Evaluation of network methods for the analysis scRNA-seq data and development of a new KNN-based method based on identified caveats

Anirudh Patir and Tom Freeman

The University of Edinburgh, Edinburgh, UK

Advances in single-cell RNA-seq (scRNA-seq) analyses have revolutionised the ability to characterise cellular heterogeneity and opened up new opportunities to study cell and tissue biology in development or disease, e.g. tracking differentiation trajectories of cells (pseudotime analysis). However, the data generated is large and technically noisy, making several methods for analysing bulk RNA-seq data unsuitable. As a result, a plethora of tools have been developed for the visualisation and analysis of scRNA-seq data, with the lack of a comprehensive comparison. Hence, in this study, we compare the non-stochastic network-based methods that are gaining popularity in the field, including Phenograph, correlation analysis, KNN, and SNN. These methods have been applied to real datasets of variable cell numbers and evaluated for their sensitivity and stability in identifying cell-subtypes of varying proportions, and in detecting outliers. Conventional correlation analysis easily identifies outliers, though unable to distinguish cell-subtypes. In contrast, KNN & SNN based approaches can detect these subtypes, however, without addressing outliers and with the stability of clusters relying on the parameter ‘k’. In all methods, datasets containing variable proportions of cell populations resulted in misclassification of subtypes. To address these two limitations, we developed a modified KNN algorithm. Firstly, outliers are revealed, as a variable set of edges are assigned to a node, depending on the respective correlations. This is in contrast to the KNN based algorithms, which does not distinguish between poor and strong correlations considered. Secondly, the current algorithm can identify cell subtypes of variable proportions by constructing the network based on the density of cell-types. Additionally, by visualising the relevant structure of the data, differentiation trajectories can also be described. Together this study identifies the caveats associated with current network approaches and proposes an algorithm to address these limitations, in particular highlighting outliers and distinguishing cell-types of varied proportions.