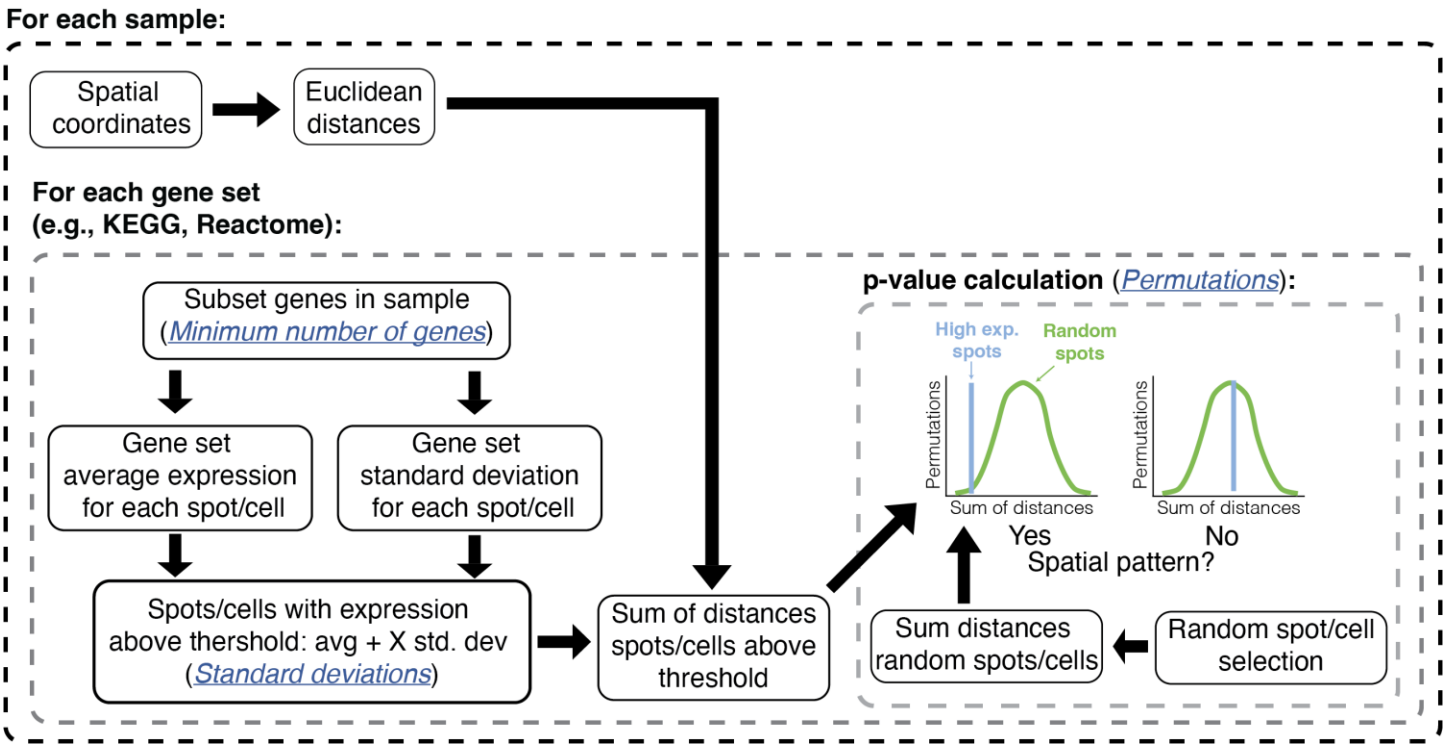
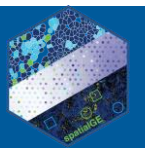
Module: Spatial gene set enrichment

In gene expression analysis, it is common to study changes not only at the gene level, but at the pathway (i.e., gene set) level. This step in the workflow is achieved by conducting a gene set enrichment analysis (GSEA). Researchers analyzing a spatial gene expression experiment might also want to detect gene sets that are "spatially enriched". In spatialGE, a spatially enriched gene set is a gene set for which Regions of Interest (ROIs), spots, or cells with high expression of the gene set are spatially aggregated (i.e., close one to the other). For example, in many tumor tissues, one might expect that certain gene sets pertaining to the cell cycle are highly expressed in tumor areas, resulting in a pattern of spatial aggregation.

The module to achieve this analysis in spatialGE is **Spatial gene set enrichment**. The module runs in the background the STenrich function from the R package spatialGE, which us a modified version of the method proposed by Hunter et al. (2021). The diagram below explains the STenrich algorithm:



In **Spatial gene set enrichment**/STenrich, each sample spatial enrichment is analyzed separately. First, Euclidean distances are calculated among all ROIs/spots/cells. Then, the gene expression of the sample is subset to the within the gene set being tested. If too few genes are left after subset (**Minimum number of genes** parameter), then the gene set is omitted for that sample. The average expression and standard deviation of those genes is calculated for each ROI/spot/cell. Next, ROIs/spots/cells with gene set expression above the average gene set expression across all ROIs/spots/cells are identified. The threshold to define these high gene set expression spots is defined by the average gene set expression plus a number X of standard deviations (**Standard deviations** parameter). The sum of the Euclidean distances between the high expression ROIs/spots/cells is calculated.

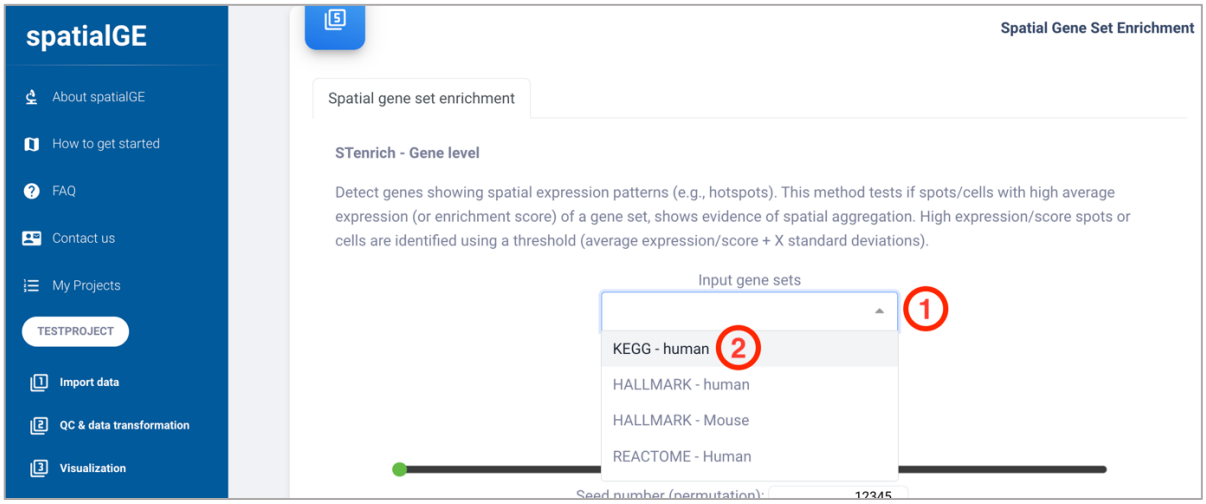


The next step involves a permutation process, in which a null distribution is generated in order to test if the (sum of) distances among high expression spots are smaller than expected. To that end, a random sample of ROIs/spots/cells (regardless of expression) is selected. The random sample has the same size as the number of high expression ROIs/spots/cells. Then, the sum of distances among the randomly sampled ROIs/spots/cells is calculated. The random selection is repeated as many times as requested (**Permutations** parameter). Finally, a p-value is calculated by noting how many times the sum of random distances was higher than the sum of distances among high expression ROIs/spots/cells. If the sum of random distances was most of the times higher than the sum of distances among high expression ROIs/spots/cells, then the null hypothesis of no spatial aggregation is rejected (i.e., ROIs/spots/cells with high gene set expression are more aggregated than expected by chance).

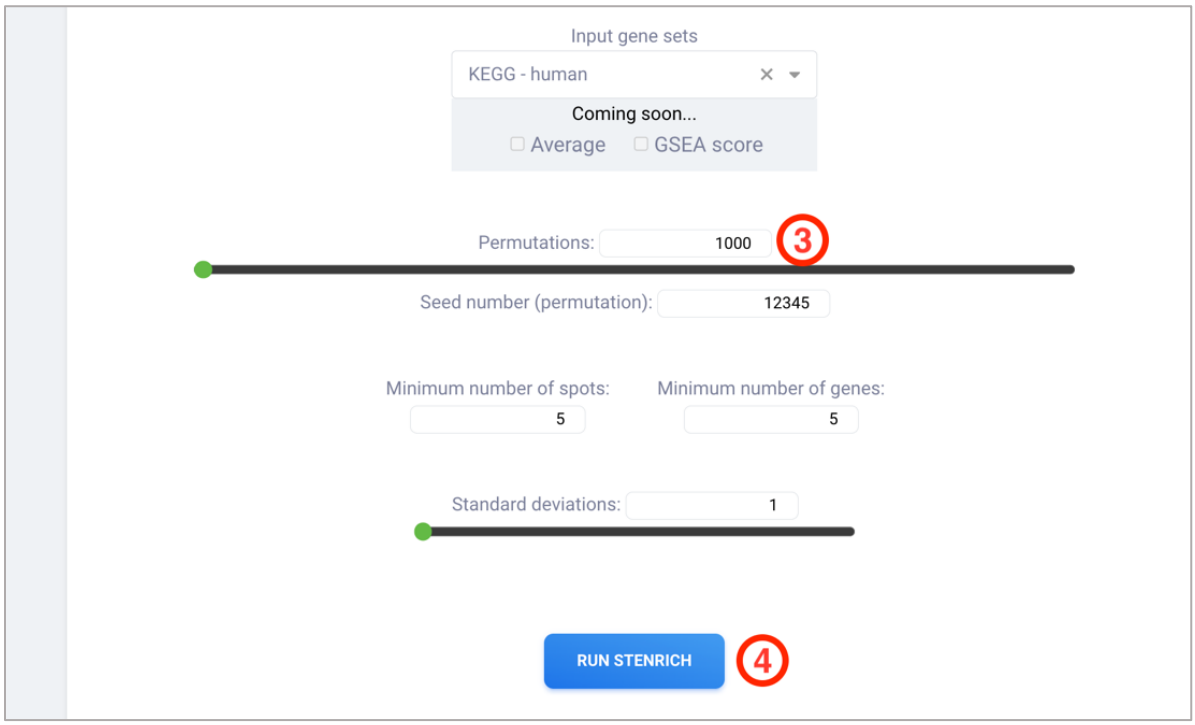
A few important notes about the algorithm to consider when conducting an analysis with the **Spatial gene set enrichment** module:

- Notice that the metric of enrichment is currently the average expression of the genes in a set within each ROI/spot/cell. Better metrics of gene set enrichment are available and will be soon available to use instead of average.
- Special attention should be paid to the reference genome used to annotate gene counts. The annotation of transcripts precedes any analysis in spatialGE. Nevertheless, if transcripts are annotated with a mouse genome (or other species), the user should use the appropriate gene set database, as gene names will likely not match to a gene set database with human gene names. Furthermore, but not less important, there might be problems with gene homology.
- The more **permutations** are requested, the longer the execution, but p-value estimates are more accurate.
- By increasing the number of **standard deviations**, fewer ROIs/spots/cells are selected as "high expression" (stringent analysis), however, it becomes more challenging to detect gene sets that show evidence of spatial aggregation. Values between 1 to 2 standard deviations are likely suitable for most studies.
- By adjusting the **Minimum number of spots**, the user has some rough control over the size of the structures to be studied. For example, small immune infiltrates could be detected with smaller number of minimum spots. Conversely, if tissue domain-level differences are sought, then a larger **Minimum number of spots** should be set.
- Take note of the random seed number used for the test (**Seed number** parameter), in case you need to generate the same results later. Nevertheless, running the analysis several times with different seed numbers is advisable to check for consistency.

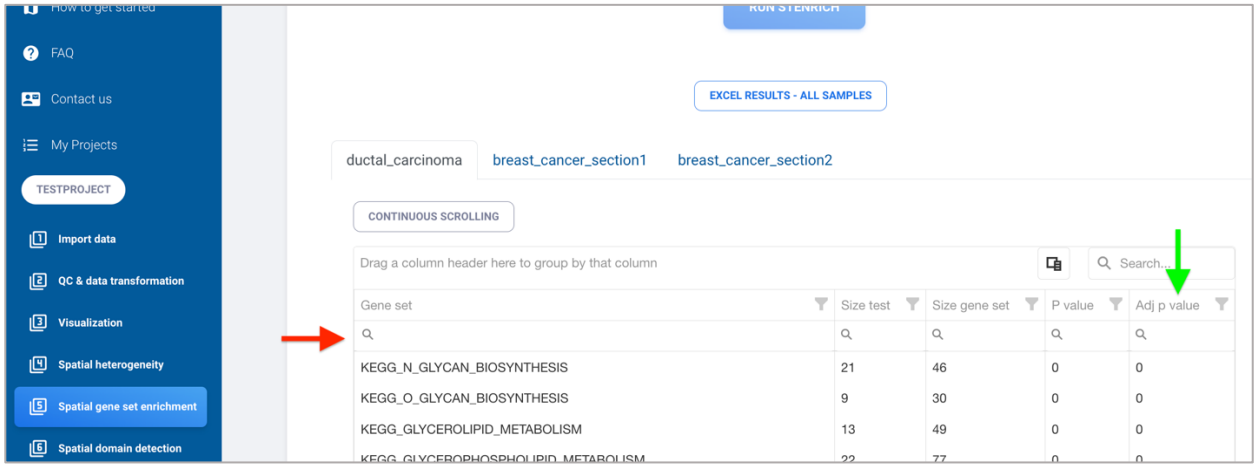
With the algorithm explained, begin an analysis by selecting a gene set in the **Input gene sets** drop-down. For this example, select the "KEGG – human" option (see also note above regarding selection of database and genome used in annotation). Additional functionality will be included to allow for custom gene sets to be entered. Currently, the Average and GSEA score checkboxes are disabled, and only the average gene set expression is calculated.



Set the number of permutations to 1000 by typing in the **Permutations** textbox. Other parameters will be left as they are; however, users are encouraged to try different combinations as changes in the detected gene sets will be observed. Finally, click the **RUN STENRICH** button. The process might take a while. It is recommended to select the radio button for e-mail notification. The radio button appears after clicking the **RUN STENRICH** button.



After completion of the process, the results are displayed in the form of a table. The table has filtering functionality, which enables the user to search results (red arrow) for a specific gene set. It also allows sorting (green arrow), in case decreasing order of p-values is desired, for example.



Note: An "NA" in the nominal and adjusted p-values of the results indicates that the test for that gene set was not conducted due to either not enough genes in a gene set or high expression spots. To enable testing of those failed gene sets, lower the requirements by setting lower values of **Minimum number of genes** and **Minimum number of spots**.