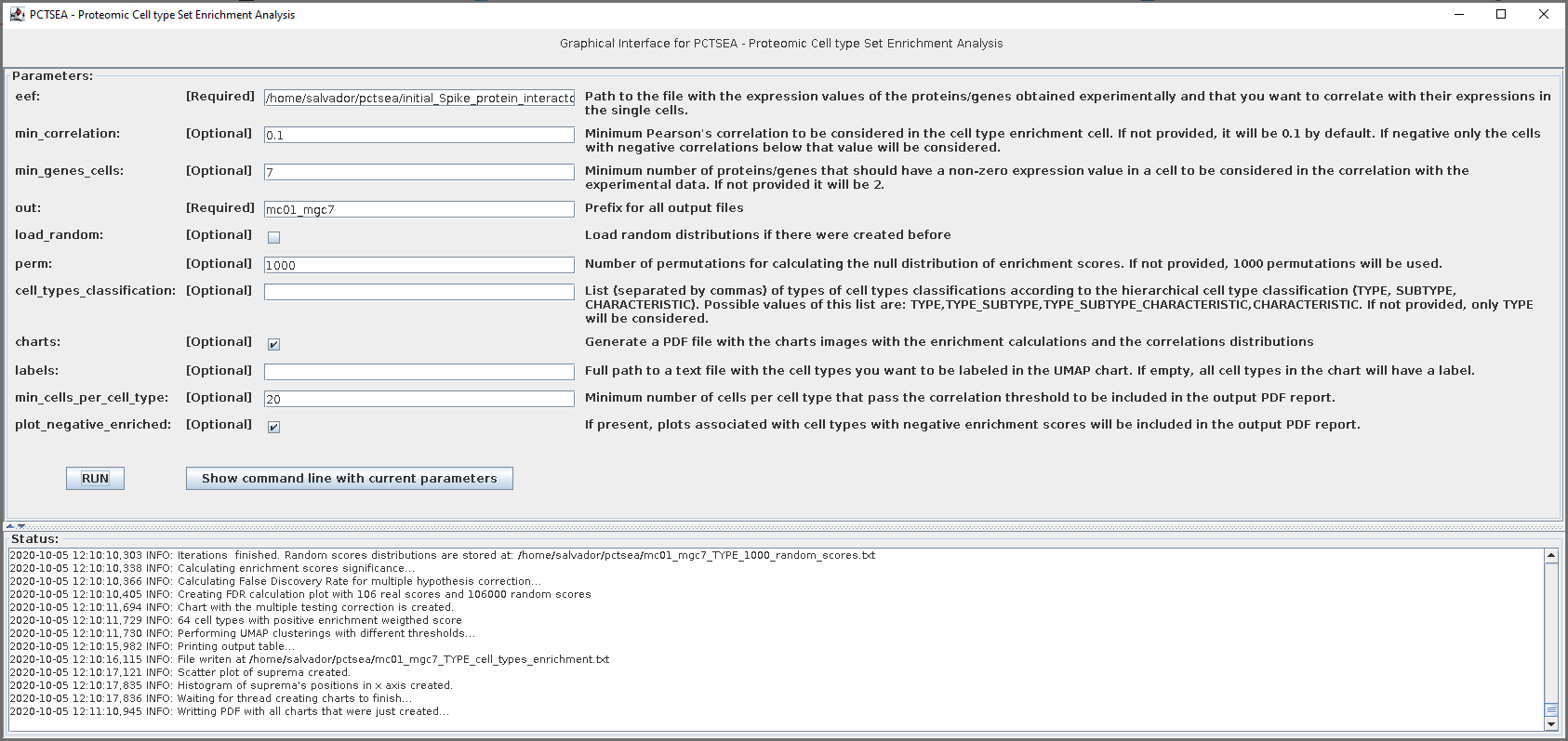
PCTSEA summary

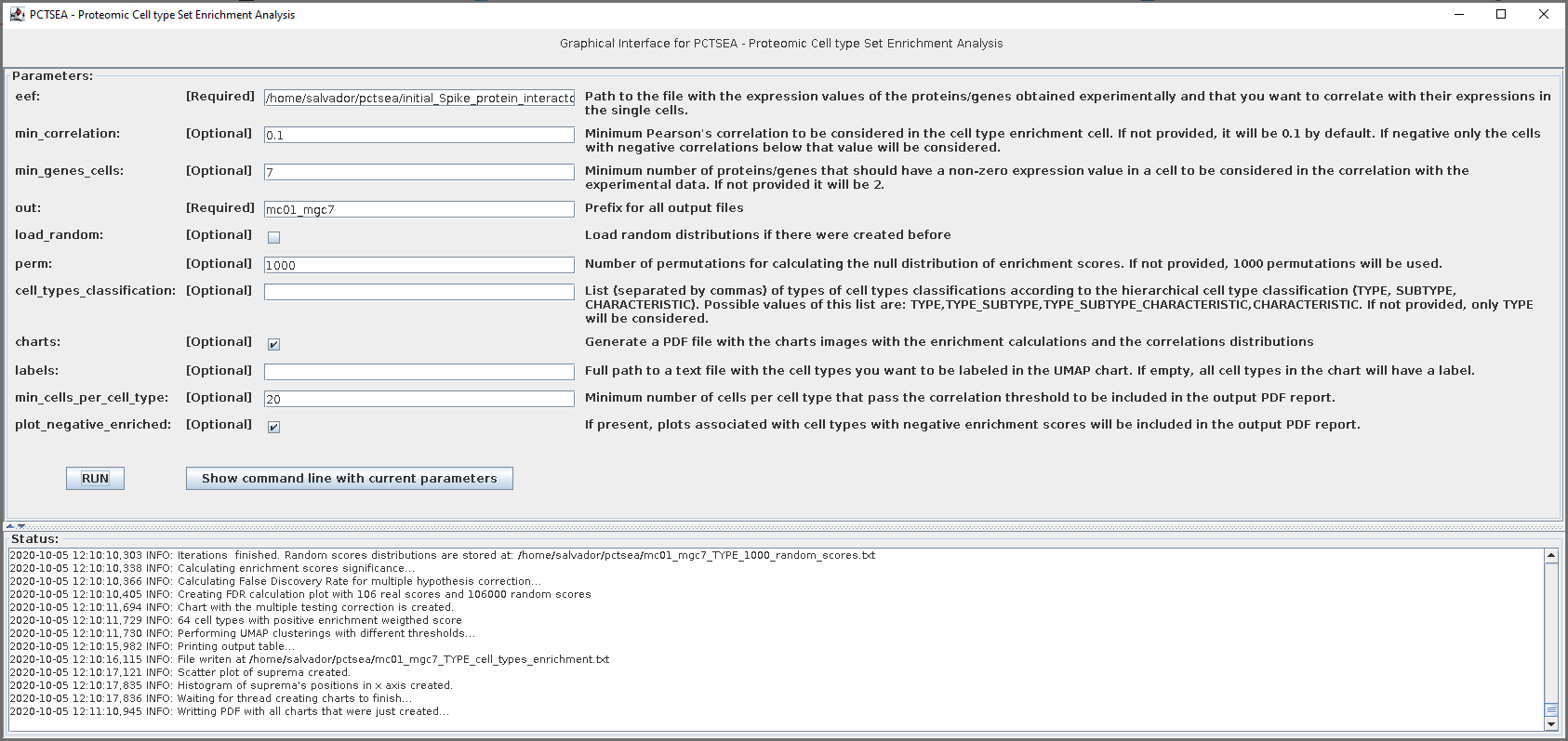
Availability

<https://github.com/proteomicsyates/pCtSEA>

Also, updates are deployed in sealion and it is accessible by typing ‘pctsea’. You will need a X11 server (for example, Xming in Windows) to be able to run it.

Input Parameters:





Datasets:

As for Oct 06 2020, there is only one dataset that can be used to analyze the data and it is the one coming from the Human Cell Landscape <https://www.nature.com/articles/s41586-020-2157-4> publication. It contains 645,246 single cell profile expressions of 245 different cell types.

*Technical note: the database in which the data is stored is located at sealion, and the connection string is mongodb://localhost:8888/?readPreference=primary&appname=MongoDB%20Compass%20Community&ssl=false. A VPN connection + SSH tunnel redirecting port 8888 to 27017 is needed.*

Analysis:

## Input:

Input file should be a two-column, tab-separated, text file, in which the first column is a UniprotKB accession or a gene name, and the second column is an abundant value associated with it.

We strongly recommend to use UniprotKB since they will be translated to gene names and their synonyms in order to be mapped to the datasets of expressions stored in the database.

## Analysis:

* First, all single cells profiles that contain at least ‘min\_genes\_cells’ genes present in the input list and with an expression value different than zero are selected.
* Then, we perform a Pearson’s correlation between the values in common from all the selected single cells, that is, between the input abundances and the single cell expression values. The single cells with a correlation value of at least ‘min\_correlation’ are kept.
* Then, grouping the selected cells for their cell type, we calculate some statistics per each cell type:
  + Hypergeometric p-value where N is the population size, K is the number of successes in the population, n is the sample size and k is the successes in the sample:
    - N = number of cells in total
    - K = number of cells that have a positive correlation with the interactors
    - n = number of cells of type X
    - k = number of cells that have a positive correlation with the interactors and that are of type X
  + Enrichment score: having the cell list that passes the thresholds described above, and sorting the cells by the correlation value in decreasing order, we calculate the enrichment score as the suprema value between the cumulative distributions of the normalized correlations of the cell of type X and cells of any other type. Some enrichment scores will be negative, and therefore, for us don’t make any biological sense, and therefore will not have any additional statistics associated.
  + Then, we calculate the significance of the enrichment score (empirical p-value) of each cell type by performing ‘perm’ iterations in which we calculate again the enrichment scores for all cell types after applying a permutation on the cell types labels. With this procedure we obtain a distribution of null enrichment scores per each cell type. Then, we calculate the significance of the enrichment score by ranking the real score of each cell type together with the corresponding null distribution. The significance probability will be the proportion of null enrichment scores that are equal or higher than the real score over the positive null enrichment score population of that cell type.
  + Then, we correct for multiple hypothesis by calculating a False Discovery Rate associated to the enrichment score. In order to do that:
    - We first account for the different cell type sizes, normalizing each enrichment score and each null enrichment score dividing by the average of their corresponding null enrichment scores of each cell type.
    - Then, we build a total enrichment score distribution with the real enrichment scores of all the cell types, and a total null enrichment score distribution with all the null enrichment scores from all the cell types. Then, for each real enrichment score of each cell type, we calculate the FDR as the number of null enrichment scores that are equal or higher than the enrichment score, divided by the number of real enrichment scores that are equal or higher than the enrichment score that we are evaluating, and we multiply that number by the number of real scores and divide it by the number of null scores.
  + Kolmogorov-Smirnov goodness-of-fit test: We calculate the p-value with the null hypothesis that the cumulative distributions of the normalized correlations are the same.
  + We correct for multiple hypothesis the previous p-values by applying a BH correction.
* We also perform an unsupervised clustering of the cell types, performing a Uniform Manifold Approximation and Projection (UMAP). In this case, for all the genes from the input file that are expressed in the cells of each cell type, we consider the number of cells in which is actually expressed. Therefore, cell types that are clustered together are likely due to the similar profiles of expression of the same genes.