Gene Ontology (GO) Term Enrichment using PANTHER Gene List Analysis tools

GO Term Enrichment is a tool commonly used to evaluate characteristics of sets of genes, such as those identified from RNA-seq or microarray experiments. The basic function takes a set of genes and compares the frequency of GO terms in the sample set with the frequency of the same set of GO terms in the a reference set, usually a whole genome set, to identify terms that are over- or underrepresented in the sample set.

TAIR uses a web service, provided by PANTHER DB, to facilitate GO term enrichment analysis for Arabidopsis, and other plant genomes represented in the PANTHER database. PANTHER's tool accesses a comprehensive list of GO annotations from the GO Consortium that is updated monthly. Because annotations are constantly being updated as new information is obtained, the monthly updating schedule ensures that analyses done using the PANTHER tool rely on the most current annotation data.

Briefly, the stats

The tool applies the binomial test (Mi, H., 2013) to identify over- or underrepresented terms in the sample gene set compared to the reference genome set. The default parameters also apply a Bonferroni correction, a common statistical method used for multiple comparisons. The Bonferroni correction is important because the tool performs multiple statistical tests (one for each pathway, or each ontology term) at the same time. This correction multiplies the single-test P-value by the number of independent tests to obtain an expected error rate.

How to Perform a Basic Term Enrichment Analysis

A. Enter your gene list (Figure 1)

1. Paste in a list of gene identifiers such as AGI Locus IDs (e.g. AT5G61160), UniProt IDs (e.g. Q9FNP9) or NCBI Entrez GeneIDs (e.g. Gene: 836237), separated by newlines or commas. PANTHER provides a complete list of acceptable identifiers.

2. Choose the plant species from which your gene list was derived. For example, if your gene list was generated from *Arabidopsis thaliana*, choose that. If your gene list is from grape, choose *Vitis vinifera* and so on.

3. Select the ontology aspect that you wish to analyze. The options are 'biological process', 'molecular function', and 'cellular component.'

4. Click Submit, to send the data to PANTHER.

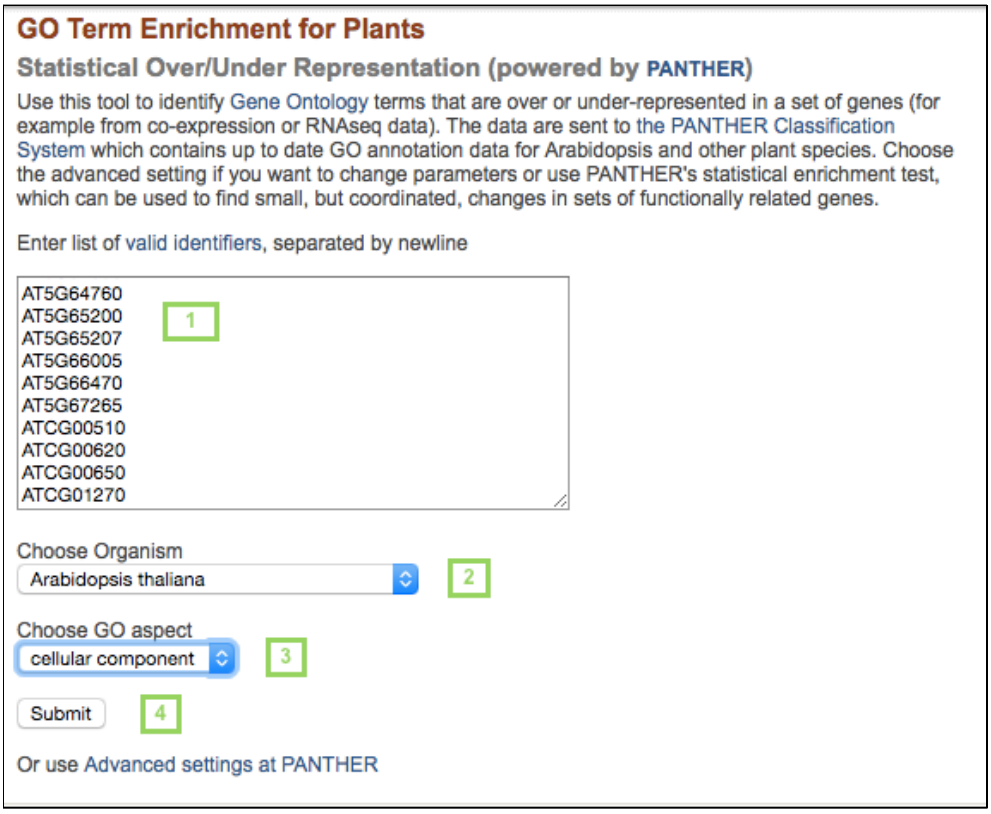

Figure 1. Entering Data

B. Understanding the results

1. Analysis Summary (Figure 2).

    A. The analysis summary box displays the analysis type (PANTHER can do several types of gene list analysis), annotation version and annotation dataset. This information is important to record and report in your publications, as the same analysis performed with different software versions and different annotation releases may yield different results.

    B. Below this section is a list of mapped and unmapped IDs. Unmapped IDs are those that could not be mapped to a corresponding protein record in the PANTHER. PANTHER also provides a list of IDs where multiple IDs mapped to the same PANTHER protein entry. Typically this occurs because more than one gene produces the same amino acid sequence. You can click on the number to review each list to see the details.

    C. Export the result set as a text file to save your data.

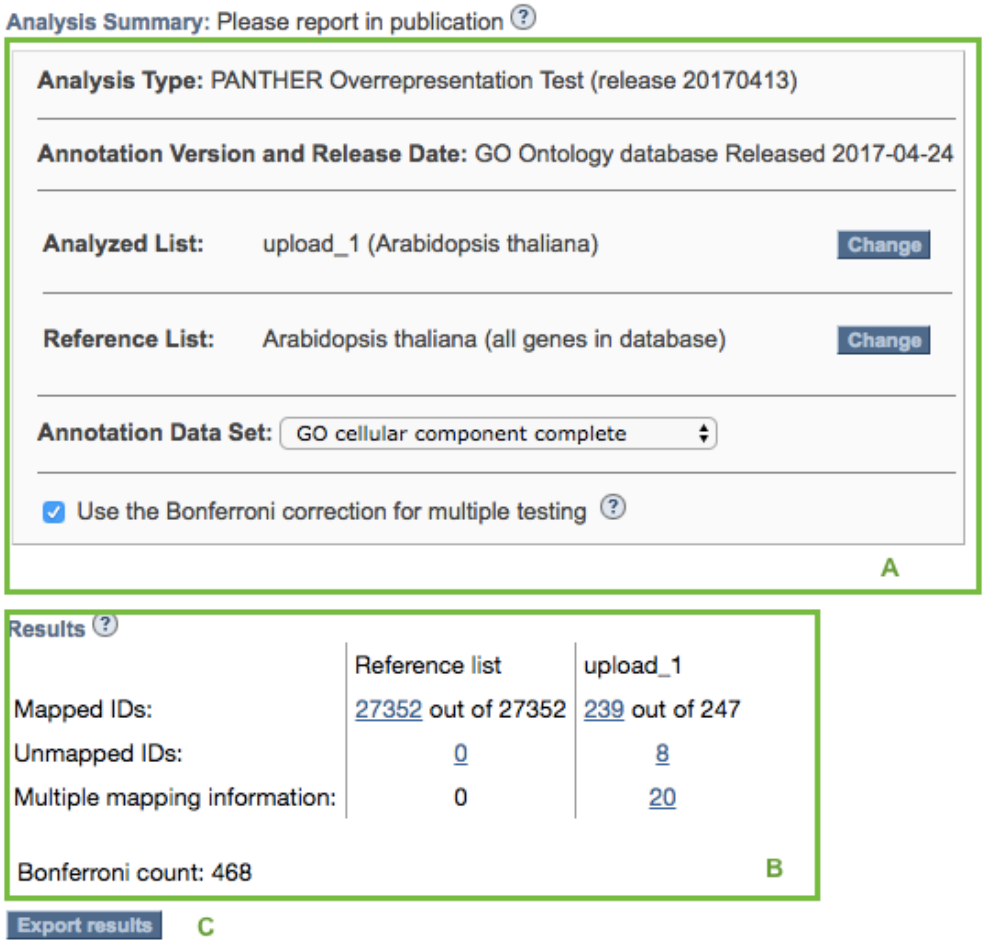

Figure 2. Analysis Summary

2. Results table (Figure 3). The default display presents a table of over- or underrepresented terms presented in an ordered hierarchy. The most granular (specific) terms are displayed first with the more general terms below. The granular terms are ordered with those having the **largest fold change** on top. Clicking any of the column headers allows you to change the sort order.

    A. The first column shows the terms. Related terms (i.e. parent-child terms in hierarchy) are grouped by background color, with the most granular term at the top. In the example table the most granular term 'cytosolic large ribosome subunit' is shown above the parent terms. If a term has multiple parents- all of paths are shown. If there

are duplications in the parent terms (as is the case in the example), they are shown only once. You can invert the sort order by clicking the term ‘Hierarchy’ in the last column header.

B. The next column shows the number of genes (#) in the reference genome dataset that map to the terms (either directly or through inheritance).

C. The third column shows the number of genes (#) in the sample gene set that map to the GO term (either directly or through inheritance).

D. The fourth column displays the number of genes mapped to the term that would be **expected** based on the whole genome representation. For the example shown in Figure 3, if 113/27,352 genes in the reference set mapped to cytosolic large ribosomal subunit, then we would expect the same proportion of genes (0.0041) to map to that term in in the sample set (0.0041 X 247=1.02). Clicking on the number will retrieve a list of the genes that map to the term.

E. The fifth and sixth columns show the fold enrichment and if the change is an increase (+) or decrease (-). Fold change is calculated by dividing the observed by expected results.

F. The seventh column displays the p value. The lower the p value, the less likely the obtained result can be explained by random distribution. By default only p values less than 0.05 are displayed.

Displaying only results with P<0.05; click here to display all results

GO cellular component complete	Arabidopsis thaliana (REF) #	upload_1 (∇ Hierarchy NEW!) #	expected	Fold Enrichment	+/-	P value
cytosolic large ribosomal subunit a	113 b	c 15	1.02 d	14.70 e	+	1.28E-10 f
↳cytosolic part	290	25	2.62	9.55	+	2.51E-14
↳cytosol	1822	55	16.45	3.34	+	1.01E-12
↳cytoplasmic part	8835	122	79.78	1.53	+	8.94E-06
↳cytoplasm	10861	138	98.08	1.41	+	9.85E-05
↳large ribosomal subunit	138	15	1.25	12.04	+	2.10E-09
↳ribosomal subunit	240	22	2.17	10.15	+	5.88E-13
↳intracellular organelle part	3582	78	32.35	2.41	+	2.04E-11
↳organelle part	3591	78	32.43	2.41	+	2.32E-11
↳intracellular ribonucleoprotein complex	544	32	4.91	6.51	+	4.42E-14
↳ribonucleoprotein complex	544	32	4.91	6.51	+	4.42E-14
↳macromolecular complex	1839	46	16.61	2.77	+	1.59E-07
↳ribosome	370	30	3.34	8.98	+	8.45E-17
↳intracellular non-membrane-bounded organelle	1087	41	9.82	4.18	+	5.72E-12
↳non-membrane-bounded organelle	1087	41	9.82	4.18	+	5.72E-12
↳cytosolic ribosome	269	24	2.43	9.88	+	5.13E-14
cytosolic small ribosomal subunit	92	7	.83	8.43	+	1.15E-02
↳small ribosomal subunit	101	7	.91	7.67	+	2.07E-02
nucleolus	279	12	2.52	4.76	+	5.36E-03
chloroplast stroma	539	20	4.87	4.11	+	6.75E-05
↳plastid stroma	555	20	5.01	3.99	+	1.07E-04

Figure 3. PANTHER Results Table

C. Further Analysis. Once you have retrieved your analysis results, there are a number of other things you can do from the PANTHER site. We list a few options below. Please see the PANTHER user guide for a complete list of functions.

1. Download the gene list from PANTHER for your sample data with added information from the PANTHER database.

    - Next to the hierarchy – click the file name (in Figure 3, it is upload_1) to view the PANTHER entries that correspond to your gene list (Figure 4). You can customize the output to display a number of different data fields.

    -

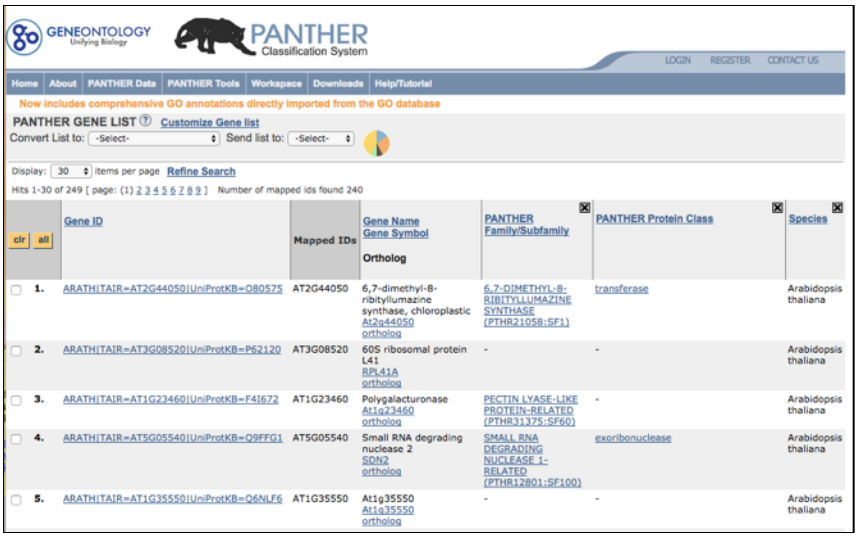

Figure 4. Panther Gene List output

2. Run GO term enrichment analysis using different aspects.

    - Go to the Annotation Data Set dropdown menu (Figure 2) and choose another annotation dataset (for example ‘molecular function complete’ or ‘biological process complete’) from the list and re-analyze your gene list. For GO term enrichment, we recommend using “GO molecular function complete” or “GO biological process complete” or “GO cellular component complete”, these are the datasets with the complete, up to date GO annotations.