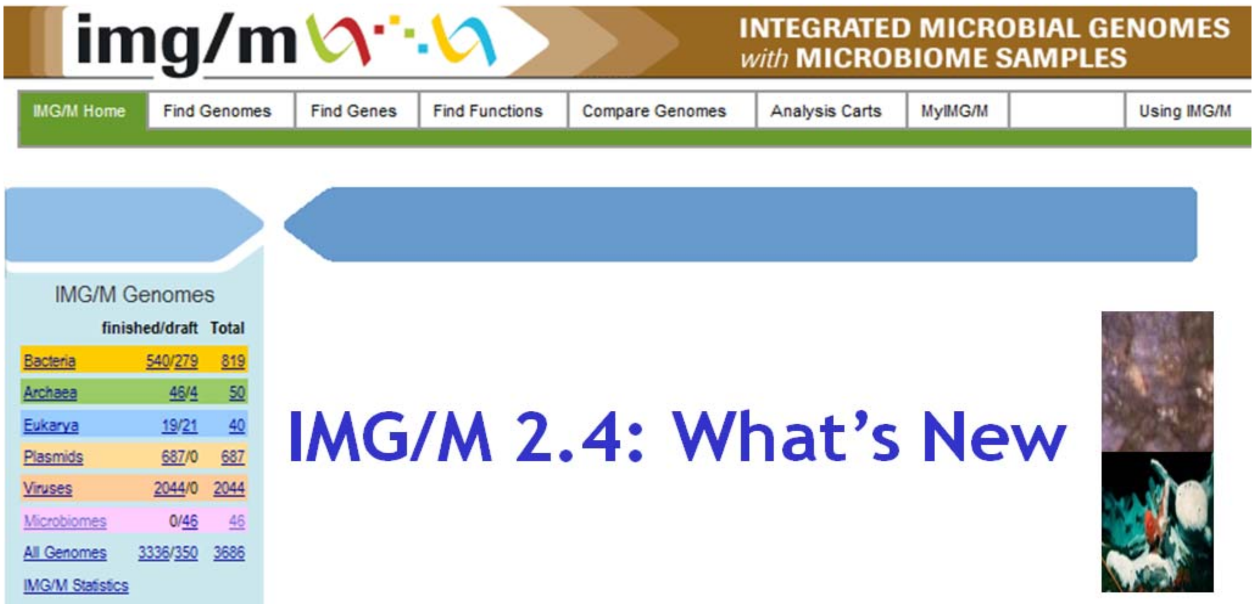


IMG/M Genomes

	finished/draft	Total
Bacteria	540/279	819
Archaea	46/4	50
Eukarya	19/21	40
Plasmids	687/0	687
Viruses	2044/0	2044
Microbiomes	0/46	46
All Genomes	3336/350	3686

IMG/M Statistics

IMG/M 2.4: What’s New

IMG 2.4.1 released on **September 1st, 2008** is an updated version of IMG/M 2.4 which was released in February 2008.

IMG/M 2.4.1 Content

IMG/M 2.4 (released in Feb 2008) contains **reference genomes** from **IMG 2.4** (released in Dec 2007) integrated with **metagenome** datasets generated from samples for the following metagenome projects:

- A methane-oxidizing archaeal community sample
- two biological phosphorus removing (EBPR) sludge samples;
- three isolated deep sea "whale fall" carcasses;
- an agricultural soil sample;
- an acid mine drainage (AMD) biofilm sample;
- two human distal gut samples;
- four gutless marine worm sample;
- five obese and lean mouse gut samples;
- three single cell TM7 samples;
- two termite hindgut samples;
- a uranium contaminated groundwater samples.

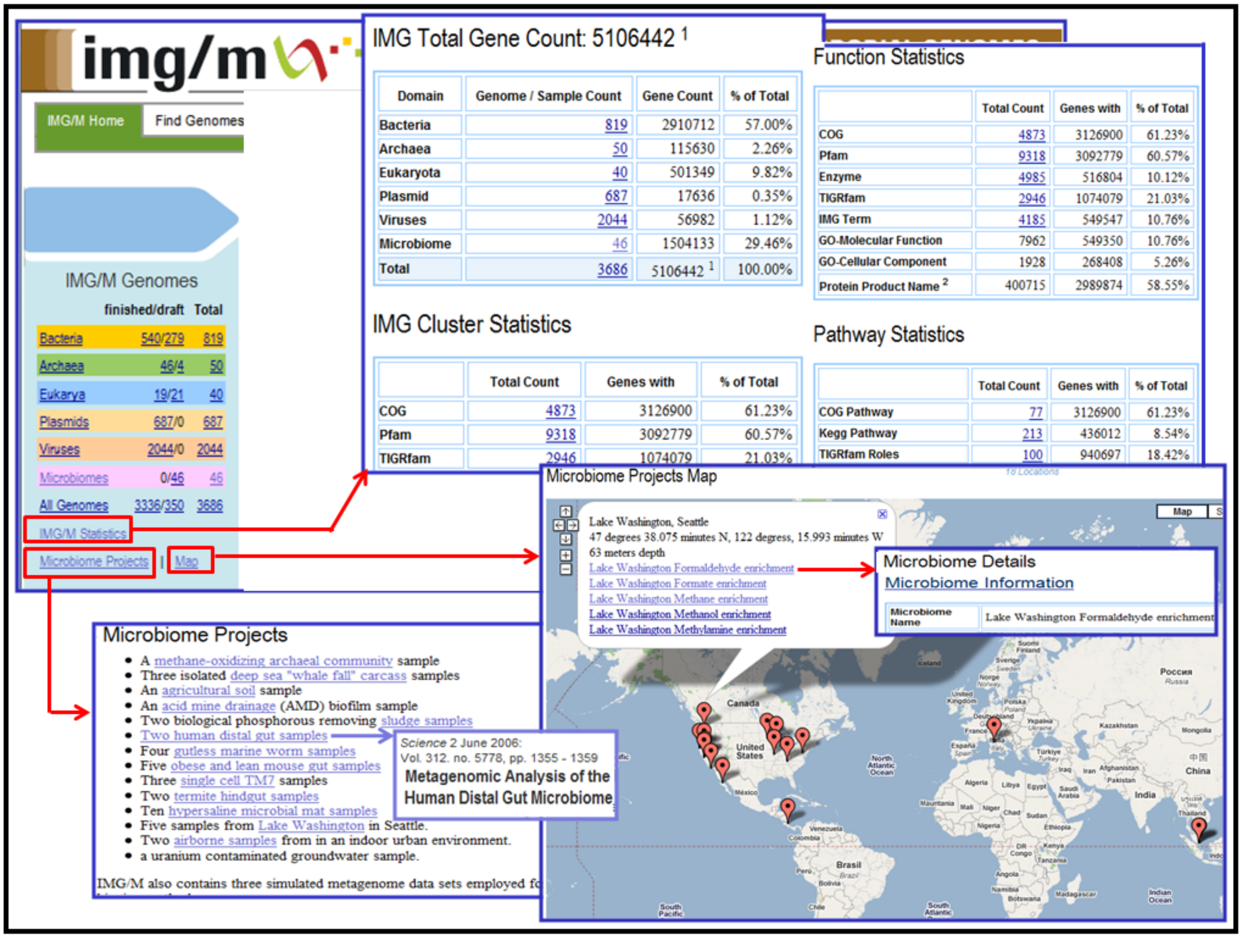
IMG/M 2.4.1, released in Sep 2008, has been updated with datasets generated from the following microbial community samples:

- ten hypersaline microbial mat samples[1],
- five samples from Lake Washington in Seattle[2], and
- two airborne samples from in an indoor urban environment[3].

[1] http://www.nature.com/msb/journal/v4/n1/pdf/msb200835.pdf
[2] http://www.nature.com/nbt/journal/vaop/ncurrent/abs/nbt.1488.html
[3] http://www.plosone.org/article/info:doi/10.1371/journal.pone.0001862

IMG/M Statistics

Various statistics are provided via the **IMG/M Statistics** link on the home page of IMG/M. The list of microbiome samples grouped by study or project is provided via the **Microbiome Samples** link, with a link to the associated publication also provided for each project. A **Map** link on the home page provides a Google Map showing the location of these samples, as shown below. For each sample, a link to its **Microbiome Details** page is provided.



IMG/M Genomes

	finished/draft	Total
Bacteria	540/279	819
Archaea	46/4	50
Eukarya	19/21	40
Plasmids	687/0	687
Viruses	2044/0	2044
Microbiomes	0/46	46
All Genomes	3336/350	3686

IMG/M Statistics

Microbiome Projects | Map

IMG Total Gene Count: 5106442 [1]

Domain	Genome / Sample Count	Gene Count	% of Total
Bacteria	819	2910712	57.00%
Archaea	50	115630	2.26%
Eukaryota	40	501349	9.82%
Plasmid	687	17636	0.35%
Viruses	2044	56982	1.12%
Microbiome	46	1504133	29.46%
Total	3686	5106442 [1]	100.00%

IMG Cluster Statistics

	Total Count	Genes with	% of Total
COG	4873	3126900	61.23%
Pfam	9318	3092779	60.57%
TIGRfam	2946	1074079	21.03%

Function Statistics

	Total Count	Genes with	% of Total
COG	4873	3126900	61.23%
Pfam	9318	3092779	60.57%
Enzyme	4985	516804	10.12%
TIGRfam	2946	1074079	21.03%
IMG Term	4185	549547	10.76%
GO-Molecular Function	7962	549350	10.76%
GO-Cellular Component	1928	268408	5.26%
Protein Product Name [2]	400715	2989874	58.55%

Pathway Statistics

	Total Count	Genes with	% of Total
COG Pathway	77	3126900	61.23%
Kegg Pathway	213	436012	8.54%
TIGRfam Roles	100	940697	18.42%

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

Microbiome Details

Microbiome Information

Microbiome Name	Lake Washington Formaldehyde enrichment

Microbiome Projects

- A methane-oxidizing archaeal community sample
- Three isolated deep sea "whale fall" carcass samples
- An agricultural soil sample
- An acid mine drainage (AMD) biofilm sample
- Two biological phosphorous removing sludge samples
- Two human distal gut samples
- Four gutless marine worm samples
- Five obese and lean mouse gut samples
- Three single cell TM7 samples
- Two termite hindgut samples
- Ten hypersaline microbial mat samples
- Five samples from Lake Washington in Seattle.
- Two airborne samples from in an indoor urban environment.
- a uranium contaminated groundwater sample.

IMG/M also contains three simulated metagenome data sets employed fo

Science 2 June 2006:
Vol. 312. no. 5778, pp. 1355 - 1359
Metagenomic Analysis of the Human Distal Gut Microbiome

IMG/M User Interface

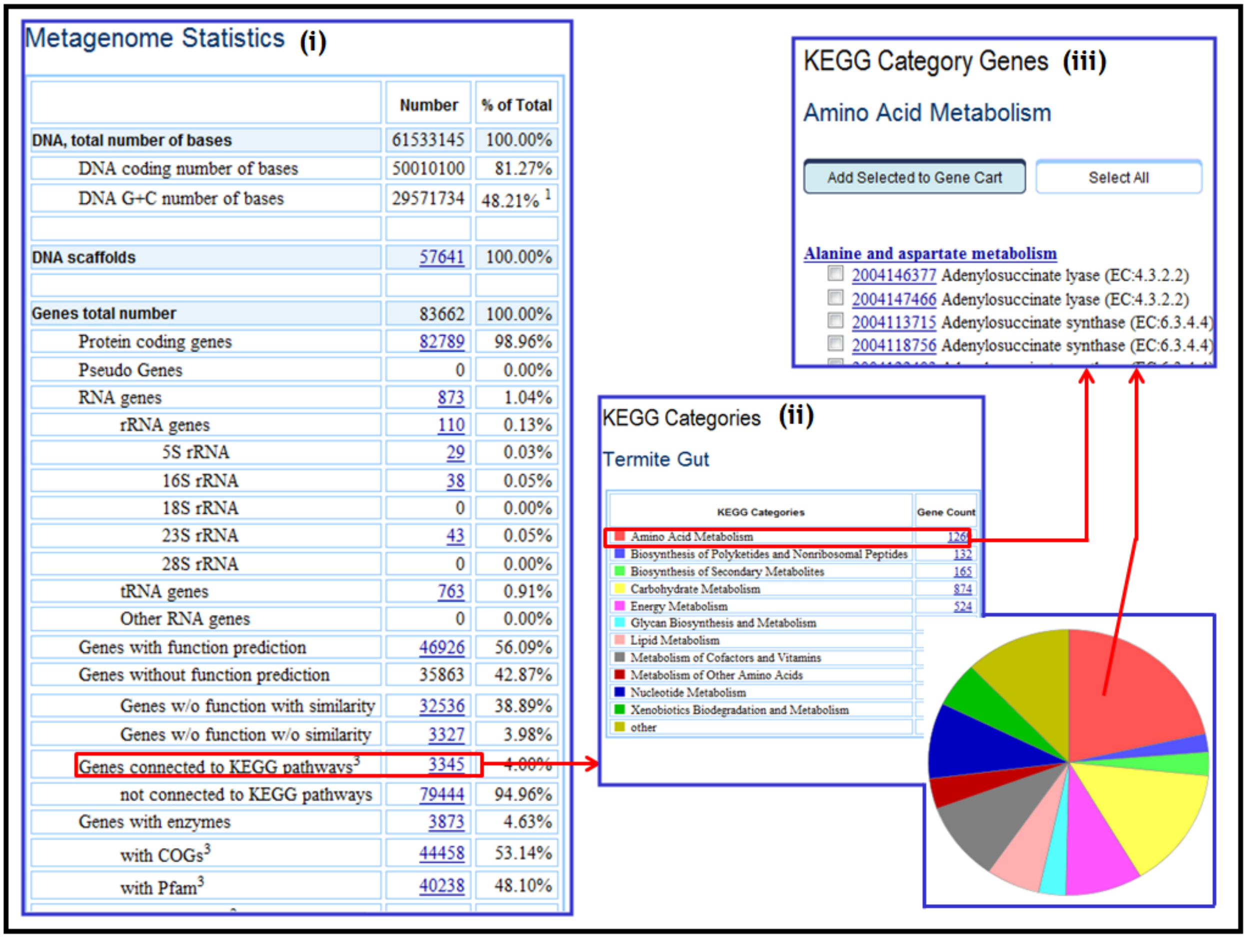
The User Interface has been extended in order to improve its overall usability. The main extensions include **graphical viewers,** such as for **protein family coverage** of single or multiple genomes and metagenomes, and an **Abundance Profile Overview** tool that extends the Abundance Profile Viewer.

Graphical Viewers

Microbiome Details – *Microbiome Information, Metagenome Statistics*

The **Microbiome Information** part of **Microbiome Details** contains additional metadata collected from scientists as well as a map display of the sample collection location using Google Map.

The **Metagenome Statistics** part of **Microbiome 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.



Metagenome Statistics **(i)**

	Number	% of Total
DNA, total number of bases	61533145	100.00%
DNA coding number of bases	50010100	81.27%
DNA G+C number of bases	29571734	48.21% [1]
DNA scaffolds	57641	100.00%
Genes total number	83662	100.00%
Protein coding genes	82789	98.96%
Pseudo Genes	0	0.00%
RNA genes	873	1.04%
rRNA genes	110	0.13%
5S rRNA	29	0.03%
16S rRNA	38	0.05%
18S rRNA	0	0.00%
23S rRNA	43	0.05%
28S rRNA	0	0.00%
tRNA genes	763	0.91%
Other RNA genes	0	0.00%
Genes with function prediction	46926	56.09%
Genes without function prediction	35863	42.87%
Genes w/o function with similarity	32536	38.89%
Genes w/o function w/o similarity	3327	3.98%
Genes connected to KEGG pathways[3]	3345	4.00%
not connected to KEGG pathways	79444	94.96%
Genes with enzymes	3873	4.63%
with COGs[3]	44458	53.14%
with Pfam[3]	40238	48.10%

KEGG Categories **(ii)**

Termite Gut

KEGG Categories	Gene Count
■ Amino Acid Metabolism	126
■ Biosynthesis of Polyketides and Nonribosomal Peptides	132
■ Biosynthesis of Secondary Metabolites	165
■ Carbohydrate Metabolism	874
■ Energy Metabolism	524
■ Glycan Biosynthesis and Metabolism	
■ Lipid Metabolism	
■ Metabolism of Cofactors and Vitamins	
■ Metabolism of Other Amino Acids	
■ Nucleotide Metabolism	
■ Xenobiotics Biodegradation and Metabolism	
■ other	

KEGG Category Genes **(iii)**

Amino Acid Metabolism

Alanine and aspartate metabolism

- □ 2004146377 Adenylosuccinate lyase (EC:4.3.2.2)
- □ 2004147466 Adenylosuccinate lyase (EC:4.3.2.2)
- □ 2004113715 Adenylosuccinate synthase (EC:6.3.4.4)
- □ 2004118756 Adenylosuccinate synthase (EC:6.3.4.4)

FIGURE 1. Organism Details - Genome Statistics.

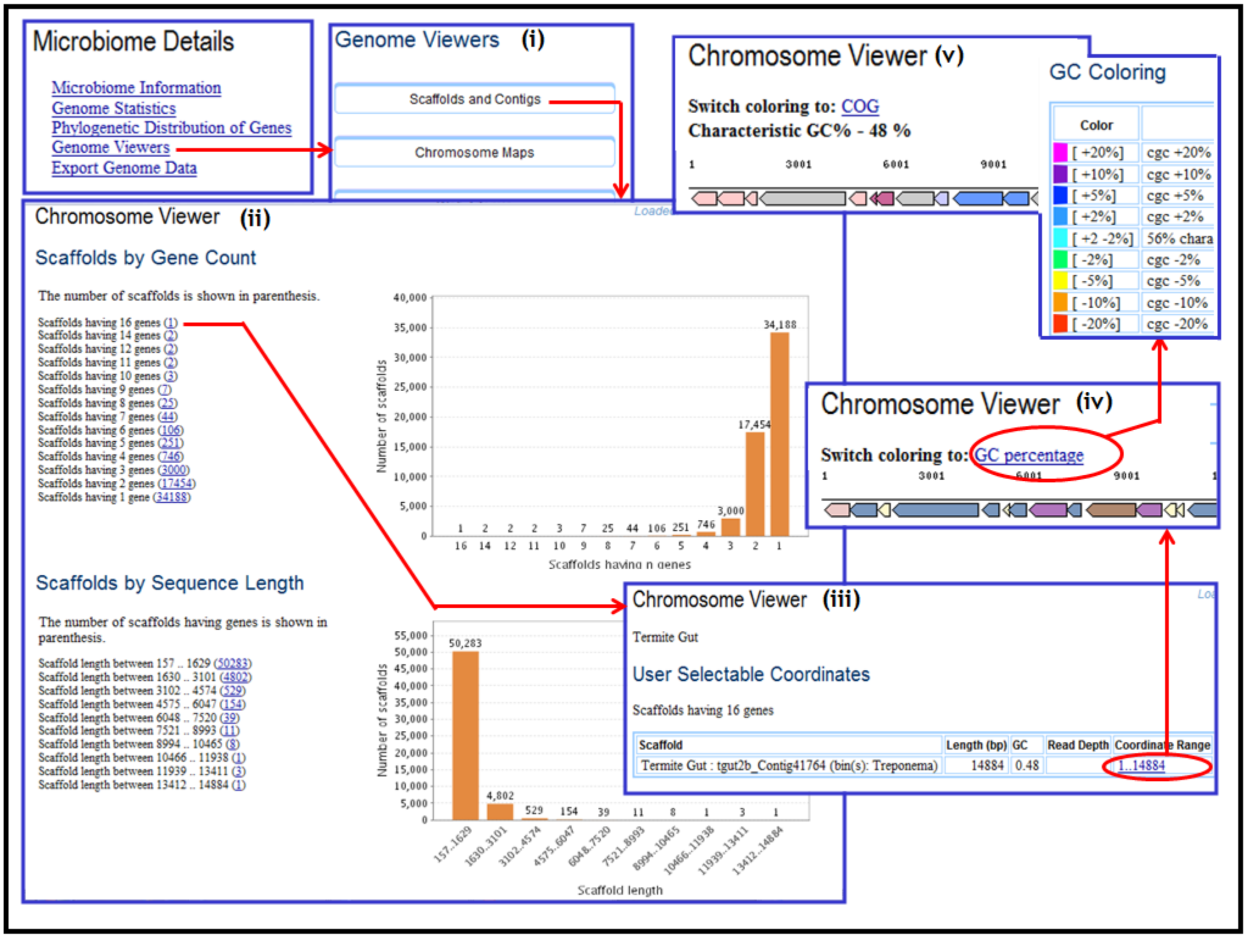
Gene counts in some categories in **Metagenome Statistics**, such as "Genes with COGs", "Genes with Pfam", "Genes with TIGRfam", and "Genes connected to KEGG pathways" are linked to tables that show these genes classified according to the corresponding functional hierarchies (e.g., COG Functional Categories, 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 the 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.

Microbiome Details – *Genome Viewers*

The **Genome Viewers** part of **Microbiome Details** includes a new display for scaffolds and an additional version of the **Chromosome Viewer**, as illustrated in Figure 2. Scaffolds are displayed in two lists sorted by gene count and sequence length, respectively. Each list is also displayed graphically using bar charts, as illustrated in Figure 2(ii).

After a scaffold and coordinate range is selected, as illustrated in Figure 2(iii), the **Chromosome Viewer** displays genes colored by their associated COG as illustrated in Figure 2(iv). Gene coloring can be then switched to reflect deviation of characteristic GC percentage for that genome, as illustrated in Figure 2(v).



The number of scaffolds is shown in parenthesis.

Scaffolds having 16 genes (1)
Scaffolds having 14 genes (2)
Scaffolds having 12 genes (2)
Scaffolds having 11 genes (2)
Scaffolds having 10 genes (3)
Scaffolds having 9 genes (7)
Scaffolds having 8 genes (25)
Scaffolds having 7 genes (44)
Scaffolds having 6 genes (106)
Scaffolds having 5 genes (251)
Scaffolds having 4 genes (746)
Scaffolds having 3 genes (3000)
Scaffolds having 2 genes (17454)
Scaffolds having 1 gene (34188)

The number of scaffolds having genes is shown in parenthesis.

Scaffold length between 157 .. 1629 (50283)
Scaffold length between 1630 .. 3101 (4802)
Scaffold length between 3102 .. 4574 (529)
Scaffold length between 4575 .. 6047 (154)
Scaffold length between 6048 .. 7520 (39)
Scaffold length between 7521 .. 8993 (11)
Scaffold length between 8994 .. 10465 (8)
Scaffold length between 10466 .. 11938 (1)
Scaffold length between 11939 .. 13411 (3)
Scaffold length between 13412 .. 14884 (1)

Scaffolds having 16 genes

Scaffold	Length (bp)	GC	Read Depth	Coordinate Range
Termite Gut : tgut2b_Contig41764 (bin(s): Treponema)	14884	0.48		1..14884

Switch coloring to: GC percentage

Switch coloring to: COG
Characteristic GC% - 48 %

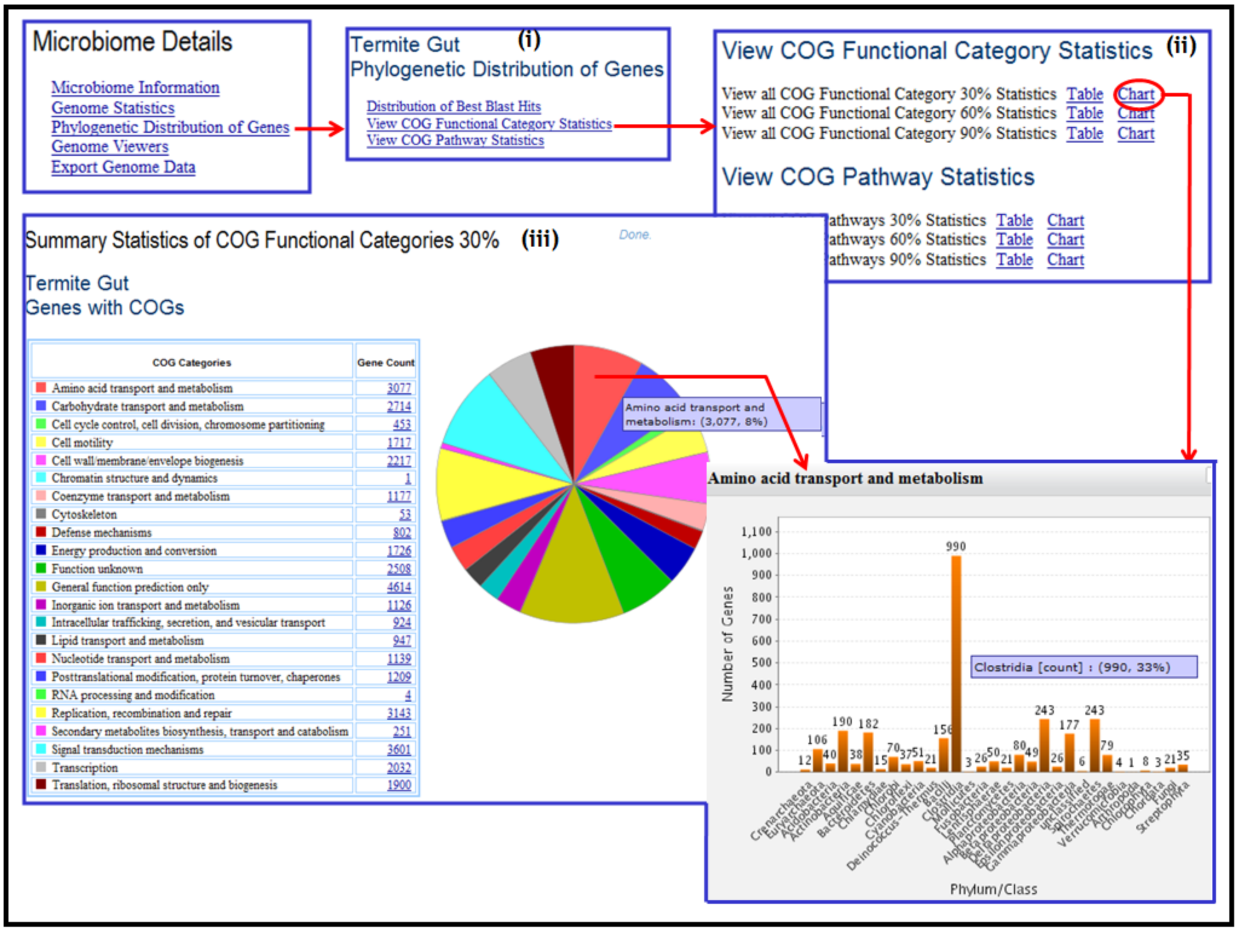
Color	
🟪 [+20%]	cgc +20%
🟪 [+10%]	cgc +10%
🟦 [+5%]	cgc +5%
🟦 [+2%]	cgc +2%
🟦 [+2 -2%]	56% chara
🟩 [-2%]	cgc -2%
🟨 [-5%]	cgc -5%
🟧 [-10%]	cgc -10%
🟥 [-20%]	cgc -20%

FIGURE 2. Microbiome Details - Genome Viewers.

Microbiome Details – *Phylogenetic Distribution of Genes*

The **Phylogenetic Distribuiton of Genes** part of **Microbiome Details** has been extended with graphical displays for viewing **COG Functional Category Statistics** and **COG Pathway Statistics,** as illustrated in Figure 3, in addition to the existing tabular displays.

Selecting the “Chart” option for viewing one of the **COG Functional Category Statistics** or **COG Pathway Statistics** (see Figure 3(ii)), will display in a tabular and ***pie chart*** format the count of genes across all phyla/classes associated with each COG category or pathway, as illustrated in Figure 3(iii). Clicking on a COG category or pathway on the pie chart or on the colored coded square for a COG category or pathway in the table, will display a ***bar chart*** with the number of genes for each phylum/class associated with that COG category or pathway, as illustrated on the lower side pane of Figure 3(iii).



Microbiome Details

Microbiome Information
Genome Statistics
Phylogenetic Distribution of Genes
Genome Viewers
Export Genome Data

Termite Gut (i)
Phylogenetic Distribution of Genes

Distribution of Best Blast Hits
View COG Functional Category Statistics
View COG Pathway Statistics

View COG Functional Category Statistics (ii)

View all COG Functional Category 30% Statistics Table Chart
View all COG Functional Category 60% Statistics Table Chart
View all COG Functional Category 90% Statistics Table Chart

View COG Pathway Statistics

athways 30% Statistics Table Chart
athways 60% Statistics Table Chart
athways 90% Statistics Table Chart

Summary Statistics of COG Functional Categories 30% (iii)

Done.

Termite Gut
Genes with COGs

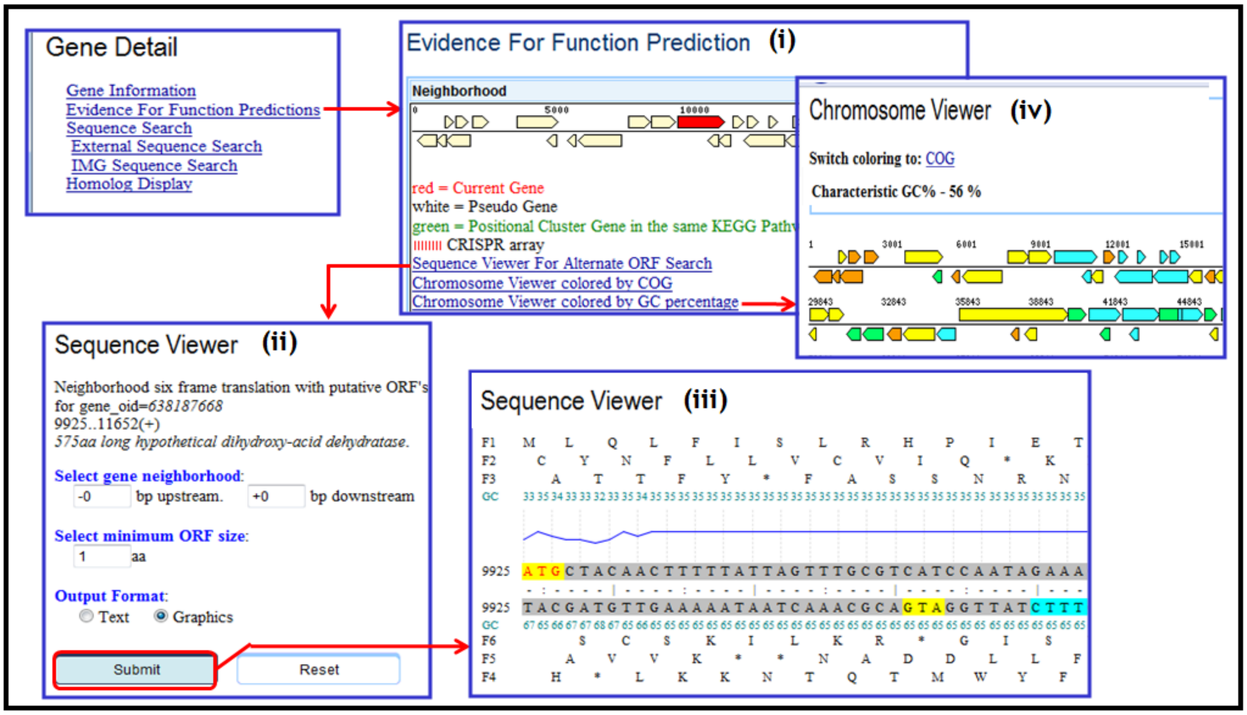
COG Categories	Gene Count
🟥 Amino acid transport and metabolism	3077
🟦 Carbohydrate transport and metabolism	2714
🟩 Cell cycle control, cell division, chromosome partitioning	453
🟨 Cell motility	1717
🟪 Cell wall/membrane/envelope biogenesis	2217
🟦 Chromatin structure and dynamics	1
🟥 Coenzyme transport and metabolism	1177
⬛ Cytoskeleton	53
🟥 Defense mechanisms	802
🟦 Energy production and conversion	1726
🟩 Function unknown	2508
🟨 General function prediction only	4614
🟪 Inorganic ion transport and metabolism	1126
🟦 Intracellular trafficking, secretion, and vesicular transport	924
⬛ Lipid transport and metabolism	947
🟥 Nucleotide transport and metabolism	1139
🟦 Posttranslational modification, protein turnover, chaperones	1209
🟩 RNA processing and modification	4
🟨 Replication, recombination and repair	3143
🟪 Secondary metabolites biosynthesis, transport and catabolism	251
🟦 Signal transduction mechanisms	3601
⬜ Transcription	2032
🟥 Translation, ribosomal structure and biogenesis	1900

FIGURE 3. Microbiome Details – Phylogenetic Distribution of Genes.

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 4(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 4(ii).The text display provides the protein sequences for the ORFs while the graphical display, illustrated in Figure 4(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 4(iv).



Gene Detail

Gene Information
Evidence For Function Predictions
Sequence Search
External Sequence Search
IMG Sequence Search
Homolog Display

Evidence For Function Prediction (i)

Neighborhood

red = Current Gene
white = Pseudo Gene
green = Positional Cluster Gene in the same KEGG Pathw
||||||| CRISPR array
Sequence Viewer For Alternate ORF Search
Chromosome Viewer colored by COG
Chromosome Viewer colored by GC percentage

Chromosome Viewer (iv)

Switch coloring to: COG

Characteristic GC% - 56 %

Sequence Viewer (ii)

Neighborhood six frame translation with putative ORF's for gene_oid=*638187668*
9925..11652(+)
575aa long hypothetical dihydroxy-acid dehydratase.

Select gene neighborhood:
-0 bp upstream. +0 bp downstream

Select minimum ORF size:
1 aa

Output Format:
○ Text ● Graphics

Sequence Viewer (iii)

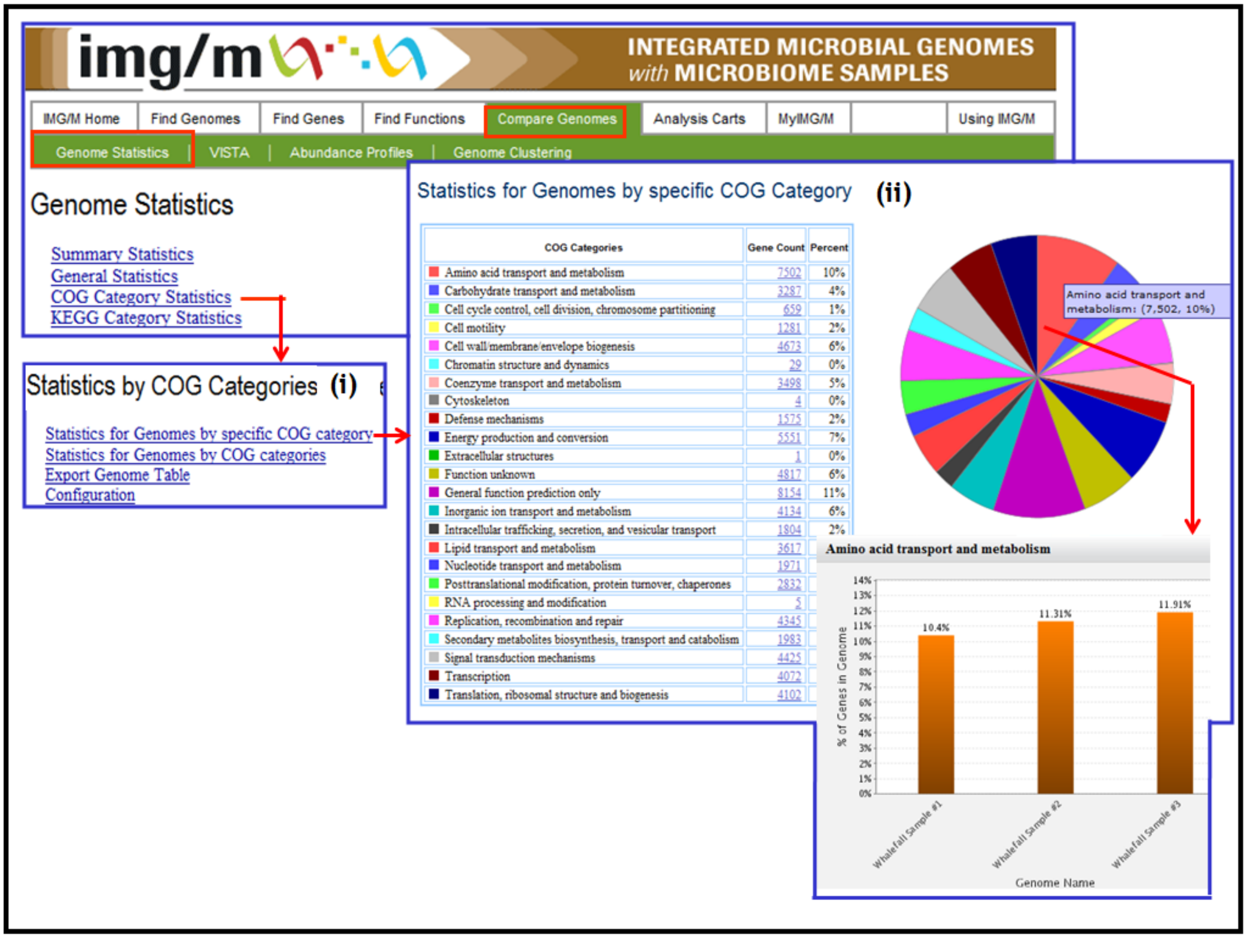
```
F1    M   L   Q   L   F   I   S   L   R   H   P   I   E   T
F2     C   Y   N   F   L   L   V   C   V   I   Q   *   K
F3      A   T   T   F   Y   *   F   A   S   S   N   R   N
GC   33 35 34 33 33 32 33 35 34 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35

9925 ATGCTACAACTTTTTATTAGTTTGCGTCATCCAATAGAAA
     -:-----|-----:-----|-----:-----|-----:-----|---
9925 TACGATGTTGAAAAATAATCAAACGCAGTAGGTTATCTTT
GC   67 65 66 67 67 68 67 65 66 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65
F6          S   C   S   K   I   L   K   R   *   G   I   S
F5        A   V   V   K   *   *   N   A   D   D   L   L   F
F4      H   *   L   K   K   N   T   Q   T   M   W   Y   F
```

FIGURE 4. Gene Details – Evidence for Function Prediction Viewers.

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 5(i)) will display in a tabular and pie chart format the count of genes associated with each COG category across all selected genomes, or metagenomes as illustrated in Figure 5(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 or metagenome associated with that COG category, as illustrated on the lower side pane of Figure 5(ii).



Genome Statistics

Summary Statistics
General Statistics
COG Category Statistics
KEGG Category Statistics

Statistics by COG Categories (i)

Statistics for Genomes by specific COG category
Statistics for Genomes by COG categories
Export Genome Table
Configuration

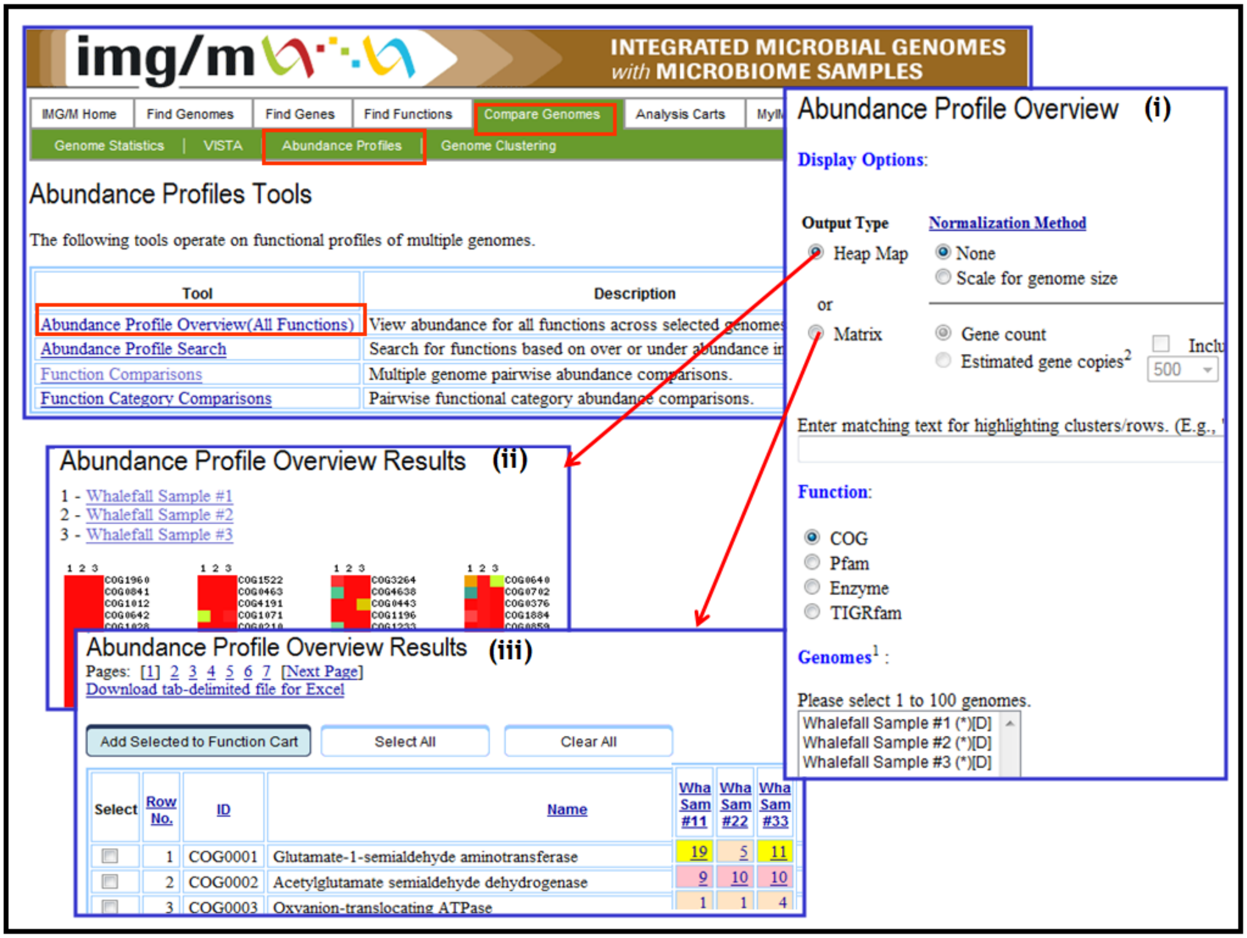
Statistics for Genomes by specific COG Category (ii)

COG Categories	Gene Count	Percent
🟥 Amino acid transport and metabolism	7502	10%
🟦 Carbohydrate transport and metabolism	3287	4%
🟩 Cell cycle control, cell division, chromosome partitioning	659	1%
🟨 Cell motility	1281	2%
🟪 Cell wall/membrane/envelope biogenesis	4673	6%
🟦 Chromatin structure and dynamics	29	0%
🟥 Coenzyme transport and metabolism	3498	5%
⬛ Cytoskeleton	4	0%
🟥 Defense mechanisms	1575	2%
🟦 Energy production and conversion	5551	7%
🟩 Extracellular structures	1	0%
🟨 Function unknown	4817	6%
🟪 General function prediction only	8154	11%
🟦 Inorganic ion transport and metabolism	4134	6%
⬛ Intracellular trafficking, secretion, and vesicular transport	1804	2%
🟥 Lipid transport and metabolism	3617	
🟦 Nucleotide transport and metabolism	1971	
🟩 Posttranslational modification, protein turnover, chaperones	2832	
🟨 RNA processing and modification	5	
🟪 Replication, recombination and repair	4345	
🟦 Secondary metabolites biosynthesis, transport and catabolism	1983	
⬜ Signal transduction mechanisms	4425	
🟥 Transcription	4072	
🟦 Translation, ribosomal structure and biogenesis	4102	

FIGURE 5. Genome Statistics: COG Category statistics.

Abundance Profile Overview

The **Abundance Profile Overview** illustrated in Figure 6 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 metagenomes or genomes in the **Abundance Profile Viewer** page, as illustrated in Figure 6(i) 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 6(ii). Each column on the map corresponds to a genome or metagenome, 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 or metagenome. 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 Profiles Tools

The following tools operate on functional profiles of multiple genomes.

Tool	Description
Abundance Profile Overview(All Functions)	View abundance for all functions across selected genomes
Abundance Profile Search	Search for functions based on over or under abundance in
Function Comparisons	Multiple genome pairwise abundance comparisons.
Function Category Comparisons	Pairwise functional category abundance comparisons.

Abundance Profile Overview (i)

Display Options:

Output Type
● Heap Map
or
○ Matrix

Normalization Method
● None
○ Scale for genome size

○ Gene count
○ Estimated gene copies^2
□ Inclu
500

Enter matching text for highlighting clusters/rows. (E.g., '

Function:

● COG
○ Pfam
○ Enzyme
○ TIGRfam

Genomes^1 :

Please select 1 to 100 genomes.
Whalefall Sample #1 (*)[D]
Whalefall Sample #2 (*)[D]
Whalefall Sample #3 (*)[D]

Abundance Profile Overview Results (ii)

1 - Whalefall Sample #1
2 - Whalefall Sample #2
3 - Whalefall Sample #3

Abundance Profile Overview Results (iii)

Pages: [1] 2 3 4 5 6 7 [Next Page]
Download tab-delimited file for Excel

Add Selected to Function Cart | Select All | Clear All

Select	Row No.	ID	Name	Wha Sam #11	Wha Sam #22	Wha Sam #33
□	1	COG0001	Glutamate-1-semialdehyde aminotransferase	19	5	11
□	2	COG0002	Acetylglutamate semialdehyde dehydrogenase	9	10	10
□	3	COG0003	Oxyanion-translocating ATPase	1	1	4

FIGURE 6. Abundance Profile Overview: Results with Heat Map and Matrix Output Format.

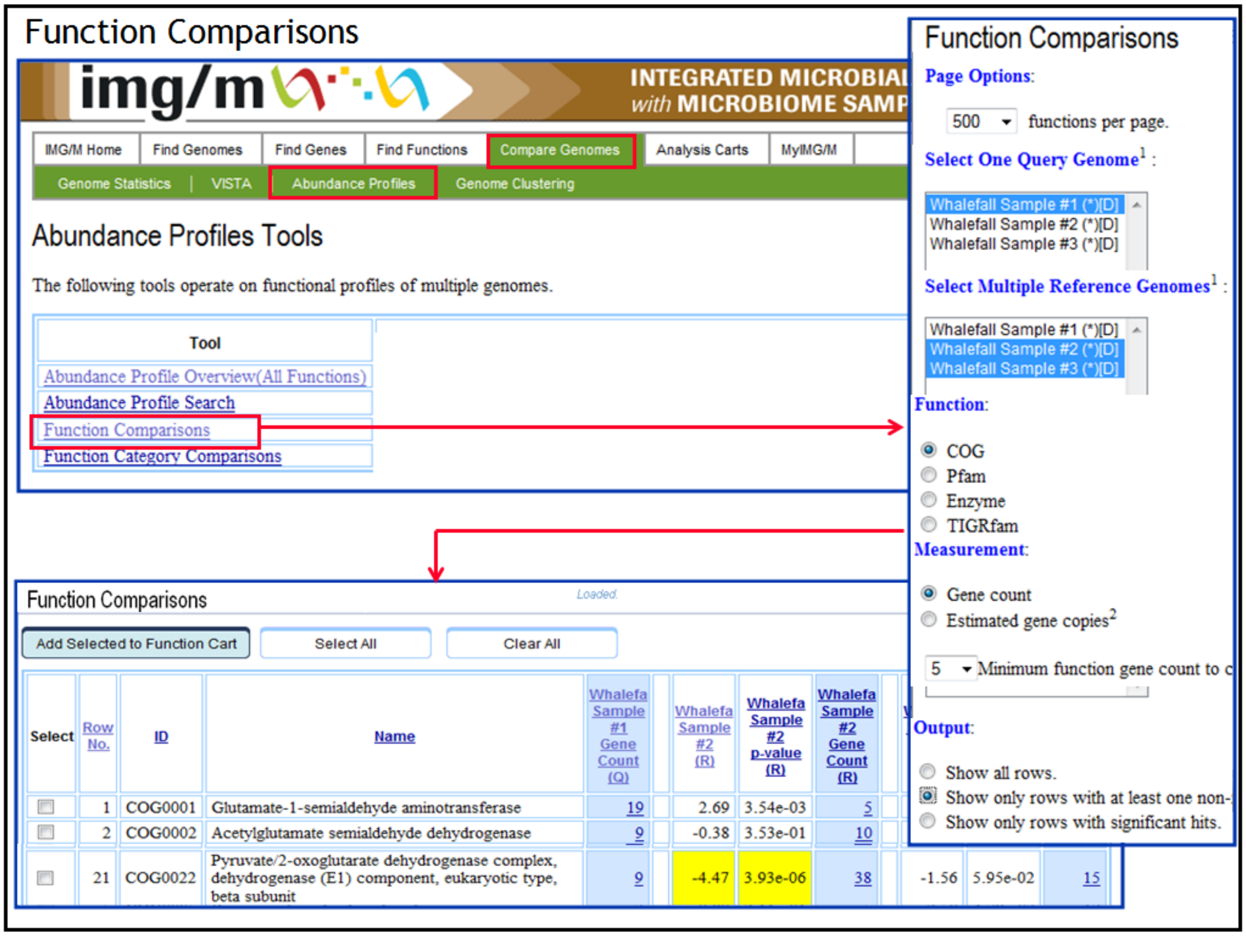
If the “Matrix” output is selected, the abundance of protein/functional families is displayed in a tabular format, with each row corresponding to a family and each cell containing the number of genes associated with a family for a specific genome or metagenome, as illustrated in Figure 6(iii). Click on the cell in order to retrieve the list of genes assigned to this particular family in a genome or metagenome. 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.

Function Comparison

The **Function Comparison** analysis tool has a new organization of the function comparison results and an updated computation of the statistical significance.

The function comparison result lists for each function, ***F***: (the number of genes or estimated gene copies in the query genome/metagenome, ***Q***, associated with ***F***, and for each reference genome/metagenome, ***R***, the number of genes or estimated gene copies in ***R*** associated with ***F***, together with the D-score and p-value associated with the comparison of ***Q*** to ***R***, as illustrated in Figure 7.

In the computation of D-scores, a False Discovery Rate (FDR) correction[4] is used prior to the display of significant hypotheses. FDR correction is meant to control the expected number of false predictions in a multiple-testing scenario. For more details, see section 3.3 of UsingIMG/M (http://img.jgi.doe.gov/m/doc/userGuide_m.pdf).



Abundance Profiles Tools

The following tools operate on functional profiles of multiple genomes.

Tool
Abundance Profile Overview(All Functions)
Abundance Profile Search
Function Comparisons
Function Category Comparisons

Function Comparisons

Page Options:

500 ▾ functions per page.

Select One Query Genome[1]:

Whalefall Sample #1 (*)[D]
Whalefall Sample #2 (*)[D]
Whalefall Sample #3 (*)[D]

Select Multiple Reference Genomes[1]:

Whalefall Sample #1 (*)[D]
Whalefall Sample #2 (*)[D]
Whalefall Sample #3 (*)[D]

Function:

● COG
○ Pfam
○ Enzyme
○ TIGRfam

Measurement:

● Gene count
○ Estimated gene copies[2]

5 ▾ Minimum function gene count to c

Output:

○ Show all rows.
● Show only rows with at least one non-
○ Show only rows with significant hits.

Function Comparisons

Select	Row No.	ID	Name	Whalefa Sample #1 Gene Count (Q)		Whalefa Sample #2 (R)	Whalefa Sample #2 p-value (R)	Whalefa Sample #2 Gene Count (R)		W		
□	1	COG0001	Glutamate-1-semialdehyde aminotransferase	19		2.69	3.54e-03	5				
□	2	COG0002	Acetylglutamate semialdehyde dehydrogenase	9		-0.38	3.53e-01	10				
□	21	COG0022	Pyruvate/2-oxoglutarate dehydrogenase complex, dehydrogenase (E1) component, eukaryotic type, beta subunit	9		-4.47	3.93e-06	38		-1.56	5.95e-02	15

FIGURE 7. Function Comparison.

[4] Benjamini Y and Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57** (1), 289–300.