CellSpectra: Query - Reference analysis from a Seurat object

Konstantin A. Klötzer

2024-12-14

Getting started

CellSpectra compares the gene expression pattern of a function or pathway in a query with the same expression pattern in a reference. A p value indicates if a significant deviation from the expected gene expression in observed.

The following example shows the "gold standard" in CellSpectra analysis starting from a Seurat object (developed with Seurat V4). We realized that differences in cell numbers / expression sparsity can have an impact on single-cell analysis results. This can be significantly improved by performing an "onthefly" analysis. Thereby, reference samples match the total counts of each individual query sample.

```
library(dplyr)
library(CellSpectra)
library(Seurat)

#we should also define an output folder for our analysis
output_folder_base = "output/" #don't forget to add the final / !
```

loading data and gene sets

To run the "onthefly" version of CellSpectra we need two things:

1.) A seurat object storing our single cell raw counts of query and reference samples (Developmental version: Seurat V4)

CellSpectra expects specific metadata columns within this object and raw counts accessible as a matrix. We will prepare the object with one of the CellSpectra functions to make sure everything runs smooth.

2.) A gene set database with genes grouped into functionally-related pathways or functions.

CellSpectra expects the gene sets in a binary matrix of genes (rows) and gene sets (columns) with 0 and 1 indicating if a gene belongs to a gene set. You can import this matrix yourself or you use one of the CellSpectra functions to create such a matrix.

```
#here we use the example ifnb dataset from Seurat.data to show basic functions
seurat_object <- readRDS("ifnb.rds")

#now we adapt the metadata columns. CellSpectra needs to find the cell type
#annotation, sample identifier, conditions (query and reference). This can be
#done with the prepare_seurat_object function

seurat_object <- prepare_seurat_object(
  seurat_obj = seurat_object, #our object to prepare
```

```
  celltype_col = "seurat_annotations", #column in which annotation is stored
  sample_id_col = "orig.ident", #sample identifier
  condition_col = "stim", #column storing the condition information
  query_list = c("STIM"), #cat in the condition_col defining query samples
  control_list = c("CTRL") #cat in the condition_col defining the reference
)
```

Our object should now be ready for further analysis and should include all necessary information to proceed.

But first, we need to load or generate our gene set database. The easiest way to do that is the process_gene_sets function and a library .txt file downloaded from the enrichR website. Here we will use GO terms but any library can be downloaded and used.

```
dir.create(output_folder_base)

#prepare database
process_gene_sets(
  filename = "GO_Biological_Process_2023.txt", #downloaded file from enrichR
  seurat_object = seurat_object,
  output_folder_base = output_folder_base,
  min_genes = 10, #we don't want to run the analysis on tiny gene sets
  max_genes = 100 #you might want to set a max cut-off as well
)
```

Note that if you want to use another database you will have to save the gene sets in the right format as a file in the output_folder_base called "go_sets_modified.rds".

Optional: We also provide a function to generate the correct input gene set database from the GO website (process_gene_sets_from_GO). Check documentation. Again, you would have to save the file manually if using the run_spectra function.

creating the onthefly reference

Now we should have everything to create the onthefly reference. This can be quite memory intensive. If you are working with a large dataset, each cell type should be submitted separately. We are working on ways to make this more efficient in the future.

The function creates pseudobulk per sample and cell type. The onthefly version makes sure that total counts of these reference pseudobulk samples is approximately the total counts of the query. So if the query is quite small we wouldn't consider all cells of the reference samples (basically wasting money).

That's where the num_replicates become important. If the query is smaller than many of the reference samples, it makes sense to create a identical number of replicates per reference to use as many cells as possible. This improves the estimation of V1 (the reference expression pattern).

Rule of thumb: If a query sample has only half the cells (or total counts) compared to a reference sample, we will only use half the cells of the reference. So basically, we can make two replicates from the one reference sample. num_replicates = 2. If the size is the same or the query has even more cells, num_replicates should be one.

In the example below we will simply simulate more replicates by setting num_replicates = 10. This doesn't make sense from the statistical view. This is simply to demonstrate the basic functions.

This function will generate subfolders within the output folder for each cell type of interest (specify in cell_types) with subfolders for each query sample. The cell_number_threshold makes sure that only samples with a sufficient number of cells will be analyzed. We recommend at least 10 cells per sample. Higher numbers might increase robustness.

```
#create references
create_references(
  seurat_object = seurat_object,
  output_folder_base = output_folder_base,
  num_replicates = 10, #this dataset doesn't include any biological replicates
  cell_types = c("CD14 Mono", "CD4 Naive T", "CD4 Memory T"),
  cell_number_threshold = 10,
  seed = 123
)
```

Note: In a real world dataset CellSpectra benefits from a wide and heterogeneous reference representing the natural (technical and biological) variance across samples. The ifnb dataset is not really well suited for this kind of analysis without any biological replicates. The gene expression of any gene set is almost identical across our 10 sampling replicates. Therefore, the query will be significantly different from this pattern in most gene sets.

Running CellSpectra

If we followed these instructions we should have an output folder now containing everything we need to run spectra. While this can take quite long for large datasets and many cell types and gene sets, splitting the reference generation from the CellSpectra analysis will save resources (creating references is memory intensive, running spectra is not).

The following function we compute p values, fdr corrected p values, and R2 values for each cell type, query sample, and gene set. We report the QC summary for each query and set the expression threshold to -1 to not remove any low expressed genes

```
run_spectra(
  output_folder_base = output_folder_base,
  cell_types = c("CD14 Mono", "CD4 Naive T", "CD4 Memory T"), #parallelize
  CHISQ.MAX = 4, #this value cuts off the contribution of individual genes
  expression_threshold = -1, #this makes sure no genes are genes are filtered
  QC_report = TRUE

)
```

You can now check the csv results and use them for further downstream analysis.

Final Remarks

We hope this helps to run CellSpectra "onthefly" starting from a Seurat object and an enrichR library text file of any database. It's also possible to run CS from any (pseudo) bulk matrix without "onthefly" reference generation. We did that to compute the coordination within conditions (loo_coordination_from_matrix) or to analyse bulk RNA-seq data (run_spectra_from_matrix).

Check our Repositories for more information.