IMG/M Genomes

	finished/draft	Total
Bacteria	805/577	1382
Archaea	59/14	73
Eukarya	19/30	49
Plasmids	974/0	974
Viruses	2524/0	2524
Microbiomes	0/65	65
All Genomes	4381/686	5067

Microbiome Projects Map

IMG/M 2.8: What’s New

IMG/M 2.8 was released on **June 7th, 2009** is an updated version of IMG/M which includes all isolate genomes from **IMG 2.8**, over 100 **GEBA** genomes, and new metagenome datasets.

IMG/M 2.8 Content

IMG/M 2.8 contains **reference genomes** from **IMG 2.8** (released in April 2009, see http://img.jgi.doe.gov) and **111 GEBA** genomes from **IMG/GEBA** (http://img.jgi.doe.gov/geba) integrated with **metagenome** datasets generated from samples for the following metagenome projects, where new metagenome datasets in IMG/M 2.8 are highlighted in red:

1. **Endobiotic**
    a. five obese and lean mouse gut samples;
    b. two human distal gut samples;
    c. three oral TM7 samples;
    d. two termite hindgut samples;
    e. four gutless marine worm sample;
    f. one endopytic microbiome from rice;
2. **Environmental**
    a. two airborne samples from in an indoor urban environment;
    b. an acid mine drainage (AMD) biofilm sample;
    c. two uranium contaminated and pristine groundwater samples (the latter is a newly added sample);
    d. two hot spring microbial communities from Yellowstone Bath Hot Spring;
    e. five hot spring microbial communities from Yellowstone Bison Hot Spring;
    f. one hot spring microbial community from Obsidian Hot Spring;
    g. seven marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)[1].

---

[1] DeLong EF et al. (2006) Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science* **27**: 496-503.

    h. ten hypersaline microbial mat samples;
    i. three fossil marine "whale fall" carcasses;
    j. six freshwater sediment methylotrophic communities samples from Lake Washington in Seattle (a new combined sample has been added);
    k. a methane-oxidizing archaeal community sample
    l. an agricultural soil sample.

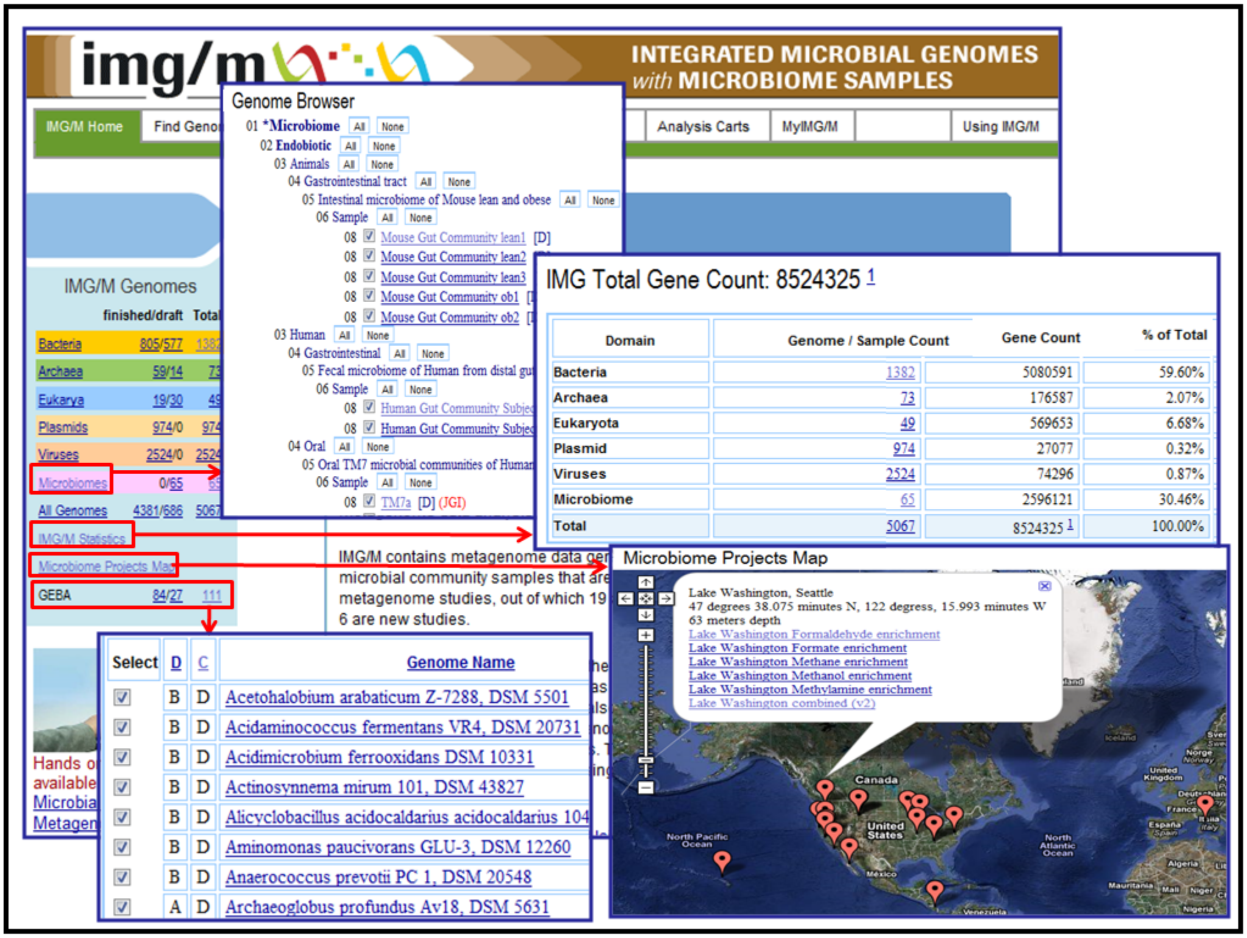
3. **Synthetic**
    a. two bioreactor wastewater biological phosphorus removing (EBPR) sludge samples;
    b. one bioreactor wastewater terephthalate-degrading community sample;
    c. three simulated microbial communities from 113 complete and draft genomes.

Similar to IMG 2.8 genomes, the functional characterization of genomes and metagenomes in IMG/M 2.8 has been extended with **KEGG Orthology (KO) terms** which serve as the main vehicle for associating IMG/M genomes and metagenomes with **KEGG pathways**. IMG/M genes are associated with KO terms as follows:

1. First, IMG/M 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/M to KEGG gene mapping was based on using NCBI's GI numbers and GeneIDs.
2. For IMG/M 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/M 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/M query gene and KEGG subject gene.

IMG/M Statistics

Various statistics are provided via the **IMG/M Statistics** link on the home page of IMG/M, as shown below, including **IMG/m Total Gene Count** which consists of counting all the genes (protein coding genes, RNA genes) except obsolete genes. Compared to the previous version of **IMG/M**, **IMG/M 2.8** contains a total of **8.5 million** genes (an increase of 3.4 million genes) consisting of about **5.9 million** reference genome genes (an increase of 2.3 million genes) and about **2.6 million** metagenome genes (an increase of 1.1 million genes).



Genome Browser

01 *Microbiome ☐ All ☐ None
  02 Endobiotic ☐ All ☐ None
    03 Animals ☐ All ☐ None
      04 Gastrointestinal tract ☐ All ☐ None
        05 Intestinal microbiome of Mouse lean and obese ☐ All ☐ None
          06 Sample ☐ All ☐ None
            08 ☑ Mouse Gut Community lean1 [D]
            08 ☑ Mouse Gut Community lean2
            08 ☑ Mouse Gut Community lean3
            08 ☑ Mouse Gut Community ob1
            08 ☑ Mouse Gut Community ob2
    03 Human ☐ All ☐ None
      04 Gastrointestinal ☐ All ☐ None
        05 Fecal microbiome of Human from distal gu
          06 Sample ☐ All ☐ None
            08 ☑ Human Gut Community Subjec
            08 ☑ Human Gut Community Subjec
      04 Oral ☐ All ☐ None
        05 Oral TM7 microbial communities of Human
          06 Sample ☐ All ☐ None
            08 ☑ TM7a [D] (JGI)

IMG/M Genomes

	finished/draft	Total
Bacteria	805/577	1382
Archaea	59/14	73
Eukarya	19/30	49
Plasmids	974/0	974
Viruses	2524/0	2524
Microbiomes	0/65	65
All Genomes	4381/686	5067
IMG/M Statistics		
Microbiome Projects Map		
GEBA	84/27	111

IMG Total Gene Count: 8524325 1

Domain	Genome / Sample Count	Gene Count	% of Total
Bacteria	1382	5080591	59.60%
Archaea	73	176587	2.07%
Eukaryota	49	569653	6.68%
Plasmid	974	27077	0.32%
Viruses	2524	74296	0.87%
Microbiome	65	2596121	30.46%
Total	5067	8524325 1	100.00%

IMG/M contains metagenome data ger
microbial community samples that are
metagenome studies, out of which 19
6 are new studies.

Microbiome Projects Map

Lake Washington, Seattle
47 degrees 38.075 minutes N, 122 degress, 15.993 minutes W
63 meters depth
Lake Washington Formaldehyde enrichment
Lake Washington Formate enrichment
Lake Washington Methane enrichment
Lake Washington Methanol enrichment
Lake Washington Methylamine enrichment
Lake Washington combined (v2)

Select	D	C	Genome Name
☑	B	D	Acetohalobium arabaticum Z-7288, DSM 5501
☑	B	D	Acidaminococcus fermentans VR4, DSM 20731
☑	B	D	Acidimicrobium ferrooxidans DSM 10331
☑	B	D	Actinosynnema mirum 101, DSM 43827
☑	B	D	Alicyclobacillus acidocaldarius acidocaldarius 104
☑	B	D	Aminomonas paucivorans GLU-3, DSM 12260
☑	B	D	Anaerococcus prevotii PC 1, DSM 20548
☑	A	D	Archaeoglobus profundus Av18, DSM 5631

The list of microbiome samples grouped by study or project is provided via the **Microbiomes** link. A **Map** link on the home page provides a Google Map showing the location of these samples, as shown BOVE. For each sample, a link to its **Microbiome Details** page is provided. Finally, a list of the **GEBA** genomes is also provided.

IMG/M 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 **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**. The **Genome Statistics** of **Microbiome Details** has been extended with counts of "Protein coding genes connected to KEGG Orthology (KO) terms".

  (ii) **Missing Enzymes** can be examined within the context of a **KEGG Map** or **Function Profile** result.

(b) **Extended features**

  (i) **Gene Details** has been extended to include 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.

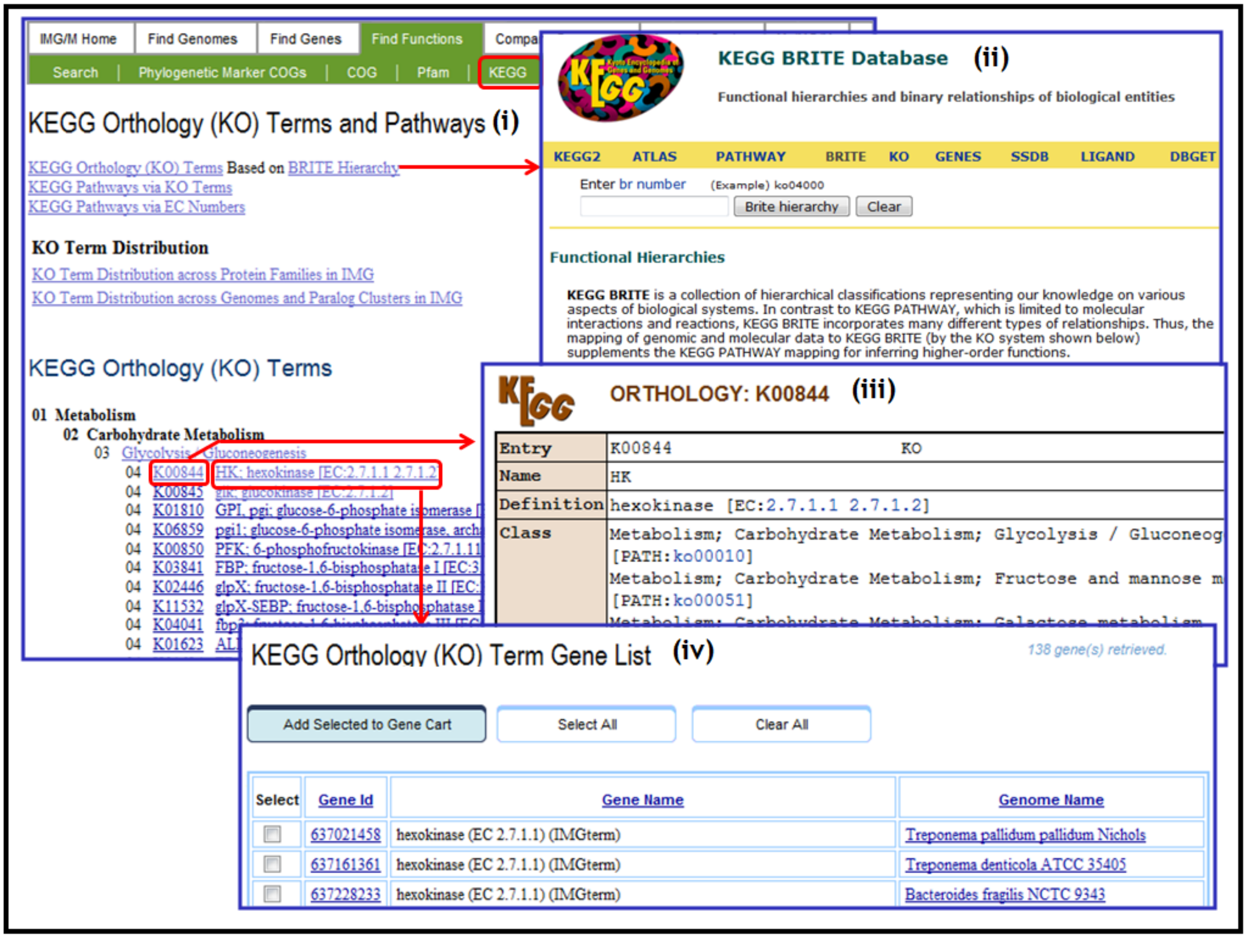
  (ii) **Phylogenetic Profile for Genes…**

New Features

Find Functions – *KEGG Orthology Terms & Pathways*

The **KEGG** collection of pathways in IMG/M has been reorganized and updated using the enhanced collection of KEGG resources[2], 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 1(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 1(ii)).



KEGG BRITE is a collection of hierarchical classifications representing our knowledge on various aspects of biological systems. In contrast to KEGG PATHWAY, which is limited to molecular interactions and reactions, KEGG BRITE incorporates many different types of relationships. Thus, the mapping of genomic and molecular data to KEGG BRITE (by the KO system shown below) supplements the KEGG PATHWAY mapping for inferring higher-order functions.

Entry	K00844 KO
Name	HK
Definition	hexokinase [EC:2.7.1.1 2.7.1.2]
Class	Metabolism; Carbohydrate Metabolism; Glycolysis / Gluconeog [PATH:ko00010] Metabolism; Carbohydrate Metabolism; Fructose and mannose m [PATH:ko00051]

Select	Gene Id	Gene Name	Genome Name
☐	637021458	hexokinase (EC 2.7.1.1) (IMGterm)	Treponema pallidum pallidum Nichols
☐	637161361	hexokinase (EC 2.7.1.1) (IMGterm)	Treponema denticola ATCC 35405
☐	637228233	hexokinase (EC 2.7.1.1) (IMGterm)	Bacteroides fragilis NCTC 9343

FIGURE 1. Find Functions – KEGG: KEGG Orthology Terms.

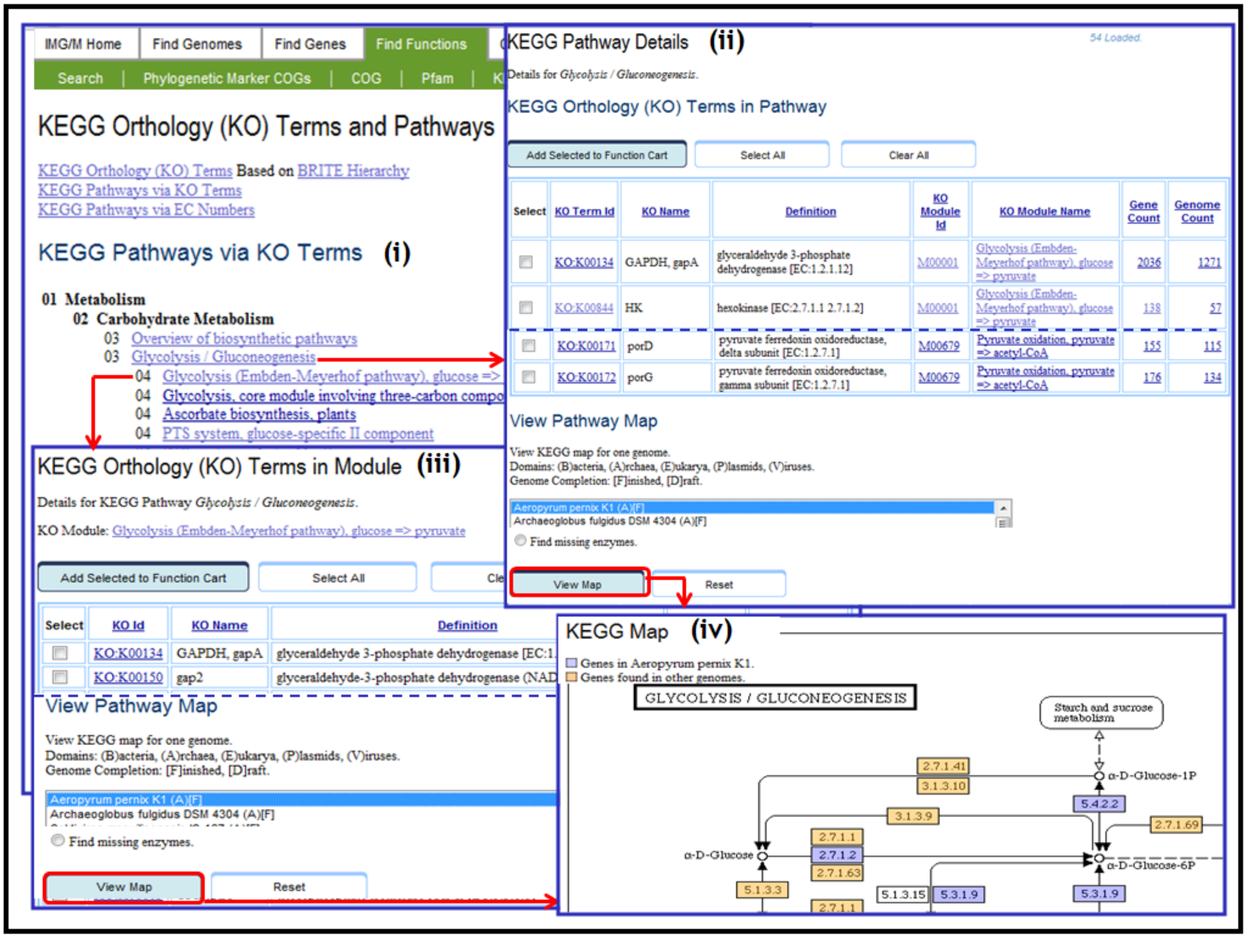
Each KO identifier (called K number) provides a link to the corresponding KEGG Orthology term specification, as illustrated in Figure 1(iii). The definition associated with a KO term provides a link to the list of IMG genes associated with that KO term, as illustrated in Figure 1(iv).

---

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

Two KEGG Pathway browsers are available. The **KEGG Pathways via KO Terms** browser, illustrated in Figure 2(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 2(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 2(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.



KEGG Orthology (KO) Terms and Pathways

KEGG Orthology (KO) Terms Based on BRITE Hierarchy
KEGG Pathways via KO Terms
KEGG Pathways via EC Numbers

KEGG Pathways via KO Terms

01 Metabolism
  02 Carbohydrate Metabolism
    03 Overview of biosynthetic pathways
    03 Glycolysis / Gluconeogenesis
      04 Glycolysis (Embden-Meyerhof pathway), glucose =>
      04 Glycolysis, core module involving three-carbon compo
      04 Ascorbate biosynthesis, plants
      04 PTS system, glucose-specific II component

KEGG Pathway Details

Details for *Glycolysis / Gluconeogenesis*.

KEGG Orthology (KO) Terms in Pathway

Select	KO Term Id	KO Name	Definition	KO Module Id	KO Module Name	Gene Count	Genome Count
☐	KO:K00134	GAPDH, gapA	glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12]	M00001	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	2036	1271
☐	KO:K00844	HK	hexokinase [EC:2.7.1.1 2.7.1.2]	M00001	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	138	57
☐	KO:K00171	porD	pyruvate ferredoxin oxidoreductase, delta subunit [EC:1.2.7.1]	M00679	Pyruvate oxidation, pyruvate => acetyl-CoA	155	115
☐	KO:K00172	porG	pyruvate ferredoxin oxidoreductase, gamma subunit [EC:1.2.7.1]	M00679	Pyruvate oxidation, pyruvate => acetyl-CoA	176	134

View Pathway Map

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

Aeropyrum pernix K1 (A)[F]
Archaeoglobus fulgidus DSM 4304 (A)[F]

○ Find missing enzymes.

KEGG Orthology (KO) Terms in Module

Details for KEGG Pathway *Glycolysis / Gluconeogenesis*.

KO Module: Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate

Select	KO Id	KO Name	Definition
☐	KO:K00134	GAPDH, gapA	glyceraldehyde 3-phosphate dehydrogenase [EC:1.
☐	KO:K00150	gap2	glyceraldehyde-3-phosphate dehydrogenase (NAD

View Pathway Map

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

Aeropyrum pernix K1 (A)[F]
Archaeoglobus fulgidus DSM 4304 (A)[F]

○ Find missing enzymes.

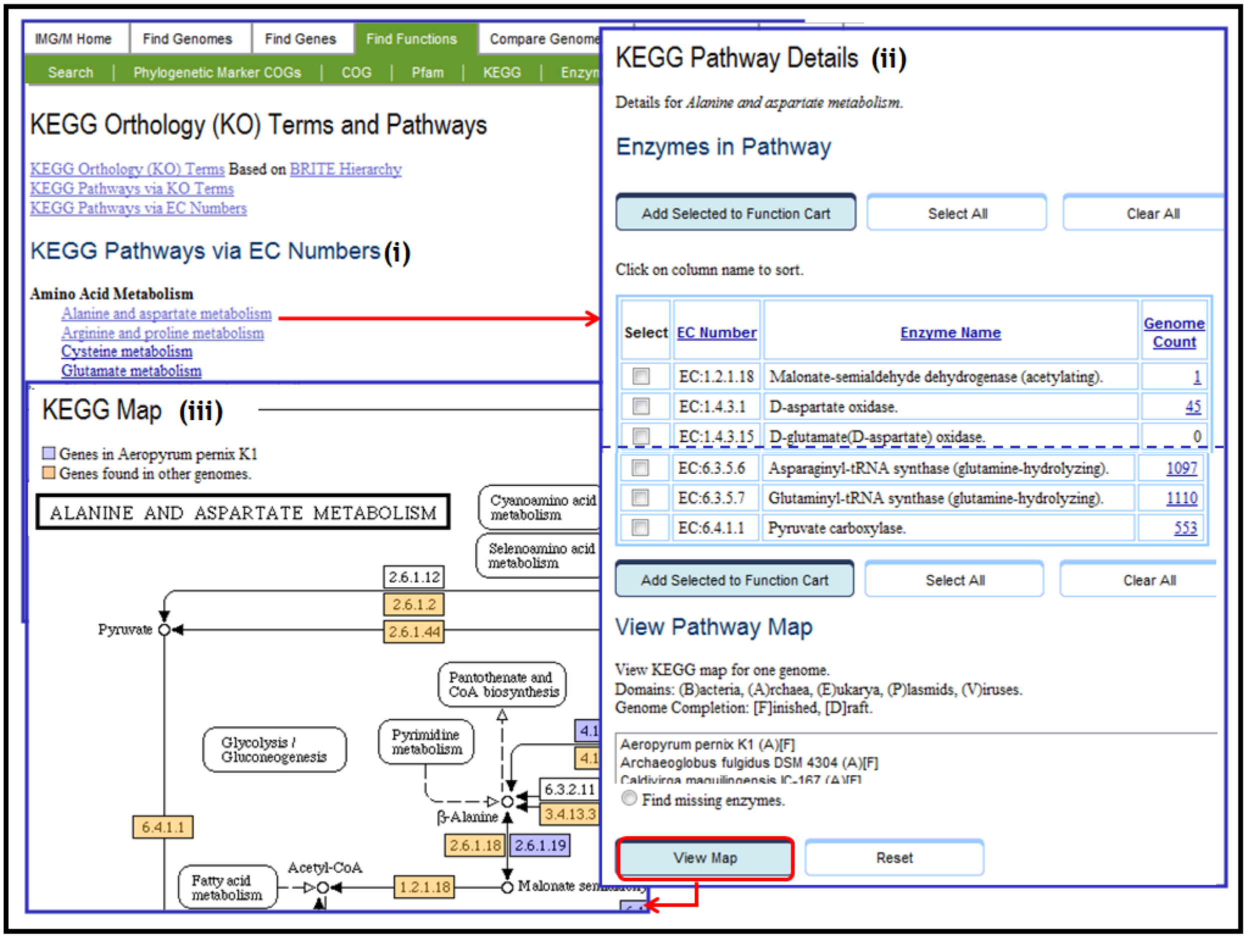
KEGG Map

▨ Genes in Aeropyrum pernix K1.
▨ Genes found in other genomes.

FIGURE 2. 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 2(iv).

The alternative **KEGG Pathways via EC Numbers** browser, illustrated in Figure 3(i), displays the KEGG pathways organized in pathways categories similar to those used in previous version of IMG/M. 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 3(ii).



KEGG Orthology (KO) Terms and Pathways

KEGG Orthology (KO) Terms Based on BRITE Hierarchy
KEGG Pathways via KO Terms
KEGG Pathways via EC Numbers

KEGG Pathways via EC Numbers (i)

Amino Acid Metabolism
Alanine and aspartate metabolism
Arginine and proline metabolism
Cysteine metabolism
Glutamate metabolism

☐ Genes in Aeropyrum pernix K1
☐ Genes found in other genomes.

KEGG Pathway Details (ii)

Details for *Alanine and aspartate metabolism*.

Enzymes in Pathway

Click on column name to sort.

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

View Pathway Map

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

Aeropyrum pernix K1 (A)[F]
Archaeoglobus fulgidus DSM 4304 (A)[F]
Caldivirga maquilingensis IC-167 (A)[F]

○ Find missing enzymes.

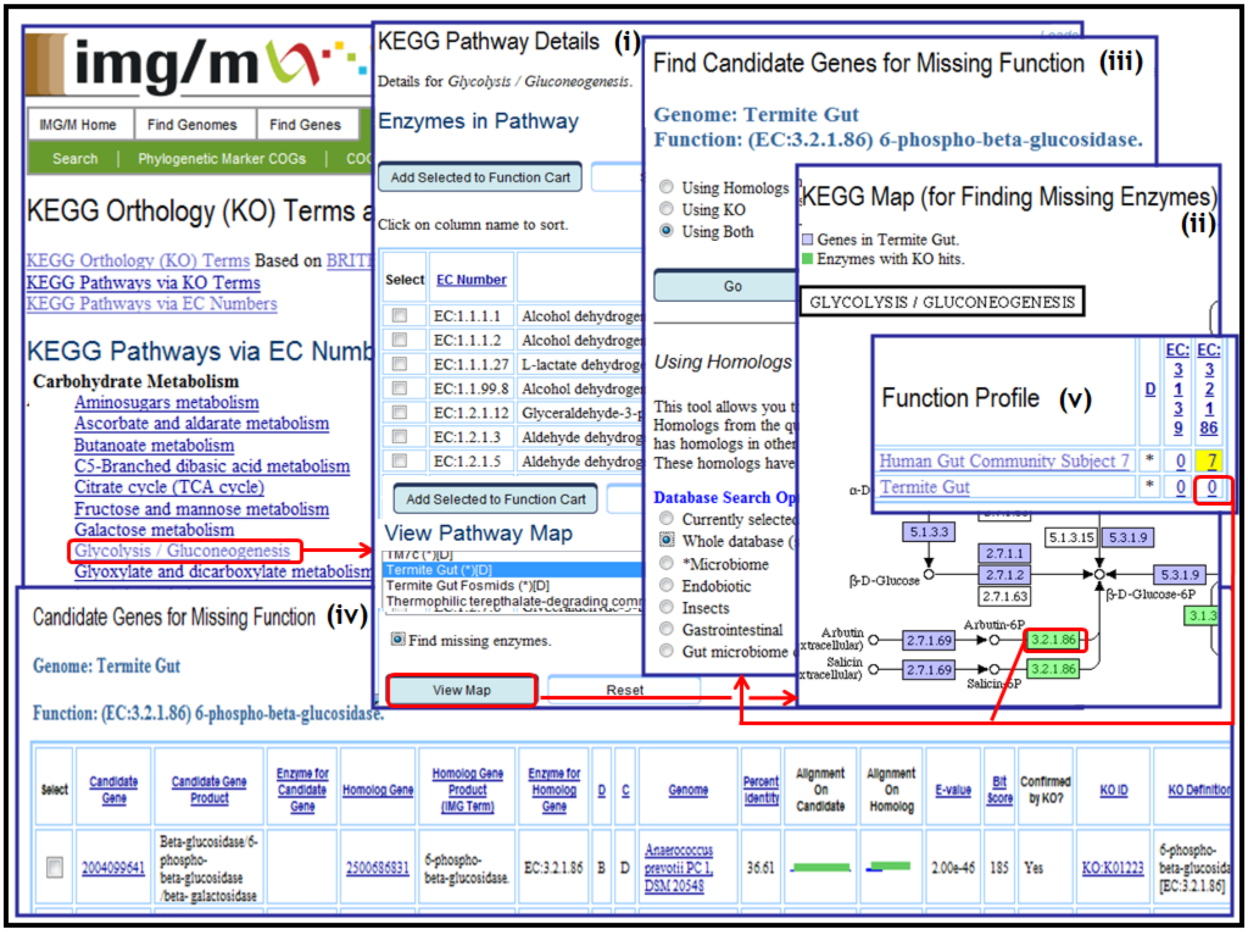
FIGURE 3. Find Functions – KEGG: KEGG Pathways via EC Numbers.

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.

Missing Enzymes – *KEGG Maps & Function Profile*

Genomes and meatgenomes 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 4.

Once a KEGG pathway is selected using the **KEGG Browser** under **Find Functions**, you can view its map for a selected genome or metagenome using the “Find missing enzymes” option, as illustrated in Figure 4(i). On the **KEGG Map**, such as that shown in Figure 4(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 4(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.



Select	Candidate Gene	Candidate Gene Product	Enzyme for Candidate Gene	Homolog Gene	Homolog Gene Product (IMG Term)	Enzyme for Homolog Gene	D	C	Genome	Percent Identity	Alignment On Candidate	Alignment On Homolog	E-value	Bit Score	Confirmed by KO?	KO ID	KO Definition
☐	2004099641	Beta-glucosidase/6-phospho-beta-glucosidase /beta- galactosidase		2500686831	6-phospho-beta-glucosidase.	EC:3.2.1.86	B	D	Anaerococcus prevotii PC 1, DSM 20548	36.61			2.00e-46	185	Yes	KO:K01223	6-phospho-beta-glucosida [EC:3.2.1.86]

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

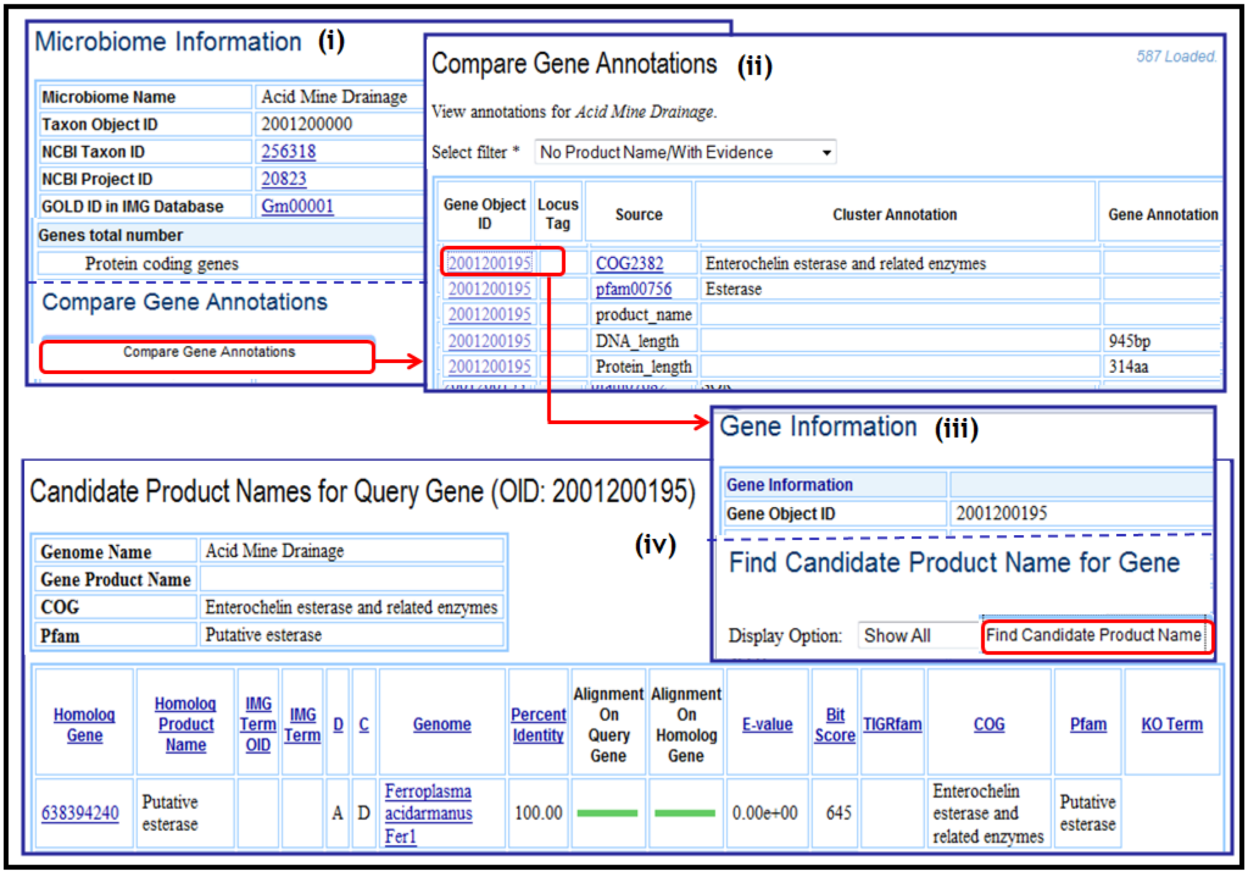
You can find candidate genes of your target genome that could be associated with a missing enzyme by searching for genes that have **homologs/orthologs** associated with the missing enzyme, as illustrated in Figure 4(iii). You can search across all the genomes available in the system, across a subset 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 4(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 4(v), will also lead to a **Find Candidate Genes for Missing Function** page.

Extended Features

Gene Details – Find Product Name

The **Metagenome Statistics** part of the **Microbiome Details** page has been extended to include the number of genes associated with KEGG Orthology (KO) terms. The list of genes provided by the **Compare Gene Annotations** tool available on the **Microbiome Details** page, as illustrated in Figure 5(i), include KO terms when available. For genes without a protein product name, such as that shown in Figure 5(ii), the **Find Candidate Product Name** tool available on the **Gene Page**, as illustrated in Figure 5(iii), provides a list of candidate protein product names from related (sequence similarity based) genes, as illustrated in Figure 5(iv).



Microbiome Information (i)

Microbiome Name	Acid Mine Drainage
Taxon Object ID	2001200000
NCBI Taxon ID	256318
NCBI Project ID	20823
GOLD ID in IMG Database	Gm00001
Genes total number	
Protein coding genes	

Compare Gene Annotations

Compare Gene Annotations (ii)

587 Loaded.

View annotations for *Acid Mine Drainage*.

Select filter * No Product Name/With Evidence

Gene Object ID	Locus Tag	Source	Cluster Annotation	Gene Annotation
2001200195		COG2382	Enterochelin esterase and related enzymes	
2001200195		pfam00756	Esterase	
2001200195		product_name		
2001200195		DNA_length		945bp
2001200195		Protein_length		314aa

Gene Information (iii)

Gene Information	
Gene Object ID	2001200195

Find Candidate Product Name for Gene

Display Option: Show All

Candidate Product Names for Query Gene (OID: 2001200195)

(iv)

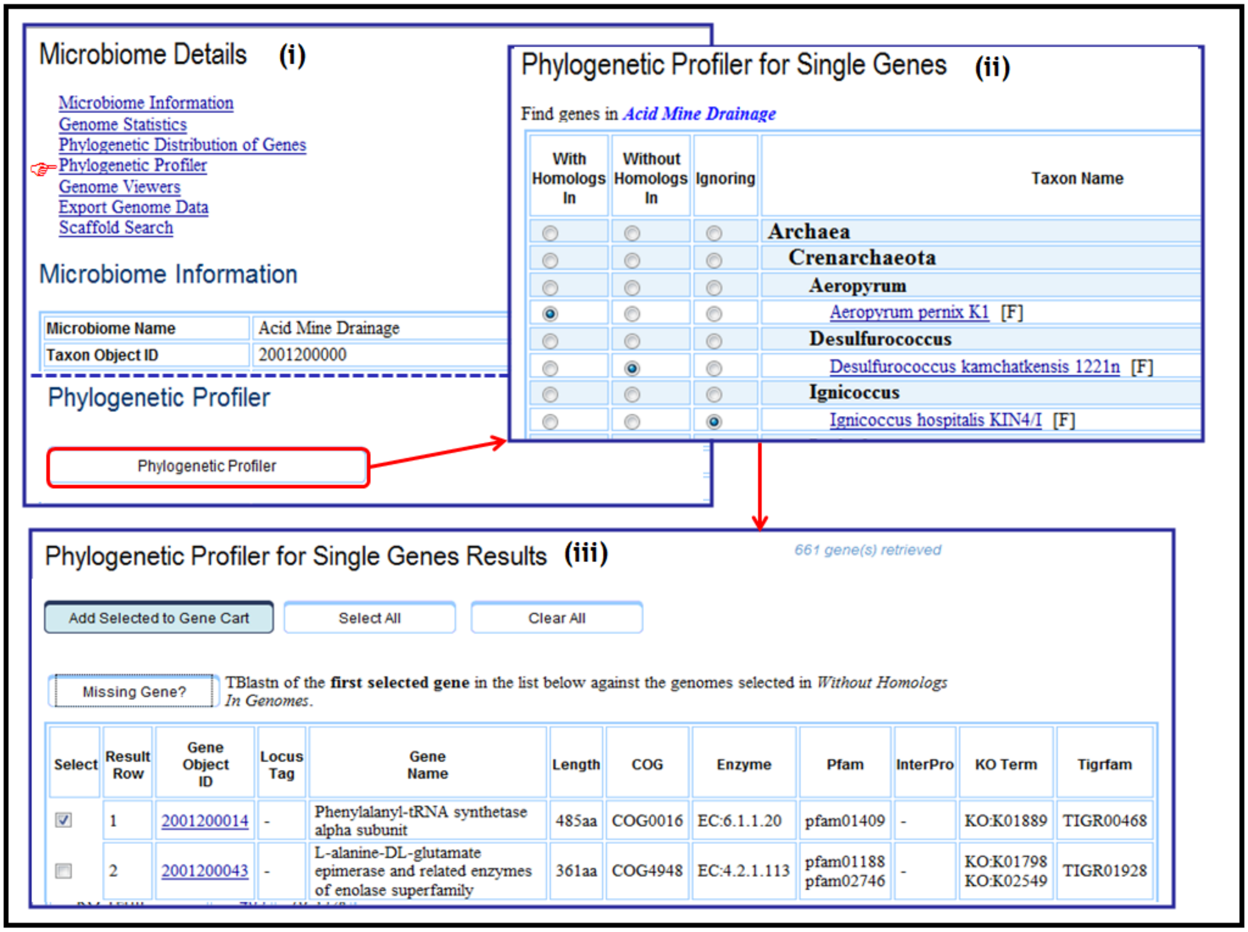
Genome Name	Acid Mine Drainage
Gene Product Name	
COG	Enterochelin esterase and related enzymes
Pfam	Putative esterase

Homolog Gene	Homolog Product Name	IMG Term OID	IMG Term	D	C	Genome	Percent Identity	Alignment On Query Gene	Alignment On Homolog Gene	E-value	Bit Score	TIGRfam	COG	Pfam	KO Term
638394240	Putative esterase			A	D	Ferroplasma acidarmanus Fer1	100.00	▬	▬	0.00e+00	645		Enterochelin esterase and related enzymes	Putative esterase	

FIGURE 5. Gene Details - Find Candidate Product Name.

Phylogenetic Profiler - Metagenomes

The **Phylogenetic Profiler** available under the **Find Genes** main menu tab, is limited only to isolate genomes in IMG/M. For metagenomes, a restricted version of the **Phylogenetic Profiler** is available on individual **Microbiome Details** pages, as illustrated in Figure 6(i).



Microbiome Details (i)

- Microbiome Information
- Genome Statistics
- Phylogenetic Distribution of Genes
- Phylogenetic Profiler
- Genome Viewers
- Export Genome Data
- Scaffold Search

Microbiome Information

Microbiome Name	Acid Mine Drainage
Taxon Object ID	2001200000

Phylogenetic Profiler

Phylogenetic Profiler for Single Genes (ii)

Find genes in *Acid Mine Drainage*

With Homologs In	Without Homologs In	Ignoring	Taxon Name
○	○	○	**Archaea**
○	○	○	**Crenarchaeota**
○	○	○	**Aeropyrum**
●	○	○	Aeropyrum pernix K1 [F]
○	○	○	**Desulfurococcus**
○	●	○	Desulfurococcus kamchatkensis 1221n [F]
○	○	○	**Ignicoccus**
○	○	●	Ignicoccus hospitalis KIN4/I [F]

Phylogenetic Profiler for Single Genes Results (iii)

661 gene(s) retrieved

TBlastn of the **first selected gene** in the list below against the genomes selected in *Without Homologs In Genomes*.

Select	Result Row	Gene Object ID	Locus Tag	Gene Name	Length	COG	Enzyme	Pfam	InterPro	KO Term	Tigrfam
☑	1	2001200014	-	Phenylalanyl-tRNA synthetase alpha subunit	485aa	COG0016	EC:6.1.1.20	pfam01409	-	KO:K01889	TIGR00468
☐	2	2001200043	-	L-alanine-DL-glutamate epimerase and related enzymes of enolase superfamily	361aa	COG4948	EC:4.2.1.113	pfam01188 pfam02746	-	KO:K01798 KO:K02549	TIGR01928

FIGURE 6. Phylogenetic Profiler for Metagenomes.

The **Phylogenetic Profiler** for a metagenome dataset such as *Acid Mine Drainage*, allows finding metagenome genes with/without homologs in one or several reference isolate genomes, as illustrated in Figure 6(ii).