Integrated Microbial Genomes data analysis system

Markowitz V.M.[1], Palaniappan K.[1], Szeto E.[1], Korzeniewski F.[1], Werner G.[2], Greiner A.[2], Padki A.[2], Zhao X.[2], Taylor K.[2], Hugenholtz P.[2], Mavrommatis K.[2], Anderson I.[2], Lykidis A.[2], Ivanova N.[2], Kyrpides N.C.[2]

1: Lawrence Berkeley National Laboratory, Biological Data Management and Technology Center, 2: DOE Joint Genome Institute

http://img.jgi.doe.gov

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

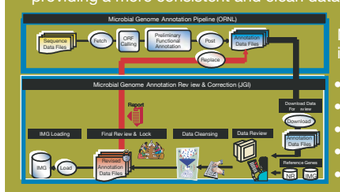
What is the purpose for this system?

JGI aims to provide a data management system that will:

- Support the analysis of all genomes
- Support community annotation of genomic sequences
- Support modeling of cellular networks
- Become a platform for understanding genomes
- Enable scientists to explore microbial community genomic data in the context of relevant environmental, geochemical and phylogenetic data
- Provide documentation and clarity of data, structural and operational semantics
- Be user friendly

Data Quality

Genomic data are manually curated before entering IMG system, providing a more consistent and clean data set



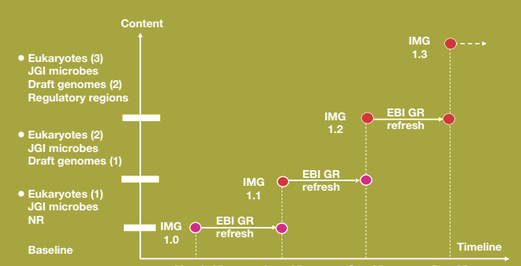
Manual curration includes identification and correction of:

- Overlapping ORFs
- ORFs with wrong 5 start sites
- Pseudogenes
- Dubious ORFs
- Missed ORFs

Future Plans

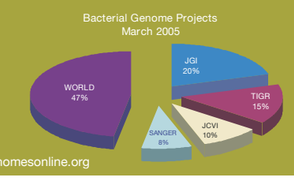
Through quarterly releases, IMG aims at providing high levels of

- genome diversity, in terms of the number of genomes integrated into the system from public resources
- data quality, in terms of the coherence of annotations, the use of sound validation and correction procedures, as well as corroboration of annotations from other public microbial genome data resources.
- annotation coverage, in terms of the breadth of the functional annotations



Why JGI?

JGI is a major worldwide contributor of bacterial genomic sequences



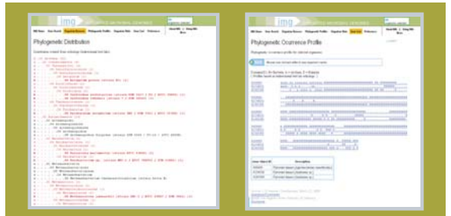
© http://www.genomesonline.org

Data Quanity

	JGI	Total
Bacteria	30/68	195/68
Archaea	0/4	20/4
Eukarya	0/0	9/0
All organisms	30/72	224/72

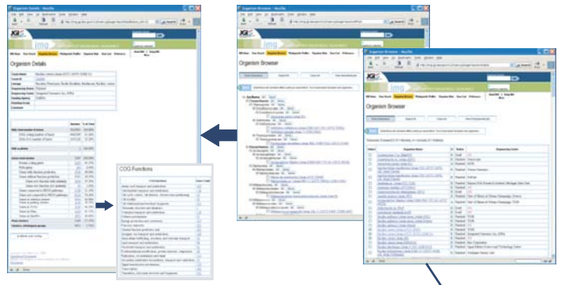
Finished/Draft

Future Extensions

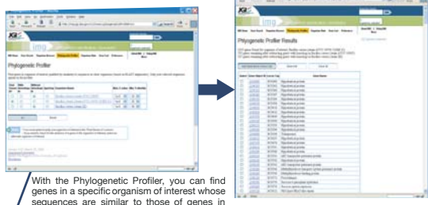


Future viewers will show the phylogenetic distribution of one (top viewer) or multiple (bottom viewer) orthologous genes across IMG genomes

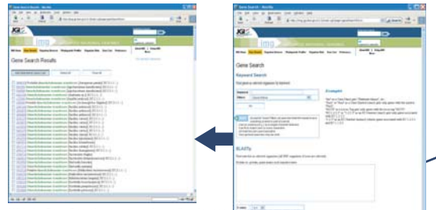
The Organism Browser provides multiple entries to the IMG system.
You can view an alphabetical list of all organisms whose genomes are available for study, or you can view the list as a phylogenetic tree.
You can then click on an individual organism to view details about its genome and follow links to useful groupings of genes within that genome (e.g., genes connected to KEGG pathways).



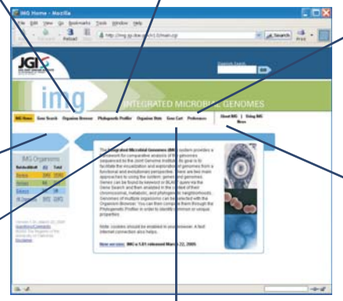
To select a subset of organisms, click to place a checkmark in the box next to each organism of interest, then click on "Save Selections." The selections you save will be used in multiple areas of IMG. These include the Gene Search, Gene Ortholog Neighborhoods (reached via the Gene Details page), Phylogenetic Profiler, Organism Stats, and the highlighting of orthologs in ortholog and homolog lists.

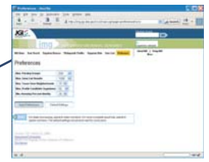


With the Phylogenetic Profiler, you can find genes in a specific organism of interest whose sequences are similar to those of genes in certain organisms and exclude those that are similar to genes in other organisms.



With the Gene Search, you can retrieve a list of genes to analyze individually or add to your gene cart. You can search by keyword or by running a BLAST query. The number of results is limited to the "Max. Gene List Results" number set in your Preferences



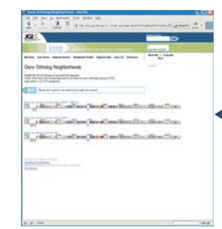


Colored EC numbers on the KEGG map are links. Red links represent the current gene and point to its Gene Details page. Green links indicate genes in a positional cluster with the current gene and link to the Gene Details page for the positional cluster gene. Blue links are other genes identified in the same organism. Click a purple link to see a list of genes in the same organism with the same EC number. Orange links are "EC equivalogs," or genes found in other organisms that have the same EC number assignment.

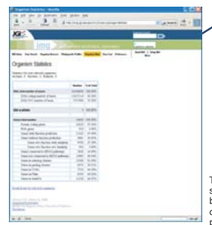
The Gene Information table includes identifying information, locus information, chemical information about the product, and KEGG information. To see more information click the "Show All Gene Information" button.
To quickly run a BLASTp of the sequence against NR (NCBI's nonredundant protein database); SwissProt; or Pfam, click the "BLASTp against External Databases" button.

The Evidence for Function Prediction provides a view of the gene's immediate neighborhood. The gene itself is shown in red. Genes in the same positional cluster that are also in the same KEGG pathway are highlighted in green. When you move the cursor over any gene, you will see a popup box with the locus tag, gene name, and scaffold coordinates (except in Internet Explorer for the Macintosh). Click the arrow representing any gene to see the Gene Details page for that gene.
To see more of the gene of interest's chromosome, click "Show in Chromosome Viewer." To see the neighborhoods of the gene and its rthologs, click "Show ortholog neighborhood regions in user-selected organisms".

The Chromosome Viewer shows all the genes in a range of sequence, with color coding by COG group. The gene whose details you were viewing is shown in red.

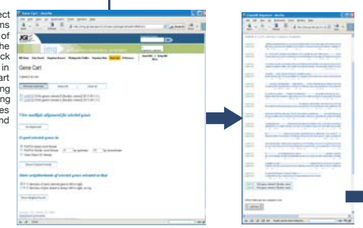


To see more of the gene of interest's chromosome, click "Show in Chromosome Vewer." To see the neighborhoods of the gene and its orthologs, click "Show ortholog neighborhood regions in user-selected organisms.!
You can simultaneously view neighborhoods for the gene of interest and its orthologs in organisms you've selected.
The Chromosome Viewer shows all the genes in a range of sequence, with color coding by COG group. The gene whose details you were viewing is shown in red.



As you work in IMG, you can collect genes from the various IMG organisms into your Gene Cart. From any list of genes, click the checkboxes for the genes you want to add and then click "Save selections." Once a gene is in your cart, you can use the Gene Cart page for doing alignments, comparing gene neighborhoods, and downloading sequence. You can also remove genes from the cart by selecting them and clicking "Remove Selected".

The Organism Stats reports statistics for the currently selected organisms. The data includes numbers of bases, scaffolds, and genes with various characteristics (e.g., genes connected to KEGG pathways, genes in InterPro).



This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the University of California, Lawrence livermore National Laboratory under Contract No. W-7405-ENG-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC0376SF00098 and Los Alamos National Laboratory under Contract No. W-7405-ENG-36.

CSO_10591