Guide for
Data Visualization and Analysis using ACSN

ACSN contains the NaviCell tool box, the intuitive and user-friendly environment for data visualization and analysis. The tool is accessible from the ACSN website and located at the bottom of the ACSN selection panel (figure 1).

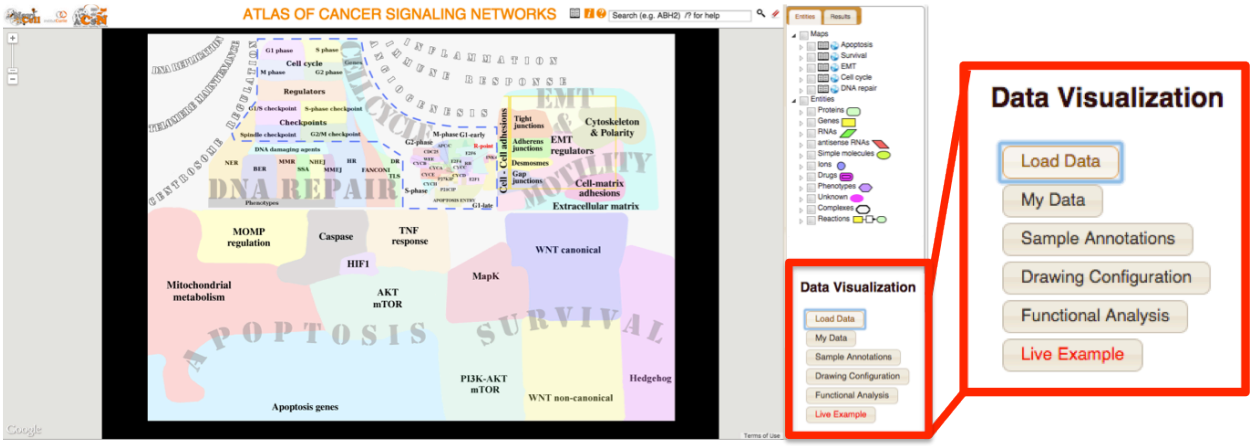


Figure 1: Localization of the data visualization and analysis menu buttons on the ACSN website.

NaviCell tool box allows users to upload several types of "omics" data and visualize them in the context of ACSN maps. The user can upload expression data for mRNA, microRNA, proteins, mutation, copy-number data, and simple gene lists using the input data format is text file. Users can also upload sample annotation files that can serve for defining groups of samples. Depending on the nature of data, different types of visualization modes can be required to achieve the informative picture. Different options are available to display data values: heat maps, bar plots, glyphs and map staining. Finally, the users can overlay several types of data and visualize them simultaneously. NaviCell tool box allows to visualize the data at different zoom levels, starting from the top level view, were patterns of integrated data can be grasped, up to the most detailed view at the level of individual molecules.

Several choices of data visualization providing an opportunity to find the best combination of zoom and data display mode to achieve the most insightful data visualization design.

Loading data

When clicking on the "Load Data" button, a window appears, with fields to specify the data files to be loaded and the display options (figure 2). The user can browse his local file system (button "browse") to choose a file, or specify in the field "URL" a web address to retrieve the file from. The user should also specify a name for the data, and choose an appropriate type of data from the drop-down menu "Type".

There are several types of data that can be imported:

1. mRNA expression data
2. microRNA expression data
3. Protein expression data

4. Discrete copy number data
5. Continuous copy number data
6. Mutation data
7. Gene list
8. Data table list

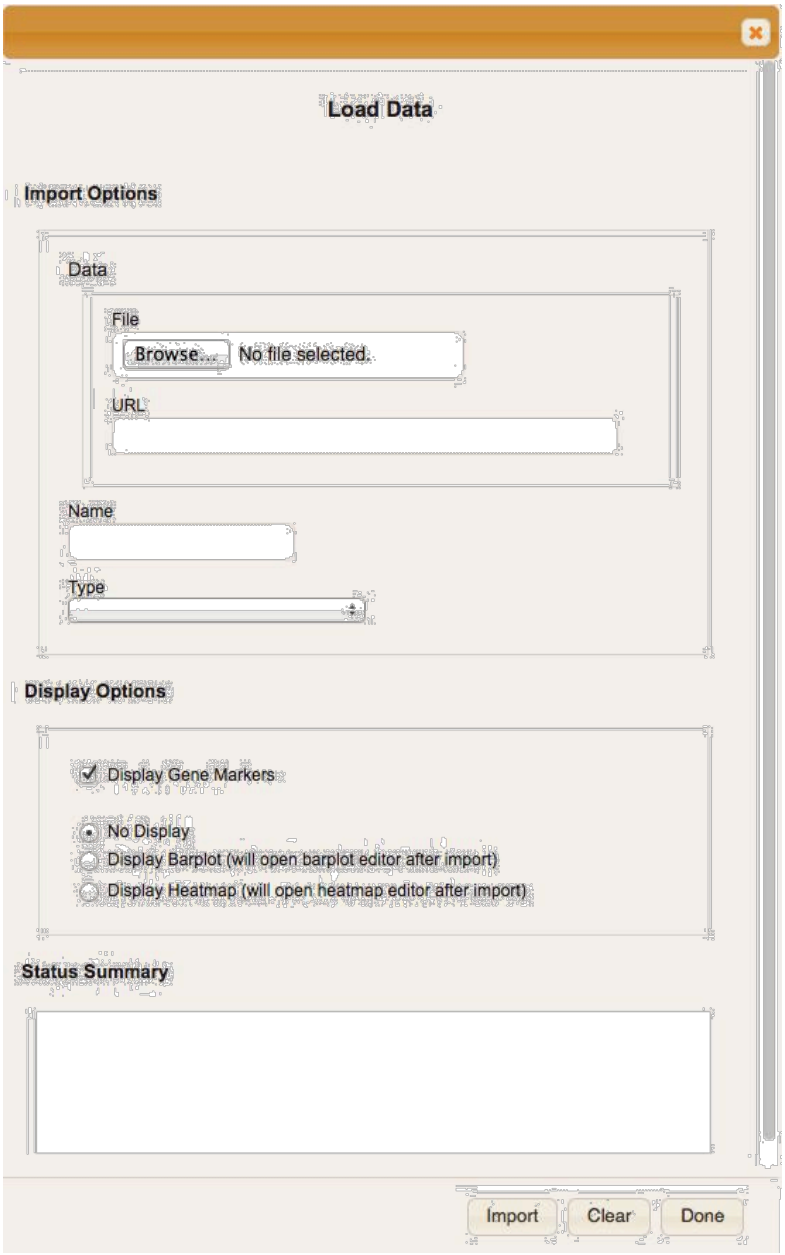


Figure 2: Dialog for loading data. The txt file can be loaded from user's computer or accessed through website and downloaded.

The format for the data types 1-6 is a simple tab-delimited tabular format, with the gene identifiers as the rows, and the samples or experiments IDs as columns. The first line contains the sample names. The data type "gene list" consist of a simple list of genes of interest in a text file. The type "data table list" is a text file containing a list of one or more web links (URLs) of different data sources. Each source of data will be loaded automatically.

For example expression data from the file “CCL_Expression.txt” has been imported by selecting the file from the local directory with the “browse” button and then specifying the name “CCL Expression” and type “mRNA Expression data”., and finally click the “Import” button.

The status summary field at the bottom of the window is displaying error messages and some statistics related to the import (Figure 3). The genes present in the data set are mapped to the entities of the map by HUGO name. The total number of genes mapped is indicated in the status window.

Figure 3: Field displaying the import report and statistics after a successful importation of a data table.

Managing imported data tables

Imported data can be managed by clicking on the “My Data” button (figure 4). The window displays the list of imported data sources (named hereafter data tables”). Data tables can be renamed or eventually removed by the user.

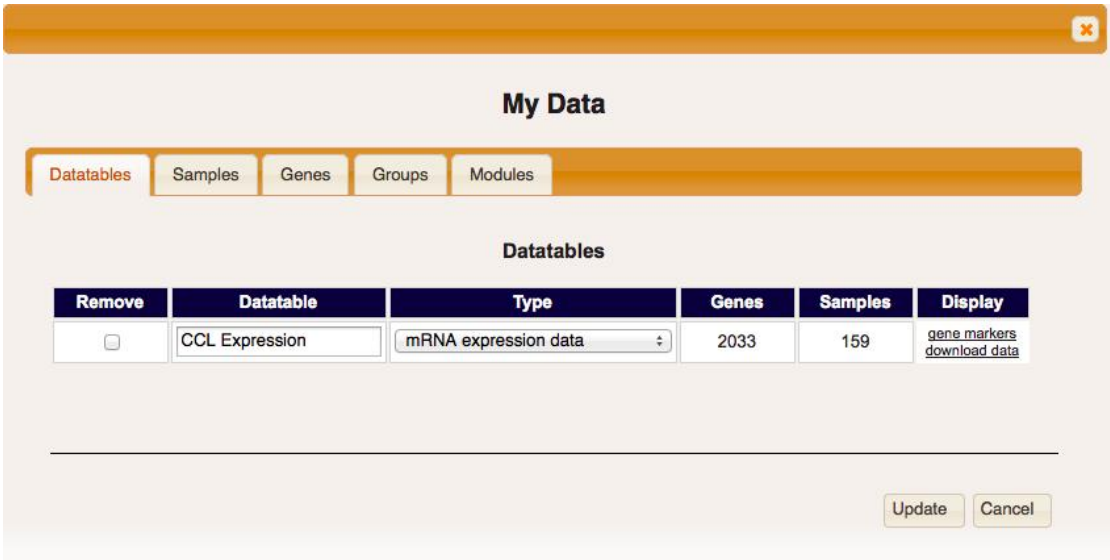
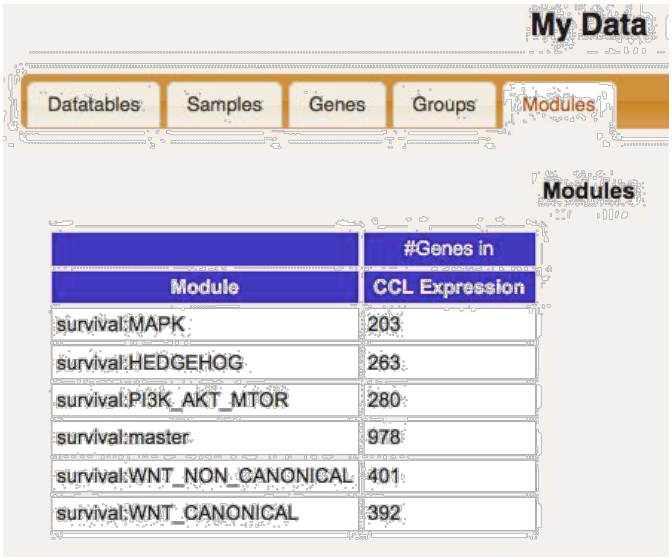


Figure 4: Management of imported data.

The tab “Samples” display the list of sample names from all imported data tables. The complete list of genes imported from all data sources can be seen on the “Genes” tab. The

tab “Groups” is used with sample annotations (see the corresponding section below). The tab “Modules” shows the number of genes per map module (annotated sub-networks within the map, see figure 5).



Module	#Genes in CCL Expression
survival:MAPK	203
survival:HEDGEHOG	263
survival:PI3K_AKT_MTOR	280
survival:master	978
survival:WNT_NON_CANONICAL	401
survival:WNT_CANONICAL	392

Figure 5: Modules tab after import of expression data for the survival map. The number of genes is displayed for the whole map (“master”) and for the different modules composing the map.

Loading and managing annotations

Sample annotations can be used to define groups of samples representing different conditions, such as disease versus normal samples. The format for the file is tab-delimited text, with sample names as rows (The IDs must match the IDs of the samples in the other data tables), and groups as columns. For example figure 6 shows an extract of the first rows and columns of the file “SampleAnnotations.txt”, with the groups “Tissue”, “Cancer Type”, “Cancer subtype” and “Sample Gender”, visible for a couple of rows.

NAME	Tissue	Cancer/Type	Cancer/Subtype	Sample/Gender
1321N1_CENTRAL_NERVOUS_SYSTEM	CENTRAL_NERVOUS_SYSTEM	glioma	astrocytoma	M
143B_BONE	BONE	osteosarcoma	NS	F
22RV1_PROSTATE	PROSTATE	carcinoma	NS	M
2313287_STOMACH	STOMACH	carcinoma	adenocarcinoma	M
42MGBA_CENTRAL_NERVOUS_SYSTEM	CENTRAL_NERVOUS_SYSTEM	glioma	astrocytoma_Grade_IV	M
5637_URINARY_TRACT	URINARY_TRACT	carcinoma	NS	M
59M_OVARY	OVARY	carcinoma	NS	F
639V_URINARY_TRACT	URINARY_TRACT	carcinoma	transitional_cell_carcinoma	M
647V_URINARY_TRACT	URINARY_TRACT	carcinoma	transitional_cell_carcinoma	M

Figure 6: Extract from the annotation file “SampleAnnotations.txt”.

The file can be loaded by clicking on the button “Sample Annotations” from the right-hand panel menu. The file can be specified by browsing the local directory or by indicating a URL (figure 7), and imported by clicking on the “Import Annotations” button.

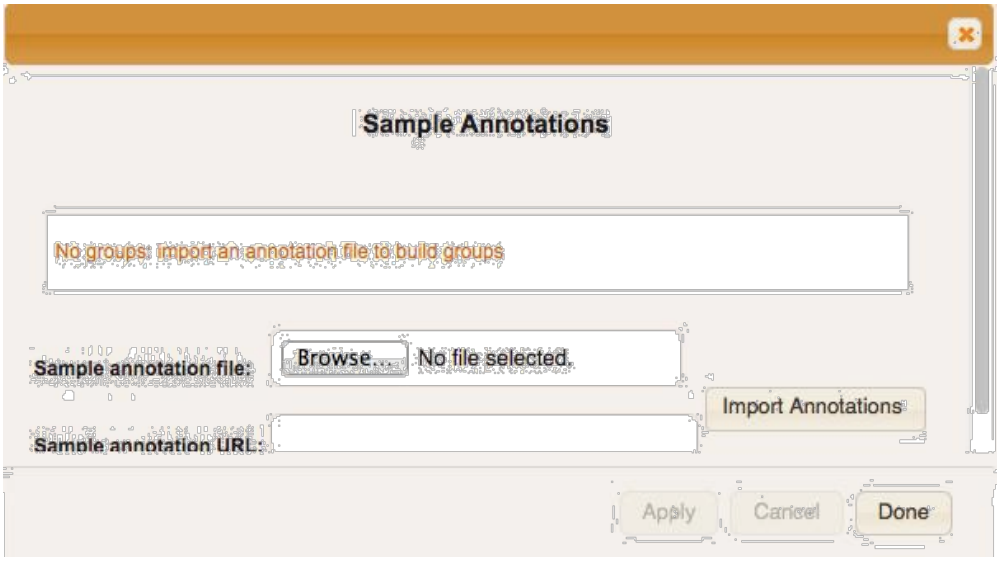


Figure 7: Sample annotation window.

After a successful import, a short report is displayed in the status field, and the annotations are showed as a table (figure 8). From there, it is possible to immediately define groups. For example, we can select the "Tissue" column and then click the button "Apply" to create groups based on the tissue type (figure 8).

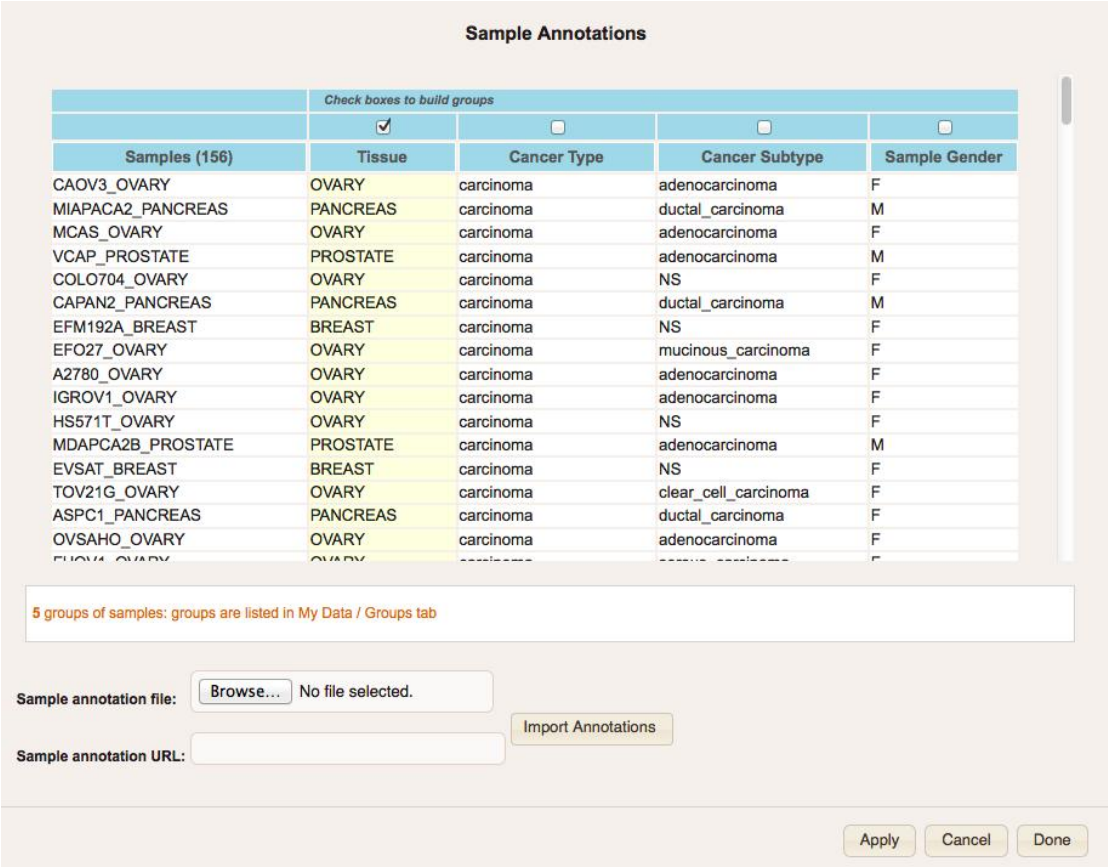


Figure 8: Annotation window after a successful import and a selection of the field "Tissue" as a grouping factor.

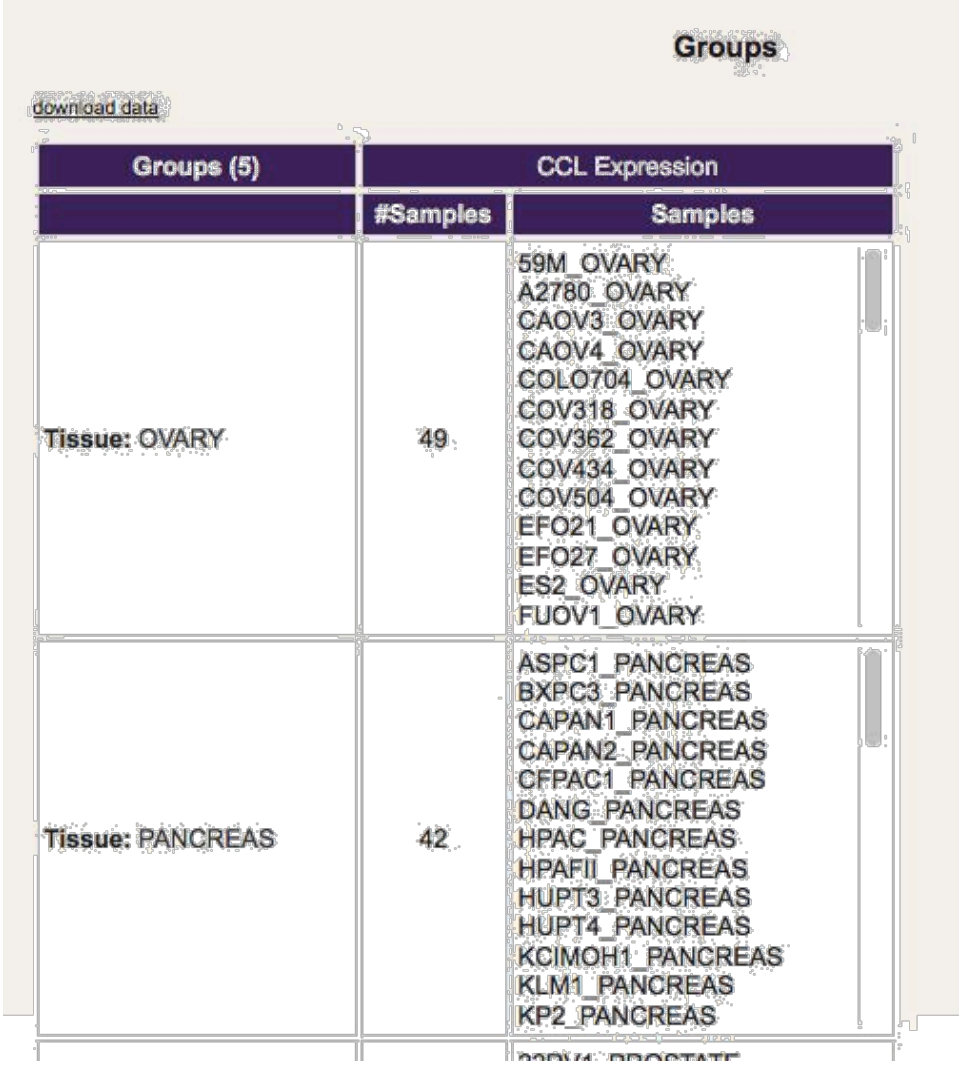


Figure 9: Visualization of tissue groups.

The groups can then be visualized and used in different contexts. For example, the figure 9 is showing the tab “Groups” of the “My Data” window, displaying the grouping of the samples by tissue type. At any moment, the use can remove or change groups by clicking on the “Sample Annotations” button.

Data visualization modes and drawing configurations

These functions allow the user to use different visualization modes for different data types, samples and groups (figure 10). The user can use chart (heatmaps and barplot), glyph and map staining types to display the data.

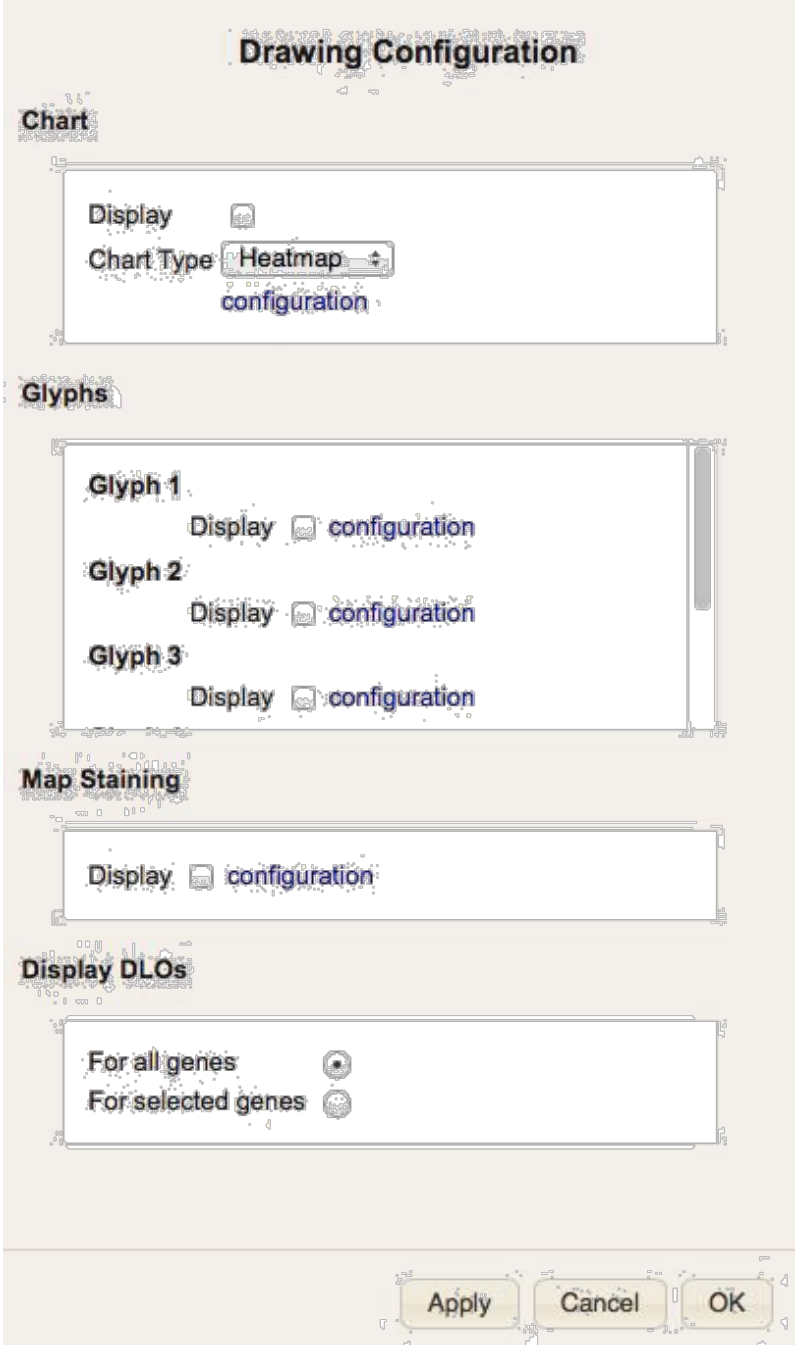


Figure 10: Drawing configuration window options.

Visualizing data using heatmap chart

For example expression data can be visualized as a heatmap chart type. In the "Drawing Configuration" window, check "display" on the 'Chart type, Heatmap', and then click on the "configuration" link. In the heatmap configuration editor window (figure 11), choose the expression data table ("CCL Expression", column "Datatables"). Then select the tissue group of samples "BREAST" and "PROSTATE" in the first and second column respectively. Finally, click on the "Apply" button to apply the changes to the map.

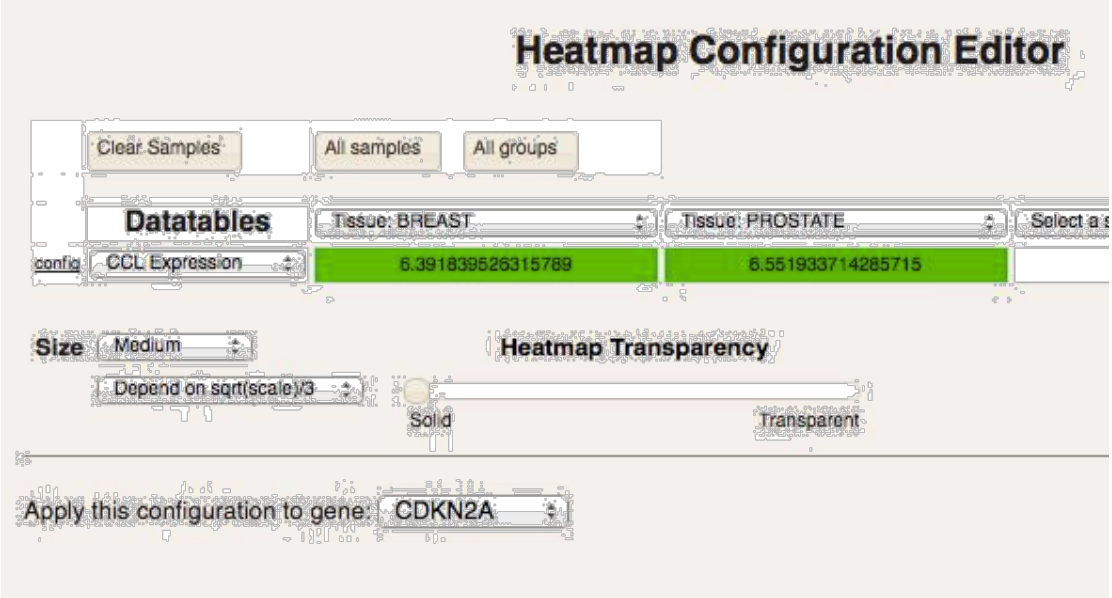


Figure 11: Heatmap configuration options.

The heatmaps are now visible on the map, as pairs of colored squares next to each entity of the map for which there is a match between the data source and the map (figure 12).

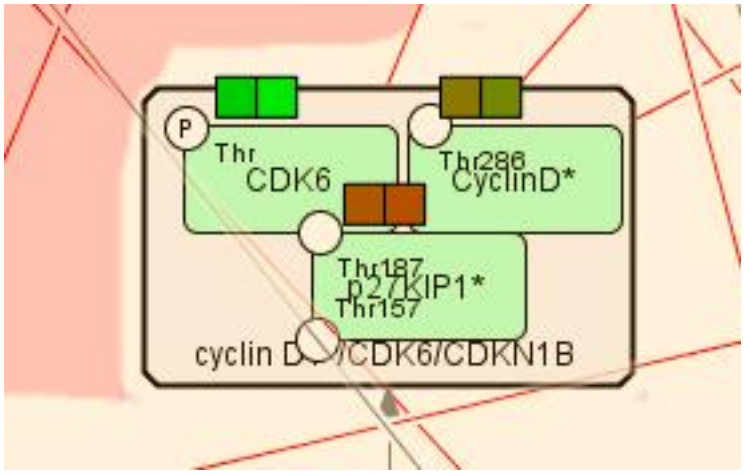


Figure 12: Zoom on three entities of the cell cycle map with heatmaps displaying expression data for the "BREAST" (left-side) and "PROSTATE" (right-side) groups of samples.

Visualizing data using glyphs

Glyph has three characteristics: shape, color and size, each one of those characteristics can be configured according to a different feature in the data. There is a possibility to display up to 5 glyphs on top of each node in the network simultaneously.

In the "Drawing Configuration" window, check "display" on the 'Glyph, and then click on the "configuration" link. In the glyph configuration editor window (figure 13), choose the sample "22RV1_PROSTATE ", choose the expression data table ("CCL Expression") for configuring the shape, color and size of the glyph and heat the "Apply" button to apply the changes to the map.

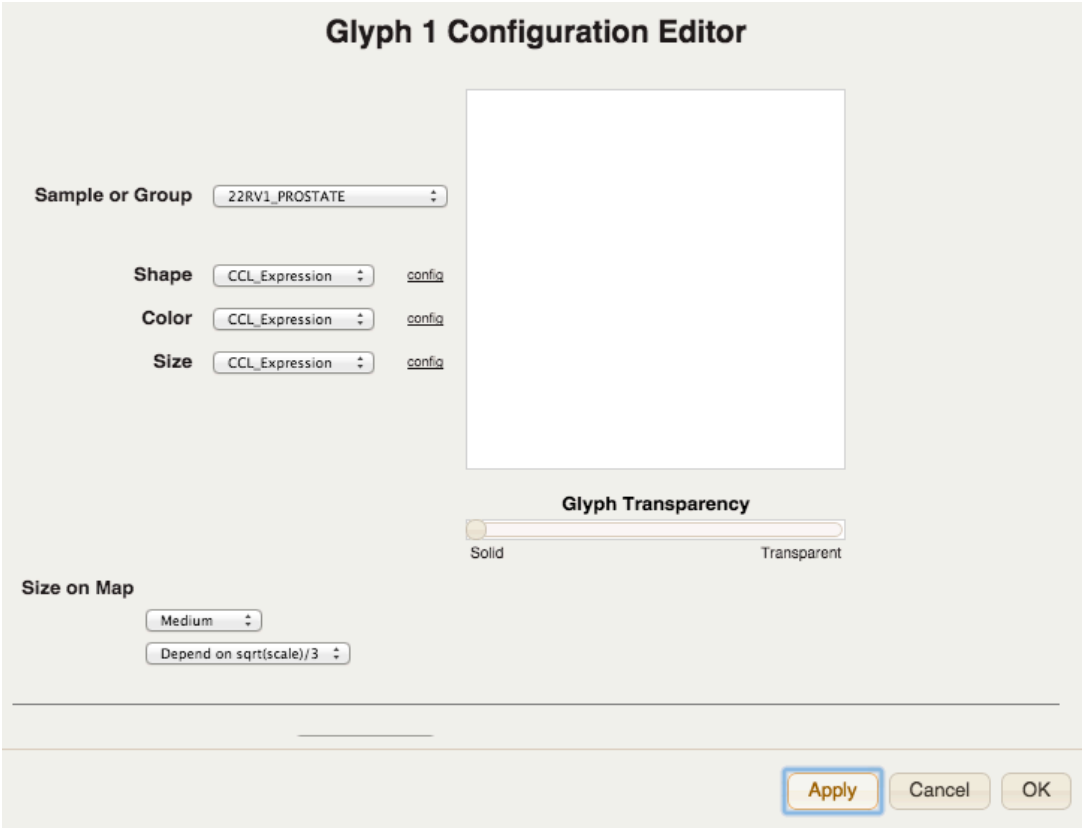


Figure 13: Glyph configuration options.

The glyphs are now visible on the map, as colored triangles next to each shape of the p53* entity of the map (figure 12).

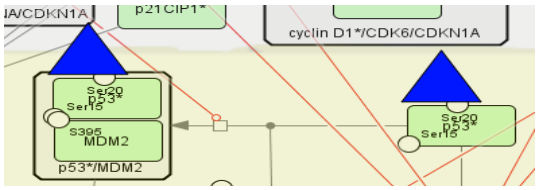


Figure 14: Zoom on the entity p53* on the cell cycle map with glyph displaying it's expression data for sample "22RV1_PROSTATE".

Visualizing data using map staining

An advanced and novel mode of data visualization is map staining. The principle of map staining is in using the background of the map for visualizing the values mapped to individual molecular entities or group of entities (modules). The resulting colorful background of the network map, provides a possibility to grasp differences in the patterns of data distribution between samples or between groups of samples.

In the "Drawing Configuration" window, check "display" on the 'Map staning', and then click on the "configuration" link. In the Map staining configuration editor window (figure 15), choose the sample "22RV1_PROSTATE ", choose the expression data table ("CCL Expression") for configuring the color and heat the "Apply" button to apply the changes to the map.

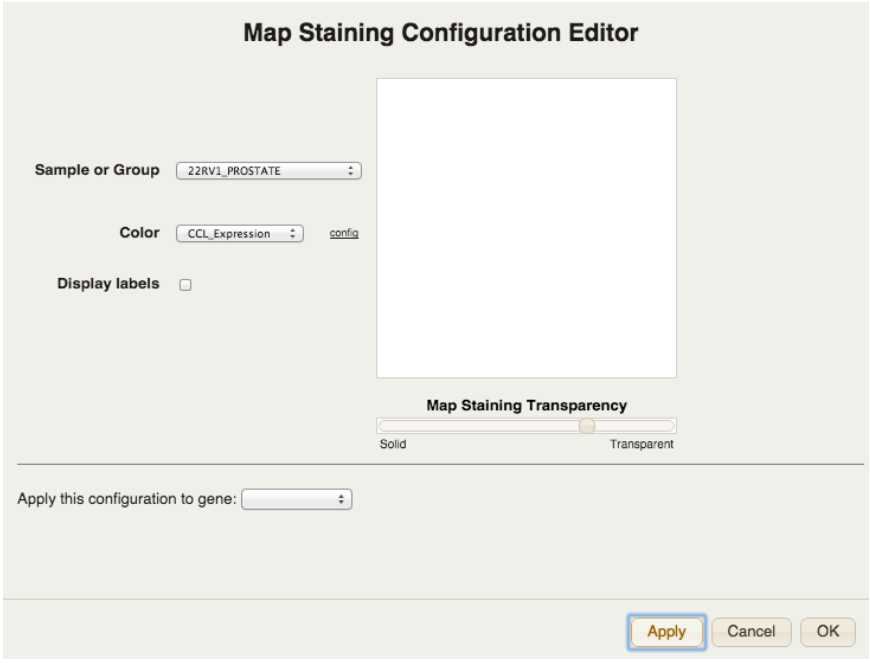


Figure 15: Map staining configuration options.

Displaying mRNA expression data in a form of map staining demonstrates that the color range and distribution patterns are different across various molecular mechanisms depicted on the ACSN.

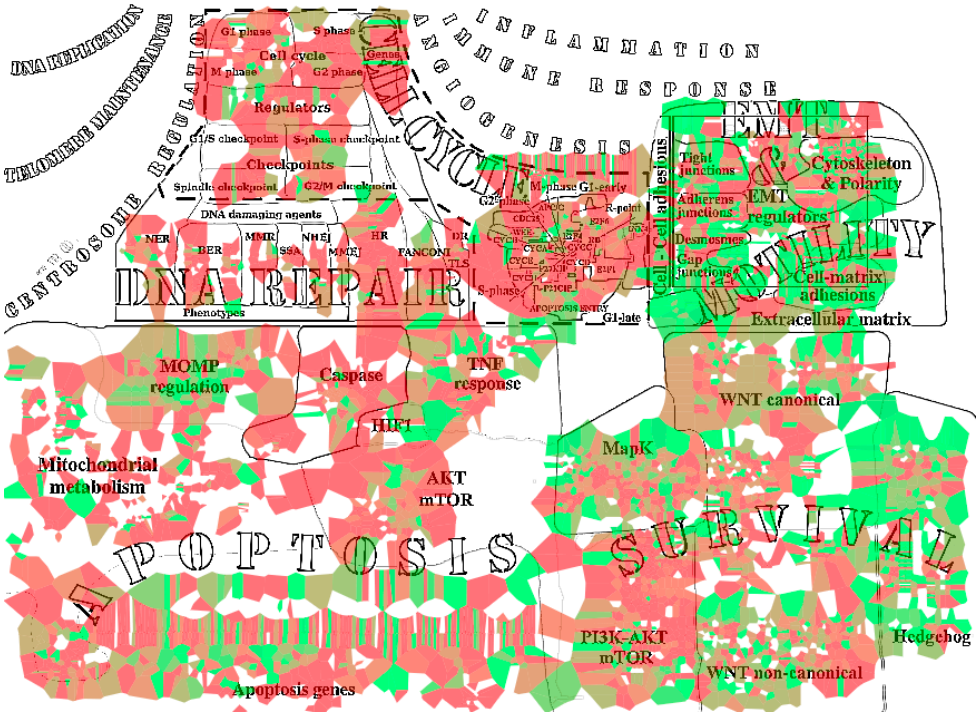


Figure 16: Top level view of ACSN displaying expression data for sample "22RV1_PROSTATE"using the map staining mode. Red-upregulated genes; green-downregulated genes.

Functional analysis

One of most common tasks in analysis of results high-throughput data is correct functional interpretation (functional annotation) of genes (proteins) lists. We have developed and integrated into ACSN some tools for functional analysis, which could be useful for our users.

Gene Enrichment Analysis quantifies the overlap of a gene (protein) list with gene lists related to different functional modules of ACSN (fig 1). The size of the overlap is characterized by a p-value, using the hypergeometric test , that it can happen by random choice of gene names.

To load the gene list of interest, the user can browse his local file system (button "browse") to choose a file, or specify in the field "URL" a web address to retrieve the file from as described in the section 'Loading Data'.

For example, a list of genes contributing most to one of the Independent Components (CIT7) calculated for bladder cancer expression data (Cell Rep. 2014 Nov 20;9(4):1235-45) has been used as the gene list of interest ("cit7.txt").

Remark: Independent Component Analysis (ICA) is a statistical method allowing decomposing a gene expression matrix into a set of independent signals which can be related to various biological factors affecting transcriptome. Each such a signal is represented by a set of weights (contributions) for each gene in the matrix. For the analysis we took one specific component and selected those genes with the largest values of contributions.

The gene list has been imported by selecting the text file from the local directory with the "browse" button a d then specifying the name "CIT7" and type "Gene list", and finally clicking the "Import" button (figure 17).

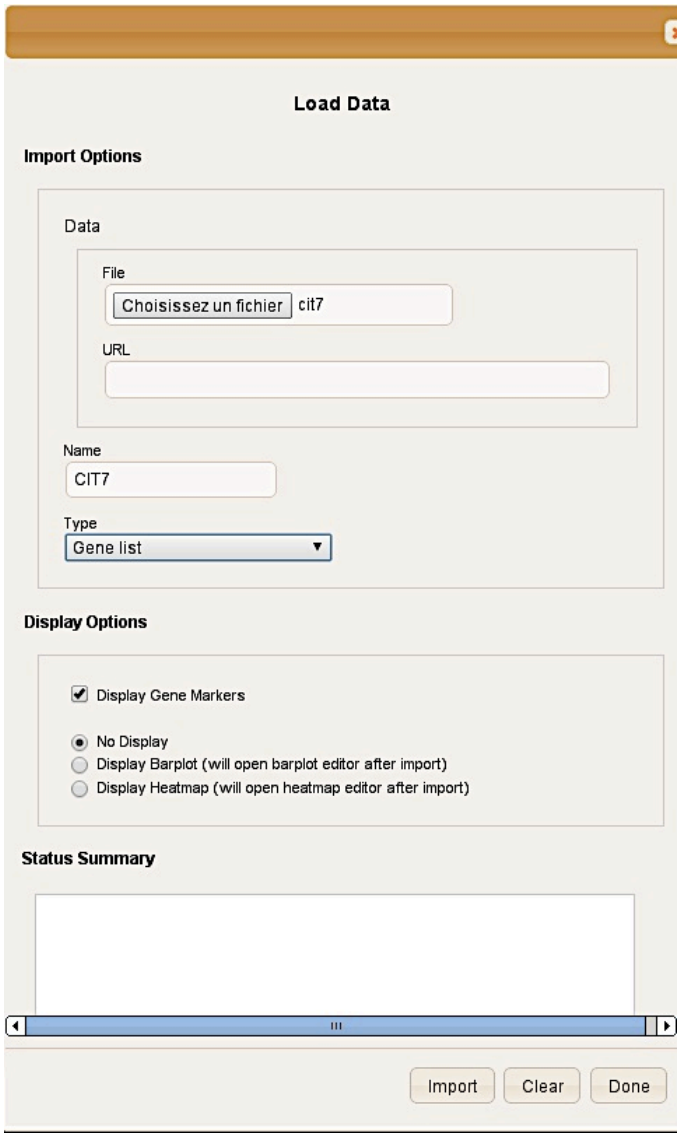


Figure 17: Dialog for loading data for functional analysis. The text file containing the gene list (HUGO names) can be loaded from user's computer or accessed through website and downloaded.

To perform functional analysis, check the button 'Functional Analysis', in the dialog 'Functional analysis' specify an analysis type (for example 'Gene Enrichment Analysis'); choose the datatable with Gene List (CIT7), define the P-value threshold and the background set ('Whole Genome' or ACSN genes 'Genes on the Map'). Check 'Correction for multiple testing' if you want to use such correction and click 'Execute' button.

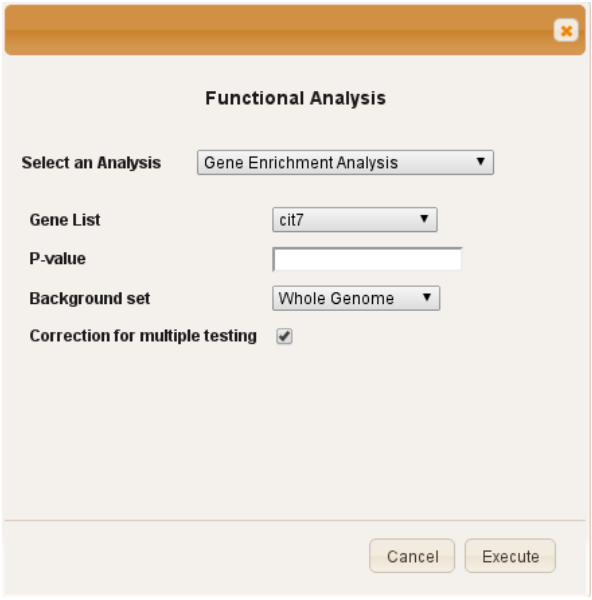


Figure 18: Dialog for loading data for functional analysis. The txt file containing the gene list (HUGO names) can be loaded from user's computer or accessed through website and downloaded.

The result of analysis is displayed in the form of a table (figure 19) demonstrating overlapping sets of genes from the gene list of interest with the functional modules of ACSN.

Gene Enrichment Analysis

Module	Module size	Nb genes in module	p-value (corrected)	Genes
cellcycle:master	256	33	2.2e-15	RRM2 CDC20 CCNB2 CDK1 MELK RAD51AP1 PBK CCNB1 NDC80 TYMS EZH2 CDCA7 MAD2L1 CHEK1 CDKN3 CCNA2 E2F8 BUB1 CDC6 CCNE2 NCAPD2 KPNA2 AURKB FBXO5 CCNE1 PCNA CDKN2C DBF4 MCM10 MCM5 POLA1 DNMT1 BIRC5
dnarepair:master	347	37	8.8e-15	GINS1 CDC20 CDK1 BUB1B TTK CCNB1 PTTG1 MAD2L1 GINS2 MCM2 UBE2T CHEK1 RFC4 BUB1 FANCI CDC6 FEN1 CDC7 BLM CCNE2 PRIM1 ESPL1 AURKB FBXO5 RFC5 CDC45 CCNE1 POLE2 CDC25B PCNA CDKN2C DBF4 CDCA8 RFC3 MCM5 POLA1 BIRC5
dnarepair:S_CC_PHASE	112	19	9.2e-11	GINS1 CDK1 CCNB1 GINS2 MCM2 CHEK1 RFC4 CDC6 CDC7 CCNE2 PRIM1 RFC5 CDC45 CCNE1 PCNA DBF4 RFC3 MCM5 POLA1
dnarepair:SPINDLE_CHECKPOINT	51	13	7.0e-10	CDC20 CDK1 BUB1B TTK CCNB1 PTTG1 MAD2L1 BUB1 ESPL1 AURKB FBXO5 CDCA8 BIRC5
cellcycle:WEE	9	6	4.0e-08	CCNB2 CDK1 CCNB1 CHEK1 CCNE2 CCNE1
dnarepair:M_CC_PHASE	43	9	2.3e-06	CDC20 CDK1 BUB1B CCNB1 MAD2L1 BUB1 ESPL1 FBXO5 CDC25B
cellcycle:CYCLINH	18	6	7.1e-06	CCNB2 CDK1 CCNB1 CCNA2 CCNE2 CCNE1
cellcycle:CYCLINB	19	6	1.0e-05	CDC20 CCNB2 CDK1 CCNB1 CHEK1 CCNA2
cellcycle:CDC25	12	5	1.3e-05	CCNB2 CDK1 CCNB1 CHEK1 CCNA2
cellcycle:P21CIP	30	7	1.6e-05	CCNB2 CDK1 CCNB1 CCNA2 CCNE2 CCNE1 PCNA
cellcycle:APC	24	6	4.5e-05	CDC20 CCNB2 CDK1 CCNB1 CHEK1 CCNA2
dnarepair:G2_CC_PHASE	15	4	7.2e-04	CDK1 CCNB1 CHEK1 CDC25B

Figure 19: Result of functional analysis for gene enrichment of the ACSN modules. The gene list from the Independent Components (CIT7) calculated for bladder cancer expression data has been used to calculate the enrichment on the ACSN modules. First column contains name of functional module of ACSN, second - number of unique genes in the module, third number of overlapped genes, fourth – p-value of overlapping calculated by hypergeometric test, fifth – list of HUGO IDs of overlapped genes.