# Single-Cell RNA Sequencing:Annotated Bibliographies

Pak Hin Yu

MSc Bioinformatic with System Biology
Department of Biological Science
Birkbeck, University of Lodon

June 2017

# Reference

Danaher, P., Warren, S., Dennis, L., D'Amico, L., White, A., Disis, M. L., Geller, M. A., Odunsi, K., Beechem, J., and Fling, S. P. Gene expression markers of tumor infiltrating leukocytes. *Journal for immunotherapy of cancer 5*, 1 (2017), 18

Danaher *et.al.*'s study proposes a set of marker genes that can be used for measuring 14 immune cell sub-populations in a tumor microenvironment: tumor infiltrating lymphocytes(TILs) in gene expression assays.The proposed list of gene marker as well as the underlying statistical methods can be potentially useful in formulating relevant gene express signatures for immune cell identification.

In tumor cell diagnostics and treatments, flow cytometry and immunohistochemsitry(ICH) are often used to quantify immune cell population, but those methods can only measure few gene markers.Gene expression profiling ,on the other hand, can provide more clinically actionable information. To determine the marker genes that can be used for profiling each cell types, Danaher's team first relied on the results of past studies of purified immune cell population, then develop a novel statistical method to select genes that exhibits maker behavior in a a tumor microenvironment.This method is based upon an adaption of Pearson correlation in which it takes into account that many biologically-related genes from different cell type may exhibit correlation. By applying this methods, only 60 genes out of 356 candidate genes are selected in which each the quality of the select gene markers is varied among each cell type. As the authors show, the cell scores derived from the expression of selected gene markers broadly agree with both flow cytometry and ICH and show good reproduciblity among 12 different tumor samples. They further demonstrate the scores can be used to access the change in cell population during immunotherapy which shows the scoring method can be useful in both discovery and clinical research.

This study provides a straightforward method to measure cell population in tumor sample which can be used in RNA sequencing.However,as the authors indicate, the marker selection is based on the expression pattern of the tumor cells from The Cancer Genome Atlas, so that adjustment may be deemed necessary for non-tumor samples .

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods 12*, 5 (2015), 453–457

In this study, Newman *et.al.* introduce a new method, known as Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts (CYBERSORT), to measure the makeup of the cells from complex tissue. Its novel use of nu-support support vector regression ($\nu$SVR), a machine learning method, for cell type classification can be applied to single-cell RNA sequencing(scRNA-seq). Comparing to other computational methonds, the result suggests that CYBERSORT performs considerably well in classifying cells from mixture of unknown content and noise , (in the case of solid tumors) and cells from closely related cell-type (in the case of mixture of naive and memory B cells).The authors claim that the superior performance is because using $\nu$SVR as a feature selection helps minimize a loss function and penalty function. A linear loss function used in $\nu$SVR gives robustness to noise and over-fitting of the data while the use of $L_2$-norm penalty function offers tolerance to multicollinearity (predictors that are inter-correlated to each other) in which the gene expression profile would not be heavily biased toward the most correlated cell-type. Apart from $\nu$SVR,the authors also stress the importance of building cell-specific expression signature, a preprocessing step that filters irrelevant features of the signature before being applied to machine learning process, in which it can speed up the computation running time and increase the signal to noise ratio of the data.Hence, the authors conclude the use of $\nu$SVR in CYBERSORT as well as various statistical refinements address the critical issues of gene expression deconvolution for nearly any tissues.

Noticeably, all gene signature profiles are obtained from microarray experiments in this study, so that it is unclear if the result would be different from the RNA-seq data.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology 33*, 5 (2015), 495–502

Satija *et.al.* introduce Seurat, a computation method, to infer cell location of single-cell RNA sequencing (scRNA-seq) data from titin situ spatial RNA pattern. Even though the focus of Seurat application is on spatial inference in this paper, its underling computational strategy, such as dealing with heterogeneous experimental data, can be applied to other purposes, such as developmental states or disease phenotype.

Satija's team uses zebrafish embryo scRNA-seq data to demonstrate the accuracy of Seurat in which 851 cells are spatially mapped.They first use *in situ* hybridization result of the "landmark" genes of the embryo to construct a spatial reference map in which tissue is divided into user-defined discrete spatial domains, known as bins. In each bin, landmark genes are either labelled as "on" or "off" to create a distinct binary expression reference profile. Since only a small set of genes are used for spatial assessment on scRNA-seq data, the resulting inference is sensitive to technical noises, such as false negative and measurement errors.Instead of relying solely on the expression level of the landmark genes, Seurat built a separate model based on the expression of co-regulated genes of those genes to minimize the noise of individual measurement.Because of the difference in the nature of the continuous scRNA-seq data and binary reference profiles, Seurat uses bimodal distributions to relate both data to build a model in which posterior probability of each cell's origin from each of the bins can be calculated based on likelihood of individual cell's gene expressions being "on". This probabilistic approach also allows cells to be assigned into multiple bins if the cell cannot be assigned to one bin exclusively.The inferred spatial pattern of the cells is largely consistently with empirical data from benchmarking experiments and literature.

Satija further applies Seurat to analyze scRNA-seq data without the use of landmark genes. The cells are first clustered by Principal Components Analysis (PCA) and k-mean clustering. Then, each cluster is identified by the expression of gene markers and analyzed by Seurat. This unsupervised approach is able to assign rare subpopulations of zebrafish embryo cells to the expected location. This demonstrates Seurat can be used as a discovery tool to identify unknown cell population within complex tissues.

Even though Seurat is proven to be a versatile and powerful tool for spatial discovery, it might not be applicable to some tissues, as Seurat relies on the spatial distinctiveness of each gene expression profile. Tissues, such as tumor cell or adult retina, where their gene expression might not have a differential spatial pattern for Seurat to use as reference.

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science 356*, 6335 (2017), eaah4573

Villani *et.al.* demonstrated the use of Seurat to discover new type of Dendritic cells(DCs) and monocytes from the single-cell RNA sequencing (scRNA-seq) data of a healthy donor's blood sample. This paper highlighted the remarkable capability of scRNA-seq in characterizing cell type, in conjunction with using traditional approaches: molecular markers,functional properties and ontogeny. With a more comprehensive gene expression profiles at single-cell level, Villani's team is able to distinguish six DC population and four monocytes populations from the the purified DCs and monocytes. Similar to the previous studies, authors first utilize fluorescence-activated cell sorting (FACS) to isolate the targeted populations based on carefully selected gene markers, followed by deep scRNA-seq. A unsupervised analysis is then carried out to group cells into distinct clusters in which gene expression profile is obtained in each cluster. The classification of the clusters is further refined by the close examination of variable genes among the clusters. With this approach, in addition to provide a more refined cell classification, the team is able to discover a rare cell type, AXL+SIGLEC6+ cells(AS DCs),in the fifth DC cluster in which its gene expression processes a spectrum of variation. However, as the authors point out, some subtypes of cells can be still be missed by various reasons, such as having only non-RNA molecule identifiers or only presented in certain physiological states. A more sophisticated approach is necessary to address those shortcomings.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications 8* (2017), 14049

In this study, Zheng *et.al.* demonstrates an effective high-throughput single-cell RNA sequencing (scRNA-seq) method in which its rapid cell encapsulation and high cell capture rate enable tens of thousands of single cell within minutes by using GemCode technology platform. Such method, particularly the analytical part , provides an useful framework to conduct scRNA-seq analysis on cell sorting.

Three major scRNA-seq analysis are carried out to show the reproducibility, sensitivity and versatility of the sequencing platform - sequencing performance of synthetic and cell lines RNAs, cell population profiling of 68K Peripheral blood mononuclear cell (PBMC) samples from a healthy donor (Donor A), scRNA-seq analysis of transplant bone marrow samples. In profiling heterogeneous population of PBMC samples, Zheng's team utilizes principal component analysis(PCA) on the top 1000 variable genes of those samples followed by k-mean clustering of the first 50 principle component to identify specific cell clusters.Each of the clusters are then classified based on the expression profile of gene markers for specific cell population. T-distributed stochastic neighbor embedding (t-SNE) is then used for visualization. By applying the prescribed method, they are able to classify the subtypes of PBMCs at expected ratios in which sub-population within certain cluster can also be identified.However, the detection of mutiplets from highly similar cell types can be difficult to detect.For further classification of 68K PBMCs,they apply the scRNA-seq reference transcriptome profiles from 10 bead-enriched sub-population of PBMC of Donor A.The reference-base classification is largely consistent with the marker-based approach, but some of the sub-populations were misclassified,possibly due to overlapping function of those population. More sophisticated clustering and classification method may be necessary to achieve higher accuracy. Apart from cell type profiling, they also successfully develop a novel approach to determine the cell origin by using single-nucleotide variants (SNVs) obtained from scRNA-seq.

In brief, this paper shows an effective analysis work flow for scRNA-seq data which offers critical insights on applying single cell sequencing data on cell population identification.