To perform a proteome-based cell type enrichment analysis, single cell transcriptomes are filtered first for the expression of a minimal number of genes that are also present in the proteome which is the host interactome of SARS-CoV-2 or, more general, any proteome or part of a proteome.  Assuming that a partial stochiometric relationship between the relative abundance of mRNA transcripts in cells is preserved in the proteome and reflected in the relative abundance of proteins in the interactome, all protein abundances are correlated with the single cell digital gene expression (dge) values (Pearson’s correlation).

Each cell type that is represented by a minimum number of cells that pass a minimal correlation threshold is further evaluated for a significant enrichment.  For each cell type the p-value of the hypergeometric test is evaluated. An enrichment score is derived from the application of the two-sample Kolmogorov-Smirnov goodness of fit test (K-S test).  Furthermore, the supremum of a cell type enrichment is differentiated from a population of suprema obtained from randomly permutating the cell types labels which yielded an empirical p-value.  Finally, the total distribution of all enrichment scores of all cell types is used to correct for multiple hypothesis and estimate a false discovery rate (FDR) associated to each cell type.

Cell types, that are significantly enriched (K-S test), are clustered based on the relative overlap with the Spike interactome using a uniform manifold approximation and projection.  pCtSEA analysis of proteomes is available at <https://github.com/proteomicsyates/pCtSEA>.