Genome Portal, Joint Genome Institute

Igor V. Grigoriev[1], Susannah Tringe[1], Inna Dubchak[1]

[1]United States Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA.

March 2013

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal 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. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of 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 any agency thereof or The Regents of the University of California.

Synonyms

Genome projects, data integration, genome analysis, comparative genomics, metagenomics

Definition

The United States Department of Energy (DOE) Joint Genome Institute (JGI) is a national user facility with massive-scale DNA sequencing and analysis capabilities dedicated to advancing genomics for bioenergy and environmental applications.

The JGI Genome Portal is an integrated genomic resource, which provides for the research community around the world access to the large collection of genomics data for plants, fungi, microbes and metagenomes and to web-based interactive tools for their analysis.

Introduction

The Department of Energy (DOE) Joint Genome Institute (JGI) was established for the Human Genome Project (Lander 2001) and later was transformed into a national user facility for genome research in the DOE mission areas of bioenergy, carbon cycling and biogeochemistry. JGI provides expertise and resources in DNA sequencing, technology development, and bioinformatics to the broader scientific community. Scientists around the world can make proposals to the JGI Community Sequencing Program (CSP; e.g., (Martin 2011)) to sequence genomes, transcriptomes and metagenomes and address important scientific questions of DOE mission relevance. Massive amounts of genomic data are assembled, annotated and delivered to users by means of integrated databases and interactive analytical tools interconnected within the JGI Genome Portal (http://genome.jgi.doe.gov;(Grigoriev 2012)).

Leading the world in the number of sequenced plants, fungi, microbes, and metagenomes (according to the Genomes Online Database (GOLD; (Pagani 2012)), JGI has dramatically increased its sequencing capabilities using new sequencing technologies. JGI projects evolved from sequencing three of the human chromosomes (Lander 2001)to the large scale “Grand Challenge” projects such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA; (Wu 2009)), the 1000 fungal genome project (Grigoriev 2011), and metagenomic projects targeting soil and rhizosphere. Since tracking individual organisms and samples at such a scale becomes critical, genomes and metagenomes sequenced or selected for sequencing are carefully catalogued and made available to the public along with their status and links to the produced data and available tools.

The sequenced data are assembled, annotated and analyzed using various computational pipelines developed for each of the products delivered by JGI to its users. The resulting annotations are available for download and also can be interactively viewed using the JGI Genome Portal offering a wide array of databases and analytical systems to interpret the data. Some systems work across multiple JGI databases, while others allow users to specifically manage datasets on plants (Phytozome; (Goodstein 2012)), fungi (MycoCosm; (Grigoriev 2012)), microbes (Integrated Microbial Genomes or IMG; (Markowitz 2012b)) and metagenomes (IMG/M; (Markowitz 2012a)).

The JGI Genome Portal provides a unified access point to all JGI genomic databases and analytical tools, as well as worldwide statistics on the usage of the JGI resources and the information about the latest genome releases and new tool development. A user can find all DOE JGI sequencing projects and their status, search for and download assemblies and annotations of sequenced genomes and metagenomes, and interactively explore those datasets and compare them with other sequenced microbes, fungi, plants or metagenomes using specialized systems tailored to each particular class of organisms. All these can serve as building blocks in comprehensive analyses of individual organisms or systems of interacting organisms.

A Catalog of Genome Sequencing Projects

Metagenomic analysis requires reference genomes for better interpretation of sequence data derived from complex microbial communities. The democratization of sequencing allows many scientists to sequence appropriate genome references in their own labs prior to approaching metagenomes. Consolidation of genomic data sequenced in different places around the world is an important step in both genomics and metagenomics.

JGI's collection of genomic projects includes over 7,000 projects of different types and is publicly available and searchable. Product types include standard or improved genome drafts, finished genomes, gene expression profiling, resequencing, metagenome projects, and others. The *Project List* (http://genome.jgi.doe.gov/genome-projects) is available from most of the Portal pages as a menu item and includes a detailed description of each project including its scope and current status, taxon, the JGI program and the project lead. The Resources column lists tools available for this project. Some of these tools, e.g.*download* are available for all genomes, while others are taxon-, project type- or stage-dependent. For example, a plant or fungal genome will be linked to Phytozome or MycoCosm, respectively.

All JGI projects are also registered in the GOLD database, which includes a larger collection of projects sequenced around the world (Pagani 2012). Currently it contains a list of about 16,000 genomes including over 3,000 that are complete and over 2,000 metagenomes. Besides utility for metagenomics, having a comprehensive list of sequencing projects from all laboratories around the world also helps to avoid redundancy when sequencing targets are selected for the large scale projects like GEBA or 1000 fungal genomes.

Annotated Genomes and Metagenomes

Finding genes in metagenomes is challenging, especially for eukaryotes with their complex intron-exon gene structure, and often relies on gene prediction based on similarity to proteins from other organisms. This requires a comprehensive collection of genes from different organisms across all domains of life. Beside the human genome (Lander 2001), JGI sequenced and annotated genomes of the first poplar tree (Tuskan 2006)and its ectomycorrhizal symbiont (Martin 2008), lignocellulose degrading fungi (Berka 2011;Eastwood 2011)and microbial communities (Hess 2011), diverse eukaryotes, often the first representatives of the Tree of Life branches (Tyler 2006;Bowler 2008;King 2008;Fritz-Laylin 2010;Colbourne 2011)and prokaryotes (Wu 2009)as well as soil (Tringe 2005)and ocean metagenomes (Walsh 2009). There are over 3,000 annotated reference genomes in the JGI database and three ways to find a particular genome of interest: using an interactive *Tree of Life*, s*earch*, and *select* functions.

The Tree of Life organizes the annotated genomes by the domains of life and links to Organism home pages. Clicking on a branch name produces a menu displaying available genomes in this kingdom, phylum, class, or order (Fig. 1), each connected to pages in different analytical resources. The same pages can be reached in a step-by-step genome selection from a hierarchical selection menu on the top of the page or searching for genomes by keyword (e.g. plants, Eukaryota), name, taxonID or projectID.

Each of the genomic datasets can be analyzed with a collection of tools linked directly to their genome databases. Each organism's home page contains a description of the project, BLAST, download and links to specialized resources as described in the next section.

Comparative Databases and Tools

Comparative genomics is more powerful approach for functional annotation and evolutionary studies of genomes than analysis of individual genome sequences. It is also a primary method for annotation and analysis of metagenomes. The JGI genome Portal includes a set of efficient comparative tools, such as gene clustering, whole-genome alignment, and building phylogenetic trees that are used across different genomic resources at JGI. *VISTA Point* is an example, designed for visualization and analysis of pairwise- and multiple DNA alignments (Frazer 2004)at different levels of resolution in three visualization modes: (a) *VISTA Browser,* for visual comparative analysis of complete genome assemblies using pairwise and multiple large-scale alignments; (b) *VISTA Synteny Viewer*, a multi-tiered graphical display of pairwise alignments at three different levels of resolution; (c) *VistaDot,* an interactive two-dimensional dot-plot genome synteny viewer across multiple chromosomes/scaffolds. Several specialized domain-specific computational systems for comparative genome analysis built at JGI include Phytozome, a comparative hub for plant genome and gene family data and analysis; MycoCosm to enable users to navigate across sequenced fungal genomes, and to conduct comparative and genome-centric analyses and community annotation; the IMG family of tools for large-scale comparative analysis of microbial genomes and metagenomes.

Phytozome (http://phytozome.net; (Goodstein 2012)) gives access to the sequences and functional annotations of a growing number of complete plant genomes (31 in release v8.0), including land plants and selected algae. Phytozome provides both organism-centric and gene family-centric views, as well as access to the BLAST, BLAT, and Search capabilities.

Phytozome provides a view of the evolutionary history of every plant and every plant gene at the level of sequence, gene structure, gene family and genome organization. The Phytozome project organizes the proteomes of green plants into gene families defined at nodes on the green plant evolutionary tree. Genes have been annotated with PFAM, KOG, KEGG, and PANTHER assignments, and publicly available annotations from RefSeq, UniProt, TAIR, and JGI are hyper-linked and searchable. The Gene Family view gives access to the information on each family and its members, organized to highlight shared attributes.

GBrowse provides genome-centric views for all genomes included in Phytozome. Each organism browser displays a number of tracks including a gene prediction track, a track of homologous sequences from related species aligned against the genome, supporting EST and VISTA tracks identifying regions of this genome that are syntenic with other plant genomes.

MycoCosm (http://jgi.doe.gov/fungi; (Grigoriev 2012)) brings together genomic data and analytical tools for diverse fungi that are important for energy and environment. Genomics data from the JGI and its users are integrated and curated via user community participation in data submission, curation, annotation, and analysis. Over 150 newly sequenced and annotated fungal genomes are available to the public through MycoCosm for *genome-centric* and *comparative* analyses. Visual navigation across the MycoCosm tree (Fig 2b), where each node represents a group of phylogenetically related fungi and is linked to analysis tools, allows user to redefine the search and analysis space from a single organism to the entire list of fungal genomes.

The Genome browser with configurable selection of tracks displays predicted gene models and annotations along with different lines of evidence in support of these predictions, such as gene and protein expression profiles. Gene models and annotations are linked to community annotation tools to revise them if needed. Functional profiles of each genome summarize gene annotations according to the GO, KEGG, and KOG classifications and can be compared with each other to study gene family expansions or contractions at different levels of granularity. Clustering using BLAST alignments of all proteins and MCL can expand these analyses to gene families even without annotation and enable side-by-side comparison of each of the cluster members for pattern of protein domains, intron-exon structure, and synteny.

MycoCosm comparative views combine the above mentioned tools to study entire groups of genomes corresponding to MycoCosm nodes. Unlike the genome-centric view, there is no reference genome in this analysis and, for example, a keyword or BLAST search for protein kinases in Basidiomycota or Ascomycota will show differences in the number of found genes or BLAST hits across different members of these phyla.

IMG, the Integrated Microbial Genomes database(http://img.jgi.doe.gov; (Markowitz 2012a;Markowitz 2012b)) is a system designed for flexible comparative analyses of microbial genomic data, which incorporates all complete public microbial genomes as well those sequenced at JGI. IMG with Microbiome samples (IMG/M) is an expanded database that includes metagenome data from diverse environments, both sequenced at JGI and submitted by external users.

In addition to importing all public genomes and their annotations from NCBI’s RefSeq, IMG curates the data by adding features missed by many annotation pipelines, such as small RNAs; assigning proteins and domains to all major protein family databases (e.g. COG, TIGRfam); and linking to organism metadata stored in GOLD, such as oxygen requirements or environment of origin. Annotations can be viewed in detailed gene pages, or summarized in genome pages that include organism metadata in addition to statistics on genome size and gene counts within various categories.

The tools available in IMG allow for analyses to be readily performed at the gene, function, or genome level, using customizable “carts” for each of these data types. Thus any given analysis can readily be performed on a single (meta) genome or several, and can be extended to many individual genes, functions or pathways. IMG/M includes a number of metagenome-specific functions, including the option to account for different organism abundances by weighting comparative analyses according to estimated gene copies, based on the contig read coverage reported in the assembly, rather than simple gene counts. It also includes a “scaffold cart” for exploring genes within a given set of contigs or scaffolds, as well as the option to categorize contigs / scaffolds into population “bins” based on oligonucleotide composition or other features.

Recent developments in IMG and IMG/M include the capacity to add and view (meta) transcriptome and (meta) proteome data in the context of a reference and compare expression profiles across experiments.

Metagenome Analysis

Analysis of metagenome data presents a number of challenges beyond those faced in isolate genome analysis, in particular the wide variation in individual organism abundances and the shallow coverage of low-abundance, but nonetheless biologically important, taxa. Both of these tend to result in highly fragmented assemblies, which are most readily interpreted when high-quality reference genome data are available.

Most metagenome analyses approach the data from either a phylogenetic perspective (i.e., who is there?) or a functional one (i.e. what are they doing?). Each of these uses a specific suite of tools, though nearly all rely on a well-curated database of genes with known phylogenies and functions. For phylogenetic analysis, genes or gene fragments are assigned to phylogenetic lineages based on homology to genes of known phylogenetic origin. This can be done for all genes from a metagenome dataset, for example using MEGAN (Huson and Mitra 2012), or for a set of conserved phylogenetic markers which can be placed onto a trees of known sequences from isolate genomes and/or amplified from uncultivated organisms, for example using pplacer (Matsen 2012). IMG/M allows for both approaches - an overall perspective of all the genes in a dataset or on a specific set of contigs is provided through the “Phylogenetic Distribution of Genes” option on the main metagenome page or in the scaffold cart, and genes with homology to particular phyla, families, genera or species can be retrieved. When there are good reference genomes available, alignments of protein-coding genes to those genomes can be viewed in a recruitment plot (Fig. 3). Phylogenetic marker genes can also be extracted and incorporated into trees using the "Phylogenetic Marker COGs" option under the "Find Functions" tab.

Functional or "gene-centric" approaches enable the comparison of metagenome datasets at the functional level, to both assess their relative similarity and identify genes or functions that are over- or under-represented in a given dataset. This type of approach is utilized by metagenome analysis systems like MG-RAST (Meyer 2008). IMG/M provides several options for whole metagenome comparisons. Metagenomes can be clustered (under the "Compare Genomes" tab) according to gene content, using either functional (e.g. COG, Pfam) or phylogenetic criteria, and the results visualized via hierarchical clustering, principal components analysis (PCA) or a correlation matrix. Relative abundances of specific gene families can be viewed via the abundance profile function also under the "Compare Genomes" tab. As mentioned above, these comparisons can be made between partly assembled genomes by taking contig read depth into account when calculating gene abundance. As mentioned above, these comparisons can be made between partly assembled genomes by taking contig read depth into account when calculating gene abundance.

Summary

Technological innovations leading to the democratization of genome sequencing have resulted in large amounts of genomic data being produced in different parts of the world. Effective analysis of genomic and metagenomic data depends on the availability of comprehensive catalogs of reference genome data for annotation and comparative genomics as well as computational tools able to process the large amounts of sequence data. The JGI Genome Portal (http://genome.jgi.doe.gov) provides a unified access point to all JGI genomic databases and analytical tools including list of sequencing projects at JGI and around the world, a comprehensive collection of annotated genomes in all domains of life, and specialized databases for comparative analysis of plant, fungal and microbial genomes and metagenomes. The latter is still in early stages of development and data generated at unprecedented scale and complexity for metagenomes will require new approaches to data processing, analysis and visualization.

References

Berka R M, Grigoriev I V, Otillar R, et al. Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris. Nat Biotechnol. 2011; 29: 922-927.

Bowler C, Allen A E, Badger J H, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008; 456: 239-244.

Colbourne J K, Pfrender M E, Gilbert D, et al. The ecoresponsive genome of Daphnia pulex. Science. 2011; 331: 555-561.

Eastwood D C, Floudas D, Binder M, et al. The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. Science. 2011; 333: 762-765.

Frazer K A, Pachter L, Poliakov A, et al. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004; 32: W273-279.

Fritz-Laylin L K, Prochnik S E, Ginger M L, et al. The genome of Naegleria gruber iilluminates early eukaryotic versatility. Cell. 2010; 140: 631-642.

Goodstein D M, Shu S, Howson R, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012; 40: D1178-1186.

Grigoriev I V, Cullen D, Goodwin S B, et al. Fueling the future with fungal genomics. Mycology. 2011; 2: 192-209.

Grigoriev I V, Nordberg H, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res. 2012; 40: D26-32.

Hess M, Sczyrba A, Egan R, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science. 2011; 331: 463-467.

Huson D H and Mitra S Introduction to the analysis of environmental sequences: metagenomics with MEGAN. Methods Mol Biol. 2012; 856: 415-429.

King N, Westbrook M J, Young S L, et al. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature. 2008; 451: 783-788.

Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409: 860-921.

Markowitz V M, Chen I M, Chu K, et al. IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res. 2012a; 40: D123-129.

Markowitz V M, Chen I M, Palaniappan K, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 2012b; 40: D115-122.

Martin F, Aerts A, Ahren D, et al. The genome of Laccaria bicolor provides insights into mycorrhizal symbiosis. Nature. 2008; 452: 88-92.

Martin F, Cullen D, Hibbett D, et al. Sequencing the fungal tree of life. New Phytol. 2011; 190: 818-821.

Matsen F A, Hoffman N G, Gallagher A, et al. A format for phylogenetic placements. PLoS One. 2012; 7: e31009.

Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMCBioinformatics. 2008; 9: 386.

Pagani I, Liolios K, Jansson J, et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 2012; 40: D571-579.

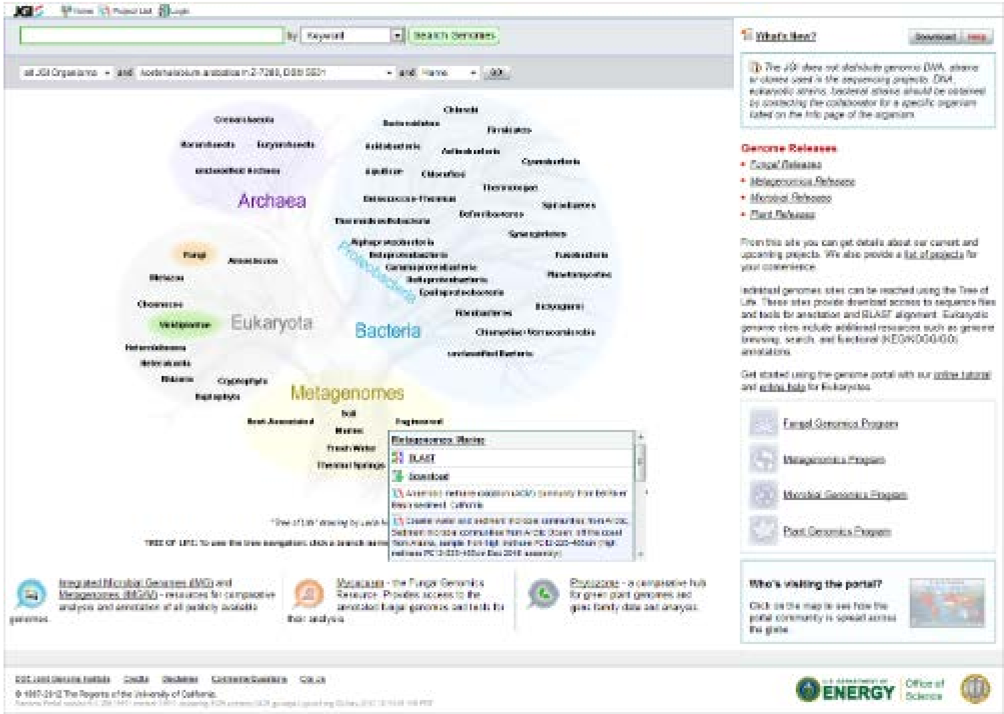
Tringe S G, von Mering C, Kobayashi A, et al. Comparative metagenomics of microbial communities. Science. 2005; 308: 554-557.

Tuskan G A, Difazio S, Jansson S, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006; 313: 1596-1604.

Tyler B M, Tripathy S, Zhang X, et al. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science. 2006; 313: 1261-1266.

Walsh D A, Zaikova E, Howes C G, et al. Metagenome of a versatile chemolithoautotrophfrom expanding oceanic dead zones. Science. 2009; 326: 578-582.

Wu D, Hugenholtz P, Mavromatis K, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature. 2009; 462: 1056-1060.



Anaerobic methane oxidation (AOM) community from Eel River Basin sediment, California

Coastal water and sediment microbial communities from Arctic: Sediment microbial communities from Arctic Ocean, off the coast from Alaska, sample from high methane PC12-225-485cm (high methane PC12-225-485cm Dec 2010 assembly)

"Tree of Life" drawing by Leslie [illegible]

TREE OF LIFE: To use the tree navigation, click a branch name [illegible]

The JGI does not distribute genomic DNA, strains or clones used in the sequencing projects. DNA, eukaryotic strains, bacterial strains should be obtained by contacting the collaborator for a specific organism listed on the Info page of the organism.

From this site you can get details about our current and upcoming projects. We also provide a list of projects for your convenience.

Individual genomes sites can be reached using the Tree of Life. These sites provide download access to sequence files and tools for annotation and BLAST alignment. Eukaryotic genome sites include additional resources such as genome browsing, search, and functional (KEGG/KOG/GO) annotations.

Get started using the genome portal with our online tutorial and online help for Eukaryotes.

Integrated Microbial Genomes (IMG) and Metagenomes (IMG/M) - resources for comparative analysis and annotation of all publicly available genomes.

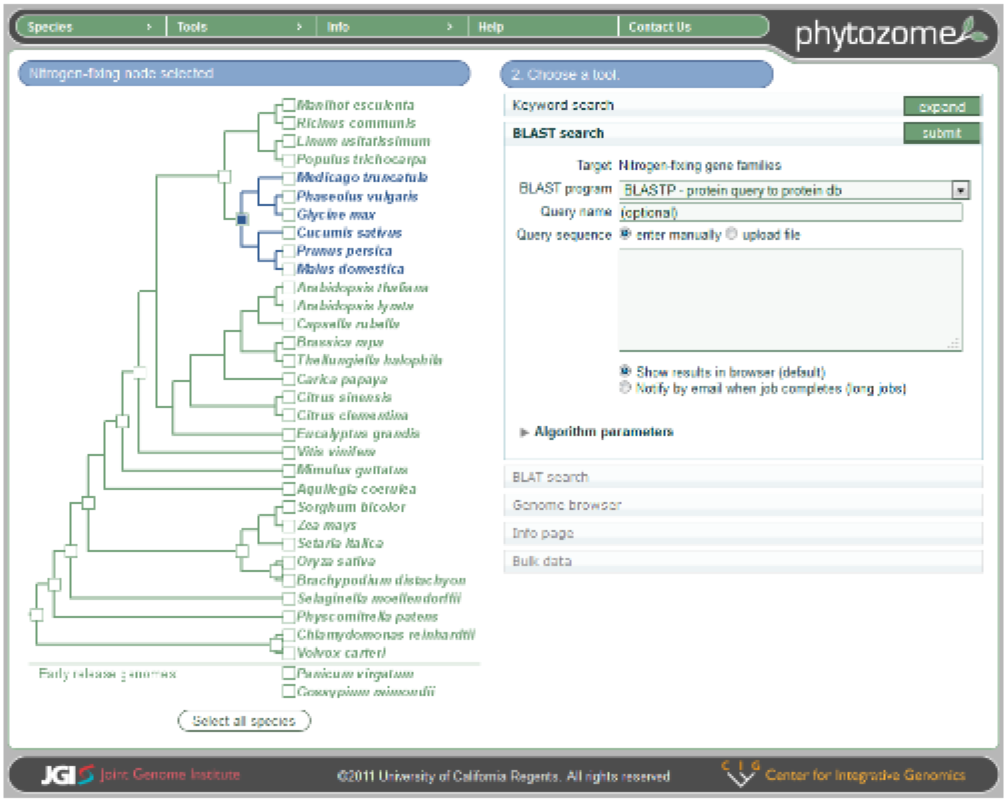
MycoCosm - the Fungal Genomics Resource. Provides access to the annotated fungal genomes and tools for their analysis.

Phytozome - a comparative hub for green plant genomes and gene family data and analysis.

Click on the map to see how the portal community is spread across the globe.

© 1997-2012 The Regents of the University of California.

***Figure 1**. The JGI Genome Portal. A pull-down menu for the 'Marine' category of Metagenomes is shown. BLAST and Download functions are available for the entire selected group. Each genome is linked to the associated resources. 'Project list' on the top leads users to the list of all sequencing projects at the DOE JGI. The bottom portion of the page connects to the specialized databases in microbes (IMG) and metagenomes (IMG/M), fungi (MycoCosm), and plants (Phytozome)*



Target: Nitrogen-fixing gene families

BLAST program: BLASTP - protein query to protein db

Query name: (optional)

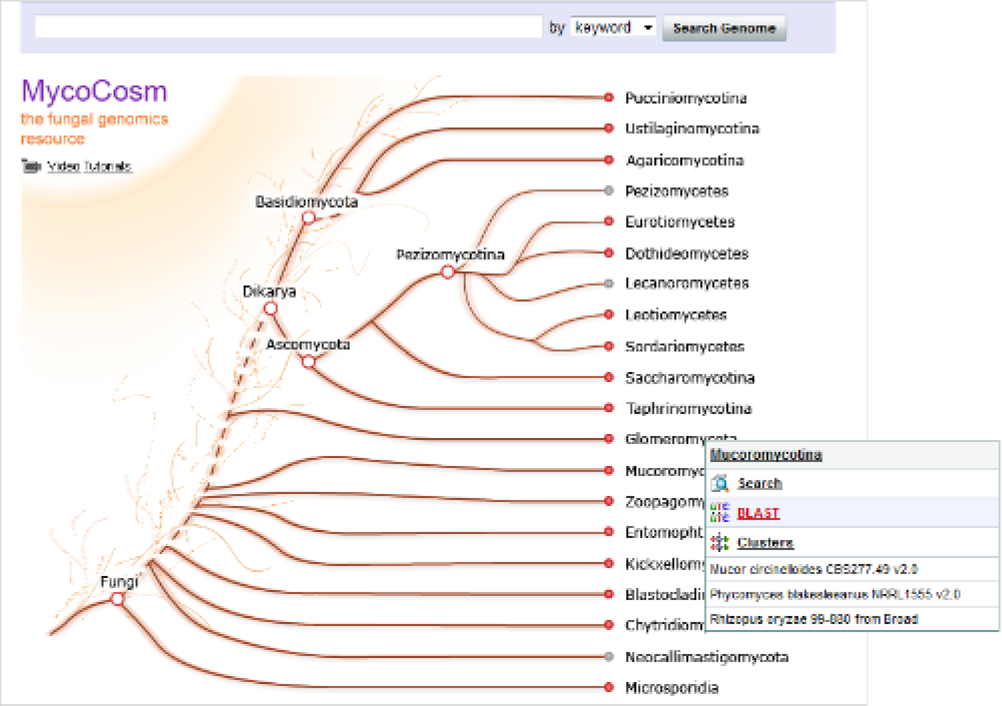
Query sequence: ● enter manually ○ upload file

● Show results in browser (default)

○ Notify by email when job completes (long jobs)

©2011 University of California Regents. All rights reserved

Figure 2 (a). *Comparative genomics resources at JGI: Phytozome for plants*

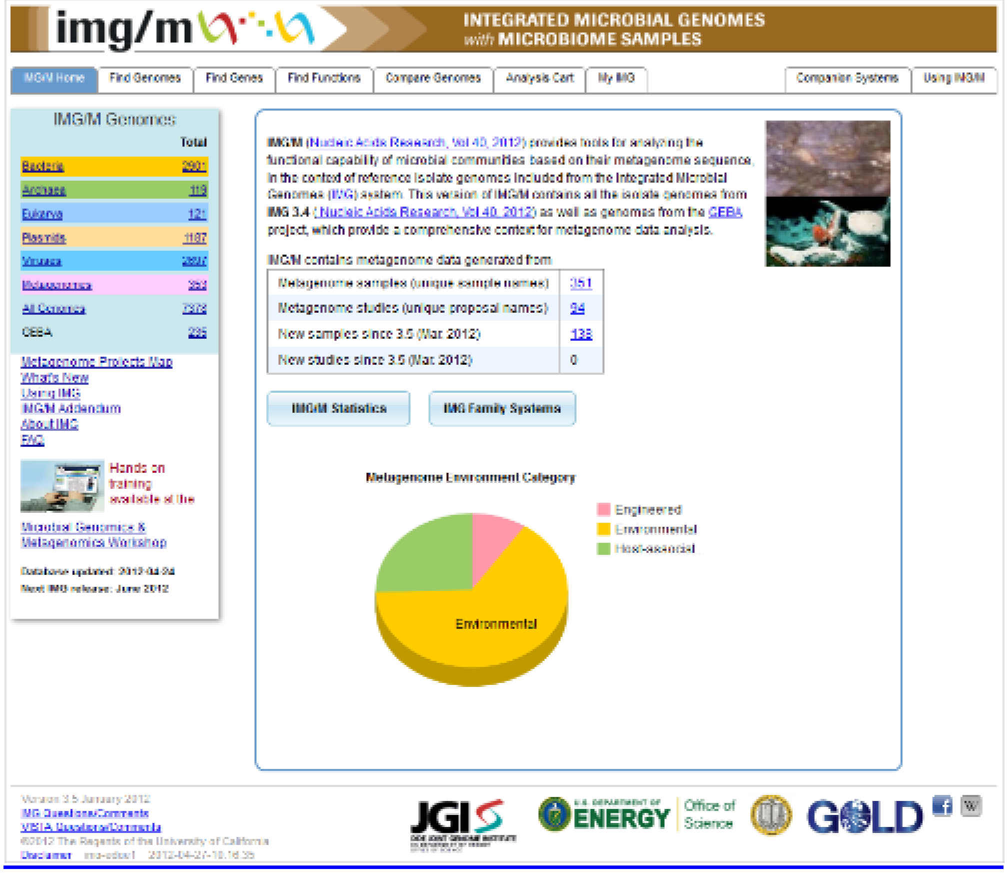


Mucor circinelloides CBS277.49 v2.0

Phycomyces blakesleeanus NRRL1555 v2.0

Rhizopus oryzae 99-880 from Broad

Figure 2 (b). *Comparative genomics resources at JGI: MycoCosm for fungi*



	Total
Bacteria	2901
Archaea	119
Eukarya	121
Plasmids	1187
Viruses	2897
Metagenomes	353
All Genomes	7378
GEBA	235

Database updated: 2012-04-24
Next IMG release: June 2012

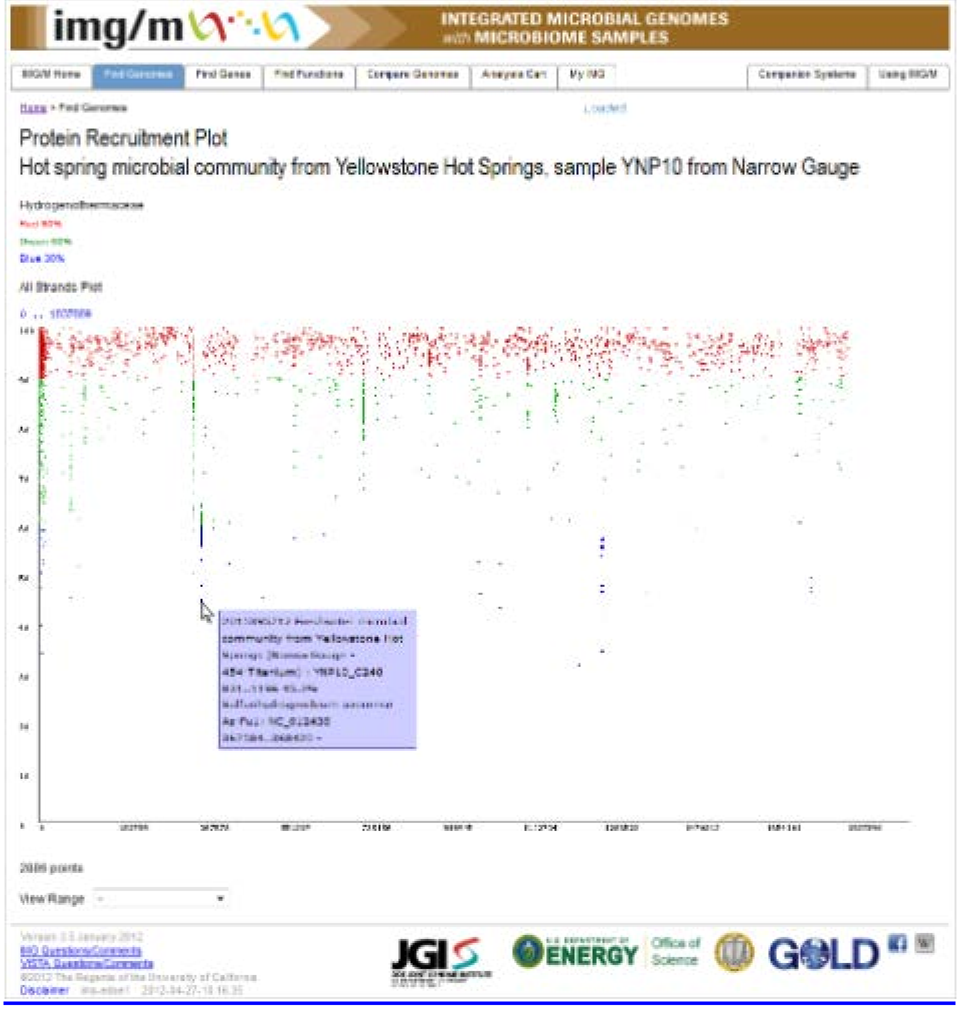
IMG/M (Nucleic Acids Research, Vol 40, 2012) provides tools for analyzing the functional capability of microbial communities based on their metagenome sequence, in the context of reference isolate genomes included from the Integrated Microbial Genomes (IMG) system. This version of IMG/M contains all the isolate genomes from IMG 3.4 (Nucleic Acids Research, Vol 40, 2012) as well as genomes from the GEBA project, which provide a comprehensive context for metagenome data analysis.

IMG/M contains metagenome data generated from

Metagenome samples (unique sample names)	351
Metagenome studies (unique proposal names)	94
New samples since 3.5 (Mar 2012)	138
New studies since 3.5 (Mar 2012)	0

Version 3.5 January 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer img-edge1 2012-04-27-10.16.35

Figure 2(c). *IMG family of tool*



Protein Recruitment Plot

Hot spring microbial community from Yellowstone Hot Springs, sample YNP10 from Narrow Gauge

Hydrogenothermaceae

Red 90%

Green 60%

Blue 30%

All Strands Plot

[illegible] community from Yellowstone Hot Springs (Narrow Gauge) - 454 Titanium) - YNP10_C240 [illegible] Sulfurihydrogenibium azorense Az-Fu1: NC_012438 [illegible]

2886 points

View Range -

Version 3.5 January 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer img-edge1 2012-04-27-18 16:35

Figure 3. *Metagenomics analysis. A protein recruitment plot showing alignment of genes from a hot spring sample to genomes from the family Hydrogenothermaceae.*