A Simple Protocol for Informative Visualization of Enriched Gene Ontology Terms

Titouan Bonnot*, Morgane B. Gillard and Dawn H. Nagel*

Department of Botany and Plant Sciences, University of California, Riverside, California, USA
*For correspondence: tbonnot@ucr.edu; dawnn@ucr.edu

[Abstract] In genome-scale datasets, Gene Ontology (GO) enrichment is a common analysis to highlight functions over-represented or under-represented in a subset of differentially expressed genes to elucidate the biological significance of the results. However, despite the diversity of existing tools to analyze GO enrichment, it is often difficult to integrate results in an article figure with sufficient clarity. This is partly due to the high number and to the redundancy of the enriched GO terms, especially when looking at large sets of differentially expressed genes. Here, we provide a simple method to plot representative enriched GO terms. The list of representative enriched GO terms is obtained using existing tools Panther and REVIGO and results are represented in different plots generated from a homemade R script and the ggplot2 R package. The generated plots are publication-quality figures. The diversity of represented parameters makes the plots highly informative (number of genes associated with the enriched GO terms, fold enrichment and level of statistical significance). Comparison of GO enrichment between different lists of genes in a single plot is possible. As proof of concept, we performed this analysis on an *Arabidopsis* heat responsive transcriptome dataset recently published.

Keywords: Omics data, Transcriptomics, Gene Ontology Enrichment, Biological Processes, R plot, ggplot2

Materials and Reagents

1. User determined list(s) of differentially expressed genes (Similar to the one provided in Supplementary Data 1). In our example, we used the lists of up-regulated and down-regulated genes in response to heat in two different genotypes [wild type (WT) and a clock double mutant (*cca1-1/lhy-20*) recently published in Blair *et al.* (2019).
2. Data file (Similar to the one provided in Supplementary Data 2)
3. R-script file (Similar to the one provided in Supplementary Data 3)
4. Data file (Similar to the one provided in Supplementary Data 4)

Equipment

1. Computer

   Computer that can run one of the following operating system:

   Microsoft® Windows® XP (or later)

   Mac® OS X® 10.4 (or later)

Ubuntu 14.04 LTS (or later)

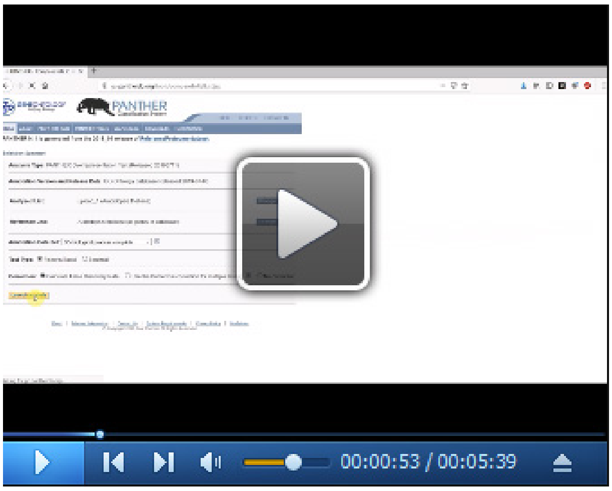
Software

1. R (https://r-project.org/)
2. RStudio (https://rstudio.com/products/rstudio/), RStudio is an optional user interface for R

Procedure

1. Obtain the list of over-represented and under-represented GO terms from a list of differentially expressed genes (DEGs). We used the interface of The Arabidopsis Information Resource (TAIR) for GO Term Enrichment for Plants, (https://www.arabidopsis.org/tools/go_term_enrichment.jsp), powered by Panther classification system (Mi *et al*., 2019).

   Note: Video 1 shows how to perform this procedure, from a list of differentially expressed genes to the plot shown in Figure 2. Other tools exist and can be used to calculate GO enrichment. This step can be skipped if the list of enriched GO terms is already obtained.



Video 1. Step by step procedure for beginners

In our example, 151 GO terms (biological processes) were significantly (Fisher's exact test, $P <$ 0.05 after False Discovery Rate (FDR) correction) over- or under-represented in the list of all DEGs (4,469 DEGs, Figure 1, Supplementary Data 1). We performed similar analyses by group (up-regulated genes in WT, down-regulated genes in WT, up-regulated genes in mutant, down-regulated genes in mutant) and 114, 281, 95 and 207 biological processes were over- or under-represented in these four different groups, respectively (Figure 1, Supplementary Data 1).

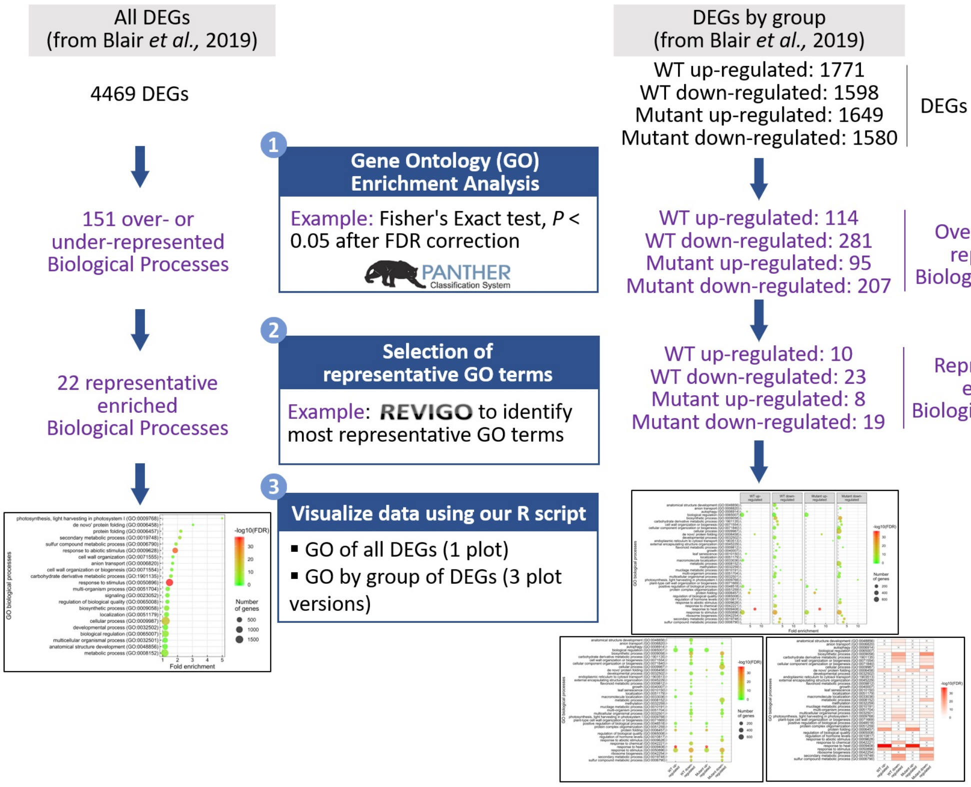

All DEGs
(from Blair *et al.*, 2019)
4469 DEGs
151 over- or under-represented Biological Processes
22 representative enriched Biological Processes
1
Gene Ontology (GO) Enrichment Analysis
Example: Fisher's Exact test, *P* < 0.05 after FDR correction
PANTHER Classification System
2
Selection of representative GO terms
Example: REVIGO to identify most representative GO terms
3
Visualize data using our R script
▪ GO of all DEGs (1 plot)
▪ GO by group of DEGs (3 plot versions)
DEGs by group
(from Blair *et al.*, 2019)
WT up-regulated: 1771
WT down-regulated: 1598
Mutant up-regulated: 1649
Mutant down-regulated: 1580
DEGs
WT up-regulated: 114
WT down-regulated: 281
Mutant up-regulated: 95
Mutant down-regulated: 207
Over- or under-represented Biological Processes
WT up-regulated: 10
WT down-regulated: 23
Mutant up-regulated: 8
Mutant down-regulated: 19
Representative enriched Biological Processes

Figure 1. Schematic summary of the procedure. The procedure presented here consists of three different steps, the last one allowing the visualization of the data in R plots. The results of this procedure applied to the data from Blair *et al.* (2019) are shown on the left for all the DEGs, and on the right for the four different groups of DEGs.

2. Identify the most representative GO terms from the list of enriched terms. The Web server REVIGO (http://revigo.irb.hr/) is particularly useful for this step because it calculates similarity between GO terms and removes redundant terms (Supek *et al*., 2011).

   In our example, we used REVIGO by providing the list of over- and under-represented biological processes and the associated FDR-corrected *P*-values (Supplementary Data 1 and 2). We selected the most representative GO terms by applying a stringent dispensability cutoff (< 0.05).

   Note: The cutoff and the method used to select representative GO terms can differ from the analysis that we have performed in our example (selection of highest fold enrichments, lowest P-values, etc.). In addition, this step can be skipped if the list of over-represented and under-represented GO terms is already known.

   Using this cutoff, 23 terms were selected. In our example, we chose to represent over-represented GO terms (GO terms with a fold enrichment > 1) only. They correspond to 22 representative enriched biological processes (Supplementary Data 1 and Video 1). We performed similar analyses by group (up-regulated genes in WT, down-regulated genes in WT, up-regulated genes in mutant, down-regulated genes in mutant) and 10, 23, 8 and 19 GO terms were selected, respectively (Figure 1, Supplementary Data 1)

3. Install the most recent version of R (https://www.r-project.org/).
4. Install the most recent version of RStudio (optional but recommended, https://rstudio.com/products/rstudio/download/).

   Note: To run RStudio, you need to have already installed R.
5. Prepare a data file similar to the one provided with this manuscript, entitled 'Supplementary Data 2'. It is necessary to keep the same table structure with identical column names and to avoid spaces and commas in the cells. In addition, this table must be a tab-delimited text file (.txt).
6. Open R (or RStudio) and open the script 'Supplementary Data 3'.
7. Install the package 'ggplot2'.

   Note: For beginners, please watch Video 1 to learn how to install an R package and how to run a script.
8. Change the working directory on line 10 ('C:/Users/Peter/Documents/GO enrichment plot/') and the name of the datafile on line 20 ('GO terms all.txt') of the script with your own (see Video 1).
9. Run the script from line 1 to line 54. This will calculate the -log10 (FDR-corrected *P*-values) and generate the plot represented in Figure 2 (Video 1).

   Note: To use other variables for the y-axis, or to change the dot sizes and dot colors, you can make the following modifications in the script:

   a. *To change the variable plotted on the y-axis, replace "Fold_enrichment" on line 36 by the column name of the variable you want to use. Please note that this is a flipped plot, so the y-axis corresponds to the horizontal axis. You would then also need to change the legend title on line 51.*
   b. *To use another variable for the dot color, replace "`|log10 (FDR)|`" on line 39 by the column name of the variable you want to use and change the legend title on line 52.*
   c. *To use another variable for the dot size, replace "Gene_number" on line 39 by the column name of the variable you want to use and change the legend title on line 52.*
   d. *To use different colors for the plots, replace the colors with your favorite colors on line 41 (see Video 1).*

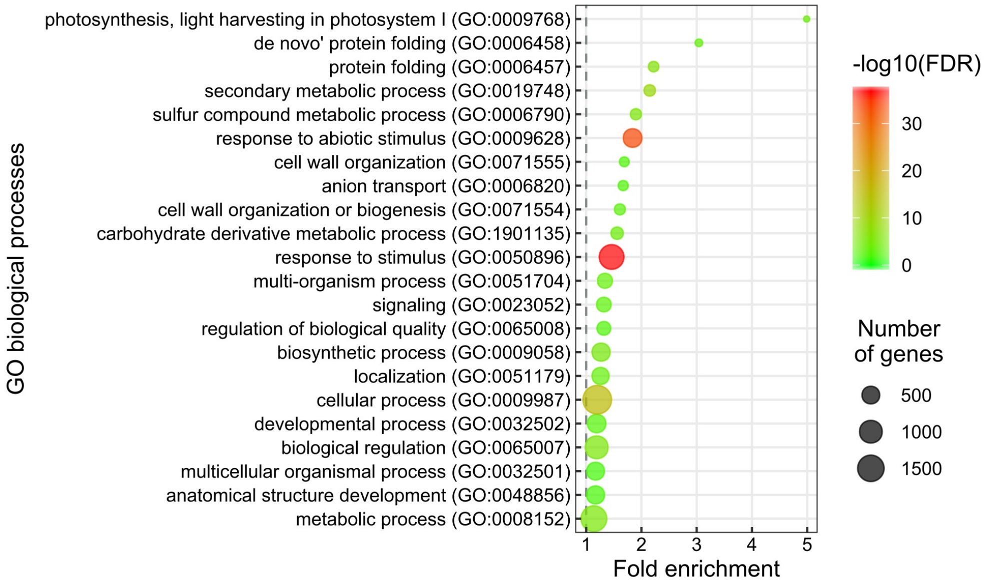


Figure 2. Over-represented biological processes in heat-responsive transcriptomes identified in Blair *et al.* (2019). The most representative and significant biological processes are represented and are sorted by fold enrichment. The dot size indicates the number of DEGs associated with the process and the dot color indicates the significance of the enrichment (-log10(FDR-corrected *P*-values)). The vertical grey dashed line represents a fold enrichment of 1.

10. Export the plot if needed (see Video 1).
11. To compare the GO enrichment between different lists of DEGs, prepare a datafile similar to the one provided with this manuscript, entitled 'Supplementary Data 4'. This table contains the selected over- and under-representative GO terms for each group of DEGs, mentioned in step 2 of the script. It is necessary to keep the same table structure with identical column names and to avoid spaces and commas in the cells. In addition, this table has to be a tab-delimited text file (.txt).
    Note: The number of groups and associated GO terms can differ from our example.
12. Replace the name of the datafile on line 66 ('Supplementary Data 4') with your own.
13. Replace the names of the different groups ("WT_up", "WT_down", "mutant_up", "mutant_down") with your own on line 84 of the script, in the order you want to see them on the plot.
14. Replace the names of the groups with your own on lines 89 to 92. This function will create new group labels on the plot and works as follows:

```
group.labs <- c (`Name of group A in the datafile` = "New name for group A, as it will appear on the plot”,  `Name of group B in the datafile` = "New name for group B, as it will appear on the plot”).
```

Notes:

a. If more than four groups, other lines should be added on the script after line 92, using the same nomenclature.

b. This step is optional. If not performed, the group names indicated in the datafile will be used on the plot by default.

15. Run the script from line 66 to 116. This will properly prepare the dataframe, calculate the -log10 (FDR-corrected P-values) and generate the plot represented in Figure 3.

    Note: To use other variables for the y-axis, or to change the dot sizes and dot colors, make similar modifications than those described in Step 8.

    This plot represents the GO enrichment results by group (four in our example), using the functionality 'facet_wrap' of the ggplot2 package. This representation is particularly useful to rapidly identify similarities and differences of enrichment between groups.

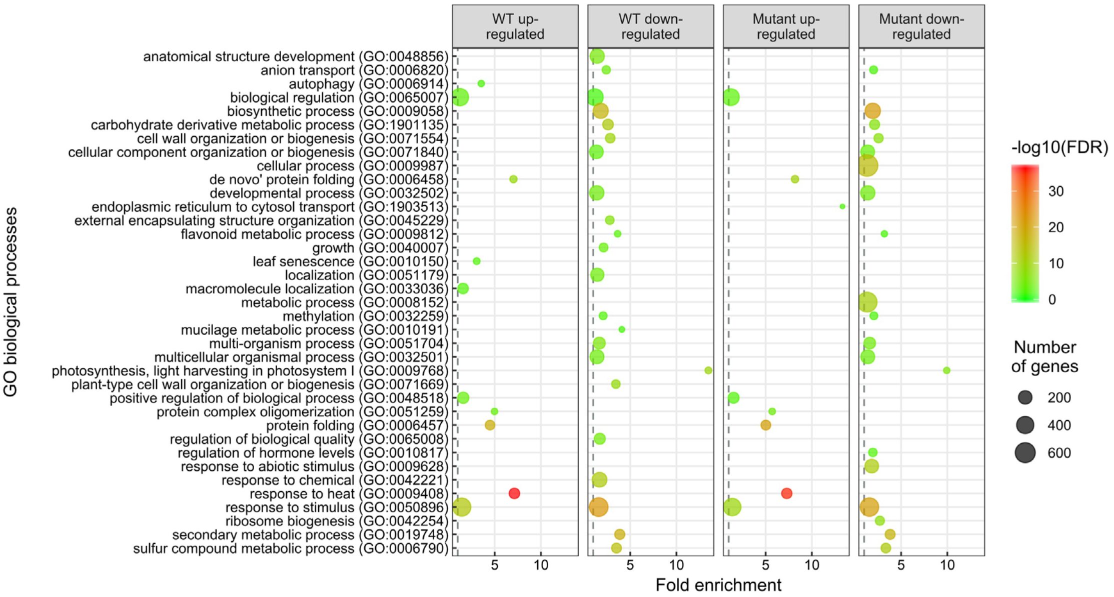


Figure 3. Plot 1 of the over-represented biological processes in upregulated and downregulated genes in response to heat in WT and clock mutant plants identified in Blair *et al.* (2019). The most representative and significant biological processes are represented and are sorted by fold enrichment. The dot size indicates the number of DEGs associated with the process and the dot color indicates the significance of the enrichment (-log10 (FDR-corrected *P*-values)). The vertical grey dashed line represents a fold enrichment of 1.

16. Run the script from line 123 to 135 to obtain another visualization of the results. You will obtain the plot represented in Figure 4. This representation does not separate the groups in different panels and is particularly adapted when comparing the GO enrichment between a high number of groups.

    Note: With this representation, the fold enrichment of the GO terms is not represented.

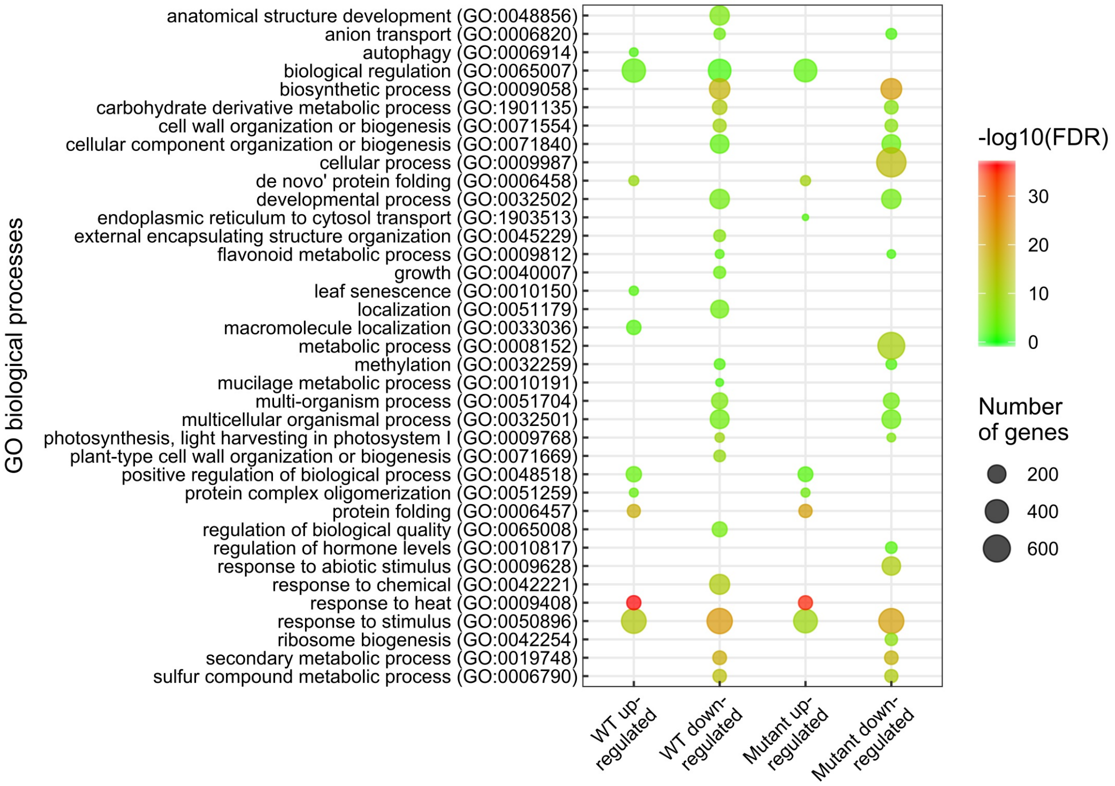


Figure 4. Plot 2 of the over-represented biological processes in upregulated and downregulated genes in response to heat in WT and clock mutant plants identified in Blair *et al.* (2019). The most representative and significant biological processes are represented. The dot size indicates the number of DEGs associated with the process and the dot color indicates the significance of the enrichment (-log10 (FDR-corrected *P*-values)).

17. Prepare a new dataframe by running the script from line 142 to 172. On lines 148 and 169, replace the names of the groups with your own.
18. Run the script from line 178 to 193 to obtain an alternative visualization of the results in a heatmap, as in Figure 5. In addition to show -log10 (FDR) values, this representation allows to easily identify the tested conditions in which a particular biological process is enriched or not.

    Note:

    a. *With this representation, the fold enrichment and the number of genes associated with the GO term are not represented.*

    b. *Please keep in mind that we represented in our example the representative enriched biological processes, all enriched terms are not represented (see Supplementary Data 1). Several GO terms can be enriched in multiple groups but are not necessarily representative in all of them, based on REVIGO results and the applied cutoff.*

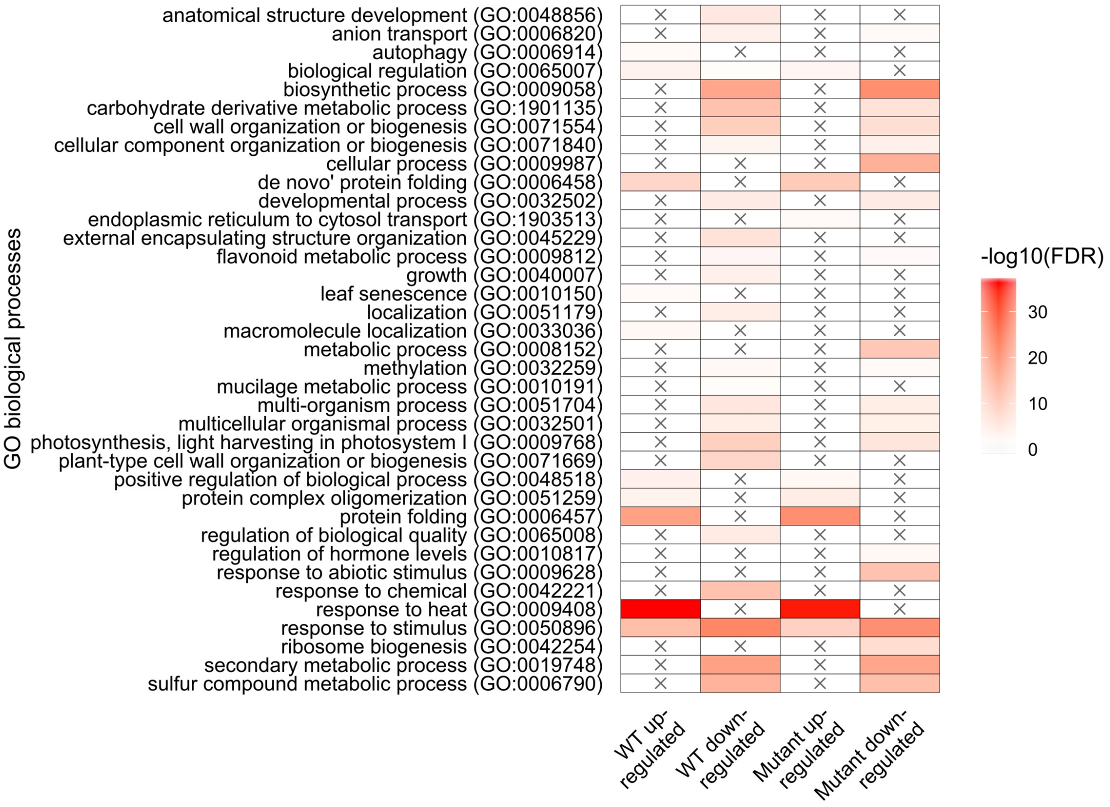


Figure 5. Plot 3 of the over-represented biological processes in upregulated and downregulated genes in response to heat in WT and clock mutant plants identified in Blair *et al.* (2019). The most representative and significant biological processes are represented. The colors indicate the significance of the enrichment (-log10 (FDR-corrected *P*-values)). An X mark in a cell means that the GO term is not a representative over- or under-represented term in the corresponding list of DEGs/in the corresponding condition/group.

Notes

In the R script, text preceded with '#' should be considered as user notes, it will not be read by the computer if the whole script is copied and pasted into the R or macro consoles.

Acknowledgments

We would like to thank members from the Nagel and Bailey-Serres lab for helpful comments and suggestions to improve the functionality of this method.

Competing interests

The authors declare that there are no conflicts of interest or competing interests.

References

1. Blair, E. J., Bonnot, T., Hummel, M., Hay, E., Marzolino, J. M., Quijada, I. A. and Nagel, D. H. (2019). Contribution of time of day and the circadian clock to the heat stress responsive transcriptome in *Arabidopsis*. *Sci Rep* 9(1): 4814.
2. Mi, H., Muruganujan, A., Ebert, D., Huang, X. and Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 47(D1): D419-D426.
3. Supek, F., Bošnjak, M., Škunca, N. and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6(7): e21800.