The MLEs under both null model and alternative model are estimated as below.

$$\hat{\lambda} = \frac{1}{|C_1| + |C_2|} \sum_{c \in C_1 \cup C_2} X_{cg}, \qquad \hat{\lambda}_1 = \frac{1}{|C_1|} \sum_{c \in C_1} X_{cg}, \qquad \hat{\lambda}_2 = \frac{1}{|C_2|} \sum_{c \in C_2} X_{cg}$$

Then, the deviance for testing these nested models is, ℓ denoting Poisson log-likelihood,

$$2 \cdot \left(\sum_{c \in C_1} \ell(X_{cg}; \hat{\lambda}_1) + \sum_{c \in C_2} \ell(X_{cg}; \hat{\lambda}_2) - \sum_{c \in C_1 \cup C_2} \ell(X_{cg}; \hat{\lambda}) \right)$$

that asymptotically follows χ^2_1 distribution.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02096-y.

Additional file 1: Supplementary figures, tables, and text to support our claims

Additional file 2: Review history.

Peer review information

Barbara Cheifet was the handling editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Acknowledgements

The authors thank the reviewers and the handling editor for their constructive comments.

Review history

The review history is available as Additional file 2.

Authors' contributions

M.C. conceived and led this work. T.K. and M.C. developed the methods and performed the analyses. T.K. implemented the HIPPO software. X.Z. participated in critically revising the draft. T.K. and M.C. wrote the paper with feedback from X.Z. All authors read and approved the final manuscript.

Funding

M.C. was supported by the National Institutes of Health (NIH) Grants R01 GM126553, a Sloan Foundation Research Fellowship and a Human Cell Atlas Seed Network grant from Chan Zuckerberg Initiative. X.Z. was supported by the NIH Grants R01HG009124 and R01GM126553 and the National Science Foundation (NSF) Grant DMS1712933.

Availability of data and materials

10X 68K (SRP073767) [3] data is available in 10X website https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/fresh_68k_pbmc_donor_a. Zhengmix, Freytag2018 (GSE115189) [17], and Tian2018 (GSE111108) [37] data can be downloaded in DuoClustering2018 package available in Bioconductor [43]. Additional analyses also include GSE114724 [27], SDY998 [39], GSE63473 [2], GSE77288 [25], and Tabula Muris data [38]. Details about the data sets are described in Methods (Table 2).

The software HIPPO is implemented in R package and is freely available under GNU General Public License v2.0 or later (GPL $\geq$2) at https://github.com/tk382/HIPPO [44].

The source code for the software and the analyses presented in the paper can be found at Zenodo repository: https://doi.org/10.5281/zenodo.3926915 [45].

Ethics approval and consent to participate

Ethics approval is not applicable to this study.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Statistics, University of Chicago, Chicago, USA. [2]Department of Biostatistics, University of Michigan, Ann Arbor, USA. [3]Department of Human Genetics and Department of Medicine, University of Chicago, Chicago, USA.

Received: 21 April 2020 Accepted: 8 July 2020
Published online: 06 August 2020

References

1. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201.
2. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5): 1202–14.
3. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
4. Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, Mazutis L. Single-cell barcoding and sequencing using droplet microfluidics. Nat Protocol. 2017;12(1):44.
5. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11(2):163.
6. Chen W, Li Y, Easton J, Finkelstein D, Wu G, Chen X. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. Genome Biol. 2018;19(1):70.
7. Sarkar AK, Stephens M. Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis. BioRxiv. 2020. https://doi.org/10.1101/2020.04.07.030007.
8. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. Genome Biol. 2019;20(1):1–16.
9. Germain P-L, Sonrel A, Robinson MD. pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single-cell RNA-seq preprocessing tools. BioRxiv. 2020. https://doi.org/10.1101/2020.02.02.930578.
10. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411.
11. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 2017;14(6):565.
12. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. Drimpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinforma. 2018;19(1):220.
13. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):1–15.
14. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JI, Raj A, Li M, Zhang NR. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018;15(7):539.
15. Andrews T. S., Hemberg M. M3drop: dropout-based feature selection for scRNASeq. Bioinformatics. 2019;35(16): 2865–7. https://academic.oup.com/bioinformatics/article/35/16/2865/5258099.
16. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1):390.
17. Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M. Comparison of clustering tools in R for medium-sized 10x genomics single-cell RNA-sequencing data. F1000Research. 2018;7:1297. https://doi.org/10.12688/f1000research.15809.2.
18. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods. 2017;14(4):414.
19. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37(5):547.
20. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381.
21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
22. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
23. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, Zhang NR. Data denoising with transfer learning in single-cell transcriptomics. Nature Methods. 2019;16(9):875–8.
24. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell rna-seq data. Nature Commun. 2018;9(1):1–17.
25. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, Gilad Y. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7:39921.
26. Choi K, Chen Y, Skelly DA, Churchill GA. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. bioRxiv. 2020. https://doi.org/10.1101/2020.03.03.974808.
27. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, Nainys J, Wu K, Kiseliovas V, Setty M, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell. 2018;174(5):1293–308.
28. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. Gencode: the reference human genome annotation for the encode project. Genome Res. 2012;22(9):1760–74.
29. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc R Soc B Biol Sci. 2010;277(1684):979–88.
30. Clivio O, Lopez R, Regier J, Gayoso A, Jordan MI, Yosef N. Detecting zero-inflated genes in single-cell transcriptomics data. BioRxiv. 2019;794875. https://www.biorxiv.org/content/10.1101/794875v2.abstract.
31. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research. 2018;7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6134335/.
32. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483.

33. Hughes AL. Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. Mol Biol Evol. 1997;14(1):1–5.
34. Hurst LD, Smith NG. Do essential genes evolve slowly?. Curr Biol. 1999;9(14):747–50.
35. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, Schoeberl B, Raue A. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun. 2017;8(1):2032.
36. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell Syst. 2016;3(4):346–60.
37. Tian L, Dong X, Freytag S, Le Cao K-A, Su S, Amann-Zalcenstein D, Weber TS, Seidi A, Naik S, Ritchie ME. scRNA-seq mixology: towards better benchmarking of single cell RNA-seq protocols and analysis methods. BioRxiv. 2018;433102. https://doi.org/10.1101/433102.
38. Consortium TM, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562(7727): 367.
39. Zhang F, Wei K, Slowikowski K, Fonseka CY, Rao DA, Kelly S, Goodman SM, Tabechian D, Hughes LB, Salomon-Escoto K, et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nat Immunol. 2019;20(7):928–42.
40. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck III WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.
41. Venables WN, Ripley BD. Modern applied statistics with S-PLUS: Springer Science & Business Media; 2013. https://cran.r-project.org/web/packages/MASS/citation.html.
42. Jackman S. pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory. Sydney: United States Studies Centre, University of Sydney; 2020. R package version 1.5.5. https://github.com/atahk/pscl/.
43. Duò A, Soneson C. DuoClustering2018: Data, Clustering Results and Visualization Functions From Duò et al 2018. 2020. R package version 1.6.0. https://bioconductor.org/packages/release/data/experiment/html/DuoClustering2018.html.
44. Kim T, Zhou X, Chen M. HIPPO (Heterogeneity Inspired Pre-Processing tOol). Zenodo. 2020. https://doi.org/10.5281/zenodo.3926915.
45. Kim T, Zhou X, Chen M. Demystifying "drop-outs" in single-cell UMI data. Zenodo. 2020. https://doi.org/10.5281/zenodo.3926915.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.