

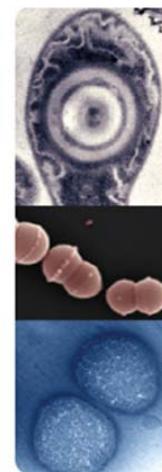


IMG Genomes

	finished/draft	Total
Bacteria	632/446	1078
Archaea	53/3	56
Eukarya	19/21	40
Plasmids	803/0	803
Viruses	2230/0	2230
All Genomes	3737/470	4207

[IMG Statistics](#)

Using IMG



Comparative Analysis with the Integrated Microbial Genomes System

Technical Report LBNL-63614

Genome Biology Program

Department of Energy Joint Genome Institute

Biological Data Management and Technology Center

Lawrence Berkeley National Laboratory

August 1, 2008

Copyright 2008 The Regents of the University of California

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.

August 1, 2008

©2008 The Regents of the University of California

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

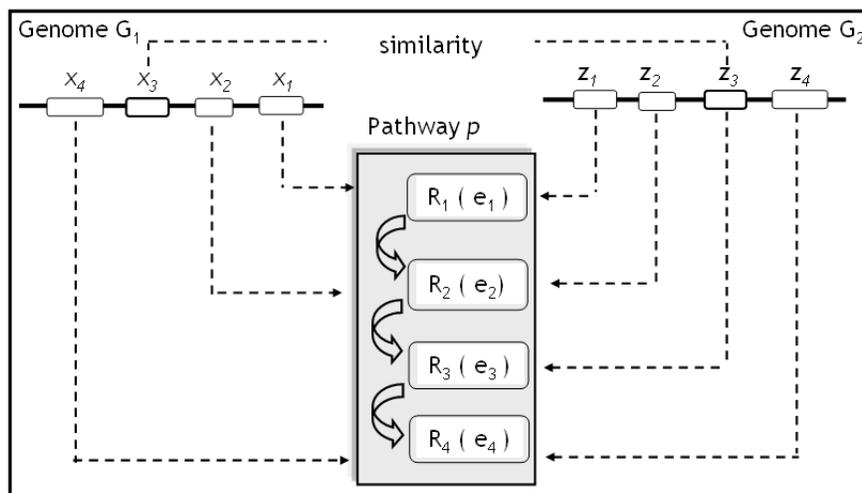
1 IMG Overview.....	1
1.1 Application View	1
1.2 IMG Data Model.....	2
1.3 IMG Data Content.....	3
1.3.1 Genomic Data	3
1.3.2 Functional Annotations.....	3
1.3.3 Gene Relationships	4
1.3.4 Parameters.....	5
1.4 IMG Data Analysis	6
References.....	8
2 Find and Examine Genomes	10
2.1 Genome Browser	10
2.2 Genome Search	11
2.3 Organism Details	12
2.3.1 Organism Information.....	12
2.3.2 Genome Statistics.....	13
2.3.3 Genome Viewers.....	14
2.3.4 Export Genome Data.....	16
3 Find and Examine Genes	17
3.1 Gene Search	17
3.2 BLAST Search.....	18
3.3 Gene Selection from Genome Statistics	19
3.4 Gene Details.....	20
3.4.1 Gene Information.....	20
3.4.2 Evidence for Function Prediction	23
3.4.3 Sequence Search	25
3.4.4 Homolog Display	26
3.5 Phylogenetic Profilers.....	28
3.5.1 Phylogenetic Profiler for Single Genes.....	28
3.5.2 Phylogenetic Profiler for Gene Cassettes	29
4 Find and Examine Functions	31
4.1 Function Search	31
4.2 Function Browsers	32
4.3 Function Details.....	33
4.3.1 COG Details.....	33
4.3.2 Pfam Details.....	34
4.3.3 KEGG Pathway Details	35
4.3.4 TIGRfam Details.....	36
4.3.5 IMG Networks Details.....	37
5 Compare Genomes.....	40
5.1 Genome Statistics.....	40
5.2 Comparing Scaffolds with VISTA.....	42
5.3 Abundance Profile Tools	43

5.3.1 Abundance Profile Overview.....	43
5.3.2 Abundance Profile Search.....	45
5.4 Genome Clustering Tools	46
6 Analysis Carts.....	47
6.1 Gene Cart	47
6.1.1 Gene List.....	47
6.1.2 Upload & Export.....	48
6.1.3 Comparison Tools.....	48
6.1.4 Sequence Alignments.....	49
6.1.5 Gene Neighborhoods	49
6.1.6 Profile Tools.....	50
6.2 Function Cart	51
6.2.1 Function List	51
6.2.2 Upload & Export.....	52
6.2.3 Profile Tools.....	52
7 MyIMG	54
7.1 Setting Preferences.....	54
7.2 MyGenomes	55
Glossary of Terms	56

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, \dots, 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 **COGs** 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.

IMG ortholog 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.

A **chromosomal cassette** is defined as a stretch of protein coding genes with intergenic distance smaller or equal to 300 base pairs. The genes must be on the same strand or divergent; convergent genes are not included in a chromosomal cassette. Groups of at least two common genes between two or more chromosomal cassettes are defined as **conserved chromosomal cassettes**. In order to identify common genes between chromosomal cassettes, genes need to be assigned to groups of equivalent genes. For this grouping, the commonly accepted clusters of orthologous genes (COG), Pfam assignments, and IMG ortholog clusters (see above) were used. If a protein consists of multiple clusters, such as in a gene fusion or multiple Pfam domains, each individual domain is included in the chromosomal cassette.

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. $\geq 70\%$ 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 with additional filtering options.
RNA all vs. all	blastall - p blastn	1e-5	None	None	250	$\geq 50\%$ 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.

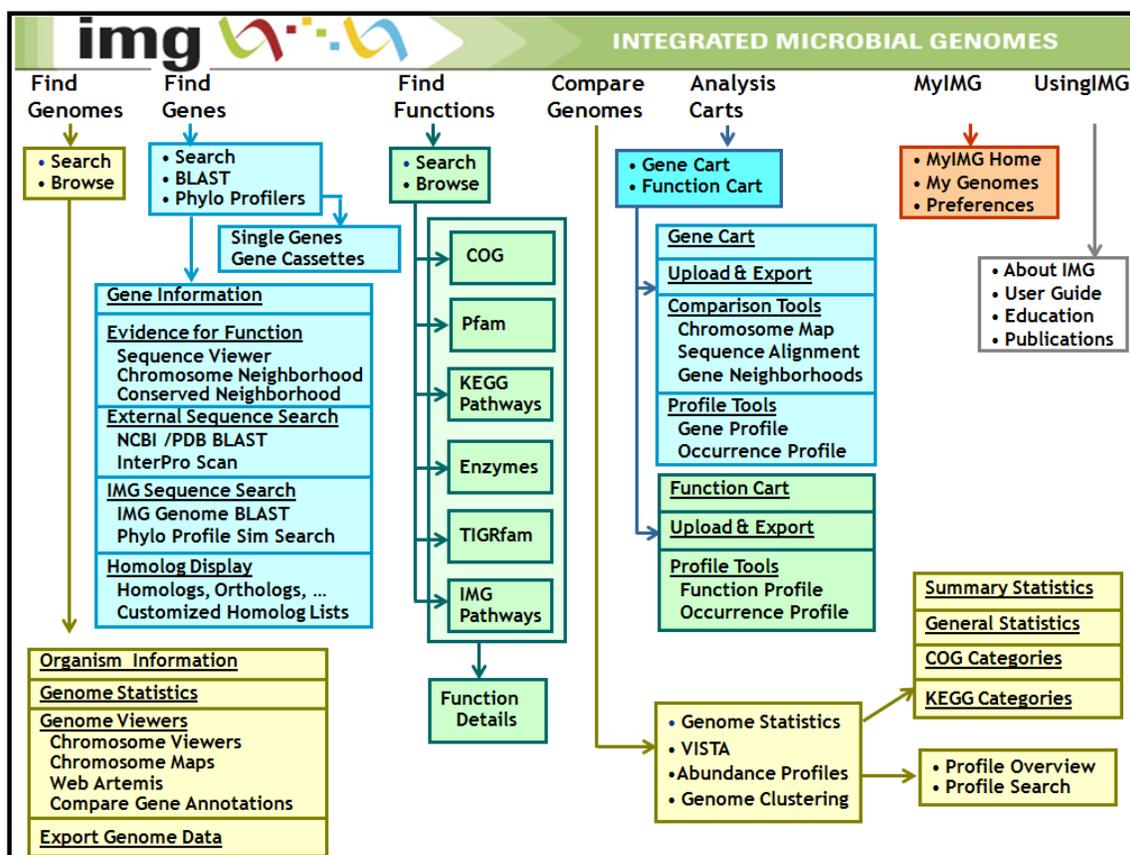


FIGURE 1.1. IMG user interface map.

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**, and **IMG Cluster** sections; **Evidence for Functional Prediction** with viewers for chromosomal neighborhood, ortholog neighborhoods, and chromosomal cassettes, and COG and Pfam hits. A gene can be searched against external database or IMG genes, and can be examined in terms of pre-computed homologs, orthologs and paralogs.

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 Overview** which provides the relative abundance for **all** protein families (COGs or Pfams) or functional families (Enzymes) across selected genomes, with the option of displaying the abundance of protein/functional families either as a heat map or in a tabular (matrix) format. Another tool, the **Abundance Profile Search**, 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 can also 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.

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**. Two **Phylogenetic Profiler** tools allow identifying *single* or *collocated* genes (i.e., genes that are part of a **gene cassette**) in a query genome

based on presence/absence of homologs in other genomes (for single genes) or presence of related collocated genes in other genomes.

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.

The screenshot displays the Integrated Microbial Genomes (IMG) interface. At the top, the 'IMG Genomes' summary table shows counts for various domains:

finished/draft	Genes	Total
Bacteria	142/74	485/243
Archaea	13/5	41/5
Eukarya	2/0	19/21
Plasmids	0/0	40/20
Viruses	0/0	1861/0
All Genomes	157/79	2809/263
Grand Total	236	2878

The 'Genome Browser' section (ii) shows a list of genomes with columns for 'Select', 'D' (Domain), 'C' (Completion), 'Genome Name', 'Sequencing Center', 'Gene Count', and 'Genome Size'. Three genomes are listed:

Select	D	C	Genome Name	Sequencing Center	Gene Count	Genome Size
<input checked="" type="checkbox"/>	A	F	Aeropyrum pernix K1	NITE	2815	1669695
<input checked="" type="checkbox"/>	A	F	Archaeoglobus fulgidus DSM 4304	Univ of Illinois at Urbana-Champaign, TIGR	2519	2178400
<input checked="" type="checkbox"/>	A	D	Caldivirga maquilungensis IC-167	JGI	1986	2077575

The 'Configuration' panel (iv) allows users to select additional output columns, with 'Sequencing Center', 'Gene Count', and 'Genome Size' currently checked.

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”.

Genome Search by Fields

Find genomes by keyword or substring.

Keyword:

Filters: Genome Name

Go Reset

Genome Search by Metadata (ii)

Select category search values.

Phenotype and

- Acetogen
- Acetotrophic
- Acidophile
- Alkaliphile
- Alkalitolerant
- Alkaliphile
- Alkane degrader
- Alpha-hemolytic
- Ammonia-oxidizer
- Amylase production

Habitat and

- Colon
- Extremotypes
- Acid mine drainage
- Antarctic
- Arctic marine sediments
- Hydrothermal vent
- Permafrost sediment
- Solfataric field
- Feces
- Food

Disease and

- Abortion
- Acrodermatitis chronica atrophicans
- Anthrax
- Arthritis
- Bacillary angiomatosis
- Bartonellosis = Carrion's disease
- Croya fever
- Blood infection
- Bloodborne infection
- Hepatitis

Relevance

- Agricultural
- Plant Pathogen
- Pest control
- Anti-tumor agent
- Astrobiological
- Avian Pathogen
- Bacterial Pathogen
- Bioenergy
- Biofuels
- Biogeochemical

Genome Metadata Search Results (iii)

2 genomes re

Save Selections Select All Clear All

hint: Selections do not take effect until you save them. You must select at least one genome. Go to Preferences to show or hide plasmids and viruses. Go to Home page statistics under IMG Genomes to select individual phylogenetic domains or all genomes.

Genome Completion: [F]inished, [D]raft.

Select	Genome Name	Phenotype	Habitat
<input type="checkbox"/>	Acidiphilium cryptum JP_5	Acidophile, Aerobe, Heterotroph, Iron reducer, Motile, Rod-shaped	Acid mine drainage, Aquatic, Hydrothermal vent
<input type="checkbox"/>	Thermoplasma volcanium GSS1	Acidophile, Facultative, Motile, Nonsporulating, Rod-shaped, Thermophile	Aquatic, Hydrothermal vent, Solfataric field

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.

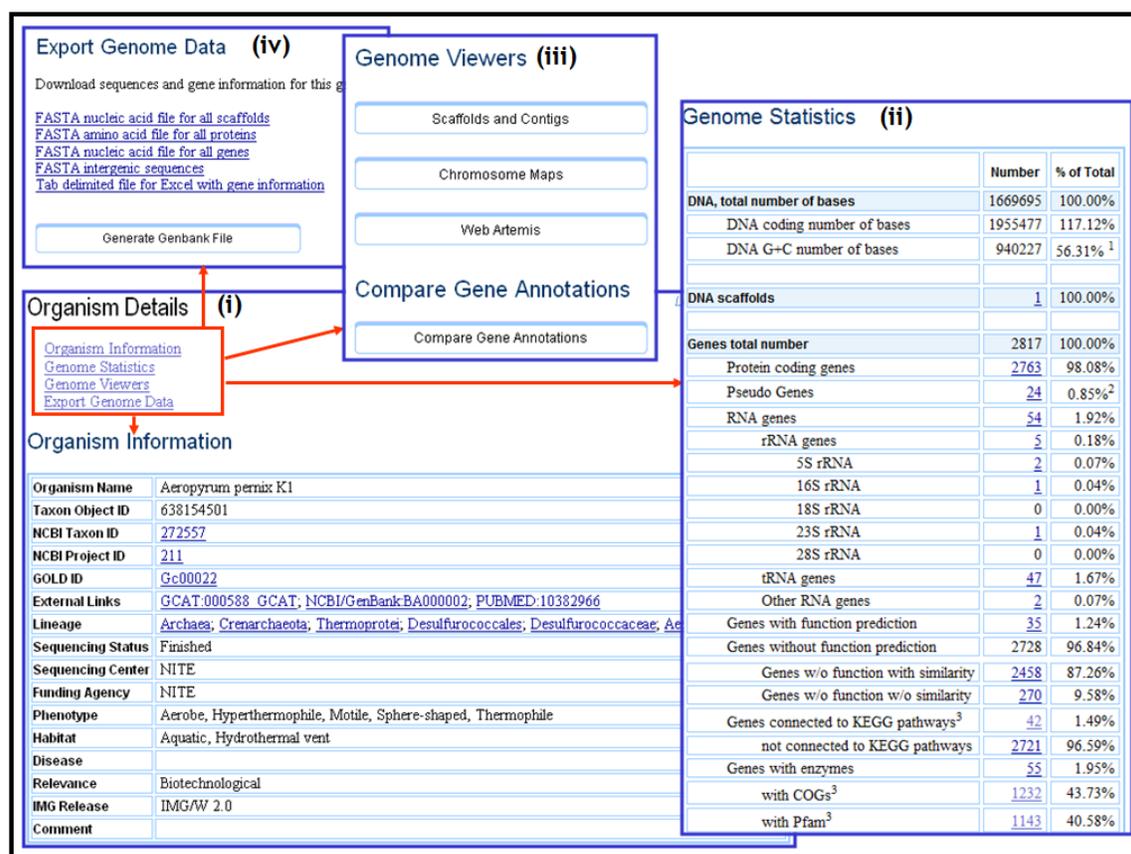


FIGURE 2.3. Organism Details: Organism Information, Genome Statistics, Exploration Tools, Export Genome Data.

Every count of genes is linked to the list of associated genes grouped, when possible, 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).

The screenshot displays the 'Genome Viewers' section of the IMG 2.6 interface. It includes several panels and tools:

- Genome Viewers Panel:** Contains links for 'Scaffolds and Contigs', 'Chromosome Maps', 'Web Artemis', 'Compare Gene Annotations', and 'Compare Gene Annotations'.
- Chromosome Viewer (i):** Shows 'Scaffolds and contigs for Aeropyrum pernix K1'. It includes a table of 'User Selectable Coordinates' and 'User Enterable Coordinates'.
- Chromosome Viewer (ii):** Shows 'Aeropyrum pernix K1 DNA: BA000002 (1669695bp gc=0.56) (coordinates 1-500000)'. It features a 'Switch coloring to: GC percentage' button and a 'Chromosome Viewer' panel with 'Switch coloring to: COG' and 'Characteristic GC% - 56 %'.
- Artemis Entry Edit: 8440.art (iv):** Shows a detailed view of a gene entry with annotations and a protein sequence.

Scaffold	Length (bp)	GC	No. Genes	Coordinate Range
Aeropyrum pernix K1 DNA: BA000002	1669695	0.56	2815	1_500000 500001_1000000 1000001_1500000 1500001_1669695

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. Gene coloring can be switched to reflect deviation from characteristic **GC percentage** for that genome, as illustrated in Figure 2.4(ii).

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:

- Genes on the forward strand colored by COG categories
- Genes on the reverse strand colored by COG categories
- RNA genes, with tRNA genes colored green, and sRNA genes colored red
- GC content
- GC skew

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

The screenshot displays the 'Compare Gene Annotations' interface for *Aeropyrum pernix K1*. It features three panels, each showing a table of gene annotations with columns for Gene Object ID, Locus Tag, Source, Cluster Annotation, and Gene Annotation. Panel (i) shows a filter dropdown menu with options: 'None', 'With product name/No Evidence', and 'No Product Name/With Evidence'. Panel (ii) shows the filter set to 'No Product Name/With Evidence'. Panel (iii) shows the filter set to 'Product Name/No Evidence'. The tables in each panel list various genes and their associated annotations, such as 'Predicted transcriptional regulators', 'PadR', and 'PIN domain protein'.

FIGURE 2.5. Organism Details: Compare Gene Annotations.

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(i), and can be downloaded into a tab-

delimited Excel file. The results can be filtered in order to display only genes (a) **without a product name**, but **with evidence** of potential function provided by association with a COG, Pfam, or TIGRfam, as illustrated in Figure 2.5(ii), or (b) **with a product name**, but **without** any other **evidence** of function provided by association with a COG, Pfam, or TIGRfam, as illustrated in Figure 2.5(iii).

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**.

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 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 3.2(i).

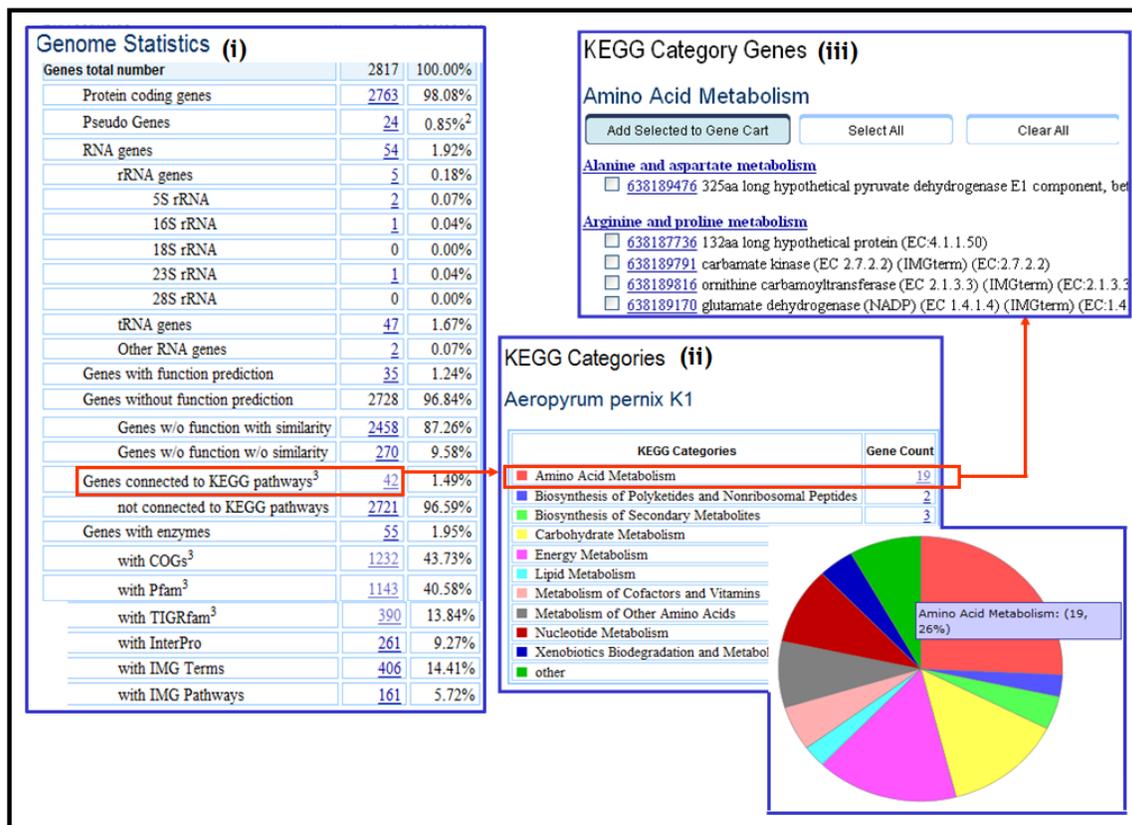


FIGURE 3.2. Selecting genes with Genome Statistics.

Gene counts in some categories in **Genome Statistics**, such as “Genes with COGs”, “Genes with Pfam”, “Genes with TIGRfam”, and “Genes connected to KEGG pathways” are linked to tables that show these genes classified according to the corresponding functional hierarchies (e.g., COG Functional Categories, KEGG Categories, etc.), as illustrated in Figure 3.2(ii), displayed both in tabular and graphical (pie chart) format. The gene counts in the table and on the pie chart representing (e.g., COG, KEGG) functional categories are linked to a table that contains groupings of genes according to individual functional groups or metabolic pathways, as illustrated in Figure 3.2(iii). Genes can be then selected and saved in the **Gene Cart** for further analysis.

3.4 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 3.3(i), that are further discussed below.

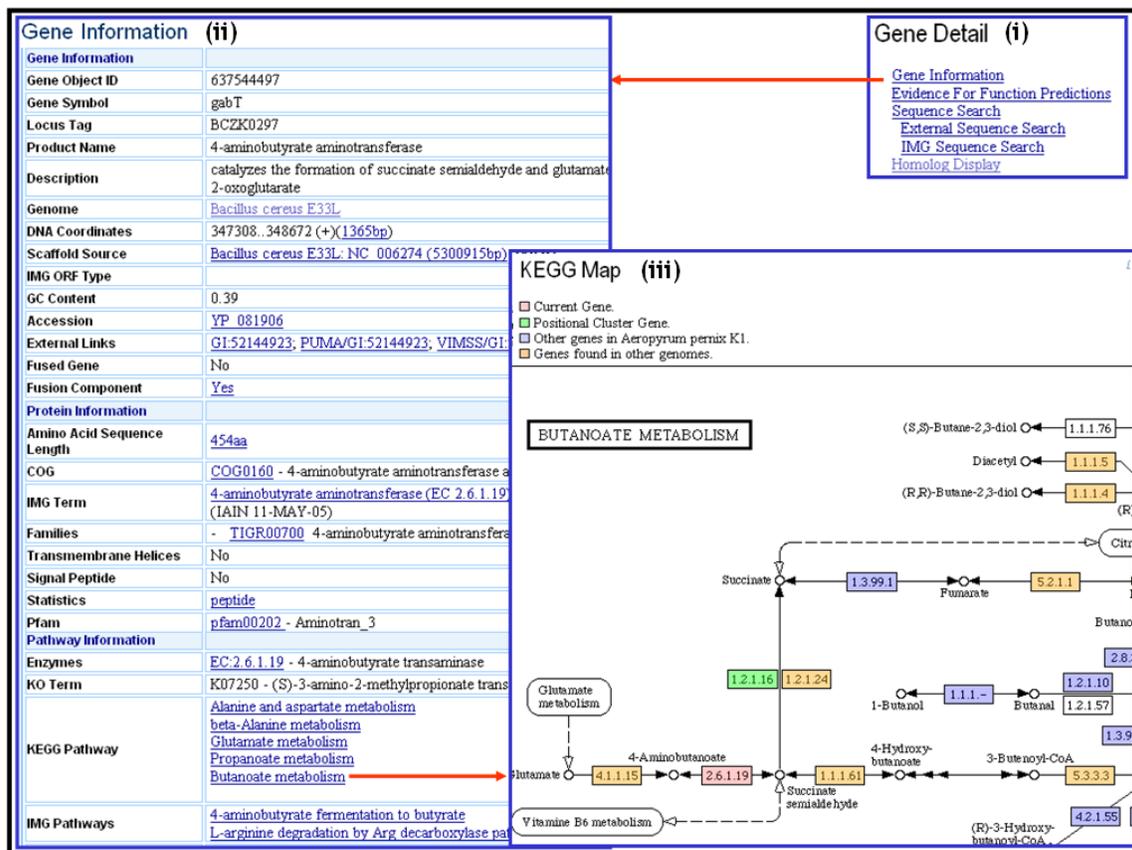


FIGURE 3.3. Gene Details: Gene Information and KEGG Map viewer.

3.4.1 Gene Information

The **Gene Information** section is divided into four parts that contain information on the gene, protein, associated pathways, and IMG clusters as shown in Figure 3.3(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 Pfam.

The third part of the **Gene Information** section contains information on KEGG and IMG pathways, and enzymes. Links to pathway and enzyme information are also provided. 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 3.3(iii). Colored EC numbers on a KEGG map are links: (i) **red** links lead to the current gene's **Gene Details** page; (ii) **green** links lead to genes in a positional cluster that includes the current gene; (iii) **blue** links lead to other genes that are in the same genome; (iv) **purple** links lead to genes in the same genome associated with the same EC number as the current gene; (v) **orange** links lead to "EC equivalogs," that is, genes in other genomes that are associated with the same EC number. Rounded rectangles are links to other pathway maps.

The fourth part of the **Gene Information** section contains data on IMG clusters, as illustrated in Figure 3.4(ii).

Gene Information (ii)

Gene Information

Protein Information

Pathway Information

IMG Clusters

Chromosomal Cassette Details By COG

Protein Cluster Context Analysis By - Select Protein Cluster -

IMG Ortholog Cluster 1094 - 4-aminobutyrate aminotransferase

Gene Detail (i)

[Gene Information](#)

[Evidence For Function Predictions](#)

[Sequence Search](#)

[External Sequence Search](#)

[IMG Sequence Search](#)

[Homolog Display](#)

Chromosomal Cassette COG (iii)

[Functions](#)

[Chromosomal Cassette Present](#)

Preferred COG Pathways

Cog Pathway Id	Pathway Name	Gene Count
8	Arginine biosynthesis	1
46	Transcriptional regulators	1

COG Functions (iv)

Select	COG Id	COG Name	Gene Id	Gene Name
<input type="checkbox"/>	COG0160	4-aminobutyrate aminotransferase and related aminotransferases	637544497	4-aminobutyrate aminotransferase
<input type="checkbox"/>	COG0546	Predicted phosphatases	637544496	haloacid dehalogenase-like hydrolase
<input type="checkbox"/>	COG1012	NAD-dependent aldehyde dehydrogenases	637544499	succinate-semialdehyde dehydrogenase (NAD(P)+)
<input type="checkbox"/>	COG2076	Membrane transporters of cations and cationic drugs	637544500	multidrug-efflux transporter (multidrug
<input type="checkbox"/>	COG2230	Cyclopropane fat related methyltras		
<input type="checkbox"/>	COG3829	Transcriptional r AAA-type ATPas domains		

Chromosomal Cassette COG Present (v)

1 - Click *Number of Gene Cluster* counts to view cassette plots with **only +** functions present

2 - Click *Number of Cassette* counts to view cassette plots with **at least +** functions present

3 - Click *Number of Organisms* counts to view Phylogenetic Distribution

Number of Gene Cluster ¹	Number of Cassette ²	Number of Organisms ³	COG0160	COG0546	COG1012	COG2076	COG2230	COG3829
6	2	2	+	+	+	+	+	+
5	11	11	+	+	+	+		+
4	12	12	+	+	+			+
4	12	12	+		+	+		+
3	13	13			+	+		+
3	13	13	+	+	+			

FIGURE 3.4. Gene Details: IMG Chromosomal Cassette Details.

Details on chromosomal cassettes can be displayed with genes labeled by their COG or Pfam association, as illustrated in Figure 3.4(iii). A **Chromosomal Cassette** details page provides information on the protein clusters (e.g., COGs) of the genes in the query cassette, as illustrated in Figure 3.4(iv). This page also provides information on other cassettes that share at least two protein clusters with the query cassette, as illustrated in Figure 3.4(v), including the number of cassettes that share at least two protein clusters with the query cassette as well as the number of organisms they come from.

Gene Information (ii)

Gene Information

Protein Information

Pathway Information

IMG Clusters

Chromosomal Cassette Details By COG

Protein Cluster Context Analysis By COG

IMG Ortholog Cluster 1094 - 4-aminobutyrate aminotransferase

Gene Detail (i)

[Gene Information](#)

[Evidence For Function Predictions](#)

[Sequence Search](#)

[External Sequence Search](#)

[IMG Sequence Search](#)

[Homolog Display](#)

Context analysis based on COGs. (iii) Loaded.

Summary

Query gene cluster	COG0160 : 4-aminobutyrate aminotransferase and related aminotransferases
Genomes with query COG	605
COGs in the same genomes and chromosomal cassettes with query COG	981
COGs in the same conserved chromosomal cassette with query COG	981
COGs fused with query COG	652

List of correlated COGs

Query COG	Correlated COG	Genomes with query COG	Genomes with correlated COG	Genomes with both COG	Genomes with both COG in cassette	Genomes with both COG in fusion	Genome cooccurrence ratio of COGs	Conserved neighborhood ratio of COGs	Fusion ratio of COGs	Genome cooccurrence correlation score	Conserved neighborhood correlation score
COG0160	COG1012 NAD-dependent aldehyde dehydrogenases	605	1020	581	308	308	96	50	50	151	130
COG0160	COG5285 Protein involved in biosynthesis of mitomycin antibiotics/polyketide fumonisin	605	258	220	30	0	85	11	0	104	52
COG0160	COG3486 Lysine/ornithine N-monoxygenase	605	231	212	20	0	91	8	0	78	49

FIGURE 3.5. Gene Details: Protein Cluster Context Analysis.

Protein Cluster Context Analysis allows examining the functional correlations of the current gene based on its COG or Pfam association, as illustrated in Figure 3.5. **Context Analysis** starts with the so called “query” protein cluster (COG, Pfam) associated with the current gene, as illustrated in Figure 3.5(ii).

For a query protein cluster A , the **Context Analysis** page contains a **summary** as illustrated in Figure 3.5(iii): for each pair (A, B) where B is a protein cluster that is associated with genes that are part of at least one cassette or fusion that also contains genes associated with A , the summary table lists the

- the **number of genomes** containing
 - genes associated with A ;
 - genes associated with B ;
 - genes associated with A and/or B ;
 - genes in the same cassette associated with A and/or B ;
 - genes in the same fusion event associated with A and B , respectively;
- the **genome co-occurrence ratio** for A and B :

$$NG1/\min(NGA, NGB)$$
, where

$NG1$ is the number of genomes in which A and B are associated with genes, NGA is the number of genomes with genes associated with A , and NGB is the number of genomes with genes associated with B ;
- the **conserved neighborhood ratio** for A and B :

$$NG1/\min(NGA, NGB)$$
, where

NG1 is the number of genomes in which *A* and *B* are associated with genes in a cassette, **NGA** is the number of genomes with genes associated with *A*, and **NGB** is the number of genomes with genes associated with *B*;

- the **fusion ratio** for *A* and *B*, computed as:

$$\text{NG1}/\min(\text{NGA}, \text{NGB}),$$
 where
NG1 is the number of genomes in which *A* and *B* are associated with genes in a fusion, **NGA** is the number of genomes with genes associated with *A*, and **NGB** is the number of genomes with genes associated with *B*.
- the **genome co-occurrence correlation score** for *A* and *B*:

$$\{ [\text{P}(\text{A}) \times \text{P}(\text{B}) \times \log(\text{P}(\text{A}, \text{B})) / (\text{P}(\text{A}) \times \text{P}(\text{B}))] + [\text{Q}(\text{A}) \times \text{Q}(\text{B}) \times \log(\text{Q}(\text{A}, \text{B})) / (\text{Q}(\text{A}) \times \text{Q}(\text{B}))] \} \times \text{maxDist}(\text{A}, \text{B}),$$
 where
 - **P(A)** and **P(B)** are the probabilities of a genome containing genes associated with *A* and *B*, respectively; **P(A,B)** is the probability of a genome containing genes in a cassette associated with *A* and *B*;
 - **Q(A) = 1-P(A)**; **Q(B) = 1-P(A)**; **Q(A,B) = 1-P(A,B)**;
 - **maxDist(A,B)** is the **maximum phylogenetic distance** between genomes that contain genes in a fusion associated with *A* and *B*;
- the **conserved neighborhood correlation score** for *A* and *B*:

$$(\text{NG1} / \text{NG2}) \times \text{maxDist}(\text{A}, \text{B}),$$
 where
NG1 is the number of genomes in which *A* and *B* are associated with genes in a cassette, **NG2** is the number of genomes in which *A* or *B* are associated with genes, and **maxDist(A,B)** is the **maximum phylogenetic distance**¹ between genomes that contain genes in a cassette associated with *A* and *B*;
- the **fusion correlation score** for *A* and *B*:

$$(\text{NG1} / \text{NG2}) \times \text{maxDist}(\text{A}, \text{B}),$$
 where
NG1 is the number of genomes in which *A* and *B* are associated with genes in a fusion, **NG2** is the number of genomes in which *A* or *B* are associated with genes, and **maxDist(A,B)** is the **maximum phylogenetic distance** between genomes that contain genes in a fusion associated with *A* and *B*.

Note that the higher a **correlation score** is, the more important the correlation is: values **greater than 500** are highly significant for **conserved chromosomal neighborhoods**, while values **higher than 200** are highly significant for fusion events.

3.4.2 Evidence for Function Prediction

The **Evidence for Function Prediction** section includes a graphic display of the gene's neighborhood, a link to the **Chromosome Viewer**, **Conserved Neighborhood Viewers**, and information on associated COG and Pfam domain, as shown in Figure 3.6.

¹ The *phylogenetic distance* between two organisms is computed based on a 16S RNA tree. The alignment of the 16S RNA genes was extracted from the **Greengenes** database, with the **Phylip DNADIST** program used to calculate the distance matrix from this alignment.

You can simultaneously view neighborhoods for the gene of interest and its orthologs in the genomes you have selected with the **Ortholog Neighborhood Viewer**, as illustrated in Figure 3.6(v). 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.

Chromosomal cassettes can be examined using the **Chromosomal Cassette Viewer**, as illustrated in Figure 3.6(vi). Chromosome cassettes can be viewed with genes labeled by their COG or Pfam association. For each chromosomal cassette, related cassettes in other genomes are also displayed. The query gene has a small red box under it. You can mouse over any gene to see its details. You can mouse over or click the red dotted line box surrounding a cassette to see the cassette details discussed above and illustrated in Figure 3.4. Genes are colored by the protein cluster (e.g., COG) association, with genes that have no protein cluster or that are outside a cassette colored yellow.

3.4.3 Sequence Search

The External Sequence Search section provides links to several resources outside IMG, such as NCBI BLAST, illustrated in Figure 3.7(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 3.7(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 3.7(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**.

The figure illustrates the workflow for external and IMG sequence searches. It starts with a 'Gene Detail' page, which provides navigation to various search tools. The 'External Sequence Search' section includes links to NCBI BLAST, EBI InterPro Scan, and Protein Data Bank BLAST. The 'IMG Sequence Search' section includes links to IMG Genome BLAST and Phylogenetic Profile Similarity Search. The 'BLAST' interface (i) shows a query sequence and search options. The 'BLAST Against Selected Genomes (ii)' interface shows a list of genomes and search parameters. The 'Phylogenetic Profile Similarity Search (iii)' interface shows search criteria like 'Min. % Occurrence Match' and 'Top N results'.

FIGURE 3.7. External and IMG Sequence Search.

3.4.4 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 3.8(i). The homolog genes are presented in multiple pages, as shown in Figure 3.8(ii).

The **Selective Homolog Display** allows examining genes that have similar sequences to your target gene, including orthologs/paralogs and top IMG homolog hits, as illustrated in Figure 3.9(i). A list of (ortholog, paralog, homolog) genes is displayed in a preset format, as shown in Figure 3.9(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 3.9(iii).

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 3.9(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.

Gene Detail

[Gene Information](#)
[Evidence For Function Predictions](#)
[Sequence Search](#)
[External Sequence Search](#)
[IMG Sequence Search](#)
[Homolog Display](#)

Customized Homolog Display

Homolog Toolkit (i)

The homolog toolkit allows you customize how homologs are retrieved and displayed in multiple ways. Find homologs for gene 638189791: 316aa long hypothetical protein.

Page Options:
500 hits per page.

Column Options:
 Show enzyme column.
 Show IMG terms.
 Show Scaffold Information.

Row Options:
 Show all homologs.
 Show homologs without IMG terms.
 Show homologs from finished genomes only.

Sort Options:
 Sort by descending bit score.
 Sort by genomes.
 Sort by descending amino acid percent identity.
 Sort by product name.

Go Reset

Homologs for query gene 638189791 316aa long hypothetical carbamate kinase. (ii)
 Sort by bit score
 527 total hits.

Pages: [1] 2 [Next Page]

Add Selected to Gene Cart Select All Clear All

Types (T): O = Ortholog, P = Paralog, - = other unidirectional hit.
 Domains(D): B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids, V=Viruses.
 Genome Completion(C): F=Finished, D=Draft.
 Click on column name to sort.

Add query gene 638189791

Select	Row No.	Homolog	Product Name	I	Percent Identity	Alignment On Query Gene	Alignment On Subject Gene	Length	E-value	Bit Score	D	C	Genome
<input type="checkbox"/>	1	640081877	Carbamate kinase	O	49.68%	<div style="width: 100%; height: 10px; background-color: green;"></div>	<div style="width: 100%; height: 10px; background-color: green;"></div>	321aa	7.0e-74	280	A	F	Hyperthermus butylicus DSM 5456
<input type="checkbox"/>	2	638173201	carbamate kinase-like carbamoylphosphate synthetase	O	47.28%	<div style="width: 100%; height: 10px; background-color: green;"></div>	<div style="width: 100%; height: 10px; background-color: green;"></div>	314aa	5.0e-76	278	A	F	Pyrococcus furiosus DSM 3638

FIGURE 3.8. Customized Homolog Display.

Gene Detail

[Gene Information](#)
[Evidence For Function Predictions](#)
[Sequence Search](#)
[External Sequence Search](#)
[IMG Sequence Search](#)
[Homolog Display](#)

Homolog Selection (i)

Homolog Selection

-- Select Homolog Type --

- Select Homolog Type --
- Paralogs / Orthologs
- Top IMG Homolog Hits
- Phenotype
- Habitat
- Disease
- Relevance

Homologs (iii)

Homolog Selection Phenotype

Summaries for Phenotype

Summaries are formed with bidirectional best hits. The genome for this gene has the following phenotype: *Sphere-shaped, Thermophile, Hyperthermophile, ...*

Phenotype	Ortholog Count
Rod-shaped	151
Pathogen	134
Facultative	130
Motile	123

Phylogenetic Distribution (iv)

Distribution of orthologs (bidirectional best hits) are shown in red.

A .01 Archaea (19)
 A . .02 Crenarchaeota (9)
 A . . .03 Thermoprotei (9)
 A04 Desulfurococcales (3)
 A05 Desulfurococcaceae (1)
 A06 Aeropyrum (1)
 A08 Aer
 A05 unclassif
 A06 Hyperth
 A08 Hyp
 A06 Staphyl
 A08 Sta
 A04 Sulfolobale
 A05 Sulfoloba
 A06 Metallo
 A08 Met
 A06 Sulfolob
 A08 Sul
 A08 Sul
 A08 Sul
 A04 Thermoprote
 A05 Thermofil
 A06 Thermof
 A08 The
 A05 Thermopro

Orthologs (ii)

(Orthologs are bidirectional best hits from BLASTP of each genomes against each other genome.)
 Domains(D): B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids, V=Viruses.
 Genome Completion(C): F=Finished, D=Draft.
 Click on column name to sort.

Select	Ortholog	Product Name	Percent Identity	Alignment On Query Gene	Alignment On Subject Gene	Length	E-value	Bit Score	Cons. Region Score ¹	D	C	Genome Name
<input type="checkbox"/>	640081877	carbamate kinase (EC 2.7.2.2) (IMGterm)	49.68	<div style="width: 100%; height: 10px; background-color: green;"></div>	<div style="width: 100%; height: 10px; background-color: green;"></div>	321aa	7.0e-74	280	0	A	F	Hyperthermus butylicus DSM 5456
<input type="checkbox"/>	638173201	carbamate kinase (EC 2.7.2.2) (IMGterm)	47.28	<div style="width: 100%; height: 10px; background-color: green;"></div>	<div style="width: 100%; height: 10px; background-color: green;"></div>	314aa	9.0e-74	278	0	A	F	Pyrococcus furiosus DSM 3638

FIGURE 3.9. Selective Homolog Display.

3.5 Phylogenetic Profilers

Genes in a target (query) genome can be selected using **Phylogenetic Profiler** tools that are available on the second-level menu of **Find Genes**, as shown in Figure 3.10(i). Genes in a query genome can be selected based on (a) presence or absence of homologs in other genomes or (ii) chromosome collocation in the target genome (as part of a *gene cassette*) as well as in other genomes.

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.

3.5.1 Phylogenetic Profiler for Single Genes

The **Phylogenetic Profiler for Single Genes** (see Figure 3.10(ii)) allows selecting genes in a query genome based on the presence or absence of homologs for these genes in other genomes. First, select your query genome by using the associated radio button in the "Find Genes In" column, as shown in Figure 3.10(ii). 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.

Phylogenetic Profiler for Single Genes (ii)

Profile

Find Genes In*	With Homologs In	Without Homologs In	Ignoring	Taxon Name
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Archaea
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Euryarchaeota
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thermoplasma
<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thermoplasma acidophilum DSM 1728 [F]
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thermoplasma volcanium GSS1 [F]

Similarity Cutoffs

Max. E-value: 1e-5
 Min. Percent Identity: 30
 Algorithm: By Present/Absent Homologs
 Min. Taxon Percent With Homologs: 100
 Min. Taxon Percent Without Homologs: 100

Summary Statistics

	Number	% of Total
Genes total number	1333	100.00%
COG	1109	83.20%
Enzyme	59	4.43%
Pfam	1053	78.99%
InterPro	1110	83.27%
No Functional Hit	155	11.63%
Unique In IMG	0	0.00%

Phylogenetic Profiler for Single Genes Results (iii)

Processing 1 comparison(s).
 1527 genes found for genome (bin) of interest, Thermoplasma acidophilum DSM 1728
 1333 genes remaining after intersecting with homologs in Thermoplasma volcanium GSS1

Select	Result Row	Gene Object ID	Locus Tag	Gene Name	Length	COG	Enzyme	Pfam	InterPro	Unique In IMG
<input type="checkbox"/>	1	638180157	Ta0001	ATP synthase (subunit E) related protein	185aa	COG1390	-	pfam01991	IPR002842	No
<input type="checkbox"/>	2	638180158	Ta0002	ATP synthase (subunit C) related protein	356aa	COG1527	-	pfam01992	IPR002843	No

FIGURE 3.10. Gene Search: Phylogenetic Profiler for Single Genes.

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 3.10(ii). Click “Go” to find the genes in the target genome that satisfy the homolog presence/absence condition you have set.

The **Phylogenetic Profiler for Single Genes** results show the genes in the query genome along with their functional characterization based on COG, Pfam, InterPro, and Enzymes, as illustrated in Figure 3.10(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. At the bottom of the results page, a summary of the genes with or without functional annotations is provided, as shown in the right side pane of Figure 3.10(iii). The count of genes is linked to the corresponding list of genes, which can be viewed by clicking on the numbers for each category.

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. 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.

3.5.2 Phylogenetic Profiler for Gene Cassettes

The **Phylogenetic Profiler for Gene Cassettes** (see Figure 3.11(i)) allows selecting genes that are part of a *gene cassette* (i.e., are collocated on the chromosome) in a query genome and are part of related (conserved part of) gene cassettes in other genomes.

First, select your query genome by using the associated radio button in the “Find Genes In” column, as shown in Figure 3.11(ii). Next, select the protein cluster used for correlating gene cassettes: COG or Pfam. Then select the genomes for gene cassette comparisons with the query genome by using the associated radio buttons in the “Collocated In”. Genomes you want to be ignored in these comparisons can be selected using the radio buttons in the “Ignoring” column.

The **Phylogenetic Profiler for Gene Cassette Results** starts with a summary of the results, as shown in the left side pane of Figure 3.11(iii), including a table with the first column listing the **size of the groups of collocated genes** in the query genome and the second column listing the **number of such groups conserved** across the other genomes involved in the selection.

The **Details** part of the **Phylogenetic Profiler for Gene Cassette Results** consists of a table that displays **groups of collocated genes** in each chromosomal cassette (identified by the Cassette Id) in the query genome that satisfy the search criterion, as illustrated in Figure 3.11(iii).

Phylogenetic Profilers (i)

Tool	Description
Single Genes	Find genes in genome (bin) of interest qualified by similarity to set alignments). Only user-selected genomes appear in the profiler.
Gene Cassettes	IMG Cassette Profiler. Find collocated genes that are part of a cassette in other genomes of interest

Phylogenetic Profiler for Gene Cassettes (ii)

Select Protein Cluster

COG
 Pfam

Find Genes In*	Collocated In	Ignoring	
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Archaea
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Euryarchaeota
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thermoplasma
<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thermoplasma acidophilum DSM 1728 [F]
<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Thermoplasma volcanium GSS1 [F]

Thermoplasma acidophilum DSM 1728 Phylogenetic Profiler for Gene Cassettes Results (iii)

Statistics

- 243 (58.98%) gene cassettes in query genome from a total of 412 gene cassettes
- 849 (55.31%) genes in query genome from a total of 1535 genes, have collocated

Details

No of Collocated Genes	Occurrences
29	1
13	1
12	1
11	3
10	3
9	2
8	1

Select	Result Row	Gene Id	Gene Name	Length	Cassette Id	Conserved Neighborhood Viewer Centered on this Gene
<input type="checkbox"/>	1	638181472	nucleolar protein Nop56 related protein	783	266638154510	Ta1243
<input type="checkbox"/>	2	638181473	dyskerin (nucleolar protein Nap57) related protein	1068		Ta1244
<input type="checkbox"/>	3	638181474	probable cytidylate kinase	567		Ta1245
<input type="checkbox"/>	4	638181475	conserved hypothetical membrane protein	696		Ta1246

Chromosomal Casette By COG (iv)

Thermoplasma acidophilum DSM 1728: RL139299 MATCHES 32
1322599 1317599 1317599 1317599

Thermoplasma volcanium GSS1 DNA: BR000011 MATCHES 29
339426 339426 344

Ta1243: nucleolar protein Nop56 related protein
1307205...
1307988
(260 aa)
(COG1498)

FIGURE 3.11. Gene Search: Phylogenetic Profiler for Gene Cassettes.

Note that:

- in each **specific group** of collocated genes in the query genome, individual genes may correspond to parts of **multiple chromosomal cassettes** in the **other genomes** involved in the profiler condition;
- the **conserved part** of a chromosomal cassette involving an individual gene in the query genome can be examined using the links provided in the "**Conserved Neighborhood Viewer Centered on this Gene**" column of results table, as shown in figure 9(iv).

You can explore the details of individual gene listed in the result 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.

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 “Alkaliphilus metalliredigenes,” as shown in Figure 4.1(i).

The screenshot shows the IMG website interface. At the top, there is a navigation bar with 'Find Functions' selected. Below it, a search box contains the keyword 'lipid' and a filter dropdown menu is open, showing 'COG' selected. The search results are displayed in a table with columns for COG ID and description. The results are as follows:

COG ID	Description
COG0416	Fatty acid/phospholipid biosynthesis enzyme (1)
COG3307	Lipid A core - O-antigen ligase and related enzymes (1)
COG0671	Membrane-associated phospholipid phosphatase (1)
COG3963	Phospholipid N-methyltransferase (1)

The interface also includes a 'Function Search Results' section with buttons for 'Add Selected to Function Cart', 'Select All', and 'Clear All'. The search results are shown in parentheses next to the descriptions.

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.

The screenshot displays the IMG 2.6 interface with several function browsers open. At the top, the navigation bar includes 'Find Functions' with sub-menus for COG, Pfam, KEGG, Enzyme, TIGRfam, and IMG Network. The COG Browser (i) shows a list of functional categories such as 'Amino acid transport and metabolism', 'Arginine biosynthesis', 'Histidine biosynthesis', and 'Leucine biosynthesis'. The Pfam Browser (ii) displays 'Amino acid transport and metabolism'. The KEGG Pathways (iii) browser shows 'Amino Acid Metabolism' and 'Alanine and aspartate metabolism'. The Enzymes (iv) browser lists EC numbers like EC:1.14.13.93 and EC:1.1.1.198. The TIGRfam Roles (v) browser shows 'Amino acid biosynthesis' categories. The IMG Network Browser (vi) displays a list of networks like 'NETWK-00065 Amino acid synthesis'. A 'Characterized COG's (vii) browser is also visible, showing a list of COG IDs and their descriptions.

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.

The screenshot displays the IMG COG Browser interface. On the left, the 'COG Browser' shows a list of COG categories under 'Amino acid transport and metabolism', with a total count of 624. The 'COG Category Details (ii)' page shows details for 'Acetylglutamate semialdehyde dehydrogenase' (COG0002) with a count of 624. The 'COG Pathway Details (iii)' page shows details for 'Arginine biosynthesis' (COG0002) with a count of 624. A 'Phylogenetic Distribution for COG0002 (v)' is also shown, listing hits in red: Archaea (40), Crenarchaeota (10), Thermoprotei (10), Desulfurococcales (1), Desulfurococaceae (1), Aeropyrum (1), Aeropyrum pernix K1 (1), unclassified, and Hyperthermus.

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.

(i) **img** Pfam Browser

Search | COG | **Pfam** | KEGG | Enzyme | TIGR

(Pfam Categories are formed from COG categories)

Pfam list

- Amino acid transport and metabolism [E]
- Arginine biosynthesis
- Histidine biosynthesis
- Isoleucine biosynthesis
- Leucine biosynthesis
- Methionine biosynthesis
- Phenylalanine/tyrosine biosynthesis
- Proline biosynthesis
- Threonine biosynthesis
- Tryptophan biosynthesis

Pfam Families (ii)

Add Selected to Function Cart

- pfam00244 14-3-3
- pfam00389 2-Hacid_dh
- pfam02826 2-Hacid_dh_C
- pfam00198 2-oxoacid_dh
- pfam04029 2-ph_phosp
- pfam02834 2_5_RNA_ligase
- pfam03171 2OG-FeII_Oxy

Pfam Category Details (iii)

Details for Pfam Category *Amino acid transport and metabolism*

Add Selected to Function Cart Select All

Click on column name to sort.

Select	Pfam ID	Pfam Name	Genome Count
<input type="checkbox"/>	pfam00217	ATP-gua_Ptrans	106
<input type="checkbox"/>	pfam00221	PAL	345
<input type="checkbox"/>	pfam00278	Orn_DAP_Arg_deC	704

Pfam Pathway Details (iv)

Details for Pfam Pathway *Arginine biosynthesis*

Add Selected to Function Cart Select All

Click on column name to sort.

Select	Pfam ID	Pfam Name	Genome Count
<input type="checkbox"/>	pfam01960	ArgJ	463
<input type="checkbox"/>	pfam02343	DUF130	2
<input type="checkbox"/>	pfam02774	Semialdehyde_dhC	745

Genomes with pfam02774 (v)

Genomes with *Semialdehyde_dhC*.

Phylogenetic Distribution

Domains(D): B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids
Genome Completion(C): F=Finished, D=Draft

D	C	Genome
A	D	<i>Caldivirga maquilgensis</i> IC-167
A	D	<i>Ferropasma acidarmanus</i> Fer1
A	D	<i>Methanococcus aeolicus</i> Nankai-3

Phylogenetic Distribution for pfam02774 (vi)

(Hits are shown in red.)

Distribution for *Semialdehyde_dhC*.

Domains(D): B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids, V=Viruses
Genome Completion: [F]inished, [D]raft.

```

A .01 Archaea (85)
A . .02 Crenarchaeota (25)
A . . .03 Thermoprotei (25)
A . . . .04 Desulfurococcales (2)
A . . . . .05 Desulfurococcaceae (2)
A . . . . . .06 Aeropyrum (2)
A . . . . . . .08 Aeropyrum pernix K1 [F](2)
A . . . . . . .05 unclassified
A . . . . . .06 Hyperthermus
A . . . . . . .08 Hyperthermus butylicus DSM 54
  
```

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).

(i) img

IMG Home Find Genomes Find Genes Find Functions
Search COG Pfam **KEGG** Enzyme TIGRFam IMG

KEGG Pathways

Amino Acid Metabolism
[Alanine and aspartate metabolism](#)
[Arginine and proline metabolism](#)
[Cysteine metabolism](#)

KEGG Map (iv)

Genes in Aeropyrum pernix K1.
 EC equivalents (genes found in other genomes that have the same EC number).

ALANINE AND ASPARTATE METABOLISM

KEGG Pathway Details (ii)

Details for Alanine and aspartate metabolism.

Enzymes in Pathway

Add Selected to Function Cart Select All Clear All

Select	EC Number	Enzyme Name	Genome Count
<input type="checkbox"/>	EC:1.2.1.18	Malonate-semialdehyde dehydrogenase (acetylating).	4
<input type="checkbox"/>	EC:1.2.1.51	Pyruvate dehydrogenase (NADP(+)).	4
<input type="checkbox"/>	EC:1.2.4.1	Pyruvate dehydrogenase (acetyl-transferring).	641
<input type="checkbox"/>	EC:6.3.5.6	Asparaginyl-tRNA synthase (glutamine-hydrolyzing).	0
<input type="checkbox"/>	EC:6.4.1.1	Pyruvate carboxylase.	231

Add Selected to Function Cart Select All Clear All

Genomes with EC:1.2.1.18 (iii)

Genomes with Malonate-semialdehyde dehydrogenase (acetylating).

Phylogenetic Distribution

View Pathway Map

View KEGG map for one genome.
 Domains: (B)acteria, (A)rchaea, (E)ukarya
 Genome Completion: [F]inished, [D]raft.

D	C	Genome	Gene Count
B	D	Bacillus thuringiensis sv israelensis ATCC 35646	1
B	D	Ralstonia solanacearum UW551	1
B	F	Bradyrhizobium japonicum USDA 110	1

Aeropyrum pernix K1 (A)[F]

Archaeoglobus fulgidus DSM 4304 (A)[F]
 Caldvirga maquilingensis IC-167 (A)[D]
 Ferroplasma acidimanus Fer1 (A)[D]
 Haloarcula marismortui ATCC 43049 (A)[F]
 Halobacterium sp. NRC-1 (A)[F]
 Haloquadratum walsbyi DSM 16790 (A)[F]
 Hyperthermus butylicus DSM 5456 (A)[F]
 Metallosphaera sedula DSM 5348 (A)[F]
 Methanobrevibacter smithii ATCC 35061 (A)[F]

View Map Reset

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).

The screenshot displays the TIGRfam Browser interface with several key components:

- (i) img**: Navigation menu with options like IMG Home, Find Genomes, Find Genes, Find Functions, Search, COG, Pfam, KEGG, Enzyme, TIGRfam, and IMG.
- TIGRfam Roles**: A tree structure showing functional subsystems such as Amino acid biosynthesis, Biosynthesis of cofactors, prosthetic groups, and other molecules, and TIGRfam's (ii).
- TIGRfam Role (iii)**: Details for the role *Aromatic amino acid family*. It includes a table of TIGRfam members:

Select	TIGRfam ID	Expanded Name	Genome Count
<input type="checkbox"/>	TIGR00033	chorismate synthase	205
<input type="checkbox"/>	TIGR00034	phospho-2-dehydro-3-deoxyheptonate aldolase	162
			219
- Genomes with TIGR00033 (iv)**: A table showing the distribution of TIGR00033 across different domains and genome completion statuses:

D	C	Genome
A	F	Methanococcoides burtonii DSM 6242
A	F	Methanosphaera stadtmanae DSM 3091
A	F	Natronomonas pharaonis DSM 2160
- Phylogenetic Distribution for TIGR00033 (v)**: A detailed distribution for *chorismate synthase*.

(Hits are shown in red.)

Distribution for *chorismate synthase*.
 Domains(D): B=Bacteria, A=Archaea, E=Eukarya, P=Plasmid
 Genome Completion: [F]inished, [D]raft.

 - A .01 Archaea (10)
 - A . .02 Crenarchaeota (2)
 - A . . .03 Thermoprotei (2)
 - A04 Desulfurococcales
 - A05 Desulfurococcaceae
 - A06 Aeropyrum
 - A08 Aeropyrum pernix K1 [F]

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).

The screenshot displays the IMG web interface with several key sections highlighted:

- (i) IMG Network:** The top navigation bar includes 'Find Functions' and 'IMG Network' (highlighted in a red box). Below it, a list of links includes 'IMG Terms' (highlighted in a red box).
- (ii) IMG Term Browser:** Shows a list of terms under '(Term Tree Root)'. The first term, '00456 2-C-methyl-D-erythritol 4-phosphate cytidyltransferase (EC 2.7.7.60) 2,4-cyclodiphosphate synthase (EC 4.6.1.12)', is highlighted in a red box.
- (iii) IMG Term Details:** A detailed view of the selected term, showing 'Term Object ID: 00456', 'Term: 2-C-methyl-D-erythritol 4-phosphate cytidyltransferase (EC 2.7.7.60) 2,4-cyclodiphosphate synthase (EC 4.6.1.12)', and 'Number of Genes: 79'.
- (iv) Genomes with Term:** A table showing the phylogenetic distribution of the term across various genomes.

D	C	Genome	Gene Count
B	D	Anaeromyxobacter sp. Fw109-5	1
B	D	Aurantimonas sp. SI85-9A1	1
B	D	Parvibaculum lavamentivorans DS-1	1

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).

The screenshot displays the IMG Pathway Details page for pathway 00341. It is divided into several sections:

- (i) img**: The main navigation header.
- IMG Pathways (ii)**: A list of pathways with checkboxes. Pathway 00341 is selected and highlighted with a red box.
- IMG Pathway Details (iii)**: Information for pathway 00341, including its name (Oxidative pentose phosphate pathway), add/modify dates, and modified by (IAIN). It includes a checkbox to add the pathway to the function cart and a button to add selected pathways.
- Pathway Reactions**: A table listing reactions with their order, IMG terms, reaction definitions, and genome counts.

Reaction Order	IMG Terms	Reaction Definition	Genome Count
1.	<input type="checkbox"/> 01492 glucose-6-phosphate 1-dehydrogenase (EC 1.1.1.49)	D-Glucose 6-phosphate + NADP+ <=> D-Glucono-1,5-lactone 6-phosphate + NADPH + H+	539
2.	<input type="checkbox"/> 01493 6-phosphogluconolactonase (EC 3.1.1.31)	D-Glucono-1,5-lactone 6-phosphate + H2O <=> 6-Phospho-D-gluconate	358
3.	<input type="checkbox"/> 01494 6-phosphogluconate dehydrogenase		
4.	<input type="checkbox"/> 01495 ribose-5-phosphate isomerase		
- Genomes with IMG Pathway Reaction (v)**: A table showing genomes associated with the pathway.

D	C	Genome	Gene Count
B	D	Actinobacillus pleuropneumoniae sv 1 4074	1
B	D	Actinobacillus succinogenes 130Z	1
- Associated Genomes (iv)**: A table showing genomes associated with the pathway.

D	C	Genome	Gene Count
A	D	Caldivirga maquilgensis IC-167	1
A	D	Ferroplasma acidarmanus Fer1	1

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 Details** page as shown in Figure 4.9(iii) and as discussed above.

The screenshot displays the IMG Network Detail page, which is divided into several sections:

- Navigation:** Includes links for IMG Home, Find Genomes, Find Genes, Find Functions, and a search bar with options for COG, Pfam, KEGG, Enzyme, TIGRfam, and IMG Network.
- IMG Network Browser:** A section for navigating through IMG networks, with a link to 'IMG Network Browser' highlighted.
- IMG Network Detail (ii):** A table providing metadata for the selected network:

Network OID	00063
Name	Amino acid synthesis
EON Grammar	
Description	
Comments	
Image ID	
Add Date	26-JUL-05
Modify Date	26-JUL-05
Modified By	IAIN (IAnderson@lbl.gov)
- IMG Network Subtree Structure:** A hierarchical tree of pathways:
 - 01 Amino acid synthesis
 - 02 Synthesis of D-amino acids
 - 03 D-alanine synthesis
 - IPWAY.00264 D-alanine synthesis from D-asparagine by transamination
 - IPWAY.00262 D-alanine synthesis from D-aspartate by transamination
 - IPWAY.00258 D-alanine synthesis from D-glutamate by transamination
 - IPWAY.00256 D-alanine synthesis from L-alanine
 - 03 D-aspartate synthesis
 - IPWAY.00263 D-aspartate synthesis from D-alanine by transamination
 - IPWAY.00260 D-aspartate synthesis from D-glutamate by transamination

- IMG Pathway Details (iii):** A table providing metadata for the selected pathway:

Pathway OID	00264
Name	D-alanine synthesis from D-asparagine by transamination
Add Date	27-JUL-05
Modify Date	27-JUL-06
Modified By	IAIN

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.

Summary Statistics (i)

	Number	% of Total
DNA, total number of bases	3149710	100.00%
DNA coding number of bases	2804128	89.03%
DNA G+C number of bases	1352399	42.94% ¹
DNA scaffolds	2	100.00%
Genes total number	3190	100.00%
Protein coding genes	3088	96.80%
RNA genes	102	3.20%
rRNA genes	6	0.19%
5S rRNA	2	0.06%
16S rRNA	2	0.06%
18S rRNA	0	0.00%
23S rRNA	2	0.06%
28S rRNA	0	0.00%
tRNA genes	92	2.88%
Other RNA genes	4	0.13%

General Statistics (ii)

D	C	Genome Name	Genes	GC Perc	Bases
A	F	Thermoplasma acidophilum DSM 1728	1580	0.46	1564906
A	F	Thermoplasma volcanium GSS1	1610	0.40	1584804

Configuration

Output	Column Name	Description
<input type="checkbox"/>	Phylum	Phylum
<input type="checkbox"/>	Class	Class
<input type="checkbox"/>	Order	Order
<input type="checkbox"/>	Family	Family
<input type="checkbox"/>	Genus	Genus
<input checked="" type="checkbox"/>	Genes	Total gene count
<input type="checkbox"/>	CDS	Number of CDS genes
<input type="checkbox"/>	CDS Perc	CDS genes (percentage)
<input type="checkbox"/>	RNA	Number of RNA genes
<input type="checkbox"/>	rRNA	Number of rRNA genes
<input type="checkbox"/>	5S	Number of 5S rRNA's
<input type="checkbox"/>	16S	Number of 16S rRNA's
<input type="checkbox"/>	18S	Number of 18S rRNA's

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. Both statistics can be **exported** into a tab-delimited file, and can be configured using a **Configuration** table.

The **COG Category Statistics** are illustrated in Figure 5.2. Selecting **Statistics for Genomes by specific COG Category** (see Figure 5.2(i)) will display in a tabular and pie chart format the count of genes associated with each COG category across all selected genomes, as illustrated in Figure 5.2(ii). Clicking on a COG category on the *pie chart* or on the colored coded square for a COG category in the table will display a *bar chart* with the percent of genes for each genome associated with that COG category, as illustrated on the lower side pane of Figure 5(ii).

Selecting **Statistics for Genomes by COG Categories** will display a statistics table, as illustrated in Figure 5.2(iii). Each genome in this table is linked to its **Genome Details** page, which provides statistics for individual genomes, as discussed above. The column headers in this table 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, as illustrated in Figure 5.2(iv).

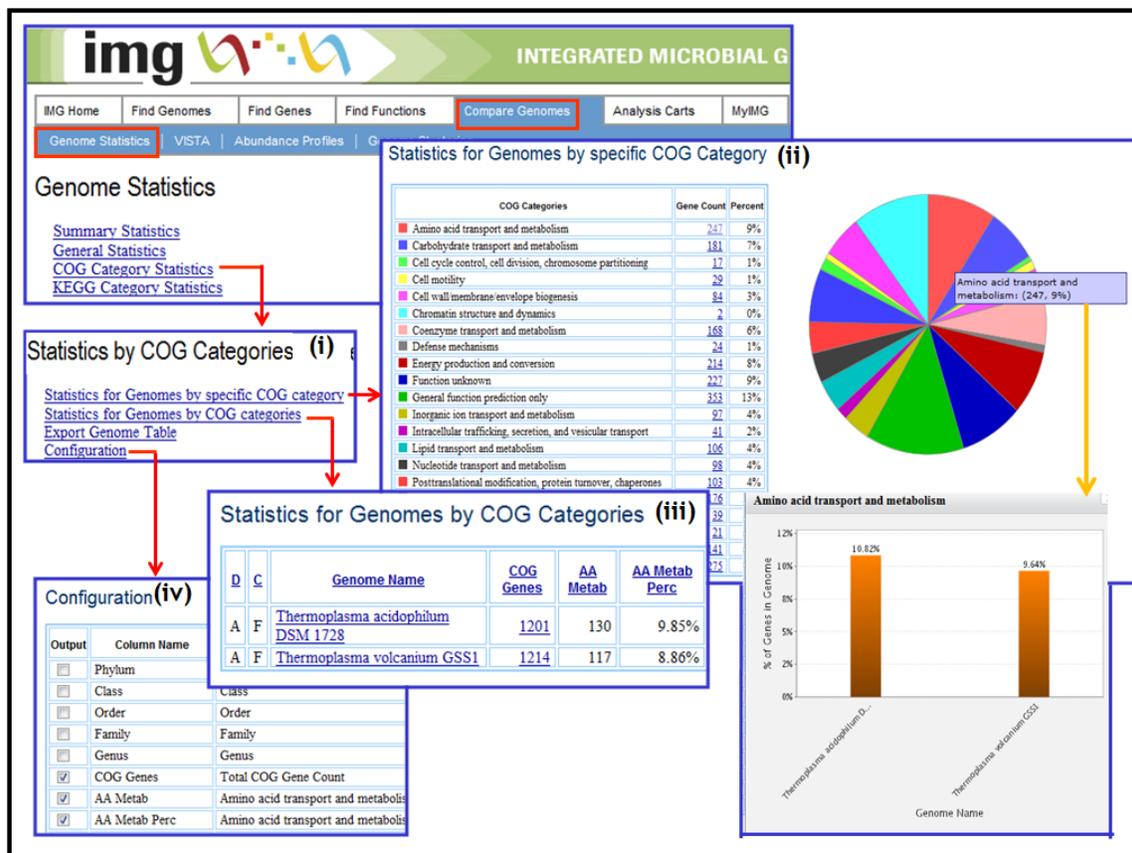


FIGURE 5.2. Genome Statistics: COG Category statistics.

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.

The figure shows two parts of the VISTA browser interface. Part (i) is the main menu where 'Compare Genomes' is selected. Part (ii) is the VISTA browser interface with the following details:

- Control Panel:**
 - Reference (Base) genome: *Bacillus cereus* ATCC 10987
 - Position (gene name or chr:start-end): NC_003909:47-9,414
 - Gene annotation: Known Genes
 - 0 more organisms: select/add
 - # Rows: 1
- Information:**
 - Cursor at: 7,355
 - Aligned contig: NZ_AAAC02000001
- Color Legend:**
 - Annotations: Gene (blue), Exon (green), UTR (red), CNS (orange)
 - Repeats: LINE (red), LTR (green), SINE (blue), RNA (orange), DNA (yellow), Other (grey)
 - Contigs: Contig (grey), Overlap (red)
 - SNPs: SNP (black)
- Main View:** A multi-track alignment plot showing conservation scores across seven different species. The x-axis represents genomic position in kilobases (kb), from 0.1k to 8.0k. The y-axis lists species: 1. *Bacillus halodurans* C-125 (SLAGAN), 2. *Bacillus anthracis* str. Sterne (SLAGAN), 3. *Bacillus anthracis* Ames Ancestor (SLAGAN), 4. *Bacillus anthracis* A2012, 5. *Bacillus anthracis* Ames, 6. *Bacillus anthracis* Sterne, 7. *Bacillus cereus* ATCC 10987. A vertical red line indicates the current cursor position at 7,355 bp.

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 Overview

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

First, select the type of format for displaying the results (“Heat Map” or “Matrix”), protein/functional families (COG, Pfam, TIGRfam, Enzyme), normalization method, and a set of genomes in the **Abundance Profile Overview** page, as illustrated in Figure 5.4(ii). For “Heat Map” output, the abundance of protein/functional families is displayed as a heat map with red corresponding to the most abundant families, as illustrated 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 illustrated in Figure 5.4(iv). Click on the identifier of the family displayed on the right of the column (e.g., COG0531) in order to include the corresponding family into the **Function Cart**, as illustrated in Figure 5.4(v).

The screenshot shows the IMG Abundance Profile Overview tool interface. The main navigation bar includes 'IMG Home', 'Find Genomes', 'Find Genes', 'Find Functions', and 'Compare Genomes'. The 'Abundance Profiles' section is highlighted. The 'Abundance Profile Overview (ii)' settings panel shows 'Output Type' set to 'Heap Map' and 'Normalization Method' set to 'None'. The 'Function' section is set to 'COG'. The 'Genomes' section shows two genomes selected: 'Thermoplasma acidophilum DSM 1728 (A)[F]' and 'Thermoplasma volcanium GSS1 (A)[F]'. The 'Abundance Profile Overview Results (iii)' section displays a heat map with columns for the two genomes and rows for various COG families. The 'Abundance Cell Gene List (iv)' section shows a list of genes for the selected families. The 'Function List (v)' section shows a list of functions in the cart, including 'COG0531' (Amino acid transporters) and 'COG2814' (Arabinose efflux permease).

FIGURE 5.4. Abundance Profile Overview: Results displayed with Heat Map Output Format.

Abundance Profile Overview (ii)

Display Options:

Output Type **Normalization Method**

Heap Map None

or **Matrix** Scale for genome size

Gene count Include all rows, in

Estimated gene copies² 500 functions per

Function:

COG

Pfam

Enzyme

TIGRFam

Genomes¹:

Please select 1 to 9 genomes.

Thermoplasma acidophilum DSM 1728 (A)[F]

Thermoplasma volcanium GSS1 (A)[F]

Abundance Profiles Tools (i)

The following tools operate on functional profiles of multiple genomes.

Tool	Description
Abundance Profile Overview(All Functions)	View abundance for all functions across
Abundance Profile Search	Search for functions based on over or

Abundance Cell Gene List (iv)

Add Selected to Gene Cart Select All

638190515 hypothetical protein (258aa) (est_cop
(BA000011) 1584804bp gc=0.40)

Abundance Profile Overview Results (iii)

Pages: [1] 2 [Next Page]

[Download tab-delimited file for Excel](#)

Add Selected to Function Cart Select All Clear All

Select	Row No.	ID	Name	The aci 178	The vol GS1
<input checked="" type="checkbox"/>	1	COG0001	Glutamate-1-semialdehyde aminotransferase	1	1
<input checked="" type="checkbox"/>	2	COG0003	Oxyanion-translocating ATPase	1	2
<input checked="" type="checkbox"/>	3	COG0004	Ammonia permease	1	1

Function List (v)

Remove Selected Select All

Selection	Function ID	Name
<input checked="" type="checkbox"/>	COG0001	Glutamate-1-semialdehyde
<input checked="" type="checkbox"/>	COG0003	Oxyanion-translocating AT
<input checked="" type="checkbox"/>	COG0004	Ammonia permease

FIGURE 5.5. Abundance Profile Overview: Results displayed with Matrix Output Format.

If the “Matrix” output is selected, as illustrated in Figure 5.5(ii), the abundance of protein/functional families is displayed in a tabular format, as illustrated in Figure 5.5(iii), with each row corresponding to a family and each cell containing the number of genes associated with a family for a specific 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 5.5(iv). Families of interest can be selected for inclusion into the **Function Cart**, as illustrated in Figure 5.5(v). The results in “Matrix” format can be exported to a tab-delimited Excel file.

By default, the **Abundance Profile Overview** 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 illustrated in Figure 5.6(ii). Abundance cutoffs can be set up for the genomes of interest. The **Abundance Profile Search** results are displayed in a tabular format, as illustrated in Figure 5.6(iii). The gene counts in the results table are linked to the corresponding lists of genes, as illustrated in Figure 5.6(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 illustrated in Figure 5.6(v).

Abundance Profile Search (ii)

Find genes in genome (bin) of interest qualified by similarity to sequences in other genomes (base user-selected genomes appear in the profiler).

Genome Completion: [F]inished, [D]raft.

Select functional classification
 COG
 Pfam

Normalization Method
 None
 Frequency

Show Results As:
 Gene Counts
 Normalized Values
 Both

More Abundant Cut-Off: 1 Less Abundant Cut-Off: 1

Enter matching text for highlighting clusters/rows. (E.g., "kinase").

Find Functions In*	More Abundant Than In	Less Abundant Than In	Ignoring	Taxon Name
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Archaea
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Euryarchaeota
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thermoplasma
<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thermoplasma acidophilum DSM 1728 [F]
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thermoplasma volcanium GSS1 [F]

Abundance Profile Search Results (iii)

Add Selected to Function Cart Select All Clear All

Red - Less abundant.
Green - More abundant.

Selection	Cog Id	Cog Name	Thermoplasma acidophilum DSM 1728	Thermoplasma volcanium GSS1
<input checked="" type="checkbox"/>	COG0006	Xaa-Pro aminopeptidase	2	1
<input checked="" type="checkbox"/>	COG0028	Thiamine pyrophosphate-requiring enzymes [acetolactate synthase, pyruvate dehydrogenase (cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase]	4	2
<input checked="" type="checkbox"/>	COG0043	3-polyprenyl-4-hydroxybenzoate decarboxylase and related decarboxylases	2	1
<input type="checkbox"/>	COG0121	Predicted glutamine amidotransferase	1	0
<input type="checkbox"/>	COG0136	Aspartate-semialdehyde dehydrogenase	1	0

Abundance Profile Search Gene List (iv)

Add Selected to Gene Cart Select All

Selection	Gene Id	Gene Name
<input type="checkbox"/>	638180669	proline dipeptidase related protein
<input type="checkbox"/>	638181252	proline dipeptidase related protein

FIGURE 5.6. 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 illustrated in Figure 5.7(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 illustrated in Figure 5.7(i).

The results of **hierarchical clustering** are displayed in a tree format, as illustrated in Figure 5.7(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 illustrated in Figure 5.7(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.

(i) **img** INTEGRATED MICROBIAL GENOMICS

IMG Home Find Genomes Find Genes Find Functions **Compare Genomes** Add Genomes

Genome Statistics VISTA Abundance Profiles **Genome Clustering**

Genome Clustering

You may cluster samples (genomes) based on similar COG, Pfam, or enzyme profiles. Proximity of grouping indicates the relative degree of similarity of samples to each other.

Select Genome(s): Please select between 2 to 90 genomes.
Domains: (B)acteria, (A)rchaea, (E)ukarya, (P)lasmiids, (V)iruses.
Genome Completion: [F]inished, [D]raft.

Ferrop^lasma aci^darmanus Fer1 (A)[D]
Picro^philus torridus DSM 9790 (A)[F]
Thermop^lasma aci^dophilum DSM 1728 (A)[F]
Thermop^lasma vol^canium GSS1 (A)[F]

Functional Profile:
 COG
 Pfam
 Enzymes

Clustering method:
 Hierarchical Clustering
 Principle Components Analysis (PCA)
 Correlation Matrix

Hierarchical Clustering Results (ii)

Hierarchical clustering shown at bottom of page. You can select on the genome name to see its details.

Find profile for *Ferrop^lasma aci^darmanus Fer1* ...
Find profile for *Picro^philus torridus DSM 9790* ...
Find profile for *Thermop^lasma aci^dophilum DSM 1728* ...
Find profile for *Thermop^lasma vol^canium GSS1* ...
Clustering genomes ...

[Ferrop^lasma aci^darmanus Fer1](#)
[Thermop^lasma vol^canium GSS1](#)
[Thermop^lasma aci^dophilum DSM 1728](#)
[Picro^philus torridus DSM 9790](#)

Correlation Matrix Results (iii)

Genome	Fer aci Fe1	Pic tor 970	The aci 178	The vol GS1
Ferrop^lasma aci^darmanus Fer1	1.00	0.83	0.83	0.85
Picro^philus torridus DSM 9790	0.83	1.00	0.80	0.79
Thermop^lasma aci^dophilum DSM 1728	0.83	0.80	1.00	0.93
Thermop^lasma vol^canium GSS1	0.85	0.79	0.93	1.00

Pearson R Color Legend

- ≤ 0
- > 0.00 - ≤ 0.10
- > 0.10 - ≤ 0.20
- > 0.20 - ≤ 0.30
- > 0.30 - ≤ 0.40
- > 0.40 - ≤ 0.50
- > 0.50 - ≤ 0.60
- > 0.60 - ≤ 0.70
- > 0.70 - ≤ 0.80
- > 0.80 - ≤ 0.90

FIGURE 5.7. 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."

The screenshot shows the IMG Gene Cart interface. The main navigation bar includes 'IMG Home', 'Find Genomes', 'Find Genes', 'Find Functions', 'Compare Genomes', 'Analysis Carts', 'MyIMG', and 'About'. The 'Analysis Carts' menu is open, showing 'Genes' and 'Functions' options. The 'Gene Cart' section has a sidebar with links: 'Gene List', 'Upload & Export', 'Comparison Tools', 'Chromosome Map', 'Sequence Alignment', 'Gene Neighborhoods', 'Profile Tools', 'Gene Profile', and 'Occurrence Profile'. The 'Gene List' section shows '4 gene(s) in cart' and buttons for 'Remove Selected', 'Select All', and 'Clear All'. Below is a table with columns: Selection, Gene Object ID, Locus Tag, Product Name, AA Seq. Length, Genome, and Batch. The table contains four rows of gene data. A 'Gene Export (iii)' window is open, showing the selected gene's FASTA sequence. A 'Show In Export Format' button is highlighted in the 'Export Genes' section.

Selection	Gene Object ID	Locus Tag	Product Name	AA Seq. Length	Genome	Batch ¹
<input checked="" type="checkbox"/>	638187736	APE0079	132aa long hypothetical protein (EC:4.1.1.50)	132aa	Aeropyrum pernix K1	1
<input checked="" type="checkbox"/>	638189791	APE1968	carbamate kinase (EC 2.7.2.2) (IMGterm)	31		
<input checked="" type="checkbox"/>	638189816	APE1992	ornithine carbamoyltransferase (EC 2.1.3.3) (IMGterm)	31		
<input checked="" type="checkbox"/>	638189170	APE1386	glutamate dehydrogenase (NADP) (EC 1.4.1.4) (IMGterm)	42		

```

>638187736 APE0079 132aa long hypothetical prote
MERREDVIVGKHVYGSLSYGVPREKATDEEYLRGVVRAAESAGATVHA
SWTIPGKGGVSVIVLVLESHLALHTWPEYDYATFDIYTCGEHTDPWF
ELLLSELKPRKRYTVHVDRSQEKTVLEAQPFR
  
```

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).

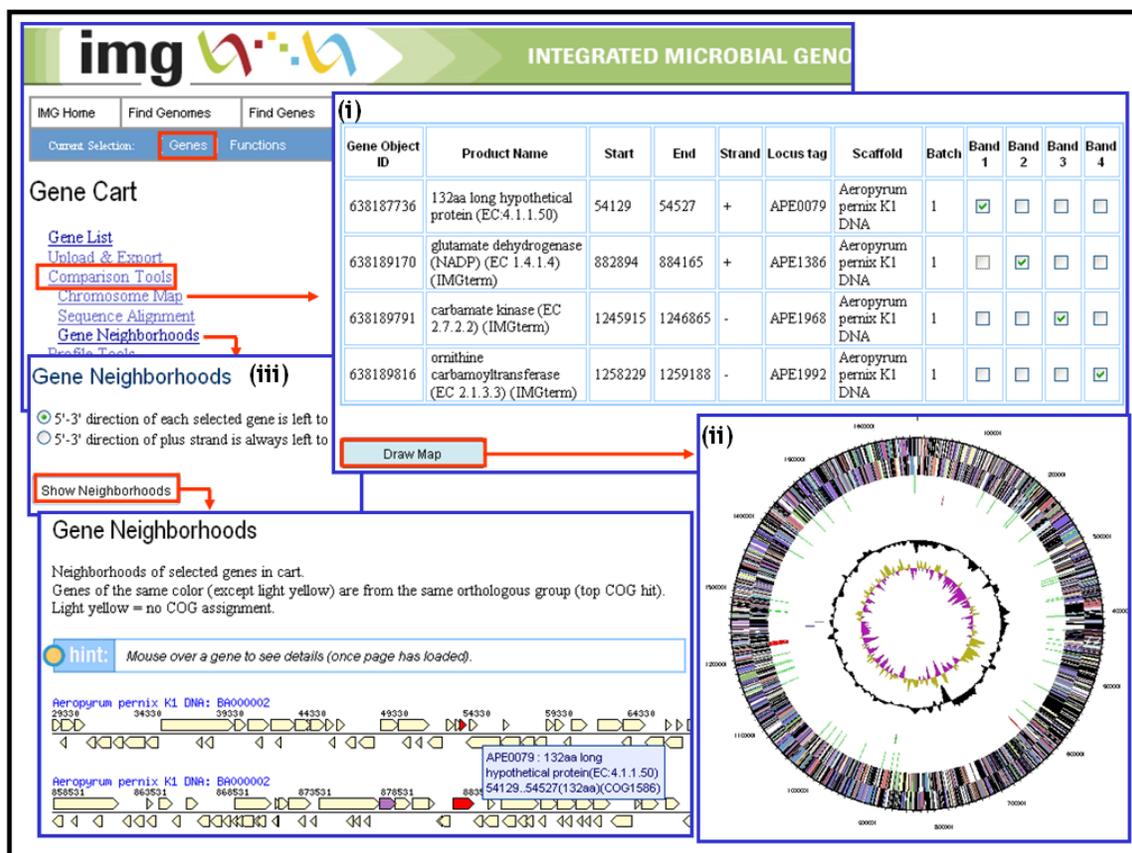


FIGURE 6.2. Gene Cart/Comparison Tools: Chromosome Map and Gene Neighborhoods.

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

- Genes on the forward strand colored by COG categories
- 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.

ClustalW Alignment (ii)

Identities computed with respect to: (1) 638187736
Colored by: consensus/80% and property

```

1 638187736 -----MERREDVIIVGKIVYGLYGVPEKATF
2 638189170 VEVLALQPTDPLEERRRQLRRRVBLLGYDYYVYEVLANPDRVLQVRVTIKKDDGTVKTEFLGWRSSQHNST
3 638189791 -----VDSGLRAVIALGGNAIAGPGMDVSVESQTARVKKRASSIADVLADGWRSVITHGNGPQVGYT
4 638189816 -----LWVEGLSLHRLRGRMLLWLADYTG--EEIRRDVELTLEMKRRYYAGERVIPVLRG--RSVGLL
consensus/100% .....p.hh.p.hhhsthh...hhG...p.h...
consensus/90% .....p.hh.p.hhhsthh...hhG...p.h...
consensus/80% .....p.hh.p.hhhsthh...hhG...p.h...
consensus/70% .....p.hh...hhh.h.tphl.h.shss...p.ht.schshp+tc.lhssctshshhG...pp.hhh

```

Mview 1.47.3, Copyright (c) Nigel P. Brown 1997-2002

Gene Object ID	Product Name	Genome	Scaffold ID
638187736	132aa long hypothetical protein	Aeropyrum pernix K1	BA000002
638189170	glutamate dehydrogenase (NADP) (EC 1.4.1.4) (IMGterm)	Aeropyrum pernix K1	BA000002
638189791	carbamate kinase (EC 2.7.2.2) (IMGterm)	Aeropyrum pernix K1	BA000002
638189816	ornithine carbamoyltransferase (EC 2.1.3.3) (IMGterm)	Aeropyrum pernix K1	BA000002

Jalview alignment editor (iii)

File Edit Font View Colour Calculate Align Help

638187736/1-423 -----MERREDVIIVGKIVYGLYGVPEKATDEVLRSQVRAKESAGATVHNSM-----IPGGGGVV
638189170/1-423 VEVLALQPTDPLEERRRQLRRRVBLLGYDYYVYEVLANPDRVLQVRVTIKKDDGTVKTEFLGWRSSQHNST
638189791/1-423 -----VDSGLRAVIALGGNAIAGPGMDVSVESQTARVKKRASSIADVLADGWRSVITHGNGPQVGYT
638189816/1-423 -----LWVEGLSLHRLRGRMLLWLADYTG--EEIRRDVELTLEMKRRYYAGERVIPVLRG--RSVGLL
Qual ity/1-423

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.

Gene Profile (i)

View selected protein coding genes against using unidirectional sequence similarities. Please select 1 to 10 genome(s).
 Domains: (B)acteria, (A)rchaea, (E)ukarya
 Genome Completion: [F]inished, [D]raft.
 Thermococcus kodakaraensis KOD1 (A)[F]
 Thermofilum pendens Hrk 5 (A)[F]
 Thermoplasma acidophilum DSM 1728 (A)[F]
 Thermoplasma volcanium GSS1 (A)[F]
 uncultured methanogenic archaeon RC-1 (A)[F]
 Acidiphilium cryptum JF-5 (B)[F]
 Acidobacteria bacterium Ellin345 (B)[F]
 Acidothermus cellulolyticus 11B (B)[F]
 Acidovorax avenae citrulli AAC00-1 (B)[F]
 Acidovorax sp. JS42 (B)[F]

Max. E-value: 0.1
 Min. Percent Identity: 10

View Genes vs. Genomes | View Genomes vs. Genes

Gene Profile (ii)

Gene Object ID	Product Name	The aci 178	The vol GS1
638187736	132aa long hypothetical protein	1	1
638189170	glutamate dehydrogenase (NADP) (EC 1.4.1.4) (IMGterm)	2	2
638189791	carbamate kinase (EC 2.7.2.2) (IMGterm)	1	1
638189816	ornithine carbamoyltransferase (EC 2.1.3.3) (IMGterm)	2	2

Gene Profile (iii)

Genome (bin) Name	638187736	638189791	638189816	638189170
Thermoplasma acidophilum DSM 1728	1	1	2	2
Thermoplasma volcanium GSS1	1	1	2	2

Profile Genes (iv)

Add Selected to Gene Cart | Select All

638190779 ornithine carbamoyltransferase (EC 2.1.3.3)
 638191163 aspartate carbamoyltransferase (EC 2.1.3.2)

Phylogenetic Occurrence Profile (v)

Domains(D): B=Bacteria, A=Archaea, E=Eukarya.
 (Profiles based on bidirectional best hit orthologs.
 A dot '.' means there are no bidirectional best hit. for the genome.)

```

638187736  AAAAAAAAAAAAAAAAAA.....AAA..AA.....
638189170  AAAAAAAAAAAAAAAAAA.A.....A.A..AAAA.AAAAAAAAAABBBB.BBBBB..B...BBBBB.
638189791  AAA.....AAAAA..A.....A.....AAAAA..B.....
638189816  AAAAAAAAAAAAAAAAAA.....BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
        [Euryarchaeota] Methanoculleus marisnigri JR1

638187736  .....B.....B.....B.....B.....
638189170  BBBB..B..BB.BBBBBB.BBB..BBBBBB.BBBBBBBBBBBB.BBBBB.BBBBBBBBBBBB..BBBB.
638189791  B.B.....BB.B.....BB.....
638189816  BBBB..BBBBBBBBBBBBBBBBBBBBBBBBBB.....BB....B....B.....BBBBBBB
  
```

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**.

The screenshot displays the IMG 2.6 web interface. At the top, the 'Analysis Carts' menu is highlighted, and the 'Functions' sub-menu is selected. The 'Function Cart' section shows a list of 5 functions with checkboxes for selection. The 'Function Profile (iii)' tool is active, displaying a list of genomes and a table of gene counts for the selected functions. The 'Phylogenetic Occurrence Profile (v)' tool is also visible, showing a dot plot of function occurrence across domains.

Function List

5 function(s) in cart

Selection	Function ID	Name
<input checked="" type="checkbox"/>	COG0002	Acetylglutamate semialdehyde dehydrogenase
<input checked="" type="checkbox"/>	COG0078	Ornithine carbamoyltransferase
<input checked="" type="checkbox"/>	COG0137	Argininosuccinate synthase
<input checked="" type="checkbox"/>	COG0160	4-aminobutyrate aminotransferase and related aminotransferases
<input checked="" type="checkbox"/>	COG0165	Argininosuccinate lyase

Function Profile (iii)

View selected function(s) against selected genomes.
Please select 1 to 50 genome(s).
Domains: (B)acteria, (A)rchaea, (E)ukarya, (P)lasmiids, (V)iruses.
Genome Completion: [F]inished, [D]raft.

Selected Genomes:

- Thermoplasma acidophilum DSM 1728 (A)[F]
- Thermoplasma volcanium GSS1 (A)[F]
- uncultured methanogenic archaeon RC-1 (A)[F]
- Acidiphilium cryptum JF-5 (B)[F]
- Acidobacteria bacterium Ellin345 (B)[F]
- Acidothermus cellulolyticus 11B (B)[F]
- Acidovorax avenae citrulli AAC00-1 (B)[F]
- Acidovorax sp. JS42 (B)[F]
- Acinetobacter haumannii ATCC 17978 (B)[F]

Max. E-value: 0.1
Min. Percent Identity: 10

Function Profile (iv)

Genome	D	COG 0002	COG 0078	COG 0137	COG 0160	COG 0165
Thermoplasma acidophilum DSM 1728	A	0	1	1	1	1
Thermoplasma volcanium GSS1	A	0	1	1	1	1

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.

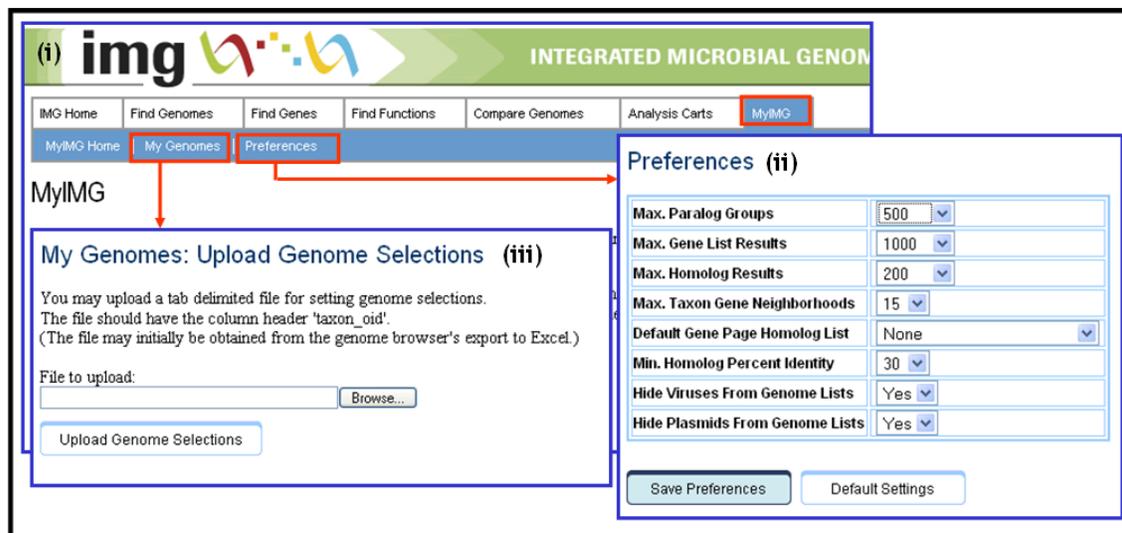


FIGURE 7.1. MyIMG.

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](http://www.ncbi.nlm.nih.gov/Genbank/genomesubmit.html#locus_tag) 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>