SUPPLEMENTARY DATA

Input data and input parameters

There are a few parameters, which must be set by the user. First the organism, ontology, background gene list source and the proper input format are selected. The input can either be expression trees of hierarchical clustering (HC) methods or multiple gene lists of data partitioning methods. If expression trees are submitted, they must be in gene tree (gtr) and clustered data table (cdt) file format, which is the output of Eisen's original hierarchical clustering package (1). The gtr / cdt-files can also be created using freeware tools, such as Tigr Mev (2). The gtr-file contains the positions and the connection order of the elements in the tree whereas the cdt-file contains the names of the genes. The cdt-file can optionally contain the expression values of the genes, in the range of –3 to 3 to enable the production of green-red heatmaps. After selecting the file format, the user sets the choice of statistical test, correction method and *p*-value cutoff. Optionally, the analysis can be limited to clusters of a certain size or to cluster that are located in a certain position of the tree. Another limitation that can be used to speed up the analysis will include only those GO-terms in the analysis that are more enriched than their expected mean.

Output listings

As a result, MultiGO produces output listings that address different questions. These listing can, for example, be used to have a rapid overview of the enriched functions. The main listing shows the most significant GO-terms of the data and clusters containing them. Below is a list of the most frequent “best GO-terms” of the cluster, *i.e* a list that shows the number of clusters having the particular GO-term as in their best GO-term. If the gene expression tree was used, then a figure is created that shows the *p*-values of the combined clusters at the different levels of the tree (supplementary figure 1).

Data interpretation tools

The web MultiGO has specific tools to ease data exploration and visualization. These tools can be used to create the gene *vs.* GO-term matrix, to draw the green-red heatmaps of the expression values and to discover clusters having certain genes or GO-terms. The visualizing tool can illustrate the expressions of genes and it can be used to explore the GO-terms of the clusters located farther from the root. This tool allows the user to focus on more specific gene sets and to examine the goodness of the clusters, *i.e.* to see if the occurrences of a GO-term listed as significant in the optimal global partition are even more highly concentrated in a sub-cluster (supplementary figure 1).

Programming language and dependencies

MultiGO is written in Perl. It requires C++ DCDFLIB to calculate the *p*-values of the chi-square test. DCDFLIB is a statistical cumulative density function library (http://www.csit.fsu.edu/~burkardt/cpp_src/dcdflib/dcdflib.html). All graphics are drawn using GnuPlot (version 4.0), which is a freeware data and function plotting utility (http://www.gnuplot.info/).

To perform analyses the tool requires gene ontology file in OBO-format, annotation file of the analyzed organism in corresponding GO-format and a background frequency file. The gene ontology and the annotation files can be downloaded from the GO homepages (http://www.geneontology.org/). The optional background frequency files can be generated from the GO-files using a script that is included in the MultiGO package. All tools, scripts and necessary files to run MultiGO and to perform the analysis, can be downloaded from the group's web pages (http://ekhidna.biocenter.helsinki.fi/poxo/multigo).

1. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U S A.*, **95,** 14863-14868.

2. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M., Sturn,A., Snuffin,M., Rezantsev,A., Popov,D., Ryltsov,A., Kostukovich,E., Borisovsky,I., Liu,Z., Vinsavich,A., Trush,V. and Quackenbush,J. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374-378.

Table 1. Experimental sample pairs used. In the table *ID* is the identifier of the experiment in heatmap pictures and *affected* is the number of genes having altered expression between the experimental sample and the corresponding control. In the *Accession* 1 is the accession codes for TAIR (http://www.arabidopsis.org) and 2 is the NASC array reference number (http://affymetrix.arabidopsis.info/).

ID	*Tissue*	*Control sample*	*Experimental sample*	*Affected*	*Author*	*Accession*
1a	shoot	Non-treated, 15 min	Drought, 15 min	1607	D`Angelo C	ME00338^1
1b	shoot	Non-treated, 30 min	Drought, 30 min	649	D`Angelo C	ME00338^1
1c	shoot	Non-treated, 1h	Drought, 1h	562	D`Angelo C	ME00338^1
1d	shoot	Non-treated, 3h	Drought, 3h	1401	D`Angelo C	ME00338^1
1e	shoot	Non-treated, 6h	Drought, 6h	2395	D`Angelo C	ME00338^1
1f	shoot	Non-treated, 12h	Drought, 12h	937	D`Angelo C	ME00338^1
1g	shoot	Non-treated, 24h	Drought, 24h	1662	D`Angelo C	ME00338^1
2a	seedling	Mock treatment, 30min	10uM ABA, 30min	129	Goda H	ME00333^1
2b	seedling	Mock treatment, 1h	10uM ABA, 1h	443	Goda H	ME00333^1
2c	seedling	Mock treatment, 3h	10uM ABA, 3h	1953	Goda H	ME00333^1
3a	leaf	Sfr2 mutant	Sfr2 mutant 24h at 4^{o}C	6460	Warren G	70^2
3b	leaf	Sfr3 mutant	Sfr3 mutant 24h at 4^{o}C	4281	Warren G	70^2
3c	leaf	Sfr6 mutant	Sfr6 mutant 24h at 4^{o}C	5159	Warren G	70^2
3d	leaf	Sfr6 mutant	Sfr6 mutant no water 2h	1566	Warren G	70^2
3e	leaf	4^{o}C 0h	4^{o}C 3h	4337	Warren G	70^2
3f	leaf	4^{o}C 0h	4^{o}C 24h	5412	Warren G	70^2
4a	leaf	No treatment	ERD15 mutant	(11875)	Helenius E	321^2
4b	leaf	ABA	ERD15 mutant - ABA	(11875)	Helenius E	321^2

Table 2. The median and most significant overall *p*-values obtained in the random analysis. In the table *Sig* is the most significant *p*-value, *Experiment* is the overall *p*-value obtained from the experimental analysis, *Higher* is the number random analyses yielding equal or more significant overall *p*-value than the overall *p*-value obtained from the experimental analysis and *Total* is the total number of randomizations performed.

Relative height	*Median p-value*	*Sig p-value*	*Experiment*	*Higher*	*Total*
0	1.E+00	1.E+00	1.E+00	1000	1000
0.01	1.E+00	1.E+00	1.E+00	1000	1000
0.02	1.E+00	1.E+00	1.E+00	1000	1000
0.03	1.E+00	1.E+00	1.E+00	1000	1000
0.04	1.E+00	1.E+00	1.E+00	1000	1000
0.05	1.E+00	1.E+00	1.E+00	1000	1000
0.06	1.E+00	1.E+00	1.E+00	1000	1000
0.07	1.E+00	1.E+00	1.E+00	1000	1000
0.08	1.E+00	1.E+00	1.E+00	1000	1000
0.09	1.E+00	1.E+00	1.E+00	1000	1000
0.1	1.E+00	1.E+00	1.E+00	1000	1000
0.11	1.E+00	1.E+00	1.E+00	1000	1000
0.12	1.E+00	1.E+00	1.E+00	1000	1000
0.13	1.E+00	1.E+00	1.E+00	1000	1000
0.14	1.E+00	1.E+00	1.E+00	1000	1000
0.15	1.E+00	1.E+00	1.E+00	1000	1000
0.16	1.E+00	1.E+00	1.E+00	1000	1000
0.17	1.E+00	1.E+00	1.E+00	1000	1000
0.18	1.E+00	1.E+00	1.E+00	1000	1000
0.19	1.E+00	1.E+00	1.E+00	1000	1000
0.2	1.E+00	1.E+00	1.E+00	1000	1000
0.21	1.E+00	1.E+00	1.E+00	1000	1000
0.22	1.E+00	1.E+00	1.E+00	1000	1000
0.23	1.E+00	1.E+00	1.E+00	1000	1000
0.24	1.E+00	1.E+00	1.E+00	1000	1000
0.25	1.E+00	1.E+00	1.E+00	1000	1000
0.26	1.E+00	1.E+00	1.E+00	1000	1000
0.27	1.E+00	1.E+00	1.E+00	1000	1000
0.28	1.E+00	1.E+00	1.E+00	1000	1000
0.29	1.E+00	1.E+00	1.E+00	1000	1000
0.3	1.E+00	1.E+00	1.E+00	1000	1000
0.31	1.E+00	1.E+00	1.E+00	1000	1000
0.32	1.E+00	1.E+00	1.E+00	1000	1000
0.33	1.E+00	1.E+00	1.E+00	1000	1000
0.34	1.E+00	1.E+00	1.E+00	1000	1000
0.35	1.E+00	1.E+00	1.E+00	1000	1000
0.36	1.E+00	1.E+00	1.E+00	1000	1000
0.37	1.E+00	1.E+00	1.E+00	1000	1000
0.38	1.E+00	1.E+00	1.E+00	1000	1000

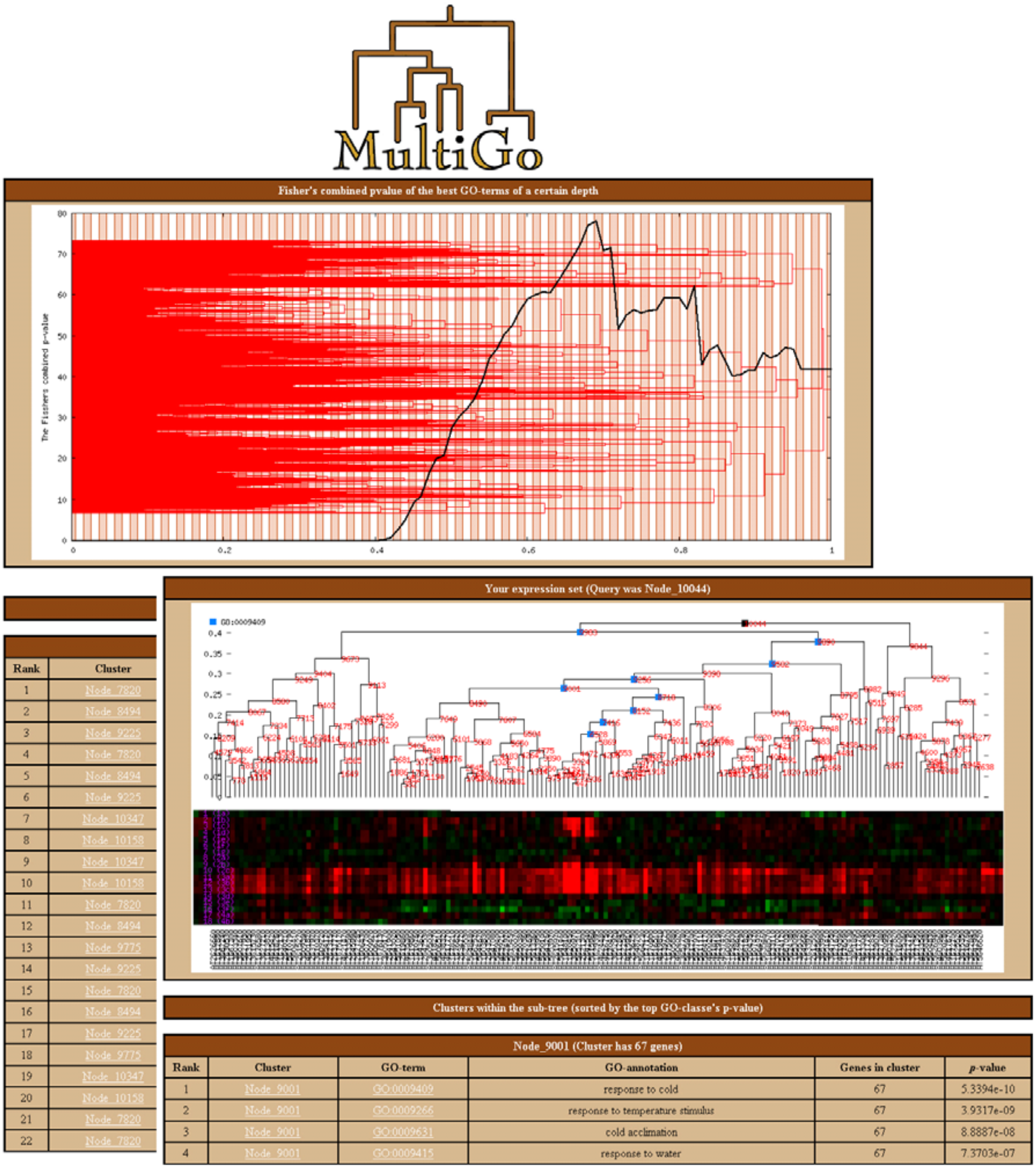
0.39	1.E+00	1.E+00	1.E+00	1000	1000
0.4	1.E+00	1.E+00	1.E+00	1000	1000
0.41	1.E+00	1.E+00	1.E+00	1000	1000
0.42	1.E+00	1.E+00	8.E-01	0	1000
0.43	1.E+00	1.E+00	2.E-01	0	1000
0.44	1.E+00	1.E+00	7.E-02	0	1000
0.45	1.E+00	1.E+00	3.E-03	0	1000
0.46	1.E+00	1.E+00	3.E-05	0	1000
0.47	1.E+00	1.E+00	4.E-07	0	1000
0.48	1.E+00	1.E+00	2.E-14	0	1000
0.49	1.E+00	1.E+00	1.E-15	0	1000
0.5	1.E+00	1.E+00	1.E-19	0	1000
0.51	1.E+00	1.E+00	1.E-22	0	1000
0.52	1.E+00	1.E+00	9.E-25	0	1000
0.53	1.E+00	1.E+00	8.E-31	0	1000
0.54	1.E+00	1.E+00	8.E-35	0	1000
0.55	1.E+00	1.E+00	5.E-38	0	1000
0.56	1.E+00	1.E+00	2.E-42	0	1000
0.57	1.E+00	1.E+00	1.E-45	0	1000
0.58	1.E+00	1.E+00	2.E-48	0	1000
0.59	1.E+00	1.E+00	2.E-54	0	1000
0.6	1.E+00	1.E+00	4.E-61	0	1000
0.61	1.E+00	1.E+00	1.E-65	0	1000
0.62	1.E+00	1.E+00	3.E-49	0	1000
0.63	1.E+00	1.E+00	3.E-48	0	1000
0.64	1.E+00	1.E+00	1.E-51	0	1000
0.65	1.E+00	1.E+00	7.E-53	0	1000
0.66	1.E+00	1.E+00	8.E-52	0	1000
0.67	1.E+00	1.E+00	5.E-53	0	1000
0.68	1.E+00	1.E+00	6.E-57	0	1000
0.69	1.E+00	1.E+00	7.E-55	0	1000
0.7	1.E+00	1.E+00	4.E-64	0	1000
0.71	1.E+00	1.E+00	5.E-67	0	1000
0.72	1.E+00	1.E+00	2.E-64	0	1000
0.73	1.E+00	1.E+00	3.E-70	0	1000
0.74	1.E+00	1.E+00	8.E-53	0	1000
0.75	1.E+00	1.E+00	7.E-56	0	1000
0.76	1.E+00	1.E+00	3.E-58	0	1000
0.77	1.E+00	1.E+00	2.E-59	0	1000
0.78	1.E+00	1.E+00	2.E-60	0	1000
0.79	1.E+00	1.E+00	2.E-57	0	1000
0.8	1.E+00	8.E-01	1.E-58	0	1000
0.81	1.E+00	7.E-01	3.E-51	0	1000
0.82	1.E+00	6.E-01	1.E-44	0	1000
0.83	1.E+00	5.E-01	1.E-45	0	1000
0.84	1.E+00	3.E-01	6.E-45	0	1000
0.85	1.E+00	2.E-01	5.E-44	0	1000
0.86	1.E+00	5.E-02	6.E-42	0	1000
0.87	1.E+00	3.E-02	5.E-44	0	1000
0.88	1.E+00	9.E-03	2.E-42	0	1000
0.89	1.E+00	9.E-03	2.E-42	0	1000

0.9	1.E+00	5.E-04	1.E-41	0	1000
0.91	1.E+00	5.E-05	3.E-43	0	1000
0.92	1.E+00	1.E-05	2.E-44	0	1000
0.93	1.E+00	1.E-05	2.E-44	0	1000
0.94	1.E+00	3.E-07	1.E-40	0	1000
0.95	3.E-02	2.E-09	9.E-35	0	1000
0.96	3.E-02	2.E-09	9.E-35	0	1000
0.97	3.E-05	1.E-10	3.E-33	0	1000
0.98	3.E-05	1.E-10	3.E-33	0	1000
0.99	7.E-14	7.E-14	7.E-14	1000	1000
1	7.E-14	7.E-14	7.E-14	1000	1000

Table 3. The top ten most significant *p*-values obtained from the experimental analysis and from the random analyses. In the table *Sig* is the most significant *p*-value and *Height* is the relative position of the cluster in the tree having the *p*-value.

	Random analyses		*Experimental analysis*	
Rank	*Sig*	*Relative height*	*Sig*	*Relative height*
1	7.E-14	0.99	5.E-84	0.18
2	7.E-14	0.99	7.E-83	0.23
3	7.E-14	0.99	1.E-80	0.30
4	7.E-14	0.99	1.E-79	0.18
5	7.E-14	0.99	1.E-78	0.23
6	7.E-14	0.99	2.E-76	0.30
7	7.E-14	0.99	2.E-76	0.14
8	7.E-14	0.99	3.E-76	0.49
9	7.E-14	0.99	2.E-72	0.14
10	7.E-14	0.99	5.E-72	0.49

Figure 1. Screenshots of MultiGO. Below is the main page showing the expression tree and the *p*-values of the Fisher's combined probability test. Above is the expression heatmap showing a selected cluster. The blue squares indicate a child cluster, whose most significant GO-term is the same as the most significant GO-term of the query cluster.



Fisher's combined pvalue of the best GO-terms of a certain depth

Your expression set (Query was Node_10044)

Rank	Cluster
1	Node_7820
2	Node_8494
3	Node_9225
4	Node_7820
5	Node_8494
6	Node_9225
7	Node_10347
8	Node_10158
9	Node_10347
10	Node_10158
11	Node_7820
12	Node_8494
13	Node_9775
14	Node_9225
15	Node_7820
16	Node_8494
17	Node_9225
18	Node_9775
19	Node_10347
20	Node_10158
21	Node_7820
22	Node_7820

Clusters within the sub-tree (sorted by the top GO-classe's p-value)

Node_9001 (Cluster has 67 genes)

Rank	Cluster	GO-term	GO-annotation	Genes in cluster	*p*-value
1	Node_9001	GO:0009409	response to cold	67	5.3394e-10
2	Node_9001	GO:0009266	response to temperature stimulus	67	3.9317e-09
3	Node_9001	GO:0009631	cold acclimation	67	8.8887e-08
4	Node_9001	GO:0009415	response to water	67	7.3703e-07

Figure 2. Calculated overall *p*-values using different clusters sizes. The y-axis is the –$\log_{10}$ value of the overall *p*-value and the x-axis is the position at the tree (relative height). Other parameters were: HD, FDR-correction, 0.001 as the *p*-value cutoff and no filtering.

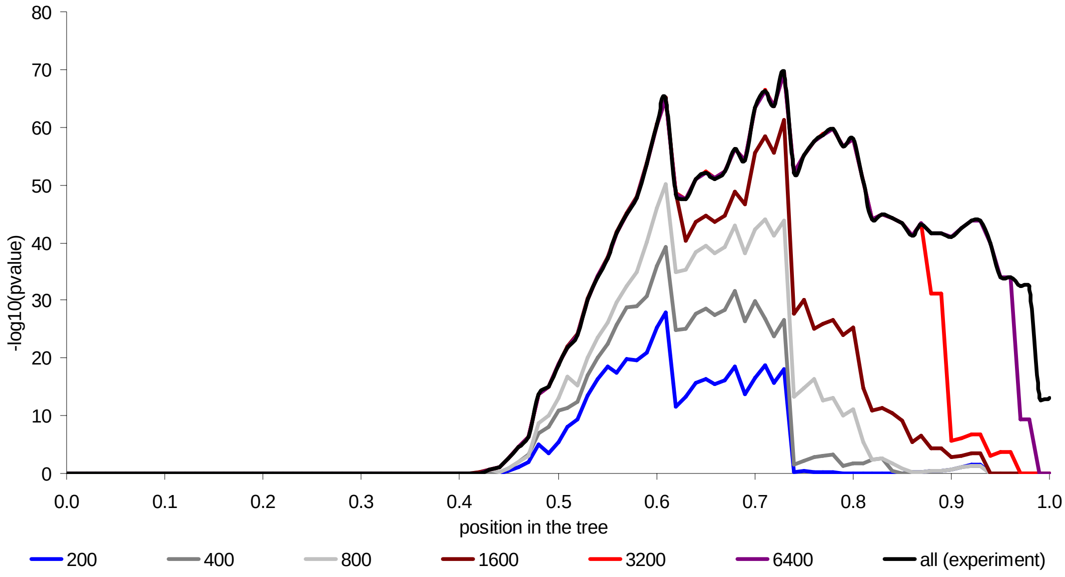


Figure 3. Calculated overall *p*-values using different statistical tests. Other parameters were: FDR-correction, 0.001 as the *p*-value cutoff, no filtering and all genes. Notations as in figure 2.

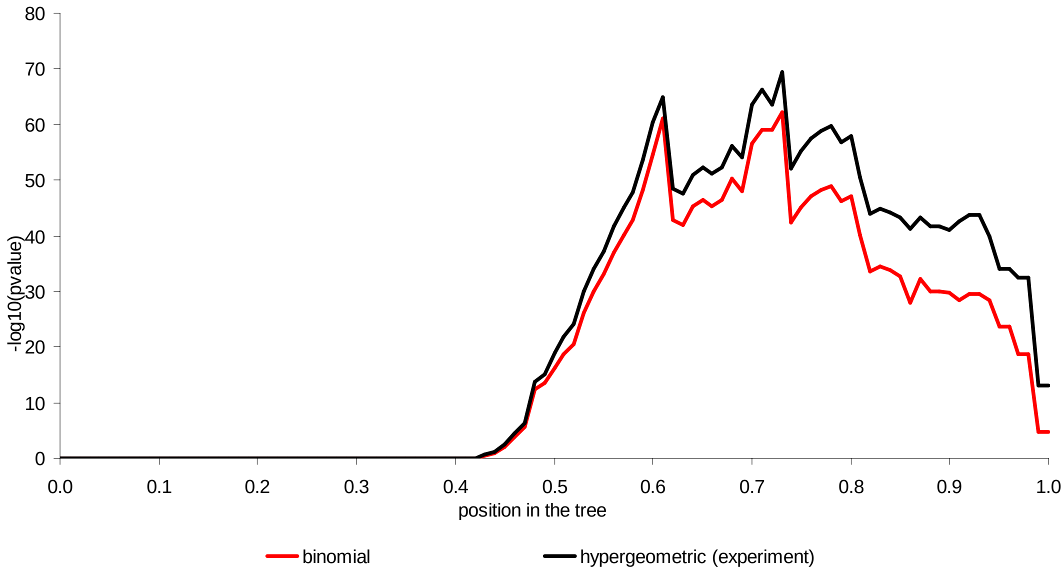


Figure 4. Calculated overall *p*-values using different *p*-value cutoffs. Other parameters were: HD, FDR-correction, no filtering and all genes. Notations as in figure 2.

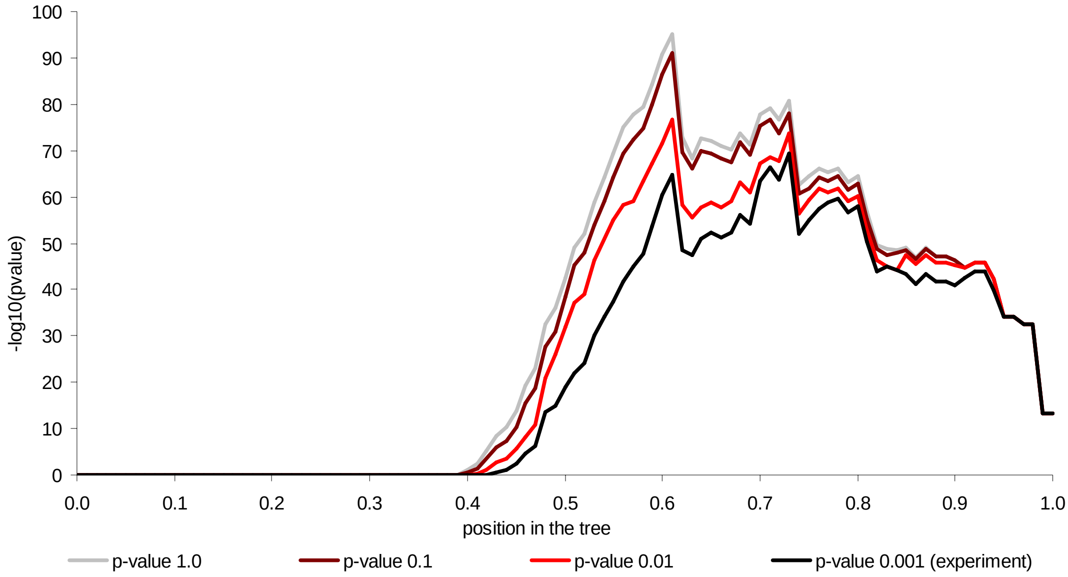


Figure 5. Calculated overall *p*-values using different multiple hypothesis correction methods. In the figure no is no correction, holm is Holm's step-down correction, bonn is Bonferroni correction and FDR is false discovery rate. Other parameters were: HD, 0.001 as the *p*-value cutoff, no filtering and all genes. Notations as in figure 2.

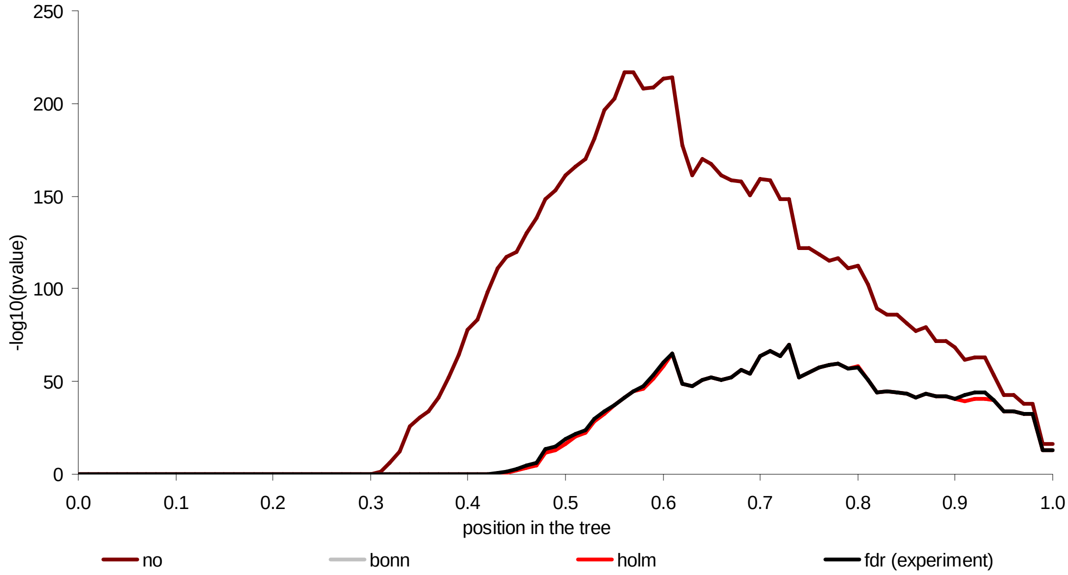


Figure 6. Calculated overall *p*-values using different filtering values. Other parameters were: In the figure 0 is no filtering, 1 is the mean plus one standard deviation and 2 is the mean plus two times the standard deviation. HD, FDR-correction, 0.001 as the *p*-value cutoff and all genes. Notations as in figure 2.

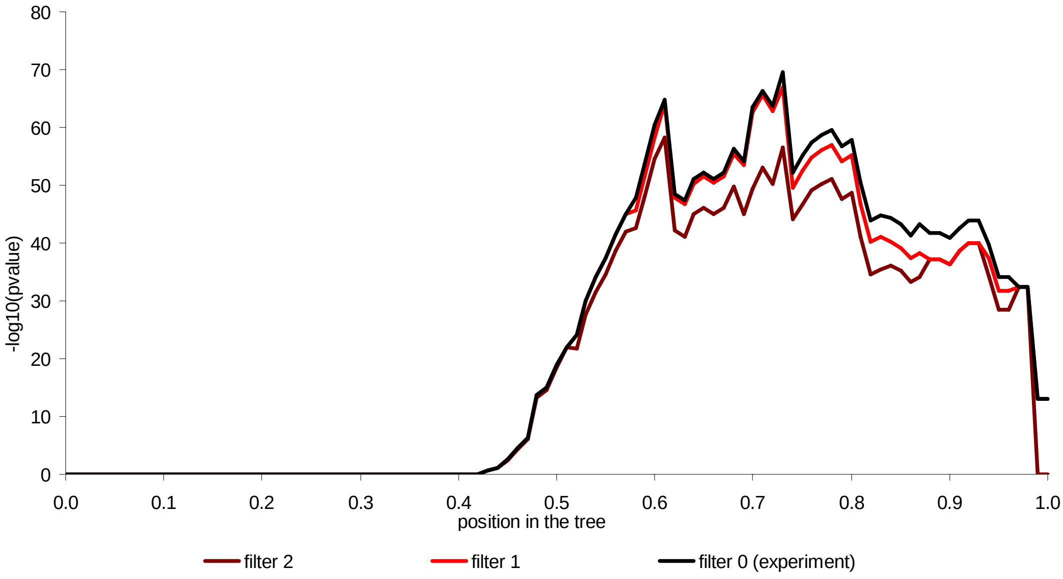


Figure 7. The most significant and the median values of the calculated overall *p*-values in randomized data. Parameters were: HD, FDR-correction, 0.001 as the *p*-value cutoff, all genes as the maximum cluster size and no filtering. Notations as in figure 2.

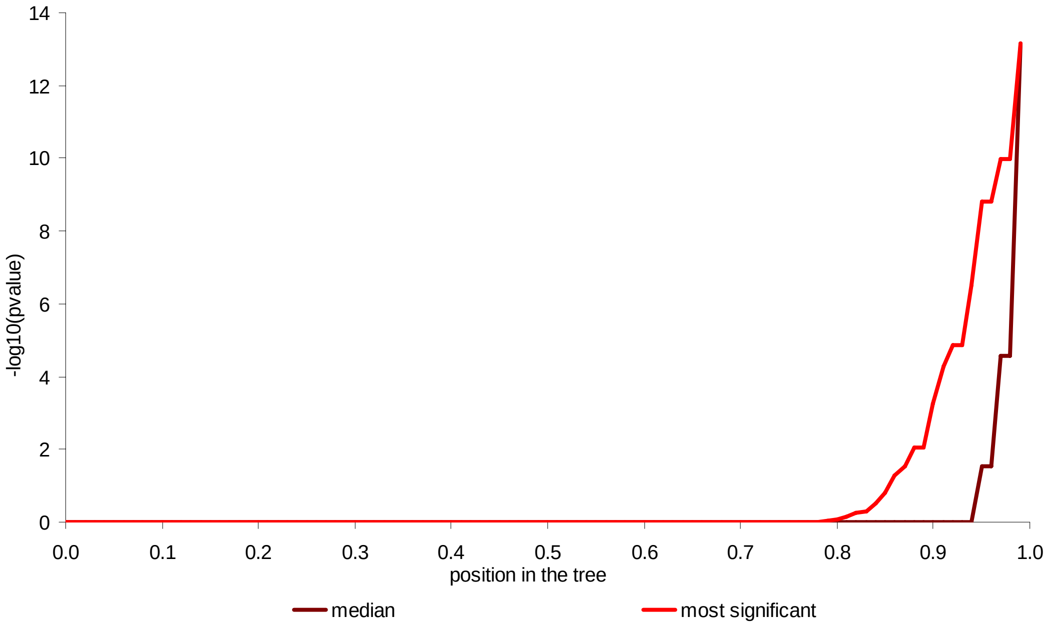


Figure 8. The expression value heatmap of the *response to wounding* cluster (GO:0009611, Node_11791, 6.E-06, genes 129). In the picture values are the log differences of the experimental pairs (the $\log_2$ value of the control subtracted from the $\log_2$ value of the corresponding sample) green corresponds to -3 and red to +3.

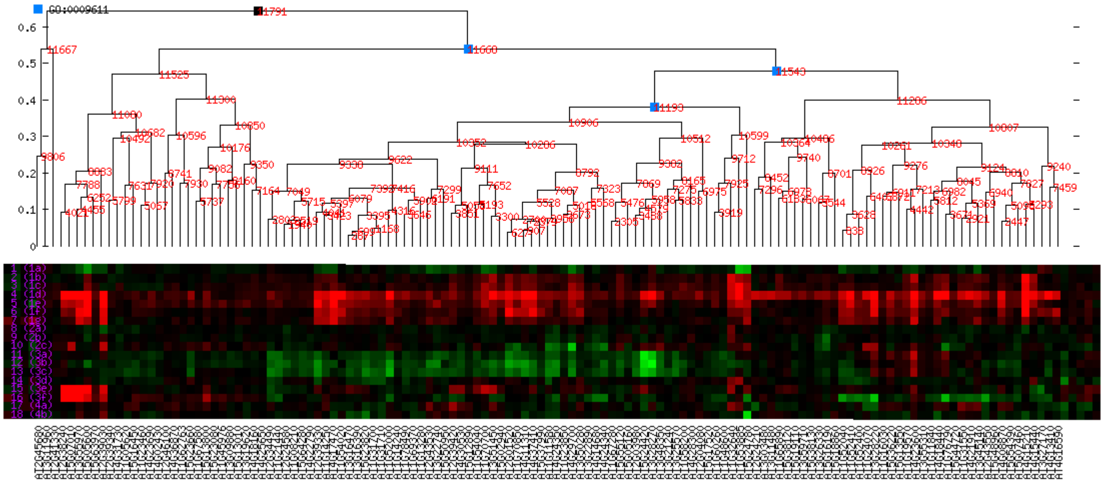


Figure 9. The expression value heatmap of the *photosynthesis* cluster (GO:0015979, Node_11801, 1.E-06, genes 80). Notations as in Figure 8.

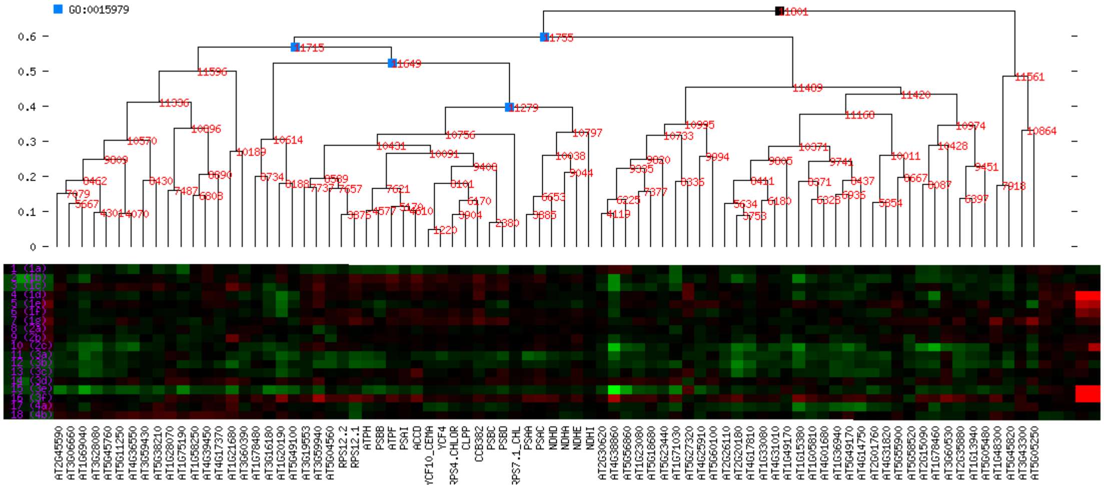


Figure 10. The expression value heatmap of the *response to chemical stimulus* cluster (GO:0042221, Node_11810, 2.E-04, genes 804). Notations as in Figure 8.

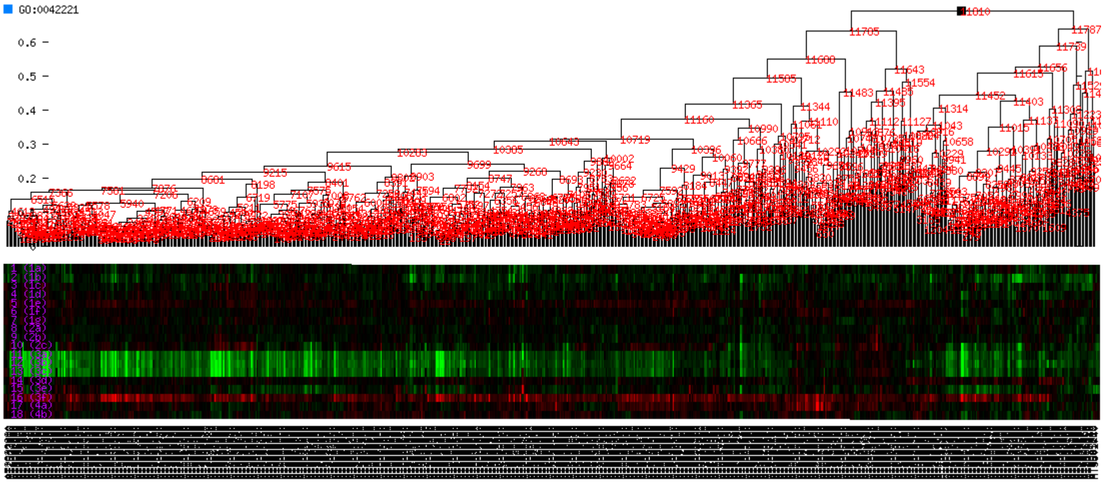


Figure 11. The expression value heatmap of the *photosynthesis* cluster (GO:0015979, Node_11822, 2.E-09, genes 1393). Notations as in Figure 8.

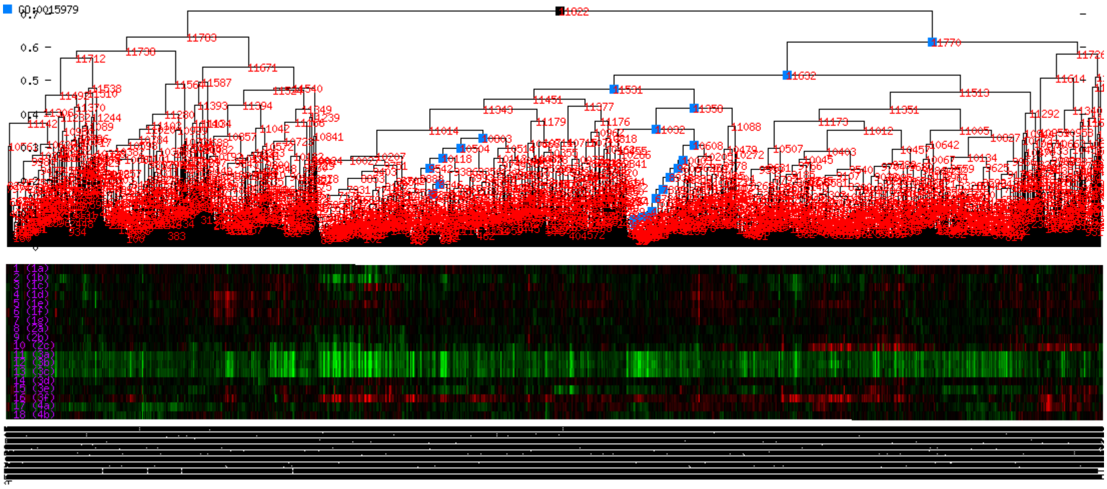


Figure 12. The expression value heatmap of the *protein catabolism* cluster (GO:0030163, Node_11800, 4.E-04, genes 180). Notations as in Figure 8.

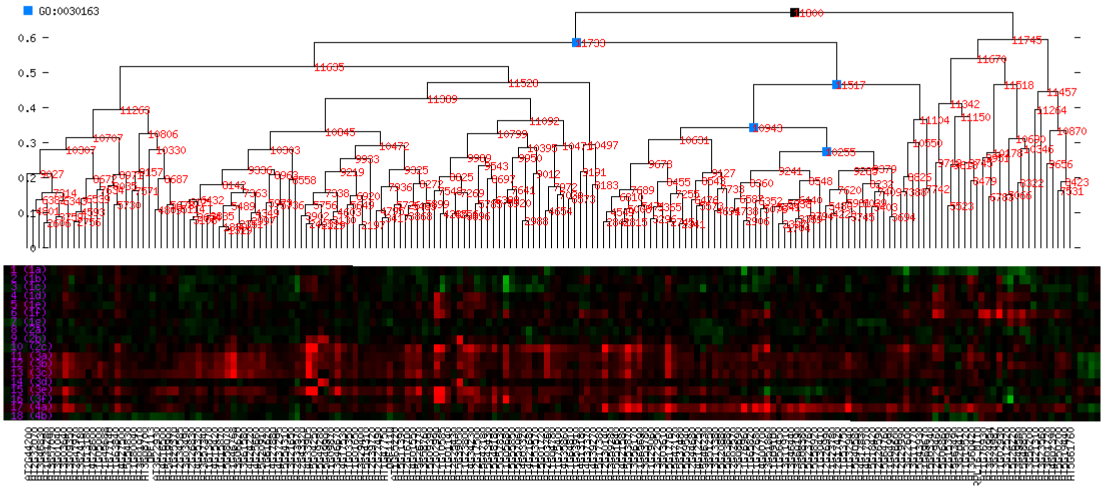


Figure 13. The expression value heatmap of the *defense response* cluster (GO:0006952, Node_11830, 3.E-10, genes 446). Notations as in Figure 8.

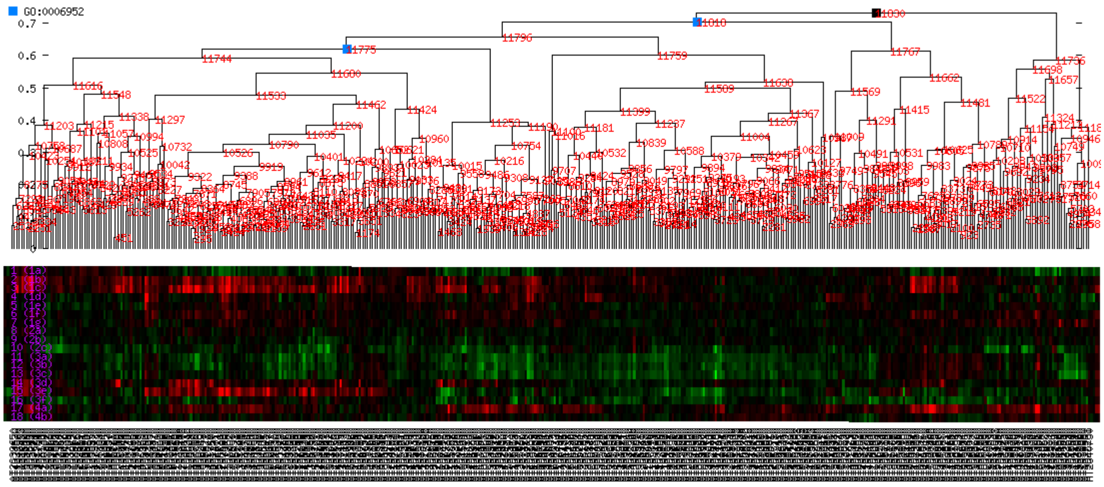


Figure 14. The expression value heatmap of the *response to heat* cluster (GO:0009408, Node_11805, 9.E-07, genes 243). Notations as in Figure 8.

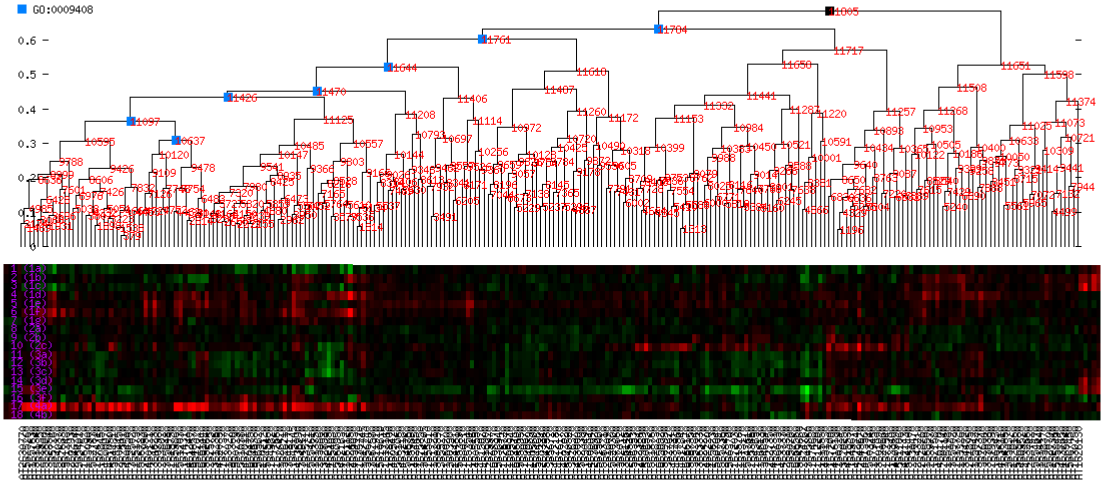


Figure 15. The expression value heatmap of the *RNA metabolism* cluster (GO:0016070, Node_11813, 2.E-06, genes 1167). Notations as in Figure 8.

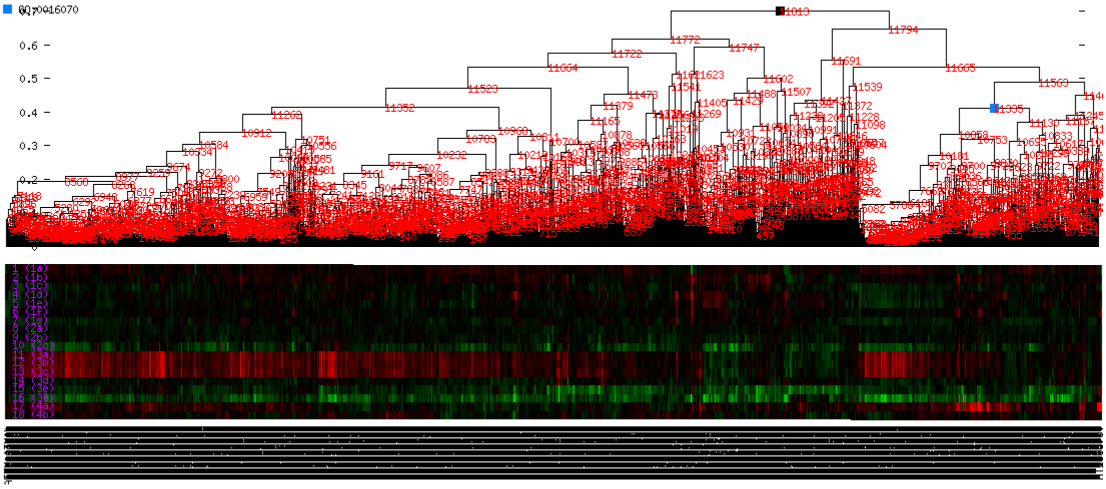


Figure 16. The expression value heatmap of the *biopolymer modification* cluster (GO:0043412, Node_11829, 9.E-04, genes 862). Notations as in Figure 8.

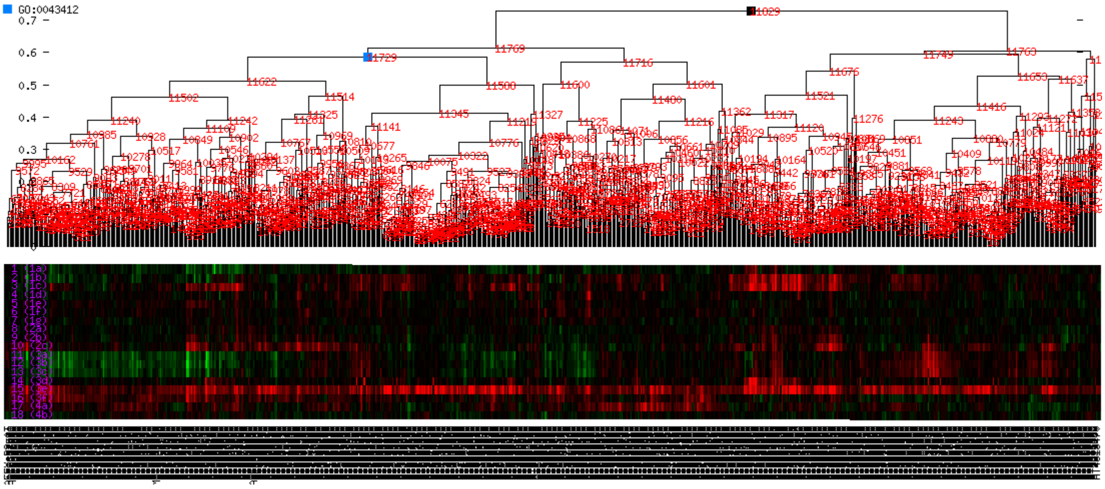


Figure 17. The expression value heatmap of the *ribosome biogenesis and assembly* cluster (GO:0042254, Node_11802, 1.E-10, genes 1664). Notations as in Figure 8.

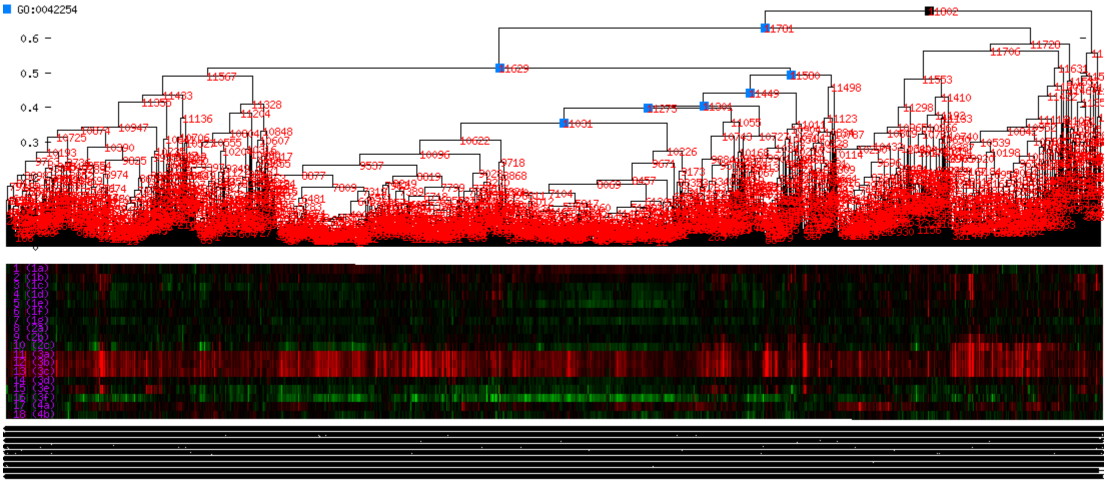


Figure 18. The expression value heatmap of the *protein biosynthesis* cluster (GO:0006412, Node_11771, 1.E-40, genes 182). Notations as in Figure 8.

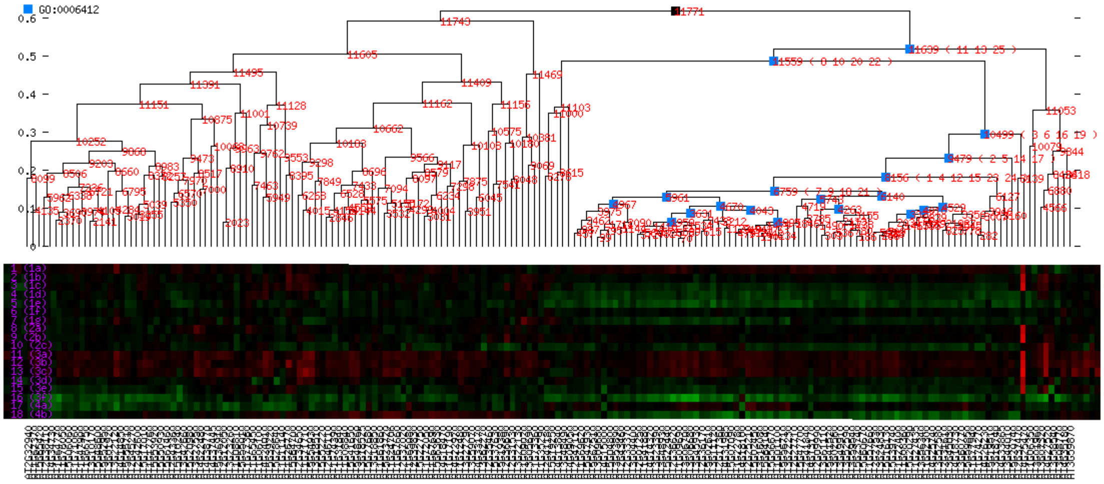


Figure 19. The expression value heatmap of the plastid *organization and biogenesis* cluster (GO:0009657, Node_11832, 1.E-07, genes 707). Notations as in Figure 8.

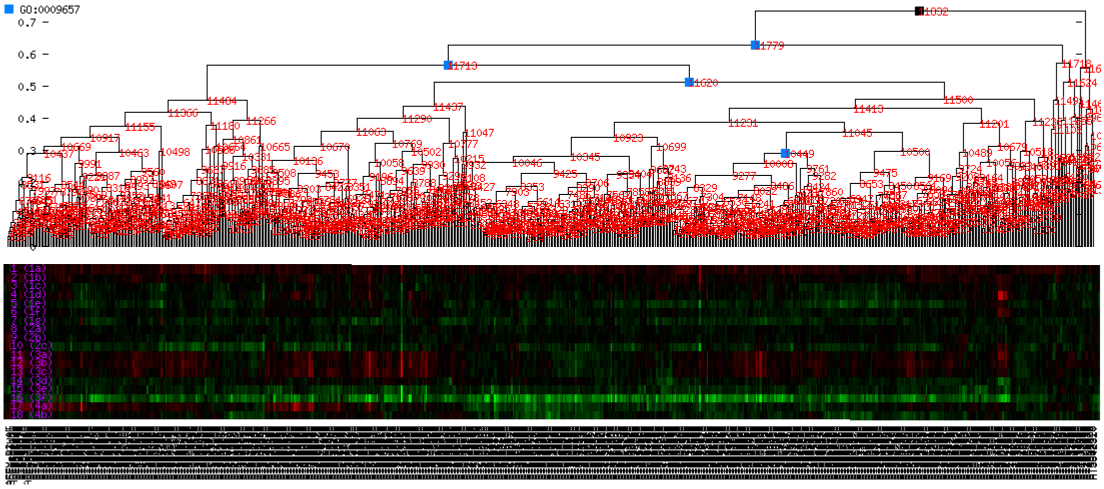


Figure 20. The expression value heatmap of the *response to abscisic acid stimulus* cluster (GO:0009737, Node_11808, 1.E-06, genes 598). Notations as in Figure 8.

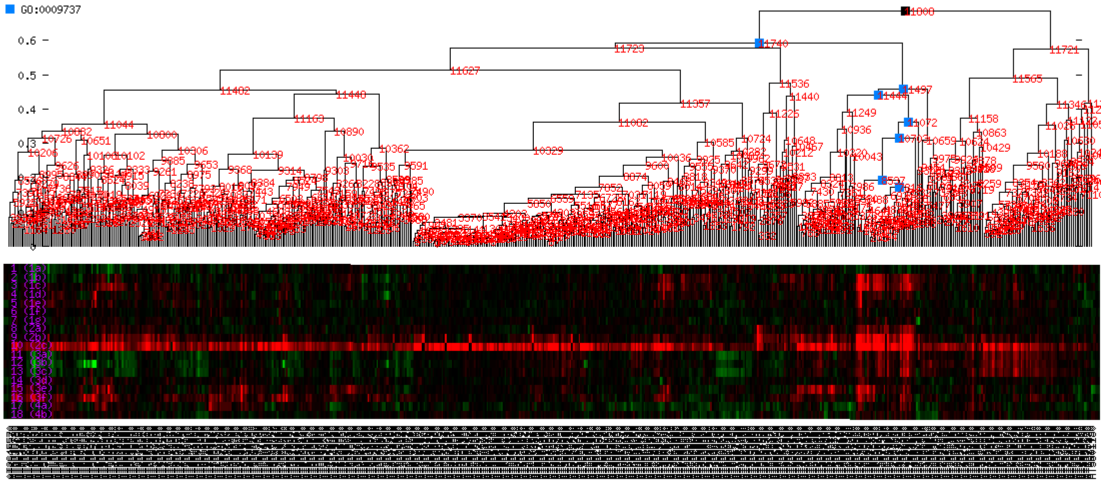


Figure 21. The expression value heatmap of the *cellular biosynthesis cluster* (GO:0044249, Node_11820, 1.E-6, genes 379). Notations as in Figure 8.

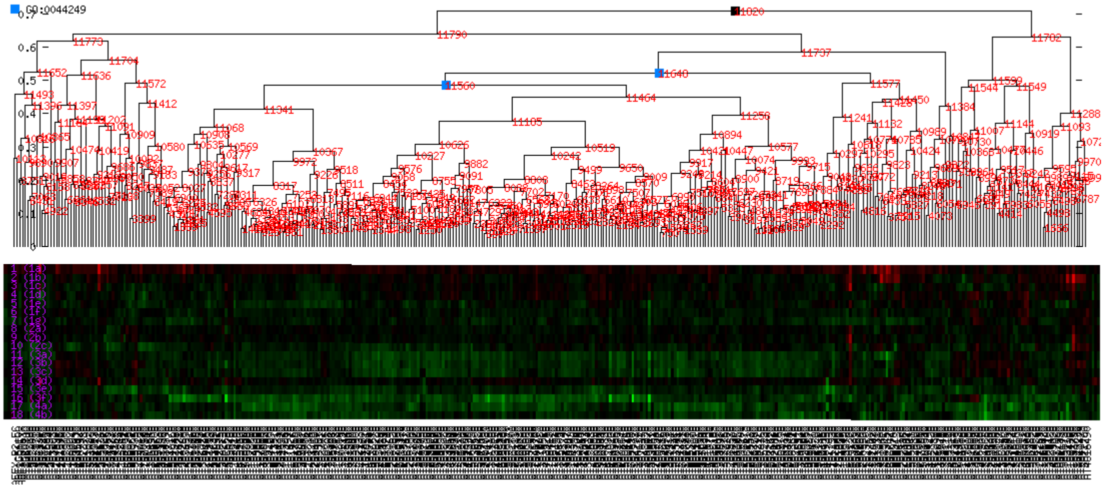
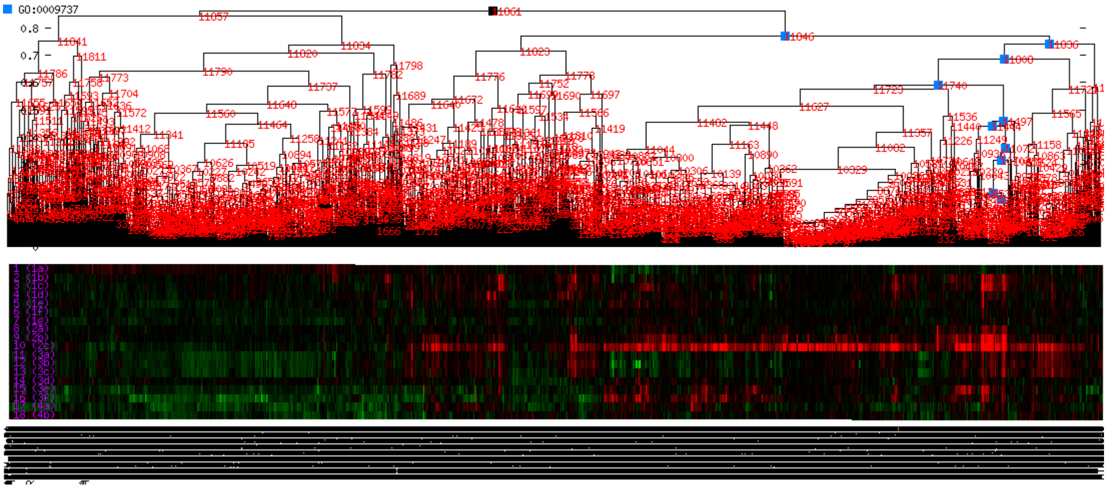


Figure 22. The expression value heatmap of the *response to abscisic acid stimulus* (GO:0009737, Node_11861, 7.E-4, genes 1348). Notations as in Figure 8.


GO:0009737
1 (1a)
2 (1b)
3 (1c)
4 (1d)
5 (1e)
6 (1f)
7 (1g)
8 (2a)
9 (2b)
10 (2c)
11 (3a)
12 (3b)
13 (3c)
14 (3d)
15 (3e)
16 (3f)
17 (4a)
18 (4b)