User Manual

Introduction:

Plant MetGenMAP is a web-based analysis and visualization package that allows the user to identify significantly changed pathways and biological processes (i.e. enriched GO terms) from gene expression and/or metabolite profile datasets and to view the profile data in the context of biochemical pathways. Biochemical pathways the system used are extracted from the BioCyc databasesor predicted using the Pathway Tools.

The lsit of the organisms and platforms currently supported by the system can be found at the homepage and the list of example input files for each organism/platform can be found at the help page.

More platforms from other plant species will be added in the future. If you have a BioCyc pathway database available and want to add a corresponding platform, please contact us.

Input Data Format

Plant MetGenMAP takes normalized and processed expression and metabolite profile data (mainly ratios and p values derived from statistical analysis programs such as LIMMA and SAM) as the input. Users can upload expression profile data and the metabolite profile data to the system separately or simultaneously. Each uploaded dataset must be in a specific format in order for it to be compatible with Plant MetGenMAP. All datasets must be in tab delimited plain text file format. The first line of the dataset must contain the labels including ID, condition names, and p values (if available) for each sample in the dataset.

ID – **Unique** identifier for each gene or metabolite included in the dataset. For gene expression dataset, the ID must be in one of the formats specified under ID type in the upload form. For metabolite profile dataset, the ID is the name of each metabolite.

Note: All available synonyms of metabolite names were collected from the corresponding BioCyc databases. A file containing these synonyms is provided on the website. The system accepts synonyms of metabolite names and converts them to metabolite common names. When preparing metabolite data, please make sure the metabolite names are consistent with the names/synonyms listed in the file.

Condition/comparison Name – The name of conditions or comparisons under which the data

(ratio or fold change) was derived. The condition/comparison names **must be unique** across the same dataset.

P value – Statistical value of probability. This is optional.

Here is an example of datasets with p values

Example:

```
ID[TAB]blue light[TAB]p value[TAB]red light[TAB]p value[TAB]white light[TAB]p value
254685_at[TAB]-45.7[TAB]0[TAB]-19.29[TAB]0[TAB]-11.89[TAB]0
263151_at[TAB]-1.89[TAB]0[TAB]-1.15[TAB]0.83[TAB]-1.24[TAB]0.36
```

Here is an example of datasets without p values

Example:

```
ID[TAB]blue light[TAB]red light[TAB]white light
254685_at[TAB]-45.7[TAB]-19.29[TAB]-11.89
263151_at[TAB]-1.89[TAB]-1.15[TAB]-1.24
```

More example input files can be found on the site.

Note: There is no limit to the number of conditions/comparisons or p values included in a dataset. However, datasets with p values must have p values for all conditions/comparisons within the dataset. In addition, corresponding expression and metabolite datasets must have the same header and format.

Data Uploading and Processing

Obtaining an account

In order to upload and analyze a dataset, you ***MUST*** have an account in the system. To obtain a new account, fill out the registration form accordingly. A confirmation email will then immediately be sent to the email address provided. To complete the registration process, the user must activate the account by clicking on the link provided in the confirmation email. Once the account is activated, the user can login to upload your dataset.

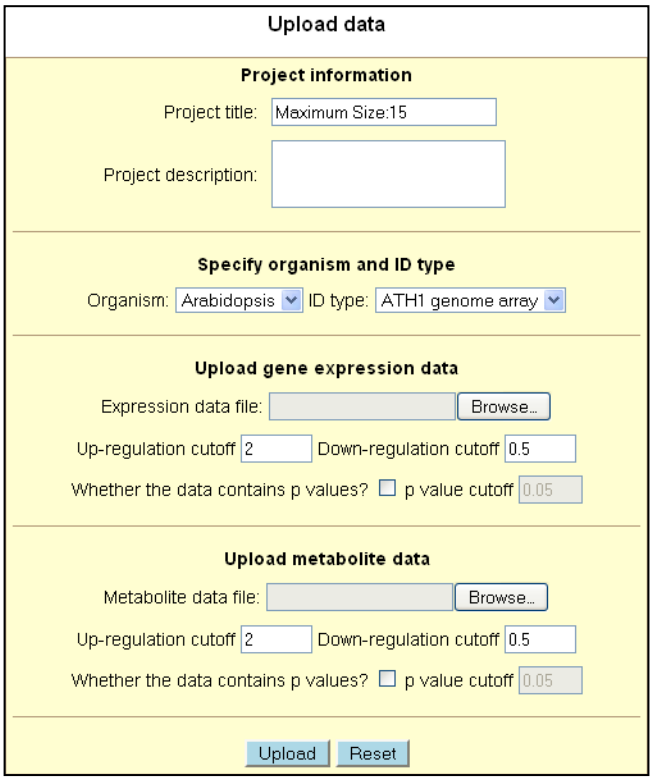
Accessing the Upload Form

1. *For existing users who have datasets stored in the system:*
   While signed into a user account, mouse over the “Data Manage” tab in the menu bar to access the drop down menu. Click on the “Upload data” option.
2. *For new users and users with no datasets loaded in the system:*

Upon signing into a user account, there will be a prompt asking the user to upload a dataset. Or the user can use the same process as described above to access the upload form.

Data uploading

Each field of the upload form is explained below in detail:



1. *Project information*

a) *Project title* – Title of the project. Cannot exceed 15 characters.

b) *Project description* – Description of the project. There is no limit to the number of characters in the description.

2. *Specify organism and ID type*

a) *Organism* – Choose the organism from which the dataset was generated. The options currently include *Arabidopsis*, rice and tomato.

b) *ID type* – Choose the identification format used in the gene expression dataset. The options for this parameter are dependent on which organism is chosen. Options for Arabidopsis include: ATH1 genome array and TAIR locus number (e.g., AT1G01040). Options for rice include: Affymetrix genome array and the genome locus number (e.g., LOC_Os10g33000). Options for tomato include: TOM1 cDNA array (e.g., 1-1-7.4.19.9), TOM2 oligo array (e.g., LE3D02), Affymetrix genome array and SGN unigene (version: Tomato_200607_build_1; e.g., SGN-U314663).

Note to tomato TOM2 oligo array users: Due to the multiple printing formats of TOM2 arrays, the system has stopped supporting the previous probe ID system for TOM2 arrays and started to use the original Plate IDs. Please contact us if there are any questions.

3. *Upload gene expression data*

a) *Expression data file* – Enter the path and file name of the expression data file to be uploaded. The file must be in the correct format as indicated above.

b) *Up-regulation cutoff* – Sets minimum value that characterizes a gene as significantly up-regulated. Default value is 2.

c) *Down-regulation cutoff* – Sets maximum value that characterizes a gene as significantly down-regulated. Default value is 0.5. If the values in the dataset are fold changes or log2 transformed ratios, a negative value is needed as a cutoff.

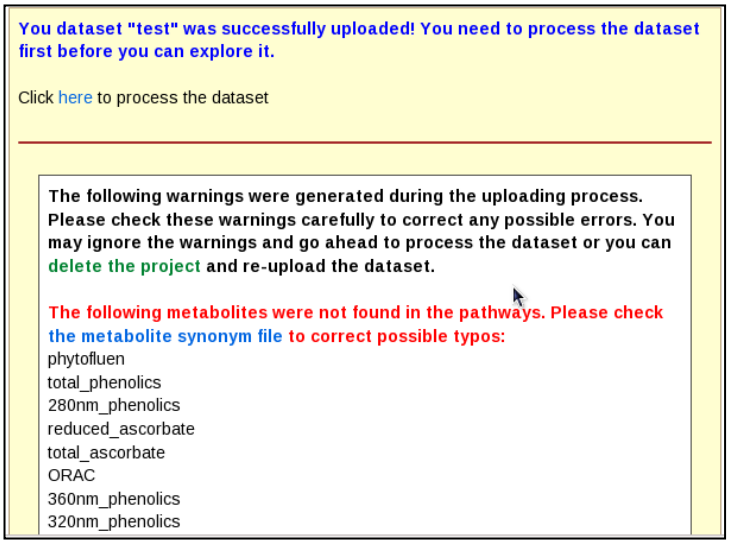
d) *Whether the data contains p values* -- Indicates whether or not the dataset contains p values. For datasets without p values, leave the check box blank.

e) *p value cutoff* – Sets value as maximum for significance. Default value is 0.05.

4. *Upload metabolite data*
   This part is largely identical to the above section - "Upload gene expression data".

The user can upload both expression and metabolite profile data generated under one project at the same time. Just leave the other "path and file name" field blank if only one file is uploaded.

During the uploading process, the system checks whether the gene/metabolite identifiers are consistent with the ones stored in the database. For metabolites, the system accepts all synonyms listed in the metabolite synonym file described above. Warnings could be generated during the uploading process if the identifiers were not found in the database. For metabolites, this could be due to that these metabolites are not in the pathways. You can ignore these warnings and the system will not include the unmatched genes/metabolites in the downstream analyses.



Note: Please check your file carefully to make sure no typo errors in gene/metabolite identifiers.

Processing Datasets

The uploaded datasets ***CANNOT*** be analyzed or explored until they are processed. Immediately after the dataset is uploaded, the system will ask the user to process the dataset. Processing can also be done at a later time at the "Project Management" page which lists the uploaded datasets in the system under the current user (for more information, see the "Project management" section). Datasets not yet processed have an option to be processed under its list of actions.

The data processing step will assign a code to each gene under each condition to indicate whether the expression of this gene is increased, decreased, or unchanged. The step will also identify changed pathways under each condition and calculate the significance of the change.

Significance of changed pathways

The significance of a changed pathway is determined using the hypergeometric distribution:

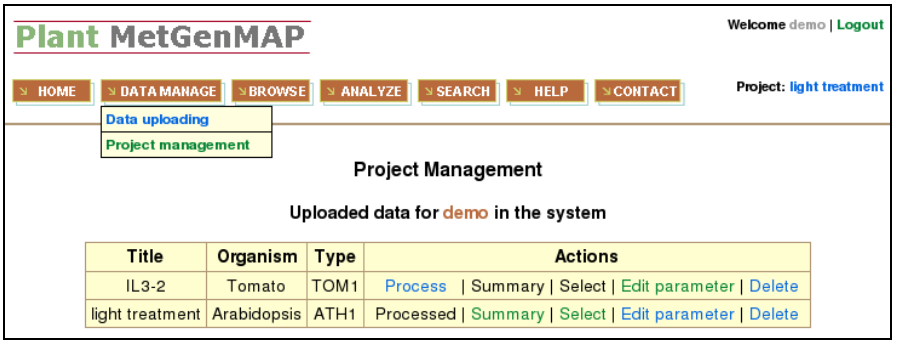
$$p\_value = \sum_{j=x}^{n} \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}$$

Where N is the total number of genes/metabolites in all the pathways, M is the total number of genes/metabolites in a particular pathway, n is the total number of significantly changed genes/metabolites in all the pathways, and x is the total number of significantly changed genes/metabolites in that particular pathway.

The p value obtained above can be explained in the following way:
Suppose that we have a total of N genes in all the pathways, and M genes belong to a particular pathway. Then the p value represents the possibility that, in a sample of n changed genes of all the pathways, we observe x or more changed genes in that particular pathway.

Project Management

To access the project management page after logging in, mouse over the “Data Manage” tab in the menu bar and select the “Project Management” option. A list of all available projects for the current user is shown. For each project, a list of action links is provided.



- *Process* – Allows the user to process the dataset. For more info on processing datasets,

see the section of “Processing Datasets” described above. If the dataset has already been processed, the message “Processed” appears instead of the link.

- *Summary* – Provides a summary of the project – the number of up- and down-regulated genes/metabolites in each condition, with each number linked to the list of the corresponding genes/metabolites. The “Summary” action is only active when the dataset has been processed.

Summary of dataset "light treatment"

Condition	Data type	# up-regulated	# down-regulated
AL	Expression	225	61
AS	Expression	346	119
BL	Expression	1058	696
BS	Expression	202	146
EL	Expression	1192	774

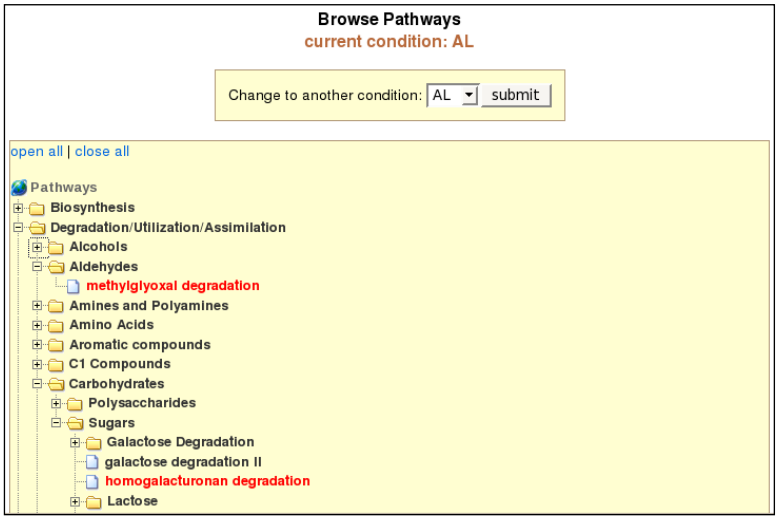
- *Select* – Selects a particular project as the current working project for analysis. Once selected, the name of the selected project appears on the top right corner of the menu bar as “*Project: XXX*”. This is visible in the menu bar no matter what page the user is on and is a useful reminder of which dataset is being explored. The “Select” action is only active when the dataset has been processed.
  Note: The “Browse”, “Analyze”, and “Search” menu options are only functional if a project dataset is selected.

- *Edit parameter* – Allows users to change the parameters for a particular dataset (e.g., up- or down-regulation cutoff, p value cutoff) which are set at the “Data uploading” step. The dataset must be re-processed after changing parameters.

- *Delete* – Permanently deletes the project from the system. Users will be asked to confirm the deletion.

Browsing Pathways

The “Browse” option on the menu bar allows users to observe or examine different biochemical pathways based on the selected expression dataset and/or metabolite dataset.



All Pathways

Selecting “All Pathways” shows all biochemical pathways for a particular condition in a folder tree format organized by the type of pathways. Options to open and close all folders are also provided. Pathways in black indicate no changes at all, while those in red indicate a change in gene expression/metabolite content.

Changed Pathways

Selecting “Changed Pathways” shows only the pathways that have any overall changes in gene expression and/or metabolite content, as well as the corresponding p values for each pathway indicating how significant the pathway changes.

Changed pathways current condition: BL

Change to another condition: AL submit

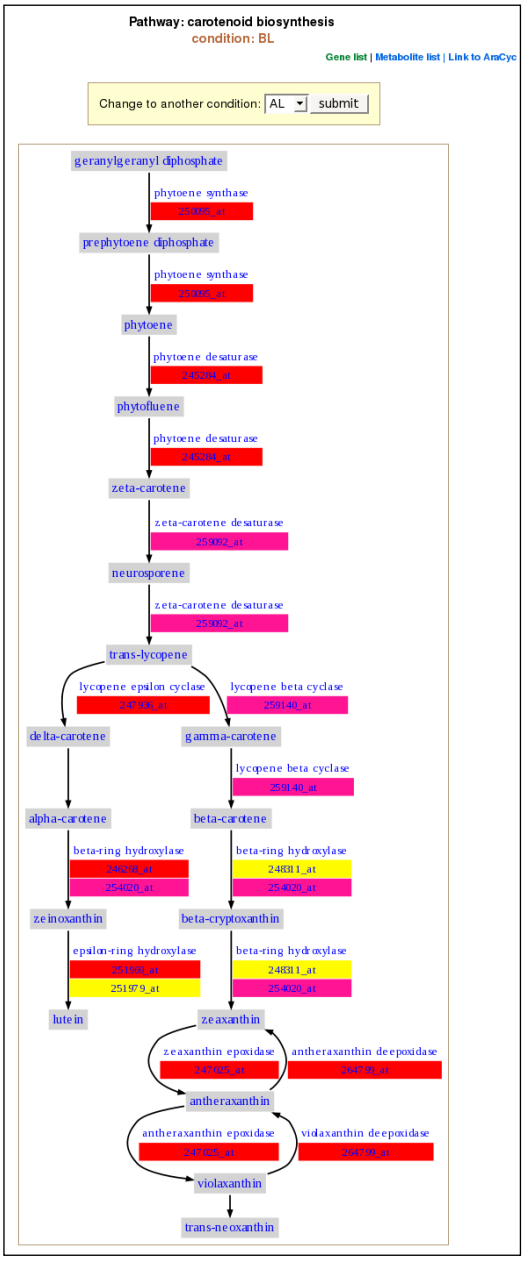
Expression data

Number	Pathway name	p value
1	photosynthesis, light reaction	5.91604e-23
2	photosynthesis	3.23026e-18
3	chlorophyllide *a* biosynthesis	2.43592e-06
4	carotenoid biosynthesis	0.000839859
5	trans,trans-farnesyl diphosphate biosynthesis	0.00339933
6	geranyldiphosphate biosynthesis	0.00339933
7	Calvin cycle	0.00368543
8	phylloquinone biosynthesis	0.00381199
9	kaempferol glucoside biosynthesis (Arabidopsis)	0.00687927
10	leucine degradation	0.0090941
11	polyisoprenoid biosynthesis	0.0139647
12	photorespiration	0.0198821
13	anthocyanin biosynthesis (pelargonidin 3-O-glucoside, cyanidin 3-O-glucoside)	0.0244792
14	xanthophyll cycle	0.0244792

Pathway details

In both “Browse All Pathways” and “Browse Changed Pathways” described above, each pathway is linked to its detail page. A pathway detail page includes a graphical visualization of the biochemical pathway with each molecule/metabolite highlighted with different colors to indicate the changes of gene expression or metabolite levels. Links to the list of all genes and metabolites in the pathway, as well as the link to the pathway in the original BioCyc database, are also provided in this page. Genes and metabolites in the pathway image are linked to pages containing their detail information.

Color legend



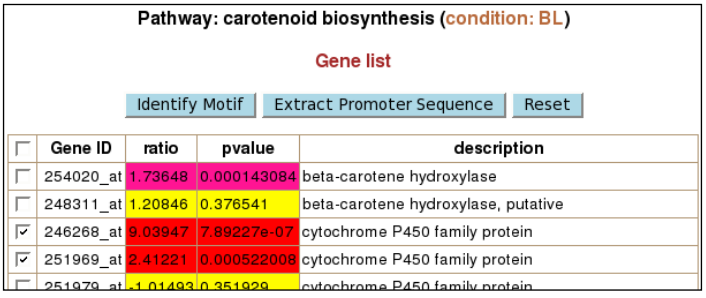
Plant MetGenMAP uses the following color legend to indicate differences of gene expression/metabolite level changes:

1) Uploaded dataset contains p values.
   - the ratio is greater than the up-regulation cutoff and the p value is significant
   - the ratio is less than the down-regulation cutoff and the p value is significant
   - the ratio does not meet either the up-regulation or down-regulation cutoff while the p value is significant
   - the ratio is greater than the up-regulation cutoff while the p value is not significant
   - the ratio is less than the down-regulation cutoff while the p value is notbsignificant
   - the ratio does not meet either the up-regulation or down-regulation cutoff and the p value is also not significant
   - not measured or filtered out

2) Uploaded dataset does not contain p values.
   - the ratio is greater than the up-regulation cutoff
   - the ratio is less than the down-regulation cutoff
   - the ratio does not meet either the up-regulation or down-regulation cutoff
   - not measured or filtered out

Promoter analysis

It has been reported that a subset of genes in the same pathway are regulated by common transcription factors. Plant MetGenMAP provides a tool to identify over-represented motifs from the promoters of co-expressed genes in the same pathway. In the

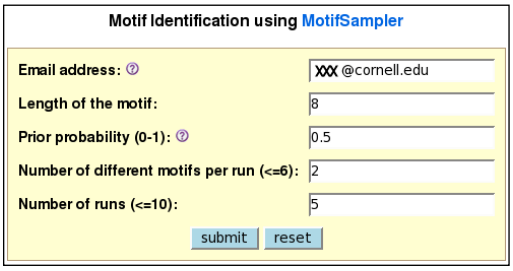


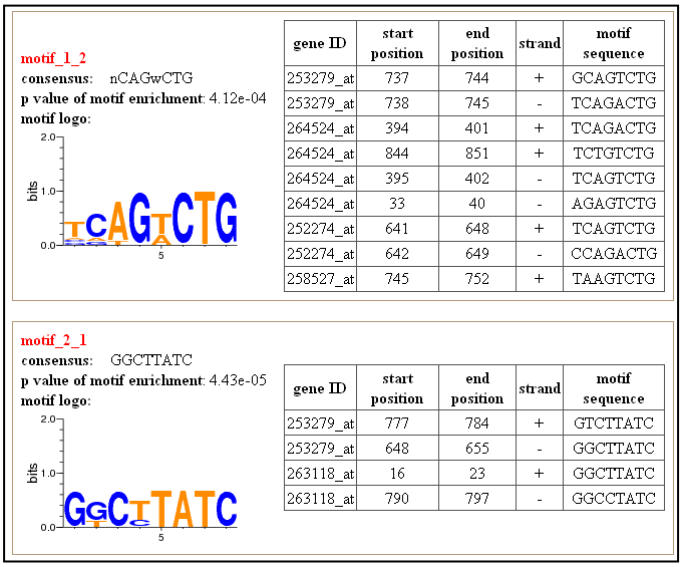
Gene ID	ratio	pvalue	description
254020_at	1.73648	0.000143084	beta-carotene hydroxylase
248311_at	1.20846	0.376541	beta-carotene hydroxylase, putative
246268_at	9.03947	7.89227e-07	cytochrome P450 family protein
251969_at	2.41221	0.000522008	cytochrome P450 family protein
251979_at	-1.01493	0.351929	cytochrome P450 family protein

"Gene list" page of each pathway, the user can extract the promoter sequences of a set of genes in the pathway (default: genes significantly up- or down-regulated) for his/her own analysis. The user can also identify motifs using the tool provided by the system (see below).

Motif identification

Plant MetGenMAP uses MotifSampler to identify motifs from the promoter sequences of co-expressed genes in a particular pathway. The user can specify parameters using the input form the system provides. For a detail explanation of these parameters, please check MotifSampler homepage.





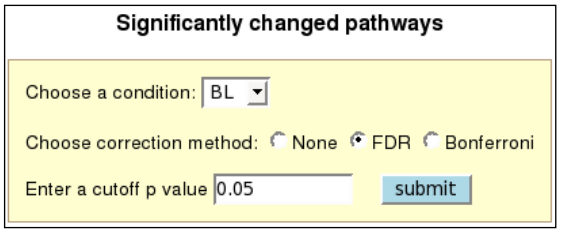
After motifs are identified by MotifSampler, each motif is screened against all promoter sequences of the organism and against the promoters of the list of co-expressed genes, respectively, using PatMatch, to identify the occurrences of the motif. Then the post-hoc test is performed to calculate p value of the motif enrichment based on the hypergeometric distribution.

After submitting the job, the job will run in background. Once the job is finished (it can take several minutes), a link to the output page will be sent to the email address the user provided (default is the email the user provided during the registration). Motifs and their sequence logos, p value of the motif enrichment, as well as the positions of the motif in the promoter of each gene, are included in the output page.
Note: Currently this tool is only available for datasets generated from Arabidopsis and rice.

Analyzing Dataset

Plant MetGenMAP provides three tools to further explore and analyze the uploaded datasets: 1) Identify significantly changed pathways; 2) Identify enriched Gene Ontology (GO) terms; 3) Functional classification of a list of genes.

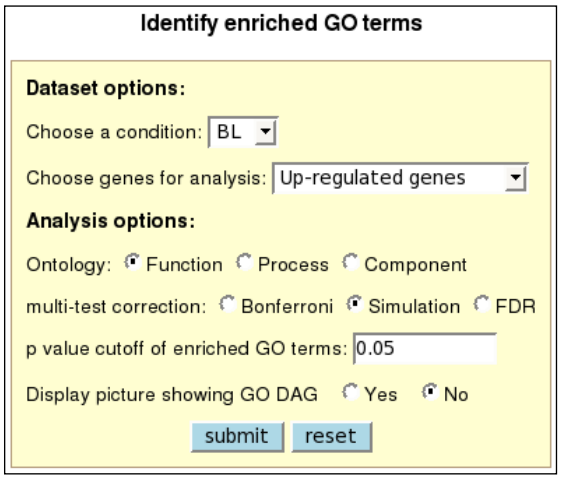
Identify Significantly Changed Pathways

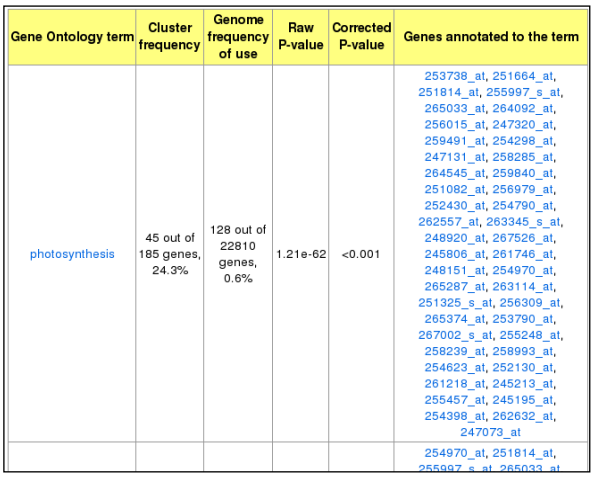


Since in each condition, the statistical test (hypergeometric test) to check the significance of pathway changes is performed on a set of pathways simultaneously, the raw p values need to be corrected for multiple testing. Plant MetGenMAP provides the most commonly used multiple test correction method - False Discovery Rate (FDR), as well as Bonferroni correction method. Based on the user specified parameters, a list of significantly changed pathways is returned.

Identify Enriched GO Terms

This tool allows the user to identify over-represented (enriched) GO terms from a list of up- and/or down-regulated genes under a specific condition. The tool was implemented based on the GO::TermFinder perl module (Boyle et al., 2004) which uses the hypergeometric distribution to calculate the significance of GO term enrichment. The tool provides three types of multiple test correction methods that come with the GO::TermFinder module: FDR, Simulation, and Bonferroni. Significantly enriched GO terms, genes annotated to each corresponding GO term,



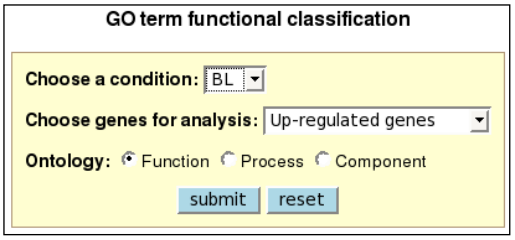


Gene Ontology term	Cluster frequency	Genome frequency of use	Raw P-value	Corrected P-value	Genes annotated to the term
photosynthesis	45 out of 185 genes, 24.3%	128 out of 22810 genes, 0.6%	1.21e-62	<0.001	253738_at, 251664_at, 251814_at, 255997_s_at, 265033_at, 264092_at, 256015_at, 247320_at, 259491_at, 254298_at, 247131_at, 258285_at, 264545_at, 259840_at, 251082_at, 256979_at, 252430_at, 254790_at, 262557_at, 263345_s_at, 248920_at, 267526_at, 245806_at, 261746_at, 248151_at, 254970_at, 265287_at, 263114_at, 251325_s_at, 256309_at, 265374_at, 253790_at, 267002_s_at, 255248_at, 258239_at, 258993_at, 254623_at, 252130_at, 261218_at, 245213_at, 255457_at, 245195_at, 254398_at, 262632_at, 247073_at
					254970_at, 251814_at, 255997_s_at, 265033_at

and the GO DAG picture (optional) are included in the output page.

Note: This program may take several minutes if the number of input genes is large.

Gene functional classification



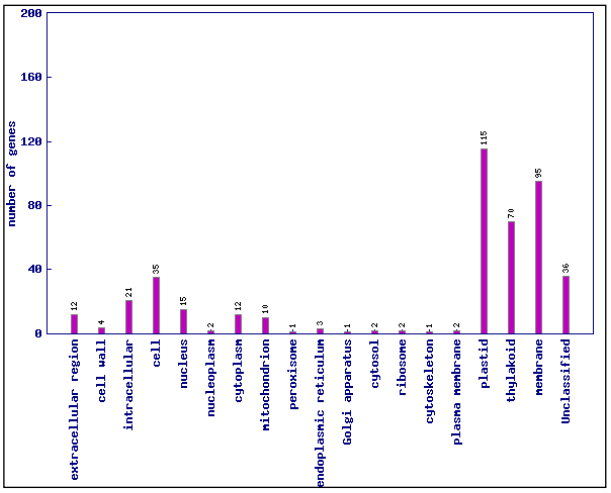
This tool classifies a list of interesting genes into different functional categories based on their GO annotations. The functional catalogue is based on the

Plant GO Slim, a cut-down version of the plant GO ontologies containing a subset of the terms in the whole GO. These GO slims give a broad overview of the ontology content without the detail of the specific fine grained terms. From a list of up- and/or down-regulated genes under a specific condition, the tool outputs the number of genes in each GO slim category with each number linked to the full list of genes belonging to the corresponding category. A clickable bar graph is also provided. It is worth noting that **one gene could be assigned to several different GO slims**, so the sum of genes assigned to each category could be more than the number of total input genes.

GO term functional classification

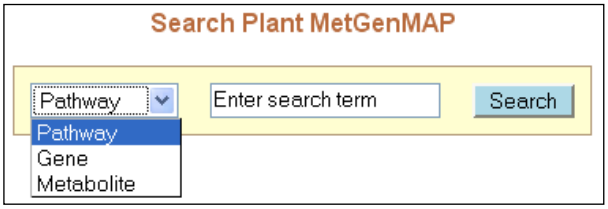
Your total input gene: 225

GO term ID	Description	Category	# of genes
GO:0005576	extracellular region	Cellular Component	12
GO:0005618	cell wall	Cellular Component	4
GO:0005622	intracellular	Cellular Component	21
GO:0005623	cell	Cellular Component	35
GO:0005634	nucleus	Cellular Component	15
GO:0005654	nucleoplasm	Cellular Component	2
GO:0005737	cytoplasm	Cellular Component	12
GO:0005739	mitochondrion	Cellular Component	10
GO:0005777	peroxisome	Cellular Component	1
GO:0005783	endoplasmic reticulum	Cellular Component	3
GO:0005794	Golgi apparatus	Cellular Component	1
GO:0005829	cytosol	Cellular Component	2
GO:0005840	ribosome	Cellular Component	2
GO:0005856	cytoskeleton	Cellular Component	1
GO:0005886	plasma membrane	Cellular Component	2
GO:0009536	plastid	Cellular Component	115



Searching

The Plant MetGenMAP search function allows users to look for information on specific terms within the selected dataset.



Users can search for three types of terms:

- Pathway – MetGenMAP will return a list of pathways that contain the entered search string. Each result entry includes a link to a more detailed pathway page.

- Gene – A list of genes containing the entered search string is returned, with genes in pathways highlighted in blue background. Each result entry includes a link to a more detailed gene annotation page.

- Metabolite – The search function will return a list of metabolites containing the entered search string. Each result entry includes a link to a more detailed metabolite page.

Adding New Platforms

New species and platforms can be easily added into our system if the pathway databases created using the Pathway Tools are available for the corresponding organisms. Gene annotations should also be available; or gene sequences must be provided so we can run our annotation pipeline against the sequences. Please contact us if you want your platforms to be included in our system.

Contacts

Any questions and comments please contact us at bioinfo@cornell.edu