**Abstract**

Droplet-based single-cell RNA sequencing (scRNA-seq) is a powerful tool for elucidating developmental, physiological, and pathological processes in biological systems. Unsupervised clustering of scRNA-seq data is a crucial(key/indispensable/necessary/obligatory) step in scRNA-seq analysis workflows that enables the identification of distinct cell types, subtypes, and states. However most clustering algorithms require an estimate of the number of distinct cell types present, as well as the dimensionality of the data. Such *a priori* knowledge is not always available, especially for poorly characterized tissues, which can result in many different interpretations of identical datasets. To address this unmet challenge, we have developed AutoClustR, a tool for automated and unbiased single-cell clustering. We compared 7 methods of dimensionality selection and 14 different ICVIs to empirically ground AutoClustR’s approach to automated clustering. AutoClustR was benchmarked and shown to accurately identify (x percent of cells?), outperforming 6 alternate scRNA-seq analysis platforms. Then, AutoClustR was applied to a real-world dataset derived from human embryonic stem cell-derived inner ear organoids to reveal a previously unappreciated diversity of cell types. AutoClustR’s approach allows researchers to characterize novel datasets, and the empirical support for this approach is a valuable resource for fellow bioinformaticians.

**Introduction**

Single-cell RNA sequencing (scRNA-seq) allows for whole transcriptome profiling at the level of individual cells, which has yielded new insights into processes such as embryonic development, tissue regeneration, and disease pathogenesis. This technology enables high-resolution interrogation of the transcriptome, revealing the full transcriptional diversity of entire organs or developmental processes. However, this high-resolution approach presents a big-data problem. Extracting meaningful information from large scRNA-seq datasets (with thousands to tens-of-thousands of cells, each expressing thousands of genes) is impossible without the help of machine learning and similarly focused computational tools [[and machine learning approaches]].

One such tool is unsupervised clustering. A crucial step in the standard workflow, unsupervised clustering groups cells based on similar (or dissimilar) gene expression profiles, without any *a priori* information on which genes are considered important. It is referred to as “unsupervised” because there are no labelled training datasets that can be used to refine the method by which cells are classified/grouped. Cells are grouped into computationally determined/defined clusters that are presumed to represent cell types, physical regions, or developmental stages. Then, downstream analysis is performed on these clusters, yielding findings more statistically robust than could be obtained by examining individual cells. Because these clusters are used as inputs for downstream analysis, false clusters (e.g., where two distinct cell types are erroneously grouped together) negatively impact results and obscure biological truths the data contain. Therefore, accurate cluster identification is of the utmost importance.

Many different platforms for scRNA-seq analysis exist, each employing a unique workflow to transform raw data into clusters of cells. Regardless of the specific algorithms the platforms employ, there are two major choices they share. Each platform requires some form of dimensional reduction, compressing the 20,000+ dimensional space defined by the genes in the human transcriptome into a smaller number of dimensions required to make subsequent transformations computationally feasible. The exact number of dimensions to retain, or **dimensionality**, is a choice that is usually left up to the user. The second choice requires users to select parameters or hyper-parameters that determine the performance of the clustering algorithm. This choice is frequently obscured, with platforms using default values unless tuning is performed by end-users, changing clustering parameters can dramatically influence the final clustering solution. In fact, a REVIEW by XYZ et al. found that changing the hyper-parameters of a platform can profoundly influence performance, with intra-platform variance equal to or greater than inter-platform variance.

The most common approach to dimensional reduction is principal component analysis, PCA, which constructs linear combinations of variables to maximize the variance in the data explained by each such principal component. After PCA is performed, researchers must decide on the dimensionality of the data, i.e., how many principal components to retain for downstream calculations. Generally, somewhere between 5 and 25 PCs are retained. The question of how many principal component to retain is well researched, with the first such discussion appearing as far back as 1950 [M.S. Bartlett.] One common approach is to plot the variance explained by each PC against the principal component number. This is referred to as a scree plot, attributed to psychologist Raymond Cattell in 1966. Several different approaches take advantage of the scree plot to automate principal component retention. Despite the large body of research that addresses this problem, it’s not uncommon for new scRNA-seq analysis platforms to implement new methods of principal component retention without considering or comparing existing methods.

After the dimensionality of the data has been reduced and the dimensionality determined, clustering the data becomes computationally feasible. Like principal component retention, the field of unsupervised clustering is well researched, and many different techniques have been developed to group n-dimensional data into clusters. A number of generic clustering algorithms have been applied to scRNA-seq data (e.g. k-means and -medoids clustering [7, 8], density-based clustering [9], graph-based clustering [10], and hierarchical clustering [11]), and many more have been developed specifically for single-cell clustering (e.g. CIDR [12], BackSPIN [13], DIMM-SC [14], and BAMM-SC [15]). While these clustering algorithms differ substantially in their approach, all require some direct (e.g., k in k-means clustering) or indirect (e.g., resolution in Louvain community detection) estimate of the cluster number. Many algorithms require additional parameters (e.g., n in n-nearest neighbors) that influence cluster solutions in more subtle ways. The impact that these parameters have on the results of the analysis are often poorly understood, and as such, the increased performance that hyper-parameter optimization could afford goes unrealized.