IMG 3.1 is the **19th** release of the Integrated Microbial Genomes (IMG) genomic data management and analysis system. **IMG 3.1** was released on **April 30th, 2010.**

IMG 3.1 Content

Genomes

The content of **IMG 3.1** has been updated with new microbial genomes available in **RefSeq version 39** (Jan 30, 2010). **IMG 3.1** contains a total of **5,648** genomes consisting of **1,912** bacterial, **84** archaeal, **76** eukaryotic genomes, **2,481** viruses (including bacterial phages), and **1,095** plasmids that did not come from a specific microbial genome sequencing project. Among these genomes, **4,638** are **finished** genomes, and **1006** are **draft** genomes, and **4** are **permanent draft** (i.e. will never be finished) genomes.

Note that **11** microbial genomes from **IMG 3.0** were **replaced** in **IMG 3.1** because (1) a **draft** genome has been replaced by its **finished** version, or (2) the composition of the genome has changed through the addition of new replicons (plasmids, chromosomes). For replaced genomes, the gene object identifiers (gene OIDs) for the protein-coding genes (CDS) were mapped to their new version in **IMG. 3.1**. See IMG Data Evolution History for details.

IMG 3.1 also contains **proteomic** data from recent *Arthrobacter chlorophenolicus* study[1], *Cryptobacterium curtum*, and *Brachybacterium faecium* studies.

[1] Unell, M., Abraham, P.E., Shah, M., Zhang, B. et al. (2009) Impact of Phenolic Substrate and Growth Temperature on the *Arthrobacter chlorophenolicus* Proteome. *J. Proteome Res.* **8** (4): 1953-1964.

IMG Statistics

Various statistics are provided via the **IMG Statistics** link on the home page of IMG, as shown below, including **IMG Total Gene Count** which consists of counting all the genes (protein coding genes, RNA genes) in IMG, except obsolete genes. Compared to **IMG 3.0**, **IMG 3.1** contains **8,114,289 genes**, an increase of **573,789** genes.

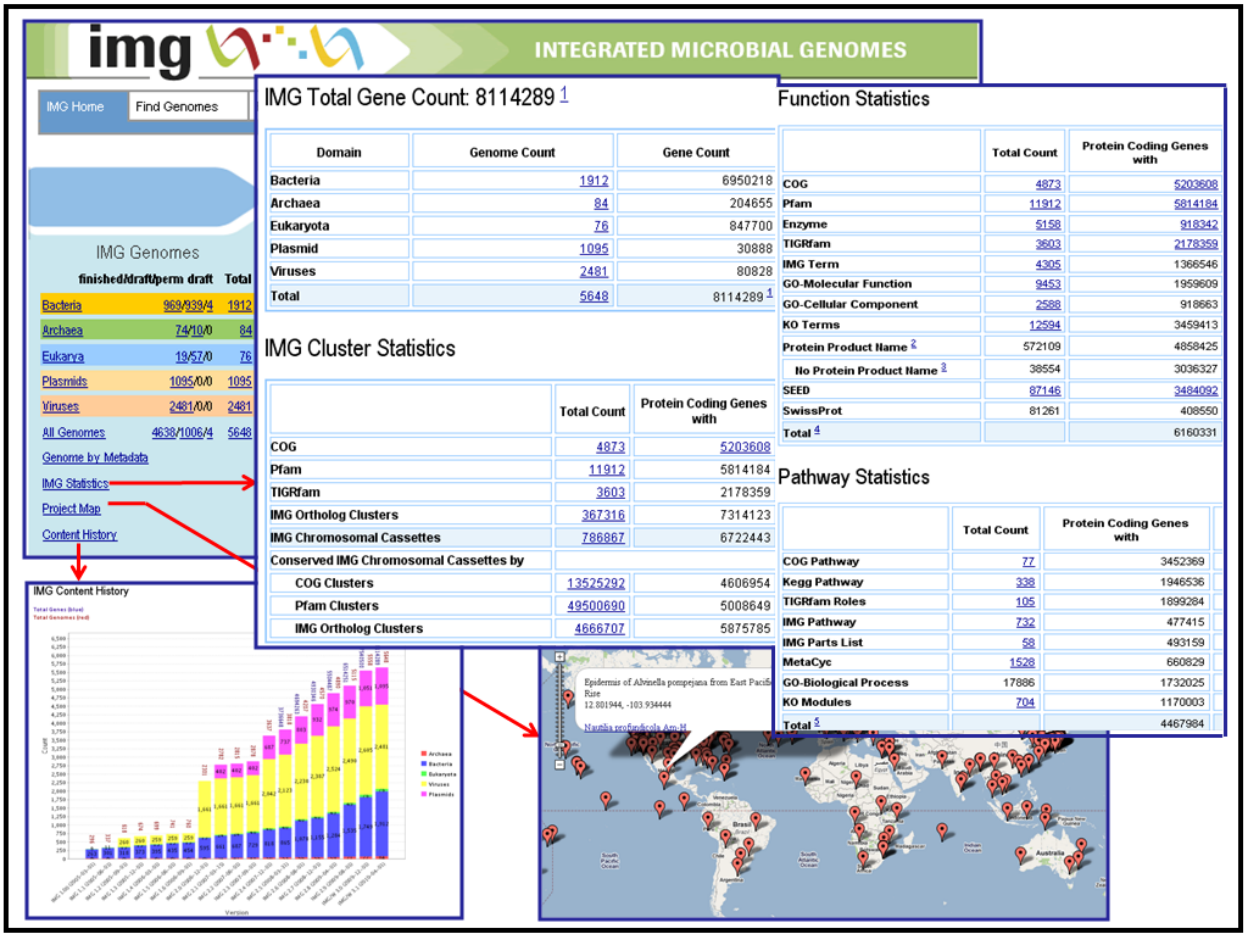


Figure 1. IMG Statistics.

The **Content History** link on the home page of IMG leads to a bar chart, as illustrated above, representing the growth of IMG in terms of number of genomes and genes since the release of its first version in March 2005. Note that the number of genes is recorded only for the last five versions of the system.

The **Project Map** link on the home page of IMG leads to a Google Map, as illustrated above, displaying the location of isolation sites for genomes that are associated with longitude/latitude coordinates in GOLD (http://www.genomesonline.org/).

IMG 3.1 User Interface

The User Interface (UI) has been extended with the following new features: **Dynamic Tables** for improved tabular display for the IMG data, **Improved Search Functions** for searching multiple terms in genomes, genes and functions, a new **Genome List Filter** for selecting and filtering genomes in the Genome List in various IMG pages, **Protein Expression Studies and Sample Clustering** for displaying and clustering protein expression and sample data, and new **Pathway Maps** for displaying KEGG pathway data with multiple genomes.

New Features

Dynamic Tables

IMG 3.1 now features a brand new dynamic layout on most tabulated information. Users may now intuitively sort columns by clicking on column headings. The sort order toggles between ascending and descending order on each click. This new layout also allows for large tables to be split into user selectable rows per page. Other features associated with this new table interface are drag and drop column reordering, resizable columns, a progress bar that dynamically indicates selected rows on all pages, and an Export button for saving any table as an Excel spreadsheet or a tab-delimited file.

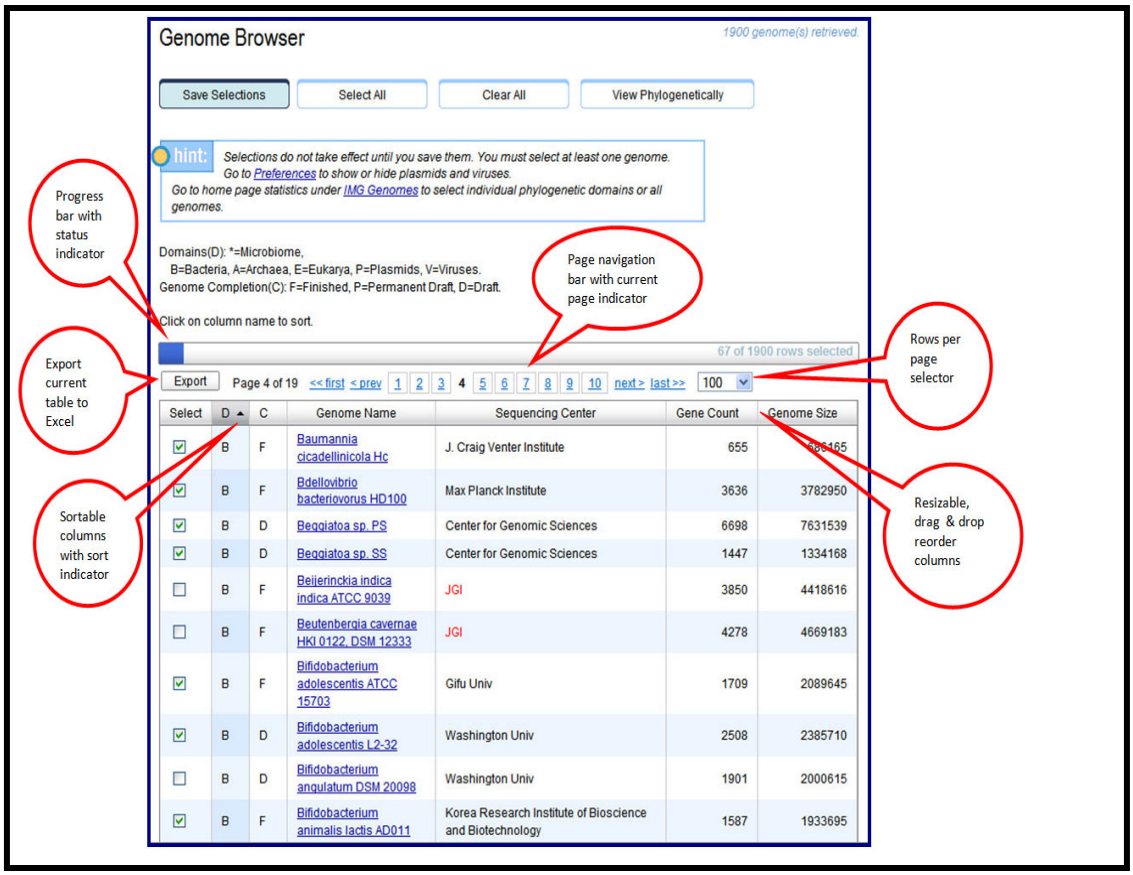


Figure 2. New Dynamic Table View.

Figure 2 above provides a visual overview of the new features on a sample table. These features are described in detail below.

1. **Sortable columns**: Most columns that contain text or numerical data can be sorted by clicking on the column title. Each mouse click on the column heading sorts the current view of the table alternately between ascending and descending order. The column by which the table is sorted is highlighted and the column heading is suffixed with a sort direction indicator. The symbol ▲ in the column heading denotes that the table is sorted by ascending order and ▼ denotes descending order. The sample table in Figure 1 is sorted by ascending order of **Domain** (the second column).
2. **Pagination**: By default, all tables are segmented into pages of 100 rows each. A pull down menu allows the user to change the number of rows per page at any time. Available choices for rows per page are *10, 25, 50, 100*, or *All*. In the case of very large tables, *All* is limited to 10,000 rows. Horizontal panels with page links above and below each table provide a mechanism for navigating to any page sequentially or randomly. The panel also has a status indicator showing the currently active page.
3. **Drag and Drop reorderable columns**: Columns in a table can be reordered by dragging them from one location and dropping them in another. To reorder a column, move the mouse to the heading of the column that needs to be moved and hover the mouse above the column heading until the cursor turns into a "move cursor" shaped "←↕→". Then drag the column to the desired position and drop. While dragging the column, the column's outline and a blue vertical line serve as guides in positioning the column.
4. **Resizable columns**: Any column in a table can be resized to the desired width by dragging the edge of the column heading to the left or to the right. To resize a column, position the mouse over the vertical edge of the column heading until the cursor changes into a "resize cursor" shaped "←||→". Then drag the edge to desired width.
5. **Dynamic progress bar**: Whenever a checkbox in a table row is selected either manually or as a result of clicking the *Select All* button, a progress bar appears with the status of rows selected. (Please refer to Figure 1.) The progress bar keeps track of the number of rows selected even upon navigating to a different page. As a result, in the case of large tables that are segmented across several pages, the progress bar provides a visual indication of the number of rows selected regardless of the active page. The progress bar remains active as long as at least one row is selected. If a table no longer has selected rows (either because they are deselected manually or as a result of clicking the *Clear All* button), the progress bar disappears. As a convenience, the progress is displayed in duplicate both above and below a table.
6. **Export table**: Data in a table can be exported for offline use or manipulation. Any table that has an *Export* button can be exported to an Excel file. (Please refer to Figure 1 for the location of the E*xport* button.) By default, the downloaded Excel

file name has the extension *.xls*. The file can either be saved to the user's desktop or be opened directly in Excel. If the browser does not detect the presence of Excel on the user's computer, it attempts to open the file in the application designated to handle *.xls* files. To export data in a table, select the desired rows and click the *Export* button. To export the entire table, click the *Select All* button to select all rows and then click the *Export* button. On tables that do not have checkboxes for selecting rows, the entire table is exported automatically. The exported file always reflects the same sort and column orders as on the displayed table.

Improved Search Functions

IMG 3.1 now extends the "exact" search into "list" search. Users not only can continue the single-item search, but also can perform multiple-items search at once. To perform multiple-items search, users must use comma to separate these items. All search terms are case-insensitive, allow leading and trailing white space, must contain alphanumeric character, and blank search will be blocked. For ID searches on COG, GO, EC, Pfam, TIGRfam, InterPro and KEGG, users are allowed to simply input digit numbers, the addition of specific ID prefix (such COG, GO:, EC:, etc.) will be taken care of by the program.

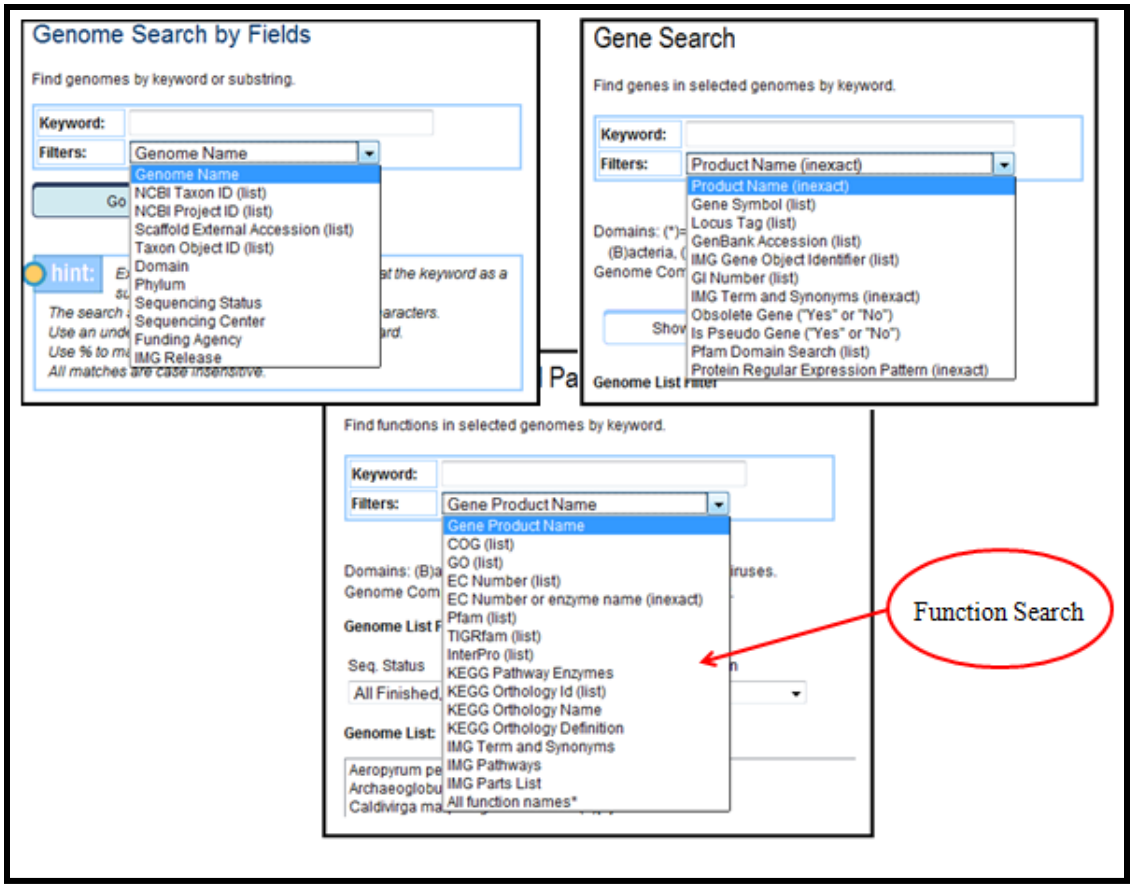


Figure 3. New Search Functions.

The following "list" searches have been included into **IMG 3.1**:

Genome Search

1. NCBI Taxon ID (list)
2. NCBI Project ID (list)
3. Taxon Object ID (list)

Gene Search

1. Gene Symbol (list)
2. Locus Tag (list)
3. GenBank Accession (list)
4. IMG Gene Object Identifier (list)
5. GI Number (list)

Function Search

1. COG (list)
2. GO (list)
3. EC Number (list)
4. Pfam (list)
5. TIGRfam (list)
6. InterPro (list)
7. KEGG Orthology Id (list)

Genome List Filter

The Genome List Filter, residing inside the pages of Gene Search, Function Search, Cassette Search, Artemis ACT, and Genome vs Metagenomes, now features client-side dynamic update. Through making choice of "seq. status" or "domain", user can notice a rapid response in the listing area of "Genome List". Microbiome became part of the "domain" selection if metagenome inclusion is desired. Further, if genomes have been selected and saved, user will be provided with the chance to make choice between the "selected" and "all" genomes through a dynamic button.

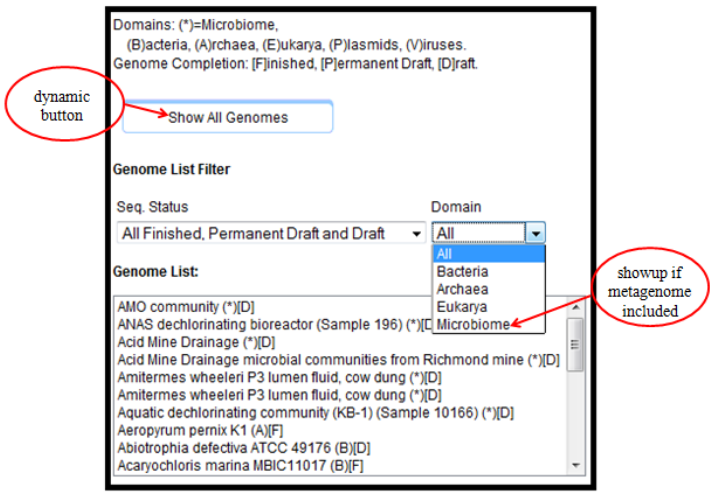


Figure 4. Genome List Filter.

KEGG Pathways – *Pathway Maps*

In KEGG Pathway Details, the user can select multiple genomes to be highlighted on the given map (see Figure 5 below). Each gene on the map is differentially colored depending on the percentage of genomes in which that gene is found (see Figure 6 below). The tooltip for a colored gene displays the number of genomes out of total selected that a gene is in.

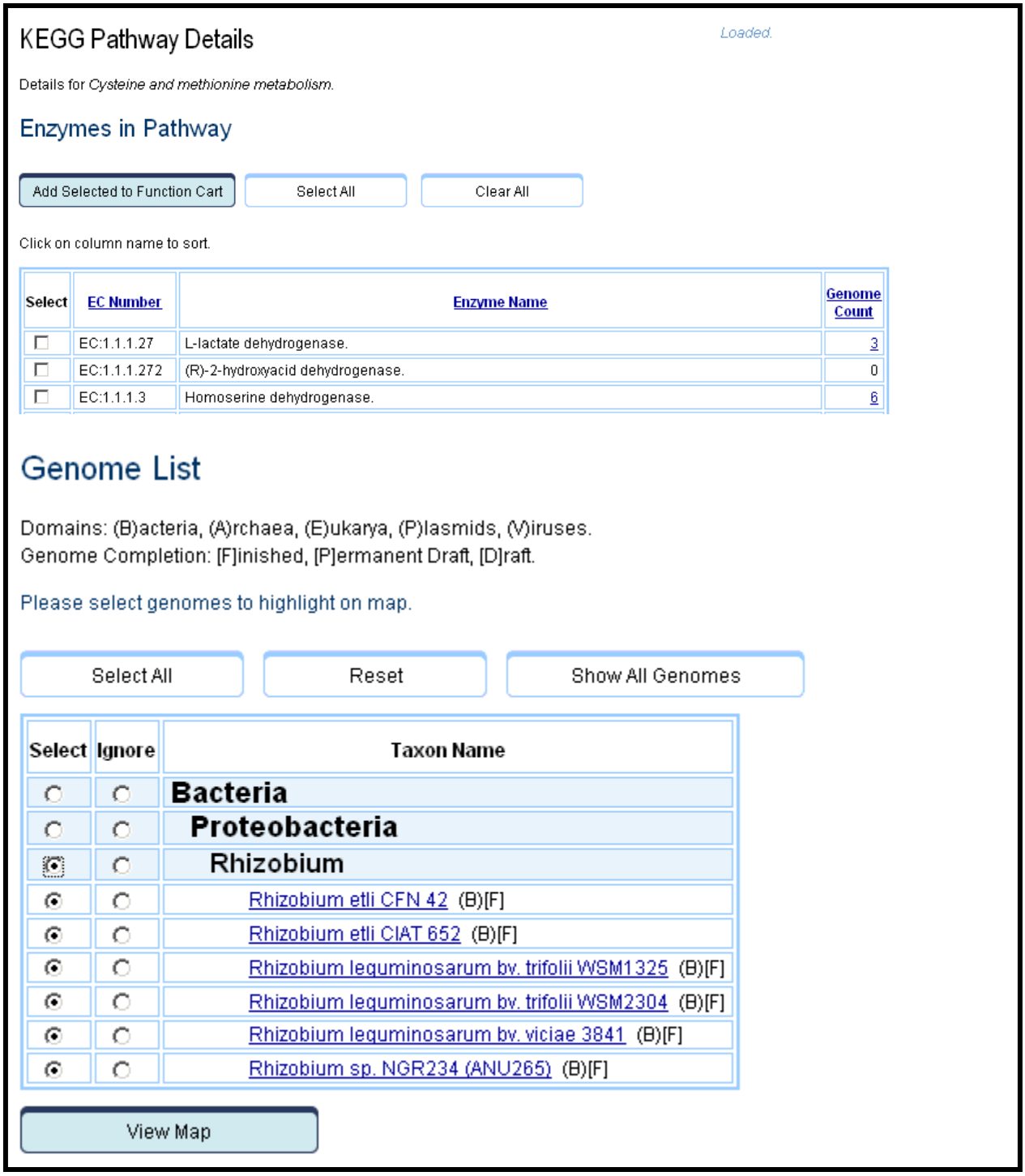


FIGURE 5. KEGG Pathway Maps.

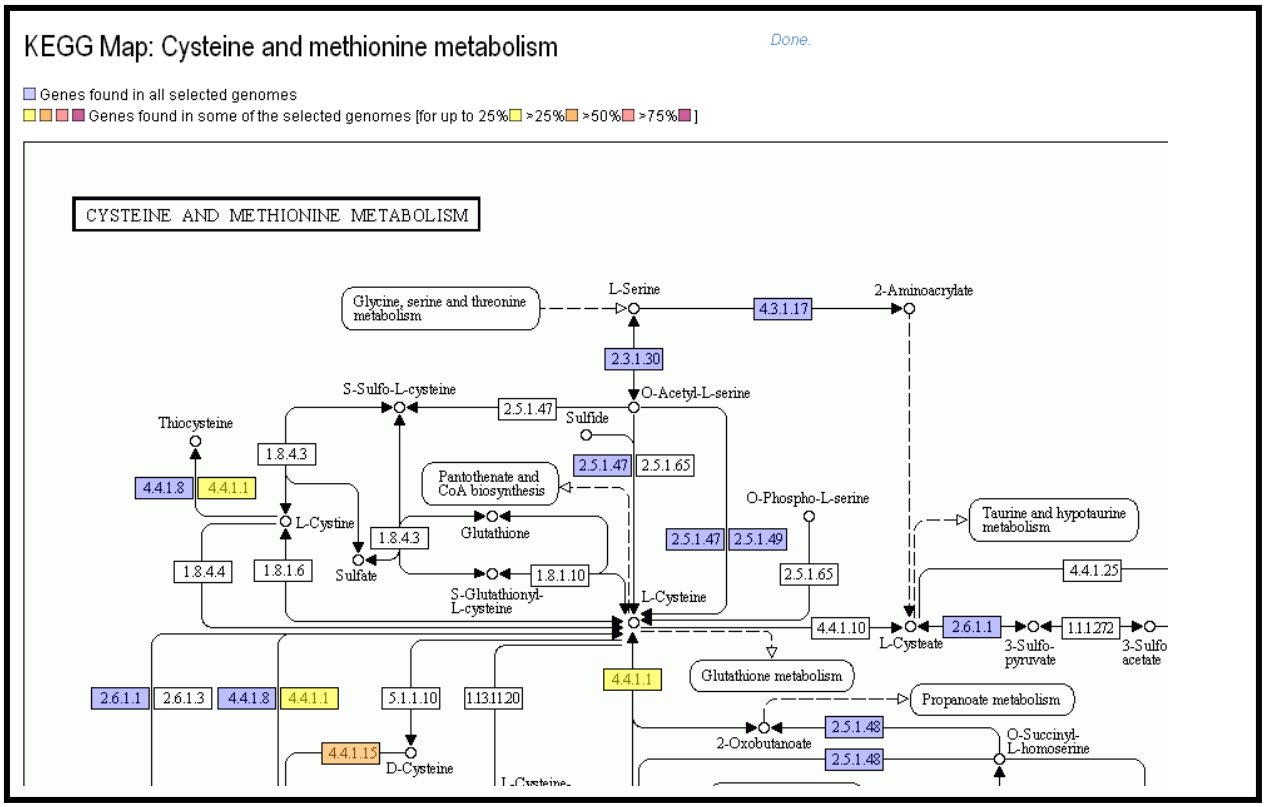


FIGURE 6. KEGG Pathway Map Coloring

Organism Details – *Protein Expression Studies*

IMG 3.1 contains **protein expression** data from three protein expression studies. The tools for displaying these data have been extended, as illustrated in Figure 7 below.

The **Protein Expression Studies** link on the **Organism Details** for a genome associated with protein expression studies, illustrated in Figure 7(i), leads to the list of studies for the genome of interest, as illustrated in Figure 7(ii). Each protein expression study is displayed together with the number of expressed genes and observed peptides. The **Study Name** provides a link to the list of **Experiments** (samples) for the study, as illustrated in Figure 7(iii).

The description for each sample consists of the experimental conditions and provides a link to the **Protein Expression Data for the Sample** organized per expressed gene, as illustrated in Figure 7(iv). For each expressed gene, the number of observed peptides and **normalized coverage** (NSAF) are provided. **NSAF** is the coverage for an expressed gene in an experiment divided by the total coverage of all the genes in that experiment, whereby the **coverage** for a gene is defined as the count of peptides divided by the size of the gene[2]. The user can select genes of interest expressed in any given sample and add them to the Gene Cart.

[2] Florens, L., Carozza, M.J., Swanson, S.K. et al. (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors, *Methods* **40**(4): 303-311.

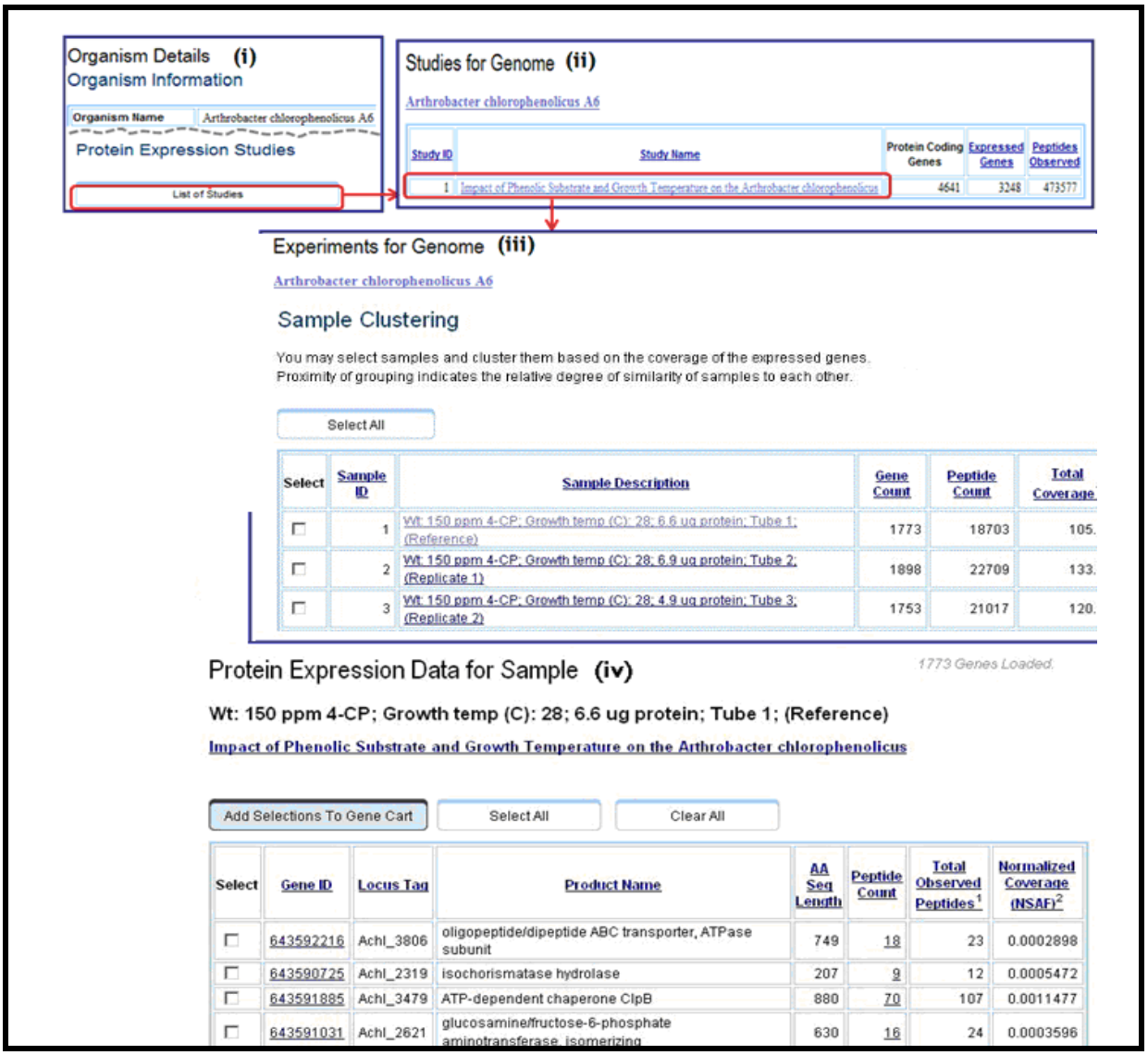


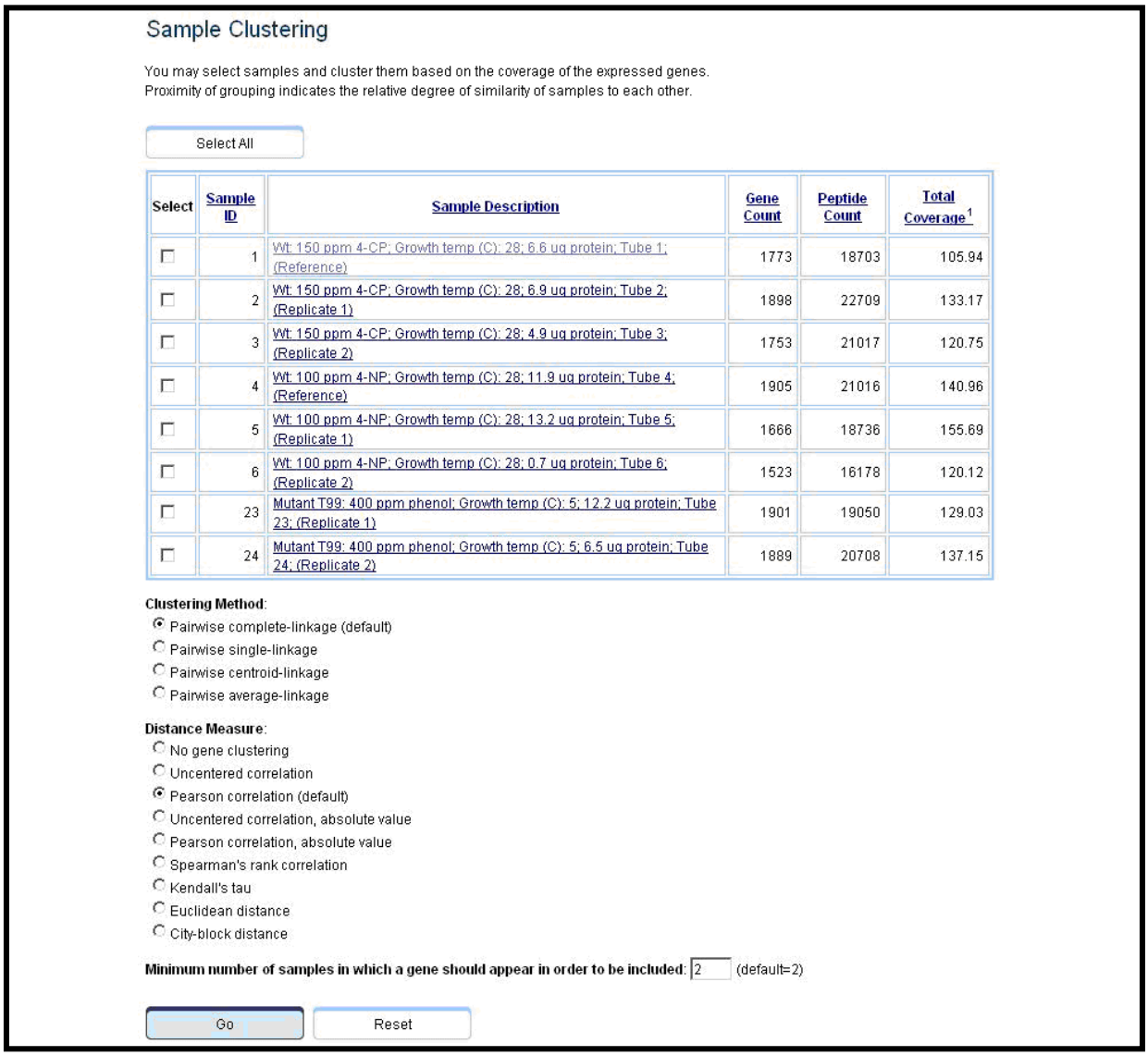
FIGURE 7. Organism Details- Protein Expression Studies.

Sample Clustering

For each study, the experiments (samples) can be selected for **clustering** based on coverage values for the genes expressed in each sample (see Figure 8 below). The user can specify the clustering method and the distance measure to use, as well as and the minimum number of samples in which a gene should appear in order to be included in the calculation. Both samples and genes are clustered. What will be displayed in IMG is a hierarchical cluster tree of samples and a normalized heat map of coverage values for each gene for each sample (see Figure 9 below). On the heat map, the samples link to sample pages and the genes link to gene pages. Clicking on a sample in the cluster tree **sorts** the heat map on descending coverage values for that sample.

In addition, the results can be viewed in Java TreeView applet (see Figure 10 below). The applet is loaded with the cluster files by clicking on the "View with Java TreeView" button. Within

the Java TreeView, the genes link to IMG gene pages and the samples link to IMG sample pages.



Sample Clustering

You may select samples and cluster them based on the coverage of the expressed genes.
Proximity of grouping indicates the relative degree of similarity of samples to each other.

Select All

Select	Sample ID	Sample Description	Gene Count	Peptide Count	Total Coverage[1]
☐	1	Wt: 150 ppm 4-CP; Growth temp (C): 28; 6.6 ug protein; Tube 1; (Reference)	1773	18703	105.94
☐	2	Wt: 150 ppm 4-CP; Growth temp (C): 28; 6.9 ug protein; Tube 2; (Replicate 1)	1898	22709	133.17
☐	3	Wt: 150 ppm 4-CP; Growth temp (C): 28; 4.9 ug protein; Tube 3; (Replicate 2)	1753	21017	120.75
☐	4	Wt: 100 ppm 4-NP; Growth temp (C): 28; 11.9 ug protein; Tube 4; (Reference)	1905	21016	140.96
☐	5	Wt: 100 ppm 4-NP; Growth temp (C): 28; 13.2 ug protein; Tube 5; (Replicate 1)	1666	18736	155.69
☐	6	Wt: 100 ppm 4-NP; Growth temp (C): 28; 0.7 ug protein; Tube 6; (Replicate 2)	1523	16178	120.12
☐	23	Mutant T99: 400 ppm phenol; Growth temp (C): 5; 12.2 ug protein; Tube 23; (Replicate 1)	1901	19050	129.03
☐	24	Mutant T99: 400 ppm phenol; Growth temp (C): 5; 6.5 ug protein; Tube 24; (Replicate 2)	1889	20708	137.15

Clustering Method:
- (●) Pairwise complete-linkage (default)
- ( ) Pairwise single-linkage
- ( ) Pairwise centroid-linkage
- ( ) Pairwise average-linkage

Distance Measure:
- ( ) No gene clustering
- ( ) Uncentered correlation
- (●) Pearson correlation (default)
- ( ) Uncentered correlation, absolute value
- ( ) Pearson correlation, absolute value
- ( ) Spearman's rank correlation
- ( ) Kendall's tau
- ( ) Euclidean distance
- ( ) City-block distance

Minimum number of samples in which a gene should appear in order to be included: 2 (default=2)

Go Reset

FIGURE 8. Clustering of protein samples.

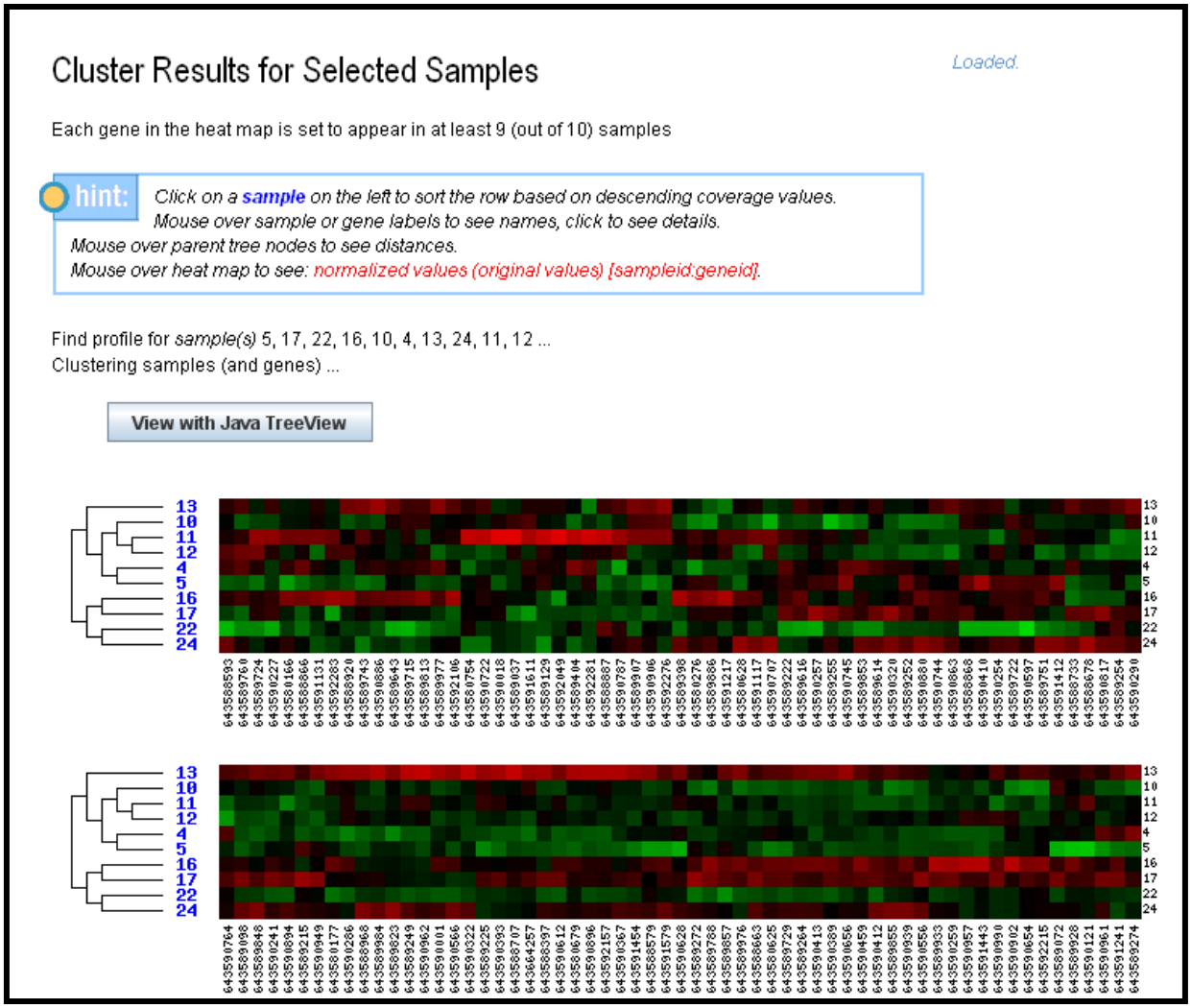


FIGURE 9. Clustering results.

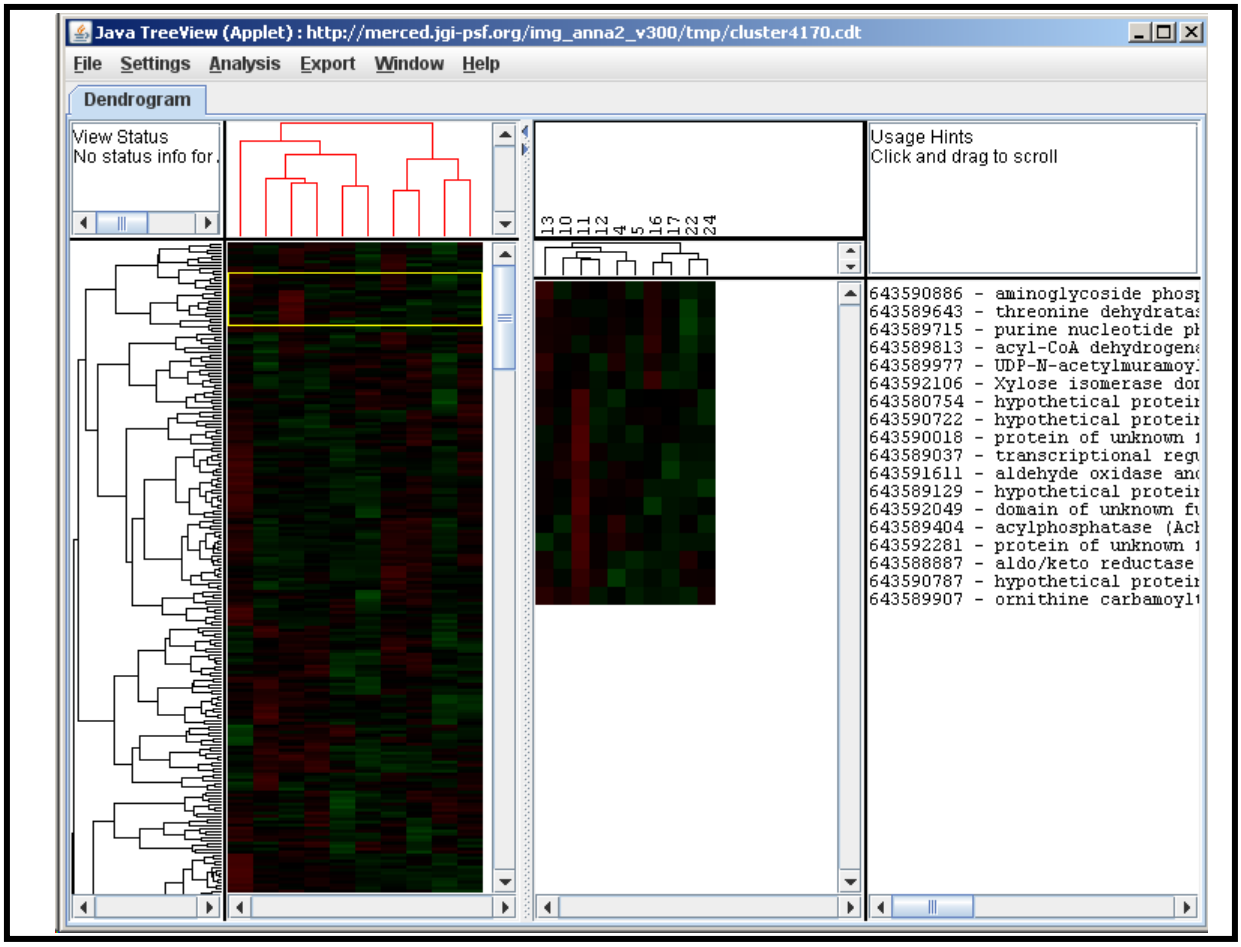


FIGURE 10. Cluster results displayed in Java TreeView application.