**Making use of permutation testing in single cell analysis - Literature Review**

Single-cell RNA Sequencing (scRNA-seq) has growing popularity in recent years. This technology has the advantage of profiling the transcriptomes of individual cells and supports the understanding of cellular heterogeneity [10] [15]. scRNA-seq produces cells and genes in a large scale, and the analysis of such data tends to be challenging. scRNA-seq data has high variabilities and is sparse and noisy. A typical workflow starts from pre-processing step of the raw count data. Various tools exist for performing quality control. In general, cell-level quality control can filter potential doublets or empty droplets, and gene-level quality control will remove genes with low expression levels. Genes and cells that pass the quality control can be used for downstream analysis.

One common cell-level analysis is to find cell types that exist in the data. Cells are often clustered, and different methods are applied to annotate the clusters. One approach to annotate the computed clusters is to identify marker genes that are representative of each cluster. Marker genes are identified with differential expression testing. As the number of marker genes tends to be large, the obtained marker genes are typically examined as a group. The group of marker genes is tested against many well-known gene sets to identify over-represented biological pathways or processes involved in. The biological pathways will aid the understanding of existing cell identities.

Manual cell-type annotation involves finding marker genes that are highly regulated for the cluster of interest. Marker genes are identified by performing differential expression testing between the cells in the interested cluster versus all the rest cells. A considerable amount of literature has been published on marker-gene identification. Tools developed originally for bulk RNA-seq data, such as limma [16], edgeR [14], DESeq [9] can be applied to scRNA-seq data. Such tools generally detect DE genes based on the mean expression level of genes. There are methods specifically designed for scRNA-seq, and the prominent ones are single-cell differential expression (SCDE), model-based analysis of single-cell transcriptomics (MAST), scDD, and D3E. MAST [3] provides a generalized linear model that models the expression rate (logistic regression) and the positive expression means (Gaussian linear model) jointly for each gene. Besides, SCDE [6] is a Bayesian framework that models each cell as a mixture of a Poisson distribution (normal) and a Negative Binomial distribution (dropout events) to accommodate high variability in scENA-seq data. Moreover, scDD [8] implements a mixture model to detect DE genes and classifies these genes into four specific differential distribution patterns. It is noticed that scDD uses the permutation test to assess the significance of the calculated Bayes factor score for each gene, which is further used for selecting marker genes. D3E [2] provides a framework that uses two non-parametric models to compare each gene's expression value distribution to identify marker genes.

The literature comparing popular DE methods has highlighted that current tools have significant variations in performance, including the number of detected marker genes, sensitivity, false discovery rate [5][17]. More importantly, tools adapted from bulk RNA-seq methods are argued to be comparable to those designed specifically for scRNA-seq.

Gene set testing requires a database that provides known gene sets for comparison. Gene Ontology [1] contains different categories of gene sets involved in biological domains and is widely used for gene set testing. Different methods exist in the literature regarding gene set testing, and most methods are originally designed under the bulk RNA sequence setting. Competitive gene set tests, such as Parametric Analysis of Gene Set Enrichment (PAGE), Gene set enrichment analysis (GSEA), and Correlation Adjusted MEan RAnk gene set test (CAMERA) aim to find enrichment of specific sets of genes. GSEA [18] calculates an enrichment score based on a weighted Kolmogorov–Smirnov-like statistic for the tested gene set and performs a permutation test to estimate the significance level. PAGE [7] computes a Z-score for a given gene set and uses standard normal distribution for statistical inference. Besides, CAMERA [21] relies on a gene-wise moderated t-statistic with an additional estimation of inter-gene correlation. The study [12] [19] comparing various gene set methods indicates the performance of gene set testing methods varies in different settings, and the applicability to scRNA-seq data is less studied [11].

Moreover, the survey [4] indicates that the absolute value of enrichment p-values calculated by GSEA or PAGE is not reliable and sensitive to the tested gene set size. One integrative method iDEA [11] that developed for scRNA-seq uses a hierarchical Bayesian framework to join the DE analysis and gene set testing. iDEA models a joint distribution of DE and gene set testing for all genes to enhance the power in both analyses.

Permutation test [13] is a non-parametric test that can be used to estimate the null distribution of the test statistic by randomly sampling the observation labels without replacement. Commonly used DE tools tend to give under-estimated p-value estimates, resulting in an overestimation in the number of up-regulated genes. scRNA-seq usually contains many cells in different cell identities, and permutation-based methods can approximate empirical distribution without specific distribution assumptions for gene expression values. The P-value for each gene is expected to be more accurate with permutation testing than model-based methods, and therefore control the false positive rate in detected marker genes. Similarly, for gene set testing, the permutation test is expected to give more reliable p values for tested genes.

There is a relatively small body of literature that is concerned with applying permutation methods to find marker genes. The time and design complexity for those methods possibly limit the practical applications. SigEMD [20] employs additional models to detect a set of influential genes for imputing the zero counts. It calculates a score for each gene based on Earth Mover’s Distance to reflect the overall differences between the expression distributions and performs the permutation test to identify marker genes. SigEMD is argued to have reduced false discovery rates with the sacrifice of running time.

**Project Plan**

The key purpose is to develop a permutation-based framework that performs marker identification to discover underlying cell types in the data. It is aimed to provide more accurate p-value estimates with the consideration of cell-to-cell and sample-to-sample variabilities in a computationally efficient framework.

This project will focus on a specific dataset with annotated cell labels. Standard data cleaning and clustering workflow will be applied. A permutation-based framework will be developed, and different test statistics will be explored for marker gene analysis and gene set testing. The effectiveness with different test statistics will be tested in simulation [22] and real data. The performance will be compared to available tools such as limma. Attention will be paid to the time efficiency of the framework. The final permutation framework will be implemented in the speckle R package.

[1] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the identification of biology. *Nat Genet,* *25*, 25–9.

[2] Delmans, M., & Hemberg, M. (2016). Discrete distributional differential expression (D3E)--a tool for gene expression analysis of single-cell RNA-seq data. *BMC bioinformatics*, *17*, 110. https://doi.org/10.1186/s12859-016-0944-6

[3] Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., Linsley, P.S. & Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data,  *Genome Biology*, 16:278. http://dx. doi.org/10.1186/s13059-015-0844-5.

[4] Huang, d., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923

[5] Jaakkola, M. K., Seyednasrollah, F., Mehmood, A., Elo, L. L. (2017). Comparison of methods to detect differentially expressed genes between single-cell populations, *Briefings in Bioinformatics*, *18*(5), 735–743.

[6] Kharchenko, P. V., Silberstein, L. & Scadden, D. T. (2014). Bayesian approach to single-cell differential expres- sion analysis. *Nat. Methods,* *11*(7), 740–742.

[7] Kim, S. Y., & Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics*, *6*, 144. https://doi.org/10.1186/1471-2105-6-144

[8] Korthauer, K. D., Chu, L. F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., & Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome biology*, *17*(1), 222. https://doi.org/10.1186/s13059-016-1077-y

[9] Love, M. I., Huber, W. & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* *15*. https://doi.org/10.1186/s13059-014-0550-8

[10] Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology*, *15*(6), e8746. https://doi.org/10.15252/msb.20188746

[11] Ma, Y., Sun, S., Shang, X., Keller, E. T., Chen, M., & Zhou, X. (2020). Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. *Nature communications*, *11*(1), 1585. https://doi.org/10.1038/s41467-020-15298-6

[12] Mathur, R., Rotroff, D., Ma, J., Shojaie, A., Motsinger-Reif, A. (2018) Gene set analysis methods: a systematic comparison. *BioData Mining*, *11*. https://doi.org/10.1186/s13040-018-0166-8

[13] Phipson, B., & Smyth, G. K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, *9*, Article39. https://doi.org/10.2202/1544-6115.1585

[14] Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

[15] Saliba, A. E., Westermann, A. J., Gorski, S. A., & Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. *Nucleic acids research*, *42*(14), 8845–8860. https://doi.org/10.1093/nar/gku555

[16] Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical applications in genetics and molecular biology*. *3.* https://doi.org/10.2202/1544-6115.1027

[17] Soneson, C., Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. *15*, 255–61. https://doi.org/10.1038/nmeth.4612

[18] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

[19] Tarca AL, Bhatti G, Romero R (2013) A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. PLOS ONE *8*(11), e79217. https://doi.org/10.1371/journal.pone.0079217

[20] Wang, T., Nabavi, S. (2018). SigEMD: a powerful method for differential gene expression analysis in single-cell RNA sequencing data. Methods San Diego Calif, *145*, 25–32.

[21] Wu, D., & Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, *40*(17), e133. https://doi.org/10.1093/nar/gks461

[22] Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome biology*, *18*(1), 174. https://doi.org/10.1186/s13059-017-1305-0