SpotXplore User Guide

M. A. Westenberg

1 Introduction

This document provides a compact user guide for SpotXplore. For relevant technical details, we refer to the scientific publications listed in the references. This manual assumes that the user has basic experience in using Cytoscape.

2 Installation

The SpotXplore plugin can be installed by unpacking the archive in the Cytoscape plugins folder. The plugin has been tested on Cytoscape versions 2.6.3 and 2.7.0, but may run on different versions as well.

3 Example data

We will be using the Cytoscape session file `bsubtsession.cys` in all examples. This workspace contains the transcription regulation network of *Bacillus subtilis*, acquired from DBTBS (DataBase of Transcriptional Regulation in *Bacillus subtilis*) [4]. Expression data from a short time series DNA microarray dataset described by Lulko and coworkers [2] are contained in the workspace as well. This data set compares the global mRNA levels at four distinct stages of growth of the bacterium *B. subtilis* strain 168 and the same strain containing a gene deletion. These four growth stages were sampled to obtain a view of the changes in gene expression during growth of this bacterium. The four time points sampled ranged from (i) the early exponential phase (the onset of fast cell growth), (ii) mid-exponential phase (fast cell growth), (iii) end-exponential phase (nutrients start slightly limiting the growth), and (iv) the stationary phase of growth (no growth of cells and start of cell death). The *B. subtilis* strain with a gene deletion has its *ccpA* gene disabled, and it is therefore called a *ccpA* deletion mutant. This comparison was performed by DNA microarrays.

4 Attribute data

4.1 Required node attribute data

SpotXplore requires gene expression ratio data to be available as node attributes. The names of these attributes have to end with the text 'exp', which is default in Cytoscape. For time series data, it is important that the time points are ordered. To ensure that SpotXplore can do this, it is necessary to include a number in the name of the attribute. The example data set uses the names 'tp1exp', 'tp2exp', 'tp3exp', and 'tp4exp'. SpotXplore orders these attributes in increasing order, using only the number in the attribute name.

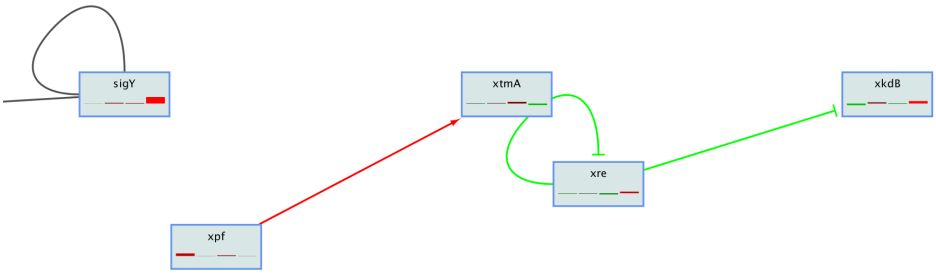


Figure 1: *The optional node attribute 'gene' containing the gene name is used as a node label. The edges are colored and decorated at their target end depending on interaction type, which can be set in the optional edge attribute 'interaction'. Positive interaction results in a red edge color, negative interaction in a green color, and unknown interaction type in a grey color.*

The Luscombe hotspot detection algorithm requires expression levels or signals in addition to ratios for its computation. Such attributes names should end with the text 'lev', so that they can be identified by SpotXplore. The example data set uses 'tp1lev' to 'tp4lev'.

4.2 Optional node and edge attribute data

The node attribute 'gene' will be used as a node label if available. The example data set has this attribute, and it contains the name of the gene. If the attribute is not available, SpotXplore will use the node ID.

The edge attribute 'interaction' is used to determine edge color and end point decoration. It should contain the text 'Positive', 'Negative', or 'ND', for positive regulation, negative regulation, or unknown regulation, respectively. A positive interaction between two genes is represented by a red edge with an arrow head at the end point. Negative interaction is represented by a green edge with a bar at the end point. Unknown interaction results in a grey edge color with no decoration at the end point. The example data set contains this attribute.

If the 'interaction' attribute is not present, a default undecorated edge of grey color will be used. Figure 1 shows part of the example data set, which contains all optional attributes.

5 Running SpotXplore

5.1 Start SpotXplore

The first step is to load the session file `bsubtsession.cys` and start SpotXplore from the Plugins menu. Fig. 2 shows SpotXplore after loading the example dataset. Cytoscape's control panel on the left has a new tab titled 'SpotXplore', which contains the main user interface of the plugin. It consists of a visualization panel and a hotspots panel, which will be described in the next sections. The data panel at the bottom also has a new tab, called 'Hotspots', which is currently empty as no hotspots have been computed yet.

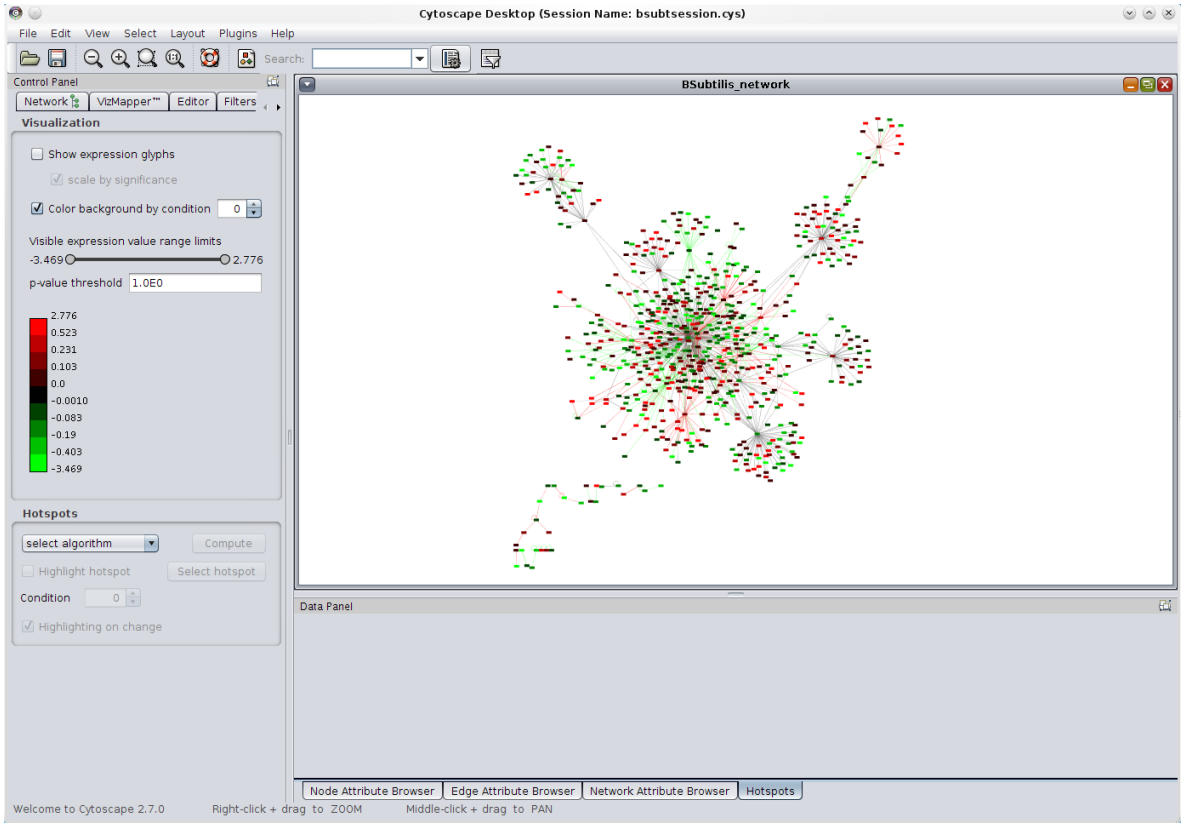


Figure 2: *Initial state of SpotXplore after loading the example session file and starting the plugin.*

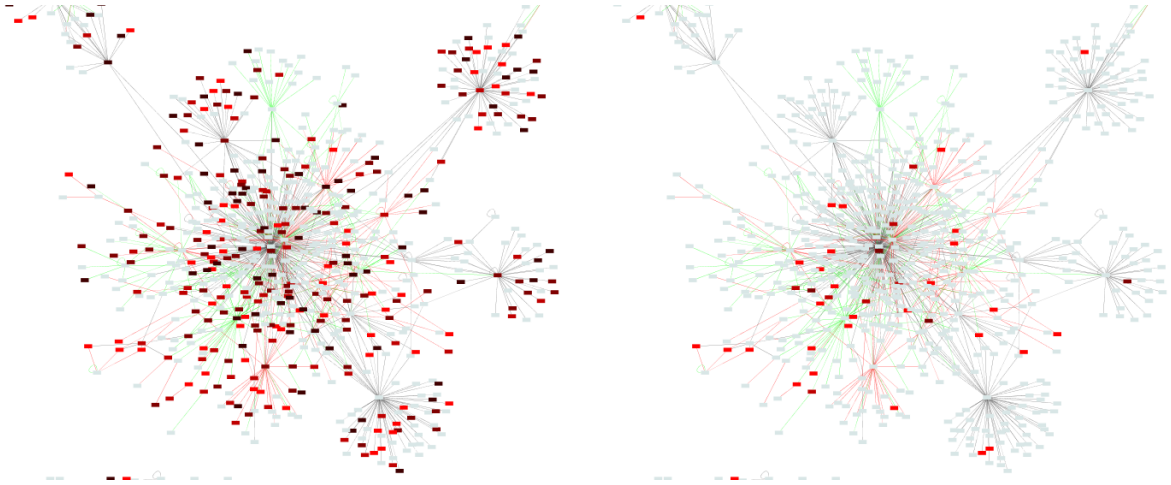
5.2 Visualization panel

The visualization panel offers possibilities to change the visualization of the expression data, and it contains the color legend at the bottom. This legend is constructed automatically based on the loaded expression ratios, and it uses a red-green color map. The data range is divided over nine quantiles, which provides insight into the distribution of ratio values.

5.2.1 Node coloring

By default, each node is colored according to expression ratio. This can be switched on and off by the checkbox. The condition or time point can be browsed by using the number spinner.

The range of expression ratios that is mapped to node color can be limited by the double slider, which sets a lower and upper threshold. Genes with expression ratio within the range defined by the slider will be colored. The other genes will be assigned a default color that is not in the color map. An example of coloring only upregulated genes in time point 0 is shown in Fig. 3(a). In addition, the p-value threshold can be set to a value lower than 1.0, so that only genes with p-value lower than the threshold are colored. This provides a way to map expression ratios to colors of only the genes that show significant change in expression. Fig. 3(b) shows upregulated genes with $p < 0.01$.



(a) (b)

Figure 3: *(a) Upregulated genes colored. (b) Upregulated genes with* $p < 0.01$ *colored.*

5.2.2 Visualize all conditions/time points

The box 'Show expression glyphs' enables visualization of all conditions or the whole time series. SpotXplore uses small rectangles that are embedded in the nodes. Each rectangle is color according to expression ratio, and scaled logarithmically in height according to p-value (Fig. 4). Reliable measurements, i.e., those with low p-value, result in taller rectangles and are therefore given more visual emphasis [5].

If desired, height scaling can be changed to scaling by expression ratio (linearly), by unselecting the 'scale by significance' box. However, this may give noisy expression ratios too much visual

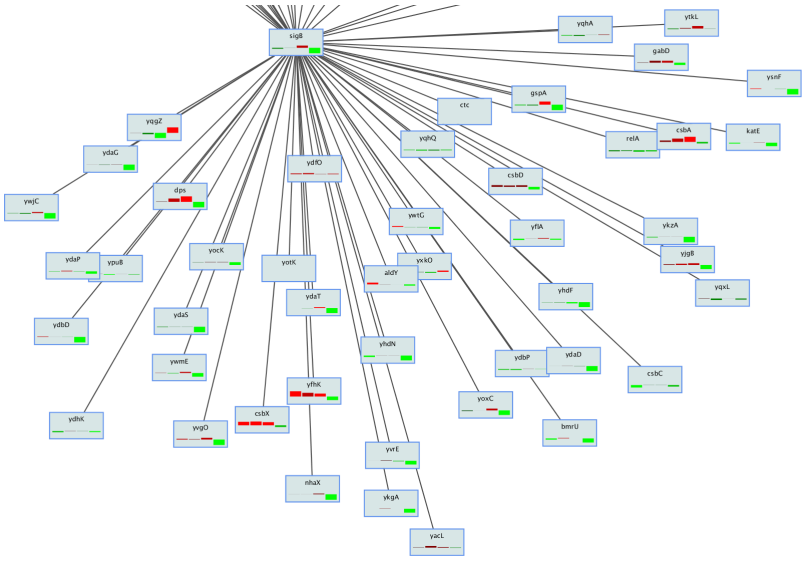


Figure 4: *Node coloring switched off. Expression glyph visualization switched on. Four time points are drawn in each node. Each time point is colored according to expression ratio, and scaled according to p-value. The view was zoomed to part of the SigB regulon, which is located in the bottom right part of the network.*

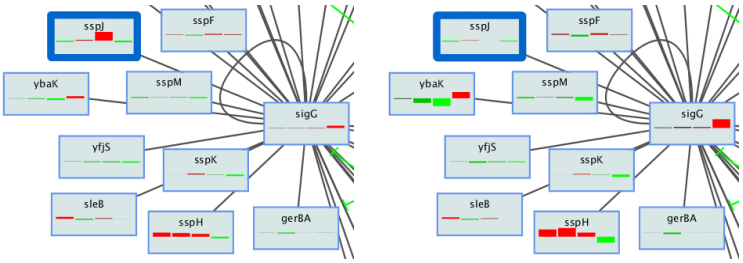


Figure 5: *Left: glyph height scaled linearly by expression ratio. Right: glyph height scaled logarithmically by significance value. Scaling by expression ratio may give the impression that the gene* sspJ *(highlighted by a wide border) is strongly upregulated at time point 3. However, the corresponding p-value is 1.0. Scaling by significance value suppresses these insignificant data points (see right). Conversely, it enhances data points that do show significant differential expression.*

emphasis, see Fig. 5.

5.3 Hotspots panel

Hotspots are network components that are differentially expressed as function of experimental parameters. SpotXplore provides two algorithms to determine such hotspots, and also allows manual construction of hotspots. The desired method can be chosen from the selection box in the hotspot panel. After selecting one of the options, a parameter panel will appear below the hotspots panel. Each method will be described in more detail in the following. The 'Compute' button will run the chosen algorithm, after which the list of detected hotspots appears in Cytoscape's data panel in the 'Hotspots' tab. Hotspots will be computed for all conditions. Various interactions are possible, which are described at the end of this section.

5.3.1 GiGA algorithm

The Graph-based iterative Group Analysis (GiGA) [1] algorithm performs the following steps. First, it assigns each gene a rank based on its expression ratio: low rank means high expression ratio. Next, local minima are identified in the network. A local minimum is a node that has a rank lower than all its direct neighbors. For each local minimum, the regulated neighborhood (hotspot) is then determined by an iterative process. At any step of this process, the maximum rank of the hotspot nodes is r. The hotspot is extended by including the neighbor with the next highest rank $m > r$, and all nodes of rank smaller than m adjacent to any of the current hotspot nodes. This means that in each step, the hotspot has n members with a maximum rank of m. The probability of observing all n nodes of the hotspot at rank m or lower in the whole network is computed by

$$p = \prod_{i=0}^{n-1} \frac{m-i}{N-i}, \tag{1}$$

where N is the total number of nodes in the graph. The process ends if either all nodes reachable from the starting point have been included, or if the hotspot reaches a user-defined maximum size N_{max}. The final hotspot associated with the local minimum is then given by the subgraph that yields the minimum p-value during the extension process.

The maximum size N_{max} of the groups as well as a p-value threshold are parameters of the algorithm, which are strongly data dependent. By default, GiGA searches for differentially expressed

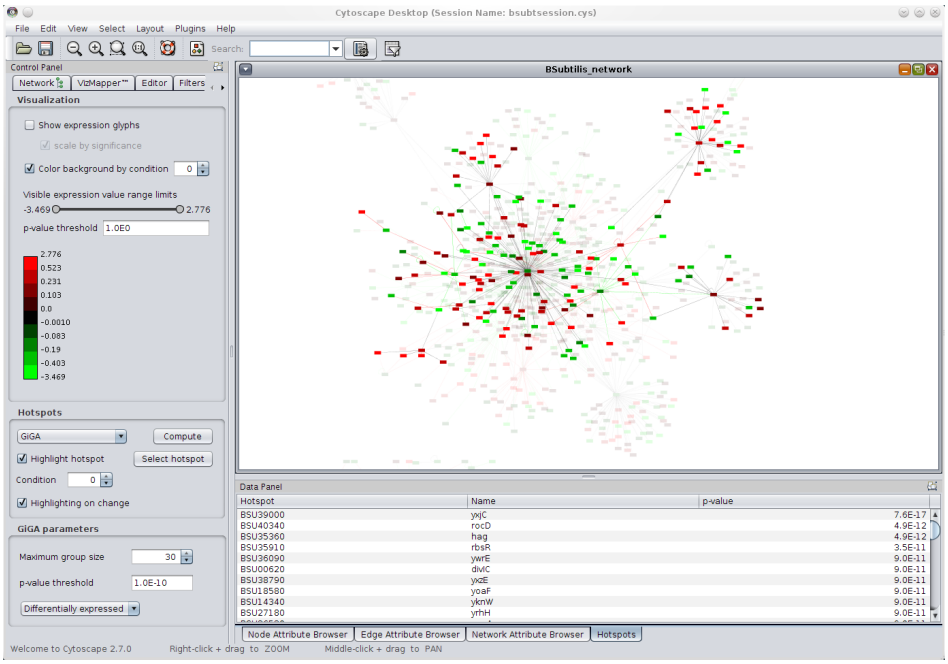


Figure 6: *Hotspots detected in time point 0 by the GiGA algorithm. Default parameter settings were used.*

groups, i.e., it considers the absolute value of the ratio when ranking the genes. If desired, it can search only for upregulated or downregulated groups by selecting one of these options in the parameter panel.

Figure 6 shows the result of running GiGA with default parameter settings on the example data set.

5.3.2 Luscombe algorithm

The Luscombe algorithm [3] identifies subnetworks that are active at a given time point. Denote by $L_{g,t}$ and $R_{g,t}$ the expression level and expression ratio (log-transformed) of the gene g at time point t, respectively. The expression level $L_{g,t}$ is *low* if $L_{g,t} < T_m$, *medium* if $T_m \leq L_{g,t} < T_h$, and *high* if $L_{g,t} \geq T_h$, where T_m and T_h are thresholds so that $0 \leq T_m < T_h$. A gene is differentially expressed if $|R_{g,t}| \geq T_r$, where $T_r > 0$ is a threshold. All thresholds can be adjusted, since they may be data dependent. The detection algorithm proceeds as follows:

1. Determine active regulators (nodes with at least one outgoing edge) in the network. A regulator gene g is active at time point t if one of the following conditions holds:

$$L_{g,t} < T_m \text{ and } R_{g,t} \geq T_r \tag{2}$$

$$T_m \leq L_{g,t} < T_h \text{ and } R_{g,t} \geq 0 \tag{3}$$

$$L_{g,t} \geq T_h \tag{4}$$

2. Determine active nonregulator genes (nodes with only incoming edges or with degree zero). A nonregulator gene g is active at time point t if

$$|R_{g,t}| \geq T_r \tag{5}$$

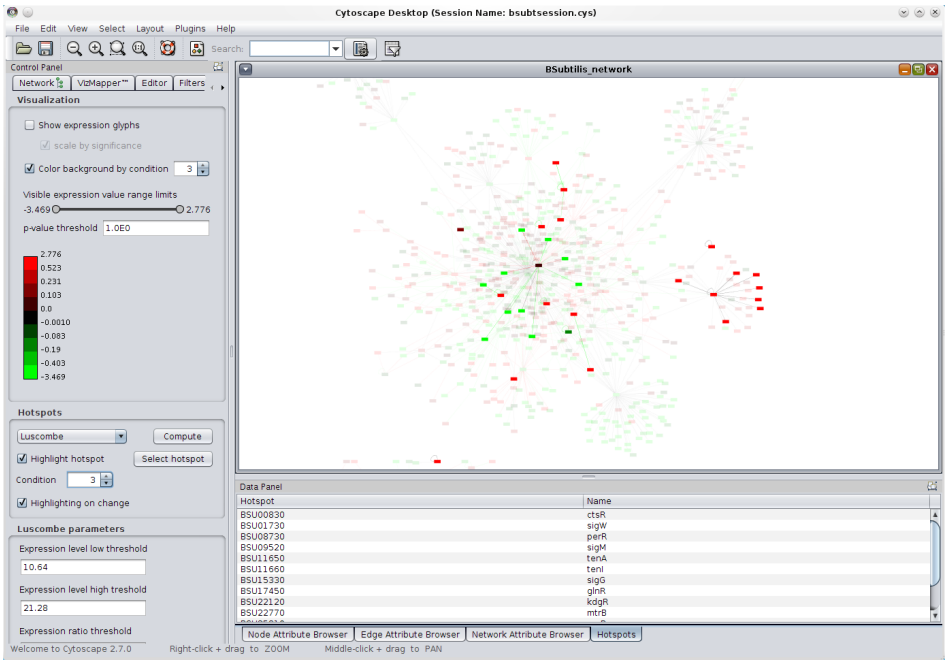


Figure 7: *Hotspots detected in time point 3 by Luscombe's algorithm. Default parameter settings were used.*

3. Determine active links. A link between a source gene and target gene is active if both endpoints are active.

The result of running Luscombe's algorithm on the example data is shown in Fig. 7.

5.3.3 Manual definition

Hotspots can be defined manually by Cytoscape's selection mechanism. First, set the condition or time point for which hotspots are to be defined by the spinner in the hotspots panel. Then select some nodes in Cytoscape, and click the 'Add selection as hotspot' button. The hotspot will show up in the data panel, and will be named '1'. Making another selection and clicking the button again will add the new selection as hotspot '2'. To define hotspots for a different condition, simply select that condition with the spinner and repeat the selection steps.

The 'Compute' button will clear *all* manually defined hotspots.

5.3.4 Interacting with hotspots

The table of detected hotspots appears in Cytoscape's data panel in the 'Hotspots' tab. Depending on the chosen algorithm, the table contains only hotspot names or also p-values. The default visualization highlights all detected hotspots in the condition chosen in the 'Hotspots' panel. An individual hotspot can be highlighted by selecting it in the table.

The checkbox 'Highlighting on change' can be unselected. This changes the default behavior of SpotXplore to highlight all the detected hotspots for a given condition when the user browses over conditions. Instead, the nodes will then be all colored according to expression ratio, and not only those that are member of a hotspot. To shows the hotspots, check the 'Highlight hotspot' checkbox,

or simply select a hotspot from the list. The checkbox will be unchecked automatically when the user switches to another condition.

The genes of a hotspot that is highlighted can be selected in Cytoscape by the 'Select hotspot' button in the panel. This allows further analysis of these genes in other plugins, or a new network may be constructed that contains only this set of genes. All edges that connect these genes will be included in the selection set as well.

5.3.5 Comparing hotspots

Cytoscape's functionality to create new networks from a selection set can be used to create multiple visualizations of hotspots for comparison.

Figure 8 shows an example. The top part of the figure shows the hotspot *gabD* selected by the user. After pressing the button 'Select hotspot', the nodes and edges of the hotspot become a selection set in Cytoscape. A new network can be created from this selection via the File menu, then New/New network/From Selected Nodes,Selected edges (or by simply pressing CTRL+SHIFT+N). To be able to see expression glyphs, it is necessary to invoke the SpotXplore plugin also on the new network. The bottom part of Fig. 8 shows a view of the newly created network (containing only the nodes of the hotspot and their interactions) and the initial network view. The latter shows the region of the network from which the hotspot was extracted, but now colored according to condition 3. The side-by-side visualization allows comparison of the members of these two gene sets.

References

[1] R. Breitling, A. Amtmann, and P. Herzyk. Graph-based iterative group analysis enhances microarray interpretation. *BMC Bioinformatics*, 5:100, 2005.

[2] A. T. Lulko, G. Buist, J. Kok, and O. P. Kuipers. Transcriptome analysis of temporal regulation of carbon metabolism by CcpA in *Bacillus subtilis* reveals additional target genes. *Journal of Molecular Microbiology and Biotechnology*, 12(1–2):82–95, 2007.

[3] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Telchmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Letters to Nature*, 431(7006):308–312, 2004.

[4] Y. Makita, M. Nakao, N. Ogasawara, and K. Nakai. DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Research*, 32:D75–77, 2004.

[5] M. A. Westenberg, S. A. F. T. van Hijum, A. T. Lulko, O. P. Kuipers, and J. B. T. M. Roerdink. Interactive visualization of gene regulatory networks with associated gene expression time series data. In L. Linsen, H. Hagen, and B. Hamann, editors, *Visualization in Medicine Life Sciences*, Visualization and Mathematics, pages 293–312. Springer Verlag, Berlin, Germany, 2007.

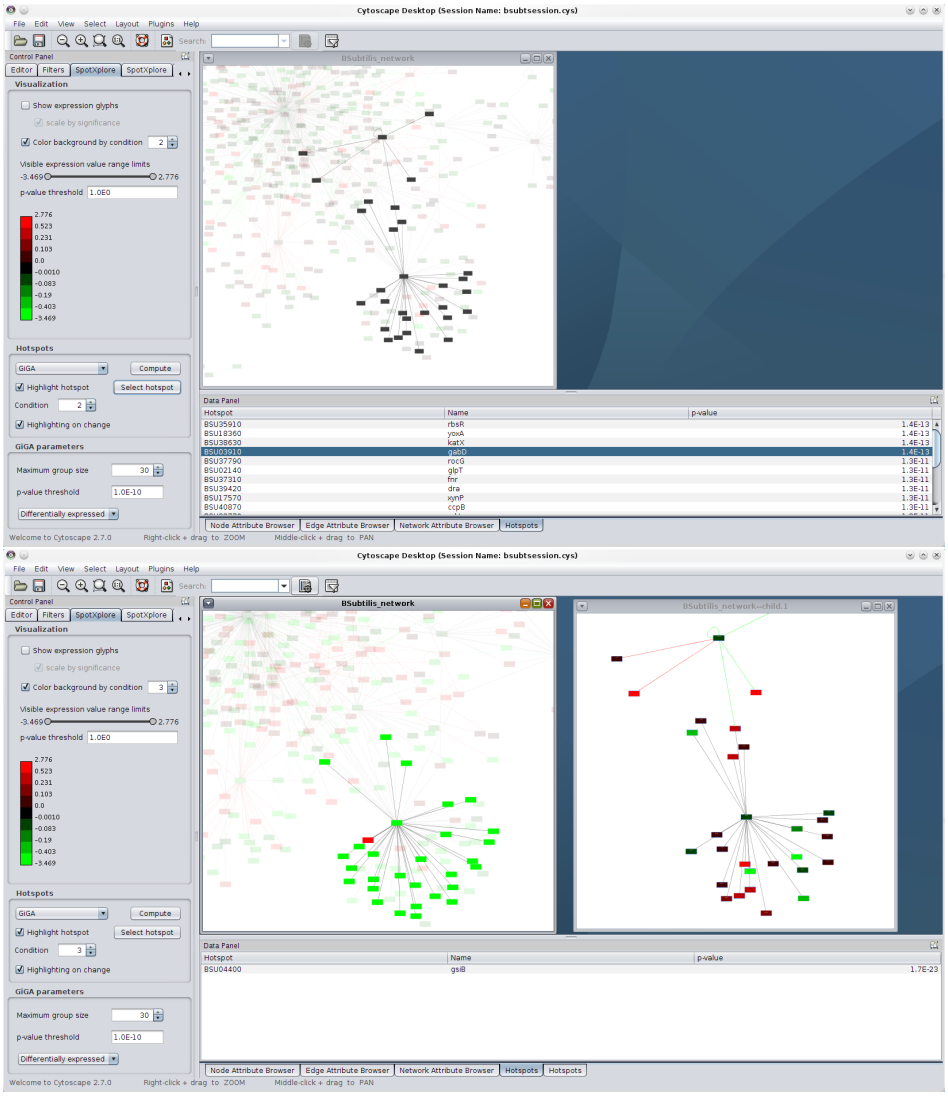


Figure 8: *Top: Hotspot* gabD, *detected in condition 2 by the GiGA algorithm. It is colored in grey, since the 'Select hotspot' button has been pressed to create a Cytoscape selection set, which is used to create a new network by standard Cytoscape facilities. Bottom: the main view shows this region of the network in condition 3. To its right, a separate view of the newly created network. Side-by-side visualization allows comparison of the members of these two gene sets.*