

img INTEGRATED MICROBIAL GENOMES

IMG Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Carts MyIMG Using IMG

IMG Genomes	finished/draft	Total
Bacteria	782/502	1284
Archaea	55/3	59
Eukarya	19/30	49
Plasmids	974/0	974
Viruses	2524/0	2524
All Genomes	4355/535	4890

Genome by Metadata
[IMG Statistics](#)
[Project Map](#)
[Content History](#)

IMG 2.8: What's New

IMG 2.8 is the **16th** release of the Integrated Microbial Genomes (IMG) genomic data management and analysis system. **IMG 2.8** was released on **April 6th, 2009**.

IMG 2.8 Content

Genomes

The content of **IMG 2.8** has been updated with new microbial genomes available in **RefSeq version 33** (January 22, 2009). **IMG 2.8** contains a total of **4,890** genomes consisting of **1,284** bacterial, **59** archaeal, **49** eukaryotic genomes, **2,524** viruses (including bacterial phages), and **974** plasmids that did not come from a specific microbial genome sequencing project. Among these genomes, **4,355** are finished genomes, and **535** are draft genomes.

The following **eukaryotic** genomes have been added to **IMG 2.8**:

1. **Fungi**
 - a. *Malassezia globosa* CBS 7966
 - b. *Enterocytozoon bieneusi* H348
 - c. *Saccharomyces cerevisiae* AWRI1631
 - d. *Aspergillus oryzae* RIB40
 - e. *Saccharomyces cerevisiae* YJM789
2. **Alveolata**
 - a. *Babesia bovis* T2Bo
 - b. *Plasmodium berghei*
 - c. *Plasmodium chabaudi chabaudi*
 - d. *Plasmodium vivax* Sal-1

Note that **28** microbial genomes from **IMG 2.7** were **replaced** in **IMG 2.8** because (1) a "Draft" genome has been replaced by its "Finished" version or (2) the composition of the genome has changed through the addition of new replicons (plasmids, chromosomes). For replaced

genomes, the gene object identifiers (gene OIDs) for the protein-coding genes (CDS) were mapped to their new version in IMG. 2.8. See IMG [Data Evolution History](#) for details.

tRNA and **rRNA** (23S, 16S, 5S) genes missing from original RefSeq genome files are added using tRNAscan-SE v1.23 for tRNA genes and similarity comparisons to existing RNA genes. In IMG 2.8 **434** tRNA, **264** rRNA and **6,662** misc_RNA genes were added in **546** genomes.

A **chromosomal cassette**, which is defined as a stretch of protein coding genes with intergenic distance smaller or equal to 300 base pairs, has been extended to include genes *on the same or different strands*¹.

The functional characterization of genomes has been extended with **KEGG Orthology (KO) terms** which serve as the main vehicle for associating IMG genomes with **KEGG pathways**. IMG genes are associated with KO terms as follows:

1. First, IMG genes that could be mapped to genes in KEGG's list of genes, were assigned the KO terms associated with the corresponding KEGG gene. The IMG to KEGG gene mapping was based on using NCBI's GI numbers and GeneIDs.
2. For IMG genes that were not mapped to KEGG genes in the first stage above, BLASTP was run against the database of KEGG genes, with soft masking (-F 'm S') for low complexity regions turned on. The results of this search are organized in a list of candidate KO assignments, where an E-value cutoff of 1e-2 for the top 25 KEGG gene hits is employed. This list of KO assignments is used for searching potentially "missing KO terms". KO terms are assigned to IMG genes using a subset of this list, where the threshold defined by an E-value cutoff of 1e-5, KO assignment rank of 5 or better, and alignment percentage of at least 70% over the length of the IMG query gene and KEGG subject gene.

¹ In IMG 2.7, only genes on the same strand or divergent genes were considered, while convergent genes were not included in a chromosomal cassette.

IMG Statistics

Various statistics are provided via the **IMG Statistics** link on the home page of IMG, as shown below, including **IMG Total Gene Count** which consists of counting all the genes (protein coding genes, RNA genes) in IMG, except obsolete genes. Compared to **IMG 2.7**, **IMG 2.8** contains **5,504,487 genes**, an increase of **574,141 genes**.

The screenshot displays the IMG Statistics interface. At the top, it shows the **IMG Total Gene Count: 5504487**. The main content is divided into several sections:

- Domain Statistics:** A table showing the distribution of genomes and genes across different domains.
- Function Statistics:** A table detailing the counts and percentages of various functional categories.
- IMG Cluster Statistics:** A table showing the distribution of clusters across different categories.
- Pathway Statistics:** A table showing the counts and percentages of various pathways.
- IMG Content History:** A bar chart showing the growth of IMG in terms of number of genomes and genes since its first version in March 2005.
- Project Map:** A Google Map displaying the location of isolation sites for genomes that are associated with longitude/latitude coordinates in GOLD.

Domain	Genome Count	Gene Count	% of Total
Bacteria	1284	4699244	85.37%
Archaea	59	134217	2.44%
Eukaryota	49	569653	10.35%
Plasmid	974	27077	0.49%
Viruses	2524	74296	1.35%
Total	4990	5504487	100.00%

Function	Total Count	Protein Coding Genes with	% of Total Protein Coding Genes
COG	4873	3520079	63.95%
Pfam	10340	3294995	59.86%
Enzyme	3317	630314	11.45%
TIGRFam	3418	1454638	26.43%
IMG Term	3920	967101	17.57%
GO-Molecular Function	9281	565551	10.27%
GO-Cellular Component	2400	272756	4.96%
KO Terms	11507	2325682	42.25%
Protein Product Name	487509	3200004	58.13%
No Protein Product Name	35762	2131202	38.72%
SwissProt	74624	316484	5.75%
Total	4019580	4019580	73.02%

Cluster Type	Total Count	Protein with	% of Total Gene
COG	4873	3520079	63.95%
Pfam	10340	3294995	59.86%
TIGRFam	3418	1454638	26.43%
IMG Ortholog Clusters	268589	4841866	87.96%
IMG Chromosomal Cassettes	535899	4532005	82.33%
Conserved IMG Chromosomal Cassettes by			
COG Clusters	8652081	4161391	75.60%
Pfam Clusters	25998158	3748667	68.10%
IMG Ortholog Clusters	2760407	4314458	78.38%

Pathway	Total Count	Protein Coding Genes with	% of Total Protein Coding Genes
COG Pathway	77	2334817	42.42%
Kegg Pathway	345	1318617	23.96%
TIGRFam Roles	106	1287192	23.38%
IMG Pathway	641	333384	6.06%
IMG Parts List	52	354774	6.45%
MetaCyc	1395	609900	11.08%
GO-Biological Process	16479	481967	8.76%
KO Modules	685	761529	13.83%
Total	2944199	2944199	53.49%

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

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

IMG 2.8 User Interface

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

The main UI changes include:

(a) **New features**

- (i) The **Genome by Metadata** link on IMG's home page provides access to a classification of the archaeal, bacterial and eukaryotic genomes by several metadata categories.
- (ii) The **KEGG** collection of **pathways** has been reorganized and updated using the enhanced collection of KEGG resources, including **KEGG Orthology (KO)** terms and **KEGG pathway modules**.

(b) **Extended features**

- (i) The **Genome Statistics** of **Organism Details** has been extended with counts of "Protein coding genes connected to KEGG Orthology (KO) terms" and "Protein coding genes connected to SwissProt Protein Products".
- (ii) **Gene Details** has been extended to include SwissProt Protein Products (when available), KEGG Orthology (KO) term, and KO Modules. For genes without a product name, a **Find Candidate Product Name** tool provides a list of candidate protein product names.
- (iii) **Missing enzymes** feature has been adapted to the new KO term based enzyme annotations. Missing enzymes can be examined using either a **KEGG Pathway Map** for a genome of interest or a **Functional Profile** involving genomes and enzymes of interest.

New Features

Genome by Metadata

The **Genome by Metadata** link on IMG's home page provides access to a classification of the archaeal, bacterial and eukaryotic genomes by several metadata categories, as illustrated in Figure 1 (i).

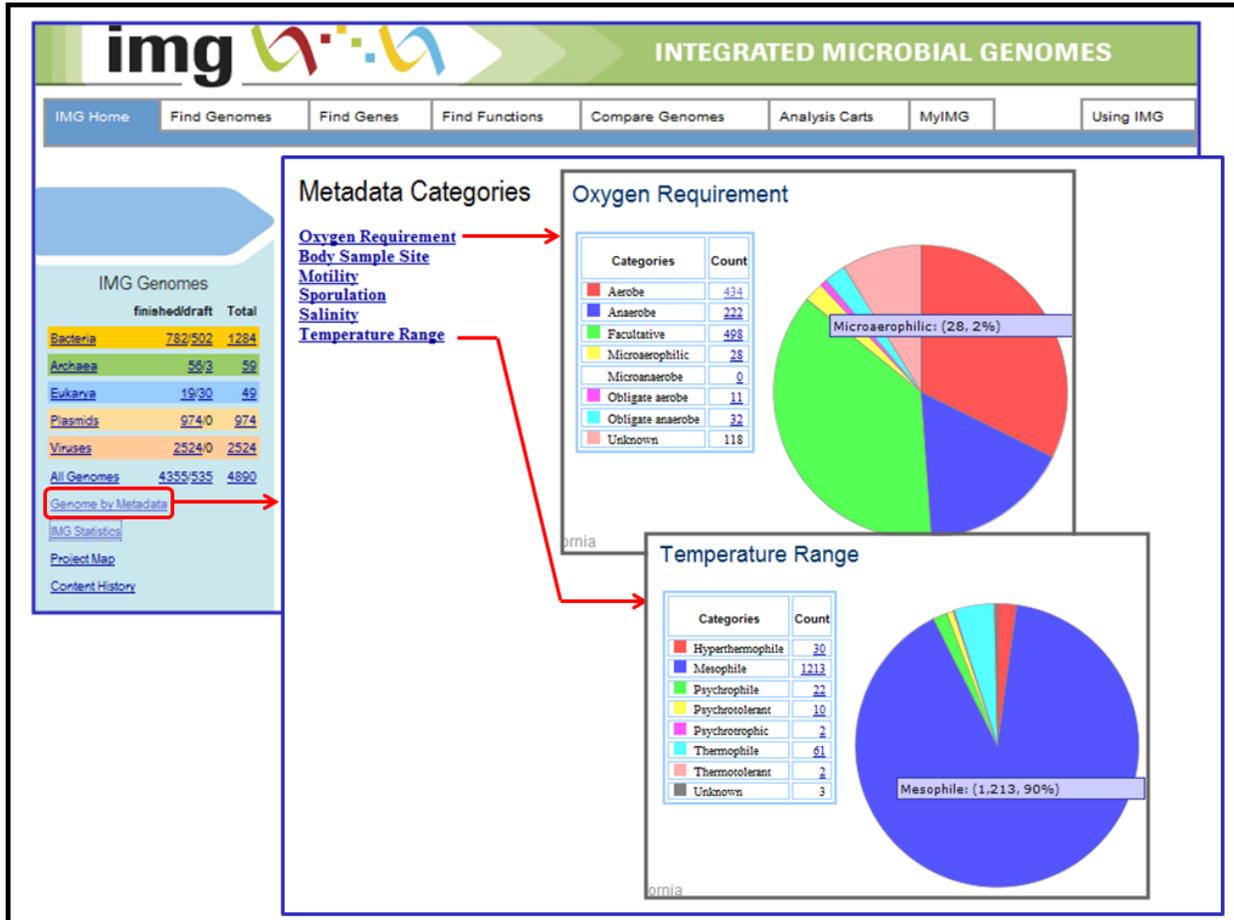


FIGURE 1. Genome by Metadata Category Classification.

Note that only a subset of the metadata categories available under **Genome Search** are provided, namely categories with controlled vocabularies of less than ten values. The metadata values are taken from GOLD (<http://www.genomesonline.org/>) and therefore reflect the level of information collection and curation in this resource.

Find Functions – KEGG Orthology Terms & Pathways

The **KEGG** collection of pathways in IMG has been reorganized and updated using the enhanced collection of KEGG resources², including KEGG Orthology (KO) terms and KEGG pathway modules.

From the **Find Function** top-level menu, the **KEGG** option on the second-level menu leads to the **KEGG Orthology Terms and Pathways** browser, as shown in Figure 2(i). KEGG Orthology (KO) terms identify orthologous groups of genes organized using the BRITE functional hierarchy (<http://www.genome.jp/kegg/brite.html> - see Figure 2(ii)).

The screenshot displays the KEGG Orthology Terms and Pathways browser interface. At the top, there is a navigation menu with 'Find Functions' selected, and a sub-menu where 'KEGG' is highlighted. The main content area is divided into several sections:

- KEGG Orthology (KO) Terms and Pathways (i):** This section includes links for 'KEGG Orthology (KO) Terms Based on BRITE Hierarchy', 'KEGG Pathways via KO Terms', and 'KEGG Pathways via EC Numbers'. It also features 'KO Term Distribution across Protein Families in IMG' and 'KO Term Distribution across Genomes and Paralog Clusters in IMG'.
- KEGG BRITE Database (ii):** This section provides a search interface for the BRITE database, including a search box for 'Enter br number' and a 'Brite hierarchy' button.
- KEGG Orthology (KO) Terms:** A hierarchical list of terms is shown, starting with '01 Metabolism' and '02 Carbohydrate Metabolism'. The term 'K00844' is highlighted, corresponding to 'HK: hexokinase [EC:2.7.1.1 2.7.1.2]'.
- KEGG Orthology: K00844 (iii):** This section provides a detailed view of the selected KO term, including its 'Entry' (K00844), 'Name' (HK), 'Definition' (hexokinase [EC:2.7.1.1 2.7.1.2]), and 'Class' (Metabolism; Carbohydrate Metabolism; Glycolysis / Gluconeogenesis).
- KEGG Orthology (KO) Term Gene List (iv):** This section displays a table of genes associated with the selected KO term. The table includes columns for 'Select', 'Gene Id', 'Gene Name', and 'Genome Name'. Three genes are listed: hexokinase (EC 2.7.1.1) from *Treponema pallidum pallidum* Nichols, *Treponema denticola* ATCC 35405, and *Bacteroides fragilis* NCTC 9343.

FIGURE 2. Find Functions – KEGG: KEGG Orthology Terms.

Each KO identifier (called K number) provides a link to the corresponding KEGG Orthology term specification, as illustrated in Figure 2(iii). The definition associated with a KO term provides a link to the list of IMG genes associated with that KO term, as illustrated in Figure 2(iv).

KO Term Distribution across protein families in IMG, genomes and paralog clusters in IMG helps assessing the consistency of protein family annotations in IMG:

1. The **KO Term Distribution across Protein Families**, illustrated in Figure 3(i), lists for each KO term the number of different COGs, Pfams, and TIGRFams it is associated with across

² Kanehisa & al. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, **36** (Database Issue): D480-484.

all the genes in IMG. The number of **unique** (COG, Pfam, TIGRfam) **combinations** associated with each KO term as part of an IMG gene annotation is also provided, whereby the details for these combinations can be examined using the **Details of KO Term Distribution across Protein Families** page, as illustrated in Figure 3(ii). For a specific (query) KO term, this page lists for each unique (COG, Pfam, TIGRfam) combination:

- the number of genes associated with the query KO term and this (COG, Pfam, TIGRfam) combination;
- the number of genes associated with this (COG, Pfam, TIGRfam) combination and a KO term different from the query KO term, including genes associated with multiple KO terms and a query KO term as one of them;
- the number of genes associated with this (COG, Pfam, TIGRfam) combination and a KO term different from the query KO term, and **not** associated with the query KO term;
- the number of genes associated with this (COG, Pfam, TIGRfam) combination and not associated with any KO term.

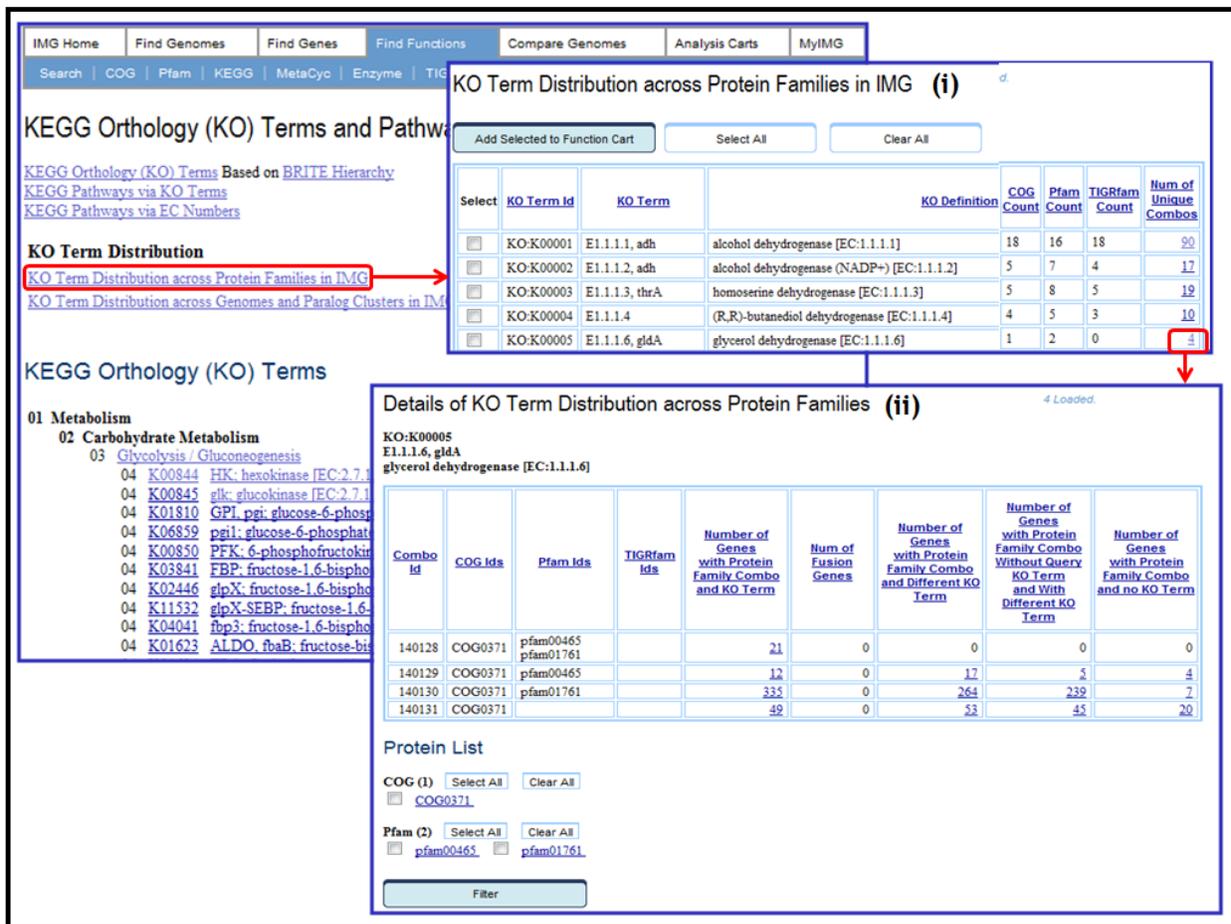


FIGURE 3. Find Functions – KEGG: KO Term Distribution across Protein Families in IMG.

2. The **KO Term Distribution across Genomes and Paralog Clusters**, illustrated in Figure 4(i), lists for each (query) KO term:

- the number of genes associated with the query KO term;
- the number of genomes that have genes associated with the query KO term;

- *average number of genes* associated with the query KO term per genome;
 - ☞ this metric helps identify KO terms that were assigned to multiple genes in the same genome, either by mistake or because these terms correspond to sequence similarity-based families rather than function-based groups;
- the number of genes associated with the query KO term that belong to paralog clusters; the list of these genes is provided in a separate page, where one can examine the paralogs annotated with the query KO term within each genome;
 - ☞ this metric indicates the likelihood of incorrect annotations due to the presence of paralogs.
- the number of genes associated with the query KO term, and that have a paralog annotated with the same KO term; the list of these genes is provided in a separate page, as illustrated in Figure 4(ii), where one can examine the paralogs annotated with the query KO term within each genome;
 - ☞ this number helps identifying incorrectly annotated paralogous genes.
- average % identity between the paralogs annotated with the same KO term;
 - ☞ if % identity is high (e.g, above 80%), annotation of both paralogs with the same term is likely correct; % identity is low (about 30%), the annotation is likely incorrect.

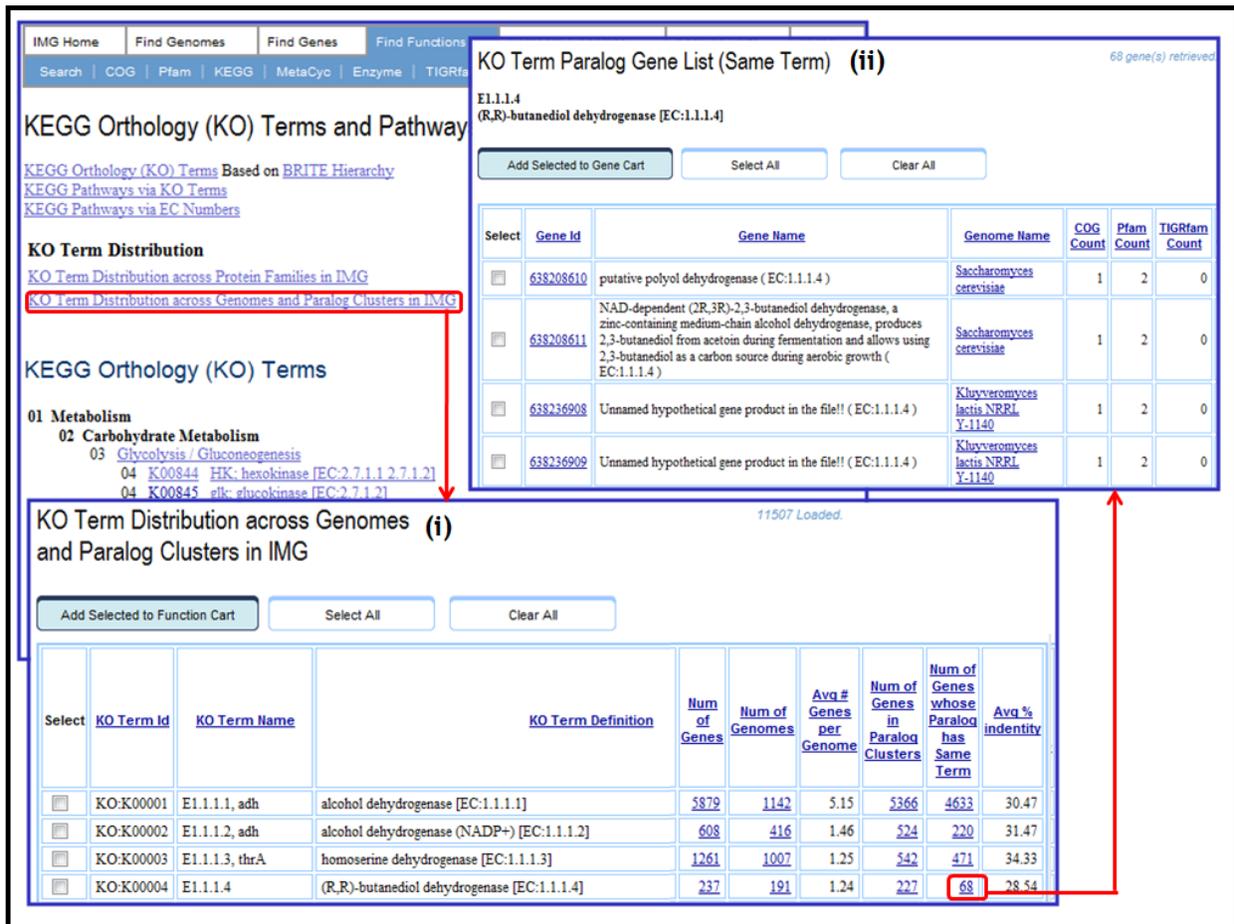


FIGURE 4. Find Functions – KEGG: KO Term Distribution across Genomes and Paralog Clusters.

Two KEGG Pathway browsers are available. The **KEGG Pathways via KO Terms** browser, illustrated in Figure 5(i), displays the KEGG pathways organized in **KEGG Modules**, which

represent smaller functional units, such as sequences of reactions and regulatory units. For a KEGG pathway, the **KEGG Pathway Details** provides the list of KO terms associated with a functional unit in each KEGG module in the pathway, as illustrated in Figure 5(ii). For a KEGG Module, a similar **KEGG Module Details** provides the list of KO terms associated with a functional unit in the KEGG Module, as illustrated in Figure 5(iii).

For each KO term, the number of genes associated with this term is also provided, together with a link that leads to the list of these genes. By clicking on the left-column checkbox for a KO term entry in the **KEGG Pathway Details** or **KEGG Module Details** page, KO terms can be added to the **Function Cart** for further analysis.

The screenshot displays the IMG 2.8 interface for finding functions. It is divided into several main sections:

- Top Navigation:** Includes links for IMG Home, Find Genomes, Find Genes, Find Functions, and a search bar with filters for COG, Pfam, KEGG, MetaCyc, Enzyme, and TIGRFam.
- KEGG Pathway Details (ii):** Shows details for the Glycolysis / Gluconeogenesis pathway. It includes a table of KEGG Orthology (KO) Terms in Pathway with columns for Select, KO Term Id, KO Name, Definition, KO Module Id, KO Module Name, Gene Count, and Genome Count.

Select	KO Term Id	KO Name	Definition	KO Module Id	KO Module Name	Gene Count	Genome Count
<input type="checkbox"/>	KO:K00134	GAPDH, gapA	glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12]	M00001	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	2036	1271
<input type="checkbox"/>	KO:K00844	HK	hexokinase [EC:2.7.1.1 2.7.1.2]	M00001	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	138	57
<input type="checkbox"/>	KO:K00171	porD	pyruvate ferredoxin oxidoreductase, delta subunit [EC:1.2.7.1]	M00679	Pyruvate oxidation, pyruvate => acetyl-CoA	155	115
<input type="checkbox"/>	KO:K00172	porG	pyruvate ferredoxin oxidoreductase, gamma subunit [EC:1.2.7.1]	M00679	Pyruvate oxidation, pyruvate => acetyl-CoA	176	134
- KEGG Pathways via KO Terms (i):** A hierarchical tree showing Metabolism, Carbohydrate Metabolism, and Glycolysis / Gluconeogenesis. Red arrows indicate navigation from the Glycolysis / Gluconeogenesis category to the pathway details.
- KEGG Orthology (KO) Terms in Module (iii):** Shows details for the KEGG Pathway Glycolysis / Gluconeogenesis. It includes a table of KO terms in the module and a 'View Pathway Map' button. A red arrow points from this button to the KEGG Map section.
- KEGG Map (iv):** Displays a metabolic map for the Glycolysis / Gluconeogenesis pathway. The map shows the conversion of alpha-D-Glucose to alpha-D-Glucose-6P and then to alpha-D-Glucose-1P. Various enzymes are represented by colored boxes with their EC numbers (e.g., 2.7.1.1, 3.1.3.10, 5.4.2.2). A legend indicates that blue boxes represent genes in Aeropyrum pernix K1 and orange boxes represent genes found in other genomes.

FIGURE 5. Find Functions – KEGG: KEGG Pathways via KO Terms.

The KEGG map associated with a KEGG pathway or KEGG module can be displayed for a selected genome, as shown in Figure 5(iv).

The alternative **KEGG Pathways via EC Numbers** browser, illustrated in Figure 6(i), displays the KEGG pathways organized in pathways categories similar to those used in IMG 2.7. For a KEGG pathway, the **KEGG Pathway Details** provides the list of EC numbers that are part of the KO terms associated with a functional unit in the KEGG pathway, as illustrated in Figure 6(ii).

The screenshot displays the 'KEGG Pathway Details (ii)' interface for 'Alanine and aspartate metabolism'. On the left, there are navigation links for 'KEGG Orthology (KO) Terms and Pathways' and 'KEGG Pathways via EC Numbers (i)'. The 'KEGG Map (iii)' section shows a metabolic pathway diagram with enzymes highlighted in yellow and orange boxes. A red arrow points from the 'Alanine and aspartate metabolism' link to the enzyme list table. The table lists enzymes with their EC numbers, names, and genome counts. A red box highlights the 'View Map' button at the bottom of the enzyme list.

Select	EC Number	Enzyme Name	Genome Count
<input type="checkbox"/>	EC:1.2.1.18	Malonate-semialdehyde dehydrogenase (acetylating).	1
<input type="checkbox"/>	EC:1.4.3.1	D-aspartate oxidase.	45
<input type="checkbox"/>	EC:1.4.3.15	D-glutamate(D-aspartate) oxidase.	0
<input type="checkbox"/>	EC:6.3.5.6	Asparaginyl-tRNA synthase (glutamine-hydrolyzing).	1097
<input type="checkbox"/>	EC:6.3.5.7	Glutaminyl-tRNA synthase (glutamine-hydrolyzing).	1110
<input type="checkbox"/>	EC:6.4.1.1	Pyruvate carboxylase.	553

FIGURE 6. Find Functions – KEGG: KEGG Pathways via EC Numbers.

For each EC number, the number of genes associated with this enzyme is also provided, together with a link that leads to the list of these genes. By clicking on the left-column checkbox for an enzyme entry in the **KEGG Pathway Details**, enzymes can be added to the **Function Cart** for further analysis.

Extended Features

Organism Details –Genome Statistics

The **Genome Statistics** of **Organism Details** has been extended with counts of “Protein coding genes connected to KEGG Orthology (KO) terms” and “Protein coding genes connected to SwissProt Protein Products”, as illustrated in Figure 7(i). SwissProt product names (see <http://ca.expasy.org/sprot/>), which are manually curated and therefore considered of high quality, can be examined via links to their SwissProt definition page, as illustrated in Figure 7(ii).

The **Compare Gene Annotations** tool available on the **Organism Details** page has been extended to include the KO term and SwissProt annotations, as illustrated in Figure 7(iii).

The screenshot displays the 'Organism Information (i)' section for *Thermoplasma volcanium* GSS1. The 'Genome Statistics' table shows the following data:

Category	Count	Percentage
Protein coding genes with function prediction		
Protein coding genes without function prediction		
Genes w/o function with similarity		
Genes w/o function w/o similarity		
Protein coding genes connected to KEGG pathways ³		
not connected to KEGG pathways		
Protein coding genes connected to KEGG Orthology (KO)	633	42.29%
not connected to KEGG Orthology (KO)	45	2.79%
Protein coding genes connected to MetaCyc pathways	1516	93.87%
not connected to MetaCyc pathways	95	5.88%
Protein coding genes connected to SwissProt Protein Product	1466	90.77%
not connected to SwissProt Protein Product	153	9.23%
Protein coding genes with enzymes		
Protein coding genes with COGs ³		
with Pfam ³		
with TIGRfam ³		
with InterPro		
with IMG Terms		
with IMG Pathways		
with IMG Parts List		
Protein coding genes in ortholog clusters		
in paralogs clusters		
in Chromosomal Cassette		
Number of Chromosomal Cassettes		
Fused Protein coding genes		
as fusion components		

The 'SwissProt Genes (ii)' section shows a table of selected genes:

Select	Gene Id	Gene Name	SwissProt Product Name
<input type="checkbox"/>	638190497	DNA-binding TFAR19-related protein (IMGterm)	DNA-binding protein TV0008
<input type="checkbox"/>	638190498	LSU ribosomal protein L39E (IMGterm)	50S ribosomal protein L39e

The 'Compare Gene Annotations (iii)' section shows a table of annotations for *Thermoplasma volcanium* GSS1:

Gene Object ID	Locus Tag	Source	Cluster Annotation	Gene Annotation	E-value
638190498	TVG0008653	pfam00832	Ribosomal_L39		1.2e-19
638190498	TVG0008653	product_name		ribosomal protein large subunit L39	
638190498	TVG0008653	ITERM:00272		LSU ribosomal protein L39E	
638190498	TVG0008653	DNA_length		156bp	
638190498	TVG0008653	Protein_length		51aa	
638190498	TVG0008653	SwissProt	50S ribosomal protein L39e		
638190498	TVG0008653	KO:K02924	large subunit ribosomal protein L39e		2.0e-22

FIGURE 7. Genome Statistics- Genes with KEGG Orthology terms, SwissProt Protein Products.

Gene Details – KO terms, KEGG Modules

The **Gene Details** page has been extended to include SwissProt Protein Products (when available), KEGG Orthology (KO) term, and KO Modules, as illustrated in Figure 8(i). For genes without a protein product name, such as that shown in Figure 8(i), the **Find Candidate Product Name** tool provides a list of candidate protein product names from related (sequence similarity based) genes, as illustrated in Figure 8(ii).

The screenshot shows two main panels: (i) Gene Information and (ii) Find Product Name Search Results.

Panel (i) Gene Information:

- Gene Information:** Gene Object ID: 638190539, Gene Symbol: TVG0051618, Locus Tag: TVG0051618, Product Name: hypothetical protein, SwissProt Protein Product: V-type ATP synthase subunit C.
- Protein Information:** Amino Acid Sequence Length: 356aa.
- COG:** COG1527 - Archaeal/vacuolar-type H⁺-ATPase subunit C.
- Families:** IPR002843 ATPase, V0/A0; IPR008266 Tyrosine protein.
- Transmembrane Helices:** No.
- Signal Peptide:** No.
- Statistics:** peptide.
- Pfam:** pfam01992 - vATP-synt_AC3.
- Pathway Information:** EC:3.6.3.14 - H(+)-transporting ATP synthase, subunit C.
- Enzymes:** EC:3.6.3.14 - H(+)-transporting ATP synthase, subunit C.
- KEGG Orthology (KO) Term:** KO:K02119 ATPase, ntpC V-type.
- KEGG Pathway:** Oxidative phosphorylation.
- KEGG Orthology (KO) Modules:** V-type ATPase (Prokaryotes).

Panel (ii) Find Product Name Search Results:

Candidate Gene (OID: 638190539): hypothetical protein

Genome Name: Thermoplasma volcanium GSS1
Gene Product Name: hypothetical protein
COG: Archaeal/vacuolar-type H⁺-ATPase subunit C
Pfam: ATP synthase (CAC39) subunit

Select	Homolog Gene	Original Product Name	IMG Term Old	IMG Term	D	C	Genome	Percent Identity	Alignment On Candidate	Alignment On Homolog	E-value	Bit Score	TIGRfam	COG	Pfam
<input type="radio"/>	638180158	ATP synthase (subunit C) related protein				A	F Thermoplasma acidophilum DSM 1728	62.78			7.00e-141	494		Archaeal/vacuolar-type H ⁺ -ATPase subunit C	ATP synthase (CAC39) subunit
<input type="radio"/>	638204220	AlAO H ⁺ -ATPase subunit C				A	F Ferrophilus torridus DSM 9730	39.20			4.00e-72	266		Archaeal/vacuolar-type H ⁺ -ATPase subunit C	ATP synthase (CAC39) subunit
<input type="radio"/>	638393667	H ⁺ -transporting ATP synthase subunit C				A	D Ferroplasma acidimanus Fer1	37.11			1.00e-69	258		Archaeal/vacuolar-type H ⁺ -ATPase subunit C	
<input type="radio"/>	640535411	A(1)A(0)-type ATP synthase, subunit C				A	F uncultured methanogenic archaeon RC-1	23.30			2.00e-21	106		Archaeal/vacuolar-type H ⁺ -ATPase subunit C	ATP synthase (CAC39) subunit
<input type="radio"/>	638201513	H ⁺ -transporting ATP synthase, subunit C (atpC)				A	F Methanocaldococcus jannaschii DSM 2661	22.54			2.00e-22	101		Archaeal/vacuolar-type H ⁺ -ATPase subunit C	ATP synthase (CAC39) subunit

At the bottom, there is a "Find Candidate Product Name" button and a "Display Option: Show All" dropdown menu.

FIGURE 8. Gene Details - KEGG Orthology term and modules, SwissProt Protein Product.

Missing Enzymes – KEGG Maps & Function Profile

Genomes may have potentially “missing” associations with functional units (e.g., reactions) on KEGG pathways. Such associations, which are based **on KO terms** assigned to genes, are called **missing enzymes**. Missing enzymes can be examined using either a **KEGG Pathway Map** for a genome of interest or a **Functional Profile** involving genomes and enzymes of interest, as illustrated in Figure 9.

Once a KEGG pathway is selected using the **KEGG Browser** under **Find Functions**, you can view its map for a selected genome using the “Find missing enzymes” option, as illustrated in Figure 9(i). On the **KEGG Map**, such as that shown in Figure 9(ii), enzymes that are associated with genes of the target genome are colored blue, while so called “missing” enzyme are colored either **green**, for enzymes that have a candidate KO term hits to genes of the target genome, or **white** for enzymes without such hits. Clicking on a missing enzyme will lead to a **Find Candidate Genes for Missing Function** page, as shown in Figure 9(iii). Note that selection of a (green colored) missing enzyme that has a KO term hit enhances the chances of finding for it good candidate genes.

The screenshot displays the IMG 2.8 interface for examining missing enzymes. It is divided into several key sections:

- KEGG Pathway Details (i):** Shows details for Methionine metabolism, including enzymes in the pathway and a list of KEGG pathways via EC numbers. The 'Methionine metabolism' pathway is highlighted with a red box.
- Find Candidate Genes for Missing Function (iii):** A search interface for the genome *Candidatus Methanoregula boonei* 6A8, targeting the function (EC:4.2.1.22) Cystathionine beta-synthase. It offers search options like 'Using Homologs or Orthologs' and 'Using KO'.
- KEGG Map (for Finding Missing Enzymes) (ii):** A metabolic map of Methionine Metabolism. Enzymes are represented by boxes with EC numbers. The enzyme EC:4.2.1.22 is highlighted in red, indicating it is missing.
- Function Profile (v):** A table showing the presence of the enzyme across different genomes. The 'Candidatus Methanoregula boonei 6A8' row shows a '0' in the 'KO' column, indicating a missing enzyme.
- Candidate Genes for Missing Function (iv):** A table listing potential candidate genes from other genomes. The top entry is from *Aeropyrum pernix K1* with a percent identity of 47.70% and an e-value of 3.00e-87.

FIGURE 9. Examining Missing Enzymes via a KEGG Pathway Map or Function Profile.

You can find candidate genes of your target genome that could be associated with a missing enzyme by searching for genes that have **homologs/orthologs** associated with the missing enzyme, as illustrated in Figure 9(iii). You can search across all the genomes available in the system, across a subset of genomes within a certain domain/phyla/class, or only across the selected genomes. You can change the default values set for percent identity and e-value cutoffs and the number of retrieved homologs. Alternatively, you can employ **KO terms** for finding genes that could be associated with the “missing” enzyme. You can change the default values set for percent identity, e-value, and percent alignment cutoffs. The result of the search for candidate genes consists of a list of genes, as illustrated in Figure 9(iv), that can be selected and included into the **Gene Cart**.

In the result for a **Function Profile** involving enzymes, missing enzymes are identified by a “0”. Clicking on the “0” identifying a missing enzyme, as shown in Figure 9(v), will also lead to a **Find Candidate Genes for Missing Function** page.