BITS :: Call for Abstracts 2019 - Oral communication

Type	Oral communication
Session	Methods for Single-cell Analysis
Title	Exploiting machine learning techniques for genes signature discovery from single cell data
All Authors	Romano G.(1), Alessandrì L.(2), Calogero A. R.(2), Beccuti M.(1), Pace L.(3), Cordero. F(1)

Affiliation

(1) Department of Computer Science, University of Torino, Torino
(2) Department of Molecular Biotechnology and Health Sciences, University of Turin, Turin, Italy
(3) Italian Institute for Genomic Medicine, Turin, Italy

Motivation

In the last century, single-cell RNA sequencing (scRNA-seq) techinque allow the researchers to analyze in deepest way the transcriptome of indivual cells. scRNA-seq is a powerful experimental technique able to examines the sequence information from single cell with optimized Next Generation Sequencing (NGS) technologies, providing a higher resolution of cellular differences. In detail, scRNA-seq is able to physically isolate the cells and, taking advantage of a barcoding method, to link the sequenced transcripts to the belonging original cells.
Since scRNA-seq technique produce rich datasets, powerful computational pipelines are required in order to be able to analyze them. Nowadays, several pipelines are available to analyze scRNA-seq data and they already include the entire computational workflow (e.g. Seurat[1], rCASC[2], scanpy[3]). The steps involved in the computational analysis usually include: quality control, mapping, quantification, normalization, clustering and identifying Differentially Expressed Genes (DEGs). The core of the of the pipelines is the clustering module since it allows to define the cell type on the basis of the transcriptome similarity. However, each pipeline include different clustering algorithms that, most of the time, are generic and they can be applied to any kind of data. Thus, the results of these pipelines can be different and usually return the whole set of cells clusterized on the basis of the cell type including, for each cluster, the set of DEGs. Thus, expression of the genes DEGs are the foundamental key that allow to establish to which cluster a particular cells belong and the choiche of them need to be as accurate as possible. In this work, we exploit machine learning technique to validate the DEGs and improve numerically the set of genes that represent the final signature of each cell type.

Methods

A dataset of 5000 immunological cells analyzed with scRNA-seq was used for the analysis [4]. The dataset was composed of purified Naïve T cells and Pentamer T cells that were clusterized in four cellular types (Naïve, Effectors, Memory, Cycling cells) by the authors using CellRanger 2.0 and then Seurat workflow. For reproducibility purpose, the dataset was initially reanalyzed using the same workflow supplied by the authors. The steps of the analysis included the use of CellRanger 2.0. (1) to demultiplex raw base call (BCL) files generated by NGS sequencers into FASTQ files and (2) to obtain the final count matrix which contains cells as columns, genes as rows and in each cell the expression value of the gene in that cell. The count matrix was used as input for Seurat [1] and rCASC [2] pipeline using default parameters. The results of the two pipelines were analyzed expoliting a machine learnings approach that take advantage of a decision tree algorithm (J48) implemented in Weka tool (University of Waikato) [5] in order to capture the signature genes that characterize each cell cluster. 600 runs of Weka were performed for each cluster and the resulting decision trees were composed of nodes (genes) and archs (expression values). In order to capture the best set of signature genes that characterize each cell cluster, we calculated an entropy measure for each level of the decision tree. Entropy is the measure of chaos of a system, thus we expected a decrease of the entropy values gradually from the root (i.e. top of the tree; max value:0.5) to the leafs (i.e. bottom of the tree; min value: 0). This meaure allow us to set a threshold for all the clusters and use it to cut the tree to the level defined by the entropy threshold, capturing the whole set of genes contained. Then, the set of genes signatures obtained for each cluster were visualized using the heatmap method.

Results

Four type of pre-analysis were conducted using the dataset of Pace et al. [4] in order to assess the method and establish putative differences among the set of genes signatures obtained for each cluster. The first analysis was performed with standard biological knowledge using Seurat pipeline [1]. Then DEGs were selected and used as input for the machine learnings approach.
The second analysis was performed with integration of biological knowledge derived from GSEA using

Seurat pipeline [1]. In detail, a list of immunological genes derived from GSEA selection was used as input. The third analysis was performed with standard biological knowledge using rCASC pipeline [2] and DEGs for a comparison with the first analysis. The fourth analysis was performed using Seurat[1] in the machine learning best contest [1].Thus, all the results were analyzed exploiting the machine learning technique explained in the Method section. The four analysis revealed a more consistent set of genes with respect to only DEGs set that characterize the immunological signature of the four cells type identified by the authors.

Info

References
[1] Butler et al., Nature Biotechnology, 2018
[2] Alessandrì et al., GigaScience, under review
[3] Guo et al., PLOS Comput. Biol., 2015
[4] Pace et al., Science, 2018
[5] https://www.cs.waikato.ac.nz

Figure

-

Availability	-

Corresponding Author

Name, Surname	Greta, Romano
Email	grromano@unito.it

Submitted on	29.04.2019