**Discovery and analysis of the C. elegans Neuronal Gene Expression Network – CeNGEN**

The CeNGEN Rshiny app provides multiple routes to explore and analyze single-cell RNA sequence expression data from the L4 *C. elegans* hermaphrodite nervous system. This README describes all the functionalities of the app, which are organized by tabs.

**1. Gene expression by cell type.**

On this tab you can obtain normalized gene expression values (in transcripts per million, TPM, see Help page for the definition of TPM) for a neuron of interest or for a gene or gene set of interest. These data are averaged expression values for a given gene (or genes) across all the cells corresponding to an individual neuron type. You may select data from one of four expression thresholds (1-4, in increasing stringency), or the unfiltered data (“Unfiltered”, see Taylor, et al., 2020 for thresholding method). We recommend the default threshold 2 as a good balance between false positives and false negatives. Users most interested in capturing as much as possible expression should use threshold 1. Users most concerned with eliminating false positive expression should use threshold 4. Gene expression values below the threshold have been converted to 0.

This tab offers three possibilities from left to right:

* Cell type: Choose a cell type and retrieve all genes expressed in that cell type under the chosen threshold, sorted by expression. The resulting table includes all genes, including those not detected and those below the threshold (set to 0).
* Gene: Choose a gene and retrieve all cell types expressing that gene under the chosen threshold, sorted by expression. Genes can be entered as WormBase Gene ID (WBGene00006744), sequence name (F26C11.2) or public name (unc-4). The resulting tables list only the cell types expressing the gene.
* Multiple genes: Retrieve the expression matrix with the chosen threshold for the subset of genes introduced in the box and separated by “,”.

**2. Find markers based on percentage of expression.**

This tab allows you to query which genes are expressed in a given percentage of the single cells within given cell types. You provide groups of cell types separated by “,” in either of the two boxes: group 1 or group 2. The first box retrieves genes expressed in a greater percentage of single cells than the provided minimum value (e.g., genes expressed in > 65% of cells in the AWC\_ON and AWC\_OFF clusters). The second box retrieves genes expressed in a smaller percentage of single cells than the provided maximum value (e.g., genes expressed in < 2% of the cells in the SMD and SIB clusters). The two boxes can be combined to retrieve genes over a threshold in one group of cells and below a different threshold in another group of cells. Alternatively, by inputting the same group of cells in both boxes, you can retrieve genes expressed in a defined percentage of the cells (e.g., genes expressed in > 50% (box 1) and < 75% (box 2) of the cells in a given neuron-type cluster). The result of each independent query and the combined query are displayed in columns 1-2 and 3, respectively.

Because this tab considers each individual cell’s expression, it uses unthresholded data. To calculate percentages of cells expressing each gene we consider a gene is expressed in a given cell if there is at least 1 UMI.

**3. Enriched Genes by Cell Type**

This tab allows you to retrieve the list of genes enriched in a given cell type compared to all other cells. You can choose between two datasets: “All cell types” and “Neurons only”. The first dataset includes non-neuronal cell types, affecting the background of cells against which the enriched genes of each cell have been calculated.

The “marker genes” in this tab are pre-calculated using the function “FindMarkers” implemented in Seurat using default parameters on the SCT-normalized data. The following metrics are reported:

* gene\_name: public gene name
* gene: WormBase Gene ID
* p\_val: p-value from Wilcoxon rank sum test
* avg\_logFC: log fold-change of the average expression between the two groups.
* pct.1: The percentage of cells where the gene is detected in the selected cell type of interest
* pct.2: The percentage of cells where the gene is detected in all other cells (either all other cells or all other neurons, depending on the dataset selected).
* p\_val\_adj: Adjusted p-value, based on Bonferroni correction using all genes in the dataset
* cluster: The selected cell type of interest.

**4. Find Differential Expression between Cell Types**

In this tab you can perform differential gene expression analysis between two individual cell types or groups of cell types. You can also calculate enriched genes for a given cell type or group of cell types against the rest of cells in the dataset. Select the cell type(s) in the “Select Group 1” and “Select Group 2” boxes. Clicking on “Calculate DEX” will calculate differential expression using one of a number of tests implemented in the Seurat “FindMarkers” function. We recommend using the default (wilcox) test. Please see the Seurat documentation for details and information about other tests (https://satijalab.org/seurat/v3.1/de\_vignette.html). In the “Select Group 2” box, two additional options are possible. Typing “NEURONS” into the Group 2 box will compare the cells in Group 1 to all other neurons. Typing “ALL” into the Group 2 box will compare Group 1 to all other cells in the dataset (including non-neuronal cells). Be aware that calculation against ALL can take up to 5 minutes to be computed.

The following metrics are returned:

* gene: WormBase Gene ID
* p\_val: p-value from Wilcoxon rank sum test
* avg\_logFC: log fold-change of the average expression between the two groups.
* pct.1: The percentage of cells where the gene is detected in cluster 1
* pct.2: The percentage of cells where the gene is detected in cluster 2
* p\_val\_adj: Adjusted p-value, based on Bonferroni correction using all genes in the dataset
* gene\_name: public gene name
* seqnames: sequence name

Because this tab considers each individual cell’s expression, it uses unthresholded data. To calculate percentages of cells expressing each gene we consider a gene is expressed in a given cell if there is at least 1 UMI.

**5. Single Cell Plot**

This tab displays multiple plots for gene expression and metadata visualization in single cells embedded in two UMAP dimensions. On the left menu the user can choose the entire CeNGEN dataset or the neuron only dataset. The second box allows to choose the plotting of metadata (i.e. Neuron Type, Tissue Type, etc.), single gene expression or the co-expression of two genes.

* Metadata plotting options
  + Detection: Highlights cells in red based on the cell-calling method that detected the cells (options: Emptydrops, default CellRanger method, or both)
  + Experiment: Highlights cells in the UMAP in red based on the single-cell experiment from which they came.
  + Neuron.type: Highlights cells in the UMAP based on the annotated neuron type.
  + Tissue.type (All cell types dataset only): Highlights cells in the UMAP based on the annotated tissue type.
* Individual genes. Typing a gene name (WBGeneID, sequence name, or public name) and clicking the plot button displays three panels.
  + Single cells colored by cell type.
  + Single cells colored by the scaled expression from the Seurat SCT-normalized data.
  + Ridge plot of the density of scaled expression sorted by cell type.

The first panel controls the zoom capabilities. Double click zooms in, double click resets zoom.

* Colocalization. Filling the two boxes with Gene 1 and Gene 2 the user obtains two panels.
  + The top panel shows the expression level of Gene 1 in red, Gene 2 in green and the colocalization in shades of yellow. The co-expression is based on a thresholded value of relative expression. The threshold for this plot can be customized using the blend bar on top of the panel from less stringent 0, to more stringent 1.
  + The bottom panel shows cell types and controls zoom. As in other plots, double click zooms in, double click resets zoom.

**6. Heatmaps of gene expression.**

This tab allows you to plot heatmaps of averaged gene expression per cell type, incorporating both relative gene expression (dot color, see below) and the percent of cells (dot size) expressing the gene within each neuron cluster. This tab uses thresholded data from threshold 2, and therefore only contains data from neuronal cell types. You can input a list of genes separated by commas, spaces, or returns. Genes can be entered as public gene names, WormBase ID or sequence names. A list of genes can also be uploaded from a file (tab separated text files and csv files are supported).

Placing the cursor over a dot will display the corresponding gene name, cell type, relative gene expression (“scaled.expr”, see below), percentage of cells expressing the gene (“prop”), and neuron modality (Sensory, Motor, Interneuron, Pharyngeal).

The scaled TPM expression value is the z-score for the gene in that neuron. The z-score represents how many standard deviations the TPM expression of a given gene in a given neuron is from the mean TPM for that gene. It is calculated by subtracting the mean TPM expression for the gene across all neurons from the TPM expression of the gene in the given neuron, then dividing by the standard deviation of TPM values for that gene. See the Help page for the calculation of TPM.