Web portal tools pipelines

Our scPlantAnnotate web service provides two analysis pipelines: annotate_and_plot, Compare_Datasets.

Users can upload their datasets to the website in different format (h5ad, rds, or 10x cellranger) and select different analysis pipelines.

Annotate_and_Plot

To use the annotate_and_plot pipeline, a user needs to select a previously uploaded dataset and a model that matches the plant species of the dataset.

The pipeline works as follows:

- Preprocessing input dataset (normalization if needed, select highly variable genes that matches the model's gene list)
- Predict each cell's type using the input scPlantAnnotate model
- Apply standard analysis and visualization to the dataset using Scanpy
    - Select highly variables genes
    - Scale the dataset to zero mean and unit variance
    - Apply PCA to reduce dimension to 40
    - Apply t-SNE and plot the graph
    - Compute KNN neighborhood graph
    - Apply UMAP and plot the graph
    - Find marker genes for each cell type group and save to files
    - Make a dot-plot of top 3 genes for each cell type group

The result package contains these files:

prediction.csv	predicted cell types for each cell
stats.csv and stats.pdf	counts for predicted cell types and plot
marker_genes.csv	full list of marker genes for all predicted cell type group
annotate_tsne.pdf, annotate_umap.pdf	t-SNE and UMAP plot of all cells grouped by predicted cell types
top3_genes_dotplot.pdf	dotplot for top 3 marker genes for each cell group
all_marker_genes.xlsx	excel file containing full marker genes for each cell type, each in a separate sheet
top25_marker_genes.xlsx	excel file containing top 25 marker genes for each cell type, each in a separate sheet

top10_marker_genes.xlsx	excel file containing top 10 marker genes for each cell type, each in a separate sheet
top5_marker_genes.xlsx	excel file containing top 5 marker genes for each cell type, each in a separate sheet

Compare_Datasets

To use the Compare_Datasets pipeline, a user needs to run anotate_and_plot pipeline beforehand to generate predictions on their datasets. Once the predictions are available, the user can select two or three datasets and their corresponding predictions to run the pipeline. User also needs to designate condition of each dataset (control, condition1 or condition2).

The pipeline works as follows:

- Load each dataset and its corresponding prediction file
- Preprocessing each input dataset (normalization if needed)
- For each common cell type between control and condition_1 data:
    - Find DEGs (differentially expressed genes) between control data (i.e. control group) and condition_1 data (i.e. condition 1 group) using Scanpy's method rank_genes_group()
        - This method will compute, for each gene, the logfold2change, scores and p_values in each group with respect to the other group
    - Find significant DEGs for control group and condition_1 group, respectively.
        - A gene is a significant DEG if, compared to the other group, its logfold2change value is greater than 1.0 and the p_value is less than 0.05
- If condition_2 data and its prediction file is provided
    - For each common cell type between control and condition_2 data:
        - Find DEGs between control data and condition_2 data
        - Find significant DEGs for control group and condition_2 group
    - For each common cell type between control, condition_1 and condition_2 data:
        - Find the common significant DEGs between control vs condition_1 and control vs condition_2, this gives the list of down-regulated genes and up-regulated genes from both conditions.

The result package contain these files:

compare_celltype_distributions.pdf	plot of cell type distributions in all conditions (control, condition1, and condition2 if provided)
control_vs_condition1	subfolder containing all, top 25, 10, 5 DEGs (differentially expressed genes) in control data and condition 1 data, respectively. Each excel file

	contains multiple sheets, one for each cell type.
control_vs_condition2(if condition2 data and predictions are provided)	subfolder containing all, top 25, 10, 5 DEGs (differentially expressed genes) in control data and condition 2 data, respectively. Each excel file contains multiple sheets, one for each cell type.
control_vs_conditions_common_sig_markers.xlsx	shared genes in significant DEGs in control_vs_condition1 comparison and control_vs_condition2 comparison, which means the list of DOWN-regulated genes in both conditions, grouped by cell type, each one on a separate sheet.
conditions_vs_control_common_sig_markers.txt	shared genes in significant DEGs in condition1_vs_control comparison and condition2_vs_control comparison, which means the list of UP-regulated genes in both conditions, grouped by cell type, each one on a separate sheet.