Gene expression

Statistical significance of cluster membership for unsupervised evaluation of cell identities

Neo Christopher Chung 1,2,3

1Institute of Informatics, Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw 02-097, Poland, 2NHLBI Integrated Cardiovascular Data Science Training Program, University of California, Los Angeles, CA 90095, USA and 3Departments of Physiology and Medicine (Cardiology), UCLA School of Medicine, Los Angeles, CA 90095, USA

Abstract

Motivation: Single-cell RNA-sequencing (scRNA-seq) allows us to dissect transcriptional heterogeneity arising from cellular types, spatio-temporal contexts and environmental stimuli. Transcriptional heterogeneity may reflect phenotypes and molecular signatures that are often unmeasured or unknown a priori. Cell identities of samples derived from heterogeneous subpopulations are then determined by clustering of scRNA-seq data. These cell identities are used in downstream analyses. How can we examine if cell identities are accurately inferred? Unlike external measurements or labels for single cells, using clustering-based cell identities result in spurious signals and false discoveries.

Results: We introduce non-parametric methods to evaluate cell identities by testing cluster memberships in an unsupervised manner. Diverse simulation studies demonstrate accuracy of the jackstraw test for cluster membership. We propose a posterior probability that a cell should be included in that clustering-based subpopulation. Posterior inclusion probabilities (PIPs) for cluster memberships can be used to select and visualize samples relevant to subpopulations. The proposed methods are applied on three scRNA-seq datasets. First, a mixture of Jurkat and 293T cell lines provides two distinct cellular populations. Second, Cell Hashing yields cell identities corresponding to eight donors which are independently analyzed by the jackstraw. Third, peripheral blood mononuclear cells are used to explore heterogeneous immune populations. The proposed P-values and PIPs lead to probabilistic feature selection of single cells that can be visualized using principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and others. By learning uncertainty in clustering high-dimensional data, the proposed methods enable unsupervised evaluation of cluster membership.

Availability and implementation: https://cran.r-project.org/package=jackstraw.

Contact: nchchung@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Single-cell RNA-seq (scRNA-seq) has enabled large-scale gene expression studies that help elucidate transcriptional heterogeneity related to cellular types, spatio-temporal contexts and environmental stimuli (Jaitin et al., 2014; Macosko et al., 2015; Patel et al., 2014). Transcriptional heterogeneity is manifested on systematic variation across gene expression, which is characterized by unsupervised clustering. Clustering-based cell identities are used in downstream feature selection, differential expression analysis and visualization (Butler et al., 2018; Guo et al., 2015; Qiu et al., 2017; Satija et al., 2015). Given that cell identities are determined in an unsupervised manner, it is critical to evaluate if they are correctly assigned. We have developed novel methods to estimate statistical significances and posterior inclusion probabilities (PIPs) of assigning cell identities to estimated subpopulations. By learning uncertainty in applying clustering to scRNA-seq data, the proposed methods enable unsupervised evaluation of cluster memberships, such as cell identities.

Clustering has been one of the most popular analysis methods for high-dimensional genomic data. Gene expression studies have long used clustering to identify co-regulated subsets of genes (Eisen et al., 1998; Gasch et al., 2000; Spellman et al., 1998) and subpopulations among samples (Alon et al., 1999; Golub et al., 1999; Sorlie et al., 2001). Recently, there have been several scRNA-seq studies where gene expression from hundreds and thousands of single cells are measured in masse (Jaitin et al., 2014; Macosko et al., 2015; Zheng et al., 2017). Identities of single cells are typically unknown a priori and characterized by unsupervised clustering. Clustering m cells to K subpopulations provides computationally defined m cell
identities. These clustering-based cell identities are of great interests, as complex phenotypes and diseases may exhibit molecular signature as yet unknown.

Single-cell analysis tools implement various clustering algorithms, including, but not limited to, K nearest neighbors in Seurat (Butler et al., 2018; Satija et al., 2015), hierarchical clustering in SINCERA (Guo et al., 2015) and density peak clustering in Monocle (Qu et al., 2017). Furthermore, a number of clustering algorithms specifically tailored to scRNA-seq data have been developed to identify subtypes of single cells (Buettner et al., 2015; Wang et al., 2017; Xu and Su, 2015; Zeisel et al., 2015). To increase computational efficiency, a number of scRNA-seq studies combine principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE; van der Maaten and Hinton, 2008) with unsupervised clustering (Guo et al., 2015; Yang et al., 2018). What is overlooked in these recent developments is how to evaluate single cells, when their cell identities are determined by clustering. To the best of our knowledge, this represents the first study on estimating statistical significance of cluster membership at a single-cell level.

The proposed non-parametric methods leverage the assumption regarding cluster structure across single cells. Clustering algorithms estimate systematic variation and identify subsets of cells that contribute to distinct patterns. Due to high dimensionality of scRNAseq, cells with ambiguous identities are artificially assigned to subpopulations, leading to weakened signals and false classifications. Our framework models and tests expression levels of cells with respect to their estimated subpopulations (Fig. 1). The jackstraw strategy accounts for overfitting characteristics of unsupervised clustering. Beyond P-values, an empirical Bayes approach is used to derive a probability that a cell truly belongs to an estimated subpopulation, which we call a PIP. This connects an unsupervised classification of high-dimensional data and a fundamental hypothesis framework in a statistically rigorous manner.

Operating characteristics of the proposed methods are demonstrated through comprehensive simulation studies. Three scRNA-seq data analyses are presented using (i) a mixture of Jurkat and 293T cell lines (Zheng et al., 2017), (ii) a Cell Hashing data of peripheral blood mononuclear cells (PBMCs) from eight independent donors (Stoeckius et al., 2018) and (iii) immune populations in 68 579 PBMCs from a single donor (Zheng et al., 2017). The reference implementation (https://CRAN.R-project.org/package=jackstraw) includes K-means clustering, partitioning around medoids (PAM; Kaufman and Rousseeuw, 1987) and mini batch K-means (MBKM) (Sculley, 2010), which are fast, robust and scalable to millions of single cells.

2 Statistical models and methods

Unsupervised clustering of m single cells into K subpopulations provides m cell identities. By modeling m cells with respect to their assigned subpopulations, we aim to evaluate cell identities. The observed data $Y_{m,n}$ contain m rows and n columns. In scRNA-seq data, we assume that single-cell samples are arranged as rows, whereas columns are genomic variables (e.g. genes). A variety of cell identities (Guo et al., 2015; McCarthy et al., 2017; Qu et al., 2017; Satija et al., 2015) are used for quality controls and normalization (Fig. 1). Furthermore, dimension reduction may be applied on genomic variables to highlight certain aspects of systematic variations or biomarkers. Therefore, n columns may be all available genes, highly variable genes, principal components or others. Nonetheless, when it is clear in context, we simply refer to n columns as genes.

Consider that m cells form K subpopulations, exhibiting distinct systematic patterns of variation. For $k = 1, ..., K$, a mutually exclusive subset of cells ($m_k$ out of m) is assigned to kth cluster. Then, $\sum_{k=1}^{K} m_k = m$. Samples within the kth cluster exhibit systematic variation that may be summarized by their center, centroid, medoid or other representative $c_k(Y)$ for $k = 1, ..., K$. In K-means clustering, the center is defined as the Euclidean mean; the nearest centers ($c_k$) are fast, robust and scalable to millions of single cells. What is overlooked in these recent developments is how to evaluate single cells, when their cell identities are determined by clustering. To the best of our knowledge, this represents the first study on estimating statistical significance of cluster membership at a single-cell level.

The proposed non-parametric methods leverage the assumption regarding cluster structure across single cells. Clustering algorithms estimate systematic variation and identify subsets of cells that contribute to distinct patterns. Due to high dimensionality of scRNA-seq, cells with ambiguous identities are artificially assigned to subpopulations, leading to weakened signals and false classifications. Our framework models and tests expression levels of cells with respect to their estimated subpopulations (Fig. 1). The jackstraw strategy accounts for overfitting characteristics of unsupervised clustering. Beyond P-values, an empirical Bayes approach is used to derive a probability that a cell truly belongs to an estimated subpopulation, which we call a PIP. This connects an unsupervised classification of high-dimensional data and a fundamental hypothesis framework in a statistically rigorous manner.

Operating characteristics of the proposed methods are demonstrated through comprehensive simulation studies. Three scRNA-seq data analyses are presented using (i) a mixture of Jurkat and 293T cell lines (Zheng et al., 2017), (ii) a Cell Hashing data of peripheral blood mononuclear cells (PBMCs) from eight independent donors (Stoeckius et al., 2018) and (iii) immune populations in 68 579 PBMCs from a single donor (Zheng et al., 2017). The reference implementation (https://CRAN.R-project.org/package=jackstraw) includes K-means clustering, partitioning around medoids (PAM; Kaufman and Rousseeuw, 1987) and mini batch K-means (MBKM) (Sculley, 2010), which are fast, robust and scalable to millions of single cells.

2.1 Jackstraw test for cluster membership

We propose to use the t-statistics to relate the single-cell samples Y and the cluster centers $c_k(Y)$ for $k = 1, ..., K$. Generally, the gene expression profiles of a given single cell y can be modeled with the cluster centers $c_k(Y)$ and other covariates $X_i$ resulting in an unrestricted full model $y_i \sim f_{full}(c_k(Y), X)$. Alternatively, a restricted
null model provides no information about \( c_k(Y) \) such that \( y_i \sim f_{null}(X_i) \). Then, an unadjusted residual sum of squares measures the discrepancy between \( y_i \) and two competing models,

\[
\text{RSS}_{null} = \sum (y_i - f_{null}(c_k(Y), X_i))^2
\]

(2)

\[
\text{RSS}_{null} = \sum (y_i - f_{null}(X_i))^2
\]

(3)

Then, the unadjusted \( F \)-statistics for the \( r \)th single cell is defined as

\[
F_r = \frac{\text{RSS}_{null} - \text{RSS}_{null}}{\text{RSS}_{null}/(n - \text{null})},
\]

(4)

where \( \text{null} \) denotes the number of parameters in the full model. However, because \( c_k(Y) \) is estimated from \( Y \), there is circular dependency resulting in artificially inflated significance (Fig. 2). To avoid circular analysis, the labels should be an independent variable that is measured externally. Using the data-dependent labels, such as cellular subpopulations derived from clustering, typically fails to control error rates. Therefore, conventional parametric or naive bootstrap-based \( F \)-tests (Supplementary Material), which expect dependent variables to be modeled by independent variables, are not valid.

We introduce a resampling-based approach to estimate the empirical distribution of \( F \)-statistics under the null model that adjusts for this circular dependency. This jackstraw approach, which was initially developed for PCA and related methods (Chung and Storey, 2015), constructs and utilizes a minimally disruptive jackstraw data \( Y^* \). Out of \( m \) observed samples, a relatively small number \((s < m)\) of samples are resampled with replacement, which we call synthetic null samples. Other \( m - s \) observed samples are unchanged. The jackstraw data \( Y^* \) combines \( s \) synthetic null samples and intact \( m - s \) observed samples. The cluster structure with \( K \) subpopulations is preserved in the jackstraw data, as \( s \) samples became independent and identically distributed (i.i.d) due to resampling with replacement.

When the jackstraw data are clustered, cluster centers \( c_k(Y^*) \) are almost identical to the original cluster centers \( c_k(Y) \) for \( k = 1, \ldots, K \) (Supplementary Fig. S2). Because of the nature of clustering algorithms, all samples in \( Y^* \), including \( s \) synthetic null samples, will be assigned to one of \( K \) clusters. When a synthetic null sample \( y_i^* \) is assigned to \( k \)th cluster, an association statistics between \( y_i^* \) and \( c_k(Y^*) \) is under the null model that assumes independence since \( y_i^* \) is i.i.d. by definition. Yet, because \( y_i^* \) is assigned to \( k \)th cluster, we learn the overfitting characteristics of clustering. Over a large number of iterations \( b = 1, \ldots, B \), the empirical distribution of null statistics is formed. This empirical distribution of null statistics is used to evaluate significance of individual samples (Algorithm 1).

The choices of \( s \) and \( B \) control the speed of computation, while the total number of null statistics \((s \times B)\) determines the overall \( P \)-value resolution. For \( B \) iterations, we need to cluster the jackstraw data \( B \) times, and for each iteration \( b = 1, \ldots, B \), we obtain \( s \) null statistics. Assuming \( s \times B \) is hold constant, a smaller \( s \) provides more accurate \( P \)-values, while increasing computational burden. Therefore, we want to ensure the original clusters are preserved as much as possible, permitting the computational power. As we increase the number of synthetic null samples \( s \) in \( Y^* \), the overall systematic variation captured by \( K \) cluster centers may be increasingly disrupted. Although we use \( s \sim 0.1 \times m \) for genomic data, the number of clusters \((K)\) and the proportion of samples assigned to them \((m_k, \ldots, m_k)\) must be considered.

The reference implementation (https://CRAN.R-project.org/package=jackstraw) uses \( K \)-means clustering (Hartigan and Wong, 1979; Lloyd, 1982; MacQueen, 1967), PAM (Kaufman and Rousseuw, 1987) and MBKM (Sculley, 2010). \( K \)-means clustering is one of the most established and popular algorithms (Hartigan and Wong, 1979; Lloyd, 1982; MacQueen, 1967). Particularly, considering a growing size of scRNA-seq data, \( K \)-means clustering is orders of magnitude more efficient than hierarchical clustering, graph-based community detection and density-based clustering (Tan et al., 2018; Xu and Su, 2015).

Furthermore, we incorporate a highly scalable mini batch version of \( K \)-means (Sculley, 2010), where a random subset of single-cell samples are used iteratively to update cluster centers and membership assignments (Steps 1 and 4 in Algorithm 1). Similarly, instead of randomly selecting cluster centers, \( K \)-means++ initialization may improve its convergence, which is available as a default option in the reference implementation (Arthur and Vassilvitskii, 2007). Because \( K \)-means clustering relies on Euclidean distance, one may be concerned about its robustness to outliers or generalizability to other distributions. By choosing observed data as cluster centers and using \( L_2 \) norm, PAM may perform more appropriately and is included in our jackstraw package.

### Algorithm 1: Jackstraw test for cluster membership

1. Apply the clustering algorithm to the observed data \( Y \), resulting in cluster centers \( c_k(Y) \) for \( k = 1, \ldots, K \) and membership assignments \( b_i,k \) for \( i = 1, \ldots, m \) and \( k = 1, \ldots, K \).
2. Compute the observed statistics \( F_1, \ldots, F_m \), where the full models include corresponding cluster centers \( c_k(Y) \).
3. Create \( s \) synthetic null samples by resampling with replacement a small proportion of samples \( s \ll m \), resulting in a jackstraw data \( Y^* \), with \( m - s \) observed samples and \( s \) synthetic null samples.
4. Apply the clustering algorithm to the jackstraw data \( Y^* \), resulting in cluster centers \( c_k(Y^*) \) and membership assignments \( b_i,k \).
5. Compute the null statistics \( F_1^*, \ldots, F_m^* \), where the full models include corresponding cluster centers \( c_k(Y^*) \).
6. Repeat the above three steps \( b = 1, \ldots, B \) times to obtain a total \( s \times B \) of null statistics.
7. Compute the \( P \)-values by empirically ranking the observed statistics among the null statistics.

### 2.2 Posterior inclusion probabilities

When clustering \( m \) samples into \( K \) subpopulations, the proposed jackstraw test estimates a probability that an individual cell may have been assigned to a given subpopulation by chance. We further propose to estimate posterior probabilities that \( m \) cells are correctly...
assigned to their clusters. This enables probabilistic feature selection and improved visualization of t-SNE, PCA and others.

Consider that the \( m \) jackstraw \( p \)-values \( p = p_1, \ldots, p_m \) are obtained for \( m \) single-cell samples that have been clustered into \( K \) subpopulations. We estimate a posterior probability that \( b_i \neq 0 \), since non-zero coefficients imply their bona fide inclusion in the clusters:

\[
\rho_i = \Pr(b_i \neq 0 | p_m) = 1 - \Pr(b_i = 0 | p_m). \tag{5}
\]

PIPs can be readily obtained by estimating \( \Pr(b_i = 0 | p_m) \) through an empirical Bayes approach (Efron, 2007; Efron et al., 2001). In multiple hypothesis testing, \( \Pr(b_i = 0 | p_m) \) is called a local false discovery rate (FDR). With a large amount of samples, it may be advantageous to consider posterior probabilities among each subpopulation or to improve estimation of FDRs and related quantities using prior biological knowledge. There also exist related Bayesian methods that could be explored for specific applications and prior distributions (Barbieri and Berger, 2004; Scott and Berger, 2006).

2.3 Feature selection and downstream uses

The proposed methods produce \( P \)-values and PIPs that are useful in downstream uses. Beyond their statistical properties, they may be used for visualization, feature selection and others.

The proposed \( m \) PIPs can be flexibly combined for downstream analyses, as to aid feature selection and dimension reduction. When applying the proposed methods to evaluate cell identities in scRNA-seq data, PIPs are used to hard-thresholding and soft-thresholding single-cell samples. First, in hard-thresholding, cells with low PIPs would be removed or masked for certain downstream analyses, achieving feature selection. For example, a subset of samples above a certain PIP threshold (e.g. > 0.8) may be visualized in reduced dimensions of t-SNE or PCA. Second, in soft-thresholding, PIPs may be used as weights for single cells for downstream analyses. In visualization, one may use PIPs to automatically control transparencies or colors that would emphasize samples with high PIPs. Our single-cell analyses demonstrate these downstream usages.

To select a threshold, one may estimate the proportion of null samples (Storey and Tibshirani, 2003). Then, samples with high PIPs or small \( P \)-values above that region would be selected accordingly. This automated procedure is used in the comparison of the proposed jackstraw to feature selection methods (Supplementary Material). Furthermore, we anticipate potential uses in weighted regression or weighted PCA in which cells with large PIPs may be considered more important than those with low PIPs. It may improve a wide range of clustering, such as improved assignments of single-cell samples to subpopulations and regularization of cluster centers.

3 Simulation studies

To demonstrate the operating characteristics of the proposed statistical tests, we conducted a comprehensive set of simulation studies, which enabled critical assessment of \( P \)-values using the ground truth. First, we generated a dataset from the model (Equation 1) while varying an amount of noise \((\sigma^2)\), a number of cells \((m)\) and a number of genes \((n)\). Second, we considered a cluster structure from gene expression profiles of 2700 PBMCs. Eight clusters with varying amounts of signals are used to simulate the data. Third, we conducted a Splatter experiment (Zappia et al., 2017) using human-induced pluripotent stem cell (iPSC) lines, in which all cells are derived from \( K = 3 \) subpopulations. Last, feature selection methods are applied and compared for cluster membership in scRNA-seq data.

First, we generated a large number of simulation configurations that may reflect scRNA-seq analyses. Generally, we investigated the operating characteristics of the proposed methods in simulated data with \( \sigma^2 = 5, 10, 15, m = 100, 1000, 2000 \) and \( n = 100, 1000, 2000 \) (Supplementary Fig. S1). Here, we focus on one scenario: \( m = 1000 \) cells, \( n = 100 \) genes and \( \sigma = 10 \). On the left, null \( P \)-values (corresponding to null hypotheses) are shown where a left skewed histogram of naive methods demonstrates an anti-conservative bias. In each of 100 simulation studies, null \( P \)-values are tested for uniformity by one-sided KS test. Then, 100 KS test \( P \)-values are plotted against a Uniform(0,1) distribution in a QQ-plot, where a downward deviation from diagonal dashed line indicates an overall anti-conservative behavior.

Fig. 3. Evaluation of the naive and proposed jackstraw methods for cluster memberships using simulation with \( m = 1000 \) cells, \( n = 100 \) genes and \( \sigma = 10 \). Theoretical Uniform(0,1)

Second, we used a dataset of 2700 PBMCs, called pbmc3k from 10x Genomics (https://s3-us-west-2.amazonaws.com/10x.files/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz) to generate scRNA-seq characteristics. Genes expressed in \( \geq 3 \) cells and cells with \( \geq 200 \) non-zero expression values are retained. After removing outliers, we log-normalized the data and regressed out technical variations due a number of unique molecular identifiers (UMIs) and a percentage of mitochondrial gene expression. Among 2638 PBMC samples, we selected 1838 highly variable genes. \( K \)-means clustering is applied on the resulting 2638 PBMC samples containing 1838 genes, using \( K = 8 \). These eight clusters contain

![Diagram](https://academic.oup.com/bioinformatics/article/36/10/3107/5788523 by guest on 20 September 2024)
346, 290, 177, 16, 186, 33, 1134 and 456 samples with diverse cluster centers. We use these eight clusters of pbmc3k data and their corresponding numbers of members to generate an identically sized dataset with 10% of i.i.d. null samples. Essentially, we simulated the PBMC dataset, where null samples are known. The proposed method was applied to evaluate cluster membership, with \( s = 264 \) and \( B = 100 \) (Supplementary Fig. S7). The jackstraw \( P \)-values corresponding to null samples follow a theoretically correct diagonal line with a KS \( P \)-value of 0.88. As expected, the true members of clusters correspond to highly significant \( P \)-values that are skewed toward to 0 (double KS test \( P \)-values < \( 2.2 \times 10^{-16} \)).

Third, we investigated the operating characteristics of the proposed methods when scRNA-seq data are simulated by Splatter (Zappia et al., 2017; Supplementary Material). Using Splatter, we investigated how the proposed jackstraw method operates when all of cells are indeed derived from in \( K = 3 \) subpopulations. The parameters for a Splat models are estimated from scRNA-seq data on human iPSC lines from the single-cell Fluidigm C1 platform (Tung et al., 2017). Following the application in Zappia et al. (2017), \( m = 400 \) cells from \( K = 3 \) subpopulations are simulated from probabilities of 0.60, 0.25 and 0.15. We applied the jackstraw for \( K \)-means clustering on a range of \( d \) eigenvectors as inspired by SC3 (Kiselev et al., 2017), which resulted in four cases using \( d = 0.04, 0.05, 0.06 \) and 0.07. We found that the \( P \)-values are highly significant such that almost all of cells are estimated to be included in their subpopulations with \( \hat{p}_0 < 0 \) (Supplementary Fig. S8).

Fourth, feature selection algorithms are compared using the main simulation scenario (\( m = 1000 \) cells, \( n = 100 \) genes and \( \sigma^2 = 10 \)). In particular, least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), max-min parents and children (Tsamardinos et al., 2003) and forward–backward selection with early dropping (Borboudakis and Tsamardinos, 2019) are applied with cross-validation for choosing hyper-parameters (Supplementary Material). For the jackstraw to automatically choose cells (e.g. ‘features’), the resulting \( P \)-values are used to estimate \( \hat{p}_0 \) (Storey and Tibshirani, 2003) and thresholded accordingly. The total number of positives, the false negative rates (FNRs) and the false positive rates are measured (Supplementary Fig. S8). Generally, the proposed jackstraw method outperforms, by considering the total number of positives, the false negative rates, the false positive rates and the FNRs. This is expected as most feature selection algorithms tend to fail at detecting the true positives, especially when the signal-to-noise ratio is low.

### 4 Single-cell analyses

Recent scRNA-seq studies obtain gene expression from single cells, in order to elucidate transcriptional heterogeneity (Jaitin et al., 2014; Macosko et al., 2015; Zheng et al., 2017). Cell identities are unknown at a single-cell level, even though heterogeneity is manifested on gene expression. Although cell identities are routinely obtained from unsupervised clustering, it may be important to test if cluster membership (e.g. placing a cell to a particular subpopulation) is correctly inferred. We applied the proposed methods on three scRNA-seq datasets.

Please note that there are a number of analytic steps prior to applying clustering to identify cellular subpopulations (Fig. 1). For example, normalization, gene selection and dimension reduction are considered to account for unwanted technical variation, to overcome a computationally bottleneck and to accentuate biological signals of interest (Brennecke et al., 2013; Hicks et al., 2018; Stegle et al., 2015). This series of challenges require understanding of study designs and goals, exploratory data analysis and sound statistical approaches. Our analyses directly utilize a number of carefully chosen choices in the original analyses.

![Cell identities in the Jurkat: 293T cell mixture data from Zheng et al. (2017). Two distinct cell lines from \( K = 2 \) cellular subpopulations. The proposed jackstraw method is applied on the top 10 PCs of UMI. (A) \( P \)-values from the proposed methods are plotted against the first PC. Two colored points correspond to two clusters. (B) At \( P < 0.80 \), 3.4% of 3381 single cells would be removed. Removing or down-weighting cells with low PIPs serve as feature selection for those with substantial association with presumed cellular subpopulations. (C) PIPs control transparency levels on the PC scatterplot. When PIP = 0, the data point is completely transparent.](https://academic.oup.com/bioinformatics/article/36/10/3107/5788523/fig}
mixture of cell lines, in which a number of putative doublets are specified by nExp = 1.3, 5% (details in Supplementary Material). The putative nulls with respect to subpopulations, identified by the jackstraw, significantly overlap with the putative doublets (mean Jaccard index $j = 0.34$ and $P$-value $< 2.2 \times 10^{-16}$; Chung et al., 2019). However, stemming from their distinct goals, their assumptions and operating characteristics are distinct (Supplementary Fig. S10). Generally, the proposed jackstraw test for cluster membership complements these multiplet detection methods, which rely on accurate estimation of cellular subpopulations.

4.2 Cell hashing and HTODemux classifications

Cell Hashing uses oligonucleotide-tagged antibodies against surface proteins to label single cells (Stoeckius et al., 2018). These labeled single cells can be pooled and sequenced together. The barcoded antibodies which correspond to different origins of cells are used to demultiplex the pooled samples, robustly identifying cell identities. In Stoeckius et al. (2018), PBMCs from eight donors were multiplexed in a single run of scRNA-seq. This Cell Hashing resulted in sequencing data of RNAs and hashtag oligonucleotides (HTOs) that are used to classify cells. Normalization and scaling were carried out as suggested in a corresponding Seurat vignette (Butler et al., 2018; Satija et al., 2015). The HTODemux algorithm in Seurat was applied for sample demultiplexing, which provide HTODemux classifications. ‘HTO-A’ through ‘HTO-H’ correspond to eight donors (i.e. ‘Singlet’).

Independent of the HTODemux classifications, the proposed methods were applied on 16,916 single cells in HTO data. The jackstraw estimates $P$-values of association between single cells and their clusters corresponding to different donors. We found that the resulting $P$-values are highly concordant with the HTODemux classifications. The distributions of $P$-values stratified by their classifications are distinct, where ‘Singlet’ (HTO-‘$s$’) cells are highly significant (Fig. 5). The mean $P$-values corresponding to ‘Singlet’ (HTO-‘$s$’), ‘Doublet’ and ‘Negative’ were 0.15, 0.53 and 0.49, respectively. Overall, Cell Hashing enables pooling different samples. Application of the jackstraw on the HTO data shows strong agreement with HTODemux classifications (Fig. 5). As Cell Hashing reports statistical errors such as FNR of $\sim 0.9\%$ (Stoeckius et al., 2018), the jackstraw $P$-values and PIPs may help demultiplexing and control overall error rates.

4.3 Immune populations among 68K PBMCs

We analyzed gene expression profiles of PBMCs from a single healthy donor (Zheng et al., 2017). PBMCs in human are consisted of heterogeneous cell types, such as lymphocytes (T, B and NK cells), monocytes and dendritic cells. The original analysis used $K = 10$ clusters to characterize transcriptional heterogeneity in this 68K PBMC dataset. Our methods identify the most relevant samples for these 10 clusters. Genes that are expressed in $> 1\%$ of observed cells and single-cell samples with $\geq 500$ genes were retained and processed using Seurat (Butler et al., 2018; Satija et al., 2015). We applied a log-normalization, followed by regressing out technical variations due to batch effects (eight channels), $%$ mitochondrial genes and numbers of UMIs. Directly reflecting the analytical choices in (Zheng et al., 2017), we selected the 1000 most variable genes by their dispersion among 40,507 PBMCs and obtained the top 50 PCs.

We applied MBKM clustering (Sculley, 2010) on the top 50 PCs obtained from this PBMC data. The proposed jackstraw test for cluster membership was applied with 10\% synthetic null samples and 100 iterations (Supplementary Fig. S12). The proportion of null samples is estimated to be $\pi_0 = 0.124$. At $P$-value $> 0.80$ and $> 0.90$, we found that 34,134 (84.2\%) and 22,407 (55.3\%) single-cell samples are assigned to their corresponding 10 clusters, respectively. Using a perplexity parameter of 30, t-SNE projection after our feature selection suggests that the proposed methods help remove cells with ambiguous identities (Fig. 6). Due to a stochastic nature of t-SNE, separate runs may result in different projections. Therefore, one may also remove a subset of cells with low PIPs using the original t-SNE projection, which is shown in Supplementary Figure S13.

Note that with $100\%$ of samples for initialization and $10\%$ batch size, MBKM clustering took $\sim 4\, s$ for 10 starts and $1000$ maximum iterations. In contrast, $K$-means clustering on this dataset required $20–21\, s$ (MacBookPro i5 2.4 GHz).

5 Discussion

scRNA-seq enables genome-wide quantification of gene expression in tens of thousands of single cells. Transcriptional heterogeneity in scRNA-seq data is routinely characterized by estimating cell identities using unsupervised clustering. We introduce a set of methods to rigorously test clustering-based cell identities, estimate PIPs and improve downstream visualization. By learning the overfitting characteristics inherent in applying clustering to high-dimensional data, the proposed methods guard against artificially inflated significances.

Our key insight is to generate and re-cluster the jackstraw data, in which a small number of synthetic null samples are used to derive the empirical null distribution. Comprehensive simulation studies demonstrated accurate operating characteristics, including rigorous error controls. Applications on three scRNA-seq datasets showcase how the proposed methods enable probabilistic feature selection and improved projections of PCA or t-SNE. Interestingly, ambiguous single cells such as multiplets are shown to contaminate high-throughput scRNA-seq data. Therefore, the proposed methods may help in quality control and identification of major molecular signatures.

When any clustering method is applied for estimation of single-cell identities, a number of clusters must be determined. Identifying an optimal number of clusters is a fundamental challenge (Akaike, 1974; Bock, 1985; Jain and Moreau, 1987; Tibshirani et al., 2001) that is beyond the scope of this study. Even if a clustering algorithm sidesteps an explicit input for a number of clusters, hyper-parameters such as a resolution, a number of nearest neighbors or a modularity are required from the user. These hyper-parameters indirectly set the number of clusters for a given scRNA-seq data. Overall, exploration of data with domain knowledge and computational analysis would help finding these parameters.
The proposed methods can aid in feature selection, biomarker identification and visualization. First, cells with low PIPs may be removed from downstream analyses, in a similar manner to quality control. Second, PIPs may be used as visual elements (e.g. alpha levels) in scatter plots and others. Third, using PIPs, one may potentially carry out weighted regression or weighted PCA. Fourth, cells with high PIPs may be used for identifying genes that are differentially expressed across conditions or labels. On the other hand, some scRNA-seq data may contain a multi-level structure. This resulted in iteratively applying a clustering algorithm, often supported by qualitative analysis and biological expertise (Macosko et al., 2015; Zheng et al., 2017). Such multi-level clustering may be improved if the proposed PIPs and feature selection are applied. In the future, the jackstraw may be further developed into an integrated method for Bayesian multilevel clustering.

The jackstraw tests for latent variables (Chung and Storey, 2015) have been used in a variety of genomic studies (e.g. alpha levels) in scRNA-seq data may contain a multi-level structure. This resulted in iteratively applying a clustering algorithm, often supported by qualitative analysis and biological expertise (Macosko et al., 2015; Zheng et al., 2017). Complementing this, the proposed tests help evaluate cluster membership, such that clustering-based subpopulations can be rigorously used in downstream analyses. This opens new possibilities for selecting canonical cluster members, shrinking cluster centers and guiding the choice of stable clusters. Because the proposed methods are not limited to scRNA-seq, we anticipate its adaptation in other data-intensive fields.

Acknowledgements

The author thanks Prof. John D. Storey of Princeton University for early discussions and encouragement.

Funding

This work was supported by the Narodowe Centrum Nauki [2016/23/D/ST6/03613].

Conflict of Interest: none declared.

References

Akaike,H. (1974) A new look at the statistical model identification. IEEE Trans. Autom. Control, 19, 716–723.

Alon,U. et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA, 96, 6745–6750.

Andrews,T.S. and Hemberg,M. (2018) Identifying cell populations with scRNA-seq. Mol. Aspects Med., 59, 114–122.

Arthur,D. and Vassilvitskii,S. (2007) K-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). Society for Industrial and Applied Mathematics Philadelphia, PA, USA, pp. 1027–1035.

Barbieri,M. and Berger,J. (2004) Optimal predictive model selection. Ann. Stat., 32, 870–897.

Bock,H.-H. (1985) On some significance tests in cluster analysis. J. Classif., 2, 77–108.

Boroudakis,G. and Tsamardinos,I. (2019) Forward-backward selection with early dropping. J. Mach. Learn. Res., 20, 276–314.

Buettner,F. et al. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol., 33, 155–160.

Butler,A. et al. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol., 36, 411–420.

Chung,N.C. and Storey,J.D. (2015) Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics, 31, 545–554.

Chung,N.C. et al. (2017) Population structure analysis of bull genomes of European and Western ancestry. Sci. Rep., 7, 40688.

Chung,N.C. et al. (2019) Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. BMC Bioinformatics, 20, 644.

Efron,B. (2007) Size, power and false discovery rates. Ann. Stat., 35, 1351–1377.

Efron,B. et al. (2001) Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc., 96, 1115–1160.

Eisen,M.B. et al. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95, 14863–14868.

Farre,P. et al. (2015) Concordant and discordant DNA methylation signatures of aging in human blood and brain. Epigenetics Chromatin, 8, 19.

Gasch,A.P. et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell, 11, 4241–4257.

George,E.I. and McCulloch,R.E. (1997) Approaches for Bayesian variable selection. Stat. Sin., 7, 339–373.

Golub,T.R. et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286, 531–537.

Guo,M. et al. (2015) Sincera: a pipeline for single-cell RNA-seq profiling analysis. PLoS Comput. Biol., 11, e1004575.

Hartigan,J.A. and Wong,M.A. (1979) Algorithm as 136: a k-means clustering algorithm. J. R. Stat. Soc. C, 28, 100–108.

Hicks,S.C. et al. (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics, 19, 562–578.

Jain,A.K. and Moreau,J.V. (1987) Bootstrap technique in cluster analysis. J. Am. Stat. Assoc., 82, 339–373.

Jaitin,D.A. et al. (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science, 343, 776–779.
