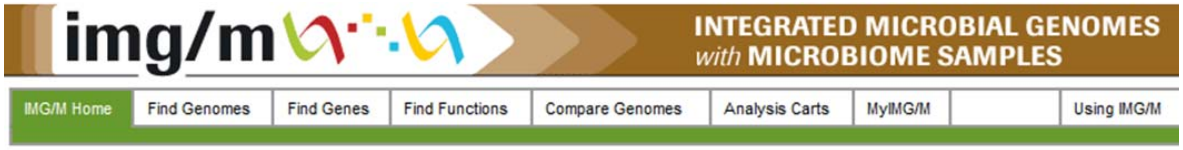
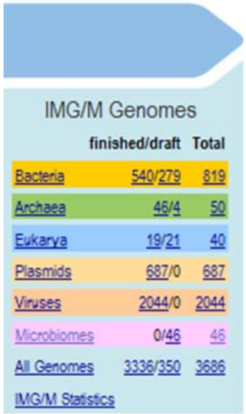



IMG/M Genomes

| | 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 | | |

# Using IMG-M

## Comparative Analysis with the IMG/M System

## Addendum to Using IMG

**Technical Report LBNL-63614**

**Genome Biology Program**
Department of Energy Joint Genome Institute

**Biological Data Management and Technology Center**
Lawrence Berkeley National Laboratory

**September 1, 2008**

## Disclaimers and Copyright

NOTICE: Information from this server resides on a computer system funded by the U.S. Department of Energy. Anyone using this system consents to monitoring of this use by system or security personnel.

### Disclaimer of Liability

With respect to documents available from this server, neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, including the warranties of merchantability and fitness for a particular purpose, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

### Disclaimer of Endorsement

Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

### Copyright Status

Joint Genome Institute authored documents are sponsored by the U.S. Department of Energy under Contracts W-7405-Eng-48, DE-AC02-05CH11231, and W-7405-ENG-36. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce these documents, or allow others to do so, for U.S. Government purposes. All documents available from this server may be protected under the U.S. and Foreign Copyright Laws and permission to reproduce them may be required. The public may copy and use this information without charge, provided that this Notice and any statement of authorship are reproduced on all copies. JGI is not responsible for the contents of any off-site pages referenced.

**September 1, 2008**

This document was prepared by:

Victor M. Markowitz*
Natalia N. Ivanova**
Iain Anderson**
Athanasios Lykidis**
Konstantinos Mavromatis**
Amrita Pati**
Ernest Szeto*
Krishna Palaniappan*
I-Min A. Chen*
Ken Chu*
Yuri Grechkin*
Nikos C. Kyrpides**

*Biological Data Management & Technology Center
Lawrence Berkeley National Laboratory

**Genome Biology Program (GBP)
Department of Energy Joint Genome Institute

# Table of Contents

# 1 Background

***Metagenome*** (environmental genome, community genome) is a sample of the collective genome of all the community members obtained directly from the natural environment, without a preliminary cultivation step.

Two alternative **metagenome sequencing** strategies are generally followed:

1) *directed sequencing*, i.e. sequencing of long-insert libraries after screening for the presence of certain phylogenetic (e.g. – 16S rRNA genes) or functional (e.g. – certain enzymatic activity) markers;
2) *shotgun sequencing* of random clones generated from aggregate DNA by Sanger sequencing or pyrosequencing of aggregate DNA without cloning.

**Metagenome data analysis** aims at addressing at least one of the following three questions:

1) Diversity and abundance of community members ("***who is there***");
2) Metabolic potential of the community and its members ("***what they are doing***");
3) Ecological relations between members of the community ("***why they are there***").

## 1.1 Definitions

***Metagenomic sample*** – usually is equivalent to the isolated aggregate DNA obtained from a certain environment; ideally should be accompanied by comprehensive *metadata* describing how this sample was obtained (e. g. – location, type of an environment, host, isolation protocol, etc.). The DNA is used to create a metagenomic clone *library*, which is further sequenced to produce *reads*. DNA isolation and cloning may introduce certain biases, so the representation of each species in metagenomic sequence may be different from that in the environment.

***Read*** – Sanger sequencing read; if left unassembled, becomes a *single-read contig* in the metagenomic dataset. Single-read contigs almost never appear in isolate genomes (even at the draft stages), but some metagenomic datasets consist almost entirely of unassembled reads (so called *shrapnel*).

***Contig*** **–** the result of assembly of *reads* based on nucleotide sequence identity; the sequence of the contig is a consensus sequence of multiple reads generated by genome assembler (such as JAZZ, PHRAP or Celera assembler [1-3]) and may not be identical to any particular *read*.

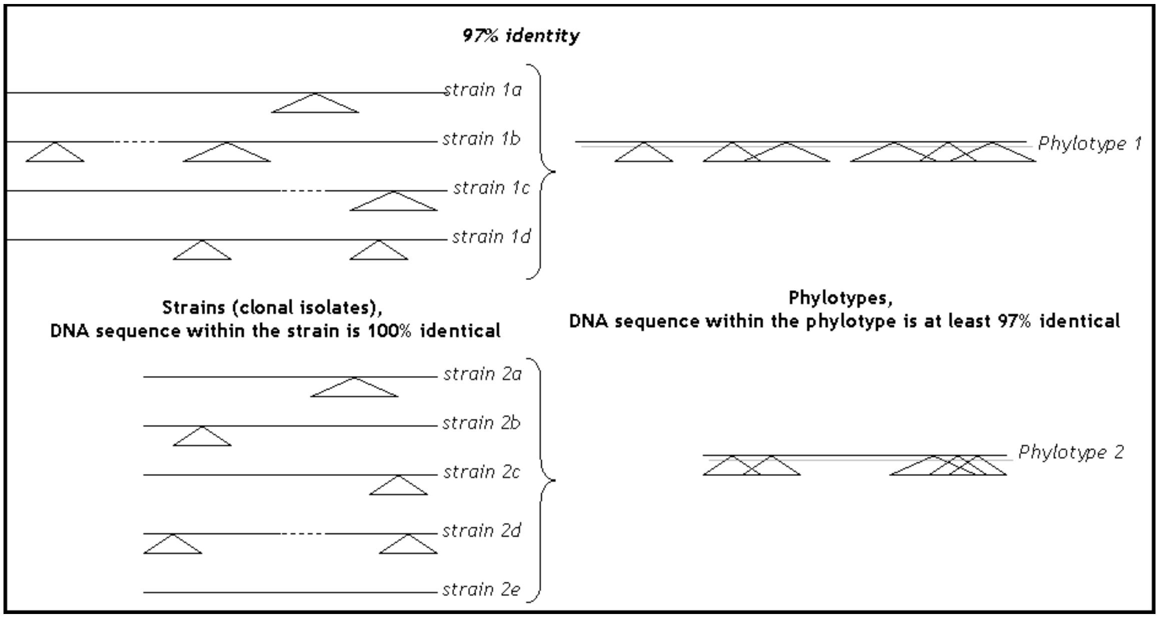
***Scaffold*** – the result of assembly of contigs joined by N-bridged gaps based on read mate-pair information; both *scaffold* and *contig* sequences are further used to identify genes.

***RNA-coding genes***: 16S and 23S rRNAs are usually identify by BLASTn against the corresponding sequences in isolate genomes; 5S rRNA is identified by BLASTn or using Rfam/INFERNAL approach [4]. *tRNA-coding genes* are identified using tRNA-Scan-SE [5]. Other stable RNAs including RNase P, SsrS RNA, SRP and riboswitches are rarely predicted in metagenomic datasets.

*CDSs* (protein-coding gene) are usually identified automatically by *ab initio* gene finding software, such as fgenesb, Glimmer or GeneMark [6-8]; alternatively, they can be predicted by running BLASTx against the protein databases.

***Functional annotations*** (protein product descriptions) are usually performed automatically using RPS-BLAST hits against the Conserved Domain Database (CDD) [9], which combines information from COG (Clusters of Orthologous Groups) and Pfam with several other minor sources; this approach is complemented by BLASTp against protein databases. In addition, hmm searches against the Pfam and TIGRfam databases can be employed.

***Bins*** are sets of metagenomic sequence fragments originating from one phylogenetic group, preferably from the same strain (or phylotype) as illustrated below.


97% identity
strain 1a
strain 1b
strain 1c
strain 1d
Phylotype 1
Strains (clonal isolates),
DNA sequence within the strain is 100% identical
Phylotypes,
DNA sequence within the phylotype is at least 97% identical
strain 2a
strain 2b
strain 2c
strain 2d
strain 2e
Phylotype 2

*Scaffolds*, *contigs* and *reads* are assigned to bins by a *binning tool*, which can use either oligonucleotide composition of DNA fragments (TETRA, PhyloPythia [10-11]) or phylogenetic affiliations of protein-coding genes (e. g. - MEGAN[12] is not a binning tool, but similar to what phylogenetic binning tools would do).

## 1.2 Data Processing

Processing of metagenomic datasets, especially those derived from high-complexity microbiomes, is characterized by significantly higher error rate than processing of isolate genomes [13]. The problems include assembly of chimeric contigs (i. e. assembly of reads originating from different taxonomic groups), under-assembly (i. e. reads that should have been assembled remain as single-read contigs), higher rate of false-positive and false-negative results of gene prediction (mostly due to gene fragmentation), and low sensitivity of binning (i. e. relatively small portion of scaffolds and contigs are assigned to bins, bins correspond to larger taxonomic groups than a species, etc.). Therefore the importance of manual inspection of the data and validation of the results of <u>any</u> analysis cannot be overestimated.

## References

1. *JAZZ assembler*: Aparicio, S., et al. 2002. Whole genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**: 1301–1310.

2. *Celera assembler*: Myers, E.W., et al., 2000. A whole-genome assembly of Drosophila. *Science*, **287**(5461), 2196-2204.

3. *PHRAP assembler*: http://www.phrap.org/phredphrapconsed.html

4. *Rfam and Infernal RNA prediction*: Griffith-Jones, S., et al. 2003. Rfam: an RNA family database. *Nucleic Acids Res*., **31**(1), 439-441.

5. *tRNAscan-SE tRNA prediction*: Lowe, T. M., & Eddy, S. R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**(5), 955-964.

6. *Fgenesb gene prediction*: http://www.softberry.com

7. *Glimmer gene prediction*: Delcher, A.L., et al. 1999. Improved Microbial Gene Identification with Glimmer, *Nucleic Acids Res.,* **27**(23), 4636-4641.

8. *GeneMark gene prediction*: Besemer, J., et al. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*. **29**(12), 2607-2618.

9. *CDD database domain analysis*: Marchler-Bauer, A., et al. 2007. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.* **35**, D237-D240.

10. *TETRA binning tool*: Teeling, H. et al. 2004. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatic***s** **5**, 163.

11. *PhyloPythia binning tool*: McHardy, A. C., et al. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods.* **4**(1), 63-72.

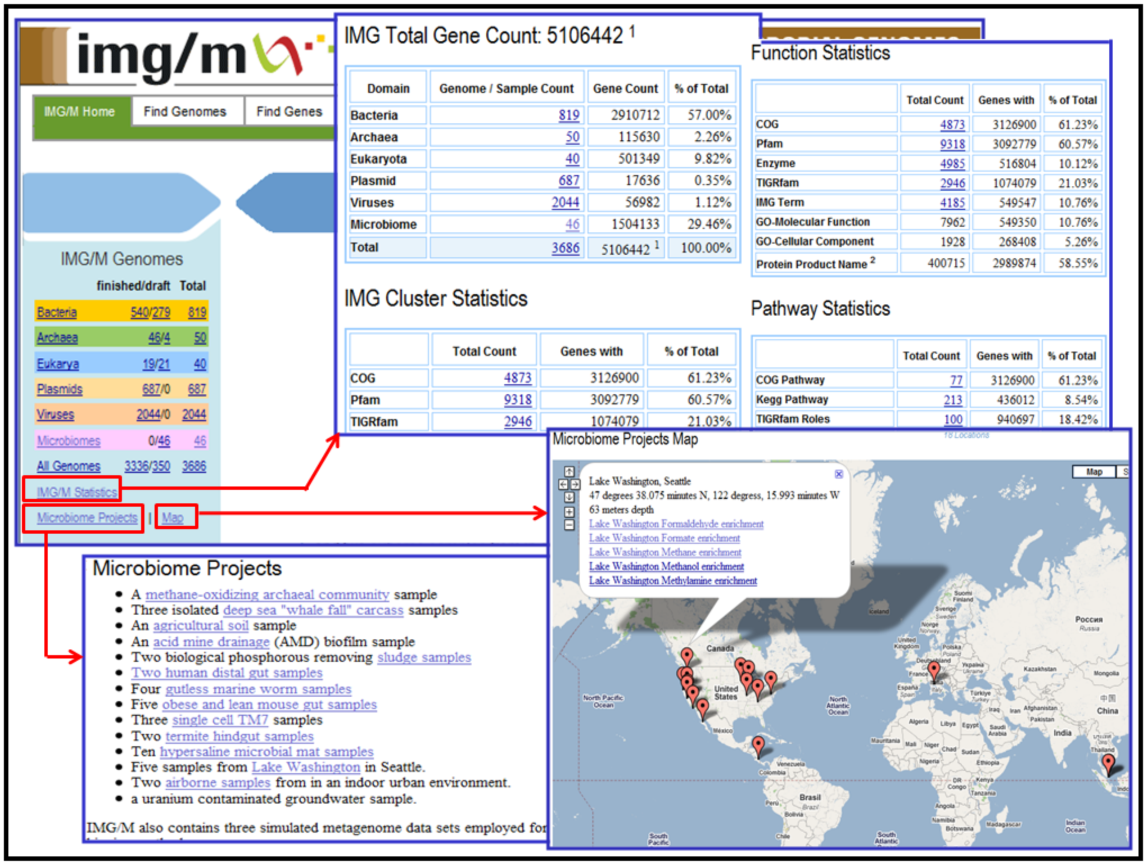
12. *MEGAN tool for phylogenetic analysis*: Huson, D. H., et al. 2007, MEGAN analysis of metagenomic data. *Genome Res.* **17**(3), 377-386.

13. *Problems of metagenome data processing*: Mavromatis, K. et al. 2007. On the fidelity of processing metagenomic sequences using simulated datasets. *Nature Methods* **4**(6), 495-500. See also http://fames.jgi-psf.org/

# 2 IMG/M Overview

## 2.1 IMG/M Data Content

IMG/M consists of microbial metagenome data integrated with isolate microbial genomes from the Integrated Microbial Genomes (IMG) system (http://img.jgi.doe.gov). The current version of IMG/M (as of September 2008) includes 3,686 genomes from IMG 2.4 (released on December 1st, 2007). IMG/M contains metagenome datasets generated using shotgun sequencing for 14 projects involving a total of 43 microbiome samples, as listed below. IMG/M also contains three simulated metagenome data sets employed for benchmarking assembly, gene prediction, and binning methods.



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

18 Locations

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 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.

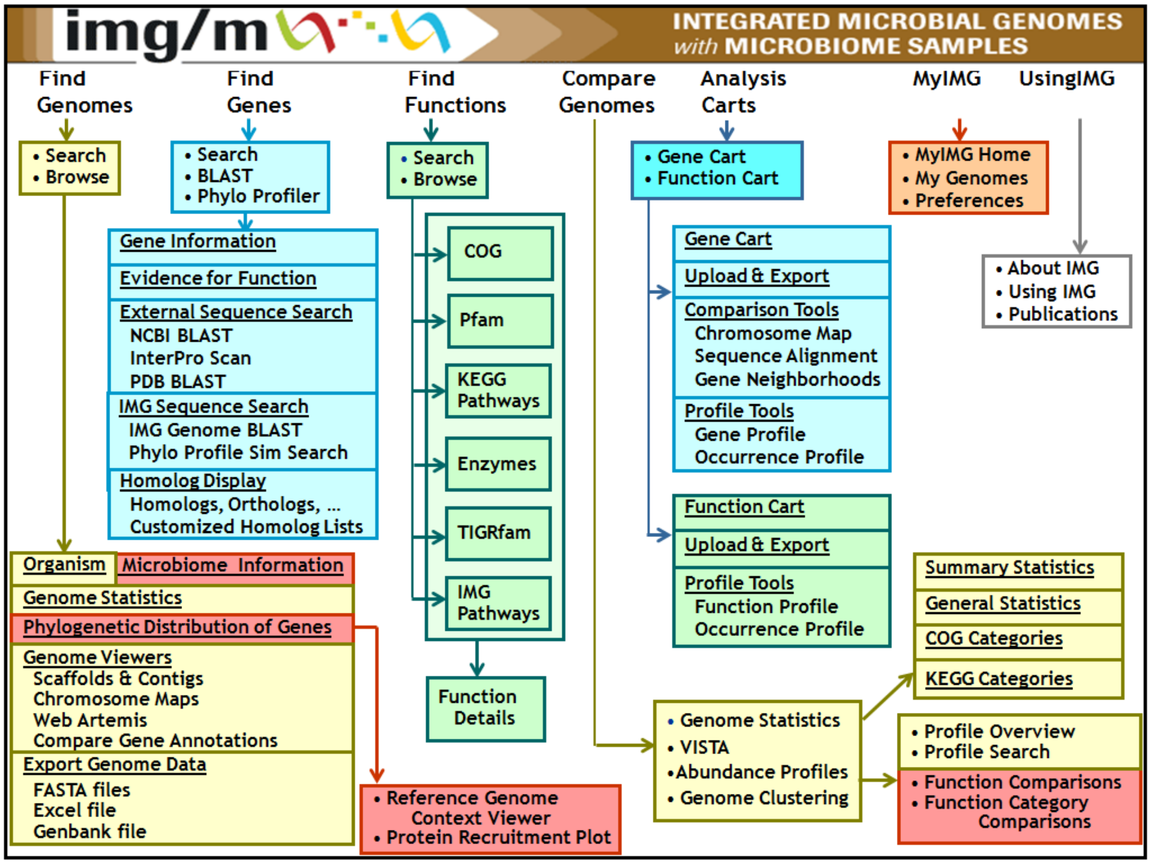
IMG/M also contains three simulated metagenome data sets employed for

Similar to IMG, the data model underlying IMG/M allows recording the primary sequence information and its organization in scaffolds and/or contigs, together with computationally predicted protein-coding sequences and some RNA-coding genes. Protein-coding genes are characterized in terms of additional annotations, such as motifs, domains, pathways and orthology relationships, which may serve as an indication of their functions. These annotations are based on diverse data sources, such as COG, Pfam, and KEGG. Genes are assigned to COGs and Pfams based on RPS-BLAST (Reverse Position Specific BLAST) and NCBI's Conserved Domain Database. Homologs are computed as unidirectional hits with an E-value of $10^{-2}$ or better, with IMG/M providing support for filtering homolog lists by percent identity, bit score, and more stringent E-values.

Isolate organisms are identified via their taxonomic lineage (domain, phylum, class, order, family, genus, species, strain), while individual microbiome samples are treated as "meta" organisms. The sequences of a microbiome sample together with their associated genes and annotations are grouped into *bins* when *binning* has been performed to assign these sequences to organism types (phylotypes). Both isolate organisms and microbiome samples are characterized by a variety of *metadata* attributes.

## 2.2 IMG/M Data Analysis

Genome data analysis in IMG/M is an extension of IMG data analysis to include metagenomes. The tools provided by IMG/M are summarized below, with metagenome specific analysis tools emphasized by the red background.



**Data exploration** tools in IMG/M help selecting and examining genomes/metagenomes, genes, and functions of interest. Similar to IMG, genes and functions can be selected using keyword searches or functional classification (e.g., COG, Pfam) browsers. Lists of genes and functional annotations of interest can be maintained and further explored using various "**Analysis Carts**".

Metagenomes and isolate genomes can be selected using a keyword based **Genome Search** tool or a **Genome Browser**. Microbiomes can be further examined using the **Microbiome Details**" where a user can find relevant metadata, such as sample site, along with various summaries of interest, such as the total number of scaffolds and genes or the number of genes associated with functional characterizations (e.g., COG, Pfam). The **Phylogenetic Distribution of Genes**, further discussed in Section 3.1, provides an estimate of the phylogenetic composition of a microbiome sample based on the distribution of the best BLAST hits of the protein-coding genes in the sample. For each phylum/class, the phylogenetic distribution of genes can be projected onto the families in that phylum/class; for each family the distribution of genes can be further projected onto the species in that family. Finally, the genes in the sample can be viewed in the context of individual reference isolate genome, using either the **Reference Genome Context Viewer** or the **Protein Recruitment Plot**.

Similar to genes of isolate genomes, metagenome genes can be examined using **Gene Details** pages, which include information on locus, biochemical properties of the product, KEGG pathways, as well as evidence for the functional prediction: gene neighborhood, COG and Pfam, and pre-computed lists of homologs, orthologs and paralogs (for isolate organisms), or intra-metagenome homologs as well as homologs to other genomes and metagenomes (for microbiomes).

For metagenomes that include contigs and scaffolds generated by assembly of individual reads and potentially comprised of sequences from multiple strains, a "**NP BLAST** tool, further discussed in Section 3.3, allows to examine the heterogeneity between the reads contributing to the composite populatioon contigs and scaffolds.

**Comparative analysis** of genomes and metagenomes is provided in IMG/M through a number of tools that allow to examine their gene content and functional capabilities. The differences in gene content of genomes and metagenomes can be examined with a profile-based selection tool (**Phylogenetic Profiler**) and further explored through gene neighborhood analysis and multiple sequence alignment tools which are similar to their IMG counterparts.

Functional capabilities of a microbial community can be examined using several occurrence and abundance profile-based tools. **Abundance Profile** tools can be used for comparing the functional capabilities of metagenomes and genomes. The **Abundance Profile Overview** tool, further discussed in Section 4.1, provides an overview of the relative abundance of protein families (COGs and Pfams) and functional families (Enzymes) across selected metagenomes and isolate genomes. The **Abundance Profile Search** tool, further discussed in Section 4.2, is similar to the **Phylogenetic Profiler** tool for gene selection, but operates on protein families rather than individual genes. This tools allows finding protein families (COGs and Pfams) in metagenomes and isolate genomes based on their relative abundance.

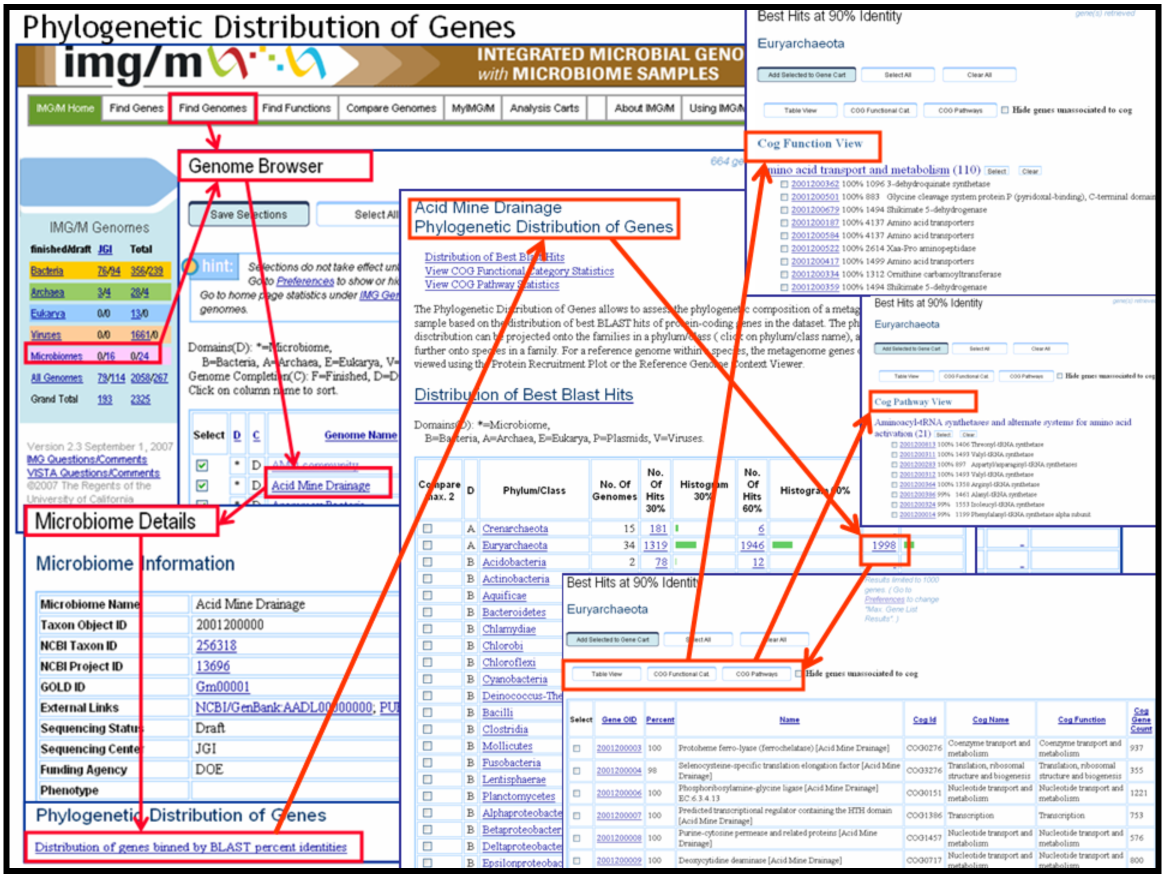
The **Abundance Profile Overview** and **Function Profile** tools provide a rough estimate of the functional capabilities of metagenomes. When metagenomes are compared to each other or to isolate genomes, statistical tests are needed for estimating the *statistical significance* of the observed differences. The **Function Comparison** tool, further discussed in Section 4.3, and **Function Category Comparison** tool, further discussed in Section 4.4, , take into account the stochastic nature of metagenome datasets and test whether the differences in abundance can be ascribed to chance variation or not. The results provided by these tools include an assessment of statistical significance in terms of associated **p-value** and **d-scores** (for Function Comparison) or **d-ranks** (for Function Category Comparison).

# 3 Analysis of Community Diversity & Abundance

## 3.1 Phylogenetic Distribution of Genes

**Purpose**. Assess phylogenetic composition of a metagenome sample based on the distribution of the best BLAST hits of the protein-coding genes found in the dataset.

**Navigation**: IMG/M Microbiomes or Find Genomes/Genome Browser → Microbiome Details → Phylogenetic Distribution of Genes.



| finished/draft | JGI | Total |
|---|---|---|
| Bacteria | 76/94 | 356/239 |
| Archaea | 3/4 | 28/4 |
| Eukarya | 0/0 | 13/0 |
| Viruses | 0/0 | 1661/0 |
| Microbiomes | 0/16 | 0/24 |
| All Genomes | 79/114 | 2058/267 |
| Grand Total | 193 | 2325 |

Version 2.3 September 1, 2007
IMG Questions/Comments
VISTA Questions/Comments
©2007 The Regents of the University of California

Selections do not take effect un
Go to Preferences to show or hi
Go to home page statistics under IMG Ger
genomes.

Domains(D): *=Microbiome, B=Bacteria, A=Archaea, E=Eukarya, V=
Genome Completion(C): F=Finished, D=D
Click on column name to sort.

| Select | D | C | Genome Name |
|---|---|---|---|
| ☑ | * | D | AMD community |
| ☑ | * | D | Acid Mine Drainage |

Microbiome Information

| | |
|---|---|
| Microbiome Name | Acid Mine Drainage |
| Taxon Object ID | 2001200000 |
| NCBI Taxon ID | 256318 |
| NCBI Project ID | 13696 |
| GOLD ID | Gm00001 |
| External Links | NCBI/GenBank:AADL00000000; PU |
| Sequencing Status | Draft |
| Sequencing Center | JGI |
| Funding Agency | DOE |
| Phenotype | |

Phylogenetic Distribution of Genes
Distribution of genes binned by BLAST percent identities

Acid Mine Drainage
Phylogenetic Distribution of Genes

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

The Phylogenetic Distribution of Genes allows to assess the phylogenetic composition of a metag sample based on the distribution of best BLAST hits of protein-coding genes in the dataset. The ph disctribution can be projected onto the families in a phylum/class ( click on phylum/class name), a further onto species in a family. For a reference genome within a species, the metagenome genes viewed using the Protein Recruitment Plot or the Reference Genome Context Viewer.

Distribution of Best Blast Hits

Domains(D): *=Microbiome, B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids, V=Viruses.

| Compare max. 2 | D | Phylum/Class | No. Of Genomes | No. Of Hits 30% | Histogram 30% | No. Of Hits 60% | Histogram 60% | | |
|---|---|---|---|---|---|---|---|---|---|
| □ | A | Crenarchaeota | 15 | 181 | ▮ | 6 | | | |
| □ | A | Euryarchaeota | 34 | 1319 | ▬ | 1946 | ▬ | 1998 | ▮ |
| □ | B | Acidobacteria | 2 | 78 | ▮ | 12 | | | |
| □ | B | Actinobacteria | | | | | | | |
| □ | B | Aquificae | | | | | | | |
| □ | B | Bacteroidetes | | | | | | | |
| □ | B | Chlamydiae | | | | | | | |
| □ | B | Chlorobi | | | | | | | |
| □ | B | Chloroflexi | | | | | | | |
| □ | B | Cyanobacteria | | | | | | | |
| □ | B | Deinococcus-The | | | | | | | |
| □ | B | Bacilli | | | | | | | |
| □ | B | Clostridia | | | | | | | |
| □ | B | Mollicutes | | | | | | | |
| □ | B | Fusobacteria | | | | | | | |
| □ | B | Lentisphaerae | | | | | | | |
| □ | B | Planctomycetes | | | | | | | |
| □ | B | Alphaproteobacte | | | | | | | |
| □ | B | Betaproteobacter | | | | | | | |
| □ | B | Deltaproteobacte | | | | | | | |
| □ | B | Epsilonproteobac | | | | | | | |

Results limited to 1000 genes. ( Go to Preferences to change "Max. Gene List Results". )

Best Hits at 90% Identity
gene(s) retrieved

Euryarchaeota

□ Hide genes unassociated to cog

Amino acid transport and metabolism (110) Select Clear
□ 2001200362 100% 1096 3-dehydroquinate synthetase
□ 2001200501 100% 883 Glycine cleavage system protein P (pyridoxal-binding), C-terminal domain
□ 2001200679 100% 1494 Shikimate 5-dehydrogenase
□ 2001200187 100% 4137 Amino acid transporters
□ 2001200584 100% 4137 Amino acid transporters
□ 2001200522 100% 2614 Xaa-Pro aminopeptidase
□ 2001200417 100% 1499 Amino acid transporters
□ 2001200334 100% 1312 Ornithine carbamoyltransferase
□ 2001200359 100% 1494 Shikimate 5-dehydrogenase

Best Hits at 90% Identity
gene(s) retrieved

Euryarchaeota

□ Hide genes unassociated to cog

Aminoacyl-tRNA synthetases and alternate systems for amino acid activation (21) Select Clear
□ 2001200813 100% 1406 Threonyl-tRNA synthetase
□ 2001200311 100% 1493 Valyl-tRNA synthetase
□ 2001200283 100% 897 Aspartyl/asparaginyl-tRNA synthetases
□ 2001200312 100% 1493 Valyl-tRNA synthetase
□ 2001200364 100% 1358 Arginyl-tRNA synthetase
□ 2001200386 99% 1461 Alanyl-tRNA synthetase
□ 2001200324 99% 1553 Isoleucyl-tRNA synthetase
□ 2001200014 99% 1199 Phenylalanyl-tRNA synthetase alpha subunit

Best Hits at 90% Identity

Euryarchaeota

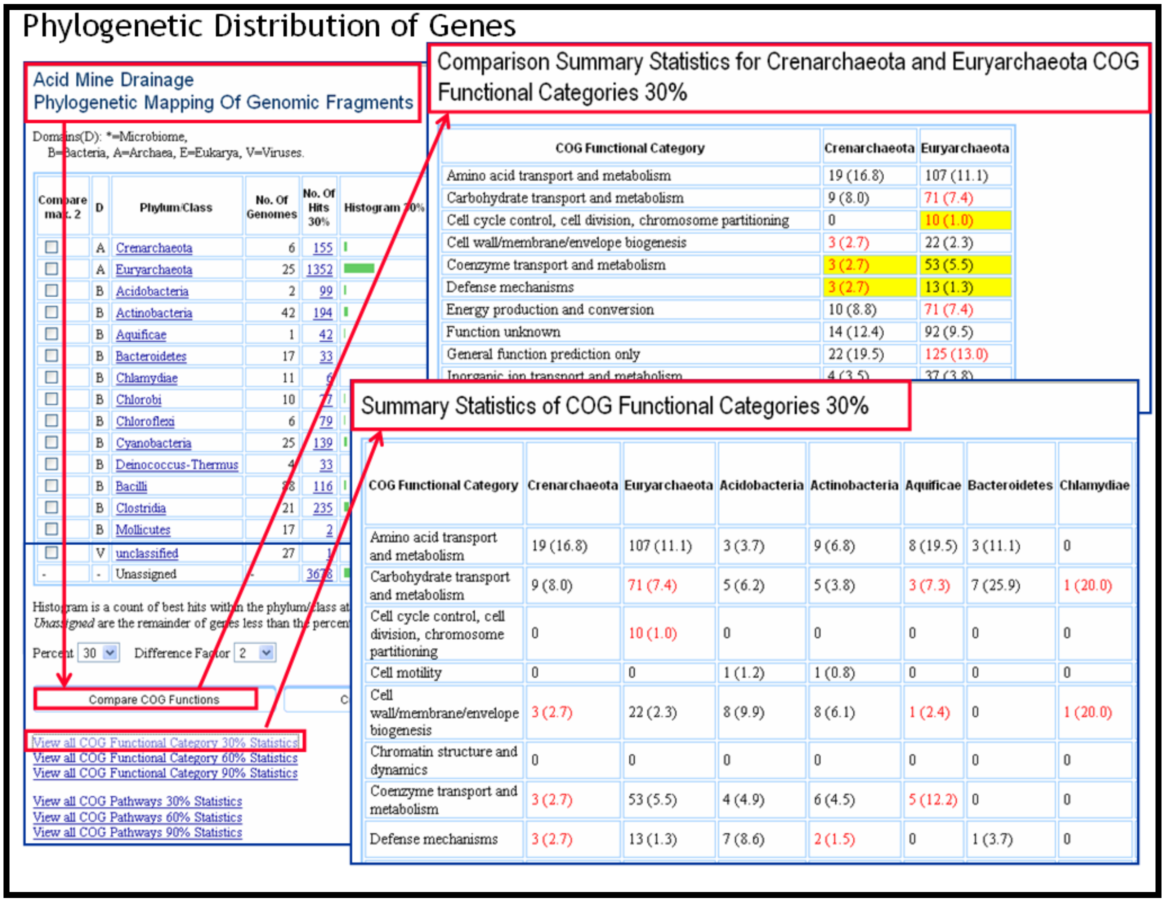
□ Hide genes unassociated to cog

| Select | Gene OID | Percent | Name | Cog Id | Cog Name | Cog Function | Cog Gene Count |
|---|---|---|---|---|---|---|---|
| □ | 2001200003 | 100 | Protoheme ferro-lyase (ferrochelatase) [Acid Mine Drainage] | COG0276 | Coenzyme transport and metabolism | Coenzyme transport and metabolism | 937 |
| □ | 2001200004 | 98 | Selenocysteine-specific translation elongation factor [Acid Mine Drainage] | COG3276 | Translation, ribosomal structure and biogenesis | Translation, ribosomal structure and biogenesis | 355 |
| □ | 2001200006 | 100 | Phosphoribosylamine-glycine ligase [Acid Mine Drainage] EC:6.3.4.13 | COG0151 | Nucleotide transport and metabolism | Nucleotide transport and metabolism | 1221 |
| □ | 2001200007 | 100 | Predicted transcriptional regulator containing the HTH domain [Acid Mine Drainage] | COG1386 | Transcription | Transcription | 753 |
| □ | 2001200008 | 100 | Purine-cytosine permease and related proteins [Acid Mine Drainage] | COG1457 | Nucleotide transport and metabolism | Nucleotide transport and metabolism | 576 |
| □ | 2001200009 | 100 | Deoxycytidine deaminase [Acid Mine Drainage] | COG0717 | Nucleotide transport and metabolism | Nucleotide transport and metabolism | 800 |

**Functionality (1)**. The phylogenetic distribution of best BLAST hits of protein-coding genes in the metagenome is displayed as a histogram; counts correspond to the number of metagenome genes that have best BLASTp hits to proteins in this phylum or class with more than 90% identity (right column), 60-90% identity (middle column) and 30-60% identity (left column). The higher the number of hits and percent identity cutoff, the more likely it is that the metagenome contains close relatives of the sequenced representatives from this phylum/class.

Gene counts in the histogram are linked to the lists of genes in the metagenome that have best BLAST hit in a certain phylum/class with specified percent identity. The genes in the list can be sorted either by their oids ("*Table View*") or by their assignment to COGs, which in turn can be classified according to COG Functional Categories ("*COG Functional Cat.*") or COG Pathways ("*COG Pathways*"). The genes in the table can be selected and added to Gene Cart or analyzed through the corresponding Gene Pages.

**Navigation (cont.)**: Phylogenetic Distribution of Genes → Compare COG Functions; View all COG Functional Category/Pathways



Phylogenetic Distribution of Genes

Acid Mine Drainage
Phylogenetic Mapping Of Genomic Fragments

Domains(D): *=Microbiome,
B=Bacteria, A=Archaea, E=Eukarya, V=Viruses.

| Compare max. 2 | D | Phylum/Class | No. Of Genomes | No. Of Hits 30% | Histogram 30% |
|---|---|---|---|---|---|
| □ | A | Crenarchaeota | 6 | 155 | ▮ |
| □ | A | Euryarchaeota | 25 | 1352 | ▮ |
| □ | B | Acidobacteria | 2 | 99 | ▮ |
| □ | B | Actinobacteria | 42 | 194 | ▮ |
| □ | B | Aquificae | 1 | 42 | ▮ |
| □ | B | Bacteroidetes | 17 | 33 | |
| □ | B | Chlamydiae | 11 | 6 | |
| □ | B | Chlorobi | 10 | 77 | ▮ |
| □ | B | Chloroflexi | 6 | 79 | ▮ |
| □ | B | Cyanobacteria | 25 | 139 | ▮ |
| □ | B | Deinococcus-Thermus | 4 | 33 | |
| □ | B | Bacilli | 38 | 116 | ▮ |
| □ | B | Clostridia | 21 | 235 | ▮ |
| □ | B | Mollicutes | 17 | 2 | |
| □ | V | unclassified | 27 | 1 | |
| - | - | Unassigned | - | 3678 | ▮ |

Histogram is a count of best hits within the phylum/class at
*Unassigned* are the remainder of genes less than the percent

Percent 30 ▾ Difference Factor 2 ▾

Compare COG Functions

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

View all COG Pathways 30% Statistics
View all COG Pathways 60% Statistics
View all COG Pathways 90% Statistics

Comparison Summary Statistics for Crenarchaeota and Euryarchaeota COG Functional Categories 30%

| COG Functional Category | Crenarchaeota | Euryarchaeota |
|---|---|---|
| Amino acid transport and metabolism | 19 (16.8) | 107 (11.1) |
| Carbohydrate transport and metabolism | 9 (8.0) | 71 (7.4) |
| Cell cycle control, cell division, chromosome partitioning | 0 | 10 (1.0) |
| Cell wall/membrane/envelope biogenesis | 3 (2.7) | 22 (2.3) |
| Coenzyme transport and metabolism | 3 (2.7) | 53 (5.5) |
| Defense mechanisms | 3 (2.7) | 13 (1.3) |
| Energy production and conversion | 10 (8.8) | 71 (7.4) |
| Function unknown | 14 (12.4) | 92 (9.5) |
| General function prediction only | 22 (19.5) | 125 (13.0) |
| Inorganic ion transport and metabolism | 4 (3.5) | 37 (3.8) |

Summary Statistics of COG Functional Categories 30%

| COG Functional Category | Crenarchaeota | Euryarchaeota | Acidobacteria | Actinobacteria | Aquificae | Bacteroidetes | Chlamydiae |
|---|---|---|---|---|---|---|---|
| Amino acid transport and metabolism | 19 (16.8) | 107 (11.1) | 3 (3.7) | 9 (6.8) | 8 (19.5) | 3 (11.1) | 0 |
| Carbohydrate transport and metabolism | 9 (8.0) | 71 (7.4) | 5 (6.2) | 5 (3.8) | 3 (7.3) | 7 (25.9) | 1 (20.0) |
| Cell cycle control, cell division, chromosome partitioning | 0 | 10 (1.0) | 0 | 0 | 0 | 0 | 0 |
| Cell motility | 0 | 0 | 1 (1.2) | 1 (0.8) | 0 | 0 | 0 |
| Cell wall/membrane/envelope biogenesis | 3 (2.7) | 22 (2.3) | 8 (9.9) | 8 (6.1) | 1 (2.4) | 0 | 1 (20.0) |
| Chromatin structure and dynamics | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Coenzyme transport and metabolism | 3 (2.7) | 53 (5.5) | 4 (4.9) | 6 (4.5) | 5 (12.2) | 0 | 0 |
| Defense mechanisms | 3 (2.7) | 13 (1.3) | 7 (8.6) | 2 (1.5) | 0 | 1 (3.7) | 0 |

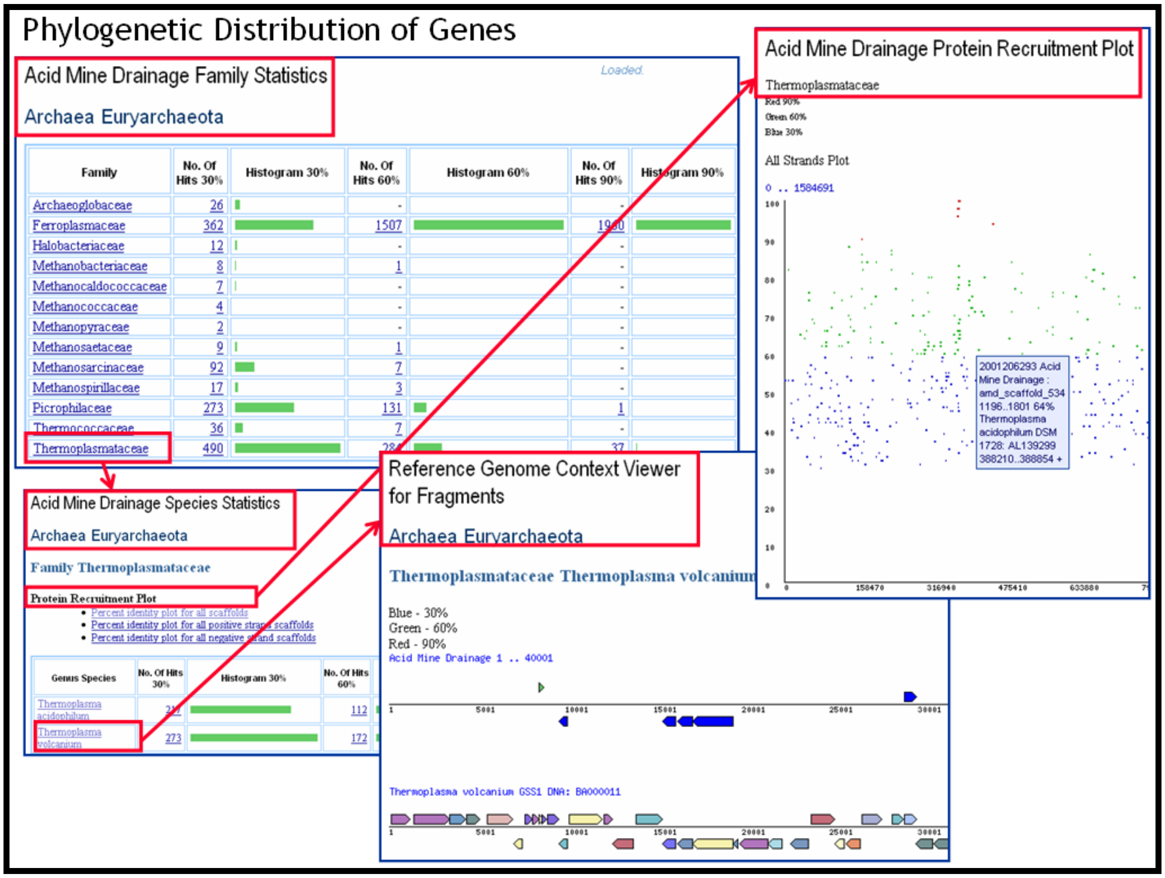
**Functionality (2)**. For metagenome genes that have best BLASTp hits with 30%, 60-90%, and 90% identity, the tool displays summary statistics of COG functional categories and pathways associated with the metagenome genes either across all phyla/classes or across two selected phyla/classes. The statistics show the number of genes associated with a specific COG category or pathway, as well as the percentage of these genes out of the total number of genes with affiliation to this particular class or phylum in the certain interval of percent identity of the best BLAST hits (shown in parenthesis).

These summary statistics tables provide a quick overview of the functional complement of metagenomic proteins with likely affiliations to different phyla/classes (e. g. which COG categories are populated with most genes). In addition, they help to identify the areas of metabolism overrepresented among the genes with different phylogenetic affiliations as defined by their best BLAST hits thus indicating possible functional specialization within the community.

**Notes.** It should be pointed out that these comparisons provide reliable estimates of functional complement only for well-populated phyla/classes and COG categories; if very few genes are affiliated with certain phylum/class it is likely that either these genes originated from a poorly sampled organism with very low abundance or that these genes belong to an abundant organism but were subject to horizontal transfer and therefore appear to have different phylogenetic affiliation than the rest of the genome.

The summary statistics can be exported to an Excel file.

**Navigation (cont.)**: Phylogenetic Distribution of Genes → (Project on Family) → (Project on Species) → Protein Recruitment Plot / Reference Genome Context Viewer



Phylogenetic Distribution of Genes

Acid Mine Drainage Family Statistics

Archaea Euryarchaeota

Loaded.

| Family | No. Of Hits 30% | Histogram 30% | No. Of Hits 60% | Histogram 60% | No. Of Hits 90% | Histogram 90% |
|---|---|---|---|---|---|---|
| Archaeoglobaceae | 26 | 🟩 | - | | - | |
| Ferroplasmaceae | 362 | 🟩 | 1507 | 🟩 | 1900 | 🟩 |
| Halobacteriaceae | 12 | 🟩 | - | | - | |
| Methanobacteriaceae | 8 | 🟩 | 1 | | - | |
| Methanocaldococcaceae | 7 | 🟩 | - | | - | |
| Methanococcaceae | 4 | | - | | - | |
| Methanopyraceae | 2 | | - | | - | |
| Methanosaetaceae | 9 | 🟩 | 1 | | - | |
| Methanosarcinaceae | 92 | 🟩 | 7 | | - | |
| Methanospirillaceae | 17 | 🟩 | 3 | | - | |
| Picrophilaceae | 273 | 🟩 | 131 | 🟩 | 1 | |
| Thermococcaceae | 36 | 🟩 | 7 | | - | |
| Thermoplasmataceae | 490 | 🟩 | 284 | 🟩 | 37 | 🟩 |

Acid Mine Drainage Species Statistics

Archaea Euryarchaeota

Family Thermoplasmataceae

Protein Recruitment Plot

- Percent identity plot for all scaffolds
- Percent identity plot for all positive strand scaffolds
- Percent identity plot for all negative strand scaffolds

| Genus Species | No. Of Hits 30% | Histogram 30% | No. Of Hits 60% |
|---|---|---|---|
| Thermoplasma acidophilum | 217 | 🟩 | 112 |
| Thermoplasma volcanium | 273 | 🟩 | 172 |

Acid Mine Drainage Protein Recruitment Plot

Thermoplasmataceae

Red 90%
Green 60%
Blue 30%

All Strands Plot

0 .. 1584691

2001206293 Acid Mine Drainage : amd_scaffold_534 1196..1801 64% Thermoplasma acidophilum DSM 1728: AL139299 388210..388854 +

Reference Genome Context Viewer for Fragments

Archaea Euryarchaeota

Thermoplasmataceae Thermoplasma volcanium

Blue - 30%
Green - 60%
Red - 90%

Acid Mine Drainage 1 .. 40001

Thermoplasma volcanium GSS1 DNA: BA000011

**Functionality (3)**. The phylogenetic distribution of best BLAST hits can be projected onto the families in a phylum/class, and then further onto species in a family. For a specific species, the metagenome genes can be displayed using the **Protein Recruitment Plot** which displays the BLASTp hits of the metagenomic CDSs against the genes of the isolate genomes within this species, with the coordinates of the scaffolds for the isolate genomes and the BLAST percent identities shown on the X and Y axis, respectively. Each gene on the plot is linked to its Gene Details page. From the group of isolate genomes in a species, one can select a reference genome and view the metagenome genes either using the **Protein Recruitment Plot** or in the context of the chromosome view for this genome using **the Reference Genome Context Viewer**: the metagenome genes are colored according to the % identity of the BLAST hits, while the reference genome genes are colored according to their COG assignments.

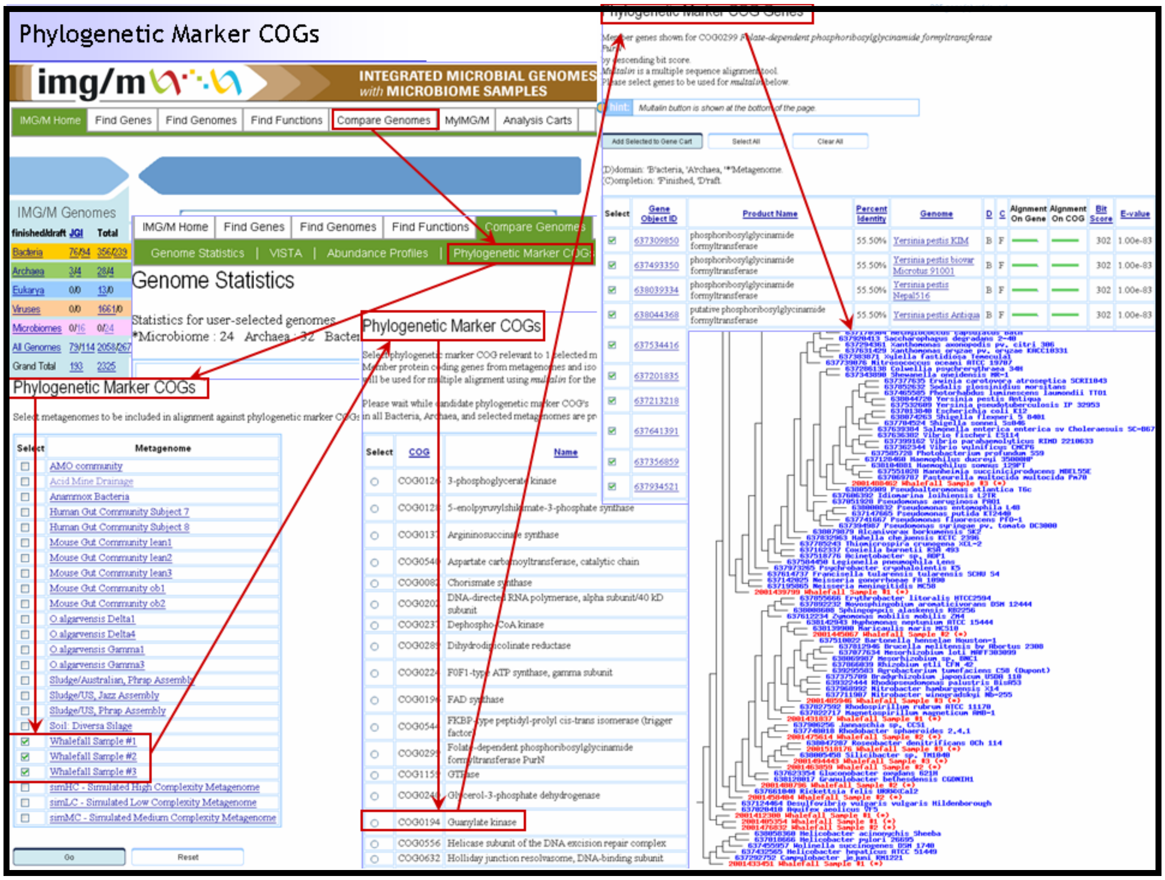
This tool allows higher granularity display of the functional content of the "bins" represented by the metagenomic CDSs with different phylogenetic affiliations according to their best BLAST hits. In addition, whenever a reference genome close to an organism in the metagenomic dataset exists, it helps identify the similarities and differences in genetic content of the two organisms.

**Notes**. This analysis is based on all protein-coding genes in the metagenome, including those protein families prone to rapid expansion/reduction and horizontal transfer. Thus the histogram cannot be used as a good estimate of the phylogenetic composition of a microbiome – especially when the phyla/classes with low counts are considered.

## 3.2 Phylogenetic Marker COGs

**Purpose**. Tool for multiple sequence alignment (MSA) of metagenome genes assigned to single-copy COGs against the representatives of these COGs from finished isolate genomes; based on Multalin MSA tool [Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res., 16(22), 10881-10890].

**Navigation**: Compare Genome → Phylogenetic Marker COGs



IMG/M Genomes

| | finished/draft JGI | Total |
|---|---|---|
| Bacteria | 76/94 | 356/239 |
| Archaea | 3/4 | 28/4 |
| Eukarya | 0/0 | 13/0 |
| Viruses | 0/0 | 1661/0 |
| Microbiomes | 0/16 | 0/24 |
| All Genomes | 79/114 | 2058/267 |
| Grand Total | 193 | 2325 |

Genome Statistics

Statistics for user-selected genomes
*Microbiome : 24 Archaea : 32 Bacter

Phylogenetic Marker COGs

Select metagenomes to be included in alignment against phylogenetic marker COGs

| Select | Metagenome |
|---|---|
| □ | AMO community |
| □ | Acid Mine Drainage |
| □ | Anammox Bacteria |
| □ | Human Gut Community Subject 7 |
| □ | Human Gut Community Subject 8 |
| □ | Mouse Gut Community lean1 |
| □ | Mouse Gut Community lean2 |
| □ | Mouse Gut Community lean3 |
| □ | Mouse Gut Community ob1 |
| □ | Mouse Gut Community ob2 |
| □ | O algarvensis Delta1 |
| □ | O algarvensis Delta4 |
| □ | O algarvensis Gamma1 |
| □ | O algarvensis Gamma3 |
| □ | Sludge/Australian, Phrap Assembly |
| □ | Sludge/US, Jazz Assembly |
| □ | Sludge/US, Phrap Assembly |
| □ | Soil: Diversa Silage |
| ☑ | Whalefall Sample #1 |
| ☑ | Whalefall Sample #2 |
| ☑ | Whalefall Sample #3 |
| □ | simHC - Simulated High Complexity Metagenome |
| □ | simLC - Simulated Low Complexity Metagenome |
| □ | simMC - Simulated Medium Complexity Metagenome |

Phylogenetic Marker COGs

Select phylogenetic marker COG relevant to 1 selected m
Member protein coding genes from metagenomes and iso
will be used for multiple alignment using *multalin* for the

Please wait while candidate phylogenetic marker COG's
in all Bacteria, Archaea, and selected metagenomes are pr

| Select | COG | Name |
|---|---|---|
| ○ | COG0126 | 3-phosphoglycerate kinase |
| ○ | COG0128 | 5-enolpyruvylshikimate-3-phosphate synthase |
| ○ | COG0137 | Argininosuccinate synthase |
| ○ | COG0540 | Aspartate carbamoyltransferase, catalytic chain |
| ○ | COG0082 | Chorismate synthase |
| ○ | COG0202 | DNA-directed RNA polymerase, alpha subunit/40 kD subunit |
| ○ | COG0237 | Dephospho-CoA kinase |
| ○ | COG0289 | Dihydrodipicolinate reductase |
| ○ | COG0224 | F0F1-type ATP synthase, gamma subunit |
| ○ | COG0196 | FAD synthase |
| ○ | COG0544 | FKBP-type peptidyl-prolyl cis-trans isomerase (trigger factor) |
| ○ | COG0299 | Folate-dependent phosphoribosylglycinamide formyltransferase PurN |
| ○ | COG1159 | GTPase |
| ○ | COG0240 | Glycerol-3-phosphate dehydrogenase |
| ○ | COG0194 | Guanylate kinase |
| ○ | COG0556 | Helicase subunit of the DNA excision repair complex |
| ○ | COG0632 | Holliday junction resolvasome, DNA-binding subunit |

Phylogenetic Marker COG Genes

*Member genes shown for COG0299 Folate-dependent phosphoribosylglycinamide formyltransferase PurN*
by descending bit score.
*Multalin* is a multiple sequence alignment tool.
Please select genes to be used for *multalin* below.

hint: *Multalin button is shown at the bottom of the page.*

(D)domain: 'B'acteria, 'A'rchaea, '*'Metagenome.
(C)ompletion: 'F'inished, 'D'raft.

| Select | Gene Object ID | Product Name | Percent Identity | Genome | D | C | Alignment On Gene | Alignment On COG | Bit Score | E-value |
|---|---|---|---|---|---|---|---|---|---|---|
| ☑ | 637309850 | phosphoribosylglycinamide formyltransferase | 55.50% | Yersinia pestis KIM | B | F | — | — | 302 | 1.00e-83 |
| ☑ | 637493350 | phosphoribosylglycinamide formyltransferase | 55.50% | Yersinia pestis biovar Microtus 91001 | B | F | — | — | 302 | 1.00e-83 |
| ☑ | 638039334 | phosphoribosylglycinamide formyltransferase | 55.50% | Yersinia pestis Nepal516 | B | F | — | — | 302 | 1.00e-83 |
| ☑ | 638044368 | putative phosphoribosylglycinamide formyltransferase | 55.50% | Yersinia pestis Antiqua | B | F | — | — | 302 | 1.00e-83 |
| ☑ | 637534416 | | | | | | | | | |
| ☑ | 637201835 | | | | | | | | | |
| ☑ | 637213218 | | | | | | | | | |
| ☑ | 637641391 | | | | | | | | | |
| ☑ | 637356859 | | | | | | | | | |
| ☑ | 637934521 | | | | | | | | | |

**Functionality.** Select one or more metagenomes; then select a COG representatives of which should be aligned. COGs in the list were selected by following criteria:

- genes assigned to this COG with at least 30% identity and aligned over at least 80% of both COG and gene length are found in at least 50% of finished genomes;
- no more than 10 finished genomes have 2 genes assigned to this COG, while all others have only 1.

The COGs are annotated with regard to whether they are strictly single-copy (i. e. there isn't a single genome with more than 1 gene assigned to this COG). Select a COG for alignment, select genes to be included in the alignment - the list of genes includes multiple representatives from the strains of the same species (e. g., *E. coli*, *S. pyogenes*, etc.), some of them can be deselected to reduce the size of the tree. The genes selected for the alignment can be added to the Gene Cart.

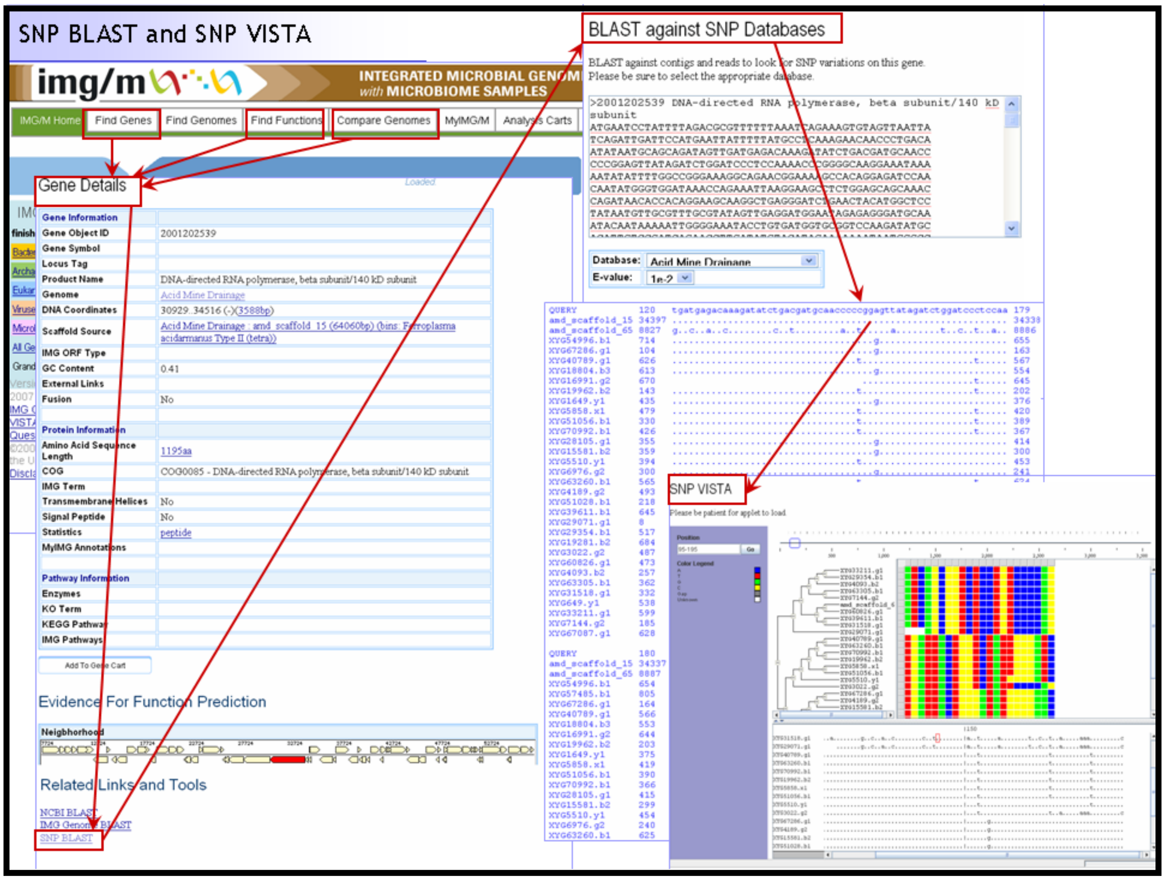
Alignment of a large number of genes may take some time; the genes on the tree are linked to the corresponding Gene Pages and the representatives of the metagenome are highlighted in red. Multiple sequence alignment in MSF format is provided in the bottom of the page.

**Notes**. The tree generated by Multalin is not a phylogenetic tree, but rather a hierarchical clustering tree. To calculate a phylogenetic tree, use the alignment in MSF format provided at the bottom of the page and the tools outside IMG (e. g. see http://au.expasy.org/tools/#align). Beware of the effects of gene fragmentation on MSA and phylogenetic trees.

## 3.3 SNP BLAST and SNP VISTA

**Purpose**. Examine strain-level heterogeneity in the metagenomes of the communities with highly abundant members by analyzing SNPs (single nucleotide polymorphisms) within multi-read contigs and scaffolds.

**Navigation.** Gene Details → SNP BLAST → SNP VISTA



SNP BLAST and SNP VISTA

Gene Details

| | |
|---|---|
| **Gene Information** | |
| Gene Object ID | 2001202539 |
| Gene Symbol | |
| Locus Tag | |
| Product Name | DNA-directed RNA polymerase, beta subunit/140 kD subunit |
| Genome | Acid Mine Drainage |
| DNA Coordinates | 30929..34516 (-)(3588bp) |
| Scaffold Source | Acid Mine Drainage : amd_scaffold_15 (64060bp) (bins: Ferroplasma acidarmanus Type II (tetra)) |
| IMG ORF Type | |
| GC Content | 0.41 |
| External Links | |
| Fusion | No |
| **Protein Information** | |
| Amino Acid Sequence Length | 1195aa |
| COG | COG0085 - DNA-directed RNA polymerase, beta subunit/140 kD subunit |
| IMG Term | |
| Transmembrane Helices | No |
| Signal Peptide | No |
| Statistics | peptide |
| MyIMG Annotations | |
| **Pathway Information** | |
| Enzymes | |
| KO Term | |
| KEGG Pathway | |
| IMG Pathways | |

Evidence For Function Prediction

Related Links and Tools

NCBI BLAST
IMG Genome BLAST
SNP BLAST

BLAST against SNP Databases

BLAST against contigs and reads to look for SNP variations on this gene.
Please be sure to select the appropriate database.

```
>2001202539 DNA-directed RNA polymerase, beta subunit/140 kD
subunit
ATGAATCCTATTTTAGACGCGTTTTTTAAATTAGAAAGTGTAGTTAATTA
TCAGATTGATTCCATGAATTATTTTTATGCCTCAAAGAACAACCCTGACA
ATATAATGCAGCAGATAGTTGATGAGACAAAGATATCTGACGATGCAACC
CCCGGAGTTATAGATCTGGATCCCTCCAAAACCCGGGGCAAGGAAATAAA
AATATATTTTGGCCGGGAAAGGCAGAACGGAAAAGCCACAGGAGATCCAA
CAATATGGGTGGATAAACCAGAAATTAAGGAAGCTTCTGGAGCAGCAAAC
CAGATAACACCACAGGAAGCAAGGCTGAGGGATCTGAACTACATGGCTCC
TATAATGTTGCGTTTGCGTATAGTTGAGGATGGAATAGAGAGGGATGCAA
ATACAATAAAAATTGGGGAAATACCTGTGATGGTGCGGTCCAAGATATGC
```

Database: Acid Mine Drainage
E-value: 1e-2

```
QUERY           120   tgatgagacaaagatatctgacgatgcaacccccggagttatagatctggatccctccaa 179
amd_scaffold_15 34397 ............................................................ 34338
amd_scaffold_65 8827  g..c..a..c........c..t........a..t.....a........t..c..t..a.. 8886
XYG54996.b1     714   .....................................g...................... 655
XYG67286.g1     104   .....................................g...................... 163
XYG40789.g1     626   ....................................t..................t..... 567
XYG18804.b3     613   .....................................g...................... 554
XYG16991.g2     670                                        ...................t..... 645
XYG19962.b2     143   ....................................t..................t..... 202
XYG1649.y1      435   .....................................g...................... 376
```

SNP VISTA

Please be patient for applet to load

**Functionality**. Under the **Evidence for Function Prediction** on the **Gene Details** page there is a list of **Related Links and Tools**, including SNP BLAST. The sequence of a particular gene can be used (with or without upstream and downstream regions); alternatively the whole contig sequence can be used or you can replace the sequence with your favorite nucleotide sequence by pasting it in the window. The databases contain the reads generated for the corresponding metagenomes; by default the database corresponding to the selected gene is chosen. The raw BLAST output shows you whether there are any SNPs among the reads corresponding to this contigs (shown by letters, while identical nucleotides are replaced by dots). On the bottom of the page with raw BLAST output a button for running SNP VISTA [16] is available.

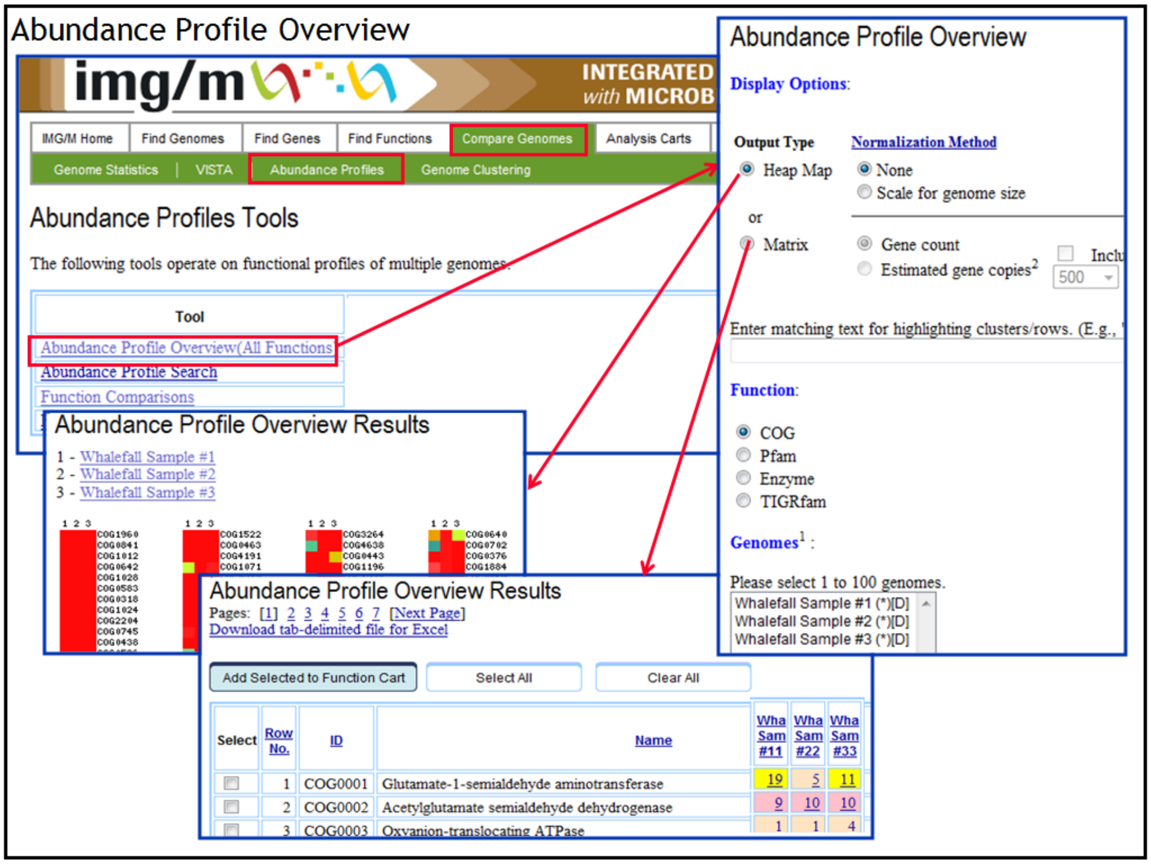
**Notes.** For some samples (AMO community, annamox bacteria, human and mouse gut microbiomes) there are no read databases. BLAST is run against all reads and contigs in the database and the output makes no distinction between the reads that were included in the contig being analyzed and other reads with sequence similarity to the contig.

# 4 Analysis Protein Family Relative Abundance

## 4.1 Abundance Profile Overview

**Purpose.** Examine relative abundance of all protein families (COGs and Pfams) and functional families (Enzymes) in metagenomes and isolate genomes.

**Navigation**: Compare Genomes → Abundance Profiles → Abundance Profile Overview



Abundance Profile Overview

Abundance Profiles Tools

The following tools operate on functional profiles of multiple genomes.

| Tool |
|---|
| Abundance Profile Overview(All Functions |
| Abundance Profile Search |
| Function Comparisons |

Abundance Profile Overview

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

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

1 2 3
COG1960
COG0841
COG1012
COG0642
COG1028
COG0583
COG0318
COG1024
COG2204
COG0745
COG0438

1 2 3
COG1522
COG0463
COG4191
COG1071

1 2 3
COG3264
COG4638
COG0443
COG1196

1 2 3
COG0640
COG0702
COG0376
COG1884

Abundance Profile Overview Results

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

| 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 |

**Functionality**. Select the type of format for displaying the results (“**Heat Map**” or “**Matrix**”), the type of protein/functional families (COG, Pfam, Enzyme), normalization method, and a set of metagenomes/isolate genomes.

For “**Heat Map**” output, the abundance of protein/functional families will be displayed as a heat map with red corresponding to the most abundant families. Each column on the map corresponds to a genome or metagenome, each row – to a family; mouse over each cell to see the count of a particular family in a genome/metagenome. Clicking on the id of the family displayed right to the column will add the corresponding family to the **Function Cart**; clicking on the cell will retrieve the list of genes assigned to this particular family in this genome or metagenome. By default the map is sorted by the abundance of families in the first sample, but can be resorted according to the abundance in other samples by clicking on the corresponding column header.

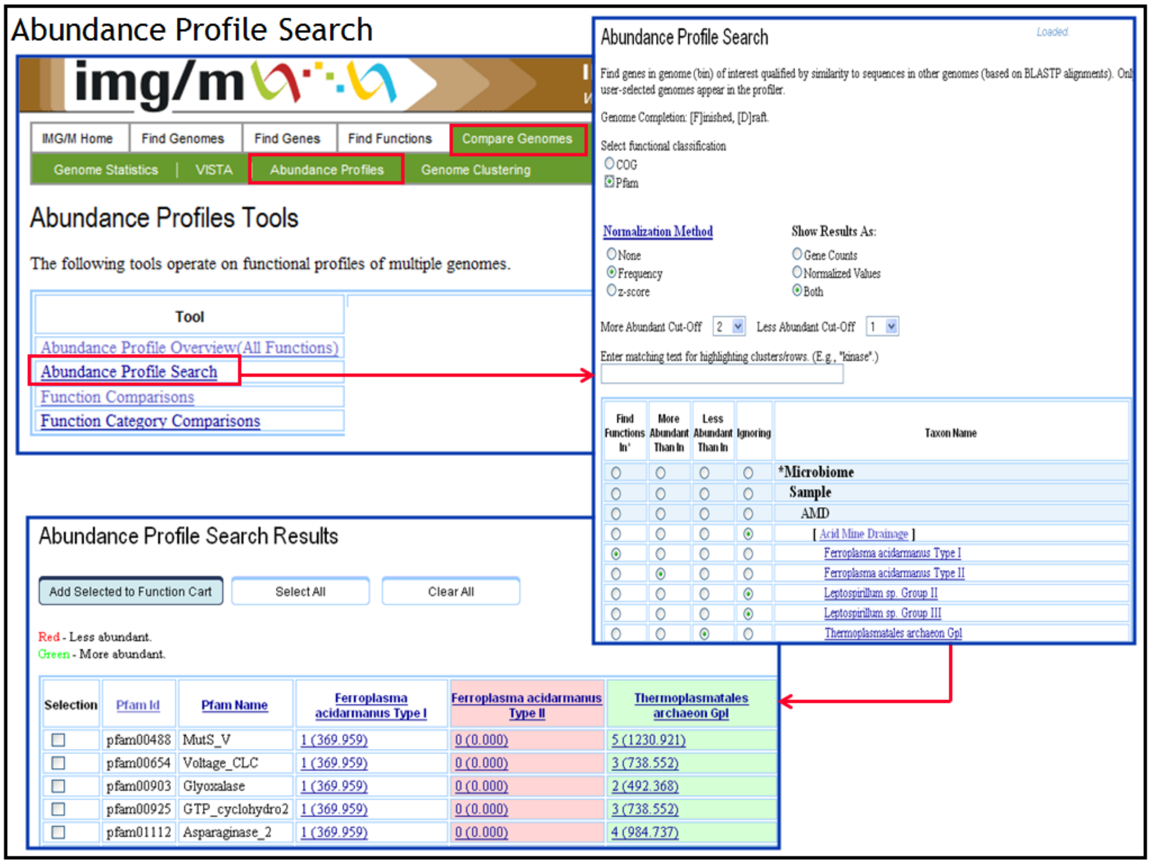
If the “**Matrix**” output is selected, the abundance of protein/functional families is displayed in a tabular format, with each row corresponding to a family and each cell containing the number of genes associated with a family for a specific genome or metagenome. Click on the cell in order to retrieve 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.

**Notes.** This analysis does not include the read depth coverage of each gene when counting family abundance. Beware when comparing high-complexity metagenomes with very low degree of assembly (e. g. soil) with low-complexity well-assembled metagenomes, such as AMD sample, since each gene in the latter may correspond to many reads.

## 4.2 Abundance Profile Search

**Purpose.** Selection of protein families (COGs and Pfams) in metagenomes and isolate genomes based on their relative abundance; similar to Phylogenetic Profiler for gene selection, but operates on protein families rather than individual genes.

**Navigation.** Compare Genomes → Abundance Profiles → Abundance Profile Search



Abundance Profile Search

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 |

Abundance Profile Search

Find genes in genome (bin) of interest qualified by similarity to sequences in other genomes (based on BLASTP alignments). Onl user-selected genomes appear in the profiler.

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

Select functional classification
○ COG
● Pfam

| Normalization Method | Show Results As: |
|---|---|
| ○ None | ○ Gene Counts |
| ● Frequency | ○ Normalized Values |
| ○ z-score | ● Both |

More Abundant Cut-Off 2 Less Abundant Cut-Off 1

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

| Find Functions In' | More Abundant Than In | Less Abundant Than In | Ignoring | Taxon Name |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | *Microbiome |
| ○ | ○ | ○ | ○ | Sample |
| ○ | ○ | ○ | ○ | AMD |
| ○ | ○ | ○ | ● | [ Acid Mine Drainage ] |
| ● | ○ | ○ | ○ | Ferroplasma acidarmanus Type I |
| ○ | ● | ○ | ○ | Ferroplasma acidarmanus Type II |
| ○ | ○ | ○ | ● | Leptospirillum sp. Group II |
| ○ | ○ | ○ | ● | Leptospirillum sp. Group III |
| ○ | ○ | ● | ○ | Thermoplasmatales archaeon Gpl |

Abundance Profile Search Results

Red - Less abundant.
Green - More abundant.

| Selection | Pfam Id | Pfam Name | Ferroplasma acidarmanus Type I | Ferroplasma acidarmanus Type II | Thermoplasmatales archaeon Gpl |
|---|---|---|---|---|---|
| □ | pfam00488 | MutS_V | 1 (369.959) | 0 (0.000) | 5 (1230.921) |
| □ | pfam00654 | Voltage_CLC | 1 (369.959) | 0 (0.000) | 3 (738.552) |
| □ | pfam00903 | Glyoxalase | 1 (369.959) | 0 (0.000) | 2 (492.368) |
| □ | pfam00925 | GTP_cyclohydro2 | 1 (369.959) | 0 (0.000) | 3 (738.552) |
| □ | pfam01112 | Asparaginase_2 | 1 (369.959) | 0 (0.000) | 4 (984.737) |

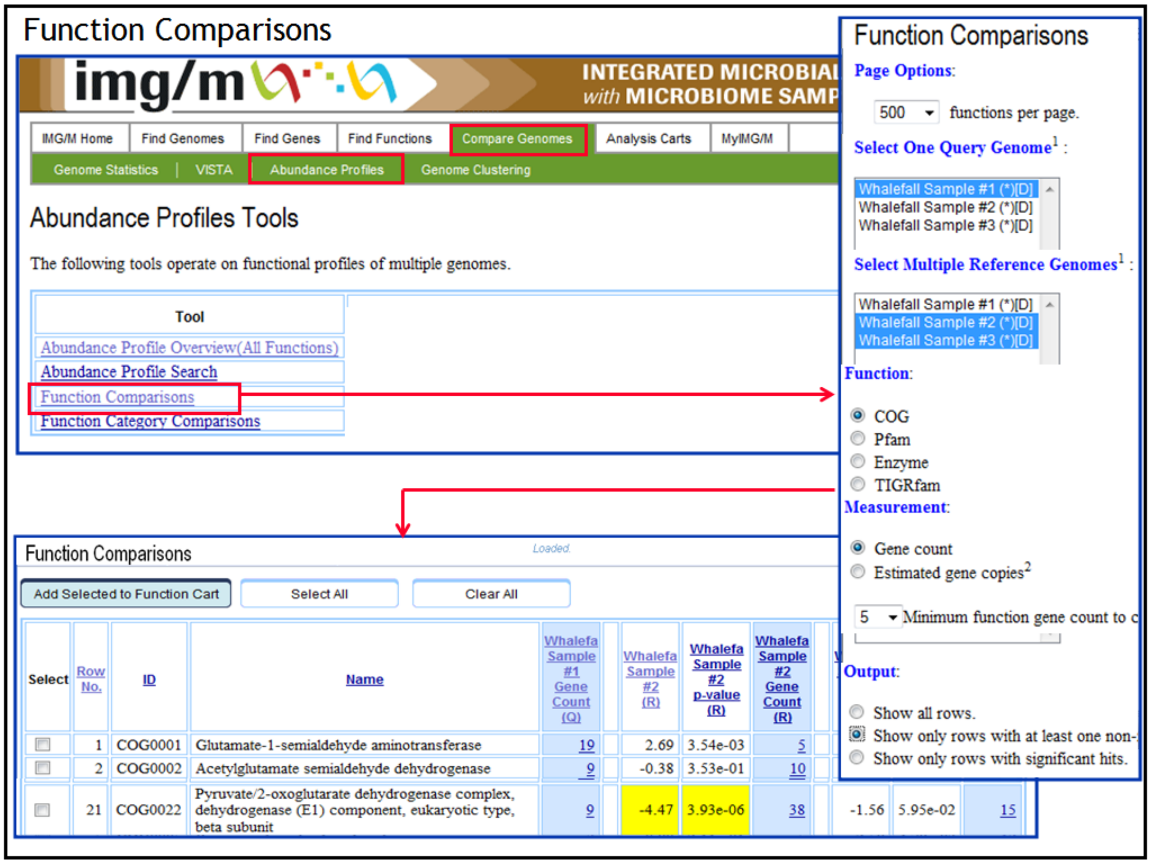
**Functionality**. Select the type of protein families (COG or Pfam), normalization method, and display of results. Abundance cut-offs can be set up for the genomes/metagenomes/bins of interest (e. g. – find all COGs in Ferroplasma Type I bin that are at least twice as abundant in this bin as in Ferroplasma Type II and are at least twice less abundant than in Thermoplasmatales archaeon). The families in the Results table can be selected and added to the Function Cart, while gene counts in the table are linked to the corresponding lists of genes, which can be also selected and added to the Gene Cart.

**Notes**. Abundance Profile Search does not take into account the degree of assembly of a metagenome/bin, i.e. the differences in read depth coverage between the genes and gene families in well assembled and poorly assembled metagenomes or bins. Beware when comparing poorly assembled metagenomes/bins with well assembled metagenomes/bins.

## 4.3 Function Comparison

**Purpose.** Comparison of a query metagenome/ isolate genome with one or several reference metagenomes/ isolate genomes, in terms of their relative abundance of protein families (COGs, Pfams, TIGRfams) and functional families (Enzymes), with estimates of the *statistical significance* of the observed differences

**Navigation.** Compare Genomes → Abundance Profiles → Function Comparisons



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]<br>Whalefall Sample #2 (*)[D]<br>Whalefall Sample #3 (*)[D]

Select Multiple Reference Genomes[1] :

Whalefall Sample #1 (*)[D]<br>Whalefall Sample #2 (*)[D]<br>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 |

**Functionality**. Select one query metagenome or genome, ***Q***, and one or several reference metagenomes or genomes, ***R***. Select the type of protein/functional families (COG, Pam, Enzyme, TIGRfam), the metric for comparison (count of genes or estimated gene copies, when read depths are available), and the type of comparison results to be included into the output, that is: (i) list all functions, even without any associated genes for the query and reference genomes/metagenomesl (ii) list the functions that have associated genes in one of the genomes/metagenomes; or (iii) list only the functions with significant differences between the query and refrence genomes/metagenomes.

The function comparison output lists for each function, ***F***: (i) the number of genes or estimated gene copies in the query genome/metagenome, ***Q***, associated with ***F***, and (ii) 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***. The computation of the D-score an p-values is further discussed below.

**Comparison of frequencies of occurrence of a protein family between a query (meta)genome and a reference (meta)genome.** Consider a query (meta)genome ***Q***, and a reference (meta)genomes ***R***. Given a protein family *f*, let $X_1$ and $X_2$ be the random variables representing the number of genes annotated with *F* in *Q* and *R*, respectively. Let $n_1$ and $n_2$ be the number of genes in *Q* and *R*, respectively, that are annotated with a protein family. Then $F_1=X_1/n_1$ and $F_2=X_2/n_2$ are the random variables representing the proportion of genes in *Q* and *R*, respectively, that are annotated with f. Then, the random variable that measures the difference between the abundances of the protein family *f* in *Q* and *R* is $F_1$-$F_2$.

Let the probabilities that a gene is annotated with protein family *f* be $p_1$ and $p_2$ in *Q* and *R*, respectively. Then, the distribution of $X_1$ and $X_2$ can be approximated by the Binomial distributions[1] $B(X_1;\ n_1,\ p_1)$ and $B(X_2;\ n_2,\ p_2)$, respectively. Recall that the binomial probability $B(X;\ n,\ p)$ is the probability that *X* out of *n* members of a sample display a property of interest. When the sample sizes $n_1$ and $n_2$ are large, the binomial distribution can be approximated by the normal distribution[2].

The null hypothesis in this scenario is:

$$H_0\textbf{:}\ F_1=F_2 \text{ or } F_1\text{-}F_2=0.$$

Under the null hypothesis, both distributions $B(X_1;\ n_1,\ p_1)$ and $B(X_2;\ n_2,\ p_2)$ are identical with $p_1=p_2=p$. Therefore, when $H_0$ is true, *p* can be approximated as follows:

$$p=(X_1+X_2)/(n_1+n_2).$$

The random variable being tested here is $F_D=F_1$-$F_2$. Under the null hypothesis, both $X_1$ and $X_2$ are binomially distributed with parameter *p*, and the expected value (mean) of $F_D$ is 0. The standard deviation (*SD*) of $F_D$ is given by

$$sqrt\ (var(F_1)+var(F_2)) = sqrt\ (p(1\text{-}p)/n_1 + p(1\text{-}p)/n_2) = sqrt\ (p(1\text{-}p)(1/n_1+1/n_2)).$$

The above result is obtained by using the standard expressions for variance and standard deviation for the binomial distribution. As an example, $var\ (F_1)=p_1(1\text{-}p_1)/n_1.$

Then the following variate, which we call the D-score

$$D = (F_1 – F_2) / sqrt\ (\ p(1\text{-}p)\ *\ (\ 1/n_1 + 1/n_2\ )\ )$$

is the *Z*-score of $F_D$ under the null hypothesis and is approximately normally distributed with mean zero and unit variance, N(0,1)[3].

The significance of individual D-scores can be computed using the standard normal distribution tables and the standard *Z*-test[1].

When the abundances of multiple protein families $f_1,\ f_2,\ ...,\ f_n$ are being tested simultaneously (in a functional category), consider the following variate:

$$D^* = \Sigma\ D_i\ /\ sqrt(n),\ i=1\ ...\ n.$$

[1] Freedman D, Pisani R, and Purves R, (1997) **Statistics**, Third Ed., W. W. Norton & Comp.
[2] Papoulis A. (1984) Probability, Random Variables, and Stochastic Processes, 2nd ed. New York: McGraw-Hill, 102-103.
[3] Hoel PG, Introduction to Mathematical Statistics (1971) Fourth Ed, Wiley.

Assuming that all the identically distributed $D_i$s above are independent, according to the Central Limit Theorem, $D^*$ is normally distributed with mean 0 and variance = square(sqrt(n))*var( $\Sigma$ $D_i/n$) = n*1/n = 1. Therefore, $D^*=N(0,1)$, too. Then using the tables for the Z-distribution, the p-value corresponding to a given $D^*$ value can be computed.

The null hypothesis is rejected when the absolute value of $D^*$ is greater than 1.96, which is equivalent to a p-value less than 0.05. Note that an absolute value of $D^*$ greater than 2.33 is equivalent to a p-value less than 0.01.

**Multiple Hypothesis Testing Correction.** When multiple hypotheses are being tested simultaneously, the likelihood of false predictions, and hence the number of type-I errors increases significantly. In order to remove the effect of testing multiple hypotheses simultaneously, 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 a given FDR 〈, the hypotheses are first ordered according to increasing p-values. For n hypotheses, let this order be $H_{(0)}$, $H_{(1)}$ ... $H_{(n)}$ and the respective ordered p-values be $p_{(0)}$, $p_{(1)}$ ... $p_{(n)}$. Then, the $k^{th}$ hypothesis $H_{(k)}$ is rejected (i.e., the alternate hypothesis is accepted as a statistically significant discovery) if $p_{(k)}$ <= k〈/n.

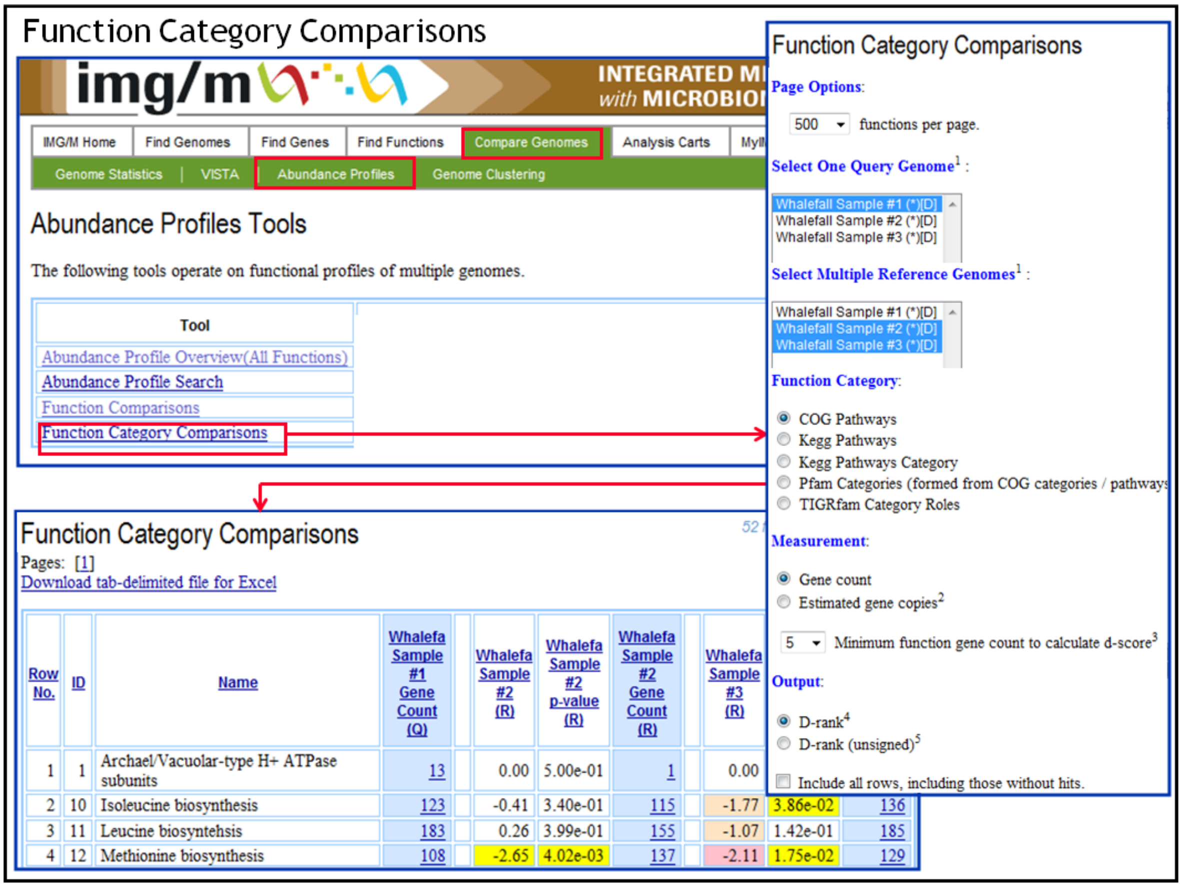
In displaying the result of a function comparison, the FDR correction is used for highlighting significant differences: non-significant differences or differences that are not significant for the number of genes annotated with protein families (see above) are not highlighted. An FDR with 〈 = 0.05 is used for selecting p-values at or better than $p_{(k).}$ The number of functions, n, represents the number of hypotheses tested.

[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.

## 4.4 Functional Category Comparison

**Purpose.** Comparison of a query metagenome/ isolate genome with one or several reference metagenomes/ isolate genomes, in terms of the relative abundance of the genes assigned to different functional categories (COG Pathway, KEGG Pathway, KEGG Pathway Category, Pfam Category, TIGRfam sub-roles), with estimates of the *statistical significance* of the observed differences

**Navigation.** Compare Genomes → Abundance Profiles → Functional Category Comparisons



Function Category Comparisons

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 Category 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 Category**:

● COG Pathways
○ Kegg Pathways
○ Kegg Pathways Category
○ Pfam Categories (formed from COG categories / pathways
○ TIGRfam Category Roles

**Measurement**:

● Gene count
○ Estimated gene copies[2]

5 ▾ Minimum function gene count to calculate d-score[3]

**Output**:

● D-rank[4]
○ D-rank (unsigned)[5]

□ Include all rows, including those without hits.

Function Category Comparisons

Pages: [1]
Download tab-delimited file for Excel

| 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) | | Whalefa Sample #3 (R) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Archael/Vacuolar-type H+ ATPase subunits | 13 | | 0.00 | 5.00e-01 | 1 | | 0.00 | | |
| 2 | 10 | Isoleucine biosynthesis | 123 | | -0.41 | 3.40e-01 | 115 | | -1.77 | 3.86e-02 | 136 |
| 3 | 11 | Leucine biosyntehsis | 183 | | 0.26 | 3.99e-01 | 155 | | -1.07 | 1.42e-01 | 185 |
| 4 | 12 | Methionine biosynthesis | 108 | | -2.65 | 4.02e-03 | 137 | | -2.11 | 1.75e-02 | 129 |

**Functionality**. Select one query metagenome or genome, ***Q***, and one or several reference metagenomes or genomes, ***R***. Select the type of functional category (COG Pathway, KEGG Pathway, KEGG Pathway Category, Pfam Categories, TIGR Role Categories), the metric for comparison (count of genes or estimated gene copies, if read depths for the genes are available), and the type of statistical significance estimate (D*-rank or D*-rank unsigned). Pfam Categories is a classification of Pfams based on their mapping to COGs through the genes assigned to both types of protein families; only those Pfams that can be unambiguously mapped onto COGs are included in classification, which mirrors COG Functional Categories and COG Pathways.

The function category comparison output lists for each function category ***F***: (i) the number of genes or estimated gene copies in the query genome/metagenome, ***Q***, associated with ***F*** and (ii) for each reference genome/metagenome, ***R***, the number of genes or estimated gene copies in ***R*** associated with ***F*** together with the D*-rank and p-value associated with the comparison of ***Q*** to ***R***. The computation of the D*-rank an p-values is further discussed below.

The comparison result includes an assessment of statistical significance of the relative frequencies of the genes assigned to different functional categories in terms of **D*-rank** which represents a normalization ranking of each pair wise comparison. D*-rank is calculated by adding the **D-scores** of all protein families (see section 3.3 above) assigned to a certain functional category normalized by the square root of the number of these categories including those with no genes assigned (zero hits). In calculating the D*-rank, we use the observation that the sum of normally distributed values is normally distributed. Similar to D-scores for individual protein families, the null hypothesis assumes that the probability of occurrence of the genes assigned to a certain functional category is the same for both metagenomes. Under the assumption of the null hypothesis D-rank values are approximately normally distributed with mean zero and unit variance, N(0,1). As in the case of D-scores of individual protein families, the null hypothesis of equal probabilities of functional categories is rejected if absolute D-rank value is greater than 1.96 at $p<0.05$ and at $p<0.01$ if absolute D*-rank value is greater than 2.33. In addition to signed or unsigned D*-rank values counts of genes assigned to a functional category in each metagenome is displayed.

**Notes**. In the computation of D*-ranks, the D-scores invalid due to low counts of genes assigned to a protein family in at least one metagenome are not excluded from the calculation but rather assigned a zero value, which represents a penalty for low gene counts. The result of this penalty is that the null hypothesis will be harder to reject as the proportion of invalid D-scores increases thus avoiding overestimation of statistical significance of the differences of gene counts in sparsely populated functional categories.

**Calculations.**

$f1 = x_1/n_1$ = frequency of functional occurrence in query group.
$f2 = x_2/n_2$ = frequency of functional occurrence in reference group.
$p = (x_1 + x_2) / (n_1 + n_2)$ = probability of occurrence.
$q = 1 - p$ = probability of non-occurrence.
$d\_score = ( f_1 - f_2 ) / sqrt( p*q * ( 1/n_1 + 1/n_2 ) )$

$$D^*\text{-}rank = \sum_{I=1}^{N} D\text{-}score_i/\sqrt{N}$$

Where:

$x_1$ = count of a given function in query group.
$x_2$ = count of a given function in reference group.
$n_1$ = total counts of all function occurrences in query group.
$n_2$ = total counts of all function occurrences in reference group.