OAT: Planteome Ontology Enrichment Analysis Tool

Botong Qu [1], Justin Elser [2], Jaden P. Diefenbaugh [1], Laurel Cooper[2], Seth Carbon[3], Chris Mungall[3], Pankaj Jaiswal[2,*], Eugene Zhang [1,*]

[1] School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, 97331, USA

[2] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, 97331, USA

[3] Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

* Corresponding author
Tel: 1-541-737-8471; Fax: 1-541-737-3573;
Email: jaiswalp@science.oregonstate.edu

Abstract

Ontology enrichment analysis of a set of genes helps biologists to identify the potential biological functions associated with a set of interesting genes. As the plant ontology categories expanding, the enriched ontology terms associated with genes become more informative since it provides knowledge from aspects of crossing categories. We introduce a tool to help biologists to discover such enriched ontology terms from the growing comprehensive ontology annotation database -- Planteome. To assist the analysis, we provide a gene annotation enrichment analysis tool which uses statistical methods to analyze all annotation data. Besides, the tool visualizes the results in three ways: 1) Highlight the enriched terms among a force-directed network graph. 2) Construct a hierarchical graph by preserving the hierarchical relationships between terms. 3) Display the correlations among enriched ontology terms and interesting genes by a matrix view.

Introduction

Gene annotations are analyzed and explored by gene curators from all over the world. Finding and visualizing the useful information from the annotations has been a hot topic for decades. The Common Reference Ontologies and Applications for Plant Biology [1] benefit biologists to enable discovery of enriched biological ontology terms among all provided ontologies (Gene Ontology (GO), Plant Ontology (PO), Trait Ontology (TO), Plant Environmental Conditions Ontology (PECO), etc.). Utilizing this comprehensive database, biologists can discover enriched biological ontology terms among all provided ontologies. This feature can help the biologist to find potential correlations between different categories of biological functions.

The Ontology Enrichment Analysis Tool (OAT) supports two ways to conduct the ontology enrichment analysis: using the planteome database as background, or the users own supplied annotations. This allows users to not be limited to whatever data happens to be loaded in the Planteome database at that time and for repeatable data sets.

OAT supplies two main statistical methods to calculate the significant p-values for each ontology terms associated with the input gene list, i.e. the Fisher's exact test and chi-square test. Besides, Yate's correction and hypergeometric distribution are supported by OAT as well. Users can select a different statistical method based on their accuracy requirement and size of the samples.

OAT allows users to input either the name or the synonyms when querying from the Planetome database. However, to get correct analysis results, OAT requires a procedure for targeting the interesting genes when the input strings are ambiguous. All the enriched terms will be shown in a table for further study.

After finding the significantly enriched ontology terms from supported ontology categories, such as GO: Cellular Component, GO: Molecular Function, GO: Biological Process, PO: Plant Anatomy: PO: Plant Growth and Development Stage, TO, and PECO, the OAT provides three types of visualizations to help users to intuitively analyze the enriched terms. The force directed network visualization is the most intuitive way to display the network structure of the enriched ontology terms. To emphasize the hierarchical structure among ontology terms, a hierarchical visualization is provided. And last, to better study the correlation among the interesting gene list and enriched ontology terms, a matrix view is provided by OAT.

In the next section, this paper will talk about the enrichment analysis used in OAT. Then, we will start to talk about the visualization methods we applied in section 3. In section 4, we will show the interface of OAT and discuss the use of it. At last, we will discuss the future work of OAT.

Ontology Enrichment Analysis

Given an annotation database, there are a number of genes (or QTL or germplasm) and their corresponding annotation information. This is also called the background data, which is the associations between genes and ontology terms.

When the biologists find an interesting gene list, they may want to know what feature or underlying biological functionalities (ontology terms) are related to this set of genes. The problem is every gene could be associated with more than one ontology terms, and most associated ontology terms could not represent the main function of this gene list. So the ontology terms which are enriched by the interesting set of gene list would become necessary. The key is finding the ontology terms which are associated with the interesting gene list and are not selected by chance. In other words, the founded ontology term should be overrepresented by the gene list. The procedure of finding out the enriched ontology terms based on the selected gene list is called enrichment analysis.

When studying the enriched ontology terms, biologists normally would like to conduct the analysis from two types of databases, i.e. a specific list of user's interesting annotations or the database provided by the system. OAT provides methods for both analyses to satisfy these two types of requirements.

OAT allows users to find the enriched ontology terms from the database of Planteome. The ontology curators update the annotation database regularly to ensure it contains the most comprehensive ontology information. Besides, users can find enriched ontology terms based on self-defined annotations. Based on the user's input data, OAT can find the ontology terms which are significantly enriched by the input gene list.

Disambiguate Input Gene Names

When doing the analysis, OAT allows biologists to input gene symbols or gene synonyms instead of the exact gene association ID to query the enriched ontology terms. However, it is possible that two different genes have the same symbols or they have the same synonyms. We call these genes ambiguous genes of the input string.

OAT allows users to select the targeting gene from the ambiguous genes. This procedure is necessary to ensure OAT return the correct analysis. A future version of the tool may allow for the joining of these ambiguous gene names.

Statistical analysis methods

After fixing the interesting gene list, the users can submit the list of genes to the server to find all the enriched ontology terms.

In our system, we create the contingency table (as table 1 shows) used in [6] and [7]. For one specific ontology term A and n interesting genes, all genes in the database (N) are classified into four categories: the genes annotated to the term and in the input gene list (m), the genes not annotated to the term and in the input gene list ($n - m$), the genes annotated to the term

and not in the input gene list ($k - m$), the genes not annotated to the term and not in the input gene list ($N - n - k + m$).

Table 1. Contingency table between one ontology term and the number of genes associated with this term

	Number of genes inside interesting genes	**Number of genes not inside interesting genes**	**Sum**
Annotated to ontology A	m	k-m	k
Not annotated to ontology A	n-m	(N-n)-(k-m)	N-k
Sum	n	N-n	N

The hypothesis that the observation is due to chance is called the null hypothesis in statistics. The constructed two-way table is the contingency table which can be used for testing the significance of the null hypothesis. Calculating p-values is a common way to measure the significance level of the observation. If the p-value is smaller than the user chosen a cut-off value (0.01 or 0.05), the term is not enriched by the input gene list. Smaller p-value represents more statistically significant, which indicates that we have stronger evidence to reject the null hypothesis.

OAT provides two main methods to test the significance level, i.e. the Fisher's exact test and Chi-square test. The users are recommended to select different methods for different size of samples.

The Fisher's exact test is better to be applied when the input genes number is small, it will provide an exact calculation of the p-value. But it also requires large computationally cost since

the factorial calculation involved in large or well-balanced data. The chi-square test would work better for large samples, but it will only give an approximation of the significance.

Fisher's exact text

With the number of genes annotated to one ontology term A inside the gene list m, the total number of genes annotated to this term in the whole database k, the number of input genes n and total number of genes in the database N, we can get the hypergeometric distribution of the observation with equation 1.

$$H_A(m,k,n,N) = \frac{\binom{k}{m}\binom{N-k}{n-m}}{\binom{N}{n}} \quad (1)$$

Then a p-value can be calculated by using Fisher's exact test with equation 2.

$$P_A = \sum_{i=m}^{k} H_A(i,k,n,N) \quad (2)$$

Chi-square test

Based on the contingency table, we calculate the expected value (E_i) of the cell that represents the number of genes annotated to the ontology term A and inside the input list by $\frac{nk}{N}$. Then we construct an expected contingency table by fixing the margin values k, n, and N and using the calculated expected value to calculate all other three cells. At last, we calculate the χ^2 value with equation 3, and transfer it to the p-value for 1 degree of freedom since a two-way table always has a freedom of 1.

$$\chi^2 = \sum_{i=1}^{4} \frac{(E_i - O_i)^2}{E_i} \quad (3)$$

Our system also supports the Yate's chi-square test (or Yate's correction for continuity) to calculate the χ^2value as Equation 4 shows,

$$\chi^2 = \sum_{i=1}^{4} \frac{(|E_i - O_i| - 0.5)^2}{E_i} \quad (4)$$

where E_i is the expected value of one distinct event. O_i is the observed value. Since our contingency table is a two-way table, we always have 4 distinct events.

Visualization of enriched ontologies

Analysis results are shown as a table in OAT. However, this table is not intuitive for biologists to study the structure of the enriched ontology terms. Also, the correlation among different ontology categories can be unclear to the users.

To solve the above questions, visualization of the enriched ontology terms become a commonly accepted solution. In OAT, we provide three types of visualizations to help users to study different aspects of the enriched ontology terms.

Force-directed network visualization

Ontology terms are intrinsically own a network structure, i.e. each ontology term may have parents and children. It inherits the properties of their parents and differs with its siblings in some functionalities. Since each ontology term can have multiple parents and siblings, the research of the enriched ontology branch of a set of genes facilitates biologists to explore the potential functions associated to the genes and can be helpful for finding featuring genes in the set.

Fig 1. The three enriched ontology terms (yellow and orange nodes) are not directly connected. Several ontology terms, which are not enriched (blue nodes), are shown to construct the connected network.

To visualize the enriched ontology terms, we apply a network visualization to the analysis results. Note that there is no guarantee that the enriched terms are always directly connected. To study the branching structure of the enriched terms, OAT show several ontology terms which are not enriched to connect the enriched terms inside each ontology category (As Fig 1 shows). The children of the enriched ontology terms can also be shown in OAT (As Fig 2 shows).

Fig 2. parts of the network visualization of the enriched ontology terms. OAT supports movements of the nodes, selecting the nodes and edges. After one node is selected, all the neighboring nodes within 2 levels will be highlighted. OAT also support the showing of all the children nodes of the enriched ontology terms.

OAT can apply a force-directed method [12] to design the layout of the networks. The initial layout of the network is a simple circular visualization, i.e. all the nodes are distributed on a circle. After starting the animation, nodes will be relocated based on the force-directed method.

Hierarchical visualization

The hierarchical visualization (as Fig 3 shows) of the analysis results is a common method to facilitate users to explore the hierarchical structure among the enriched ontology terms([3], [2], [4]). This kind of visualization emphasizes inherited properties and relationships between ontology terms.

Fig 3. the hierarchical visualization of the enriched ontology terms. Each branch of the hierarchical distributed graph corresponds to one ontology category. All the branches are evenly distributed horizontally. Different colors of the edges correspond to the different types of relationships between ontology terms.

Note that in graph theory, the enriched ontology terms do not actually construct a tree graph, i.e. an undirected acyclic graph. Also, when considering all types of relationships among ontology terms, calculation of the hierarchical structure, which requires a deciding of the level of each node will become over-complicated. So OAT constructs the hierarchical structure by only considering the relationship “is a”.

For each category of the ontology terms, find the root node and assign it as level 0, then all its children used to connect enriched ontology terms and construct the connected network are assigned to level 1. Iterate this process for each of the node in level 1 until all the nodes are assigned a level. Then nodes on each level will be horizontally evenly distributed. Each branch of the hierarchical visualization will also be horizontally evenly distributed. The color scheme for nodes and edges for the hierarchical visualization is the same as the one used in the force-directed network visualization.

Matrix visualization

To facilitate the study and provide an intuitive overview of the correlation between the input genes and enriched ontology terms. OAT provides the third type of visualization, i.e. the matrix view (as Fig 4 shows) of the enriched terms.

Fig 4. the matrix visualization of the enriched ontology terms. The association between the enriched ontology terms and input genes are displayed in a matrix. Each colored cell of the matrix corresponds to one significant enrichment. The colors are assigned based on the significance of the association.

A table is constructed by considering each column as one ontology term and each row as one input gene. Each cell of the table is colored based on whether the input gene list is enriching the corresponded ontology terms and whether the gene is associated with that term. So all the

colored cells in each column will be assigned to the same color based on the significance level. The color scheme of each cell is as same as the one used in the other two visualizations.

OAT supports several methods to sort rows or columns so that users can study the correlation between the genes and enriched ontology terms from different aspects. For example, users can sort rows by a number of ontology terms associated with the genes, or users can sort the columns based on p-value which represents the extent of the significance of this enriched ontology term.

Results

Fig 5. The interface of OAT.

The user interface of the system is as Fig 5 shows. The users can select the statistical analysis method and cut-off p-value for the analysis. Also, the ontology categories and taxon for the input genes can be changed. Users can indicate the annotation database to allow OAT to conduct the analysis from different sets of annotations. When selecting the static analysis, the query will be conducted with only considering the input annotations.

Fig 6. disambiguate the input genes. The system allows users to select the target genes from the set of genes which using the same string as symbol or synonym. For example, the string "GR1" is used as a symbol by both TAIR:gene:1005714586 and TAIR:gene:1009021737. It is also used as one of the synonyms of TAIR:gene:2094517 and TAIR:gene:1009021925.

When selecting the dynamic analysis, OAT will ask users to disambiguate the input genes if the input strings do not uniquely appear in the whole database. As Fig 6 shows, one input string can

be either the symbol or the synonym of a gene. Selecting the targeting genes or modify the unrecognized input strings can help users to get more accurate analysis results.

After the procedure of disambiguating input strings. Users can submit interesting genes to the server to analyze the enriched ontology terms. The results will be shown as Fig 7 shows. OAT allows users to search, to sort and to download the analyzed results.

Fig 7. analysis results table. The input # and ref# correspond to the number of genes inside the interesting gene list (m in table 1) and the number of genes (k in table 1). The p-value is calculated based on the user's selected statistic analysis method.

OAT uses three ways to visualize the enriched ontology terms. The network visualization (Fig 8 and 2), hierarchical visualization (Fig 3), and matrix visualization (Fig 4). In all these three types of visualizations, different color of nodes represents different levels of significance. Nodes with darker colors mean the p-values decrease, in other words, more significant.

Fig 8. the network visualization of the enriched ontology terms. Note that this is one branch of the whole visualization results. Each branch normally corresponds to one ontology category.

In both the network visualization and hierarchical visualization, different colors of edges represent a different type of relationships between ontology terms, i.e. is a, part of, regulate, positively regulate, negatively regulate, and occurs in. Besides, quick search among all the nodes of visualizations, highlighting the selected nodes and edges, and filter the results based on categories of ontologies are supported by these two types of visualization.

Discussion and future direction

The current version of OAT is implemented using JavaScript. Several js libraries such as d3.js [10] and vis.js [11] are incorporated into the system to facilitate the implementation of the

visualization and analysis. The framework is easy to be expanded for future modifications. All code is open-source and available on Github [13].

The tool is useful for studying the relationships between genes and ontology terms.

In the future, we would like to develop more useful visualizations to assist the study of the enriched ontology terms. Also, more statistical methods could be added to the system to provide more options for the analysis.

Commented [1]: We should add more here, just not sure what. Maybe PJ or LC could help?

Acknowledgments

The authors would like to thank the anonymous reviewers and for all the suggestions. This work was supported by the National Science Foundation award [grant number IOS #1340112]

References

1. Jaiswal P, Cooper L, Elser JL, Meier A, Laporte MA, Mungall C, Smith B, Johnson EK, Seymour M, Preece J, Xu X. Planteome: a resource for common reference ontologies and applications for plant biology.

2. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC bioinformatics. 2009 Dec;10(1):48.

3. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic acids research. 2010 Apr 30;38(suppl_2):W64-70.

4. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005 Jun 21;21(16):3448-9.

5. Zhou X, Su Z. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC genomics. 2007 Dec;8(1):246.

6. Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test?. Bioinformatics. 2006 Dec 20;23(4):401-7.

7. Mi G, Di Y, Emerson S, Cumbie JS, Chang JH. Length bias correction in gene ontology enrichment analysis using logistic regression. PloS one. 2012 Oct 2;7(10):e46128.

8. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2008 Nov 25;37(1):1-3.

9. Yates F. Contingency tables involving small numbers and the χ 2 test. Supplement to the Journal of the Royal Statistical Society. 1934 Jan 1;1(2):217-35.

10. Bostock M, Ogievetsky V, Heer J. D^3 data-driven documents. IEEE transactions on visualization and computer graphics. 2011 Dec;17(12):2301-9.

11. Almende BV, Thieurmel B, Robert T. visNetwork: Network Visualization using “vis. js” Library. R package version. 2016;1(1).

12. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PloS one. 2014 Jun 10;9(6):e98679.

13. Cooper L, Meier A, Laporte MA, Elser JL, Mungall C, Sinn BT, Cavaliere D, Carbon S, Dunn NA, Smith B, Qu B. The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic acids research. 2017 Nov 23;46(D1):D1168-80.

Figures

Fig 1.

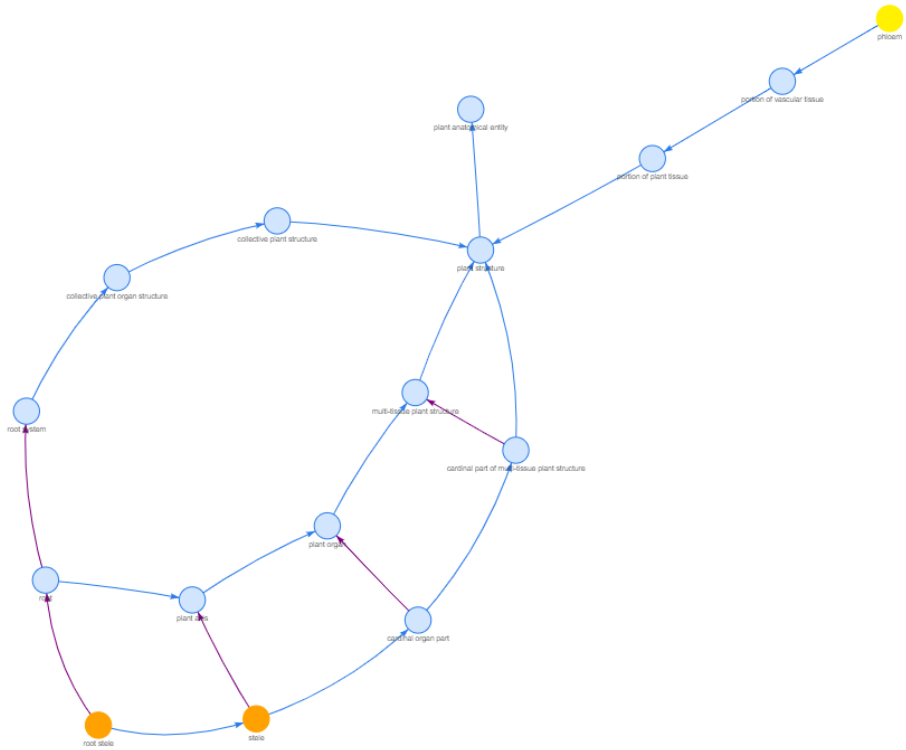
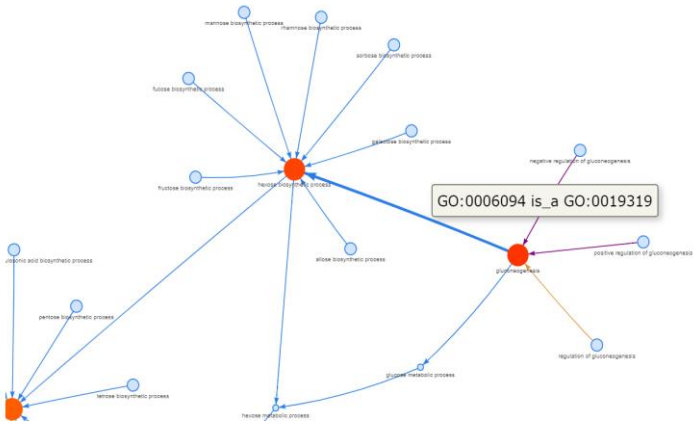


Fig 2.



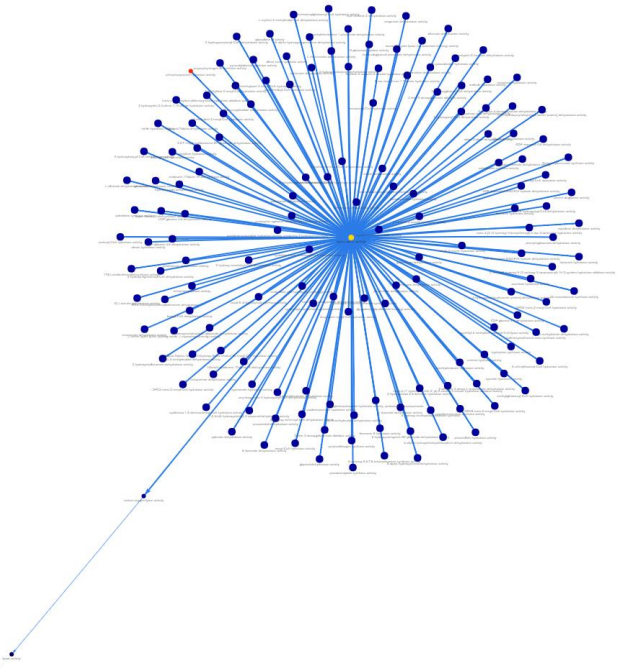
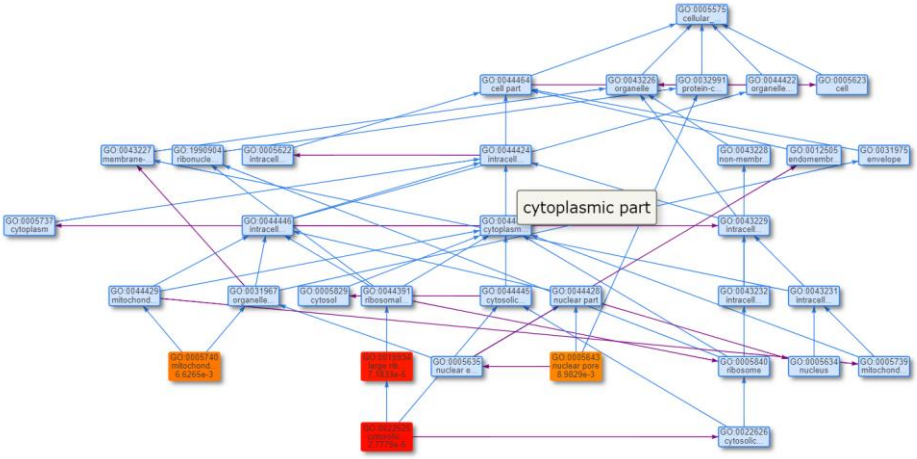


Fig 3.


cytoplasmic part

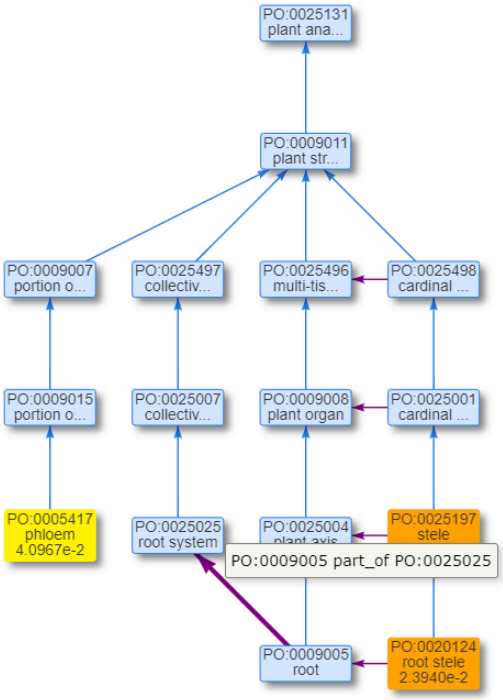

PO:0025131
plant ana...
PO:0009011
plant str...
PO:0009007
portion o...
PO:0025497
collectiv...
PO:0025496
multi-tis...
PO:0025498
cardinal ...
PO:0009015
portion o...
PO:0025007
collectiv...
PO:0009008
plant organ
PO:0025001
cardinal ...
PO:0005417
phloem
4.0967e-2
PO:0025025
root system
PO:0025004
PO:0025197
stele
PO:0009005 part_of PO:0025025
PO:0009005
root
PO:0020124
root stele
2.3940e-2

Fig 4.


GO:0022625
GO:0015934
GO:0004634
GO:0006094
GO:0019319
GO:0015510
GO:0046364
GO:0000059
GO:0057740
GO:0005643
PO:0020124
PO:0025197
GO:0009846
GO:0000741
GO:0048284
GO:0016836
GO:0009553
PO:0005417
GO:0042254
XPO1B
LOS2
RPL16A
AT2G44860
AT1G74060
GR1
ENO1
ENOC
SAC52
AT1G18540
RPL10B
PRPL11
RP1
RPL3B
SAG24
AT1G74050
RPL24A
AT2G37190
RPL23AA

330 **Fig 5.**

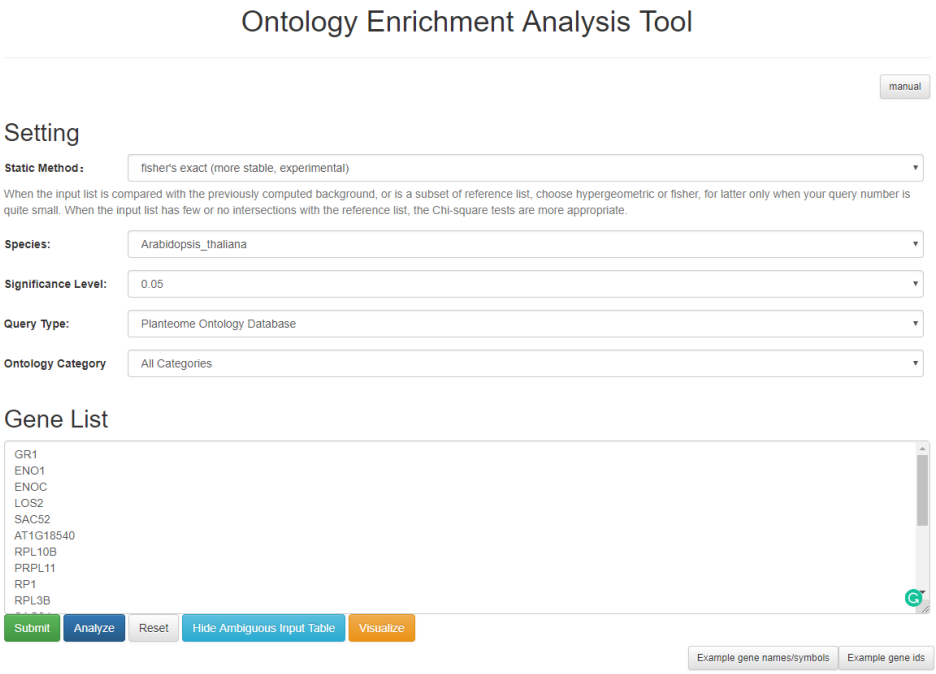

Ontology Enrichment Analysis Tool
manual
Setting
Static Method:
fisher's exact (more stable, experimental)
When the input list is compared with the previously computed background, or is a subset of reference list, choose hypergeometric or fisher, for latter only when your query number is quite small. When the input list has few or no intersections with the reference list, the Chi-square tests are more appropriate.
Species:
Arabidopsis_thaliana
Significance Level:
0.05
Query Type:
Planteome Ontology Database
Ontology Category
All Categories
Gene List
GR1
ENO1
ENOC
LOS2
SAC52
AT1G18540
RPL10B
PRPL11
RP1
RPL3B
Submit
Analyze
Reset
Hide Ambiguous Input Table
Visualize
Example gene names/symbols
Example gene ids

Fig 6.

Name	Ambiguous IDs	Match type
GR1	TAIR:gene:1005714586	bioentity_label
	TAIR:gene:1009021737	bioentity_label
	TAIR:gene:2094517	synonym
	TAIR:locus:1005716561	bioentity_label
	TAIR:gene:2093690	bioentity_label
	TAIR:gene:1009021925	synonym
	TAIR:locus:2094518	synonym
	TAIR:locus:2093691	bioentity_label
ENO1	TAIR:gene:2031475	bioentity_label
	TAIR:At1g74030	bioentity_label
	TAIR:locus:2031476	bioentity_label
ENOC	TAIR:gene:2043066	bioentity_label
	TAIR:locus:2043067	bioentity_label
LOS2	TAIR:locus:2044851	bioentity_label
	TAIR:gene:2044850	bioentity_label

Fig 7.

the number of input genes is: 19
the number of background genes is: 40025

Show 100 entries Search:

Ontology ID	Name	Category	Description	Input#	Ref#	P-val
GO:0000059	obsolete protein import into nucleus, docking	biological_process	OBSOLETE. A protein complex assembly process that contributes to protein import into the nucleus, and that results in the association of a cargo protein, a carrier protein such as an importin alpha/beta heterodimer, and a nucleoporin located at the periphery of the nuclear pore complex.	1	12	5.6824e-3
GO:0000741	karyogamy	biological_process	The creation of a single nucleus from multiple nuclei as a result of fusing the lipid bilayers that surround each nuclei.	1	56	2.6257e-2
GO:0001510	RNA methylation	biological_process	Posttranscriptional addition of a methyl group to either a nucleotide or 2'-O ribose in a polyribonucleotide. Usually uses S-adenosylmethionine as a cofactor.	2	179	3.2350e-3
GO:0004634	phosphopyruvate hydratase activity	molecular_function	Catalysis of the reaction: 2-phospho-D-glycerate = phosphoenolpyruvate + H2O.	1	3	1.4235e-3
GO:0005643	nuclear pore	cellular_component	Any of the numerous similar discrete openings in the nuclear envelope of a eukaryotic cell, where the inner and outer nuclear membranes are joined.	1	19	8.9629e-3
GO:0005740	mitochondrial envelope	cellular_component	The double lipid bilayer enclosing the mitochondrion and separating its contents from the cell cytoplasm; includes the intermembrane space.	1	14	6.6265e-3
GO:0006094	gluconeogenesis	biological_process	The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol.	2	164	2.7257e-3

Fig 8.