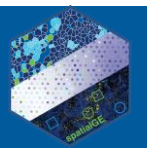
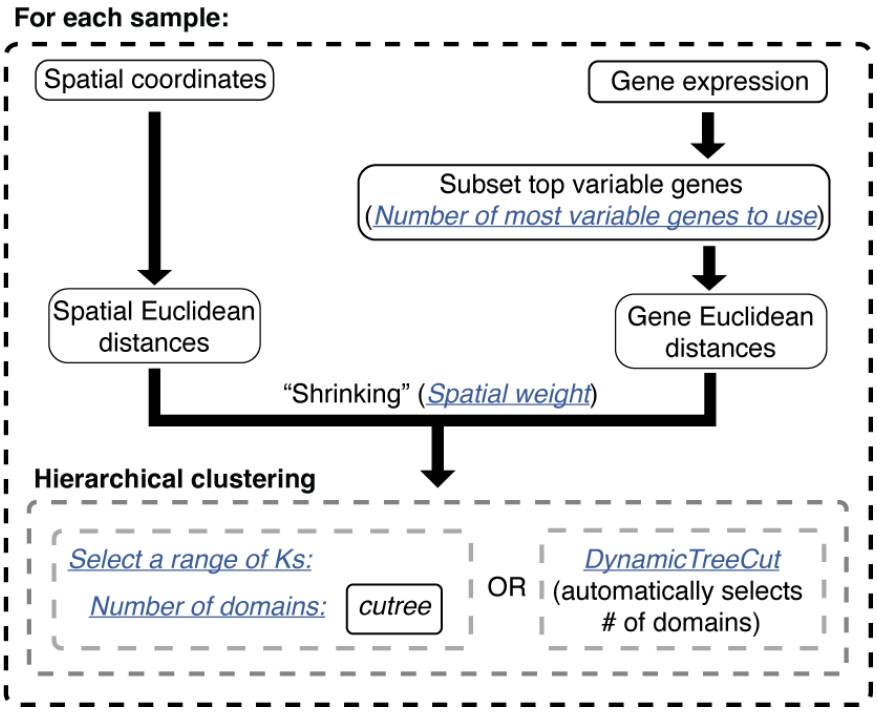


Researchers studying single-cell gene expression commonly use some flavor of clustering algorithm to find cells based on their transcriptomic similarity. When studying spatial transcriptomics (ST) data, this type of analysis acquires an additional purpose: Identify tissue niches or domains. Multiple 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. The clustering algorithm in spatialGE, named STclust, falls in the second category.

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 "shrank" with the spatial information in the form of Euclidean distances (details in Ospina et al. 2022). The diagram below explains the STclust algorithm:

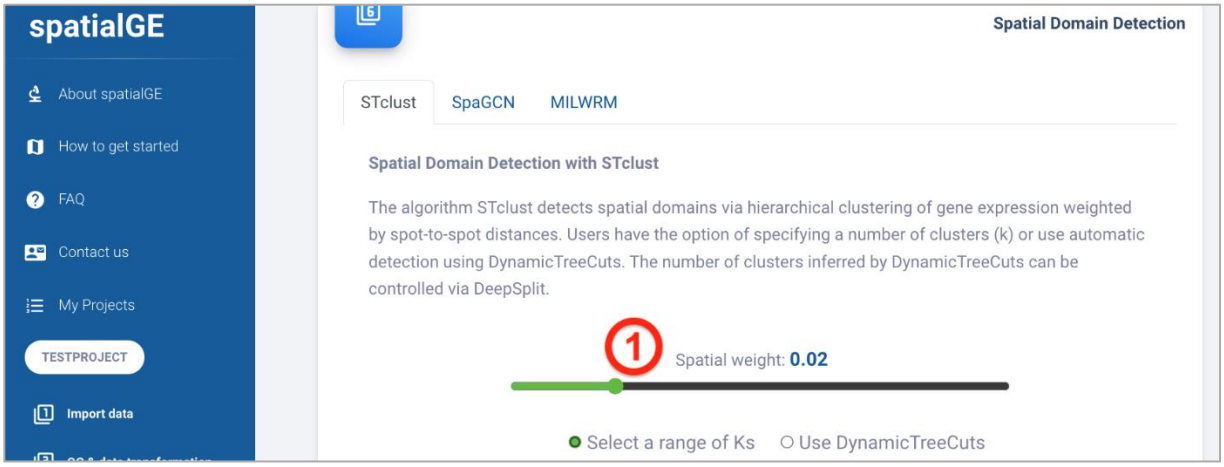


For each sample, STclust calculates Euclidean distances among the ROIs/spots/cells derived from the spatial (x, y) locations. STclust also calculates Euclidean distances derived from the gene expression using the top variable genes (**Number of most variable genes to use**) within that sample. Next, 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 methods SpaGCN and MILWRM will be implemented soon.

To begin domain detection, begin by defining a **Spatial weight**. For the purpose of this example, the default 0.2 will be kept. In addition to running STclust with **Spatial weight**=0.2, 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.

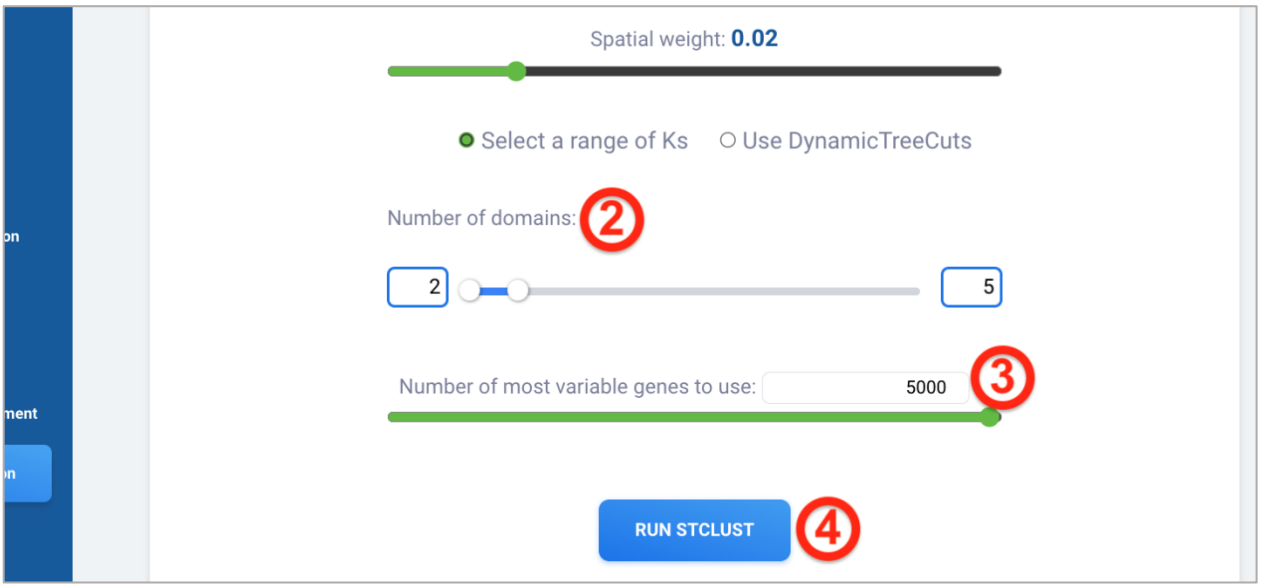


Next, the method for selection of detected domains is selected. Two options are available: The first allows the user to specify a range of K values (i.e., number of expected domains). This option is activated by clicking the **Select a range of Ks** radio button. The second option allows automatic prediction 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**=0.2. 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