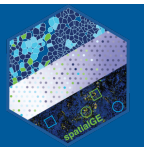


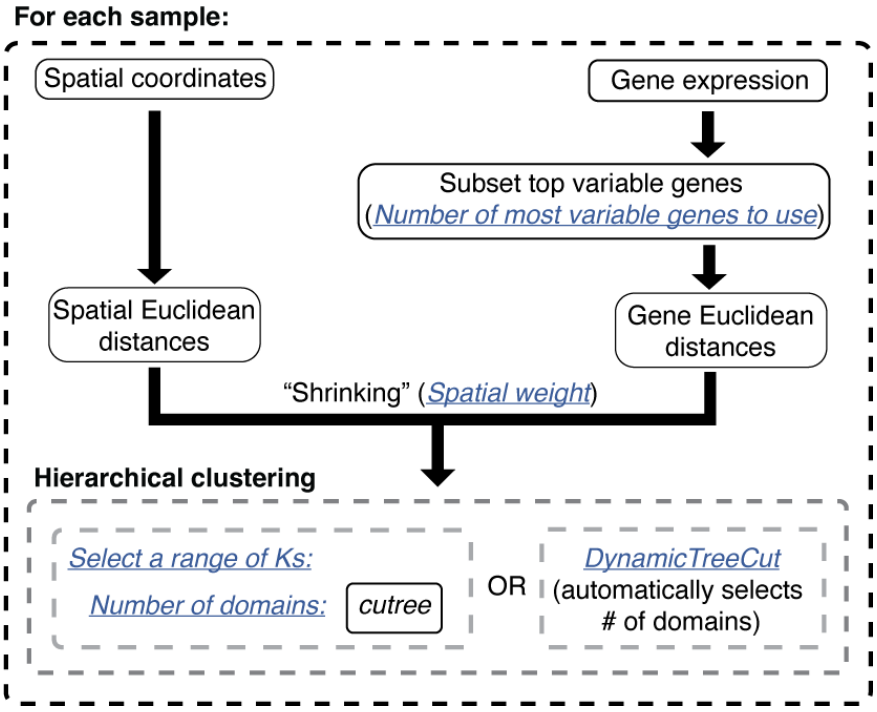
Researchers studying single-cell gene expression commonly use some flavor of clustering algorithm to group cells based on their transcriptomic similarity. When studying spatial transcriptomics (ST) data, this type of analysis acquires an additional purpose: Identify tissue niches or domains. Multiple options are available for clustering of ST data, with some treating Regions of Interest (ROIs), spots, or cells as spatially independent, and others explicitly using the spatial information. In the spatialGE web application, users can apply two clustering/domain detection methods:

1. **STclust:** The built-in method in spatialGE, which uses groups spots/cells based on gene expression weighted by spatial Euclidean distances (*See page 2*).

2. **SpaGCN:** A domain detection method using graph convolutional networks (GCNs) to integrate gene expression and spatial coordinates and conduct clustering of the spots/cells. (*See page 6*).

STclust

The method STclust is implemented in the **Spatial domain detection** module. The use of spatial distances among ROIs/spots/cells to inform clustering tends to yield continuous clusters that can be considered tissue domains. The method applies hierarchical clustering on gene expression that has been weighed ("shrinking") by the spatial information in the form of Euclidean distances (details in Ospina et al. 2022). The diagram below explains the STclust algorithm:



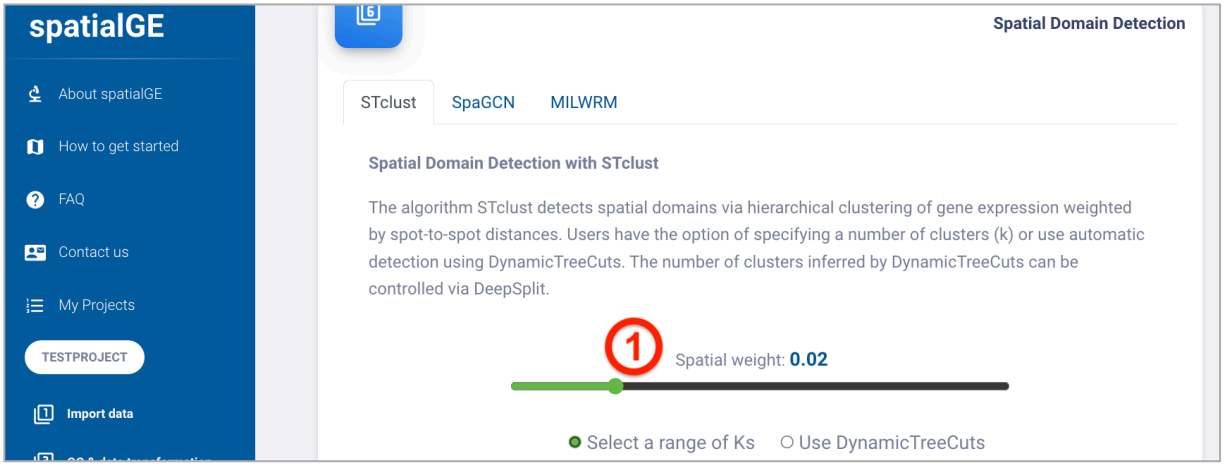
For each sample, STclust calculates Euclidean distances among the ROIs/spots/cells derived from the spatial (x, y) locations. STclust also calculates Euclidean distances derived from the gene expression using the top variable genes (**Number of most variable genes to use**) within that sample. Next, the weighted average between the spatial distances and the gene expression distances is calculated. The weights are derived from the **Spatial weight** parameter, which indicates how much importance should be given to the spatial distances. The weighted distances are used in hierarchical clustering. Finally, the user has the option to allow DyamicTreeCuts (Langfelder et al. 2008), to automatically decide the number of domains in the sample or define domains across a range of values (e.g., find two domains (k=2), three domains (k=3), four domains (k=4), etc.).

Some aspects to consider when running the **Spatial domain detection** module:

- The STclust algorithms works on a sample-by-sample basis and no cross-sample domains are identified. For example, cluster/domain "1" in sample_1 may not be the same as cluster/domain "1" in sample_2. Differential expression analysis can be conducted to assign biological identities to clusters/domains and ascertain similarities across samples.

- Selecting a value for the **Spatial weight** parameter is largely arbitrary. As a general rule, the user should try different combinations of **Spatial weight** with broad range of Ks (**Select a range of Ks**) or **DeepSplit** (if using **DynamicTreeCuts**). The decision on the best number of domains can also be facilitated by inspection of corresponding tissue image. For reasonable results that resemble the overall tissue organization in the tissue images, we have seen that **Spatial weights** as low as 0.025 are enough. **Spatial weights** over 0.1 might lead to non-biologically informative results, as domains will reflect mostly the physical distances among ROIs/spots/cells.
- The **DeepSplit** parameter provides the user with a rough control on the number of domains automatically detected with **DynamicTreeCuts**. Generally speaking, the larger this value is, the more domains are obtained.
- The alternative method SpaGCN is also provided. An additional method, MILWRM, will be implemented soon.

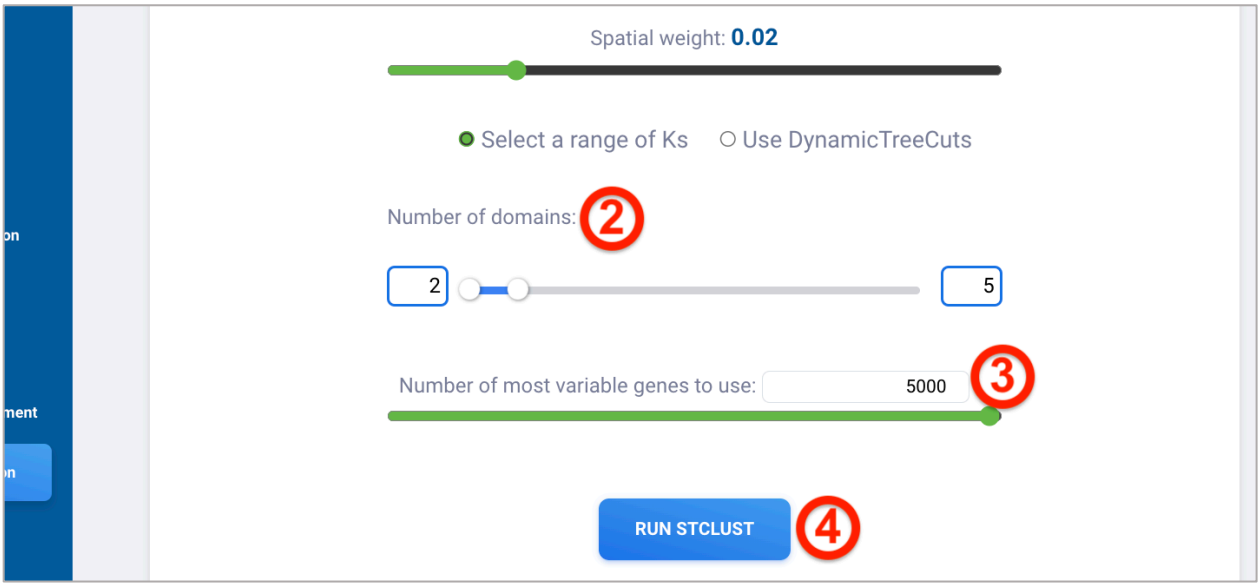
To begin domain detection, begin by defining a **Spatial weight**. For the purpose of this example, the default 0.02 will be kept. In addition to running STclust with **Spatial weight**=0.02, the **Spatial domain detection** module will always run **Spatial weight**=0 too. This behavior allows the user to assess whether adding a spatial weight is necessary or not.



Next, the method for detection of domains is selected. Two options are available: The first allows the user to specify a range of K values (i.e., number of expected domains). This option is activated by clicking the **Select a range of Ks** radio button. The second option allows automatic prediction of the number of domains via the R package DynamicTreeCut and is activated by clicking the Use **DynamicTreeCuts** radio button. For this tutorial, the **Select a range of Ks** radio option will be selected, which enables **Number of domains** slider. The range from 2 to 5 domains will be kept. This setting results in STclust detecting two, three, four, and five clusters/domains in each sample. It is up to the user to decide which number of clusters better explains the biology of their samples. For a more informed decision, users should perform Differential expression analysis (provided as another spatialGE module).

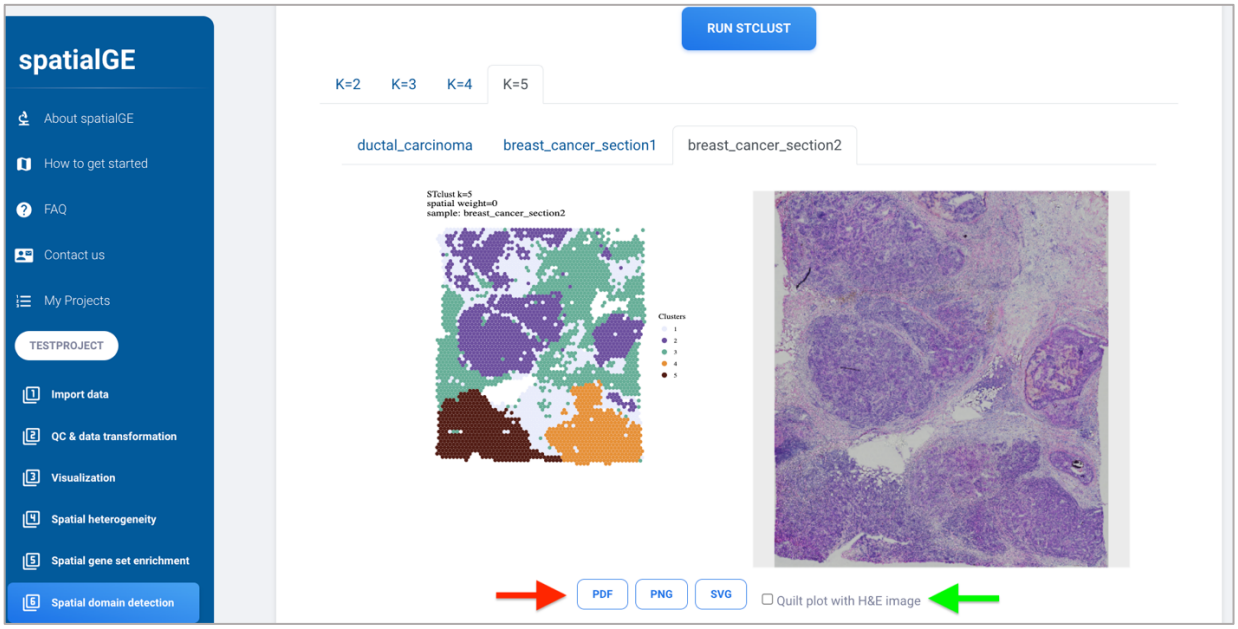
Note: If the user decides to use DynamicTreeCut, then a single clustering solution is presented per sample. Once the **Use DynamicTreeCuts** radio button is clicked, the **DeepSplit** slider is presented. The parameter **DeepSplit** controls the level of "splitting" that DynamicTreeCut performs. The larger **DeepSplit** is, the more domains will be predicted per sample.

Now, the number of genes to be input to STclust is specified in the **Number of most variable genes to use**. Genes are selected based on variance (estimated using the vst method from the R package Seurat). The default (3000 genes) should work for most cases, however, for this tutorial the number will be set to 5000 genes to improve accuracy in the prediction of domains. Note that more genes do not always imply better predictions. Spatial transcriptomics data is generally zero-inflated, and by inputting too many genes, there is the risk of adding more noise, dampening the biological differences between tissue domains. Please, write the number 5000 in the textbox on top of the **Number of most variable genes to use** slider. Finally click on the **RUN STCLUST** button to perform tissue domain detection.



Once the **Spatial domain detection** (STclust) procedure is finished, a series of tabs (K=2 to K=5) are presented in the interface below the **RUN STCLUST** button. Each tab contains a representation of the spatial locations of domains within the tissue samples (and the tissue images if available). Within each tab, additional sub-tabs contain the plots for each sample. These tabs contain the results from STclust when using a **Spatial weight**=0 and **Spatial weight**=0.02. Plots can be downloaded by clicking the appropriate file format button (**PDF/PNG/SVG**, green arrow). The tissue images (if available) can be downloaded along the tissue domain plots by checking the **Quilt plot with H&E image** checkbox (red arrow).

Module: Spatial domain detection – STclust



SpaGCN

The method SpaGCN (Hu et al. 2021) is implemented in the **Spatial domain detection** module. SpaGCN uses graph convolutional networks (GCNs) to conduct unsupervised clustering of the spots or cells in an ST sample. In brief, SpaGCN constructs a graph from the spatial coordinates to describe the spatial relationships between the spots/cells. A GCN is constructed to summarize gene expression of neighboring spots/cells. Domains are identified from the expression graph using Louvain clustering followed by iterative clustering. If the data set contains Visium samples, the user can optionally refine the domain assignments, resulting in domains that are more spatially continuous (see Hu et al. 2021 for details).

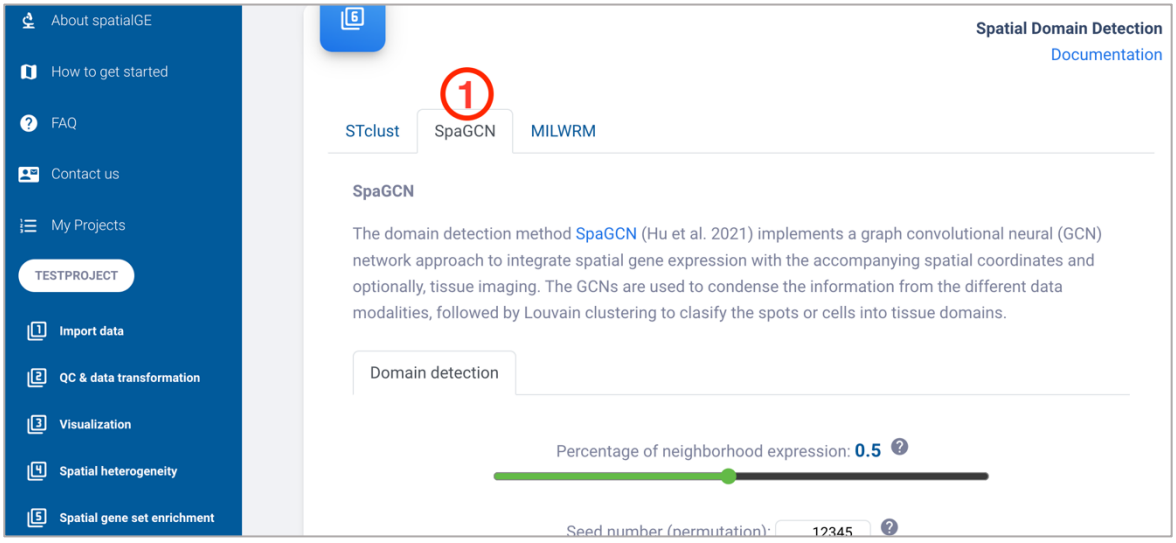
The SpaGCN tool also allows for detecting spatially variable genes (SVGs) for each domain. Users can detect SVGs in the **Spatially variable genes** tab after running domain detection with SpaGCN.

Some additional aspects to consider about the spatialGE implementation of SpaGCN:

- SpaGCN supports the use of tissue images to inform spatial relationships better. This feature is not implemented yet in the **Spatial domain detection** module of spatialGE but will be available in the future.
- A **seed number** argument is available in this implementation. The user is encouraged to run SpaGCN using different seed values and saving the results to check for consistency in tissue domain predictions.

SpaGCN – Domain detection

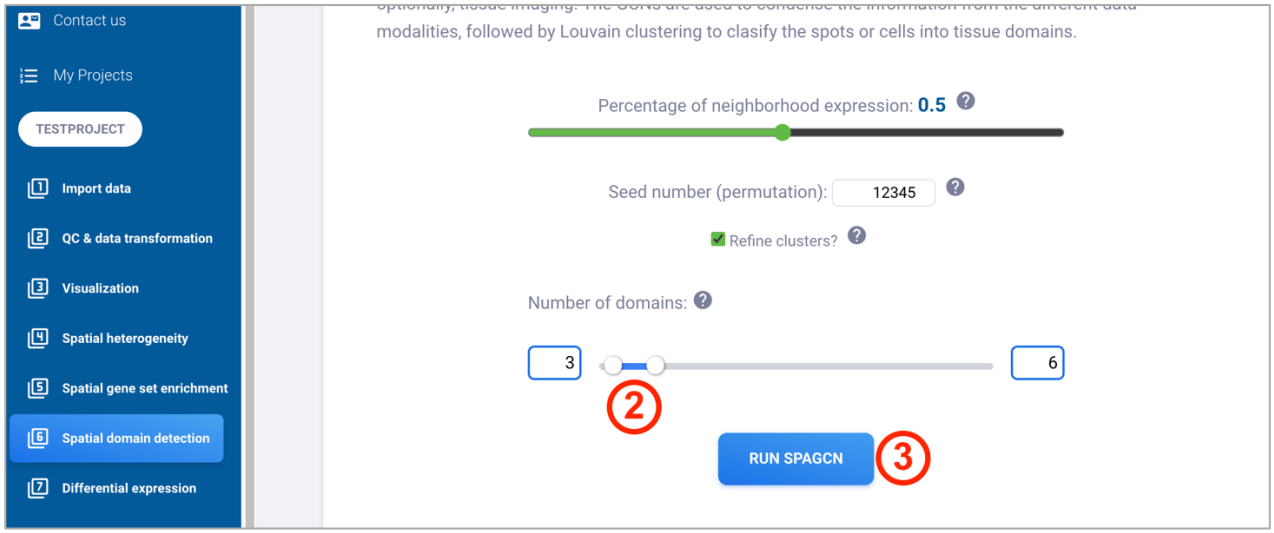
To begin domain detection, click the **SpaGCN** tab next to the STclust tab.



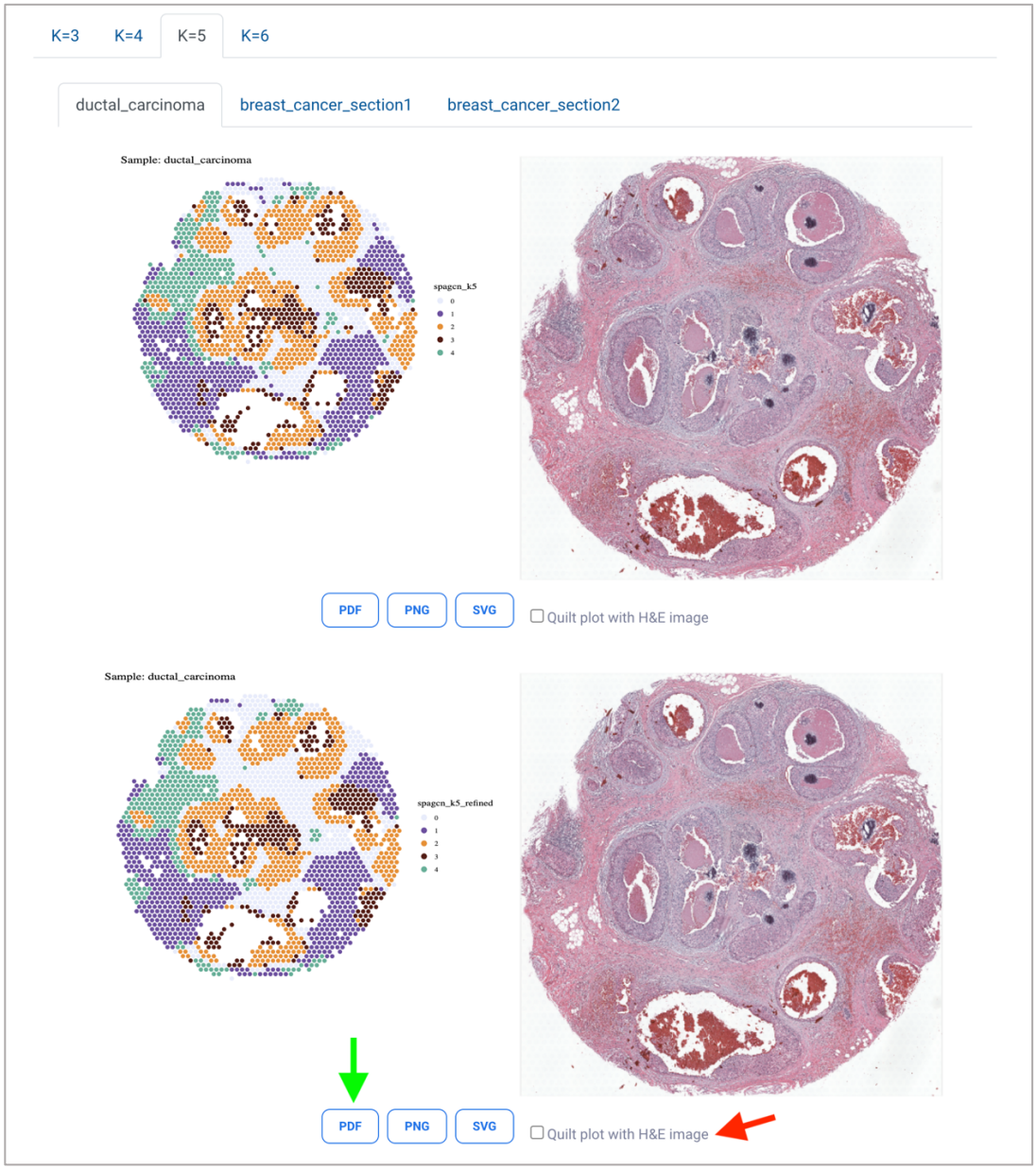
One of the parameters to define when running SpaGCN is the **Percentage of neighborhood expression** contributing to the expression profile of a given spot/cell. The authors of SpaGCN state that an appropriate setting of this parameter for Visium experiments is 0.5 (default). If data from single-cell spatial transcriptomics is used, this parameter should be set to a higher value, so that the summarized expression is more affected by neighboring

cells. Since the data set used here was generated with Visium (see Import data), the **Percentage of neighborhood expression** parameter will be left in its default (0.5).

Next, use the **Number of domains** slider to specify a range of K values (i.e., number of expected domains). SpaGCN will be run for each of the selected K values. For this tutorial, move the slider so that the range of K values goes from 3 to 6. Finally click on the **RUN SPAGCN** button to perform tissue domain detection.



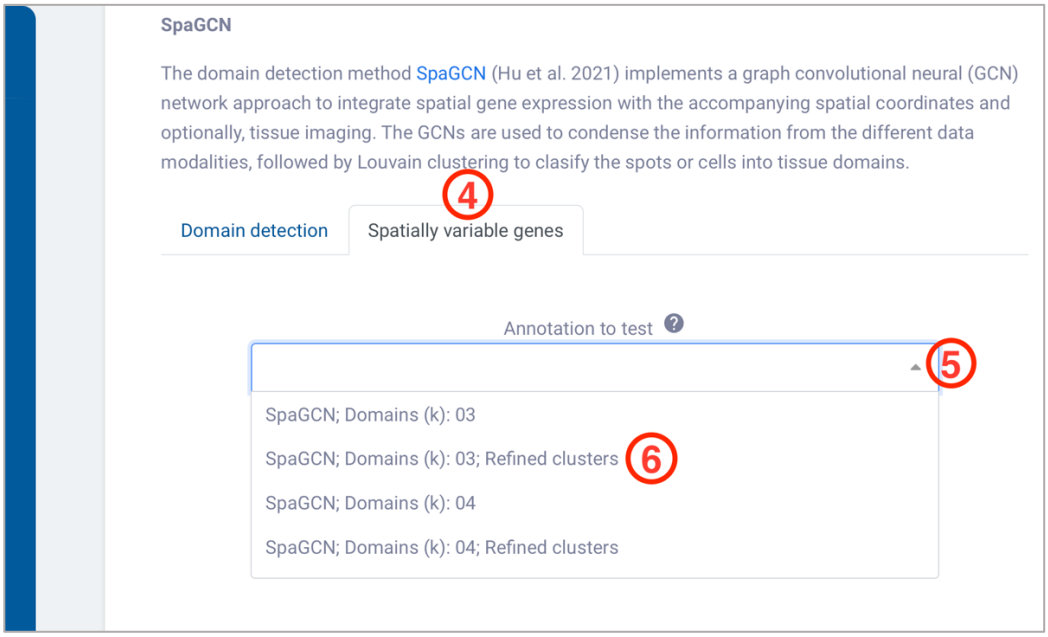
Once the SpaGCN has been completed, a series of tabs (K=3 to K=6) are presented in the interface below the **RUN SPAGCN** button. Each tab contains a representation of the spatial locations of domains within the tissue samples (and the tissue images if available). Within each tab, additional sub-tabs contain the plots for each sample. These tabs contain the results from SpaGCN when for the K value in the first tab level, as well as the *refined* clustering if the **Refine clusters** checkbox was activated. Plots can be downloaded by clicking the appropriate file format button (**PDF/PNG/SVG**, green arrow). The tissue images (if available) can be downloaded along the tissue domain plots by checking the **Quilt plot with H&E image** checkbox (red arrow).



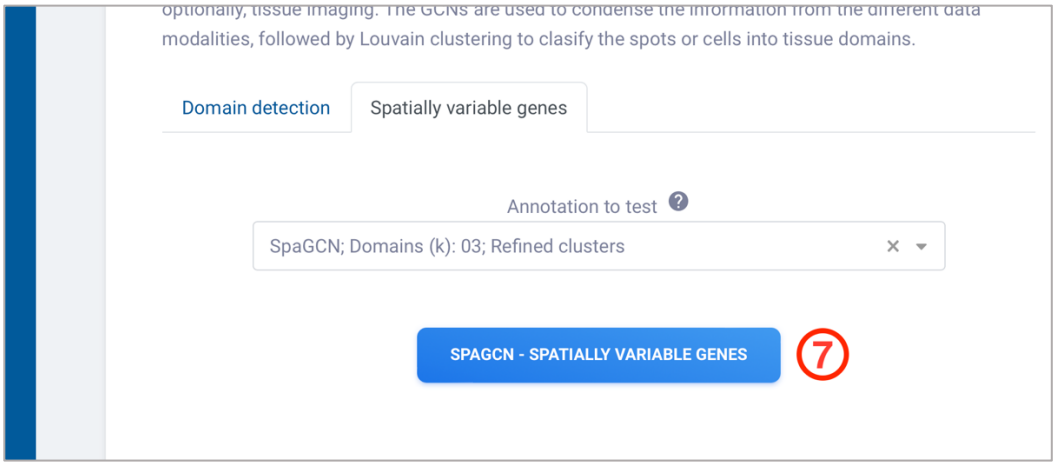
SpaGCN – Spatially variable genes

Now that spatial domains have been inferred, the **Spatially variable genes** tab is enabled. To begin the detection of SVGs, click the **Spatially variable genes** tab. The section includes a single drop-down menu (**Annotation to test**) that allows the user to select the domain detection solution that the user thinks best reflects the biology of the tissue. In practice, users may want to run the analysis several times selecting a different domain detection solution and saving the results each time.

In this example, the SVGs will be detected using the "*SpaGCN; Domains (k): 03; Refined clusters*" option in the **Annotation to test** dropdown. By selecting this option, SpaGCN will find SVGs for each of the clusters previously inferred and refined (Domains 0, 1, and 2).



Next, click the **SPAGCN – SPATIALLY VARIABLE GENES** button to initiate the analysis.



After SVG detection has been completed, a series of tabs appear below the **SPAGCN – SPATIALLY VARIABLE GENES** button. The first level of tabs contains SVGs for each sample. Within each sample tab, SVGs for each domain are shown (three tabs in this case, since K=3 was selected in the **Annotation to test** dropdown). For convenience, the gene symbols are links to the GeneCards database displaying information about each gene.

Module: Spatial domain detection – SpaGCN