LITERATURE REVIEW: NEW MEDICAL APPLICATIONS OF MULTIVARIATE CLASSIFICATION METHODS

1. Classification of single cell transcriptomic profiles

Transcriptomics refers to the study of the RNA molecules in a cell, where. The study of gene expression profiles on a single-cell basis can be extremely useful in medical research. This is especially true in oncological research, for example, which often entails the analysis of tumour tissue samples – comprising of an assortment of different and mutating cell-types. Fan, Slowikowski and Zhang (2020) highlight the advantages of single-cell transcriptomics when compared to analysis conducted on pools of heterogenous cell-types. In many downstream applications, the ability to not only distinguish between different cell-populations but also to accurately classify them is of great importance.

Major advancements have been seen in the field of computational transcriptomics since the development of single-cell RNA sequencing [sc-RNASeq]. This technique yields isolated and sequenced gene expressions of individual cells, allowing cell-type to be determined through analysing the heterogeneity between the transcriptome of different cell populations. Challenges are presented in analysing this data since the dimensionality is extremely high (mammalian samples often have $\sim 10^4$ genes), and thus most analyses begin with reducing the feature space. This can be done through two general approaches (Andrews & Hemberg, 2018). The first method involves generating a lower-dimensional representation of the data which preserves some underlying characteristic (such as between-cell heterogeneity) through some statistical procedure such as PCA. The second method involves implementing some feature selection algorithm to exclude uninformative genes from the data. It is usually beneficial a combination of these approaches to reduce the feature space optimally.

Given the reduced feature space, the data can then be used to classify cell populations. Fan, Slowikowski and Zhang (2020) detail several contemporary computational methods enabling this classification from high-throughput sc-RNASeq data. The authors further outline several other analyses relevant to oncological research. Andrews and Hemberg (2018) provide a more specific review of cell identification from sc-RNASeq data. The identification of cell populations is most often implemented through an unsupervised clustering algorithm. These methods can be broadly grouped into non-heirarchical and hierarchical clustering algorithms. The former assumes a predetermined number of cell populations of similar shape and size. To this extent, the k-means algorithm is commonly used. When this is not appropriate, hierarchical algorithms such as Ward’s (Ward, 1963) can be used. Further methods can be explored, such as density-based clustering algorithms like DBSCAN (Ester et. al., 1996) – however these also require a pre-specified number of clusters and assume all cell populations are equally homogeneous. These approaches all thus have some obvious limitations relating to the identification of less frequent cell types and cell types with different between-cell homogeneity.

Adaptations of these algorithms have been developed for the specific application of single cell classification, which often entail some combination of dimension reduction and unsupervised clustering. Mahalanabis et al (2022) compares the performance of 15 commonly used algorithms over 8 datasets containing gene expressions of various tumour samples. The researchers identify the superior performance of three algorithms on non-malignant samples (Seurat, bigSCale and Cell Ranger) and two on malignant samples (Monocle and SC3).

One of the major challenges of unsupervised clustering approaches is that, even if the chosen algorithm correctly distinguishes cell populations, the assignment of cell cluster to cell population is heavily reliant on the clinical expertise of the investigator. Xie et al (2019) thus develop a supervised classifier, SuperCT, which uses mouse-cell training data to enable the supervised learning of an artificial neural network. The authors argue that this solves the many limitations of the unsupervised approach. The authors acknowledge that, in an intuitive sense, one would expect better performance from a model trained on human data but argue (with evidence) that the performance of the classifier should be applicable to other mammalian species.

Andrews, S. & Hemberg, M. 2018. Identifying cell populations with scRNASeq, Molecular Aspects of Medicine, Volume 59:114-122. ISSN 0098-2997. <https://doi.org/10.1016/j.mam.2017.07.002>.

Fan, J., Slowikowski, K. & Zhang, F. 2020. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp Mol Med* **52**, 1452–1465. <https://doi.org/10.1038/s12276-020-0422-0>

Mahalanabis, A., Turinsky, A., Husić, M., Christensen, E., Luo, P., Naidas, A., Brudno, M., Pugh, T., Ramani, A., & Shooshtari, P. 2022. Evaluation of single-cell RNA-seq clustering algorithms on cancer tumor datasets. Computational and Structural Biotechnology Journal. Volume 20: 6375-6387. ISSN 2001-0370. <https://doi.org/10.1016/j.csbj.2022.10.029>.

Xie, P., Gao, M., Wang, C., Zhang, J., Noel, P., Yang, C., Von Hoff, D., Han, H., Zhang, M. & Lin, W. 2019. SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles, Nucleic Acids Research, Volume 47, Issue 8:48. <https://doi.org/10.1093/nar/gkz116>