IMG 2.7 is the **15th** release of the Integrated Microbial Genomes (IMG) genomic data management and analysis system. **IMG 2.7** was released on **December 10th, 2008.**

IMG 2.7 Content

Genomes

The content of **IMG 2.7** has been updated with new microbial genomes available in **RefSeq version 31** (August 30, 2008).

IMG 2.7 contains a total of **4,570** genomes consisting of **1,155** bacterial, **56** archaeal, **40** eukaryotic genomes, **2,387** viruses (including bacterial phages), and **932** plasmids that did not come from a specific microbial genome sequencing project. Among these genomes, **4,091** are finished genomes, and **479** are draft genomes; **328** are **JGI** sequenced genomes: **243** are finished and **63** are draft genomes.

Note that **31** microbial genomes from **IMG 2.6** were **replaced** in **IMG 2.7** because (1) a "Draft" genome has been replaced by its "Finished" version or (2) the composition of the genome has changed through the addition of new replicons (plasmids, chromosomes).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. 2.7. See IMG Data Evolution History for details.

Plasmid names were curated by adding strain names to organism name when available from publications or other sources.

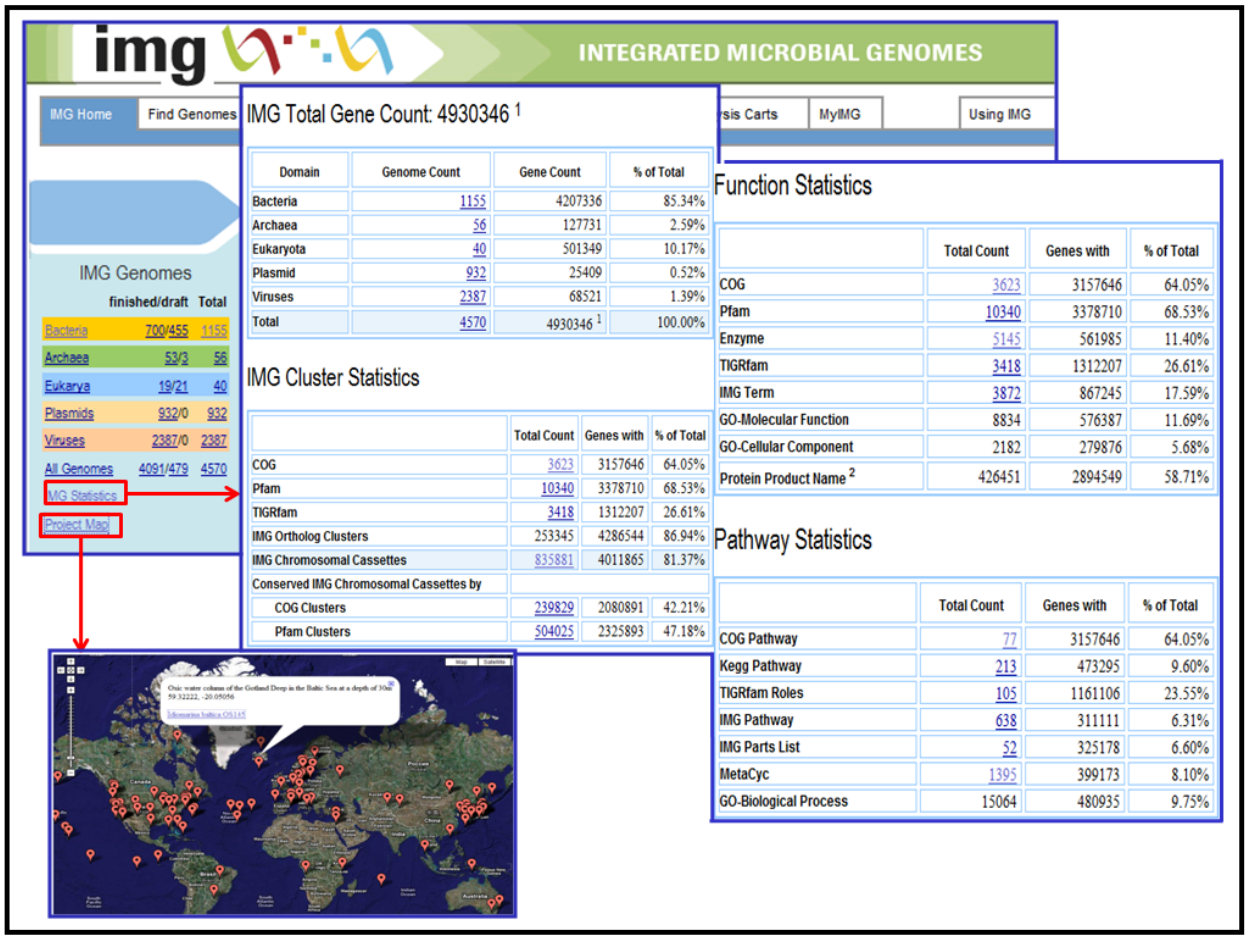
tRNA and **rRNA** genes (23S, 16S and 5S) missing from the original RefSeq genome files are added using tRNAscan-SE v1.23 for tRNA genes and similarity comparisons to existing RNA genes. In IMG 2.7 **928** tRNA and **202** rRNA genes were added in **167** genomes. Furthermore, the existing RNA genes were checked and the orientation of 24 rRNA genes and 50 tRNA genes in public genomes was corrected.

IMG 2.7 includes pathways from **MetaCyc** (http://www.metacyc.org/) version 12.5 (Oct 27, 2008).

IMG Statistics

Various statistics are provided via the **IMG Statistics** link on the home page of IMG, as shown below, including: (i) **IMG Total Gene Count** which consists of counting all the genes (protein coding genes, RNA genes) in IMG, except obsolete genes, and (ii) **Protein Product Names** which consists of counting all distinct protein product names associated with (predicted for) protein coding genes (CDSs); note that this count does not include RNA or obsolete genes.

Compared to **IMG 2.6**, **IMG 2.7** contains **4,930,346 genes**, an increase of **236,083** genes.



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 2.7 User Interface

The User Interface (UI) has been extended in order to improve its overall functionality and usability.

The main UI changes are marked on the UI Map diagram above and include:

- **New features**
  - **(i)** Each **Organism Details** page contains a **Phylogenetic Distribution of Genes** that allows examining potential horizontally transferred genes of a genome based on the distribution of best BLAST hits of its protein-coding genes.
  - (ii) **MetaCyc** pathways have been included into IMG 2.7 and are used for characterizing all genomes.
  - (iii) An additional version of **Function Profile** is provided under **Compare Genomes** with a more flexible genome selection capability.
  - (iv) **Missing Enzymes** can be examined within the context of a **KEGG Map** or **Function Profile** result.
- **Extended features**
  - (i) **Genome Search** and **Organism Details** extended to include additional **metadata** fields.
  - (ii) **Genome Statistics** part of **Organism Details** extended with genes without enzymes but with PRIAM hits.
  - (iii) **Gene Details** for pseudogenes show their individual components.
  - (iv) **Function Search** extended with keyword search across all functional name filters.
  - (v) **Phylogenetic Profiler** for single genes extended with **Missing Gene** search.
  - (vi) **Summary Statistics** table for comparing genomes includes links to the **graphical viewers** for displaying the distribution of genes based on their COG and KEGG, Pfam, and TIGRfam associations.

New Features

Organism Details – *Phylogenetic Distribution of Genes*

Phylogenetic Distribution of Genes provides a glimpse into the evolutionary history of the genes in the genome based on the distribution of best BLAST hits of its protein-coding genes. The genes that were likely vertically inherited are expected to have higher sequence similarity to the genes in the genomes within the same taxonomic group, while those horizontally transferred may have their best BLAST hits to the genes in distantly related organisms. Since this tool considers best BLAST hits and does not perform phylogenetic tree reconstruction and analysis, the results can be used as a first approximation of the evolutionary history of the genes and require manual analysis to establish whether the genes of interest were indeed horizontally transferred.

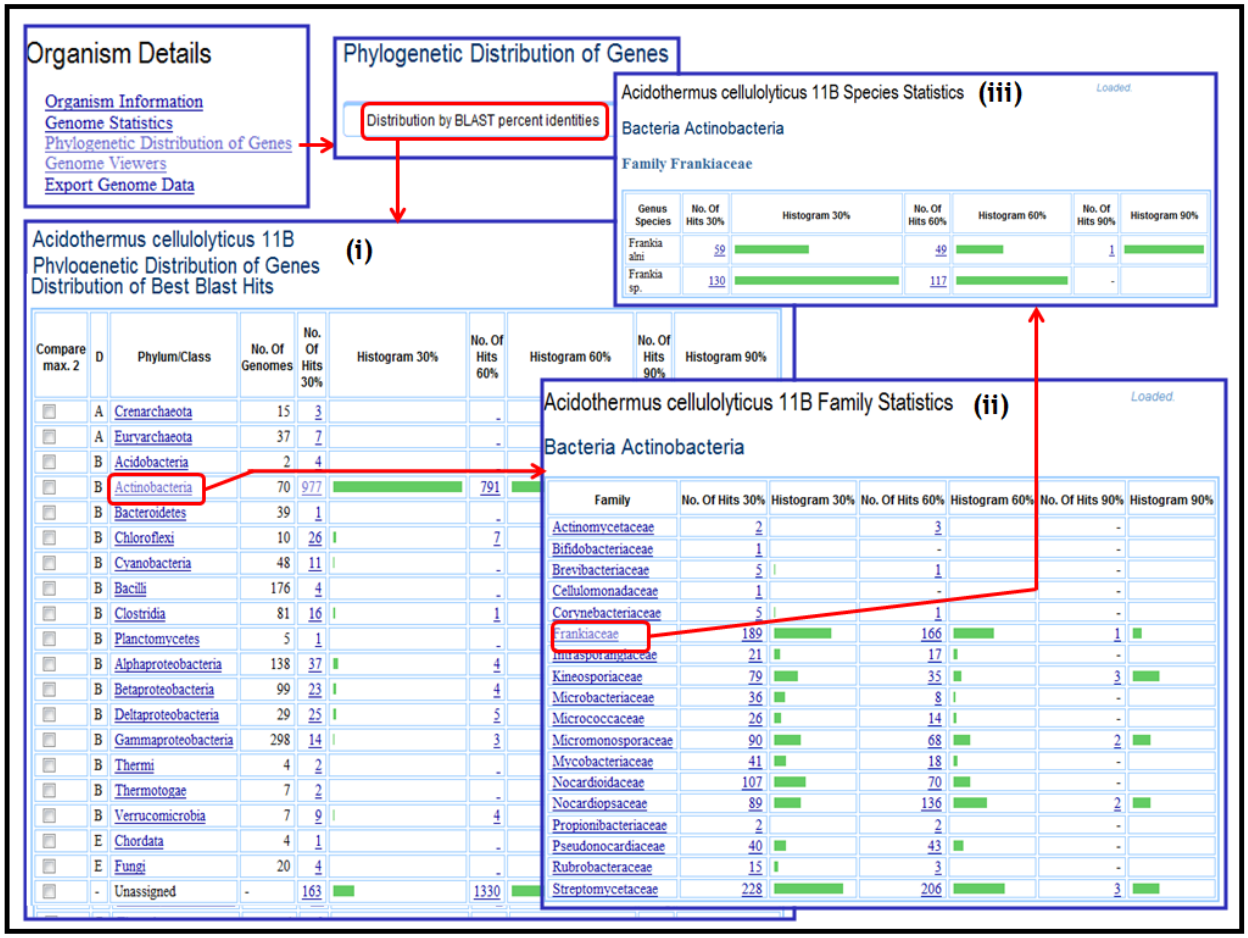


FIGURE 1. Organism Details – Phylogenetic Distribution of Genes.

The phylogenetic distribution of best BLAST hits of protein-coding genes in a selected genome is displayed as a histogram, as shown in Figure 1(i); counts correspond to the number of genes that have best BLASTp hits to proteins of other genomes in a specific phylum or class with more than 90% identity (right column), 60-90% identity (middle column) and 30-60% identity (left column). Gene counts in the histogram are linked to the lists of genes in the selected genome that have best BLAST hit in a certain phylum/class with specified percent identity. The genes in the list can be sorted either by their oids ("*Table View*") or by their assignment to COGs, which

in turn can be classified according to COG Functional Categories ("*COG Functional Cat.*") or COG Pathways ("*COG Pathways*"). The genes in the table can be selected and added to **Gene Cart** or analyzed through the corresponding **Gene Details**.

For genes that have best BLASTp hits with 30%, 60-90%, and 90% identity, the tool displays summary statistics of COG functional categories and pathways associated with these genes either across all phyla/classes or across two selected phyla/classes. The statistics show the number of genes associated with a specific COG category or pathway, as well as the percentage of these genes out of the total number of genes with affiliation to this particular class or phylum in the certain interval of percent identity of the best BLAST hits (shown in parenthesis). The phylogenetic distribution of best BLAST hits can be **projected** onto the **families** in a phylum/class (see Figure 1(ii)), and then further onto species in a family (see Figure 1(iii)).

Find Functions – *MetaCyc Pathways*

IMG includes the **MetaCyc** collection of pathways available from http://www.metacyc.org/. IMG genomes are associated with MetaCyc pathways via enzymes predicted using PRIAM.

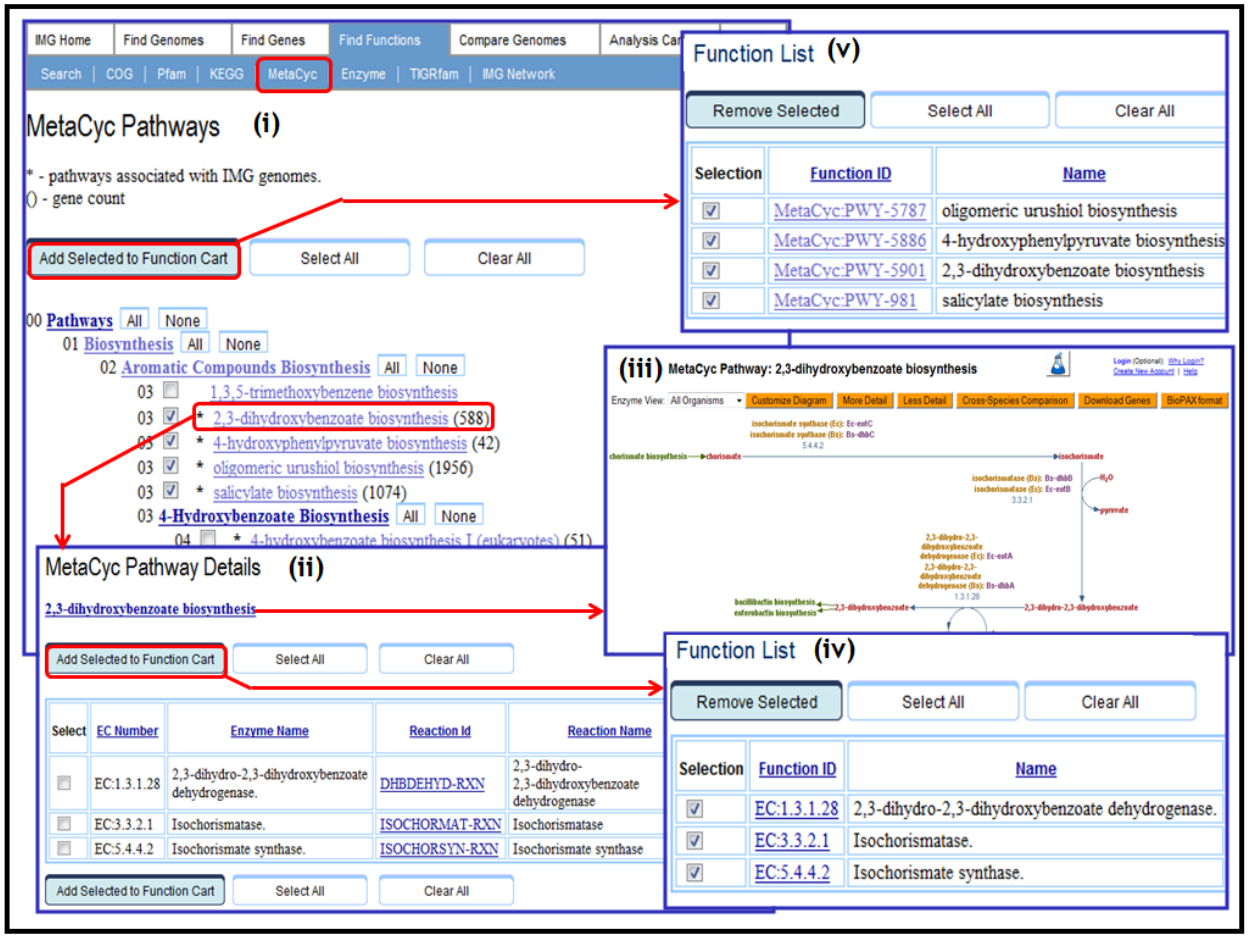


FIGURE 2. Find Functions – MetaCyc Pathways.

From the **Find Function** top-level menu, the **MetaCyc** option on the second-level menu leads to the **MetaCyc Pathways** browser, as shown in Figure 2(i). MetaCyc pathways associated with IMG genomes are marked and associated with the number of related IMG genes. MetaCyc

pathways are organized hierarchically within four broad categories: biosynthesis, degradation/utilization/assimilation, generation of precursor metabolites, and energy and detoxification. Each category is organized in sub-categories (classes) of pathways.

MetaCyc pathways can be examined using the **MetaCyc Pathway Details** pages, as shown in Figure 2(ii). The **MetaCyc Pathway Details** provides a link to the specification of the pathway at the MetaCyc site, as illustrated in Figure 2(iii), together with a list of the enzymes associated with a specific reaction in the pathway. For each enzyme, the number of genes associated with this enzyme is also provided, together with a link that leads to the list of these genes. By clicking on the left-column checkbox for an enzyme entry in the MetaCyc Pathway Details page, enzymes can be added to the Function Cart for further analysis, as illustrated in Figure 2(iv). Pathways can be also added directly to the Function Cart from the MetaCyc Pathways browser, as illustrated in Figure 2(v).

Compare Genomes – *Function Profile*

A new version of **Function Profile** has been added under the **Compare Genomes** main menu tab. Similar to the version available under **Analysis Carts/ Functions**, the **Function Profile** displays the number (abundance) of genes associated with a function selected from the **Function List** across selected genomes, as shown in Figure 3.

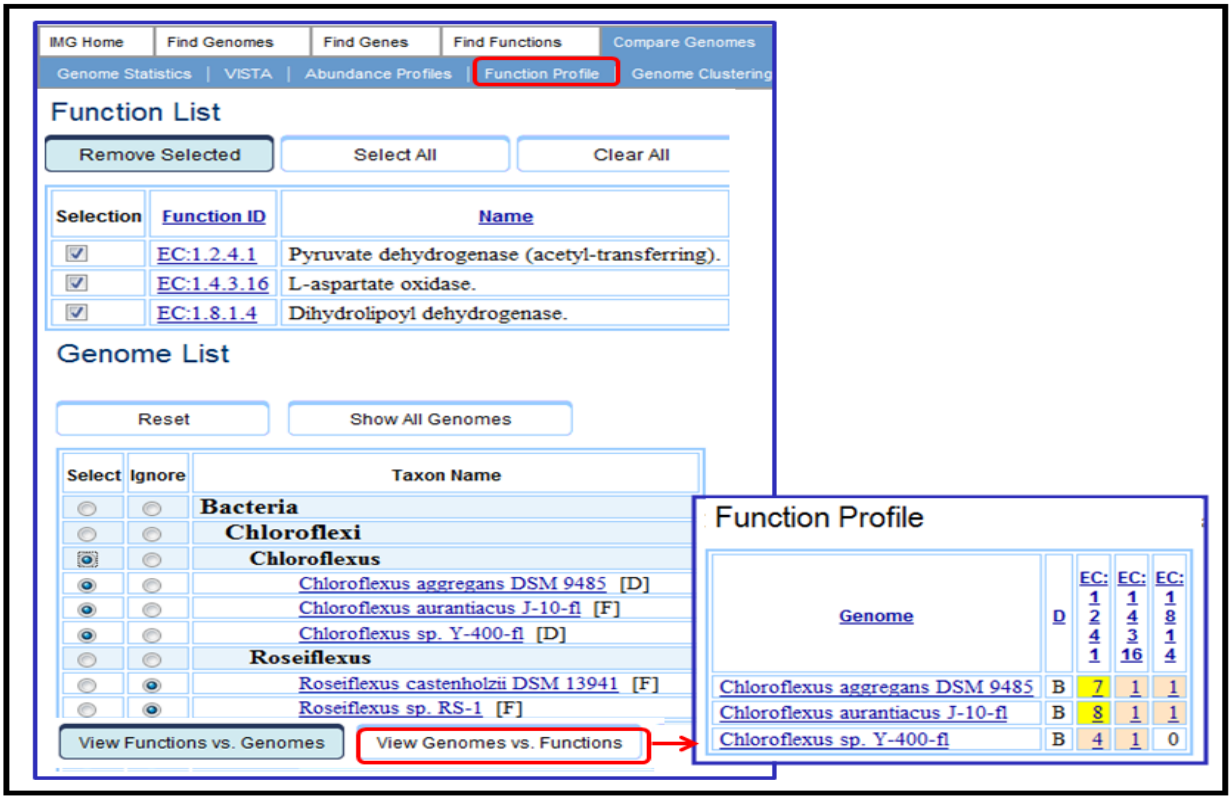


FIGURE 3. Compare Genomes – Function Profile.

Genome selection is provided via a phylogenetically organized list of genomes, whereby genomes can be selected either individually or as phylogenetic groups. The list of genomes displays the genomes that have been selected and saved using the **Genome Browser**. The entire list of genomes can be displayed using the "Show All Genomes" button.

The **Function Profile** results are displayed in a tabular format with each column displaying the profile of a specific function across the genomes, as illustrated in the right side pane of Figure 3. Each cell in the profile-result table displays the count (abundance) of genes in a specific genome associated with a specific function, and contains a link to the associated list of genes.

Missing Enzymes – *KEGG Maps & Function Profile*

Genomes may have potentially "missing" associations between enzymes and their genes. We call such associations ***missing enzymes.*** Missing enzymes can be examined using either a **KEGG Pathway Map** for a genome of interest or a **Functional Profile** involving genomes and enzymes of interest, as illustrated in Figure 3.

Once a KEGG pathway is selected using the **KEGG Browser** under **Find Functions**, you can view its map for a selected genome using the "Find missing enzymes" option, as illustrated in Figure 4(ii). On the **KEGG Map**, such as that shown in Figure 4(iii), enzymes that are associated with genes of the target genome are colored blue, while so called "missing" enzyme are colored either ***green***, for enzymes that have a PRIAM hit to genes of the target genome, or ***white*** for enzymes without PRIAM hits. Clicking on a missing enzyme will lead to a **Find Candidate Genes for Missing Function** page, as shown in Figure 4(iv). Note that selection of a (green colored) missing enzyme that has a PRIAM hit enhances the chances of finding for it good candidate genes.

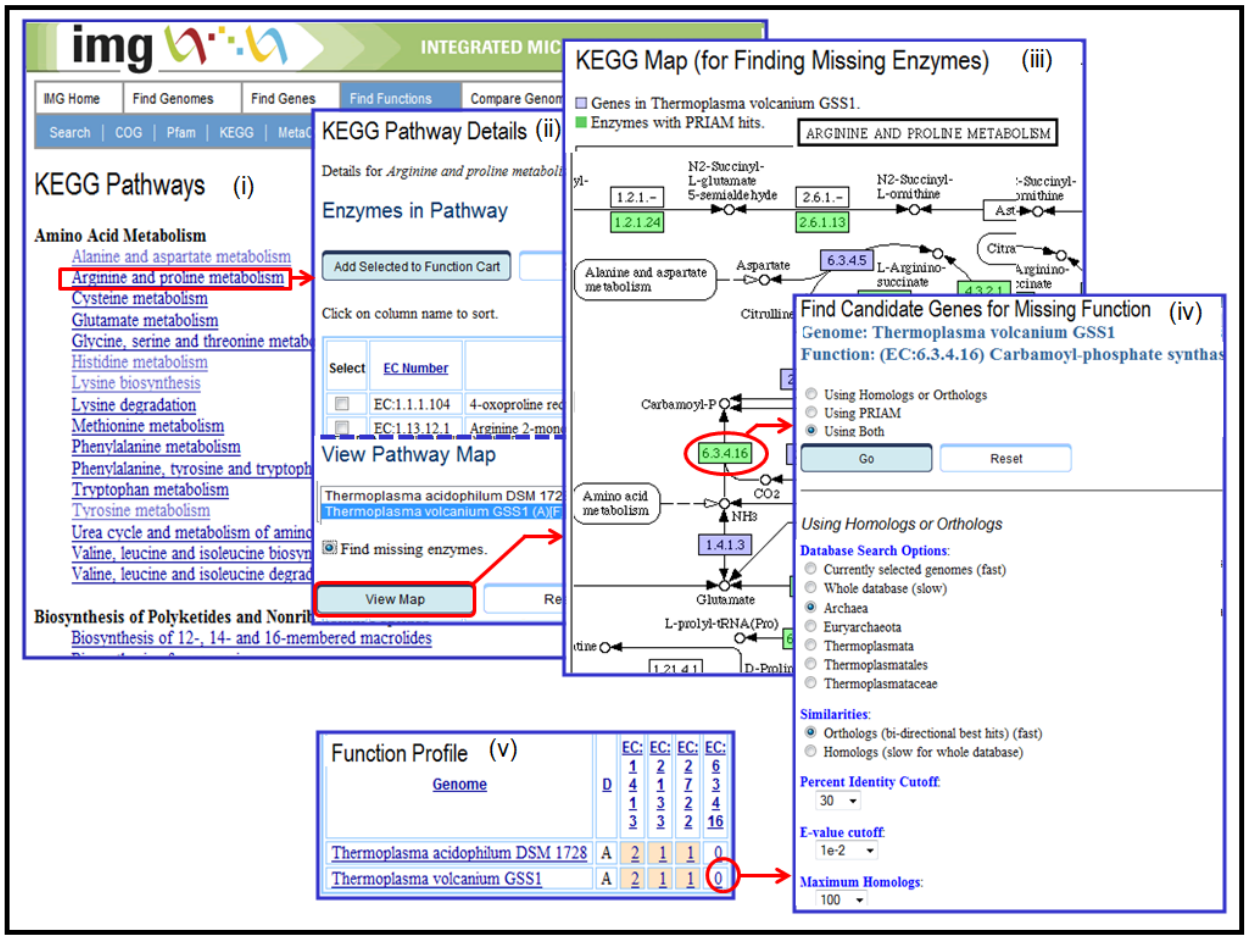


FIGURE 4. Examining Missing Enzymes via a KEGG Pathway Map or Function Profile.

You can find candidate genes of your target genome that could be associated with a missing enzyme by searching for genes that have **homologs**/**orthologs** associated with the missing enzyme, as illustrated in Figure 4(iv). You can search across all the genomes available in the system, across a subset of genomes within a certain domain/phyla/class, or only across the selected genomes. You can change the default values set for percent identity and e-value cutoffs and the number of retrieved homologs. Alternatively, you can employ **PRIAM** for finding genes that could be associated with the "missing" enzyme. You can change the default values set for percent identity, e-value, and percent alignment cutoffs.The result of the search dor candidate genes consists of a list of genes that can be selected and included into the **Gene Cart.**

In the result for a **Function Profile** involving enzymes, missing enzymes are identified by a "0". Clicking on the "0" identifying a missing enzyme, as shown in Figure 4(v), will also lead to a **Find Candidate Genes for Missing Function** page.

Extended Features

Find Genomes – *Genome Search*

Genome Search based on metadata has been extended to allow searches based on Oxygen Requirement, Motility, Sporulation, Salinity, Temperature Range, Phenotype, Disease, Relevance, Habitat, Cell Arrangement, Energy Source, and Metabolism, as shown in Figure 5.

Metadata based genome searches are carried out directly on the IMG-GOLD database that underlies the GOLD catalog of genome projects (http://www.genomesonline.org/). Search results include the metadata values for the fields involved in the search, as illustrated in the right pane of Figure 5.

Metadata in the **Organism Details** page for a genome has been also extended to include the metadata available for the corresponding project in GOLD (http://www.genomesonline.org/). Projects with longitude/latitude coordinates are associated with a Google Map of their location and are included into the **Project Map** of all IMG genomes, as mentioned above.

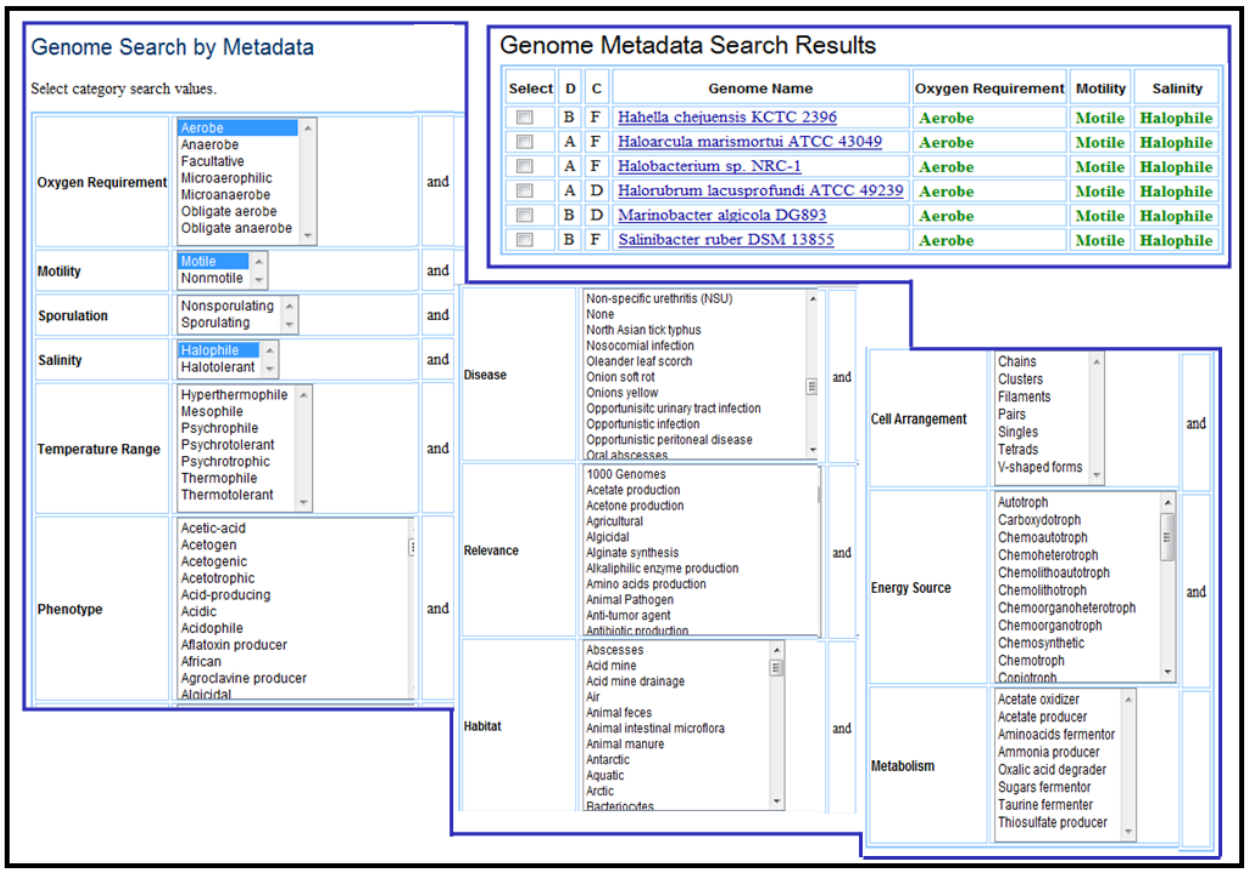


FIGURE 5. Find Genomes – Genome Search.

Organism Details –*Genome Statistics*

The **Genome Statistics of Organism Details** has been extended with a count of "Genes without enzymes, but with PRIAM hits" which leads to a list of genes that could be associated with enzymes predicted by PRIAM, as illustrated in Figure 6. Note that only the top PRIAM hits are provided in this list.

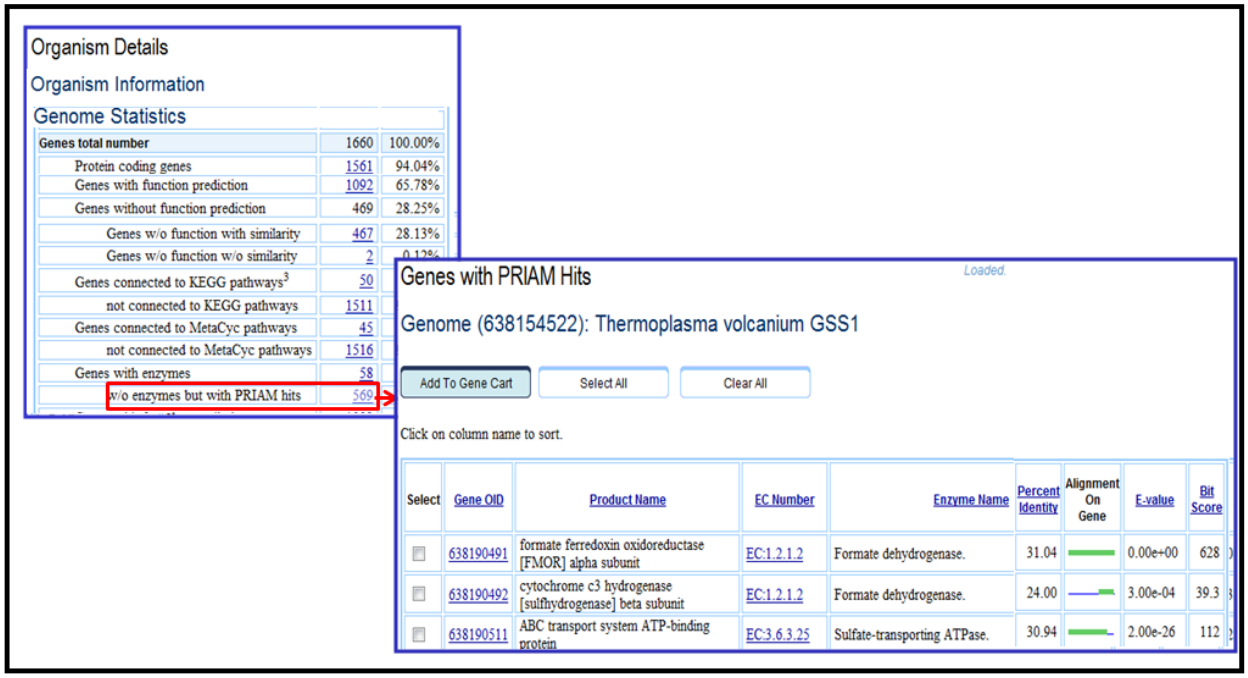


FIGURE 6. Genome Statistics- Genes without Enzymes, but with PRIAM hits.

Gene Details – *Pseudogenes*

The chromosome and gene neighborhood viewers provided in the **Gene Details** page has been extended to display the components of pseudogenes, as illustrated in the right side pane of Figure 7. The previous versions of these viewers were displaying pseudogenes as one contiguous region, as shown in the left pane of Figure 7.

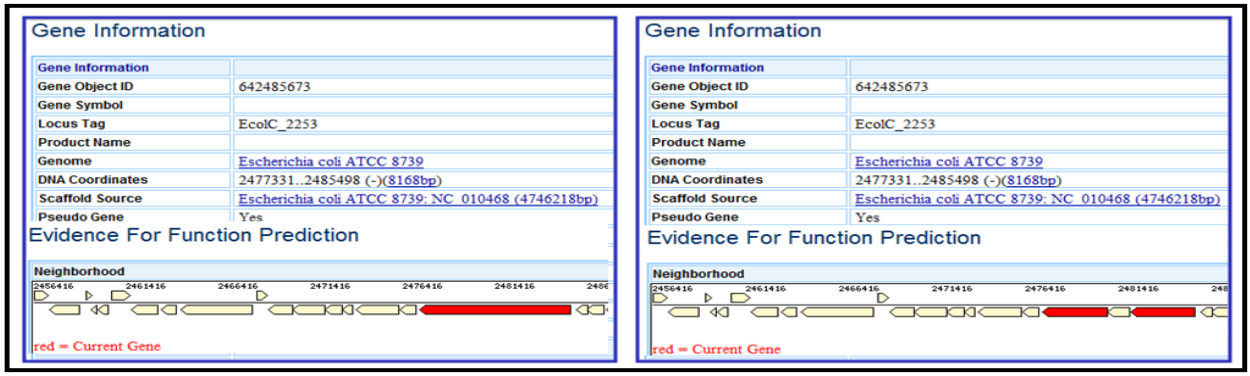


FIGURE 7. Gene Details – Pseudogenes.

Find Functions –*Function Search*

The **Function Search** has been extended with keyword search across all functional name filters, as illustrated in Figure 8(i).

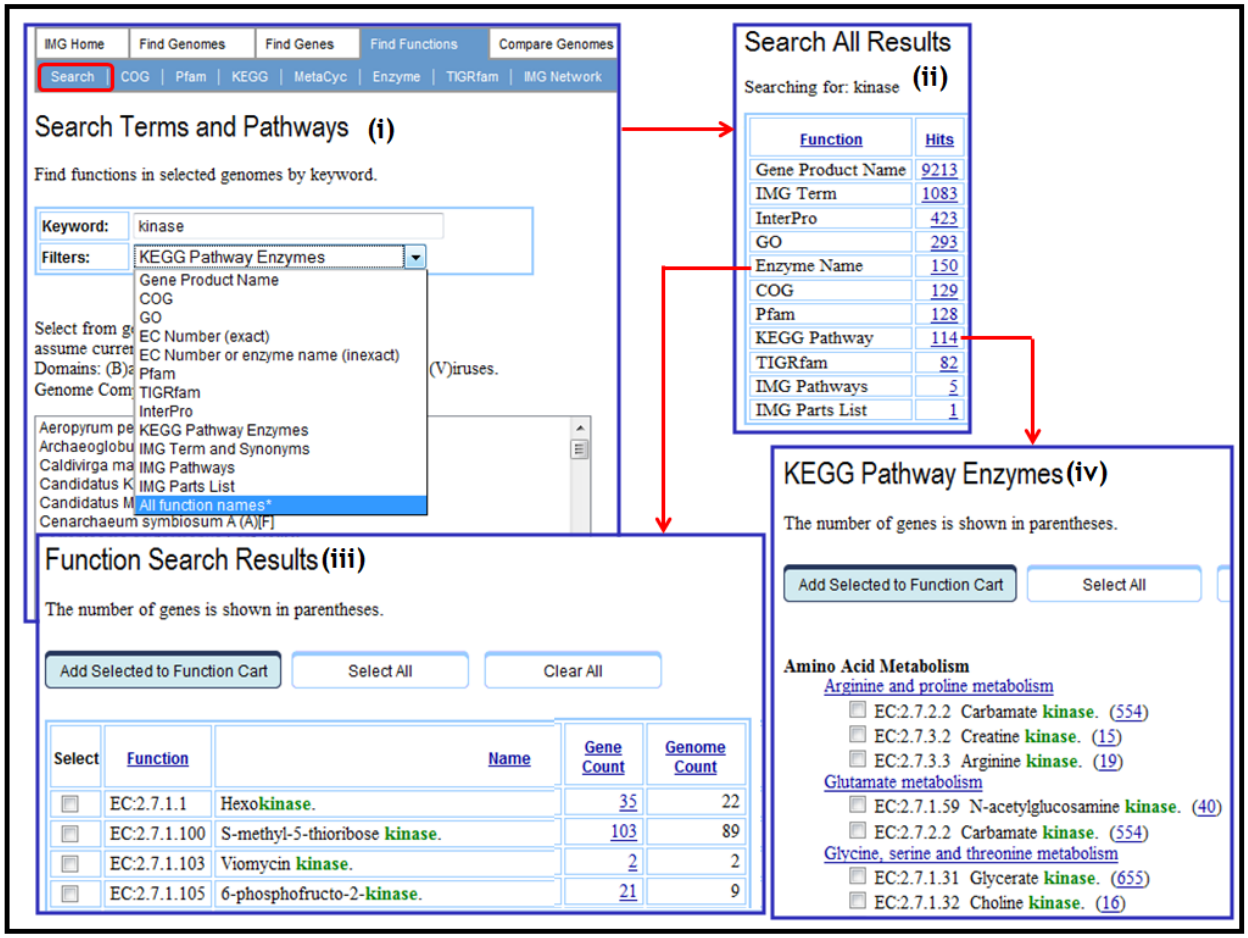


FIGURE 8. Find Functions – Function Search across all Function Name Filters.

The result of this search, shown in Figure 8(ii), lists the number of hits for each function name filter, which in turns leads to detailed results for each filter, as illustrated in Figures 7(iii) and 8(iv).

Find Genes – *Phylogenetic Profiler for Single Genes*

The **Phylogenetic Profiler for Single Genes** has been extended with a **Missing Gene** search capability provided on the **Phylogenetic Profiler for Single Genes Results** page, as illustrated in Figure 9(i).

Missing Gene search is useful for checking whether genes in a query genome (e.g., *Thermoplasma volcanium* in Figure 9) that have no homologs in other genomes (e.g., *Thermoplasma acidophilum* in Figure 9) are missing in these genomes. The search involves TBLASTn of the potentially missing gene against the sequence of these genomes, as illustrated in Figure 9(ii), where the missing gene is found in *Thermoplasma acidophilum.*

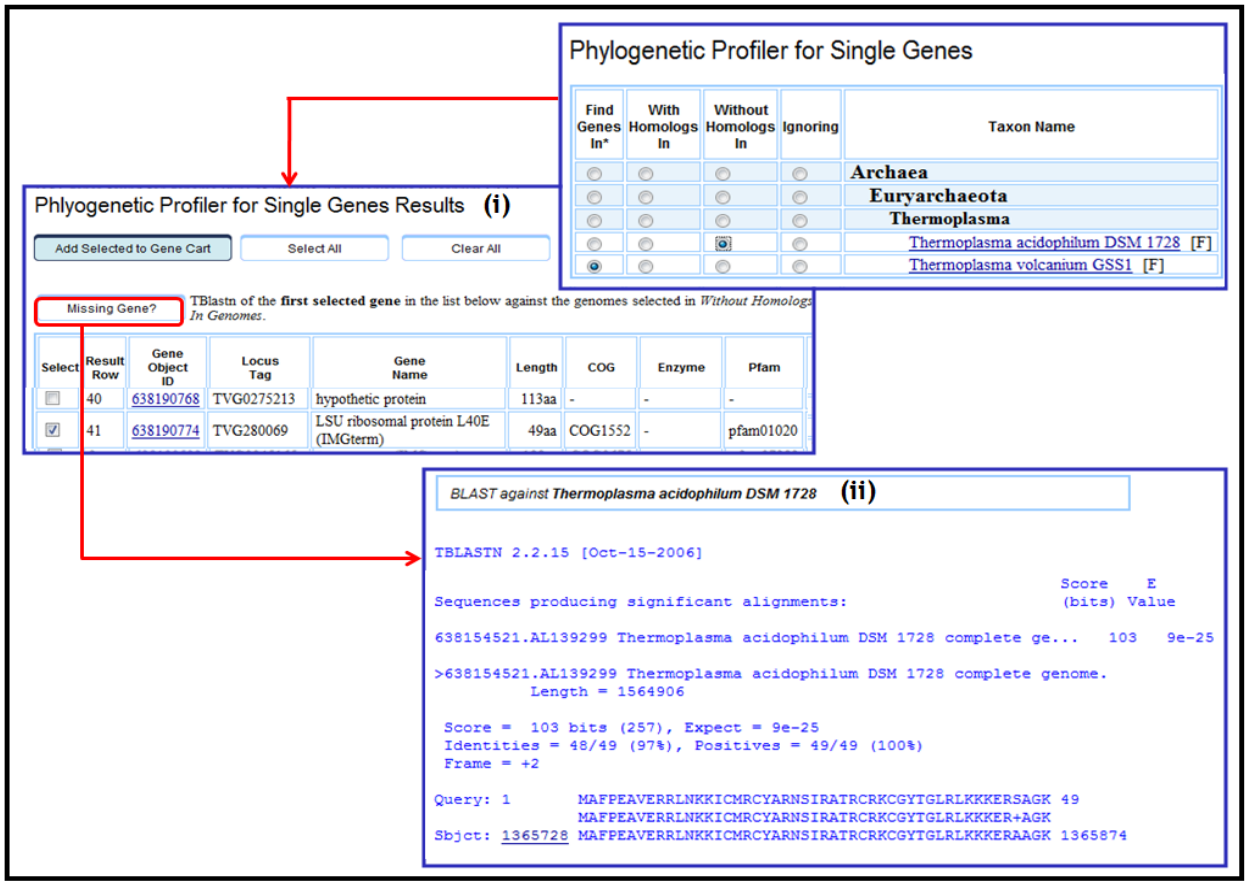


FIGURE 9. Find Genes – Phylogenetic Profiler for Single Genes.

Compare Genomes – *Summary Statistics*

The **Summary Statistics** table for comparing genomes selected using the **Genome Browser** includes links to **graphical viewers** for displaying the distribution of genes based on their COG and KEGG, Pfam, and TIGRfam associations, as illustrated in Figure 10.

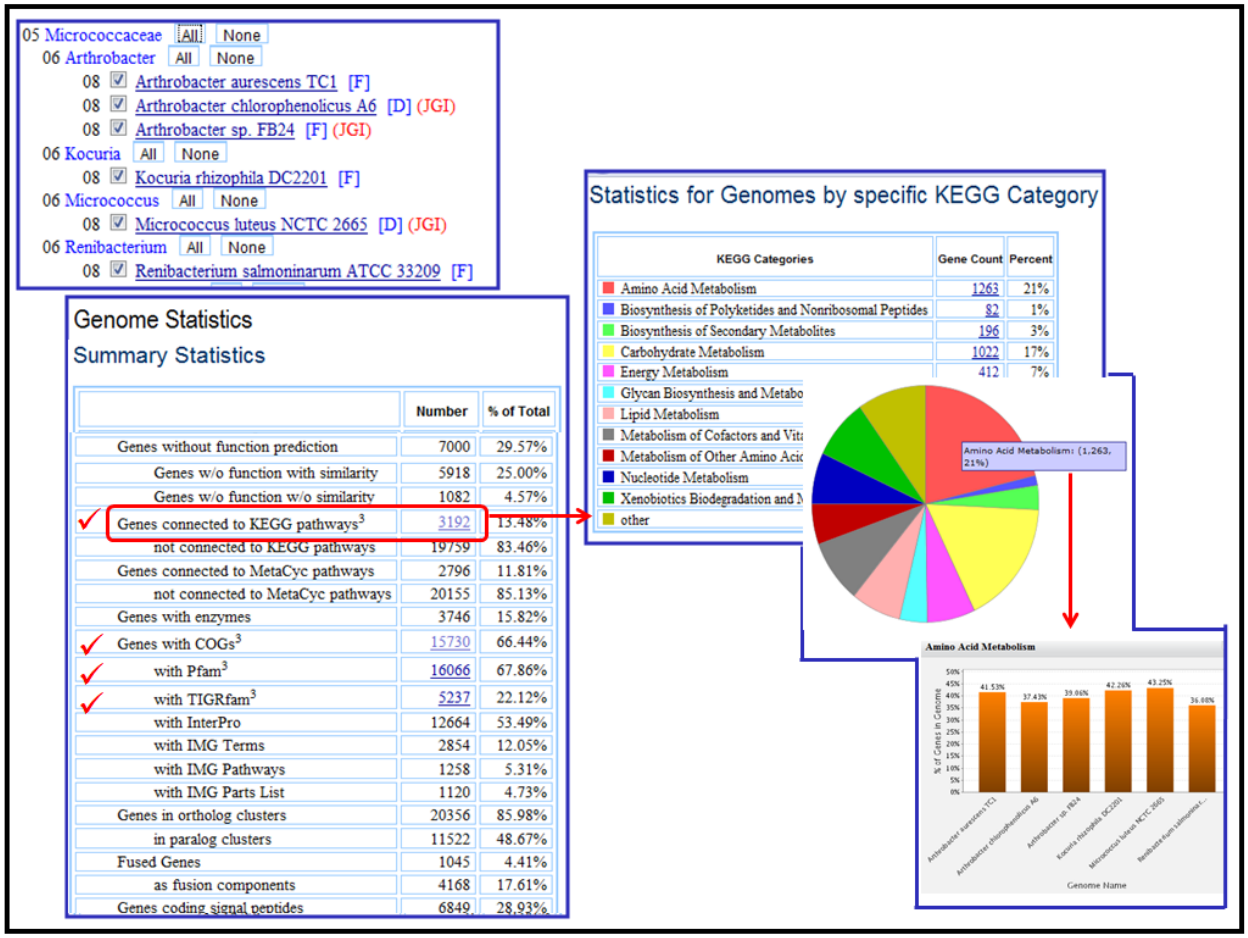


FIGURE 10. Compare Genomes – Summary Statistics.

For example, clicking on the number of genes connected to KEGG pathways, as shown in Figure 10, will display in a tabular and pie chart format the count of genes associated with each KEGG category across all selected genomes, as illustrated in the right pane of Figure 10. Clicking on a KEGG category on the *pie chart* or on the colored coded square for a KEGG category in the table will display a *bar chart* with the percent of genes for each genome associated with that KEGG category, as illustrated on the lower side left pane of Figure 10.