Tutorial for the R package - “maGUI”

The “maGUI” package aims to provide the user with graphical user interface for analysing microarray data produced on various platforms such as Affymetrix, Agilent, Illumina, Nimblegen and so on. It allows the user to preprocess and analyze the microarray data automatically. It also allows the user to identify functional categories and pathways for differentially expressed genes. Further, the user can build a co-expression network of differentially expressed genes.

The present tutorial describes the version 2.1 of the maGUI package. It relies on gWidgets (John Verzani et al. (2014)) package for generating GUIs. In order to generate and export different types of plots, it uses functions from graphics (Becker et al. (1988)) package and grDevices respectively. For smooth processing during export and import of graphs and tables, tcltk interface is utilized. Further, it uses various packages during the analysis viz., RSQLite, GEOquery, GEOmetadb, RBGL, WGCNA, Rgraphviz, simpleaffy, beadarray, lumi, oligo, pdInfoBuilder, Gostats, globaltest, ssize, etc.

In order to install the he maGUI package, use the following command on the R console.

```
> install.packages("maGUI",dependencies=TRUE)
```

With this command, all the required functions and dependencies of the maGUI package will also get installed automatically. In order to use the package, the installed packages are loaded as usual with library or require functions

```
> library(maGUI)
```

Load the maGUI GUI using the function below

```
> maGUI:::maGUI()
```

Microarray data can be imported from File → Load. Files with CEL extensions will be used to load Affymetrix data. Raw files with foreground mean signal and background median signal values will be used Agilent-one color data while loading of Agilent-two color data require raw files with Agilent source. Nimblegen requires raw files with .xys or .pair extensions while non-normalized files are used to load both Illumina beadarray and lumi data. Series matrix file along-with platform file is required to load Series Matrix data. GSE number is required to load On-line data. Once the data is loaded successfully, it prompts for automated analysis of microarray data. Selecting "OK" results in normalization, QC check, filtering with 2 fold over-expression in at least 50 percent of the arrays or filtering with standard deviation of at least 2, differential gene expression, PCA, clustering of samples and classification of loaded microarray data.

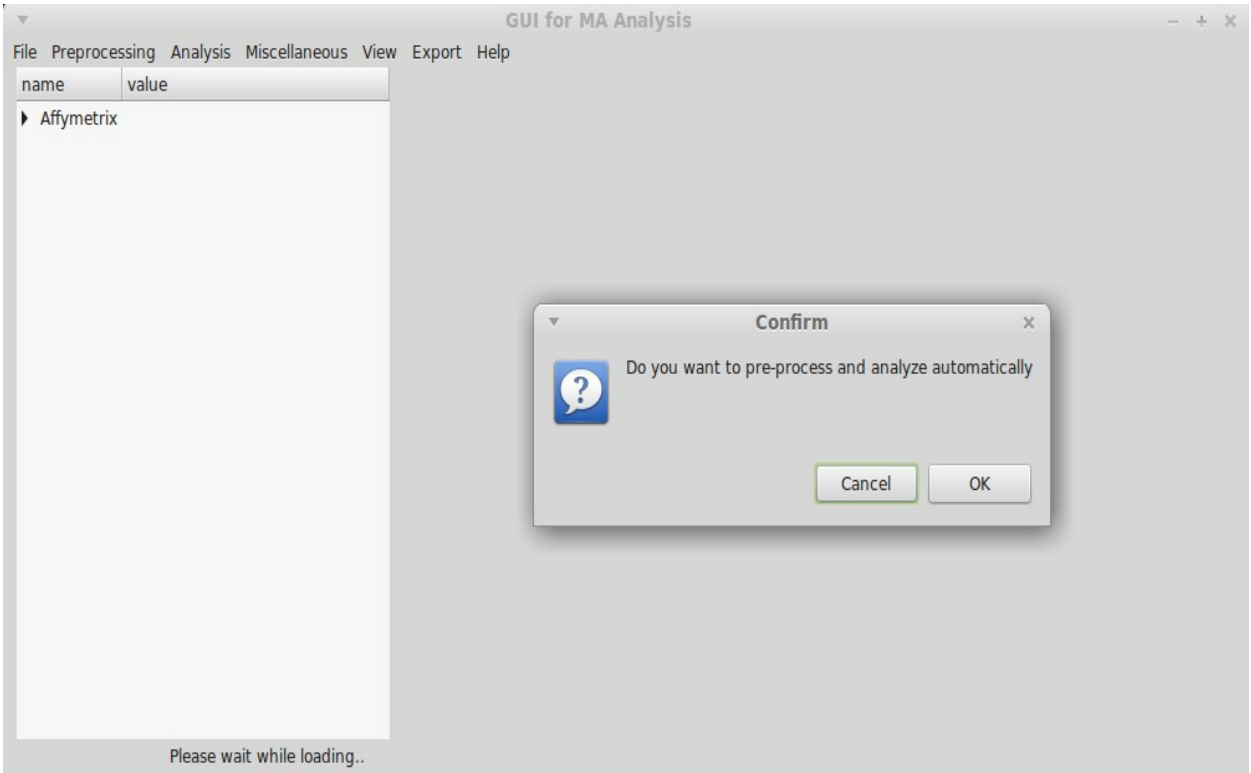

Figure 1. Initiation of GUI for automatic analysis of loaded Affymetrix data.

The loaded microarray data can be normalized from Preprocessing → Normalization. Quality assessment of the normalized data can be made from Preprocessing → Quality_Control through qc plot, boxplot or scatterplot. The following figure is a qc plot of normalized microarray data for the experiment number GSE68613.

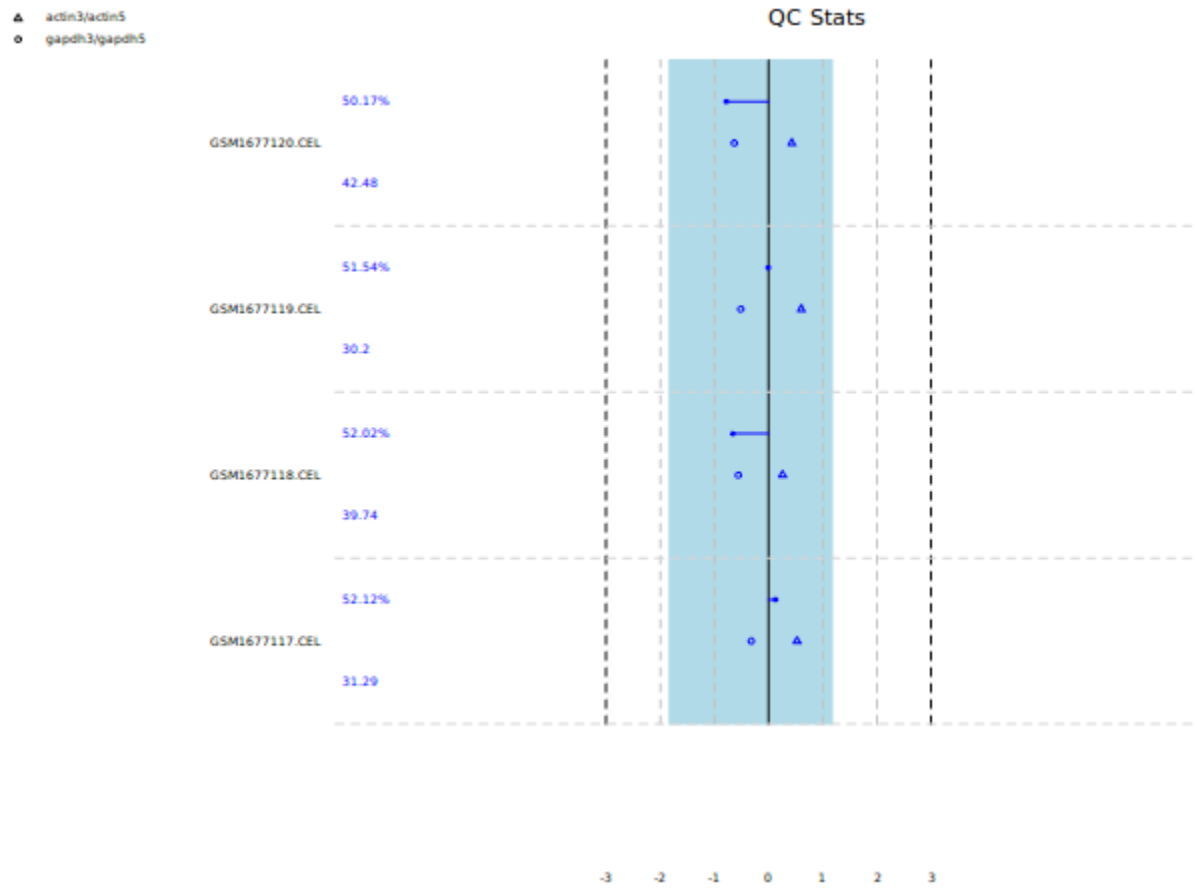

Figure 2. QC plot of loaded Affymetrix data.

Normalized data can be used to plot principle component analysis (PCA) with singular value decomposition method from Analysis → Principal_Component_Analysis_Unsupervized. Samples in microarray experiment can also be clustered using normalized data with pearson correlation coefficient and complete linkage methods from Analysis → Clustering_and_Visualization_Unsupervized. The following figures represent PCA and clustering of samples in the microarray experiment.

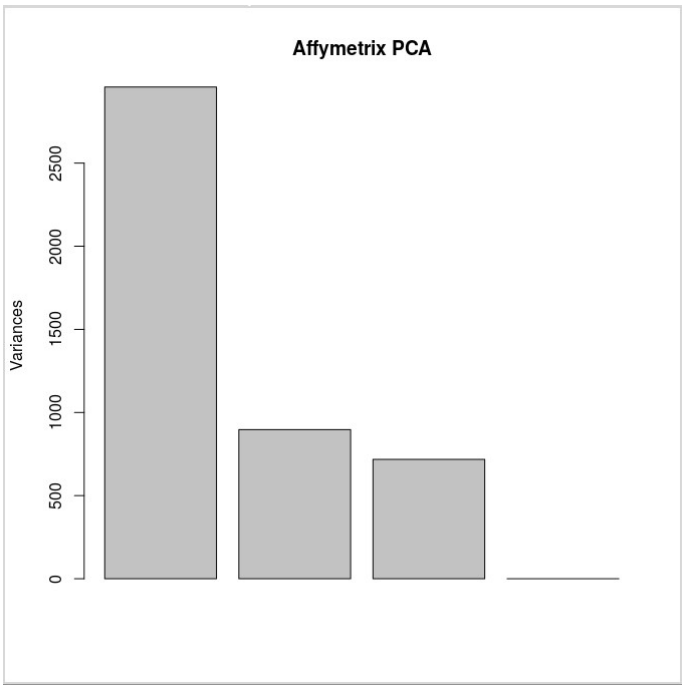

Figure 3. PCA

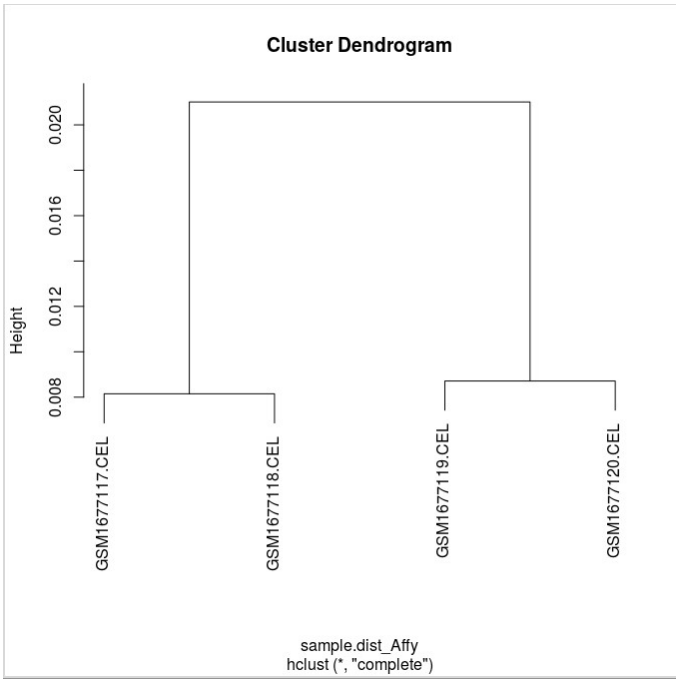

Figure 4. Clustering of Samples

Normalized microarray data can be filtered to exclude uninteresting genes or genes with very low expression levels from Analysis → Filtering_and_Statistical_Analysis → UnSpecific. Genes can also be filtered out using group knowledge from Analysis → Filtering_and_Statistical_Analysis → Specific. User can specify the names of control samples and test samples. The user can also add extra group of control and test samples using “Add” button as shown below.

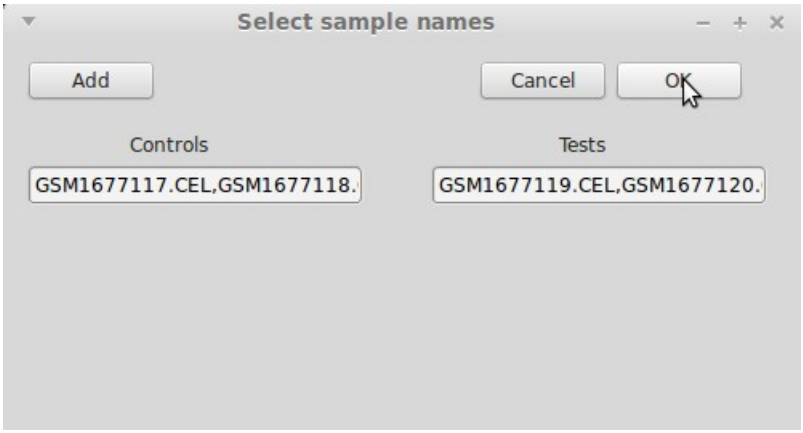

Figure 5. Providing Control and Test sample names of two different groups.

Using specific or unspecific filtering data, top differentially expressed genes can be obtained from Analysis → Differential_Gene_Expressions.

Figure 6. Parameters for Differentially Expressed Genes

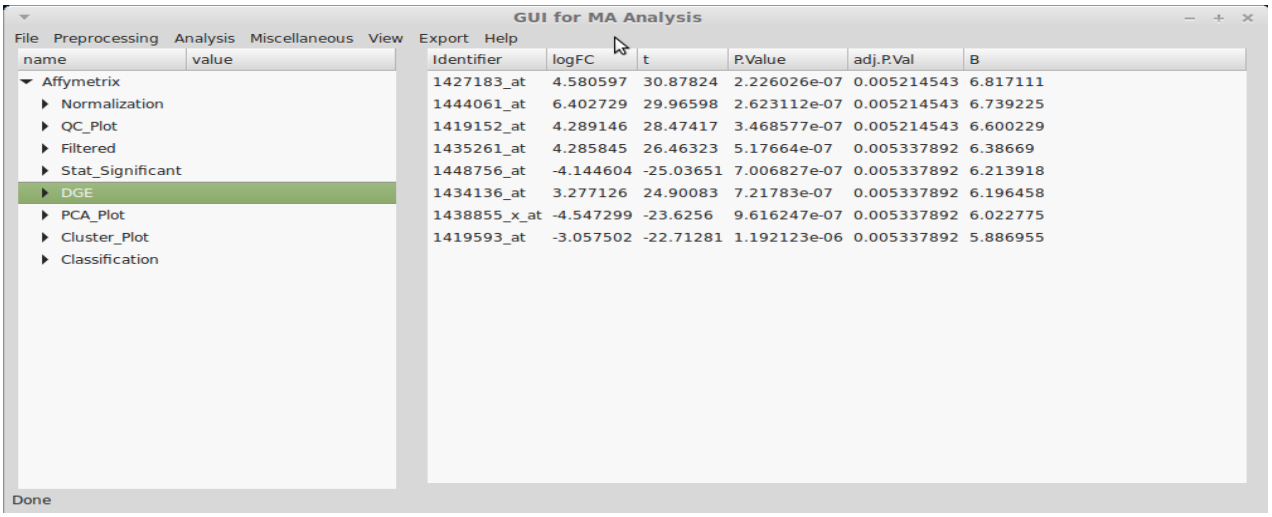

Figure 7. Table of DEGs predicted from Affymetrix data.

Classification of differentially expressed genes can be performed from Analysis → Classification_and_Visualization_Supervized.

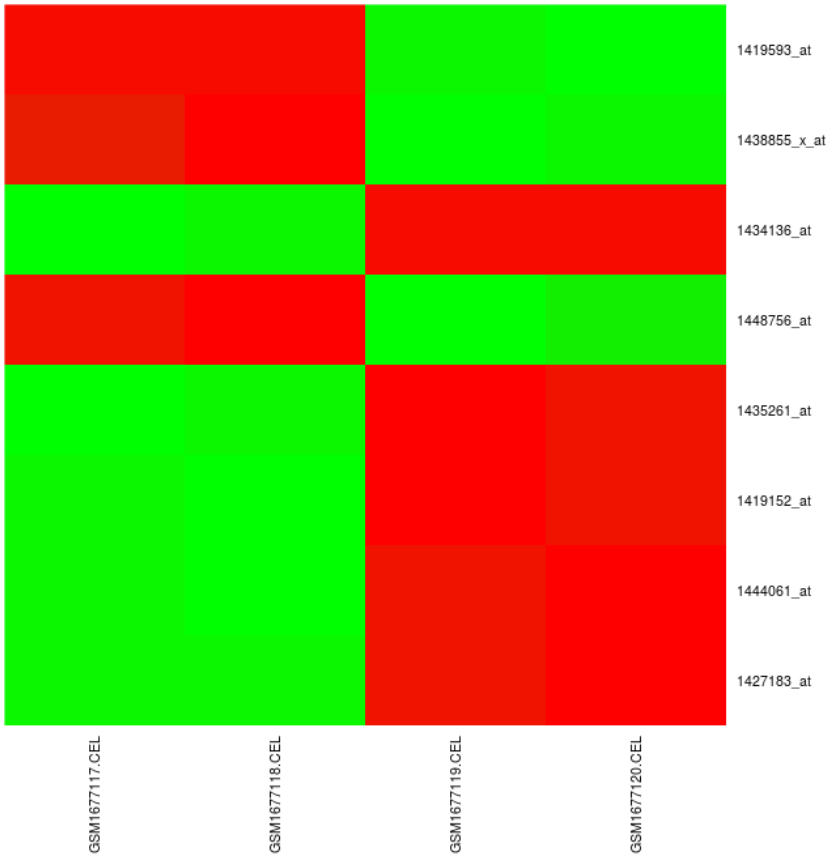

Figure 8. Classification able of DEGs.

It can be visualized as heatmap of expression profiles of differentially expressed genes where red color represents upregulation and green color represents downregulation.

Gene set enrichment analysis (GSEA) can be performed on statistically significant genes of microarray data from Miscellaneous menu. User can select GO categories such as Biological Process from Miscellaneous → Gene_Set_Enrichment_Analysis → GO_Biological_Process, Molecular Function from Miscellaneous → Gene_Set_Enrichment_Analysis → GO_Molecular_Function and Cellular Component from Miscellaneous → Gene_Set_Enrichment_Analysis → GO_Cellular_Component and set the p-value to get the corresponding GO terms for statistically significant genes. Alternatively, user can opt for KEGG pathways to get enrichment of the significant genes in some pathways from Miscellaneous → Gene_Set_Enrichment_Analysis → KEGG_Pathways. GSEA is achieved with hyperGTest function from GOstats package. Results are viewed and saved as tables. The following figure represents GO terms of GSEA Biological Process with p-value less than 0.001.

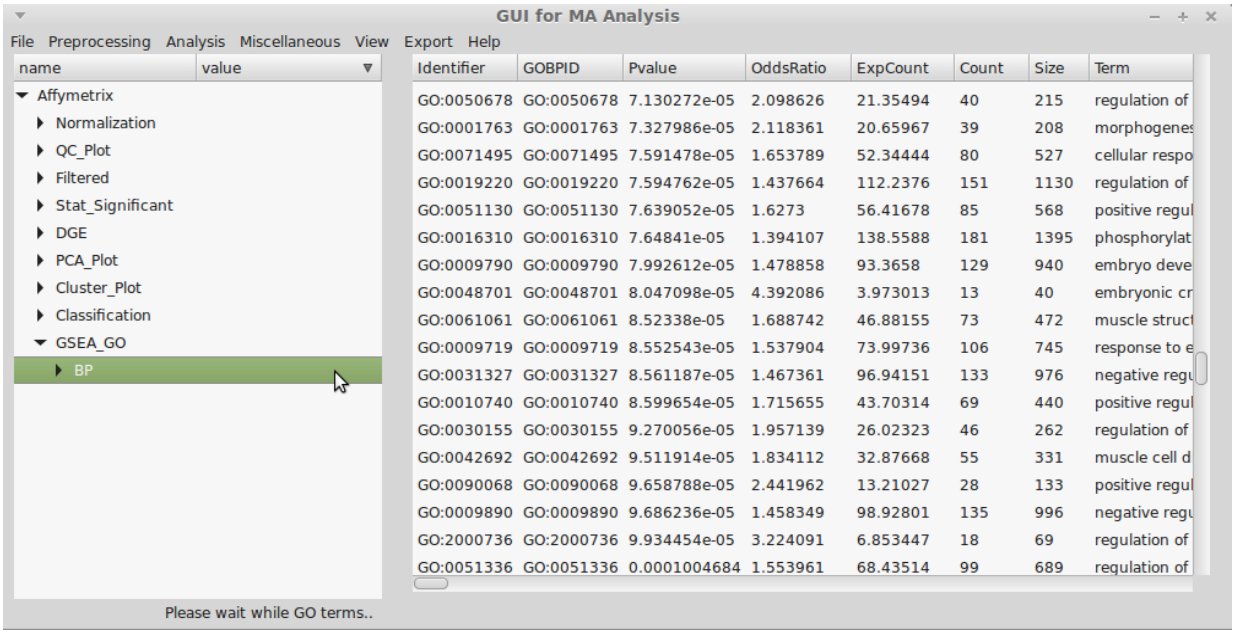

Figure 9. GSEA GO terms of Biological Process

Similar tables can be generated for Molecular Function and Cellular Component categories of GSEA. The following figure represents KEGG pathways with p-value less than 0.001.

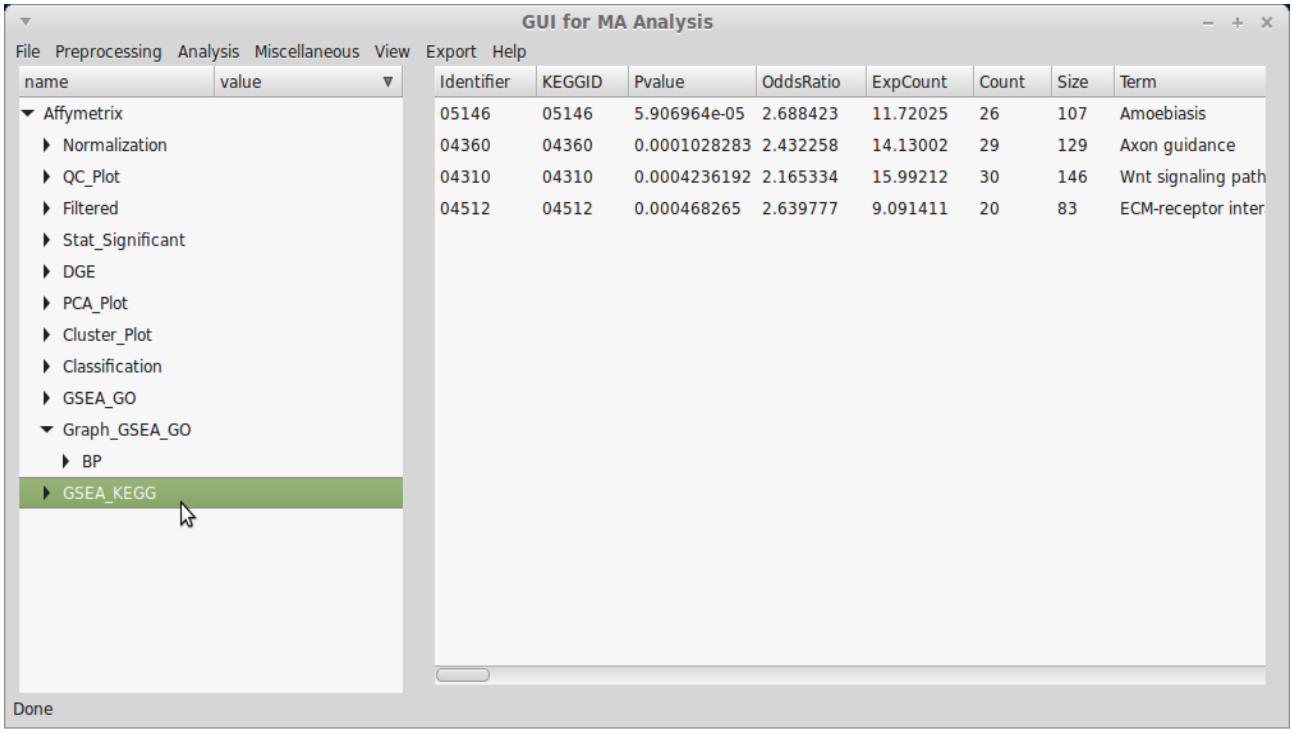

Figure 10. GSEA KEGG Pathways

Gene set test analysis (GSTA) is used to assign GO terms and KEGG pathways for all genes in the normalized data based on control and test samples groups. The following figure represents control and test sample names in the microarray experiment for GSTA. It is performed for Biological Process from Miscellaneous → Gene_Set_Test_Analysis → GO_Biological_Process, for Molecular Function from Miscellaneous → Gene_Set_Test_Analysis → GO_Molecular_Function and for Cellular Component from Miscellaneous → Gene_Set_Test_Analysis → GO_Cellular_Component. Alternatively, user can opt for KEGG pathways from Miscellaneous → Gene_Set_Test_Analysis → KEGG_Pathways.

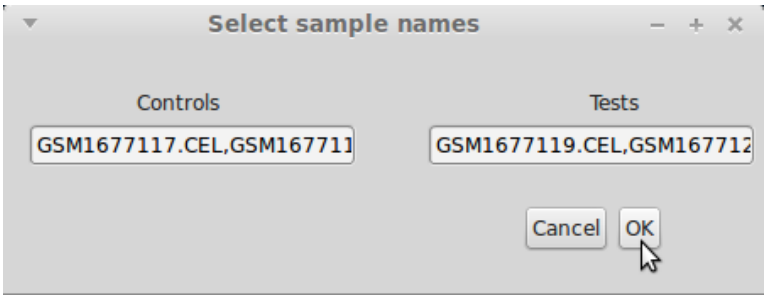

Figure 11. Control and Test sample names for GSTA

Enrichment of genes for GO terms can be visualized as graphs from Miscellaneous → Graphs → GSEA_GO_Biological_Process for Biological Process, from Miscellaneous → Graphs → GSEA_GO_Molecular_Function for Molecular Function and from Miscellaneous → Graphs → GSEA_GO_Cellular_Component for Cellular Component. User can set the cut off p-value to get the desired GO terms. The following figures are the graphs for GSEA GO terms and KEGG pathways and their legends with p-value less than 0.001. Yellow colored nodes represent the most significant ones while the white nodes are their parents.

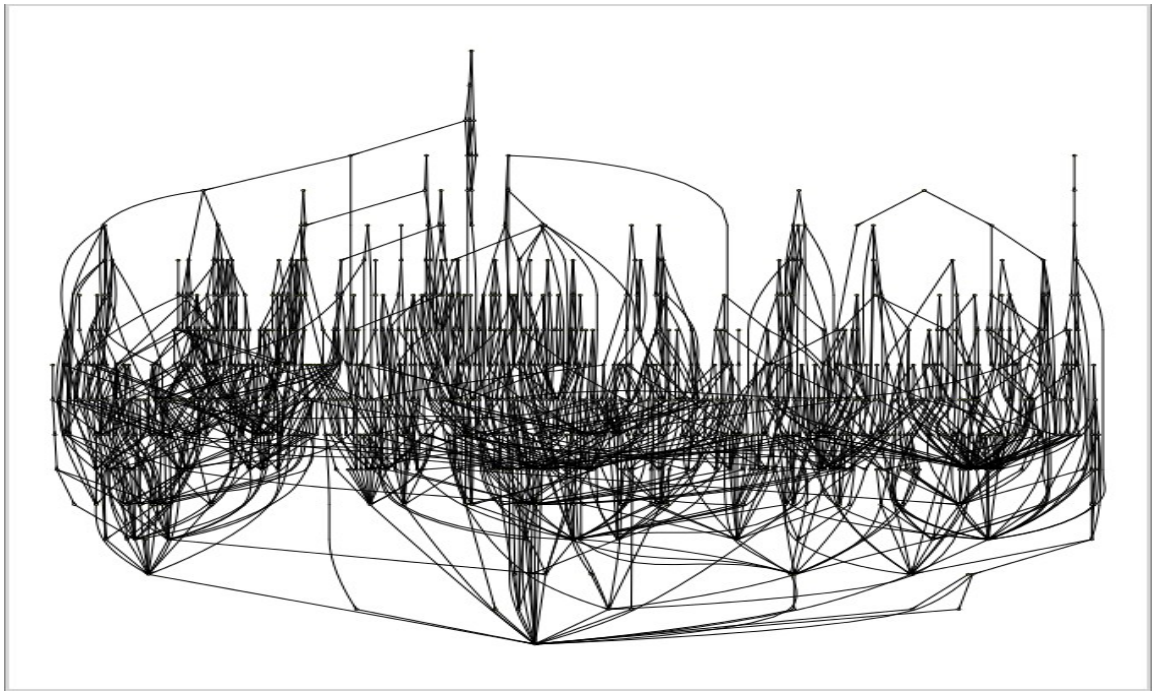

Figure 12. Graph of GSEA GO Biological Process

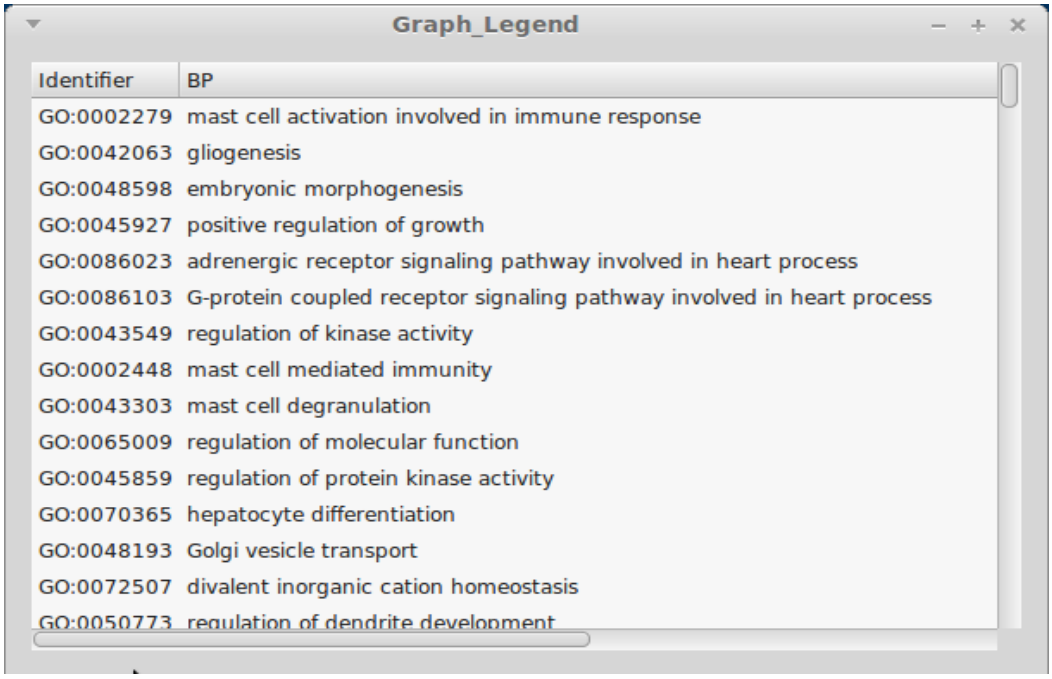

Figure 13. Legend for Graph of GSEA GO Biological Process

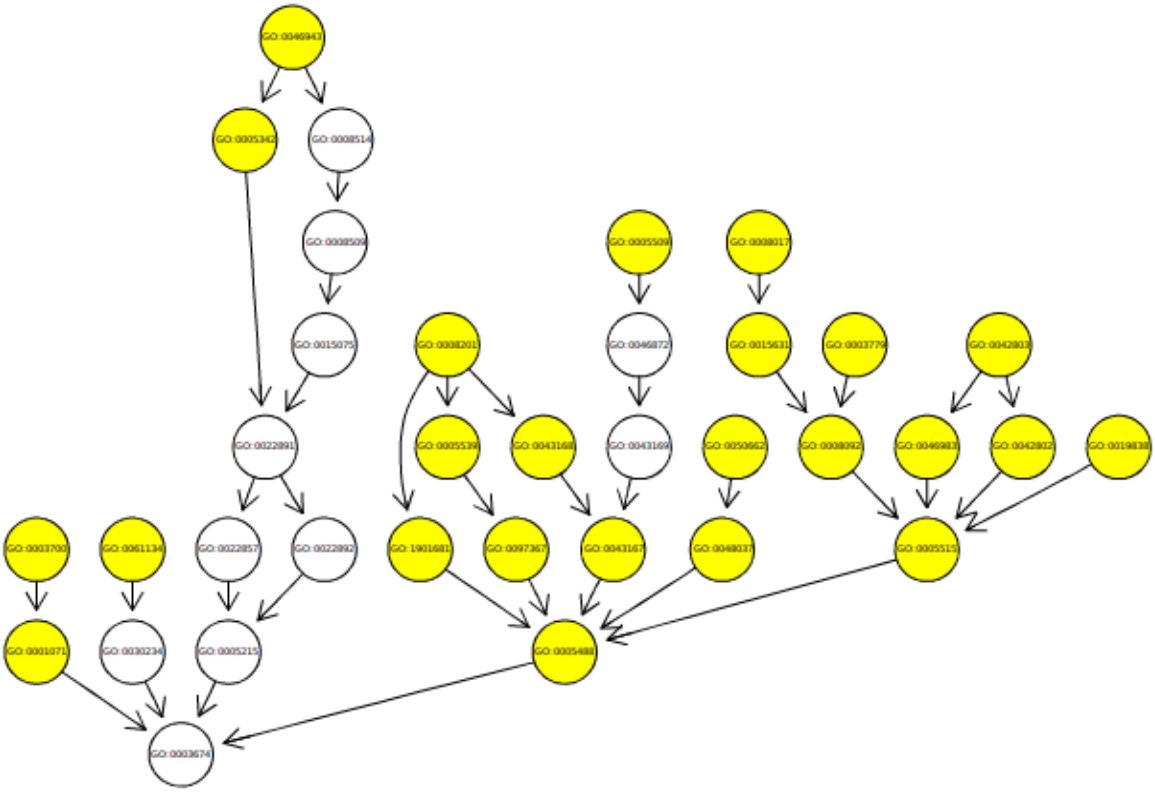

Figure 14. Graph of GSEA GO Molecular Function

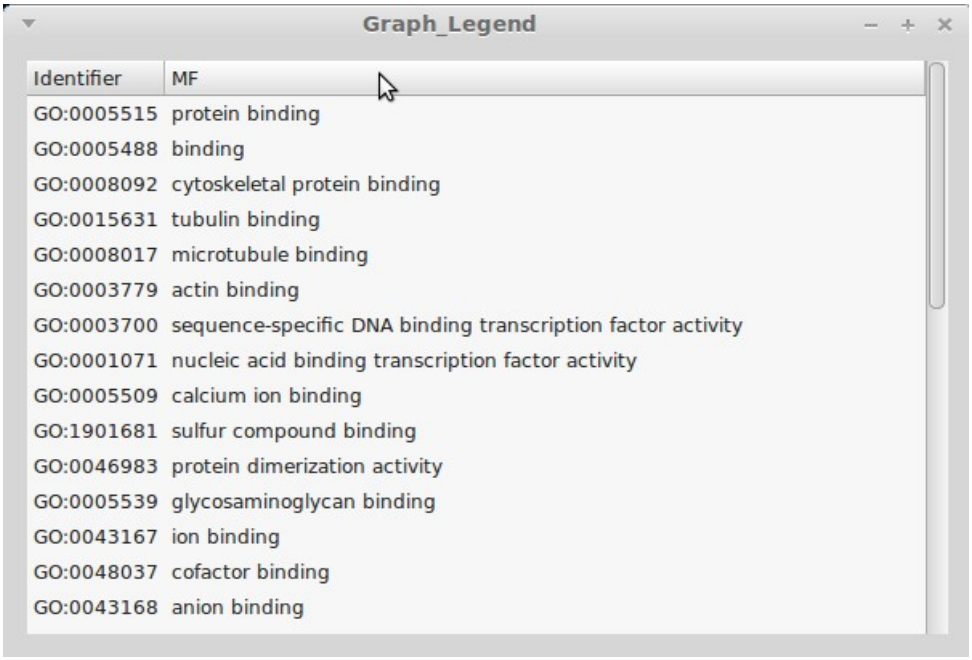

Figure 15. Legend for Graph of GSEA GO Molecular Function

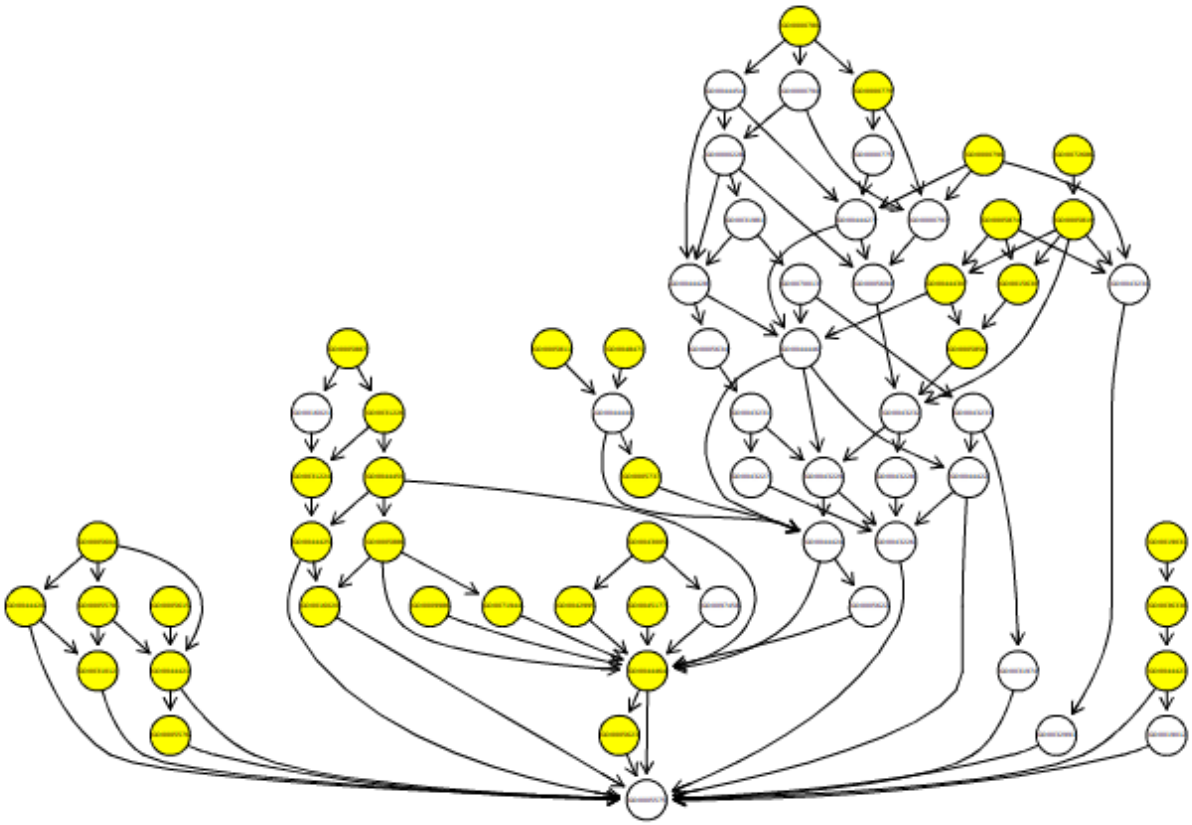

Figure 16. Graph of GSEA GO Cellular Component

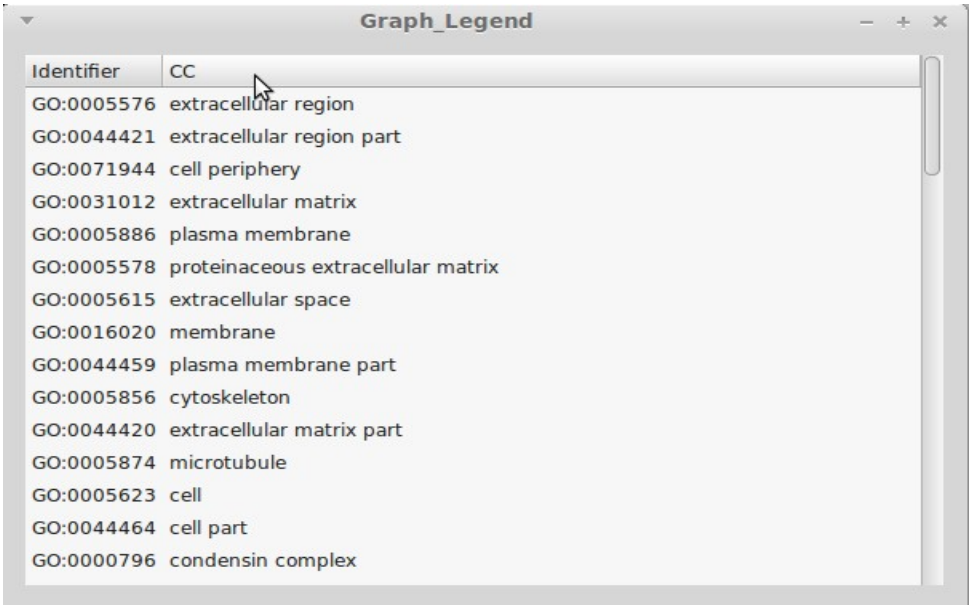

Figure 17. Legend for Graph of GSEA GO Cellular Component

Similar graphs can be generated for GSTA GO terms of Biological Process from Miscellaneous → Graphs → GSTA_GO_Biological_Process, Molecular Function from Miscellaneous → Graphs → GSTA_GO_Molecular_Function and Cellular Component from Miscellaneous → Graphs → GSTA_GO_Cellular_Component.

Visualization of KEGG Pathways

Visualization of KEGG pathways for significant genes can be achieved from Miscellaneous → Graphs → GSEA_KEGG_Pathways setting the p-value and selecting the KEGG ID. Nodes in the pathway of KEGG ID are mapped to differentially expressed genes color coding of which is in the range of Red and Green based on log fold change value. Red colored nodes in the graph are upregulated genes, while Green colored nodes represent downregulated genes. Dark grey colored nodes in the graph are the genes that are not differentially expressed while the white nodes represent genes absent in the microarray data. The following figures represent pathway of KEGG ID “04360” of GSEA with p-value less than 0.001 and its legend.

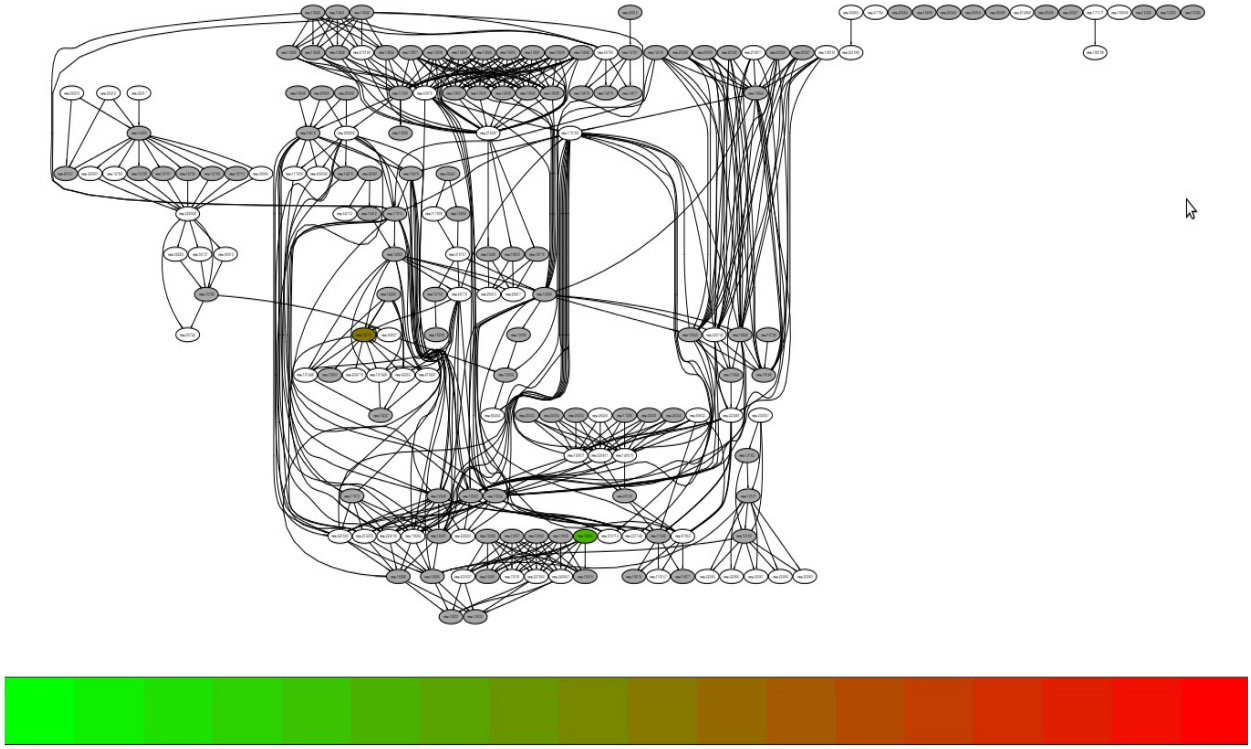

Figure 18. Graph of KEGG Pathway of ID “04310” of GSEA. Nodes mmu:13176 and mmu:19055 are downregulated.

Graph_Legend

Node	Node Color	logFC
mmu:93737	white	0
mmu:93742	white	0
mmu:14677	darkgrey	0
mmu:14678	darkgrey	0
mmu:14679	darkgrey	0
mmu:12767	darkgrey	0
mmu:26413	white	0
mmu:26417	white	0
mmu:170758	white	0
mmu:19353	darkgrey	0
mmu:19354	darkgrey	0
mmu:13176	#877800	-1.54499219505475
mmu:18019	darkgrey	0
mmu:18021	darkgrey	0
mmu:73181	white	0
mmu:19055	#4BB300	-3.12564068130478

Figure 19. Legend for Graph of GSEA KEGG Pathway of ID “04360”. Node mmu:13176 is downregulated with log fold change of 1.5 while node mmu:19055 is downregulated with log fold change of 3.1

Similar graphs of KEGG pathways can be generated for GSTA genes. This is achieved from Miscellaneous → Graphs → GSTA_KEGG_Pathways setting the p-value and selecting the KEGG ID.

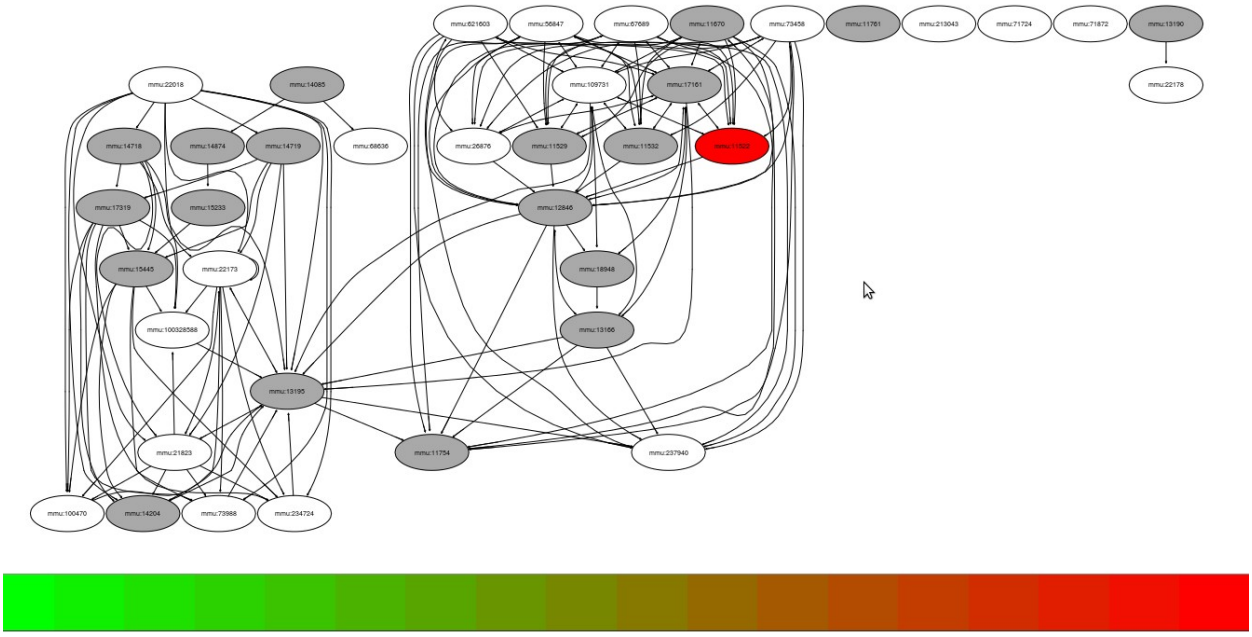

Figure 20. Graph of KEGG Pathway of ID “00350” of GSTA. Node mmu:11522 in the graph is upregulated

Graph_Legend

Node	Node Color	logFC
mmu:11761	darkgrey	0
mmu:213043	white	0
mmu:71724	white	0
mmu:71872	white	0
mmu:14085	darkgrey	0
mmu:13190	darkgrey	0
mmu:12846	darkgrey	0
mmu:11522	#FF0000	4.28584455866481
mmu:11529	darkgrey	0
mmu:11532	darkgrey	0
mmu:26876	white	0
mmu:18948	darkgrey	0
mmu:13166	darkgrey	0
mmu:21823	white	0
mmu:22178	white	0

Figure 21. Legend for Graph of GSEA KEGG Pathway of ID “00350”. Node mmu:11522 is upregulated with log fold change of 4.28

The nodes in the pathway of KEGG ID are mapped to differentially expressed genes color coding of which is in the range of Red and Green based on log fold change value. Red colored nodes are upregulated genes, Green colored nodes are downregulated genes, Dark grey colored nodes are the genes that are not differentially expressed while the white nodes are the genes absent in the microarray data as shown below for . The following figures represent pathway of KEGG ID “00350” of GSTA with p-value less than 0.001 and its legend.

All the identifiers in a microarray experiment can be mapped to corresponding gene symbol from Miscellaneous → Identifier_Symbol. It requires annotation package from Bioconductor resource. GUI uses the inbuilt table to map the microarray experiment to its corresponding annotation database available at Bioconductor, nonetheless user can also provide updated GEOmetadb database.

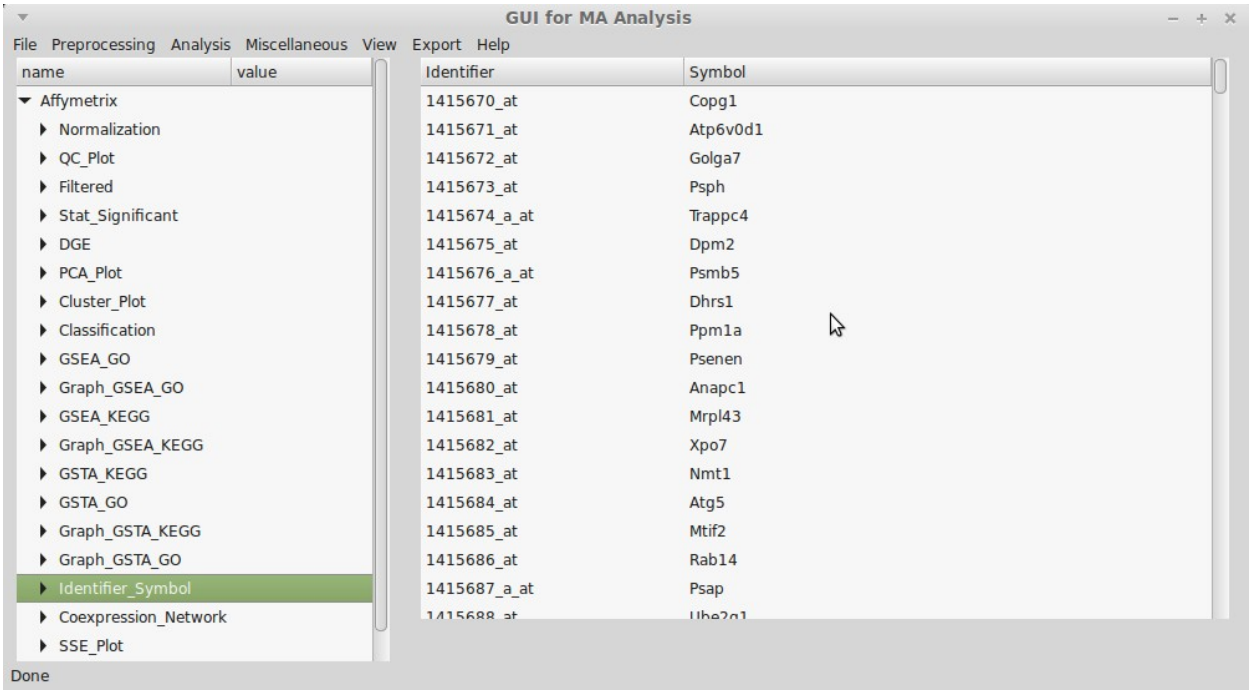

Figure 22. Identifiers of Affymetrix microarray data mapped to gene symbols**.**

Sample size estimation for microarray experiment can be achieved from Miscellanoues → Sample_Size_Estimation.

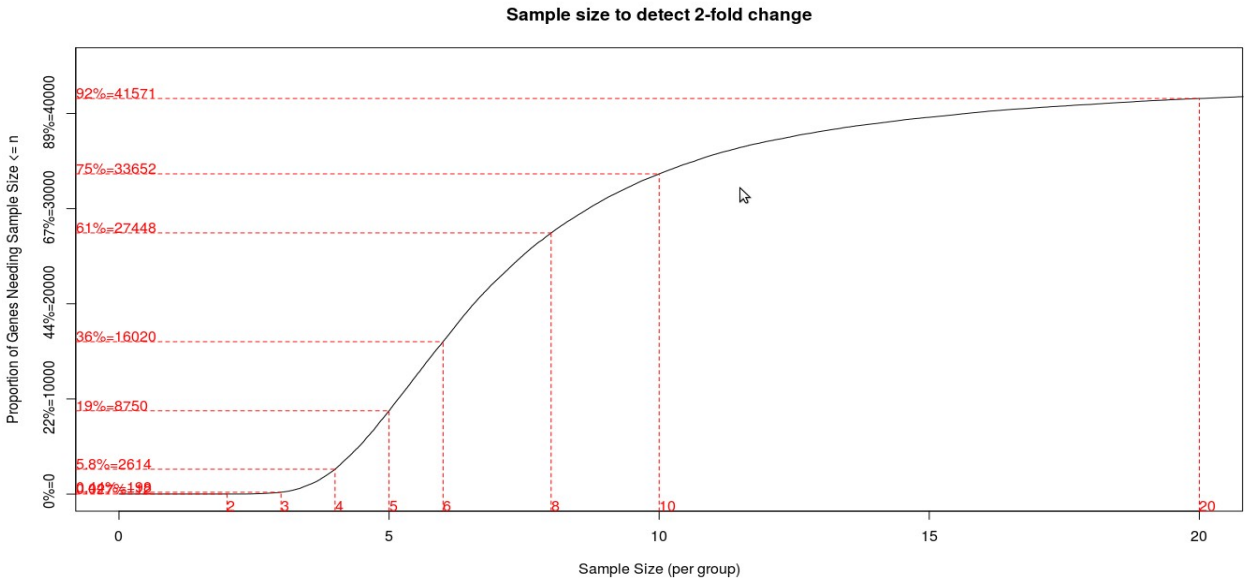

Figure 23. Graph of Sample size estimation

Co-expression network can be built for differentially expressed genes from Miscellaneous → Coexpression_Network. When the expression correlation between genes is 70 percent and above, it forms a link between the genes. A co-expression network is built using such links.

Figure 24. Co-expression Network of Differentially Expressed Genes

Protein-protein associations can be identified from Miscellanoues → PPIs_Prediction. It uses correlation of co-expression profiles among the genes of two different normalized data. When the correlation is 80 percent and above, two genes are considered to be associated.

Tasks performed on microarray data can be viewed from workspace of hierarchical nature in the left hand side of the GUI. Content can be viewed in the graphical area by double clicking any task in the workspace or selecting any task from View menu. The following figure represents all the tasks performed on Affymetrix microarray data. Similar tasks can be performed on microarray data of different experimental platforms such as Agilent of One and Two color types, Illumina of Beadarray and Lumi types and Nimblegen. Series matrix file along with platform soft file can also be used. Further Online procedure is also available.

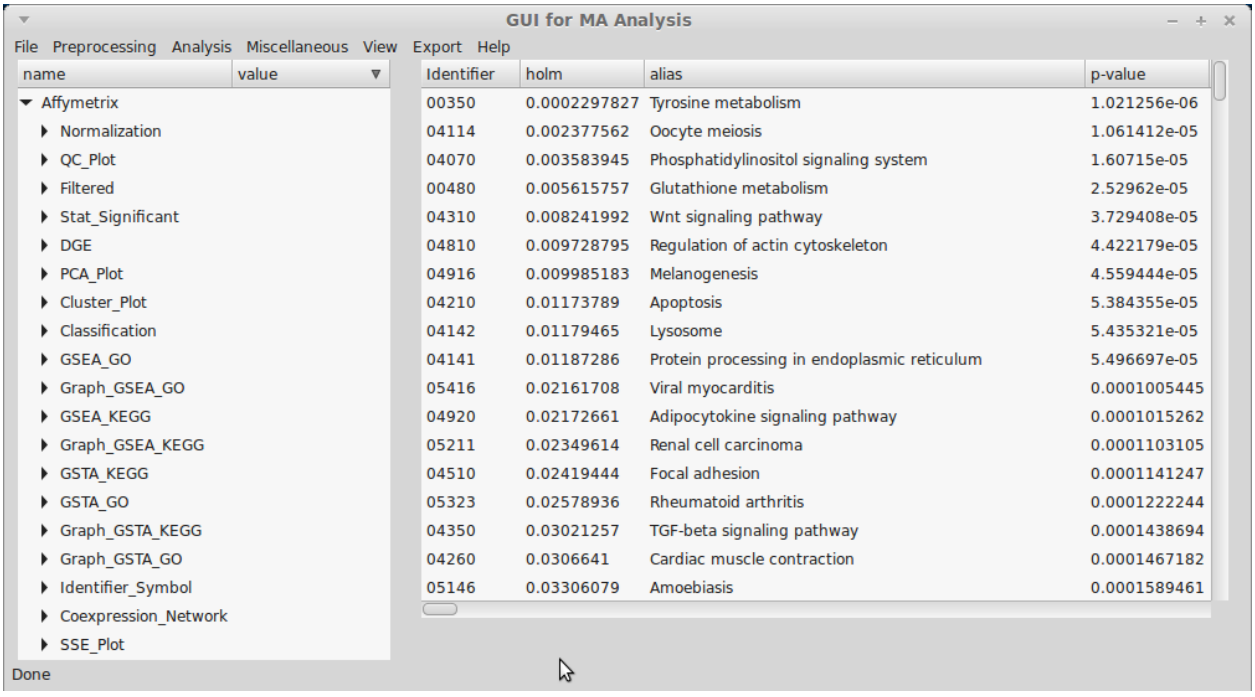

Figure 25. Tasks performed on Affymetrix microarray data

All the tables and figures generated during microarray data analysis and annotation can be exported from Export menu. User can save images directly from the graphical region with mouse Right click. User can also search any identifier in the tables by pressing Ctrl + F in the graphical region. “ls” function can be used to identify the objects created at R terminal in background. These objects can be utilized for any sort of analysis at R terminal. Further, they can be saved as Rdata file from File → Save. The Rdata file can be loaded at any R terminal with load function.