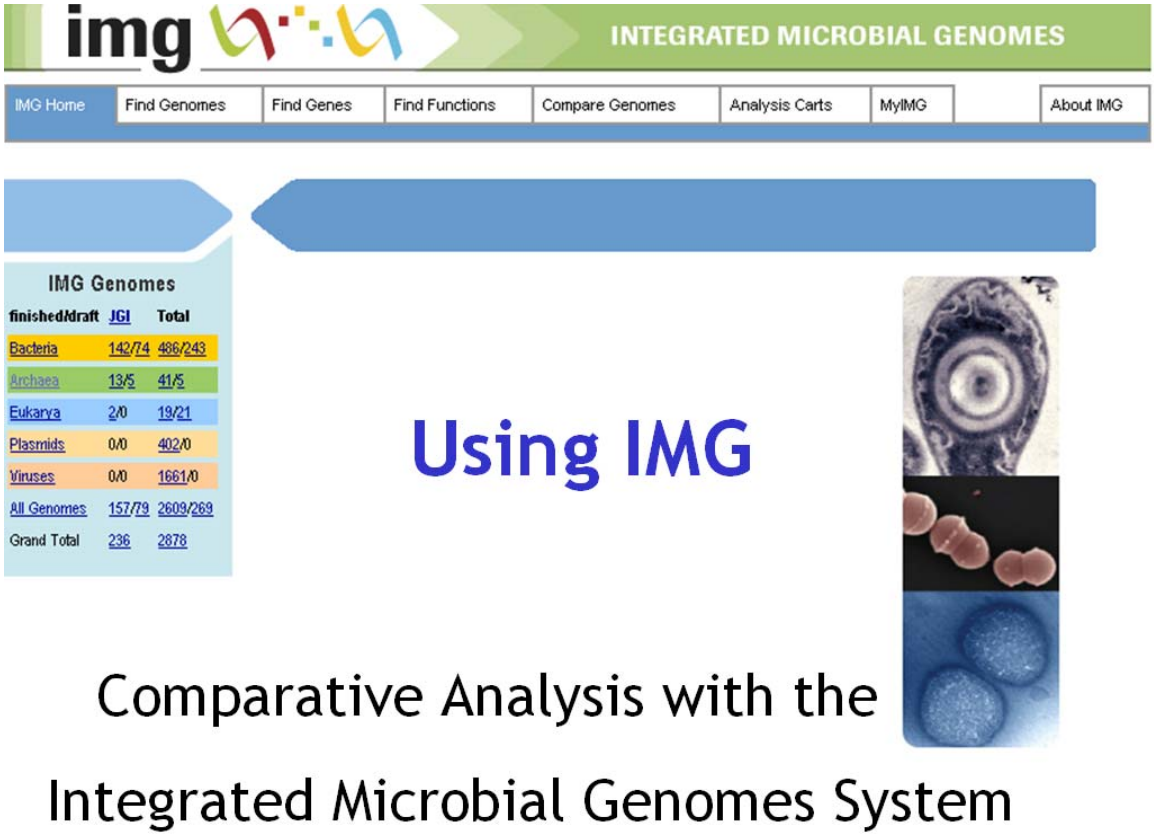


Technical Report LBNL-63614

Genome Biology Program
Department of Energy Joint Genome Institute

Biological Data Management and Technology Center
Lawrence Berkeley National Laboratory

December 1, 2007

Disclaimers and Copyright

NOTICE: Information from this server resides on a computer system funded by the U.S. Department of Energy. Anyone using this system consents to monitoring of this use by system or security personnel.

Disclaimer of Liability

With respect to documents available from this server, neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, including the warranties of merchantability and fitness for a particular purpose, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

Disclaimer of Endorsement

Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Copyright Status

Joint Genome Institute authored documents are sponsored by the U.S. Department of Energy under Contracts W-7405-Eng-48, DE-AC02-05CH11231, and W-7405-ENG-36. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce these documents, or allow others to do so, for U.S. Government purposes. All documents available from this server may be protected under the U.S. and Foreign Copyright Laws and permission to reproduce them may be required. The public may copy and use this information without charge, provided that this Notice and any statement of authorship are reproduced on all copies. JGI is not responsible for the contents of any off-site pages referenced.

December 1, 2007

This document was prepared by:

Victor M. Markowitz*
Natalia N. Ivanova**
Iain Anderson**
Athanasios Lykidis**
Konstantinos Mavromatis**
Ernest Szeto*
Krishna Palaniappan*
I-Min A. Chen*
Ken Chu*
Yuri Grechkin*
Nikos C. Kyrpides**

*Biological Data Management & Technology Center
Lawrence Berkeley National Laboratory

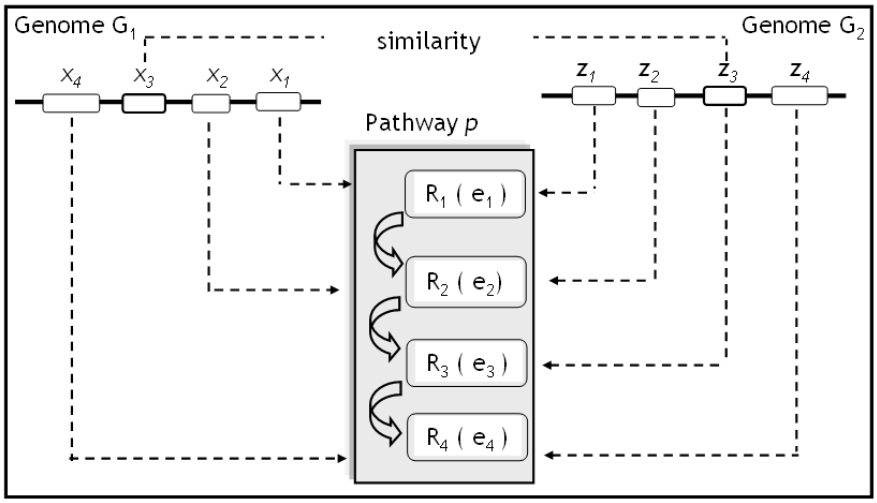
**Genome Biology Program (GBP)
Department of Energy Joint Genome Institute

Table of Contents

1 IMG Overview

1.1 Application View

Microbial ***genomes*** have ***genes*** that are characterized in terms of ***functions*** and (networks of) ***pathways***, as illustrated in the diagram below.



Genes (e.g., x_1, x_2, x_3, x_4 in the diagram above), usually identified using gene prediction methods, are located with start/end coordinates along the genome sequence (e.g., G_1), and are associated with functions (e.g., e_1, e_2, e_3, e_4) and pathways (e.g., *p*) using functional characterization methods. Genes are related to each other in terms of their sequence similarity (e.g., x_3 and z_3) or associations with functions and pathways (e.g., x_1 and z_1). A genome (e.g., G_1) is associated with a specific function (e.g., e_1) or pathway if its genome has a gene (e.g., x_1) that is associated with this function or pathway.

A typical microbial genome data analysis involves comparing ***profiles*** of subsets of objects of a specific type (e.g., genomes, genes, functions) across objects of another type. Profiles can be viewed as two-dimensional matrices, where each cell in the matrix is associated with objects x_i and y_j and represents a set of genes associated with these objects. For example, the profile of a function x_i across genomes y_1 to y_n consists of sets of y_j genes that are characterized by x_i. Similarly, the profile for a gene x_i across genomes y_1 to y_n consists of sets of y_j genes that are associated with x_i, where the association of y_j genes with x_i is based on a specific sequence similarity.

The count of genes in each set of genes, k_{ij}, in a profile is called *gene abundance*, and a profile that consists of gene abundance counts are called *abundance profiles*. *Presence* (or occurrence) *profiles* are a special case of abundance profiles, where each gene count is represented by either "a" (absent) if the corresponding gene abundance is zero, or by "p" (present) if otherwise.

The following example shows the abundance profiles for genes (or functions) x_1 to x_4 across genomes y_1 to y_8.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
x_1	*2*	*1*	*1*	*3*	*0*	*0*	*1*	*0*
x_2	*1*	*1*	*2*	*2*	*0*	*0*	*1*	*0*
x_3	*0*	*1*	*1*	*0*	*0*	*0*	*0*	*0*
x_4	*1*	*1*	*1*	*1*	*2*	*1*	*2*	*1*

Examining the profiles of the genes (or functions) of a specific genome, y, in the context of other related genomes allows determining what y may have in "common" with y_1,…, y_k. For the example above, genome y_1 has gene (or function) x_4 in "common" with genomes y_1 to y_8; and genes (or functions) x_1 and x_2 have the same presence profile across genomes y_1 to y_8.

1.2 IMG Data Model

The data model underlying IMG allows recording microbial and selected eukaryotic genomic data. The main data types modeled are: (a) primary genomic sequence information, (b) computationally predicted and curated gene models, (c) pre-computed gene relationships, and (d) functional annotations and pathway information.

Genomes are identified in IMG using an internally generated unique object identifier. In addition, individual genomes are associated with the NCBI Genomes Project Identifier and taxonomic lineage via NCBI's Taxonomy (domain, phylum, class, order, family, genus, species, and strain). For every genome, IMG incorporates its primary genome sequence information recorded in RefSeq (1), including its organization into chromosomal replicons (for finished genomes) and scaffolds and/or contigs (for draft genomes), cross-referenced with their RefSeq Accession Identifiers.

Genes are identified in IMG using an internally generated unique object identifier. IMG also employs RefSeq's Gene Identifiers to link to other NCBI resources, such as Entrez Gene [2]. Genes in IMG model predicted coding sequences (CDSs) and functional RNAs, and are recorded with start/end coordinates. Genes have attributes containing data on their protein product name, functional role and participation in pathways, and relationship (sequence similarity) to other genes.

Functional roles are assigned to genes based on different controlled vocabularies, such as Enzyme Nomenclature (3), COG clusters (4), Pfam (5), TIGRfam (6), InterPro (7), Kegg Ortholog (KO) terms (8), and Gene Ontology (GO) terms (9). Functional roles are further defined by their association with functional classifications including COG functional categories, TIGR role categories, and the KEGG pathway collection. The assignment of specific functional roles is further discussed below in the Functional Annotations section.

In order to address problems with the inconsistencies of the protein product names as well as with the current functional classifications (10), genes are further annotated in IMG using a native collection of generic (protein cluster–independent, i.e., not sequence similarity–based protein clusters) functional roles called **IMG terms** that are further defined by their association with generic (organism-independent) functional hierarchies, called **IMG pathway**. IMG terms and pathways are currently specified by domain

experts at DOE-JGI as part of the process of annotating specific genomes of interest, and are subsequently propagated throughout the system.

IMG Terms form a hierarchy, whereby the leaves of this hierarchy consist of functional roles for gene products (protein product descriptions) assigned to individual genes. These lower-level IMG Terms of type “Gene Product” can be directly associated with reactions, whereby they function as either “Catalysts” or “Reactants”. Alternatively, they can be assigned recursively as “children” of IMG Terms of type “Protein Complex”, thus indicating that they constitute subunits of a multi-subunit protein complex. A detailed discussion of the rationale for IMG terms and pathways and their specification is available at http://img.jgi.doe.gov/pub/doc/imgterms.html, as part of IMG’s online documentation. Note that, despite somewhat similar nomenclature, IMG Terms are not equivalent to GO terms . A mapping of IMG terms to GO terms is currently developed by the GO consortium in collaboration with DOE-JGI scientists.

1.3 IMG Data Content

1.3.1 Genomic Data

Genomic data in IMG consists of publicly available genomic sequence data integrated with JGI sequence data for microbes and selected model eukaryotes. All JGI microbial genomes are sequenced and assembled at the JGI Production Genomics Facility, with subsequent finishing at various JGI partner or collaborator institutions. Automated annotation for archaeal and bacterial genomes sequenced at JGI is provided by the Genome Analysis Pipeline at Oak Ridge National Laboratory.

NCBI’s RefSeq is IMG’s primary data source for publicly available finished and draft microbial, viral (including phages), and isolate (i.e., not already part of a sequenced microbe) plasmid genomes. RefSeq contains curated versions of entries in the Genbank nucleotide sequence database representing the complete sequences of chromosomes and plasmids. RefSeq is updated regularly, thus ensuring continuous improvement and standardization of gene annotations.

IMG currently contains genomic data for several lower eukaryotes (fungi, protozoa). Primary genomic sequence data are collected from NCBI's RefSeq. For higher level eukaryotes (such as human, mouse, fly, etc.), RefSeq and Entrez/Gene serve as the primary data sources mainly for their currency, consistency, and uniformity of cross-references with functional resources, such as UniProt, InterPro, KEGG, and model organism databases.

1.3.2 Functional Annotations

Protein product names are available from RefSeq and typically consist of the function prediction provided by sequence genome centers.

Genes are associated with **COG**s using RPS-BLAST (Reverse Position Specific BLAST) computations against NCBI's Conserved Domain Database (CDD) (11).

Genes are associated with **Pfam** and **TIGRfam** using hmmsearch from the HMMER package (http://hmmer.janelia.org/). A BLAST pre-filter is used for quickly narrowing

down the candidate HMM models. BLAST is run with a nonstringent e-value cutoff in order to pick up subsequences from the seed sequences of an HMM model that could serve as candidates for full HMM scoring. Low-complexity masking is turned off. For this step, the BLAST e-value cutoff is set to 10,000 for Pfam, and 1,000 for TIGRfam seed-sequence databases, respectively. hmmsearch is then applied to the candidate models with a per family noise cutoff (--cut_nc). The scoring for domain-level hits are recorded for Pfam, while scoring for the full model is recorded for TIGRfam.

EC numbers are computed using RPS-BLAST against the PRIAM database (12), as a complement to the (often sparse) native EC numbers collected via RefSeq.

UniProt (13) is used to associate genes with additional annotations, such as InterPro, TIGRfam, and GO terms, while KEGG is used to establish KO term associations. In addition, CRISPR repeats (14), signal peptides using SignalP (15), and transmembrane helices using TMHMM (16) are computed, and potentially missing data from the original RefSeq data files (such as various RNAs) are added. RNA gene models are synchronized with Rfam (17).

1.3.3 Gene Relationships

Sequence similarities for identifying **candidate homologs** are computed using NCBI BLASTp with 1e-2 E-value cutoff, and low-complexity soft masking (-F 'm S') turned on. IMG provides support for filtering candidate homolog lists by percent identity, bit score, and more stringent E-values, as well as with a variety of metadata such as phenotype, habitat, etc.

Homolog, paralog, and ortholog relationships are established through BLAST hits between genes, with genes in all genomes compared to genes in all other genomes ("all vs. all"). The computations are carried out incrementally for different versions of IMG.

Orthologous pairwise relationships are computed as bidirectional best hits between genomes. **Paralogous** pairwise relationships are computed as reciprocal hits within the same genome. **Homologs** are unidirectional hits. The E-value of 10^{-2} or better is used for these pairwise relationships. Additional filtering by percent identity, bit score, and more stringent E-values can be applied as needed through the Web UI application filtering or database queries.

Orthologous groups are formed with the Markov Cluster Algorithm (MCL). A *conservation score* is calculated to normalize the strength of similarity, by dividing the bit score between two sequences by the bit score of the sequences when BLASTed against itself (self bit score): $cons\_score_{xy} = bit\_score_{xy} / max(bit\_score_{xx}, bit\_score_{yy})$, where x and y are two separate sequences. MCL is applied with default parameters. **Paralog groups** are formed by using the same procedure.

Fusions (**fused genes)** in a specific genome are defined as genes that are formed via the composition (fusion) of two or more previously separate genes (component genes) from other genomes. The identification of fusions is based on BLAST similarities between genes. In IMG, only genes from **finished** genomes are considered as putative components, in order to avoid false predictions from fragmented genes in draft genomes. Furthermore, genes that frequently appear fragmented in finished genomes, such as

transposases and *integrases*, as well as pseudogenes, are excluded from fusion calculations.

Fusions are computed as follows:

1. Starting from a *candidate fused gene*, *x*, in a given genome, *G*, similarities to all other genes in other genomes, G_i, are examined: for each genome G_i, *candidate component genes* are identified from G_i genes that align to more than 80% of gene *x*. Candidate component genes are kept only if they overlap less than 10% of the shortest candidate component gene. Additionally, candidate components should not be paralogs.
2. For each *candidate fused gene*, *x*, in a given genome, *G*, that has *candidate component genes* in genomes, G_i, *x* is *accepted* as a *valid fusion* only if (i) the candidate components found in each genome G_i cover more than 80% of *x*; (ii) the same combination of component genes is found in at least two other genomes, and in at least one of these genomes the components are *not in tandem*, i.e., in at least one genome one or more genes are found between the components. The second condition eliminates cases of consistent frameshifts in a group of genomes.

1.3.4 Parameters

The following table summarizes the parameters used for the BLAST computations.

Computation	Tool	Maximum E-value	Minimum %Identity	Low-Complexity Filtering	Top Hits Limit	Notes
COG	rpsblast	1e-2	None	None	1	Top hit only.
PRIAM	rpsblast	1e-10	45%	Soft Masking (-F 'm S')	1	Top hit only. ≥ 70% alignment length on query gene and PRIAM sequence.
Protein all. vs. all	blastall -p blastp	1e-2	None	Soft Masking (-F 'm S')	2500	These similarities are employed by various tools, such as the Phylo Profiler, which provide additional filtering options.
RNA all vs. all	blastall -p blastn	1e-5	None	None	250	≥ 50% alignment over length of both sequences.

1.4 IMG Data Analysis

Genome data analysis in IMG consists of operations involving **genomes**, **genes**, and **functions** that can first be ***selected*** and then ***examined*** individually. **Genomes** also can be ***compared*** in terms of various statistics, gene content, function capabilities, and sequence conservation. **Genes** and **functions** also can be compared using a variety of comparative tools. The tools provided by IMG are outlined in Figure 1.1.

In order to perform comparative analysis in IMG, **genomes**, **genes** or **functions** are first **selected** using browsers or search tools. **Browsers** are provided for selecting genomes and functions, organized as alphabetical lists or hierarchically (e.g., based on a phylogenetic tree for genomes). **Keyword search** tools allow identifying genomes, genes, and functions of interest using a variety of keyword filters. Genomes also can be selected using a search tool that allows specifying conditions involving phenotype, habitat, disease, and relevance metadata fields, while genes also can be selected using BLAST search tools against various datasets. The genomes that result from search operations are displayed as a list from which they can be selected and saved for further analysis. In a similar manner, the genes and functions that result from search operations are displayed as lists from which genes and functions can be added to the **Gene Cart** and **Function Cart**, respectively.

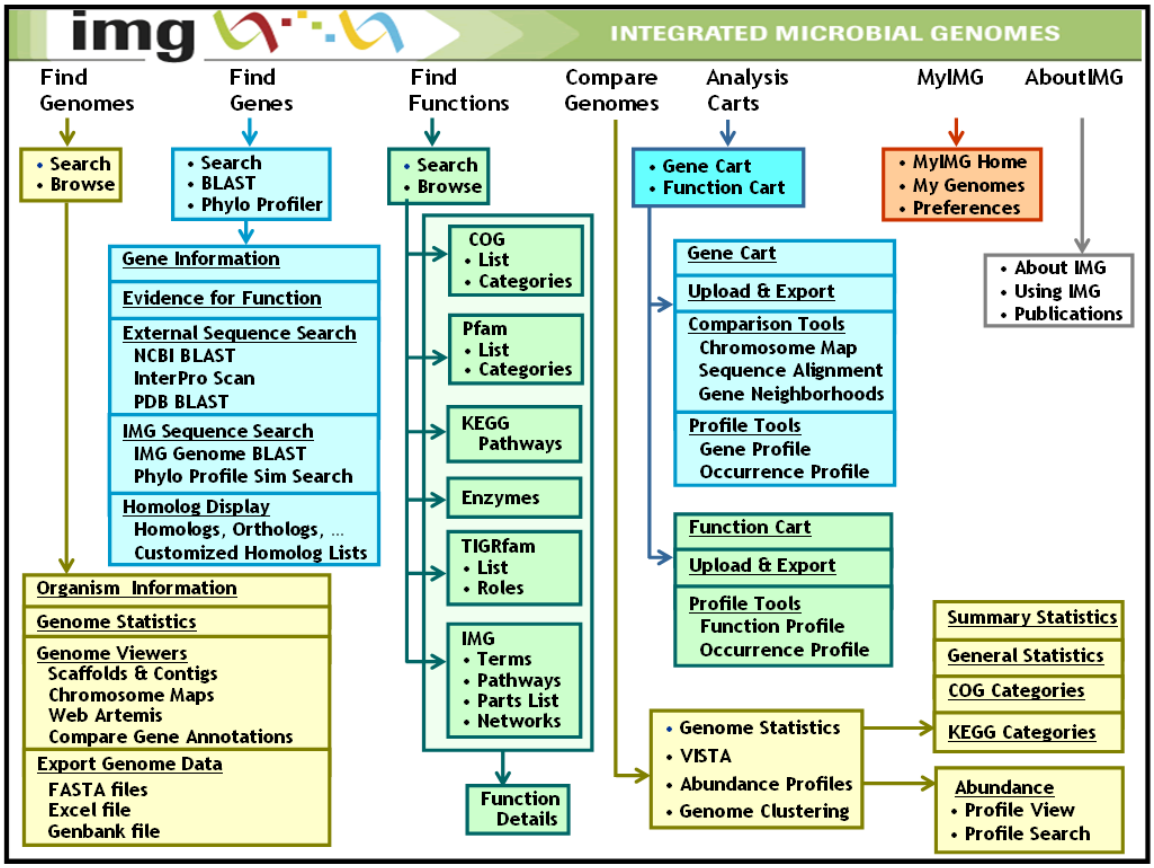


FIGURE 1.1. IMG user interface map.

Individual genomes can be examined using the **Organism Details** page, which includes information on the organism together with various genome statistics of interest, such as the number of genes that are associated with KEGG, COG, Pfam, InterPro, or enzyme information. For each genome, one can also examine the associated list of scaffolds and contigs using the **Chromosome Viewer**, or can generate circular chromosomal maps on which a variety of data can be projected.

Individual genes can be examined using the **Gene Details** page, which includes Gene Information, Protein Information, and Pathway Information tables; evidence for functional prediction; and COG, Pfam, and pre-computed homologs. A gene can be examined in the context of its location on the chromosome using the **Chromosome Viewer**.

Individual functional groups, such as COG categories, can be further examined using summary pages, such as the **COG Category Details** page, which lists the COGs of a given category and the number of organisms that have genes belonging to each COG, where the "organism counts" are linked to a list of organisms and their associated "gene counts."

Comparative analysis of genomes is provided in IMG through a number of tools that allow genomes to be compared in terms of various statistics, gene content, function capabilities, and sequence conservation.

Genome Statistics provides statistics across the genomes that have been previously selected and saved as discussed above. The display can be configured by including a variety of genome attributes, such as GC content, number of protein coding genes, and various functional annotations.

Genomes can be compared in terms of functional capabilities using the **Abundance Profile Search** tool, which allows defining a *profile* for functions (COGs, Pfams) in a query genome in terms of their abundance compared to other related genomes.

The functional capabilities of genomes also can be compared using a number of additional **functional profile** tools. First, functions of interest, such as protein families, enzymes, and IMG terms, are selected with the **Function Cart**. For these functions a profile across genomes can be computed, with the results displayed in a tabular format. Each cell in the profile result table displays the count (*abundance*) of genes in an organism and contains a link to the associated list of genes. The genes associated with a specific function can be saved using the **Gene Cart**, and further examined using various tools, such as gene neighborhood analysis and multiple sequence alignment tools.

Another functional profile tool, the **Abundance Profile Viewer**, provides an ***overview*** of the relative abundance of protein families (COGs and Pfams) and functional families (Enzymes) across selected genomes, with abundance of protein/functional families displayed as a heat map. **Comparative analysis of genes** includes gene neighborhood analysis, phylogenetic occurrence profile analysis, and multiple sequence alignment, which can be applied to genes collected into the **Gene Cart**.

Finally, DNA conservation can be explored for closely related organisms in IMG using the **VISTA** comparative genome analysis tools. Selecting an organism from a predefined list invokes the VISTA browser that can be then used for examining conservation.

MyIMG allows users to set system-wide preferences, such as the maximum number of rows displayed for lists of homologs, and to upload their genome selections.

Users who are not familiar with IMG and its analytical tools can start by reviewing the following documents:

- An introduction to IMG provided by the OpenHelix IMG tutorial available at: http://www.openhelix.com/downloads/img/img_home.shtml
- Details on various IMG system components and content available at: http://img.jgi.doe.gov/pub/doc/about_index.html

References

1. Pruitt, K.D., Tatusova, T., Maglott, D.R. (2007) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acid Research* **35**, D61-D65.
2. Maglott, D.R., Ostell, J., Pruitt, K.D., Tatusova, T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acid Research* **35**, D26-D31.
3. Bairoch A. (2000). The ENZYME database in 2000. *Nucleic Acids Research* **28**, 304-305.
4. Tatusov, R.L., Koonin, E.V., and Lipman, D.J., A. (1997) Genomic Perspective on Protein Families, *Science* **278**, 631-637.
5. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. (2004) The Pfam Protein Families Database. *Nucleic Acids Research* **32**, D138-D141.
6. Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R., White O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes *Nucleic Acids Research* **35**, D260-D264.
7. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al. (2005) InterPro, Progress and Status in 2005. *Nucleic Acids Research* **33**, D201-D205.
8. Kanehisa, M., Goto, S., Kawashima, S. Okuno, Y., and Hattori, M. (2004) The KEGG Resource for Deciphering the Genome. *Nucleic Acids Research* **32**, D277-D280.
9. Gene Ontology Consortium. (2004) The Gene Ontology Database and Informatics Resource. *Nucleic Acids Research* **32**, 258-261.
10. Ivanova, N.N., Anderson I., Lykidis A., Mavrommatis K., Mikhailova, N., Chen, I.A., Szeto, E., Palaniappan, K., Markowitz, V.M., Kyrpides N.C. (2007) Metabolic Reconstruction of Microbial Genomes and Microbial Community Metagenomes. *Technical Report 62292*, Lawrence Berkeley National Laboratory.
11. Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., Bryant, S.H. (2002) CDD: A Database of Conserved Domain Alignments with Links to Domain Three-Dimensional Structure. *Nucleic Acids Research* **30** (1), 281-283.

12. Claudel-Renard, C., Chevalet, C., Faraut, T., Daniel Kahn, D. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Research* **31** (22), 6633-6639.
13. The UniProt Consortium. (2007) The universal protein resource (UniProt). *Nucleic Acids Research* **35**, D193-D197.
14. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., Hugenholtz, P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
15. Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. *Nature Protocols* **2**, 953-971.
16. Moller, S., Croning, M.D.R., Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. (2001) *Bioinformatics*, **17**(7), 646-653.
17. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Research* **31** (1), 439-441.

2 Find and Examine Genomes

Genomes in IMG can be selected by browsing through the list of genomes in IMG with the **Genome Browser** or by using **Genome Search**. Individual genomes can be examined using **Organism Details**.

2.1 Genome Browser

You can select the **Genome Browser** via the second-level menu under **Find Genomes** or via the summaries on the **IMG home page,** as shown in Figure 2.1(i). The genome summaries on the IMG home page provide links to subgroups of genomes that can be viewed in the **Genome Browser**. Click on an underlined value to load the list of genomes it represents. For example, you can quickly retrieve a list of all the archaeal genomes by clicking on Archaeal label in the summary table.

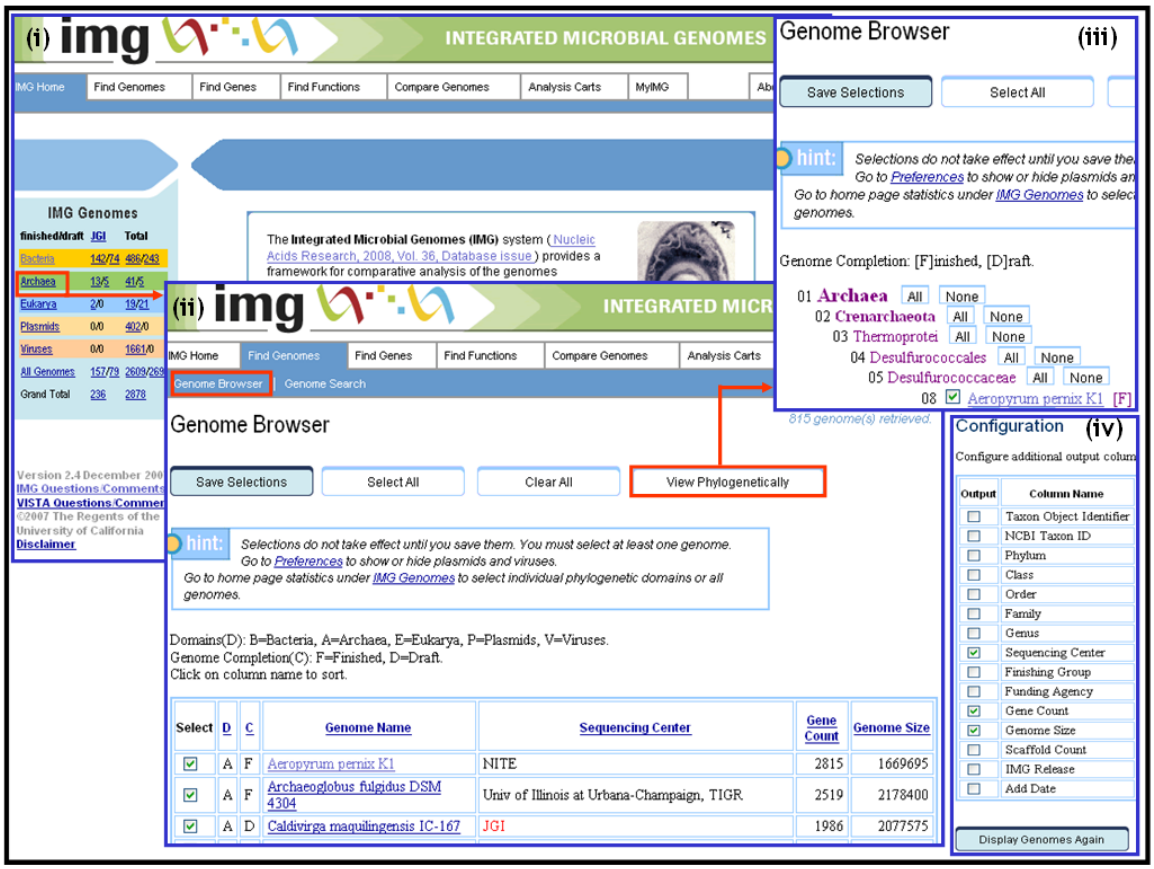


FIGURE 2.1. IMG home page with genome summaries and two Genome Browser views of the list of genomes.

The **Genome Browser** allows you to select genomes from a list of genomes organized ***alphabetically***, as shown in Figure 2.1(ii), or as a ***phylogenetic tree***, as shown in Figure 2.1(iii). Genomes can be selected or de-selected individually or collectively. You can click on the name of an individual genome to view the associated **Organism Details**, which is described later in this manual. The alphabetical list of genomes includes

information on completion status, genome size, and sequencing center. With the phylogenetic tree view, you can select or deselect all of the genomes in a particular taxonomic group.

You can sort the alphabetical list of genomes by clicking on a column name. Alphabetical columns are usually sorted in ascending order. Numeric columns are usually sorted in descending order (most to least significant).

You can configure the columns in the alphabetical list of genomes with the **Configuration** selector at the bottom of the **Gene Browser** page, shown in Figure 2.1(iv). You can add a column by first clicking the Output box next to the column name in the Configuration selector, and then selecting the "Display Genomes Again" button.

You can ***save*** genomes selected with the **Genome Browser** by clicking the checkmark box next to each genome, and then clicking on "Save Selections." You also have the option to "Select All" of the genomes or "Clear All" of the selections.

The genome selections you save define the set of genomes for the IMG analysis tools such as **Gene Search**, **Gene Ortholog Neighborhoods**, **Phylogenetic Profiler**, **Genome Statistics**, and for highlighting in ortholog and homolog lists. These tools will be applied only on this set of saved genomes, unless you *override* this selection via the list of genomes provided as part of some of the IMG tools.

The box in the upper-right-hand corner of the browser window displays how many genomes are selected for the current analysis tool.

2.2 Genome Search

Genomes can be searched via implicit attributes using the **Quick Genome Search** box at the top-right corner of every IMG page, as shown in Figure 2.2(i).

A detailed search of genomes can be carried out using **Genome Search** on the second-level menu of **Find Genomes**. **Genome Search** allows finding genomes by using ***name*** or ***metadata*** as search filters, as shown in Figure 2.2(i), or by selecting specific ***Phenotype***, ***Habitat***, ***Disease***, and ***Relevance*** metadata values, as shown in Figure 2.2(ii). Keywords or substrings are used to search names and descriptions. Genome identifier (ID) values should be exact. Searchable fields are available in the pull-down filter list, including "Genome Name," "Taxon ID," "Sequencing Status," "Sequencing Center," "Domain," "Phylum," and "Funding Agency."

As an example, in order to perform a general search of all genomes added in a particular IMG release, such as 1.2, select the "IMG Release" filter, type "1.2" in the keyword box, and click on the "Go" button. To view all genomes sequenced at JGI, select "Sequencing Center" filter, type "JGI" in the keyword box, and click on the "Go" button.

Genomes also can be searched using specific ***Phenotype***, ***Habitat***, ***Disease***, and ***Relevance*** metadata values. Click on one or more values for each category from each metadata attribute list of values. Metadata attributes without a selected value will be ignored in the search. A logical "OR" is used when searching multiple values selected for a single metadata attribute. A logical "AND" is used when searching for values across multiple metadata attributes. Once the selection is completed, click on the "Go" button to

see the search results, as illustrated in Figure 2.2(iii), which shows the result of searching for genomes with phenotype “Acidophile” and habitat “Hydrothermal vent”.

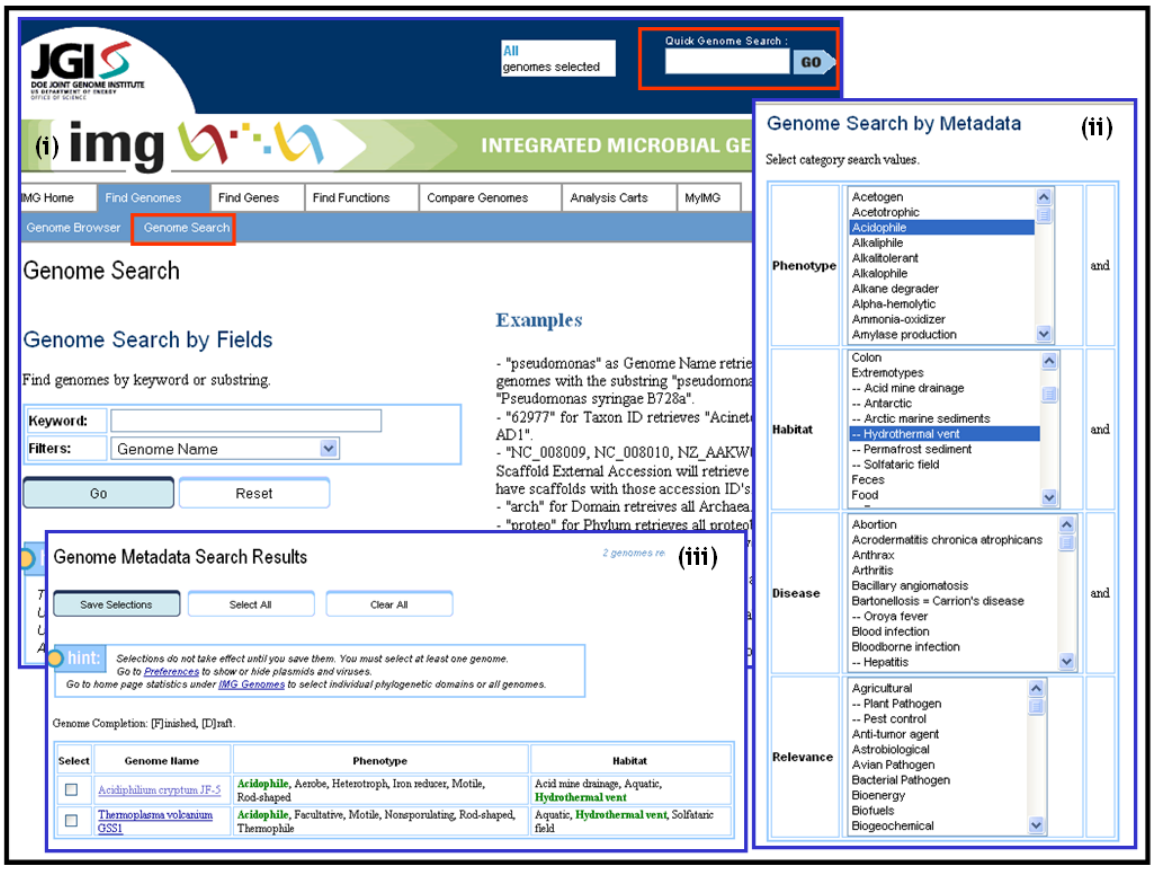


FIGURE 2.2. Genome Search: quick search, keyword search, and metadata attribute search.

The result of **Genome Search** is a **list of genomes**, such as that shown in Figure 2.2(iii), which can be examined individually, or selected and saved for reducing the genome context for future analysis.

2.3 Organism Details

Genomes can be explored using **Organism Details** pages that can be accessed by clicking on a genome name in the list of genomes in the **Genome Browser**, the list of genomes returned by a Genome Search, or anywhere else a genome name is provided in IMG. The **Organism Details** page contains four sections, as shown in Figure 2.3, which are further discussed below.

2.3.1 Organism Information

The Organism Information section contains taxonomic, sequencing, and metadata (e.g., phenotype, habitat, disease, relevance) information on the organism, as shown in Figure 2.3(i). Links to a variety of external resources, such as GOLD, Genbank, Refseq, Pubmed, are also provided.

2.3.2 Genome Statistics

The **Genome Statistics** section provides information on the DNA sequence (GC content, number of bases and scaffolds), and statistics regarding the functional annotation of the genes, as shown in Figure 2.3(ii). The following statistics are provided for the genes of the genome:

1. Count of genes for each type of gene, such as various RNA genes, genes associated with a product name, protein coding genes, pseudogenes;
2. Count of genes associated with each protein family, functional category, and pathway collection available in IMG, such as COG, Pfam, TIGRfam, KEGG, IMG terms, and IMG pathways; the number of genes that are not associated with each protein family (functional category and pathway collection are also provided);
3. Count of genes associated with various gene clusters, such as ortholog and paralog clusters computed in IMG.

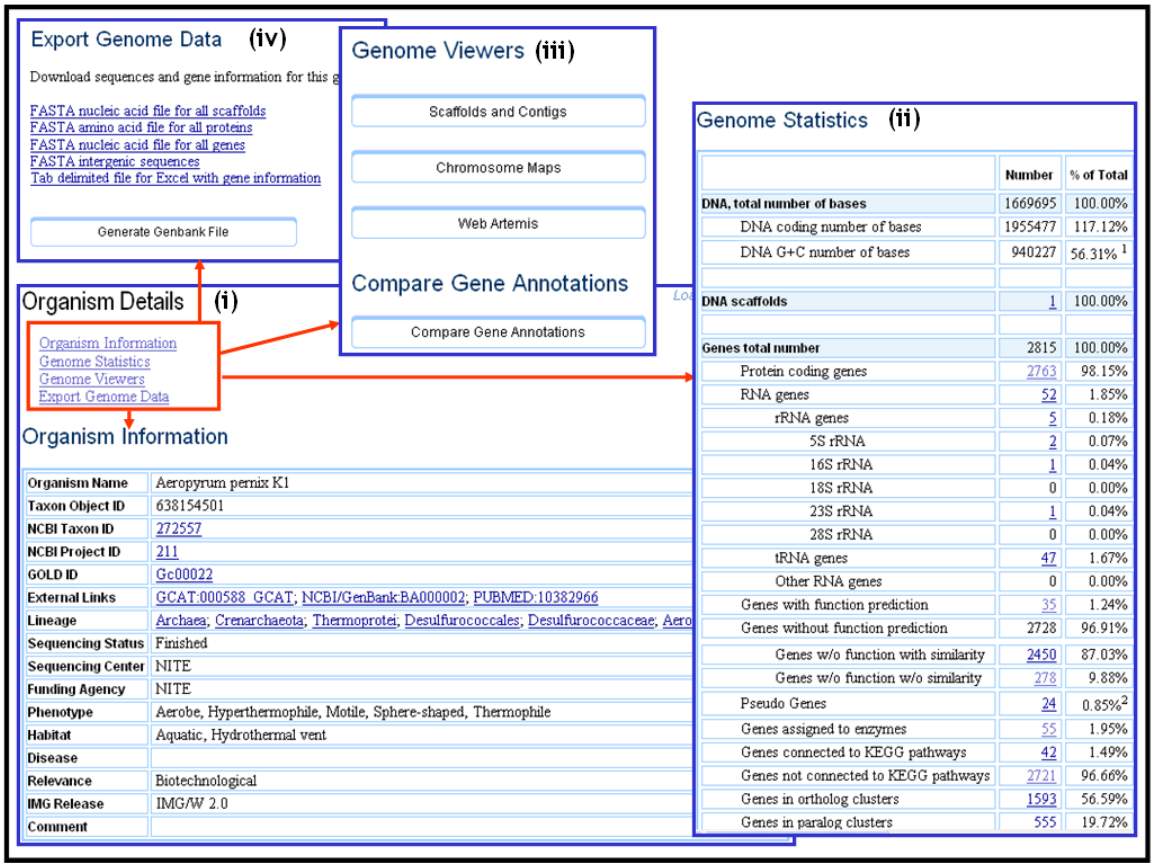


FIGURE 2.3. Organism Details: Organism Information, Genome Statistics, Exploration Tools, Export Genome Data.

Every count of genes is linked to the list of associated genes grouped into categories that are specific to each functional classification. For example, when you click on the count of genes associated with KEGG pathways, first a list of KEGG pathways will be displayed

together with the count of genes for each pathway. This count of genes provides a link to the list of genes associated with a specific pathway, as discussed below (see Section 3.4 (Gene Selection from Genome Statistics) and Figure 3.3). From the list of genes, you can click a "Gene Object ID" to display details about that gene, including COG, Pfam, and InterPro data, and links to maps of the KEGG pathways in which the gene product participates. You can also click the checkbox for a gene to select it and add it to the Gene Cart.

2.3.3 Genome Viewers

The **Genome Viewers** section provides tools for examining the genome and its annotations, as shown in Figure 2.3(iii).

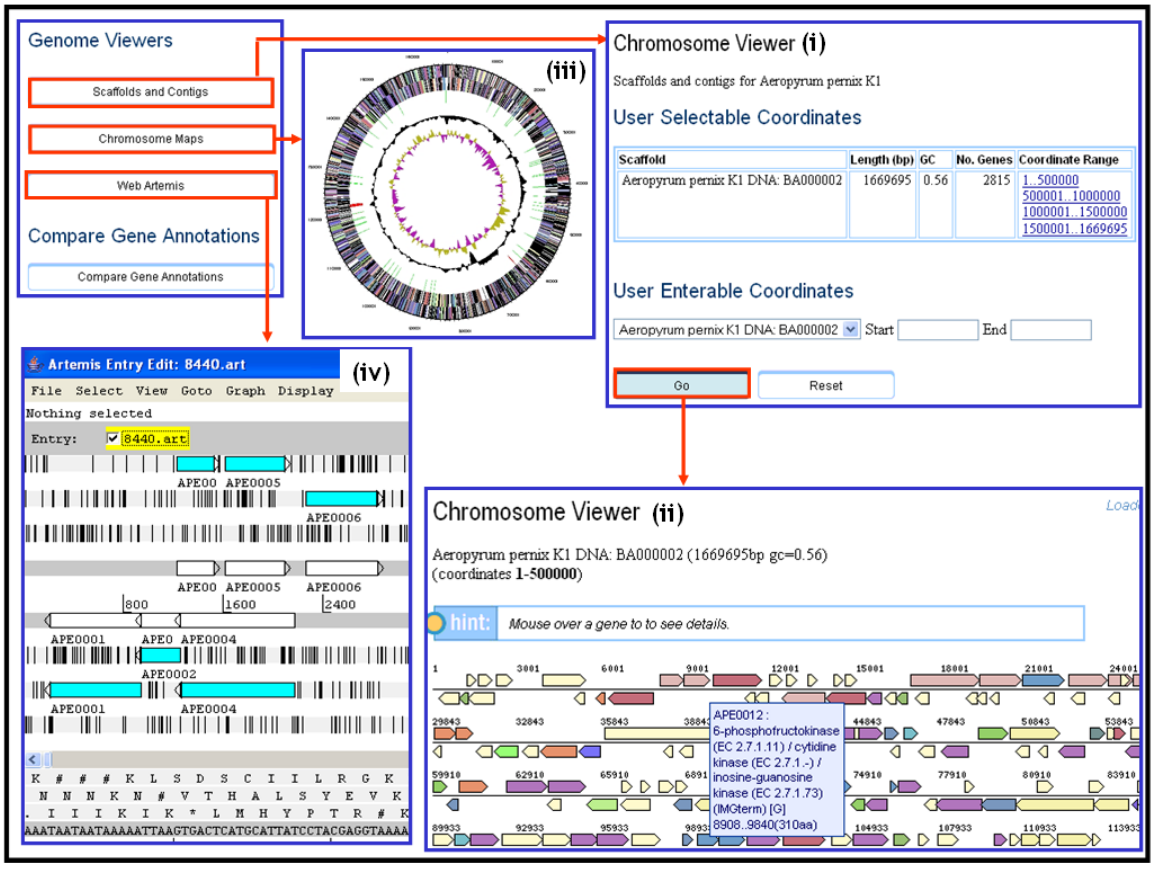


FIGURE 2.4. Organism Details: Genome Viewers.

Scaffold and Contig Viewer. Scaffolds and contigs can be viewed using a **Chromosome Viewer,** as shown in Figure 2.4. First, a list of scaffolds and contigs is displayed together with the length in base pairs, and the number of genes in each scaffold or contig, as shown in Figure 2.4(i). Predefined links to coordinate ranges are also displayed, and can be selected to view a specific coordinate range. You can enter a larger range than presented by the links. Be careful not to overwhelm the resources of your browser when selecting a large range.

The **Chromosome Viewer** displays genes within a given coordinate range, as shown in Figure 2.4(ii). The genes are colored by high-level COG function categories. You can move the next or previous coordinate ranges using the "Next>" or "<Previous" buttons. You can mouse over the gene to see more details about the gene in the chromosome viewer. COG coloring can be configured using the table at the bottom of the Chromosome Viewer page.

Chromosome Map. You can view a chromosome or scaffold as a linear or circular map, as illustrated in Figure 2.4(iii). A publication-quality postscript file can be downloaded. Genes are displayed on the **Chromosome Map** from outside to the center, as follows:

(a) Genes on the forward strand colored by COG categories

(b) Genes on the reverse strand colored by COG categories

(c) RNA genes, with tRNA genes colored green, and sRNA genes colored red

(d) GC content

(e) GC skew

Web Artemis. You can view a scaffold using Web Artemis, as illustrated in Figure 2.4(iv).

Compare Gene Annotations. The functional annotations, including product name, associated COG, Pfam, TIGRfam, and enzyme, for all the genes in a genome can be viewed in a tabular format, as shown in Figure 2.5, and can be downloaded into a tab-delimited Excel file.

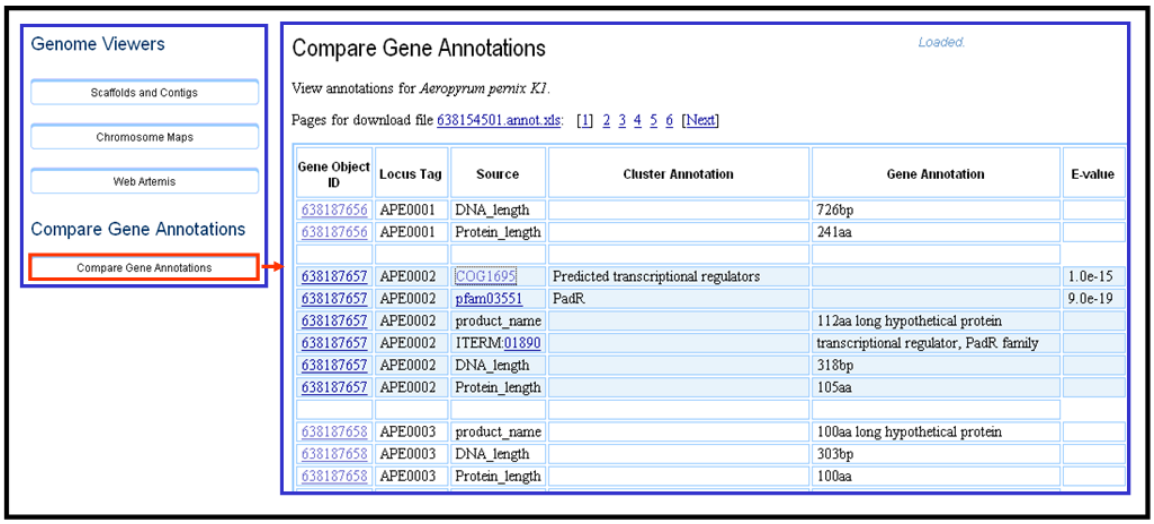


FIGURE 2.5. Organism Details: Compare Gene Annotations.

2.3.4 Export Genome Data

The **Export Genome Data** section provides tools for downloading genome data using a variety of file formats, as shown in Figure 2.3(iv). You can export genome sequences, Scaffold DNA, ORF amino acid and DNA sequences, intergenic sequences, and gene

information in tab-delimited format, readable using **Excel**. Genomes also can be downloaded in **Genbank** file format, which can be used for editing while using **Artemis**.

3 Find and Examine Genes

Genes in IMG can be selected using **Gene Search, BLAST** search, or with the **Phylogenetic Profiler**. The number of results is limited to the "Max. Gene List Results" number set with **MyIMG Preferences**. The default is 1,000 genes and can be changed to larger numbers with **MyIMG**. You can explore individual genes by clicking on the gene in the resulting gene list, or add them to the **Gene Cart** for further analysis.

3.1 Gene Search

Gene keyword search can be carried out using **Gene Search** on the second-level menu of **Find Genes**, as shown in Figure 3.1(i).You can search for genes containing a keyword in a specific field of the IMG database. The search will be carried out only for the genomes you have selected with the **Genome Browser** (all genomes by default) unless you override the selection via the genome selection list in the **Gene Search** page. The default filter searches genes by gene product name. You can select other filters to find genes by gene symbol, locus tag, GenBank or UniProt accession ID, IMG's gene-object identifier, or other attributes available in the pull-down list of filters. You can also find genes by entering a regular expression match against proteins.

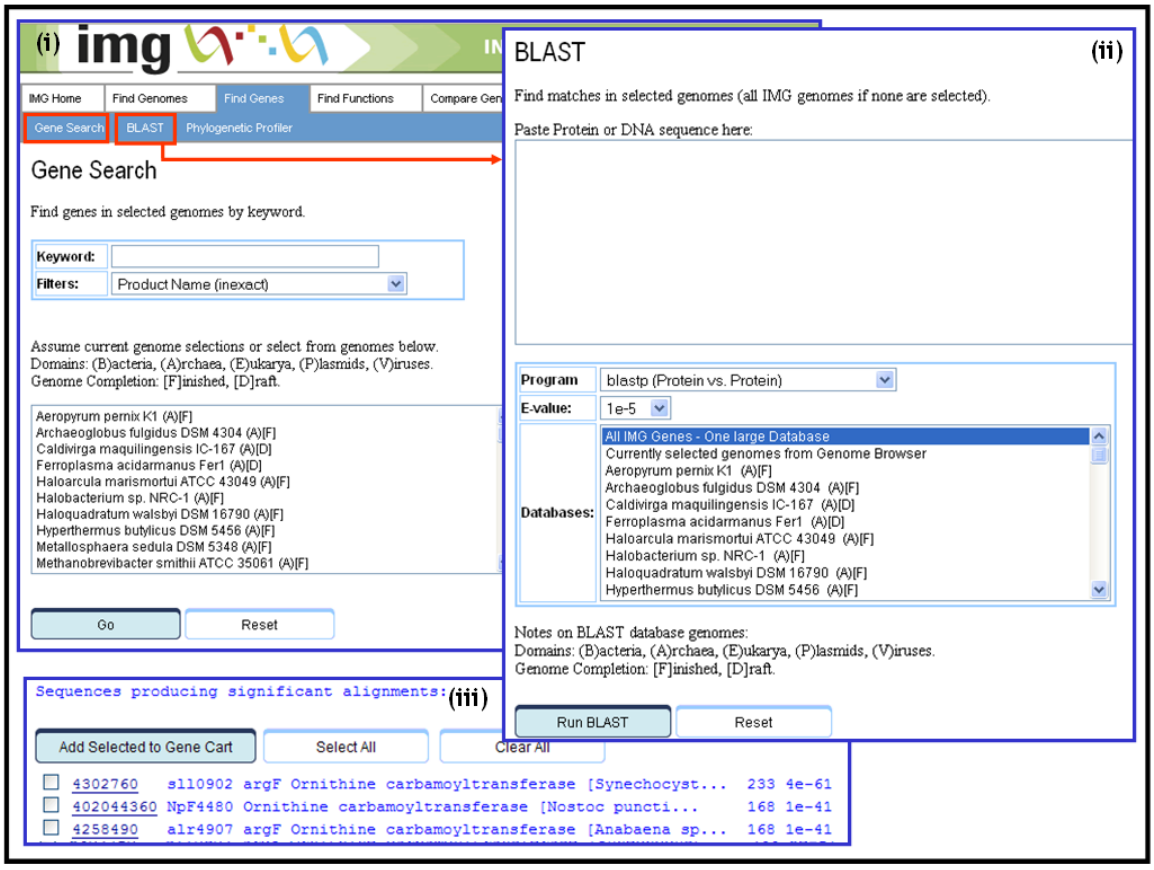


FIGURE 3.1. Gene Search: keyword search and BLAST search.

Most filtering options involve partial matches, that is, the keyword is used as a word or part of a word. You can use the "exact" option to limit the results to gene symbols, or locus tags that your keyword matches exactly. The Gene Object Identifier filter always uses exact matching. Exact matches or matches involving identifiers are case sensitive. For example, (i) searching for "kin" with the filter set to "Product Name" gets "Shikimate kinase," etc.; (ii) searching for "fusA" with the filter set to "Gene Symbol (exact)" only gets genes with the symbol "fusA"; searching for "bsu02690" with the filter set to "Locus Tag" only gets genes with locus tag "BSU02690."

You can use a percent sign (%) as a wildcard in the middle of a keyword. The results will include any genes with zero or more additional characters at that position. For example, "hydro%ase" will get results with "hydrolase" and "hydrogenase." If you want only a single character of your keyword to be variable, type an underscore (_) in that position. Searching for "hydro_ase" will get results with "hydrolase," not "hydrogenase."

For regular-expression protein searches, standard regular expression constructs are used against a sequence of amino acid residues. Common constructs include:

- . Matches and single amino acid
- $ Matches end of sequence.
- ^ Matches beginning of sequence.
- Matches zero or more occurrences of the preceding residue.
- + Matches one or more occurrences of the preceding residue.
- [ ] Matches any of residues between brackets.
- [^] Matches any of residues not between brackets.

3.2 BLAST Search

BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs that look for local, as opposed to global, sequence alignments among IMG genomes, thus detecting relationships among sequences that share only isolated regions of similarity.

A BLAST search can be carried out using **BLAST** on the second-level menu of **Find Genes**, as shown in Figure 3.1(ii).You can run a BLASTp (protein-vs-protein), BLASTx (DNA-vs-protein), BLASTN (DNA-vs-DNA), or TBLASTn (protein-DNA-vs-DNA-protein) search to find genes in IMG that match a query sequence. Copy and paste your sequence into the BLAST text field, select an E-value cutoff and a database option, and click "Run BLAST." The database list allows you to run BLAST against all IMG genomes or against the genomes you have selected. The optimum E-value will depend on the size of the BLAST database you select. (For a larger database, use a larger E-value.) The BLAST result consists of a list of genes, as shown in Figure 3.1(iii), that can be selected and saved using the “Add Selected to Gene Cart” button.

3.3 Gene Selection with the Phylogenetic Profiler

Genes in a specific genome can be selected based on presence or absence of homologs in other genomes using **Phylogenetic Profiler** on the second-level menu of **Find Genes**, as shown in Figure 3.2(i).

Before you start, select the IMG genomes that will be used in finding genes with the Genome Browser or Genome Search, and save your selections. Selecting fewer genomes will speed up the computations underlying the Phylogenetic Profiler by reducing the genome context for the search.

First, select your target genome by using the associated radio button in the "Find Genes In" column, as shown in Figure 3.2(i). Next, select the genomes for homolog comparisons with the target genome by using the associated radio buttons in the "With Homologs In" and "Without Homologs In" columns. Genomes you want to be ignored in these comparisons can be selected using the radio buttons in the "Ignoring" column.

You can set the Maximum E-value and Minimum Percent Identity for which results are reported using the “Similarity Cutoffs” available at the bottom of the Phylogenetic Profiler page, as shown in Figure 3.2(ii). Click "Go" to find the genes in the target genome that satisfy the homolog presence/absence condition you have set.

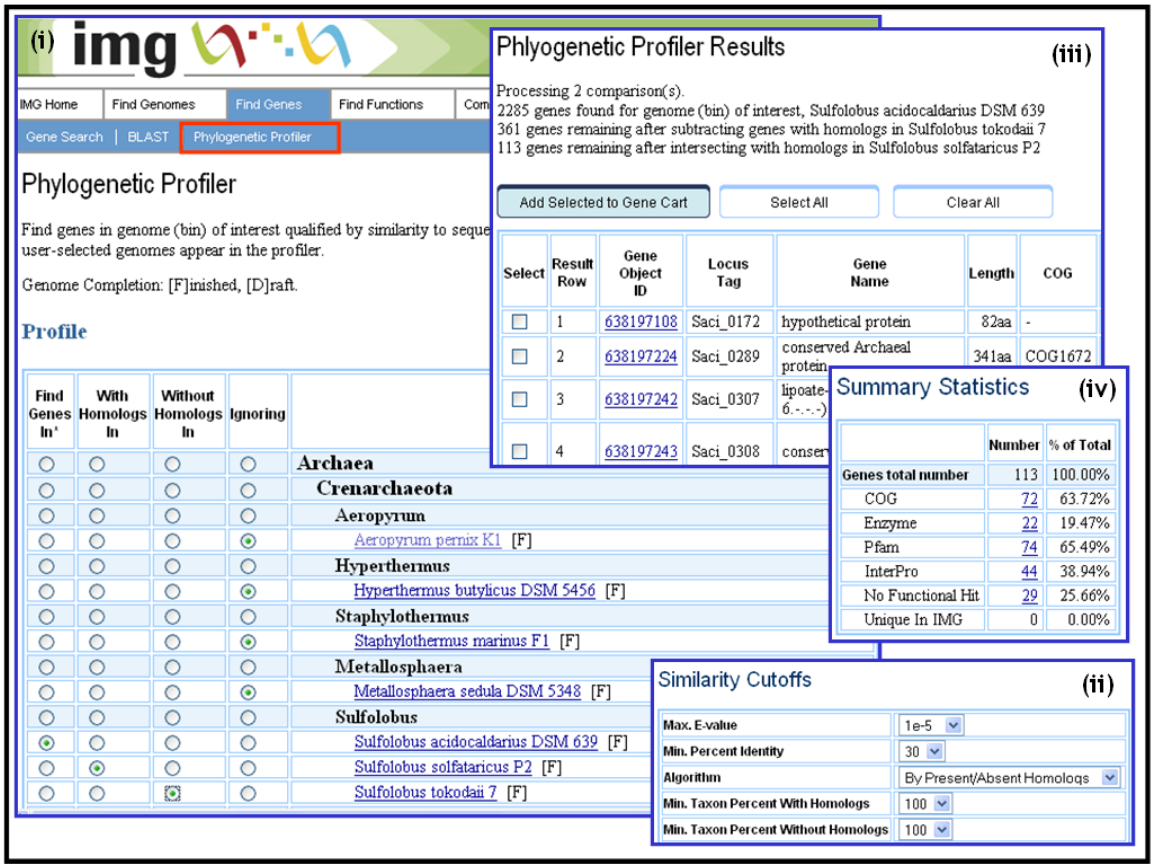


FIGURE 3.2. Gene Search: Phylogenetic Profiler.

The **Phylogenetic Profiler** results show the genes in the target genome along with their functional characterization based on COG, Pfam, InterPro, and Enzymes, as illustrated in

Figure 3.2(iii). The results also show whether the gene is **unique** in the IMG database. Note that gene uniqueness should be considered only in the context of specific **Phylogenetic Profiler** search parameters. You can explore individual gene details from the results list by clicking on the associated "Gene Object ID." By clicking on the radio buttons in the "Select" column, you can select genes that will be added to the **Gene Cart** for further analysis. If you want to select all the genes, click on the "Select All" button. To clear all selections, click on "Clear All" button. After the genes are selected, click on the "Add Selected to Gene Cart" button.

At the bottom of the results page, a summary of the genes with or without functional annotations is provided, as shown in Figure 3.2(iv). The count of genes is linked to the corresponding list of genes, which can be viewed by clicking on the numbers for each category.

By default, the presence/absence homolog search is based on set operations involving sets of genes in the target genome that have or do not have homologs in other genomes. In some situations, some genomes may be missing gene calls, in which case, a more fault-tolerant search based on the percentage of matching genomes can be used.

3.4 Gene Selection from Genome Statistics

Genes can be selected from a specific genome from one of the gene categories provided in the **Genome Statistics** section of **Organism Details** (see Figure 2.3), as shown in Figure 3.3(i).

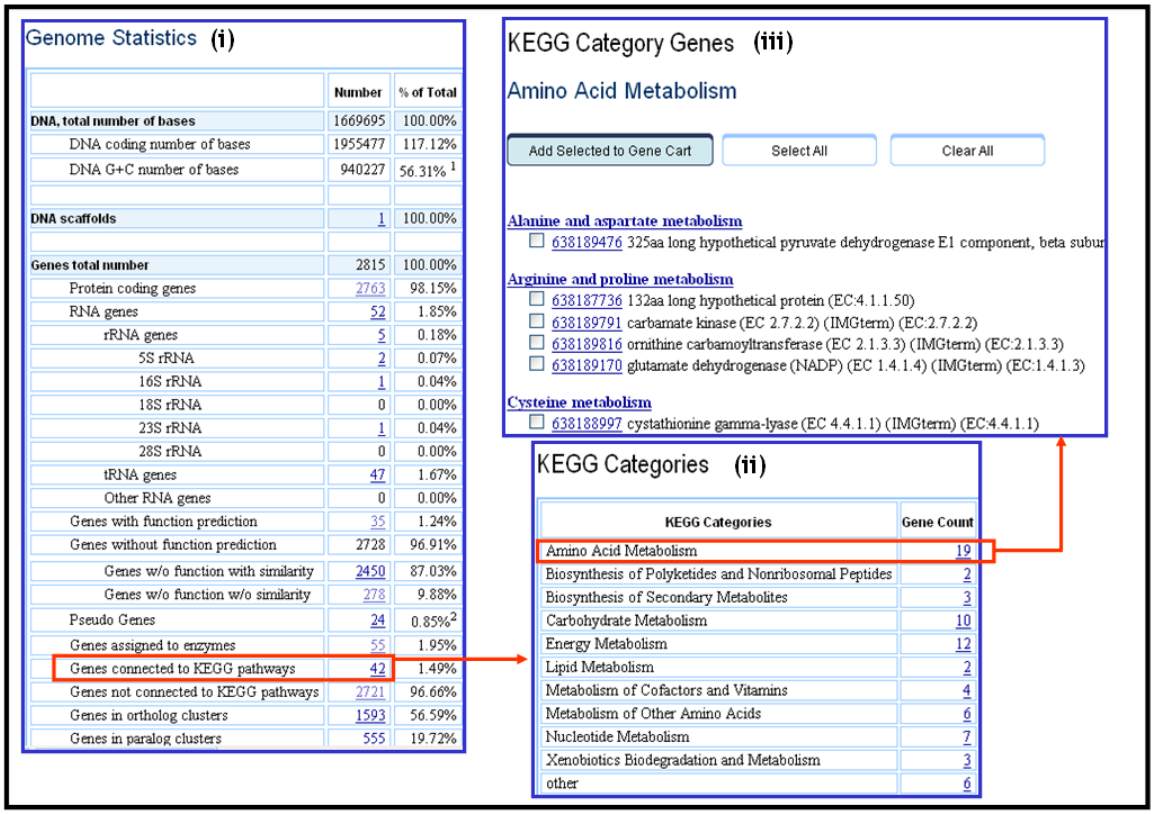


FIGURE 3.3. Selecting genes with Genome Statistics.

Gene counts in some categories, such as “Genes in COGs” and “Genes connected to KEGG pathways” are linked to tables that show these genes classified according to the corresponding functional hierarchies (COG Functional Categories and KEGG Categories, respectively, with the possibility of further breakdown into COG Pathways and KEGG Maps). The gene counts in the “KEGG Categories” table are linked to the table that contains groupings of genes according to individual KEGG Maps corresponding to collections of metabolic pathways, as illustrated in Figure 3.3(ii). For a specific KEGG category, a link is provided to the list of genes organized in specific pathways, such as those shown in Figure 3.3(iii). Genes can be then selected and saved in the **Gene Cart** for further analysis.

3.5 Gene Details

Genes can be explored using Gene Details pages that can be accessed by clicking on a gene object identifier in any list of genes, such as one resulting from a Gene Search. All gene lists include checkboxes that can be used to add genes to the Gene Cart. Gene lists include the object identifier (Object ID) of the gene, its name, and the name of the genome in which it was found. These lists also show an EC number assignment for the gene's product, if one exists.

The Gene Details page contains several sections, as shown in Figure 3.4(i), that are further discussed below.

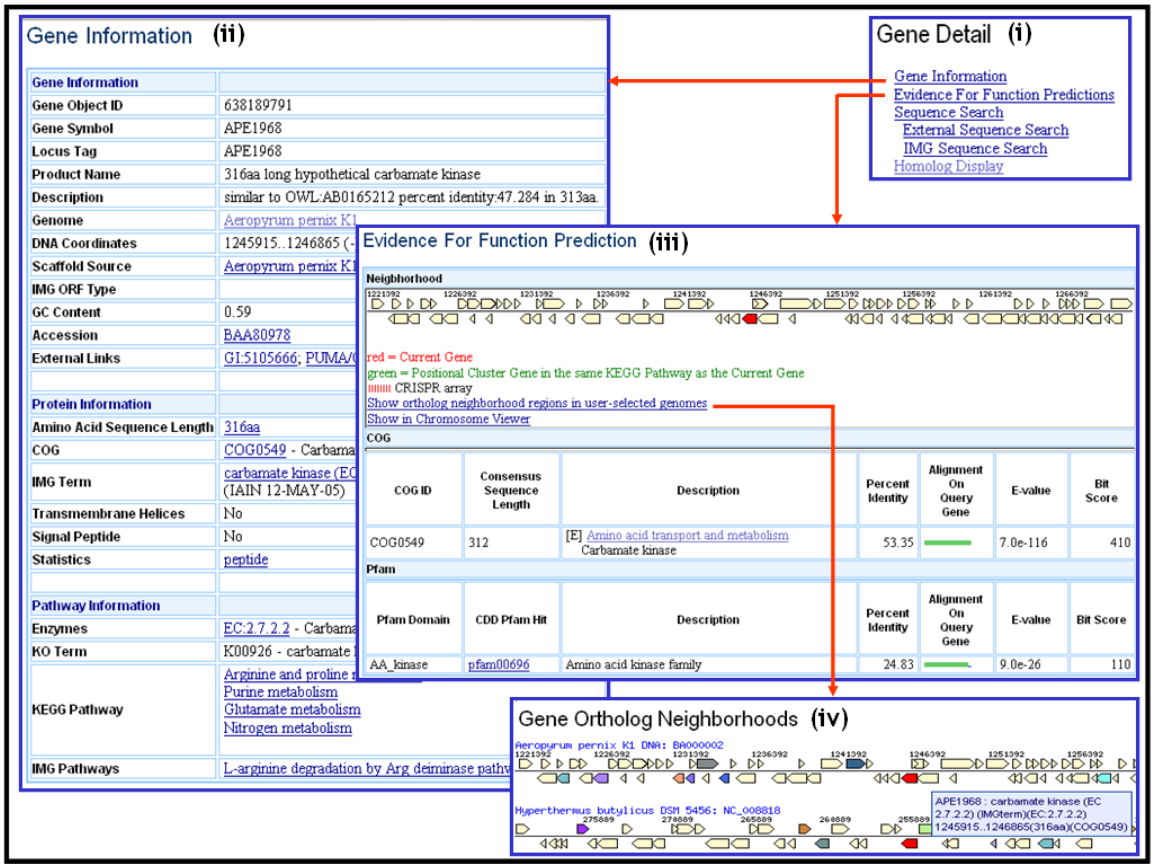


FIGURE 3.4. Gene Details: Gene Information and Evidence for Function Prediction sections.

3.5.1 Gene Information

The Gene Information section is divided into three parts that contain information on the gene, protein, and associated pathways, as shown in Figure 3.4(ii). The first part includes the gene identifier, gene symbol, product name, coordinates on the scaffold, and links to external resources, such as NCBI's Entrez Gene. The second part of the **Gene Information** section contains information on the protein, such as the amino acid sequence length, signal peptide, and associated protein family based on various classifications, such as TIGRfam and InterPro.

The third part of the Gene Information section contains information on associated KEGG and IMG pathways, and enzymes. Links to the pathways and enzyme information are also provided. For example, in order to view the KEGG map for a pathway that a particular gene is connected to, click the pathway name. The map will be displayed as illustrated in Figure 3.5.

Colored EC numbers on the KEGG map are links: (i) **red** links represent the current gene and point to its **Gene Details** page; (ii) **green** links indicate genes in a positional cluster that includes the current gene, and point to the corresponding **Gene Details** page; (iii) **blue** links indicate other genes that are in the same genome; (iv) **purple** links indicate genes in the same genome associated with the same EC number as the current gene; (v) **orange** links indicate "EC equivalogs," that is, genes in other genomes that are associated with the same EC number. Rounded rectangles are links to other pathway maps.

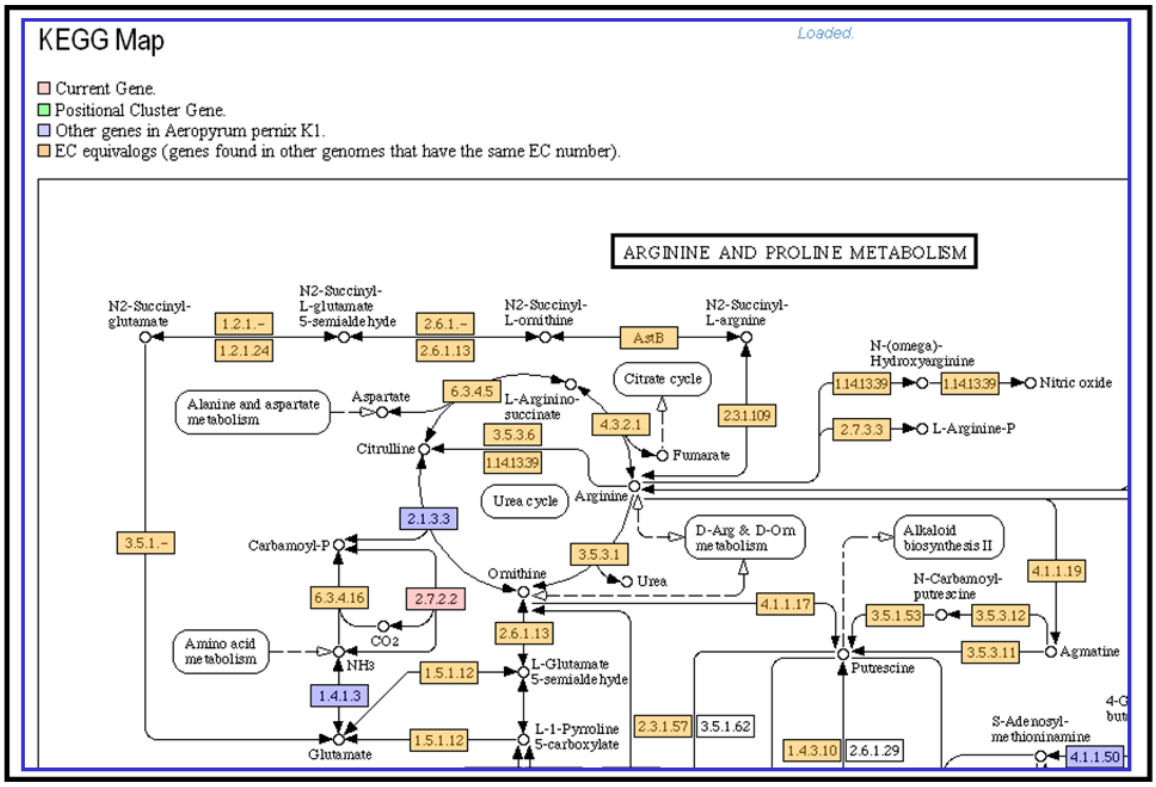


FIGURE 3.5. Gene Details: displaying a KEGG map.

3.5.2 Evidence for Function Prediction

The Evidence for Function Prediction section includes a graphic display of the gene's neighborhood, and information on associated COG and Pfam domain, as shown in Figure 3.4(iii).

The gene's immediate neighborhood is displayed graphically on a single line with the genes reading in one direction on top, and those reading in the opposite direction on the bottom. The target gene is shown in red. Genes in the same positional cluster that are also in the same KEGG pathway are highlighted in green. When you move the cursor over any gene, you will see a popup box with the locus tag, gene name, and scaffold coordinates (except in Internet Explorer for Macintosh). Click the arrow representing a gene to see the Gene Details page for that gene.

You can see more details on the gene of interest's chromosome, via the **Show in Chromosome Viewer** link. The **Chromosome Viewer** shows all the genes in a range of sequence, with color coding by COG group, as illustrated in Figure 2.4(ii). The gene whose details you were examining is shown in red. If you move the cursor over a gene, you will see the gene name, COG code, and coordinates of that gene (except in Internet Explorer for Macintosh). The COG code appears in brackets. Click an arrow to view details for the gene it represents. The COG function coloring can be customized.

You can simultaneously view neighborhoods for the gene of interest and its orthologs in the genomes you have selected via the **Show ortholog neighborhood regions in user-selected genomes** link, as illustrated in Figure 3.4(iv). You may want to avoid this option if you have all the genomes selected, since the page will be slow to load, and the orthologs you would like to compare are not likely to be shown, because the number of displayed neighborhoods is limited. You can change the number of neighborhoods shown by entering a new setting for "Max. Taxon Gene Neighborhoods" on the **Preferences** page. The gene whose details you were examining is shown in red at the center of the top neighborhood. Each ortholog is also shown in red in the center of its own neighborhood. Genes with the same color, except for the default light yellow, are from the same COG group. When you move the cursor over any gene, you will see a popup box with the locus tag, scaffold coordinates, and COG group number (except in Internet Explorer for Macintosh). Click the arrow representing a gene to see the **Gene Details** page for that gene.

3.5.3 Sequence Search

The External Sequence Search section provides links to several resources outside IMG, such as NCBI BLAST, illustrated in Figure 3.6(i).

IMG Sequence Search provides the **IMG Genome BLAST** tool that allows you to carry out a BLAST search using the gene you examine against specific IMG genomes you can select from a list, as shown in Figure 3.6(ii). The result of the BLAST search consists of a table of genes that can be selected and included into the **Gene Cart**.

Another **IMG Sequence Search** tool is the **Phylogenetic Profile Similarity Search**, illustrated in Figure 3.6(iii). In order to search for genes with similar phylogenetic occurrence profiles, you need to select the *minimum percent occurrence match* and

number of results (“*Top N results*”) to be retrieved. The percent occurrence match determines the similarity threshold—the minimum required number of gene occurrence matches when comparing the profiles of two genes out of the total number of occurrences across all genomes. The search returns a list of genes, in descending percentage order, with profiles that satisfy the minimum percent occurrence match condition. Genes in this list can be selected and included into the **Gene Cart**.

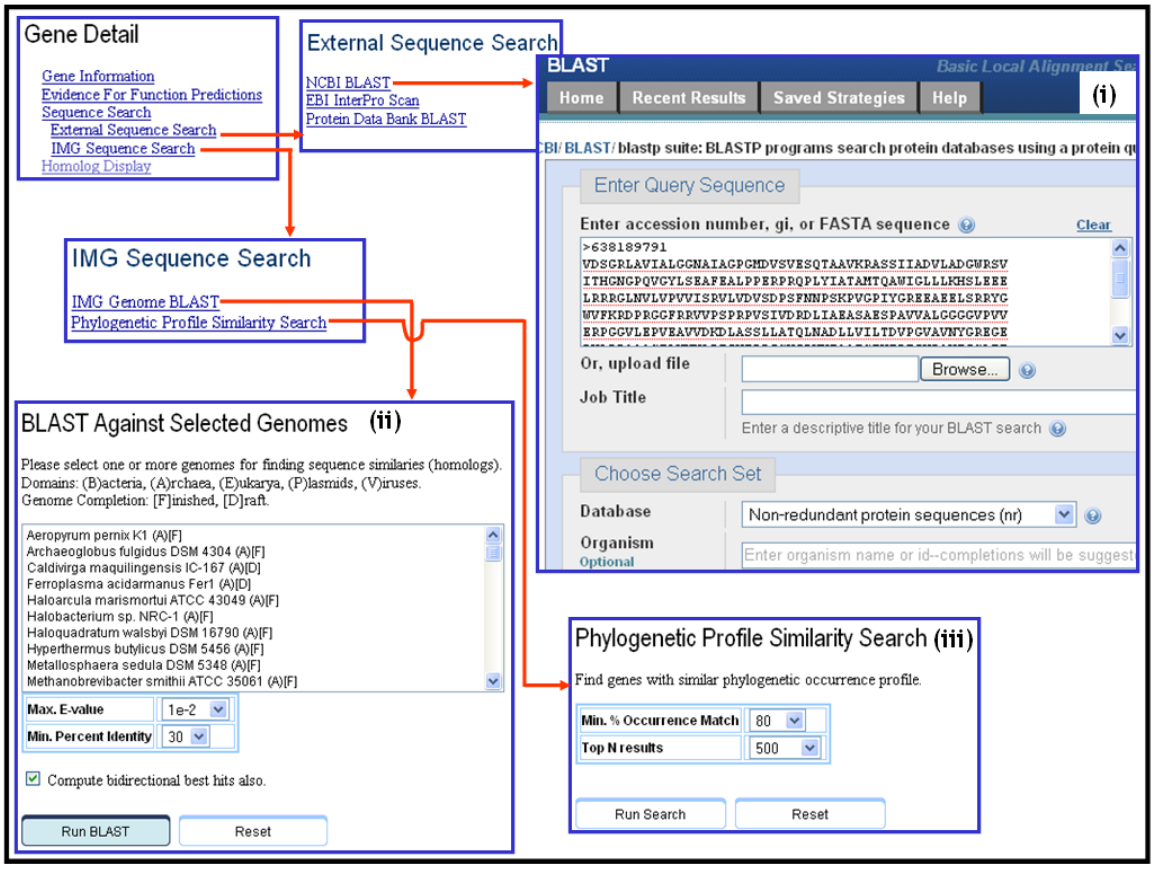


FIGURE 3.6. External and IMG Sequence Search.

3.5.4 Homolog Display

The Homolog Display section provides support for examining genes that have similar sequences to your target gene.

The **Customized Homolog Display** allows examining genes that are homologs of your target gene using a customized format, as illustrated in Figure 3.7(i). The homolog genes are presented in multiple pages, as shown in Figure 3.7(ii).

The **Selective Homolog Display** allows examining genes that have similar sequences to your target gene, including othologs/paralogs and top IMG homolog hits, as illustrated in Figure 3.8(i). A list of (ortholog, paralog, homolog) genes is displayed in a preset format, as shown in Figure 3.8(ii). If you choose one of the metadata attributes in the pull-down menu, such as Phenotype, then the homologs are first listed in metadata-attribute-specific groups, as illustrated in Figure 3.8(iii).

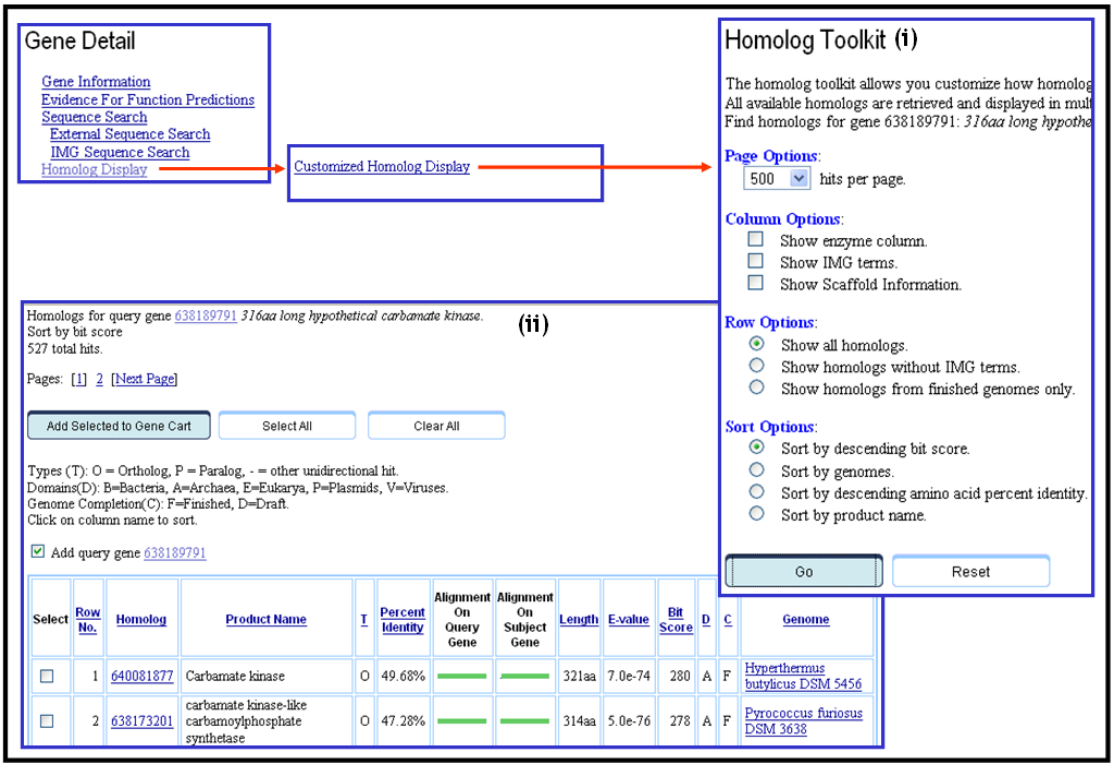


FIGURE 3.7. Customized Homolog Display.

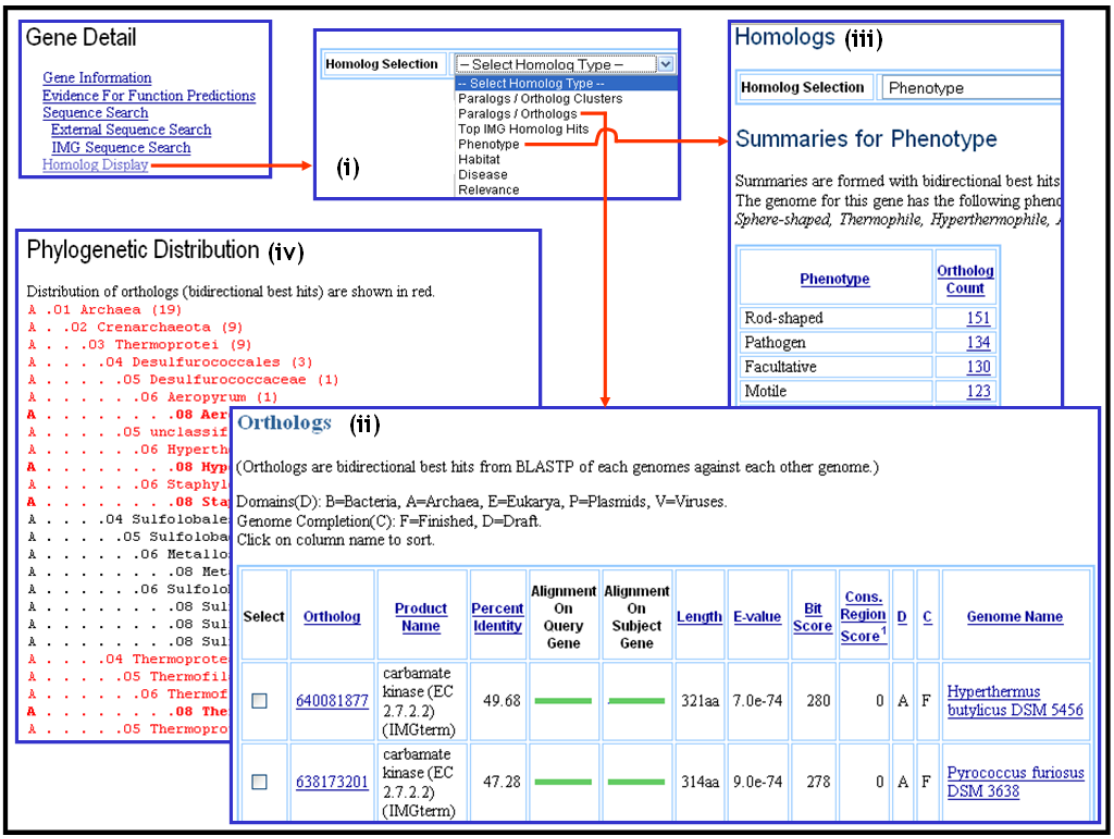


FIGURE 3.8. Selective Homolog Display.

You can sort on column names to show most significant to least significant entries based on the column parameter. You can view the homologs in the context of the phylogenetic list of genomes, as illustrated in Figure 3.8(iv).

A **Phylogenetic Distribution** viewer is available at the end of the list of homolog genes. In the phylogenetic list of genomes, genomes containing a homolog of your target gene are highlighted in red. The count of homologous genes at each taxonomic level is shown in parentheses.

4 Find and Examine Functions

Functions in IMG can be selected using **Search Terms and Pathways** or with browsers for specific functional classifications, such as COG, Pfam, TIGRfam, KEGG, and IMG terms, pathways, and networks.

4.1 Function Search

Functional terms and pathway search can be carried out using **Search** on the second-level menu of **Find Functions**, as shown in Figure 4.1(i).

Select a functional classification filter from the pull-down list, such as “COG” or “Pfam,” then enter a search term corresponding to a COG category or COG name, GO term, EC number, Pfam ID or name, InterPro ID or name, or KEGG category or KEGG pathway name. The search term for a name can be a word in a term, or a substring matching the values in the filter attribute.

The genome context of the search can be limited by selecting genomes with the **Genome Browser,** or by selecting one or more genomes from the scroll-down list provided in the **Search** page.

For example, you can search for COGs that contain “lipid” in their name, and restrict the search to genome “Alkaliphillus metalliredigenes,” as shown in Figure 4.1(i).

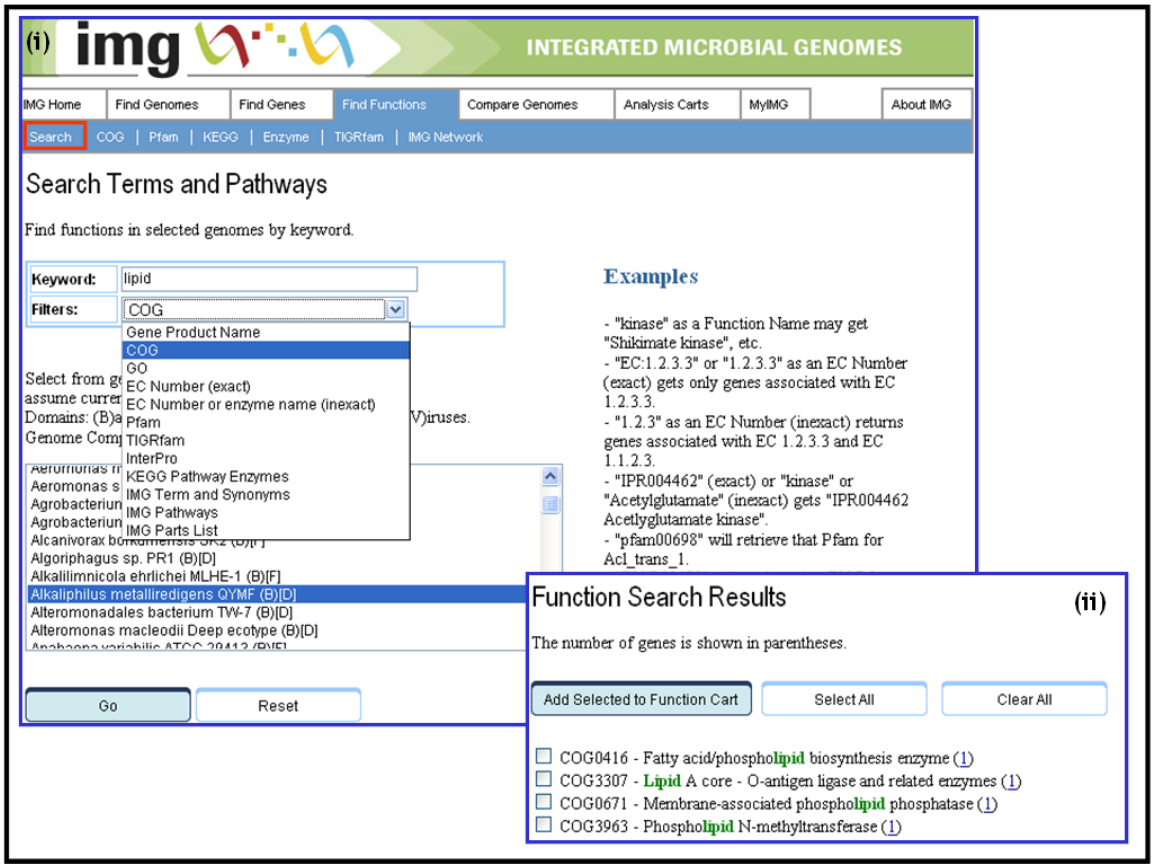


FIGURE 4.1. Function search.

The result of a function search is a list of functional terms or pathways, as shown in Figure 4.1(ii). The number shown for each functional term or pathway is the list of genes associated with that term or pathway in the genomes that were selected with the **Genome Browser** or in the **Search** page. Clicking on such a number leads to the list of genes. Terms and pathways can be added to the **Function Cart** by clicking on the check boxes next to the terms or pathways of interest, then clicking on the "Add Selected to Function Cart" button.

4.2 Function Browsers

Browsers for functional terms and pathways are provided on the second-level menu of **Find Functions**, as shown in Figure 4.2(i). Browsers are provided for **COG** (see Figure 4.2(i)), **Pfam** (see Figure 4.2(ii)), **KEGG** (see Figure 4.2(iii)), **Enzymes** (see Figure 4.2(iv)), **TIGRfam** (see Figure 4.2(v)), and **IMG terms**, **pathways, parts list,** and **networks** (see Figure 4.2(vi)).

Each browser displays the functional categories specific to each functional classification, such as COG categories, TIGRfam roles, and IMG networks. You can examine a functional category by clicking on its name and opening a functional category **details** page. Such pages are discussed below in the **Function Lists and Details** section.

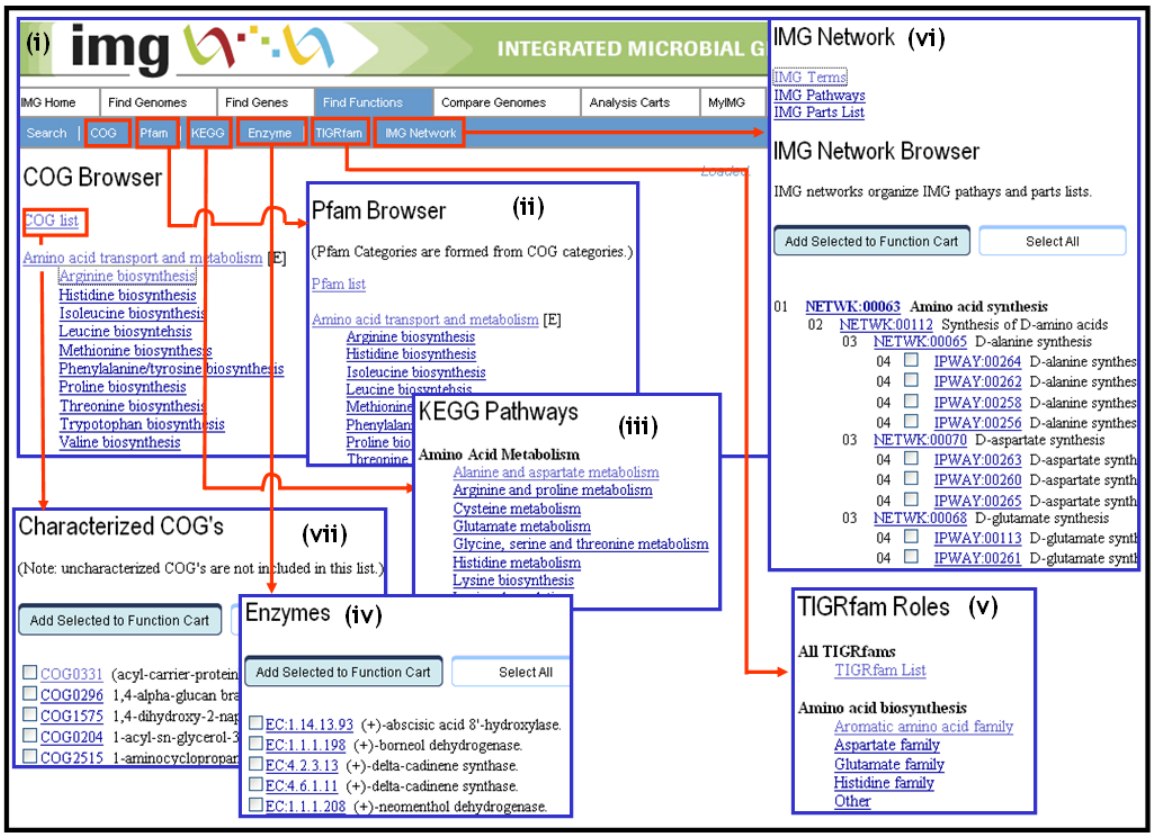


FIGURE 4.2. Function Browsers.

Each browser also provides a link to a **list of terms**, such as the list of COGs shown in Figure 4.2(vii), specific to each functional classification. From every list of functional

terms and pathways you can look at details of an individual term or pathway by clicking on its name, or select the term or pathway for inclusion into the **Function Cart**.

4.3 Function Details

Functions can be explored using details pages specific to each functional or pathway classification available under the **Find Function** menu option.

4.3.1 COG Details

From the **Find Function** top-level menu, the **COG** option on the second-level menu leads to the **COG Browser** page, as shown in Figure 4.3(i). COGs are listed in a two-level hierarchy consisting of COG categories and COG pathways, respectively.

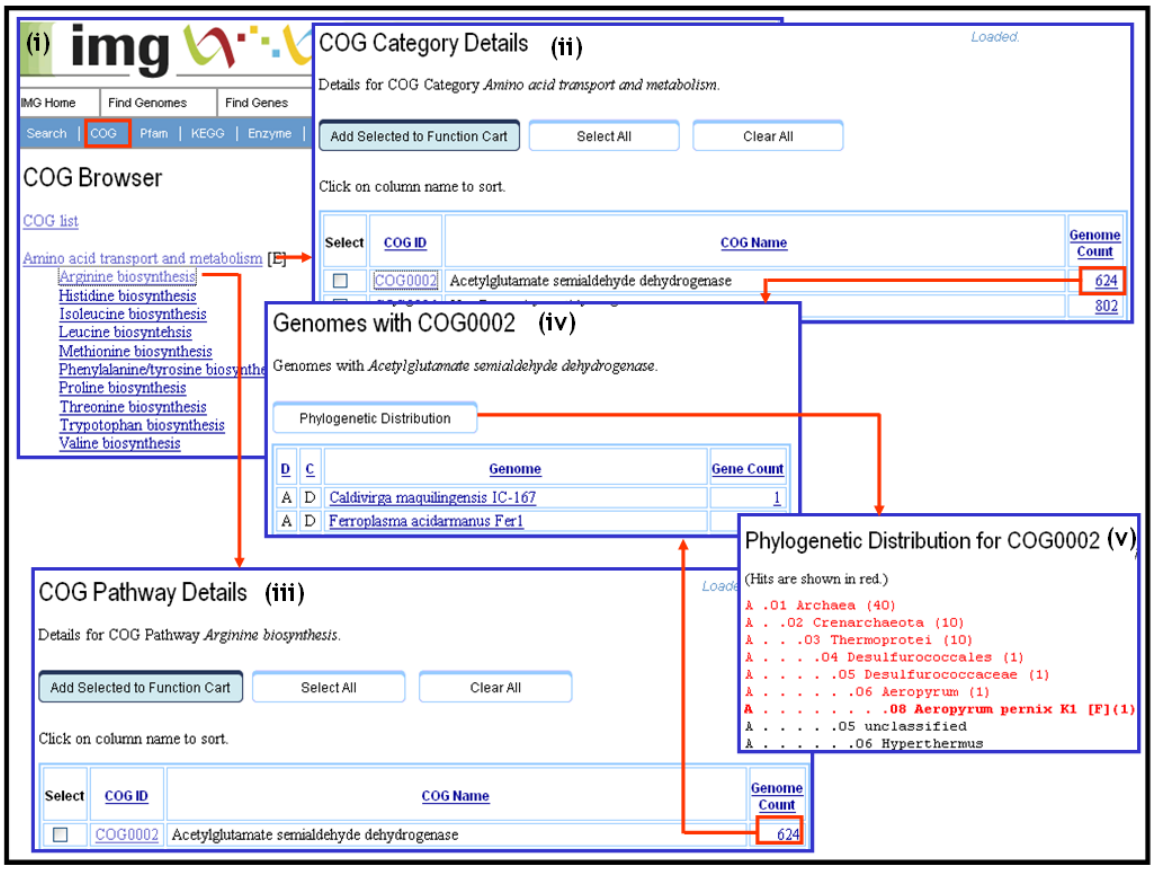


FIGURE 4.3. COG Category and Pathway Details.

COG categories and pathways can be examined using the **COG Category Details** and **COG Pathway Details** pages, respectively, as shown in Figure 4.3(ii) and Figure 4.3(iii). Both **COG Category Details** and **COG Pathway Details** pages list all the COG members for a given COG category or pathway. COGs can be selected from these lists and included into the **Function Cart** for further analysis. For each COG, the number of genomes that have genes associated with this COG member is also provided. This genome number provides a link that leads to the list of genomes associated with this COG, as well as the number of genes associated with this COG for each genome, as

shown in Figure 4.3(iv). For genomes associated with a specific COG, the **Phylogenetic Distribution** of the COG across IMG genomes can be displayed as shown in Figure 4.3(v): genomes containing a gene associated with this COG are highlighted in red, and the count of genes associated with this COG is shown in parentheses at each taxonomic level.

4.3.2 Pfam Details

From the **Find Function** top-level menu, the **Pfam** option on the second-level menu leads to the **Pfam Browser** page, as shown in Figure 4.4(i).

A list of Pfam families is provided as shown in Figure 4.4(ii). Pfams can be selected from this list and included in the **Function Cart** for further analysis. Pfam domains are organized, when possible, into Pfam categories and Pfam pathways that are based on COG categories and pathways, respectively. Accordingly, Pfams are listed in a two-level hierarchy consisting of Pfam categories and Pfam pathways, respectively.

Pfam categories and pathways can be examined using the **Pfam Category Details** and **Pfam Pathway Details** pages, respectively, as shown in Figure 4.4(iii) and Figure 4.4(iv). Both **Pfam Category Details** and **Pfam Pathway Details** pages list all the Pfam members for a given Pfam category or pathway. Pfams can be selected from these lists and included into the **Function Cart** for further analysis.

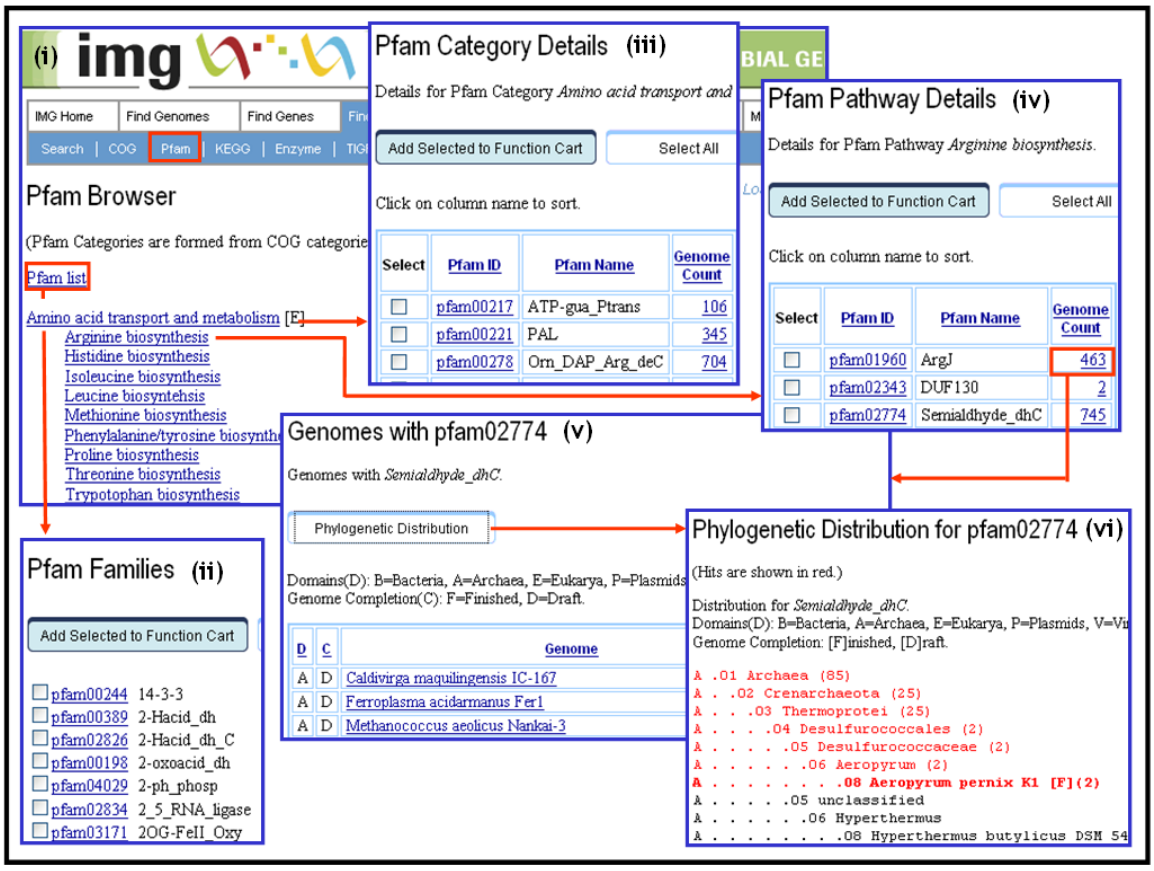


FIGURE 4.4. Pfam Category and Pathway Details.

For each Pfam member, the number of genomes that have genes associated with this Pfam member is also provided. This genome number provides a link that leads to the list of genomes associated with this Pfam, as well as the number of genes associated with this Pfam for each genome, as shown in Figure 4.4(v). For genomes associated with a specific Pfam, the **Phylogenetic Distribution** of the Pfam across IMG genomes can be displayed as shown in Figure 4.4(vi): genomes containing a gene associated with this Pfam are highlighted in red, and the count of genes associated with this Pfam is shown in parentheses at each taxonomic level.

4.3.3 KEGG Pathway Details

From the **Find Function** top-level menu, the **KEGG** option on the second-level menu leads to the **KEGG Browser** page, as shown in Figure 4.5(i).

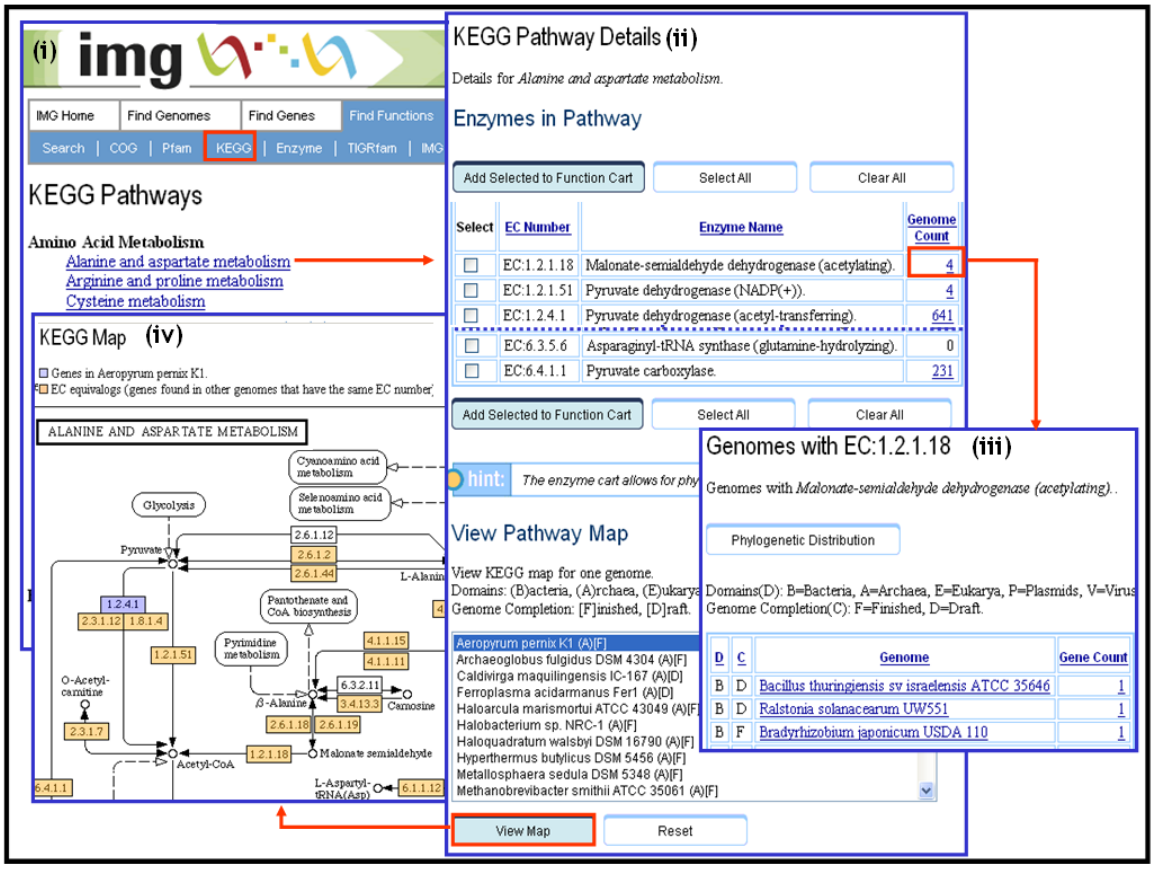


FIGURE 4.5. KEGG Pathway Details.

KEGG pathways are organized in KEGG categories. Each KEGG pathway can be examined with a **KEGG Pathway Details** page, as shown in Figure 4.5(ii), which lists the enzymes associated with a specific KEGG pathway. For each enzyme, the number of genomes that have genes associated with this enzyme is also provided. This genome number provides a link that leads to the list of genomes associated with this enzyme, as well as the number of genes associated with this enzyme for each genome, as shown in Figure 4.5(iii). By clicking on the left-column checkbox for an enzyme entry in the

KEGG Pathway Details page, enzymes can be added to the **Function Cart** for further analysis.

The **KEGG Pathway Details** contains a **View Pathway Map** section that can display the KEGG map associated with a pathway for a genome selected from a list, as shown in Figure 4.5(ii). In a KEGG pathway map, such as that shown in Figure 4.5(iv), the enzymes associated with genes for the selected genome are displayed in blue; enzymes associated with genes for other genomes in IMG are displayed in orange.

4.3.4 TIGRfam Details

From the **Find Function** top-level menu, the **TIGRfam** option on the second-level menu leads to the **TIGRfam Browser** page, as shown in Figure 4.6(i).

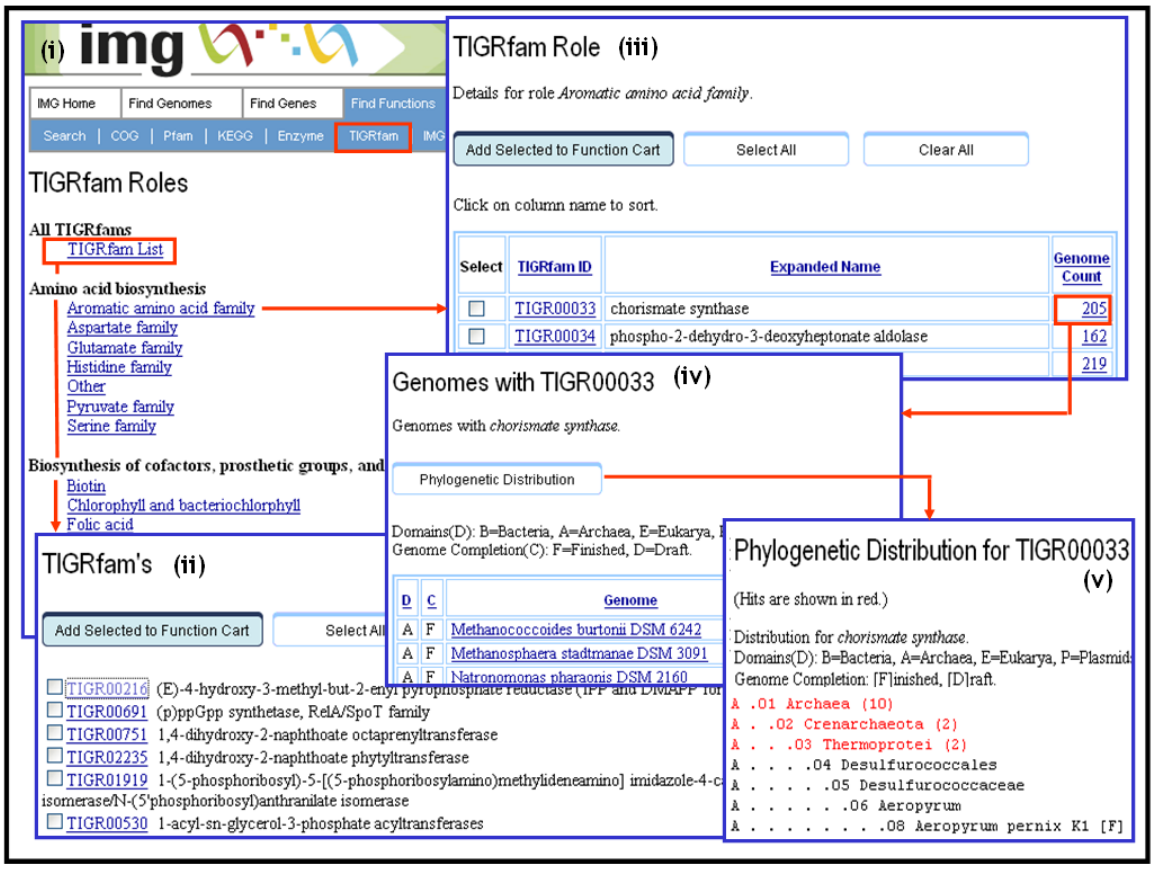


FIGURE 4.6. TIGRfam Details.

A list of TIGRfam families is provided as shown in Figure 4.6(ii). TIGRfams can be selected from this list and included into the **Function Cart** for further analysis. TIGRfams are organized into functional subsystem roles. Each TIGRfam role can be examined with a **TIGRfam Role Details** page, as shown in Figure 4.6(iii), which includes the list of TIGRfam members for a given role. TIGRfams can be selected from this list and included into the **Function Cart** for further analysis. For each TIGRfam member, the number of genomes that have genes associated with this TIGRfam member is also provided. This genome number provides a link that leads to the list of genomes

associated with this TIGRfam, as well as the number of genes associated with this TIGRfam for each genome, as shown in Figure 4.6(iv). For genomes associated with a specific TIGRfam, the **Phylogenetic Distribution** of the TIGRfam across IMG genomes can be displayed as shown in Figure 4.6(v): genomes containing a gene associated with this TIGRfam are highlighted in red, and the count of genes associated with this TIGRfam is shown in parentheses at each taxonomic level.

4.3.5 IMG Networks Details

From the **Find Function** top-level menu, the **IMG Networks** option on the second-level menu leads to the **IMG Network** page, as shown in Figure 4.7(i).

4.3.5.1 IMG Terms

IMG terms provide a controlled vocabulary for the functional roles of genes within IMG. The list of IMG terms can be examined with the **IMG Term Browser**, as shown in Figure 4.7(ii). IMG terms can be selected from this list and included into the **Function Cart** for further analysis.

For each IMG term, the **IMG Term Details** page provides information on the term, related (parent or child) terms, related IMG pathways, reactions, and parts list, and the number of genes associated with the term, as shown in Figure 4.7(iii). The genomes that have genes associated with the term are listed as shown in Figure 4.7(iv).

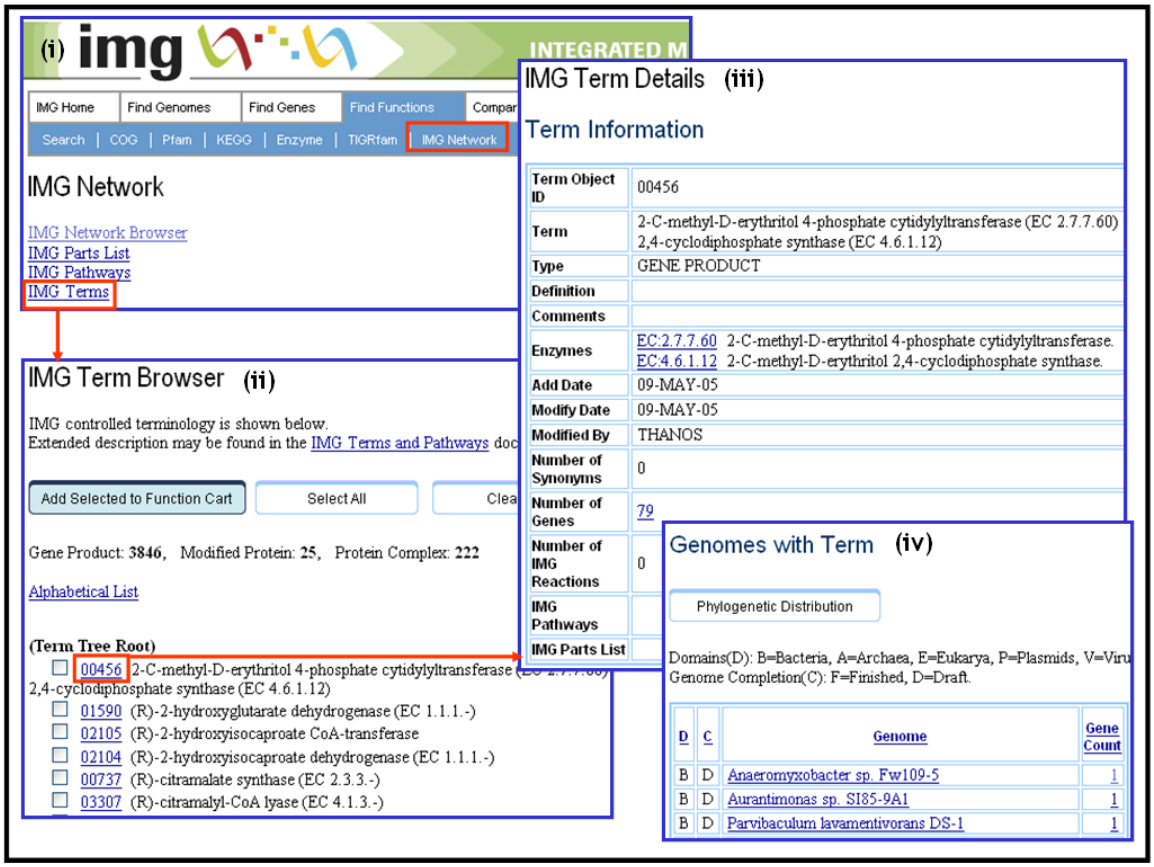


FIGURE 4.7. IMG Term Details.

4.3.5.2 IMG Pathways

A list of IMG pathways is provided as shown in Figure 4.8(ii). IMG pathways can be selected from this list and included into the **Function Cart** for further analysis.

For each IMG pathway, the **IMG Pathway Details** page provides information on the pathway and its reactions, as shown in Figure 4.8(iii). Reactions are listed with a reaction order number, with alternates within a reaction shown with a suffix letter. The genomes that have genes associated with the pathway are listed as shown in Figure 4.8(iv). For each reaction in the pathway, the IMG terms associated with the reaction are also listed. Such terms usually describe an enzyme catalyzing the reaction or subunits involved in the formation of a protein complex. For terms associated with reactions, the number of genomes that have genes associated with the term is provided; this number provides a link to the list of these genomes, as shown in Figure 4.8(v).

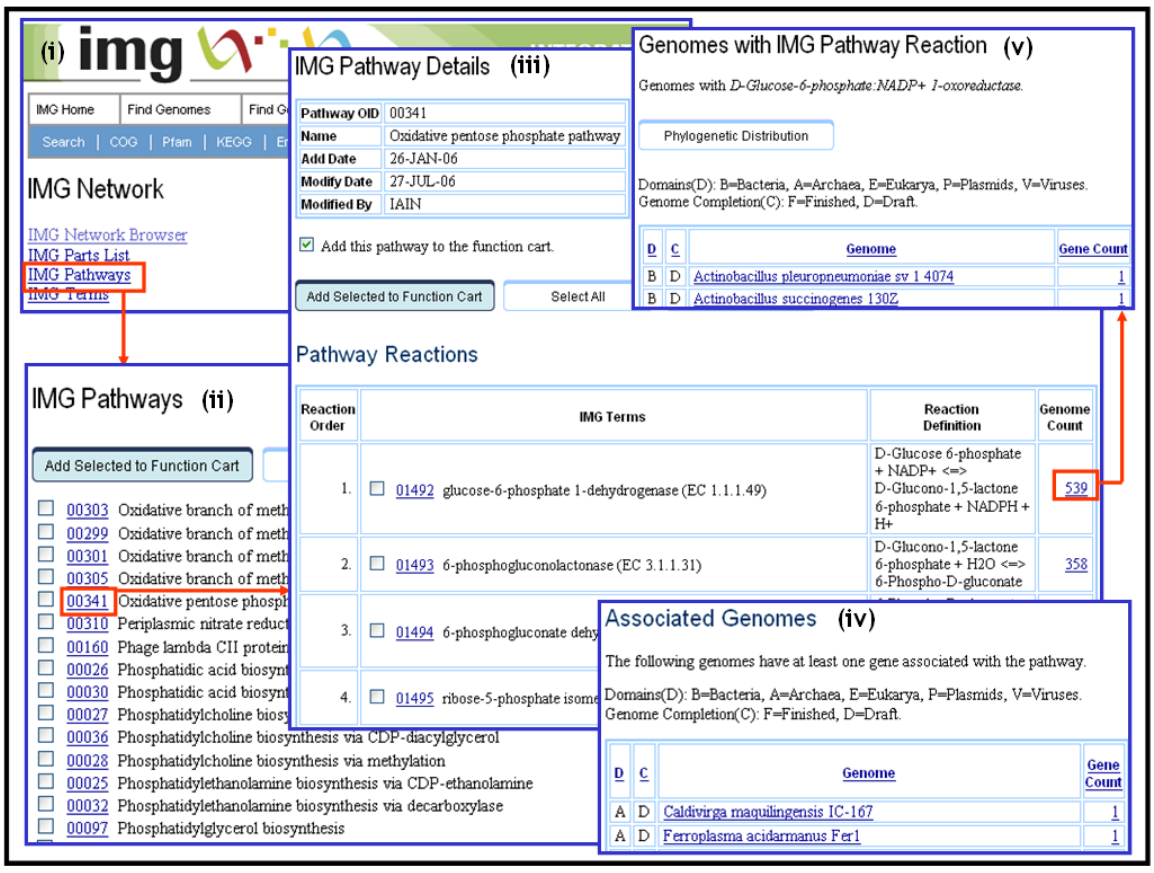


FIGURE 4.8. IMG Pathway Details.

4.3.5.3 IMG Networks

IMG networks are organized hierarchically with each network consisting of component (child) networks or IMG pathways, as shown in Figure 4.9(i). Component IMG pathways can be selected and included into the **Function Cart** for further analysis.

For each IMG network, the **IMG Network Details** page provides information on the network and its component pathways, as shown in Figure 4.9(ii). Each pathway can be

examined using the **IMG Pathway Detail**s page as shown in Figure 4.9(iii) and as discussed above.

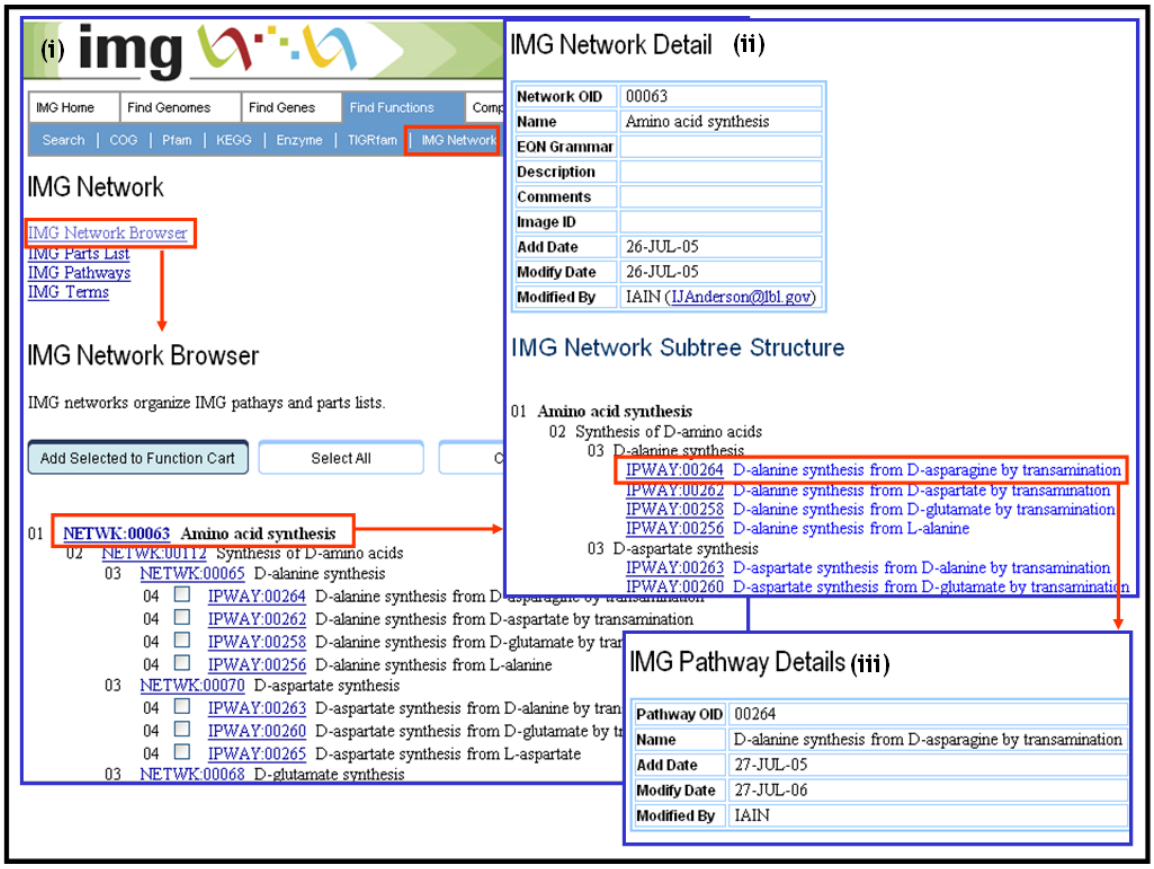


FIGURE 4.9. IMG Networks Detail.

5 Compare Genomes

Genomes in IMG can be compared in terms of various statistics, function capabilities, and sequence conservation by using tools available under IMG's **Compare Genomes** main menu option.

5.1 Genome Statistics

Genome Statistics provide statistics for the genomes that have been selected using the **Genome Browser**, as shown in Figure 5.1.

The **Summary Statistics** part of **Genome Statistics** consists of cumulative statistics for the selected genomes, including the numbers of bases, scaffolds, and genes with various characteristics, as illustrated in Figure 5.1(i).

The **General Statistics** part of **Genome Statistics** provides statistics for the selected genomes side by side in a tabular format, as shown in Figure 5.1(ii). The statistics can be **exported** into a tab-delimited file and can be configured using a **Configuration** table. Each genome in the **General Statistics** table is linked to its **Genome Details** page, which also provides statistics for individual genomes, as discussed above.

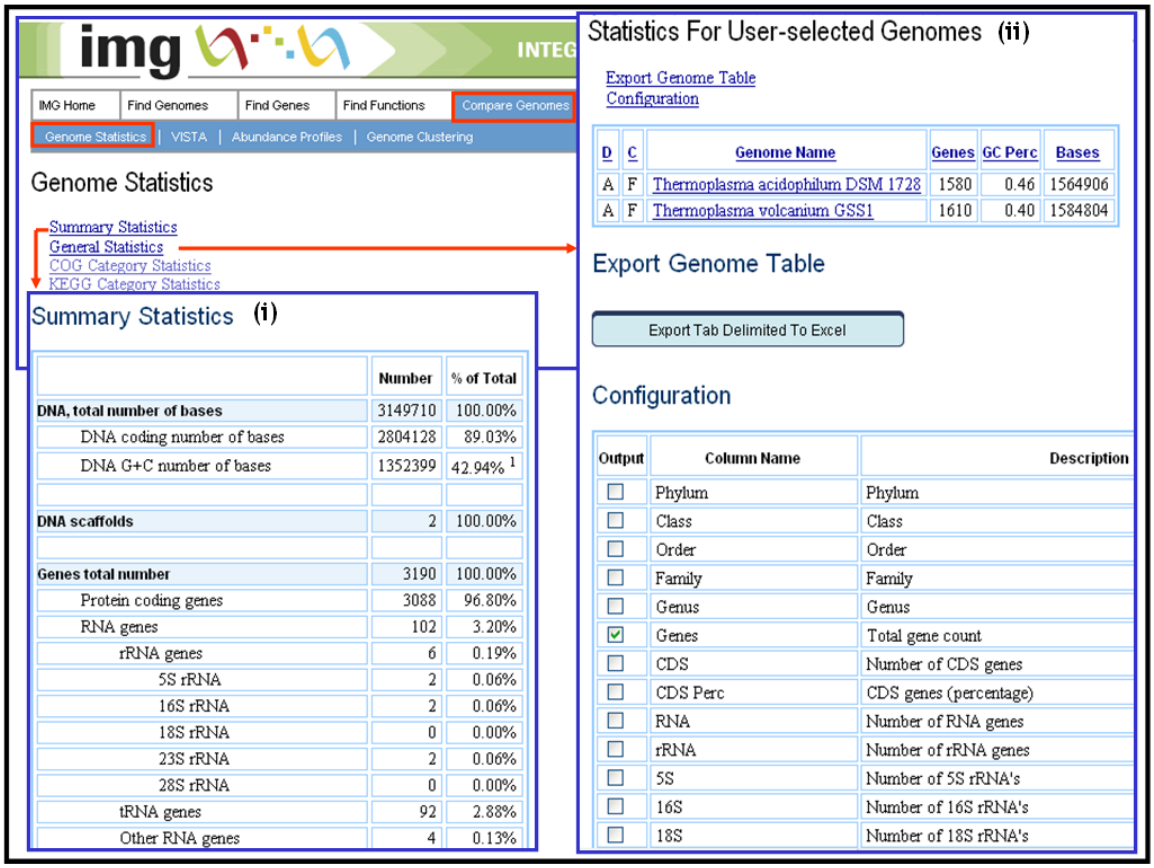


FIGURE 5.1. Genome Statistics: Summary Statistics and Statistics by Genome.

The **COG Category Statistics** and **KEGG Category Statistics** parts of **Genome Statistics** provide statistics in a tabular format for the selected genomes across top-level COG and KEGG categories, respectively, as shown in Figure 5.2(i) and Figure 5.2(ii). Both statistics can be **exported** into a tab-delimited file, and can be configured using a **Configuration** table. Each genome in the statistics table is linked to its **Genome Details** page, which provides statistics for individual genomes, as discussed above. The column headers in these statistics can be used for sorting. Alphabetical columns are sorted in ascending order. Numerical values are usually sorted in descending order (most significant to least significant). Parameter columns can be added or deleted using the **Configuration** selector at the bottom of the page.

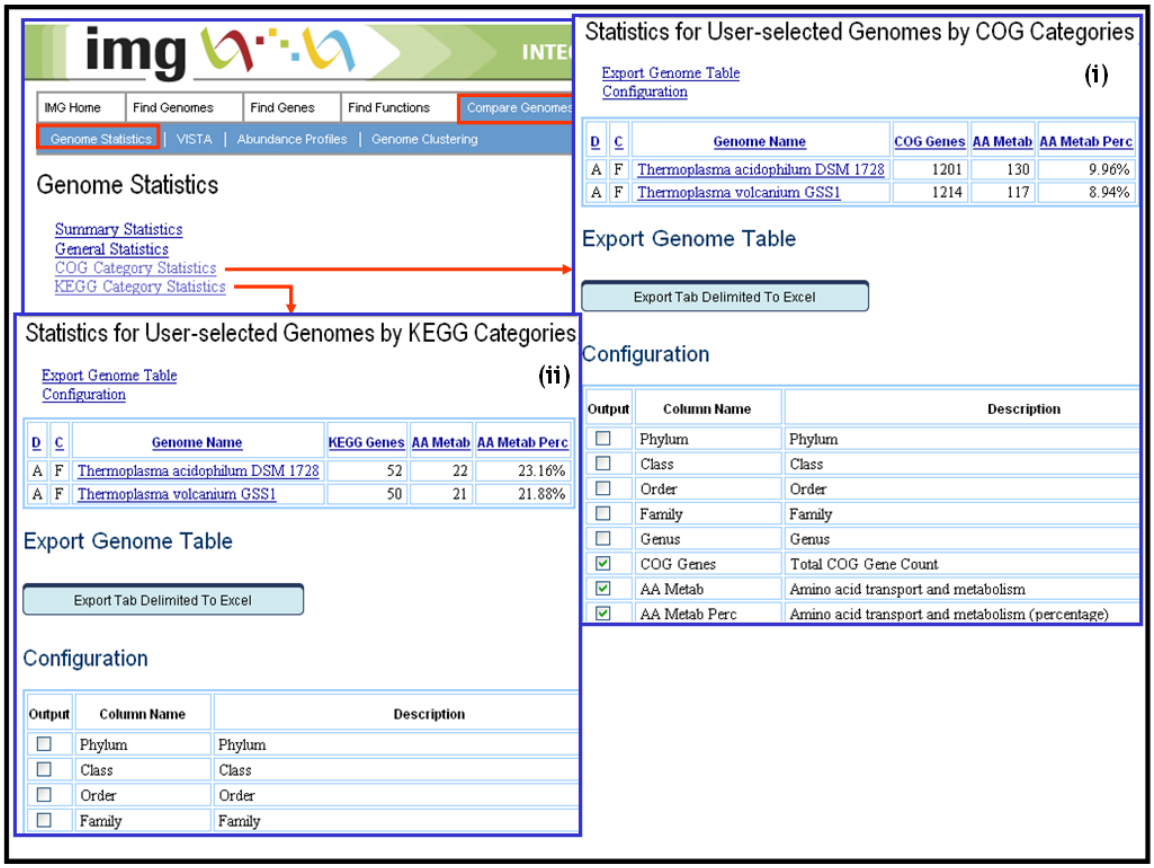


FIGURE 5.2. Genome Statistics: COG Category and KEGG Category statistics.

5.2 Comparing Scaffolds with VISTA

DNA conservation for closely related organisms can be explored in IMG using the **VISTA** comparative genome analysis tool available under the **Compare Genomes** main menu, as shown in Figure 5.3(i). Pre-computed alignments are available for a list of genomes in IMG; selecting a genome invokes the **VISTA** browser that can be then used for examining conservation, as illustrated in Figure 5.3(ii).

The **VISTA** Browser is an interactive Java applet for viewing multiple large-scale alignments, simplifying the identification of regions of high conservation across multiple species. The **VISTA** Browser requires Java 1.2 or better.

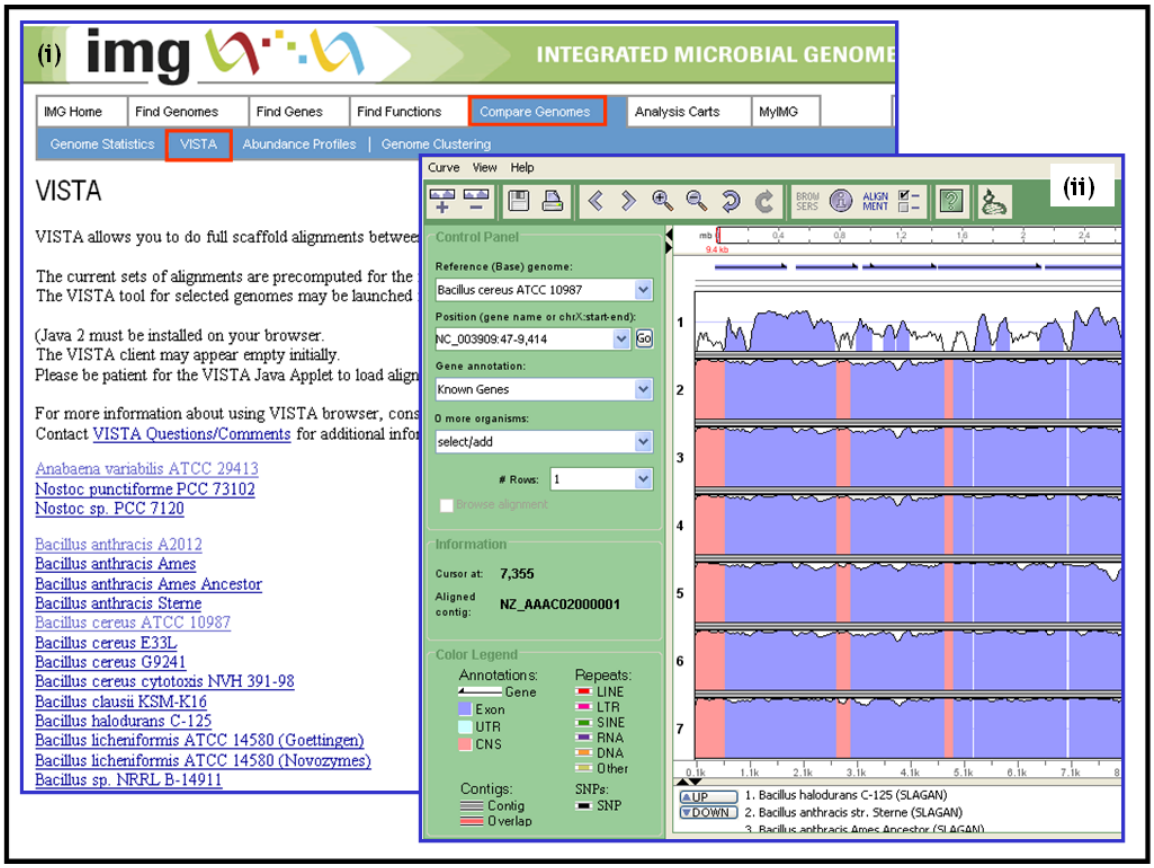


FIGURE 5.3. Comparing Scaffolds with VISTA.

5.3 Abundance Profile Tools

Genomes in IMG can be compared in terms of function abundance using two abundance profile tools available under IMG's **Compare Genomes** main menu option, as shown in Figure 5.4(i).

5.3.1 Abundance Profile Viewer

The **Abundance Profile Viewer** allows comparing selected genomes in terms of their relative abundance across *all* protein families (COGs and Pfams) and functional families (Enzymes).

First, select the type of protein/functional families (COG, Pfam, TIGRfam, Enzyme), normalization method, and a set of genomes in the **Abundance Profile Viewer** page, as shown in Figure 5.4(ii). The abundance of protein/functional families is displayed as a heat map with red corresponding to the most abundant families, as shown in Figure 5.4(iii). Each column on the map corresponds to a genome, and each row corresponds to a family; mouse over each cell to see the count of a particular family in a genome. Click on the cell in order to retrieve the list of genes assigned to this particular family in a genome, as shown in Figure 5.4(iv). Click on the identifier of the family displayed on the right of the column (e.g., COG05331) in order to include the corresponding family into the **Function Cart**, as shown in Figure 5.4(v).

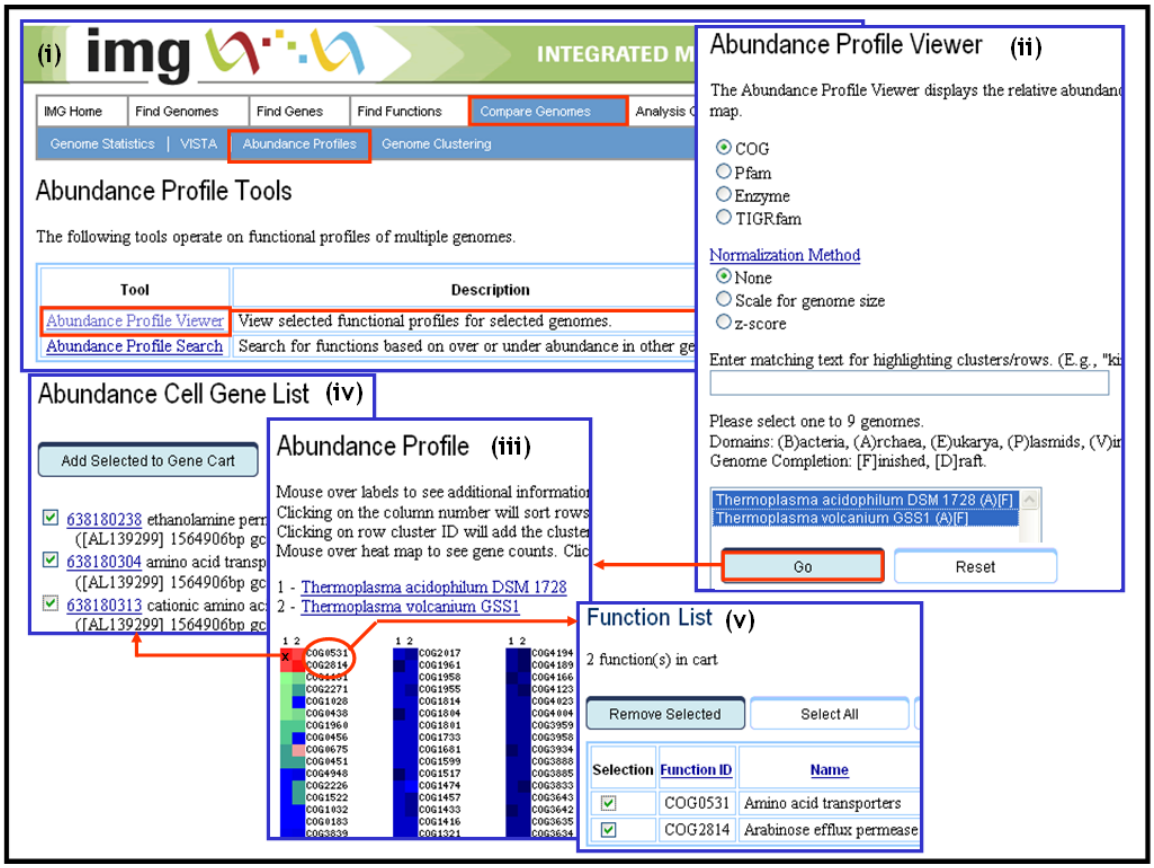


FIGURE 5.4. Abundance Profile Viewer.

By default, the **Abundance Profile Viewer** results are sorted by the abundance of families in the first genome. These results can be resorted according to the abundance in other genomes by clicking on the corresponding column header.

Genomes can be compared using raw gene counts, or can be **normalized** in order to take into account the genome size. For *z-score* normalization, a z-score is computed for each protein/functional family x in a given genome: $z_x = (x - mean_x) / standard.deviation_x$

5.3.2 Abundance Profile Search

The **Abundance Profile Search** tool allows comparing genomes in terms of the relative abundance of protein families (COGs and Pfams).

First, select the type of protein families (COG or Pfam), normalization method, and the type of results (gene counts, normalized values, or both) in the **Abundance Profile Search** page, as shown in Figure 5.5(ii). Abundance cutoffs can be set up for the genomes of interest. The **Abundance Profile Search** results are displayed in a tabular format, as shown in Figure 5.5(iii). The gene counts in the results table are linked to the corresponding lists of genes, as shown in Figure 5.5(iv). Genes can be then selected from these lists and included into the **Gene Cart**. Protein families in the results table can be selected and included into the **Function Cart**, as shown in Figure 5.5(v).

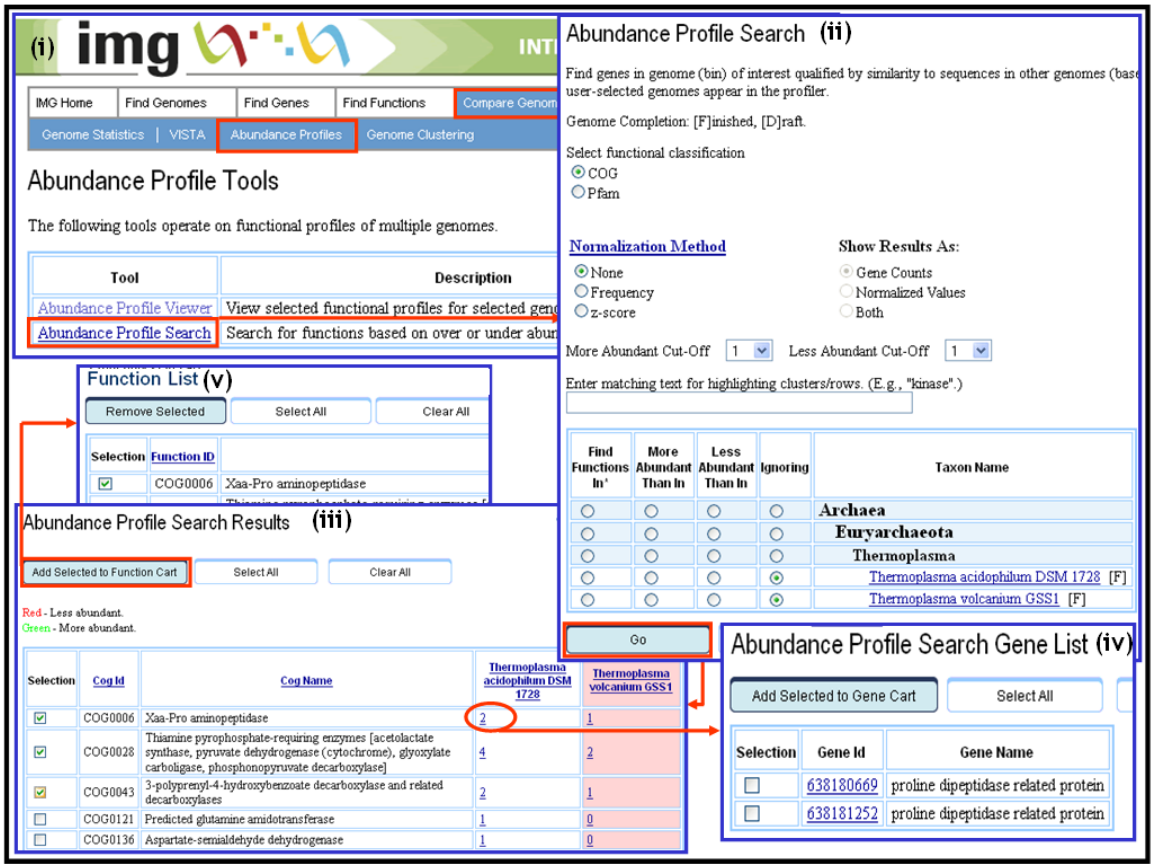


FIGURE 5.5. Abundance Profile Search.

5.4 Genome Clustering Tools

Genomes in IMG can be compared in terms of clusters by using the three clustering tools available under IMG's **Compare Genomes** main menu option, as shown in Figure 5.6(i). Genomes can be clustered by using hierarchical clustering, principal component analysis, or correlation matrix.

Select first the type of protein/functional families (COG, Pfam, Enzyme), clustering method and the genomes you want to compare in the **Genome Clustering** page, as shown in Figure 5.6(i).

The results of **hierarchical clustering** are displayed in a tree format, as shown in Figure 5.6(ii). The placement in the tree reflects the distance between genomes, whereby the computed distance is based on the similarity of the functional characterization of genomes in terms of a specific protein/functional family.

The results of **correlation clustering** are displayed in a matrix format, as shown in Figure 5.6(iii), where each cell of the matrix displays the correlation coefficient between the genomes on the corresponding row and column. The correlation coefficient is computed based on the similarity of the functional characterization of the genomes. The diagonal correlations (of genomes with themselves) are always 1.00. The genomes listed in the clustering results are linked to their associated **Organism Details** pages.

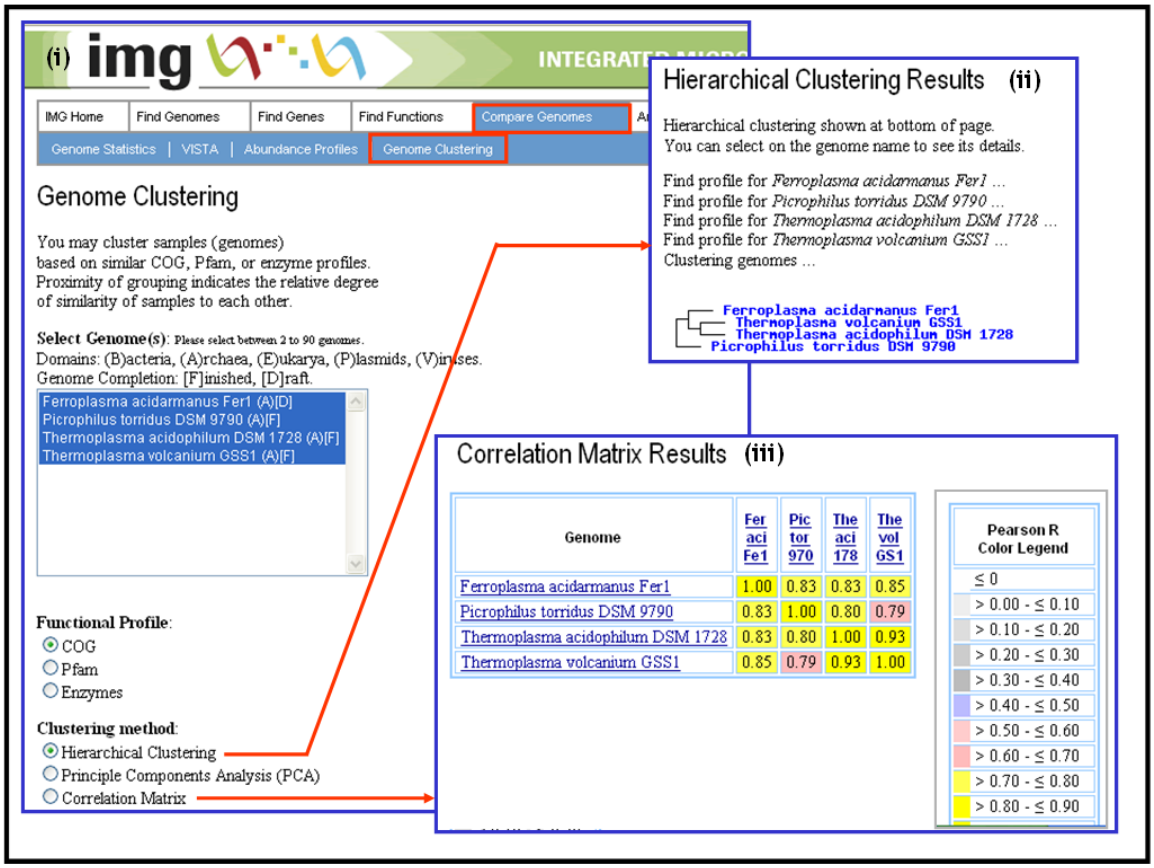


FIGURE 5.6. Genome Clustering.

6 Analysis Carts

Genes and functions in IMG can be managed by using the **Gene Cart** and **Function Cart**, respectively, and compared by using tools associated with these carts.

6.1 Gene Cart

Genes can be managed using the **Gene Cart** available under IMG's **Analysis Carts** main menu option, as shown in Figure 6.1(i).

6.1.1 Gene List

The **Gene Cart** allows users to maintain a **Gene List** that consists of genes selected from the results of various IMG analysis or search tools. When the result of such a tool includes a list of genes, you can click on the checkboxes associated with the genes you want to include into the **Gene Cart** and then click "Add Selected to Gene Cart."

Separate sets of genes can be added to the **Gene List** in **batches**. Each time a set of genes in IMG is added to the **Gene Cart**, a new batch number is generated for this set. The batch number appears in the right-hand column of the **Gene List** table. You can **remove** genes from the cart by selecting them and clicking "Remove Selected."

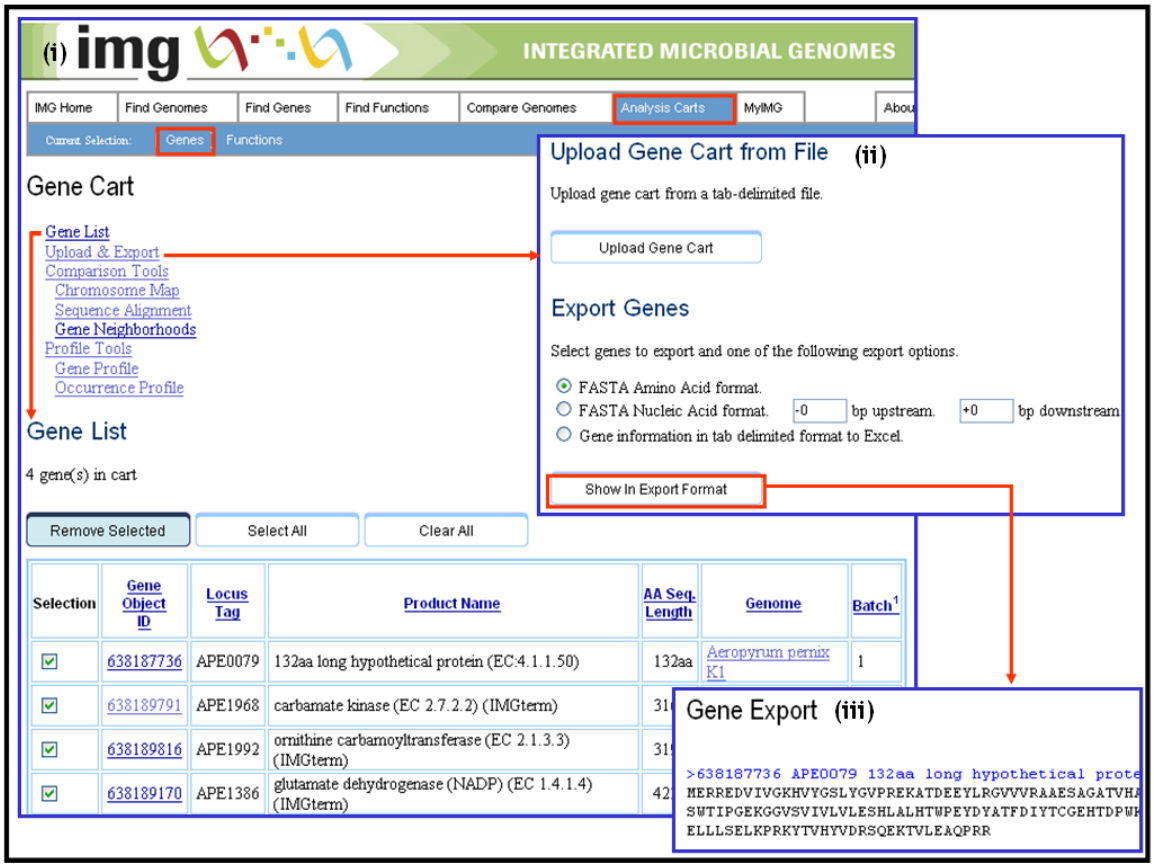


FIGURE 6.1. Gene Cart: Gene List, Upload & Export Genes.

6.1.2 Upload & Export

Genes can be exported to external files in different formats, as shown in Figure 6.1(ii). The export format options include FASTA amino acid format, and FASTA nucleic acid format, with customized upstream and downstream padding.

Genes also can be exported to Excel files in tab-delimited format, where the exported data consists of the gene object identifier, locus tag, gene symbol, and gene name or gene description. Tab-delimited files with this format can be used for uploading genes into the **Gene Cart**. Genes selected for export can be viewed, as illustrated in Figure 6.1(iii), by using the "Show in Export Format" button.

6.1.3 Comparison Tools

The **Gene Cart** provides several tools for comparing genes in its **Gene List**.

Genes selected from the **Gene List** can be displayed on a **Chromosome Map**, as shown in Figure 6.2(ii). Before displaying the map, you can select up to four bands on which the selected genes are drawn, as shown in Figure 6.2(i).

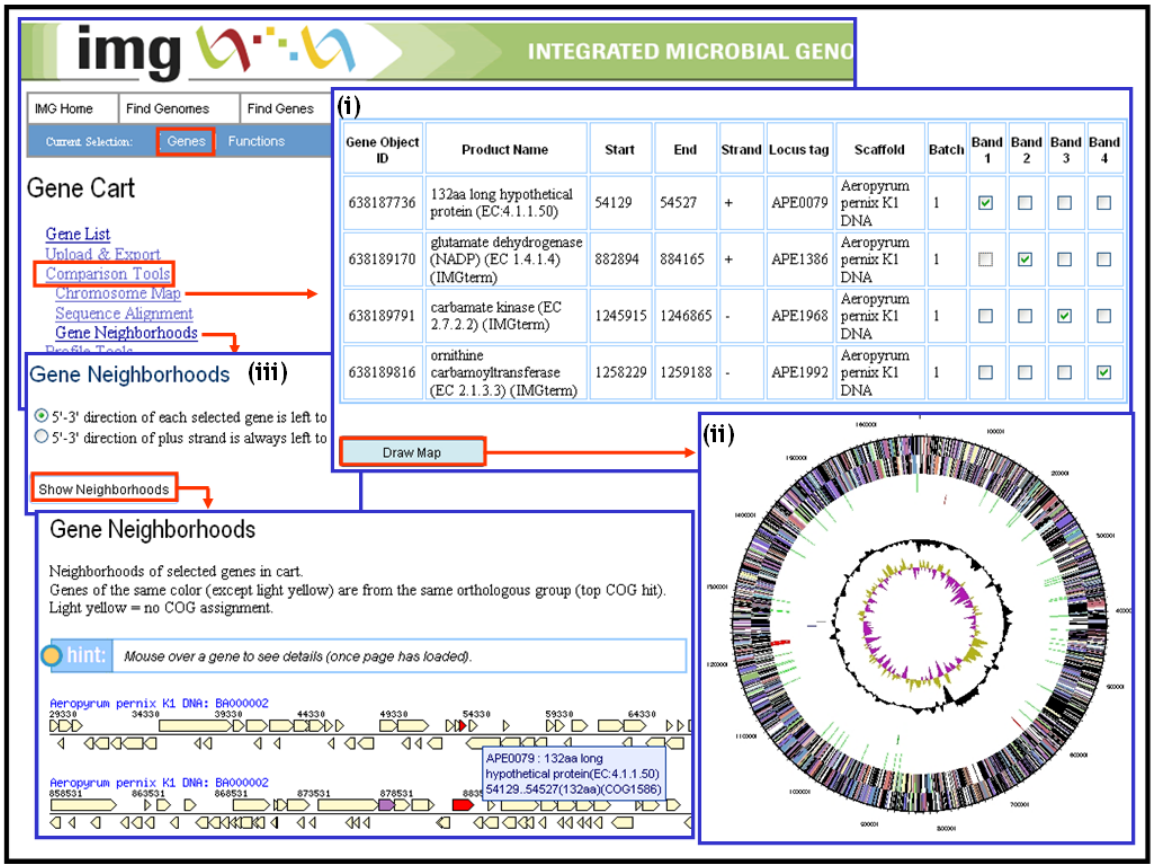


FIGURE 6.2. Gene Cart/Comparison Tools: Chromosome Map and Gene Neighborhoods.

Genes are displayed on the **Chromosome Map** from outside to the center, as follows:

(a) Genes on the forward strand colored by COG categories

(b) Genes on the reverse strand colored by COG categories
(c) RNA genes, with tRNA genes colored green, and sRNA genes colored red
(d) User-selected genes for band 1
(e) User-selected genes for band 2
(f) User-selected genes for band 3
(g) User-selected genes for band 4
(h) GC content
(i) GC skew

6.1.4 Sequence Alignments

The sequences of genes selected from the **Gene List** can be aligned using **ClustalW** alignment, as shown in Figure 6.3(i). You have the option to align either the DNA or the protein sequence. For DNA, you can also extend the region upstream and downstream for the alignment. By default, the ClustalW output is formatted by **Mview**, a multiple-sequence alignment formatting tool, as illustrated in Figure 6.3(ii). This option can be turned off. The "JalView" button at the bottom of the ClustalW output page allows examining a graphical representation of the alignment quality at each amino acid position using **JalView**. Note that you need to have Java enabled in your browser to use JalView.

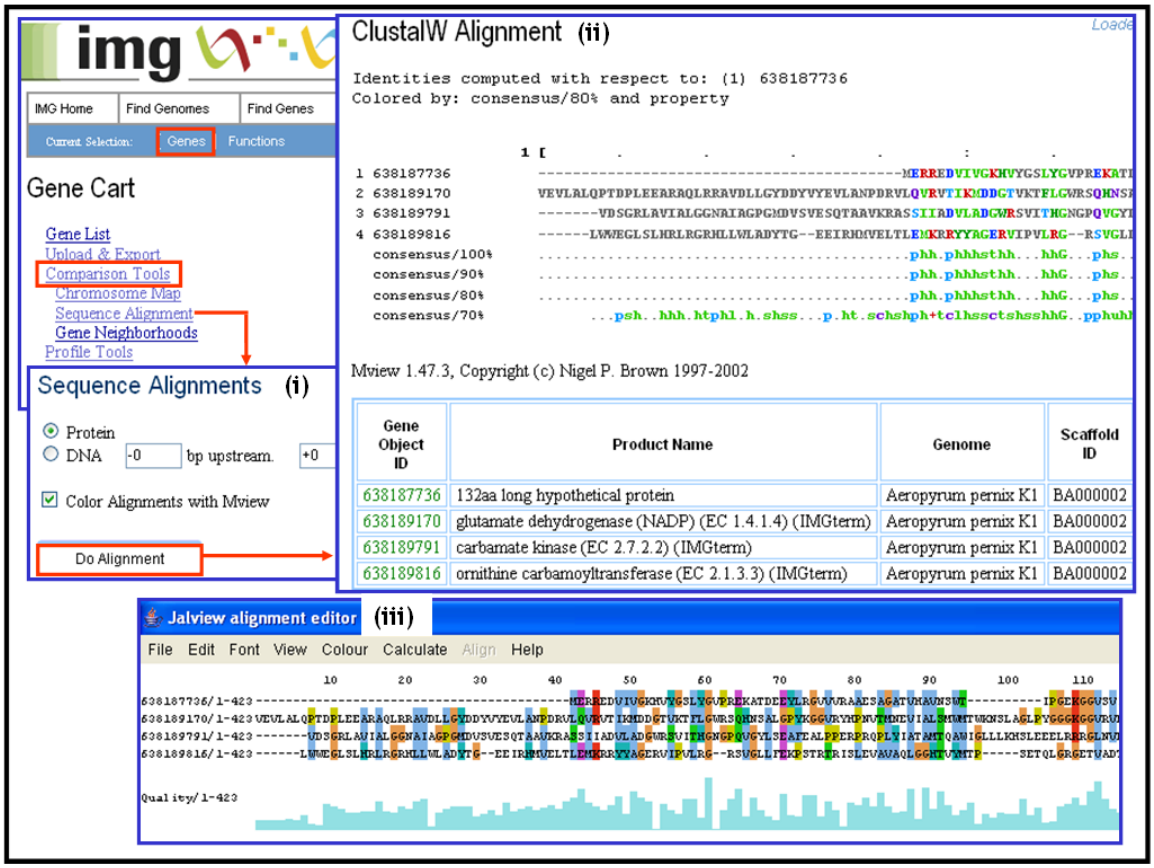


FIGURE 6.3. Gene Cart/Comparison Tools: Sequence Alignment.

6.1.5 Gene Neighborhoods

Genes selected from the **Gene List** can be displayed using the **Gene Neighborhoods** viewer, as shown in Figure 6.2(iii). The neighborhood orientation option determines whether a given gene's neighborhood is shown with the plus strand reading left to right or right to left. Selecting the "5'-3' direction of each selected gene is left to right" option will show all the genes oriented in the same direction. Selecting the "5'-3' direction of plus strand is always left to right, on top" option will show all the strands oriented in the same direction.

The number of neighborhoods that can be shown is limited; to change the limit, enter a new value under "Max. Taxon Gene Neighborhoods" with **MyIMG Preferences**.

The **Gene Neighborhoods** viewer shows each target gene and other genes within 25 kilobases from the target gene. Each gene's neighborhood appears above and below a single line showing the genes reading in one direction on top, and those reading in the opposite direction on the bottom. Genes with the same color, except for the default light yellow, are from the same COG group. The target gene always appears in red at the center of the neighborhood. When you move the cursor over a gene, you will see a popup box with the locus tag, scaffold coordinates, and COG group number (except in Internet Explorer for Macintosh). You can click the arrow that represents a gene to bring up the **Gene Details** page for that gene.

6.1.6 Profile Tools

Gene profile tools can be applied on genes selected from the **Gene List.**

The **Gene Profile** tool displays the number (abundance) of homologs of the genes selected from the **Gene List** across genomes selected from a list, as shown in Figure 6.4(i). The percent identity and e-value cutoffs can be set to determine the homologs.

The **Gene Profile** results are displayed in a tabular format with each row displaying the profile of a specific gene across the selected genomes, as illustrated in Figure 6.4(ii). Alternatively, you can display the profile of a specific gene across selected genomes in each column, as illustrated in Figure 6.4(iii). Each cell in the profile-result table displays the count (abundance) of homolog genes in a genome and contains a link to the associated list of genes, as shown in Figure 6.4(iv). Colors are used to represent gene abundance, whereby white, bisque, and yellow represent gene counts of 0, 1–4, and over 4, respectively.

The **Occurrence Profile** tool displays the pattern of occurrence for the genes selected from the **Gene List** across multiple genomes, as illustrated in Figure 6.4(v). For each gene, a fixed-length ordered vector is displayed in a BLAST-like alignment format. The positions of the vector correspond to the list of genomes selected with the **Genome Browser** (all genomes by default), with the genomes ordered phylogenetically. This ordering facilitates examining similar profiles in phylogenetically close genomes.

Presence of a gene or its homolog in a given genome is indicated by a domain letter—“B” for Bacteria, “A” for Archaea, and “E” for Eukarya—while the absence of the gene is indicated by a dot (“.”). You can mouse over the letter or dot to see the genome name along with its phylum.

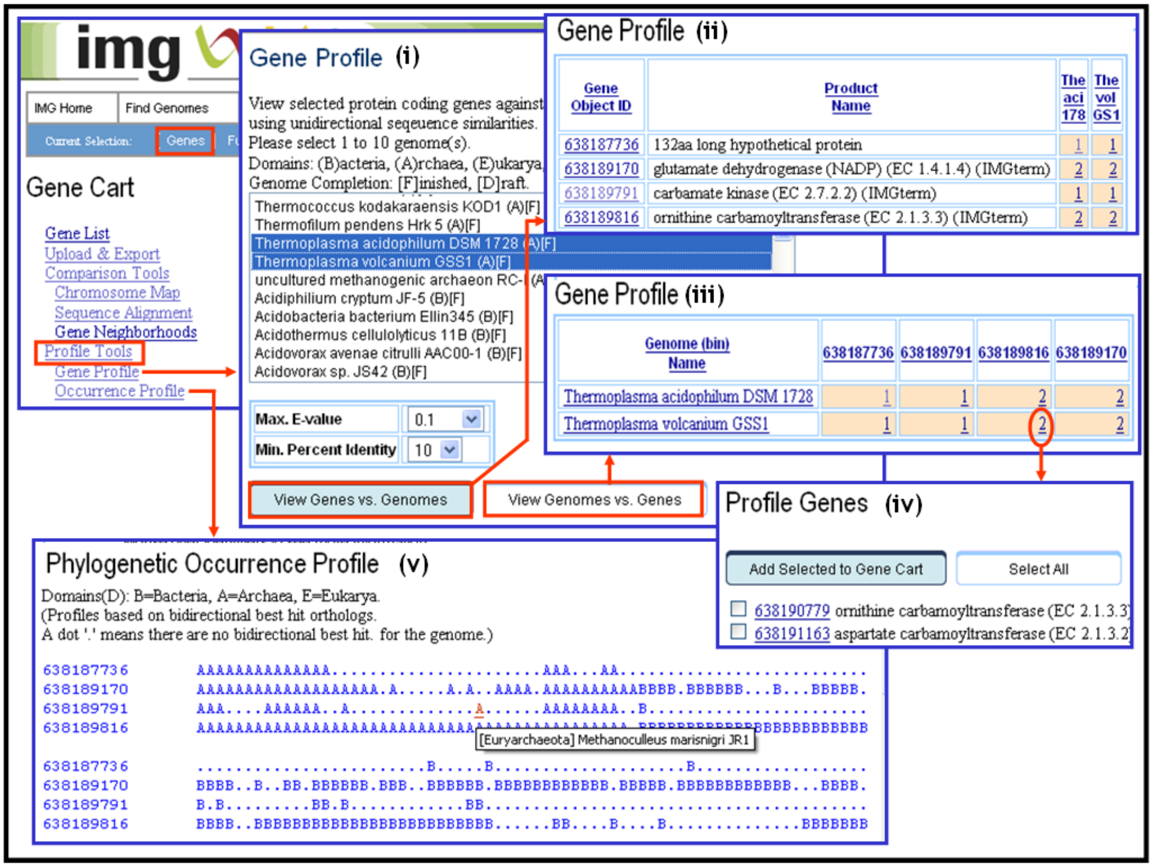


FIGURE 6.4. Gene Cart: Profile Tools.

6.2 Function Cart

Functions can be managed using the **Function Cart** available under IMG's **Analysis Carts** main menu option, as shown in Figure 6.5(i).

6.2.1 Function List

The **Function Cart** allows the user to maintain a **Function List** that consists of functions, such as COGs, Pfams, Enzymes, IMG terms, and IMG pathways, that have been selected from results of various IMG analysis or search tools. When the result of such a tool includes a list of functions, you can click on the checkboxes associated with functions you want to include into the **Function Cart,** and then click "Add Selected to Function Cart."

Separate sets of functions can be added to the **Function List** in **batches**. Each time a set of functions in IMG is added to the **Function Cart**, a new batch number is generated for this set. The batch number appears in the right-hand column of the **Function List** table. You can **remove** functions from the cart by selecting them and clicking "Remove Selected."

6.2.2 Upload & Export

Functions can be exported to an external file in tab-delimited format, as shown in Figure 6.5(ii). Tab-delimited files can be used for uploading functions into the **Function Cart**.

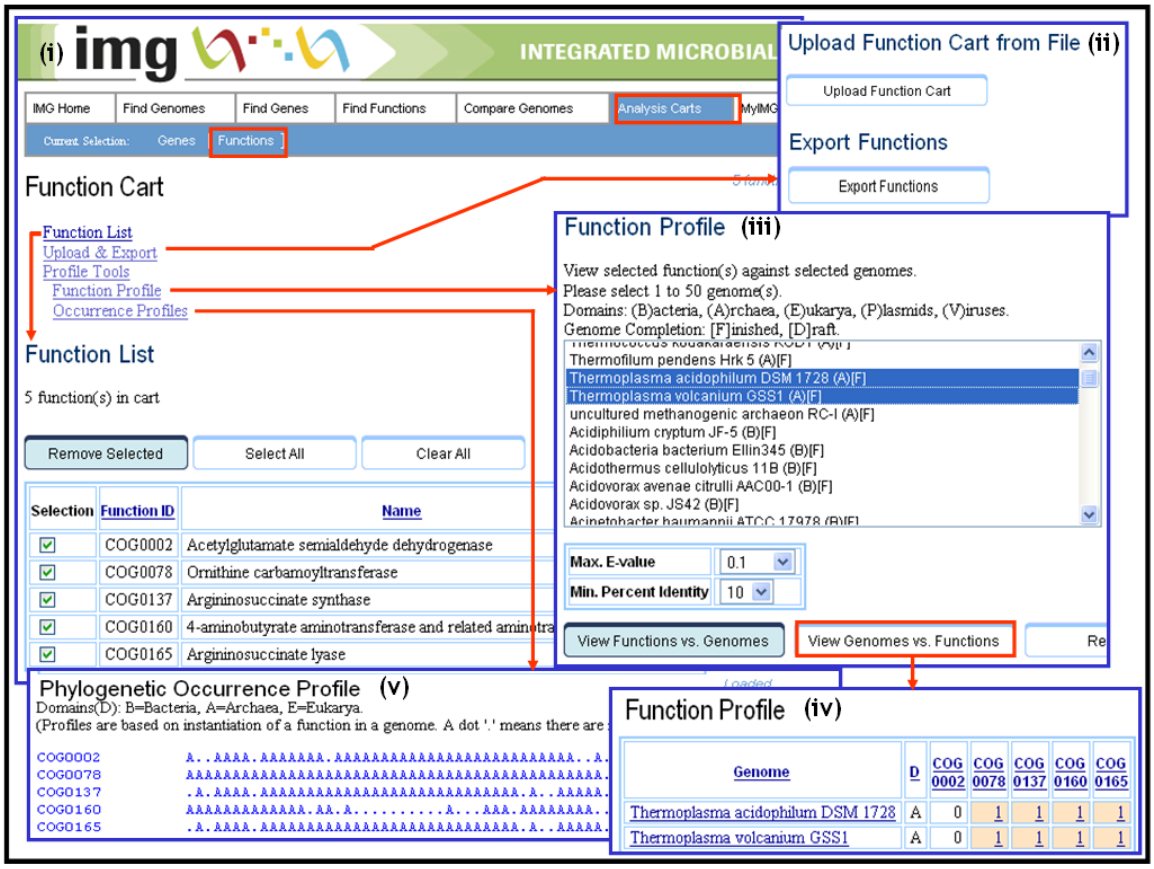


FIGURE 6.5. Function Cart.

6.2.3 Profile Tools

Function profile tools can be applied on functions selected from the **Function List.**

The **Function Profile** tool displays the number (abundance) of genes associated with a function selected from the **Function List** across genomes selected from a list, as shown in Figure 6.5(iii).

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 Figure 6.5(iv). Alternatively, you can display the profile of a specific function across genomes using rows. 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, similar to the list shown in Figure 6.4(iv). Colors are used to represent gene abundance, whereby white, bisque, and yellow represent gene counts of 0, 1–4, and over 4, respectively.

The **Occurrence Profile** tool displays the occurrence pattern for the functions selected from the **Function List** across multiple genomes, as illustrated in Figure 6.5(v). For each function, a fixed-length ordered vector is presented in a BLAST-like alignment format. The positions of the vector correspond to the list of genomes selected with the **Genome Browser** (all genomes by default), with the genomes ordered phylogenetically.

Presence of a function in a given genome is indicated by a domain letter—"B" for Bacteria, "A" for Archaea, and "E" for Eukarya—while the absence of the function is indicated by a dot (".”). You can mouse over the letter or dot to see the genome name along with its phylum.

7 MyIMG

MyIMG is available on IMG's top menu, as shown in Figure 7.1(i). **MyIMG** has three sections. The first section allows you to set system-wide preferences. The second section allows you to upload your genome selections, previously exported from the **Genome Browser** or **Genome Statistics**. The third section requires login to a specialized version of IMG, called **IMG ER** (Expert Review), which provides support for user annotations. This section is not covered in this manual.

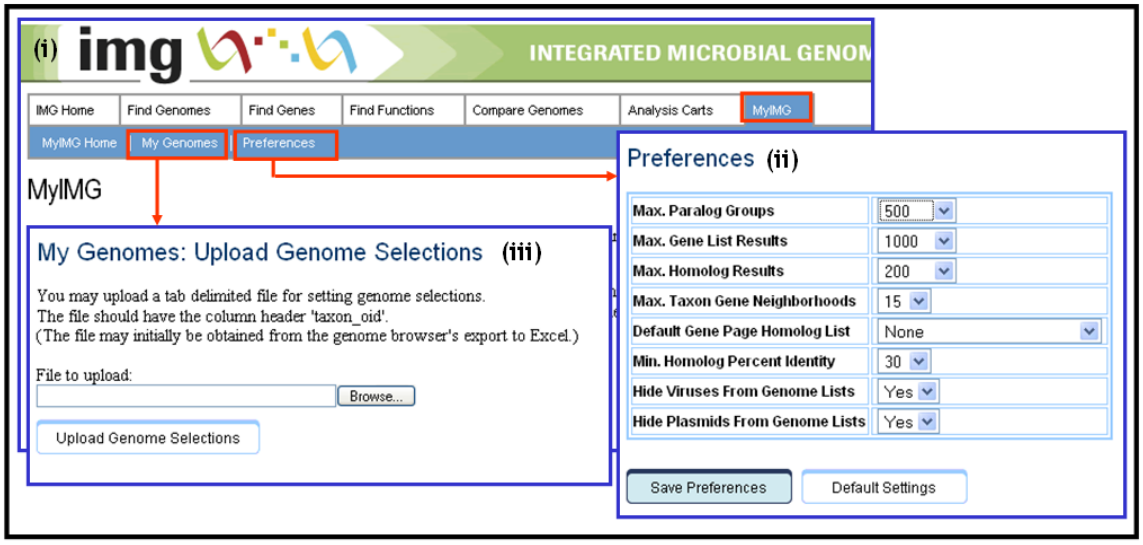


FIGURE 7.1. MyIMG.

7.1 Setting Preferences

Results of analysis or search tools in IMG may consist of potentially very long lists of genes, genomes, or functions. The **Preferences** section in **MyIMG,** shown in Figure 7.1(ii), allows you to set the following system-wide defaults:

1. "Max. Paralog Groups" sets the maximum number of rows displayed on the "Genes in Paralog Clusters" page, accessible from the **Genome Details** page. The default number is 500.
2. The "Max. Gene List Results" sets the maximum number of rows displayed in a gene list, such as the result of a **Gene Search**. The default number is 1000.
3. The "Max. Homolog Results" sets the maximum number of rows displayed in the list of homologs in the **Gene Details** page. The default is top 200 hits.
4. The "Max. Taxon Gene Neighborhoods" sets the maximum number of neighborhoods of orthologous genes displayed with a **Gene Neighborhood** viewer. The default number is 15.
5. The "Default Gene Page Homolog List" determines how the homolog list is presented in the **Gene Details** page. The default value is "Paralogs/Ortholog Clusters," showing the ortholog list clustered and summarized with ortholog counts displayed for each cluster. This default allows a more compact and faster-

loading page. If you want the homolog list displayed a different way, select one of the other options available: “Paralogs/Orthologs,” “Homolog Alignments,” “Phenotype,” “Ecotype,” “Disease,” and “Relevance.”

6. "Min. Homolog Percent Identity" sets the minimum threshold value for similarity between homologs that are shown. The default value is 30.
7. "Hide Viruses From Genome Lists" hides viruses from the **Genome Browser** and genome selection lists. The default is "Yes."
8. "Hide Plasmids From Genome Lists" hides isolate plasmids from **Genome Browser** and genome selection lists. The default is "Yes."

You must click "Save Preferences" for your selections to take effect.

7.2 MyGenomes

You can load your list of genomes into IMG, as illustrated in Figure 7.1(iii). Click on the "Browse" button to find the local file that was previously exported in tab-delimited format from the **Genome Browser** or **Genome Statistics**. The file should have at least the column header "taxon_oid." Click on the "Upload Genome Selections" button to set the current genome selection to the list of genomes in your tab-delimited file.

Glossary of Terms

Enzyme – A protein catalyzing a biochemical transformation (i.e., accelerating a chemical reaction).

Fusion – A hybrid gene formed from two previously separate genes; components of a fusion gene are the separate genes, while a composite gene is the result of the gene fusion.

Gene (or **protein**) **annotation** – A description of the gene or protein product in the molecular, cellular, and phenotypic context (e.g., interactions of a protein with other proteins or metabolites, participation of a protein in a biochemical pathway, or the effect of a gene knockout on the phenotype of an organism).

Genome context of a gene – A set of parameters defining the spatial position of a gene on the chromosome or a plasmid in a certain genome, including its co-localization with other genes, regulatory elements in its proximity, location of a gene on the leading or lagging DNA strand, etc.

Gene symbol – A unique abbreviation of a gene name consisting of italicized upper-case Latin letters and Arabic numbers, assigned after a gene has been identified.

Locus tag – A systematic gene identifier that is assigned to each gene in a Genbank file. For details see http://www.ncbi.nlm.nih.gov/Genbank/genomesubmit.html#locus_tag

Metabolism – A set of chemical transformation taking place within a living cell, multicellular organism, or a microbial community.

Metabolic network – A representation of metabolism as a graph with nodes corresponding to metabolites and edges representing the reactions (or enzymes catalyzing the reactions).

Metabolic pathway – A set of consecutive biochemical transformations (enzymatic and spontaneous reactions) taking place in a living cell.

Homologous genes (homologs) – Genes with sequence similarity (either at the level of nucleotide sequence or at the level of amino acid sequence of their protein products) due to their shared ancestry.

Orthologous genes (orthologs) – Genes with sequence similarity separated by speciation events or vertically inherited genes: if a gene existed in a species, which gave rise to two species, then the divergent copies of the gene in the resulting two species are orthologous.

Paralogous genes (paralogs) – Genes with sequence similarity separated by duplication events.

Operon – A group of genes sharing the common regulatory elements (promoter, operator, terminator) and transcribed as a unit to produce a single messenger RNA.

Regulon – A group of genes and operons in an organism under regulation of the same regulatory protein.

A detailed **Glossary of Terms** is available at: http://ghr.nlm.nih.gov/ghr/page/Glossary