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
Copyright 2008 The Regents of the University of California
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

©2008 The Regents of the University of California
This document was prepared by:

Victor M. Markowitz*
Natalia N. Ivanova**
Iain Anderson**
Athanasiou 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 ................................................................................................................................. 1
   1.1 Definitions ......................................................................................................................... 1
   1.2 Data Processing ............................................................................................................... 2
   References .............................................................................................................................. 3

2 IMG/M Overview ....................................................................................................................... 4
   2.1 IMG/M Data Content ....................................................................................................... 4
   2.2 IMG/M Data Analysis ...................................................................................................... 5

3 Analysis of Community Diversity & Abundance ................................................................. 7
   3.1 Phylogenetic Distribution of Genes ................................................................................. 7
   3.2 Phylogenetic Marker COGs ............................................................................................. 10
   3.3 SNP BLAST and SNP VISTA .......................................................................................... 11

4 Analysis Protein Family Relative Abundance .................................................................... 13
   4.1 Abundance Profile Overview ......................................................................................... 13
   4.2 Abundance Profile Search ............................................................................................... 14
   4.3 Function Comparison ........................................................................................................ 15
   4.4 Functional Category Comparison ...................................................................................... 18
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.

![Diagram of bins](image)

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 any analysis cannot be overestimated.
References


6. **Fgenesb gene prediction**: [http://www.softberry.com](http://www.softberry.com)


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.

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 population 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.

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.
**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.
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

**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](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

**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

**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

**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

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/metagenomes; (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 reference 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 and 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:

\[ H_0: F_1 = F_2 \text{ or } F_1 - F_2 = 0. \]

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

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

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

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

\[ D = (F_1 - F_2) / \sqrt{p(1-p) * (1/n_1 + 1/n_2)} \]

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:

\[ D^* = \sum D_i / \sqrt{n}, i=1 \ldots n. \]

---


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 $= \text{square}(\sqrt{n}) \cdot \text{var}(\Sigma D_i/n) = n^*1/n = 1$. Therefore, $D^* \sim N(0,1)$, 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 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 $\zeta$, 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 $p(k) \leq \zeta/n$.

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 $\zeta = 0.05$ is used for selecting p-values at or better than $p(k)$. The number of functions, n, represents the number of hypotheses tested.

---

4.4 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

**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 \textbf{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, \( \text{N}(0,1) \). 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.

\textbf{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.

\textbf{Calculations}.

\[ f_1 = \frac{x_1}{n_1} = \text{frequency of functional occurrence in query group.} \]
\[ f_2 = \frac{x_2}{n_2} = \text{frequency of functional occurrence in reference group.} \]
\[ p = \frac{(x_1 + x_2)}{(n_1 + n_2)} = \text{probability of occurrence.} \]
\[ q = 1 - p = \text{probability of non-occurrence.} \]
\[ d\text{-score} = \frac{(f_1 - f_2)}{\sqrt{pq \cdot (1/n_1 + 1/n_2)}} \]
\[ N \]
\[ D^*\text{-rank} = \sum_{l=1}^{N} \frac{D\text{-score}}{\sqrt{N}} \]

Where:

\[ x_1 = \text{count of a given function in query group.} \]
\[ x_2 = \text{count of a given function in reference group.} \]
\[ n_1 = \text{total counts of all function occurrences in query group.} \]
\[ n_2 = \text{total counts of all function occurrences in reference group.} \]