SC1: A Web-based Single Cell RNA-Seq Analysis Pipeline

Marmar Moussa and Ion I. Mandoiu, CSE Department, University of Connecticut

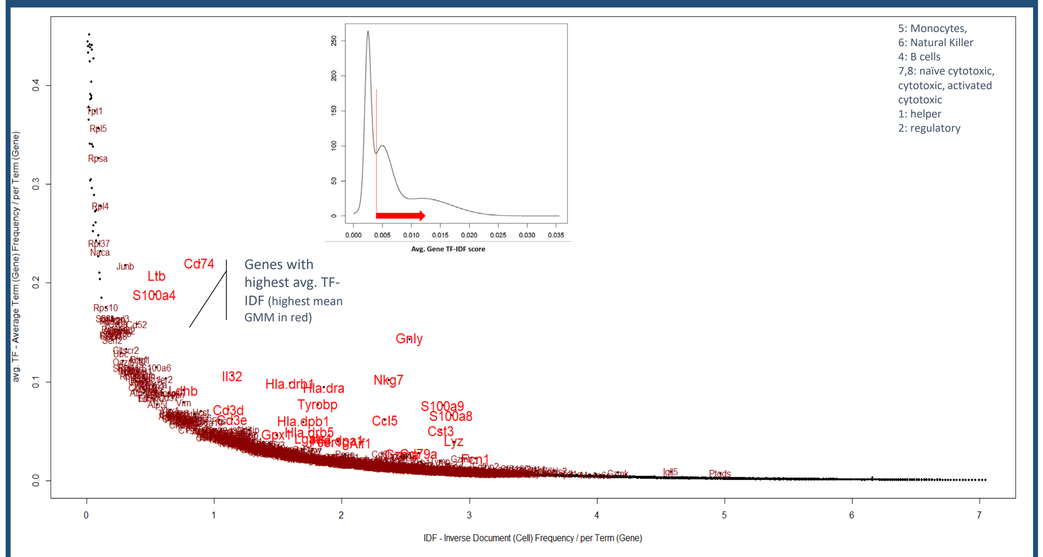
Abstract

There are currently only a few software packages, mostly in R, available for single cell RNA-Seq (scRNA-seq) data analysis. Most of them require considerable programming knowledge and are not easy to use by biologists. Here, we present a web-based interactive scRNA-seq analysis pipeline publicly accessible at **https://sc1.engr.uconn.edu**.

The SC1 workflow is implemented in the R programming language, with an interactive web-based front-end built using the Shiny framework. To facilitate interactive data exploration, time consuming computations such as computing PCA and t-SNE projections of the data following basic QC are performed in a preprocessing step. Preprocessed data is saved in the tool's .scDat file format that can be uploaded for interactive analysis. The SC1 tool incorporates commonly used quality control (QC) options, including filtering cells based on number of detected genes, fraction of reads mapping to mitochondrial genes, ribosomal protein genes, or synthetic spike-ins. The analysis workflow also employs a novel method for gene selection based on Term-Frequency Inverse-Document-Frequency (TF-IDF) scores [1], and provides a broad range of methods for cell clustering, differential expression analysis, gene enrichment, visualization, and cell cycle analysis [3].

Correspondence: {marmar.moussa,ion}@uconn.edu

TF-IDF Based Gene Selection



Term Frequency-Inverse Document Frequency (TF-IDF) for scRNA-Seq data is defined as

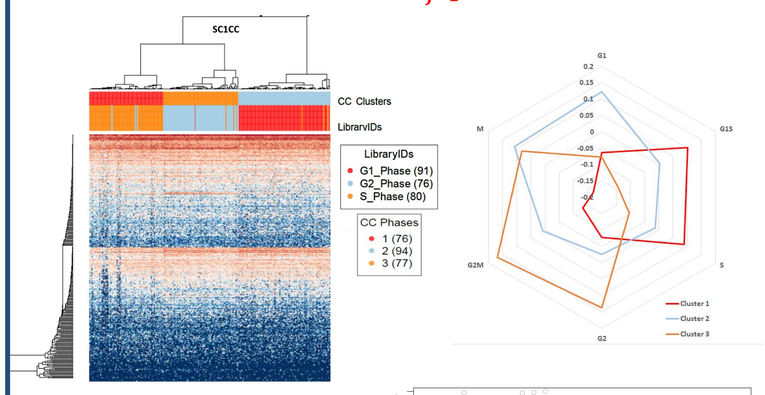
$$TF_{ij} \times IDF_i = (x_{ij} / \max_k x_{kj}) \times (log_2(N/n_i))$$

where x_{ij} is the UMI count of gene i in cell j, n_i is the number of cells expressing gene I, and N is the total number of cells. The above figure shows the TF-IDF based gene selection for the PBMCs from [6]. Genes that are highly expressed (high TF) and expressed in a small subset of cells (high IDF) contribute most to the segregation of the clusters. The genes with highest average TF-IDF score provide pre-clustering heterogeneity insights with as few as 50 top genes.

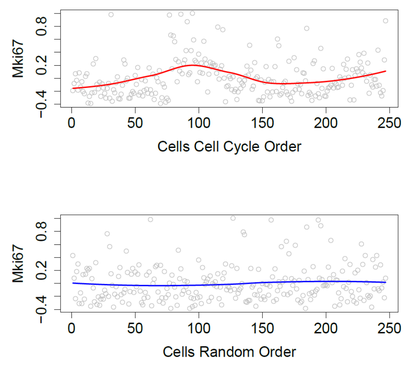
Cell Cycle Analysis

In SC1CC approach, a PCA step is followed by performing a 3-dimensional t-SNE projection using the first few PCs to capture the local similarity of the cells without sacrificing global variation. Next, the cells, represented by their t-SNE coordinates, are clustered into a hierarchical structure (dendrogram) based on their cosine similarity. Finally, to generate an order of cells consistent to their position along the cell cycle, we reorder the leaves of the obtained dendrogram by using the Optimal Leaf Ordering (OLO) algorithm. We also propose a novel metric, referred to as Gene-Smoothness Score (GSS), based on serial correlation to assess the cells' order:

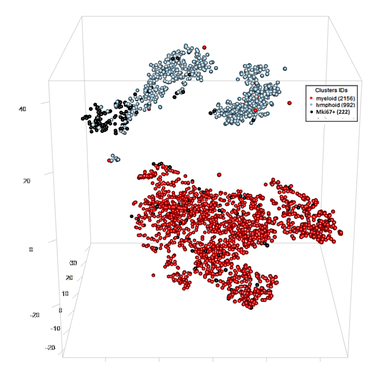
$$GSS(c) = Median\{|SC_{ord}(g_i) - \frac{1}{R}\sum_{j=1}^{R} SC_{rand_j}(g_i)| : i = 1, \ldots, N\}$$



Results for labeled hESC cell line from [7] (200+ single undifferentiated human embryonic stem cells isolated by FACS into 91, 80 and 76 cells in G1, S and G2/M).



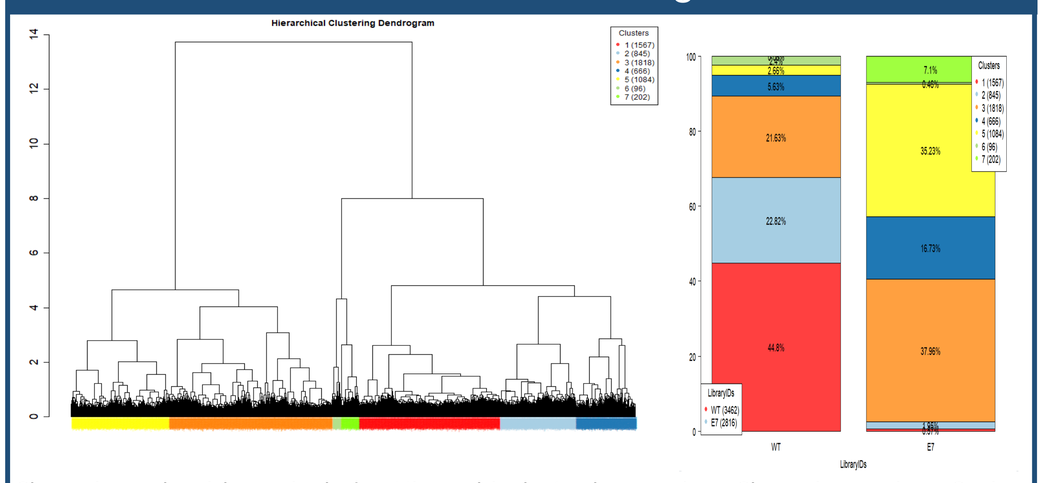
Heat map and cell order and GSS scores ((a) and (b) below) for the clusters inferred by SC1CC on the αCTLA-4 dataset [5] (right). The ~200 cells in cluster 7 are further partitioned by SC1CC into 3 sub-clusters (c), all of which are marked as actively dividing based on GSS scores (d).



References

1. Moussa, M., and Mandoiu, I.: Single Cell RNA-seq Data Clustering using TF-IDF based Methods. (BMC Genomics, 2018).
2. Moussa, M., Mandoiu, I.: LSImpute: Locality sensitive imputation for single-cell RNA-seq data. (Journal of Computational Biology, 2018).
3. Moussa, M. Computational cell cycle analysis of single cell RNA-seq data. 2018 IEEE 8th ICCABS. (2018).
4. Lukowski, S.W et al.: Detection of hpv e7 transcription at single-cell resolution in epidermis. Journal of Investigative Dermatology 138, (2018)
5. Gubin, M.M et al.: High-dimensional analysis delineates myeloid and lymphoid compartment remodeling during successful immune-checkpoint cancer therapy. Cell 175(4), 1014{1030 (2018)
6. Zheng, G.X et al.: Massively parallel digital transcriptional profiling of single cells. bioRxiv p. 065912 (2016)
7. Leng, N., Chu et al.: Oscope identifies oscillatory genes in unsynchronized single-cell rna-seq experiments. Nature methods 12(10), 947 (2015)
8. Li, C.L et al.: Somatosensory neuron types identified by high-coverage single-cell rna-sequencing and functional heterogeneity. Cell research 26(1),83 (2016)

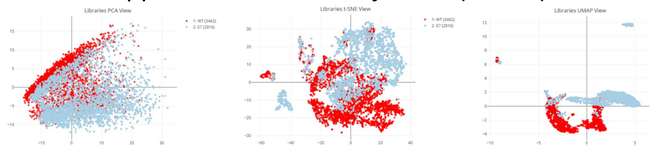
Cell Clustering



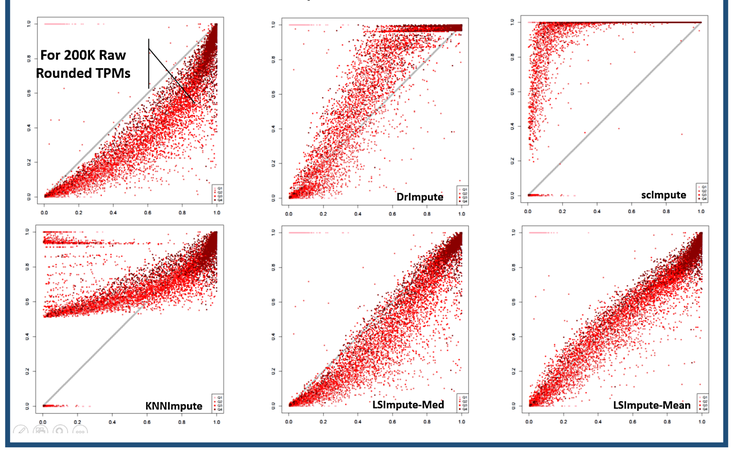
Clustering algorithms include Hierarchical Agglomerative Clustering using Cosine distance applied to log2(x + 1) expression levels of most informative genes selected using top average TF-IDF scores. SC1 also provides Spherical K-means clustering using the top average TF-IDF genes as features and graph-based clustering using binarized TF-IDF data as described in [1]. The number of clusters can be specified by the user or automatically selected using gap statistics. In the above plot, Ward's Hierarchical Agglomerative Clustering using the top average TF-IDF genes was applied to the HPV data set from [4].

Preprocessing and Imputation

Dimensional reduction techniques include principal component analysis (PCA), a faster PCA version using randomized singular value decomposition, *t*-distributed Stochastic Neighborhood Embedding (t-SNE); and Uniform Manifold Approximation and Projection (UMAP).



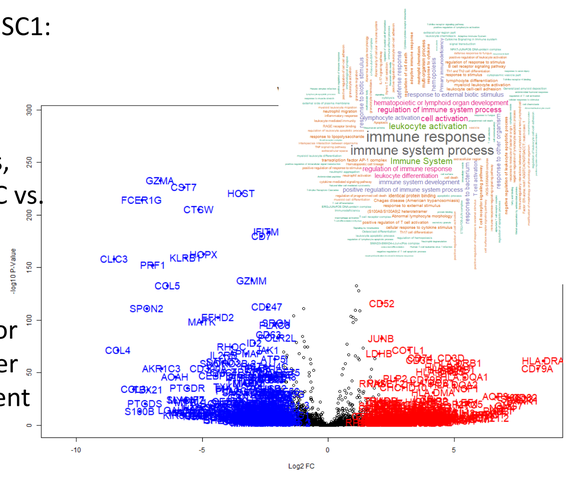
SC1 supports optional imputation using Locality Sensitive Imputation (LSImpute) [2], a novel imputation algorithm that iteratively selects cells with highest similarity level using locality sensitive hashing. The figures below show the Gene Detection Fraction (GDF) of an ultra-deep scRNA-Seq data of 209 somatosensory neurons from the mouse dorsal root ganglion [8] (≈31.5M mapped reads ;≈10,950 +/-1,218 genes per cell) vs. a down-sampled version simulating drop-out effect at 200 K reads per cell.



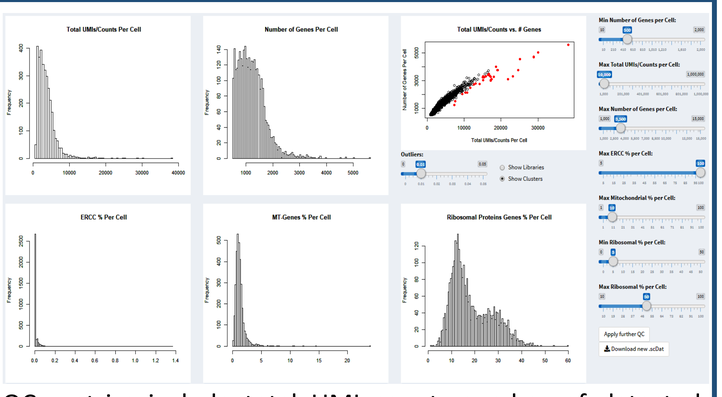
Differential Expression

Differential expression (DE) analysis in SC1:

- One cluster vs. the rest (t-test using Welch approximation with 0.95 confidence interval)
- Selected clusters/groups DE analysis, volcano plots visualization of Log2FC vs. p-values for DE genes
- Genes enrichment analysis/cluster annotation: functional enrichment analysis performed using gProfiler for the differentially expressed genes per cluster, terms with highest enrichment scores visualized as word clouds

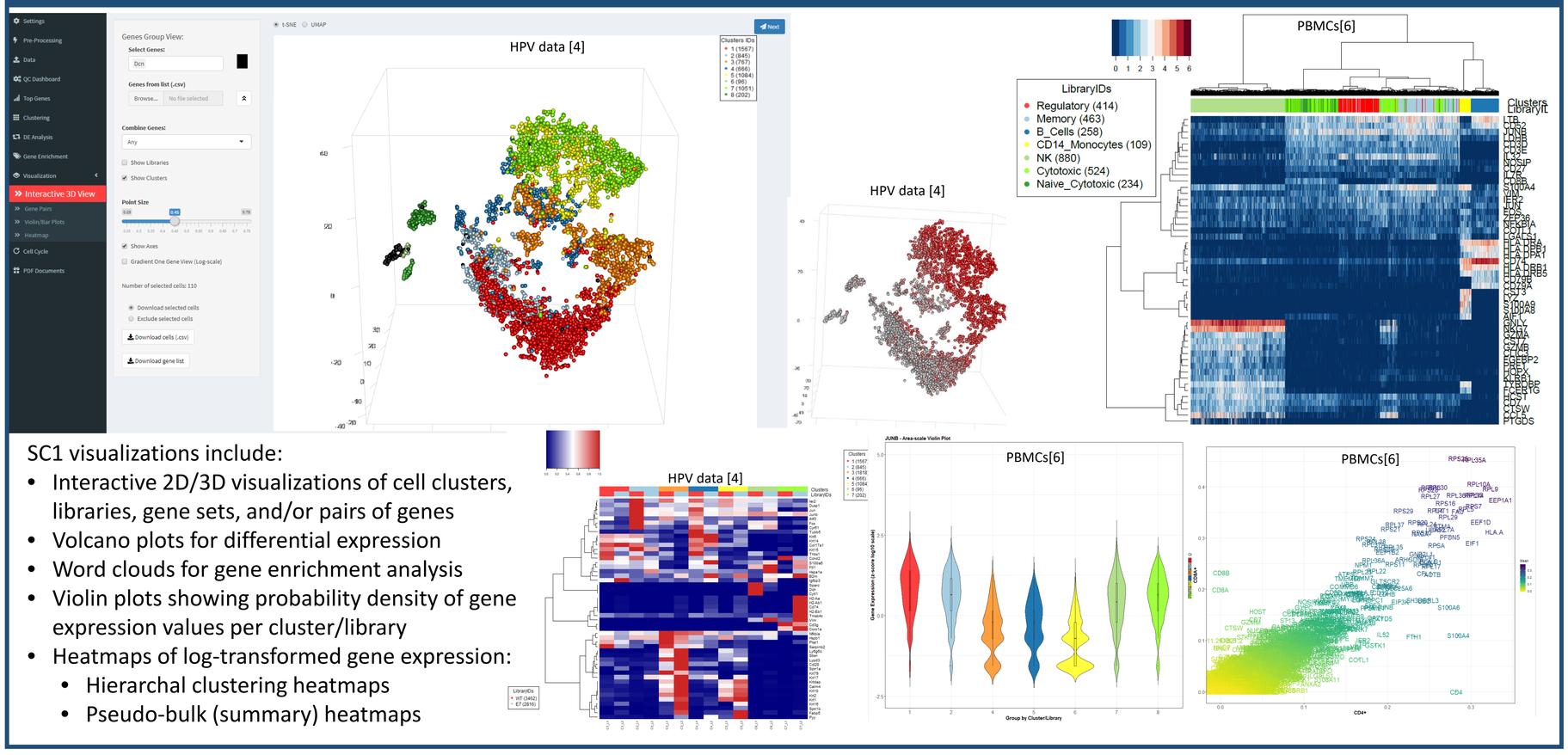


Quality Control Dashboard



QC metrics include total UMI count, number of detected genes, ratio between the number of detected genes and total UMI count per cell, fraction of reads mapping to mitochondrial genes, and number of ribosomal protein genes detected.

Interactive Visualization



SC1 visualizations include:

- Interactive 2D/3D visualizations of cell clusters, libraries, gene sets, and/or pairs of genes
- Volcano plots for differential expression
- Word clouds for gene enrichment analysis
- Violin plots showing probability density of gene expression values per cluster/library
- Heatmaps of log-transformed gene expression:
  - Hierarchal clustering heatmaps
  - Pseudo-bulk (summary) heatmaps