

img/er & img/m er  **EXPERT REVIEW DATA SUBMISSION
MICROBIAL GENOMES & METAGENOMES**

Expert Review Submission Home	Genome Submissions	Metagenome Submissions	User Guide
-------------------------------	--------------------	------------------------	------------

IMG Databases
You can check your loaded genomes in IMG by clicking the following links:

- [IMG ER](#)
- [IMG/M ER](#)

IMG/ER & IMG/M ER Data Submission Guide



Towards Microbial Genome & Metagenome Data Expert Review with the IMG Systems

Genome Biology Program
Department of Energy Joint Genome Institute
Biological Data Management and Technology Center
Lawrence Berkeley National Laboratory

July 21, 2008

Copyright 2008 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.

July 21, 2008

©2008 The Regents of the University of California

This document was prepared by:

I-Min A. Chen*
Victor M. Markowitz*
Konstantinos Mavromatis**
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. Getting Started	3
3. Preparing Dataset Files for Submission	4
4. IMG ER Dataset Submission	6
4.1 Associating a Genome Dataset Submission with a Project	6
4.2 New Genome Dataset Submission	8
4.3 Genome Dataset Submission Status	9
5. IMG/M ER Dataset Submission	10
5.1 Associating a Metagenome Dataset Submission with a Project	10
5.2 New Metagenome Dataset Submission	12
5.3 Metagenome Dataset Submission Status	13
5.4 New Genome Dataset Submission to IMG/M ER	14
References	16

1. Synopsis

The **IMG/ER & IMG/M ER Submission Site** provides support for submitting (a) **microbial genome datasets** for inclusion into IMG/ER and (b) **microbial metagenome datasets** for inclusion into IMG/M ER; when needed, reference genome datasets can be also submitted for inclusion into IMG/M ER. Datasets included into IMG/ER or IMG/M ER are available only to (i.e., can be viewed, analyzed, curated by) the scientists who have submitted these datasets. Datasets submitted for inclusion into IMG/ER or IMG/M ER need to be associated with a **project** specified in **IMG-GOLD** or **GOLD**. After inclusion into IMG/ER or IMG/M ER, these datasets are analyzed in the context of all other public genomes and metagenomes available in IMG/ER and IMG/M ER, respectively.

- **IMG/ER & IMG/M ER Submission Site** is at: <http://img.jgi.doe.gov/submit>
- **IMG/ER** is at: <http://img.jgi.doe.gov/er>
- **IMG/M ER** is at: <http://img.jgi.doe.gov/mer>
- **IMG-GOLD** is at: <http://img.jgi.doe.gov/gold>
- **GOLD** is at: <http://www.genomesonline.org/>

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 brief introduction to IMG/M is provided in (Markowitz, Ivanova & al 2008).
- A user guide for IMG is available at: http://img.jgi.doe.gov/pub/doc/using_index.html.
- A user guide for IMG/M is available at: http://img.jgi.doe.gov/m/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>.

The purpose of this document is to serve as a use guide for submitting microbial genome and metagenome datasets to IMG/ER and IMG/M ER, respectively.

1.1 IMG ER & IMG/M ER Content

The IMG ER genome content baseline consists of all the public isolate genomes in IMG 2.5 (as of July 1st, 2008). The IMG/M ER genome and metagenome content baseline consists of all the public isolate genomes and metagenomes in IMG/M 2.4 (as of July 1st, 2008). The baseline for IMG ER and IMG/M ER is updated 2-3 times per year.

Genome and metagenome datasets submitted for inclusion into IMG/ER and IMG/M ER may contain gene models and/or product names generated using a genome or metagenome annotation pipeline. Alternatively, these datasets may contain only DNA sequences, whereby gene prediction and product name assignment are carried out using the IMG annotation pipeline (Ivanova et al. 2008)

1.2 Submission Summary

In order to submit a dataset to IMG ER or IMG/M ER, you need to:

1. Get account (section 2) via **Request Account** at: <http://img.jgi.doe.gov/request>.
Only one account is needed to access the IMG ER & IMG/M ER submission site, IMG ER and IMG/M ER.
2. Prepare files for submission (section 3).
3. Select submission type (IMG ER or IMG/M ER)
 - a. For IMG ER (section 4):
 - i. Find or define in IMG-GOLD the genome project to be associated with the submitted dataset (section 4.1);
 - ii. Submit the genome dataset and provide data processing requirements, including gene prediction, EC prediction, product name assignment (section 4.2);
 - iii. Check the status of your submission (section 4.3).
 - b. For IMG/M ER (section 5):
 - i. Find or define in IMG-GOLD the metagenome project and sample to be associated with the submitted dataset (section 5.1);
 - ii. Submit the metagenome dataset and provide data processing requirements, including gene prediction, EC prediction, product name assignment (section 5.2);
 - iii. Check the status of your submission (section 5.3).

2. Getting Started

In order to be able to submit dataset to IMG/ER and IMG/M ER and then gain access to these systems, users need to have an IMG account which can be requested by filling the **Request Account** form at: <http://img.jgi.doe.gov/request>, as illustrated in Figure 1.

The screenshot shows a web form titled "Request Account" for "img/er & img/m er". The header includes the text "EXPERT REVIEW DATA SUBMISSION MICROBIAL GENOMES & METAGENOMES". Below the header are navigation tabs: "Expert Review Submission Home", "Genome Submissions", "Metagenome Submissions", and "User Guide". The form contains the following fields, all marked with an asterisk (*):

- *Your Name
- *Title
- *Department
- *Your Email
- *Phone Number
- *Organization
- *Address
- *City
- *State
- *Country
- *Preferred Login Name
- *Group (if known)
- *Reason(s) for Request (large text area)

A "Submit" button is located at the bottom left of the form.

Figure 1. Request Account for IMG ER & IMG ER.

All fields marked with “*” are mandatory. The reasons for requesting an account need to be explained in detail, including the type of datasets to be submitted and the nature of the planned analysis or studies.

3. Preparing Dataset Files for Submission

Data files submitted to IMG ER and IMG/M ER need to be **Genbank** or **FASTA** format. Note that data files in other formats, such as EMBL are not accepted for submission. Genbank files contain the nucleotide sequence of the genome or metagenome as well as the coordinates, translation and annotation of the predicted genes. FASTA files contain only the nucleotide sequence of the genome and are used with IMG ER or IMG/M ER gene prediction and functional annotation.

Check that your data file is in Genbank format by checking the first line which should start with the word "LOCUS". A thorough validation of the file can be carried out using a data viewer or editor such as Artemis (<http://www.sanger.ac.uk/software/artemis>)

If there are **multiple dataset components**, such as multiple contigs of a draft genome, or an organism with **multiple replicons**, then these components must be concatenated into a single file¹.

For submissions in FASTA format there are no other requirements. For submissions in Genbank format, please check the following in order to avoid errors at later stages and subsequent delays in the processing of a submission please make sure that the ORGANISM and the first SOURCE line for each contig contain the same information. For example:

```
SOURCE          my favorite sequence XXX
                ORGANISM    my favorite sequence XXX
                Bacteria; Proteobacteria; ...
```

There is a "source" section under the FEATURES line, where the "organism" and mol_type" should be included, as illustrated below:

```
FEATURES             Location/Qualifiers
    source            1..807
                     /organism="my favorite sequence XXX"
                     /mol_type="genomic DNA"
                     /isolation_source="dinner table"
                     /collection_date="19-Apr-2005"
```

Prior to inclusion into IMG ER or IMG/M ER, a dataset needs to contain genes (CDSs). Contigs that do not contain genes can be included in the submitted dataset as long as some of the other contigs contain genes. If only the DNA sequence is available for submission, **gene prediction** can be carried out by the **IMG ER data processing pipeline** prior to inclusion into IMG ER or IMG/M ER.

Note that each gene (CDS) has a **unique locus tag**, *product name* and *valid translation*, as illustrated below:

```
CDS    642..2006
        /locus_tag="locus_0001"
```

¹ In a Unix environment "> cat contig1.gb contig2.gb > allContigs.gb" would concatenate files "contig1.gb" with file "contig2.gb". The same concatenation in a Windows environment is accomplished by typing "more contig1.gb contig2.gb > allContigs.gb".

```
/product="chromosomal replication initiator protein DnaA"  
/translation="MVDYDVNKVWEDIKEVIKKELNPSPTDISYNTWVETLVPICFDE  
NDTFILKAFADFHRDIVINRYSLLILNALRQLYSPHLSLKVILPNEVEKYKKYIKQKQ  
EEKVEVTTTLLNPKYTFETFVVGNNRLAHAAALAVAETPPGEKTYNPLFIYGGVGLGK  
THLMHAIGHHVLKLYPDTKVMYVTSEIFTNELIAAIRDEKTDEFMKYRNVDVLLIDD  
IQFLGGKERTQEEFFHTFNTLYEANKKIILSSDRPPKEINTLEDRLRSRFEWGLITDI  
QPPDFETRIAILSKKCQLEGTPVPQHILEFIASKIETNIRELEGALNKILAYSKLMAP  
DKEITLELAEKALKEFIDTNTKKELTIEDIQAEVAGYFNIKLEDFKSSRRSRNVAFPR  
QIAMYLARELTNVS LPKIGEAFGGKDHTTVLHACEKIKELINKDTNIRNTVENLKKRL  
INRE"
```

If a locus tag or a product name is not provided the file will be loaded, however it will be difficult to proceed with data analysis in IMG-ER.

A valid translation of the gene is necessary to proceed: if one CDS does not have a valid translation, the submission validation tool will generate an error and the submission will be cancelled.

4. IMG ER Dataset Submission

After login into the IMG/ER & IMG/M ER Submission Site (img.jgi.doe.gov/submit), select **IMG ER Submission** from the main menu and then “**Submit Dataset to IMG ER**” on the IMG ER Submission page (Figure 2.i).

The screenshot shows the 'Expert Review Data Submission' page for 'MICROBIAL GENOMES & METAGENOMES'. The user is logged in as 'User gbp'. The page has several sections:

- Navigation:** Expert Review Submission Home, IMG ER Submissions (highlighted), IMG/M ER Submissions, User Guide.
- IMG ER Submissions (i):** Includes buttons for 'Check Status', 'Cancel Submission', and 'Filter Submission'.
- New Submission:** Includes a 'Submit Dataset to IMG ER' button.
- Search Project for Your Submission (ii):** A search form with fields for 'GOLD Stamp ID', 'Project Type' (Genome), 'Domain' (BACTERIAL), 'Project Display Name' (Haloth), 'Genus', and 'Species'. A 'Search Projects' button is at the bottom.
- Project Search Result (iii):** A table showing search results.

Selection	ER Project ID	Project Display Name	GOLD ID	Phylogeny	Add Date	Last Mod Date
<input type="radio"/>	11874	Halothiobacillus neapolitanus c2, ATCC 23641	Gi02105	PROTEOBACTERIA-GAMMA	27-NOV-07	
<input type="radio"/>	12201	Halothermothrix orenii H 168	Gi01049	FIRMICUTES	27-NOV-07	02-APR-08

Figure 2. Submission of a new genome dataset to IMG ER.

4.1 Associating a Genome Dataset Submission with a Project

Each dataset submission has to be associated with a project that has been specified in IMG-GOLD. The “**Search Project for Your Submission**” page allows finding a project using the GOLD Stamp ID or by searching on one or several fields as illustrated in Figure 2.ii. The search returns a list of projects that satisfy the search criteria, as illustrated in Figure 2.iii. If no project has been yet specified, follow the link to **IMG-GOLD** in order to specify a new project, as illustrated in Figure 3.

Project specification with IMG-GOLD is outside the scope of this document. Briefly, such a specification involves information on **Organism** (Figure 3.i), **Project** (Figure 3.ii), **Links** (Figure 3.iii) and various **Metadata** fields (Figure 3.iv). The metadata fields follow the Minimum Information about a Genome Sequence (MIGS) guideline² of the Genomic Standards Consortium (Filed & al 2008). An example of a project specified with IMG-GOLD is shown in Figure 6 below.

² The MIGS/MIMS checklist is at http://gensc.org/gc_wiki/index.php/MIGS/MIMS.

Figure 3. Specifying a new project in IMG-GOLD.

Project Search Result (i)

If you cannot find a project for your submission, go to [IMG-GOLD](#) to define a new project.

Select Project

Selection	ER Project ID	Project Display Name	GOLD ID	Phylogeny	Add Date	Last Mod Date
<input type="radio"/>	11874	Halothiobacillus neapolitanus c2, ATCC 23641	G02105	PROTEOBACTERIA-GAMMA	27-NOV-07	
<input checked="" type="radio"/>	12201	Halothermothrix orenii H 168	G01049	FIRMICUTES	27-NOV-07	02-APR-08

New Genome Dataset Submission (ii)

Submit isolate genome for inclusion into IMG ER system.

Submission

Target ER System (*) IMG ER

ER Submission Project ID 12201

Genbank Fasta File (*) Browse...

Or, file name in JGI file system (**)

JGI Project ID

EC computation by PRIAM needed? (*) Yes

Gene calling needed? (*) GeneMark

Product name computation needed? (*) Yes

Is public in IMG database? (*) No

Unique species code or locus tag prefix (' required if gene calling is needed)

Topology

Comments

Figure 4. Specifying data processing requirements for a genome dataset submission.

Note that archaeal, bacterial and eukaryal genomic projects specified with IMG-GOLD are associated with a GOLD stamp and are listed in the GOLD project catalogue. Projects for viruses and plasmids that are not part of another genomic project can be specified with IMG-GOLD but are not associated with a GOLD stamp nor listed in the GOLD catalogue.

From the list of projects returned by the project search, such as the search illustrated in Figure 2.ii, the project associated with the submitted dataset can be selected as illustrated in Figure 4.i.

4.2 New Genome Dataset Submission

Once a project is selected, the “**New Genome Dataset Submission**” allows specifying the location of the dataset files as well as the type of data processing required for including the dataset into IMG ER, including **gene calling** using GeneMark (for submission in FASTA format), **EC#** derivation using PRIAM, and **product name** assignment (Ivanova et al. 2008), as illustrated in Figure 4.ii.

The screenshot shows the 'img/er & img/m er' interface for 'EXPERT REVIEW DATA SUBMISSION MICROBIAL GENOMES & METAGENOMES'. The main area displays 'IMG ER Submissions (i)' with a table of submissions. Submission 38 is highlighted, and a 'Filter Submission' button is visible. A detailed view for 'Submission 38 (ii)' is shown below, listing submission details. To the right, a 'Set Submission Filter for gbp (iii)' panel shows a status list with checkboxes for various submission stages.

Submission	ER Submission ID	Target ER System	ER Submission Project ID	Genbank-Fasta File	Submission Status	EC computation by PRIAM needed?	Gene calling needed?	Product name computation needed?	Is public in IMG database?	IMG Taxon OID	Database Info	Submitter	IMG Contact Name(s)	Submission Date	Last Modify Date
38	12201	Halothermothrix orenii H 168	Gi01049	2500395326	kostas	28-JAN-08	Finished with product name assignment.					kostas (KMavrommatis@lbl.gov)	gbp (nckyrpides@lbl.gov)	28-JAN-08	31-MAY-08

Figure 5. List of genome dataset submissions and associated information.

Datasets submitted in Genbank format are checked for **Genbank** format consistency. An email with a report of this validation, as illustrated below, is subsequently sent to the email associated with the submitter.

Submission report for genome X, submission ID: 11_58_2471_16_975. Submitted by: Jane Doe (jane). User's e-mail: jdoe@somewhere.com.

File has been processed.
 0 errors reported.
 0 warnings issued.

This is an automated message. Please do not reply.

4.3 Dataset Submission Status

If a user has already submitted genome datasets to IMG/ER, then after login into the IMG/ER & IMG/M ER data submission site and selecting **IMG ER Submission** from the main menu, the list of these submissions is displayed as illustrated in Figure 5.1.

The status of a submission can be checked either by clicking on the “**Check Status**” button or directly on the Submission ID, as shown in Figure 5.i. A submission summary is then displayed, as illustrated in Figure 5.ii. The status of the submission is displayed in the “**Submission Status**” field and can have one the 11 values shown in Figure 5.iii. If the list of submission is long, the list of displayed submissions can be reduced by applying a filter on the submission status and/or submission date, as illustrated in Figure 5.iii.

The information on the project associated with a specific submission can be reviewed by clicking on **ER Submission Project ID**. For example, by clicking on “12201” project identifier shown in Figure 5.i, will result in displaying information on Project 12201, as shown in Figure 6, including information specific to the organism (Figure 6.i), project (figure 6.ii), associated links (Figure 6.iii) and metadata (Figure 6.iv).

Project 12201 (Halothermothrix orenii H 168)	
Organism (i)	
Display Name	Halothermothrix orenii H 168
NCBI Taxon ID	373903
NCBI Kingdom	Bacteria
NCBI Phylum	Firmicutes
NCBI Class	Clostridia
NCBI Order	Halanaerobiales
NCBI Family	Halanaerobiaceae
NCBI Genus	Halothermothrix
NCBI Species	Halothermothrix orenii
NCBI Project ID	16377
Domain	BACTERIAL
Phylogeny	FIRMICUTES
Genus	Halothermothrix
Species	orenii
Strain	H 168
Culture Collection	DSM 9562
Isolation	Salted lake sediment
Project (ii)	
ER Submission Project OID	12201
GOLD Stamp ID	Gi01049
Project Web Page	1
GOLD Web Page Code	1
Project Type	Genome
Project Status	complete
Availability	Proprietary
Contact Name	Mavrommatis Kostas
Contact Email	KMavrommatis@lbl.gov
GC Percent	39.6
Sequencing Status	Complete
Sequencing Depth	12.5x
Gene Calling Method	GeneMark
Estimated Size	2463
Units	Kb
Gene Count	2273
Sequencing Country	USA
IMG Contact	kostas (KMavrommatis@lbl.gov)
Add Date	27-NOV-07
Last Modify Date	02-APR-08
Last Modified By	nikos (NCKyrpides@lbl.gov)
Metadata (iv)	
Oxygen Requirement	Anaerobe
Temperature Range	Thermophile
Salinity	Halophile
Diseases	None
Habitat	Aquatic
Project Relevance	Biotechnological
Sequencing Methods	454, Sanger
ER Submission Info	
Submissions	38
Links (iii)	
IMG Object ID	638341104
GCAT ID	001788_GCAT
Data Links (URLs)	Data, Refseq, NZ_AAOZ00000000, URL Database, NCBI, 5456, URL Funding, DOE, , URL Information, Entrez, 16377, URL Information, Isolation, 7520742, URL Information, Taxonomy, 373903, URL Seq Center, Joint Genome Institute, , URL

Figure 6. Project information associated with a genome data submission.

5. IMG/M ER Dataset Submission

After login into the IMG/ER & IMG/M ER Submission Site (img.jgi.doe.gov/submit), select **IMG M ER Submission** from the main menu and then “**Submit Metagenome Dataset to IMG/M ER**” on the IMG ER Submission page (Figure 7.i). Note that in certain cases isolate genome datasets can be also submitted for inclusion into IMG/M ER. This is discussed later below.

The screenshot shows the 'IMG/M ER Metagenome Submissions' page. A red box highlights the 'Submit Metagenome Dataset to IMG/M ER' button. A blue box highlights the search form with the following details:

- Target Database: IMG/M ER
- Go to [IMG-GOLD](#) if you need to define new projects or samples.
- Text searches are based on case-insensitive substring match.
- GOLD Stamp ID:
- Project Type: Metagenome
- Domain: MICROBIAL
- Project Display Name:
- Search Projects button

A blue box highlights the 'Project Search Result' section, which includes a 'Show Samples' button and a table of results:

Selection	ER Project ID	Project Display Name	GOLD ID	Project Category	Add Date	Last Mod Date
<input type="radio"/>	10736	Extreme microbial communities from Yellowstone National Park	Gm00095	ENVIRONMENTAL-EXTREME	27-NOV-07	
<input type="radio"/>	10739	Aquatic microbial communities from Yellowstone Bison Hot Spring Pool	Gm00098	ENVIRONMENTAL-AQUATIC	27-NOV-07	11-FEB-08

Figure 7. Submission of a new metagenome dataset to IMG/M ER.

5.1 Associating a Metagenome Dataset Submission with a Project and Sample

Each metagenome dataset submission has to be associated with a project and a sample within that project, both previously specified in IMG-GOLD. The “**Search Metagenome Projects for Your Submission**” page allows finding a metagenome project using the “GOLD Stamp ID” or “Project Display Name” as illustrated in Figure 7.ii. Note that “Project Type” and “Domain” are already preset to “Metagenome” and “MICROBIAL” respectively, following GOLD conventions for specifying metagenome projects. The search returns a list of metagenome projects that satisfy the search criteria, as illustrated in Figure 7.iii.

If no project has been specified yet, follow the link to IMG-GOLD in order to specify a new metagenome project as illustrated in Figure 8. Metagenome project specification with IMG-GOLD is outside the scope of this document. Briefly, such a specification involves information on **Metagenome**, **Project**, **Links**, and various project specific

Metadata fields, as illustrated in Figure 8. The metadata fields follow the Minimum Information about a Metagenome Sequence (MIMS) guideline of the Genomic Standards Consortium (Filed & al 2008). An example of a project specified with IMG-GOLD is shown in Figure 12.i below.

Figure 8. Specifying a new metagenome project in IMG-GOLD.

From the list of projects returned by the project search, such as the search illustrated in Figure 7.iii, the project for the submitted dataset can be selected using “**Select Sample**”, (see Figure 7.iii), which will return its associated list of samples displayed on the **Select Sample to Submit** page as illustrated in Figure 10.ii.

If for a specific project the appropriate sample has not been yet specified, then on the **Select Sample to Submit** page (see Figure 10.ii) either:

1. follow the link to **IMG-GOLD** in order to specify a new sample, as illustrated in Figure 9, or
2. Create a default sample using the “**Create Default Sample**” button which will add a new sample to the list of samples on this page.

Sample specification with IMG-GOLD is outside the scope of this document. Briefly, such a specification involves information on **Sample Collection** (Figure 9.i), **Metagenome** (Figure 9.ii), and sample specific **Metadata** fields (Figure 9.iii). Similar to the project specific metadata fields, the sample specific metadata fields follow the Minimum Information about a Genome Sequence (MIMS) guideline of the Genomic Standards Consortium. An example of a sample specified with IMG-GOLD is shown in Figure 12.ii below. Note that a sample created using “**Create Default Sample**” can be further edited or revised using IMG-GOLD.

The screenshot shows the 'New Sample' form in IMG-GOLD. The form is for a metagenome project and includes the following sections and fields:

- Navigation:** 'Sample Collection (*)', 'Metagenome', 'Metadata' tabs.
- Project Information:** 'Project 10739: (Aquatic microbial communities from Yellowstone Bison Hot Spring Pool)'. Sub-tabs: 'Sample Collection (*)', 'Metagenome', 'Metadata'.
- Sample Collection (*) Section:**
 - Required fields are marked with (*).
 - Fields: Sample Display Name (*), Sample Site, Date Collected, Geo Location, Latitude, Longitude, Altitude, Sampling Strategy, Sample Isolation, Sample Volume, Biomass, Diversity, Temperature Range, Temperature, Salinity, Pressure, pH, Sample Link.
- Habitat Section:**
 - Text: 'Please edit the following list of values:'
 - Field: Habitat Type (dropdown menu).
- Energy Sources Section:**
 - Text: 'Please edit the following list of values:'
 - Field: Energy Source (dropdown menu).
- Misc Metadata Section:**
 - Text: 'Please edit the following list of values:'
 - Field: Meta Tag (input field).
- Sequencing Methods Section:**
 - Fields: Host NCBI Taxon ID, Host Name, Host Gender, Host Age, Host Health Condition, Library Method, Binning Method, Sequencing Depth, Gene Calling Method, GC Percent, Chromosome Count, Plasmid Count, Estimated Size, Units, Contig Count, Singlet Count, Gene Count, Comments.

Figure 9. Specifying a new metagenome sample in IMG-GOLD.

5.2 New Metagenome Dataset Submission

Once a project is selected from the list of projects returned by a metagenome project search, as illustrated in Figure 10.i, the list of samples associated with the selected project are displayed as illustrated in Figure 10.ii. A sample from this list is then selected for the metagenome dataset submission. Note that the **“New Metagenome Dataset Submission”** form for a sample (Figure 10.iii) is similar to the form for a project (Figure 4.ii) except the second field which displays the *Submission Sample ID* rather than the *Submission Project ID*.

Once a sample is selected, the **“New Metagenome Dataset Submission”** form allows specifying the location of the dataset files as well as the type of data processing required for including the dataset into IMG ER, including **gene calling**, **EC#** derivation using PRIAM, and **product name** assignment (Ivanova et al. 2008), as illustrated in Figure 10.iii.

Note that there are two gene calling options available for metagenomes that are submitted in FASTA format. GeneMark can be employed for metagenome sequences generated using Sanger sequencing, similar to isolate genomes. For short metagenome reads generated using 454 pyrosequencing, gene identification can be carried out using the proxygene method described in (Dalevi et al. 2008). Note also that for metagenome datasets additional data can be included into IMG/M ER, such as contigs, singlets, bins, and READS (see lower part of Figure 10.iii).

Project Search Result (i)

If you cannot find a project for your submission, go to [IMG-GOLD](#) to define a new project.

[Show Samples](#)

Selection	ER Project ID	Project Display Name	GOLD ID
<input type="radio"/>	10736	Extreme microbial communities from Yellowstone National Park	Gm000
<input checked="" type="radio"/>	10739	Aquatic microbial communities from Yellowstone Bison Hot Spring Pool	Gm000
<input type="radio"/>	10742	Aquatic microbial communities from Yellowstone Bath Hot Springs	Gm001

Select Sample to Submit (ii)

Project 10739: Aquatic microbial communities from Yellowstone

If you cannot find a sample for your submission, either create a default sample or go to [IMG-GOLD](#) to define a new sample.

Selection	ER Sample ID	Sample Display Name	Sample Site
<input checked="" type="radio"/>	10022	2_050719S	Bison Pool, 6 m downstream of boiling source
<input type="radio"/>	10023	5_050719P	Bison Pool, downstream photosynthetic mats
<input type="radio"/>	10024	4_050719Q	Bison pool, downstream edge of photosynthetic mats
<input type="radio"/>	10025	1_050719N	Bison Pool boiling source pool
<input type="radio"/>	10026	3_050719R	Bison Pool, 11 m downstream of boiling source

(Number of rows displayed: 5)

New Metagenome Dataset Submission (iii)

Submit isolate genome or metagenome for inclusion into IMG/M ER system.

Submission	
Target ER System (*)	IMG/M ER
ER Submission Sample ID	10022
Genbank/Fasta File (*)	
Or, file name in JGI file system (**)	
JGI Project ID	
EC computation by PRIAM needed? (*)	Yes
Gene calling needed? (*)	No
Product name computation needed? (*)	GeneMark Prokaryotes
Is public in IMG database? (*)	No Unknown
Unique species code or locus tag prefix (* required if gene calling is needed)	
Topology	
Comments	
Additional Metagenome Files	
Contigs File	
Singlets File	
Binning File	
READS Files	

Figure 10. Specifying data processing requirements for a metagenome dataset associated with a sample of a specific project.

Similar to genome dataset submissions, a metagenome dataset submitted to IMG/M ER is checked for **Genbank** format **consistency** and an email with a report of this validation is subsequently sent to the email associated with the submitter, as discussed above for genome dataset submission.

5.3 Metagenome Dataset Submission Status

If a user has already submitted a metagenome datasets to IMG/M ER, then after login into the IMG/ER & IMG/M ER data submission site and selecting **IMG M ERSubmission** from the main menu, the list of these submissions is displayed as illustrated in Figure 11.

The status of a submission can be checked either by clicking on the **“Check Status”** button or directly on the Submission ID, as shown in Figure 11.i. A submission summary is then displayed, as illustrated in Figure 11.ii. The status of the submission is displayed in the **“Submission Status”** field and can have one of the 11 values shown in Figure 5.iii. If the list of submission is long, the list of displayed submissions can be reduced by applying a filter on the submission status and/or submission date, as illustrated in Figure 5.iii.

Expert Review Submission Home | IMG ER Submissions | **MG/M ER Submissions** | User Guide

IMG/M ER Metagenome Submissions (i)

Check Status | Cancel Submission | Filter Submission

Click the column name to have the data order by the selected column. Click (Rev)

Selection	Submission ID (Rev)	ER Submission Project ID (Rev)	Project Name (Rev)
<input type="radio"/>	7	9	US sludge - combined Sanger 454 assembly
<input type="radio"/>	8	11288	Candidatus Endomicrobium trichonymphae
<input type="radio"/>	9	10739	Aquatic microbial communities from Yellowstone Bison Hot Spring Pool
<input checked="" type="radio"/>	33	10739 (Sample: 10022)	Aquatic microbial communities from Yellowstone Bison Hot Spring Pool
<input type="radio"/>	34	10739 (Sample: 10026)	Aquatic microbial communities from Yellowstone Bison Hot Spring Pool
<input type="radio"/>	53	48	TM7e
<input type="radio"/>	54	10739 (Sample: 10026)	Aquatic microbial communities from Yellowstone Bison Hot Spring Pool

Submission 33 (ii)

Submission	
ER Submission ID	33
Target ER System	IMG/M ER
ER Submission Project ID	10739
ER Submission Sample ID	10022
Genbank/Fasta File	/home/img2/imachen/submission/33.gb
Submission Status	10 - Finished with product name assignment.
JGI Project ID	4003086
EC computation by PRIAM needed?	No
Gene calling needed?	GeneMark
Product name computation needed?	Yes
Is public in IMG database?	No
Unique species code or locus_tag prefix	BISONS
Comments	Original file: /home/img2/imachen/submission/33.fna IMG Taxon OID: 2009439000
Database Info	
Submitter	soil (sgtringe@lbl.gov)
IMG Contact Name(s)	bison (eshock@asu.edu), soil (sgtringe@lbl.gov)
Submission Date	14-JAN-08
Last Modify Date	11-FEB-08
Last Modified By	AMY (IMACHen@lbl.gov)
Additional Metagenome Files	
Contigs File	/psf/project/assorted1/microbe/community/4003086/edit_dir.12Nov/4003086.lucy.pga.assem.contigs
Singlets File	/psf/project/assorted1/microbe/community/4003086/edit_dir.12Nov/4003086.lucy.pga.assem.singletons.noN
READS Files	/psf/project/assorted1/microbe/community/4003086/edit_dir.12Nov
Samples	10022 - 2_050719S

Figure 11. List of metagenome dataset submissions and associated information.

The information on the metagenome project associated with a specific submission can be reviewed by clicking on **ER Submission Project ID**. For example, by clicking on “10739” project identifier shown in Figure 11.i, will result in displaying information on Project 10739, as shown in Figure 12.i, including information specific to the project, associated links and metadata.

The information on the metagenome sample associated with a specific submission can be reviewed by clicking the **Sample ID** part of the **ER Submission Project ID**. For example, by clicking on “10022” sample identifier shown in Figure 11.i, will result in displaying information on Sample 10022, as shown in Figure 12.ii, including information specific to the sample, metagenome, and metadata.

5.4 New Genome Dataset Submission to IMG/M ER

If the baseline of public isolate genomes in IMG/M ER does not contain reference genomes of interest for analyzing specific metagenome datasets³, individual genome datasets can be submitted for inclusion into IMG/M ER by following the “**Submit Genome Dataset to IMG/M ER**” link provided on the IMG/M ER submission page (see

³ The isolate genome baseline of IMG/M ER is updated once or twice each year, therefore it may not contain isolate genomes available in Genbank, RefSeq or even the public version of IMG which is updated more frequently.

bottom part of Figure 7.i). The submission of such datasets to IMG/M ER is identical to the submission of genome datasets for IMG ER.

Metagenome Project 10739 (Aquatic microbial communities from Yellowstone Bison Hot Spring Pool) (i)		Sample 10022 (2_050719S) (ii)	
Metagenome		Sample	
Display Name	Aquatic microbial communities from Yellowstone Bison Hot Spring Pool	Sample ID	10022
Domain	MICROBIAL	Sample Display Name	2_050719S
Category (Phylogeny)	ENVIRONMENTAL-AQUATIC	Sample Site	Bison Pool, 6 m downstream of boiling source pool
Common Name	Bison Pool thermal gradient	Latitude	44.5696298
Habitat (Genus)	Aquatic	Longitude	-110.8651817
Community (Species)	microbial communities	Temperature Range	Thermophile
Location (Strain)	from Yellowstone Bison Hot Spring Pool	Temperature	85 C
Identifier (Serovar)	thermal gradient	Contact Info	gbp (nckyrpides@lbl.gov)
Isolation	Lower Geyser Basin in Yellowstone National Park	Add Date	04-DEC-07
Isolation Country	USA	Last Modify Date	08-FEB-08
Geographic Location	Bison Pool / Rosette Geyser, Yellowstone National Park	Last Modified By	soil (sgringe@lbl.gov)
Location Coordinates	Lat: 44.5696298, Lon: -110.8651817	Metagenome	
Project		Library Method	short-insert
ER Submission Project OID	10739	Contig Count	12033
GOLD Stamp ID	Gm00098	Singlet Count	14263
Project Web Page	4	Habitat	Aquatic
GOLD Web Page Code	4	Links	
Project Type	Metagenome	GCAT ID	002936_GCAT
Project Status	incomplete	Data Links (URLs)	Funding, DOE, , URL Seq Center, Joint Genome Institute, , URL
Availability	Public	Metadata	
Contact Name	Shock E	Temperature Range	Thermophile
Contact Email	eshock@asu.edu	Diseases	None
Sequencing Status	Incomplete	Habitat	Aquatic
Library Method	short-insert	Sequencing Methods	Dideoxysequencing
Assembly Method	lucy / pga	Samples	10022 - 2_050719S 10023 - 5_050719P 10024 - 4_050719Q 10025 - 1_050719N 10026 - 3_050719R

Figure 12. Project information associated with a metagenome data submission.

References

- Dalevi D., Ivanova, N.N., Mavromatis, K., Hooper, S., Szeto, E., Hugenholtz, P., Kyrpides, N.C., and Markowitz, V.M. (2008) Annotation of metagenome short reads using proxygenes. Accepted for publication. *Bioinformatics*, **24** (18).
- Field, D., Garrity, G.M., Gray, T., Morrison, N., Selengut, J.D., Sterk, P., et al. (2008). The Minimum Information about a GenomeSequence (MIGS) specification. *Nat Biotechnol*.
- Ivanova, N.N., Mavromatis, K., Chen, I.A., Markowitz, V.M., and Kyrpides, N.C. (2008) Standard operating procedure for the annotations of genomes and metagenomes submitted to the Integrated Microbial Genomes Expert Review Systems. http://img.jgi.doe.gov/w/doc/img_er_ann.pdf
- Markowitz, V.M., Szeto E., Palaniappan K., Grechkin Y., Chu K., Chen I-M, Dubchak I., Anderson I., Lykidis A., Mavromatis K., Ivanova N.N., Kyrpides N.C. (2008) The Integrated Microbial Genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Research*, 36.
- Markowitz, V.M., Ivanova, N., Szeto, E., Palaniappan K., Grechkin Y., Chu K., Chen I-M, Anderson I., Lykidis A., Mavromatis K., Ivanova N.N., Hugenholtz, P., Kyrpides N.C. (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Research*, 36.