



## Instruction Manual

# Table of Contents

<b>Table of Contents</b> .....	<b>2</b>
<b>Section 1: Introduction</b> .....	<b>3</b>
Software Description.....	3
Documentation.....	3
Interfaces .....	3
Installation Requirements .....	3
Program Overview .....	4
<b>Section 2: Running GO-Elite</b> .....	<b>5</b>
GO-Elite Input Files.....	5
Preparing Your Gene Lists .....	5
Running GO-Elite Online .....	6
Running GO-Elite Locally Using the Graphical User Interface .....	8
Running GO-Elite Locally Using the Command-Line Option .....	15
<b>Section 3: Interpreting GO-Elite Results</b> .....	<b>20</b>
GO-Elite Output Files.....	20
Downstream Analyses .....	23
<b>Section 4: Algorithms</b> .....	<b>25</b>
GO and Pathway Over-Representation Analysis .....	25
Multiple Identifier Mapping for ORA .....	25
Filtering of ORA Results.....	27
Pruning of Gene Ontology Hierarchical Relationships .....	28
<b>Section 5: GO-Elite Gene and Pathway Databases</b> .....	<b>30</b>
Database Files Overview .....	30
Downloading Official GO-Elite Databases .....	30
Updating WikiPathways Relationships .....	33
Addition and Update of Species Databases .....	35
Updating Gene Ontology Structure Annotations .....	42
Updating Gene Ontology Relationships from EntrezGene .....	43
Updating Affymetrix Relationships .....	45
Rebuilding Species Ensembl Databases (Advanced) .....	47
System Codes.....	50
<b>References</b> .....	<b>52</b>

# Section 1: Introduction

## **Software Description**

GO-Elite is an application designed to identify a non-redundant set of Gene Ontology (GO) terms (1) or pathways (WikiPathways (2)) to describe a particular set of genes. This application is able to calculate advanced over-representation analysis (ORA) statistics from user gene lists, determine the minimal set of biologically distinct GO terms and pathways from these results and summarize these results at multiple levels. GO-Elite version 1 Beta is currently provided as a cross-platform stand-alone application, which can be run as a compiled Windows executable file (GO\_Elite.exe), Mac OS X application (GO\_elite.app) or as cross-platform source code in python for any operating system. To download the program or use the GO-Elite web interface, go to:

[http://www.genmapp.org/go\\_elite/](http://www.genmapp.org/go_elite/).

## **Documentation**

In addition to the information provided in this document, instructions, tutorials, program update information and user questions are posted at our Google Groups page at:

<http://groups.google.com/group/go-elite>. You can also contact us at:

[genmapp@gladstone.ucsf.edu](mailto:genmapp@gladstone.ucsf.edu) with any questions or problems.

## **Interfaces**

Three main interfaces are available for GO-Elite: (1) graphical user interface (GUI), (2) command-line and (3) online. The GUI and command line options are available from the downloaded stand-alone application, whereas the online version can be accessed independently. Please note, the latter option lacks the ability to update or modify the GO-Elite gene systems and species configurations.

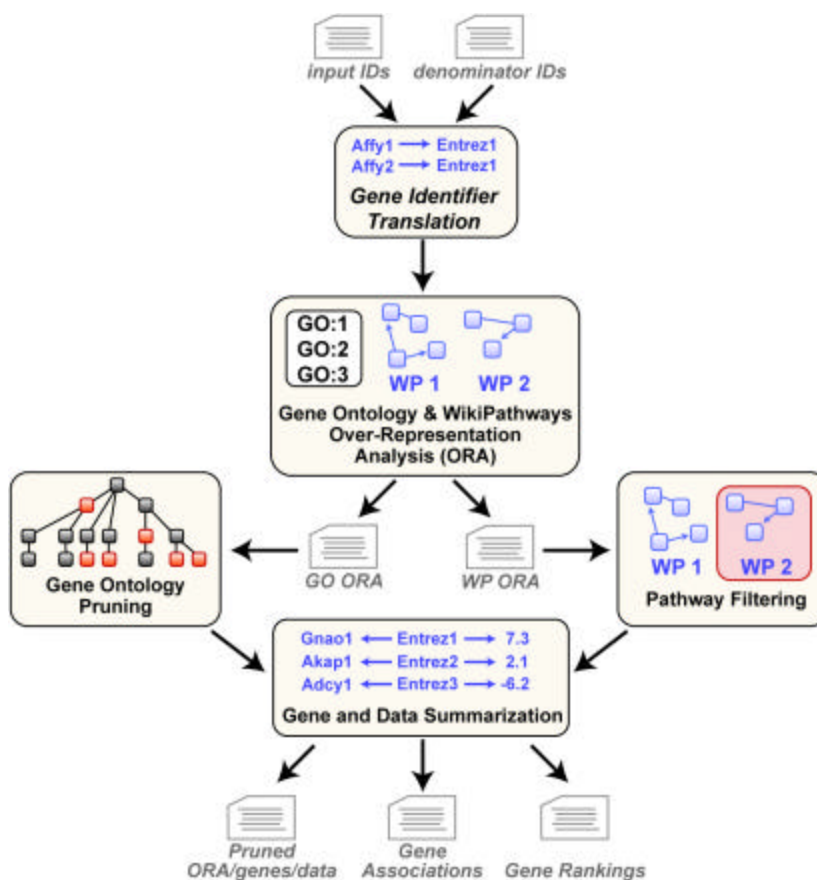
## **Installation Requirements**

When installing the compiled OS specific installers, no additional software is required

(python components are bundled with the application). If running GO-Elite directly from the source code (e.g., on Linux or Unix), Python 2.3 or greater and Tkinter (python Tk) are required, however, these are usually installed by default on Linux and Mac OS X.

## **Program Overview**

When analyzing user data, GO-Elite begins by reading in at least one input and one denominator file. The input file contains all gene IDs to be examined for over-representation (e.g., Affymetrix probesets, Ensembl or EntrezGene) along with the gene system code, whereas the denominator contains all gene IDs examined (e.g., all array probesets). The input file can also contain any number of additional columns with numeric data for pathway level summarization. These lists are submitted to GO-Elite for over-representation analysis (ORA) along with parameters for downstream filtering and pruning of the resulting GO-terms and pathways. A diagram illustrating these steps along with the major output files is shown below.



## Section 2: Running GO-Elite

### GO-Elite Input Files

The user is only required to provide two files: (A) input gene lists and (B) denominator list (containing all genes examined, including all input genes). Alternatively, an existing over-representation analysis (ORA) results file from GO-Elite or its sister program MAPPFinder (3) (a component of the GenMAPP application (4)), can be used as input.

### Preparing Your Gene Lists

To perform ORA in GO-Elite, you need to provide at least one input gene list and denominator list. The input list is a subset of the denominator list which consists of gene IDs or probesets that are highlighted from a user analysis (e.g., up and downregulated genes). Both input and denominator lists consist of a column of gene IDs (column 1), system code for each ID (column 2) and any other data you may wish to summarize at the pathway level (e.g. fold change). To find out what system code corresponds to your gene data, start GO-Elite, select the option “**Analyze Gene Lists**” and select the button named “**GO-Elite Supported System Codes**” at the bottom of the interface. A list of all current supported system codes for GO-Elite can be found at the end of Section 5 in this document. Commonly used system codes are “X” for Affymetrix, “L” for EntrezGene and “En” for Ensembl. If your array is not one of the supported array types (Affymetrix, Agilent, Codelink and Illumina), the system “Ma” for Miscellaneous Array may contain your array system (see <http://www.ensembl.org/biomart/martview>). Currently, GO-Elite can only accept one gene system per file. If the system code column is not present, GO-Elite will may be able to guess what type of ID system it is, but this is not recommended. For the denominator file, only the first two columns are used (gene IDs and system code). If you have multiple input files in a single directory that corresponds to different denominator files, GO-Elite will properly match these up if you place a unique number, letter or name before the name of each input file, separated by a period (e.g. **exp1**.input.txt and **exp2**.input.txt) and denominator file that matches it (e.g., **exp1**.denominator.txt and

**exp2.denominator.txt).**

### *Sample Input Gene List*

<b>Probeset (REQUIRED)</b>	<b>SystemCode (RECOMMENDED)</b>	<b>Fold TP1 (OPTIONAL)</b>	<b>Fold TP2 (OPTIONAL)</b>
j05479_s_at	X	1.23	2.31
L49502_s_at	X	-1.92	-1.85
Msa.33069.0_s_at	X	-2.41	-1.33
Msa.37566.0_s_at	X	3.03	0.25
AF028071_s_at	X	-1.91	0.85
Aa462409_s_at	X	0.25	4.35

### **Running GO-Elite Online**

In addition to the stand-alone GO-Elite program, a simple query interface is available over the internet that does not require the user to download any software. This interface has all basic analysis features for GO-Elite, however, it is not possible to modify or add new database information (e.g., new gene systems or species support). The online interface supports a limited number species and ID systems (typically Affymetrix probeset IDs, EntrezGene and Ensembl). If performing different analyses from these, we recommend downloading the GO-Elite software to your computer.

The screenshot displays the 'Opal Dashboard' with three tabs: 'Summary Home', 'Statistics', and 'List of Applications'. The 'Summary Home' tab is active. Below the tabs is a yellow box titled 'Submission form for go-elite'. Inside this box, there are several input fields and radio buttons. The first field is 'Insert number of CPU (only for parallel application):' with an empty text box. Below it is the text 'Ungrouped input fields...'. The 'Species to analyze; default: Mm' section has two radio buttons: 'Mm' (selected) and 'Hs'. The 'Input denominator file' and 'Input gene list file' sections each have a text box and a 'Browse...' button. The 'Primary gene system linked to GO/pathway databases; default EntrezGene' section has two radio buttons: 'Ensembl' and 'EntrezGene' (selected). The 'Number of permutations for over-representation analysis; default: 2000' section has a text box containing '2000'.

Opal Dashboard

Summary Home Statistics List of Applications

Submission form for go-elite

Insert number of CPU (only for parallel application):

Ungrouped input fields...

Species to analyze; default: Mm ☒ Mm ☐ Hs

Input denominator file  Browse...

Input gene list file  Browse...

Primary gene system linked to GO/pathway databases; default EntrezGene ☐ Ensembl ☒ EntrezGene

Number of permutations for over-representation analysis; default: 2000

**Figure – GO-Elite Web Interface**

From the GO-Elite website ([http://www.genmapp.org/go\\_elite](http://www.genmapp.org/go_elite)), select the link under **“GO-Elite Web Services”**. This link will take you to a web page where you can select the species code corresponding to your species (e.g., Mm = Mus musculus and Hs = Homo sapiens), the location of the file containing your input and denominator gene lists and the analysis parameters used by GO-Elite. These parameters are discussed in the following section on **“Using the Stand-Alone Program Graphical User Interface”**. Once these options are selected, click “Submit” to start the analysis. A window displaying the job status will appear. Once finished, links to the results files will be provided. Download these to your hard-drive and see Section 3 of this manual (Interpreting GO-Elite Results).

Please note that the length of the analysis depends on the size of the input gene list and number of permutations selected. An input list of 4,000 gene IDs and 2,000 permutations can take up to 25 minutes.

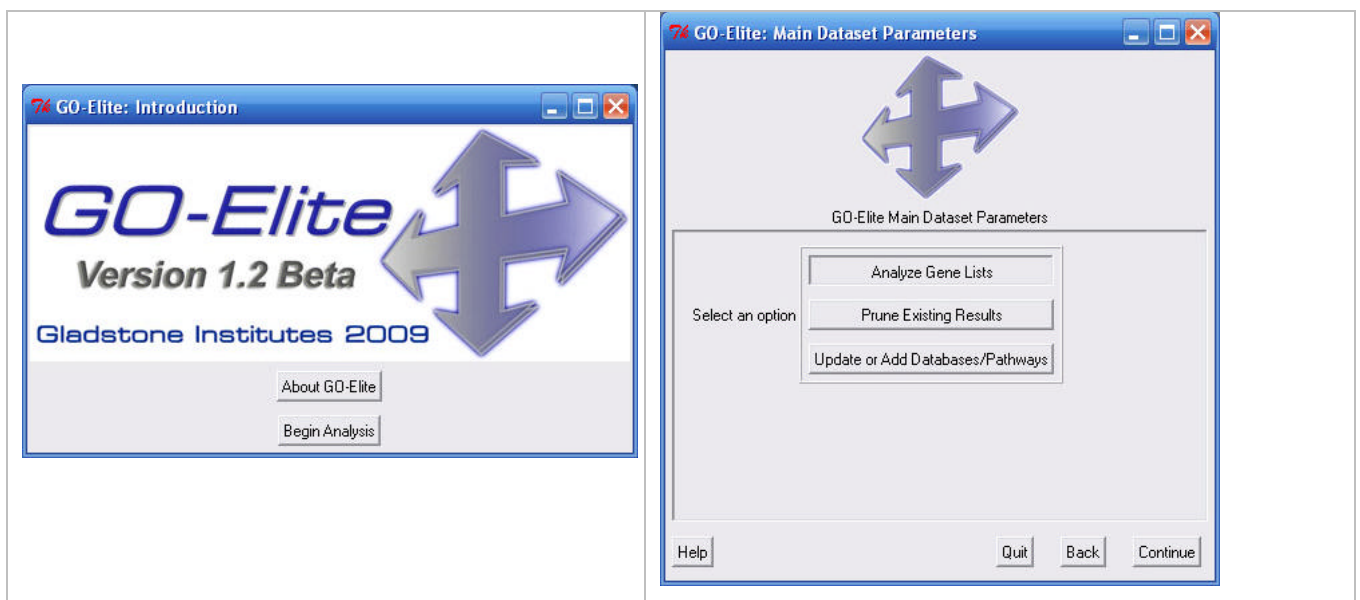
## **Running GO-Elite Locally Using the Graphical User Interface**

When installed locally through the stand-alone application, GO-Elite has a simple, intuitive graphical user interface intended to be run by novices and experts alike. To run GO-Elite, follow the appropriate instructions .

- ? **PC** – Double click on the file “GO\_Elite.exe”.
- ? **Mac** – Double click on the file “GO\_Elite”. If an OS specific version is unavailable, follow the source code instructions.
- ? **Linux or source code** – Installation of python is required, but is typically present with most Linux installations and Mac OSX. To run, open a terminal window on Linux or Mac or a DOS prompt on PCs and go to the GO-Elite main folder (e.g. cd GO-Elite\_120beta). Once in this directory typing “python GO\_Elite.py” in the terminal window will begin to run GO-Elite (you will see the below interface screen). For command line options, please read the instructions at the end of this section.

When GO-Elite is launched, the user is presented with an introduction screen, where they can get version information or begin the analysis. After selecting “Begin Analysis” the user is presented with three options:

- 1) Analyze Gene Lists
- 2) Prune Existing Results
- 3) Update or Add Databases/Pathways





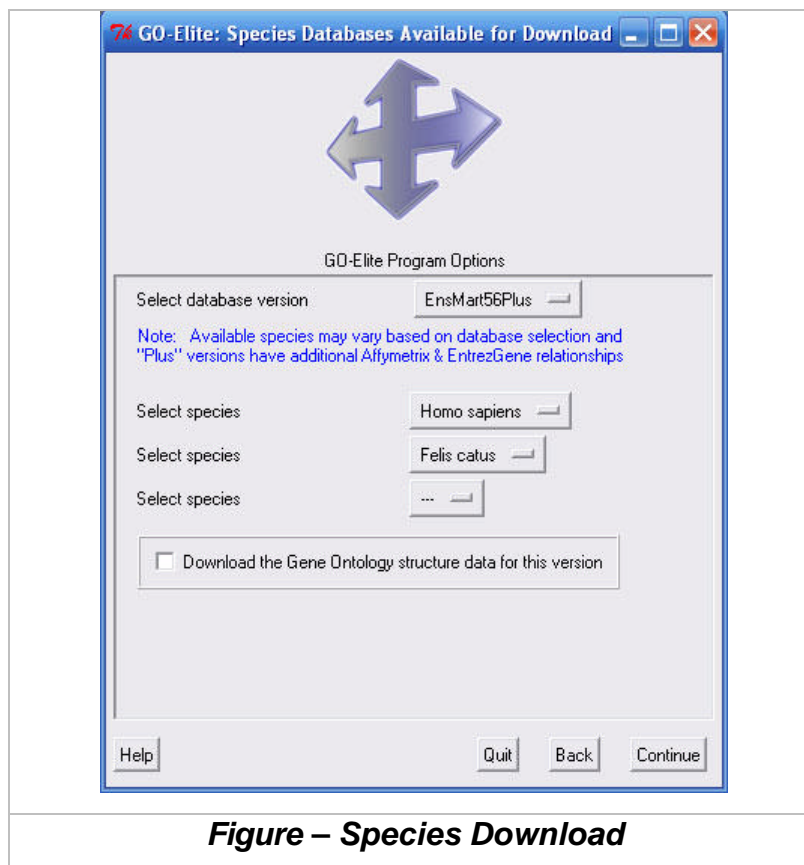
**Figure - Introduction Window**

**Figure - Main Data Analysis Window**

Selecting **“Analyze Gene Lists”** is used when starting with an input and denominator gene list, whereas **“Prune Existing Results”** is only necessary when starting with existing GO or pathway results, which can include previously produced GO-Elite outputs. Selecting **“Update or Add Databases”** is selected if you want to add currently unsupported species gene relationships or update existing relationships. More on this option is available in Section 5.

### Downloading Species Gene Databases

When beginning any analysis for the first time, the user will be prompted to download a species database, after selecting either **“Analyze Gene Lists”** or **“Prune Existing Results”**. Over 60 species gene databases are supported by GO-Elite corresponding to over 50 gene identifier systems. Gene databases are built from either exclusively from the Ensembl database or are further augmented from Affymetrix annotation files.



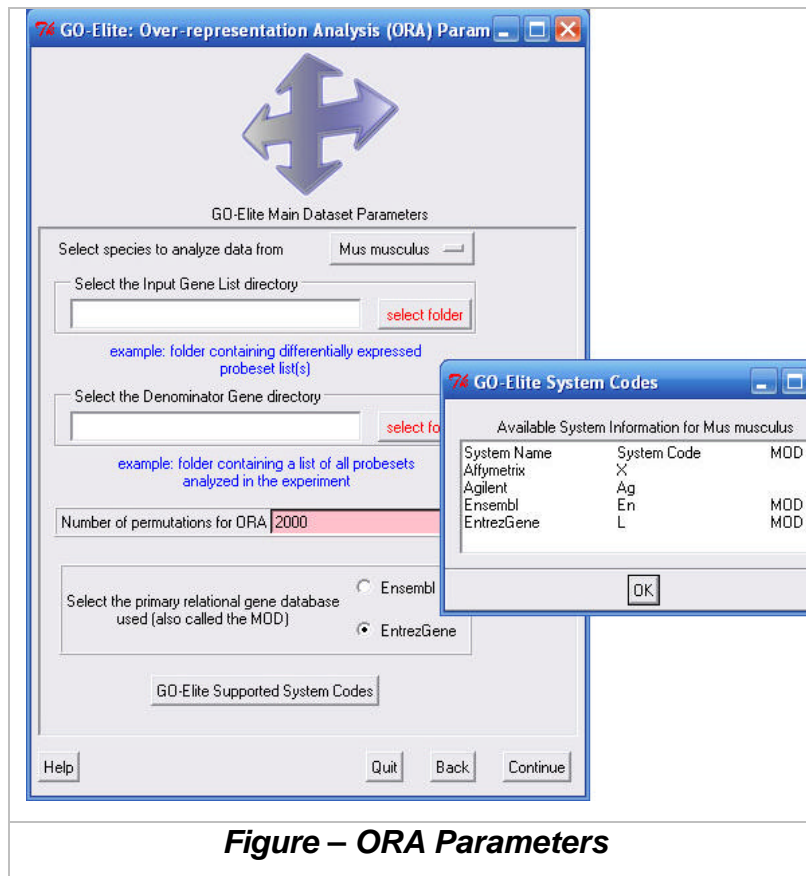
**Figure – Species Download**

When prompted to select a species, find the species and genus name of the organism to be analyzed under the drop-down menu along with the corresponding version of Ensembl. Some species databases will only be available for specific Ensembl versions and augmented databases. Exclusively built Ensembl databases have the name "EnsMart" followed by the Ensembl version number, while augmented databases have the suffix "Plus". Most users will wish to download the "Plus" databases, which contain richer gene annotations for Affymetrix and EntrezGene, however, users of GenMAPP and PathVisio may wish to download the Ensembl exclusive versions, since these match the databases from those programs. If the user selects a version of Ensembl that is not 56, select "Download the Gene Ontology structure data for this version". Select "Continue" to download the databases and proceed with analysis of your data.

This menu can be accessed again by selecting **"Update or Add Databases"** from the previous menu. If your species or gene system of interest is not supported, see section 5 for creating these.

### **Analyze Gene Lists**

When selecting the analysis option **"Analyze Gene Lists"**, the user will be presented with a series of options for performing an over-representation analysis (ORA) on your gene/array ID lists. All options have pre-defined default parameters (defined in Config/options.txt).



This analysis computes an over-representation z-score and permutation p-value for each GO-term/pathway. The options for this menu are:

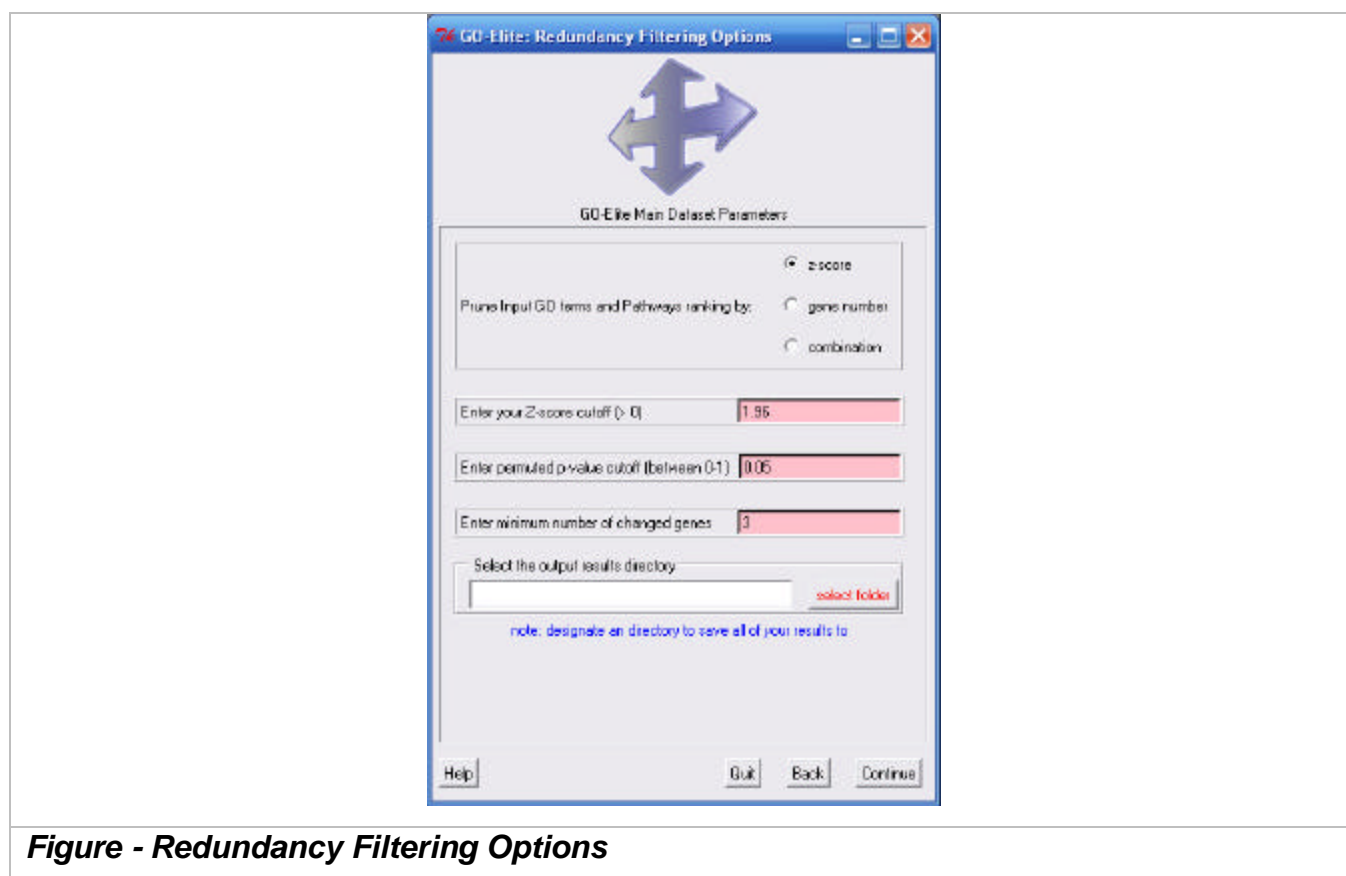
- 1) **Selecting species to analyze data from:** Tells GO-Elite what species your gene lists correspond to.
- 2) **Select the Input Gene List directory:** Tells GO-Elite where your input gene list(s) are located. Multiple input files can be placed in this directory. From the selected menu, double-click on the folder with the input gene files.
- 3) **Select the Denominator Gene List directory:** Tells GO-Elite where your denominator gene list(s) are located.
- 4) **Select the number of permutations for ORA:** Indicates the number of permutations to run for ORA. With increased permutations, you have more statistical confidence but also much longer run-times. The default value of 2000 is recommended for most analyses, but 0 is also acceptable.

- 5) **Select the primary relational gene database:** This option tells GO-Elite which gene ID system to use when linking data to GO terms and pathways. If gene IDs from Ensembl or EntrezGene are used in the input files, simply select the appropriate gene system, but if you are working with other gene IDs (such as Affymetrix), then you typically can choose between either Ensembl or EntrezGene.

Once these options are established, GO-Elite will prompt the user on which filtering statistics to apply to the computed ORA results (**see following section**).

### Pruning ORA Results

This window establishes criterion for filtering pre-processed ORA results from GO-Elite or GenMAPP's MAPPFinder program. Filtering is based on the relationship of GO-terms in the GO-hierarchy, relative ORA scores of such GO-terms and redundancy in gene content for both GO-terms and pathways (see Section 4 - Algorithms). The options are:

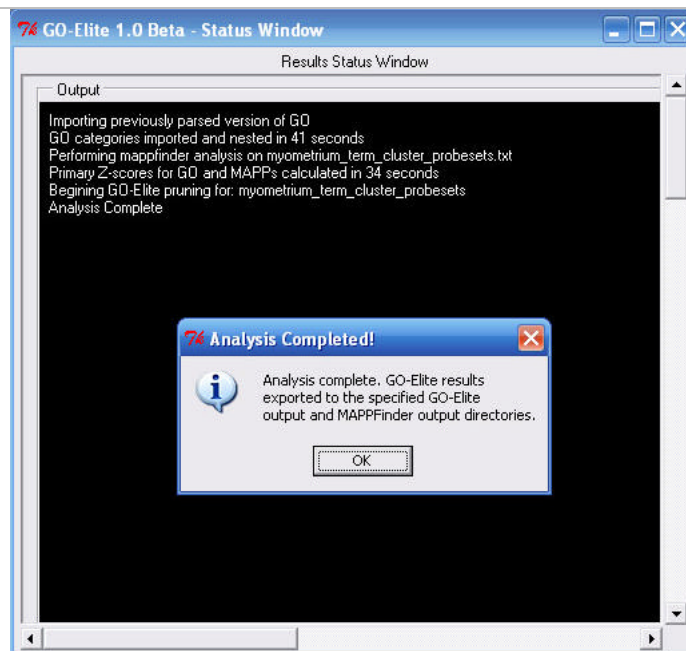


- 1) **Pruning Options for GO terms:** This option instructs GO-Elite which pruning option to use on the output ORA data (this is the Elite step). The options are “**z-score**”, “**gene number**” and “**combination**”. These comparison statistics are used to compare related GO terms that have a **minimum z-score**, **maximum permuted p-value** and **minimum number of genes** designated by the user in the below entry fields. Only those GO-terms or pathways that meet these filters are used for filtering in GO-Elite. The **z-score** option will instruct the program to compare related GO terms based on their relative over-representation z-score, which is an indicator of degree of over-representation. **Gene number** compares the number of genes changed in the pathway/GO term. **Combination** uses the z-score times the log2 number of genes changed to rank related GO-terms. When two or more related GO terms are compared, GO-Elite chooses the GO-term with the highest scoring statistic based on the relative position of the terms in the GO hierarchy. The user is encouraged to try different methods and compare the results.
- 2) **Select the output results directory:** Selecting this button opens a folder selection menu, where the user can choose which directory to save your GO-Elite pruned results to. The results will be stored in a new sub-directory of this folder named “GO-Elite\_results”. If using the “**Analyze Gene Lists**” option, the user’s ORA results will also be saved to this folder in the directory named “GO-Elite\_results/CompleteResults/ORA”. You will notice the result files will be saved to a sub-directory of “ORA”, named “archived-\*time-stamp\*” where the time-stamp indicates the date and time the analysis was run. If the user selected “**Prune Existing Results**”, this directory is the one containing your previously generated ORA file(s).
- 3) **Select the Input Gene List directory:** This option is only present if “**Prune Existing Results**” was selected. It allows the user to locate the directory containing a set of input gene lists that corresponds to the ORA results file. The array IDs/genes in this file will be linked to the results file and used to annotate the results based on which actual genes are changed. The name of the input files should be identical to the ORA results file, where the ORA results may also be

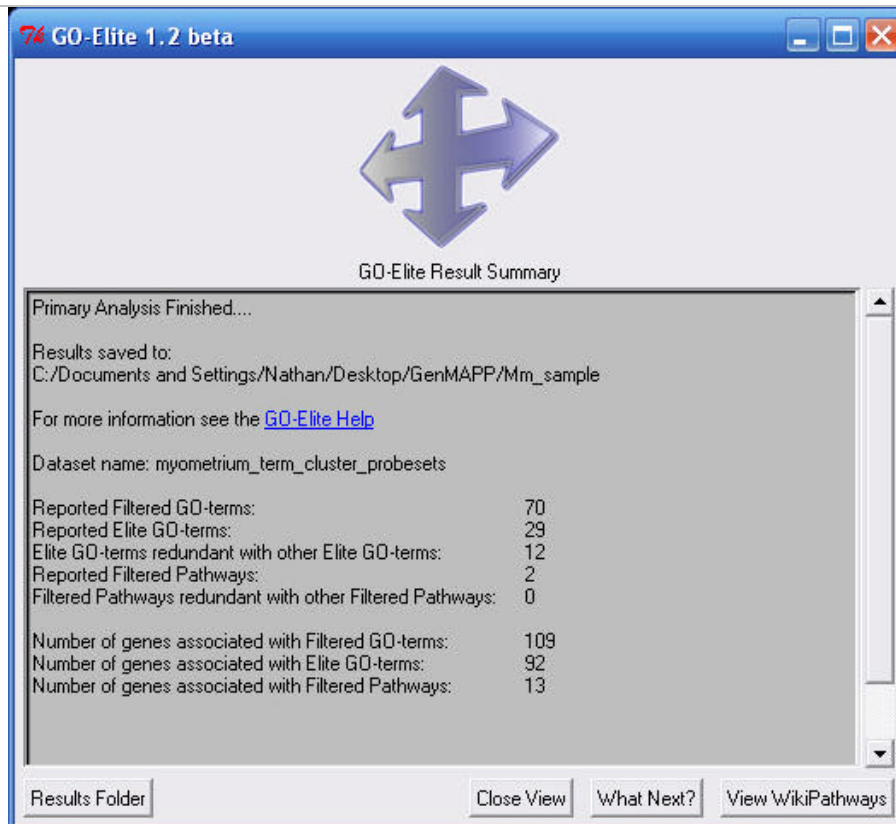
preceded by “-GO.txt” or “-local.txt”.

## Calculating Results

Once the filtering options are selected GO-Elite will present a status window displaying the results of the analysis. Processing can take several minutes to 45 minutes per input gene list, depending on the size of the input gene list and the number of permutations. Once finished, the number of GO terms and WikiPathways matching the user filters will be displayed along with the corresponding number of unique genes (for the selected gene system). This information is exported to the file "GO-Elite\_report.log" in the user output directory. Selecting the option “**Results Folder**” will open the output folder designated by the user. Additional options are present in this window providing link-outs to more information.



**Figure - ORA Calculation**



**Figure - ORA Summary**

## **Running GO-Elite Locally Using the Command-Line Option**

GO-Elite can also be executed using the command line in a terminal or DOS prompt window. To do this, follow the below steps:

- ? Move the GO-Elite source code (GO-Elite application folder/Source\_code) to the GO-Elite application folder
- ? Open a terminal or DOS prompt window
- ? In the terminal, change directories to the GO-Elite application folder
- ? Input the command line options.

For example, on a PC, given a directory of input and denominator text files:

```
python GO_Elite.py --species Mm --mod EntrezGene --permutations
2000 --method "z-score" --zscore 1.96 --pval 0.05 --num 3 --input
```

```
"C:/Mm_sample/input_list_small" --denom "C:/Mm_sample/denominator"  
--output "C:/Mm_sample"
```

These flags instruct GO-Elite to analyze an input and a denominator list using specified statistical parameters (see below). Note that in this example the 2<sup>nd</sup> through 8<sup>th</sup> option represent default values, and as such they can be omitted. Thus, an alternative to the example above would be:

```
python GO_Elite.py --species Mm "C:/Mm_sample/input_list_small" --  
denom "C:/Mm_sample/denominator" --output "C:/Mm_sample"
```

As illustrated, options are presented as flags, preceded by "--". Flags for analysis functions in GO-Elite are:

**--species string** Two letter species code corresponding to the genus and species recognized by GO-Elite (see GO-Elite application folder/Config/species.txt). Examples are "Hs" (human), "Rn" (rat), "Bt" (Bos taurus).

**--mod string** Primary gene system linked to Gene Ontology (GO) or pathways used by GO-Elite. By default this is EntrezGene. The other default mod for GO-Elite is "Ensembl"

**--permutations integer** Number of permutations performed for over-representation analysis (ORA).

**--method string** Pruning Method used by GO-Elite ("z-score", "gene number", combination)

**--zscore float** Remove GO or pathways reported from the ORA with a z-score less than this threshold.

**--pval float** Remove GO or pathways reported from the ORA with a permutation p-value (non-adjusted) greater than this threshold.

**--num float** Remove GO or pathways reported from the ORA with the number of genes changed less than this threshold.

**--input string** Full hard-drive path of the folder containing the text file(s) (list of gene IDs and system code for ORA - e.g., Affymetrix).



`--denom` **string** Full hard-drive path of the folder containing the denominator text file(s) for input list.

`--output` **string** Full hard-drive path of the folder that GO-Elite will save the result folders and files to.

The result files produced by these command line options are identical to those produced by the graphical user interface. In addition to instructions for analyzing data in GO-Elite, the user can pass flags to the program for updating the database or adding new species support. Most of these commands are intended for advanced users/developers wanting to build entire databases from scratch. Below are several examples:

-Download and integrate the latest Gene Ontology OBO format files:

```
python GO_Elite.py --update GO
```

-Download and integrate an official GO-Elite species database version

```
python GO_Elite.py --update Official --species Dr --version 56
```

-Download and integrate the most current EntrezGene-GO relationships:

```
python GO_Elite.py --update EntrezGene --species all
```

-Update Affymetrix-EntrezGene, Affymetrix-Ensembl and EntrezGene and current WikiPathways relationships. Affymetrix files must be in BuildDBs/Affymetrix/\*species-code\*:

```
python GO_Elite.py --update Affymetrix --update WikiPathways --species all --uaffygo no --replaceDB yes
```

-Incorporate the most recent WikiPathways relationships:

```
python GO_Elite.py --update WikiPathways --species all
```

-Add new species information (not necessary when adding Ensembl support for a new species):

```
python GO_Elite.py --addspecies yes --speciesfull "Ciona intestinalis" --species Ci --taxid 7719
```

-Re-build a specific version of Ensembl for a selected species from scratch:

```
python GO_Elite.py --update Ensembl --species Dr --system  
EntrezGene --system UniGene --system "Uniprot/SPTREMBL" --system  
"AFFY_Zebrafish" --system "AGILENT_G2518A" --replaceDB yes --force  
no --version 56
```

-Simultaneously update multiple relationships, ID systems and species:

```
python GO_Elite.py --update Ensembl --update Affymetrix --update  
EntrezGene --speciesfull all --system all --replaceDB no --  
delfiles yes --version current
```

As demonstrated above, there are multiple options when downloading, importing, processing and exporting various gene relational databases. Some flags are specific to certain update options and will be ignored if used in the wrong situation (e.g., `--uaffygo` and `--version`). Flags for update functions in GO-Elite are:

**--update string** This flag directs GO-Elite to update the user provided system relationships. No default arguments for this flag. Arguments for this flag are "Ensembl", "EntrezGene", "GO", "Affymetrix" and "WikiPathways". When choosing the update option "Affymetrix", the Affymetrix CSV files must be saved in the appropriate species folders in the directory "BuildDBs/Affymetrix" prior to update.

**--force string** Default value is "yes". With the argument "yes", this flag directs GO-Elite to download the latest version of the database. If "no" is indicated, GO-Elite will use any previously downloaded build files. If no build files are present and "no" is chosen then GO-Elite will set force equal to "yes".

**--replaceDB string** Default value is "no". With the argument "yes", this flag directs GO-Elite to replace rather than update the existing relationship table. Note: If downloading multiple tables of the same system, such as Affymetrix array platforms, if this option is set to no, then only the first Affymetrix table will be over-written by the other.

**--version string/integer** Indicates the version of Ensembl to download. Default value is "current" (most recent version). Other valid version arguments are numeric (greater than 46 and less than the current version number).

**--system** **string** Corresponds to the name of the gene system related to Ensembl to be exported. For example, the argument “EMBL” will direct GO-Elite to export an Ensembl-EMBL relationship file for use in GO-Elite. These tables are saved to the “uid-gene” folder for that species and are considered secondary gene systems that relate to the primary. As such, the user can construct the input and denominator lists with these secondary system IDs (e.g., EMBL) for GO-Elite analysis. Multiple systems can be designated in a single command line operation. The argument “all” is used to export all related systems whereas the argument “arrays” will export all array systems (e.g., Affymetrix, Illumina, Agilent). See the file “Config/EnsExternalDBs.txt” for a list of compatible systems. After a run, additional external DBs will be provided in “Config/external\_db.txt” and “Config/array.txt”.

**--speciesfull** **string** Alternative to the flag --species, required when the Ensembl species to update is not currently in the database. Rather than the species code (e.g., Ci), the argument for this flag is the full species name (e.g., “Ciona intestinalis”). If the argument “all” is provided, all supported species in Ensembl will be used. Quotations around the species name is required.

**--uaaffygo** **string** Default value is “no”. When the argument equals “yes”, GO-Elite will also extract the Ensembl-GO and EntrezGene-GO relationships from the Affymetrix CSV file(s).

**--addspecies** **string** Default value is “no”. When the argument equals “yes”, GO-Elite will add new species information to the database. Required flags for this option are **--speciesfull**, **--species** and **--taxid**.

**--taxid** **integer** Taxonomy identifier from NCBI (<http://www.ncbi.nlm.nih.gov/Taxonomy/>).

**--delfiles** **string** Default value is “no”. When the argument equals “yes”, GO-Elite will delete any downloaded files used to extract out Ensembl or EntrezGene-GO relationships. These files can take up substantial hard-disk space (up to 3GB for Ensembl, depending on the species), and are not required after data extraction is complete.

Note: All database update functions are available in command-line mode except for “Manually Add New Relationships” functions, which are only supported in the GUI.

## Section 3: Interpreting GO-Elite Results

### GO-Elite Output Files

Four primary file types are exported by GO-Elite with the completion of each run. These include:

- 1) Over-representation analysis (ORA) results (aka ORA)
- 2) Pruned ORA results (aka ORA\_pruned)
- 3) Gene annotation file
- 4) Gene ranking file

Each one of these files is created for each input user gene list Gene Ontology (GO) and local (WikiPathways) results. Since many files can be created when a user is analyzing multiple gene lists, an additional combined file is created for files 2, 3, and 4, which will contain all criterion analyzed in a single run, including GO and local results together. These combined files begin with the name “pruned-results”. Most users will just want to get these combined results. See the folder **GO-Elite\_results**.

### **ORA Results**

These files are similar to those produced by the program MAPPFinder 2.0, a component of the application GenMAPP version 2. For each GO-Elite analysis, both a GO and local ORA results file are created and saved to the folder “CompleteResults/ORA”, in a sub-folder with the appropriate time stamp (e.g., “archived-20091203-162240”). These files have the same name as the input gene list with the addition of the suffix “-GO.txt” or “-local.txt”. Information on the date of the OBO files used, number of gene identifiers (IDs) in the user input file, number of genes linked to GO or pathways, and input file name are reported at the top of this file.

1.GE-cardiac-specific-expanded-non-adjp-up-GO.txt				
	A	B	C	D
1	GO-Elite MAPPFinder Results			
2	File:			
3	Table:			
4	Database: Based on OBO-Database version: 12/26/2007			
5	colors:			
6	5/24/2008			
7	Homo sapiens			
8	Pvalues = true			
9	Calculation Summary:			
10	527 Ensembl source identifiers supplied in the input file:1.GE-cardiac-specific-expa			
11	527 source identifiers meeting the filter linked to a Ensembl ID.			
12	357 genes meeting the criterion linked to a GO term.			
13	29151 source identifiers in this dataset.			
14	29151 source identifiers linked to a Ensembl ID.			
15	17438 Genes linked to a GO term.			
16	The z score is based on an N of 17438 and a R of 357 distinct genes in the GO.			

**Figure - ORA Result Headers**

The main contents of this file are GO or pathway names, IDs, number of genes changed, number of genes associated with each term, z-scores, permutation p-values, and Benjamini-Hochberg adjusted (5) p-values (Figure 2). There are also a number of blank columns. These would typically contain non-nested GO gene associations, which are not currently calculated in GO-Elite. Both the GO and local files have the same format that you would get from running MAPPFinder and have similar results.

These files list GO and [WikiPathways](#) pathways (local) gene associations to genes from your input and denominator files along with calculated statistics for ORA (z-score, permutation p-value and Benjamini-Hochberg false discovery rate p-values). To determine which pathways might be of most interest, open either of these files in MS-Excel. Select the row containing the column headers (e.g. row 18 for the local file) and then select the menu item “Data”, “Filter”, and “Autofilter”. A series of dropdown menus appear above each field in this row. We recommend filtering based on the “Number Changed” (>2), “Z Score”(>1.96), and “PermuteP”(<0.05). Performing this filter or other user defined filters will reduce the set of results from hundreds to a handful of pathways that are over-represented in your analysis.

GOID	GO Name	GO Type	N	N	N	P	P	Number Changed	Number Measured	Number in GO	Percent Changed	Percent Present	Z Score	PermuteP	AdjustedP
31014	troponin T binding	F						2	2	2	100	100	9.782503	0	0
42574	retinal metabolic process	P						2	3	3	66.66667	100	7.904137	0	0
43288	apocarotenoid metabolic process	P						2	3	3	66.66667	100	7.904137	0	0
43462	regulation of ATPase activity	P						2	3	3	66.66667	100	7.904137	0	0
30049	muscle filament sliding	P						3	5	5	60	100	9.151974	0	0
33275	actin-myosin filament sliding	P						3	5	5	60	100	9.151974	0	0
5220	inositol 1\,4\,5-triphosphate-sensitive calcium-rel	F						3	6	6	50	100	8.295782	0	0
5861	troponin complex	C						3	6	6	50	100	8.295782	0	0
15278	calcium-release channel activity	F						3	6	6	50	100	8.295782	0	0
2027	regulation of heart rate	P						3	8	8	37.5	100	7.082523	0	0
5790	smooth endoplasmic reticulum	C						4	8	8	50	100	9.579693	0	0
55010	ventricular cardiac muscle morphogenesis	P						4	8	8	50	100	9.579693	0	0
48644	muscle morphogenesis	P						5	9	9	55.55556	100	11.3383	0	0
55008	cardiac muscle morphogenesis	P						5	9	9	55.55556	100	11.3383	0	0
35050	embryonic heart tube development	P						3	10	11	30	90.90909	6.243707	0	0
2026	regulation of the force of heart contraction	P						4	12	12	33.33333	100	7.655697	0	0

**Figure - ORA Statistics**

In these result files, you will likely have many terms that are related to each other and thus have largely redundant gene content. For example, “negative regulation of translation”, “negative regulation of cellular biosynthetic process”, and “negative regulation of biosynthetic process” are all top regulated terms in this results file. Thus, these files are further pruned by GO-Elite to report a minimally redundant set of ORA terms.

### Pruned GO-Elite ORA Summary Results

The pruned ORA files are similar to that of the MAPPFinder output except in that the reported terms are pruned to only include non-redundant information. These files can be opened in Microsoft Excel or similar spreadsheet program and include a summary of the gene symbols for any changed genes and the mean of any numeric data in the input gene list file. The files are saved to the folder “GO-Elite\_results” with the named “pruned-results” and to “CompleteResults/ORA\_pruned” with the same name as the original input gene ID file with the suffix pruning algorithm + “\_elite.txt”.

### Pathway Gene Association Files

These files contain all GO and pathway terms listed in the pruned ORA summary result’s file along with associated input gene IDs and gene annotation information. These files are saved in the folder “GO-Elite\_results” to “pruned-gene-associations.txt “ and to the folder “CompleteResults/ORA\_pruned/gene\_associations” with the suffix “-gene-

associations.txt". Each line in this file consists of a gene name, symbol, primary ID, source IDs, GO term or pathway and associated user-data in the input file.

### **Gene Ranking Files**

These files contain all genes listed in the GO-Elite gene-association file, the number of pruned ORA summary terms associated with that gene, percentage of terms the gene is associated with, and associated terms are listed on each line. These files are also saved to the "gene\_associations" file directory. These genes are listed in descending order, based on the number of terms each is associated with this. Unlike the gene-association files, each gene is only listed once. This file can be used determine which sets of genes may be over-represented among GO-Elite terms.

### **Combined Files**

The combined files for pruned ORA results, gene association, and gene ranking files are stored into a single file for each of these three categories. These include all input gene lists analyzed in a single GO-Elite run, GO, and local pathway results. Each line is identical to that in the original input file, except in that the first column indicates the filename of the data it was derived from. These are typically the most useful files, since they contain all of the data combined into a single spreadsheet.

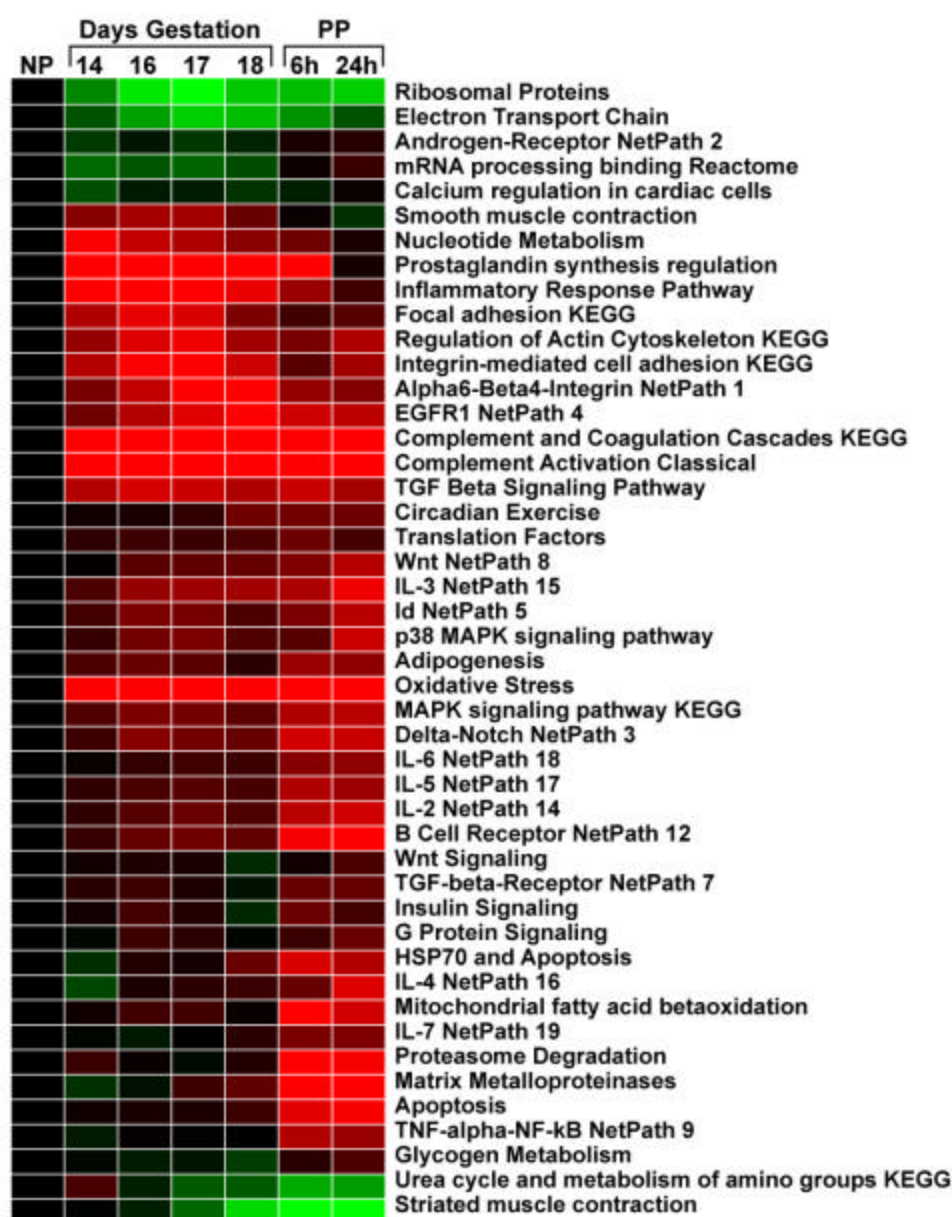
### **Downstream Analyses**

As you've seen, GO-Elite is a useful tool to summarize gene-level changes from a microarray experiment for both pathways and GO terms. In addition to these options, if the user includes gene expression changes directly in their input file, these will also be summarized at the level of GO terms and pathways. This analysis can be useful if your input file contains data for multiple time-points or conditions and contains more than one predicted expression pattern.

### **Expression Summarization of Pathways**

If your input gene file contains numeric data from your experiment, this will also be

summarized in the GO-Elite output files. For example, if there are fold changes for each probeset in the input file, all fold changes for probesets associated with genes in a particular pathway will be averaged for that pathway, likewise for GO-terms. If your input file contains such data, the “pruned-results” file, will contain similar fold changes to those found in the input gene list, but summarized for each gene (where there are multiple probesets associating) and each pathway. These expression results can be graphed in a graphing program or clustered in an expression clustering and visualization program like Cluster and TreeViewer.





## Section 4: Algorithms

### GO and Pathway Over-Representation Analysis

A z-score and permutation p-value are calculated to assess over-representation of GO terms and pathways (local). The z-score is calculated by subtracting the expected number of genes associated with biological term from the observed number (input list) of genes and dividing by the standard deviation of the observed number of genes. This z-score is a normal approximation to the hypergeometric distribution. This equation is expressed as:

$$z = \frac{(\text{observed} - \text{expected})}{\text{std.deviation}(\text{observed})} \quad z = \frac{\left(r - n \frac{R}{N}\right)}{\sqrt{n \left(\frac{R}{N}\right) \left(1 - \left(\frac{R}{N}\right)\right) \left(1 - \frac{n-1}{N-1}\right)}}$$

n = All genes associated with a given element

r = Alternatively regulated genes associated with a given element

N = All genes examined

R = All alternatively regulated genes

Once z-scores have been calculated for all domains/motifs and miR-BSs linked to alternatively regulated probe sets, a permutation analysis is performed to determine the likelihood of observing these z-scores by chance. This is done by randomly selecting the same number of regulated probe sets from all probe sets examined and recalculating z-scores for all terms 2000 times. The likelihood of a z-score occurring by chance is calculated as the number of times a permutation z-score is greater than or equal to the original z-score divided by 2000. A Benjamini-Hochberg (BH) correction (5) is used to transform this p-value to adjusted for multiple hypothesis testing.

### Multiple Identifier Mapping for ORA

Some primary identifier (ID) systems, such as Affymetrix probesets can have a single primary ID that maps to multiple genes or multiple primary IDs which maps to a single gene (e.g. Ensembl). One of the main functions of GO-Elite's mappfinder function is to not

count a primary ID, such as probeset, more than once in to a GO-term or pathway. This philosophy minimizes falsely over-weighted results.

Below is the schema which GO-Elite's mappfinder function uses to build unique primary ID and gene to GO/pathway associations. These methods are largely similar to the program MAPPFinder 2.0's (a component of GenMAPP version 2), implementation. The below schema is illustrated for Affymetrix probesets.

- 1) Import all probesets in the criterion (input file in GenesToQuery directory) and denominator file.
- 2) Link the probesets to gene IDs (e.g. Ensembl) via the 'Ensembl-Affymetrix.txt' file in the 'uid-gene' Database species directory.
- 3) Import the gene to GO relationships (e.g. 'Ensembl\_to\_Nested-GO.txt').
- 4) Send the denominator gene-probeset relationships (where the gene is the unique ID or key) and gene-nested GO relationships to the mappfinder.py function 'countGenesInPathway'. For each gene in the gene-nested GO table, store the unique set of probsets corresponding to the gene ID for that gene with that GO term. Repeat the same process with just those IDs found in the input criterion file. This should produce the following theoretical results: If gene 1 maps to Probe X and Probe Y, store 'GO\_count\_db[GO\_term] = [Probe X, Probe Y]'.  
This should produce the following theoretical results: If gene 1 maps to Probe X and Probe Y, store 'GO\_count\_db[GO\_term] = [Probe X, Probe Y]'.
- 5) For each GO term examined determine the number of unique sets of probesets linked to genes (rather than count unique gene IDs count unique probeset lists). Also store this probeset list ([Probe X, Probe Y]) for each gene as the key in another database to find how many unique sets of probesets there are corresponding to genes (unique gene count linked to GO).

### **Scenario 1**

If you have the genes 1 and 2, that both match to probesets X and Y, the list of probesets X,Y will be counted once for each GO term and in the total number of genes linked to GO.

### **Scenario 2**

If you have two genes, 1 and 2 and X associates with 1 and X & Y associate with 2, then two unique genes are counted.

For scenario 1, if only probeset X is in your input list, then only one gene is count (since X becomes the unique ID for both genes, and doesn't consider Y). For scenario 2, although gene 2 links to X & Y, since only X is present genes 1 and 2 are counted only once, since both ONLY link to X.

GenMAPP's MAPPFinder implement a similar strategy (we've compared results but not the Visual Basic code directly), but only for the criterion probeset-gene relationships. For the denominator, it appears that the program counts the number of unique Ensembl's linked to probesets, even if the associations are redundant. Some people I prefer the GenMAPP method, however, for GO-Elite we decided to calculate the numerator and the denominator using the same method.

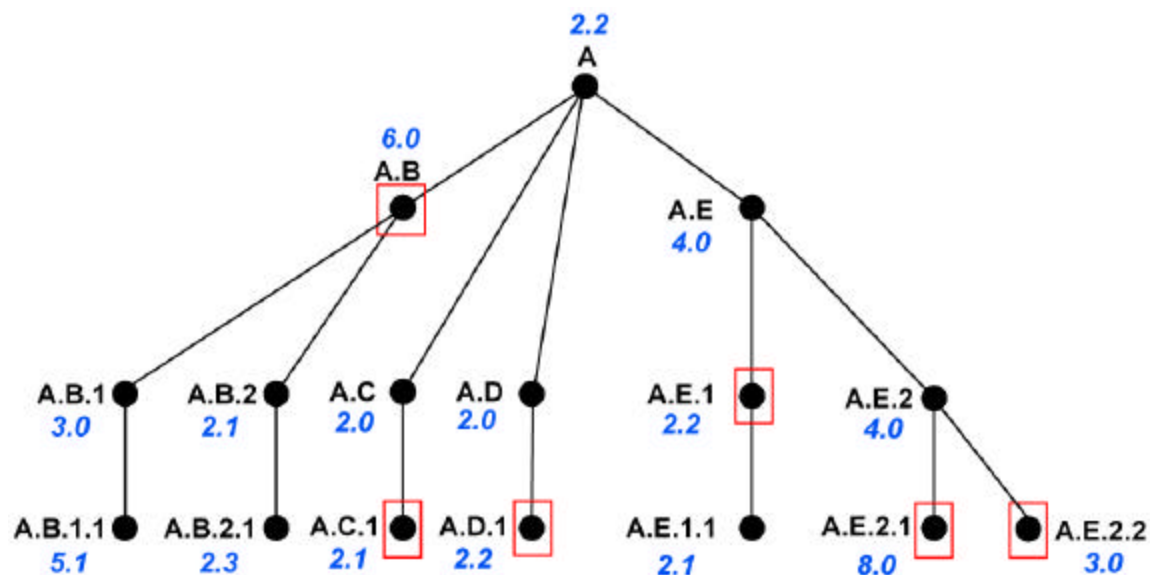
### **Filtering of ORA Results**

The Fisher's exact z-score, number of genes changed and non-adjusted permutation p-value generated by either the mappfinder module or GenMAPP are the default statistics used for pruning GO terms and pathways in GO-Elite. When ORA data from other GO and pathway analysis programs are used as input for GO-Elite filtering, analogous statistics are recommended. Upon import of ORA data, only those GO terms and pathways that meet the user defined minimum filters (by default, permuted p-value < 0.05, z-score > 2 and number of genes change > 2) are processed for redundancy. Once imported, GO-Elite will compare related GO terms based one of three possible options: 1) z-score, 2) number genes changed or 3) combination. The z-score option ranks GO terms only based on the calculated z-score, ranked from higher to lowest.

The gene number option allows the user to rank based on the number of genes changed in the GO hierarchy, again from highest to lowest. The combination option is a weighted metric based on both number of genes changed and z-score, generated by multiplying the z-score times the log base 2 of the number of genes changed for a given GO term. These scores are used for selecting which GO terms will be reported by GO-Elite.

## Pruning of Gene Ontology Hierarchical Relationships

GO-Elite can process different types of ORA files, corresponding to either GO results (file suffix “-GO.txt”) or pathway results (file suffix “-local.txt”). For GO-level results, after GO terms are initially filtered based on user defined statistics (permutep, number of genes changed and z-score), all possible parent-child relationships are built and stored for these GO terms, where each parent is the key in the database (Python dictionary object) and all of its children are the value. This full database is stored for later queries, while the full parent-child paths (agglomerated path relationships) for all entries are generated by iterating this process. The program then searches these relationships in a hierarchical nature to identify the most significant scoring GO term that either has a higher score than all of it's children (along that branch of the tree) or sibling terms (children of a single parent, each representing distinct branches) where at least one of the sibling terms on a branch scores greater than the parent. For these sibling terms, if one sibling branch scores higher than the parent and another branch does not, the highest scoring term from the latter sibling branch is also selected for the GO-Elite output, but the parent term is not. A visual representation of this pruning strategy is shown for a theoretical set of parent-child relationships with corresponding z-scores below.



Step 1) Build all possible parent-child relationships.

Step 2) Find parents from this list more significant (see score options) than all of their children

Step 3) Find the most significant child terms (downstream of the last bifurcation)

Step 4) Eliminate terms from step 3 that are children of any other term from step 3

Step 5) Report the most significant parent OR child terms

This process allows the user to view the highest scoring term(s) for a particular branch of GO terms and eliminates redundancy of GO terms within the same global category (e.g. biological process, molecular function and cellular component), without needing to consider associated gene content. Since some terms and branches are replicated within the GO hierarchy (redundant), already eliminated or selected GO terms are removed from the results from these other branches.

## Section 5: GO-Elite Gene and Pathway Databases

### Database Files Overview

GO-Elite stores all of its gene, pathway and Ontology relationships locally as text files within the directory named “Databases” of the GO-Elite program folder. If the user downloads the “Official Database” releases, these files will be stored in a folder within the “Databases” directory with the name of the database version (e.g., “EnsMart56Plus”). These databases can be easily updated, modified or replaced by the user within GO-Elite. In general there are three ways to update GO-Elite gene databases; 1) Download of the official GO-Elite databases, 2) selection of specific gene systems and relationships to update from files online and 3) addition or update of species tables from custom user tables. The following sections address how to perform these update functions for any database relationship. **Note: most relationships needed to run GO-Elite for a species are already present in the pre-packaged species databases.**

### Downloading Official GO-Elite Databases

With each new version of the Ensembl database, two official GO-Elite databases are released. The first is exclusively built from the Ensembl database (e.g., EnsMart56) and the second contains additional relationships and species support from Affymetrix (EnsMart56Plus). **Most users will wish to download the Plus database, which supports more species and more gene relationships.** If a user wished to, they can fully recreate these databases in GO-Elite using the existing automated tools (see following sections).

In release 56 of the Ensembl database over 50 species are supported (EnsMart56). The EnsMart56 database contains only gene relationships extracted from Ensembl and includes all supported gene associations in the Ensembl database (e.g., Gene Ontology-to-Ensembl, Affymetrix-to-Ensembl, Agilent-to-Ensembl, RefSeq-to-Ensembl). An augmented version of this database (EnsMart56Plus) additionally has support for 14 other species, not found in Ensembl but for which there is support in

Affymetrix annotation files. These augmented databases are built on top of the Ensembl specific versions, but also contain Affymetrix-to-EntrezGene and Affymetrix-to-Ensembl relationships from annotations provided by <http://www.affymetrix.com> as well as EntrezGene-to-Gene Ontology annotations from NCBI. These additional Affymetrix relationships can be useful, since EntrezGene often annotates organisms where Ensembl does not and Ensembl can exclude some valid probeset-gene relationships. However, since the non-augmented database contains the same content found in databases for the programs GenMAPP and PathVisio's, users might want to restrict their analysis to the Ensembl only database. Both database versions contain gene to WikiPathways associations, obtained from the WikiPathways website.

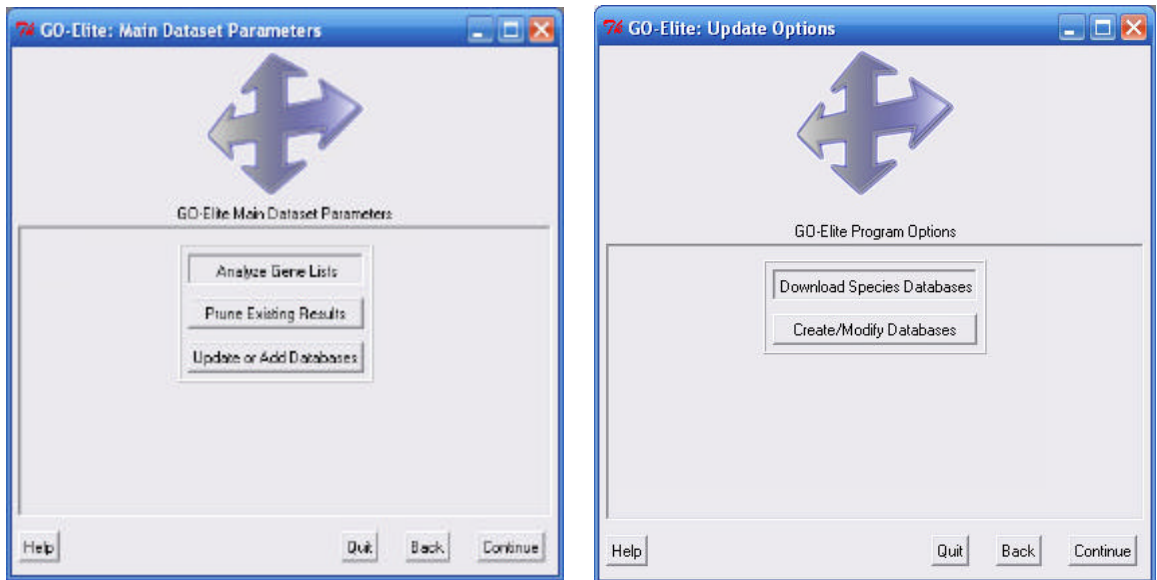
These databases can be downloaded when the user first begins GO-Elite or at any time from the user interface. To update databases:

a) Start GO-Elite



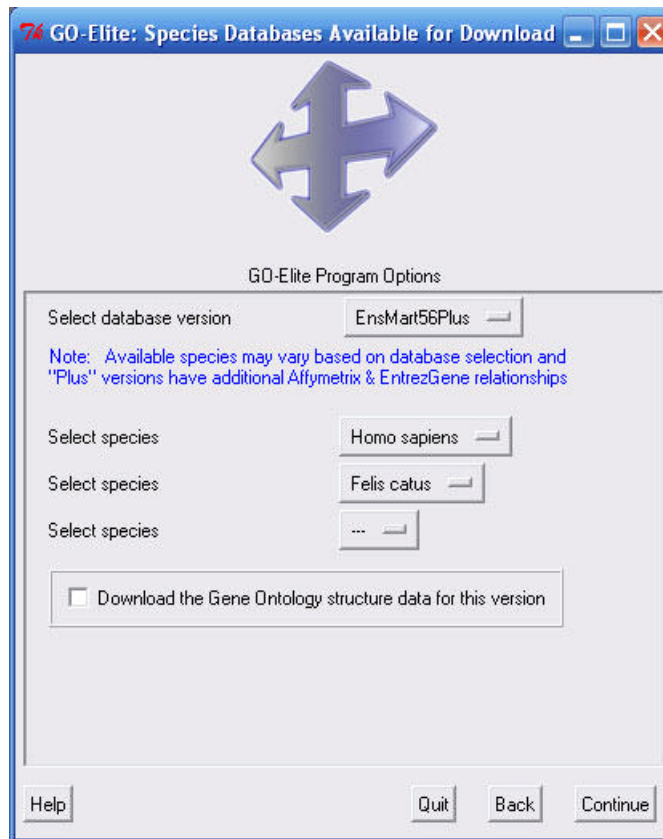
b) Select "Update or Add Databases" and select "Continue".

c) Select "Download Species Databases" and select "Continue".



- d) Select the GO-Elite database version and species you wish to analyze. If you select a database version with the suffix “Plus”, it will include additional relationships for Ensembl and EntrezGene to Affymetrix and EntrezGene to Gene Ontology, mentioned above.
- e) If selecting a database version other than Ensembl 56, the user may also want to download the Gene Ontology structure relationships that correspond to the specific database version. To change this Gene Ontology data, select “Download the Gene Ontology structure data for this version” and select “Continue”.





- f) GO-Elite will then download and extract these annotations to the appropriate folders within the “Databases” directory. These relationships are now available for use in GO-Elite.

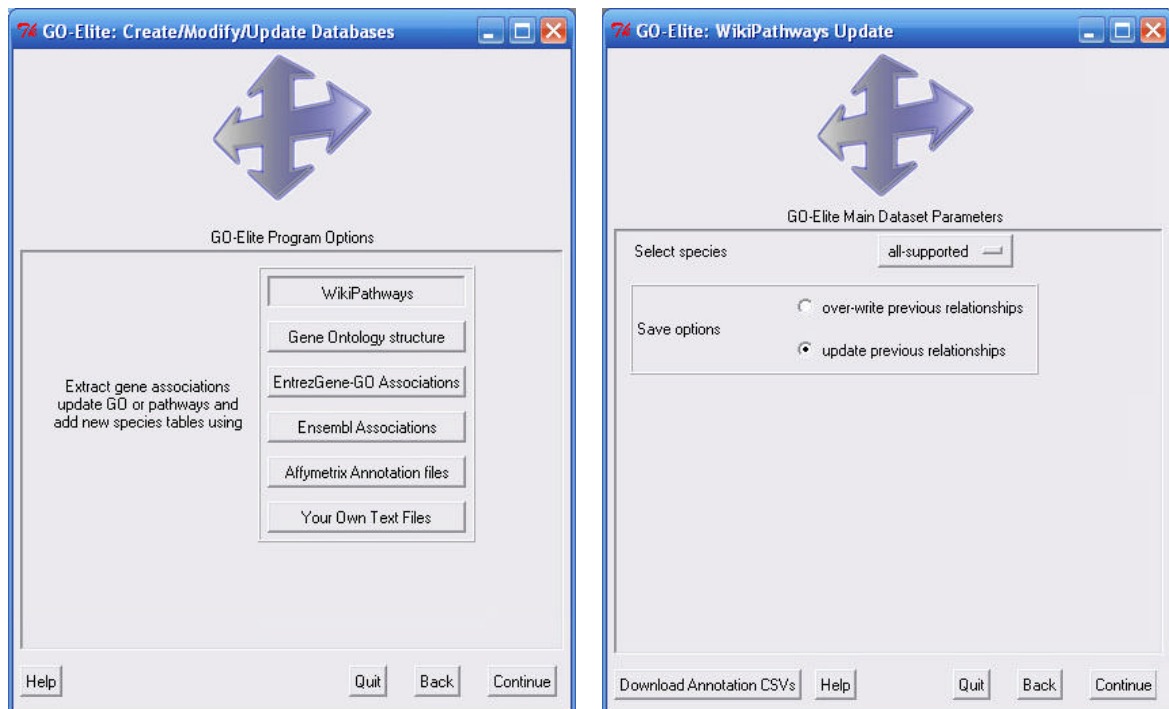
If a user has multiple official database versions, they will be able to select the desired version when analyzing data or updating databases. Once the databases are installed, the user can choose to keep these databases, update or revert to a previous version if available or build new or modify existing databases using built in tools (see following sections).

### **Updating WikiPathways Relationships**

WikiPathways is an online resource for contribution of biological pathways by the scientific community. This information is constantly updated with pathway data available for many species (typically those supported by Ensembl), often via species inference. GO-Elite can use pathway information from GO-Elite or custom pathway relationships provided by the user (see following section). Gene to WikiPathway relationships are

downloaded from a flat-file provided at [wikipathways.org](http://wikipathways.org), that contains the primary gene IDs for one of several gene systems (e.g., Ensembl, EntrezGene, UniProt) for each gene object, as well as related IDs in other gene systems. Although most gene relationships are extracted directly from the downloaded file, GO-Elite also tries to infer missing gene relationships using a class of tables known as “meta” files that contain gene relationships already in the database. To update WikiPathways:

- a) From the main menu select “Update or Add Databases”, “Create/Modify Databases” and then “WikiPathways”.



- b) Select your species of interest (e.g., Bos Taurus). Selecting “all-supported” will update all species in your database.
- c) Keep or change the default save options. These are:
  - a. Over-write previous relationships OR update previous relationships. The first option will replace the existing table entirely with the new information while the second option will add to the existing annotations.

## **Addition and Update of Species Databases**

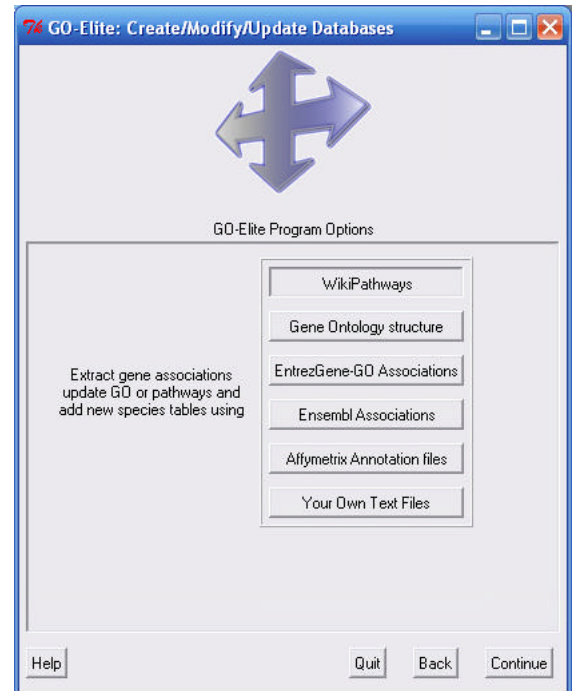
While the GO-Elite Official databases supports a large number species and gene relationships, the user may wish to add support for new species or add additional relationships not included in the existing databases. Users may also wish to add relationships to existing tables. All of these options are supported in GO-Elite through the “Create/Modify Databases” menu. Options include:

- 1) Register any new species
- 2) Add/replace gene annotation files (e.g., Ensembl)
- 3) Add/augment/replace gene to Gene Ontology or pathway relationship files.
- 4) Add/augment/replace gene relationship files (e.g., Affymetrix-EntrezGene).

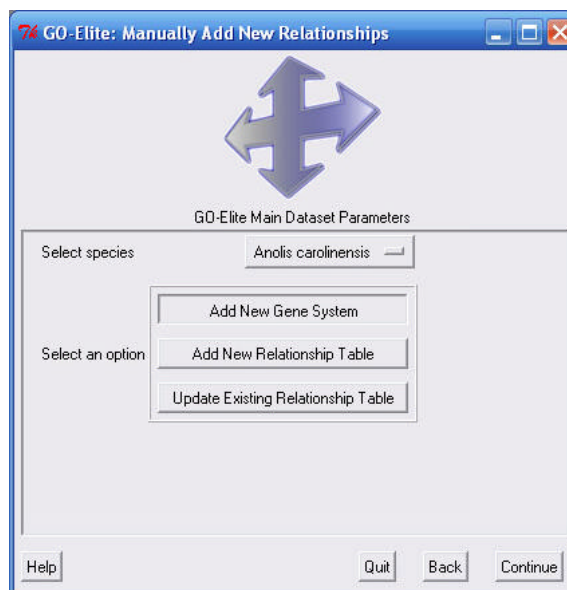
When augmenting or replacing any existing gene relationship tables, this process is simple and straightforward. When building support for an unsupported species, this process is also straightforward, but requires the addition of all essential tables prior to data analysis (gene and gene-GO or gene-MAPP). Both strategies are described herein.

### **Augmenting or Replacing Existing Relationship Tables**

- a) From the main menu select “Update or Add Databases”, “Create/Modify Databases” and then “Your Own Text Files”.

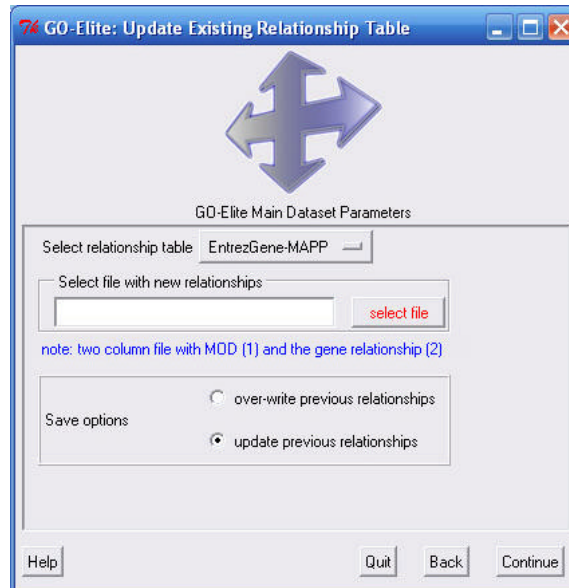


- b) Select the species you want to analyze from the dropdown menu. If this is a new species, skip ahead to the next section “Add Custom Species Support”.
- c) Select “Update Existing Relationship Table” and select Continue”.



- d) In the selected window, the user can select from any existing relationship table in the GO-Elite database. This includes gene-WikiPathway annotations (e.g.,

EntrezGene-MAPP), gene-GO (e.g., Ensembl-GeneOntology) and uid-gene (e.g., Ensembl-Affymetrix).



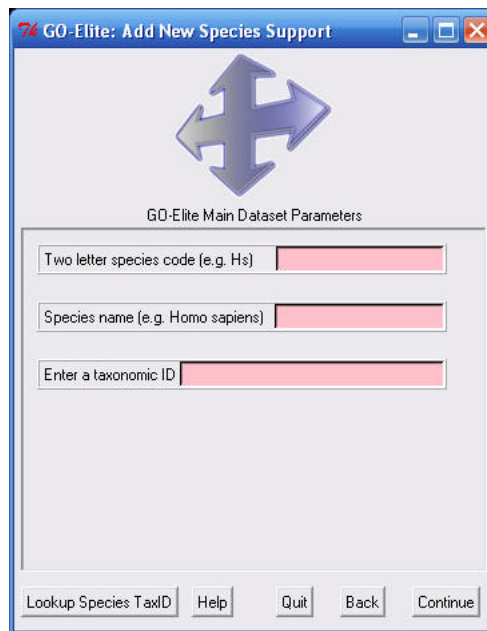
- e) Select the relationship table and select the file from your computer that contains the relationships to add/replace. This table has two columns; the first needs to have the primary gene ID (aka MOD) and the second is the related ID.
- f) Keep or change the default save options. These are:
  - a. Over-write previous relationships OR update previous relationships. The first option will replace the existing table entirely with the new information while the second option will add to the existing annotations.
- g) Once completed, the existing relationship table will contain the new information from your two-column text file. To verify, you can open this relationship file in a spreadsheet editor, like Microsoft Excel (saved to the folder "Databases/EnsMart\*build\*/\*species\_code\*" in the designated output directory).

### Add Custom Species Support

If your species or gene relationships of interest are not available from either the Official GO-Elite gene databases or can be integrated using various means. If the user is analyzing data from an unsupported species, where an Affymetrix microarray exists, skip ahead to the section **"Updating Affymetrix Relationships"**. If the user knows that EntrezGene-GeneOntology information exists at NCBI for that species, skip ahead to the

section “**Updating Gene Ontology Relationships from EntrezGene**”. To add new species support:

- a) From the main menu select “Update or Add Databases”, “Create/Modify Databases” and then “Your Own Text Files”.
- b) From the species selection dropdown menu, select “New Species” and “Continue”.



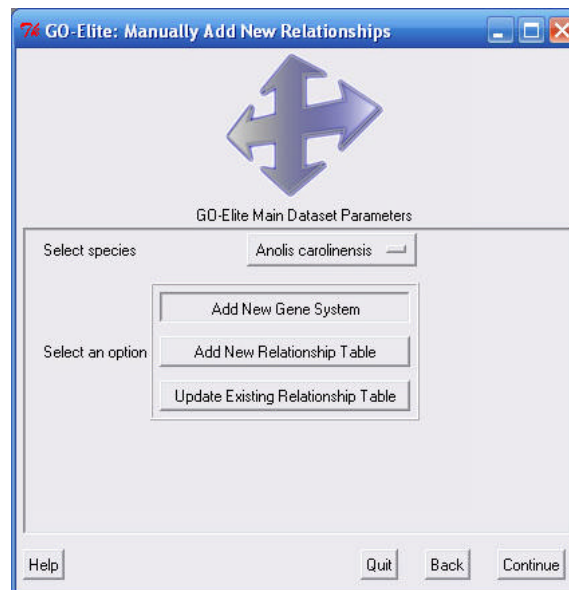
- c) In the new window that appears, add the two-letter species code (e.g., Bt for Bos taurus), species name (e.g., Bos taurus) and taxonomic ID from NCBI for that species. To figure out which taxonomic ID applies to your species, select the button named “Lookup Species TaxID” at the bottom of the menu, enter the species name in the web browser and select the first link. For Bos Tuarus this is “9913”.
- d) Select “Continue” once these fields are filled in.
- e) Now your full species name will appear in the species dropdown menu for the update selections.

### Add New Gene System

Adding a new gene system allows the user to add any relationship files that include the new gene system. For example, if add support for a new system named “Bob’s gene IDs”

you can then create additional tables that either directly or indirectly link these new IDs to existing or user designated primary gene relationships (aka MOD). Thus, “Bob’s gene IDs” could link directly to GeneOntology terms and pathways or indirectly through an existing MOD, such as Ensembl.

- a) From the main menu select “Update or Add Databases”, “Create/Modify Databases” and then “Your Own Text Files”.
- b) Select a species from the dropdown menu and the option “Add New Gene System” and “Continue”.



- c) Once selected, you will be asked to enter a gene system name and gene system acronym. It is recommended that the acronym be two letters, but can be longer. You can add gene systems that are already in the database if you want to change the system name, system code or MOD status.

GO-Elite: Add New Gene System

GO-Elite Main Dataset Parameters

Enter the new system name (e.g., EMBL)

Create a new system code (e.g., Em for EMBL)

Is this a primary gene system (aka MOD) linking to the GO? ☐ yes ☒ no

Select file with new gene annotations (OPTIONAL)

note: three column file with gene ID (1), gene symbol (2) and description (3)

- d) If the gene system will be used to directly link to Gene Ontology or pathways, you can designate this system as a MOD (Model Organism Database), in order to later upload these relationships.
- e) Finally, if this is a MOD that will be directly linked to Gene Ontology or pathways in your gene database, you can choose to upload a three column file containing gene annotations for that gene system. Although this interface can be used to add gene systems not currently in the gene database, for any species, it can be used to replace an existing gene annotation file.
- f) Once completed, the existing relationship table will contain the new information from your three-column text file. To verify, you can open this annotation file in a spreadsheet editor, like Microsoft Excel (saved to the folder "Databases/EnsMart\*build\*/\*species\_code\*/gene).

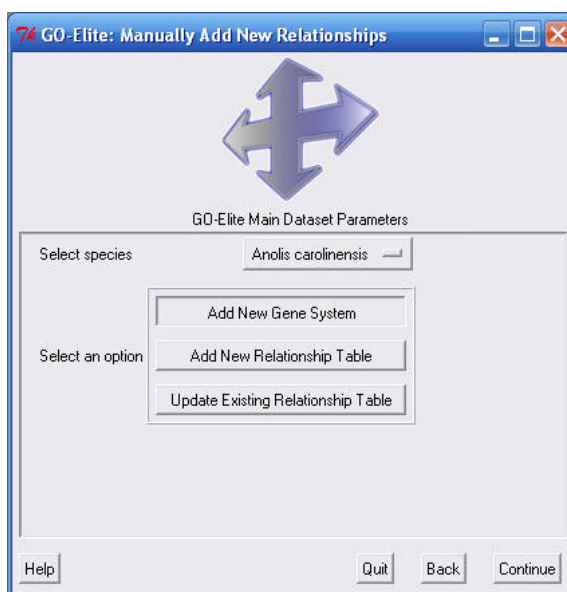
### Add New Gene Relationship Files

Gene relationship files are essential for performing over-representation analysis (ORA) along Gene Ontology terms and pathways. There are three types of gene relationships files; 1) gene-GO, 2) gene-MAPP and 3) uid-gene. This first type contains relationships between your primary gene system and Gene Ontology. This primary gene system can



be an existing one, such as Ensembl, or one added in the previous menu. The second type contains non-Gene Ontology associations, to which ever “pathway” annotations the user has. These can be from any desired source, including custom annotations. The third type of relationship file contains unique ID to primary gene system relationships. An example is Affymetrix to EntrezGene relationships. If the user’s data is Affymetrix IDs these can be analyzed for ORA when this table and EntrezGene-GeneOntology relationships are present. Typically, each gene ID within a primary gene system (MOD) represents a single genomic gene. This is important when assessing over-representation of genes in a GO term or pathway, however, in specific cases where the user is not interested in gene level results a MOD can be used where each gene is linked to multiple MOD IDs. To add a new relationship file:

- a) From the main menu select “Update or Add Databases”, “Create/Modify Databases” and then “Your Own Text Files”.



- b) Select a species from the dropdown menu, the option “Add New Relationship Table” and “Continue”.



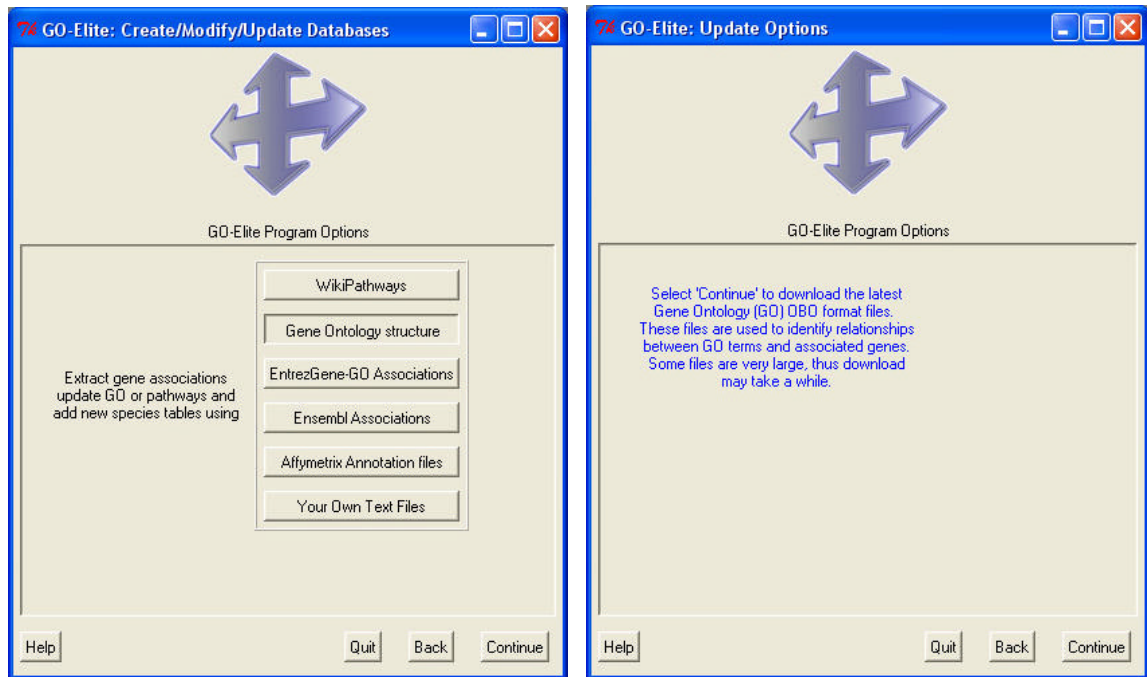
- c) Two dropdown menus will be presented in the selected window, one for the MOD and the other for the related gene IDs. Select the MOD (existing or added) and related IDs corresponding to the file you wish to upload. The file must be a tab-delimited text file with two columns; MOD (1) and gene relationship (2) for the designated systems. If adding support for a new species, you will need to add MOD-GeneOntology or MOD-MAPP associations as a minimum to perform ORA.
- d) Repeat this process until all GeneOntology/Pathway and unique ID -MOD associations have been added. You can later modify these tables using the "Update Existing Relationships Table" menu if needed.
- e) Once complete, you are ready to analyze your data (see Section 2).

### **Updating Gene Ontology Structure Annotations**

GO tree structure annotations can be downloaded for each version of the GO-Elite database downloaded using the "Download Species Database" menu. These files annotations are extracted from the much larger OBO format files from the Gene Ontology website. If the user wishes to download the very most recent annotations, the large OBO format files can be directly downloaded in GO-Elite. The version date of the Gene Ontology OBO files can be found in the file named "version.txt" in the program folder "OBO", to see what version you have. This version date is also written to the intermediate

ORA result files in the user output folder "GO-Elite\_results/CompleteResults/ORA ". To update these files from GO-Elite:

- a) From the main menu select "Update or Add Databases", "Create/Modify Databases" and then "Gene Ontology Structure".



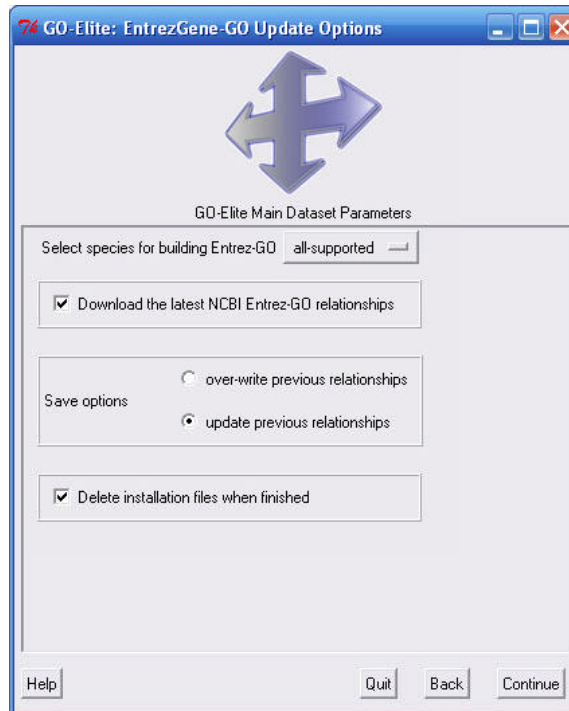
- b) When ready to begin the file download, select "Continue". This will replace any existing Gene Ontology OBO format files with the latest. The OBO format files are very large (up to 400MB), thus download time will likely be lengthy.

Once the OBO files are written, GO-Elite will recognize that new files have been downloaded and rebuild the OBO summary files needed to run GO-Elite as well as Nested gene-GeneOntology associations. This process will take up-to 10 minutes, but is performed only once with each new Gene Ontology structure file download.

### **Updating Gene Ontology Relationships from EntrezGene**

This function imports Gene Ontology to EntrezGene associations for any supported species at NCBI. To run this function:

- a) From the main menu select "Update or Add Databases", "Create/Modify Databases" and then "EntrezGene-GO Associations".



- b) Select your species of interest (e.g., Bos Taurus). ). Selecting “all-supported” will update all species in your database. Otherwise, if this is a new species, select “New Species” and refer to the previous section “Add Custom Species Support”.
- c) Keep or change the default save options. These are:
  - a. Download the latest NCBI Entrez-GO relationships. If this check-box is selected, GO-Elite will make sure to get the latest annotation files, even if you have previously downloaded and extracted annotations from EntrezGene. If not checked, GO-Elite will use the previously downloaded annotations.
  - b. Over-write previous relationships OR update previous relationships. The first option will replace the existing table entirely with the new information while the second option will add to the existing annotations.
  - c. Delete installation files when finished. This option will delete the large downloaded database file from NCBI containing all species EntrezGene-GO associations, when the check box is selected. Since these files are annotated frequently, it is recommended they are deleted after each run.
- d) Once these options are chosen, select “Continue”.

- e) GO-Elite will download the necessary annotations and add support new EntrezGene to Gene Ontology relationships for that species.

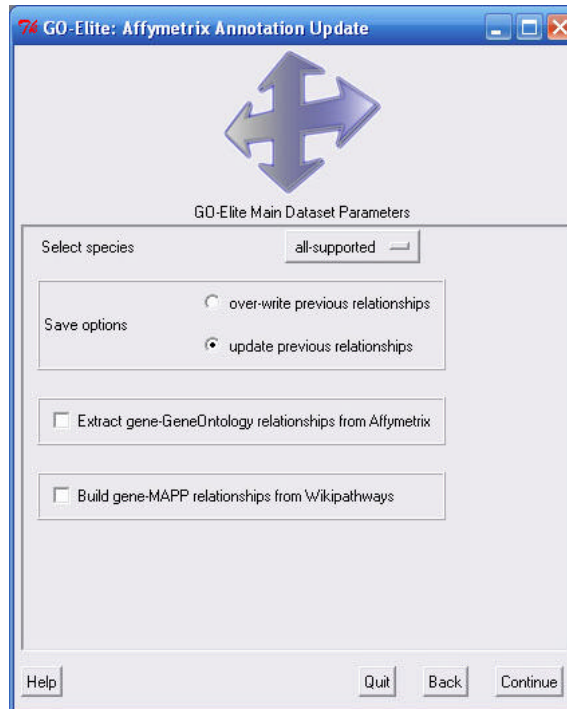
### **Updating Affymetrix Relationships**

In the GO-Elite databases with the suffix “Plus”, EntrezGene-to-Affymetrix, Ensembl-to-Affymetrix and EntrezGene basic annotations are extracted from Affymetrix provided annotation files for all available species and added to the database. However, a user may need to extract and integrate additional Affymetrix annotations on their own if:

- 1) The user wants to add archival Affymetrix annotations to an existing database
- 2) The user wants to add newer Affymetrix annotations to an existing database
- 3) A user wants to build a new species database from Affymetrix annotations provided in a custom microarray, not currently supported by GO-Elite

Although the Affymetrix annotations do not provide direct links between EntrezGene or Ensembl and Gene Ontology, this function can also infer these relationships from the Affymetrix annotation file(s). However, this option is only recommended with GeneOntology relationships can not be directly gathered from either Ensembl or NCBI using the described GO-Elite methods. To run this function:

- a) From the main menu select “Update or Add Databases”, “Create/Modify Databases” and then “Affymetrix Annotation Files”.



- b) Select your species of interest (e.g., Bos Taurus). Selecting “all-supported” will update all species in your database. Otherwise, if this is a new species, select “New Species” and refer to the previous section “Add Custom Species Support”.
- c) Keep or change the default save options. These are:
  - a. Over-write previous relationships OR update previous relationships. The first option will replace the existing table entirely with the new information while the second option will add to the existing annotations.
  - b. Extract gene-GeneOntology information from Affymetrix. By default, this variable is set to “no”. Selecting “yes” will retrieve Ensembl and EntrezGene to Gene Ontology relationships and add these to the database. This option is only necessary when neither EntrezGene nor Ensembl to Gene Ontology relationships are already in the existing gene database.
  - c. Build gene-MAPP Associations from Wikipathways . By default, this variable is set to “no”. Selecting “yes” will retrieve WikiPathway gene relationships from the internet to include in the gene-MAPP relationship files. If relationships between gene systems need to be inferred, the Affymetrix array annotations will be used to assist in this process (see following section).

- d) Once these options are chosen, select “Continue”.

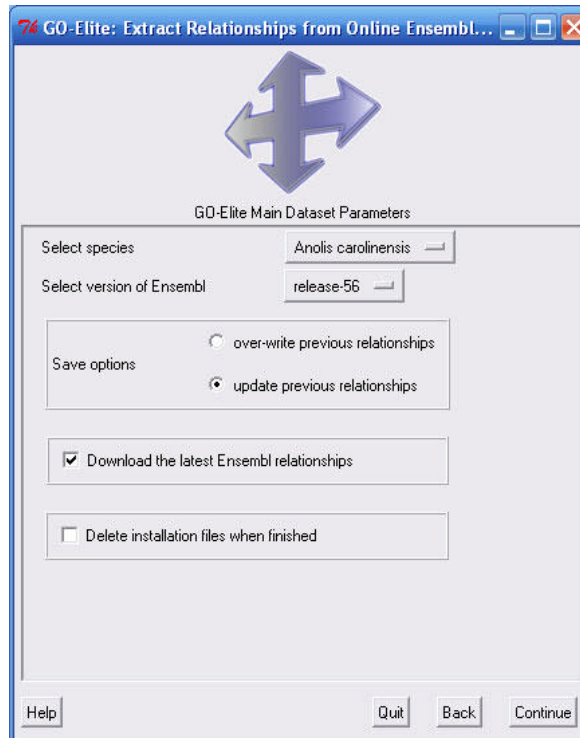
### **Rebuilding Species Ensembl Databases (Advanced)**

Existing or new GO-Elite databases can be built on the fly using the Ensembl update menu. ***This function should only be required if.***

- 1) You require immediate access to the most recent version of Ensembl, prior to release by GO-Elite developers
- 2) You are an open-source developer building your own custom versions of GO-Elite databases

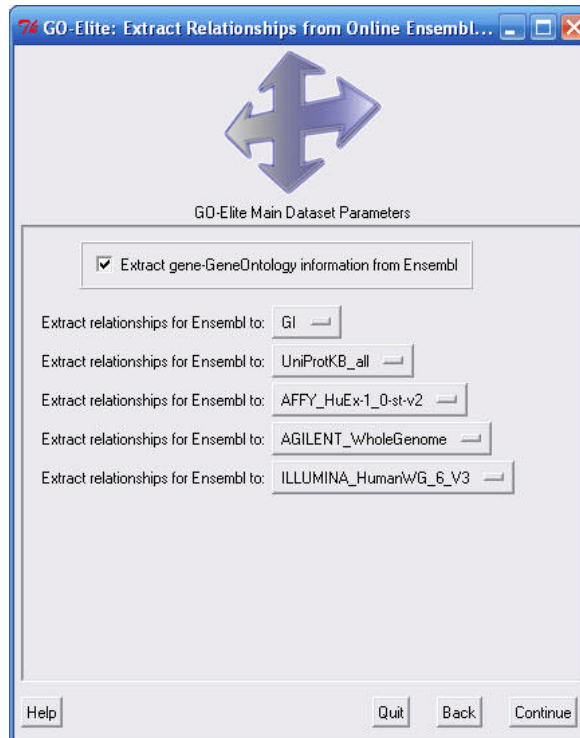
The Ensembl update function imports most critical data available from Ensembl for all supported species. This function downloads Ensembl SQL relationship files from the Ensembl website (<http://www.ensembl.org>). This includes any gene annotations in the Ensembl database for that species, including all annotations for all supported microarrays. Thus, depending on the species, the combined downloads can be very large (>1GB). To run this function:

- a) From the main menu select “Update or Add Databases”, “Create/Modify Databases” and then “Ensembl Annotations”. This option requires a web connection (may take 10 seconds to look up supported species).



- b) Select your species of interest (e.g., *Bos Taurus*).
- c) Options:
  - a. Over-write previous relationships OR update previous relationships. The first option will replace the existing table entirely with the new information while the second option will add to the existing annotations.
  - b. Download the latest Ensembl relationships. If this check-box is selected, GO-Elite will make sure to get the latest annotation files, even if you have previously downloaded and extracted annotations from Ensembl (these are stored locally in the GO-Elite program directory under “BuildDBs”. If not checked, GO-Elite will use the previously downloaded annotations.
  - c. Delete installation files when finished. This option indicates whether to delete the downloaded Ensembl SQL files when finished. This is most typically desired, since these annotations are frequently updated and the downloaded files can be very large for mammalian model organisms.
- d) Once these options are chosen, select “Continue”.





- e) Select up-to five different annotation resources that you want to include in the GO-Elite database. The first option, “Extract gene-GeneOntology information from Ensembl”, retrieves Gene Ontology annotations for all of the latest Ensembl genes. The five dropdown menus contain a listing of all available annotation resources. Some examples are RefSeq, UniProt, Affymetrix, EntrezGene, UCSC, Unigene and GI. For example, if the data you are analyzing are Unigene IDs, by selecting to include Unigene to Ensembl, GO-Elite will add these relationships to the database along with the system code “Ug”. When the user then analyzes Unigene data, they will then add the System Code “Ug” to the second column of their input and denominator files for all IDs. To see the currently included Gene ID systems, go to the “Analyze Gene Lists” menu under “GO-Elite Supported System Codes” option.
- f) Select “Continue” and GO-Elite will proceed to download the major components of the Ensembl database, for the current build for the selected species only. This process can take up-to two-hours and two hundred megabytes in downloaded files for species with diverse annotations. However, for most organisms, this process will be completed in under 10 minutes.

## **System Codes**

While the user can add new gene systems to the GO-Elite database and add their own unique system code, by default, GO-Elite supports a variety of gene systems and existing system codes. The majority of these gene systems are extracted from Ensembl during the Official GO-Elite gene database build process. These system codes are typically identical to those used by the programs GenMAPP (<http://www.genmapp.org>) and PathVisio (<http://www.pathvisio.org>). For the current release of Ensembl (build 56), the following system codes are available among the different 64 support species. Add the appropriate system code to the second column of your input and denominator gene files (see section 2 of this document). The systems with the named "MOD" in the column "MOD\_status" indicate that this gene system links directly to GeneOntology or pathways (aka MAPPs) in the gene database. This system information can be modified through the "Update or Add Databases" menu. Note: The Ensembl database contains annotations for many microarray types. By default all non-genomic tiling array platforms in the Ensembl database that are not Affymetrix, Agilent, Codelink nor Illumina are assigned the system name MiscArray and the system code Ma.

<b>System</b>	<b>SystemCode</b>	<b>MOD_status</b>
Affymetrix	X	
Agilent	Ag	
BDGP_insitu_expr	Bd	
Cint	Cj	
CioInt	Cio	
CloneID	Clb	
Codelink	Co	
DEDb	De	
EMBL	Em	
Ensembl	En	MOD
ENST	Enst	
EntrezGene	L	MOD
EPD	Epd	
Fantom	Fa	
FlyBase	F	
FlyGrid	Fg	
GadFly	Gf	
Genoscope_pred_gene	Gen	

goslim_goa	gos	
GPCR	Gpcr	
HGNC	Hg	
Illumina	Il	
IMGT	Im	
IPI	Ip	
Kyotograil	Ky	
MEROPS	Merops	
MGI	M	
MIM	Mi	
miRBase	Mb	
MiscArray	Ma	
modCB	Mg	
Osford_FGU	Of	
OTT	Ot	
PDB	Pd	
ProteinID	Pi	
RefSeq	Q	
RFAM	Rf	
Sanger	Sh	
sharesCDS	Sc	
TakRub	Tak	
TetNig	Tet	
Tgut_symbol	Ts	
TransFac	Transfac	
UniGene	Ug	
UniProt	S	
Vega	Vg	
Vega_transcript	Vr	
WikiGene	Wg	
Wormbase	Wb	
Xenopus_Jamboree	Xj	
XenTro	Xen	
Zfin	Z	

## References

1. Ashburner M, *et al.* (2000) Gene Ontology: Tool for the unification of biology. *Nat. Genet.* 25(1):25–29.
2. Pico AR, *et al.* (2008) WikiPathways: pathway editing for the people. (Translated from eng) *PLoS Biol* 6(7):e184 (in eng).
3. Doniger SW, *et al.* (2003) MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* 4:R7–R7.12.
4. Salomonis N, *et al.* (2007) GenMAPP 2: new features and resources for pathway analysis. (Translated from eng) *BMC Bioinformatics* 8:217 (in eng).
5. Benjamini Y & Hochberg Y (1995) Controlling the False Discovery Rate—a new and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57:289-300.
6. van Iersel MP, *et al.* (2008) Presenting and exploring biological pathways with PathVisio. (Translated from eng) *BMC Bioinformatics* 9:399 (in eng).