

Toxygates/AdjuvantDB

User Guide

2019 Update

Rodolfo Allendes
Johan Nyström-Persson

This is a task-oriented guide to using Toxygates, an interactive platform for omics data access and analysis. Toxygates currently contains Open TG-GATEs and adjuvant data, and users may also upload and analyse their own data. In this document we walk through the core analysis functions in depth, in the hope that users will also gain a sufficient intuitive understanding to explore the system freely.

© Toxygates authors

The National Institutes of Biomedical Innovation, Health and Nutrition, Japan
rallendes@nibiohn.go.jp, johan@lifemetics.co.jp

Contents

Contents	ii
1 Basic Concepts and Data Viewing	1
1.1 Displaying expression data	1
1.2 Gene sets	2
1.3 Ranking compounds	3
2 Sample search	5
3 Clustering and heat maps	7
4 miRNA-mRNA network analysis	9
4.1 Choosing datasets	9
4.2 Defining sample groups	9
4.3 Choosing miRNA-mRNA associations	11
4.4 Dual table analysis	11
4.5 Visual network analysis	12
5 Uploading your own data	17
6 Other tasks	21
6.1 Orthologous data inspection	21
6.2 Import/export to/from InterMine instances	22
6.3 Enrichment testing	22
6.4 Downloading data	22
6.5 Viewing pathways and other annotations	23
6.6 Inspecting sample attributes and biochemical data	24
6.7 Viewing pathologies	25

Chapter 1

Basic Concepts and Data Viewing

Toxygates is split into several different **screens**, such as *Start*, *Sample groups*, *View data* and so on. The navigation links to move between screens (if they are available) are located at the top of the page.

1.1 Displaying expression data

To inspect data (such as Open TG-GATEs) in Toxygates, it is necessary to first define **Sample groups**. After defining the data source, it is possible to select the **species**, **test type**, **repeat type** and **tissue** you are interested in. This can be done at the top left corner of the *Sample groups* screen, as shown in Figure 1.1. Then, select the compounds you are interested in from the compound list.

When you have selected one or several compounds, you can select **time** and **dose** combinations on the right hand side, as shown in Figure 1.2. In Open TG-GATEs, usually, each time and dose combination will correspond to 3 treated and 3 control samples (in some cases less). Check the boxes you wish to inspect and save the group. The selected samples will be saved together and later averaged when displayed. If you wish to contrast two different time points or dose levels, etc, you can save them as two groups with different names.

A **group name** will be automatically suggested, but you may enter another name if you wish. Saved groups will be stored in your browser, and will remain the next time you use Toxygates.

Once at least one group has been defined and saved, it is possible to proceed to the *View data* screen. Here you can see the data you have selected. Your sample groups will correspond to columns in the table, and genes (probes) will correspond to rows.

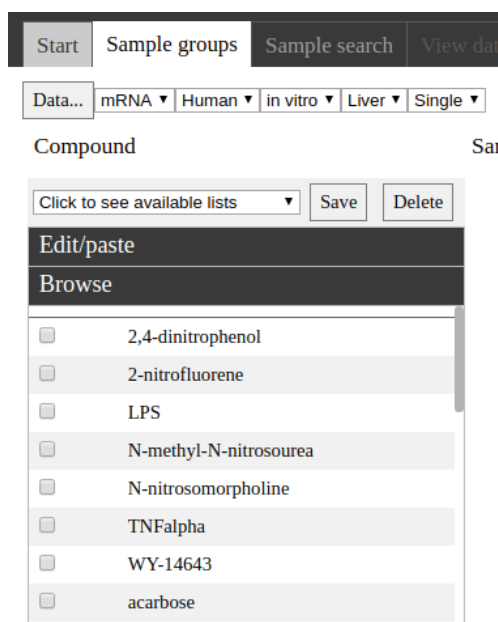


Figure 1.1: List of the first available compounds in the Open TG-Gates dataset.

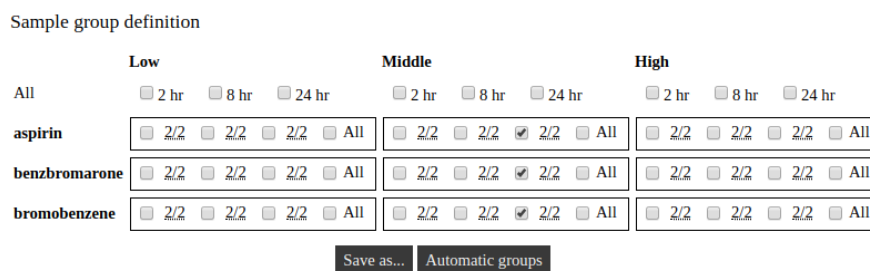


Figure 1.2: Example time and dose selection for three different compounds found in the Open TG-Gates dataset.

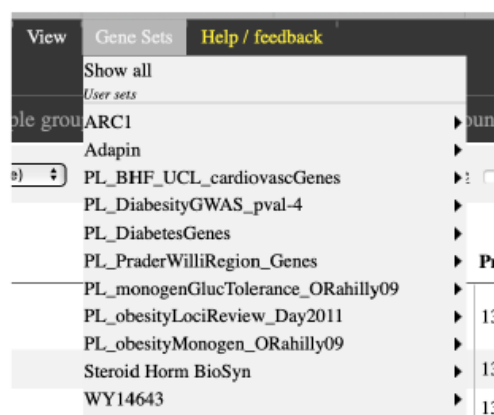
Genes may be *filtered* by clicking on the filter icon next to each column header. Figure 1.3 shows the filter modification dialog; here you can enter a numerical threshold for the column. This will exclude all genes that do not pass the threshold. A blue icon indicates that the filter is active.

1.2 Gene sets

The currently displayed genes may be saved as a *gene set* and given a name. A size limit applies - currently, no more than 1000 probes can be saved in this way. Such a gene set can also be edited before being saved (for example, by adding GO terms or pathways of interest). Saved gene sets may be selected

Figure 1.3: *Filter definition dialog box.*

from the **Gene sets** menu, which is accessible on the *View data* screen. If you have previously carried out any clusterings and saved them, you will also find those here. The gene sets will be stored in your browser, and remain the next time you return to Toxygates. If you have selected a gene set from this menu and wish to return to a display of all genes, you may select "Show all" at the top of the **Gene sets** menu. To see all genes, you may also need to clear any column filters you have activated.

Figure 1.4: *A list of previously saved gene sets.*

1.3 Ranking compounds

You can also perform a **Compound ranking** on the screen with this name. This function allows you to rank compounds according to their behaviour with respect to a set of genes, as seen in expression **time series**. To use this function, first go to the *compound ranking* screen (shown in Figure 1.5), then select the organism, tissue and so on that you are interested in (similar to

the sample groups screen). Then type in, in the field "Gene/probe", the gene that you are interested in, for example *Gss* for glutathione synthetase. You can enter any number of genes (within reasonable limits). Many match types (ranking criteria) are available; the default is "total upregulation", which ranks the compounds according to how much they upregulate the gene. (For reference, please use the **Help** menu and select "Help for this screen" while you are performing compound ranking).

Compound	Score	Gene/probe	Match type	User ptn.	Ref. con
<input type="checkbox"/> butylated hydroxyanisole	1.099 (1)	<input checked="" type="checkbox"/> <i>gss</i>	Total upregulation		2,4-din
<input type="checkbox"/> phorone	1.083 (2)	<input checked="" type="checkbox"/> <i>gclc</i>	Total upregulation		2,4-din
<input type="checkbox"/> 2-nitrofluorene	1.073 (3)	<input type="checkbox"/>	Total upregulation		2,4-din
<input type="checkbox"/> diethyl maleate	1.066 (4)				

Figure 1.5: *The compound ranking screen.*

When you have specified the rules, click the "Rank" button. When the ranking is complete, you will see the results in the compound list on the left. Each compound will have a number assigned to indicate its relative score. (This number is a relative measure only and has no biological meaning.) By clicking the chart icon next to each compound name, you can see the various time series for the relevant genes for that compound. The dose level that best matches your rules will be highlighted in blue.

In addition to ranking by time series, it is now also possible to rank by **dose series**. In the former case, the time point is the independent variable, and gene expression against time is considered in order to identify compounds and dose levels of interest. In the latter case, the dose level is the independent variable, and gene expression against dose is considered in order to identify compounds and time points of interest. In other words, the roles of the time and dose parameters are exchanged.

This concludes the discussion of basic concepts. Many screens in Toxygates also have a help function, which will give you additional information. You can access it via the menu item "Help for this screen" on the **help** menu.

Chapter 2

Sample search

It is possible to search for samples (and by extension compounds or other sample treatments) in terms of how they affect biological measurements on the *Sample search* screen, shown in Figure 2.1. As on the column definition screen, first the appropriate sample space must be selected, e.g. mRNA/rat/in vivo/liver/single dose. Then a search may be performed inside this space.

The screenshot shows the 'Sample search' interface. At the top, there are navigation tabs: Start, Sample groups, Sample search (active), View data, Compound ranking, Pathologies, Sample details, and My data. Below the tabs, there are dropdown menus for 'Data...' (mRNA), 'Rat', 'in vivo', 'Liver', and 'Single'. The search criteria are defined in three rows: 'Kidney weight left (g)' with a 'Low' condition, 'Creatinine (mg/dL)' with a 'High' condition, and an 'Undefined' condition. The search is performed using 'Individual samples' (selected) and 'Sample units'. The results show 58 results found. A table displays the search results with columns for Compound, Dose level, Exposure Time, Sample ID, Kidney weight left (g), and Creatinine (mg/dL). The table lists five rows, with the last two rows checked.

Compound	Dose level	Exposure Time	Sample ID	Kidney weight left (g)	Creatinine (mg/dL)
<input type="checkbox"/> 3-methylcholanthrene	Low	24 hr	003017916022	0.85	4.5
<input type="checkbox"/> 3-methylcholanthrene	Middle	24 hr	003017916025	0.82	4.5
<input type="checkbox"/> 3-methylcholanthrene	High	24 hr	003017916028	0.83	4.56
<input checked="" type="checkbox"/> acetazolamide	Low	24 hr	003017262017	0.87	0.2
<input checked="" type="checkbox"/> acetazolamide	Low	24 hr	003017262019	0.98	0.2

Figure 2.1: *Sample search screen.*

A search is specified by indicating parameters of interest and selecting whether they should be low, high, normal range, below a given limit or above a given limit. Currently, “low” and “high” are defined as more than 2 standard deviations from the mean, and normal range as being within 2 standard deviations. In the example shown in Fig. 2.1, the search query is “kidney weight left is low AND creatinine is high”. Expressions can be arbitrarily large: or-conditions expand horizontally and and-conditions expand

vertically. Thus, expressions such as ((a OR b) AND (c OR d)) can easily be input.

When samples of interest have been discovered, they can be selected using checkboxes on the left hand side, and then saved as a group. This has the same effect as saving a sample group from the *Sample groups* screen, discussed in Chapter 1.

Searching for sample units instead of individual samples is similar, except that conditions match on samples that belong together (e.g. as replicates under the same experimental conditions) instead of on individual samples.

It is possible to reset the search condition and/or the selection by using menu commands on the **Edit** menu. Search results can also be downloaded as CSV files by using the corresponding command from the **File** menu.

Chapter 3

Clustering and heat maps

To perform a hierarchical clustering, you must first select the samples (compounds) and genes you are interested in. The Open TG-GATEs dataset is too large to be clustered all at once by Toxygates. To select compounds and samples, follow the steps detailed in Chapter 1, save your sample groups and then go to the *View data* screen. At least two groups must be defined.

Most likely, you will at first see too many genes being displayed. For example, the default number is 31,042 genes (probes) for rat samples. There are several ways to select a more specific gene set. Here we will show how to perform p -value filtering. This option is only available when your groups contain exactly one time/dose/compound combination, so for this example, please do not combine multiple doses, times or compounds into one group.

On the *View data* screen, check the “ p -value columns” box at the top of the screen. Additional columns will appear, grouped by sample group. The first two columns correspond to the \log_2 fold change and p -values of sample group 1, the third and fourth columns to the \log_2 fold change and p -values of group 2, and so on.

Click the filter icon to set a filter. Type in a value, such as 0.1. Genes with a larger associated p -value will now be excluded. The filter icon will become blue to indicate that the filter is active.

Your total gene count may still be more than 1000. (If you cluster more than 1000 genes, an automatic cutoff will be applied). You may filter the second group in the same way as the first group, or you may tighten the filter on the first group, for example to 0.05.

When you are happy with the total number of genes, select "Show heat map..." on the **Tools** menu. The result should be similar to the graph shown in Figure 3.1.

As shown in Figure 3.1, other than inspecting the results, it is also possible to change parameters such as distance metric and clustering method on the right hand side. If you change the parameters, please click "Update" to see the result. To define clusters, select a cutoff point by clicking on the dendrogram

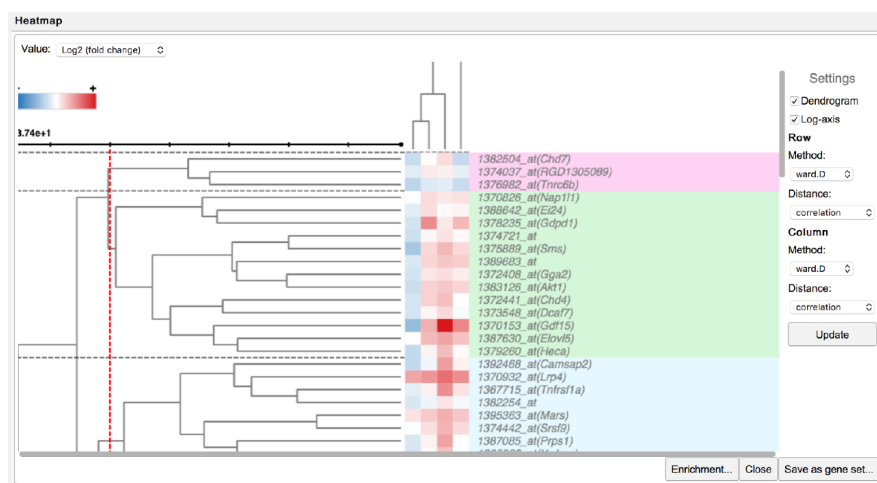


Figure 3.1: Sample graph showing the clustering of four sample groups (4 columns).

on the y-axis. This will partition the genes into gene sets corresponding to your selected clusters.

You can test your selected clusters for enrichment of various kinds, for example of pathways and GO terms, by clicking the **Enrichment** button. This may take a little while to compute. The result table will show the best matching annotation for each cluster. The enrichment is performed by using TargetMine (see Chapter 4).

Such enrichment testing may, in some situations, indicate whether your clustering is meaningful. A sample of clusters' enrichment is shown in Figure 3.2.

Best enrichment results					
Cluster	Size	ID	Description	p-value	Matches
Cluster 1	181	R-RNO-72689	Formation of a pool of free 40S subunits	0.000183	8
Cluster 2	293	(No result)			
Cluster 3	142	(No result)			

Figure 3.2: Table showing the results of performing enrichment on the generated clusters.

If you wish, you may repeat the steps above. Once you are happy with your clusters, you may click the **Save** button to save the clusters as gene sets. They will then become accessible on the **Gene sets** menu on the data screen.

Chapter 4

miRNA-mRNA network analysis

miRNA-mRNA networks can be analysed, in tabular (numerical) form as well as visually, by simultaneously activating sample groups of two different types. At the time of writing, a large mRNA-miRNA dataset is in preparation, but as it cannot yet be released, we have inserted the GSE29250 dataset from GEO. This is a set of non-small cell lung cancer (NSCLC) samples. In the discussion below we will show how to perform a basic dual table interaction analysis on this data, and how to visualise the associated interaction network. Similar steps will apply to any other mRNA-miRNA data that we release in the future. We recommend familiarising yourself with analysis of single sample groups (under Chapter 1, basic concepts, above) before proceeding with this section.

4.1 Choosing datasets

By clicking the "Data..." button on the *Sample groups* screen, it is possible to select active datasets. For a network to be constructed, both mRNA and miRNA samples must be available. At this time of writing, it is necessary to select the GEO dataset in order to see the NSCLC data, as shown in Figure 4.1.

4.2 Defining sample groups

As shown in Figure 4.1, next to the *Data...* button in the top left corner, it is possible to select the assay type of samples. Currently, mRNA and miRNA types are available. A network can be analysed if groups of both types have been saved and are active simultaneously. For example, we may define an mRNA group and a miRNA group. Note that as long as the groups contain samples for the same species, it is possible to simultaneously analyse any

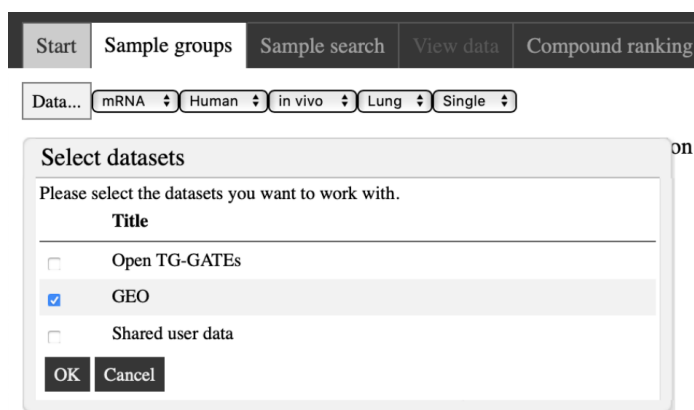


Figure 4.1: *Data selection dialog. The GEO dataset needs to be selected in order to define sample groups required for network analysis.*

number of mRNA and miRNA groups. This makes it possible to compare and contrast different sample conditions on the mRNA side, miRNA side, or both.

Select mRNA/Human/in vivo/Lung/Single, check the NSCLC adenocarcinoma *compound* (which in this dataset corresponds to a sample treatment rather than a compound), and check the 1 day middle “*dose*” compound that appears on the right. (The dose level and time point have no meaning for this dataset; age at cancer diagnosis, severity etc may vary). Click “*save as*” and accept the default name. Then change the sample class to miRNA/Human/in vivo/Lung/Single (miRNA instead of mRNA) and repeat the same procedure, saving again. In your group list, you should now have two sample groups, one of mRNA type and one of miRNA type, which might look like the ones shown in Figure 4.2.

Active	Group	Type	#Treated samples	#Control samples
<input checked="" type="checkbox"/>	NSCLC Ad/M/1 day	mRNA	3	3
<input checked="" type="checkbox"/>	NSCLC Ad/M/1 day 1	miRNA	3	3

Figure 4.2: *Sample graph showing the clustering of four sample groups (4 columns).*

Activate at least one mRNA group and at least one miRNA group simultaneously as to have enough information for the definition of a network and to enable the dual table/network display. Next, proceed to the “*View data*” screen.

4.3 Choosing miRNA-mRNA associations

mRNA-miRNA networks are defined by associations, which can be predicted or experimentally verified. Several well-known databases of such associations exist. At the time of writing, we provide access to the associations found in the miRTarBase, MirDB, and MiRAW (human only) databases.

The first time you try to load a network, no miRNA-mRNA associations will be enabled by default, and you will be asked to select some. This is done by using the "Select miRNA sources..." command on the **Tools** menu. A dialog will open, in which you can select miRNA sources (it is possible to enable several at the same time) and optionally select a cutoff for each source, when association scores are available (see Figure 4.3). For example, for MirDB, associations are predicted, and a numerical score of 90 would indicate a higher confidence than one of 80. More information is available at the official web site of each miRNA source (these are linked from the dialog). When you have selected association sources here, they will be automatically persisted in your browser. You can edit these settings at any time to refine a network.

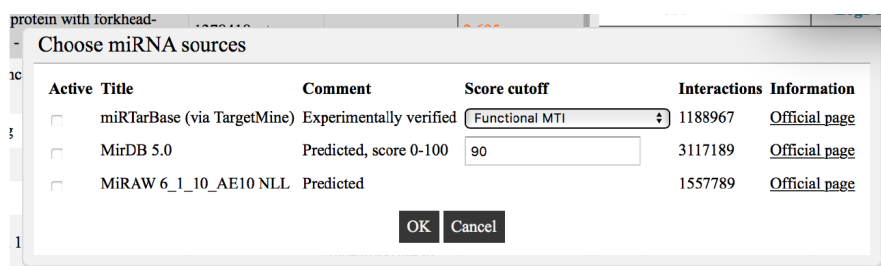


Figure 4.3: List of the currently available sources of predefined mRNA-miRNA associations.

In general, the best choice of miRNA source would depend on the biological problem being studied. For the purposes of working through this example, we recommend using MirDB and a cutoff value of 90.

4.4 Dual table analysis

When columns of two different types have been defined, and miRNA sources have been selected, as above, a dual table network will load, such as the one shown in Figure 4.4. The left hand "main" table is controlled by the paging buttons and works exactly like the data table normally does in the regular single-type mode. The right hand "side" table "follows" the main table by updating its contents to reflect what is being displayed. As the probe (gene) set in the main table changes, the side table set will update to reflect the associations of the currently displayed set. Thus, the two tables together display a partial view of an interaction network. Either mRNA or miRNA type

columns can be the main type. Note that a single miRNA probe typically has more associations than a single mRNA probe. Thus, a network defined by 50 miRNAs may be structurally very different from a network defined by 50 mRNAs.

Gene S...	Probe	MicroRNA	NSCLC Ad/M1 day
FAM71E1	229289_at		13.598
MUC5B	222268_x_at		13.218
MUC5B	213432_at		13.218
IGSF9	229276_at	hsa-miR-8485	10.772
PDZD3	220303_at	hsa-miR-4319	9.932
PDZD3	1555490_s_at	hsa-miR-4319	9.932
UNC79	229550_at	hsa-miR-3651 hsa-miR-2115-3p hsa-miR-660-5p hsa-miR-4261 hsa-miR-101-3p hsa-miR-708-3p hsa-miR-4317 hsa-miR-3121-3p hsa-miR-137 hsa-miR-7844-5p	8.001

miRNA	Probe	NSCLC Ad/M1 day 1	P-value	Count
hsa-miR-297		0.082	0.73	19
hsa-miR-1305		0.057	0.915	26
hsa-miR-137		0.056	0.886	32
hsa-miR-1283		0.035	0.937	49
hsa-miR-500a-5p		-0.004	0.998	93
hsa-miR-186-3p		-0.075	0.836	16
hsa-miR-643		-0.151	0.451	26
hsa-miR-590-3p		-0.168	0.791	13
hsa-miR-301b-3p		-0.193	0.53	46
hsa-miR-708-3p		-0.212	0.528	10
hsa-miR-23b-3p		-0.212	0.063	64

Figure 4.4: Dual table inspection of sample groups.

Figure 4.4 shows an example of such a dual table network with mRNA on the left hand side. The relative width of each table can be adjusted by dragging the handle in the middle. When a mRNA or miRNA is selected, the corresponding associated items on the opposite side are highlighted automatically. Sorting and filtering is possible in both tables. For example, to find miRNAs associated with low p -value genes, the p -value columns may be enabled (this affects only the main table; the side table always shows p -values), and the filter set to 0.05. This will limit the genes displayed in the main table and as a consequence, the miRNA table display will be updated as well. If a gene set were to be saved in this example, it would be treated as a mRNA set (since mRNA is the main table). The "Flip mRNA/microRNA" button, which is only visible in this mode, may be used to switch the main type and the side types. If a gene set is saved after switching in this fashion, it would instead be a miRNA set.

4.5 Visual network analysis

To analyse a network visually, it is first necessary to load it on the dual table data screen, as described above. In order to focus on only the most significant data in this example, we will apply additional filters as follows:

1. Main table \log_2 fold $|x| \geq 0$
2. Side table \log_2 fold $|x| \geq 0$

3. Main table p -value < 0.005

The first two filters filter out gene values with no measurable expression, and the third selects only the most differentially expressed mRNA.

When displaying a dual table view, on the **Network** menu, the "Visualize network" command will be available (as shown in Figure 4.5). This opens the network visualisation dialog. The network being visualised will be based on the data shown in the dual tables, and will initially respect the same sorting and filtering settings. However, because very large networks are slow to render visually, to speed up processing a size limit may apply, and only the top items from a network (according to sorting and filtering preferences) may be displayed. At the time of writing, we limit the number of items from the main table ("main type nodes") to 100, and there is no restriction on the number of associated side type nodes. If such a size limitation is applied, a message will be displayed to inform the user. We expect that users will want to gradually refine and experiment with networks until they find a reasonably sized subset of interest.

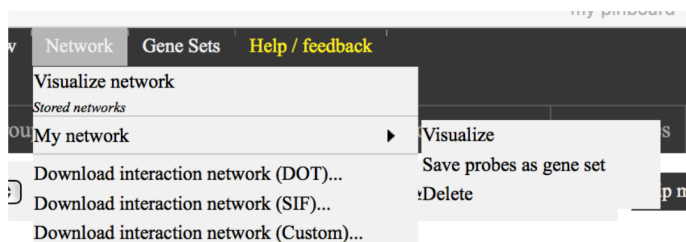


Figure 4.5: *Network menu. Different options allow visualizing the current network, previously saved networks, and downloading the network to various formats.*

A layout type may be chosen to adjust the visual appearance of the network. The default layout type being used is concentric, which can be rendered quickly. Users may wish to try other layout types including force directed, which can be slower to render but may be more visually intuitive in some cases. If we after displaying the network right-click the white background and select "apply colour scale" and accept the default settings (i.e. concentric layout), we should see something a display similar to the one shown in Figure 4.6.

In the default colour scale shown in Figure 4.6, blue nodes indicate up-regulation and red indicate down-regulation. Other colours (default colours) would indicate that no gene expression is available for that mRNA or miRNA node. By hovering the mouse cursor over a node, more information about that node and its expression levels may be displayed. The network visualisation is based on `cytoscape.js` and allows you to perform manipulations such as

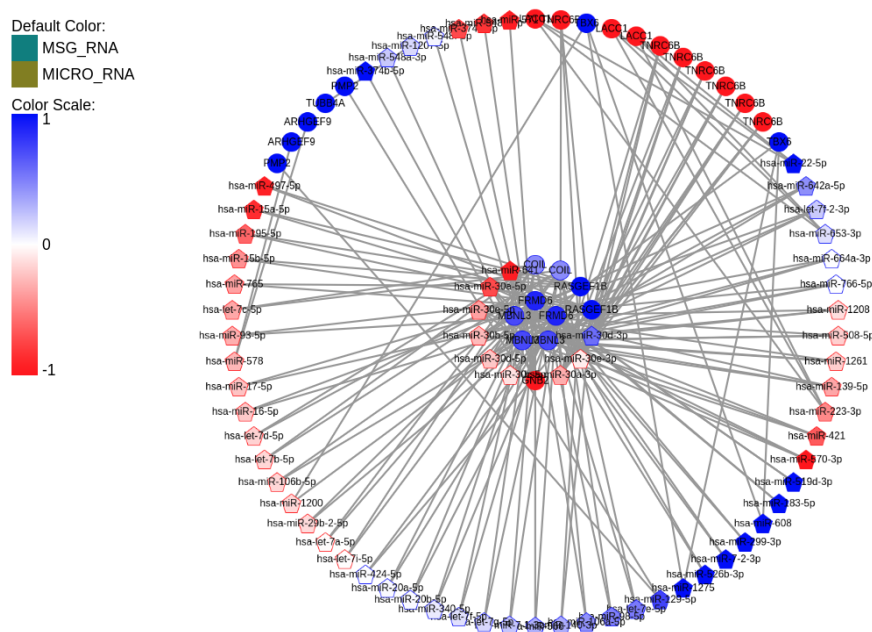


Figure 4.6: *Sample network visualization.*

panning and zooming (e.g. with the mouse scroll wheel) and adjusting the positions of nodes.

It is possible to change the filtering (but not sorting) underlying the network directly from the network dialog. To do so, select a column in the drop-down box at the bottom of the screen, and click the **Edit** button. For example, selecting a tighter filter bound would reduce the size of the underlying dual table network, and a looser filter bound would increase it. Such filtering applies prior to the maximum size cap for visualisation described above.

By using the "save and close" button, it is possible to save the network with a name. The network then becomes available on the **Network** menu on the *View data* screen for visualisation at any time, similar to the **Gene sets** menu. It also becomes available for cross-network comparison.

After being saved, a network can also be converted to a simple gene set by using the corresponding command on the **Network** menu on the *View data* screen. When this is done, you may select which type of probe to extract: mRNA or miRNA. The resulting set appears on the **Gene sets** menu and may then be treated as any other gene set. By using this method, gene sets and (partial) networks can freely be converted back and forth between each other.

Finally we will discuss how to contrast two networks. We assume that the first network has been defined as above and saved with a name. After closing

the network display, we clear the filters and instead set up the following filters:

1. Main table $\log_2\text{-fold} < -1$
2. Main table $p\text{-value} < 0.01$
3. Side table $\log_2\text{-fold } |x| \geq 0$

This will show only mRNAs down-regulated by at least half, with a somewhat looser p -value bound than before. We will still filter out miRNAs that have no discernible expression value. After visualising this network, we apply the default colour scale, and then click "load additional network". Here we select the network we saved previously. This should show the two networks side by side, as depicted in Figure 4.7.

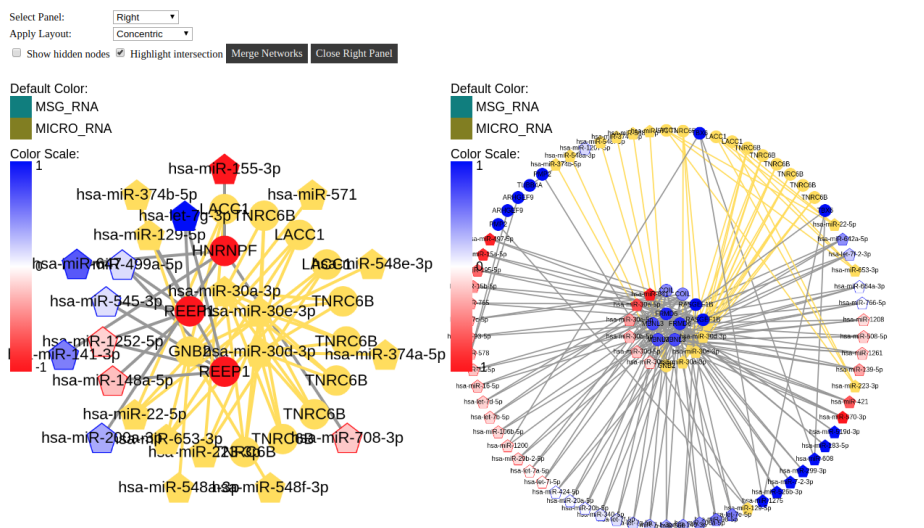


Figure 4.7: Side by side visualization of two networks. Highlighted are the intersecting elements among both networks.

From here we can perform various contrast analyses, such as merging the two networks, highlighting shared nodes (as we have done in Figure 4.7) by using the "highlight intersection" checkbox, and so on. Selecting a node will also highlight it on the opposite side if it is available in both networks. The button "close right panel" can be used to remove the side network from the display.

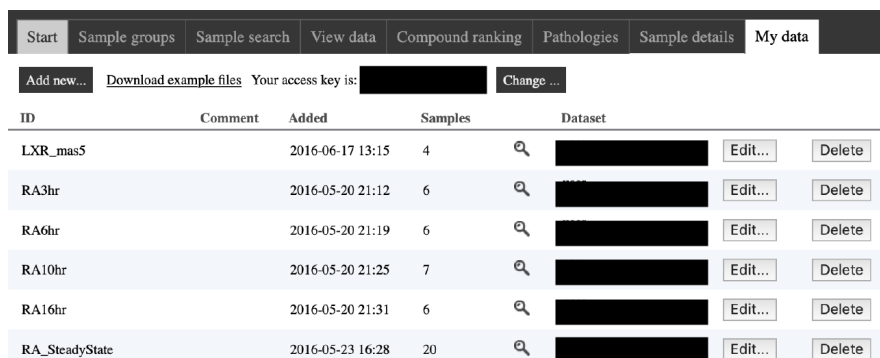
This concludes the discussion of interaction network analysis. We recommend interested readers to explore the various functions presented here on their own, to gain a deeper understanding of how they behave.

Chapter 5

Uploading your own data

Toxygates supports data upload of pre-normalised microarray data from rat, human and mouse platforms. Normalisation can be carried out using the `affy` package in R (bioconductor).

Uploaded data is managed from the *My data* screen, shown in Figure 5.1. We recommend that you first download the example files (linked from that screen) to understand the necessary data format.



ID	Comment	Added	Samples	Dataset		
LXR_mas5		2016-06-17 13:15	4	🔍 [redacted]	Edit...	Delete
RA3hr		2016-05-20 21:12	6	🔍 [redacted]	Edit...	Delete
RA6hr		2016-05-20 21:19	6	🔍 [redacted]	Edit...	Delete
RA10hr		2016-05-20 21:25	7	🔍 [redacted]	Edit...	Delete
RA16hr		2016-05-20 21:31	6	🔍 [redacted]	Edit...	Delete
RA_SteadyState		2016-05-23 16:28	20	🔍 [redacted]	Edit...	Delete

Figure 5.1: *My data* screen. The different rows in the table represent previously uploaded batches of data.

Once you have prepared the necessary files, you can upload them. Data is uploaded as batches (sets of samples). To begin an upload, click the "Add new..." button.

Next, you will need to supply a unique ID to identify the batch, and proceed to upload the necessary files, as shown in Figure 5.2. If there is any error in the files, you should get a warning message indicating what is wrong. Once you click "OK", the upload process starts. It may take several minutes to finish.

After uploading, you can manage your batches from the *My data* screen. You can inspect the sample parameters of a batch by clicking the magnifying

Edit batch

ID
VitaminA

Private comments

Visibility
Private

Metadata file (TSV)
Choose Files no files selected
Please upload a file

Normalized data file (CSV)
Choose Files no files selected
Please upload a file

Affymetrix calls file (CSV) (optional)
Choose Files no files selected
Please upload a file

OK Cancel

Figure 5.2: Batch setting dialog used for the uploading of user defined data.

glass. Batches may be edited or deleted by clicking the respective buttons.

When data has been uploaded, you can examine it as follows. Go to the *Sample group* definitions screen (as in Chapter 1) and select the necessary dataset, by using the "Data..." button. Your uploaded samples will be in a dataset called *My data*, which you should enable (see Figure 5.3).

You can optionally also enable other datasets, such as Open TG-GATEs, if you wish to perform a comparison with this data. Your compound names will be prefixed with the word [user]. For example, if you uploaded Acetaminophen data, it will appear as [user] Acetaminophen to distinguish it from the pre-existing data in Toxygates.

By including it in sample groups, you can now proceed as before to view your data (see Chapter 1), cluster it (see Chapter 3) or, if it includes both mRNA and miRNA, use it for the generation and visualization of networks (see Chapter 4).

Your data will be associated with an access key, which you can see at the top of the *My data* screen. The access key is a password that protects your data from other users. The key is automatically saved in your browser, but we recommend that you write it down, to avoid losing access to your data. If you know your key, you can access your data on a different computer or in a different web browser by clicking the **Change...** button and using the key.

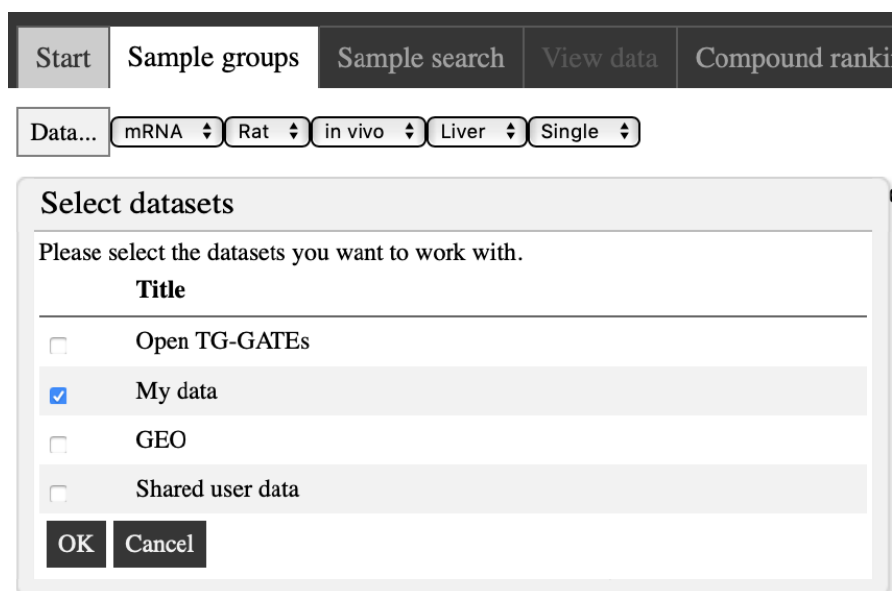


Figure 5.3: *User defined data is identified as My data in the Sample group screen.*

Chapter 6

Other tasks

6.1 Orthologous data inspection

If you define sample groups (see Chapter 1) from multiple species (for example, one group with human samples and one with rat samples), then the Affymetrix probes (and corresponding genes) will be combined in the data table, based on orthologous amino acid sequences in the relevant proteins. This allows for easier cross-species data analysis. In the example shown in Figure 6.1, three probes have been combined in the first row (1 human, and 2 rat), and 6 in the second row (3 human, 3 rat). Rat probes are prefixed with “Rat” and human probes with “HG”. For each row, the data values displayed will be the median value of all the underlying probes’ expression values.




Gene S...	Probe ...	Probe	RefSeq... ▼	caffeine/M/24 hr	caffeine/M/24 hr 1
				▼ caffei... ▼	caffei... ▼
 HG-...P2RX5 (1 probe) Rat...P2rx5 (2 probes)	purinergic receptor P2X - ligand-gated ion channel - 5 (3 probes)	1369673_at 1369674_at 210448_s_at	NM_001204519 NM_001204520 NM_002561 NM_080780 NM_175080 NM_175081	2.539	-0.241
 HG-...VAMP1 (3 probes) Rat...Vamp1 (3 probes)	vesicle-associated membrane protein 1 (3 probes) vesicle-associated membrane protein 1 (synaptobrevin 1) (3 probes)	1370556_at 1373510_at 1387519_at 207100_s_at 207101_at 213326_at	NM_013090 NM_014231 NM_016830 NM_199245	1.722	0.283
 Rat...LOC100910934 (1 probe) Rat...RGD1309362 (1 probe)	similar to interferon-inducible GTPase (1 probe)	1377950_at	NM_001024884 XM_003751792	1.71	(absent)

Figure 6.1: Combined display of sample groups defined for different species.

6.2 Import/export to/from InterMine instances

Toxygates can synchronise data with data warehouses built on the InterMine framework, such as HumanMine, TargetMine, RatMine and so on. You can synchronise your gene lists, through either importing or exporting them. This is done on the **Tools** menu on the *View data* screen, as shown in Figure 6.2. InterMine applications such as TargetMine provide many analysis functions not available in Toxygates and may allow you to obtain additional insight about your gene sets. Every supported InterMine instance will have its own submenu on the Tools menu.

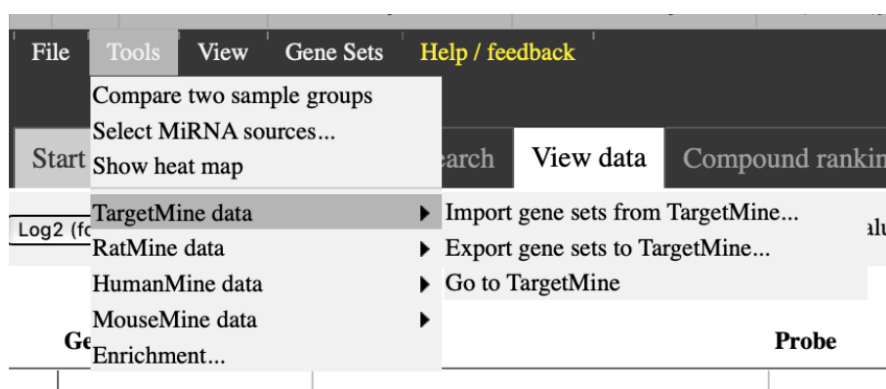


Figure 6.2: *Data selection includes the data uploaded by the user, which should be selected before being used.*

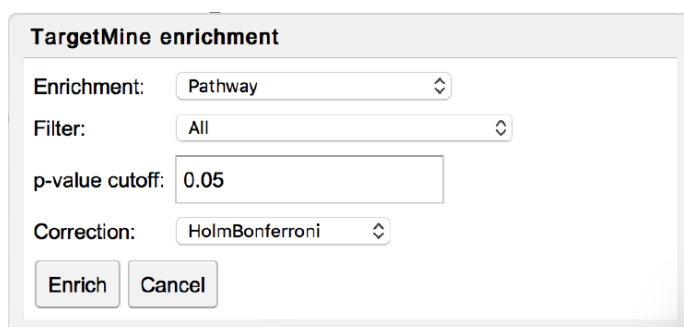
6.3 Enrichment testing

By using this function, for example on the **Tools/TargetMine** data menu on the *View data* screen (shown in Figure 6.2), you can test your currently displayed gene set for enrichment. Various enrichment types are available, such as pathways, GO terms, GOSlim terms and integrated pathway clusters (IPCs). For each type, sub-categories and other parameters are also available. The dialog used to define all these parameters is shown in Figure 6.3. Please note that no more than 1000 genes can be tested simultaneously.

The result of enrichment testing is a list of enriched objects, the number of matching genes, and the corresponding p -value for each entry, as shown in Figure 6.4.

6.4 Downloading data

From the **File** menu on the *View data* screen (and similarly on the *Sample search* and *Sample details* screens), it is possible to download the data



The dialog box titled "TargetMine enrichment" contains the following fields and controls:

- Enrichment:** A dropdown menu with "Pathway" selected.
- Filter:** A dropdown menu with "All" selected.
- p-value cutoff:** A text input field containing "0.05".
- Correction:** A dropdown menu with "HolmBonferroni" selected.
- Buttons for "Enrich" and "Cancel" at the bottom.

Figure 6.3: Dialog available to set the parameters used when performing gene set enrichment through TargetMine.

Enrichment results			
ID	Description	p-value	Matches
mo03010	Ribosome	1.19e-13	35
mo03050	Proteasome	0.000509	11
mo04919	Thyroid hormone signaling pathway	0.0220	15
mo04141	Protein processing in endoplasmic reticulum	0.0296	18
mo04120	Ubiquitin mediated proteolysis	0.0381	16
mo04520	Adherens junction	0.0459	11

Figure 6.4: Sample list of enrichment results.

displayed in the data table as a CSV file, as shown in Figure 6.5. Downloading the data as grouped samples will show it as averaged values, as it is displayed in Toxygates. Downloading the data as individual samples, on the other hand, will break out each value of each sample separately, which may be useful to carry out further statistical testing (for example, to compute your own p -values).

6.5 Viewing pathways and other annotations

By using the **View** menu on the *View data* screen, it is possible to enable or disable additional columns in the data table. This will show you additional information about your probes and genes. In many cases, these additional items will also have hyperlinks, taking you to external web pages with further information about them. A list of all the available options at the time of writing is shown in Figure 6.6.

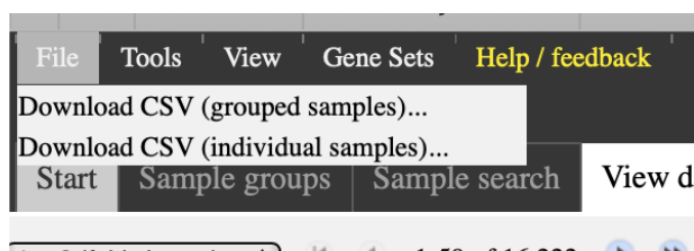


Figure 6.5: *Samples can be downloaded as CSV files, both as grouped or individual values.*

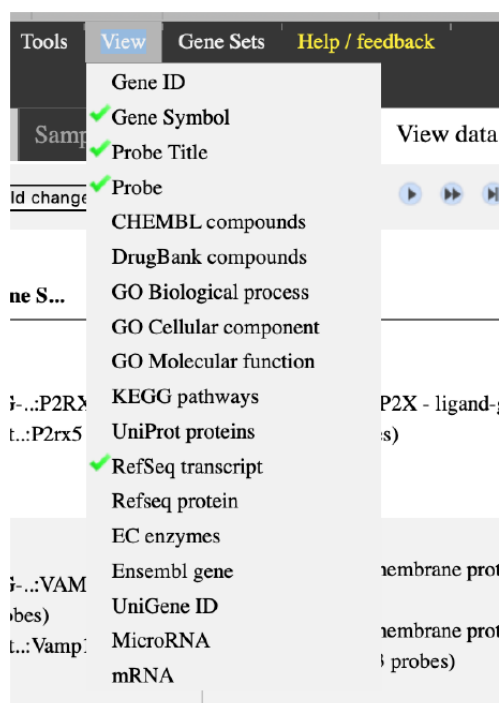


Figure 6.6: *Additional columns can be added to the table display of data.*

6.6 Inspecting sample attributes and biochemical data

On the *Sample details* screen you can see all the attributes for all the samples in the groups you have defined (see Chapter 1). This includes blood biochemical data as well as various measurements.

6.7 Viewing pathologies

Pathological findings are associated with samples in Open TG-GATEs. They may be displayed on the *Pathologies* screen. Only pathologies for your currently defined sample groups will be visible (see Chapter 1).