

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](#)

IMG 2.6: What's New

IMG 2.6 is the **14th** release of the Integrated Microbial Genomes (IMG) genomic data management and analysis system. **IMG 2.6** was released on **August 11th, 2008**.

IMG 2.6 Content

Genomes

The content of **IMG 2.6** has been updated with new microbial genomes available in **RefSeq version 29** (May 9, 2008).

IMG 2.6 contains a total of **4,207** genomes consisting of **1,078** bacterial, **56** archaeal, **40** eukaryotic genomes, **2,230** viruses (including bacterial phages), and **803** plasmids that did not come from a specific microbial genome sequencing project. Among these genomes, **3,737** are finished genomes, and **470** are draft genomes; **308** are **JGI** sequenced genomes: **237** finished and **71** draft genomes. JGI genomes are also available through individual microbial portals at <http://genome.jgi-psf.org/microbial>.

Note that **16** microbial genomes from **IMG 2.5** were **replaced** in **IMG 2.6** because (1) a "Draft" genome has been replaced by its "Finished" version or (2) the composition of the genome has changed through the addition of new replicons, that is, plasmids or chromosomes. For replaced genomes, whenever possible, the gene object identifiers (gene OIDs) for the protein-coding genes (CDS) were mapped to their new version in IMG. 2.6. See IMG [Data Evolution History](#) for details.

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

tRNA and **rRNA** genes (23S, 16S and 5S) missing from the original RefSeq genome files are added using tRNAscan-SE v1.23 for tRNA genes and similarity comparisons to existing RNA genes. In IMG 2.6 **928** tRNA and **202** rRNA genes were added in **167** genomes.

In terms of total number of genes, **IMG 2.6** contains **4,694, 263 genes**, an increase of 957,615 genes compared to IMG 2.5.

IMG Statistics

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

The screenshot displays the IMG (Integrated Microbial Genomes) website interface. The main header shows the IMG logo and the text "INTEGRATED MICROBIAL GENOMES". A navigation bar includes "IMG Home" and "Find Genomes". A sidebar on the left lists "IMG Genomes" with a table of finished/draft counts for Bacteria, Archaea, Eukarya, Plasmids, and Viruses, along with "All Genomes" and a highlighted "IMG Statistics" link. The main content area is divided into several sections:

- IMG Total Gene Count: 4694263¹**: A table showing domain breakdown.

Domain	Genome Count	Gene Count	% of Total
Bacteria	1078	3979688	84.78%
Archaea	56	127781	2.72%
Eukaryota	40	501349	10.68%
Plasmid	803	21105	0.45%
Viruses	2230	64340	1.37%
Total	4207	4694263¹	100.00%
- Function Statistics**: A table showing the distribution of genes across functional categories.

	Total Count	Genes with	% of Total
COG	3623	3000793	63.92%
Pfam	9318	3120571	66.48%
Enzyme	4985	534138	11.38%
TIGRfam	2946	1167397	24.87%
IMG Term	3379	733123	15.62%
GO Term	26095	644465	13.73%
GO-Molecular Function	8834	585929	12.48%
GO-Cellular Component	2182	283996	6.05%
Protein Product Name	446321	4512279	96.12%
- IMG Cluster Statistics**: A table showing the distribution of genes across cluster types.

	Total Count	Genes with	% of Total
COG	4873	3000793	63.92%
Pfam	9318	3120571	66.48%
TIGRfam	2946	1167397	24.87%
IMG Chromosomal Cassettes	791307	3789797	80.73%
Conserved IMG Chromosomal Cassettes by			
COG Clusters	606945	3119612	66.46%
Pfam Clusters	1731444	3284382	69.97%
- Pathway Statistics**: A table showing the distribution of genes across pathway categories.

	Total Count	Genes with	% of Total
COG Pathway	77	3000793	63.92%
Kegg Pathway	213	450791	9.60%
TIGRfam Roles	100	1040696	22.17%
IMG Pathway	539	291692	6.21%
IMG Parts List	49	289606	6.17%
GO-Biological Process	15064	488474	10.41%

IMG 2.6 User Interface

The User Interface has been extended in order to improve its overall usability. The main extensions include: (1) **graphical viewers**, such as for **protein family coverage** of single or multiple genomes; (2) an **Abundance Profile Overview** tool that extends the Abundance Profile Viewer; (3) new tools for examining and searching **gene cassettes**.

Graphical Viewers

Organism Details – Genome Statistics

The **Genome Statistics** part of **Organism Details** has been reorganized in order to improve clarity. Graphical viewers have been added for displaying the distribution of genes associated with COG, Pfam, TIGRfam, and KEGG, as illustrated below.

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 1(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 1(iii). Genes can be then selected and saved in the **Gene Cart** for further analysis.

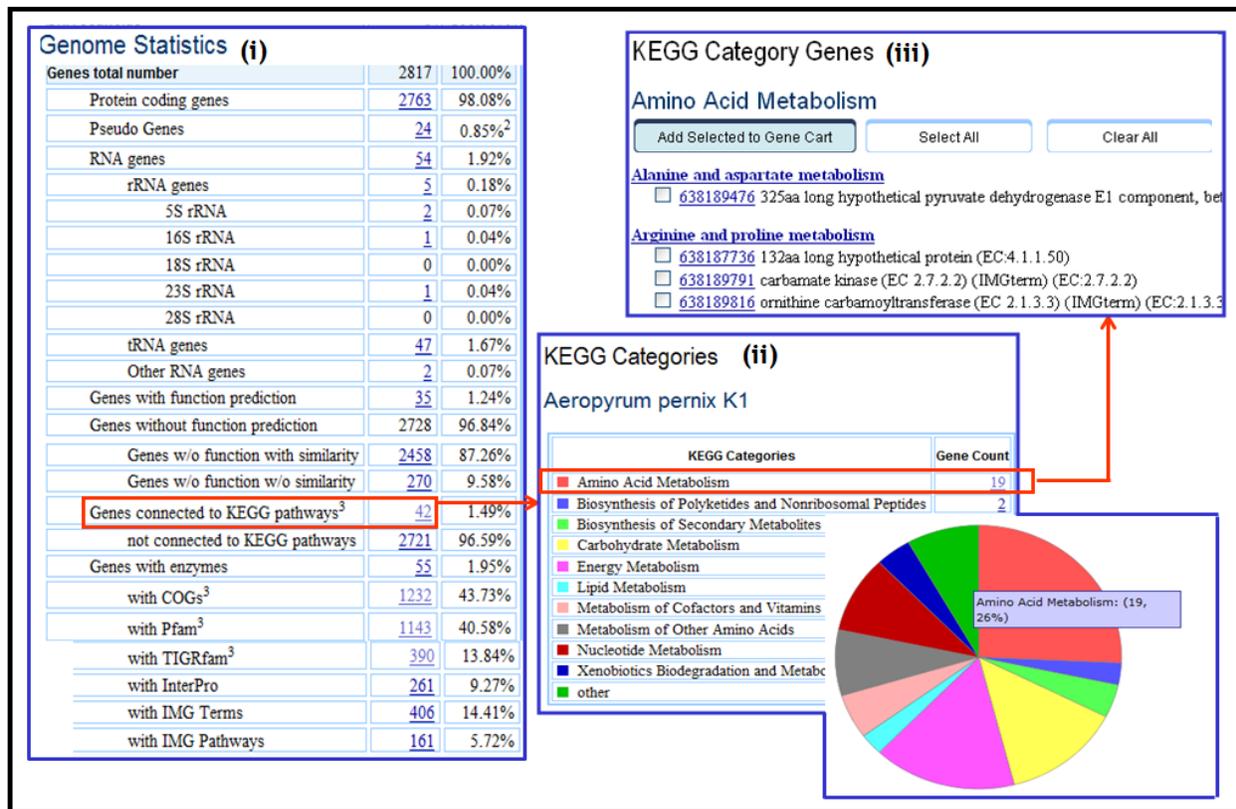


FIGURE 1. Organism Details - Genome Statistics.

Gene Details – Evidence for Function Prediction

The **Evidence for Function Prediction** part of **Gene Details** includes a new **Sequence Viewer** and an additional version of the **Chromosome Viewer**, as illustrated in Figure 3(i).

For a specific gene, the **Sequence Viewer** displays the six frame translation with putative ORF's, potential start codons, and potential Shine-Delgano regions. The gene neighborhood, minimum ORF size and type of display (graphic or text) can be selected, as illustrated in Figure 3(ii). The text display provides the protein sequences for the ORFs while the graphical display, illustrated in Figure 3(iii), includes a GC plot. The additional **Chromosome Viewer** displays the neighborhood of the target gene with genes colored to reflect deviation of characteristic GC percentage for that genome, as illustrated in Figure 3(iv).

Compare Genomes – Genome Statistics

Graphical viewers have been added for displaying the results of **COG and KEGG Category Statistics**, as illustrated below for COG Category Statistics.

Selecting **Statistics for Genomes by specific COG Category** (see Figure 4(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 4(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 4(ii).

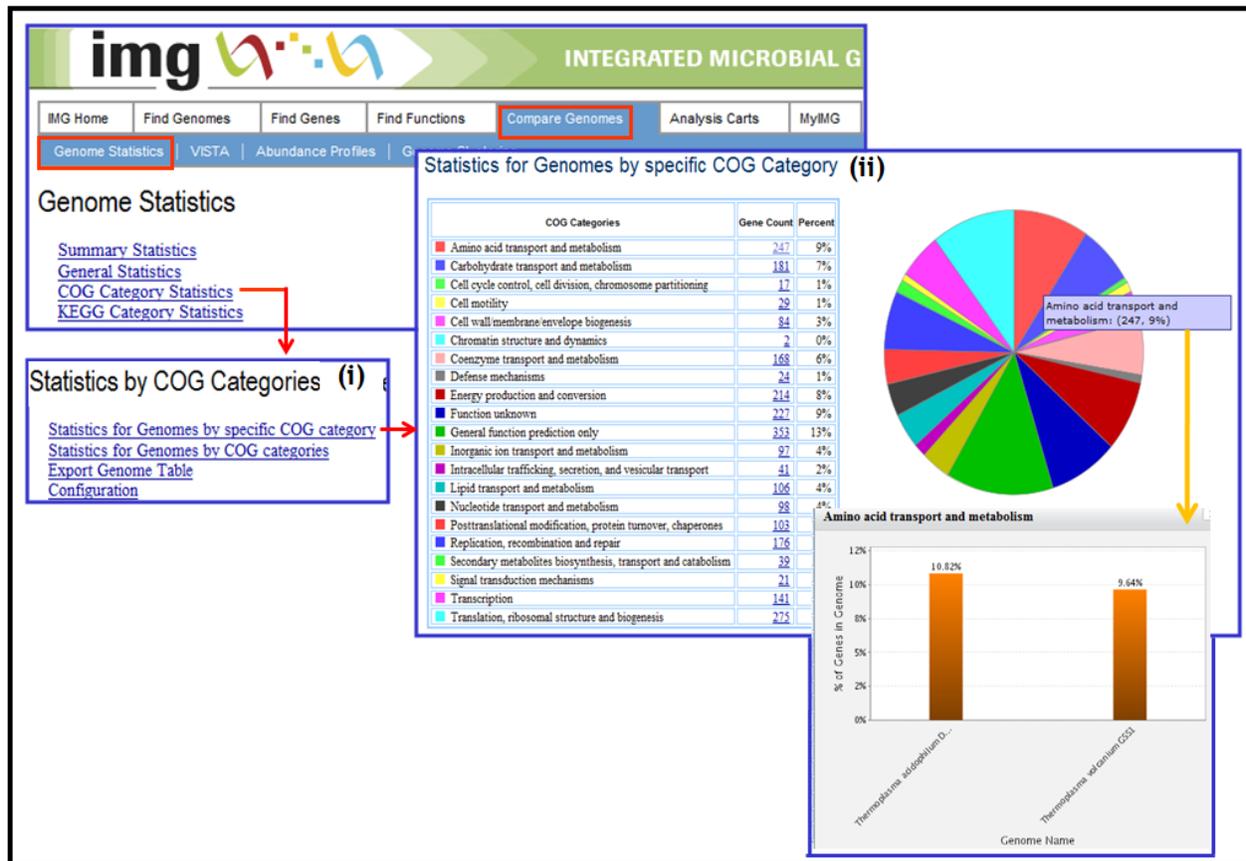


FIGURE 4. Genome Statistics: COG Category statistics.

Abundance Profile Overview

The **Abundance Profile Overview** illustrated in Figure 5 extends the **Abundance Profile Viewer**. 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. For “Heat Map” output, the abundance of protein/functional families is displayed as a heat map with red corresponding to the most abundant families. Each column on the map corresponds to a genome, and each row corresponds to a family. Click on the cell in order to retrieve the list of genes assigned to this particular family in a genome. Click on the identifier of the family displayed on the right of the column in order to include the corresponding family into the **Function Cart**.

Abundance Profile Overview Results

1 - *Aeropyrum pernix* K1
2 - *Archaeoglobus fulgidus* DSM 4304

COG ID	Aeropyrum pernix K1	Archaeoglobus fulgidus DSM 4304
COG1131	2	1
COG0642	1	1
COG0589	1	1
COG2294	1	1
COG1969	1	1
COG1361	1	1
COG0491	1	1
COG5550	1	1
COG3466	1	1
COG0410	1	1
COG0395	1	1
COG0378	1	1
COG0347	1	1
COG0318	1	1
COG0144	1	1
COG0095	1	1
COG0076	1	1
COG1456	1	1
COG1444	1	1
COG1418	1	1
COG1392	1	1
COG1387	1	1
COG1378	1	1
COG1359	1	1
COG1341	1	1
COG1271	1	1
COG0145	1	1
COG0128	1	1
COG0123	1	1
COG0119	1	1
COG0086	1	1
COG0071	1	1
COG0066	1	1
COG0045	1	1
COG0053	1	1

Abundance Profile Overview Results

Pages: [1] 2 3 [Next Page]
Download tab-delimited file for Excel

Add Selected to Function Cart Select All Clear All

Select	Row No.	ID	Name	Aer per K11	Arc ful 434
<input type="checkbox"/>	1	COG0001	Glutamate-1-semialdehyde aminotransferase	2	1
<input type="checkbox"/>	2	COG0002	Acetylglutamate semialdehyde dehydrogenase	1	1
<input type="checkbox"/>	3	COG0003	Oxyanion-translocating ATPase	2	0

FIGURE 5. Abundance Profile Overview: Results with Heat Map and Matrix Output Format.

If the “Matrix” output is selected, the abundance of protein/functional families is displayed in a tabular format, 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. Families of interest can be selected for inclusion into the **Function Cart**. The results in “Matrix” format can be exported to a tab-delimited Excel file.

Gene Cassettes

Definitions

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 allowed to participate in the formation of 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) and Pfam assignments 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.

Gene Details - Gene Information

The **Gene Information** section of **Gene Details** is divided into four parts that contain information on gene, protein, associated pathways, and IMG clusters. The fourth part of the **Gene Information** section contains data on IMG clusters, as shown in Figure 6(ii).

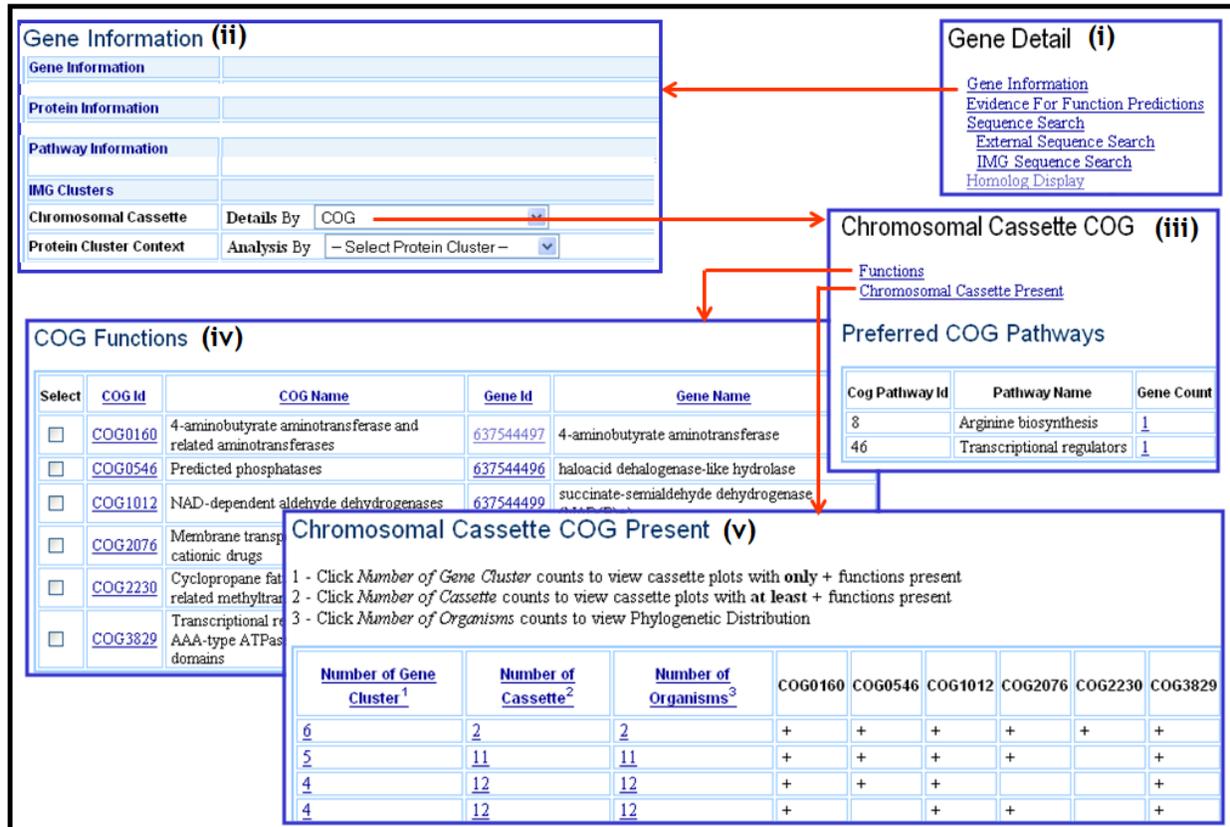


FIGURE 6. Gene Details: IMG Chromosomal Cassette Details.

Details on **chromosomal cassettes** can be displayed with genes labeled by their COG or Pfam association, as illustrated in Figure 6(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 6(iv). This page also provides information on other cassettes that share at least two protein clusters with the query cassette, as illustrated in Figure 6(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.

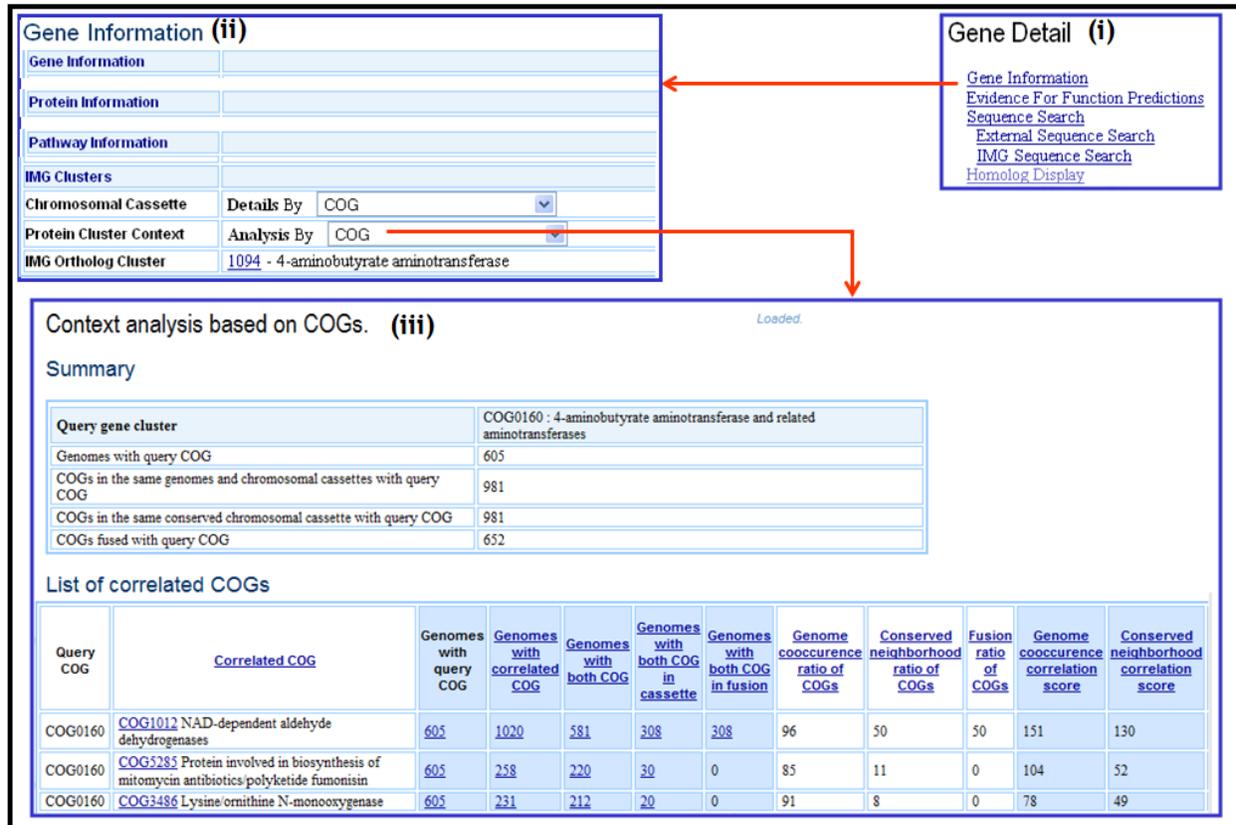


FIGURE 7. Gene Details: Protein Cluster Context Analysis.

Protein Cluster Context Analysis allows examining the functional correlations of the current gene based on its COG or Pfam association, as illustrated in Figure 7. **Context Analysis** starts with the so called “query” protein cluster (COG, Pfam) associated with the current gene, as illustrated in Figure 7(ii).

For a query protein cluster **A**, the **Context Analysis** page contains a **summary** as illustrated in Figure 7(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**:

$$NG1/\min(NGA, NGB), \text{ where}$$

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:
 $NG1/\min(NGA, NGB)$, 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**:
 $\{ [P(A) \times P(B) \times \log(P(A,B))/(P(A) \times P(B))] + [Q(A) \times Q(B) \times \log(Q(A,B))/(Q(A) \times Q(B))] \} \times \maxDist(A, B)$, 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**, respectively;
 - **Q(A) = 1-P(A)**; **Q(B) = 1-P(B)**; **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**, respectively;
- the **conserved neighborhood correlation score** for **A** and **B**:
 $(NG1/ NG2) \times \maxDist(A,B)$, 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**, respectively;
- the **fusion correlation score** for **A** and **B**:
 $(NG1/ NG2) \times \maxDist(A,B)$, 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**, respectively.

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.

Gene Details - Evidence for Function Prediction

The **Evidence for Function Prediction** section of **Gene Details** 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 8.

Chromosomal cassettes can be examined using the **Chromosomal Cassette Viewer**, as illustrated in Figure 8(iv). Chromosome cassettes can be viewed with genes labeled by their

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.

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 6. 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.

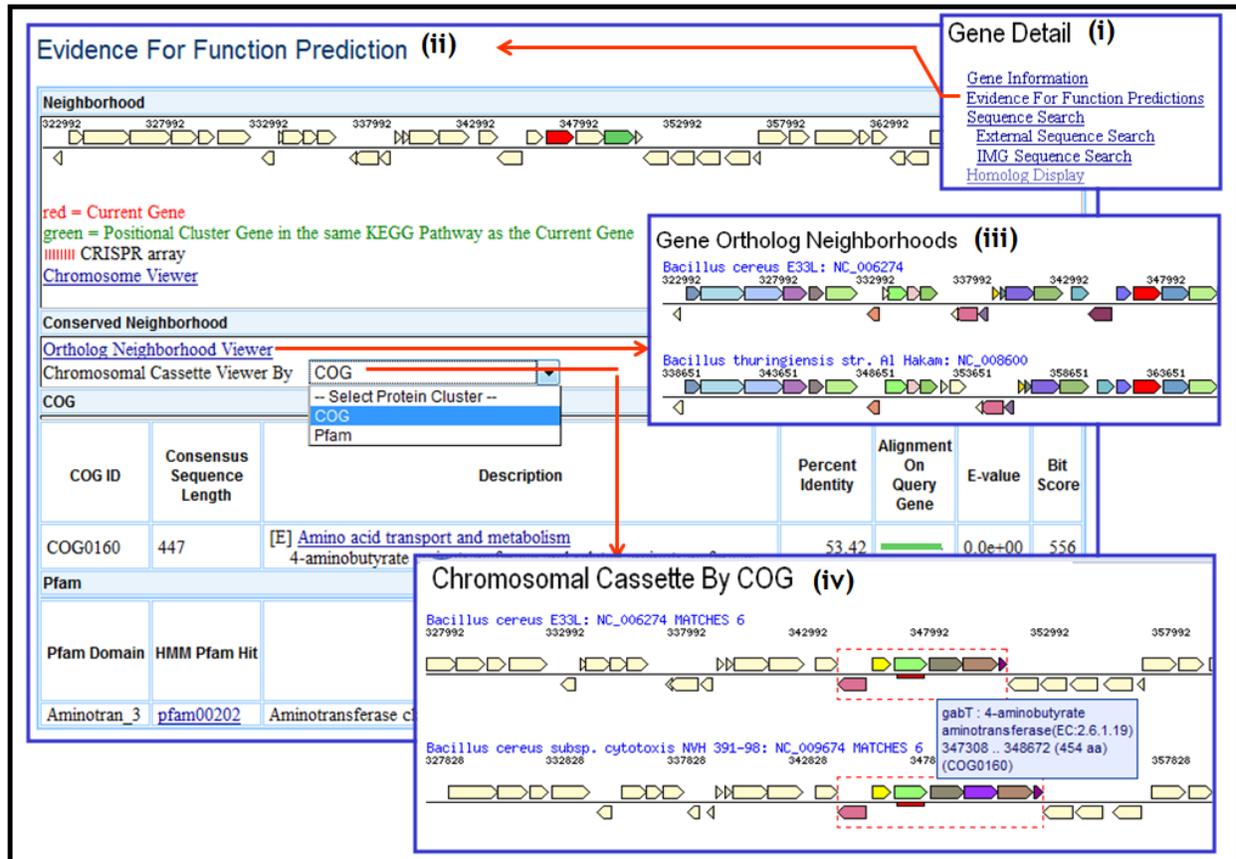


FIGURE 8. Gene Details: Evidence for Function Prediction.

Phylogenetic Profiler for Gene Cassettes

In addition to the **Phylogenetic Profiler** tool that allows selecting **single genes** based on presence or absence of homologs in other genomes, The **Phylogenetic Profiler for Gene Cassettes** (see Figure 9) 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 9(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 9(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.

Phylogenetic Profilers (i)

Tool	Description
Single Genes	Find genes in genome (bin) of interest qualified by similarity to set of 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)
By COG Conserved Cassettes

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

Chromosomal Cassette By COG (iv)

Thermoplasma acidophilum DSM 1728: AL139299 MATCHES 32

Thermoplasma volcanium GSS1 DNA: BA000011 MATCHES 29

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

FIGURE 9. Gene Search: Phylogenetic Profiler for Gene Cassettes.

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 9(iii). Note that:

- i. 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;
- ii. 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 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.