The University of North Carolina at Chapel Hill

Gillings School of Global Public Health

Departments of

Biostatistics and

Environmental Sciences and Engineering

Dose-response pathway analysis for gene expression: Graphical User Interface

User Manual

Contents

1. About DR Pathway Analysis

Dose-response (DR) modeling is a key component of quantitative risk assessments. When performing microarray studies of transcriptional dose-response, it is of interest to test the coordinated involvement of transcripts (genes) from known biological pathways or functional categories. However, existing gene-set testing methods are not well-suited for DR pathway analysis, which requires careful control of false positive rates and fast DR curve fitting.

DR Pathway Analysis GUI was made for DR analysis, with fast curve fitting procedures and the introduction of a *pathway* dose-response profile. A bootstrap procedure is used to obtain confidence envelopes for the pathway DR profile.

* DR GUI is written in Java, and for the testing phase uses ideas from the Significance Analysis of Function and Expression (SAFE) package (Barry et al., 2005) written in R (R Development Core Team, 2006).

2. Installation

2.1 Requirements

In order to use DR GUI user need to have R and the latest Java installed on your PC. **NOTE: R software should not be older than version 2.15.1.** R software could be downloaded from many mirrors around the world. You could find these mirrors here http://www.r-project.org/. Please select Windows operation system and then “base” as a subdirectory. One can find Java software here: http://www.java.com/en/ .

After you have successfully installed R and Java you need to download DR Pathway Analysis installation package. The most recent version of DR Pathway Analysis is located here http://comptox.unc.edu/resources.html .

Before installation make sure that you have Full Control over your R installation directory.

1. To check that right click on the R folder (usually it's here: C://Program Files/R).
2. Select "Properties" from menu that will appear.
3. Navigate to "Security" tab.
4. Click "Edit" button.
5. Select Users in the list of Group or user names.
6. Check "Full control" box under Allow.

7. Click OK.

2.2 Installation process

Please unzip DR GUI archive to any place on your PC after you have downloaded **DRGUI.zip** file. Then open DRGUI folder and double click on **DRGUIInstaller.jar** file. The Installation window should appear.

DR Pathway installer will look for installed versions of R inside default locations and if you have administrative rights on your PC, Java and R properly installed into default directories you'll be able to use the "Automatic installation" function.

If you don't have administrative rights on your PC you will need to install R into any folder under your user account where you have rights to read/write/modify/execute files. In order to install DR GUI you will need to use the "Manual installation" option. One will need to click "Select file" button and select the folder where R was installed. The next step is click "Manual installation" button.

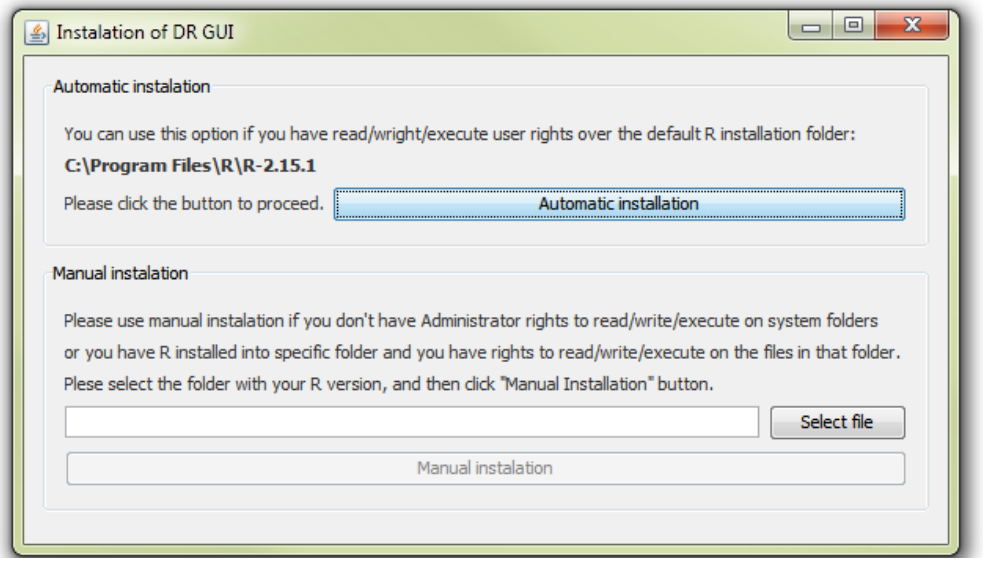


Figure 1. Installer main screen

DR Pathway GUI will try to download and install all necessary R libraries from the Internet, so Internet connection is required at this stage. During the installation you will see the progress window and occasionally the R progress window.

Figure 2. Progress window

If installation was successful you will see a successful install message displayed. Installer will place startup and uninstall link inside **R\DRGUI** directory in case it's unable to create startup menu items and Desktop shortcuts for DR Pathway GUI.

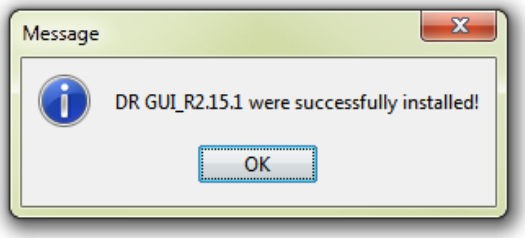


Figure 2. Installer success message

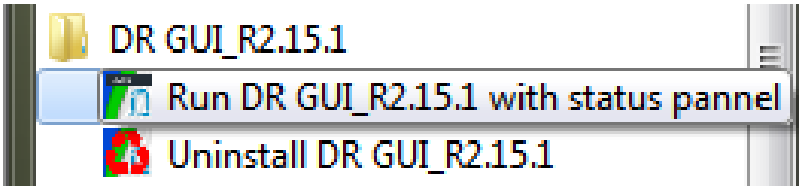


Figure 4. DR GUI folder in Windows Start menu

If installation fails with error message below, please follow installation instructions for User Access Control below.

2.2.1 Installation under Windows Vista

As Windows Vista has UAC (User Access Control) which doesn't allow copy files in Windows you should undertake some additional steps.

1) Before installation make sure that you have Full Control over your R installation directory.

8. To check that right click on the R folder (usually it's here: C://Program Files/R).

9. Select "Properties" from menu that will appear.
10. Navigate to "Security" tab.
11. Click "Edit" button.
12. Select Users in the list of Group or user names.
13. Check "Full control" box under Allow.
14. Click OK.

2) Double click on **DRGUIInstaller.jar** file.

3) Click "Install" button to proceed

4) If you will see message "Cannot create Start menu! Shortcuts will be created in R directory" the program will not create Desktop shortcut and Start menu items.

But you'll find the shortcuts inside your R directory (usually it's here: C://Program Files/R/SAFEGUI).

***You can avoid changing user rights by switching off UAC before installation. <u>But we do not recommend you to do this.</u>**

3. Main window

3.1. DR Pathway Analysis Settings

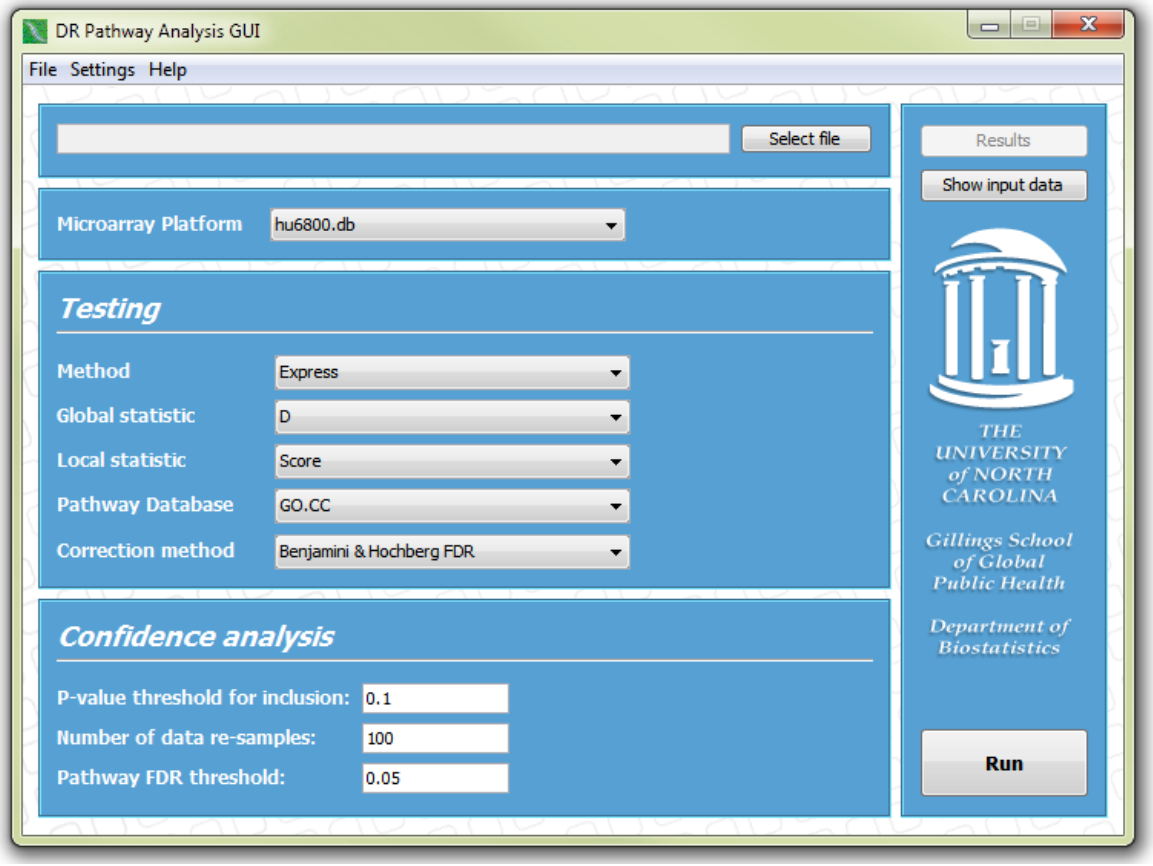


Figure 3. DR Pathway Analysis main screen

On Figure 3 you can see the DR Pathway Analysis main screen which allows you to set input parameters for your data analysis.

First of all you need to set a path to the input data file. To do that, click on the "Select file" button. File browsing dialog should appear. Browse to your data file. Click "Open".

Select following parameters when you have selected input data file:

- **Microarray platform** - name of a Bioconductor annotation package to build gene pathways that corresponds to the array platform that was used to collect the data.

TESTING OPTIONS

"Testing" refers to testing pathways for a significant dose-response relationship among numerous genes in the pathway

- **Method** - Currently "express" is the only option. This is a mathematical approximation to permutation, described in Zhou et al. (2013). Other options may be added in future releases.

- **Global statistic** - Specifies the global statistic for a pathway (gene set or category). The statistic D is currently the only option. It is a competitive statistic that compares the mean squared score statistic within the pathway to the mean squared score statistic among the remaining genes.

- **Local statistic** - Specifies the gene-specific statistic - “Score” is the current default.

- **Pathway Database** - "GO.MF", "GO.BP", "GO.CC" specify the three Gene Ontologies. "GO.CC" will form categories from Cellular Compartment ontology, while GO.BP and GO.MF will work with Biological Processes and Molecular Function, respectively. It is important to note that in the hierarchical structure of the GO vocabularies, a gene category is generally thought of as containing the set of genes directly annotated to a term, and also to any terms beneath it in the ontology.

- **Correction method** - Specifies the method for computing false positive error rates, accounting for multiple comparisons across numerous pathways. A Bonferroni, ("Bonferroni FWER"), Holm’s step-up ("Holm FWER"), and Benjamini- Hochberg step down ("Benjamini & Hochberg FDR") adjustment can be specified.

CONFIDENCE ANALYSIS OPTIONS

"Confidence analysis" refers to quantifying uncertainty in the *pathway* dose-response curve. This is performed only for pathways meeting the "Pathway FDR threshold." The pathway dose response curve is created by

(i) selecting genes meeting "threshold for inclusion." For each of these genes, a four-parameter logistic curve is fit to represent the gene. For the included genes, bootstrap resampling of the gene expression and dose values is performed. In contrast to the testing phase, the confidence phase treats the entire expression profile and dose for each sample as a combined unit for bootstrap resampling.

(ii) For the observed data and each resample, a pathway dose response curve is created either by a) summing the individual dose-response curves (rescaled to the unit interval) or b) fitting a single logistic curve to all of the genes, which have been rescaled to the unit interval.

- **P-value threshold for inclusion** - gene-specific p-value threshold to select genes that are to be included in the confidence analysis. Default is 0.1.

- **Number of data re-samples** – An integer value to specify the number of bootstrap resamples performed. The entire curve fitting procedure to create the pathway dose-response curve is performed for each bootstrap.

- **Pathway FDR (or Bonferroni/Holm) threshold** - The criterion for deciding which pathways should be included in the confidence analysis. Setting this value too high will result in performing the confidence analysis step for a large number of pathways, and thus will take a long time for computations to complete. By default, value is 0.05.

BUTTONS

- **Results** *(button)* – will show the results window on click. Note that this button became enabled after you run DR Pathway Analysis.

- **Show input data** *(button)* – will show your input data file in a table format.

- **RUN** *(button)* – runs DR Pathway Analysis method with selected parameters

3.2. Main menu

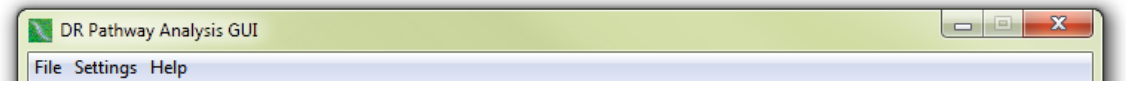


Figure 6. DR Pathway Analysis menu

On the main DR Pathway GUI window you will see a top positioned menu with three main items – "File", "Settings" and "Help".

3.2.1. File Menu

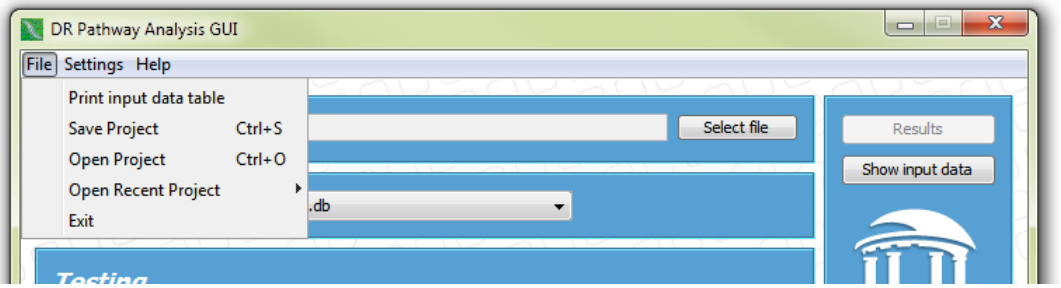


Figure 7. DR Pathway Analysis File menu

- **Print input data table** – prints the data in table format from selected data file.
- **Save Project** – saves current project.
- **Open Project** – allow user to load previously saved project.
- **Open Recent Project** – gives user a list of recent projects to load.
- **Exit** – closes DR Pathway Analysis.

3.2.2. Settings Menu

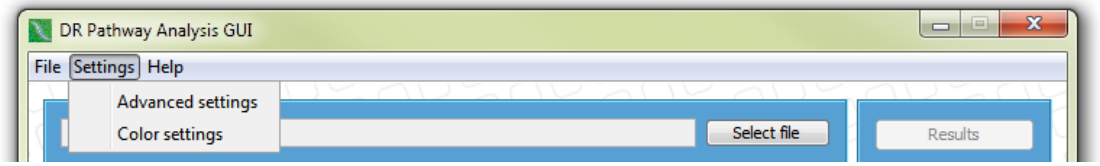


Figure 8. DR Pathway Analysis Settings menu

- **Advanced settings** – will open Advanced Settings window
- **Color settings –** will open Color Settings window

3.2.2.1. Advanced settings

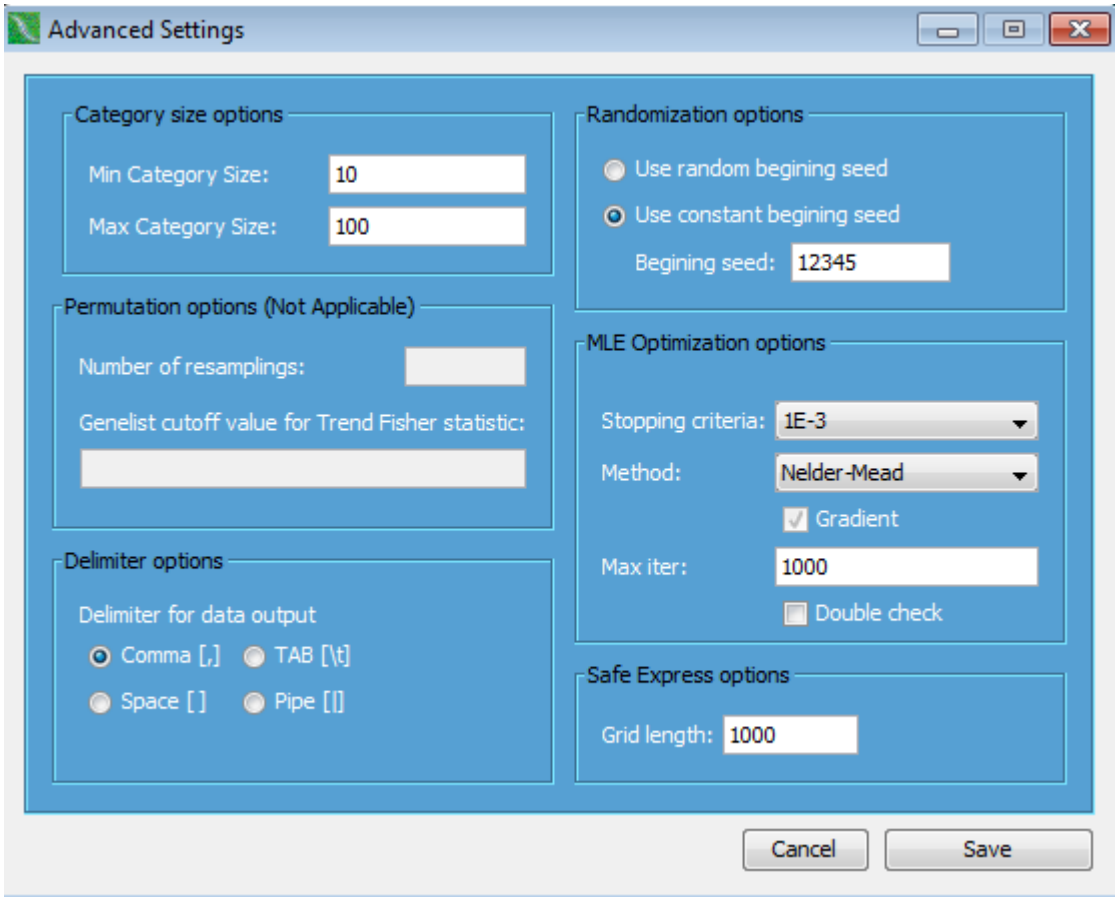


Figure 9. DR Pathway Analysis Advanced Settings window

- **Category size options** – allows user to set minimum and maximum category (pathway) size for the pathways to be included in the results.
- **Randomization options –** allows user to set random seed or constant seed for the bootstrap resmapling. With constant seed option selected you'll need to input sthe eed value, and expect that you will obtain reproducible results.
- **Permutation options –** these options are intended for future versions of DR Pathway.
- **MLE Optimization options** – allows user to set options for MLE optimization for logistic curve fitting
- **Delimiter options -** allows users to set delimiter for the tables which are produced when user tries to save the result data table. Usually delimiter is tab (“ ”), space (“ ”), comma(“,”) or pipe (“|”).
- **Safe express options** – allows user to set “grid length” to be used for approximating permutation p-values. Higher grid values produce more accurate p-values. The default of 1000 should suffice unless the sample size is large (>200).

3.2.2.2. Color settings

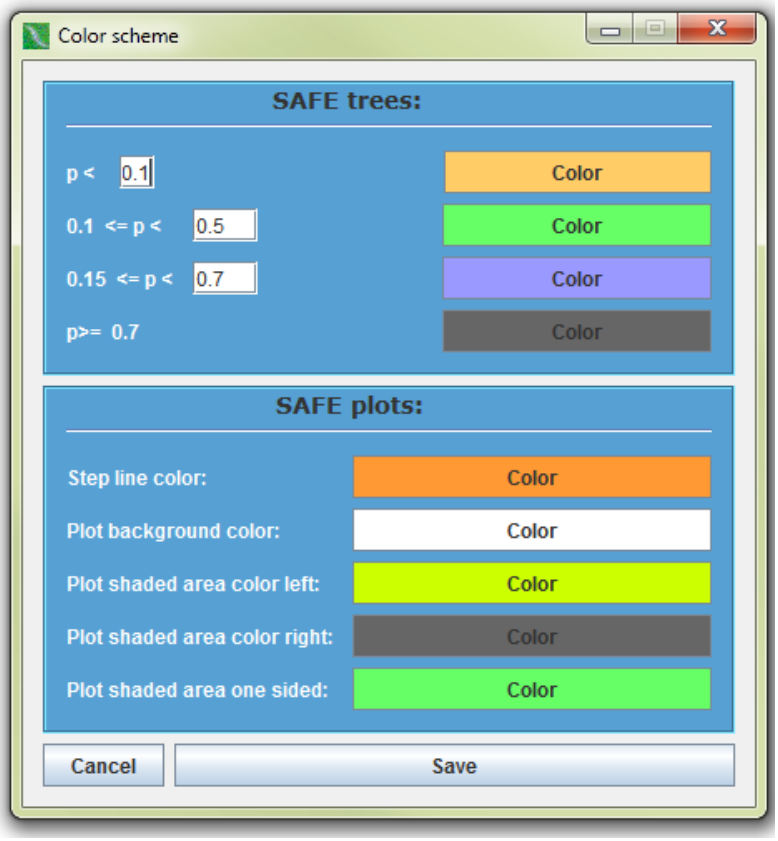


Figure 10. DR Pathway Analysis Color Settings window

3.2.3. Help menu

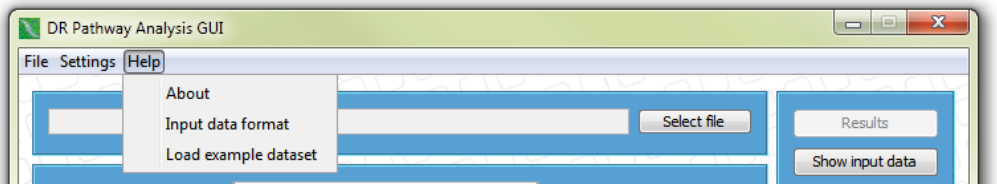


Figure 11. DR Pathway Analysis Color Settings window

- **About** – contains information about the DR GUI software
- **Input data format** – shows examples of the input data file structure.

- **Load example dataset** – loads example dataset of Rat expression data, on the RAE230 platform, for demonstration purposes.

4. Input Data Format

Data in the input data file should be separated using **"," (comma)** delimiter. File extension usually should be .csv. Input data file should be in the following format:

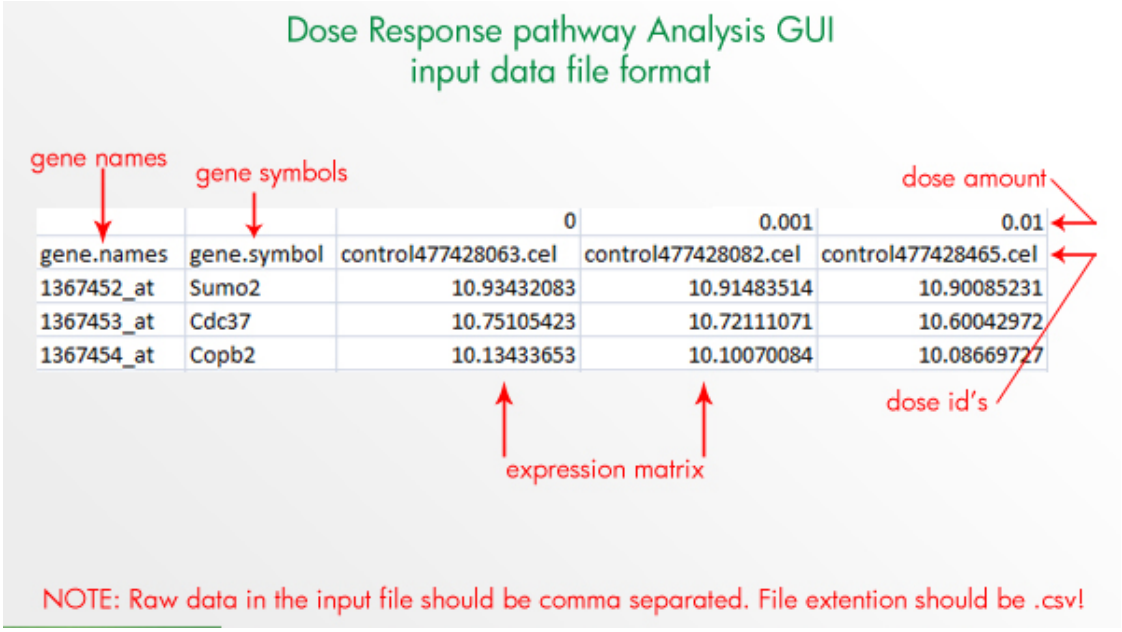
First row – sample names,

Second row through n-row – expression data.

Dose values should be on an original meaningful scale, but will be analyzed on the log scale. To avoid difficulties with log(0), zero values will be transformed on the log scale to be an increment *x* lower than log(lowest positive dose), where *x* is the average increment between successive log(positive dose) values.

The expression data should be in the form of an mXn matrix, where appropriate normalization and other pre-processing steps have been taken. It should be noted that in the current version of DR Pathway Analysis, missing values are not allowed in the expression data, and must be imputed prior to analysis.

Example of the input data file:



5. Running DR Pathway Analysis

After you click "RUN" button main window will disappear and Status window will appear.

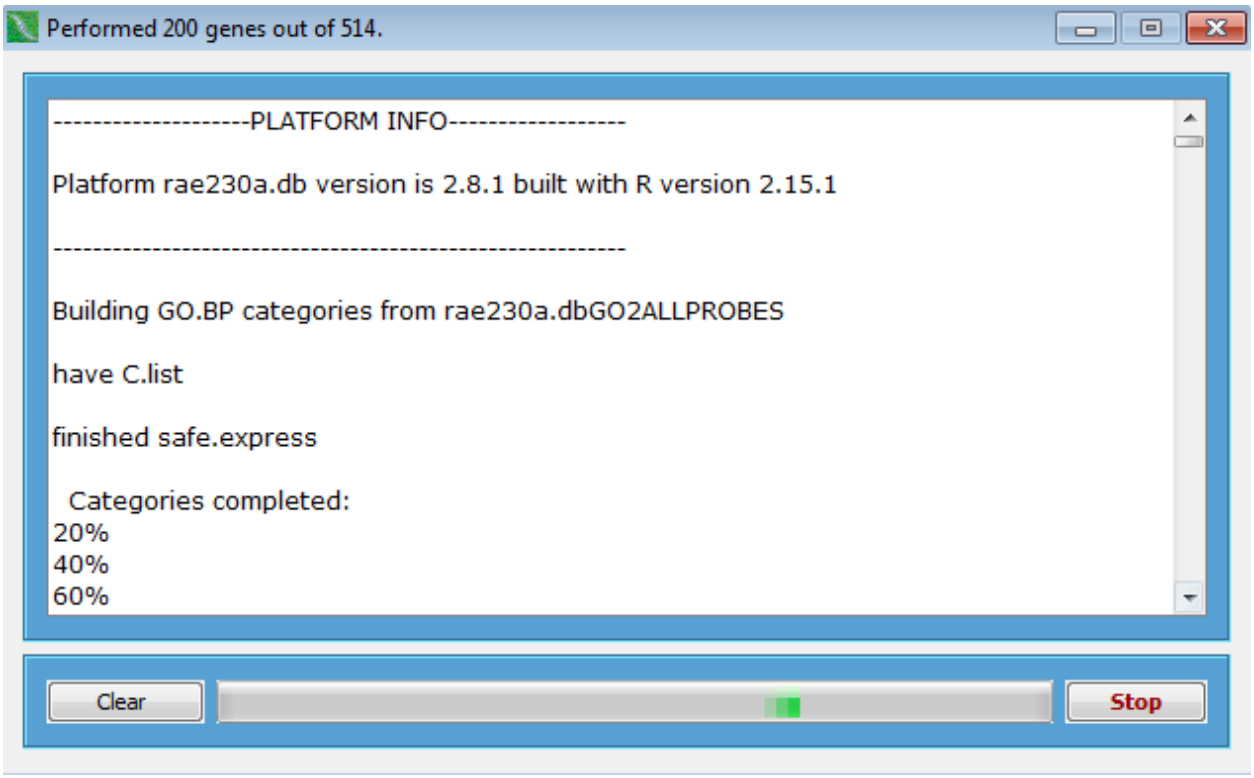


Figure 12. DR Pathway Analysis Progress window

You can see how Status window looks like on Figure 6. In this window DR Pathway Analysis displays status of the current process. You can halt execution by clicking **"Stop"** button. You can clear status window by clicking **"Clear".**

6. Results

Results window will appear after DR Pathway GUI finishes with all permutations.

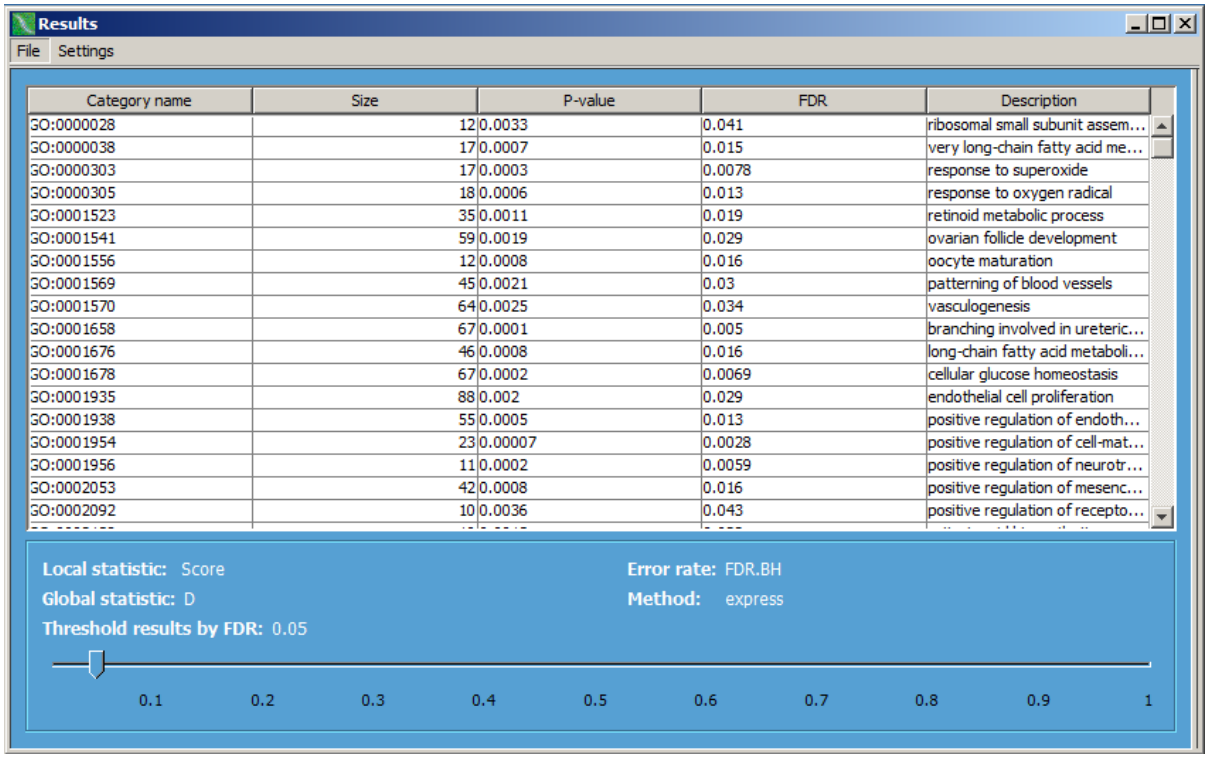


Figure 13. DR Pathway Analysis Results window

Result window is divided by two logical parts – Categories part and Specific category part. Categories part consists of the categories table, which we will describe more precisely below, and Analysis parameters, such as:

- **Local statistic** used for analysis
- **Global statistic** used for analysis
- **Error rate**
- **Method** (Express is the only option)
- **Threshold results by FDR**. User could change this value by dragging the slider

6.1. Categories (pathways) table

The "Categories" table consists of categories that attain a certain level of significance. Figure 14 shows significant results for categories that have empirical p-values <= 0.25. For each category, the category name, and number of annotated genes in the dataset is displayed along with the p-value, FDR value and Category description.

Specific category information including genes contained in selected category will appear when user clicks on any category. Processing times for specific results depend on category size. Results for smaller categories will appear immediately, while results for bigger categories will take some time but only the first time a user click on the category.

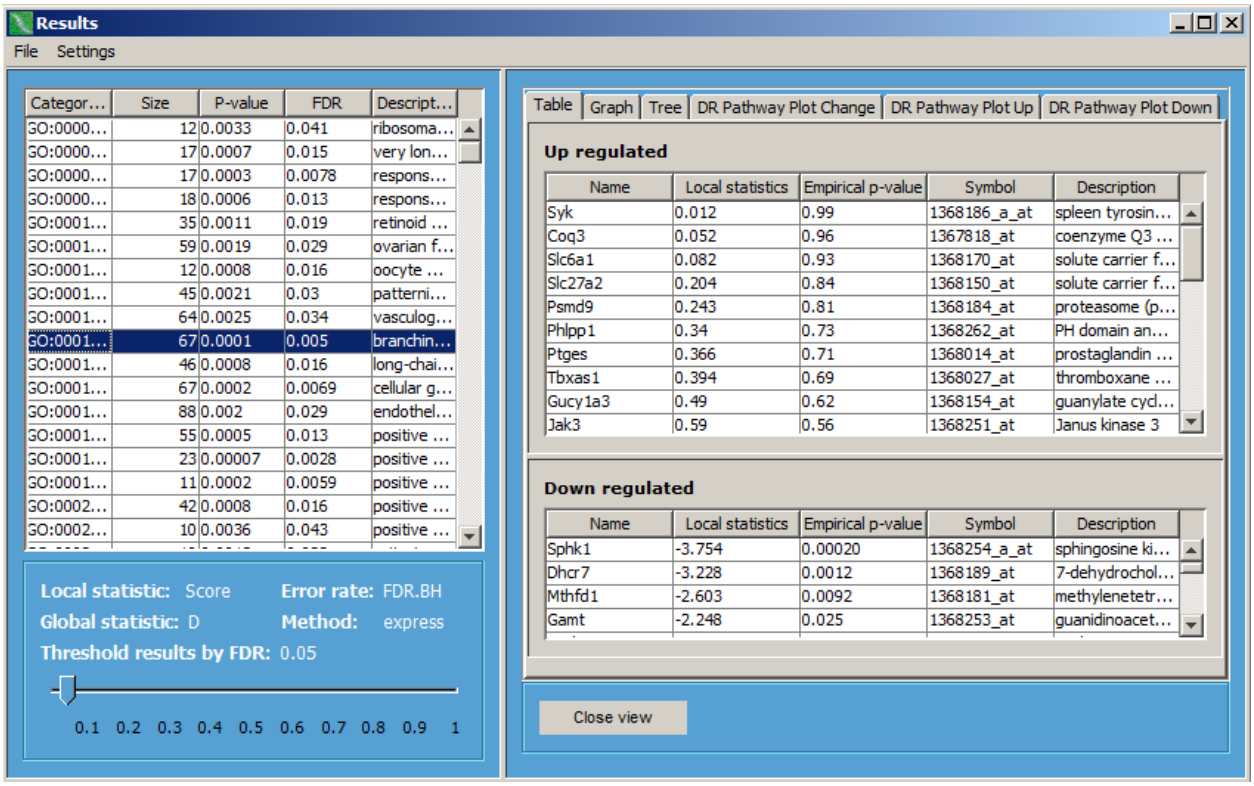


Figure 14. DR Pathway Analysis Results window

6.2. Genes table

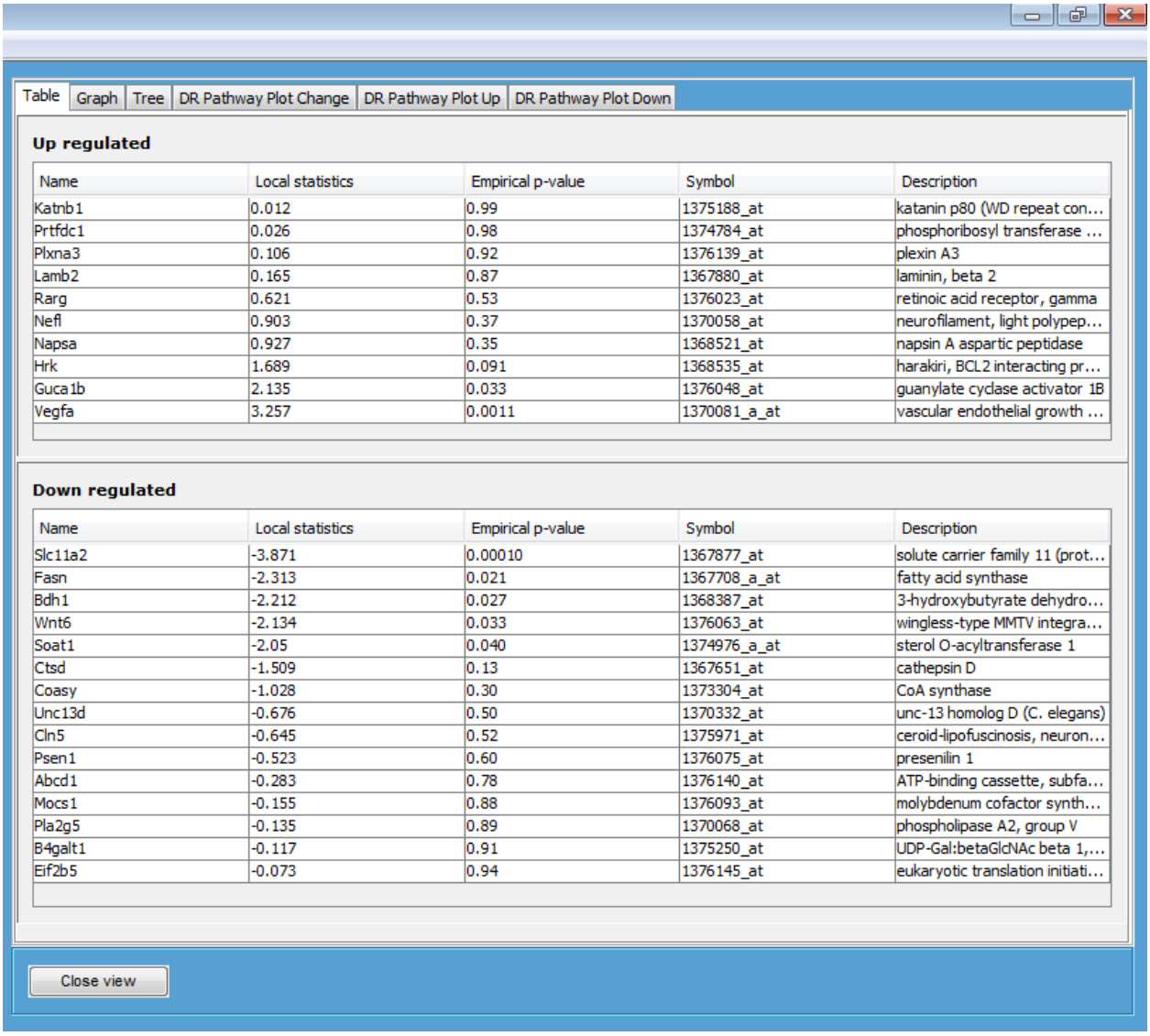


Figure 15. DR Pathway Analysis Results window (Genes table)

The "Genes" table is located under the Table tab and could consist of one or two tables. On Figure 15 you can see *Up regulated* genes table versus *Down regulated* genes table. Up regulated are genes with local statistic values more than 0. Down regulated are genes with local statistic value under 0.

Each genes table contains the following information:

- Gene name
- Gene local statistic
- Empirical p-value
- Gene symbol
- Gene description

6.3. Graph

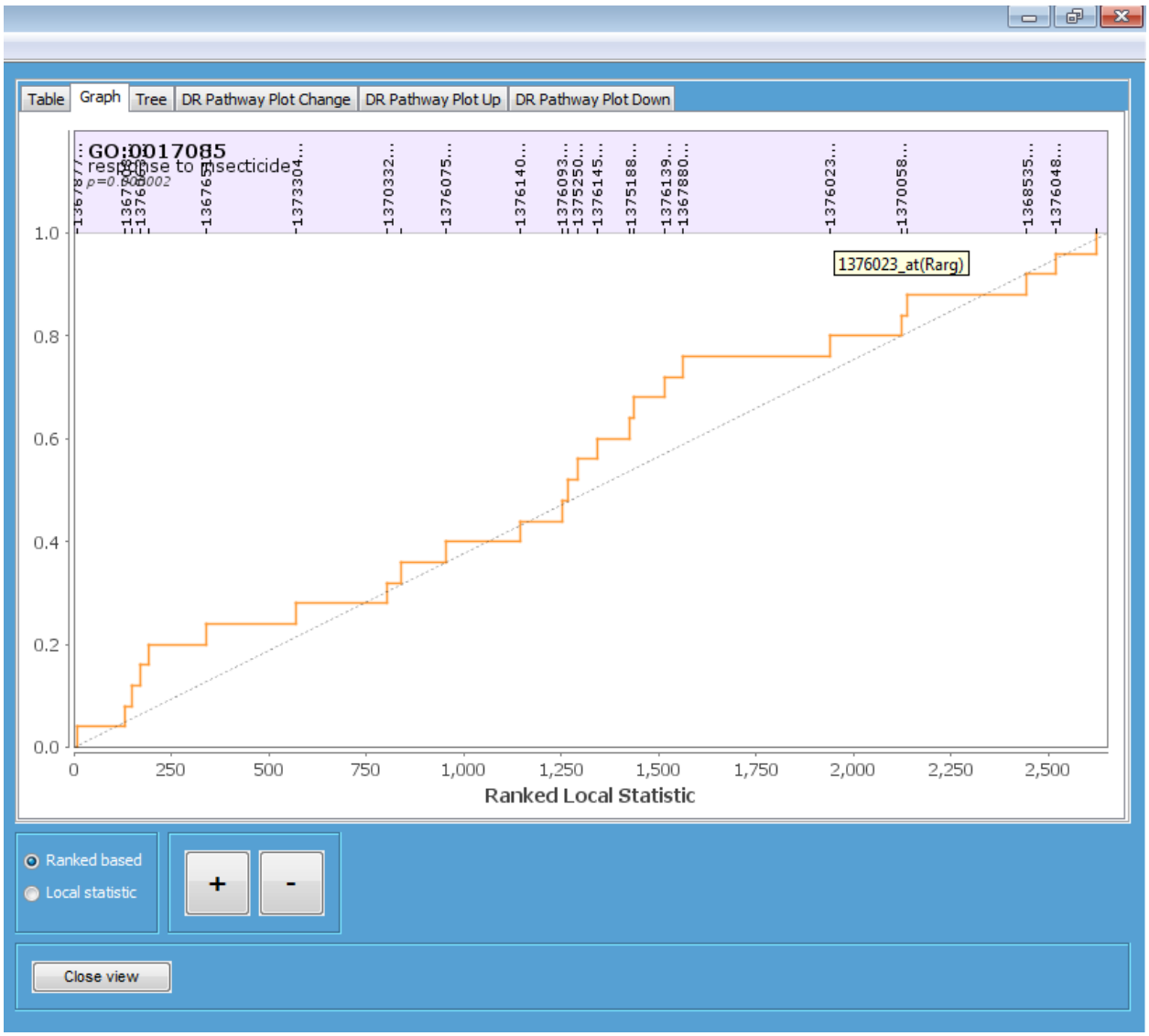


Figure 16. DR Pathway Analysis Results window (Genes plot)

The graph tab contains a plot for a single pathway. We have proposed that the differential expression of genes be plotted as a SAFE-plot (Barry et al., 2005). SAFE-plots show the cumulative distribution function (CDF) for the ranked local statistics from a given category (solid line). A significant category will have more extreme associations to the response of interest than its complement, resulting in a rightward, leftward, or bidirectional shift in the CDF away from the unit line (dashed line). The shaded regions of the plot correspond to the genes that pass a nominal level of significance (empirical p-values <= 0.05 by default). A user can select SAFE-plot representation by switching from ranked local statistic to simple local statistic.

To zoom in/out the plot a user can use "+"/"-" buttons or by rotating mouse scroll. To move zoomed plot, push and hold left mouse button and drag the plot right or left.

If you click on the gene symbol, DR Pathway Analysis will open extended gene information in your default browser.

Figure 17. DR Pathway Analysis Results window (Genes Plot)

6.4. Tree

The Tree tab utilizes the structured vocabulary whereby genes are annotated in GO from broad to narrow levels of classification in a directed acyclic graph (DAG). As such, many categories are highly related in their gene membership, and visualizing results across the ontology can be useful in ascertaining the relationship among multiply significant categories.

By default, nodes with unadjusted p-values less than 0.1 are drawn in brown; less than 0.5 are drawn in green; and less than 0.7 are drawn in purple, more than 0.7 in grey, all other categories are colored in cyan. User-defined cutoffs for these colors can be specified using the menu **Options->Color scheme**.

By clicking on any tree, the node graph will be refreshed and the selected node will become the top of the tree showing child categories. You can easily undo that operation by pushing **"Undo" button.**

By pushing the **"Big tree"** button, a new window will open showing the tree containing all the categories in selected package.

By checking **"Show all children"** DR Pathway Analysis will show all the categories which are not in threshold results.

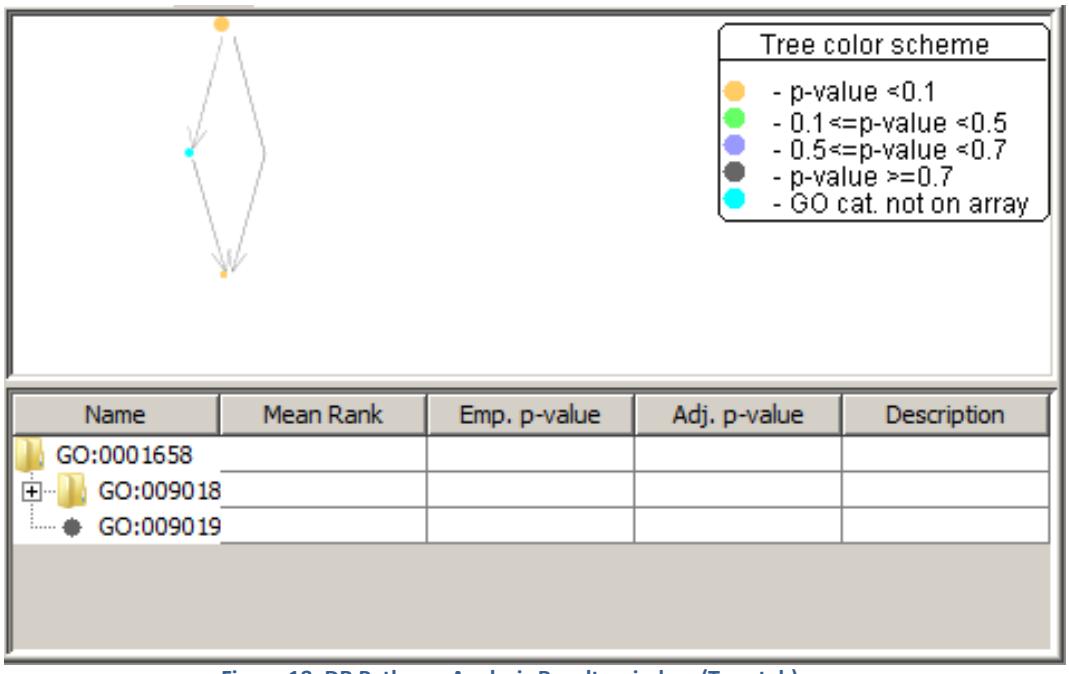


Figure 18. DR Pathway Analysis Results window (Tree tab)

6.5. DR Path Detector Plot tabs

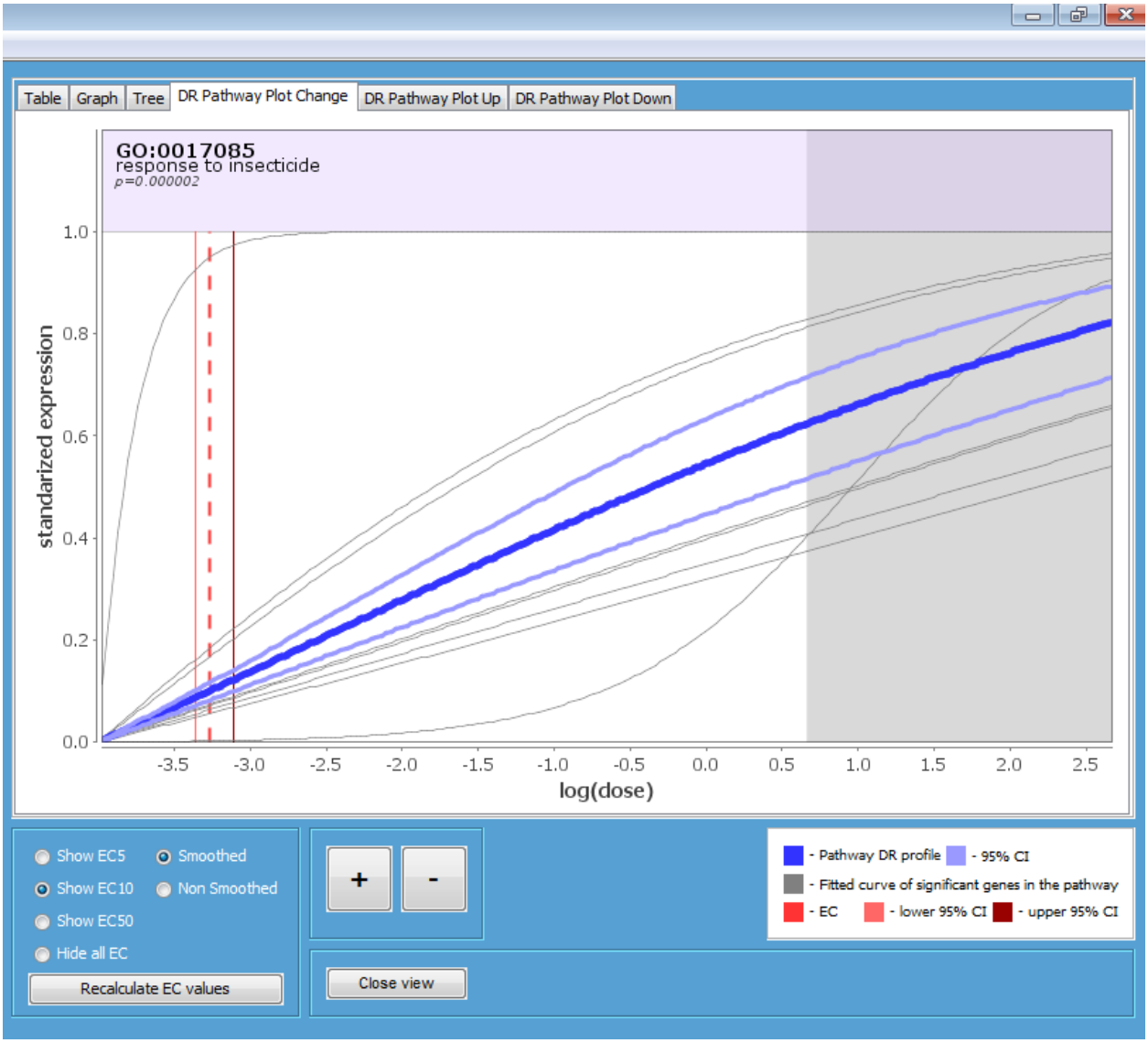


Figure 19. DR Pathway Analysis Results window (DR Path Change tab)

DR Path Detector (Change, Up, Down) Plot tab contains information and graphical representation of the pathway dose response profiles. Users can see the EC5, EC10 and EC50 values on the table and on the plot by selecting one of the radio buttons. The EC values and 95% confidence intervals are in red. Clicking on the EC buttons will also cause them to appear in the table of all significant pathways.

The shaded areas indicate extrapolation of results beyond the actual range tested. As with any extrapolation, these should be interpreted with caution.

Zooming ability is available and behaves in the same manner as Gene plots.

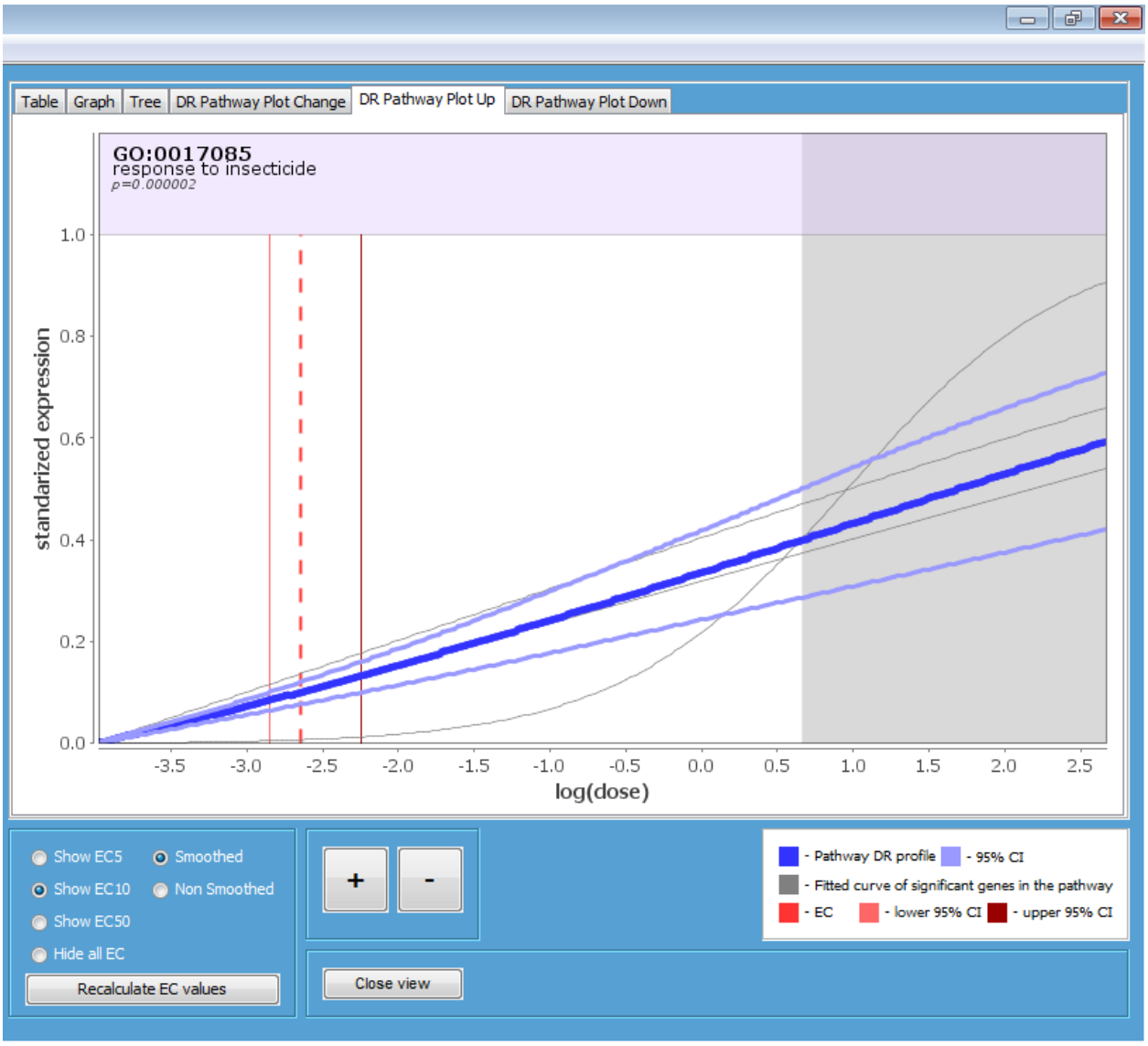


Figure 20. DR Pathway Analysis Results window (DR Path Up tab)

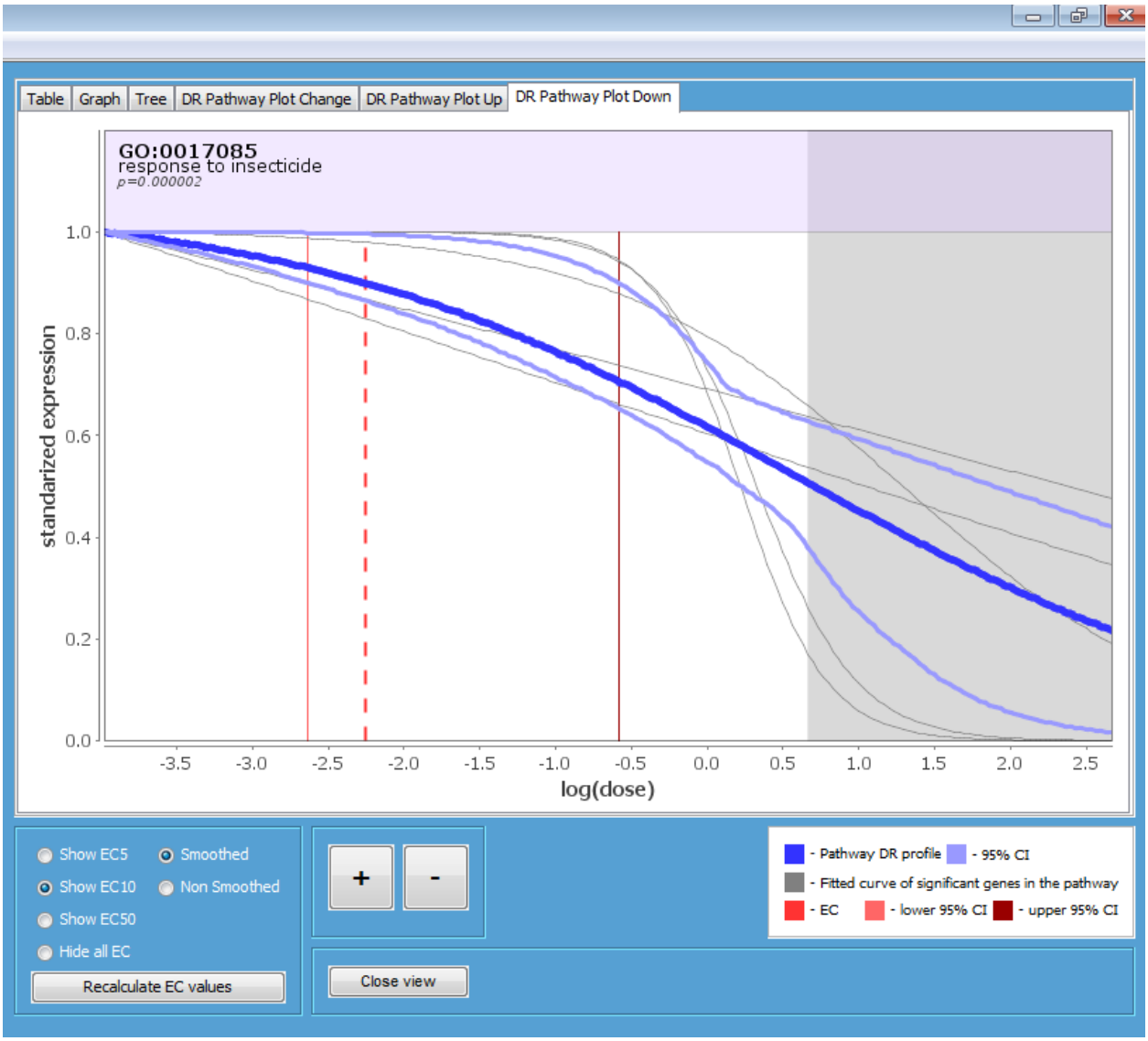


Figure 20. DR Pathway Analysis Results window (DR Path Down tab)

Results can be printed or saved from the Results window.