

Microbial Genome Annotation & Analysis Tutorial



Technical Report LBNL-63615

Genome Biology Program
Department of Energy Joint Genome Institute

Biological Data Management and Technology Center
Lawrence Berkeley National Laboratory

October 15, 2007

Copyright 2007 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.

October 15, 2007

©2007 The Regents of the University of California

This document was prepared by:

Victor M. Markowitz*
Natalia N. Ivanova**
Iain Anderson**
Athanasios Lykidis**
Konstantinos Mavromatis**
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

Glossary of Terms	1
1. Synopsis	2
2. Getting Started	2
3. Genome Context for Annotation	3
4. Select Target Genes for Annotation	5
4.1 Select Genes using Gene Search	5
4.2 Select Genes from a Specific Genome Statistics	6
4.3 Select Genes by Comparing Gene Annotations	7
4.4 Select Genes using the Phylogenetic Profiler	8
4.5 Select Genes using Abundance Profile Search	9
4.6 Select Genes using a Functional Profile	10
5. Review Functional Annotation for Genes of Interest	12
6. Manual Curation of Functional Annotations	15
6.1 MyIMG Curation	15
6.2 Product Name Specification	17
6.3 Missing Genes	19
7. Examine Different Gene Models and Product Names ..	21
7.1 Annotation Overview	21
7.2 Individual Genome Model Comparison	24
7.3 Gene Model Differences Examined with the Phylogenetic Profiler	27
7.4 Examining Product Names	28
References	29

Glossary of Terms.

Enzyme – a protein catalyzing a biochemical transformation (i.e. accelerating a chemical reaction).

Fusion - hybrid gene formed from two previously separate genes. **Component** of a Fusion gene are the separate genes, while **composite** gene is the result of the gene fusion.

Gene (or protein) annotation – description of the gene or protein product in the molecular, cellular and phenotypic context (e.g., interactions of a protein with other proteins or metabolites, participation of a protein in a biochemical pathway or effect of gene knock-out on the phenotype of an organism).

Genome context of a gene – a set of parameters defining the spatial position of a gene on the chromosome or a plasmid in a certain genome, including its co-localization with other genes, regulatory elements in its proximity, location of a gene on the leading or lagging DNA strand, etc.

Gene symbol – a unique abbreviation of a gene name consisting of italicized uppercase Latin letters and Arabic numbers, assigned after a gene has been identified.

Locus tag – a systematic gene identifier that is assigned to each gene in a [Genbank](#) file. For details see http://www.ncbi.nlm.nih.gov/Genbank/genomesubmit.html#locus_tag

Metabolism – a set of chemical transformation taking place within a living cell, multicellular organism or a microbial community.

Metabolic network – a representation of metabolism as a graph with nodes corresponding to metabolites and edges representing the reactions (or enzymes catalyzing the reactions).

Metabolic pathway – a set of consecutive biochemical transformations (enzymatic and spontaneous reactions) taking place in a living cell.

Homologous genes (homologs) – genes with sequence similarity (either at the level of nucleotide sequence or at the level of amino acid sequence of their protein products) due to their shared ancestry.

Orthologous genes (orthologs) – genes with sequence similarity separated by speciation events or vertically inherited genes: if a gene existed in a species, which gave rise to two species, then the divergent copies of the gene in the resulting two species are orthologous.

Paralogous genes (paralogs) – genes with sequence similarity separated by duplication events.

Operon – a group of genes sharing the common regulatory elements ([promoter](#), [operator](#), [terminator](#)) and transcribed as a unit to produce a single [messenger RNA](#).

Regulon – a group of genes and operons in an organism under regulation of the same regulatory protein.

A detailed **Glossary of Terms** is available at: <http://ghr.nlm.nih.gov/ghr/page/Glossary>

1. Synopsis

IMG EDU is a stand alone system that provides support for undergraduate and graduate level courses in microbial genome analysis and annotation. The IMG EDU genome content baseline consists of all the isolate genomes in IMG 2.3. In addition, users can submit for inclusion into IMG EDU a new genome of interest in which the genes were predicted and/or product name assignments have been applied using one of the available public microbial genome annotation services, such as Glimmer (Salzberg & al 1998) and GeneMark (Besemer & Borodovsky 2005) available at NCBI's Microbial Genome Annotation sites¹.

Genomes can be loaded into IMG EDU with product names assigned to genes prior to inclusion into IMG EDU by a specific annotation pipeline. Alternatively, the genomes can be loaded with no product assignments, whereby all proteins are annotated as "hypothetical protein". IMG EDU provides support for curation of protein product names and a number of associated functional annotations, using IMG's MyIMG capabilities. The purpose of this document is to present IMG's comparative analysis and annotation capabilities that support curation of product descriptions (functional annotations) associated with genes.

IMG EDU is available at: <http://img.jgi.doe.gov/edu>

2. Getting Started

Users not familiar with IMG's analytical tools can peruse the following papers and documents:

- A brief introduction to IMG is provided in (Markowitz & al 2008).
- Details on various IMG system components and content are available at: http://img.jgi.doe.gov/pub/doc/about_index.html.
- A user guide for IMG is available at: http://img.jgi.doe.gov/pub/doc/using_index.html.

Genome annotation problems are discussed in (Salzberg 2007). Additional documents on various aspects of microbial genome sequence data processing and analysis are available at: <http://www.jgi.doe.gov/education/microbialworkshop/>.

¹ See http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi or <http://www.ncbi.nlm.nih.gov/genomes/MICROBES/genemark.cgi>.

3. Genome Context for Annotation

Microbial genome annotation generally refers to the process of interpreting the raw sequence data with respect to the biological properties of an organism, by identifying protein-coding sequences and other genome features and determining their likely physiological functions.

Gene annotation is usually based on a combination of (i) automated methods that generate a “preliminary” annotation in terms of predicted protein-coding genes (also called Coding Sequences or CDSs) and (ii) assignment of gene product names. The latter is generally based on sequence similarity searches and may describe biological functions of gene products, such as enzymatic activity or participation in certain macromolecular complex, or merely indicate its membership in a certain sequence-similarity based protein family.

In addition to assignment of product names, “preliminary” annotation may suggest the placement of a gene product in various pathways and/or functional categories. While it is possible to perform manual assignment of product descriptions for all proteins in the genome, in most cases this is not necessary, since many protein families, especially those representing the housekeeping functions and core biosynthetic machinery, can be annotated with sufficient accuracy by virtually any annotation pipeline. Conversely, genes encoding members of certain enzymatic families with common catalytic mechanisms (aminotransferases, dehydrogenases, glycosyltransferases, etc.) are notoriously difficult in terms of their functional annotation and may require careful manual analysis and editing of their product descriptions.

In general microbial genome data analysis and annotation rely on comparison (sequence, chromosomal context, etc.) of the genes and genomes of interest against other genes and genomes. Although microbial genomes can be analyzed in the context of all other genomes available in IMG, it is often useful to limit this context to a certain subset of genomes. Genome (organism) selections help focus the analysis on a subset of interest, especially in terms of phylogenetic relationships.

Setting a genome context involves the following two stages:

(a) **Select genomes**

- a.1. Start with **Find Genomes** and select **Genome Browser**;
- a.2. Select **View Alphabetically** or **View Phylogenetically**;
- a.3. Select the genomes of interest, potentially after a **Clear All** on existing or default selections.

Example 1(i). Under **Find Genomes** in the Main Menu, use the **Genome Browser** with **View Phylogenetically** the list of genomes. First clear all default selections and then select from *Archaea* genomes, the two *Thermoplasmataceae* strains, *Thermoplasma volcanium* GSS1 (*T. volcanium*) and *Thermoplasma acidophilum* DSM 1728 (*T. acidophilum*), shown in Figure 1(i). Save these selections.

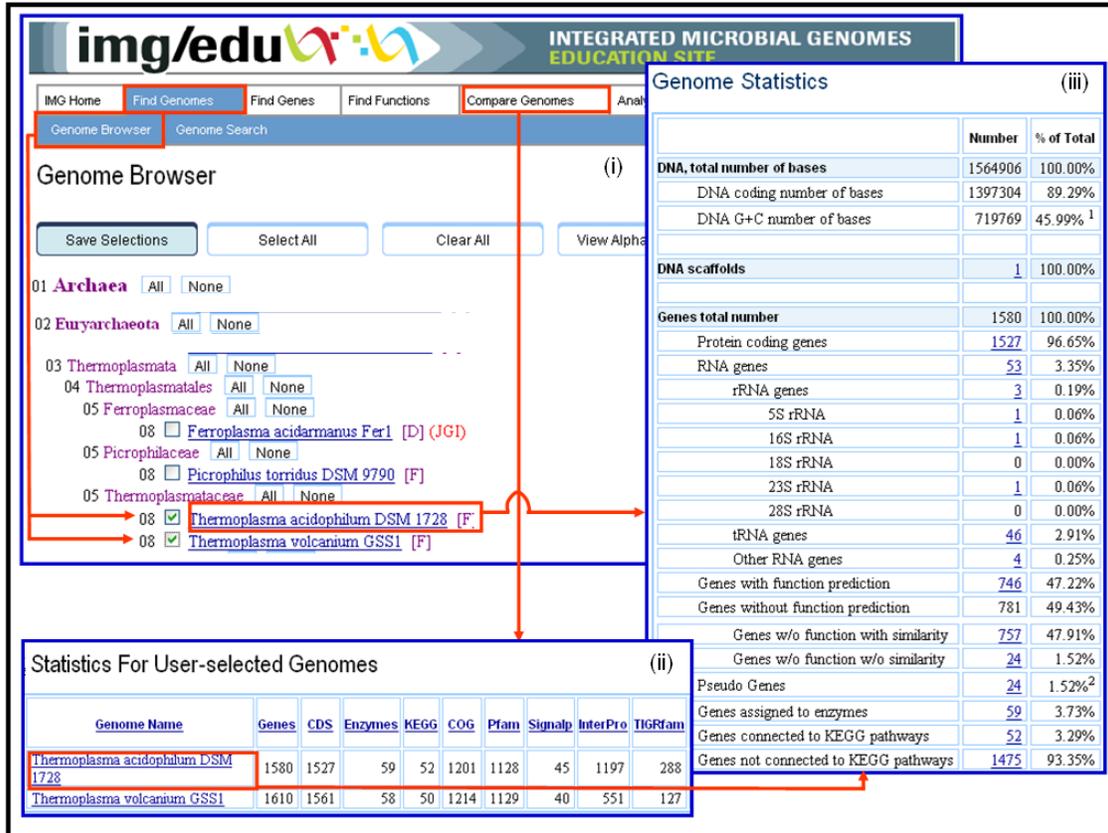


Figure 1. Setting a genome context and examining statistics for selected genomes.

(b) When analyzing a genome with “preliminary” automated annotation it is often useful to assess the quality of this annotation by comparing the general statistics of annotations (such as the total number of predicted protein-coding genes, the number of proteins with functional annotation or with membership in different protein families) for the genome of interest to that of its closest phylogenetic neighbors, since gross discrepancies in this statistics may indicate certain problems of automated annotation.

Assess annotation coverage for selected genomes via

b.1. **Statistics for User Selected Genomes** which can help in identifying differences in the total number of genes, CDSs, as well as in functional annotations based on various functional classifications, such as COG, Pfam.

Example 1(ii). Select **Compare Genomes** in the Main Menu, and then select **Statistics for User Selected Genomes**. You can adjust the columns that are displayed in the comparative statistics table. Notice the differences in genes, CDSs, and functional annotations between *T. volcanium* and *T. acidophilum*, as shown in Figure 1(ii).

b.2. **Genome Statistic** of individual genomes, available under Organism Details which can be reached from the list of genomes in the **Genome Browser** or **Statistics for User Selected Genomes**.

Example 1(iii). In the **Statistics for User Selected Genomes** (see Figure 1(ii)) select *T. acidophilum* in order to view the **Genome Statistics** table for this genome, as shown in Figure 1(iii).

4. Select Target Genes for Annotation

A number of tools are available in IMG-EDU for selecting target genes for further analysis and functional annotation; some of them do not require setting up a genomic context for analysis, while others operate on subsets of genomes only and require setting the genome context prior to their use, as described in the previous section.

4.1 Select Genes using Gene Search

Genes can be selected from one or several specific genomes using **BLAST** or **Gene Search**.

BLAST similarity searches are implemented via BLASTp (protein-vs.-protein), BLASTx (DNA-vs.-protein), BLASTn (DNA-vs.-DNA) or tBLASTn (protein-DNA-vs.-DNA-protein). You can define similarity thresholds and select the target database or genomes. For more details, refer to **Using IMG** manual.

Gene Search allows selecting genes based on partial or exact matches to a string of characters in specified fields, including product name, gene symbol, and a variety of other gene identifiers, as illustrated in Figure 2. Note that if a genome context is set, then this search will be performed on the genomes in this subset only.

The screenshot shows the IMG/EDU Gene Search interface. The search keyword is 'NADH oxidase'. The filters are set to 'Product Name (inexact)'. The search results are displayed in a table with columns for Selection, Gene Object ID, Match Text, and Genome. Two genes are selected: 638180327 (NADH oxidase related protein) and 638190735 (NADH oxidase). The Gene Cart (iii) shows 2 gene(s) in cart, with columns for Selection, Gene Object ID, Locus Tag, Product Name, AA Seq. Length, and Genome. The selected genes are 638180327 (Ta0162, NADH oxidase related protein, 536aa) and 638190735 (TVG0244368, NADH oxidase, 547aa).

Gene Product Name Results (ii)

Selection	Gene Object ID	Match Text	Genome
<input checked="" type="checkbox"/>	638180327	NADH oxidase related protein	Thermoplasma acidophilum DSM 1728
<input checked="" type="checkbox"/>	638190735	NADH oxidase	Thermoplasma volcanium GSS1

Gene Cart (iii)

Selection	Gene Object ID	Locus Tag	Product Name	AA Seq. Length	Genome
<input checked="" type="checkbox"/>	638180327	Ta0162	NADH oxidase related protein	536aa	Thermoplasma acidophilum DSM 1728
<input checked="" type="checkbox"/>	638190735	TVG0244368	NADH oxidase	547aa	Thermoplasma volcanium GSS1

Figure 2. Select genes for further analysis and functional annotation with **Gene Search**.

Example 2. Under **Find Genes** in the Main Menu, select **Gene Search**. If you have selected and saved genomes *T. volcanium* and *T. acidophilum* using the Genome Browser as discussed in Example 1, the gene search is restricted by default to these genomes. Gene search can be also restricted to specific genomes via the genome list provided in the **Gene Search** page, as shown in Figure 2(i). The search for product names containing “NADH oxidase” returns two genes, as shown in Figure 2(ii). The genes can be save into the Gene Cart, as shown in Figure 2(iii), for further analysis or curation.

4.2 Select Genes from a Specific Genome Statistics Table

Genes can be selected from a specific genome from one of the gene categories provided in the **Genome Statistics** table. Gene counts in some categories, such as “Genes in COGs” and “Genes connected to KEGG pathways” are linked to tables that show these genes classified according to the corresponding functional hierarchies (COG Functional Categories and KEGG Categories, respectively, with the possibility of further breakdown into COG Pathways and KEGG Maps). Grouping genes into such functional categories provides a convenient separation of the housekeeping genes that are quite likely to be accurately annotated by an automated pipeline from those participating in other metabolic activities and allows focusing further analysis on one or several functional categories of interest.

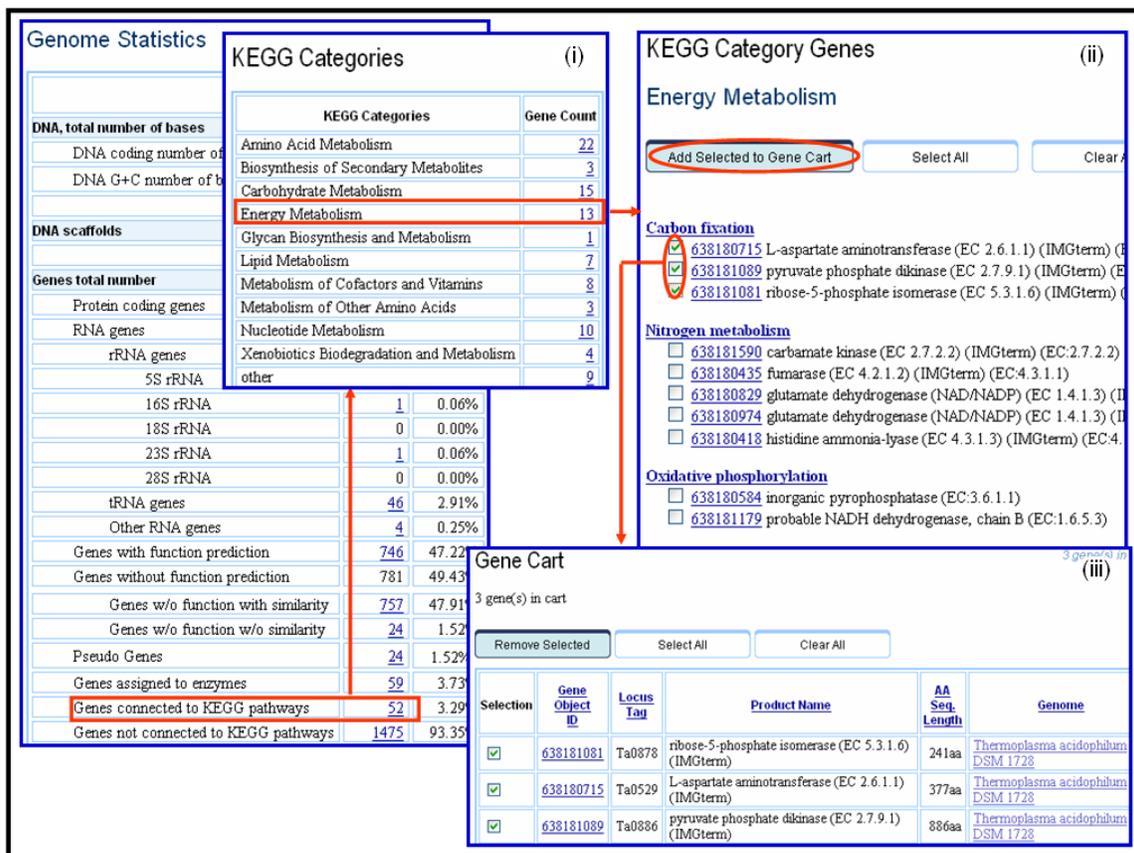


Figure 3. Select genes for further analysis and functional annotation via the gene categories available in the **Genomics Statistics** table.

The gene counts in the “KEGG Categories” table are linked to the table that contains groupings of genes according to individual KEGG Maps corresponding to collections of metabolic pathways. These maps can be examined for “completeness” of coverage with genes from the genome of interest. For example, you can check whether all consecutive reactions in a pathway are connected to the genes in a specific genome.

Example 3. Follow the link for **Genes connected to KEGG pathways** from the **Genome Statistics** table shown in Figure 1(iii) (see also left pane of Figure 3), which leads to a further classification of genes based on their association with specific KEGG categories, as shown in Figure 3(i). For a specific KEGG category, such as *Energy Metabolism*, a link is provided to the list of genes organized in specific pathways, such as *Carbon Fixation*, as shown in Figure 3(ii). Genes can be then selected and saved in the **Gene Cart** for further analysis, as shown in Figure 3(iii).

4.3 Select Genes by Comparing Gene Annotations

Genes can be selected from a specific genome by reviewing the various gene annotations using the **Compare Gene Annotations** page that can be reached from the **Organism Details** via the Compare Gene Annotations button located below **Genome Statistics** (see Figure 4).

Genome Statistics

Compare Gene Annotations (i) Loaded

View annotations for *Thermoplasma acidophilum DSM 1728*.

Pages for [download file 638154521.annot.xls](#) 1 [2] 3 4 [Next]

Gene ID	Locus Tag	Source	Cluster	Annotation
638180701	Ta0515	COG0541	Signal recognition particle GTPase	
638180701	Ta0515	pfam00448	SRP54	
638180701	Ta0515	pfam02881	SRP54_N	
638180701	Ta0515	pfam02978	SRP_SPB	
638180701	Ta0515	product_name		probable signal recognition particle protein
638180701	Ta0515	ITERM_01969		signal recognition particle subunit FFH/SRP54 (srp54)
638180701	Ta0515	DNA_length		1371bp
638180701	Ta0515	Protein_length		456aa
638180702	Ta0516	COG0644	Dehydrogenases (flavoproteins)	
638180702	Ta0516	pfam00070	Pyr_redox	
638180702	Ta0516	pfam01360	Monooxygenase	
638180702	Ta0516	product_name		geranylgeranyl reductase related protein
638180702	Ta0516	DNA_length		1170bp
638180702	Ta0516	Protein_length		389aa

Paralogous groups 190 100.00%

Orthologous groups 1434 0.77%

Chromosomal Viewer Tools

Scaffolds and Contigs

Chromosome Maps

Compare Gene Annotations

[Compare Gene Annotations](#)

638154521.annot (ii)

	A	B	C	D	E
1	gene_oid	Locus Tag	Source	Cluster	Annotation
2	6.38E+08	Ta0001	COG1390	Archaeal/vacuolar-type H+-ATPase subunit E	
3	6.38E+08	Ta0001	pfam01991	vATP-synt_E	
4	6.38E+08	Ta0001	product_name		ATP synthase (subunit E) related protein
5	6.38E+08	Ta0001	DNA_length		568bp
6	6.38E+08	Ta0001	Protein_length		185aa

Figure 4. Select genes for further analysis and functional annotation using the **Compare Gene Annotations** table that can be reached via the **Organism Details** page.

Compare Gene Annotations provides a list of protein-coding genes from the genome of interest and their product names together with the information about their membership in various protein families (COG, Pfam, TIGRfam) and descriptions of their functions based on this membership. This tool provides a quick way of assessing the quality of automated annotation by comparing the names of protein products assigned by an automated pipeline to the likely functions suggested by the membership of a protein in protein families and identifying the most obvious discrepancies between the two (e. g., a product name of “probable signal recognition particle protein” assigned to a member of COG0644 “Dehydrogenases (flavoproteins)”). The genes with such discrepancies between product names and functional annotations based on protein family membership can be further analyzed through the individual **Gene Details** pages.

Example 4. Follow the link for **Compare Gene Annotations** via the button provided below the **Genome Statistics** table shown in Figure 1(iii) (see also left pane of Figure 4), which leads to a list of all the genes for the selected genome, with every available annotation provided for each gene, as shown in Figure 4(i). Examine the product name in the context of other available functional annotations, such as COG, pfams, and IMG term (when available). The annotations for all the genes can be downloaded into a local excel file, as shown in Figure 4(ii).

4.4 Select Genes using the Phylogenetic Profiler

In many cases the differences in physiology, phenotypic properties and ecology of different organisms can be attributed to the differences in their gene content, i.e. the differences in abundance of various gene families, including the ultimate case of certain genes being present in one genome but not in another genome(s) and vice versa. Therefore the genes identified as more or less abundant (or present or absent) when comparing the genome of interest to its genome context, often become the focus of microbial genome analysis and may require special attention from the annotator.

The **Phylogenetic Profiler** tool allows finding genes in a specific genome that have / do not have homologs in other related genomes. There are two steps involved in such a selection:

- (a) Start with **Find Genes** in the Main Menu and select **Phylogenetic Profiler**.
- (b) Select a target genome and set the condition for selecting its genes with respect to presence or absence of homologs in other related genomes.

Example 5. After setting the genome context to two genomes, *T. volcanium* and *T. acidophilum*, as discussed above, use the **Phylogenetic Profiler** to find *T. volcanium* genes that have no homologs in *T. acidophilum*, as shown in Figure 5. Similarity cutoffs can be used to fine-tune the selection. The list of genes with the specified profile are then provided as a selectable list as shown in Figure 5.

The **Phylogenetic Profiler** can be used, for example for finding *unique, conserved, or gained* genes in the target genome with respect to other genomes of interest. In the example shown in Figure 5, 237 genes are found to be unique in *T. volcanium* with respect to *T. acidophilum*.

Phylogenetic Profiler Results 237 gene(s) retrieved

237 genes remaining after subtracting genes with homologs in *Thermoplasma acidophilum* DSM 1728

Add Selected to Gene Cart Select All Clear All

Select	Result Row	Gene Object ID	Locus Tag	Gene Name	Length	COG	Enzyme	Pfam	InterPro
<input type="checkbox"/>	1	638190495	TVG0007048	hypothetical protein	168aa	COG0675	-	pfam07282	-
<input type="checkbox"/>	2	638190515	TVG0025913	hypothetical protein	258aa	COG0003	-	pfam02374	-
<input type="checkbox"/>	3	638190532	TVG0045165	transposase (IMGterm)	188aa	COG0675	-	pfam07282	IPR010095 IPR012337
<input type="checkbox"/>	4	638190537	TVG0049983	H ⁺ -transporting ATP synthase subunit K	75aa	-	-	-	-

Profile

Find Genes In'	With Homologs In	Without Homologs In	Ignoring
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Archaea

Euryarchaeota

Thermoplasma

[Thermoplasma acidophilum DSM 1728](#) [F]

[Thermoplasma volcanium GSS1](#) [F]

Similarity Cutoffs

Max. E-value: 1e-5

Min. Percent Identity: 30

Figure 5. Find genes with the *Phylogenetic Profiler* tool.

4.5 Select Genes using Abundance Profile Search

Genes can be selected using the *Abundance Profile Search* tool that allows finding genes in a specific genome that are associated with a function that is more / less abundant than in other related genomes. There are three steps involved in such a selection:

- Start with *Compare Genomes* in the Main Menu and select *Abundance Profiles*. Select the *Abundance Profile Search* tool.
- Select a target genome and set the condition for finding functions associated with its genes in terms of their relative abundance with regard to other related genomes.
- From the functions listed as results of the *Abundance Profile Search*, select functions of interest and save them in the *Function Cart* or follow the link for a specific function to the list of genes associated with it.

Example 6. After setting the genome context to two genomes, *T. volcanium* and *T. acidophilum*, as discussed above, use the *Abundance Profile Search* tool to find COGs that are more abundant in *T. volcanium* than in *T. acidophilum*, as shown in Figure 6(i). The result displays the abundance counts for all COGs across the two genomes, as shown in Figure 6(ii). Each abundance count provides a link to the associated list of genes, such as the genes associated with COG 1522, as shown in Figure 6(iii). Some of the COG representatives found in *T. volcanium*, such as COG1552 (*Ribosomal protein L40E*) shown in Figure 6(ii), have no match in *T. acidophilum*, which may be of evolutionary significance or explained by the fact that the genes were missed by the original annotation.

Abundance Profile Search Results (ii)

Selection	Cog Id	Cog Name	Thermoplasma volcanium GSS1	Thermoplasma acidophilum DSM 1728
<input type="checkbox"/>	COG0003	Oxyanion-translocating ATPase	2	1
<input type="checkbox"/>	COG0119	Isopropylmalate/homocitrate/citramalate synthases	1	0
<input type="checkbox"/>	COG0145	N-methylhydantoinase A/acetone carboxylase, beta subunit	1	0
<input type="checkbox"/>	COG0146	N-methylhydantoinase B/acetone carboxylase, alpha subunit	1	0
<input type="checkbox"/>	COG1522	Transcriptional regulators	0	5
<input type="checkbox"/>	COG1552	Ribosomal protein L40E	1	0

Abundance Profile Search Gene List (iii)

Cog Id: COG1522
Cog Name: Transcriptional regulators

Selection	Gene Id	Gene Name
<input type="checkbox"/>	638190805	leucine-responsive regulatory protein [Lrp]
<input type="checkbox"/>	638190892	leucine-responsive regulatory protein [Lrp]
<input type="checkbox"/>	638191191	leucine-responsive regulatory protein [Lrp]
<input type="checkbox"/>	638191727	leucine-responsive regulatory protein [Lrp]
<input type="checkbox"/>	638191805	leucine-responsive regulatory protein [Lrp]
<input type="checkbox"/>	638191954	leucine-responsive regulatory protein [Lrp]

Figure 6. Find genes with the *Abundance Profile Search* tool.

In addition to focusing on individual genes, annotation and analysis can also focus on certain protein families, since many of them may include multiple paralogs with different activities and biological roles that require manual curation.

4.6 Select Genes using a Functional Profile

The *Function Profile* tool allows to find all the genes in the genome(s) of interest associated with a specific protein family (COG, Pfam, etc.) and to compare the abundance of this family across multiple genomes. Many of these protein families can be unambiguously associated with certain activities or functions, although in many cases this functional description lacks the necessary precision. For instance, most members of Pfam00155 (Aminotransferase class I and II) catalyze transfer of amino group between two reactants. However, these reactants may be quite different (e. g., aromatic amino acids, aspartate, histidinol phosphate, etc.) and in addition to *bona fide* aminotransferases, this protein family also includes L-threonine O-3-phosphate decarboxylase and some other enzymes that share the overall structure, cofactor requirement and substrate-binding pocket with aminotransferases, yet catalyze very different reactions, such as decarboxylation instead of transamination. This example illustrates the need for manual analysis and comparison of the individual members of protein families, which can be selected using the *Function Profile* tool and their annotations can be reviewed through the corresponding *Gene Details* pages.

There are three steps involved in such a selection:

- (a) Start with **Find Functions** in the Main Menu and select either **Search** or one of the browsing options for the available functional classifications. Examine individual functional roles or categories.
- (a) Select functions of interest and save them in the **Function Cart**.
- (b) The **Function Cart** is available under **Analysis Carts** in the Main Menu. Select the functions and the genomes of interest in the **Function Cart**, and compute a **Function Profile**. The results will be displayed in a tabular format, with functions listed in either columns or rows.

Example 7. Under **Find Functions**, select the **COG** browser, as shown in Figure 7(i). Scroll down to *Ribosomal proteins – large subunits* COG pathway and examine it using the **COG Pathway Details**, as shown in Figure 7(ii). From this COG pathway select COG1552 and COG1631 and save them in the **Function Cart**, as shown in Figure 7(iii). In the **Function Cart**, select both COGs and the genomes (*T. volcanium* and *T. acidophilum*) available in the genome list, and compute a **Function Profile** with the option of displaying the results in the *Genomes vs. Function* format, as shown in Figure 7(iv). The count of genes associated with a specific COG in a given genome is shown in a cell of the tabular result: the count serves as a link that leads to the actual list of genes. Similar to the result of the Abundance Profile Search (see Figure 6(ii)), COG1552 is present in *T. volcanium* but not in *T. acidophilum*, which may indicate a gene missed by the original annotation.

The screenshot shows the IMG/EDU interface with several key components:

- Navigation Bar:** Includes 'img/edu' logo and 'INTEGRATED MICROBIAL GENOMICS EDUCATION SITE'. Menu items include 'Find Genomes', 'Find Genes', 'Find Functions' (highlighted), 'Compare Genomes', and 'Analysis Carts'.
- COG Browser (i):** A list of COG categories such as 'Amino acid transport and metabolism [E]', 'Translation, ribosomal structure and biogenesis [J]', and 'Ribosomal proteins - large subunit'.
- COG Pathway Details (ii):** Shows details for 'Ribosomal proteins - large subunit'. A table lists COG IDs and names with genome counts:

Select	COG ID	COG Name	Genome Count
<input type="checkbox"/>	COG0080	Ribosomal protein L11	2
<input type="checkbox"/>	COG0081	Ribosomal protein L1	2
<input type="checkbox"/>	COG0087	Ribosomal protein L3	2
<input type="checkbox"/>
<input checked="" type="checkbox"/>	COG1552	Ribosomal protein L40E	1
<input checked="" type="checkbox"/>	COG1631	Ribosomal protein L44E	2
- Function Cart (iii):** Shows '2 function(s) in cart'. A table lists selected COGs:

Function ID	Name	Batch ¹
COG1552	Ribosomal protein L40E	1
COG1631	Ribosomal protein L44E	1
- Function Profile (iv):** Displays a table comparing the selected COGs across two genomes:

	COG 1552	COG 1631
Thermoplasma acidophilum DSM 1728	0	1
Thermoplasma volcanium GSS1	1	1

Figure 7. Find genes with a **Function Profile** tool.

5. Review Functional Annotation for Genes of Interest

The functional annotation for individual genes selected using a variety of tools described in the previous section can be reviewed using the **Gene Details** page (see Figure 8) that is available via the link provided from individual gene identifiers (so called OIDs). The **Gene Details** page consists of:

- (a) a **Gene Information** section that includes gene identification, locus information, product name, and related information; of special interest in this section are the **Product Name** which is usually assigned by the sequencing center that has processed the genome, and various **External Links** to public resources (when available) that allow viewing different representations of the gene;
- (b) a **Protein Information** section that includes functional characterization based on COG, Pfam, InterPro, and native IMG terms;
- (c) a **Pathway Information** section that includes associated Enzymes (EC numbers), KO term, and KEGG and IMG native pathways; when a specific functional annotation is available, links to details, such as KEGG maps or IMG pathways, are provided;
- (d) an **Evidence for Function Prediction** section that includes a **Chromosome Viewer**, **Gene Ortholog Neighborhood** viewer, alignment of the gene sequence to the COG and Pfam representative sequences (centroids), and pre-computed lists of homologs, orthologs and paralogs.
- (e) A variety of **BLAST** and similarity **search** tools.

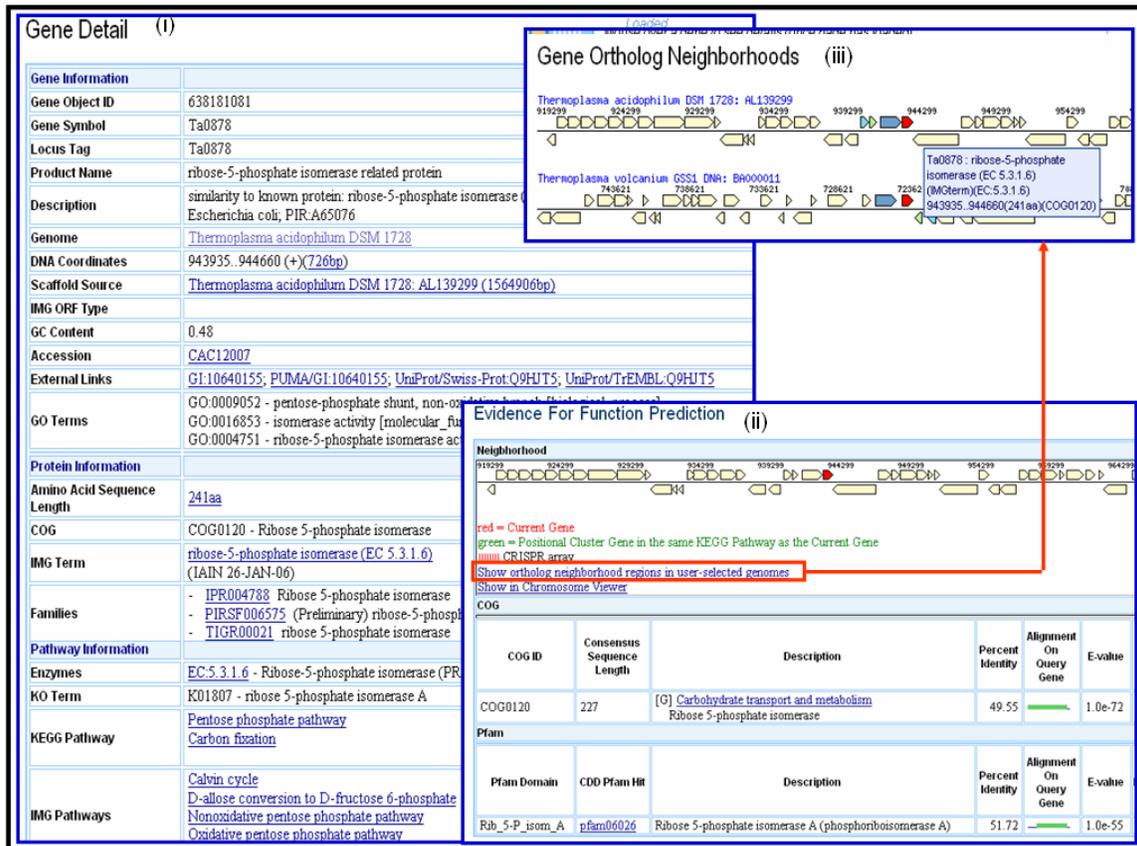


Figure 8. Review functional annotations for a gene with **Gene Details**.

The purpose of reviewing the information provided on the **Gene Details** page is similar to that of the **Compare Gene Annotations** tool, i.e. it aims at identifying discrepancies between the product name assigned to a particular gene, functions suggested by the membership of this protein in various protein families and/or chromosomal clusters, and product names of its closest homologs, as these may be assigned not based on bioinformatics inference, but according to experimental data about the activity and function of these proteins.

Example 8. Start with the steps in Example 1, which lead to the selection of *T. acidophilum* and its **Genome Statistics**, as shown in Figure 1(iii). Follow the link for **Genes connected to KEGG pathways** from the **Genome Statistics** page shown in Figure 1(iii) which leads to a further classification of genes based on their association with specific KEGG categories, as shown in Figure 3(i). Follow the link *Energy Metabolism* to the list of genes organized in specific pathways, and select *Carbon Fixation*, as shown in Figure 3(ii).

For the **gene** with identifier **638181081**, follow the link to its **Gene Details** page, as shown in Figure 8(i). Examine the various sections of the **Gene Details** page, follow links to see related functional annotations in resources such as UniProt.

Examine **Evidence for Function Prediction** section: the gene neighborhood panel displays the target gene with its neighboring genes in a 25kb chromosomal window with the target gene in the center highlighted in red (see Figure 8(ii)).

Select *Show ortholog neighborhood regions in user-selected genomes* in order to display the *Gene Ortholog Neighborhoods* viewer, as shown in Figure 8(iii). This viewer displays the gene neighborhood of gene **638181081** in *T. acidophilum* aligned with its ortholog in *T. volcanium*: each gene's neighborhood appears above and below a single line showing the genes reading in one direction on top and those reading in the opposite direction on the bottom; genes with the same color indicate association with the same COG group. For each gene, locus tag, scaffold coordinates, and COG group number are provided locally (by placing the cursor over the gene), while additional information is available in the **Gene Details** that can be reached from each gene.

Example 9. The **Phylogenetic Profiler** tool mentioned above allows identifying quickly the unique and common genes between *T. volcanium* and *T. acidophilum* and the result indicates that *T. volcanium* has 237 unique genes (see Example 5). This high number of unique genes (about 15% of the total number of its predicted genes) suggests that a large percentage of the coding capabilities of *T. volcanium* is distinct compared to *T. acidophilum*. However, examining two of these genes using IMG's **Ortholog Neighborhoods**, as illustrated in Figure 9, shows that some of the differences in gene content between *T. volcanium* and *T. acidophilum* are due to inconsistencies of the gene models:

- (a) The first gene (see Figure 9(i)) is a subunit of ATP synthase, which is an integral membrane protein complex and is a function that is essential in almost all organisms. This gene is missing in *T. acidophilum*. tBLASTn of the *T. volcanium* gene can be used to find the missing gene in *T. acidophilum*, whereby Artemis is used to reveal that the gene is missing the first two amino acids, probably because it occurs at the very beginning of the genome sequence.



Figure 9. Using *Gene Ortholog Neighborhoods* to examine two *T. volcanium* genes that seem to be missing in the *T. acidophilum* genome.

(b) The second gene (see Figure 9(ii)) is a 50S ribosomal protein L40E (COG 1552) which is also an essential gene. Note that the *Abundance Profile Search* tool can be also used to detect such missing genes, as illustrated in Figure 6. Subsequently, tBLASTn of the *T. volcanium* gene can be used to identify the missing gene in *T. acidophilum*.

6. Manual Curation of Functional Annotations

In this section we will discuss the tools provided by IMG EDU for the manual curation of protein products and other related annotations.

6.1 MyIMG Curation

The functional annotation for individual genes can be curated using the **MyIMG Annotations** features of **MyIMG** (see Figure 11). In addition to curation of functional annotations, **MyIMG** provides support for uploading user genome selections that have been saved earlier from **Genome Browser** or **Genome Statistics** and set system wide user preferences.

MyIMG Annotations requires a login and password which can be requested through IMG's "Questions/Comments" link at left side lower corner of each IMG page. **MyIMG Annotations** provides support to:

1. **Edit** the functional annotation for one or several related genes. The following annotations can be manually edited:
 - Product Name
 - Protein Description
 - EC Number
 - PUBMED ID
 - Inference
 - Is Pseudo Gene?
 - Notes

Changes can be applied to genes that have been saved in the **Gene Cart** and have been selected for curation.

2. **Review** the functional annotations. Annotations can be reviewed:
 - For individual genes within the Gene Details MyIMG Annotation section.
 - For all the genes that have been curated, with
 - the annotations displayed in a tabular format, where each row consists of the annotations for a gene (see Figure 10(ii)).
 - In groups of genes per genomes, where annotation summaries per genome are displayed in a tabular format; select the list of genes for a specific genome to review their annotations (see Figure 10(iii)).
3. **Export** to/ **upload** from a tab-delimited file functional annotations for genes identified by their IMG identifier (OID). The file has the following column headers:
 - gene_oid: Gene Object ID;
 - MyIMG_Annotation: annotated product name;
 - MyIMG_Prot_Desc: annotated prot desc;
 - MyIMG_EC_Number: annotated enzyme EC number(s);
 - MyIMG_PUBMED_ID: annotated PUBMED ID(s);
 - MyIMG_Inference: annotated inference;
 - MyIMG_Is_Pseudogene: is pseudo gene?
 - MyIMG_Notes: user annotated free text notes.

Only "gene_oid" and "MyIMG_Annotation" columns are mandatory, with the rest optional.

Example 9. Consider for review gene PF1186 (IMG identifier 638173757) of genome *Pyrococcus furiosus* whose details are shown in Figure 10(i). This gene is associated with product name *NADH oxidase*, as shown in Figure 10(i), and as recorded in GenBank and RefSeq which can be viewed by following the appropriate External Links. The list of top homologs for the gene under review can be displayed via the Homologs section of its **Gene Details**, as shown in Figure 10(ii). Based on a recent study², it has been determined that the function for this gene is *NADPH:sulfur oxidoreductase*, and an expert review of the best homologs of this gene indicated that this product name also may be confidently applied to the top three homologs. The gene under review and these top homologs are added to the **Gene Cart**, as shown in Figure 10(iii). Next, the product name where is changed to *NADPH:sulfur oxidoreductase* using the **MyIMG Annotation** tool accessed from **Gene Cart**, as shown in Figure 10(iv). Other annotations (e.g., EC number) can be also modified. User annotations are stored in IMG and can be reviewed at any time using **MyIMG** viewing options, as shown in Figure 11.

The screenshot displays the MyIMG interface for gene PF1186. It is divided into several panels:

- Gene Details (i):** Shows gene information for PF1186, including its locus tag, product name (*NADH oxidase*), and description. A red circle highlights the product name.
- Top IMG Homolog Hits (ii):** A table listing top homologs with columns for Select, Homolog ID, Product Name, Percent Identity, Alignment On Query Gene, Alignment On Subject Gene, and Length. Three homologs are checked for selection.
- Gene Cart (iii):** A table showing the selected genes (PF1186, TK1299, PH0572, PAB0936) with their gene object IDs and locus tags.
- MyIMG Annotation for Selected Genes (iv):** A form where the product name for the selected genes is being updated to *NADPH:sulfur oxidoreductase*. Other fields like EC Number, PUBMED ID, and Inference are also visible.

Red arrows indicate the workflow: from the product name in Gene Details to the Gene Cart, and then to the MyIMG Annotation form.

Figure 10. Review and curation of product name for a gene of *Pyrococcus furiosus* using **MyIMG Annotation**.

² Schut, G.J., Bridger, S.L., Adams, M.W. (2007) Insights into the metabolism of elemental sulfur by the hyperthermophilic archaeon *pyrococcus furiosus*: characterization of a coenzyme a-dependent NAD(P)H Sulfur Oxidoreductase. *Journal of Bacteriology*.

The screenshot displays the IMG/EDU MyIMG Annotations interface. At the top, there is a navigation bar with tabs for 'IMG Home', 'Find Genomes', 'Find Genes', 'Find Functions', 'Compare Genomes', 'Analysis Carts', and 'MyIMG'. Below this, a secondary navigation bar includes 'MyIMG Home', 'My Genomes', 'Annotations', 'Preferences', and 'Logout'. The main content area is divided into two primary sections: 'View My Annotations' and 'Upload Annotations from File'. The 'View My Annotations' section includes a 'View My Annotations' button and a 'View My Annotations by Genomes' button. The 'Upload Annotations from File' section includes an 'Upload Annotations' button. A pop-up window titled 'My Annotations by Genomes (iii)' is open, showing a table of annotations with columns for 'Select', 'Taxon OID', 'Genome Name', and 'Genes'. Below the pop-up, a larger table titled 'My Annotations (ii)' is visible, with columns for 'Select', 'Gene OID', 'Genome', 'Original Product Name', 'Annotated Product Name', 'Annotated Prot Desc', 'Annotated EC Number', 'Annotated PUBMED ID', 'Inference', 'Is Pseudo Gene?', 'Notes', and 'Last Modified Date'. The table contains four rows of annotation data.

Select	Gene OID	Genome	Original Product Name	Annotated Product Name	Annotated Prot Desc	Annotated EC Number	Annotated PUBMED ID	Inference	Is Pseudo Gene?	Notes	Last Modified Date
<input type="checkbox"/>	638200851	Pyrococcus abyss GE5	NADH oxidase (noxA-1)	NADPH:sulfur oxidoreductase							Tue 07-Aug-2007 16:29:06
<input type="checkbox"/>	638173757	Pyrococcus furiosus DSM 3638	NADH oxidase	NADPH:sulfur oxidoreductase							Tue 07-Aug-2007 16:29:06
<input type="checkbox"/>	638186008	Pyrococcus horikoshii OT3	445aa long hypothetical NADH oxidase	NADPH:sulfur oxidoreductase							Tue 07-Aug-2007 16:29:06
<input type="checkbox"/>	638207469	Thermococcus kodakaraensis KOD1	NADH oxidase	NADPH:sulfur oxidoreductase							Tue 07-Aug-2007 16:29:06

Figure 11. View *MyIMG* Annotations.

6.2 Product Name Specification

The main challenge for assigning product names to genes is **consistency** of product names across genes and genomes. Product names are usually free-text descriptions of a gene's physiological role and may also refer to other characteristics of a protein, such as subcellular localization, genetic locus, etc. It is commonly accepted that proteins with the same function should be assigned with the same product name in each and every genome, and a number of **controlled vocabularies** (enumerated lists of unambiguous, non-redundant terms with definitions) are being used for this purpose.

An example of **controlled vocabularies** is the **Enzyme Nomenclature** (Bairoch 2000) which assigns a recommended name and a number (*EC number*) to each enzymatic activity based on the substrates and products of the reaction and its catalytic mechanism; these attributes of enzymatic reactions are also recorded in the Enzyme Nomenclature and can be viewed as definitions of *enzyme names*, which correspond to terms in this controlled vocabulary. Other controlled vocabularies include **UniProtKB/Swiss-Prot** controlled vocabularies (Bairoch & al 2005), **Gene Ontology** (GO Consortium 2004), annotations of **COG** clusters (Tatusov & al 1997), **Pfam** families (Sonnhammer et al. 1998) or **TIGRFAMs** (Haft & al 2001). Different resources take **different approaches** to the construction of their controlled vocabularies; for instance, Swiss-Prot maintains a controlled vocabulary of *keywords* that are classified into 10 categories, including *biological process*, *cellular component*, *developmental stage*,

domain and *molecular function*; these keywords are then assigned to proteins in the UniProt knowledgebase. Gene Ontology associates genes with *GO terms* that describe them in terms of *biological processes*, *cellular components* and *molecular functions*. Some but not all Swiss-Prot keywords are equivalent to certain GO terms and are mapped to them; similarly, mappings of EC numbers and annotations of COG clusters, Pfam and TIGRFAMs to GO terms are available, although there is no one-to-one correspondence between any of these controlled vocabularies.

In order to address problems with the inconsistencies of the protein product names as well as with the current functional classifications, genes in IMG are further annotated in IMG using a native collection of generic (protein cluster-independent) functional roles called **IMG terms** that are further defined by their association with generic (organism-independent) functional hierarchies, called **IMG pathway** (Ivanova & al 2007).

IMG Terms form a hierarchy, whereby the leaves of this hierarchy consist of functional roles for gene products (protein product descriptions) assigned to individual genes. These lower-level IMG Terms of type “*Gene Product*” can be directly associated with reactions, whereby they function as either “*Catalysts*” or “*Reactants*”. Alternatively, they can be assigned recursively as “children” of IMG Terms of type “*Protein Complex*”, thus indicating that they constitute subunits of a multi-subunit protein complex. A detailed discussion of the rationale for IMG terms and pathways and their specification is available at <http://img.jgi.doe.gov/pub/doc/imgterms.html>, as part of IMG’s online documentation. An example of an IMG term, together with its phylogenetic distribution across all the genomes in IMG, is shown in Figure 12.

img/edu

IMG Home Find Genomes Find Genes Find Functions Compare Genomes

Search COG Pfam KEGG Enzyme TIGRFam IMG Network

IMG Term Details

Term Information

Term Object ID	00112
Term	5-aminolevulinate synthase (EC 2.3.1.37)
Type	GENE PRODUCT
Definition	
Comments	
Enzymes	EC:2.3.1.37 5-aminolevulinate synthase.
Add Date	14-MAR-05
Modify Date	14-MAR-05
Modified By	THANOS
Number of Synonyms	3
Number of Genes	147
Number of IMG Reactions	1
IMG Pathways	00077 - 5-aminolevulinate synthesis via
IMG Parts List	

Genomes with Term

Phylogenetic Distribution

Domains(D): B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids, V=Virus
Genome Completion(C): F=Finished, D=Draft.

D	C	Genome	Gene Count
B	D	Aurantimonas sp. SI85-9A1	1
B	D	Caulobacter sp. K31	1
B	D	Oceanicola batsensis HTCC2597	1
B	D	Oceanicola granulosus HTCC2516	1
B	D	Parvibaculum lavamentivorans DS-1	1
B	D	Parvularcula bermudensis HTCC2503	1
B	D	Rhodobacterales bacterium HTCC2150	1

Phylogenetic Distribution

```

B . . .02 Proteobacteria (127)
B . . . .03 Alphaproteobacteria (125)
B . . . . .04 Caulobacterales (2)
B . . . . . .05 Caulobacteraceae (2)
B . . . . . . .06 Caulobacter (2)
B . . . . . . . .08 Caulobacter crescentus CB15 [F](1)
B . . . . . . . . .08 Caulobacter sp. K31 [D](1)
B . . . . . . .04 Parvularculales (1)
B . . . . . . .05 Parvularculaceae (1)
B . . . . . . .06 Parvularcula (1)
B . . . . . . . .08 Parvularcula bermudensis HTCC2503 [D](1)
B . . . . . .04 Rhizobiales (42)
B . . . . . .05 Aurantimonadaceae (3)
B . . . . . .06 Aurantimonas (1)
B . . . . . . .08 Aurantimonas sp. SI85-9A1 [D](1)
B . . . . . .06 Fulvimarina (2)
B . . . . . . .08 Fulvimarina pelagi HTCC2506 [D](2)
B . . . . . .05 Bartonellaceae
    
```

Term Hierarchy

Only terms with no sub-components are selectable.

00112 5-aminolevulinate synthase (EC 2.3.1.37)

Figure 11. An Example of an IMG Term.

IMG terms and pathways are currently specified by domain experts at DOE-JGI as part of the process of annotating specific genomes of interest, and are subsequently propagated throughout the system.

6.3 Missing Genes

The review of genes and their functional annotations may lead to the identification of missing genes, as discussed in Example 9 above.

After determining that a gene x of a genome G is missing because of a similar gene, x' in a closely related genome G' , you can use **Artemis** (Rutherford & al 2000) to fill in the missing gene as follows:

1. Pick the sequence for gene x' from and run TBLASTn against genome G where you want to find the missing gene.
2. If you get a TBLASTn hit, copy part of the sequence and paste it into the Artemis navigator in the box labeled "Find Amino Acid String", as illustrated in Figure 12(i). The navigator is under the "Go to" menu. Then click on the "Goto" button.
3. The amino acid sequence is now highlighted, as illustrated in Figure 12(ii). Go to the "Create" menu and select "Create feature from base range", as illustrated in Figure 12(iii).

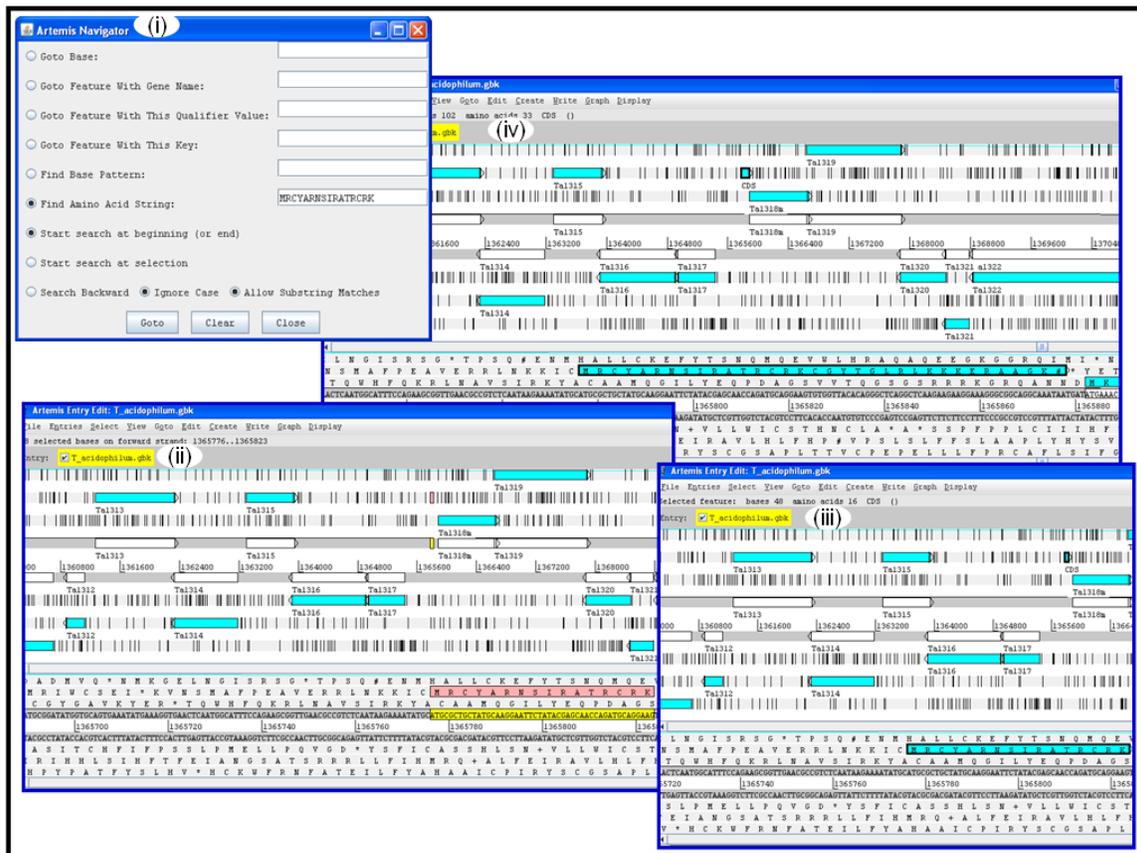


Figure 12. Using Artemis to define a missing gene.

4. To extend the gene, go to the "**Edit**" menu and select "*Extend to next stop codon*", then select "*Fix stop codons*".
5. To find the 5' end, under the "**Edit**" menu, click on "*Extend to previous stop codon*" (you can also use Control-Q for this), as illustrated in Figure 13(i).
6. To get the amino acid sequence, go to the View menu and select "View amino acid sequence as FASTA". BLAST the sequence against NCBI or IMG. Based on the aligned sequences, find where the start codon should be, as illustrated in Figure 13(ii).
7. Select the start codon by pressing "Control-Y". "Control-Y" moves the 5' end to the next potential start codon, as illustrated in Figure 13(iii).



Figure 13. Using Artemis to define a missing gene (cont.).

7. Examine Different Gene Models and Product Names

IMG/EDU contains four versions of the *Salinispora arenicola* (*S.arenicola*) genome that have identical nucleotide sequence, but differ in their annotation (gene finding and functional predictions). These versions of *S. arenicola* annotation were generated as follows:

- (1) **Sare-ORNL** version was generated using the automated data processing pipeline at Oakridge National Laboratory (ORNL), which uses a combination of CRITICA and Glimmer for predicting genes (Hauser et al 2004) followed by IMG native product name assignment tool;
- (2) **GeneMark** version was generated using automated gene prediction with GeneMark (Besemer and Borodovsky 2005) followed by IMG native product name assignment;
- (3) **Sare-RAST** version was generated using the RAST pipeline for gene prediction and assigning product names to genes based on SEED functional roles (<http://rast.nmpdr.org/>);
- (4) **Sare-manual** version was generated using automated gene prediction with GeneMark (2) followed by manual curation of gene models following the evidence-based procedure described at http://durian.jgi-psf.org/img_i_er/doc/dataprep.html and product name assignment using ORNL native tool (Hauser et al. 2004).

These four versions of the same genome enable users to compare the results of automated gene prediction using three different gene prediction tools with manually curated gene models, as well as evaluate three different pipelines for automated assignment of protein product names.

In this section we will show how the tools in IMG EDU can be used for comparing different genome annotations, including predicted stable RNAs and protein-coding gene models, and for examining product names generated using different pipelines.

7.1 Annotation Overview

Annotation pipelines differ in terms of both the type of objects they identify and the methods they employ to identify the same type of objects, such as identifying protein-coding genes. The prediction of protein-coding genes (CDSs) and assignment of their product names is common to all annotation pipelines. Most annotation pipelines identify stable RNA-coding genes, including tRNAs and rRNAs (5S, 16S and 23S). Some pipelines also identify other functional RNAs, including RNase P RNA component, signal recognition particle (SRP) RNA component, SsrS and many other regulatory RNAs (see RFAM, <http://www.sanger.ac.uk/Software/Rfam/>).

Genomes can be compared in terms of various annotation statistics using **Genome Statistics** available under **Compare Genomes** main menu option, as illustrated in the example below.

The screenshot shows the IMG/EDU website interface. Region (i) shows the 'Genome Search' page with a search bar containing 'Salinispora arenicola' and a 'Go' button. Region (ii) shows the search results for 'Salinispora arenicola CNS205' with five entries, four of which are selected. Region (iii) shows the 'Genome Statistics' page with 'General Statistics' selected. Region (iv) shows the 'General Statistics' page with 'Breakdown by selected genomes, general statistics' selected. Region (v) shows the 'Configuration' table with checkboxes for 'Genes', 'CDS', 'RNA', '5S', '16S', '23S', 'tRNA', and 'Other RNA' checked.

Figure 14. Select genomes and configure general statistics for comparing annotations.

Example 10. Select four versions of *Salinispora arenicola* using **Find Genomes** which provides several options for finding genomes of interest, including **Genome Search** (see Figure 14(i)). Enter “*Salinispora arenicola*” for the “Genome Name” filter and you will retrieve a list of 5 genomes, as shown in Figure 14(ii). One of these genomes represents a draft version of the same genome (on top), while the other four genomes are versions of the finished genome with different annotations. Select these four genomes and save the selections by clicking on “Save Selections” button.

Under the **Compare Genomes** main menu option select **Genome Statistics**, then follow the **General Statistics** link, as shown in Figure 14(iii). Select **Breakdown by selected genomes, general statistics** (see Figures 14(iv)) which will provide the default statistics table for the selected genomes. These statistics can be configured using the **Configuration** table, as shown in Figure 14 (v). This table allows you to choose the information you are interested to examine, ranging from the taxonomic lineage of an organism to detailed statistics of genome annotation. For this example select the following column names:

- 1) **Genes**, which represents the total number of genes in the genome including protein- and RNA-coding genes.
- 2) **CDS**, which represents the total number of protein-coding genes including pseudogenes.

- 3) **RNA**, which represents the total number of stable RNA-coding genes including rRNAs, tRNAs and other RNAs as discussed above.
- 4) **5S, 16S, 23S, tRNA and Other RNA**, which provide a breakdown of RNA-coding genes into different categories.
- 5) **Pseudo**, which represents the number of pseudogenes in the genome including those for protein-coding and RNA-coding genes.
- 6) **Unchar**, which represents the number of uncharacterized CDSs, that is CDSs without definitive functional prediction including all proteins with names “hypothetical”, “predicted” and “unknown”.
- 7) **w/Func Pred**, which represents the number of CDSs with definitive functional assignment including protein family annotations, such as “dehydrogenase family protein”.
- 8) **Enzymes**, which represents the number of CDSs with assigned EC numbers including incomplete EC numbers, such as EC:1.1.1.-.
- 9) **COG**, which represents the number of CDSs assigned to COG protein families.
- 10) **Pfam**, which represents the number of CDSs assigned to Pfam protein families.
- 11) **TIGRfam**, which represents the number of CDSs assigned to TIGRfam protein families.
- 12) **Bases**, which represents the total number of nucleotides in the genome sequence.
- 13) **Coding bases**, which represents the total number of nucleotides in predicted genes – if expressed as percentage of “Bases” gives so called “coding density” of the genome.

Note that for the columns 6-11 pseudogenes are excluded from the counts.

The table with the new configuration for general genome statistics can be displayed by clicking the “Display Genomes Again” button, as shown in Figure 15.

Loaded.

Statistics For User-selected Genomes

[Export Genome Table Configuration](#)

D	C	Genome Name	Genes	CDS	RNA	5S	16S	23S	tRNA	Other RNA	Pseudo	Unchar	w/ Func Pred	Enzymes	COG	Pfam	TIGRfam	Bases	Coding Bases
B	F	Salinispora arenicola CNS205 (GeneMark version)	5238	5177	61	3	3	3	52	0	0	0	3006	188	3258	3505	1062	5786361	4991013
B	F	Salinispora arenicola CNS205 (Sare-ORNL version)	4967	4906	61	3	3	3	52	0	0	0	3031	183	3279	3543	1064	5786361	5144934
B	F	Salinispora arenicola CNS205 (Sare-RAST version)	4644	4644	0	0	0	0	0	0	0	0	3747	1057	3136	3366	1021	5786361	5052705
B	F	Salinispora arenicola CNS205 (Sare-manual version)	5172	5111	61	3	3	3	52	0	190	0	3516	496	3256	3509	1064	5786361	5098757

Figure 15. Annotation statistics of four versions of *Salinispora arenicola* genomes.

The annotation statistics generated for the four versions of *Salinispora arenicola* genome indicate that different methods produce substantially different results, especially with regard to the prediction of protein-coding genes. For example, note that:

- (a) The total number of protein-coding genes varies by more than 500 CDSs, from 4,644 in the Sare-RAST version to 5,177 in the GeneMark version (see “CDS” column in Figure 15). However, it is unclear from these CDS counts whether the variation can be attributed to **higher sensitivity** (i.e. finding more true positive protein-coding genes) or to **lower specificity** (finding more false positive CDSs) for GeneMark compared to RAST.
- (b) **True positive CDSs** are expected to belong to sequence similarity-based protein families, such as COGs (Clusters of Orthologous Groups), Pfams or TIGRfams. The low count of genes assigned to COGs, Pfams and TIGRfams in the Sare-RAST version (e.g., see “COG” column in Figure 15), suggests lower sensitivity in RAST rather than a higher rate of false positives in other pipelines.
- (c) Despite having the highest number of predicted CDSs, the GeneMark version has the lowest number of **coding bases** (“Coding Bases” column in Figure 15). Thus, the Sare-ORNL version seems to have an extra 145 kb of coding sequence as compared to the GeneMark version despite having much fewer CDSs. The higher number of coding bases combined with the lower number of CDSs suggests that the ORNL annotation pipeline tends to predict longer genes than GeneMark, i.e. the two pipelines differ in their determination of the correct start codon with the ORNL pipeline using a more “greedy” approach. However, it is not clear from these statistics which pipeline identifies more correct start codons.

General statistics provide a preliminary comparison of different annotations. More detailed comparisons can be performed using **Chromosomal neighborhood** available on individual **Gene Pages** in IMG, as discussed below.

7.2 Individual Gene Model Comparison

Individual gene models can be examined using IMG’s **Gene Details**, as discussed in section 5 (see Figure 8). The Gene Details for a specific gene can be accessed from any list of genes, such as those discussed in section 4 (see Figures 2-7). In order to compare gene predictions generated by different annotation pipelines, genes can be selected from the **Organism Details** for a specific genome (e.g., see Figure 1), as discussed in section 4.2 and illustrated in Figure 3.

Example 11. Using **Genome Browser** or **Genome Search**, select the Sare-manual version of *Salinispora arenicola* genome and go to its **Organism Details** page (see Figure 16(i)). Follow the **Genome Viewers** link and select “Scaffolds and Contigs”, as shown in Figure 16(ii) in order to get to the **Chromosome Viewer** selection page, as shown in Figure 16 (iii). Selecting a range of coordinates will lead to the graphical Chromosome Viewer, as shown in Figure 16(iv). The interactive linear map of the genome is displayed in 500 kb fragments.

Organism Details (i)

[Organism Information](#)
[Genome Statistics](#)
[Genome Viewers](#)
[Export Genome Data](#)

Organism Information

Organism Name	Salinispora arenicola CNS205 (Sare-manual version)
Taxon Object ID	5000000002
NCBI Taxon ID	391037
NCBI Project ID	17109
GOLD ID	Gi01541
External Links	
Lineage	Bacteria; Actinobacte
Sequencing Status	Finished
Sequencing Center	JGI
Funding Agency	
Phenotype	
Habitat	Aquatic, Extremotype
Disease	
Relevance	
IMG Release	
Comment	JGI's Genome Porta

Genome Viewers (ii)

Scaffolds and Contigs
Chromosome Maps
Web Artemis

Chromosome Viewer (iii)

Scaffolds and contigs for Salinispora arenicola CNS205 (Sare-manual version)

User Selectable Coordinates

Scaffold	Length (bp)	GC	No. Genes	Coordinate Range
Salinispora arenicola (Sare-manual version) : sareManual_ctg2085	5786361	0.70	5172	1..500000 500001..1000000 1000001..1500000 1500001..2000000

Chromosome Viewer (iv)

Salinispora arenicola (Sare-manual version) : sareManual_ctg2085 (5786361bp gc=0.70)
(coordinates 1-500000)

hint: Mouse over a gene to see details.

Sare_0008 : DNA gyrase, A subunit [L]
9641..12160(839aa)

Figure 16. Selecting genes with the **Chromosome Viewer**.

Mouse over individual genes to see more details about each gene, as shown in Figure 16(iv). Click on a specific gene, such as gene “Sare_0008 DNA gyrase, A subunit”, to get to its **Gene Detail** page, as shown in Figure 17(i), which provides information about the gene, as discussed in section 5.

Follow the link “Evidence For Function Predictions” which leads to the graphical display of the chromosomal neighborhood of the query gene, as shown in Figure 17(ii). Next, follow the link “Show neighborhood regions with the same top COG hit”, which leads to a graphical display of the chromosomal neighborhoods of the genes with the same COG in this and the other selected *Salinispora* genomes, as shown in Figure 17(iii).

Gene Neighborhoods viewer shown in Figure 17(iii) displays the same chromosomal region near the origin of replication, centered around the gene “Sare_0008” (colored red), which was selected for exploration. While the genes in the four versions of the *Salinispora* genomes seem to be similar, there are notable discrepancies indicated by the red arrows. For instance:

- (a) the Sare-ORNL version contains a gene on the opposite strand of the query gene, “Sare_008”. This gene (so called “shadow” gene) is likely to be a false positive.
- (b) As seen from the examination of the general annotation statistics, the Sare-RAST version has fewer genes than the other versions of the genome.

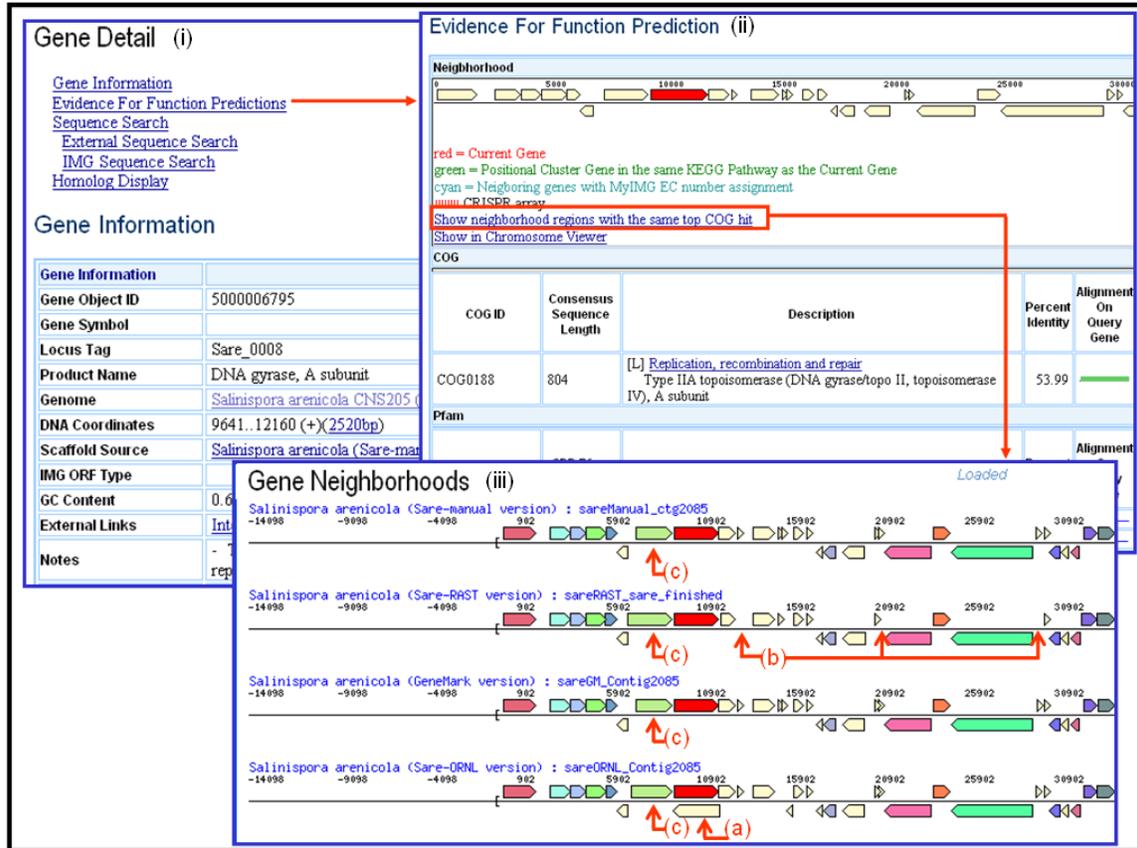


Figure 18. Exploring gene model differences with **Gene Neighborhoods**.

- (c) The gene models generated by the automated pipelines have different start codons for the gene next to “Sare_0008” (green colored gene), with the genes in Sare-ORNL and Sare-RAST versions longer than the analogous gene in the GeneMark version. Note that this gene in the GeneMark version is consistent with analogous the gene in the manual version, in which alignment of this gene against its best homologs has been examined and the start codon has been selected accordingly.

7.3 Gene Model Differences Examined with the *Phylogenetic Profiler*

The gene models generated using different methods and pipelines can be compared with the *Phylogenetic Profiler* as illustrated by the example shown in Figure 19(i).

Example 12. Select and save four versions of *Salinispora* genomes, as discussed in Example 10 above.

Select *Phylogenetic Profiler* under the *Find Genes* main menu option. *Phylogenetic Profiler* allows selection of the genes present/absent in one or more versions of the *Salinispora* genomes. Thus, you can find the genes that were predicted similarly in all the versions of the *Salinispora* genomes or you can find the genes that were predicted in one but not in the other versions of the genome.

In order to find the genes predicted by GeneMark but not by the ORNL pipeline, first select the GeneMark version as the query genome by using the “Find Genes In” radio button, and then select the Sare-ORNL version as a reference genome with no homologs to the genes in GeneMark version by using the “Without Homologs In” radio button. With the default settings for similarity cutoffs, as shown in Figure 19(i), the *Phylogenetic Profiler* tool finds 380 genes in the GeneMark version of the *Salinispora* genome with no match in the Sare-ORNL version of the genome.

The screenshot shows the IMG/EDU Phylogenetic Profiler interface. The main window displays 'Phylogenetic Profiler Results' for 380 genes. A table lists genes with columns for Select, Result Row, Gene Object ID, Locus Tag, Gene Name, Length, COG, Enzyme, and Pfam. Gene 324 (Gene Object ID 5000016484) is highlighted with a red box. Below the table, a 'Salinispora' section lists four genome versions: GeneMark, Sare-ORNL, Sare-RAST, and Sare-manual. The 'Gene Information' panel for gene 5000016484 shows details such as Locus Tag (sare_4525), Product Name (phenylacetate-CoA oxygenase, PaaH subunit), and Genome (Salinispora arenicola CNS205 (GeneMark version)).

Select	Result Row	Gene Object ID	Locus Tag	Gene Name	Length	COG	Enzyme	Pfam
<input type="checkbox"/>	1	5000011970	sare_11	hypothetical protein	59aa	-	-	-
<input type="checkbox"/>	2	5000011971	sare_12	hypothetical protein	77aa	-	-	-
<input type="checkbox"/>	323	5000016443	sare_4484	hypothetical protein	113aa	-	-	-
<input type="checkbox"/>	324	5000016484	sare_4525	phenylacetate-CoA oxygenase, PaaH subunit	106aa	COG3460	-	pfam06243

Find Genes In'	With Homologs In	Without Homologs In	Ignoring
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Max. E-value	1e-5
Min. Amino Acid Percent Identity	30
Algorithm	By Present/Absent Homologs
Min. Taxon Percent With Homologs	100
Min. Taxon Percent Without Homologs	100

Gene Information	
Gene Object ID	5000016484
Gene Symbol	
Locus Tag	sare_4525
Product Name	phenylacetate-CoA oxygenase, PaaH subunit
IMG Product Source	TIGR.fam
Genome	Salinispora arenicola CNS205 (GeneMark version)

Figure 19. Comparing gene models with the *Phylogenetic Profiler* and examining product names with *Gene Details*.

The genes found with the **Phylogenetic Profiler** can be further explored through their **Gene Details** pages as discussed above. Note that if your query gene does not have a COG or if the gene has no homologs in any other version of the *Salinispora* genomes, the link “Show neighborhood regions with the same top COG hit” will be inactive. In this case a neighbor gene in the chromosomal neighborhood should be selected for display of **Gene Neighborhoods** graphical overview.

7.4 Examining Product Names

Product names are assigned to genes using various methods based on gene similarity and involving one or several functional classifications, such as COG, Pfam, and TIGRfam. For example, the IMG product name assignment procedure involves a sequence of stages, each applied on the genes that have no product name assigned at the end of the previous stage. Thus, a gene *x* that has no product name (i.e. is associated with a default “hypothetical protein” name) is assigned one of the following product names: (a) IMG term(s) associated consistently with at least two close³ homologs of *x*; (b) the name of TIGRfams, COGs, or Pfams that were associated with *x* by IMG’s functional annotation pipeline; or (c) the product name(s) associated consistently with at least two close homologs of *x*. Multiple components (e.g., multiple IMG terms, Pfam family descriptions, etc.) in a product name are concatenated using “/” as a separator.

The product names associated with genes can be examined using the **Gene Details** as discussed in Example 8 and illustrated in Figure 8 above. In particular, the consistency between a gene’s product name and its functional annotations (e.g. IMG term, COG, Pfam, TIGRfam, etc.) can be examined using the various **Gene Details** sections. For product names assigned using the IMG procedure mentioned above, **Gene Details** provides an “IMG Product Source” field as shown in Figure 14(ii).

Example 13. After using the **Phylogenetic Profiler** to find genes in the GeneMark version of the *Salinispora* genome that have no homologs in the Sare-ORNL version of the genome, as discussed in Example 12 above, examine the list of **Phylogenetic Profiler Results**. Most of these genes have no assigned product name (i.e., have “hypothetical protein” name) with a few exceptions, such as the gene with locus tag “sare_4525” shown in Figure 19(i). Follow the link provided by this gene’s Object ID to its **Gene Details** page. The “IMG Product Source” field in the **Gene Information** section of this page, shown in Figure 19(ii), shows that the product name is based on the TIGRfam associated with the gene.

³ A homolog is considered “close” to a gene *x* if it has at least 50% sequence identity and 70% sequence alignment with *x*.

References

- Bairoch A. (2000) The ENZYME database in 2000, *Nucleic Acids Research* **28**, 304-305. See also <http://au.expasy.org/enzyme/>.
- Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., et al. (2005) The Universal Protein Resource (UniProt), *Nucleic Acids Research* **33**, D154-159. See also UniProtKB/Swiss-Prot Documentation at: <http://au.expasy.org/sprot/sp-docu.html>.
- Besemer, J. and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes, and viruses, *Nucleic Acids Research* **33**, Web Server Issue, 451-454.
Available at <http://www.ncbi.nlm.nih.gov/genomes/MICROBES/genemark.cgi>.
- Gene Ontology Consortium (2004) The Gene Ontology Database and Informatics Resource. *Nucleic Acids Research*, **32**, 258-261.
- Haft, D. H., et al. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research* **29**, 41-43.
See also: <http://www.tigr.org/TIGRFAMs/>.
- Hauser, L., Larimer, F., Land, M., Shah, M., and Uberbacher, E. (2004) Analysis and Annotation of Microbial Genome Sequences, Genetic Engineering, 26, Kluwer Academic/Plenum Publishers, 225-238.
- 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, *Lawrence Berkeley National Laboratory Technical Report* LBNL-62292.
- Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Ken, C., Chen, I-M., Dubchak, I., Anderson I., Lykidis A., Mavrommatis K., Ivanova N.N., et al. (2008) The Integrated Microbial Genomes (IMG) System, *Nucleic Acids Research* **38**, Special Database Issue.
- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, Barrell B. (2000) Artemis: sequence visualisation and annotation. *Bioinformatics* **16** (10): 944-945.
- Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. (1998) Microbial gene identification using interpolated Markov models, *Nucleic Acids Research*, **26**(2), 544-548.
Available at http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi
- Salzberg, S.L. (2007) Genome re-annotation: a wiki solution?, *Genome Biology*, **8**:102.
- Sonnhammer, E. L. L. et al. 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research* **26**, 320-322.
See also: <http://www.sanger.ac.uk/Software/Pfam/>.
- Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A Genomic Perspective on Protein Families, *Science*, **278**, 631-637. See also: <http://www.ncbi.nlm.nih.gov/COG/>.