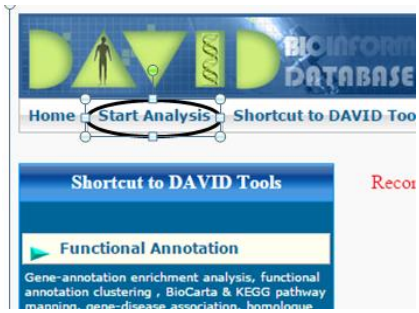
DAVID hands-on

by Ester Feldmesser, June 2017

1. Go to the DAVID website (http://david.abcc.ncifcrf.gov/)

2. Press on Start Analysis:

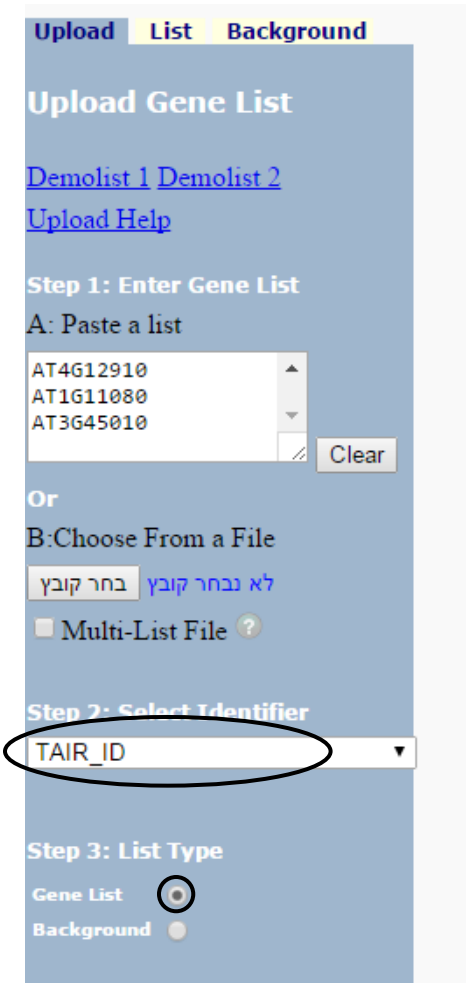


3. Choose the Upload tab in the left panel:

4. Download the k-means5_arabidopsis.txt file to your computer. Choose cluster 1 from the k-means clustering with 5 clusters, it includes genes down regulated in day 16.

5. Copy the gene symbols of the relevant cluster and paste the list into the David upload.

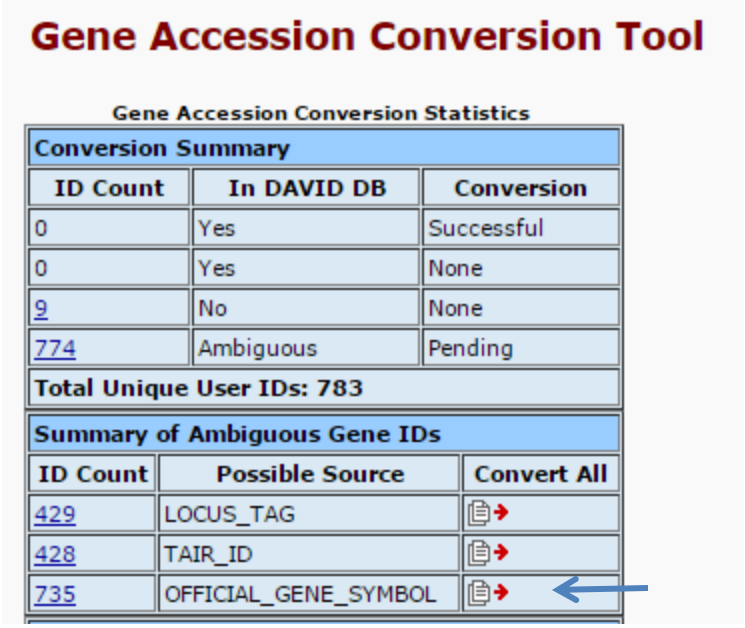
6. Choose the identifier (TAIR_ID) and the list type and submit.



7. Some of the gene IDs are not recognized and you need to submit them to the Conversion Tool. Click on the submission botton:

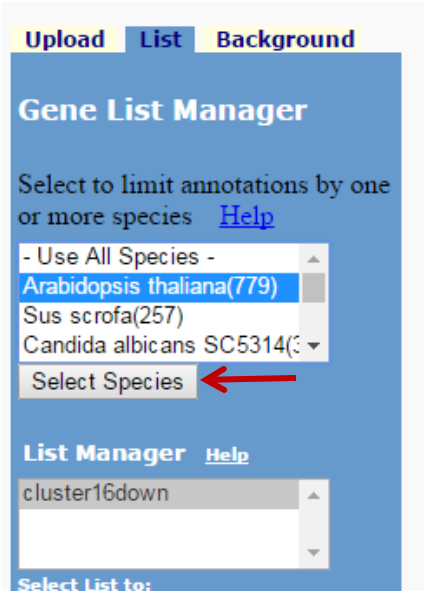
8. Be patient. It will take some time…

9. Click on red arrows of the column "**Convert All**" by the "OFFICIAL_GENE_SYMBOL" (see blue arrow below) and then by "TAIR_ID":

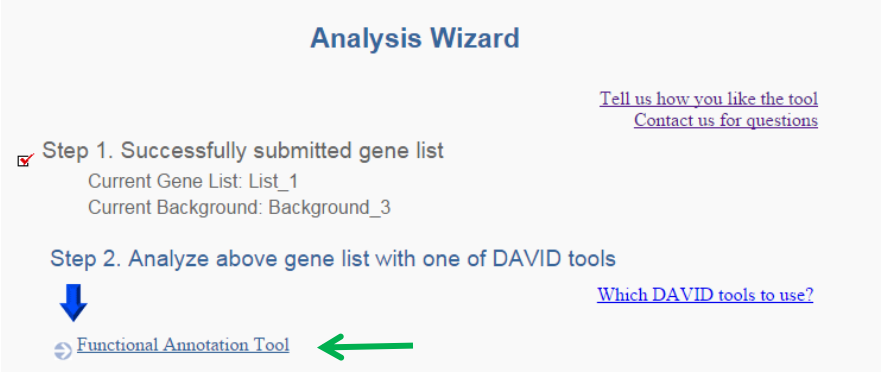


10. Click on the green button: “Submit Converted List to DAVID as a Gene List”. Give a name to the list and press OK.

11. Return to the previous DAVID tab and see the List. Some of the IDs match more than one species. You need to select Arabidopsis (red arrow).



12. We will use as a background the automatically detected Arabidopsis thaliana. Press on the Functional Annotation Tool button (green arrow below).



13. Press the + by Gene_Ontology (3 selected). Look at the Enrichment results for Biological Process FAT (GOTERM_BP_FAT) by clicking on the Chart button. Change the chosen term from DIRECT to FAT in the three GO categories. GO FAT filters out very broad GO terms based on a measured specificity of each term (not level-specificity) and this way reduces the dependency between the terms.

14. Keep this window open for further comparisons. If you want to save the data, select the text in the table you want to keep copy and paste to Excel. Be patient it takes time. Alternatively click on Download File, then also the genes related to each term will appear in the table.

15. What is the first term (the most significant one? Are there similar terms in the results? Click on the name of the first term. You get to a website for a browser for Gene Ontology terms and annotations. Read the term description and look at its ancestors and child terms.

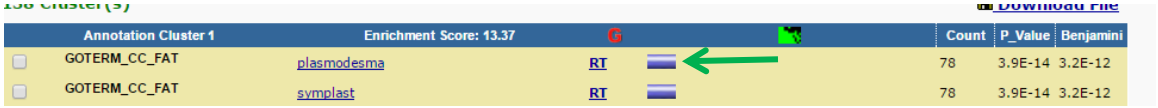
16. Return to the results window. Now open the Pathways window. Look at the Kegg results in the Pathways window by clicking on Chart. Is the first pathway related to the GO_BP result?

17. Click on the pathway name to follow the link to the pathway web site. The blinking red stars and the red names when scrolling down are the genes that appeared in cluster 2. Save the name of the pathway to compare it with the results of other tool.

18. Return to the Annotation Summary Results; choose the terms for the functional clustering. Open the Functional categories window and unselect all the terms (they are not very informative) and click on Functional Annotation Clustering.

19. The clusters are divided by blue strips. Each cluster has a title with a score and additional links. What is the score of the first cluster? How would you define the function that this cluster represents?

20. You can click on the red G to get the list of genes in the functional cluster or on the blueish rectangle (green arrow) to see the genes related to only one term. Try both.



21. Obtain a graphical description of the terms- genes relationship in the cluster by clicking on  . Allow Java to run if you are asked. Click on the Help button at the top of the page and read an explanation of the plot.

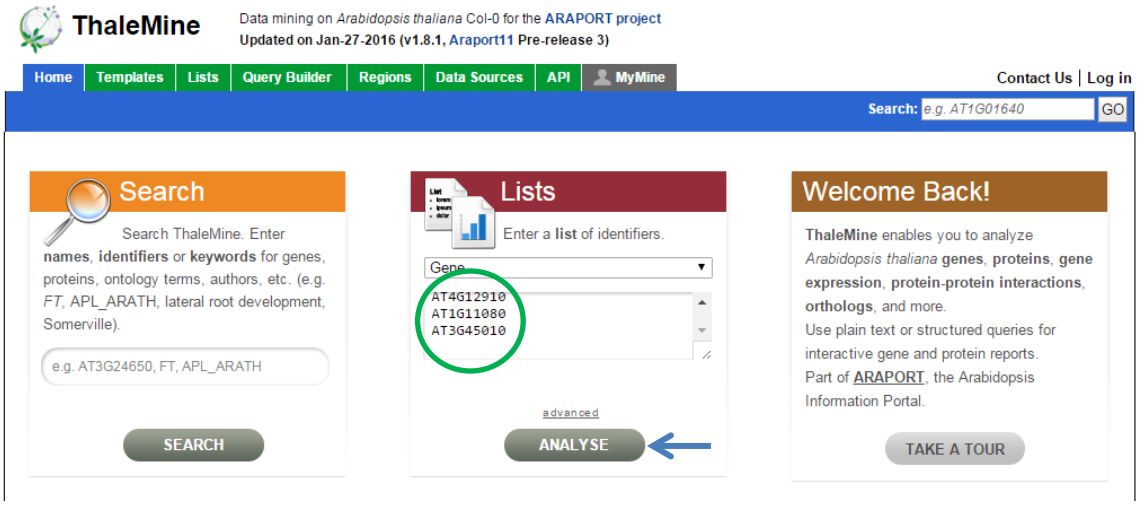
ThaleMine and Revigo hands-on

by Ester Feldmesser, June 2017

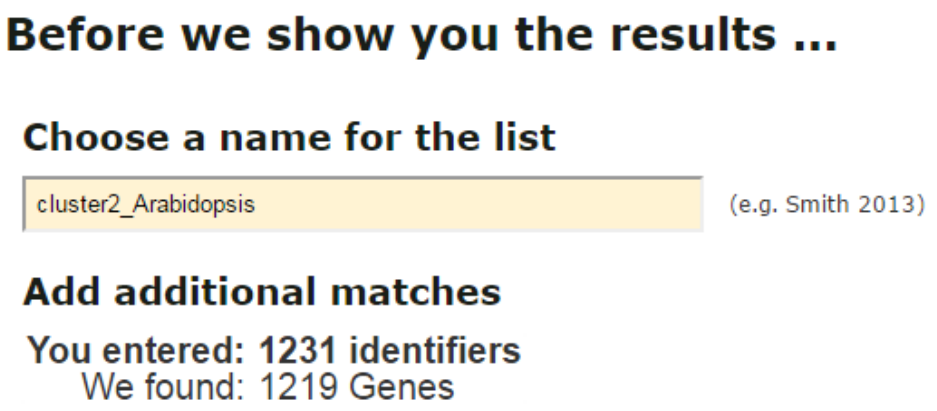
Part1. ThaleMine

22. InterMine warehouse was developed with the complexity of biological data in mind. The data model is flexible and extensible, and a range of data parsers is provided to facilitate the data loading. Go to the InterMine website (http://intermine.org/). A number of different data warehouses powered by InterMine already exist. We will work with ThaleMine. Click on its link (at the bottom) to go to their website.

23. Paste the list of genes symbols of cluster 1 to the box under Lists (green circle) and click on Analyze (blue arrow).



24. Give a name to the gene list:

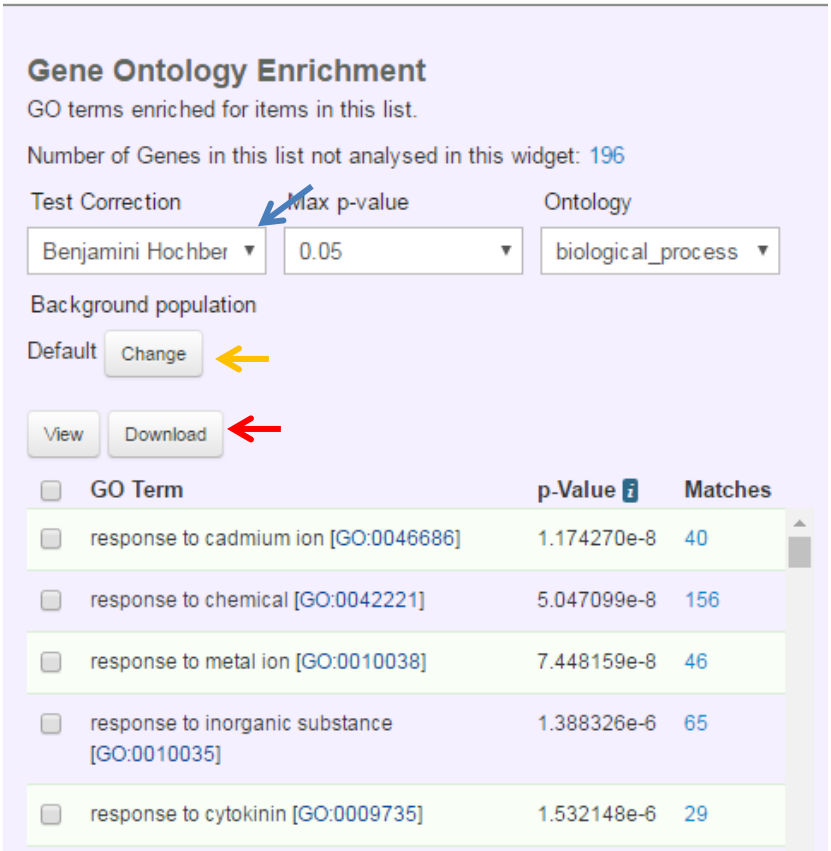


25. Press Enter in your keyboard.
26. Wait until the analysis finishes. The first part of the results includes your gene list with annotations; it can be filtered and downloaded.

27. Scroll down to see additional results. You can find Chromosome Distribution, Publication Enrichment, Gene Ontology Enrichment, Protein Domain Enrichment, Pathways Enrichment and Orthologues.

28. Look at the most significant pathway enriched. Is it the same one as in David? Open its link in an additional browser tab. Here the genes in your list appear in green (no blinking stars).

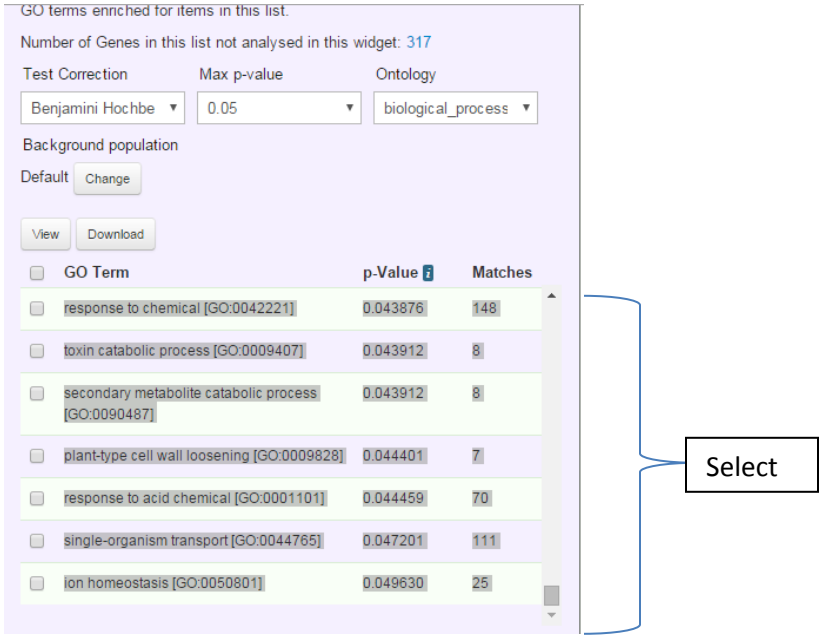
29. Return to the results window. We will now focus on the GO analysis. We will stay with the Biological process section of GO and we will change the Test Correction to Benjamini Hochbert, that is more permissive than Holm-Bonferroni.



30. Click on download (red arrow) and save the file. Open the file to view the genes that are related to each term.

31. Under Background population click the Change button (orange arrow). You can choose one of the options, but not upload your custom list. Since the suggested backgrounds are irrelevant for our experiment, leave the default (the default are all Arabidopsis genes with annotation in ThaleMine).

32. To save the text inside the results box, select it, copy its content and paste it into an Excel file.



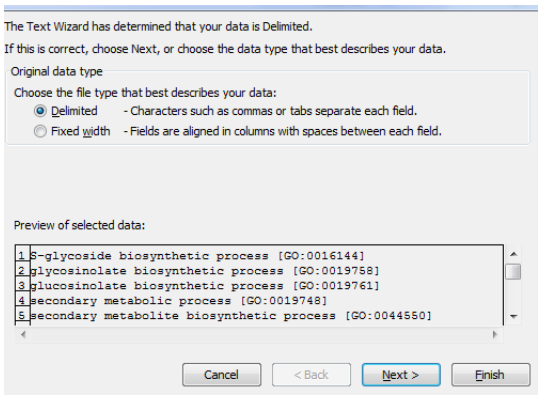
33. The number of terms found significant (Benjamini <0.05) is different in ThaleMine and in David. Are the first terms the same in both tools? Why are there differences in the number of significant terms?

Part2. Revigo

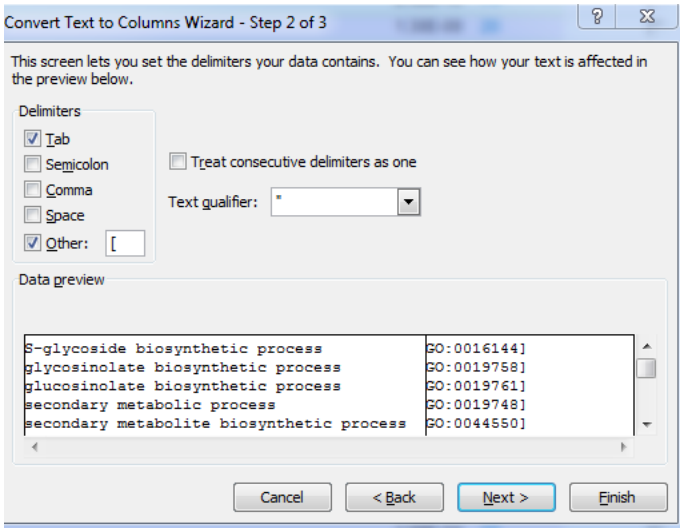
REViGO can take long lists of Gene Ontology terms and summarize them by removing redundant GO terms. The remaining terms can be visualized in semantic similarity-based scatterplots, interactive graphs, or tag clouds.

Long redundant lists of GO terms are difficult to interpret, but are likely to contain clusters of semantically similar GO terms. To mitigate the problem of large and redundant lists, REViGO finds a single representative GO term for each of these clusters.

1. First, we will prepare the input. The format should be tab delimited, with three columns: the GO term identifier (example: GO:0016144), the p-value and an optional column, in this case the number of genes matched to the term. Use the ThaleMine GO results in your Excel file. If it is difficult to manipulate the Excel file, select all and remove hyperlinks. Add a column after the first one. Select the first column a use the Text to columns function under the Data menu. In the dialog window choose Delimited.

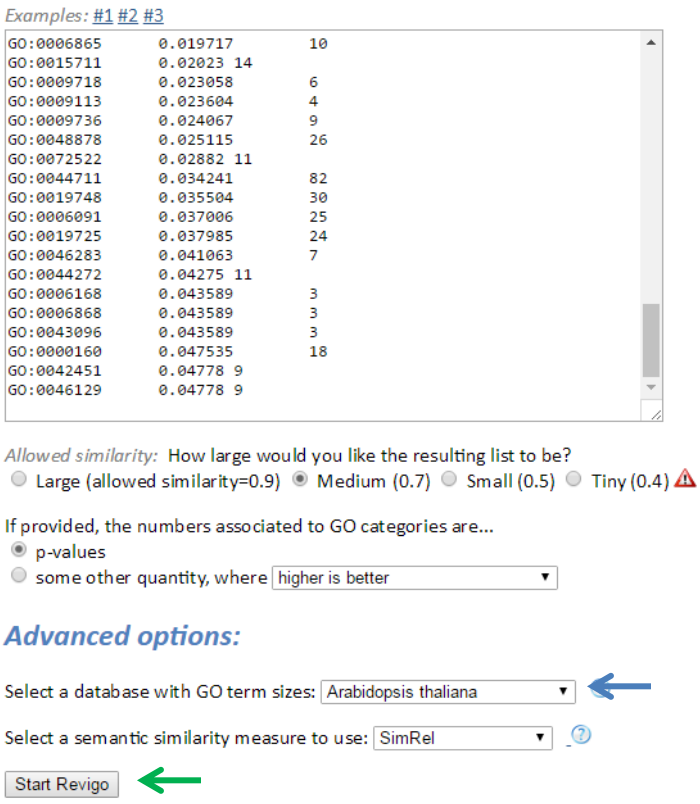


Click on Next. Select Other and write "[" in the box.

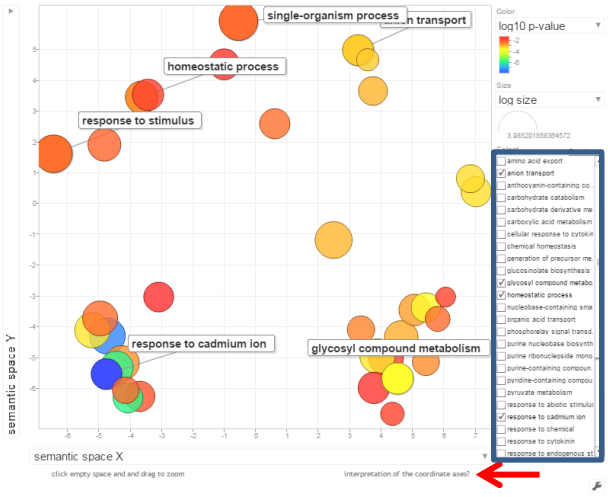


Click on Finish. Delete the first column. Remove the square parenthesis using Find and Replace.

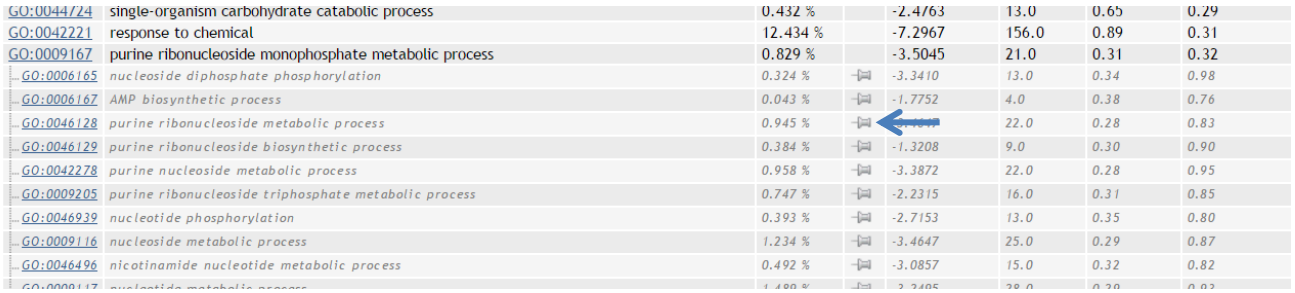
2. Go to the Revigo website (http://revigo.irb.hr/).
   Copy the data from the Excel that you prepared to the box.



3. In the Advanced Options change the database with GO terms sizes to Arabidopsis (blue arrow) and click on Start Revigo (green arrow).
4. Wait until the results finish running. The graphic results can be saved using PrintScreen or a similar command and data tables can be exported.
5. The first result tab that you see is Scatterplot & Table. If you do not see the scatterplot, you need to find where to allow flash to be activated. To understand the coordinate axes click on interpretation of the coordinate axes? (red arrow). Names of GO terms can be added or removed by clicking on the checkbox near their names, inside the Select box (blue rectangle).



6. Note that REVIGO generally does not prioritize higher-level or lower-level GO terms as cluster representatives – instead, the user-supplied p-values/enrichments are used to guide the selection. Very general GO terms, however, are always avoided as cluster representatives as they tend to be uninformative.
   It is also possible to manually override the choice of the representative GO term using the 'pin' option. Look at the table below the scatterplot. Terms in light grey were clustered under the term in black. Let's change a representative GO term and instead of purine ribonucleoside monophosphate metabolic process choose *purine ribonucleoside metabolic process* by clicking on its pin (blue arrow).

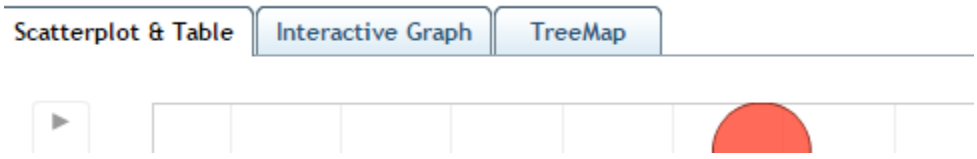


GO:0044724	single-organism carbohydrate catabolic process	0.432 %		-2.4763	13.0	0.65	0.29
GO:0042221	response to chemical	12.434 %		-7.2967	156.0	0.89	0.31
GO:0009167	purine ribonucleoside monophosphate metabolic process	0.829 %		-3.5045	21.0	0.31	0.32
GO:0006165	*nucleoside diphosphate phosphorylation*	*0.324 %*		*-3.3410*	*13.0*	*0.34*	*0.98*
GO:0006167	*AMP biosynthetic process*	*0.043 %*		*-1.7752*	*4.0*	*0.38*	*0.76*
GO:0046128	*purine ribonucleoside metabolic process*	*0.945 %*			*22.0*	*0.28*	*0.83*
GO:0046129	*purine ribonucleoside biosynthetic process*	*0.384 %*		*-1.3208*	*9.0*	*0.30*	*0.90*
GO:0042278	*purine nucleoside metabolic process*	*0.958 %*		*-3.3872*	*22.0*	*0.28*	*0.95*
GO:0009205	*purine ribonucleoside triphosphate metabolic process*	*0.747 %*		*-2.2315*	*16.0*	*0.31*	*0.85*
GO:0046939	*nucleotide phosphorylation*	*0.393 %*		*-2.7153*	*13.0*	*0.35*	*0.80*
GO:0009116	*nucleoside metabolic process*	*1.234 %*		*-3.4647*	*25.0*	*0.29*	*0.87*
GO:0046496	*nicotinamide nucleotide metabolic process*	*0.492 %*		*-3.0857*	*15.0*	*0.32*	*0.82*

   The scatterplot and table will be changed accordingly. See the green pin below.

term ID	description	frequency	pin?	$\log_{10}$ p-value	userVal_2	uniqueness	dispensability
GO:0006820	anion transport	1.524 %		-3.1018	26.0	0.86	0.00
GO:0042592	homeostatic process	2.637 %		-2.5774	38.0	0.94	0.00
GO:0044699	single-organism process	41.709 %		-2.1634	363.0	0.99	0.00
GO:0046128	purine ribonucleoside metabolic process	0.945 %		-3.4647	22.0	0.28	0.00
GO:0006165	*nucleoside diphosphate phosphorylation*	*0.324 %*		*-3.3410*	*13.0*	*0.34*	*0.98*
GO:0006168	*adenine salvage*	*0.022 %*		*-1.3606*	*3.0*	*0.42*	*0.85*

7. Let's look at the results at a different display. Click on the tab called Interactive Graph.



8. The nodes are colored according to the p-values and they can be moved with the mouse. Try it. The network can be downloaded for use in software that can read the appropriate format.

9. The last tab TreeMap displays the results in the treemap format. Each rectangle is a single cluster representative. The representatives are joined into 'superclusters' of loosely related terms, visualized with different colors.

10. Answers for this exercise can be found in http://dors.weizmann.ac.il/course/Answers%20to%20David%20and%20ThaleMine%20hands.pdf