The integrated microbial genomes (IMG) system: an expanding comparative analysis resource

Victor M. Markowitz[1], I-Min A. Chen[1], Krishna Palaniappan[1], Ken Chu[1], Ernest Szeto[1], Yuri Grechkin[1], Anna Ratner[1], Iain Anderson[2], Athanasios Lykidis[2], Konstantinos Mavromatis[2], Natalia N. Ivanova[2] and Nikos C. Kyrpides[2]

[1]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, USA, [2]Genome Biology Program, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

ABSTRACT

The Integrated Microbial Genomes (IMG) system serves as a community resource for comparative analysis of publicly available genomes in a comprehensive integrated context. IMG contains both draft and complete microbial genomes integrated with other publicly available genomes from all three domains of life, together with a large number of plasmids and viruses. IMG provides tools and viewers for analyzing and reviewing the annotations of genes and genomes in a comparative context. Since its first release in 2005, IMG's data content and analytical capabilities have been constantly expanded through regular releases. Several companion IMG systems have been set up in order to serve domain specific needs, such as expert review of genome annotations. IMG is available at http://img.jgi.doe.gov.

INTRODUCTION

The Integrated Microbial Genomes (IMG) system serves as a community resource for comparative analysis of publicly available genomes in a comprehensive integrated context. IMG employs NCBI's RefSeq resource (1) as its main source of public genome sequence data, and "primary" annotations consisting of predicted genes and protein products. IMG genomes are classified using NCBI's (domain, phylum, class, order, family, genus, species, strain) taxonomy. For every genome, IMG records its primary genome sequence information from RefSeq including its organization into chromosomal replicons (for finished genomes) and scaffolds and/or contigs (for draft genomes), together with predicted protein-coding sequences (CDSs), some RNA-coding genes, and protein product names that are provided by the genome sequence centres. Every genome included in IMG is associated with metadata attributes, available from GOLD (2).

IMG's data integration pipeline computes CRISPR repeats (3), signal peptides using SignalP (4) and transmembrane helices using TMHMM (5), and associates genes with "secondary" functional annotations and lists of related (e.g., homolog, paralog) genes. IMG generated annotations consist of protein family and domain characterizations based on COG clusters and functional categories (6), Pfam (7), TIGRfam and TIGR role categories (8), InterPro domains (10), Gene Ontology (GO) terms (11), and KEGG Ortholog (KO) terms and pathways (9)[1]. Genes are further characterized using an IMG native collection of generic (protein cluster-independent) functional roles called IMG terms that are defined by their association with generic (organism-independent) functional hierarchies, called IMG pathways (12). IMG terms and pathways are specified by domain experts at DOE-JGI as part of the process of annotating specific genomes of interest, and are subsequently propagated to all the genomes in IMG using a rule based methodology (13).

Gene relationships in IMG are based on sequence similarities computed using NCBI BLASTp for protein coding genes and BLASTn for RNA genes). For each gene, IMG provides lists of related (e.g., candidate homolog, paralog, ortholog) genes that can be filtered using percent identity, bit score, and more stringent E-values, or using metadata attributes such as phenotype and habitat[1].

IMG has expanded regularly its collection of genomes and aims at improving gradually the coverage and consistency of its functional annotations. IMG's analytical tools have been continuously enhanced in terms of their usability, analysis flow, and performance. Several companion IMG systems have been set up in order to serve domain specific needs, including expert review of

[1] For more details see the Data Processing section of About IMG at: http://img.jgi.doe.gov/w/doc/dataprep.html.

genome annotations prior to their publication (IMG/ER: http://img.jgi.doe.gov/er), teaching courses and training in microbial genome analysis (IMG/EDU: http://img.jgi.doe.gov/edu), and analysis of genomes related to the Human Microbiome Project (IMG/HMP: http://www.hmpdacc-resources.org/img_hmp)[2]. We review below IMG's data content and analysis tool extensions since the last published report on IMG (14).

IMG DATA CONTENT GROWTH

IMG's initial collection of **296** bacterial, archaeal, and eukaryotic genomes in its first version (March 2005) grew to **825** genomes in IMG 2.3 (September 2007) and then more than doubled to **1,655** genomes in IMG 2.9 (August 2009). In addition, IMG 2.9 includes 2,490 virus genomes and 970 plasmids that did not come from a specific microbial genome sequencing project, bringing its total genome content to **5,115** genomes with over 6.5 million genes[3].

Prior to their inclusion into IMG, RefSeq genomes undergo a review process. First, the taxonomic classification for genomes and the names and host information for plasmids are reviewed. In particular plasmid names are curated by adding strain names to organism name when available from publications or other sources, and plasmid sequences are added to host genome sequences when appropriate. Next, missing RNAs are identified using tRNAS-can-SE-1.23 (15) for tRNAs, RNAmmer (16) for rRNAs, and Rfam (17) and INFERNAL (18) for small RNAs. Finally, for genomes without any functional annotation in RefSeq, protein product names are assigned to genes using the procedure described in (13): such annotations are performed only by request, for example from a centre such as HMP-DACC (http://www.hmpdacc.org/).

The functional annotations generated by IMG's data integration pipeline are regularly reviewed by scientists in JGI's Genome Biology Program with the goal of improving their coverage. Following such a review, the KEGG collection of pathways in IMG has been reorganized and updated using the enhanced collection of KEGG resources, including KEGG Orthology (KO) terms and KEGG pathway modules (9). The association of KEGG pathways with IMG genomes is based on the assignment of KEGG Orthology (KO) terms to IMG genes via a mapping of IMG genes to KEGG genes. The MetaCyc collection of pathways (19) has been also included into IMG, whereby the association of MetaCyc pathways with IMG genomes is based on correlating enzyme EC numbers in MetaCyc reactions with EC numbers associated with IMG genes via KO terms.

Two interactive reports regarding the KO term distribution in IMG across protein families, genomes and paralog clusters, are provided for assessing the consistency of protein family annotations in IMG. For a specific (query) KO term, the first report lists: (i) the number of genes associated with the query KO term and the number of genomes that have genes associated with this KO term; (ii) the *average number of genes* associated with the query KO term per genome, whereby this metric helps identify KO terms that were assigned to multiple genes in the same genome either by mistake or because these terms correspond to sequence similarity-based families rather than function-based groups; (iii) the number of genes associated with the query KO term that belong to paralog clusters, whereby this metric indicates the likelihood of incorrect annotations due to the presence of paralogs; and (iv) the number of genes associated with the query KO term and that have a paralog annotated with the same KO term, whereby this number helps identifying incorrectly annotated paralogous genes.

The second report lists for each unique (COG, Pfam, TIGRfam) combination: (i) the number of genes associated with the query KO term and this combination; (ii) the number of genes associated with this combination and a KO term different from the query KO term, including genes associated with multiple KO terms and a query KO term as one of them; (iii) the number of genes associated with this combination and a KO term different from the query KO term, and not associated with the query KO term; and (iv) the number of genes associated with this combination and not associated with any KO term.

The gene correlations computed by IMG's data integration pipeline have been extended from pair-wise relationships to include gene fusions and cassettes. A fused gene (*fusion*) is defined as a gene that is formed from the composition (fusion) of two or more previously separate genes (component genes). The identification of fusions employs well established methods based on pairwise similarities

[2] The Human Microbiome Project is part of NIH's Roadmap for Medical Research: http://nihroadmap.nih.gov/hmp/.
[3] A Content History link on IMG's home page provides an overview of its content growth.

between genes (20)[4]. Genes, such as transposases and integrases, pseudogenes, and genes from draft genomes are not considered as putative fusion components in order to avoid false positives caused by gene fragmentation.

A chromosomal neighbourhood, also known as *chromosomal cassette*, is defined as a stretch of genes with intergenic distance smaller or equal to 300 base pairs (21), whereby the genes can be on the same or different strands. Chromosomal cassettes with a minimum size of two genes common in at least two separate genomes are defined as *conserved chromosomal cassettes*. The identification of common genes across organisms is based on three gene clustering methods, namely participation in COG, Pfam, and IMG ortholog clusters. The computation of gene cassettes and their support for context analysis in IMG is described in detail in (22).

IMG DATA ANALYSIS TOOL EXTENSIONS

Genome data analysis in IMG consists of operations involving genomes, genes, and functions which can be selected, explored individually, and compared. The composition of analysis operations is facilitated by gene and function "carts" that handle lists of genes and functions, respectively.

Data Selection Tools

Genomes, genes and functions can be selected using browsers and search tools. Browsers allow users to select genomes and functions organized as alphabetical lists or using domain specific hierarchical classifications. Keyword search tools allow identifying genomes, genes, and functions of interest using a variety of selection filters. Genomes can be also selected using a search tool which allows specifying conditions involving metadata attributes, while genes can be also selected using BLAST search tools against various datasets.

IMG's data selection tools have been extended in order to improve their efficiency and usability. In particular genomes can be selected using a new phylogenetic tree based "**Genome Browser**", a geographical location based project map, and a metadata based classification, as illustrated in Figure 1. The phylogenetic tree based "**Genome Browser**" starts with a display of the three genome domains, as illustrated in Figure 1(i), which can be expanded using open/close icons available at each level of the tree, as illustrated in Figure 1(ii). Genomes can be selected either individually or in groups using the green dot "select all" icons available at each level of the tree. For example, clicking the "select all" (green dot) icon associated with *Crenarchaeota*, as illustrated in Figure 1(ii), will both expand the sub-tree under this phylum down to individual genomes and select all these genomes, as illustrated in Figure 1(iii). Genomes can be unselected (cleared) either individually or in groups using the red dot "clear all" icons available at each level of the tree.

The "**Genome by Metadata**" link on IMG's home page provides access to a classification of the archaeal, bacterial and eukaryotic genomes by several metadata attributes, as illustrated in Figure 1 (iv). Note that only a subset of the metadata attributes available in IMG are provided, namely attributes associated with controlled vocabularies of less than ten values, while additional attributes are available in "**Genome Search**", as illustrated in Figure 1(v). The metadata attributes and values are taken from GOLD (2) and reflect the continuously increasing level of information collection and curation in this resource.

Individual genomes can be explored using the "**Organism Details**" page which includes information on the organism together with various genome statistics of interest, such as the number of genes that are associated with KEGG, COG, Pfam, InterPro or enzyme information. Individual genes can be analyzed using the "**Gene Details**" page which includes Gene Information, Protein Information, and Pathway Information tables, evidence for functional prediction, COG, Pfam, and pre-computed homologs. New graphical viewers, such as graphical displays of the distribution of genes associated with COG, Pfam, TIGRfam, and KEGG for each genome, have been added to "**Organism Details**" and "**Gene Details**" in order to facilitate genome and gene exploration. Individual functional categories, such as KEGG Orthology terms and pathways, MetaCyc pathways, can be explored using functional category specific browsers.

Several new IMG tools allow users to search and explore gene cassette information. A chromosomal cassette involving a specific (query) gene can be examined using a "**Chromosomal Cassette Details**" page available via the "**Gene Information**" section of "**Gene Details**" for that gene.

[4] Fusion computation is described at: http://img.jgi.doe.gov/w/doc/fusions.html.

This page provides information on the protein clusters (e.g., COGs) of all the genes in the cassette, as well as information on other cassettes that share at least two protein clusters with the cassette that includes the query gene. Gene cassettes can be searched using "**Cassette Search**" and "**Phylogenetic Profiler for Gene Cassettes**". "**Cassette Search**" allows users to find genes that are part of chromosomal cassettes involving specific protein clusters, as illustrated in Figure 2(i), where the search involves COG clusters. By default, the search is carried out across all the genomes in IMG, with various filters provided for limiting the search to specific genomes. The result of "**Cassette Search**" consists of genes that satisfy the search condition, together with the identifiers of the cassettes they are part of, their associated protein cluster identifiers and names, and their genomes, as illustrated in Figure 2(ii). Cassette identifiers provide links to the "**Chromosomal Cassette**" details page, as illustrated in Figure 2(iii).

The genomes that result from browsing and search operations are displayed as a list from which they can be selected and saved for further analysis. The genes and functions that result from search operations are displayed as lists from which genes and functions can be selected for inclusion into the "**Gene Cart**" and "**Function Cart**", respectively.

Comparative Analysis Tools

IMG comparative analysis tools allow comparing genomes in terms of gene content, functional and metabolic capabilities, and sequence conservation.

Genomes can be compared in terms of ***gene content*** using the "**Phylogenetic Profiler**" tool which allows users to identify genes in a query genome in terms of presence or absence of homologs in other genomes. This tool can be used, for example for finding *unique* genes in the query genome with respect to other genomes of interest. The "**Phylogenetic Profiler for Gene Cassettes**" extends its counterpart for single genes by allowing users to find genes that are part of a gene cassette in a query genome as well as part of related (conserved part of) gene cassettes in other genomes, as illustrated in Figure 2(iv). The result of such a search includes a summary, as shown in the left side pane of Figure 2(v), and a details part that displays groups of collocated genes in each chromosomal cassette in the query genome that satisfy the search condition, as illustrated in Figure 2(v). The conserved part of a chromosomal cassette involving an individual gene in the query genome can be examined using the links provided in the "**Conserved Neighbourhood Viewer Centred on this Gene**" column of results table, as shown in Figure 2(vi). More details on context analysis based on IMG's gene cassettes can be found in (22).

The gene content of a genome can be examined from an *evolutionary* point of view using tools available as part of a genome's "**Organism Details**". The "**Phylogenetic Distribution of Genes**" provides a glimpse into the evolutionary history of the genes in a genome based on the distribution of best BLAST hits of its protein-coding genes. The genes that were likely vertically inherited are expected to have higher sequence similarity to the genes in the genomes within the same taxonomic group, while those horizontally transferred may have their best BLAST hits to the genes in distantly related organisms. Since this tool considers best BLAST hits and does not perform phylogenetic tree reconstruction and analysis, the results can be used as a first approximation of the evolutionary history of the genes and require manual analysis to establish whether the genes of interest were indeed horizontally transferred. The phylogenetic distribution of best BLAST hits of protein-coding genes in a selected genome is displayed as a histogram, as shown in Figure 3(i); counts correspond to the number of genes that have best BLASTp hits to proteins of other genomes in a specific phylum or class with more than 90% identity (right column), 60-90% identity (middle column) and 30-60% identity (left column). The phylogenetic distribution of best BLAST hits can be further projected onto the families in a phylum/class. Gene counts in the histogram are linked to the lists of genes in the selected genome that have best BLAST hit in a certain phylum/class with specified percent identity. The genes in the table can be selected and added to "**Gene Cart**" or analyzed through the corresponding "**Gene Details**".

"**Putative Horizontally Transferred Genes**", also available as part of a genome's "**Organism Details**", allows users to explore genes in a query genome that are likely horizontally transferred from genomes in phylogenetic groups that are different than the group the query genome belongs to. Putative horizontally transferred genes are defined as genes that have best hits (best bitscores) to genes that don't belong to the phylogenetic group of the query genome. In this calculation we use not only the best hit (i.e. the hit with the best bitscore) but all the hits that have bitscore equal or greater than 95% of the best hit. For a query genome, such as *Methanosaeta thermophila PT*, two lists of

genes are provided, as illustrated in Figure 3(ii). The first list consists of genes with best hits (best bit score) to genes of genomes within a phylogenetic group (domain, phylum, class, etc.) that is different than the analogous group *the query genome* belongs to. For example, as an archaeal genome, *Methanosaeta thermophila PT* has 228 genes with best hits to bacterial genomes, 17 genes with best hits to eukaryotic genomes, and 1 gene with best hits to viral genomes. These genes may be horizontally transferred genes from bacterial, eukaryotic, or viral genomes, respectively. The second lists consists of genes with best hits to genomes within a phylogenetic group (domain, phylum, class, etc.) that is different than the analogous group the query genome belongs to, and no hits to genes of genomes within the same phylogenetic group (domain, phylum, class, etc.) as the group the query genome belongs to. For example, *Methanosaeta thermophila PT* has 2 genes with best hits to bacterial genomes and no hits to other archaeal genomes, as illustrated in Figure 3(iii), with a higher likelihood of being horizontally transferred from bacterial genomes.

Genomes can be compared in terms of ***functional capabilities*** using a number of functional profile tools. The "**Abundance Profile Overview**" allows users to compare the relative abundance of protein families (COGs, Pfams, TIGRfams) and functional families (enzymes) across selected genomes, as illustrated in Figure 4(i) where the *T. volcanium* and *T. Acidophilum* genomes are compared in terms of enzymes assigned to their genes. The abundance of protein/functional families is displayed either as a heat map or a matrix, as illustrated in Figure 4(ii), where each column corresponds to a genome, and each row corresponds to a family. The abundance of protein/functional families is displayed either as a heat colour map with red corresponding to the most abundant families, or in a tabular format, where each cell contains the number of genes associated with a family for a specific genome. Cells in the heat map and matrix are linked to the list of genes assigned to a particular family in a genome. Families of interest can be selected for inclusion into the "**Function Cart**". The results in matrix format can be exported to a tab-delimited Excel file. The functional capabilities of genomes can be also compared using the "**Function Profile**", which is a selective version of the "**Abundance Profile Overview**", with functions of interest first selected with the "**Function Cart**". The "**Function Profile**" result is displayed in a matrix format, as illustrated Figure 4(iii), similar to the matrix display for "**Abundance Profile Overview**" results.

The metabolic capabilities of genomes can be analyzed using functional profile tools applied on enzymes (e.g., the enzymes involved in a pathway of interest) together with a tool for finding "missing" enzyme that are marked by a null abundance in the function profile result. Such a null abundance for an specific "missing" enzyme leads to the "**Find Candidate Genes for Missing Function**" tool, as illustrated in Figure 4(iv), which allows users to search for candidate genes that could be associated with this missing enzyme either via KO terms or homolog/ortholog genes associated with it. The result of the search for candidate genes, illustrated in Figure 4(v), consists of a list of genes that can be selected and included into the "**Gene Cart**" and further examined using various tools, such as gene neighbourhood analysis and multiple sequence alignment tools.

Sequences of genomes can be compared using VISTA tools (23) and a "**Dotplot**" tool. Users can select an organism from a predefined list in order to invoke the VISTA browser that can be then employed for examining the sequence conservation of closely related organisms in IMG. "**Dotplot**", a recent addition to IMG's comparative analysis toolkit, employs the program *Mummer* to generate dotplot diagrams between two genomes, whereby nucleotide sequences are used for genomes with fairly similar sequences and protein sequences are used for genomes with less similar nucleotide sequences.

IMG FAMILY OF SYSTEMS

The initial IMG system has expanded into a family of four related systems covering two application domains: microbial genome analysis (IMG, IMG ER) and metagenome analysis (IMG/M, IMG/M ER).

The "**Expert Review**" version of IMG (IMG/ER) allows individual scientists or groups of scientists to review and curate the functional annotation of microbial genomes in the context of IMG's public genomes. Scientists include their genome datasets into IMG ER prior to their public release either with their original annotations or with annotations generated by IMG's annotation pipeline (13). IMG ER provides tools for identifying and correcting annotation anomalies, such as dubious protein product names, and for filling annotation gaps detected using IMG's comparative analysis tools, such as genes that may have been missed by gene prediction tools or genes without predicted functions (24). The development of the IMG ER tools was driven by and applied to the genome analysis and

curation needs of over 150 microbial genomes, such as *Halothermothrix orenii* (25). In addition to individual genome reviews, the annotations of a group of 56 Genomic Encyclopedia for Bacteria and Archaea (GEBA) genomes (http://www.jgi.doe.gov/programs/GEBA/pilot.html) were revised by JGI scientists using IMG ER (26). Gene annotations that result from expert review and curation are captured in IMG ER as so called "MyIMG" annotations associated with individual scientist or group accounts. Genomes curated with IMG ER are included into Genbank either as new submissions or as revisions of previously submitted datasets, thus contributing to a coordinated improvement of the public genome data resources.

The "**Integrated Microbial Genomes with Microbiome Samples**" (IMG/M) system provides support for the comparative analysis of metagenomic sequences generated with various sequencing technology platforms and data processing methods in the context of the reference isolate genomes from IMG. IMG/M's analysis tools extend IMG's comparative analysis tools with metagenome-specific analysis tools (27). Similar to IMG ER, an "**Expert Review**" version of IMG/M (IMG/M ER) provides support for annotation review and curation of metagenome datasets prior to their public release.

IMG HMP is an auxiliary resource based on IMG focusing on analysis of genomes related to the Human Microbiome Project (HMP) in the context of all publicly available genomes in IMG. IMG-HMP is part of the HMP Data Analysis and Coordination Center (DACC) funded by the National Institutes of Health (http://www.hmpdacc.org/).

FUTURE PLANS

IMG's genome sequence data content is maintained through regular updates from RefSeq and other public sequence data resources. IMG's functional annotations are gradually extended by including annotations from systems, such as SEED (http://www.theseed.org/wiki/Home_of_the_SEED), or by providing links to systems such as CMR (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi), thus providing extensive corroboration of annotations from multiple microbial genome data resources.

IMG has been recently extended to include protein expression data from a recent *Arthrobacter chlorophenolicus* study (28). Protein expression studies for a genome of interest are provided via the genome's "**Organism Details**", whereby each study is associated with the number of expressed genes, observed peptides, and a list of experiments/samples. The description for each sample consists of the experimental conditions and provides a link to the protein expression data for the sample organized per expressed gene. For each expressed gene, the number of observed peptides leads to the peptide details page, where the peptide sequences are displayed aligned on the gene's protein sequence. For an expressed gene, the "**Protein Information**" section of its "**Gene Detail**" provides a link to a "**Proteomic Data**" page which displays the list of experiments/ samples involving the expressed gene and the peptides observed for the expressed gene as part of each experiment. We plan to follow a similar strategy for including into IMG results from microarray experiments, as well as information on transcriptional regulatory binding sites.

In order to facilitate the exploration of a rapidly increasing number of genomes, genes, and annotations, IMG will be extended with pangenomes (29), where a pangenome represents the sum of all the genes present in the genomes of different strains belonging to a given species. Pangenome analysis tools and viewers will allow users to explore individual pangenomes and compare pangenomes and genomes.

ACKNOWLEDGEMENTS

We thank Philip Hugenholtz, Alla Lapidus, Amrita Pati, Sean Hooper, and Inna Dubchak for their contribution to the development and maintenance of IMG. The work of JGI's production, cloning, sequencing, assembly, finishing and annotation teams is an essential prerequisite for IMG. Eddy Rubin and James Bristow provided, support, advice and encouragement throughout this project. The work presented in this paper was supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

1. Pruitt, K.D., Tatusova, T., Maglott, D.R. (2007) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acid Research* **35**: D61-D65.
2. Liolios, K., Mavrommatis, K., Tavernarakis, N., and Kyrpides, N. (2008) The genomes online database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **36**, D475-D479.
3. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., Hugenholtz, P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**: 209.
4. Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. *Nature Protocols* **2**, 953-971.
5. Moller, S., Croning, M.D.R., Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. (2001) *Bioinformatics*, **17**(7), 646-653.
6. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R. Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**:41.
7. Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., Bateman, A. (2008) The Pfam Protein Families Database. *Nucleic Acids Research* **36**: D281-D288.
8. Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R., White O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes *Nucleic Acids Research* 35, D260-D264.
9. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, K., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, **36** (Database Issue**)**: D480-484.
10. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al. (2005) InterPro, Progress and Status in 2005. *Nucleic Acids Research* **33**, D201-D205.
11. Gene Ontology Consortium. (2008) The Gene Ontology project in 2008. *Nucleic Acids Research* **36**: D440-D444.
12. Ivanova, N.N., Anderson I., Lykidis A., Mavrommatis K., Mikhailova, N., Chen, I.A., Szeto, E., Palaniappan, K., Markowitz, V.M., Kyrpides N.C. (2007) Metabolic Reconstruction of Microbial Genomes and Microbial Community Metagenomes. Technical Report 62292, Lawrence Berkeley National Laboratory. See also: http://img.jgi.doe.gov/w/doc/imgterms.html.
13. Mavromatis, K., Ivanova, N.N., Chen, I.A., Szeto, E., Markowitz, V.M., Kyrpides, N.C. (2009) The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes. *SIGS* **1**(1): 68-71. http://standardsingenomics.org/index.php/sigen/article/view/sigs632.
14. Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I.A., Dubchak, I. Anderson, I., Lykidis, A., Mavromatis, K. et al. (2008) The Integrated Microbial Genomes (IMG) System, *Nucleic Acids Research* **36**, D528-D533.
15. 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**: 955-964.
16. Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W. (2007) RNAmmer: con-sistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**: 3100-3108.
17. Griffiths-Jones, S., Moxon, S., Marshall, M., Khan-na, A., Eddy, S.R., Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**: D121-124.
18. Nawrocki, E.P., Kolbe, D.L., Eddy S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**: 1335-1337.
19. Caspi, R. Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M. Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., et al. (2008) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* **36**: D623-D631.
20. Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. **Nature 402**(6757): 86-90.

21. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *PNAS* **96**(6): 2896-901.
22. Mavromatis, K., Chu, K., Ivanova, N., Hooper, S.D., Markowitz, V.M., and Kyrpides, N.C. (2009) Gene context analysis in the Integrated Microbial Genomes (IMG) data management system, accepted for publication, *PLoS ONE*.
23. Frazer K.A, Pachter, L., Poliakov, A., Rubin, E.M., Dubchak, I. (2004) VISTA: Computational Tools for Comparative Genomics. *Nucleic Acids Research* **32**, W273-W279.
24. Markowitz, V.M., Mavromatis, K., Ivanova, N.N., Chen, I.A., Chu, K., Kyrpides, N.C. (2009) IMG ER: a system for microbial annotation expert review and curation. *Bioinformatics* **25**(17): 2271-2278.
25. Mavromatis, K., Ivanova, N.N., Anderson I., Lykidis, A., Hooper, S.D., Sun, H., Kunin, V., Lapidus, A., Hugenholtz, P., Patel, B., Kyrpides, N.C. (2009a) Genome analysis of the anaerobic thermohalophilic bacterium Halothermothrix orenii. *PLoS ONE* **4**(1).
26. Wu, D., Goodwin, L., Pukall, R., Mavromatis, K., Kunin, V. et al. (2009) A phylogeny-driven genomic encyclopedia of Bacteria and Archaea. (submitted for publication).
27. Markowitz, V. M., Ivanova, N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.A., Grechkin, Y., Dubchak, I., Anderson, I., et al. (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* **36**, D534-538.
28. Unell, M., Abraham, P.E., Shah, M., Zhang, B., Ruckert, C., VerBerkmoes, N.C., Jansson, J.K. Impact of phenotic substrate and growth temperature on the *Arthrobacter chlorophenolicus* Proteome. *J. Proteome Res.* **8** (4): 1953-1964.
29. Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S. et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *PNAS* **102**(39): 13950-13955.

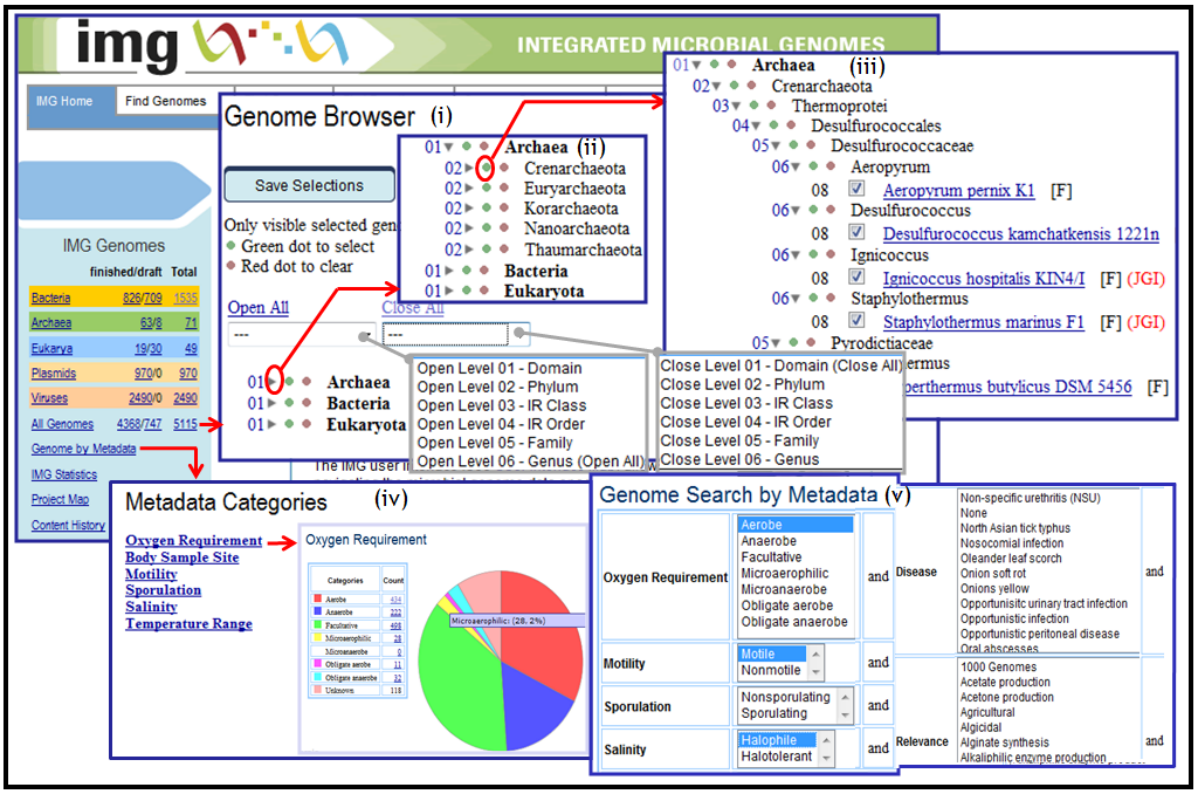

Figure 1. Genome Browser and Search Tools. The "**Genome Browser**" (i) initially displays the three genome domains, whereby the genome display can be modified using the "Open All" and "Close All" options or (ii) using the open/close icons available at each level of the tree. (iii) Genomes can be selected either individually using the select boxes associated with each strain, or in groups using the green dot "select all" icons available at each level of the tree. Metadata genome selection is provided by (iv) a "**Metadata Categories**" based genome classification and (v) a "**Genome Search**" tool based on a variety of metadata attributes.

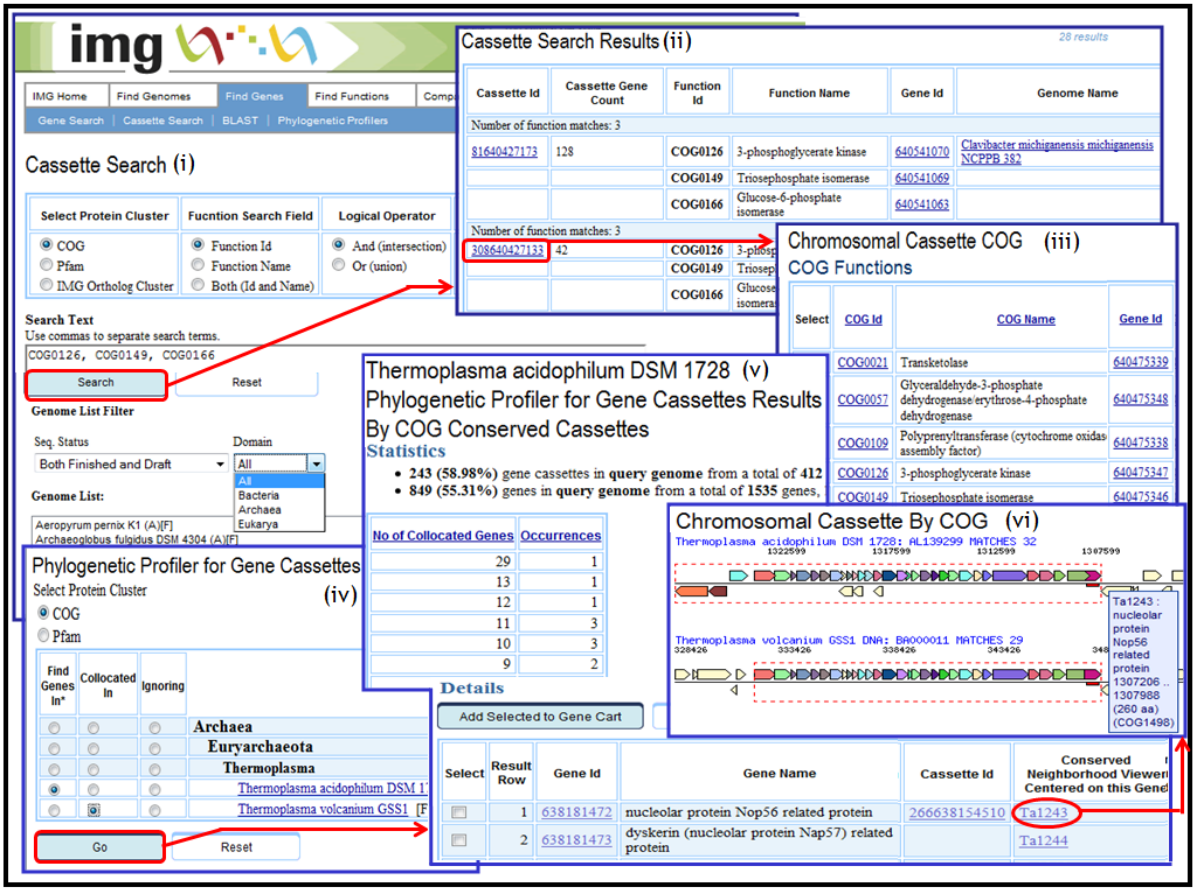

Figure 2. Gene Cassette Search Tools. "**Cassette Search**" allows users to find genes that are part of chromosomal cassettes involving specific protein clusters. First, users (i) select the protein cluster underlying the cassettes, the protein cluster identifier for the search, the logical operator used for the search expression and the order of presenting the search results. The search is carried out across all the genomes in IMG (default) or can be limited only to a subset of genomes using various filters or selecting genomes from the "**Genome List**". (ii) The "**Cassette Search Result**" lists the genes that satisfy the search condition, together with the identifiers of the cassettes they are part of, their associated protein cluster identifiers and names, and their genomes. (iii) The cassette identifiers provide links to the "**Chromosomal Cassette**" details page. (iv) The "**Phylogenetic Profiler for Gene Cassettes**" allows users to find genes that are part of a gene cassette in a query genome and are part of related gene cassettes in other genomes: users select the query genome by using the associated radio button in the "Find Genes In" column, the protein cluster used for correlating gene cassettes, and the genomes for gene cassette comparisons with the query genome by using the associated radio buttons in the "Collocated In". (v) The "**Phylogenetic Profiler for Gene Cassette Results**" starts with a summary of the results, including a table with the first column listing the size of the groups of collocated genes in the query genome and the second column listing the number of such groups conserved across the other genomes involved in the selection. The Details part of the results consists of a table that displays groups of collocated genes in each chromosomal cassette in the query genome that satisfy the search criterion. (vi) The conserved part of a chromosomal cassette involving an individual gene in the query genome can be examined using the links provided in the "**Conserved Neighborhood Viewer Centered on this Gene**" column of results table.

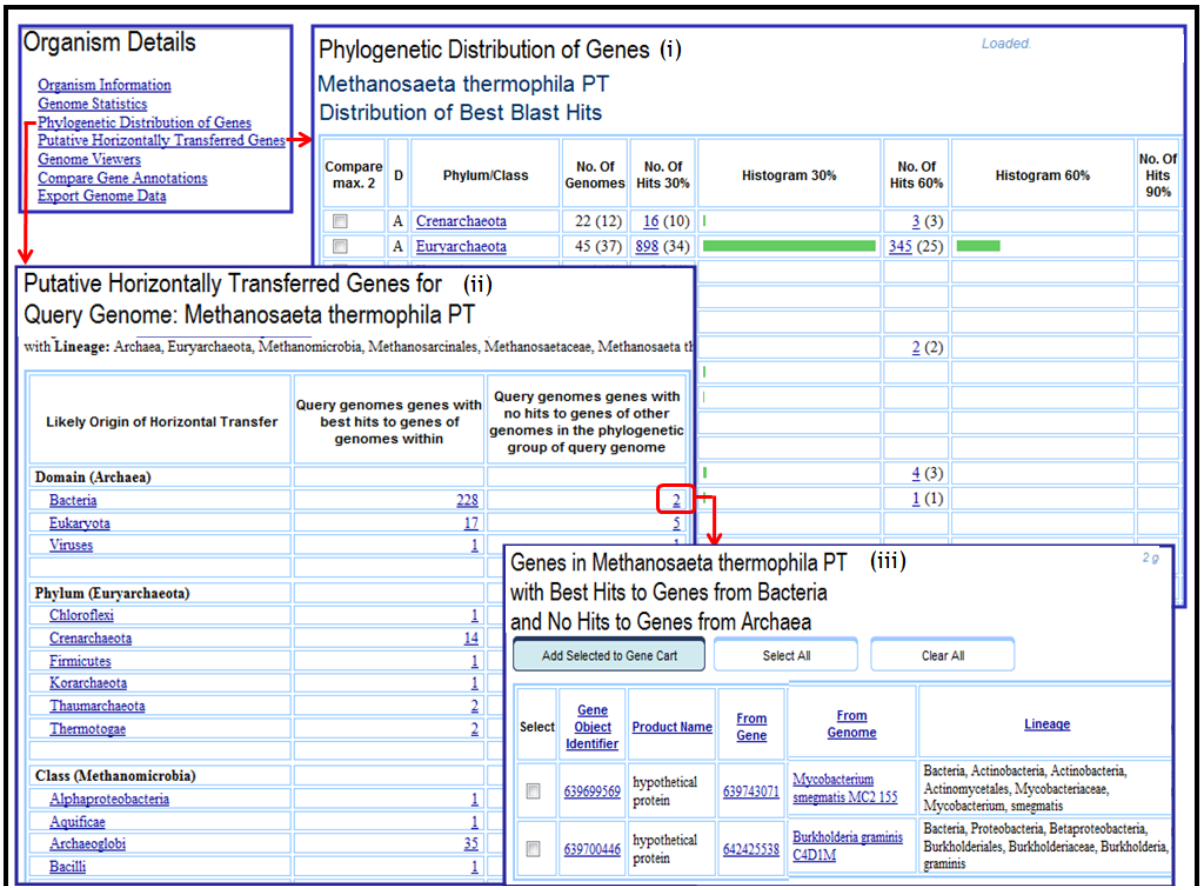

Figure 3. Phylogenetic Distribution of Genes and Putative Horizontally Transferred Genes. The "**Phylogenetic Distribution of Genes**" is available as part of a genome's **Organism Details** and (i) displays the distribution of best BLAST hits of protein-coding genes in the genome as a histogram: counts correspond to the number of genes that have best BLASTp hits to proteins of other genomes in a specific phylum or class with more than 90% identity (right column), 60-90% identity (middle column) and 30-60% identity (left column). Gene counts in the histogram are linked to the lists of genes in the selected genome that have best BLAST hit in a certain phylum/class with specified percent identity. "**Putative Horizontally Transferred Genes"** allows users to explore genes in a query genome that are likely horizontally transferred via (ii) two lists of genes: genes with best hits to genes of genomes within a phylogenetic group (domain, phylum, class, etc.) that is different than the analogous group the query genome belongs to, and genes with best hits to genomes within a phylogenetic group that is different than the analogous group the query genome belongs to, and no hits to genes of genomes within the same phylogenetic group as the group the query genome belongs to. (iii) *Methanosaeta thermophila PT* has 2 genes with best hits to bacterial genomes and no hits to other archaeal genomes, which may indicate a higher likelihood of being horizontally transferred from bacterial genomes.

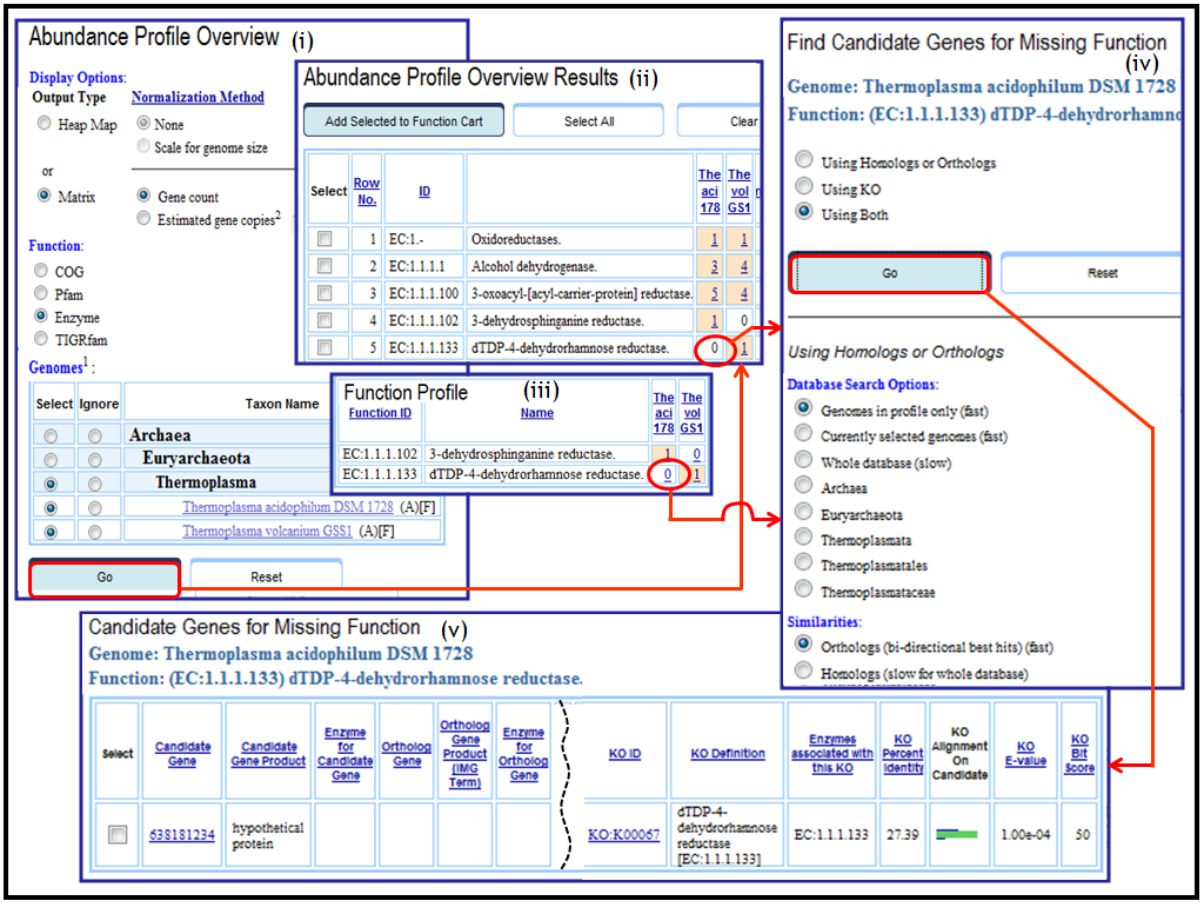

Figure 4. Function Profile Tools. (i) The "**Abundance Profile Overview**" allows users to compare genomes across all the terms of a functional or protein family. Users select the type of format for displaying the results ("Heat Map" or "Matrix"), protein/functional families (COG, Pfam, TIGRfam, Enzyme), normalization method, and a set of genomes. (ii) If the "Matrix" option 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. (iii) The "**Function Profile**" allows users to compare genomes across functional or protein family terms selected using the "**Function Cart**". (iii) The result of a "**Function Profile**" is displayed in a tabular format similar to the "Matrix" format of the "**Abundance Profile Overview**". Users can click on a cell of an "**Abundance Profile Overview**" or "**Function Profile**" result in order to retrieve the list of genes assigned to a particular family in a genome. For profiles involving enzymes, a zero abundance ("missing") enzyme leads to (iv) the "**Find Candidate Genes for Missing Function**" tool that allows users to find candidate genes of a target genome that could be associated with the missing enzyme. The search can be conducted across all IMG genomes, across a subset of genomes within a certain domain/phyla/class, or only across the selected genomes. The search can be based on homologs, orthologs, or KO terms for finding genes that could be associated with the "missing" enzyme. (v) The result of the search for candidate genes consists of a list of genes that can be selected and included into the "**Gene Cart**".