Making use of permutation testing in single cell analysis

Single cell RNA Sequencing (scRNA-seq) has rapidly gained popularity over the last few years for profiling the transcriptomes of thousands to millions of individual cells. This new technology has enhanced our understanding of complex tissues and enabled the discovery of novel cell types. However, the analysis of scRNA-seq data can be challenging as the data is sparse and prone to technical artifacts such as batch effects. One of the first challenges in analysing this data is discovering which cell types are present in sequenced samples. As such, a crucial step in scRNA-seq analysis is clustering cells based on the similarity of their transcriptional profiles. With these groupings of cells specified, it then becomes a matter of annotating the cell type that is represented by the cluster. While clustering has received a lot of attention in the literature, areas that have been less well explored are (1) appropriate statistical tests for discovering genes that are enriched in each cluster, i.e. marker gene analysis, and (2), the interpretation of cell type specific marker genes, to aid in cluster identification, in the form of gene set testing. In particular, cell-to-cell and sample-to-sample variability are often ignored in these types of analyses, with cells within samples treated as independent entities. Ignoring the correlation structure of the data leads to p-values that are over optimistic, inflating the false discovery rate. Permutation based methods, while generally inappropriate for bulk RNA-seq due to small sample sizes, have great potential for application in single cell analysis as there are typically thousands of cells per experiment. This allows for accurate p-value estimation from empirical distributions, while preserving the correlation structure inherent in the data. The goal of this project is to develop a computationally efficient permutation-based testing framework that will simultaneously perform marker gene analysis and gene set testing to discover the cell types that are present in the data. These methods will provide more accurate p-value estimates and appropriately take into account the different levels of variability in the data. The methods will be implemented in the speckle R package and will be publicly available for the research community.