Microbial Genome Annotation & Analysis with the IMG/ER System

Technical Report LBNL-63615

Genome Biology Program

Department of Energy Joint Genome Institute

Biological Data Management and Technology Center

Lawrence Berkeley National Laboratory

April 20, 2009

Copyright 2009 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.

April 20, 2009

©2009 The Regents of the University of California

This document was prepared by:

Victor M. Markowitz*
Natalia N. Ivanova**
Konstantinos Mavromatis**
Iain Anderson**
I-Min A. Chen*
Ken Chu*
Krishna Palaniappan*
Nikos C. Kyrpides**

*Biological Data Management & Technology Center
Lawrence Berkeley National Laboratory

**Genome Biology Program (GBP)
Department of Energy Joint Genome Institute

Table of Contents

1 Synopsis 1
2 Setting the Genome Context for Annotation 2
2.1 Select Genomes 2
2.2 Assess Genome Annotation Coverage 4
3 Find Genes of Interest via Genome Exploration 7
3.1 Select Genes using Gene Search 7
3.2 Select Genes from a Specific Genome Statistics Table 8
3.3 Select Genes by Reviewing Gene Annotations 9
3.4 Select Genes using the GC Based Chromosome Viewer 10
4 Find Genes of Interest via Genome Comparison 12
4.1 Select Genes using the Phylogenetic Profiler 12
4.2 Select Genes using Abundance Profile Overview 13
4.3 Select Genes using a Functional Profile 15
5 Review Functional Annotations for Genes 17
6 MyIMG Annotations 21
6.1 Product Names 21
6.2 Missing Enzymes 24
6.2.1 Missing Enzymes for Specific Genomes and Genes 24
6.2.2 Missing Enzymes within a KEGG Pathway or Function Profile 26
6.3 Missing Genes 29
6.3.1 Handling Missing Genes within IMG ER 30
6.3.2 Handling Missing Genes outside IMG ER 30
6.4 Deleting or Merging Genes 33
7 Reviewing MyIMG Annotations 34
References 37
Glossary of Terms 38

1 Synopsis

IMG ER is part of the IMG family of systems that provides support for reviewing and curating the annotation of both public and so called "private" microbial genomes. The IMG ER public genome content baseline consists of all the isolate genomes in IMG 2.6.

Users can submit their "private" genomes for inclusion into IMG ER prior to their public release via the **IMG ER submission** site. Such genomes can be loaded into IMG ER with genes predicted and product names assigned to genes by a specific annotation pipeline. Alternatively, gene prediction and/or product name assignment can be carried out using the **IMG ER annotation pipeline**.

IMG ER provides support for curation of protein product names and a number of associated functional annotations, using IMG ER'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.

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

- A brief introduction to IMG is provided in (Markowitz, Szeto & al 2008).
- A user guide for IMG is available at: http://img.jgi.doe.gov/pub/doc/using_index.html.
- Materials on various aspects of microbial genome sequence data processing and analysis are available at: http://img.jgi.doe.gov/pub/doc/education.html.

In order to be able to submit genome data for loading into IMG/ER and access IMG ER users need to have an IMG account which can be requested by filling the **Request Account** form at: http://img.jgi.doe.gov/request.

IMG ER is at: http://img.jgi.doe.gov/er.

IMG ER annotation SOP is described at: http://img.jgi.doe.gov/w/doc/img_er_ann.pdf

IMG/ER Submission Site is at: http://img.jgi.doe.gov/submit

2 Setting the Genome Context for Annotation

Microbial genome annotation generally refers to the process of interpreting the raw sequence data 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.

2.1 Select Genomes

Genomes can be **selected** using the **Genome Browser** as follows:

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

Example 2.1(i). Under **Find Genomes** in the Main Menu, use the **Genome Browser** with **View Phylogenetically** to see 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 2.1(i). Save these selections.

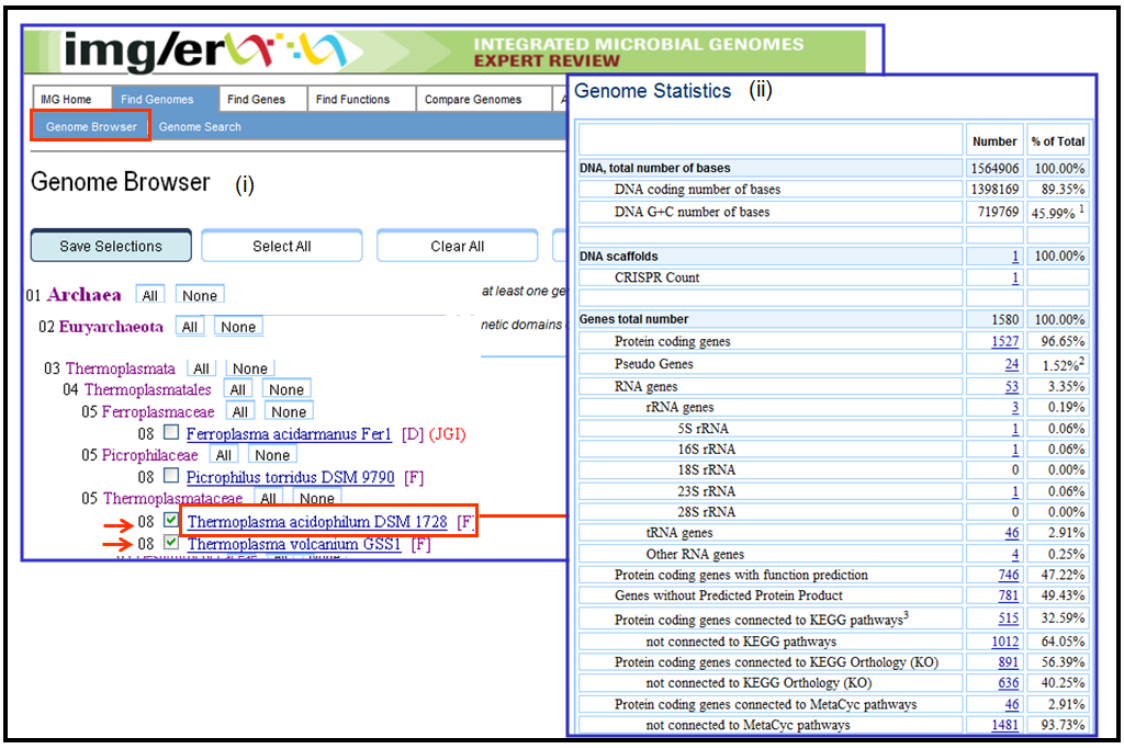


Figure 2.1. Selecting genomes: **Genome Browser**.

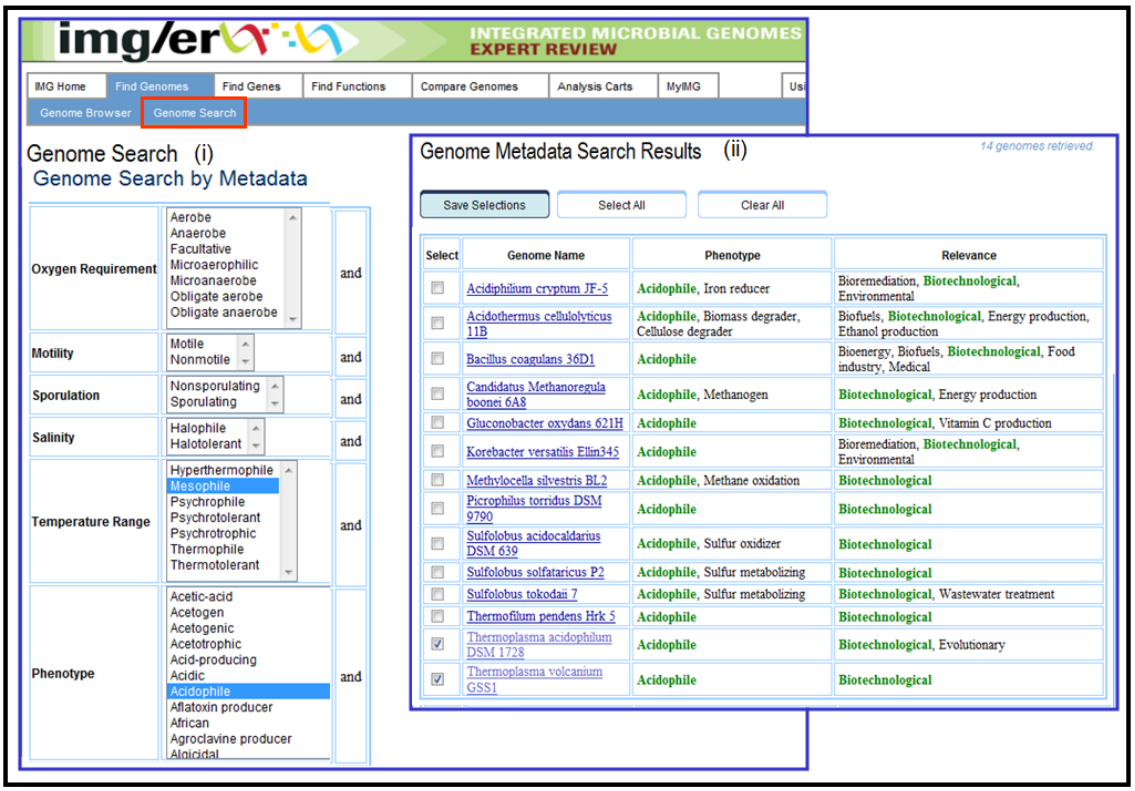


Figure 2.2. Selecting genomes: **Genome Search by Metadata**.

Genomes can be also **selected** using **Genome Search** as follows:

1. Start with **Find Genomes** and select **Genome Search**.
2. Select **Genome Search by Fields** or **Genome Search by Metadata.**
3. Select a **Filed/Filter** for the keyword search (e.g., Genome Name) or set a condition involving the **Metadata** fields (e.g., Phenotype, Relevance)

Example 2.1(ii). Under **Find Genomes** in the Main Menu, use **Genome Search** and then go to **Genome Search by Metadata**. Set the search condition by selecting "Acidophile" for **Phenotype** and "Mesophile" for **Temperature Range**, as shown in Figure 2.2(i). The list of genomes satisfying the search condition is returned as shown in Figure 2.2(ii). Save these selections.

2.2 Assess Genome Annotation Coverage

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 **individual** genomes using the **Genome Statistics section** the **Organism Details** which can be reached from any genome name, such as those listed in the **Genome Browser**.

Example 2.2. In the list of genomes displayed by the Genome Browser (see Figure 2.1(i)) select *T. acidophilum* in order to view its **Genome Statistics** table, as shown in Figure 2.1(ii).

Assess annotation coverage for **multiple** genomes using the **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. These statistics are available via **Genome Statistics** under the **Compare Genomes** main menu option.

The **Breakdown by selected genomes, general statistics** option provide the default statistics table for the selected genomes. These statistics can be configured using a **Configuration** table, which allows you to choose the information you are interested to examine, including:

a) **Genes**, which represents the total number of genes in the genome including protein- and RNA-coding genes.
b) **CDS**, which represents the total number of protein-coding genes including pseudogenes.
c) **RNA**, which represents the total number of stable RNA-coding genes including rRNAs, tRNAs and other RNAs as discussed above.

d) **5S**, **16S**, **23S**, **tRNA** and **Other RNA**, which provide a breakdown of RNA-coding genes into different categories.
e) **Pseudo**, which represents the number of pseudogenes in the genome including those for protein-coding and RNA-coding genes.
f) **Unchar**, which represents the number of uncharacterized CDSs, that is CDSs without definitive functional prediction including all proteins with names "hypothetical", "predicted" and "unknown".
g) **w/Func Pred**, which represents the number of CDSs with definitive functional assignment including protein family annotations, such as "dehydrogenase family protein".
h) **Enzymes**, which represents the number of CDSs with assigned EC numbers including incomplete EC numbers, such as EC:1.1.1.-.
i) **COG**, which represents the number of CDSs assigned to COG protein families.
j) **Pfam**, which represents the number of CDSs assigned to Pfam protein families.
k) **TIGRfam**, which represents the number of CDSs assigned to TIGRfam protein families.
l) **Bases**, which represents the total number of nucleotides in the genome sequence.
m) **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.

Example 2.3. Select **Compare Genomes** in the Main Menu, and then select **Genome Statistics.**

(a) **Summary Statistics** provide cumulative statistics for the selected genomes, including the numbers of genes and various functional annotations, as illustrated in Figure 2.3(i). 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*.

(b) **COG Category Statistics** and **KEGG Category Statistics** provide statistics in a tabular format for the selected genomes across top-level COG and KEGG categories, respectively. Select **COG Category Statistics**, and then select

i. **Statistics for Genomes by COG Categories**, as illustrated in Figure 2.3(ii), in order to display a statistics table, as shown in Figure 2.3(iii). Each genome in this table is linked to its **Genome Details** page, which provides statistics for individual genomes, as discussed above. Parameter columns can be added or deleted using the **Configuration** selector at the end of the table.

ii. **Statistics for Genomes by specific COG Category** in order to display in a tabular and pie chart format the count of genes associated with each COG category across all selected genomes, as illustrated in Figure 2.3(iv). Click on a COG category on the *pie chart* or on the colored coded square for a COG category in the table to display a *bar chart* with the percent of genes for each genome associated with that COG category, as illustrated on the lower side pane of Figure 2.3(iv).

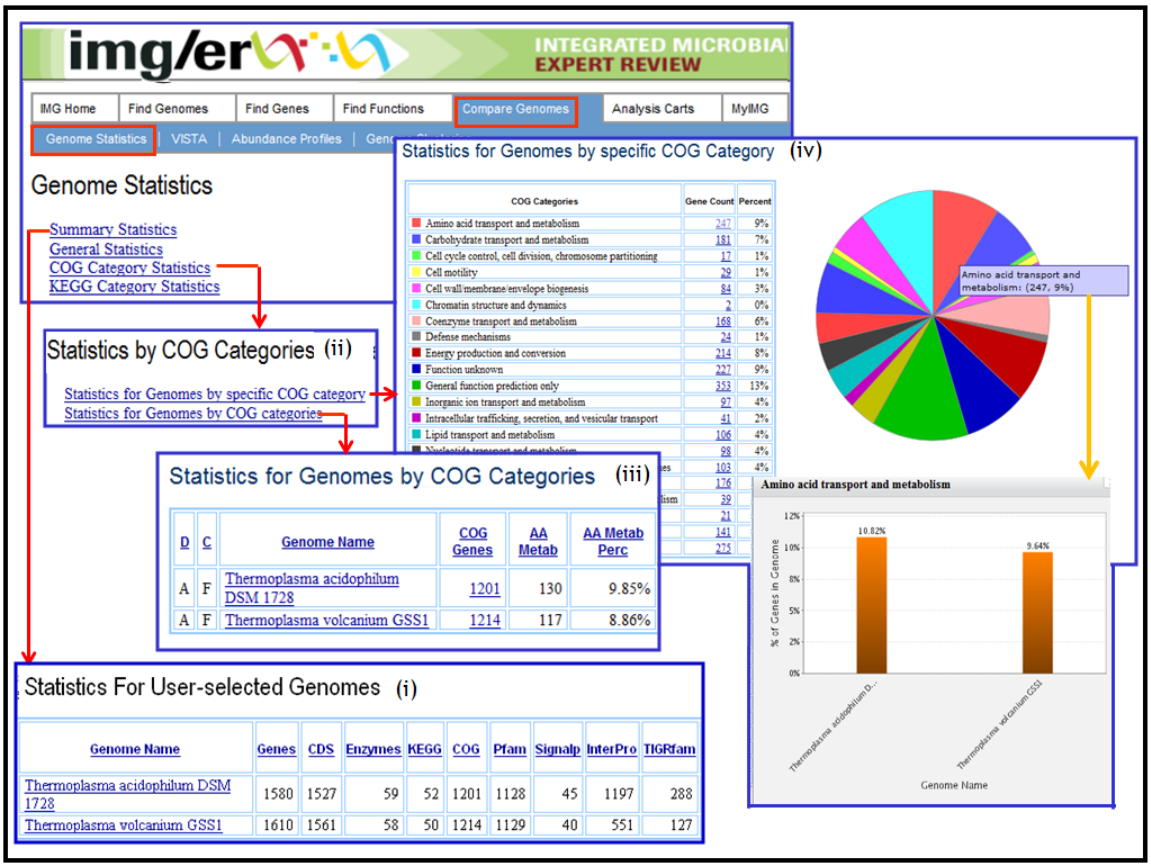


Figure 2.3. Assessing genome annotation coverage with **Genome Statistics**.

3 Find Genes of Interest via Genome Exploration

A number of tools are available for selecting genes of interest for further analysis and functional annotation. We review below gene selection that involves exploration of individual genomes.

3.1 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 **UsingIMG** 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 3.1. Note that if a genome context is set, then this search will be performed on the genomes in this subset only.

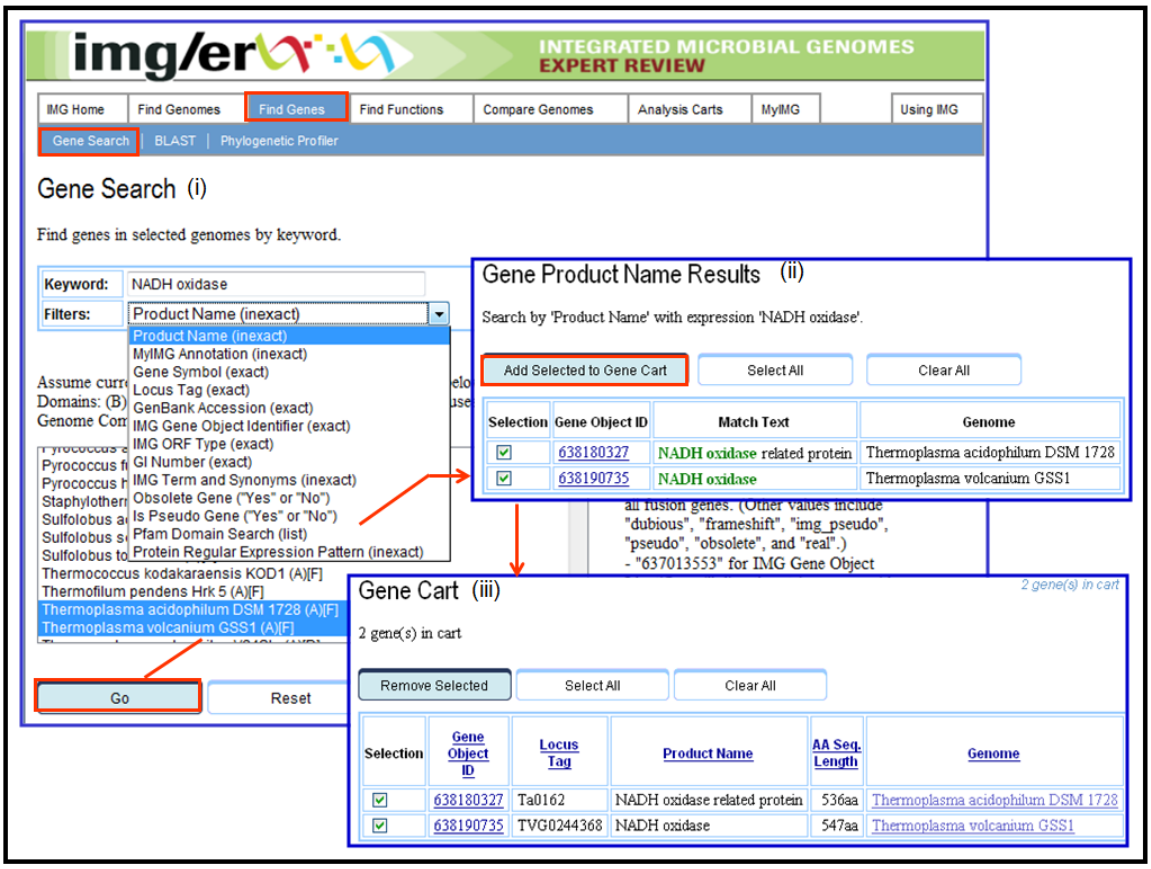


Figure 3.1. Select genes for analysis and functional annotation with **Gene Search**.

Example 3.1. 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 2.1, 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 3.1(i). The search for product names containing "NADH oxidase" returns two genes, as shown in Figure 3.1(ii). The genes can be save into the Gene Cart, as shown in Figure 3.1(iii), for further analysis.

3.2 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, as illustrated in Figure 3.2. Gene counts in some categories, such as "Genes in COGs" and "Genes connected to KEGG pathways" are linked to viewers 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.

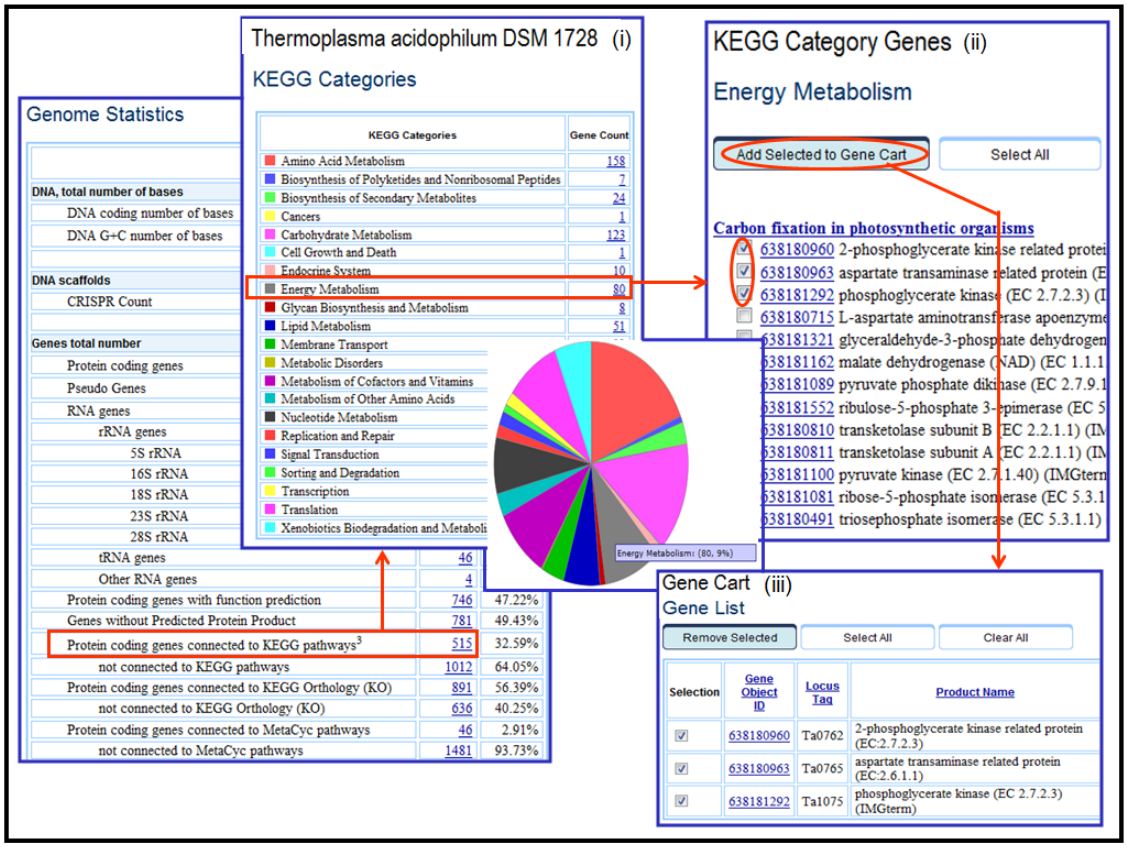


Figure 3.2. Select genes for further analysis and functional annotation via the gene categories available in the **Genomics Statistics** table.

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 3.2. Follow the link for **Genes connected to KEGG pathways** from the **Genome Statistics** table (Figure 3.2) to a classification of genes based on association with specific KEGG categories, as shown in Figure 3.2(i). For a specific KEGG category, such as *Energy Metabolism*, links (via the number of genes, the corresponding area of the pie chart) are provided to the list of genes organized in specific pathways, such as *Carbon Fixation*, as shown in Figure 3.2(ii). Genes can be then selected and saved in the **Gene Cart** for further analysis, as shown in Figure 3.2(iii).

3.3 Select Genes by Reviewing 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 3.3).

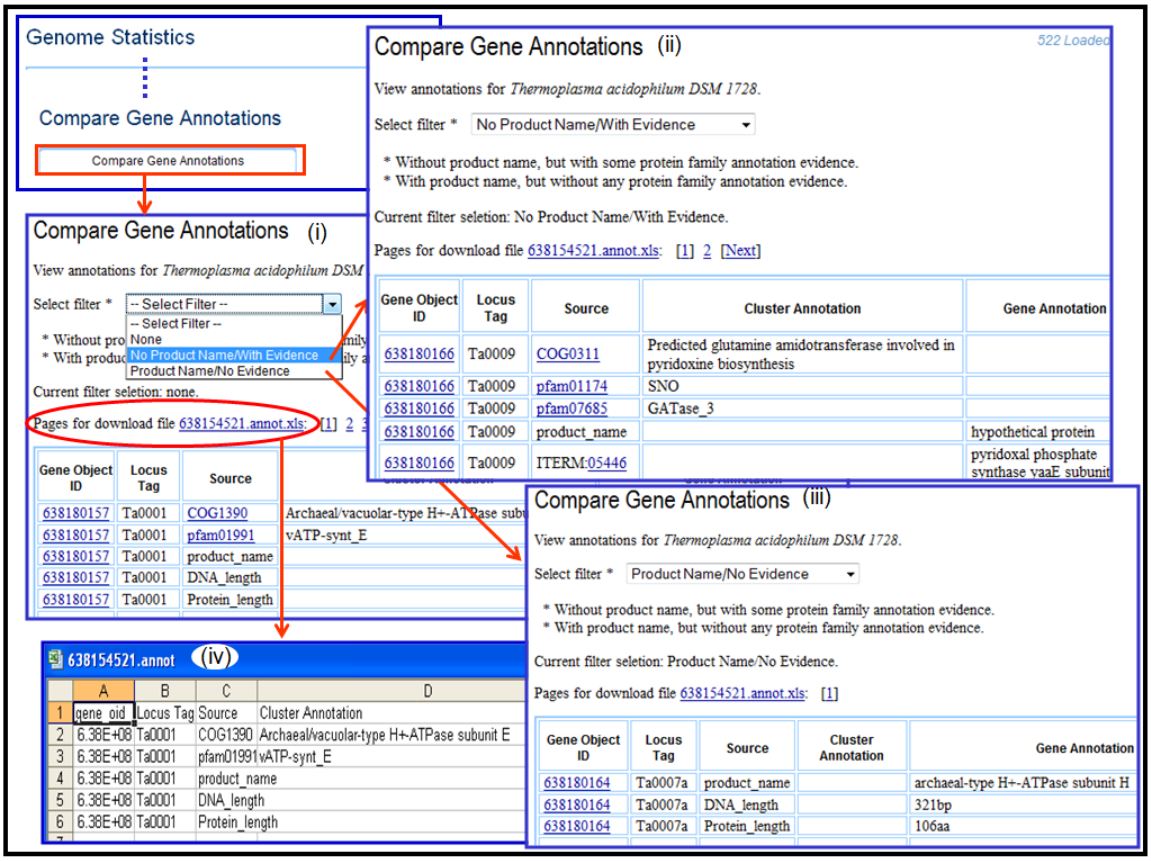


Figure 3.3. Select genes for analysis and functional annotation using the **Compare Gene Annotations** table that can be reached via the **Organism Details** page.

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 3.3. Follow the link for **Compare Gene Annotations** via the button provided below the **Genome Statistics** table (see Figure 3.3) 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 3.3(i). Examine the product name in the context of other available functional annotations, such as COG, pfams, and IMG term (when available). The results can be filtered in order to display only genes (a) **without a product name**, but **with evidence** of potential function provided by association with a COG, Pfam, or TIGRfam, as illustrated in 3.3(ii), or (b) **with a product name**, but **without any other evidence** of function provided by association with a COG, Pfam, or TIGRfam, as illustrated in Figure 3.3(iii). The annotations for all the genes can be downloaded into a local excel file, as shown in Figure 3.3(iv).

3.4 Select Genes using the GC Based Chromosome Viewer

Bacterial chromosomes for the most part have a uniform GC content but may contain regions with GC content that differs from the average. The regions with distinct GC content are thought to originate from other organisms, viruses, or transposable elements. For example, pathogenicity islands are regions of a genome that encode virulence genes and commonly have a different GC content than the rest of the genome (Guy 2006). Such regions are found in pathogenic bacteria but commonly are absent from nonpathogenic relatives.

There are three steps involved in such a selection:

1. Select an organism/ genome of interest using the **Genome Browser** or **Genome Search**.
2. Go to the **Genome Viewers** of the **Organism Details** page and select **Scaffolds and Contigs**.
3. Select the desired coordinate range of a scaffold of interest or enter the coordinate range of interest. In the Chromosome Viewer, select the **GC percentage** coloring option.

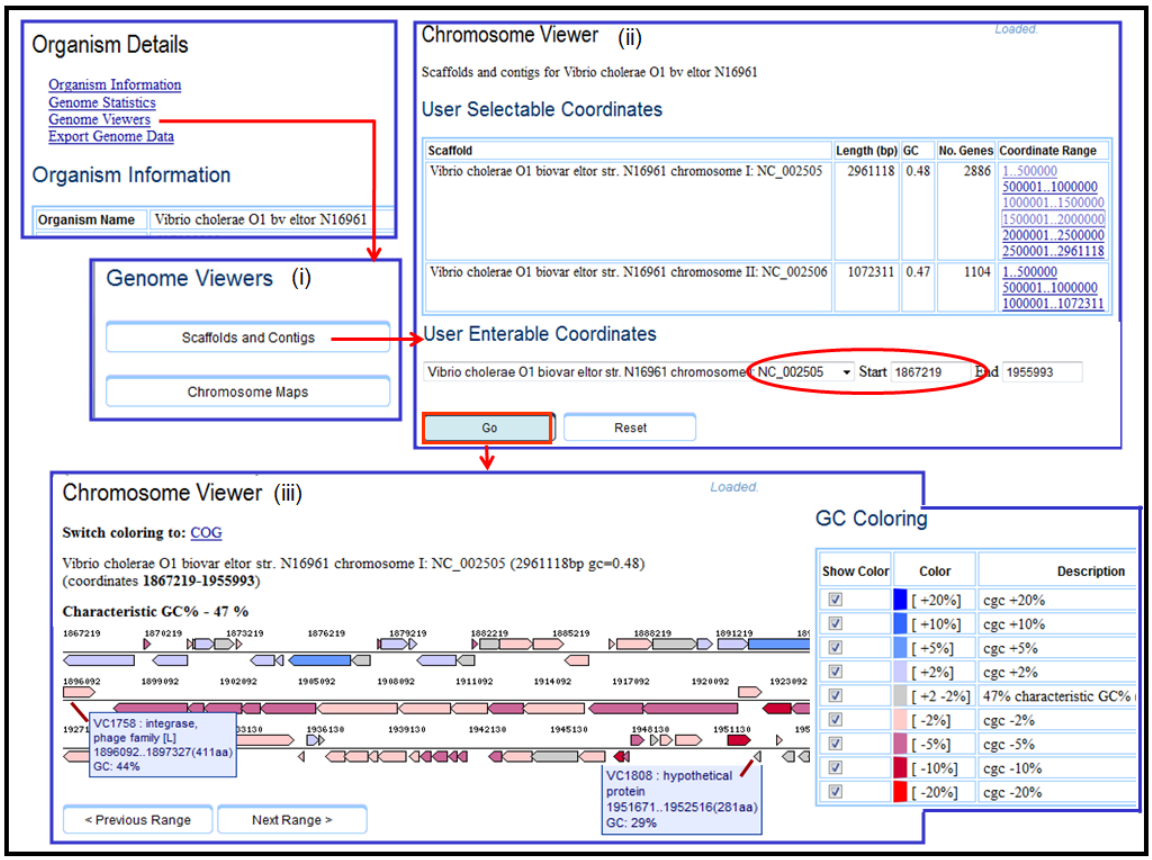


Figure 3.4. Find genes with the **GC** colored **Chromosome Viewer**.

<u>Example 3.4</u>. Select the *Vibrio cholera cholerae* O1 biovar eltor str. N16961 genome using the **Genome Browser** or **Genome Search**, go to the **Genome Viewers** section of **Organism Details** page, and select **Scaffolds and Contigs**, as shown in Figure 3.4(i). Enter 1867219 and 1955993 as start and end coordinates and launch the **Chromosome Viewer**, as shown in Figure 3.4(ii), then switch to **GC percentage coloring** version of the **Viewer**. Using the Chromosome Viewer, examine the genes that deviate from the characteristic GC percentage for this genome (47%), as shown in Figure 3.4(iii).

An example is the island VPI-2 from *Vibrio cholerae*, which in the *V. cholerae* O1 biovar eltor str. N16961 genome (Jermyn & Boyd 2002). This pathogenicity island extends from VC1758 to VC1809 and has a lower GC content than the rest of the chromosome, as shown in Figure 3.4(iii).

4 Find Genes of Interest via Genome Comparison

Genes that may require further analysis and functional annotation can be identified using comparative analysis tools within a specific genomic context.

4.1 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 for Single Genes** 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 for Single Genes** under the **Phylogenetic Profiler**s second level menu bar.

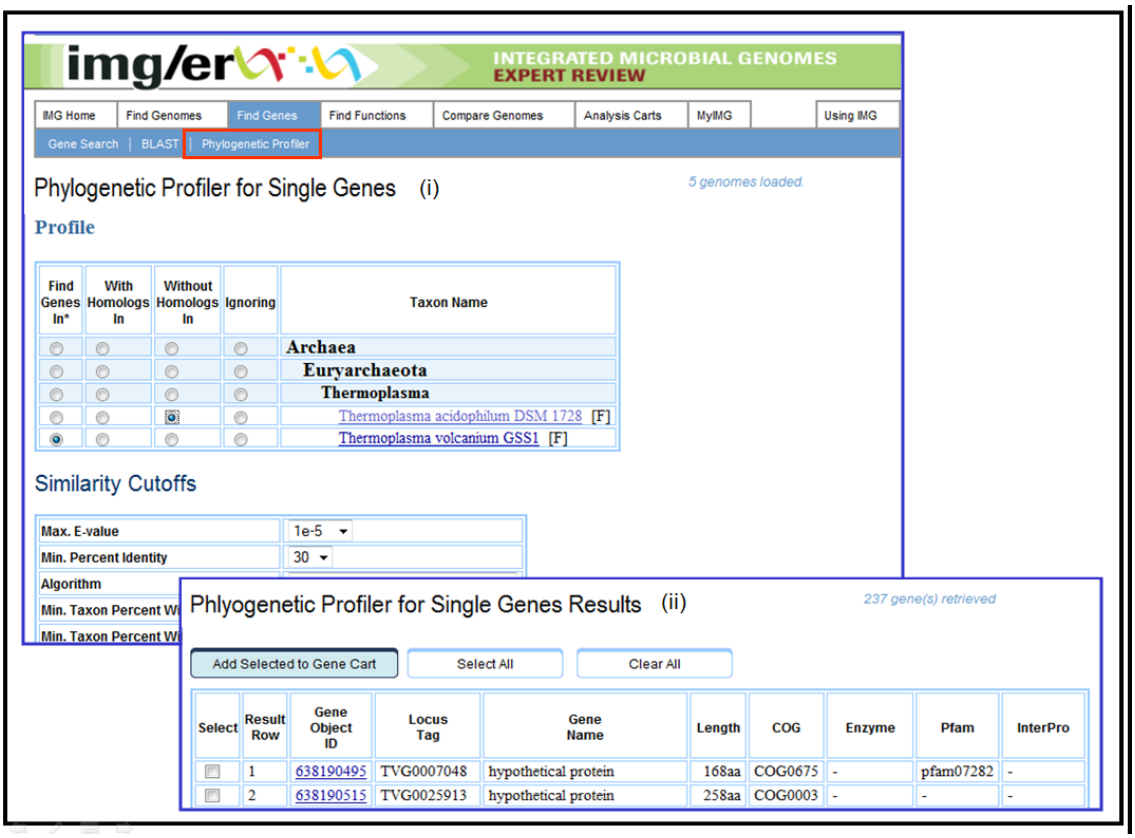


Figure 4.1. Find genes with the **Phylogenetic Profiler for Single Genes** tool.

(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 4.1. 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 4.1(i). 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 4.1(ii).

The **Phylogenetic Profiler for Single Genes** 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 4.1, 237 genes are found to be unique in *T. volcanium* with respect to *T. acidophilum.*

4.2 Select Genes using Abundance Profile Overview

Genes can be selected using the **Abundance Profile Overview** tool that allows comparing selected genomes in terms of their relative abundance across *all* protein families (COGs, Pfams, and TIGRfams) and functional families (Enzymes). There are three steps involved in such a selection:

(a) Start with **Compare Genomes** in the Main Menu and select **Abundance Profiles**. Select the **Abundance Profile Overview** tool.

(b) Select the type of format for displaying the results ("Heat Map" or "Matrix"), protein/functional families (COG, Pfam, TIGRfam, Enzyme), normalization method, and the genomes that will be compared, as illustrated in Figure 4.2(i).

(c) From the functions listed as results of the **Abundance Profile Overview**, 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.

For "Heat Map" output, the abundance of protein/functional families is displayed as a heat map with red corresponding to the most abundant families, as illustrated in Figure 3.5(ii). Each column on the map corresponds to a genome, and each row corresponds to a family; mouse over each cell to see the count of a particular family in a genome. Click on the cell in order to retrieve the list of genes assigned to this particular family in a genome, as illustrated in Figure 4.2(iii). Click on the identifier of the family displayed on the right of the column (e.g., COG0675) in order to include the corresponding family into the **Function Cart**, as illustrated in Figure 4.2(iv).

If the "Matrix" output is selected, the abundance of protein/functional families is displayed in a tabular format, as illustrated in Figure 4.2(v), with each row corresponding to a family and each cell containing the number of genes associated with a family for a specific genome. Click on the cell in order to retrieve the list of genes assigned to this particular family in a genome, as shown in Figure 4.2(iii). Families of interest can be

selected for inclusion into the **Function Cart**, as illustrated in Figure 4.2(iv). The results in "Matrix" format can be also exported to a tab-delimited Excel file.

By default, the **Abundance Profile Overview** results are sorted by the abundance of families in the first genome. These results can be resorted according to the abundance in other genomes by clicking on the corresponding column header.

Example 4.2. After setting the genome context to two genomes, T. volcanium and T. acidophilum, as discussed above, use the **Abundance Profile Overview** tool to find COGs that are more abundant in T. volcanium than in T. acidophilum, as shown in Figure 4.2(i). The result displays the abundance counts for all COGs across the two genomes, as shown in Figure 4.2(ii) and 4.2(v). Each abundance cell or count provides a link to the associated list of genes, such as the genes associated with COG 0675, as shown in Figure 4.2(iii).

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.

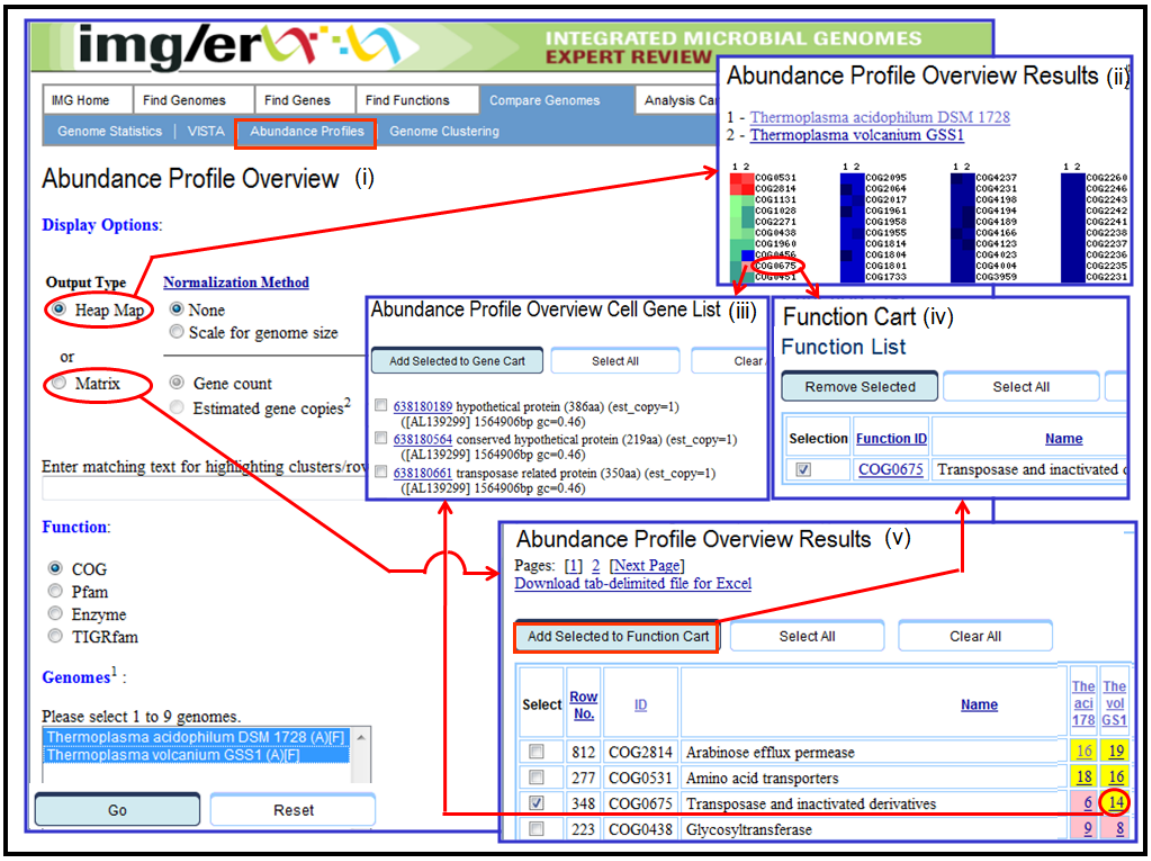


Figure 4.2. Find genes with the **Abundance Profile Overview** tool.

4.3 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:

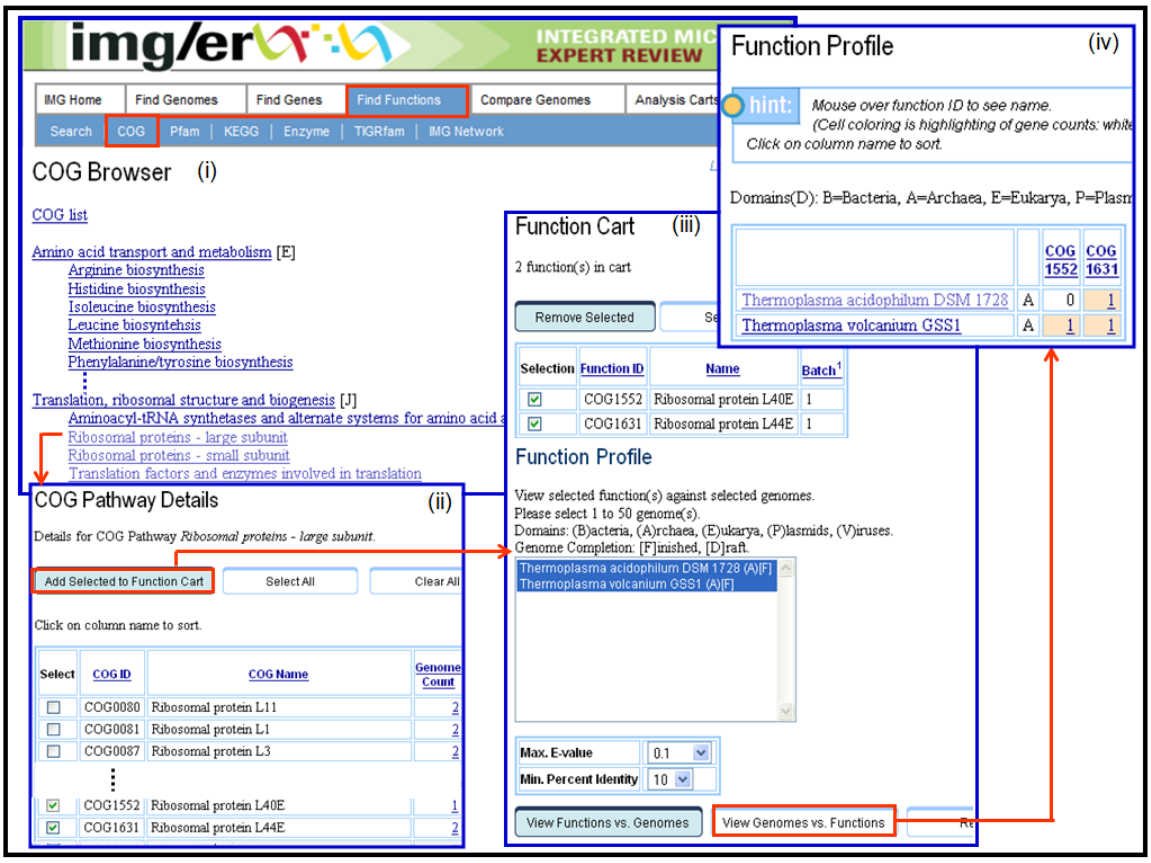


Figure 4.3. Find genes with a **Function Profile** tool.

(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 4.3. Under **Find Functions**, select the **COG** browser, as shown in Figure 4.3(i). Scroll down to Ribosomal proteins – large subunits COG pathway and examine it using the **COG Pathway Details**, as shown in Figure 4.3(ii). From this COG pathway select COG1552 and COG1631 and save them in the **Function Cart**, as shown in Figure 4.3(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 4.3(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. Note that COG1552 is present in T*. volcanium* but not in *T. acidophilum*, which may indicate a gene missed by the original annotation.

5 Review Functional Annotations for Genes

The functional annotation for individual genes can be reviewed using the **Gene Detail** page (see Figure 5.1) that is available via the link provided from individual gene identifiers (so called OIDs). The **Gene Detail** 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 **Sequence Viewer**, two **Chromosome Viewers**, **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.

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.

Product names are **assigned** to genes using various methods based on gene similarity and involving one or several functional classifications, such COG, Pfam, and TIGRfam. For example, the **IMG product name assignment** procedure[1] 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 close[2] 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.

[1] The IMG ER annotation procedure is described at: http://img.jgi.doe.gov/w/doc/img_er_ann.pdf.
[2] A homolog is considered "close" to a gene *x* if it has at least 50% sequence identity and 70% sequence alignment with *x*.

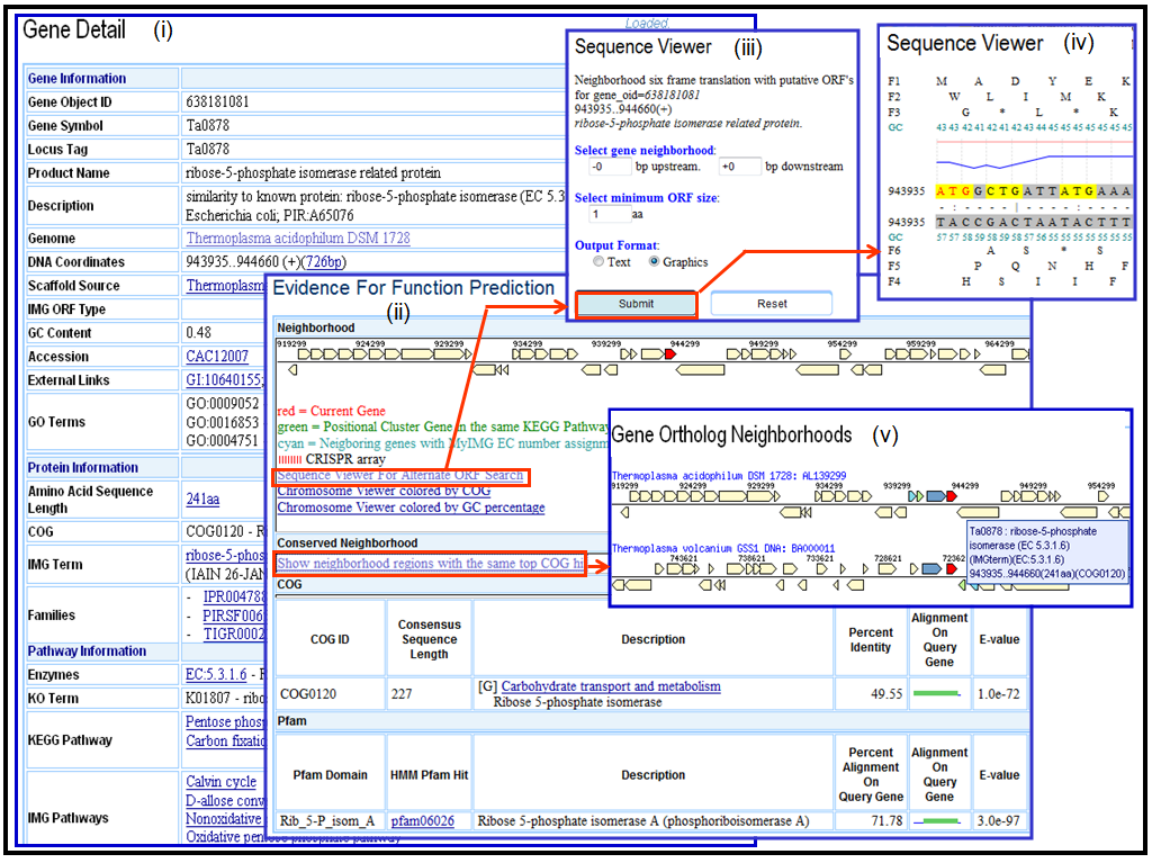


Figure 5.1. Review functional annotations for a gene with **Gene Detail**.

The product names associated with genes can be examined using the **Gene Details**. 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.

Example 5.1. Start with the steps in Example 2.1, which lead to the selection of *T. acidophilum* and its **Genome Statistics**, as shown in Figure 2.1(iii). Follow the link for **Genes connected to KEGG pathways** from the **Genome Statistics** page shown in Figure 2.1(iii) which leads to a further classification of genes based on their association with specific KEGG categories, as shown in Figure 3.2(i). Follow the link *Energy Metabolism* to the list of genes organized in specific pathways, and select *Carbon Fixation*, as shown in Figure 3.2(ii).

For the **gene** with identifier **638181081**, follow the link to its ***Gene Details*** page, as shown in Figure 5.1(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 5.1(ii)). Select the **Sequence Viewer** to display the six frame translation with putative ORF's, potential start codons, and potential

Shine-Delgano regions. The gene neighborhood, minimum ORF size and type of display (graphic or text) can be selected, as illustrated in Figure 5.1(iii).The text display provides the protein sequences for the ORFs while the graphical display, illustrated in Figure 5.1(iv), includes a GC plot. Note that the Sequence Viewer does not provide all the details nor the capabilities of tools such as Artemis (Rutherford et al 2000). You can use **Web Artemis** viewer on the **Organism Details** page to examine the details of a sequence region of interest.

Select *Show neighborhood regions with the same top COG hit* in order to display the **Gene Ortholog Neighborhoods** viewer, as shown in Figure 5.1(v). 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 5.2. 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 4.1). 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 5.2, 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 5.2(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.

(b) The second gene (see Figure 5.2(ii)) is a 50S ribosomal protein L40E (COG 1552) which is also an essential gene. Subsequently, tBLASTn of the *T. volcanium* gene can be used to identify the missing gene in *T. acidophilum*.

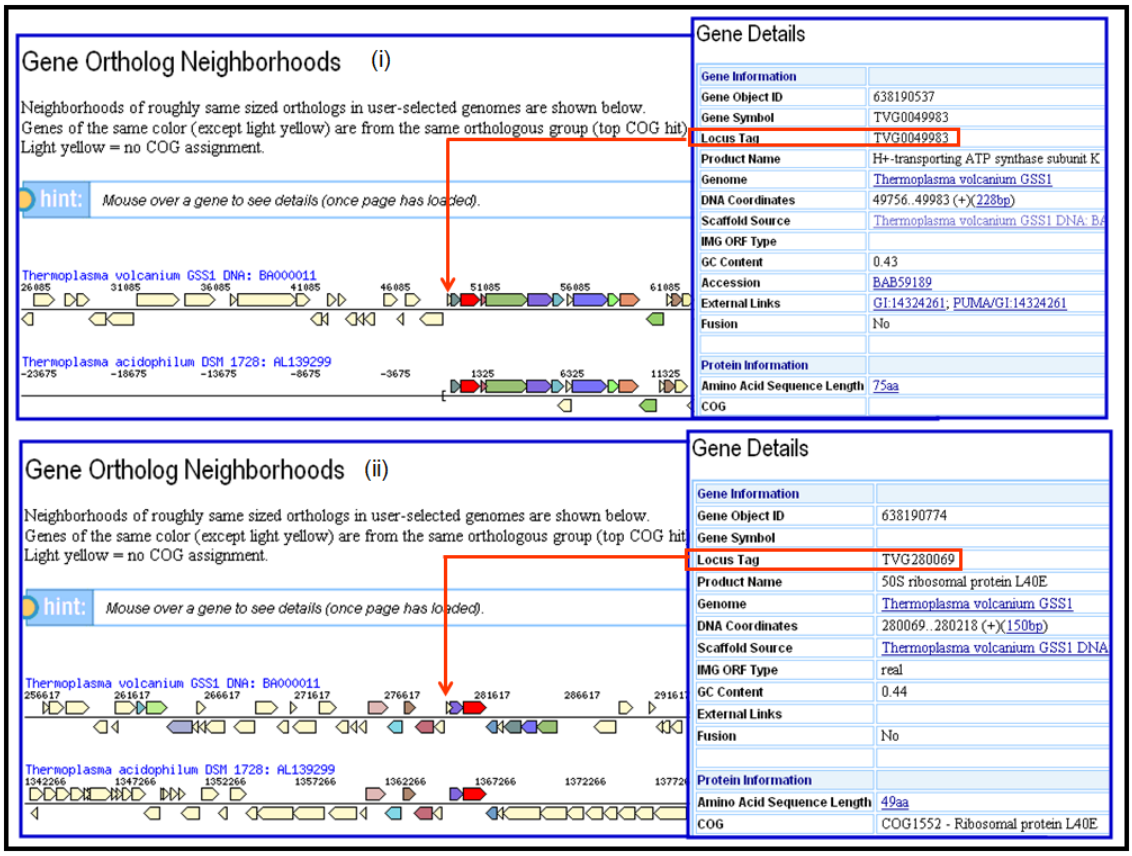


Figure 5.2. Using **Gene Ortholog Neighborhoods** to examine two *T. volcanium* genes that seem to be missing in the *T. acidophilum* genome.

6 MyIMG Annotations

The functional annotation for individual genes can be curated using the **MyIMG Annotations** features of **MyIMG**. 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 by filling the **Request Account** form at: http://img.jgi.doe.gov/request.

MyIMG Annotations provides support for: (1) **editing** the product name and several associated fields for one or more related genes that have been previously selected and saved in the **Gene Cart**; (2) marking a gene as **deleted** (removed) from a genome; (3) finding genes associated with **enzymes** missing in specific genomes; (4) **reviewing** the functional annotations for individual genes or group of genes; (5) **exporting** to/ **uploading** from a tab-delimited file functional annotations for genes identified by their IMG identifier (OID).

6.1 Product Names

Genes that may require **product name review** and **curation** can be identified using various analysis tools, such as those discussed in section 3.

MyIMG provides support for editing the product name and associated information for one or several related genes. The following annotation fields can be manually edited:

- Product Name
- Protein Description
- EC Number
- PUBMED ID
- Inference
- Is Pseudo Gene?
- Notes
- Gene Symbol
- Remove Gene from Genome

First, select and save the target genes for product name curation in the **Gene Cart**. Next, use the **Annotate Selected Genes** link in the **MyIMG Annotation** section of the **Gene Cart** to access the tool that allows editing the fields listed above.

Example 6.1. Consider for review gene PF1186 (IMG identifier 638173757) of genome *Pyrococcus furiosus* whose details are shown in Figure 6.1(i). This gene is associated with product name NADH oxidase, as shown in Figure 6.1(i), and as recorded in GenBank. The list of top homologs for the gene under review can be displayed display via the Homologs section of its **Gene Details**, as shown in Figure 6.1(ii).

Based on a recent study[3], 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 6.1(iii).

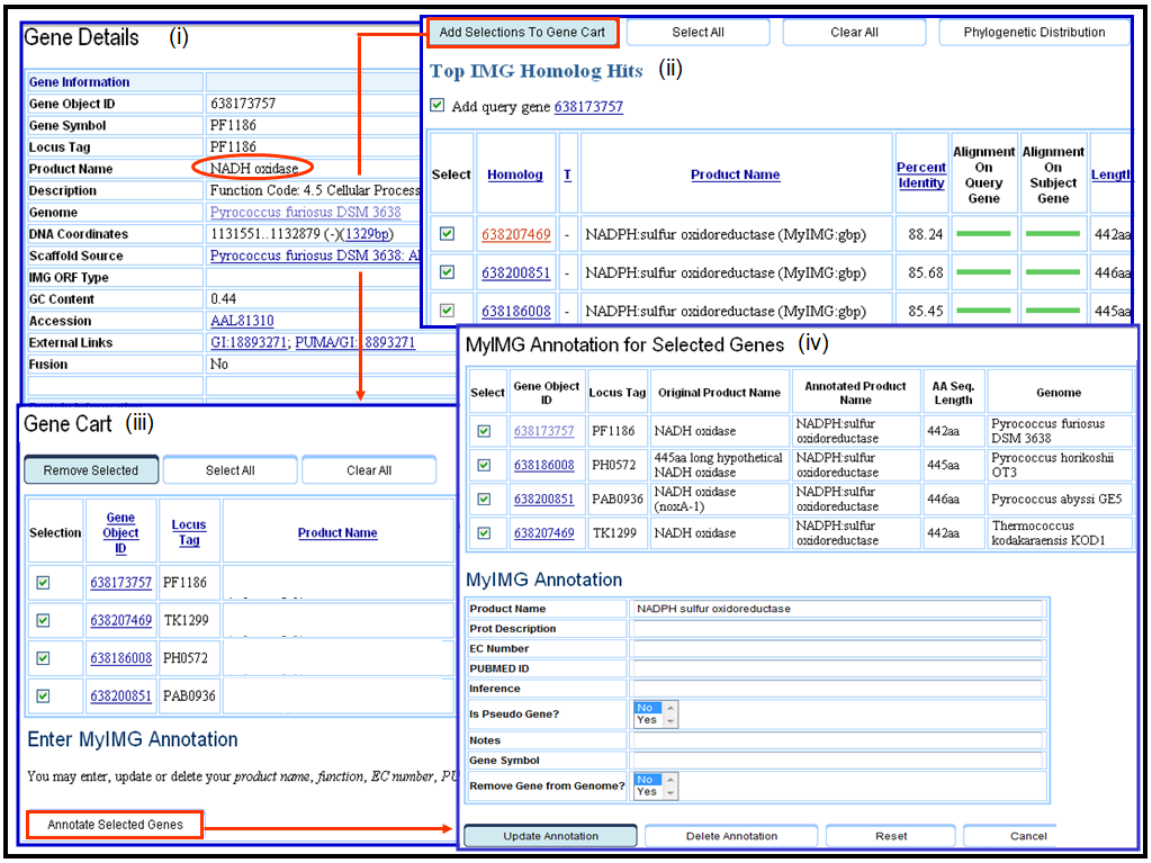


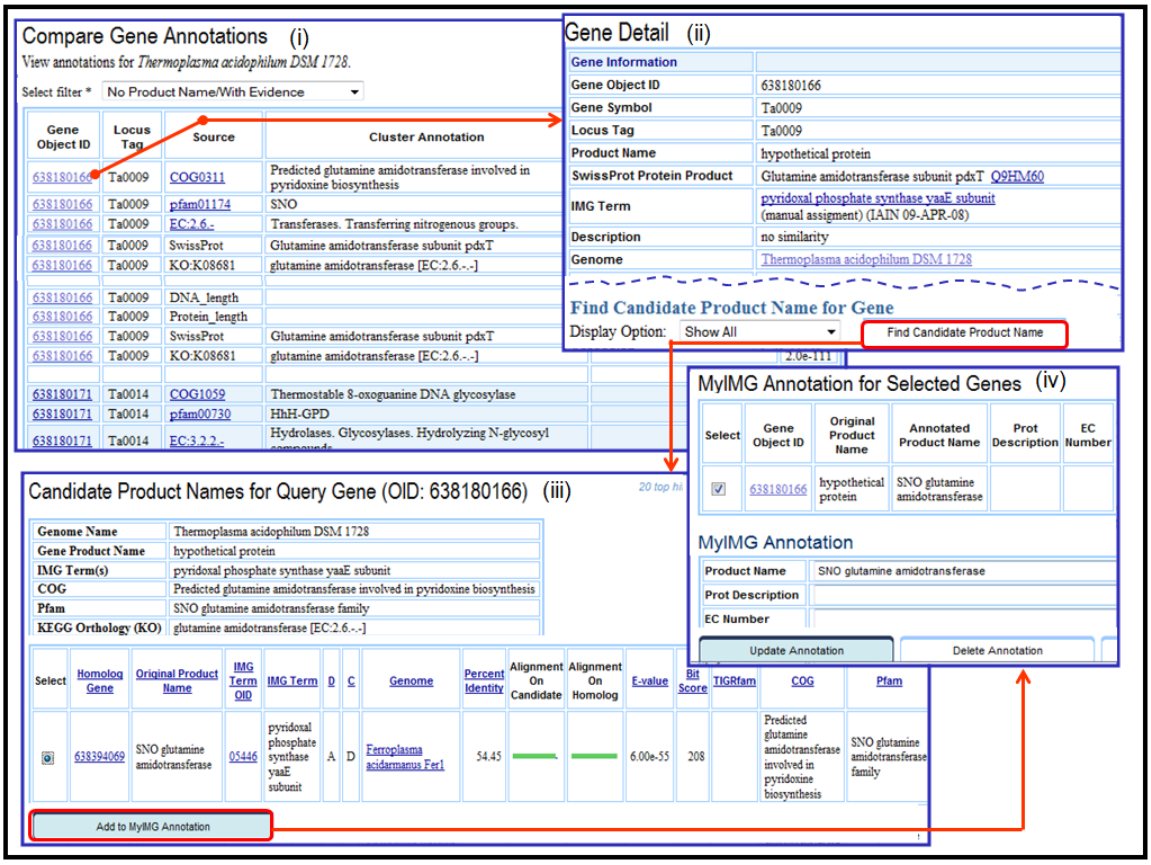
Figure 6.1. Review and curation of the product name for a gene of *Pyrococcus furiosus* using **MyIMG** Annotation.

Next, the product name where is changed to NADPH:sulfur oxidoreductase using the **MyIMG Annotation** tool accessed from **Gene Cart**, as shown in Figure 6.1(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 6.2.

Finding missing or problematic protein products across an entire genome in IMG is facilitated by the **Compare Gene Annotations** tool. For example, genes without a product name but with evidence of potential functional annotation or with product name but without any evidence of functional annotation (see example 3.3) are candidates for product name review and curation.

[3] 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*.

Example 6.2. Consider for review the genes of genome *Thermoplasma acidophilum*. **Compare Gene Annotations** for *T. acidophilum* allows focusing the review on genes *without a product name*, but *with evidence* of potential function provided by association with a COG, Pfam, or TIGRfam, as illustrated in Figure 6.2(i), or *with a product name*, but without any other evidence of function provided by association with a protein family. Potentially missing or inconsistent product names can be further reviewed through



individual **Gene Details**, as illustrated in Figure 6.2(ii), as discussed in Example 6.1 above. Alternatively, candidate product names for a query gene of interest can be identified using the **Find Candidate Product Names** tool which retrieves the product names of the gene's closest homologs, as illustrated in Figure 6.2(iii). For each candidate product name, protein family and alignment information is provided in order to assist in choosing the appropriate product name.

Figure 6.2. Review and curation of the product name for a gene of *Pyrococcus furiosus* using **MyIMG** Annotation.

For genes for which a product name has been identified with the **Find Candidate Product Names,** the **MyIMG Annotation** tool is accessed via the **Add to MyIMG Annotation** link available on the **Candidate Product Names** results page, as illustrated in Figure 6.2(iii). **MyIMG Annotation** allows editing the product name and associated information, such as protein description and EC number, as illustrated in Figure 6.2(iv).

6.2 Missing Enzymes

The metabolic capacity of a genome is defined by its association with pathways via gene products that function as enzymes serving as catalysts for reactions in these pathways. A genome-pathway association may be partial, with "missing" associations between enzymes for reactions on the pathway with genes on the genome. We call such associations ***missing enzymes.***

6.2.1 Missing Enzymes for Specific Genomes and Genes

MyIMG provides support for revising ***missing enzymes*** for specific genomes and genes. For each genome, the **Genome Statistics** section of its **Organism Details** page contains a count of "Genes without enzymes, but with candidate KO based enzymes" which leads to a list of genes that could be associated with enzymes predicted via KEGG Orthology (KO) terms[4]. These predicted enzymes can be examined for accuracy and then associated with genes using **MyIMG Annotation** tools.

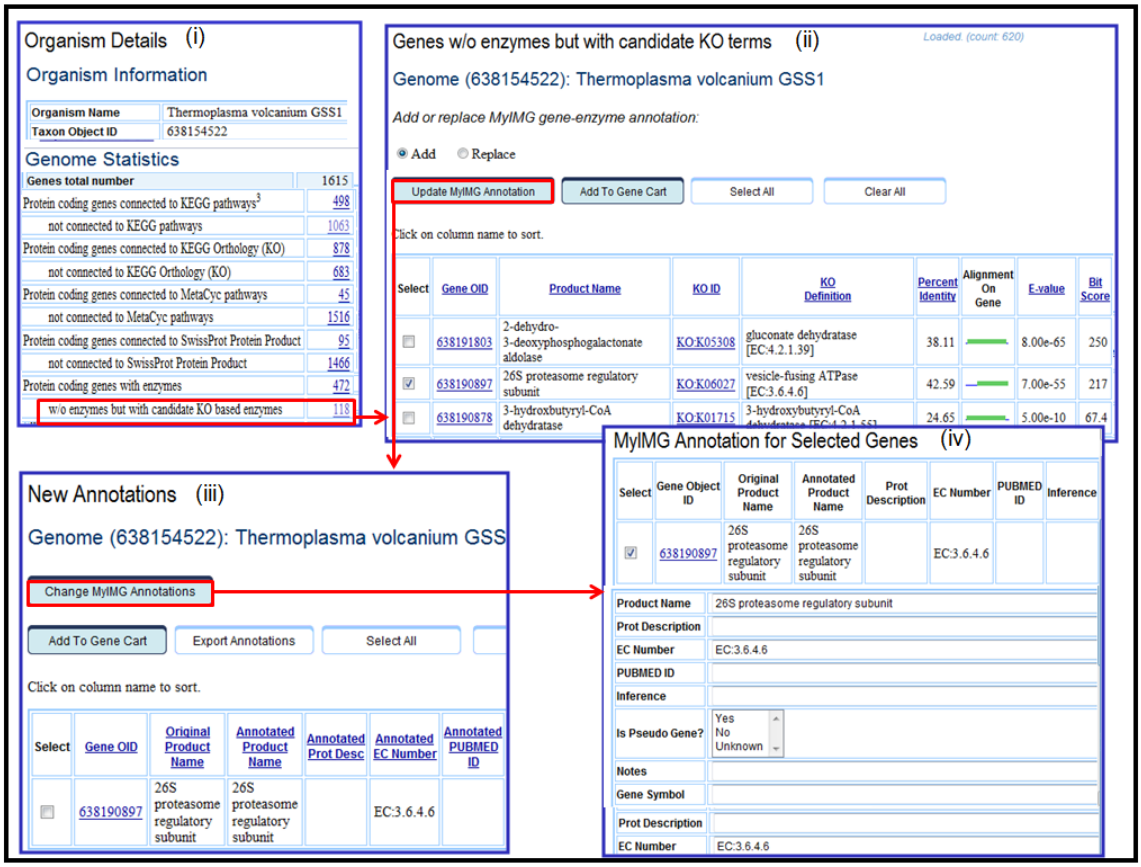


Figure 6.3. Examining Missing Enzymes for a Specific Genome.

[4] IMG genes are assigned KO terms associated with KEGG genes to which they correspond via NCBI GI numbers. For IMG genes that have no KEGG correspondents, BLASTP is run against KEGG genes, with the results organized in a list of candidate KO assignments, using 1e-2 cutoff for the top 25 KEGG gene hits. A subset of this list (1e-5 cutoff, KO assignment rank of 5 or better, and alignment percentage of at least 70% over the length of the IMG query gene and KEGG subject gene) is used to assign KO terms and enzymes to IMG genes, while the rest of the list is used for searching potentially "missing enzymes".

Example 6.4. Use the **Genome Browser** for selecting *Thermoplasma volcanium GSS1 (T. volcanium)* genome and examine the **Genome Statistics** section of its **Organism Details**, as shown in Figure 6.3(i). Follow the "Genes w/o enzymes, but with candidate KO term based enzymes" link to the list of genes that have enzymes predicted via KO term associations, as shown in Figure 6.3(ii).

Select the genes you want to associate with the predicted enzymes and then use **Update MyIMG Annotation**, as shown in Figure 6.3(ii). You can either **Add** the predicted enzyme to or **Replace** an existing enzyme in your MyIMG annotation. The new gene-enzyme associations are listed for review, as shown in Figure 6.3(iii). **MyIMG Annotations** can be further revised as shown in Figure 6.3(iv).

Instead of associating lists of genes with the top predicted enzyme, one can examine all enzyme predictions for individual genes via **Find Candidate Enzymes** available in the **Gene Information** section of the gene's **Gene Detail** page, as illustrated in Figure 6.4(i).

Example 6.4. In the list of genes of *Thermoplasma volcanium GSS1 (T. volcanium)* genome that have with KO term based predicted enzymes (see Figure 6.3 (ii) go to the **Gene Detail** page for the gene with IMG identifier 638191803. Use **Find Candidate Enzymes** immediately before the **Evidence for Function Prediction** section, as shown in Figure 6.4(i), to get to the list predicted genes on the **Candidate Enzymes Using KEGG Orthology** page, as shown in Figure 6.4(ii). Examine the enzymes in the list and select the best predicted enzyme(s).

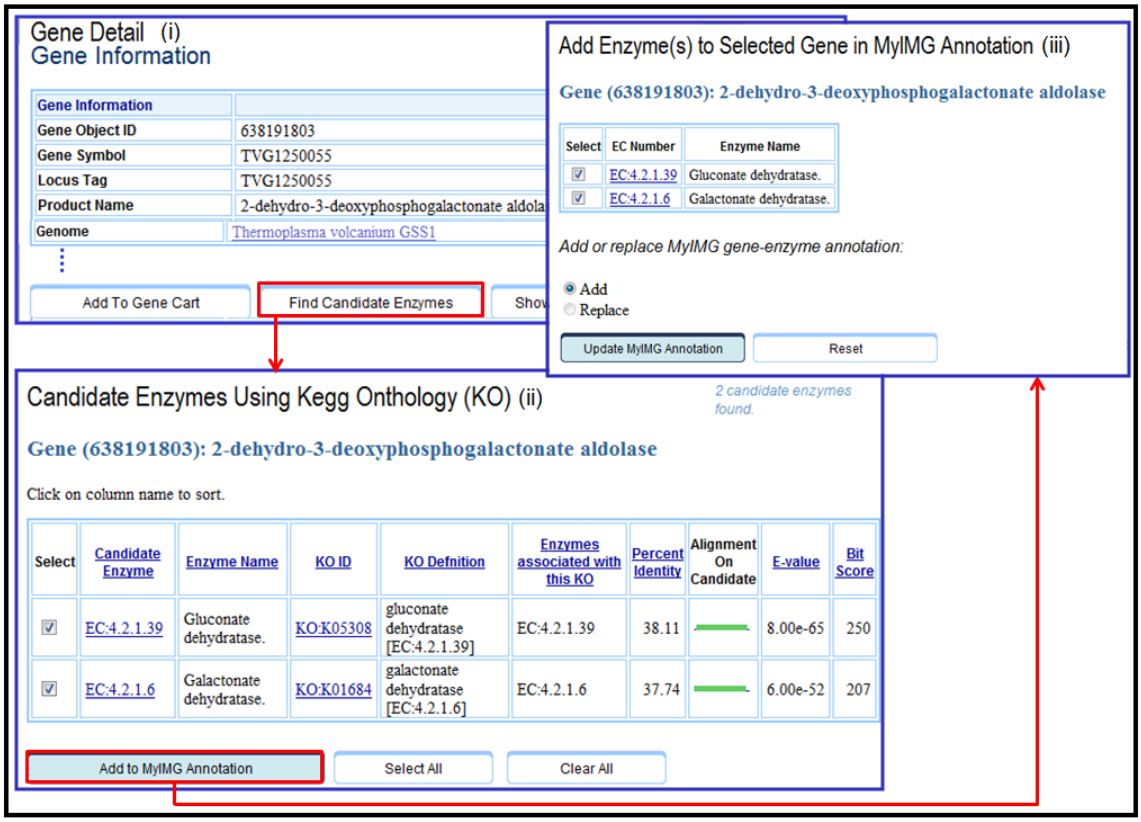


Figure 6.4. Examining Missing Enzymes for a Specific Gene.

Use **Add to MyIMG Annotation**, as shown in Figure 6.4(ii), and then either select **Add** or **Replace**, as shown in Figure 6.4(iii), to update the MyIMG enzyme annotation for this gene.

6.2.2 Missing Enzymes within a KEGG Pathway or Function Profile

MyIMG provides support for ***examining missing enzymes*** for a specific genome, ***G***, using either a **KEGG Pathway Map** for ***G*** or a **Functional Profile** involving ***G*** and enzymes of interest, as illustrated in Figure 6.5.

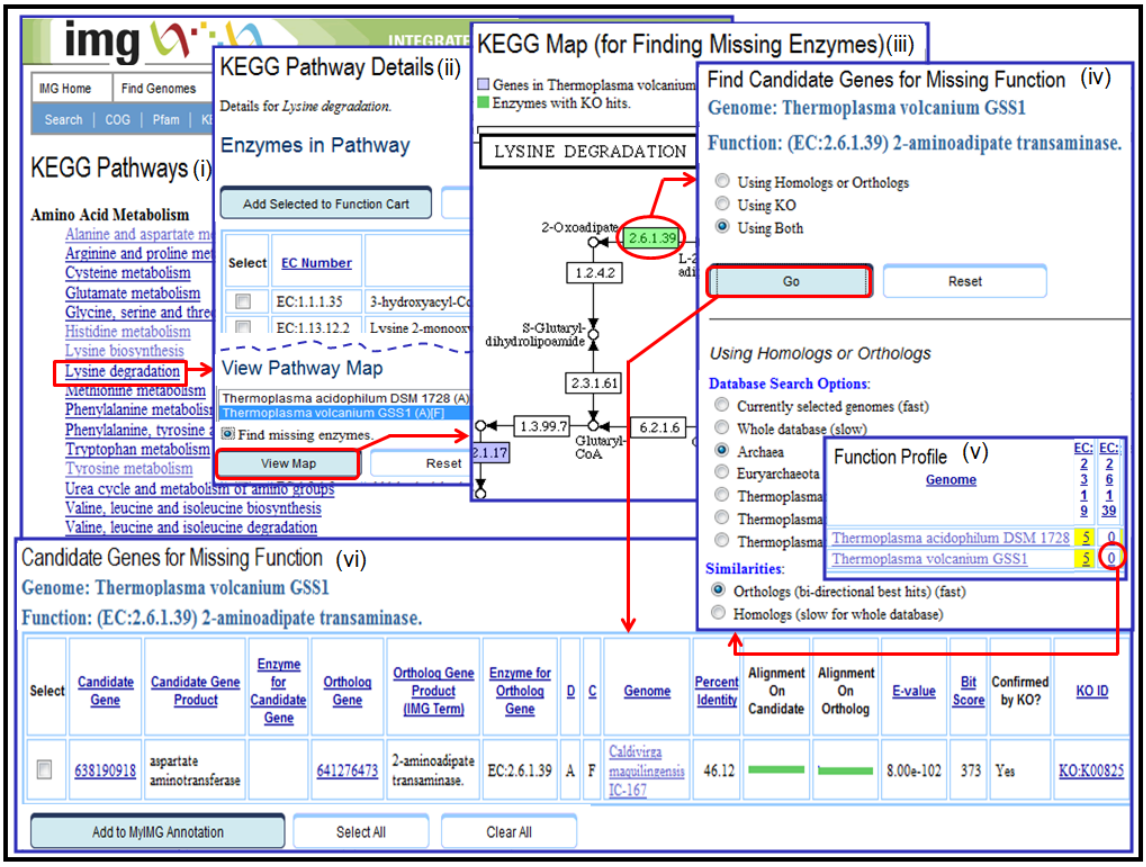


Figure 6.5. Examining Missing Enzymes with **KEGG Pathway Map** and **Function Profile**.

<u>Example 6.5(a)</u>. Use the **Genome Browser** for selecting *Thermoplasma volcanium GSS1 (T. volcanium)* and *Thermoplasma acidophilum DSM 1728 (T. acidophilum)* genomes. Save these selections. Select **Find Functions** in the Main Menu and then select the **KEGG** browser with the **KEGG Pathways via EC Numbers** option, as illustrated in Figure 6.5(i). Selecting the *Lysine degradation* pathway will lead to the **KEGG Pathway Details** page for this pathway, as shown in Figure 6.5(ii). You can either:

(i) Select **View Map** at the bottom of the **KEGG Pathway Details**, and use the "Find Missing Enzymes" option to display the KEGG map for the *Lysine degradation* pathway, as shown in Figure 6.5(iii), or

(ii) Select enzymes of interest from the list of enzymes (e.g., EC:2.3.1.9, EC:2.6.1.39) provided by the **KEGG Pathway Details,** save them in **Function Cart**, and compute a **Function Profile** for these enzymes across *T. volcanium* and *T. acidophilum* which will result in the profile shown in Figure 6.5(v). On the **KEGG Map**, enzymes that are associated with a *T. volcanium* gene are colored blue, while so called "missing" enzyme are colored either ***green***, for enzymes that have a KO term based prediction for a *T. volcanium* gene, or ***white***, for enzymes without any enzyme predictions for *T. volcanium* genes. Clicking on a missing enzyme such as EC:2.6.1.39, as illustrated in Figure 6.5(iii), will lead to a **Find Candidate Genes for Missing Function** page, as shown in Figure 6.5(iv). Note that selection of a (green colored) missing enzyme that has a KO term based enzyme prediction enhances the chances of finding for it good candidate genes.

In the **Function Profile** result, enzymes that are associated with *T. volcanium* or *T. acidophilum* genes are identified by positive integer numbers which represent the count of genes associated with the enzymes, while so called "missing" enzyme are identified by a "0". Clicking on the "0" identifying a missing enzyme such as EC:2.6.1.39, as illustrated in Figure 6.5(v), will lead to a **Find Candidate Genes for Missing Function** page, as shown in Figure 6.5(iv).

For missing enzymes of interest, **MyIMG** provides support for associating ***candidate genes*** with these enzymes.

Example 6.5(b). Consider "missing" enzyme **EC:2.6.1.39** discussed above, and leading to **Find Candidate Genes for Missing Function** page, as shown in Figure 6.5(iv).

You can find candidate genes of *T. volcanium* that could be associated with this enzyme as follows:

(i) **Search** for *T. volcanium* genes that have **homologs/orthologs,** associated with enzyme EC:2.6.1.39, or **KO** term based predictions for enzyme EC:2.6.1.39**,** or **both**, as illustrated in Figure 6.5(iv).

(ii) You can search across all the genomes available in the system, across a subset of genomes within a certain domain/phyla/class, or only across the selected genomes (i.e. *T. acidophilum).* If you have started from a **Function Profile**, you can also restrict the search to the genomes involved in the profile. You can change the default values set for percent identity and e-value cutoffs and the number of retrieved homologs. When the combination of **homolog/ortholog** based search and **KO** term based predictions are used jointly, the results are listed together. The result of this search lists *T. volcanium* candidate genes, as illustrated in Figure 6.5(vi). You can select a candidate gene and associate it with the "missing" enzyme using the **MyIMG Annotation** tool.

From the list of candidate genes you can select a candidate gene and go to the **Add Enzyme to Candidate Gene in MyIMG Annotation** tool shown in Figure 6.6(i). You can either add or replace the "missing" enzyme for the selected candidate gene. If you started your search for a candidate gene via a **Function Profile**, you can rerun the profile and see the effect of your annotation, as shown in Figure 6.6(ii) and 6.6(iii).

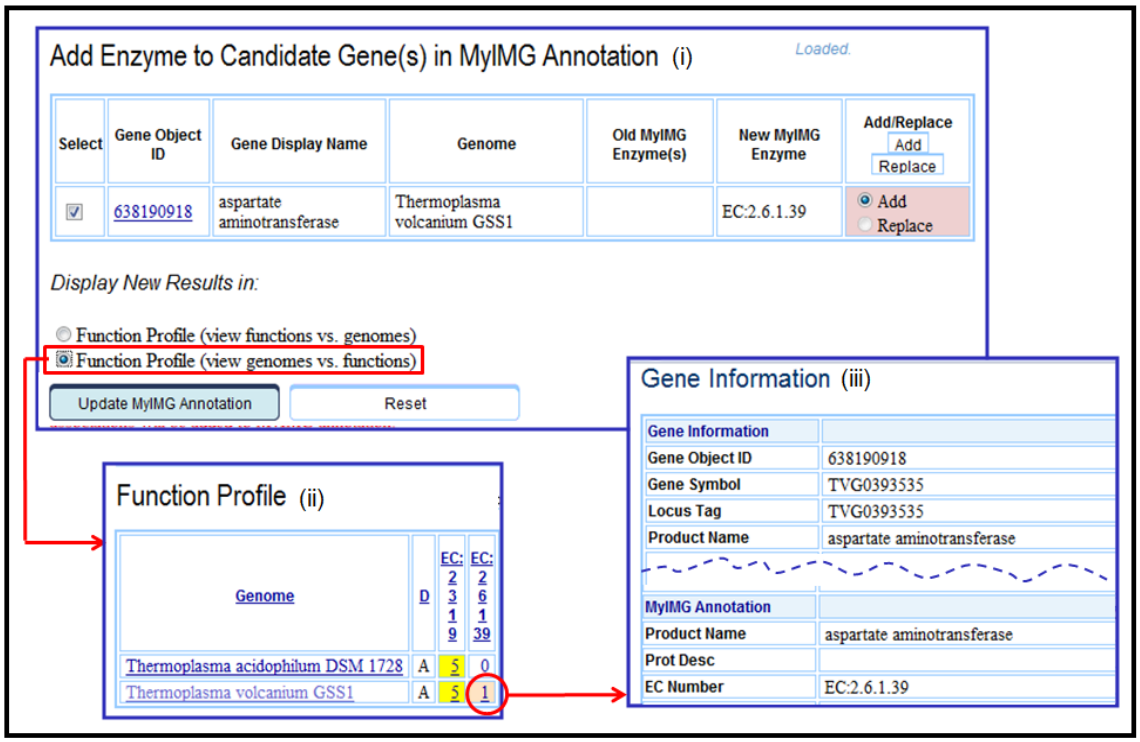


Figure 6.6. Associating a Candidate Gene with an Enzyme using **MyIMG Annotation**.

6.3 Missing Genes

Reviewing the genes and functional annotations for genomes may reveal genes that were missed by the gene prediction pipeline. Such potentially "***missing genes***" are usually found in IMG by comparing the gene content of related genomes with the **Phylogenetic Profiler for Single Genes**. This tool allows finding genes in a genome of interest that are present or missing (i.e., with or without homologs) in other genomes.

Example 6.6. Employ the **Phylogenetic Profiler for Single Genes** to find genes in the *T. volcanium* genome that are missing in its closely related genome *T. acidophilum*, as discussed in Exampl 4.1 and as shown in Figure 6.7(i). Examining the potentially unique genes shown in Figure 6.7(ii) reveals a 50S ribosomal protein L40E which is known as an essential gene, probably missed by the gene prediction pipeline.

Further review of potentially missing genes is provided by a new **Missing Gene** function that has been added to the **Phylogenetic Profiler** tool. In the example shown in Figure 6.7(ii), **Missing Gene** is applied on the 50S ribosomal protein L40E, which involves running TBLASTn of this *T. volcanium* gene's protein sequence against the *T. acidophilum* DNA sequence in order to determine whether it is missing in this genome, as shown in Figure 6.7(iii).

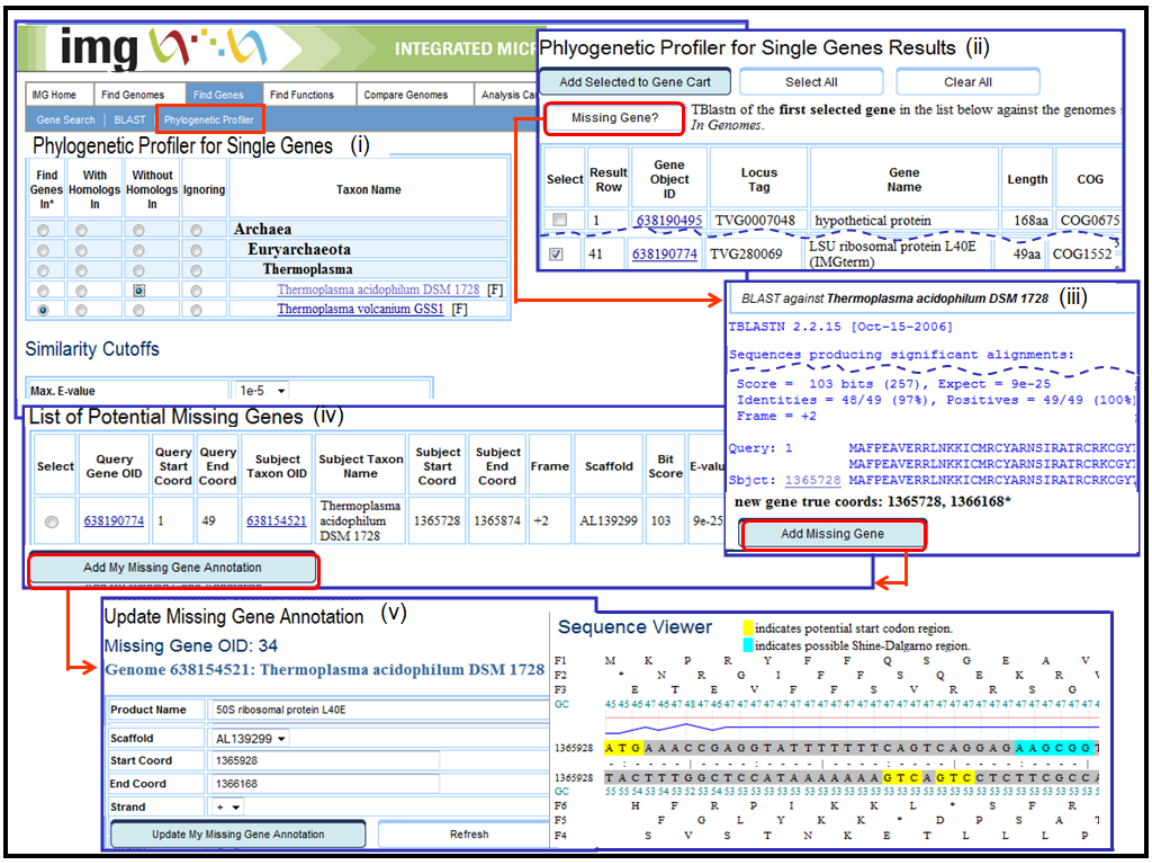


Figure 6.7. Finding and reviewing missing genes in IMG ER.

6.3.1 Handling Missing Genes within IMG ER

Missing gene curation in IMG ER is provided by the **Add Missing Gene Annotation** tool available in **Missing Gene**'s TBLASTn result.

Example 6.7. Consider the potentially missing gene discussed in Example 6.6 and shown in Figure 6.7(iii). The start and end coordinates of potentially missing genes are computed and provided at the end of the TBLASTN result. Use **Add Missing Genes** to display the list of the missing genes, as shown in Figure 6.7(iv). Each gene in this list can be examined using the **Update Missing Gene Annotation** tool, as illustrated in Figure 6.7(v). A **Sequence Viewer** which displays the six frame translation with putative ORF's, potential start codons, Shine-Delgano regions, and associated GC plot helps review the gene coordinates which can be adjusted.

Following the review of its coordinates, a missing gene can be recorded in IMG ER using **Update My Missing Gene Annotation**. While the homologs, paralogs, and orthologs of predicted genes are computed as part of a genome's inclusion into IMG ER, such computations are not performed for missing genes since they would affect all the genomes in the system. These computations are carried out by reloading the revised genome into IMG ER.

6.3.2 Handling Missing Genes outside IMG ER

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 review the missing gene outside IMG ER 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 shown in Figure 6.8(i). The navigator is under the "***Go to***" menu. Then click on "*Goto*" button.
3. The amino acid sequence is highlighted, as shown in Figure 6.8(ii). Go to "*Create*" menu and select "*Create feature from base range*", as illustrated in Figure 6.8(iii).
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 6.9(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 6.9(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 6.9(iii).

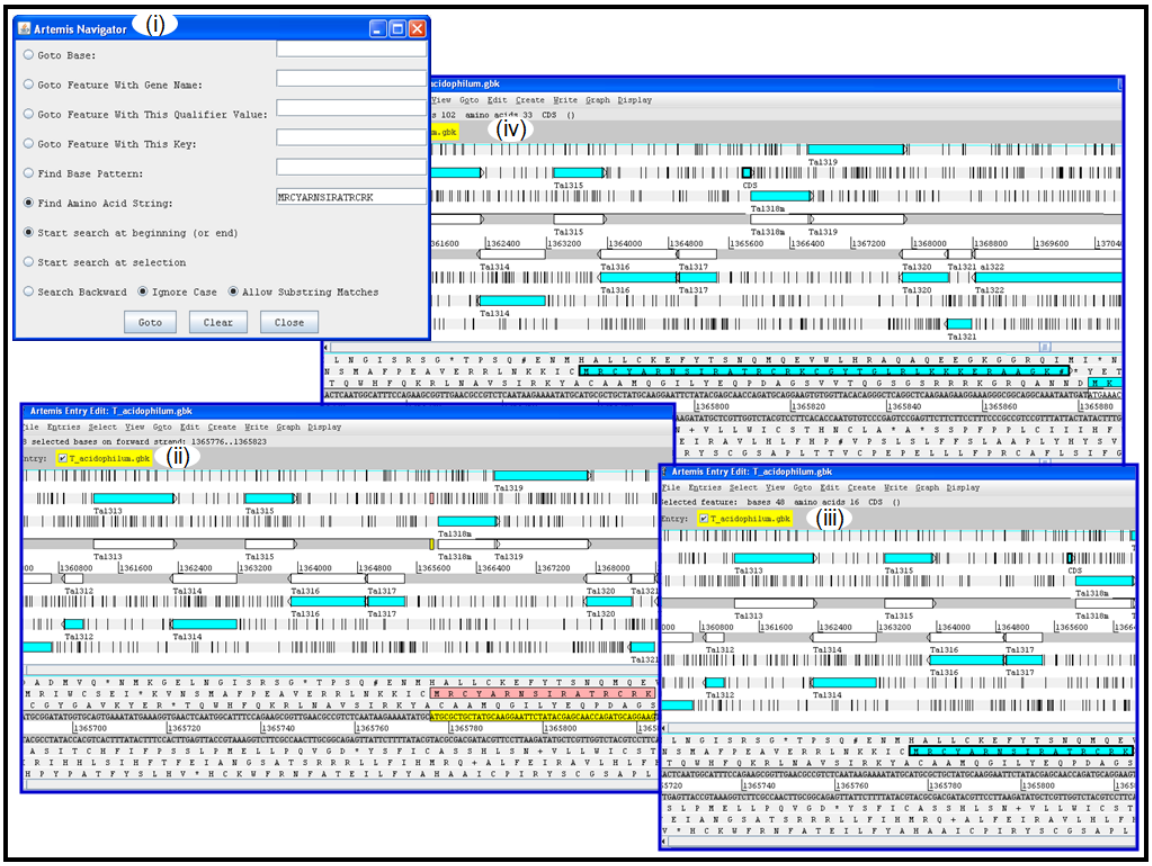


Figure 6.8. Using **Artemis** to define a missing gene.

Figure 6.9. Using **Artemis** to define a missing gene (cont.).

After determining the coordinates of a missing gene using a sequence editor such as **Artemis**, the gene can be recorded into IMG ER using **MyMissing Gene Annotations** as illustrated in Figure 6.10.

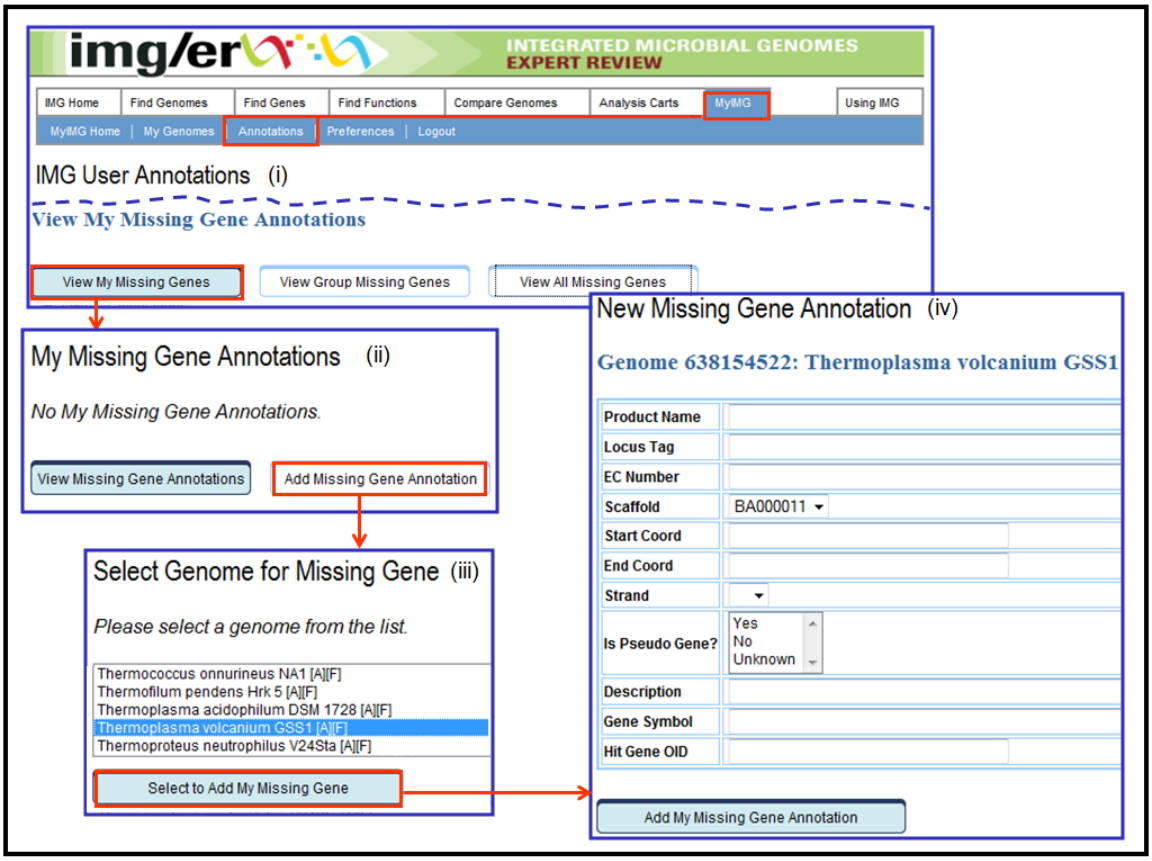


Figure 6.10. Entering a missing gene using **MyMissing Gene Annotations**.

<u>Example 6.8</u>. Consider a missing gene edited using Artemis. You can record this gene in IMG ER by first selecting **View My Missing Genes** available at the end of the IMG User Annotations page, as shown in Figure 6.10(i). Next, select Add Missing Gene Annotation, as shown in Figure 6.10(ii). Select the target genome, as shown in Figure 6.10(iii), then fill in the information regarding the gene using **New Missing Gene Annotation**, as shown in Figure 6.10(iv).

6.4 Deleting or Merging Genes

Reviewing the genes and functional annotations for genomes may reveal mispredicted genes, namely: (1) genes that need to be **deleted** (removed) from a genome; and (2) genes that need to be **merged** into a single gene.

In order to ***delete*** a gene, first include it into the **Gene Cart**. Next, use the **Annotate Selected Genes** link in the **MyIMG Annotation** section of the **Gene Cart** to access the **MyIMG Annotation** tool that allows editing the gene, and set to "Yes" the **Remove Gene from Genome** field.

In order to ***merge genes*** into a single gene, first delete all the genes as discussed above, then create a new gene using **MyMissing Gene Annotation** tool as discussed above in Example 6.8 and illustrated in Figure 6.10.

7 Reviewing MyIMG Annotations

MyIMG annotations for all the genes that you have curated can be reviewed using the **View My Annotations** section of the **IMG User Annotations** page. This page can be accessed using the **Annotations** sub-menu of the **MyIMG** main menu tab, as illustrated in Figure 7.1(i). Two review alternatives are available:

(i) All the genes can be displayed in a tabular format, where each row consists of the annotations for an individual gene, as illustrated in Figure 7.1(ii).

(ii) The genes are first displayed grouped per genomes, as illustrated in Figure 7.1(iii). You can select the list of genes for a specific genome to review their annotations.

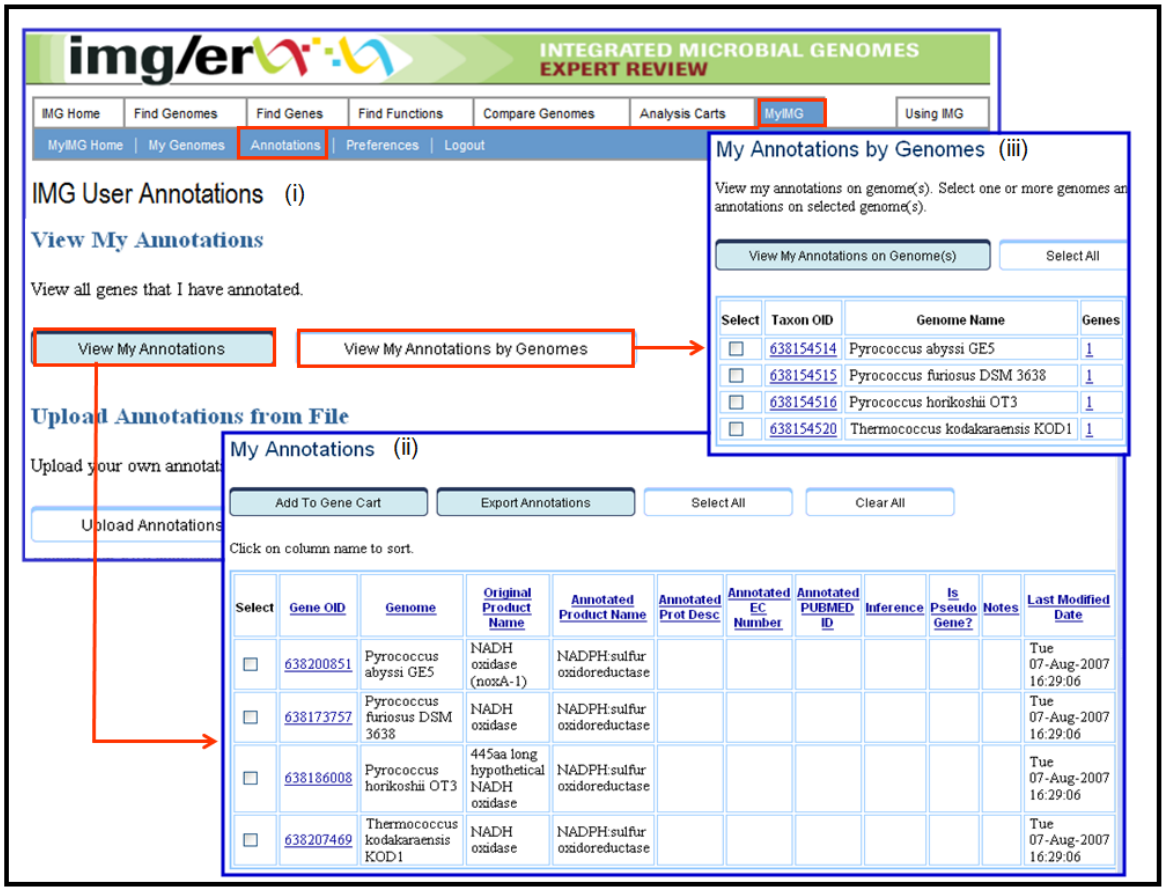


Figure 7.1. Review **MyIMG Annotations**.

The page displaying **MyIMG Annotations** for curated genes provides support for **exporting** these annotations to a tab-delimited file using **Export Annotations** (see Figure 7.1(ii)). 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.

Annotations can be uploaded using the **Upload Annotations** section of the **IMG User Annotations** page from a tab delimited file with the structure specified above. Only "gene_oid" and "MyIMG_Annotation" columns are mandatory, with the rest optional.

MyMissing Genes for the genes that you have curated can be reviewed using the **View My Missing Gene Annotations** section of the **IMG User Annotations** page, as illustrated in Figure 7.2(i). Your missing genes will be displayed in a tabular format, as illustrated in Figure 7.2(ii). You can select a gene of interest and review the information associated with it, including its start and end coordinates, its product name, etc., as illustrated in Figure 7.2(iii). You can update the information association with a missing gene using the Update Missing Gene Annotation, as illustrated in Figure 7.2(iv).

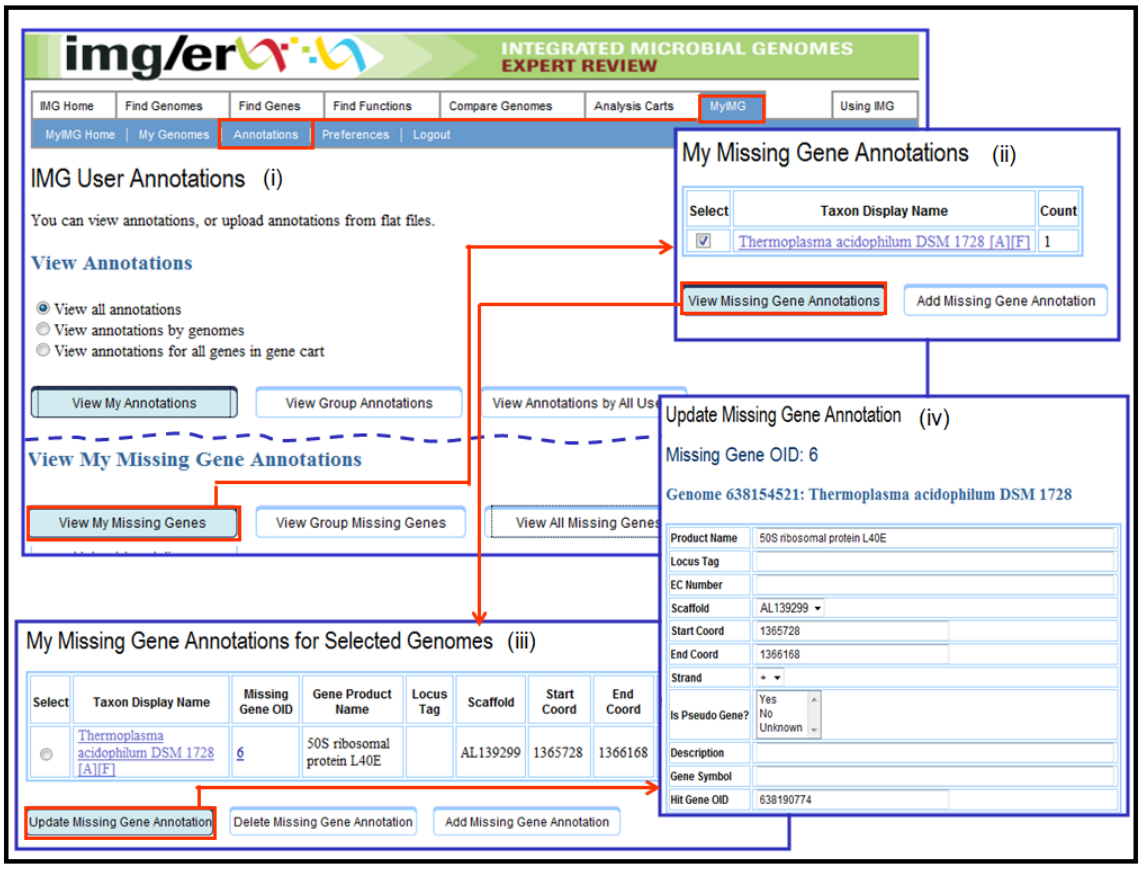


Figure 7.2. Review **MyMissing Gene Annotations**.

MyIMG annotations for all the genes of a specific genome can be also reviewed starting from the Organism Details page for that genome, as illustrated in Figure 7.3(i).

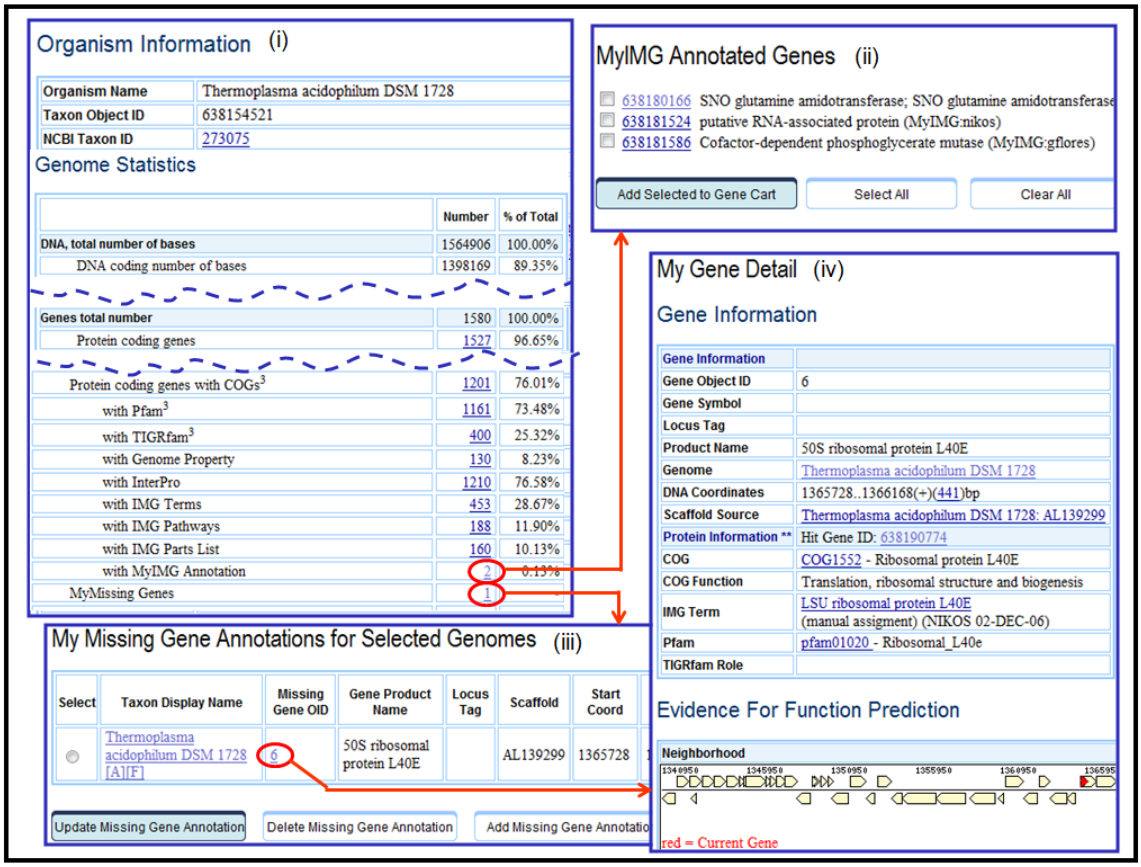


Figure 7.3. Review **MyIMG Annotations** for a Specific Genome.

Genes with MyIMG annotations are presented as a list, as shown in Figure 7.3(ii) and can be included into the **Gene Cart** or reviewed individually through their **Gene Details**. Missing genes, such as that shown in Figure 7.3(iii), can be reviewed using its **MyGene Details** page, as shown in Figure 7.3(iv), which includes its location on the scaffold as well as the functional annotation it has "inherited" from the gene used to identify it.

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.

Guy, L. (2006) Identification and characterization of pathogenicity and other genomic islands using base composition analyses. *Future Microbiol*. **1**, 309-316.

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.

Jermyn, W.S. and Boyd, E.F. (2002) Characterization of a novel Vibrio pathogenicity island (VPI-2) encoding neuraminidase (nanH) among toxigenic Vibrio cholerae isolates. *Microbiology* **148**, 3681-3693.

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 vsualisation 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/.

Glossary of Terms

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

Fusion – A hybrid gene formed from two previously separate genes; components of a fusion gene are the separate genes, while a composite gene is the result of the gene fusion.

Gene (or **protein**) **annotation** – A 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 the effect of a gene knockout 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 upper-case 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