

1 Synopsis

IMG supports MyIMG and missing gene annotations for users to add or update existing gene information and to facilitate community annotations.

MyIMG and Missing Gene Annotations requires a login and password which can be requested by filling the **Request Account** form at: <http://img.jgi.doe.gov/request>.

In order to view, add, or modify MyIMG or missing gene annotations, a user should first log into IMG ER to find the **Annotations** menu item under the **MyIMG** tab (Figure 1.1):

IMG User Annotations

You can view annotations, or upload annotations from flat files.

View Annotations

View all annotations
 View annotations by genomes
 View annotations for all genes in gene cart

[View My Annotations](#) [View Group Annotations](#)

Upload Annotations from File

Upload your own annotations from a tab-delimited file.

[Upload Annotations](#)

View My Missing Gene Annotations

[View My Missing Genes](#) [View Group Missing Genes](#)

Figure 1.1: IMG User Annotations

There are 3 sections on this page:

- **View Annotations:** view, add, or update user's MyIMG annotations. There is also a "View Group Annotations" button if the user belongs to an IMG group.
- **Upload Annotations from File:** update MyIMG annotations from a tab-delimited file.
- **View My Missing Gene Annotations:** view, add, or update user's missing gene annotations. There is also a "View Group Missing Genes" button if the user belongs to an IMG group.

2 MyIMG Curation of Functional 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 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) finding genes associated with **enzymes** missing in specific genomes; (3) **reviewing** the functional annotations for individual genes or group of genes; (4) associating genes with additional **IMG terms**; (5) marking genes **obsolete**; (6) **exporting to/ uploading** from a tab-delimited file functional annotations for genes identified by their IMG identifier (OID).

2.1 Editing Product Names

Genes that may require **product name review** and **curation** can be identified using various analysis tools. 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 are candidates for product name review and curation.

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

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 2.1. Consider for review gene PF1186 (IMG identifier 638173757) of genome *Pyrococcus furiosus* whose details are shown in Figure 2.1(i). This gene is associated with product name NADH oxidase, as shown in Figure 2.1(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 display via the Homologs section of its **Gene Details**, as shown in Figure 2.1(ii).

The screenshot displays the MyIMG web interface for gene review and curation. It is divided into several sections:

- Gene Details (i):** A table showing gene information for Gene Object ID 638173757. The product name is "NADH oxidase", which is circled in red. Other details include Locus Tag PF1186, Genome *Pyrococcus furiosus* DSM 3638, and Accession AAL81310.
- Top IMG Homolog Hits (ii):** A table listing three homologous genes with their product names, percent identity, and alignment scores. All three are checked for selection.

Select	Homolog	I	Product Name	Percent Identity	Alignment On Query Gene	Alignment On Subject Gene	Length
<input checked="" type="checkbox"/>	638207469	-	NADPH:sulfur oxidoreductase (MyIMG:gbp)	88.24	100%	100%	442aa
<input checked="" type="checkbox"/>	638200851	-	NADPH:sulfur oxidoreductase (MyIMG:gbp)	85.68	100%	100%	446aa
<input checked="" type="checkbox"/>	638186008	-	NADPH:sulfur oxidoreductase (MyIMG:gbp)	85.45	100%	100%	445aa
- Gene Cart (iii):** A table showing the selected genes from the homolog hits, including their Gene Object IDs, Locus Tags, and Product Names.

Selection	Gene Object ID	Locus Tag	Product Name
<input checked="" type="checkbox"/>	638173757	PF1186	NADH oxidase
<input checked="" type="checkbox"/>	638207469	TK1299	NADH oxidase
<input checked="" type="checkbox"/>	638186008	PH0572	NADH oxidase
<input checked="" type="checkbox"/>	638200851	PAB0936	NADH oxidase (noxA-1)
- MyIMG Annotation for Selected Genes (iv):** A table showing the current MyIMG annotations for the selected genes.

Select	Gene Object ID	Locus Tag	Original Product Name	Annotated Product Name	AA Seq. Length	Genome
<input checked="" type="checkbox"/>	638173757	PF1186	NADH oxidase	NADPH:sulfur oxidoreductase	442aa	<i>Pyrococcus furiosus</i> DSM 3638
<input checked="" type="checkbox"/>	638186008	PH0572	445aa long hypothetical NADH oxidase	NADPH:sulfur oxidoreductase	445aa	<i>Pyrococcus horikoshii</i> OT3
<input checked="" type="checkbox"/>	638200851	PAB0936	NADH oxidase (noxA-1)	NADPH:sulfur oxidoreductase	446aa	<i>Pyrococcus abyssi</i> GES
<input checked="" type="checkbox"/>	638207469	TK1299	NADH oxidase	NADPH:sulfur oxidoreductase	442aa	<i>Thermococcus kodakarensis</i> KOD1
- MyIMG Annotation:** A form for editing the product name and other details for the selected genes. The product name is currently "NADPH:sulfur oxidoreductase".

Figure 2.1. Review and curation of the product name for a gene of *Pyrococcus furiosus* using MyIMG Annotation.

Based on a recent study¹, 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 2.1(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 2.1(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 2.2.

¹ 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*.

2.2 Finding 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**.

2.2.1 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 PRIAM hits” which leads to a list of genes that could be associated with enzymes predicted by PRIAM. These predicted enzymes can be examined for accuracy and then associated with genes using **MyIMG Annotation** tools.

Example 2.2. 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 2.2(i). Follow the “Genes w/o enzymes but with candidate KO based enzymes” link to the list of genes that have candidate KO based enzymes, as shown in Figure 2.2(ii).

The figure shows two panels from the MyIMG Genome Browser interface for the genome *Thermoplasma volcanium* GSS1 (638154522).

Panel (i): Genome Statistics

Category	Number	% of Total
DNA, total number of bases	1584804	100.00%
DNA coding number of bases	1399584	88.31%
DNA G+C number of bases	632630	39.92% ¹
DNA scaffolds	1	100.00%
CRISPR Count	2	
genes total number	1615	100.00%
Protein coding genes	1561	96.66%
Pseudo Genes	64	3.96% ²
RNA genes	54	3.34%
rRNA genes	3	0.19%
5S rRNA	1	0.06%
16S rRNA	1	0.06%
23S rRNA	1	0.06%
IRNA genes	46	2.85%
Other RNA genes	5	0.31%
Protein coding genes with function prediction	1032	67.62%
without function prediction	469	29.04%
Protein coding genes connected to SwissProt Protein Product	95	5.88%
not connected to SwissProt Protein Product	1466	90.77%
Protein coding genes connected to SEED	931	57.65%
not connected to SEED	630	39.01%
Protein coding genes with enzymes	462	28.61%
w/o enzymes but with candidate KO based enzymes	15	0.93%

Panel (ii): Genes w/o enzymes but with candidate KO based enzymes

Select	Gene ID	Product Name	KO ID	KO Definition	Percent Identity	Alignment On Gene	E-value	Bit Score
<input checked="" type="checkbox"/>	638192022	DNA adenine modification methylase	KO:K07319	putative adenine-specific DNA-methyltransferase [EC:2.1.1.72]	50.00		0.00e+00	255.4
<input checked="" type="checkbox"/>	638191956	hypothetical protein	KO:K14182	choismate mutase / prephenate dehydrogenase [EC:5.4.99.5 / 1.3.1.12]	30.10		1.20e-07	57
<input checked="" type="checkbox"/>	638191927	hypothetical protein	KO:K12405	2-hydroxyacyl-CoA dehydrogenase / 3a, 7a, 12a-2-hydroxy-3-oxocholesterol 24-enoyl-CoA hydratase [EC:1.1.1.35 / 4.2.1.107]	33.30		3.50e-08	58.9
<input checked="" type="checkbox"/>	638191747	ABC transport system protein, SB1929A	KO:K01652	acetolactate synthase III/III large subunit [EC:2.2.1.6]	90.00		0.00e+00	226.5

Figure 2.2. Examining Missing Enzymes for a Specific Genome.

Select the genes you want to associate with the predicted enzymes and then use **Update MyIMG Annotation**, as shown in Figure 2.2(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 2.2(iii).

Instead of associating lists of genes with the candidate KO based enzymes, one can examine all enzyme predictions for individual genes via **Find Candidate Enzymes with PRIAM** available in the **Gene Information** section of the gene's **Gene Detail** page, as illustrated in Figure 2.3(i).

Example 2.3. In the list of genes of *Thermoplasma volcanium* GSS1 (*T. volcanium*) genome that have candidate KO based enzymes (see Figure 2.2 (ii) go to the **Gene Detail** page for the gene with IMG identifier 638191966. Use **Find Candidate Enzymes** immediately before the **Find Candidate Product Name** section, as shown in Figure 2.3(i), to get to the list predicted genes on the **Candidate Enzymes Using Kegg Ontology (KO)** page, as shown in Figure 2.3(ii). Examine the enzymes in the list and select the best predicted enzyme(s). Use **Add to MyIMG Annotation**, as shown in Figure 2.3(ii), and then either select **Add** or **Replace**, as shown in Figure 2.3(iii), to update the MyIMG enzyme annotation for this gene.

The figure shows three panels illustrating the workflow for finding and adding candidate enzymes to a gene's MyIMG annotation.

Panel (i): Gene Information

Gene Information	
Gene ID	638191966
Gene Symbol	TVG1420982
Locus Tag	TVG1420982
Product Name	hypothetical protein
SwissProt Protein Product	
SEED	
IMG Term	chorismate mutase (EC 5.4.99.5) (Iain Anderson 2005-04-28)
Genome	Thermoplasma volcanium GSS1
DNA Coordinates	1420446..1420982 (+)(537bp)
Scaffold Source	Thermoplasma volcanium GSS1 DNA: BA000

Buttons: Add To Gene Cart, Find Candidate Enzymes, Show All

Panel (ii): Candidate Enzymes Using Kegg Ontology (KO)

Gene (638191966): hypothetical protein

Filter column: Candidate Enzyme | Filter text: | Apply

Export | Page 1 of 1 | << first < prev 1 next > last >> | All

Select	Candidate Enzyme	Enzyme Name	KO ID	KO Definition	Enzymes associated with this KO
<input checked="" type="checkbox"/>	EC:1.3.1.12	Prephenate dehydrogenase	KO:K14187	chorismate mutase / prephenate dehydrogenase [EC:5.4.99.5 1.3.1.12]	EC:1.3.1.12, EC:5.4.99.5
<input checked="" type="checkbox"/>	EC:5.4.99.5	Chorismate mutase	KO:K14187	chorismate mutase / prephenate dehydrogenase [EC:5.4.99.5 1.3.1.12]	EC:1.3.1.12, EC:5.4.99.5

Buttons: Add to MyIMG Annotation, Select All, Clear All

Panel (iii): Add Enzyme(s) to Selected Gene in MyIMG Annotation

Gene (638191966): hypothetical protein

Select	EC Number	Enzyme Name
<input checked="" type="checkbox"/>	EC:1.3.1.12	Prephenate dehydrogenase
<input checked="" type="checkbox"/>	EC:5.4.99.5	Chorismate mutase

Add or replace MyIMG gene-enzyme annotation:
 Add Replace

hint: Click "Update MyIMG Annotation" to save your change(s) to the database. Only gene-enzyme associations will be added to MyIMG annotation.

Buttons: Update MyIMG Annotation, Reset

Figure 2.3. Examining Missing Enzymes for a Specific Gene.

2.2.2 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 2.4.

Example 2.4. 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 and **Pathways via EC Numbers** submenu item, as illustrated in Figure 2.4(i). Selecting the *Fatty acid metabolism* pathway under **Lipid Metabolism** will lead to the **KEGG Pathway Details** page for this pathway, as shown in Figure 2.4(ii). You can either:

- (i) Select the **View Map for Selected Genomes** tab, and then select *T. volcanium* and "Find missing enzymes" option, and click **View Map** at the bottom of the **KEGG Pathway Details** to display the KEGG map for the *Fatty Acid metabolism* pathway, as shown in Figure 2.4(iii), or
- (ii) Select enzymes of interest from the list of enzymes (e.g., EC:1.3.3.6, EC:1.3.99.3, and EC:6.2.1.3) 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 2.4(iv).

Lipid Metabolism (i)

- Arachidonic acid metabolism
- Biosynthesis of unsaturated fatty acids
- Cutin, suberine and wax biosynthesis
- Ether lipid metabolism
- Fatty acid biosynthesis
- Fatty acid elongation
- Fatty acid metabolism**
- Glycerolipid metabolism
- Glycerophospholipid metabolism
- Linoleic acid metabolism
- Primary bile acid biosynthesis
- Secondary bile acid biosynthesis
- Sphingolipid metabolism
- Steroid biosynthesis

Function Profile (iv)

3 functions and 2 genomes are selected.

Note: MyIMG gene-enzyme annotations (if any)

hint: Mouse over function ID to see name. (Cell coloring is highlighting of gene c

Click on column name to sort.

Domains(D): * =Microbiome, b=bin, B=Bacteria, A=Archaea, E=Eukarya, P=Plasmid

Show Slim View

Filter column: Genome Filter text:

Export Page 1 of 1 << first < prev 1 next > last >>

Genome	Domain	EC:1.3.3.6	EC:1.3.99.3	EC:6.2.1.3
Thermoplasma acidophilum DSM 1728	A	0	3	0
Thermoplasma volcanium GSS1	A	0	2	0

Export Page 1 of 1 << first < prev 1 next > last >> All

KEGG Pathway Details (ii)

Pathway: Fatty acid metabolism

Enzymes in Pathway Save to My Workspace View Map for Selected Genomes

Genome List

Domains: (*)=Microbiome, (B)acteria, (A)rchaea, (E)ukarya, (P)lasmids, (G)ragment, (V)iruses. Genome Completion: [F]inished, [P]ermanent Draft, [D]raft.

Please select genomes to highlight on KEGG Pathway Map.

Select All Clear All Show All Genomes

Select	Ignore	Taxon Name
<input type="radio"/>	<input type="radio"/>	Archaea
<input type="radio"/>	<input type="radio"/>	Euryarchaeota
<input type="radio"/>	<input type="radio"/>	Thermoplasma
<input type="radio"/>	<input type="radio"/>	Thermoplasma acidophilum DSM 1728 (A)[F]
<input checked="" type="radio"/>	<input type="radio"/>	Thermoplasma volcanium GSS1 (A)[F]

EC equivalents and genes in selected genome.
 Find missing enzymes.
 None

View Map Reset

KEGG Map (for Finding Missing Enzymes) (iii)

Current Genome: Thermoplasma volcanium GSS1

- Genes in Thermoplasma volcanium GSS1
- MyIMG annotated EC numbers
- Enzymes with KO hits
- Missing Enzymes

FATTY ACID METABOLISM

Fatty acid biosynthesis Fatty acid elongation Glycerolipid metabolism

Hexadecanoate (Fatty acid) 1.2.1.48 1.1.1.92 16-Hexadecanol

6.2.1.3

2.3.1.21 Hexadecanoyl-CoA

L-Palmitoyl-carnitine 1.3.3.6 1.3.99.3 1.3.99.3 1.3.99.3 1.3.99.3 1.3.99.3

trans-Hexadec-2-enoyl-CoA 4.2.1.17 4.2.1.74

(S)-3-Hydroxyhexadecanoyl-CoA 1.1.1.35 1.1.1.211

trans-Tetradec-2-enoyl-CoA 4.2.1.17 4.2.1.74

(S)-3-Hydroxytetradecanoyl-CoA 1.1.1.35 1.1.1.211

trans-Dodec-2-enoyl-CoA 4.2.1.17 4.2.1.74

(S)-3-Hydroxydodecanoyl-CoA 1.1.1.35 1.1.1.211

Figure 2.4. Examining Missing Enzymes with **KEGG Pathway Map** and **Function Profile**.

On the **KEGG Map**, enzymes that are associated with a *T. volcanium* gene are colored **blue**, enzymes that are associated with any *T. volcanium* MyIMG gene annotations are colored **cyan**, while so called “missing” enzyme are colored either **green**, for enzymes that have candidate enzymes with KO hits to a *T. volcanium* gene, or **yellow**, for enzymes without any KO hits to *T. volcanium* genes. Clicking on a missing enzyme such as EC:6.2.1.3, as illustrated in Figure 2.5(i), will lead to a **Find Candidate Genes for Missing Function** page, as shown in Figure 2.5(iii). Note that selection of a (green colored) missing enzyme that has a KO hit 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”.

For missing enzymes of interest, **MyIMG** provides support for finding **candidate genes** that could be associated with these enzymes, as illustrated in Figures 2.3 and 2.4.

Example 2.5. Consider “missing” enzyme **EC:6.2.1.3** identified as discussed in Example 2.4 above, and leading to **Find Candidate Genes for Missing Function** page, as shown in Figure 2.5(iii).

(i) KEGG Map (for Finding Missing Enzymes)

Current Genome: [Thermoplasma volcanium GSS1](#)

- Genes in Thermoplasma volcanium GSS1
- MyIMG annotated EC numbers
- Enzymes with KO hits
- Missing Enzymes

(ii) Column Selector

Genome	Domain	EC:1.3.3.6	EC:1.3.99.3	EC:6.2.1.3
Thermoplasma acidophilum DSM 1728	A	0	3	0
Thermoplasma volcanium GSS1	A	0	2	0

(iii) Find Candidate Genes for Missing Function

[Thermoplasma volcanium GSS1](#)
 Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Thermoplasmataceae;Thermoplasma;volcanium

Function: (EC:6.2.1.3) Long-chain-fatty-acid--CoA ligase.

Using Homologs
 Using KO
 Using Both

Go **Reset**

(iv) Candidate Genes for Missing Function

Genome: [Thermoplasma volcanium GSS1](#)

Function: (EC:6.2.1.3) Long-chain-fatty-acid--CoA ligase.

Domains(D): *Microbiome.
 B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids, G=Fragment, V=Viruses.
 Genome Completion(C): F=Finished, P=Permanent Draft, D=Draft.

Filter column: Candidate Gene Filter text: Apply

Export Page 1 of 1 << first < prev 1 next >> last >> All

Select	Candidate Gene	Candidate Gene Product	Enzyme for Candidate Gene	Homolog Gene	Homolog G (IMG)
<input type="checkbox"/>	638190519	hypothetical protein	EC:6.2.1.1	2509440882	Long-chain-fatty-acid--CoA ligase
<input type="checkbox"/>	638191204	medium-chain acyl-CoA ligase	EC:6.2.1.3	2502086732	Long-chain-fatty-acid--CoA ligase
<input type="checkbox"/>	638191734	acetyl-CoA synthetase	EC:6.2.1.1	2502089732	Long-chain-fatty-acid--CoA ligase

Using Homologs

This tool allows you to find genes associated with enzymes through homologs in other genomes. Homologs from the other genome [Thermoplasma volcanium GSS1](#) has homologs in other genomes associated with EC:6.2.1.3. These homologs have a reciprocal hits in the query genome, which are listed as candidates for gene-enzyme association.

Database Search Options:

Currently selected genomes (fast)
 Whole database (slow)
 Archaea
 Euryarchaeota

Figure 2.5. Finding Candidate Genes for **Missing Enzymes** on a **KEGG Pathway Map** using **Homologs**.

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:6.2.1.3, as illustrated in Figure 2.5(iii). 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. The result of this search lists *T. volcanium* candidate genes, as illustrated in Figure 2.5(iv). You can select a candidate gene and associate it with the “missing” enzyme using the **MyIMG Annotation** tool.
- (ii) Employ **KO** for finding *T. volcanium* genes that could be associated with the “missing” enzyme, as illustrated in Figure 2.6(ii). You can change the default values set for percent identity, e-value, and percent alignment cutoffs.

The result of this search lists *T. volcanium* candidate genes, as illustrated in Figure 2.6(iii). You can select a candidate gene and associate it with the “missing” enzyme using the **MyIMG Annotation** tool.

KEGG Map (for Finding Missing Enzymes) (i)

Current Genome: [Thermoplasma volcanium GSS1](#)

- Genes in Thermoplasma volcanium GSS1
- MyIMG annotated EC numbers
- Enzymes with KO hits
- Missing Enzymes

Column Selector (ii)

Genome	Domain	EC:1.3.3.6	EC:1.3.99.3	EC:6.2.1.3
Thermoplasma acidophilum DSM 1728	A	0	3	0
Thermoplasma volcanium GSS1	A	0	2	0

Using KO (iii)

Percent Identity Cutoff: 10

E-value cutoff: 1e-2

Bit-score cutoff: 0

Percent Alignment Cutoff: 10

Candidate Genes for Missing Enzyme Using KO (iv)

Genome: [Thermoplasma volcanium GSS1](#)

Function: (EC:6.2.1.3) Long-chain-fatty-acid--CoA ligase.

Filter column: Candidate Gene Filter text: Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Select	Candidate Gene	Candidate Gene Product	Enzyme for Candidate Gene	Confirmed by KO?	KO ID	KO Definition	Enzymes associated with this KO	KO Percent Identity	KO Alignment On Candidate	KO E-value
<input type="checkbox"/>	638190626	hypothetical protein		Yes	KO:K01897	long-chain acyl-CoA synthetase [EC:6.2.1.3]	EC:6.2.1.3	23	<div style="width: 100%; height: 10px; background-color: green;"></div>	1.80e-19

Figure 2.6. Finding Candidate Genes for Missing Enzymes on a KEGG Pathway Map using KO.

(iii) Employ a combination of **homolog/ortholog** based **search** and **KO**, which will list the results together.

From the list of candidate genes (e.g., see Figure 2.6(iv)), you can select a candidate gene and go to the **Add Enzyme to Candidate Gene(s) in MyIMG Annotation** tool shown in Figure 2.7(i). You can either add or replace the “missing” enzyme for the selected candidate gene. After the addition of MyIMG annotation, KEGG Map display will show enzyme EC:6.2.1.3 being color **cyan** instead of the original **green** (Figure 2.7 (ii)). Gene Detail of gene 638190626 (TVG0135483) page will display the new MyIMG annotation (Figure 2.7(iii)).

Add Enzyme to Candidate Gene(s) in MyIMG Annotation (i)

Select	Gene ID	Gene Display Name	Genome	Old MyIMG Enzyme(s)	New MyIMG Enzyme	Add/Replace
<input checked="" type="checkbox"/>	638190626	hypothetical protein	Thermoplasma volcanium GSS1		EC:6.2.1.3	<input checked="" type="radio"/> Add <input type="radio"/> Replace

hint: Click 'Update MyIMG Annotation' to save your change(s) to the database. Only selected gene-enzyme associations will be added to MyIMG.

Update MyIMG Annotation
Reset

Gene Detail (iii)

MyIMG Annotation	
Product Name	hypothetical protein
EC Number	EC:6.2.1.3
Is Public?	No

KEGG Map (for Finding Missing Enzymes) (ii)

Current Genome: [Thermoplasma volcanium GSS1](#)

- Genes in Thermoplasma volcanium GSS1
- MyIMG annotated EC numbers
- Enzymes with KO hits
- Missing Enzymes

FATTY ACID METABOLISM

The diagram shows a metabolic pathway starting with Hexadecanoate (Fatty acid). It is converted to Hexadecanoyl-CoA by the enzyme EC:6.2.1.3, which is highlighted in red. Other enzymes shown include 1.2.1.48 (Hexadecanoate to 16-Hexadecanal), 1.1.1.192 (16-Hexadecanal to 16-Hexadecanol), 2.3.1.21 (Hexadecanoyl-CoA to Hexadecanoyl-), and 6.2.1.2 (Hexadecanoyl- to Tetradecanoyl-). The pathway also branches into Tetradecanoyl- and Dodecanoyl-CoA.

Figure 2.7. Associating a Candidate Gene with an Enzyme using **MyIMG Annotation**.

2.3 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 2.8(i). Three 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 2.8(ii).
- (ii) The genes are first displayed grouped per genomes, as illustrated in Figure 2.8(iii). You can select the list of genes for a specific genome to review their annotations.
- (iii) Display is limited to only genes in gene cart.

The screenshot shows the 'My Annotations' section of the IMG User Annotations page. It is divided into three main areas:

- (i) Overview:** Includes options to 'View all annotations', 'View annotations by genomes' (selected), and 'View annotations for all genes in gene cart'. There are buttons for 'View My Annotations' and 'View Gro...'. Below this is the 'Upload Annotations from File' section with an 'Upload Annotations' button.
- (ii) Gene-level Annotations:** A table with columns: Select, Gene ID, Genome, Original Product Name, Annotated Product Name, Annotated Prot Desc, and Annotated EC Number. It lists genes from four genomes: *Dechloromonas aromatica* RCB, *Fusobacterium nucleatum* ATCC 25586, *Halobacterium* sp. NRC-1, and *Methanocaldococcus jannaschii* DSM 2661.
- (iii) Genome-level Annotations:** A table with columns: Select, Taxon ID, Genome Name, and Genes. It lists genomes: *Dechloromonas aromatica* RCB, *Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586, *Halobacterium* sp. NRC-1, *Methanocaldococcus jannaschii* DSM 2661, *Methanosarcina acetivorans* C2A, and *Methanosarcina mazei* Go1.

Figure 2.8. View MyIMG Annotations.

The page displaying **MyIMG Annotations** for curated genes provides support for **uploading** annotations to a tab-delimited file using **Upload Annotations** (see Figure 2.8(i)). The file has the following column headers:

- Gene ID or OID (**required**): gene object ID

- Annotated Product Name (**required**): my annotated product name(s); separate multiple product names using ','
- Annotated Prot Desc (optional): my annotated prot description
- Annotated EC Number (optional): my annotated EC number(s); separate multiple EC numbers using space or ','
- Annotated PUBMED ID (optional): my annotated PUBMED ID(s); separate multiple PUBMED ID's using space or ','
- Inference (optional): my annotated inference
- Is Pseudo Gene? (optional): is pseudo gene? (Yes, No)
- Notes (optional): my annotated free text notes
- Annotated Gene Symbol (optional): my annotated gene symbol
- Remove Gene from Genome? (optional): remove gene from genome? (Yes, No)

2.4 Group and Community Annotations

If a user belongs to an IMG group, then the user can view all MyIMG annotations by group members with the following restrictions:

- The user must have access permission to the genomes. All MyIMG annotations on private genomes will only be visible to other group members that have access permission to the private genomes. MyIMG annotations on public genomes are not restricted.
- The user can only view, but not modify, MyIMG annotations by other group members.

To view MyIMG annotations by group members, click the "View Group Annotations" button in the IMG User Annotations page shown in Figure 1.1. Many IMG users have used this feature for group annotations with colleagues.

MyIMG annotations are private by default. However, IMG users can change any of their MyIMG annotations to public. Public MyIMG annotations are visible to all IMG users provided that users have access permissions to the corresponding genomes. Public MyIMG annotations on public genomes can be viewed by all users.

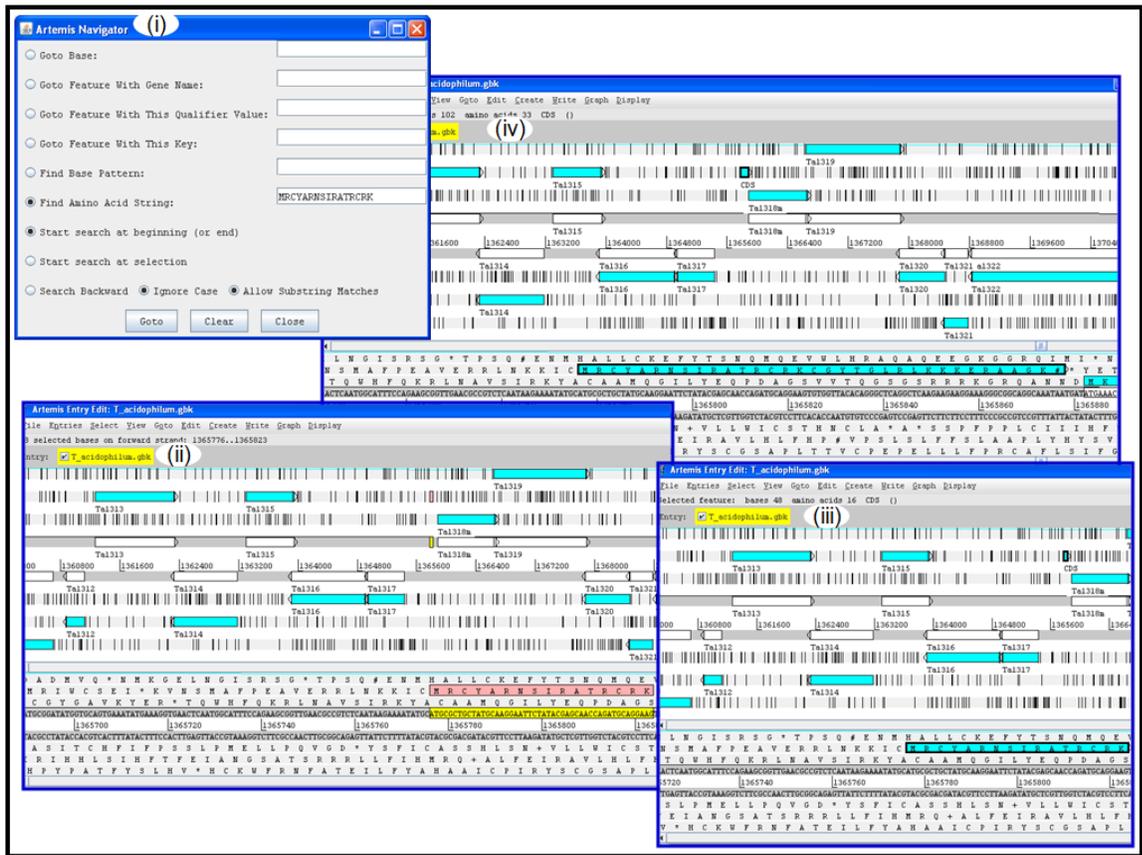
3 Missing Genes

The review of genes and their functional annotations may lead to the identification of missing genes.

3.1 Using Artemis to Find Missing Genes

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**² 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 3.1(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 3.1(ii). Go to the "Create" menu and select "Create feature from base range", as illustrated in Figure 3.1(iii).
4. To extend the gene, go to the "Edit" menu and select "Extend to next stop codon", then select "Fix stop codons".



² Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, Barrell B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 16 (10): 944-945.

Figure 3.1. Using **Artemis** to define a missing gene.

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 3.2(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 3.2 (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 3.2(iii).



Figure 3.2. Using **Artemis** to define a missing gene (cont.).

3.2 Viewing and Adding Missing Genes into IMG

After users identify missing genes in certain genomes, they can add missing genes into IMG. In the **IMG User Annotations** page (Figure 1.1), there is a **View My Missing Gene Annotations** section. Click the "View My Missing Genes" button to view all your missing genes (Figure 3.3(i)).

My Missing Gene Annotations (i)

Select	Taxon Display Name	Count
<input checked="" type="checkbox"/>	Hyperthermus butylicus DSM 5456 [A][F]	2
<input checked="" type="checkbox"/>	Methanococcoides burtonii DSM 6242 [A][F]	1

View Missing Gene Annotations

My Gene Detail (iii)

Gene Information

Gene ID	343
Gene Symbol	
Locus Type	CDS
Locus Tag	test
Product Name	(1->4)-alpha-D-glucan branching enzyme/(1->4)-alpha-D-glucan synthase (ADP-glucose)
Is Pseudogene?	No
Genome	Hyperthermus butylicus DSM 5456
DNA Coordinates	<2000..2300,3000..>3500 (+) (602bp) -- partial gene
Scaffold Source	Hyperthermus butylicus DSM 5456: NC_008818
Is Public?	No
Protein Information	
Amino Acid Sequence Length	267aa

My Missing Gene Annotations for Selection (ii)

Filter column: Genome Name Filter: text

Export Page 1 of 1 << first < prev 1 next > last

Select	Genome Name	Missing Gene ID	Gene Name
<input type="radio"/>	Hyperthermus butylicus DSM 5456 [A][F]	343	brer D (A
<input type="radio"/>	Hyperthermus butylicus DSM 5456 [A][F]	622	te
<input type="radio"/>	Methanococcoides burtonii DSM 6242 [A][F]	227	tRNA-Trp

Export Page 1 of 1 << first < prev 1 next > last >> All

Update Missing Gene Annotation Delete Missing Gene Annotation Add Missing Gene Annotation Add/Update IMG Term(s)

Figure 3.3. View My Missing Genes

Click "View Missing Gene Annotations" in **My Missing Gene Annotations** page will show a list of existing missing genes (Figure 3.3(ii)). Click on a specific Missing Gene ID will lead to the **My Gene Detail** page of that missing gene (Figure 3.3(iii)).

To add a missing gene, click the "Add Missing Gene Annotations" button to select the target genome, and then click the "Select to Add My Missing Gene" button.

A user can define the following properties of a missing gene:

- Product Name (required)
- Locus Type (required): CDS, tRNA, rRNA, miscRNA, or misc_feature
- Locus Tag (required)
- EC number: for enzyme(s)
- Scaffold
- DNA Coordinates (required):
 - Use comma to separate DNA coordinate ranges; e.g., 3146..3680,5982..8922

- Use '<' or '>' to indicate DNA coordinates in partial gene; e.g., <1..30,25..>75
- Strand: + or -
- Is Pseudo Genes? Yes or No
- Description
- Gene Symbol
- Hit Gene ID: if gene has any hits
- Is Public? Yes or No
- Replacing Gene(s): existing genes on the same scaffold to be replaced by this new missing gene

New Missing Gene Annotation

hint: (1) Fields with (*) are required.
 (2) Use comma to separate DNA coordinate ranges; e.g., 3146..3680,5982..8922
 (3) Use '<' or '>' to indicate DNA coordinates in partial gene; e.g., <1..30,25..>75
 (4) Missing genes with 'Is Public?' set to Yes will be visible to all users, *and may be modified by JGI experts.*
 (5) If this missing gene is replacing one or more existing genes, enter Gene OID(s) in the 'Replacing Gene(s)' field (comma delimited).

Aquiflexum balticum BA160, DSM 1650

Product Name (*)	hypothetical p
Locus Type (*)	CDS
Locus Tag (*)	Aquba_3556
EC Number	
Scaffold	Aquba Con
DNA Coordinates (*)	4139895..414
Strand	+ ▼
Is Pseudo Gene?	No Yes ▼
Description	
Gene Symbol	
Hit Gene ID	
Is Public?	No Yes ▼
Replacing Gene(s)	

My Missing Gene Neighborhood

My Gene Id: 0
 Start Coord. 4139895
 End Coord. 4144045
 Strand. +

red = Current Gene
 cyan or dashes = My Gene
 white = Pseudo Gene
 ||||| CRISPR array

Figure 3.4. Adding New Missing Gene Annotation

The "View Neighborhood" button in the **New Missing Gene Annotation** page enables users to graphically view the new missing gene on the selected scaffold in **My Missing Gene Neighborhood** popup (Figure 3.4).

Users can also add IMG terms to the missing genes. To add or update IMG terms of a missing gene, first select the missing gene in the **My Missing Annotations for Selected Genomes** window (Figure 3.3 (ii)) and then click the "Add/Update IMG Term(s)" button. Select proper IMG term(s), and then click the "Update Term Association" button to update.

3.3 Community Missing Gene Annotations in IMG

Missing gene annotations are private by default. A user can change his/her missing gene annotations to public. Public missing gene annotations are visible to all users who have access permission to the corresponding genomes. Moreover, since IMG ER is an "Expert Review" site, all public missing gene annotations can be reviewed and modified by JGI experts.

(i)

Aquiflexum balticum BA160, DSM 16537

Buttons: Add to Genome Cart, Browse Genome, BLAST Genome, Download Data

About Genome

- Overview
- Statistics
- Genes

Overview

Proposal Name, Organism Name, Taxon ID, IMG Submission ID, NCBI Taxon ID, GOLD ID in IMG Database, External Links, Genome Type, Lineage, Sequencing Status, IMG Release, Comment, Release Date, Add Date

Genome Statistics

Number of Chromosomal Cassettes	513	-
All Public Missing Genes	42	-
Fused Protein coding genes	42	0.82%
Protein coding genes coding signal peptides	546	10.67%

(ii)

All Public Missing Gene Annotations for Selected Genomes

Filter column: IMG Contact Filter text: Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector

Genome Name	IMG Contact	Missing Gene ID	Gene Product Name	Locus Type	Locus Tag	Scaffold	DNA
Aquiflexum balticum BA160, DSM 16537 [R][F]	Galina Ovchinnikova (GOvchinnikova@lbl.gov)	759	hypothetical protein	CDS	Aquba_3556	Aquba_Contig381	4139895..4142
Aquiflexum balticum BA160, DSM 16537 [R][F]	Galina Ovchinnikova (GOvchinnikova@lbl.gov)	760	hypothetical protein	CDS	Aquba_4317	Aquba_Contig381	5045483..5046

Figure 3.5. View Public Missing Genes from Genome Detail page.

If there are any public missing genes defined in a genome, in the **Genome Statistics** table of the **Genome Detail** page, there is "All Public Missing Genes" with count (see Figure 3.5 (i)). Click on the count will show a list of all public missing genes with additional information.

Public missing genes can be added into the IMG database after JGI expert review. After the database update, those public missing genes will become official genes in IMG.