BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 17 2009, pages 2271–2278
doi:10.1093/bioinformatics/btp393

Databases and ontologies

IMG ER: a system for microbial genome annotation expert review and curation

Victor M. Markowitz[1,*], Konstantinos Mavromatis[2], Natalia N. Ivanova[2], I-Min A. Chen[1], Ken Chu[1] and Nikos C. Kyrpides[2]

[1]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 and [2]Genome Biology Program, DOE Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, CA 94598, USA

Received on April 15, 2009; revised on June 20, 2009; accepted on June 22, 2009
Advance Access publication June 27, 2009
Associate Editor: Alex Bateman

ABSTRACT

Motivation: A rapidly increasing number of microbial genomes are sequenced by organizations worldwide and are eventually included into various public genome data resources. The quality of the annotations depends largely on the original dataset providers, with erroneous or incomplete annotations often carried over into the public resources and difficult to correct.

Results: We have developed an Expert Review (ER) version of the Integrated Microbial Genomes (IMG) system, with the goal of supporting systematic and efficient revision of microbial genome annotations. IMG ER provides tools for the review and curation of annotations of both new and publicly available microbial genomes within IMG's rich integrated genome framework. New genome datasets are included into IMG ER prior to their public release either with their native annotations or with annotations generated by IMG ER's annotation pipeline. IMG ER tools allow addressing annotation problems detected with IMG's comparative analysis tools, such as genes missed by gene prediction pipelines or genes without an associated function. Over the past year, IMG ER was used for improving the annotations of about 150 microbial genomes.

Contact: vmmarkowitz@lbl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A rapidly increasing number of microbial genomes are sequenced by organizations worldwide, undergo similar annotation procedures, and are eventually included into public genome data resources. First, raw ('read') sequences of microbial genomes are assembled into longer 'contigs' (contiguous sequences) in order to produce 'draft' genome sequences, with draft genomes sometimes 'finished' by closing gaps between contigs. Next, annotation pipelines are used for predicting genes and determining their functional roles in draft or finished genomes. Subsequently, annotated microbial genome sequences are submitted to/collected by primary archival public sequence data repositories, such as Genbank (Benson *et al.*, 2009), which perform data validation on genome datasets in order to ensure consistency of their format and, to a certain degree, their content. Datasets in these resources have different degrees of precision and resolution due to diverse annotation methods employed by individual data providers. Secondary public resources, such as NCBI's RefSeq (Pruitt *et al.* 2007), further process microbial genome data from primary resources with the dual goals of providing the most current view on microbial genome sequences and of gradually increasing the quality and completeness of their associated functional annotations via manual curation and computation. In addition to public primary and secondary resources, microbial genome datasets are incorporated into a variety of tertiary resources, such as SEED (Overbeek *et al.*, 2005) and IMG (Markowitz *et al.*, 2008a), which further revise microbial genome annotations that may be inaccurate and sparse.

While the combined validation and curation procedures of various data resources improve to a certain degree the quality and completeness of microbial genome annotations, erroneous or incomplete annotations are often carried over into the public resources and are difficult to correct (Salzberg, 2007). This problem is compounded by the rapid increase in the number of sequenced microbial genomes with incomplete and rarely curated annotations.

We have developed an Expert Review (ER) version of the Integrated Microbial Genomes (IMG) system, with the goal of supporting systematic and efficient revision of microbial genome annotations. IMG ER provides support for the review and curation of annotations for both new and publicly available microbial genomes in the broad integrated context of IMG's genomes. New genome datasets are usually included into IMG ER prior to their public release either with their original annotations or with annotations generated by IMG ER's annotation pipeline (http://img.jgi.doe.gov/w/doc/img_er_ann.pdf). IMG ER shares the goals of systems such as Manatee (http://manatee.sourceforge.net), MaGe (Vallenet *et al.*, 2006), PeerGAD (D'Ascenzo *et al.*, 2004), PseudoCAP (Winsor *et al.*, 2009), and ASAP (Glasner *et al.*, 2006), which provide mechanisms with different degrees of complexity for manual review of annotations, usually for specific organisms (PeerGAD, PseudoCAP), or groups of related organisms (ASAP). In contrast to organism specific annotation systems, IMG ER sets the curation within IMG's rich comparative genome context and diverse protein family and domain characterizations that are based on a variety of functional resources (see IMG Statistics at http://img.jgi.doe.gov).

*To whom correspondence should be addressed.

Comparative analysis in IMG allows detecting potential annotation gaps, namely genes that may have been missed by gene prediction tools and genes without predicted functions. IMG ER provides tools for filling such annotation gaps. A key goal in devising these tools was to provide seamless composition of analysis, review and curation operations. IMG's comparative analysis framework supports an effective revision process, whereby groups of related genes are handled jointly across multiple genomes. The development of IMG ER tools was driven by and applied to the genome analysis and curation needs of over 150 microbial genomes included into IMG ER since November 2007, such as *Halothermothrix orenii* (Mavromatis *et al*., 2009) and *Methanococcoides burtonii* (Allen *et al*., 2009).

A secondary goal of IMG ER is to provide support for enhanced metadata characterization of microbial genomes. Genomes in both IMG and IMG ER are characterized by metadata attributes, such as phenotype and habitat, which are based on the recommendations of the Genome Standards Consortium (Field *et al*., 2008). Metadata attribute values are hard to recover after genome sequences have been published, therefore they are collected at the time of genome submission to IMG ER and subsequently shared with public genome project resources such as GOLD (Liolios *et al*., 2008).

Gene annotations that result from expert review and curation are captured in IMG ER as so called 'MyIMG' annotations associated with individual scientist or group accounts. Genomes curated with IMG ER are included into Genbank either as new submissions or as revisions of previously submitted datasets, thus contributing to a coordinated improvement of the public genome data resources.

IMG ER is available at http://img.jgi.doe.gov/er and requires acquiring first an account at http://img.jgi.doe.gov/request.

2 METHODS

Microbial genome annotation is usually based on a combination of automated methods that generate a 'preliminary' annotation in terms of predicted protein-coding genes, also called Coding Sequences or CDSs, and assigning to genes protein product names that may describe the biological functions of gene products, such as enzymatic activity. A preliminary annotation may also suggest the placement of a gene product in various biological pathways and functional categories.

In general, microbial genome annotation reviews rely on the comparison of the genes and genomes of interest with other genes and genomes, whereby the comparison is based on gene (sequence) similarities, genome chromosomal context, and functional annotations.

2.1 Comparative genome context

IMG ER includes all publicly available genomes in IMG. IMG contains draft and complete genomes from all three domains of life integrated with a large number of plasmids and viruses. For example, IMG 2.8 (as of April 2009) contains a total of 4890 genomes consisting of 1284 bacterial, 59 archaeal, 49 eukaryotic genomes, 2524 viruses and 974 plasmids that did not come from a specific genome sequencing project. IMG ER is updated every four months with new genomes from IMG which in turn is updated with new genomes from RefSeq. This continuously growing set of genomes serve as reference for the annotation review and curation of unpublished (so called 'private') genomes in IMG ER.

Private genome datasets and associated metadata are submitted for inclusion into IMG ER via a web-based site (http://img.jgi.doe.gov/submit). Scientists have the option of including their genomes into IMG ER either with their native predicted genes and associated protein product names, or with genes and product names generated by IMG ER's annotation pipeline (http://img.jgi.doe.gov/w/doc/img_er_ann.pdf). For every genome included into IMG ER, candidate homologs for its genes are computed using BLASTp with 1*e*–2 *e*-value cutoff, and low complexity soft masking turned on. Candidate homolog lists are available for filtering by percent identity, bit score, and more stringent *e*-values. Genes are assigned various annotations based on functional resources, such as COG clusters (Tatusov *et al*., 2003), Pfam (Finn *et al*., 2008), TIGRfam (Selengut *et al*., 2007), and Gene Ontology (Gene Ontology Consortium, 2008). Genes are also associated with product names from the curated part of the SwissProt database (Gattiker *et al*., 2003), which is considered a reliable source of experimentally characterized protein products. EC numbers (Fleischmann *et al*., 2004) are assigned to genes using the KEGG Orthology (Kanehisa *et al*., 2008). Functional annotations are further characterized by their association with functional classifications including COG functional categories and the KEGG (Kanehisa *et al*., 2008) and MetaCyc (Caspi *et al*., 2008) pathway collections.

Protein product review is usually a time-consuming process because product names often consist of free-text descriptions which lack structure and are exposed to inconsistencies across genes and genomes. While the controlled vocabularies (enumerated lists of well defined and non-redundant terms) provided by COG clusters, Pfam, TIGRfam, KEGG Orthology, Swiss-Prot and Gene Ontology help reviewing protein products in IMG ER, the review process may still require examining and reconciling differences between different annotations and/or vocabularies. In order to facilitate this process, IMG ER provides a native collection of functional roles called 'IMG terms' that help mediate between diverse functional annotations. IMG terms form a hierarchy, where the leaves consist of *protein products* that are initially assigned manually to genes of genomes already in IMG by expert scientists in Joint Genome Institute's Genome Biology Program, and subsequently are propagated to genes of new genomes included into IMG ER via a conservative rule based term assignment mechanism.[1] The IMG term vocabulary consists currently of over 3900 protein products, with about 18% of genes associated with IMG terms (see the IMG Statistics section on IMG's home page). The gradual growth of the IMG term vocabulary and of the number of genes annotated with IMG terms is expected to help improve the efficiency of the protein product curation process.

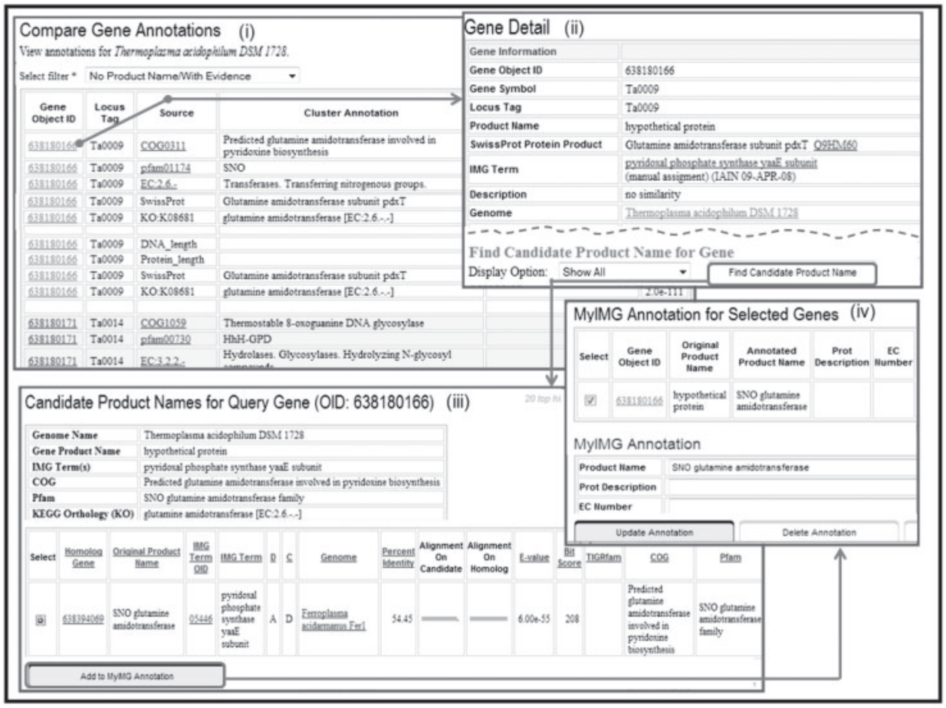
2.2 Genome annotation review

Genome data analysis in IMG consists of operations involving genomes, genes and functions which can be first selected and then explored individually (Markowitz *et al*., 2008a). Genomes, genes and functions can be selected using browsers and search tools. Various analysis tools allow comparing genomes in terms of gene content, functional capabilities and sequence conservation. The composition of search and comparative analysis operations is facilitated by gene and function 'carts' employed for recording and managing lists of genes and functions, respectively.

IMG data analysis operations have provided the foundation for devising the genome *annotation review and revision workflow* which involves: (i) *finding missing or problematic* annotations; (ii) *identifying candidates* to address missing or problematic annotations; (iii) *reviewing candidate* annotations; and (iv) *revising (curating) annotations*. Several IMG analysis tools were extended in order to support the first three stages of this workflow which involve only reading from the database underlying IMG. These tools are discussed in this section. IMG ER curation tools were then devised to support the fourth stage of the workflow which involves writing into the IMG database. These tools are discussed in the next section.

Protein products predicted for genes are one of the main targets for genome annotation review. The product name associated with a specific gene can be examined using IMG's 'Gene Details' in the context of the gene's protein

[1] A detailed discussion of the rationale for IMG terms is available at: http://img.jgi.doe.gov/pub/doc/imgterms.html.



Compare Gene Annotations (i)

View annotations for *Thermoplasma acidophilum DSM 1728.*

Select filter * No Product Name/With Evidence

Gene Object ID	Locus Tag	Source	Cluster Annotation
638180166	Ta0009	COG0311	Predicted glutamine amidotransferase involved in pyridoxine biosynthesis
638180166	Ta0009	pfam01174	SNO
638180166	Ta0009	EC:2.6.-	Transferases. Transferring nitrogenous groups.
638180166	Ta0009	SwissProt	Glutamine amidotransferase subunit pdxT
638180166	Ta0009	KO:K08681	glutamine amidotransferase [EC:2.6.-.-]
638180166	Ta0009	DNA_length	
638180166	Ta0009	Protein_length	
638180166	Ta0009	SwissProt	Glutamine amidotransferase subunit pdxT
638180166	Ta0009	KO:K08681	glutamine amidotransferase [EC:2.6.-.-]
638180171	Ta0014	COG1059	Thermostable 8-oxoguanine DNA glycosylase
638180171	Ta0014	pfam00730	HhH-GPD
638180171	Ta0014	EC:3.2.2.-	Hydrolases. Glycosylases. Hydrolyzing N-glycosyl compounds

Gene Detail (ii)

Gene Information	
Gene Object ID	638180166
Gene Symbol	Ta0009
Locus Tag	Ta0009
Product Name	hypothetical protein
SwissProt Protein Product	Glutamine amidotransferase subunit pdxT Q9HM60
IMG Term	pyridoxal phosphate synthase yaaE subunit (manual assigment) (IAIN 09-APR-08)
Description	no similarity
Genome	Thermoplasma acidophilum DSM 1728

Find Candidate Product Name for Gene

Display Option: Show All

MyIMG Annotation for Selected Genes (iv)

Select	Gene Object ID	Original Product Name	Annotated Product Name	Prot Description	EC Number
☑	638180166	hypothetical protein	SNO glutamine amidotransferase		

MyIMG Annotation

Product Name	SNO glutamine amidotransferase
Prot Description	
EC Number	

Candidate Product Names for Query Gene (OID: 638180166) (iii)

Genome Name	Thermoplasma acidophilum DSM 1728
Gene Product Name	hypothetical protein
IMG Term(s)	pyridoxal phosphate synthase yaaE subunit
COG	Predicted glutamine amidotransferase involved in pyridoxine biosynthesis
Pfam	SNO glutamine amidotransferase family
KEGG Orthology (KO)	glutamine amidotransferase [EC:2.6.-.-]

Select	Homolog Gene	Original Product Name	IMG Term OID	IMG Term	D	C	Genome	Percent Identity	Alignment On Candidate	Alignment On Homolog	E-value	Bit Score	TIGRfam	COG	Pfam
◉	638394069	SNO glutamine amidotransferase	05446	pyridoxal phosphate synthase yaaE subunit	A	D	Ferroplasma acidarmanus Fer1	54.45			6.00e-55	208		Predicted glutamine amidotransferase involved in pyridoxine biosynthesis	SNO glutamine amidotransferase family

Fig. 1. Missing protein product analysis in IMG and IMG ER often starts with (i) 'Compare Gene Annotations' for a genome of interest and usually involves (ii) examining the 'Gene Details' of genes without product names but with other functional annotation evidence, and (iii) finding 'Candidate Product Names' for these genes. The product names of the genes under review can be then (iv) curated in IMG ER using the 'MyIMG Annotation' tool.

family and domain characterization based on COGs, Pfams, TIGRfams, and its role within pathways, such as the KEGG and MetaCyc metabolic pathways. When available, IMG terms, SwissProt product names, KEGG Orthology (KO) terms, and Gene Ontology (GO) terms provide additional context for examining product names. Various viewers (e.g. for displaying the alignment of the gene sequence on the COG and Pfam representative sequences), and pre-computed lists of homologs, orthologs and paralogs, provide support for reviewing product names.

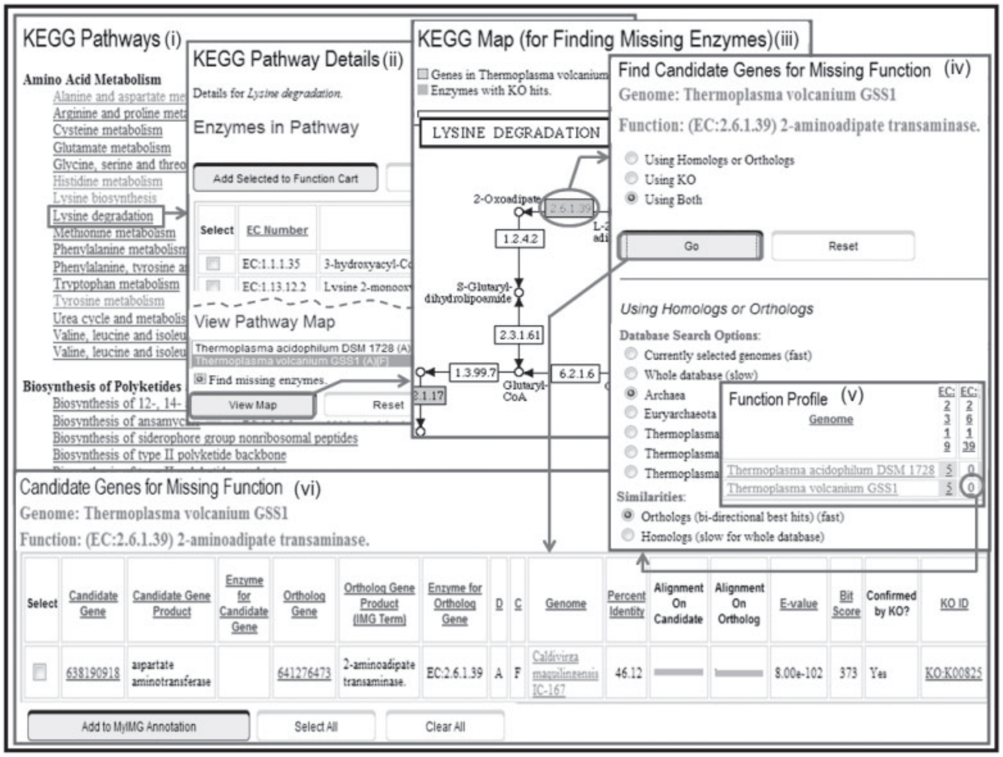
Finding missing or problematic protein products across an entire genome in IMG is provided with a new 'Compare Gene Annotations' tool. For a genome of interest, this tool provides the list of protein-coding genes and their predicted protein products together with the information about their membership in various protein families and descriptions of their functions based on this membership. This tool provides a quick way of assessing the quality of a gene's predicted protein product by comparing it with functions suggested by the gene's membership in protein families and identifying the most obvious discrepancies between the two. 'Compare Gene Annotations' allows focusing the review on genes *without a product name*, but *with evidence* of potential function provided by association with a COG, Pfam, TIGRfam, KO terms, IMG terms, and SwissProt product names, as illustrated in Figure 1(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 1(ii).

Identifying candidate product names for a query gene of interest is supported by a new 'Find Candidate Product Names' tool which retrieves the product names of the gene's closest homologs, as illustrated in Figure 1(iii). For each candidate product name, protein family and alignment information is provided in order to assist in choosing the appropriate product name. Subsequently, product name revision can be carried out with IMG ER curation tools, as discussed in the next section.

Enzymes predicted for genes are important for reviewing the metabolic capability of a genome. A genome is said to "miss an enzyme" on a pathway if it does not have any genes associated with an enzyme needed to catalyze a reaction on that pathway. Missing enzymes for a specific genome can be examined in IMG using the 'Genome Statistics' section of 'Organism Details' summary page which provides both the list of genes associated with enzymes predicted with conservative cutoff criteria, and the list of genes that could be associated with enzymes predicted using less restrictive criteria.[2] These enzyme predictions can be reviewed for accuracy and then associated with genes using IMG ER's 'MyIMG Annotation' tools.

Missing enzymes for a specific genome can be also examined in the context of a specific biological pathway, as illustrated in Figure 2. KEGG pathways can be selected using the 'KEGG' option of 'Find Functions' in IMG's Main Menu, as illustrated in Figure 2(i) where the *Lysine degradation* pathway is selected in order to examine its 'KEGG Pathway Details', as shown in Figure 2(ii). For a selected genome, the 'View Map' function of the 'KEGG Pathway Details' provides a graphical display of the pathway that highlights the enzymes (colored blue) that are associated with genes in this genome as well as with genes in other genomes. Highlighted enzymes are hyperlinked to 'Gene Details' of the associated genes. Missing enzyme review can be then carried out using the lists of homologs and orthologs provided by the 'Gene Details' for these genes via a sequence of analysis steps that is often complex and time consuming. A new 'Find Missing Enzymes' option added to the 'View Map' function facilitates the review of missing enzymes. Such enzymes are represented on the KEGG pathways either colored green, for enzymes that have a potential KO prediction for a gene in the target genome, or white otherwise, as shown in Figure 2(iii). For a missing enzyme, the 'Find Candidate Genes' tool, illustrated in Figure 2(iv), allows searching for candidate genes in the target genome

[2] The criteria used for predicting enzymes based on KEGG Orthology (KO) terms are available at: http://img.jgi.doe.gov/w/doc/dataprep.html.



KEGG Pathways (i)

Amino Acid Metabolism

- Alanine and aspartate me
- Arginine and proline meta
- Cysteine metabolism
- Glutamate metabolism
- Glycine, serine and threo
- Histidine metabolism
- Lysine biosynthesis
- Lysine degradation
- Methionine metabolism
- Phenylalanine metabolism
- Phenylalanine, tyrosine a
- Tryptophan metabolism
- Tyrosine metabolism
- Urea cycle and metabolis
- Valine, leucine and isoleu
- Valine, leucine and isoleu

Biosynthesis of Polyketides

- Biosynthesis of 12-, 14-
- Biosynthesis of ansamyci
- Biosynthesis of siderophore group nonribosomal peptides
- Biosynthesis of type II polyketide backbone

KEGG Pathway Details (ii)

Details for *Lysine degradation.*

Enzymes in Pathway

Add Selected to Function Cart

Select	EC Number	
□	EC:1.1.1.35	3-hydroxyacyl-C
□	EC:1.13.12.2	Lysine 2-monoox

View Pathway Map

Thermoplasma acidophilum DSM 1728 (A)
Thermoplasma volcanium GSS1 (A)(F)

☑ Find missing enzymes.

View Map | Reset

KEGG Map (for Finding Missing Enzymes)(iii)

□ Genes in Thermoplasma volcanium
▨ Enzymes with KO hits.

LYSINE DEGRADATION

Find Candidate Genes for Missing Function (iv)

Genome: Thermoplasma volcanium GSS1

Function: (EC:2.6.1.39) 2-aminoadipate transaminase.

○ Using Homologs or Orthologs
○ Using KO
● Using Both

Go | Reset

Using Homologs or Orthologs

Database Search Options:

○ Currently selected genomes (fast)
○ Whole database (slow)
● Archaea
○ Euryarchaeota
○ Thermoplasma
○ Thermoplasma
○ Thermoplasma

Similarities:

● Orthologs (bi-directional best hits) (fast)
○ Homologs (slow for whole database)

Function Profile (V)

Genome	EC: 2 3 1 9	EC: 2 6 1 39
Thermoplasma acidophilum DSM 1728	5	0
Thermoplasma volcanium GSS1	5	0

Candidate Genes for Missing Function (vi)

Genome: Thermoplasma volcanium GSS1

Function: (EC:2.6.1.39) 2-aminoadipate transaminase.

Select	Candidate Gene	Candidate Gene Product	Enzyme for Candidate Gene	Ortholog Gene	Ortholog Gene Product (IMG Term)	Enzyme for Ortholog Gene	D	C	Genome	Percent Identity	Alignment On Candidate	Alignment On Ortholog	E-value	Bit Score	Confirmed by KO?	KO ID
□	638190918	aspartate aminotransferase		641276473	2-aminoadipate transaminase.	EC:2.6.1.39	A	F	Caldivirga maquilingensis IC-167	46.12			8.00e-102	373	Yes	KO:K00825

Add to MyIMG Annotation | Select All | Clear All

Fig. 2. Missing enzyme analysis in IMG and IMG ER starts with (i) the selection of a KEGG pathway and (ii) selection of a genome of interest in order to (iii) display its missing enzymes on the pathway map. Missing enzyme links lead to a tool for (iv) 'Finding Candidate Genes' that could be associated with the missing enzyme. The analysis can also start with a (v) 'Function Profile' involving enzymes across genomes of interest, with links from missing enzymes to the 'Find Candidate Genes' tool. Candidate genes can be then (vi) examined and associated with the missing enzyme using the 'MyIMG Annotation' tool in IMG ER.

based on available KO based predictions and/or based on homologs or orthologs that are associated with the missing enzyme, such as EC:2.6.1.39 in Figure 2(iii). Homolog or ortholog based searches can be carried out across all the genomes available in IMG, across a subset of genomes within a certain domain/phyla/class, or only across a previously selected subset of genomes. The search can be adjusted for percent identity and *e*-value cutoffs and the number of retrieved homologs. For each candidate gene, alignment information is provided in order to assist in choosing the appropriate candidate.

Missing enzymes for a specific genome can be also examined using the 'Function Profile' tool which compares the abundance of specific enzymes across multiple genomes, as illustrated in Figure 2(v) for genomes *T. volcanium* and *T. acidophilum*. The positive integer numbers in each cell of the profile result represent the count of genes associated with a specific enzyme, while a missing enzyme is identified by a '0'. Clicking on the '0' identifying a missing enzyme, such as EC:2.6.1.39 in Figure 2(v), will lead to the 'Find Candidate Genes' tool discussed above and shown in Figure 2(iv). When a candidate gene for a missing enzyme is considered reliable, IMG ER provides the curation tools needed for revising its annotation, as discussed in the next section.

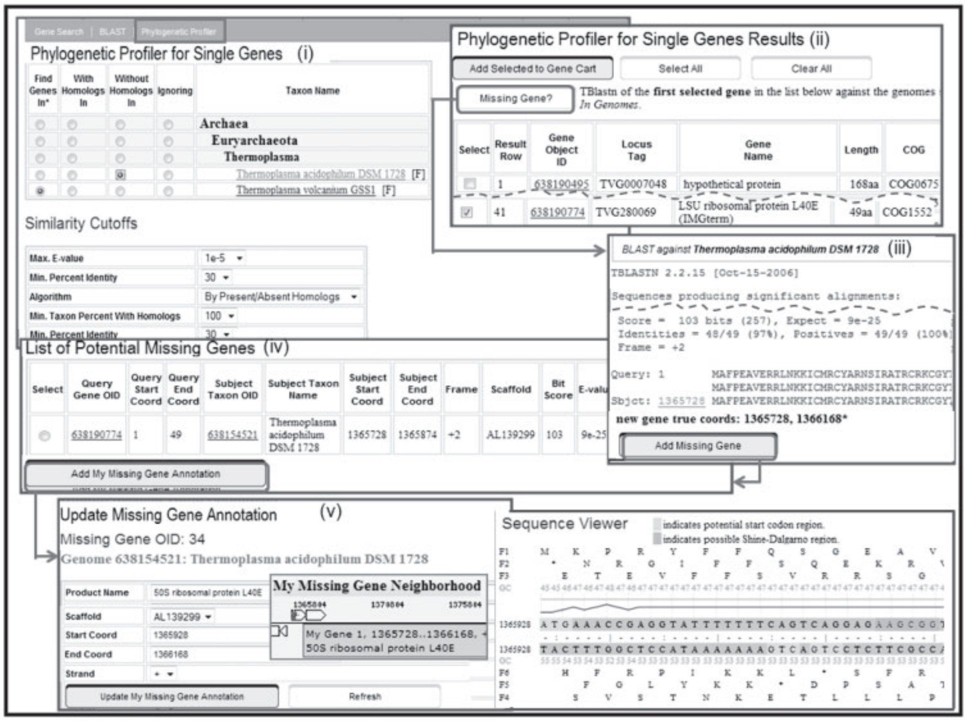
Annotation review may also reveal genes that may have been missed by the gene prediction pipeline. In IMG such potentially "missing genes" are usually found by comparing the gene content of related genomes with the 'Phylogenetic Profiler for Single Genes'. This tool allows finding genes in a genome of interest that are present or missing (i.e. with or without homologs) in other genomes. For example, the 'Phylogenetic Profiler' can be used to find genes in the *T. volcanium* genome that are missing in its closely related genome *T. acidophilum*, as shown in Figure 3(i), with similarity cutoffs available for fine-tuning the search. Examining the potentially unique genes shown in Figure 3(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 3(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 3(iii). When a potentially missing gene is found, IMG ER provides the curation tools needed for reviewing and placing this gene on the genome, as discussed in the next section.

3 RESULTS

IMG analysis tools, in particular the tools discussed in the previous section, are effective in revealing gaps in microbial genome annotations, namely genes with missing or inconsistent protein product names, missing enzymes in the context of biological pathways, and genes missed during gene prediction. IMG ER provides the curation tools needed for addressing such annotation gaps, whereby these tools are coupled seamlessly with the analysis tools used for reviewing annotations.

Annotation review and curation in IMG ER is carried out prior to a genome's public release. Users get password-protected access to their 'private' genomes and all publicly released genomes. Private



Phylogenetic Profiler for Single Genes (i)

Find Genes In*	With Homologs In	Without Homologs In	Ignoring	Taxon Name
○	○	○	○	**Archaea**
○	○	○	○	**Euryarchaeota**
○	○	○	○	**Thermoplasma**
○	○	●	○	Thermoplasma acidophilum DSM 1728 [F]
●	○	○	○	Thermoplasma volcanium GSS1 [F]

Similarity Cutoffs

Max. E-value	1e-5
Min. Percent Identity	30
Algorithm	By Present/Absent Homologs
Min. Taxon Percent With Homologs	100
Min. Percent Identity	30

Phylogenetic Profiler for Single Genes Results (ii)

TBlastn of the **first selected gene** in the list below against the genomes *In Genomes*.

Select	Result Row	Gene Object ID	Locus Tag	Gene Name	Length	COG
□	1	638190495	TVG0007048	hypothetical protein	168aa	COG0675
☑	41	638190774	TVG280069	LSU ribosomal protein L40E (IMGterm)	49aa	COG1552

BLAST against ***Thermoplasma acidophilum DSM 1728*** (iii)

```
TBLASTN 2.2.15 [Oct-15-2006]

Sequences producing significant alignments:

 Score =  103 bits (257), Expect = 9e-25
 Identities = 48/49 (97%), Positives = 49/49 (100%)
 Frame = +2

Query: 1       MAFPEAVERRLNKKICMRCYARNSIRATRCRKCGY
                MAFPEAVERRLNKKICMRCYARNSIRATRCRKCGY
Sbjct: 1365728 MAFPEAVERRLNKKICMRCYARNSIRATRCRKCGY
```

new gene true coords: 1365728, 1366168*

List of Potential Missing Genes (IV)

Select	Query Gene OID	Query Start Coord	Query End Coord	Subject Taxon OID	Subject Taxon Name	Subject Start Coord	Subject End Coord	Frame	Scaffold	Bit Score	E-valu
○	638190774	1	49	638154521	Thermoplasma acidophilum DSM 1728	1365728	1365874	+2	AL139299	103	9e-25

Update Missing Gene Annotation (V)

Missing Gene OID: 34

Genome 638154521: Thermoplasma acidophilum DSM 1728

Product Name	50S ribosomal protein L40E
Scaffold	AL139299
Start Coord	1365928
End Coord	1366168
Strand	+

My Missing Gene Neighborhood

My Gene 1, 1365728..1366168,
50S ribosomal protein L40E

Sequence Viewer

indicates potential start codon region.
indicates possible Shine-Dalgarno region.

```
F1       M   K   P   R   Y   F   F   Q   S   G   E   A   V
F2         *   N   R   G   I   F   F   S   Q   E   K   R
F3          E   T   E   V   F   F   S   V   R   R   S   G
GC
1365928 A T G A A A C C G A G G T A T T T T T T T C A G T C A G G A G A A G C G G
1365928 T A C T T T G G C T C C A T A A A A A A A G T C A G T C C T C T T C G C C
GC
F6        H   F   R   P   I   K   K   L   *   S   F   R
F5          F   G   L   Y   K   K   *   D   P   S   A
F4       S   V   S   T   N   K   E   T   L   L   L   P
```

Fig. 3. Missing gene analysis in IMG And IMG ER starts with (i) the 'Phylogenetic Profiler for Single Genes' for finding genes in a reference genome without homologs (i.e. missing) in the genome of interest. Subsequently, the (ii) 'Missing Gene' search can be applied on a gene selected from the 'Phylogenetic Profiler' result, with (iii) the result provided in TBLASTn format. Missing genes can be then (iv) examined and (v) curated using 'MyMissing Gene Annotation' tool in IMG ER.

genome datasets are included into IMG ER together with metadata attributes provided at the time of their submission.

3.1 Protein product curation

IMG analysis tools allow identifying genes that may require protein product curation, in particular genes without a product name but with evidence of potential functional annotation or with product name but without evidence of functional annotation. Protein product curation in IMG ER is provided by 'MyIMG Annotation' tool.

For genes, for which a product name has been identified with the 'Find Candidate Product Names' tool discussed above and illustrated in Figure 1(ii), 'MyIMG Annotation' tool is accessed via the 'Add to MyIMG Annotation' link available on the 'Candidate Product Names' results page, as illustrated in Figure 1(iii). 'MyIMG Annotation' allows editing the product name and associated information, such as protein description and EC number, as illustrated in Figure 1(iv). User annotations for a specific genome are recorded in the IMG ER database and can be further revised by the scientists who have access to the genome.

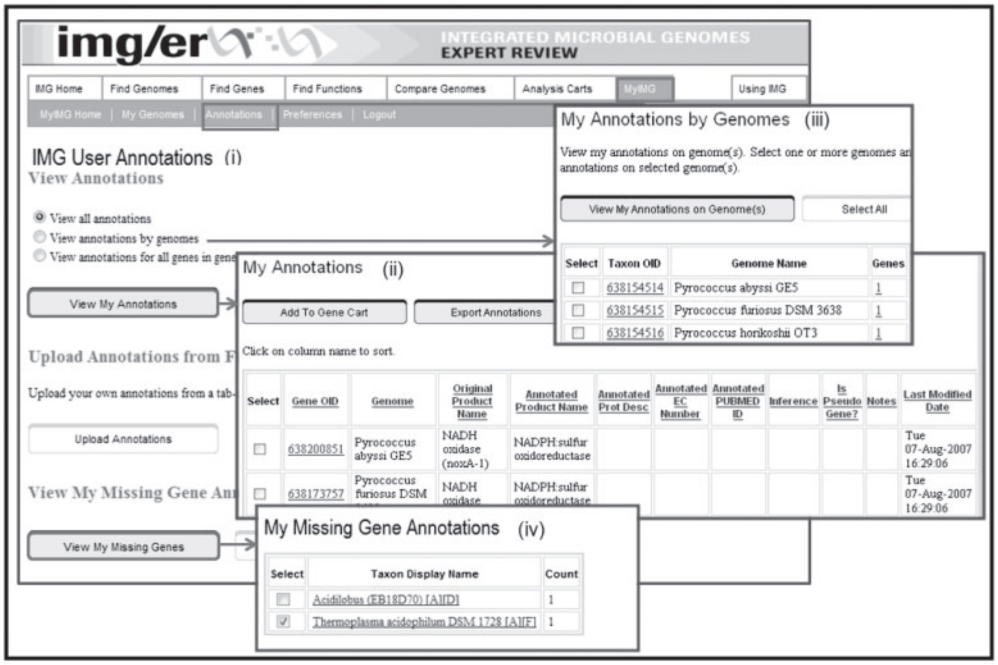
Protein product review can also be applied on a set of related genes that are included into IMG's 'Gene Cart'. Instead of searching for candidate product names using 'Find Candidate Product Names' tool of 'Gene Details', a scientist would start by reviewing the homologs of the query gene available in its 'Gene Details 'Homologs' section. Joint product name curation can be subsequently applied to the query gene and some of its homologs by first including them into the 'Gene Cart', and then with 'MyIMG Annotation' tool that is available via the 'Annotate Selected Genes' link in 'Gene Cart's MyIMG Annotation section.

Note that 'MyIMG Annotation' is not available in IMG where candidate product names can be reviewed for individual genes or groups of related genes, but cannot be curated.

3.2 Missing enzymes

Missing enzymes for a genome can be reviewed in IMG in the context of a specific KEGG pathway, as illustrated in Figure 2(iii) or using a 'Function Profile' involving specific enzymes across the genome of interest together with other genomes serving as reference context for comparison, as illustrated in Figure 2(v). For a missing enzyme of interest, the review results in a list of candidate genes that could be associated with the enzyme as discussed in the previous section and illustrated in Figure 2(vi). Missing enzyme curation in IMG ER is provided by 'Add Enzyme to Candidate Gene' tool available via the 'Add to MyIMG Annotation' link on the 'Candidate Genes for Missing Function' results page, as illustrated in Figure 2(vi). Note that in the example shown in Figure 2(vi), the candidate gene is identified using both a KO based prediction and a search for homologs across archaeal genomes. Note also that 'Add to MyIMG Annotation' is not available in IMG where candidate genes for missing enzymes can be reviewed but cannot be curated.

After a candidate gene is selected, the missing enzyme can be either added to the list of enzymes associated with the gene or can replace an existing enzyme. If the missing enzyme analysis was carried out in the context of a KEGG pathway, this pathway is



IMG User Annotations (i)

View Annotations

● View all annotations
○ View annotations by genomes
○ View annotations for all genes in gene

Upload Annotations from F

Upload your own annotations from a tab-

View My Missing Gene An

My Annotations (ii)

Click on column name to sort.

Select	Gene OID	Genome	Original Product Name	Annotated Product Name	Annotated Prot Desc	Annotated EC Number	Annotated PUBMED ID	Inference	Is Pseudo Gene?	Notes	Last Modified Date
□	638200851	Pyrococcus abyssi GE5	NADH oxidase (noxA-1)	NADPH:sulfur oxidoreductase							Tue 07-Aug-2007 16:29:06
□	638173757	Pyrococcus furiosus DSM	NADH oxidase	NADPH:sulfur oxidoreductase							Tue 07-Aug-2007 16:29:06

My Annotations by Genomes (iii)

View my annotations on genome(s). Select one or more genomes an annotations on selected genome(s).

Select	Taxon OID	Genome Name	Genes
□	638154514	Pyrococcus abyssi GE5	1
□	638154515	Pyrococcus furiosus DSM 3638	1
□	638154516	Pyrococcus horikoshii OT3	1

My Missing Gene Annotations (iv)

Select	Taxon Display Name	Count
□	Acidilobus (EB18D70) [A][D]	1
☑	Thermoplasma acidophilum DSM 1728 [A][F]	1

Fig. 4. Annotations for the genes of a genome undergoing revision within IMG ER can be reviewed using (i) 'View My Annotations' section of the 'IMG User Annotations'. Genes can be examined (ii) in a tabular format, where each row consists of the annotations for an individual gene, or (iii) grouped by individual genomes. (iv) Missing gene annotations can be reviewed separately.

redisplayed with the added enzyme colored in light blue indicating a user (MyIMG) annotation. If the missing enzyme analysis involved a 'Function Profile', this profile is recomputed in order to confirm the effect of the curation.

3.3 Missing genes

The 'Missing Gene' function of IMG's 'Phylogenetic Profiler for Single Genes' tool allows examining potentially missing genes, as discussed in the previous section. Missing gene curation in IMG ER is provided by 'Add Missing Gene Annotation' tool available in 'Missing Gene TBLASTn result. Thus, potentially missing genes, such as that shown in Figure 3(iii), can be reviewed and recorded. First, the start and end coordinates of potentially missing genes are computed and the list of these genes is provided for further review, as illustrated in Figure 3(iv). Each gene in this list can be then examined using 'Update Missing Gene Annotation', as illustrated in Figure 3(v). 'MyMissing Gene Neighborhood' viewer allows reviewing the new gene in the context of its chromosomal neighborhood, while a 'Sequence Viewer' helps review the gene coordinates by displaying the six frame translation with putative ORF's, potential start codons, Shine-Delgano regions, and associated GC plot. Following the review of its coordinates, a missing gene can be recorded in IMG ER's database.

Reviewing the genes predicted for a genome may also reveal problematic genes that need to (i) have their coordinates adjusted, or (ii) deleted (removed) from the genome, or (iii) be merged into a single gene. For all these cases, the 'Remove Gene from Genome' field of 'MyIMG Annotation' allows deleting the problematic genes after they are included in the 'Gene Cart' for annotation. Pseudogenes can be marked using a 'Pseudogene' binary field in 'MyIMG Annotation'. In order to adjust the coordinates of a gene or merge genes into a single gene, a new gene is then specified using a 'New Missing Gene Annotation' tool that is similar in structure and functionality to the 'Update Missing Gene Annotation' discussed above.

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 or new genes since they would affect all the genomes in the system. These computations are carried out by reloading the revised genome into IMG ER.

3.4 Annotation review

User annotations can be reviewed using the 'View My Annotations' section of the 'IMG User Annotations' page, as illustrated in Figure 4(i). Two review alternatives are provided: genes can be displayed in a tabular format, where each row consists of the annotations for an individual gene, as illustrated in Figure 4(ii); or genes are displayed grouped by individual genomes, as illustrated in Figure 4(iii). Missing genes can be reviewed separately, as illustrated in Figure 4(iv). User annotations can be exported to/ importer from tab-delimited files.

Gene annotations are revised either by individual scientists or groups of scientists working jointly. For group reviews, each scientist can see all the annotations within the group, but cannot override other scientist's annotation. Groups usually have a leader with editorial privileges and coordination responsibilities. In most cases scientists within a group work on different sets of genes or different areas, such as examining different metabolic pathways. Potential conflicts are resolved through direct interactions between scientists or through the group leader. This revision strategy seems to work without problems, mainly because scientists within a group know and trust each other.

From September 2007 to June 2009, the annotations of about 20 500 genes were revised across 380 genomes with IMG ER. For 9 of these genomes the annotation review lead to publications, including (Anderson *et al*., 2009; Alen *et al*., 2009; Herlemann *et al*., 2009 and Mavromatis *et al*., 2009). In addition to individual genome reviews, the annotations of a group of 56 Genomic Encyclopedia for Bacteria and Archaea (GEBA) genomes (http://www.jgi.doe.gov/programs/GEBA/pilot.html) were revised by JGI scientists using IMG ER (Wu *et al*., submitted for publication).

Comprehensive annotation reviews, such as that conducted for *M. burtonii* (Alen *et al*., 2009) where a group of scientists revised the annotations for 2431 genes over a period of several years require a major investment of time and therefore are not frequent. Such large-scale annotation reviews sometime involve an evidence rating system set up by the reviewers. For example, the protein products of *M. burtonii* genes are associated with an evidence rating (ER) ranging from ER1 (experimentally characterized function) to ER5 (no evidence for function). IMG ER' 'MyIMG Annotation' provides both an inference field for entering such ratings and a field for recording bibliographic references when available.

Once a genome's annotation review is completed, IMG ER provides tools for generating either the submission file required for including a new genome into Genbank or the revision file required for updating the annotations for a genome that is already available in Genbank.

4 DISCUSSION

IMG ER curation tools were devised as extensions of the IMG analysis tools generally employed for reviewing microbial genome annotations. Such reviews are usually performed in the context of phylogenetically related genomes, whereby discrepancies between the annotations of the target genome to that of close phylogenetic neighbors serve as warnings for potential problems. Accordingly, the 'Phylogenetic Profiler' tool that allows examining differences in gene content across related genomes provides the basis for missing gene curation, while the 'Function Profile' tool that allows examining differences in protein and functional families across genomes, provides the basis for missing enzyme curation.

The IMG ER curation tools aim at improving the efficiency of the annotation revision process. These tools were gradually extended by evaluating them in the context of microbial genome studies, such as *Methanococcoides burtonii* (Alen *et al*., 2009). The protein product curation tools, which were part of the first IMG ER version released to users in 2007, have been applied to the manual annotation of tens of genomes. The development of the missing enzyme curation tools was driven by the *Halothermothrix orenii* study (Mavromatis *et al*., 2009), whereby these tools helped reduce substantially the time required for the analysis of potentially missing enzymes in metabolic pathways.[3] The missing gene curation tools provide a streamlined version of protein coding gene analysis available in sequence annotation systems such as Artemis (Rutherford *et al*., 2000). While users have the option of using such systems in conjunction with IMG ER, IMG ER missing gene curation allows them to save time by remaining within the framework of a single system where they can record all their annotations.

IMG ER genome annotation curation capabilities continue to be extended. For missing enzyme curation, finding candidate genes currently is carried out in the context of KEGG pathways and relies on KEGG Orthology based enzyme predictions. The breadth of candidate gene searches will be extended through additional PathoLogic enzyme predictions (Green and Karp, 2004) generated in the context of MetaCyc/BioCyc pathways (Caspi *et al*., 2008). This extension has been enabled through the inclusion of MetaCyc pathways into IMG, with the addition of PathoLogic predicted enzymes planned for future versions of IMG and IMG ER.

A metagenomic specific counterpart for IMG ER, IMG/M ER, provides the same capabilities as IMG ER for the review and curation of metagenome datasets. IMG/M ER is based on the IMG/M metagenome data management and analysis system (Markowitz *et al*., 2008b) and contains the same reference baseline of isolate genomes as IMG ER, integrated with 65 public metagenome datasets from IMG/M (http://img.jgi.doe.gov/m). In addition, IMG/M ER contains 275 unpublished (i.e. 'private') metagenome datasets that are part of 85 ongoing metagenome studies. Metagenome datasets are included into IMG/M ER via the same submission site as that employed for IMG ER, and involves collecting the values for a comprehensive set of metagenome specific metadata attributes.

Metagenome datasets are substantially more complex and large than isolate genome datasets, and are inherently fragmented and incomplete. While the missing enzyme tools are used to expand the annotation coverage for metagenome datasets, a review and curation process similar to that for isolate genomes is seldom used for such datasets. Instead, IMG/M ER is mainly employed for examining the functional capabilities of metagenome datasets in the context of isolate reference genomes. IMG/M ER's comparative analysis tools allow detecting assembly or gene prediction errors (Martin *et al*., 2006) which may lead to reprocessing the datasets and then reloading them into IMG/M ER as part of an iterative review process. Since the sequencing technology platforms and data processing methods employed for generating metagenome datasets keep evolving, we will continue to observe how scientists review such datasets in order to extend IMG/M ER with additional analysis and curation capabilities.

ACKNOWLEDGEMENTS

The authors thank Krishna Palaniappan, Ernest Szeto, Yuri Grechkin, Iain Anderson and Athanasios Lykidis, for their contribution to the development and maintenance of IMG ER.

Funding: Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, US Department of Energy, under Contract No. DE-AC02-05CH11231.

Conflict of Interest: none declared.

REFERENCES

Alen,M.A. *et al*. (2009) The genome sequence of the psychrophilic archaeon, *Methanococcoides burtonii*: the role of genome evolution in cold adaptation. *ISME J.*, [Epub ahead of print, doi: 10.1038/ismej.2009.45., April 30, 2009]

Anderson,I. *et al*. (2009) Genome analysis of the sulfur-reducing Crenarchaeote Staphylothermus marinus. *BMC Genomics*, **10**, 145.

[3] Analyzing *H.orenii* with IMG ER is described in IMG's documentation at http://img.jgi.doe.gov/w/doc/Halothermothrix_orenii_case_study.pdf.

Benson,D.A. *et al*. (2009) Genbank. *Nucleic Acids Res.*, **37**, D26–D31.

Caspi,R. *et al*. (2008) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **36**, D623–D631.

D'Ascenzo,M.D. *et al*. (2004) PeerGAD: a peer-review-based and community-centric web application for viewing and annotating prokaryotic genome sequences. *Nucleic Acids Res.*, **32**, 3124–3135.

Field,D. *et al*. (2008) The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnol.*, **26**, 541–547.

Finn,R.D. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

Fleischmann,A. *et al*. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.*, **32**, D434–D437.

Gattiker,A. *et al*. (2003) Automatted annotation of microbial proteomes in Swiss Prot. *Comput. Biol. Chem.*, **27**, 49–58.

Gene Ontology Consortium. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.

Glasner,J.D. *et al*. (2006) ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. *Nucleic Acids Res.*, **34**, D41–D45.

Green,M.L. and Karp,P. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, **5**, 76.

Herlemann,D.P.R. *et al*. (2009) Genomic analysis of 'Elusimicrobium minutum,' the first cultivated representative of the phylum 'Elusimicrobia' (formerly termite group 1). *Appl. Environ. Microbiol.*, **75**, 2841–2849.

Kanehisa,M. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Liolios,K. et al. (2008) The genome on line database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **36**, D475–D479.

Markowitz,V.M. *et al*. (2008a) The integrated microbial genomes (IMG) system. *Nucleic Acids Res.*, **36**, D528–D533.

Markowitz,V.M. *et al*. (2008b) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res*., **36**, D534–D538.

Martin,H.G. *et al*. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol*., **24**, 1263–1269.

Mavromatis,K. *et al*. (2009) Genome analysis of the anaerobic thermohalophilic bacterium *Halothermothrix orenii*. *PLoS ONE*, **4**, e4192.

Overbeek,R. *et al*. (2005) The subsystems aproach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acid Res.*, **33**, 5691–5702.

Pruitt,K.D. *et al*. (2007) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acid Res.*, **35**, D61–D65.

Rutherford,K. *et al*. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

Salzberg,S.L. (2007) Genome re-annotation: a wiki solution? *Genome Biol.*, **8**,102.

Selengut,J.D. *et al*. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes *Nucleic Acids Res.*, **35**, D260–D264.

Tatusov,R.L. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

Vallenet,D. *et al*. (2006) MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res.*, **34**, 53–65.

Winsor,G.L. *et al*. (2009) Pseudomonas genome database: facilitating user-freindly, comprehensive comparisons of microbial genomes. *Nucleic Acids Res*., **37**, D483–D488.