WebGestalt Manual

Introduction

WebGestalt is a "WEB-based GEne SeT AnaLysis Toolkit". WebGestalt incorporates information from different public resources and provides an easy way for biologists to make sense out of large sets of genes. It enables biologists to manipulate integrated information and find patterns that are not detectable otherwise. WebGestalt is designed for functional genomic, proteomic and large scale genetic studies from which high-throughput data are continuously produced. It currently works for human and mouse. WebGestalt is free for academic use after registration.

There are four key modules in WebGestalt:

- **Gene set management module** provides tools for uploading, saving, retrieving and deleting gene sets, as well as tools for the Boolean operation to generate union, intersection or difference between gene sets.
- **Information retrieval module** retrieves up to 20 attributes for all genes in a gene set.
- **Organization/visualization module** organizes and visualizes gene sets using eight sub-modules: *GOTree, Tissue Expression Bar Chart, Chromosomal Distribution Chart, KEGG Table and Maps, BioCarta Table and Maps, Protein Domain Table, PubMed Table* and *GRIF Table*.
- **Statistics module** automatically chooses appropriate statistical methods to suggest important biological areas that warrant further study.

This manual will guide you through the toolkit. The first part is the introduction to the user interface. The second part is the introduction to the gene set management module. The third part is the introduction to the information retrieval module. The forth part is the introduction to the organization/visualization module. The fifth part is the introduction to the statistics module. The last part illustrates two sample applications: 1) Boolean operation between gene sets to answer complex biological questions, and 2) Third party application in WebQTL.

I. User interface

All of the tools in WebGestalt can be accessed through a simple and intuitive user interface (Figure 1). The interface can be divided into five areas. Area A provides gene set management tools for uploading, retrieving, deleting and Boolean operation. Area B displays the name and description of the currently active gene set. Area C provides the gene set information retrieval tool, where the user can choose to retrieve information for up to 20 attributes. Area D provides the gene set organization and visualization tools that help the users to explore large gene sets. Area E displays a table for the genes in the currently active gene set, including the ID used in the input file, LocusLink ID, gene symbol and gene name. Each LocusLink ID is hyperlinked to a gene information record with detailed information retrieved from our local database GeneKeyDB.

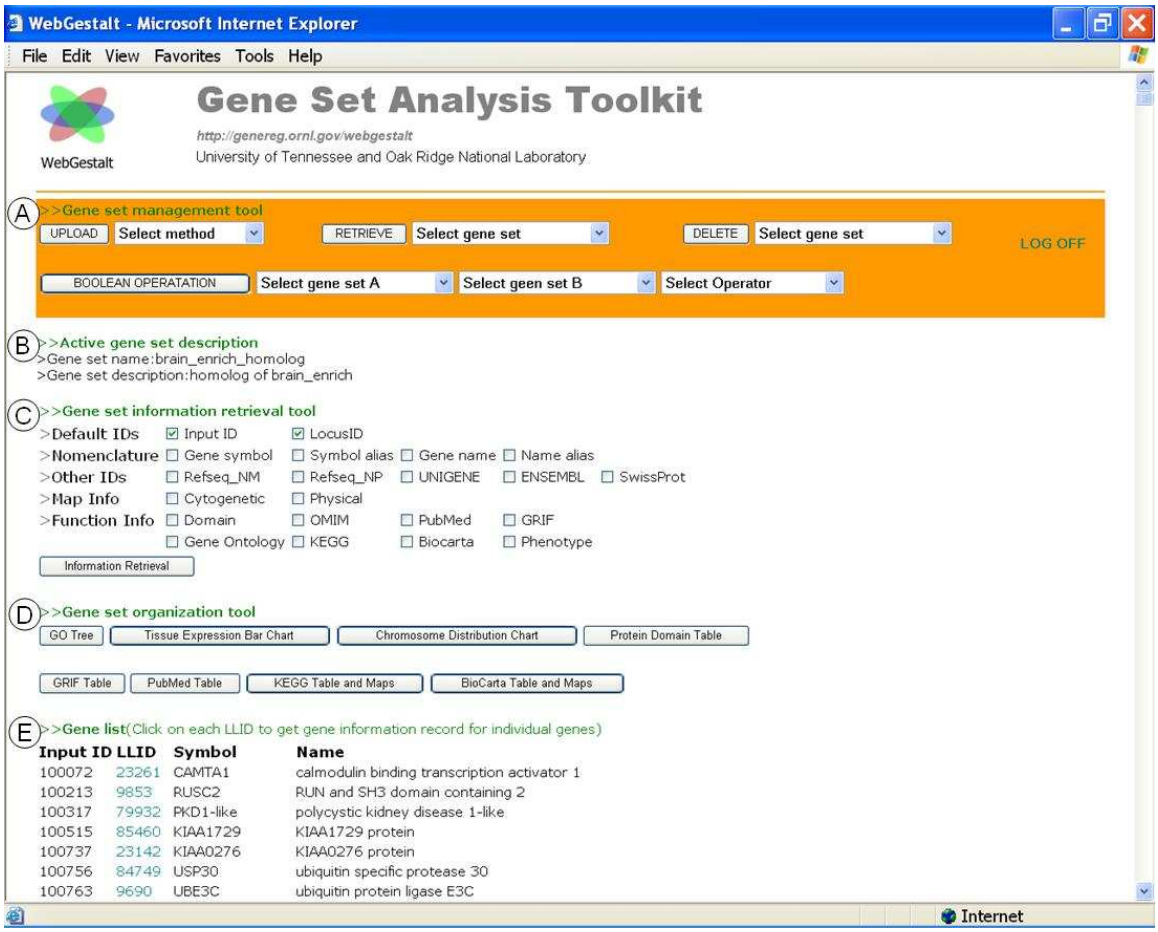


Figure 1 User interface of WebGestalt. The User interface of WebGestalt can be divided into five areas: A. Gene set management tools; B. Active gene set description; C. Gene set information retrieval tool; D. Gene set organization tools; E. Gene list for the active gene set.

II. Gene set management

2.1 Uploading

Webgestalt provides three options for the uploading of gene sets, which can be selected from the dropdown menu near the UPLOAD button (Figure 1A). The options are "*From file*", "*By location*", and "*By Gene Ontology*". Each option will give a slightly different uploading interface as shown in Figure 2. When the "From file" option is selected, gene identifiers (ID type) that can be recognized by the uploading tool are *LocusLink IDs, gene symbols, Affymetrix probe set IDs, Unigene IDs, Swiss-Prot IDs* and *Ensembl IDs*. WebGestalt only accepts txt files, please don't use excel files. The format of the file should be one ID per row. Sample ID lists can be found in the following table:

LocusLink ID	Gene symbol	Affy ID	Unigene ID	Swiss-prot ID	Ensembl ID
28989	ADRBK2	100128_at	Hs.284250	Q8TE60	ENSG00000101093
55830	ADRM1	100156_at	Hs.297304	Q8TE59	ENSG00000169252
56900	ADSL	100355_g_at	Hs.82933	O95450	ENSG00000100357
57409	ADSS	100631_r_at	Hs.325631	P59510	ENSG00000164025
57413	ADSSL1	100697_at	Hs.28847	O15072	ENSG00000164025
83935	ADTB1L1	100921_at	Hs.44004	O75173	ENSG00000164025
84230	ADTB1L2	100971_at	Hs.412836	Q9UKP2	ENSG00000164025

When uploading a gene set in WebGestalt, the uploading tool also automatically uploads a companion set from the human or mouse orthologs. The orthologs are defined by the HomoloGene from NCBI. We hope that this could help the comparative genomics studies. A unique analysis name is given for each gene set by the user and can be used to retrieve or delete the gene set in the future.

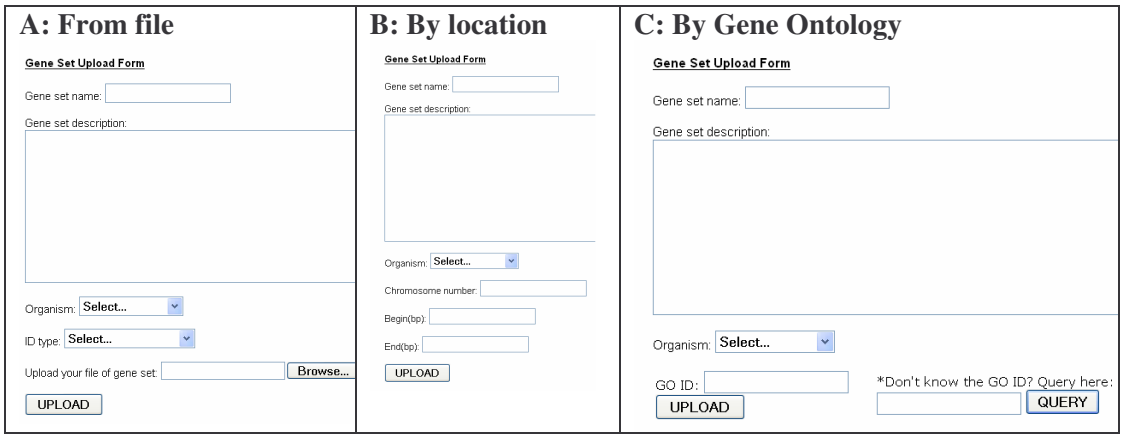


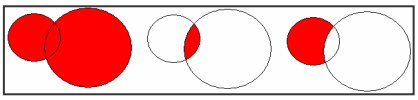
Figure 2. Uploading interface. The Organism dropdown menus in A, B and C have two options: Homo sapiens and Mus musculus. A) By file. ID type can be selected from LocusLink ID, gene symbol, Affymetrix probe set ID, Unigene ID, Swiss-Prot ID and Ensembl ID. B) By chromosome location. C) By GO category. GO ID can be queried for the interesting terms.

2.2 Retrieving and Deleting

Stored gene sets can be retrieved or deleted using the "**RETRIEVE**" or "**DELETE**" button. (Figure 1A).

2.3 Boolean operation

Boolean operations can be performed between two selected gene sets. You need to choose "Gene set A", "Gene set B" and an "Operator" (Figure 1A). The options of the operator are "*Union*"(A or B), "*Intersection*"(A and B), and "*Difference*"(A minus B).



Union **Intersection** **Difference**

III. Information retrieval

Existing information for all genes in a gene set can be retrieved by the "**Information Retrieval**" button (Figure 1C). The attributes that can be retrieved include *nomenclature information, identifiers reference into different databases, Map information* and *functional information.* Table 1 lists all of the 20 attributes, their sources, and associated websites. The retrieved information for all genes in a gene set can be downloaded as a tab-delimited file or opened directly in the web browser using Microsoft Excel (Figure 3).

Table 1. Gene and gene product attributes that can be retrieved by WebGestalt

Attribute		Source	Website
Nomenclature information	Gene symbol	LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
	Symbol alias	LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
	Gene name	LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
	Name alias	LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
IDs reference into different databases	LocusLink ID	LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
	Refseq_NM	Refseq	http://www.ncbi.nlm.nih.gov/RefSeq/
	Refseq_NP	Refseq	http://www.ncbi.nlm.nih.gov/RefSeq/
	UNIGENE ID	UNIGENE	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene
	ENSEMBL ID*	ENSEMBL	http://www.ensembl.org
	Swissprot ID	SwissProt	http://us.expasy.org/sprot/
Map information	Cytogenetic*	LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
	Physical*	UCSC	ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/
Functional information	Domain name	CDD	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=cdd
	OMIM ID	OMIM	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
	PubMed ID	PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
	GRIF record	LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
	GO term	GO	http://www.geneontology.org
	KEGG*	KEGG	http://www.genome.ad.jp/kegg
	BioCarta*	BioCarta	http://www.biocarta.com/
	Phenotype	LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/

ENSEMBL ID, ENSEMBL gene stable ID; Cytogenetic, Cytogenetic map location; Physical, Physical map location; KEGG, KEGG pathway name; BioCarta, BioCarta pathway name

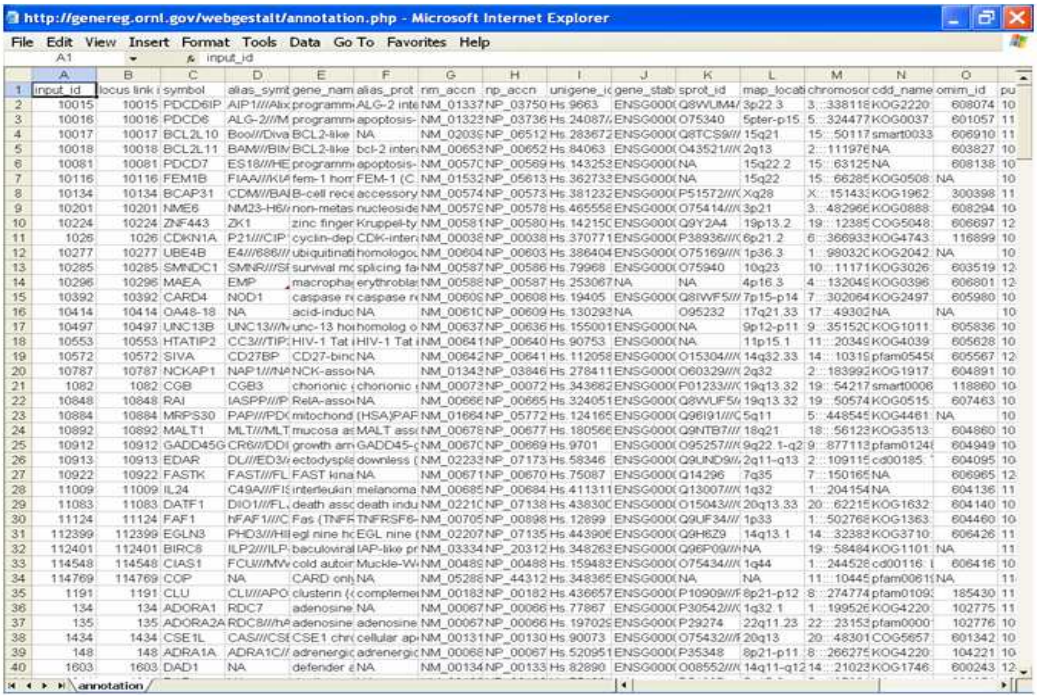


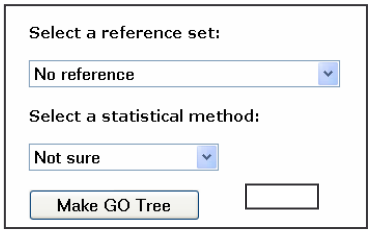
Figure 3. Retrieved information for all genes in a gene set can be opened directly in the web browser using Microsoft Excel.

IV. Organization/visualization

The organization/visualization module in WebGestalt is intend to assist biologists in exploring large gene sets by organizing and visualizing the genes based on various common features using eight sub-modules: *GOTree, Tissue Expression Bar Chart, Chromosomal Distribution Chart, KEGG Table and Maps, BioCarta Table and Maps, Protein Domain Table, PubMed Table* and *GRIF Table.*

4.1 GO Tree

The GO Tree module can be accessed from the "**GO Tree**" button (Figure 1D). A new window will be opened, in which you can add a reference gene set for the GO category enrichment analysis (right).



If you just want to organize the active gene set by GO, you can press the "**Make GO Tree**" button directly. However, if you want to do statistical analysis to identify GO categories with enriched gene numbers, you need to select a reference gene set from the list. The list includes some prestored gene sets in WebGestalt, including *human genome, mouse genome* and *all affy arrays* (prefixed with "WEBGESTALT_"), as well as *all of your uploaded gene sets*. WebGestalt provides two statistical methods, *hypergeometric test* and *Fisher's exact test*. If you know which method is better for your analysis, you can choose it from the dropdown menu. If you are not sure which method is better for you, WebGestalt will try to choose an appropriate one for you based on the reference gene set you select (see the statistics section for more details).

Figure 4 shows the GO Tree window. The main part in the window is an exapandable tree, in which each node is a GO category. Each GO category is followed by three parameters: O (Observed gene number in the category); E (Expected gene number in the category) and R (Ratio of enrichment for the category). The fourth parameter P (p value calculated from the statistical test) is given for the categories with R>1 to indicate the significance of enrichment. Categories with P<0.01 are colored red. The "**GO Term Search**" and "**Key Word Search**" buttons can be used to search for interesting GO categories.

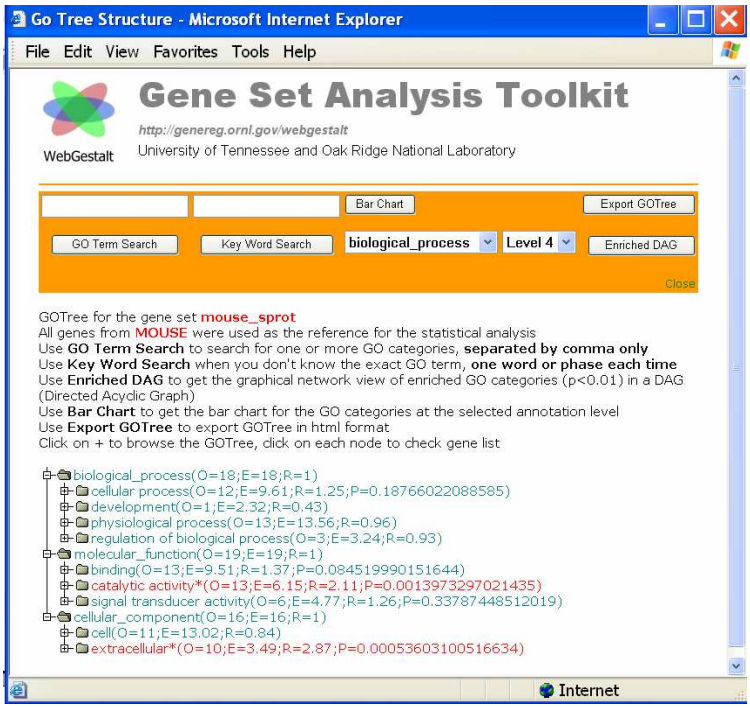
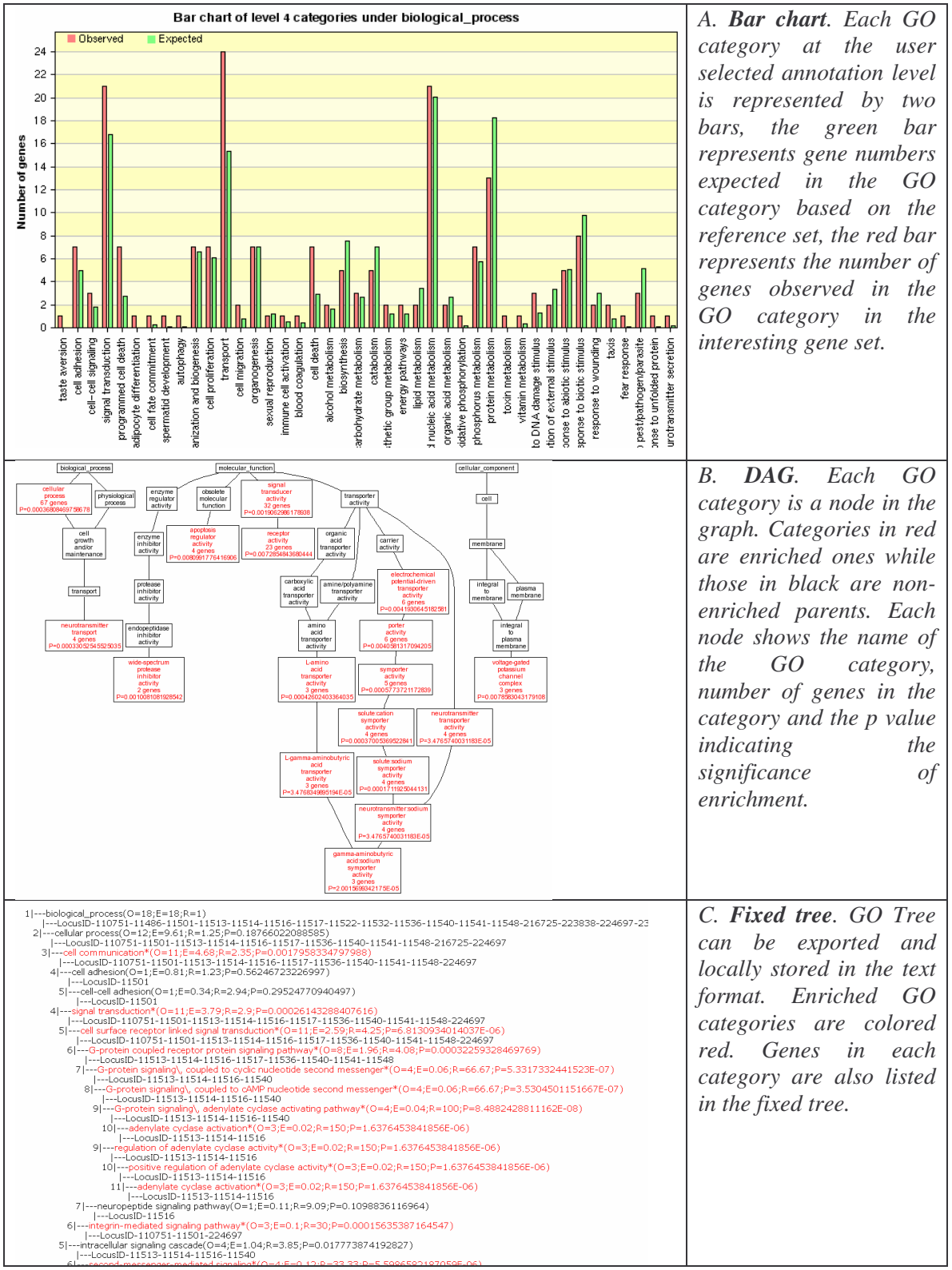


Figure 4. The GO Tree window

Other visualizations include: 1) The bar chart (Figure 5A), which can be generated by clicking the "**Bar Chart**" button (Figure 4); 2) The enriched DAG (Directed Acyclic Graph, Figure 5B),

which can be generated by clicking on the "**Enriched DAG**" button (Figure 4), and 3) The fixed tree (Figure 5C), which can be generated by clicking on the "**Export GO Tree**" button (Figure 4).



*A. **Bar chart**. Each GO category at the user selected annotation level is represented by two bars, the green bar represents gene numbers expected in the GO category based on the reference set, the red bar represents the number of genes observed in the GO category in the interesting gene set.*

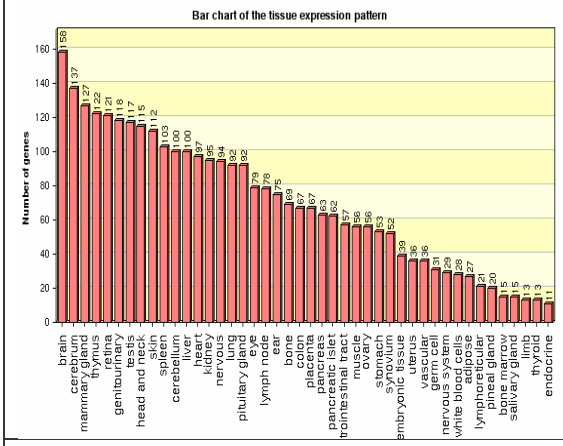
*B. **DAG**. Each GO category is a node in the graph. Categories in red are enriched ones while those in black are non-enriched parents. Each node shows the name of the GO category, number of genes in the category and the p value indicating the significance of enrichment.*

*C. **Fixed tree**. GO Tree can be exported and locally stored in the text format. Enriched GO categories are colored red. Genes in each category are also listed in the fixed tree.*

Figure 5. Different visualizations generated by the GO Tree module

4.2 Tissue Expression Bar Chart

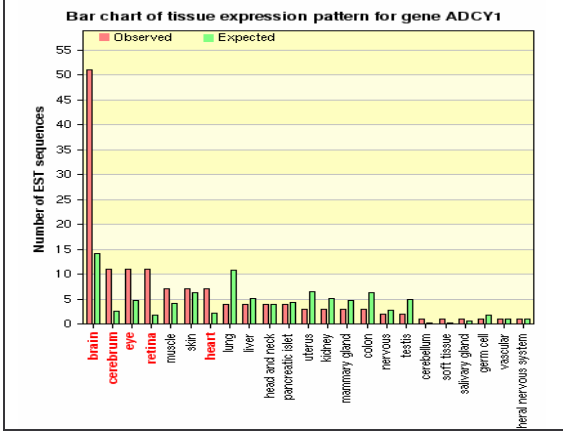
Tissue expression bar chart can be generated through the "**Tissue Expression Bar Chart**" button (Figure 1D). The Tissue Expression Bar Chart is based on the data from the CGAP (Cancer Genome Anatomy Project) expressed sequence tag (EST) project, in which the cDNA libraries were derived from a wild variety of human or mouse tissue types. As of September 2004 when this manual was written, CGAP had archived 9408 cDNA libraries, which had been prepared from more than 50 different tissue types in human and mouse. These libraries include more than 7,637,483 ESTs representing 21,207 unique human genes and 21,502 unique mouse genes. The broad coverage of both tissue types and genes makes CGAP an excellent resource for the tissue expression pattern study. WebGestalt can organize and visualize genes in a gene set based on the tissue expression pattern using the CGAP data (Figure 6).



*A. **Tissue expression bar chart for a gene set**. Each tissue is represented by a bar in the chart. The height of the bars represents the number of genes that are expressed in the tissue. Clicking on individual tissue will generate a table (6B) with a sub-set of genes expressed in the tissue*

LLID	Symbol	Name	Over/Under representation
117	ADCYAP1R1	adenylate cyclase activating polypeptide 1 (pituitary) receptor type I	1.386e-36
8038	ADAM12	a disintegrin and metalloproteinase domain 12 (meltrin alpha)	-1.092e-18
2	A2M	alpha-2-macroglobulin	-3.575e-18
107	ADCY1	adenylate cyclase 1 (brain)	8.158e-17
8745	ADAM23	a disintegrin and metalloproteinase domain 23	4.671e-16
134	ADORA1	adenosine A1 receptor	5.683e-12
123	ADFP	adipose differentiation-	-1.646e-09

*B. **Sub-set of genes expressed in a specific tissue**. The over/under-representation of individual genes in the tissue is evaluated by hypergeometric test. The p value is provided in the last column of the table. Clicking on individual p value will create a tissue expression bar chart for the corresponding gene (6C).*



*C. **Tissue expression bar chart for a gene.** Each tissue is represented by two bars. The height of the red bars represents the observed number of EST sequences for the selected gene in the tissue. The height of the green bars represents the expected number of EST sequences (Expected number of EST sequences for a specific gene in a specific tissue = Total number of EST sequences for the gene in all tissues* Total number of EST sequences for all genes in the tissue/Total number of EST sequences for all genes in all tissues).*

Figure 6. Tissue expression bar chart

4.3 Chromosome Distribution Chart

The Chromosome Distribution Chart can be generated through the "**Chromosome Distribution Chart**" button (Figure 1D). The chromosome location information comes from the UCSC genome annotation databases (ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/). In the Chromosome Distribution Chart (Figure 7), each chromosome is represented by a bar in the chart. Each gene is represented by a red cross symbol and located on the chromosome based on its location. Clustered genes from a gene set can be easily visualized in the chart. By clicking on individual chromosome, the sub-set of genes located on the chromosome will be generated. The sub-set of genes can be saved as a new gene set.

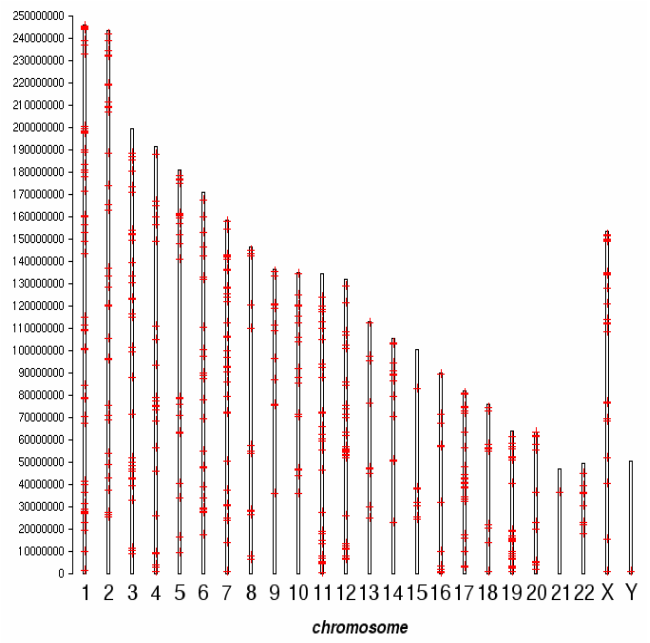


Figure 7. Chromosome Distribution Chart. Each chromosome is represented by a bar in the chart. Each gene is represented by a red cross symbol and located on the chromosome based on its location. By clicking on individual chromosome, the sub-set of genes located on the chromosome will be generated.

4.4 KEGG and BioCarta Tables and Maps

One of the most important tasks in the high-throughput experiments is to identify the pathways that are involved in the biological process studied. WebGestalt can organize genes based on the KEGG biochemical pathways in a KEGG Table (Figure 8A). The KEGG Table shows KEGG pathways associated with the gene set, the number of genes in each pathway and the LocusLink IDs for the genes. Each pathway name in the KEGG table is hyperlinked to the KEGG Map, in which genes in the gene set are highlighted in red (Figure 8B). WebGestalt can also organize genes based on another popular pathway database, BioCarta into a BioCarta Table (Figure 9A). The BioCarta Table has the same structure as the KEGG Table. Each pathway name in the BioCarta Table is hyperlinked to the BioCarta Map (Figure 9B).

A. KEGG Table

KEGG map	Gene number	Locus link IDs
Phosphatidylinositol Signaling System	10	1607 1608 1609 5291 5294 5330 5777 8525 8527 9162
Glycerolipid Metabolism	6	1607 1608 1609 8525 8527 9162
Inositol Phosphate Metabolism	2	5291 5330
Purine Metabolism	1	5168
Nicotinate and Nicotinamide Metabolism	1	5168
Pantothenate and CoA Biosynthesis	1	5168
Riboflavin Metabolism	1	5168
Starch and Sucrose Metabolism	1	5168

B. KEGG Map

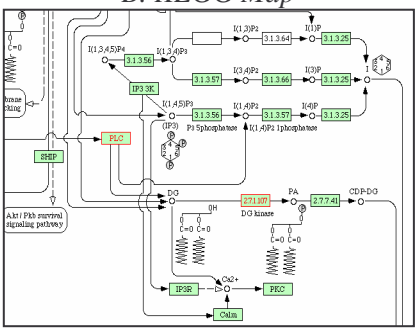


Figure 8. KEGG TABLE and MAP

A. BioCarta Table

BioCarta map	Gene number	Locus link IDs
Signaling of Hepatocyte Growth Factor Receptor	4	109905 16412 218397 26419
Control of Gene Expression by Vitamin D Receptor	4	12848 18519 21974 22385
HIV-I Nef: negative effector of Fas and TNF	4	11783 12675 218397

B. BioCarta Map

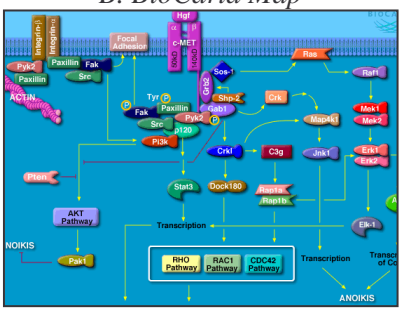


Figure 9. BioCarta TABLE and MAP

4.5 Domain Table

The Domain Table in WebGestalt organizes the genes based on the PFAM protein domains (Figure 10). The Domain Table shows the name of the PFAM domains associated with the gene set, the number of genes having each domain and the LocusLink IDs for the genes. Each domain name is linked to the Conserved Domain Database of the NCBI, where domain function, structure and sequence are available.

Domain name	Gene number	Locus link IDs
Ubiquitin carboxyl-terminal hydrolase	4	170822 22284 230484 59025
Ion transport protein	4	16438 16508 18764 20265
RecF/RecN/SMC N terminal	3	13006 14211

Figure 10. Domain Table

4.6 PubMed Table and GRIF Table

WebGestalt can organize genes according to their co-occurrence in publications, based on the gene-publication association information retrieved from in the LocusLink database. LocusLink provides two types of gene-publication indices, one is computed from the PubMed, and the other is GRIF. WebGestalt organizes genes based on both indices and generate PubMed Table (Figure 11) or GRIF Table (Figure 12) respectively. The PubMed Table shows PubMed IDs for the publications associated with the gene set, the number of genes in each publication and the LocusLink IDs for the genes. Each PubMed ID is hyperlinked to corresponding PubMed record, where the abstract for the paper is available. The GRIF Table is similar to the PubMed Table, except for one additional column showing the GRIF comments.

PUBMED ID	Gene number	Locus link IDs
10349636	6	106795 14866 1
11042159	6	106795 14866 1
11076861	6	106795 14866 1
11217851	6	106795 14866 1
12477932	5	106795 14866 5
12466851	2	53313 67397
12119294	1	53313
8607970	1	14866

Figure 11. PubMed Table

PUBMED ID	GRIF comment	Gene number	Locus link IDs
12145348	Brain somatostatin receptors 1,2,4 and 5 are up-regulated in somatostatin-deficient mice, and SSTR3 is down-regulated.	5	20604 20606 20607 20608 20609
11850187	Developmental expression profiles of Celsr (Flamingo) genes in the mouse	3	107934 12614 53883
11850418	mediates factor Xa signaling in endothelial cells	3	14062 14063

Figure 12. GRIF Table

V. Statistics

Statistical analysis is necessary for biological discovery from large gene sets. Statistical tests that have been used for identifying enriched GO categories include the Fisher's exact test, the χ2 test, the T test, the binomial test and the hypergeometric test. Although being implemented in several existing tools, the T test, the χ2 test and the binomial test require certain distributions that are usually violated. Both the Fisher's exact test and the hypergeometric test require a great number of combinatorial calculations and are disfavored in some existing tools. However, the speed of these tests can be greatly increased by using pre-calculated factorials. When analyzing the functional significance of the interesting gene sets produced by microarray experiments, all of the genes on the microarray, which represent the population from which the interesting genes are drew, should be used as the reference. This becomes a sampling without replacement problem and can be appropriately modeled by the hypergeometric distribution. Suppose that we have n genes in the interesting gene set (A) and N genes in the reference gene set (B). Suppose further that there are k genes in A and K genes in B are in a given category (C). If B represents the population from which the genes in A are drew, WebGestalt uses the hypergeometric test to evaluate the significance of enrichment for category C in gene set A, $P = \sum_{i=k}^{n} \frac{\binom{N-K}{n-i}\binom{K}{i}}{\binom{N}{n}}$.

Alternatively, the user may want to compare two independent gene sets for functional significance, for example, genes that are up-regulated at two different time points in a microarray study. In this case, A and B are two independent gene sets, neither of the gene sets can represent the population, thus the hypergeometric test is inappropriate, WebGestalt uses the Fisher's exact test instead, $P = \sum_{i=k}^{n} \frac{\binom{n}{i}\binom{N}{K}}{\binom{N+n}{K+i}}$. Some of the existing tools provide several statistical tests and allow the users to select any of the implemented tests. This may be dangerous for biologists who do not have enough statistics background. Appropriate statistical test should be chose based on different purpose, as inappropriate statistical test may be misleading. In our opinion, the hypergeometric test is the right choice for microarray data analysis when all the genes on the microarray were used as the reference, while the Fisher's exact test is appropriate for comparing two independent gene sets. To ensure correct biological discoveries, WebGestalt analyzes the input gene sets and recommends the right statistical test for the users who are not able to make a decision by themselves.

The hypergeometric test is also used for the evaluation of the over/under-representation of individual genes in a selected tissue. Suppose that we have n EST sequences for a selected gene in all tissues and N EST sequences for all genes in all tissues. Suppose further that there are k EST sequences for the selected gene in a selected tissue and K EST sequences for all genes in the tissue. If $k > n*K/N$, we consider that the gene is over-represented in the tissue, and the P value indicating the significance of over-representation is calculated by this formula:

$$P = \sum_{i=k}^{n} \frac{\binom{N-K}{n-i}\binom{K}{i}}{\binom{N}{n}}$$. If $k < n*K/N$, we consider that the gene is under-represented in the tissue, and the P value indicating the significance of under-representation is calculated by this formula:

$$P = \sum_{i=0}^{k} \frac{\binom{N-K}{n-i}\binom{K}{i}}{\binom{N}{n}}.$$

VI. Sample applications

6.1 Boolean operation between gene sets

One unique feature of WebGestalt is the Boolean Operation between the gene sets. It will help to answer simple questions such as: "show me all genes identified through experiment A or experiment B" (union), "show me the genes that are consistently up-regulated in two microarray experiments" (intersection) or "show me the genes that are expressed in brain but not skin" (difference). Recursively applying the Boolean operation makes it possible to combine information from any number of gene sets. Putting the organization module and the gene set management module together, Figure 13 shows that WebGestalt is able to answer complex questions such as "give me all genes in my gene set that are expressed in the brain or cerebellum, located on chromosome 5 and involved in signal transduction".

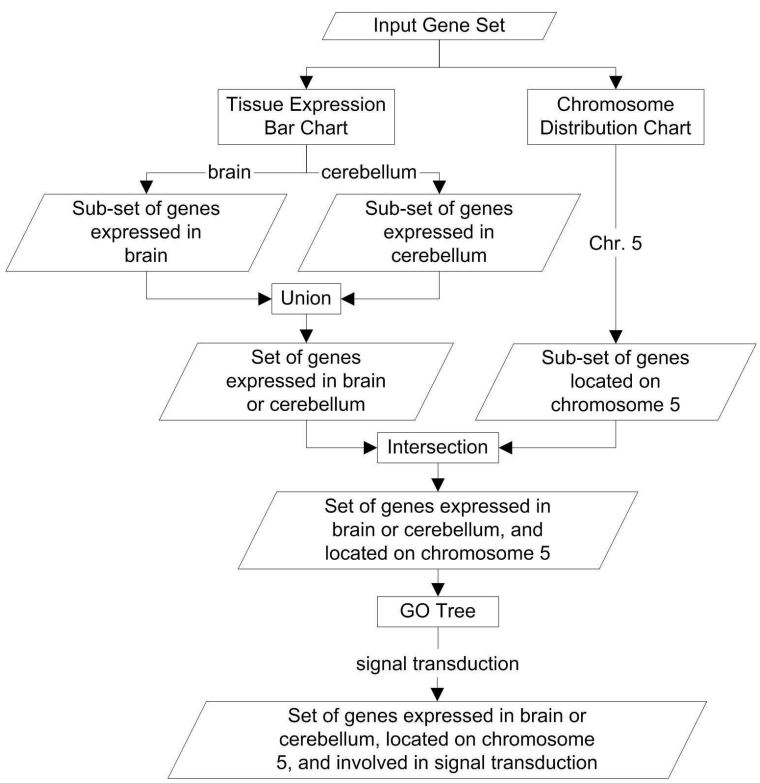


Figure 13. Solving complex biology questions using the organization module and the gene set management module in WebGestalt. The parallelograms in the flow chart represent the gene sets, the rectangles represent tools in WebGestalt.

6.2 Third party application

WebGestalt can be easily used by third party applications when requested. For example, WebGestalt has been implemented in WebQTL, which is a unique service that allows biologists to rapidly identify and map genes and quantitative trait loci (QTL). The WebGestalt modules are used to analyze sets of genes that are highly correlated with various phenotypes in WebQTL (Figure 14).

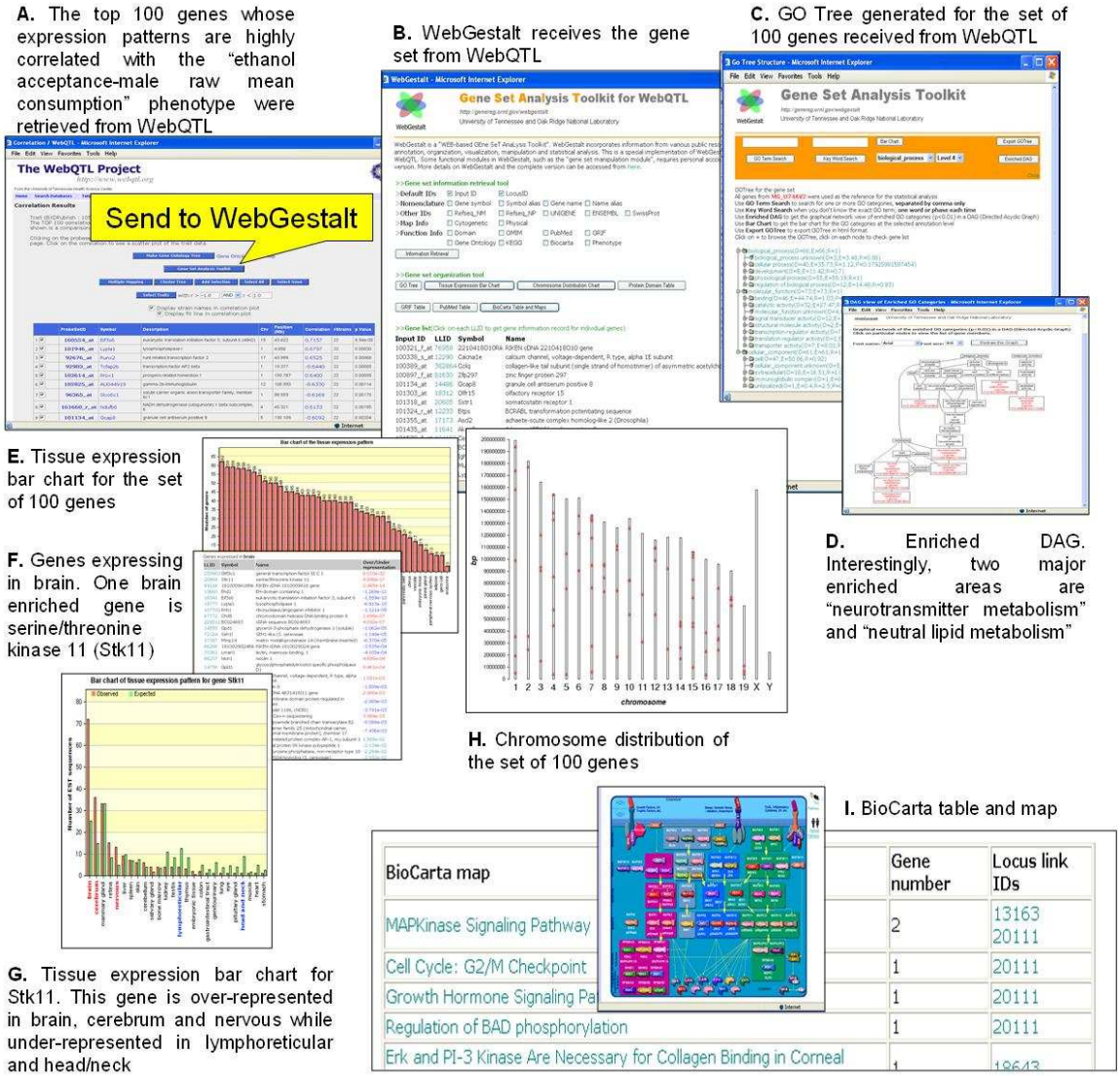


Figure 14. The WebGestalt modules are used to analyze sets of genes that are highly correlated with various phenotypes in WebQTL.