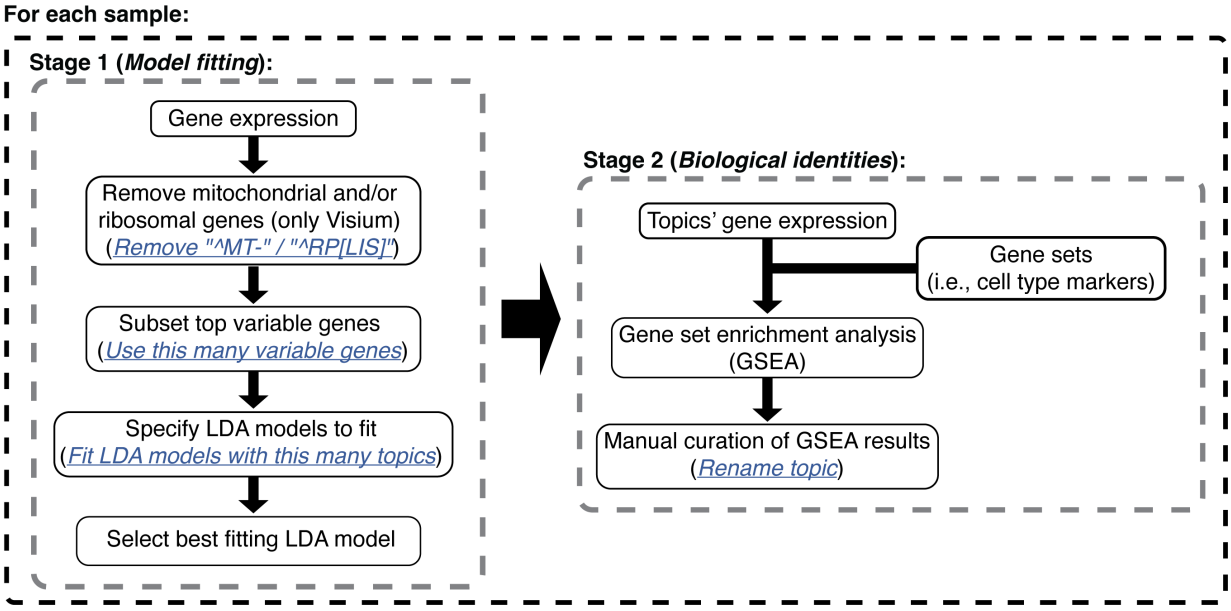
Identifying cell types or functional states is a fundamental part of analyzing single-cell and spatial gene expression experiments. Researchers apply algorithms developed for non-spatial single-cell data. While these algorithms may be appropriate for some spatial transcriptomics (ST) technologies providing data at the single-cell level (e.g., MERFISH, CosMx, Xenium), they may not provide reliable results for multicellular-level technologies (e.g., Visium). Methods to phenotype multicellular-level ST data are currently available, which, rather than assigning a single classification of each spot, provide a percentage of each possible class (e.g., cell type) within each spot.

In the spatialGE **Phenotyping** module, the STdeconvolve (Miller et al. 2022) algorithm has been implemented to assign biologically meaningful classes (e.g., cell types, tissue niche types, functional states) to spots in multicellular-level ST experiments. Briefly, STdeconvolve finds latent topics within the gene expression data by using Latent Dirichlet Allocation (LDA). The topics represent gene expression patterns that may correspond to cell types, tissue niches, or functional states within the tissue. To ascertain the biological identity of the topics, STdeconvolve features built-in gene set enrichment analysis (GSEA) to match combinations of genes highly expressed in the topics with genes belonging to gene sets or pathways. *Currently, spatialGE only supports the use of built-in gene sets. Soon, users will be able to upload custom gene sets.*

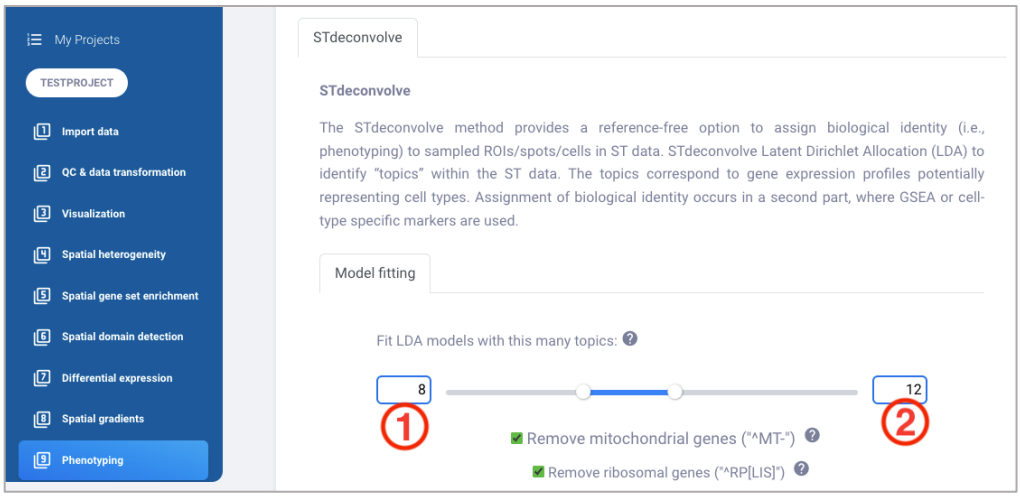
The STdeconvolve algorithm has been implemented in spatialGE as a two-stage procedure. In the first stage, users fit LDA models including a varying number of topics to the gene expression data with a predetermined number of latent topics. For example, a user may choose to test three different models: One with seven, one with eight, and one with nine topics. With the help of certain model metrics, which will be addressed later, the user decides which model fits the data better. In the second stage, GSEA is completed to assign the biological identity (i.e., cell types, tissue niche type, functional state) to each topic in the model with the best fit.



Before starting analysis in the **Phenotyping** module with STdeconvolve, analysis with the **Spatial domain detection** module must be completed. The **Phenotyping** module can be accessed by clicking the button in the sidebar on the left of the window. Next, options to conduct the first stage of the analysis are explained.

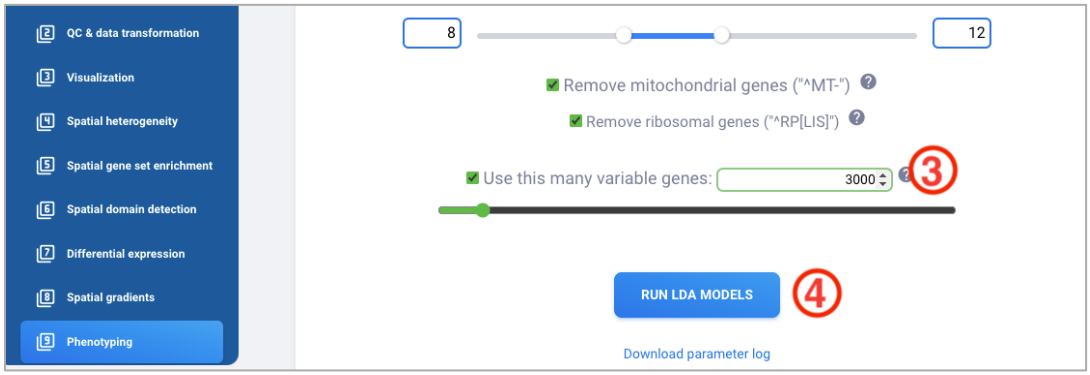
STdeconvolve – Model fitting

The first step in this module is to select the number of LDA models to be fitted and the number of topics within each model. In this implementation, both the number of models and topics per model are controlled by the slider **Fit LDA models with this many topics**. For this example, the slider will be set with a minimum of 8 and a maximum of 12 topics. Consequently, five LDA models will be fitted: One with eight topics, one with 9 topics, one with 10 topics, one with 11 topics, and one with 12 topics. The number of topics defines the possible number of cell types to be detected in the sample. The more models to be tested, the more time completing STdeconvolve will take. Nonetheless, much can be gained by testing a wide range of models because more options are available to select the LDA model that represents the tissue biology of the sample in the best way. In general, ST data is limited in capturing the fine-grained differences among cell subpopulations, and hence, a suggestion is first to attempt to recover the major cell types.

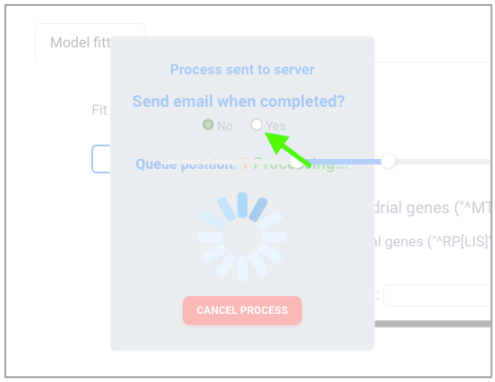


Mitochondrial and ribosomal genes will be removed by keeping checked the **Remove mitochondrial genes ("^MT-")** and **Remove ribosomal genes ("^RP[L|S]")** controls. This filter currently only works in Visium ST data. Often, LDA models define clusters based on the high abundance of mitochondrial and/or ribosomal genes in certain regions of the tissues (especially diseased tissues). Nonetheless, it is up to the user to decide if this is a good practice to keep or remove these genes. *Importantly, this filter only affects STdeconvolve*. If the user did not remove mitochondrial and/or ribosomal genes in the **QC & data transformation** module, these genes will remain in the data set when running other analyses.

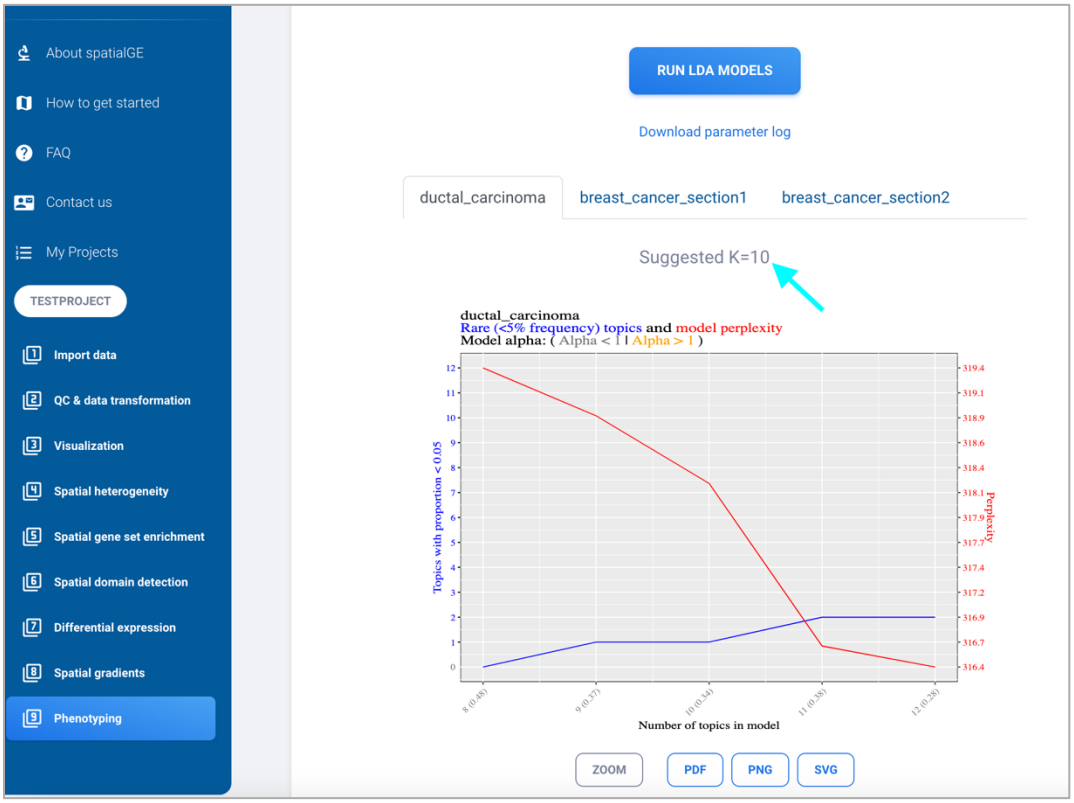
Now, the number of top variable genes must be specified by writing 3000 in the **Use this many variable genes** text box. The more genes are selected, the more accurate fit is obtained. However, execution time also increases. An appropriate number for Visium data is 5000 genes. At this point, all the required parameters to fit the LDA models have been specified. Click the **RUN LDA MODELS** button to initiate the analysis.



Computationally intensive analyses such as LDA model fitting take time. In spatialGE, users can opt to receive an email notification when the analysis is finished. Click the *Yes* radio button (green arrow) below **Send email when completed?** to receive a notification.

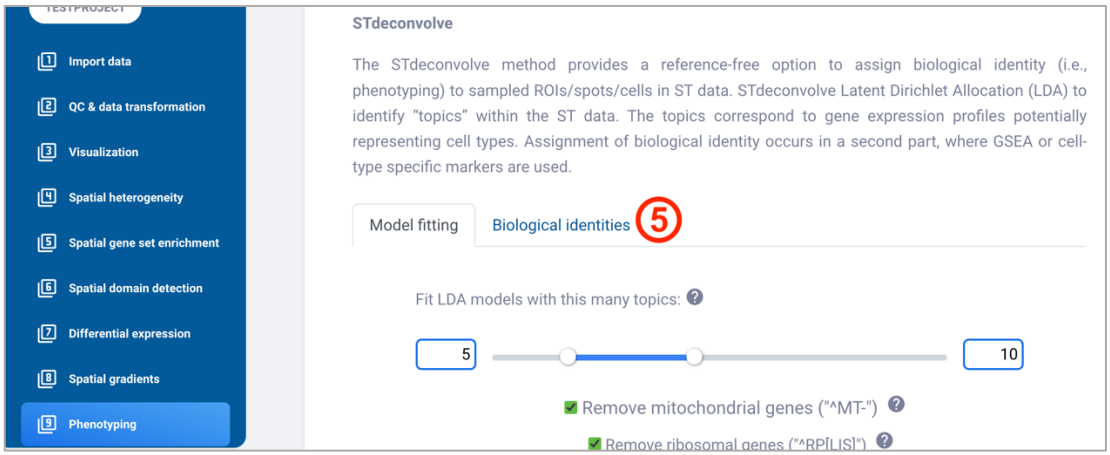


After the model fitting is completed, a series of plots (one per sample) are shown under the **RUN LDA MODELS** button. The plots summarize the fit of the LDA models to the data and are critical to selecting the model to be used during the biological identification of the topics. Three metrics are shown in the plots: perplexity (red line), number of rare topics (blue line), and model alpha (gray numbers in parentheses on the x-axis). In broad terms, perplexity measures how well topics represent the data. The lower the perplexity the better the model's topic captures the data. Often, as perplexity goes down, the number of topics that are rare (i.e., present in a few spots) increases. That the number of rare topics increases is not necessarily detrimental, but they may also suggest overfitting. In addition, many topics that are rare usually contribute little to understanding the overall tissue architecture. For these reasons, a suggestion is to keep the number of rare topics to a minimum. The alpha parameter measures how different each topic is from other topics. A rule of thumb is to keep models with alpha closer to 1. Models with alpha > 1 are undesirable because with alpha >> 1, topics tend to contain almost the same information. In biological terms, this means that topics do not capture the differences among cell types. Based on these metrics, spatialGE makes a recommendation about which is the best model. This recommendation is above the plot for each sample. In this example, "*Suggested K=10*" (turquoise arrow), indicates that the LDA model with 10 topics is the suggested model. If the user considers this is not the best model, options are available in the next part of the analysis to select a different model.



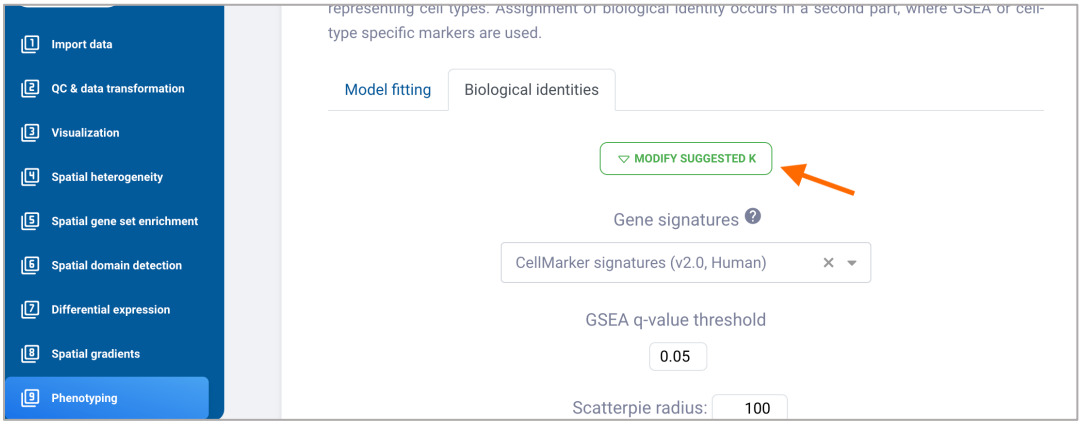
STdeconvolve – Biological identities

After LDA models have been fitted, GSEA can be run on the gene expression of the topics in the best fitting model. Currently, users are only able to test built-in gene sets, but soon functionality will be implemented to allow the input of custom signatures. To begin biological identification of the topics, click on the **Biological identities** tab that have appeared above the parameter controls.

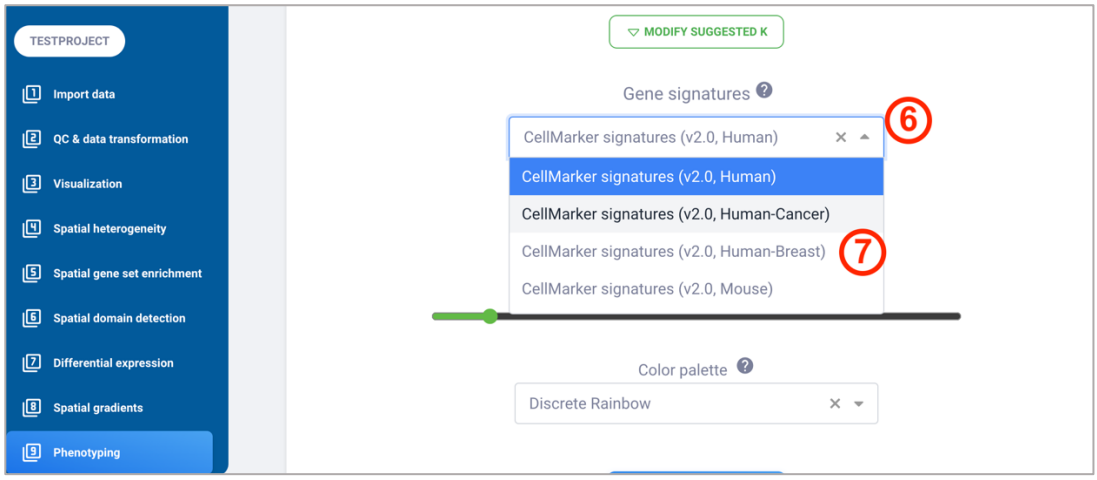


The first element in the **Biological identities** tab is the **MODIFY SUGGESTED K** button (orange arrow). Although no modification will be made to the automatic suggestions, if the user considers that the K suggested in the previous step does not correctly take into account the model fitting parameters (perplexity, number of rare

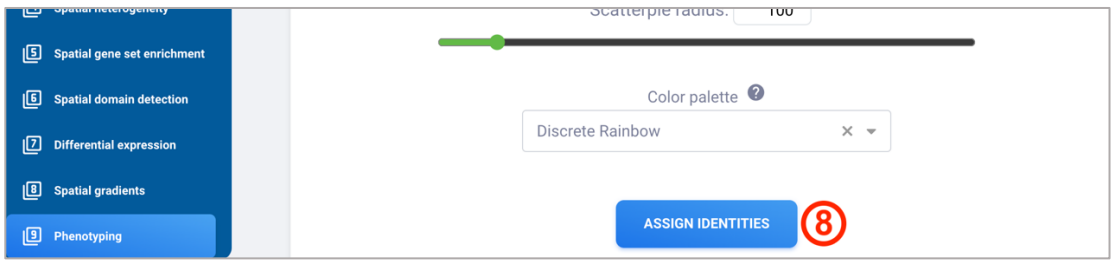
topics, and alpha), or it does not represent the tissue biology (e.g., number of cell types), a different number can be entered for one or several samples in the data set.



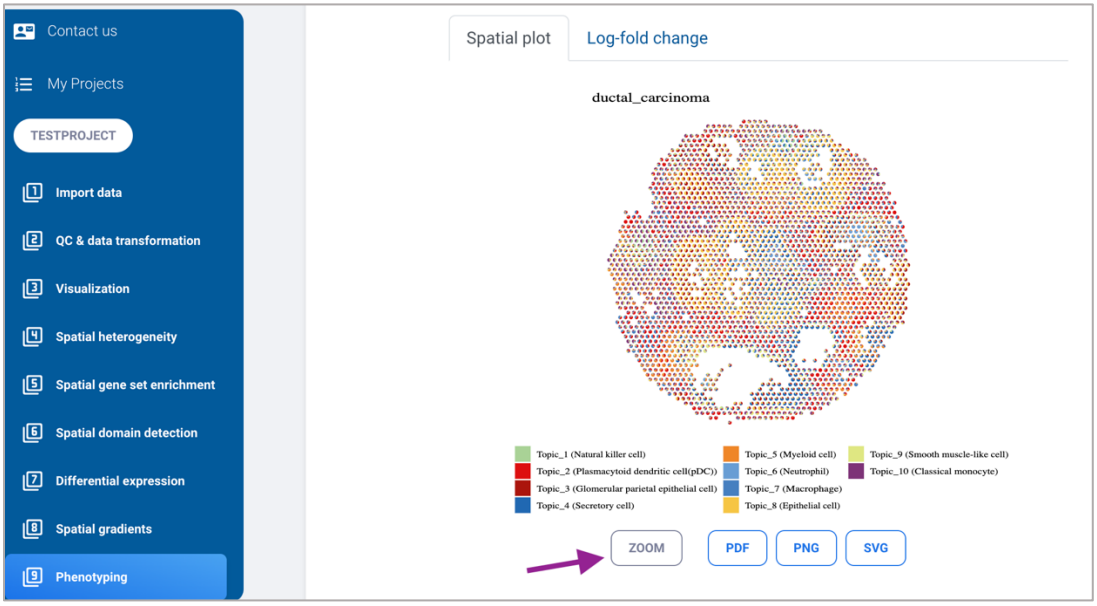
For now, click the **Gene signatures** dropdown and select the *CellMarker signatures (v2.0, Human-Breast)* option. These gene sets have been extracted from the CellMarker 2.0 database (Hu et al. 2023).



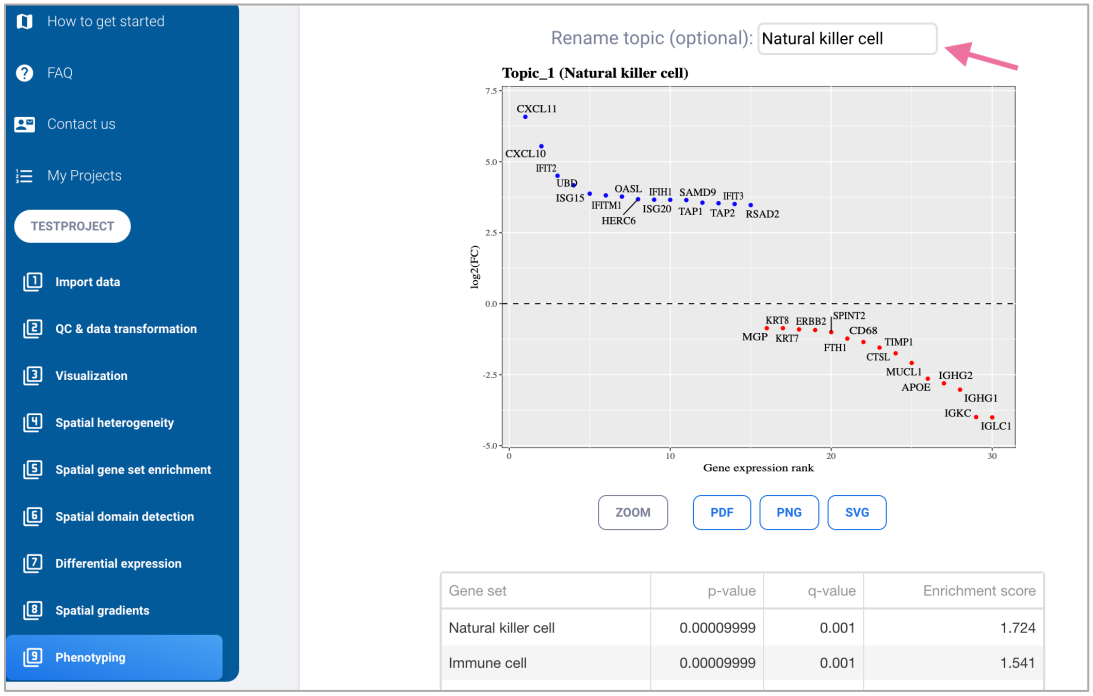
All other parameters will be left in their default values; however a brief explanation of these parameters will be given. The **GSEA q-value threshold** controls the q-value at which a gene set score will be considered as enriched. The GSEA scores with q-values higher than this threshold will not be considered. The **Scatterpie radius** and **Color palette** control the appearance of the resulting plots. To start the analysis, click the **ASSIGN IDENTITIES** button.



After completion, results are presented in tabs, one per sample. Within each tab, two additional tabs are available. The first one, called **Spatial plot**, shows the relative abundance of each LDA topic at each spot in the sample (also called a "scatterpie"). For a better observation of the individual piecharts, click the **ZOOM** button (purple arrow). Make sure to deactivate the zoom feature to be able to scroll the page.



The second tab (**Log-fold change**) contains a series of plots, one per topic, showing the genes with the highest and lowest log-fold change. This is the log-fold change of a gene in a given topic compared to others. Following each plot, a table is presented with the results of the GSEA.



In this section, the name assigned to each topic by GSEA can be modified if the user considers that genes with the highest (or lowest) log-fold change represent a different cell type. This is the manual curation step, which is a frequent practice during cell phenotyping tasks. Changes can be made by entering a different name in the **Rename topic** textbox (pink arrow). A new button will appear (**RENAME TOPICS**). We suggest that all changes are made across all topics and samples, and then the **RENAME TOPICS** button be clicked once.