Title

IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes.

Permalink

https://escholarship.org/uc/item/7hc8b5xx

Journal

Nucleic acids research, 47(D1)

ISSN

0305-1048

Authors

Chen, I-Min A
Chu, Ken
Palaniappan, Krishna
et al.

Publication Date

2019

DOI

10.1093/nar/gky901

Peer reviewed

D666–D677 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 5 October 2018*
doi: 10.1093/nar/gky901

IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes

I-Min A. Chen[1,*], Ken Chu[1], Krishna Palaniappan[1], Manoj Pillay[1], Anna Ratner[1], Jinghua Huang[1], Marcel Huntemann[1], Neha Varghese[1], James R. White[2], Rekha Seshadri[1], Tatyana Smirnova[1], Edward Kirton[1], Sean P. Jungbluth[1], Tanja Woyke[1], Emiley A. Eloe-Fadrosh[1], Natalia N. Ivanova[1,*] and Nikos C. Kyrpides[1,*]

[1]Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA and [2]Resphera Biosciences, Baltimore, MD, USA

Received September 05, 2018; Revised September 17, 2018; Editorial Decision September 23, 2018; Accepted September 24, 2018

ABSTRACT

The Integrated Microbial Genomes & Microbiomes system v.5.0 (IMG/M: https://img.jgi.doe.gov/m/) contains annotated datasets categorized into: archaea, bacteria, eukarya, plasmids, viruses, genome fragments, metagenomes, cell enrichments, single particle sorts, and metatranscriptomes. Source datasets include those generated by the DOE's Joint Genome Institute (JGI), submitted by external scientists, or collected from public sequence data archives such as NCBI. All submissions are typically processed through the IMG annotation pipeline and then loaded into the IMG data warehouse. IMG's web user interface provides a variety of analytical and visualization tools for comparative analysis of isolate genomes and metagenomes in IMG. IMG/M allows open access to all public genomes in the IMG data warehouse, while its expert review (ER) system (IMG/MER: https://img.jgi.doe.gov/mer/) allows registered users to access their private genomes and to store their private datasets in workspace for sharing and for further analysis. IMG/M data content has grown by 60% since the last report published in the 2017 NAR Database Issue. IMG/M v.5.0 has a new and more powerful genome search feature, new statistical tools, and supports metagenome binning.

INTRODUCTION

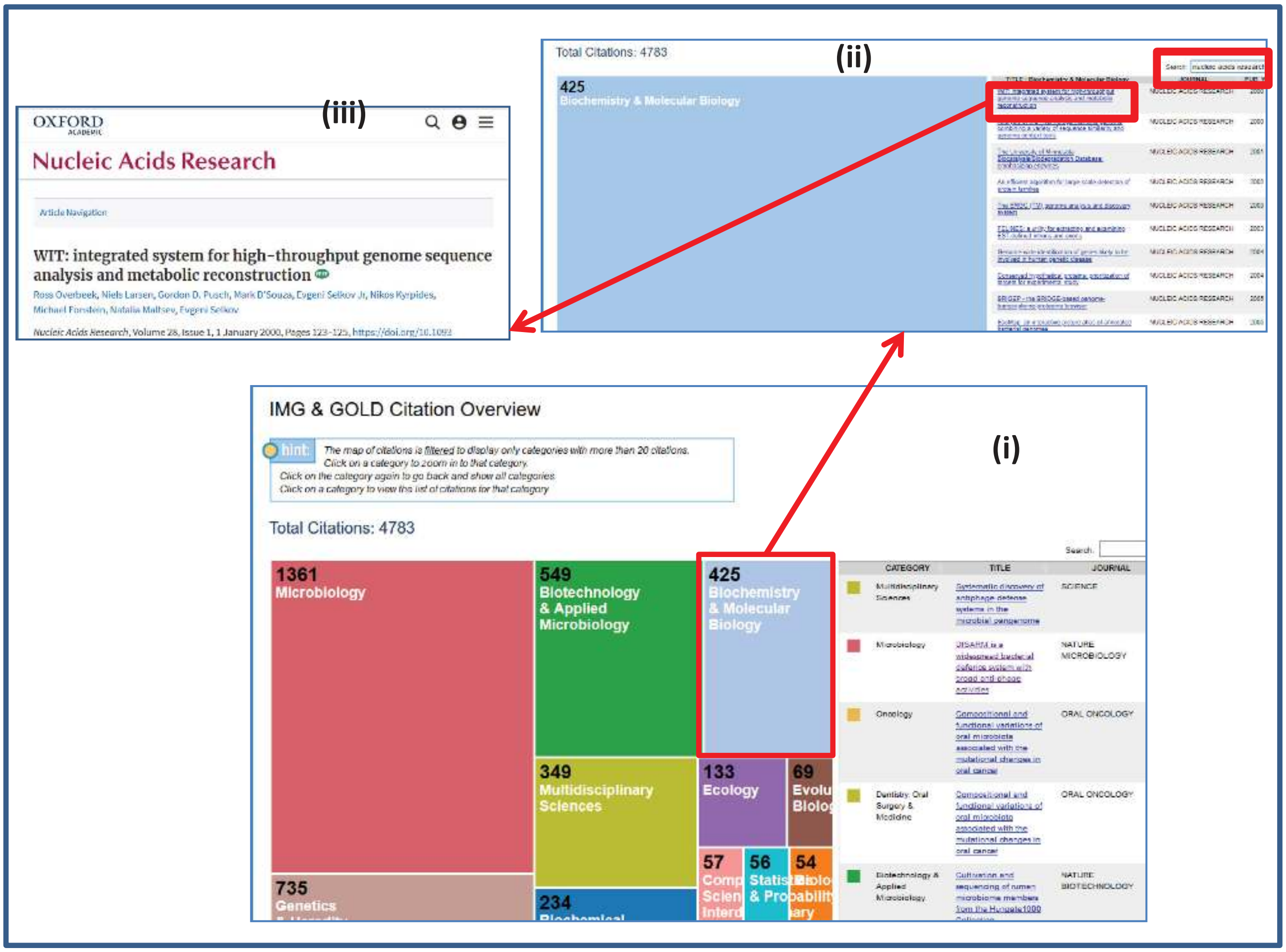
The Integrated Microbial Genomes & Microbiomes data management system v.5.0 (IMG/M: https://img.jgi.doe.gov/m/) includes archaea, bacteria, eukarya, plasmids, viruses, genome fragments (genomic regions of interest generated by targeted sequencing), as well as genomes of uncultured organisms represented by single-cell amplified genomes (SAGs) and metagenome-assembled genomes (MAGs), metagenomes and metatranscriptome datasets. In addition to the JGI-generated sequences, GenBank (1) serves as IMG's major source of genome sequence data for cultured and uncultured organisms. After associated metadata is incorporated into the Genomes OnLine Database (GOLD) (2), genome sequence data retrieved from GenBank are processed through the IMG submission system (https://img.jgi.doe.gov/submit/) and IMG annotation pipeline (3) before being integrated into the IMG data warehouse.

IMG continues to support external submissions of assembled genome and metagenome data generated by any sequencing technology. Each genome or metagenome submission must be associated with a sequencing and analysis project identifier, and its associated metadata in GOLD. This supports the availability of extensive metadata cataloging for data versioning and output of complex analysis pipelines, as well as integration of metadata standards as defined by the Genomics Standards Consortium (4).

IMG supports two types of isolate genome submissions: files in GenBank format with predicted features and unannotated sequences in FASTA format. Submissions of annotated genomes in GenBank format mainly support eukaryotic genome processing since the IMG annotation pipeline incorporates only prokaryotic gene finders, even though it is also possible for users to submit prokaryotic genomes in GenBank format to bypass the gene calling process done by IMG. *De novo* annotation of genomes submitted in FASTA format is performed by the JGI's Microbial Genome Anno-

[*]To whom correspondence should be addressed. Tel: +1 925 296 5697; Fax: +1 925 296 5850; Email: IMAChen@lbl.gov
Correspondence may also be addressed to Natalia N. Ivanova. Email: NNIvanova@lbl.gov
Correspondence may also be addressed to Nikos C. Kyrpides. Email: nckyrpides@lbl.gov

Published by Oxford University Press on behalf of Nucleic Acids Research 2018.



OXFORD ACADEMIC

Nucleic Acids Research

Article Navigation

WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction

Ross Overbeek, Niels Larsen, Gordon D. Pusch, Mark D'Souza, Evgeni Selkov Jr, Nikos Kyrpides, Michael Fonstein, Natalia Maltsev, Evgeni Selkov

Nucleic Acids Research, Volume 28, Issue 1, 1 January 2000, Pages 123–125, https://doi.org/10.1093

Search: nucleic acids research

TITLE : Biochemistry & Molecular Biology	JOURNAL	PUB. Y
WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction	NUCLEIC ACIDS RESEARCH	2000
[illegible] combining a variety of sequence similarity and genome context tools	NUCLEIC ACIDS RESEARCH	2000
The University of Minnesota Biocatalysis/Biodegradation Database [illegible]	NUCLEIC ACIDS RESEARCH	2001
An efficient algorithm for large-scale detection of protein families	NUCLEIC ACIDS RESEARCH	2003
The ERGO (TM) genome analysis and discovery system	NUCLEIC ACIDS RESEARCH	2003
[illegible]	NUCLEIC ACIDS RESEARCH	2003
[illegible] involved in human genetic disease	NUCLEIC ACIDS RESEARCH	2004
Conserved hypothetical proteins: prioritization of targets for experimental study	NUCLEIC ACIDS RESEARCH	2004
BRIGEP - the BRIDGE-based genome-[illegible] browser	NUCLEIC ACIDS RESEARCH	2005
[illegible]	NUCLEIC ACIDS RESEARCH	[illegible]

IMG & GOLD Citation Overview

hint: The map of citations is filtered to display only categories with more than 20 citations.
Click on a category to zoom in to that category.
Click on the category again to go back and show all categories
Click on a category to view the list of citations for that category

Total Citations: 4783

Search:

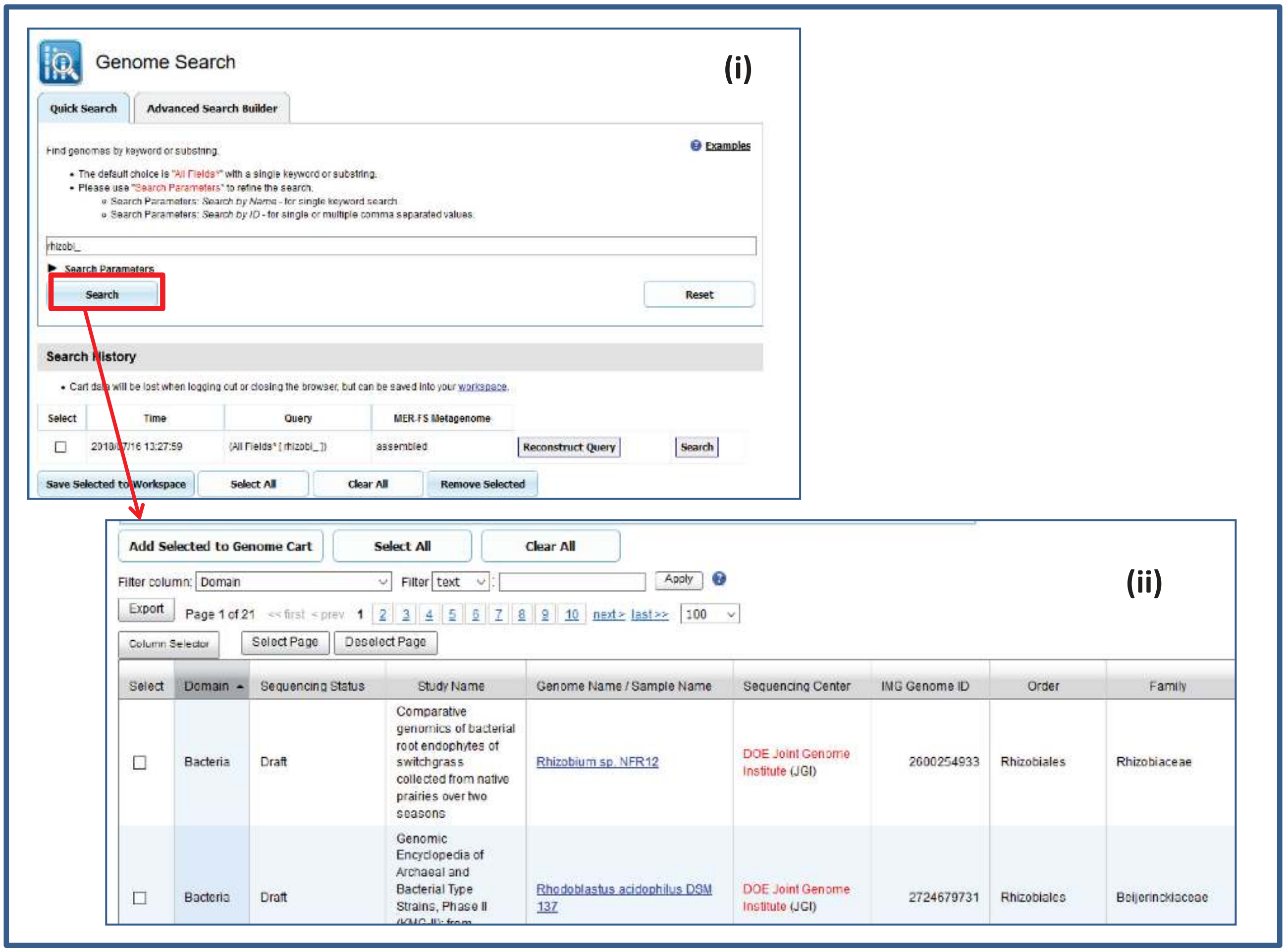
	CATEGORY	TITLE	JOURNAL
🟨	Multidisciplinary Sciences	Systematic discovery of antiphage defense systems in the microbial pangenome	SCIENCE
🟥	Microbiology	DISARM is a widespread bacterial defence system with broad anti-phage activities	NATURE MICROBIOLOGY
🟧	Oncology	Compositional and functional variations of oral microbiota associated with the mutational changes in oral cancer	ORAL ONCOLOGY
🟨	Dentistry, Oral Surgery & Medicine	Compositional and functional variations of oral microbiota associated with the mutational changes in oral cancer	ORAL ONCOLOGY
🟩	Biotechnology & Applied Microbiology	Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection	NATURE BIOTECHNOLOGY

Figure 1. IMG & GOLD Citation Overview. (**i**) IMG & GOLD Citation Overview page shows all papers citing IMG or GOLD. Publications are divided into different categories. (**ii**) Users can click on the category to see a set of papers in this category. The list can be searched or sorted. The example shows searching on 'nucleic acids research' (case insensitive) from the Search field at the right upper corner. (**iii**) Clicking on a title will lead to the actual publication.

tation Pipeline (MGAP v. 4) (3), which includes identification of protein-coding genes, non-coding RNAs, regulatory RNA features and binding motifs, as well as CRISPR elements. Briefly, CRISPR elements are detected using a modified CRT (5); tRNAs, ribosomal RNAs, non-coding RNAs and RNA regulatory features are predicted using Rfam covariance models and Infernal tools (6–8); and protein-coding genes are called by Prodigal v2.6.2 (9).

Next, protein-coding genes undergo prediction of signal peptides and transmembrane regions using SignalP (10) and TMHMM (11), respectively, followed by protein family assignments and functional annotation steps. These involve assignment to COG (12) using position-specific scoring matrices and RPS-BLAST, comparing to Pfam-A (13) and TIGRfam (14) Hidden Markov Models using HMMER 3.1 (15), assignment to a subset of InterPro families using a customized version of InterProScan5 (16) and associating proteins with KEGG Orthology (KO) terms (17) using LAST (18). Proteomes are associated with KEGG pathways based on KO term assignments, and are associated with MetaCyc pathways (19) based on gene annotations with Enzyme Commission numbers derived from KO terms. In addition, proteomes may be associated with IMG Pathways via IMG term assignment (20), leading to pathway assertions and phenotype predictions for sequenced organisms (21). In addition, Bidirectional Best Hits (BBH) between proteomes and best hits against IMG reference proteomes are computed using LAST (18). IMG reference proteomes are derived from high-quality public genomes and are used for placing the genomes in phylogenetic context through Phylogenetic Distribution of Best Hits tool. Isolate genomes also undergo Average Nucleotide Identity (ANI) (22) distance matrix computations, prediction of gene cassette regions (23), gene fusions (24), and biosynthetic clusters (25), as previously described.

Starting in 2016, IMG supports only assembled data for newly added metagenomes and metatranscriptomes, even though unassembled sequences are still available for legacy datasets. *De novo* feature prediction and functional annotation of metagenomes and metatranscriptomes are similar to the process for isolate genomes described above. Main differences include the use of hmmsearch (15) for assignment of COGs to metaproteomes, and omission of several computationally intensive steps, such as non-coding RNA pre-



Find genomes by keyword or substring.

- The default choice is "All Fields*" with a single keyword or substring.
- Please use "Search Parameters" to refine the search.
    - Search Parameters: *Search by Name* - for single keyword search.
    - Search Parameters: *Search by ID* - for single or multiple comma separated values.

rhizobi_

Search History

- Cart data will be lost when logging out or closing the browser, but can be saved into your workspace.

Select	Time	Query	MER-FS Metagenome
□	2018/07/16 13:27:59	(All Fields* [rhizobi_])	assembled

Filter column: Domain Filter text : 100

Select	Domain ▲	Sequencing Status	Study Name	Genome Name / Sample Name	Sequencing Center	IMG Genome ID	Order	Family
□	Bacteria	Draft	Comparative genomics of bacterial root endophytes of switchgrass collected from native prairies over two seasons	Rhizobium sp. NFR12	DOE Joint Genome Institute (JGI)	2600254933	Rhizobiales	Rhizobiaceae
□	Bacteria	Draft	Genomic Encyclopedia of Archaeal and Bacterial Type Strains, Phase II (KMG-II): from	Rhodoblastus acidophilus DSM 137	DOE Joint Genome Institute (JGI)	2724679731	Rhizobiales	Beijerinckiaceae

Figure 2. Quick Search option for the new Genome Search feature. (**i**) The Quick Search option allows users to type in a keyword to search all IMG genomes. (**ii**) The search results can be added to Genome Cart.

diction and IMG term assignment and pathway assertion. The detailed description of IMG metagenome and metatranscriptome processing can be found in (26).

The IMG data warehouse content has experienced substantial growth in the past two years. IMG UI also has several new analysis features. We will discuss these in the following section.

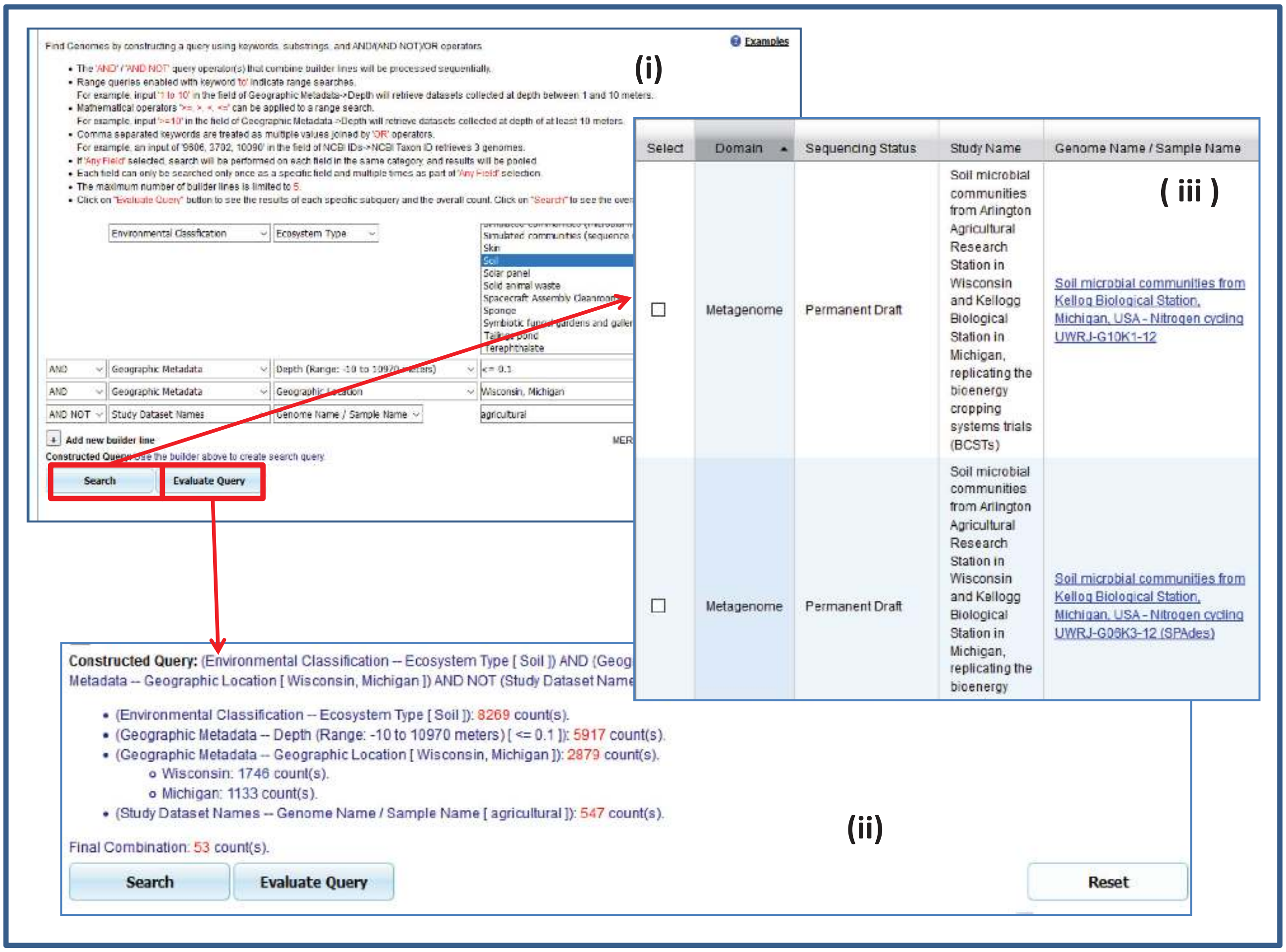
DISCUSSION

Data content

The current IMG v.5.0 (as of July 2018) contains 77 821 (64 446 public) archaeal, bacterial and eukaryotic genomes, 9674 (8388 public) viruses and 1215 (1190 public) plasmids. Overall, IMG includes circa 272 million genes from isolate genomes, SAGs and MAGs, which represents ∼60% data growth since July 2016 (27).

As compared to July 2016, IMG has about 80% more metagenome and metatranscriptome datasets. There are 28 799 microbiome datasets in total (as of July 2018) including 18 907 (13 232 public) metagenomes, 4605 (2423 public) metatranscriptome datasets, and two new types of reduced complexity ‘microbiomes’ - cell enrichments (1333 including 801 public), and single particle sorts (3954 including 3486 public), produced by physically separating microcolonies or isolating cell aggregates by flow cytometry, respectively. About two thirds (66%) of the IMG metagenome and metatranscriptome datasets are from environmental samples (e.g. aquatic, terrestrial, air); ∼22% are derived from host-associated samples and about 12% come from engineered environments. IMG now contains >54 billion metagenome genes.

As of July 2018, IMG had more than 16,000 external isolate genome submissions and 11 400 external metagenome submissions including 6500 genomes and 6200 metagenomes submitted in the last two years. All public isolate genome and metagenome datasets are available from the public IMG/M site (https://img.jgi.doe.gov/m/) that does not require login. However, ∼16% of the isolate genomes and 30% of metagenomes remain private and password protected in the IMG Expert Review site (https://img.jgi.doe.gov/mer/). Public availability of the datasets depends on their source: all genomes imported from NCBI are public, and all JGI-generated datasets are



Find Genomes by constructing a query using keywords, substrings, and AND/AND NOT/OR operators

- The 'AND' / 'AND NOT' query operator(s) that combine builder lines will be processed sequentially.
- Range queries enabled with keyword 'to' indicate range searches.
  For example, input '1 to 10' in the field of Geographic Metadata->Depth will retrieve datasets collected at depth between 1 and 10 meters.
- Mathematical operators '>=, >, <, <=' can be applied to a range search.
  For example, input '>=10' in the field of Geographic Metadata->Depth will retrieve datasets collected at depth of at least 10 meters.
- Comma separated keywords are treated as multiple values joined by 'OR' operators.
  For example, an input of '9606, 3702, 10090' in the field of NCBI IDs->NCBI Taxon ID retrieves 3 genomes.
- If 'Any Field' selected, search will be performed on each field in the same category and results will be pooled.
- Each field can only be searched only once as a specific field and multiple times as part of 'Any Field' selection.
- The maximum number of builder lines is limited to 5.
- Click on "Evaluate Query" button to see the results of each specific subquery and the overall count. Click on "Search" to see the over

Environmental Classification | Ecosystem Type | Simulated communities (sequence / Skin / **Soil** / Solar panel / Solid animal waste / Spacecraft Assembly Cleanroom / Sponge / Symbiotic fungal gardens and galler / Tailings pond / Terephthalate

AND | Geographic Metadata | Depth (Range: -10 to 10970 meters) | <= 0.1

AND | Geographic Metadata | Geographic Location | Wisconsin, Michigan

AND NOT | Study Dataset Names | Genome Name / Sample Name | agricultural

Constructed Query: Use the builder above to create search query

Select	Domain	Sequencing Status	Study Name	Genome Name / Sample Name
□	Metagenome	Permanent Draft	Soil microbial communities from Arlington Agricultural Research Station in Wisconsin and Kellogg Biological Station in Michigan, replicating the bioenergy cropping systems trials (BCSTs)	Soil microbial communities from Kellog Biological Station, Michigan, USA - Nitrogen cycling UWRJ-G10K1-12
□	Metagenome	Permanent Draft	Soil microbial communities from Arlington Agricultural Research Station in Wisconsin and Kellogg Biological Station in Michigan, replicating the bioenergy	Soil microbial communities from Kellog Biological Station, Michigan, USA - Nitrogen cycling UWRJ-G06K3-12 (SPAdes)

Constructed Query: (Environmental Classification -- Ecosystem Type [Soil]) AND (Geog
Metadata -- Geographic Location [Wisconsin, Michigan]) AND NOT (Study Dataset Name

- (Environmental Classification -- Ecosystem Type [Soil]): 8269 count(s).
- (Geographic Metadata -- Depth (Range: -10 to 10970 meters) [<= 0.1]): 5917 count(s).
- (Geographic Metadata -- Geographic Location [Wisconsin, Michigan]): 2879 count(s).
    - Wisconsin: 1746 count(s).
    - Michigan: 1133 count(s).
- (Study Dataset Names -- Genome Name / Sample Name [agricultural]): 547 count(s).

Final Combination: 53 count(s).

Figure 3. Example application of the Advanced Search Builder option of the new Genome Search feature. (**i**) Users can build a complex query to find all soil metagenome datasets sampled at depth of up to 10 cm in Wisconsin or Michigan that are not classified as agricultural soils. (**ii**) Users can click the Evaluate Query button to see statistics information. (**iii**) Query result is displayed after the Search button is clicked.

released following the JGI Data Release Policy as detailed at the JGI website (https://jgi.doe.gov/user-program-info/pmo-overview/policies/). For externally submitted datasets, IMG suggests that they remain private for 18 and 24 months for genomes and metagenomes, respectively. However, submitters can publicly release their private data prior to the deadline by requesting their release or using the tools in the IMG submission site.

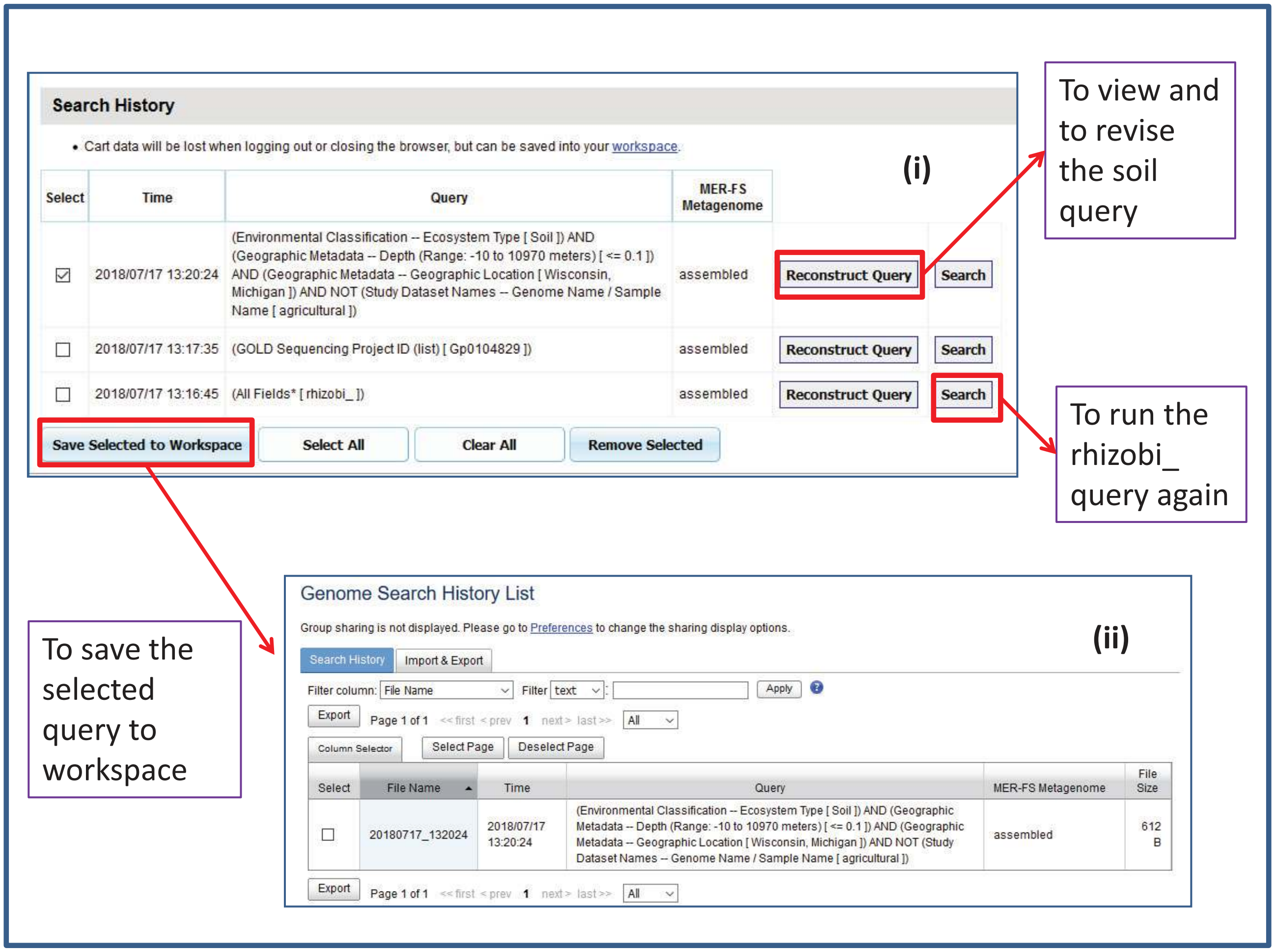
In addition to the main data warehouse serving the data via public IMG/M and Expert Review IMG, IMG also hosts two data marts, IMG/ABC (25) and IMG/VR (28), which provide more detailed data and unique tools for the analysis of biosynthetic gene clusters and viruses, respectively. IMG/ABC (https://img.jgi.doe.gov/abc/) continues to provide experimentally validated and predicted biosynthetic clusters and currently includes >1.3 million biosynthetic clusters predicted from isolate genomes and metagenomes. IMG/VR (https://img.jgi.doe.gov/vr/) provides sequence data, clustering and host information for viruses and viral fragments derived from metagenomic samples, as well as tools for their analysis. IMG/VR currently contains >700 000 isolate viruses and metagenomic viral contigs.

Data analysis

IMG users can query the data and perform comparative analysis through the IMG User Interface (UI) (https://img.jgi.doe.gov/m/). Most of the UI layout remains the same as described in the previous IMG publications (27). Several new features added in the last two years are described in this section.

In the IMG/M Home Page, there is a new **IMG & GOLD Citation Overview** icon. Clicking on the icon will lead to a page showing all publications since year 2000 citing IMG and/or GOLD (see Figure 1(i)). Publications are divided into different categories such as Microbiology, Genetics & Heredity, and Biochemistry & Molecular Biology. Clicking on the link leads to a zoomed-in view of the grid and the list of publications, which can be searched and sorted (Figure 1(ii)). The titles are linked to the corresponding publications (Figure 1(iii)).

The UI **Main** page lists the summary of IMG content and provides links to the detailed **IMG Statistics** and its **Data Usage Policy**. Main menu categories include **Find Genomes**, **Find Genes**, **Find Functions**, **Compare Genomes**, **OMICs**, **My IMG**, **Data Marts** and **Help**. In the Expert Review ver-



Search History

- Cart data will be lost when logging out or closing the browser, but can be saved into your workspace.

Select	Time	Query	MER-FS Metagenome		
☑	2018/07/17 13:20:24	(Environmental Classification -- Ecosystem Type [Soil]) AND (Geographic Metadata -- Depth (Range: -10 to 10970 meters) [<= 0.1]) AND (Geographic Metadata -- Geographic Location [Wisconsin, Michigan]) AND NOT (Study Dataset Names -- Genome Name / Sample Name [agricultural])	assembled	Reconstruct Query	Search
☐	2018/07/17 13:17:35	(GOLD Sequencing Project ID (list) [Gp0104829])	assembled	Reconstruct Query	Search
☐	2018/07/17 13:16:45	(All Fields* [rhizobi_])	assembled	Reconstruct Query	Search

To view and to revise the soil query

To run the rhizobi_ query again

To save the selected query to workspace

Genome Search History List

Group sharing is not displayed. Please go to Preferences to change the sharing display options.

Select	File Name	Time	Query	MER-FS Metagenome	File Size
☐	20180717_132024	2018/07/17 13:20:24	(Environmental Classification -- Ecosystem Type [Soil]) AND (Geographic Metadata -- Depth (Range: -10 to 10970 meters) [<= 0.1]) AND (Geographic Metadata -- Geographic Location [Wisconsin, Michigan]) AND NOT (Study Dataset Names -- Genome Name / Sample Name [agricultural])	assembled	612 B

Figure 4. Search History of the new Genome Search feature. (**i**) All searches done in a session will be saved. Users can also reconstruct and search any of the selected queries. (**ii**) Expert Review users have the additional ability to save any queries into the Workspace.

sion, there is an additional **Workspace** category, where registered users can access their private datasets stored in IMG and share them with collaborators (29).

The **Find Genomes** menu now provides a new powerful **Genome Search** capability, which has two options represented by different tabs: Quick Search and Advanced Search Builder. The Quick Search option allows users to quickly find datasets of interest by typing a keyword or a comma-separated list of identifiers. These will retrieve all datasets in IMG that have the corresponding keyword in the metadata fields, such as genome name or taxonomy, or are associated with the corresponding identifier, such as IMG genome id or NCBI taxonomy id. Figure 2 shows an example of the search with the keyword 'rhizobi_', where '_' is the wildcard character. This search retrieves the datasets containing 'rhizobi' in their metadata including study name, genome name, taxonomic lineage (order, family, genus, species) or any other field.

Since the Quick Search option often retrieves hundreds, if not thousands of datasets, an Advanced Search Builder option was introduced, allowing users to build complex, but very precise queries involving many sub-conditions connected by AND/AND NOT/OR clauses. For example, Figure 3 shows construction of a query that retrieves all soil metagenomes and metatranscriptomes sampled from the depth of up to 10 cm in Wisconsin or Michigan, that are not classified as agricultural soils. In order to assist in query construction, two new features were added: first, a user can view all IMG fields available for query construction by clicking on 'Show All Categories' link. The same expanded list can be used to select specific fields for inclusion in the query builder. Second, by clicking on 'Evaluate Query' button, a user can see the counts of datasets that will be retrieved based on each of the sub-conditions, and the total count of datasets that will be retrieved by the entire query. Based on these counts, users can revise their queries in order to achieve the desired results.

All queries submitted in the course of one UI session are saved and displayed in the Search History as shown in Figure 4(i). Users can use the History to rerun any selected query or, in the case of Advanced Search Builder, reconstruct the query and modify it to retrieve a different set of results. Like the Analysis Carts in IMG the query history is transient, and will disappear after UI session ends. However, on the Expert Review site of IMG, users can select any queries and save them into their Workspace (see Figure 4(ii)).

Although the new **Genome Search** includes all of the capabilities of the old genome search function, which was available in IMG since its inception, the old search function is still listed in **Find Genomes** menu under the title of **Original Genome Search**.

The BLAST function under **Find Genes** menu has been drastically expanded to support five options: Selected Genomes, 16S RNA, Virus (28), CRISPR Spacers, and All Isolates. For the Selected Genomes option, IMG enables protein, nucleotide and translated BLAST search of a user-submitted sequence against a dynamically-created BLAST database based on the user selection of genomes and/or metagenomes. For the rest of the options, BLAST databases are pre-generated and refreshed regularly using the most recent IMG database content available at the time. 16S RNA BLAST allows users to search their sequence of interest against a collection of small subunit rRNA sequences extracted from IMG genomes, metagenomes, and metatranscriptomes. This reference database is updated on a quarterly basis. Virus BLAST supports protein and nucleotide BLAST searches against a database of publicly available isolate viruses and metagenomic viral contigs, which is also updated on a quarterly basis. Nucleotide BLAST database of CRISPR Spacers and protein and nucleotide database for BLAST All Isolates are updated approximately twice per month. In addition to real-time BLAST searches, the Expert Review site of IMG allows users to submit larger BLAST jobs via 'Submit a Computation Job' mechanism, which runs BLAST jobs on the background and notifies users of job completion via email.

Due to the significant growth of IMG content, statistical analyses of genomes and metagenomes have become increasingly popular. Therefore we have implemented a new set of statistical test tools enabling users to quantitatively compare the differences between functional genes of communities (metagenomes) or groups of isolates (genomes) and assign statistical significance to these findings. Since these statistical tools operate on two or more predefined sets of genomes or metagenomes, which can include as many as several thousand datasets, they are too computationally intensive to finish in real time. Therefore they are made available only on the Expert Review site of IMG, as an option in **Workspace**, which is generally used to support saved datasets and computations on demand. New tools are intended to cover all major scenarios of statistical analysis and are based on five statistical methods: Fisher's Exact (30), Mann–Whitney (31) and Welch's *T*-test (32) are used to compare two groups or sets of genomes, while Analysis of Variance or ANOVA (33) and Kruskal–Wallis (34) are used to compare 3–10 groups.

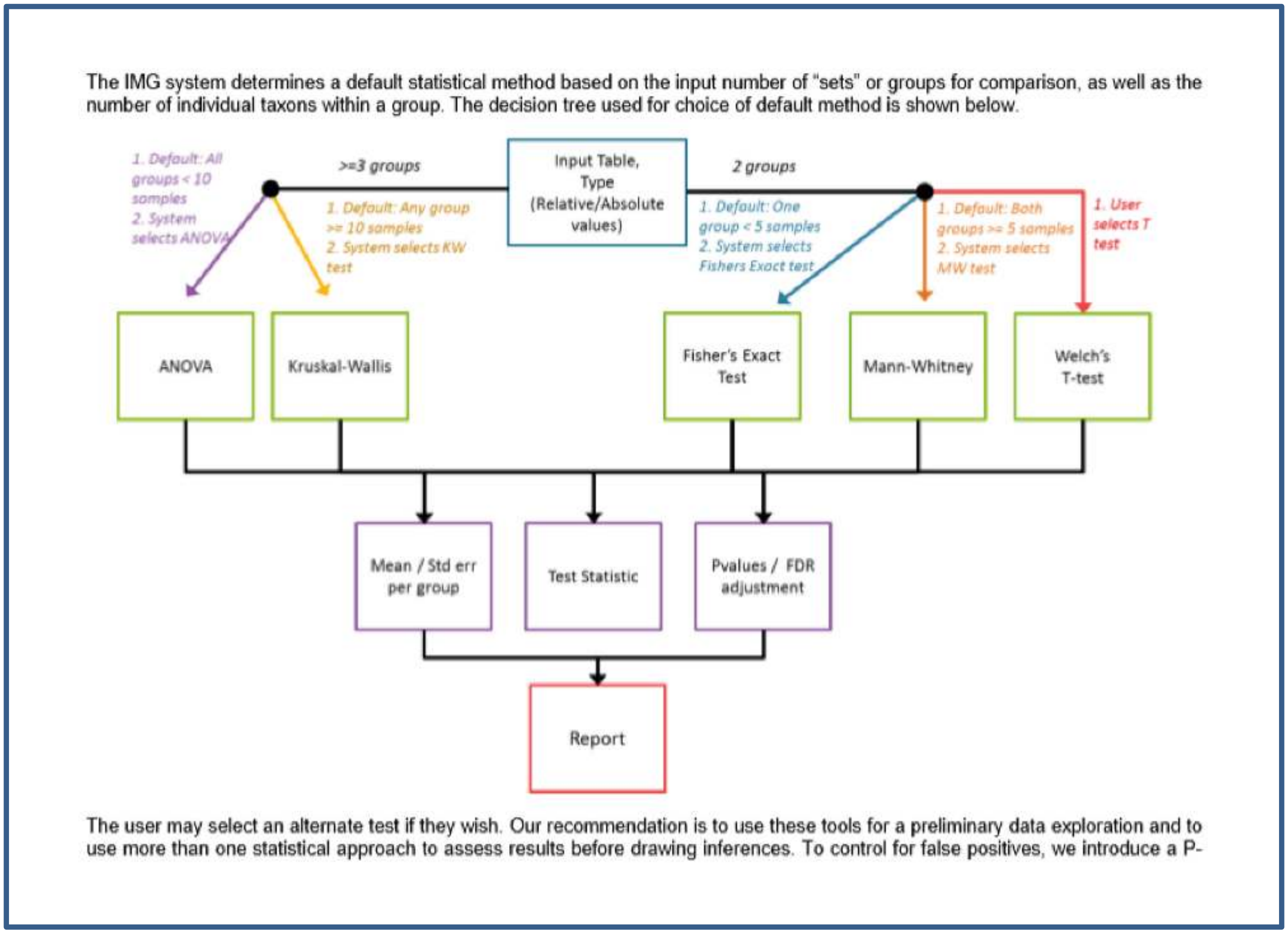
For these new tools, IMG attempts to guide the users to the most judicious method and parameter choices, with the caveat that experimental protocols, data processing, and QC methods can influence functional profiles and introduce artifacts in the analysis. Our recommendation therefore is to use these tools for a preliminary data exploration and to use more than one statistical approach to assess the results before drawing any biological inferences. The abundance of genes assigned to a function (KO, TIGRfam, Pfam or COG) may be compared in terms of their relative abundance between communities or in two or more populations or 'sets' of isolates. For isolate comparisons, absolute abundance may also be compared. The choice of a default statistical method is informed by the number of groups and number of individuals per group. For example, to handle expected errors due to small group sizes, the system defaults to Fisher's Exact test. The decision tree for choice of default statistical test is depicted in Figure 5. The users may select an alternate test if they disagree with the default test or wish to compare results from multiple tests. For function category comparisons (KO modules, Pfam or COG category), gene counts assigned to all functions within each category are summed to aggregate, and then compared. For metagenome comparisons alone, estimated gene copies (calculated by multiplying with average read depth of the scaffold the gene resides on) may be compared. To control for false positives, we introduce a *P*-value adjustment for multiple corrections using a conventional Benjamini-Hochberg method (35) to control false discovery rate (FDR), and the functions with an adjusted FDR *P*-value of <0.05 are deemed 'significant.' A FDR of 0.05 indicates that approximately 5% of significant tests will be 'false discoveries,' that is, incorrectly reject the null hypothesis. We further advise the users to consult the many varied reviews and guidelines related to this topic (36). The user interface, which makes calls to a combination of custom Perl and R libraries, enables the download of the displayed output table, as well as a complete output table including raw and normalized gene counts for every function in every input dataset. Figure 6 shows an example of selecting two genome sets to check gene count distribution based on Pfams using the default selected method (Mann-Whitney in this particular example). Users are notified of job completion by email, and analysis results are available from **My Jobs** in Workspace. Users can select to export the result data table or to download a full report (see Figure 7). A future version of the software will allow creation of on-the-fly genome sets for comparisons, as well as enable a taxonomy-based comparison.

The User Guide in the **Help** menu now includes all the updated IMG user guides. **Site Map** lists all essential UI functions together with the corresponding user guides indicated by 'book icons' at the right.

New metagenome binning

Advances in next-generation sequencing technologies coupled with *de novo* metagenome assembly developments have enabled the reconstruction of population genomes directly from an environmental sample (37). Genome-resolved metagenomics has become a powerful approach in microbiome research to link metabolic and functional potential with phylogenetic information, providing genomic context for uncultivated microbes (38). Recent large-scale recovery of metagenome-assembled genomes (MAGs) from diverse environments has provided significant insights into evolutionary and metabolic properties of uncultivated bacteria and archaea (39,40).

Until recently, genome-resolved metagenomic approaches have proved challenging due to high microbial diversity and a sizeable fraction of uncharacterized


The IMG system determines a default statistical method based on the input number of "sets" or groups for comparison, as well as the number of individual taxons within a group. The decision tree used for choice of default method is shown below.

1. Default: All groups < 10 samples
2. System selects ANOVA
>=3 groups
Input Table, Type (Relative/Absolute values)
2 groups
1. Default: Any group >= 10 samples
2. System selects KW test
1. Default: One group < 5 samples
2. System selects Fishers Exact test
1. Default: Both groups >= 5 samples
2. System selects MW test
1. User selects T test
ANOVA
Kruskal-Wallis
Fisher's Exact Test
Mann-Whitney
Welch's T-test
Mean / Std err per group
Test Statistic
Pvalues / FDR adjustment
Report

The user may select an alternate test if they wish. Our recommendation is to use these tools for a preliminary data exploration and to use more than one statistical approach to assess results before drawing inferences. To control for false positives, we introduce a P-

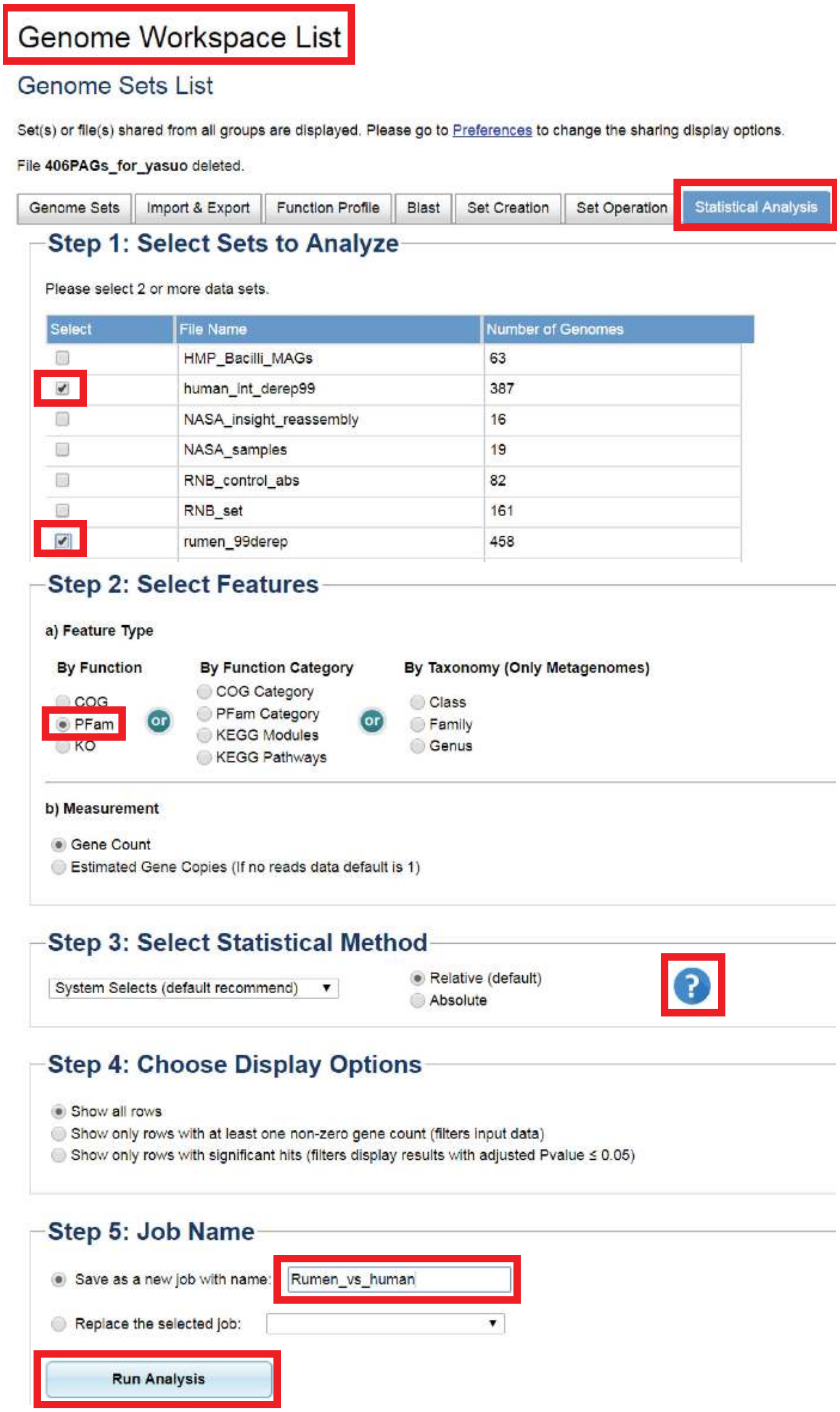
Figure 5. Decision tree for selection of default statistical test method. FDR is false discovery rate.

genomes compared to reference databases (41). Methods to link contigs to their respective genomes – termed binning – are required post-assembly. Binning methods exploit sequence composition, species abundance, chromosome organization, or other inherent properties of the shotgun metagenome data to associate groups of contigs as deriving from a seemingly coherent microbial species. A myriad of binning tools and approaches are available (for a review, see (37)), with a recent assessment providing important benchmark comparisons across a subset of binning algorithms and tools (41). Following binning, estimates for genome completeness and contamination are typically evaluated by recovery of a set of core single-copy marker genes and can be reported according to the recently described Genomic Standards Consortium (GSC)-compliant standards (42).

We have incorporated the automated metagenome binning tool, MetaBAT (43), along with CheckM (44) and other quality assessment metrics for compliance with the MIMAG standards, to provide MAGs on a per-sample basis for 4,789 metagenomes in IMG. These datasets were selected from assembled public metagenomes in IMG with adequate coverage information. A new 'Genome Bin Count' field has been added to the Data Statistics menu in the Metagenome Table Configuration for users to identify which metagenomes have associated bins. A total of 42 484 bins (4425 high-quality and 38 059 medium-quality bins) are publicly available. MetaBAT (version 0.32.4) was run with a 3000 bp minimum contig cutoff, contig coverage information, and parameter '-superspecific' for maximum specificity.

In the **Metagenome Statistics** section of the metagenome detail page, there will be a **Metagenome Bins** count if there are bins associated with this metagenome (Figure 8(i)). Users can click on the count to view the bins (Figure 8(ii)). Each bin shown on the Metagenome Bins page has a Bin ID, Bin Quality (high, medium or low), associated estimates for genome completeness and contamination, and predicted taxonomic lineage assignment based on majority ruling for contig-level affiliation. Expert Review users have the additional capability to select one or more bins to save as workspace scaffold sets for further analysis. To view all scaffolds included in a bin, simply click on the scaffold count of the bin (Figure 8(iii)); additional analysis options (e.g. analysis of nucleotide oligomer composition or functional profile of the bin) can be accessed by adding all scaffolds from the bin to the Scaffold Cart.

We are currently working on incorporating metagenome binning into the IMG Annotation Pipeline for all metagenomes. All new metagenomes going through the annotation pipeline starting from late 2018 will



Genome Workspace List

Genome Sets List

Set(s) or file(s) shared from all groups are displayed. Please go to Preferences to change the sharing display options.

File **406PAGs_for_yasuo** deleted.

Step 1: Select Sets to Analyze

Please select 2 or more data sets.

Select	File Name	Number of Genomes
□	HMP_Bacilli_MAGs	63
☑	human_int_derep99	387
□	NASA_insight_reassembly	16
□	NASA_samples	19
□	RNB_control_abs	82
□	RNB_set	161
☑	rumen_99derep	458

Step 2: Select Features

a) Feature Type

By Function
○ COG
● PFam
○ KO

or

By Function Category
○ COG Category
○ PFam Category
○ KEGG Modules
○ KEGG Pathways

or

By Taxonomy (Only Metagenomes)
○ Class
○ Family
○ Genus

b) Measurement

● Gene Count
○ Estimated Gene Copies (If no reads data default is 1)

Step 3: Select Statistical Method

System Selects (default recommend) ▼

● Relative (default)
○ Absolute

?

Step 4: Choose Display Options

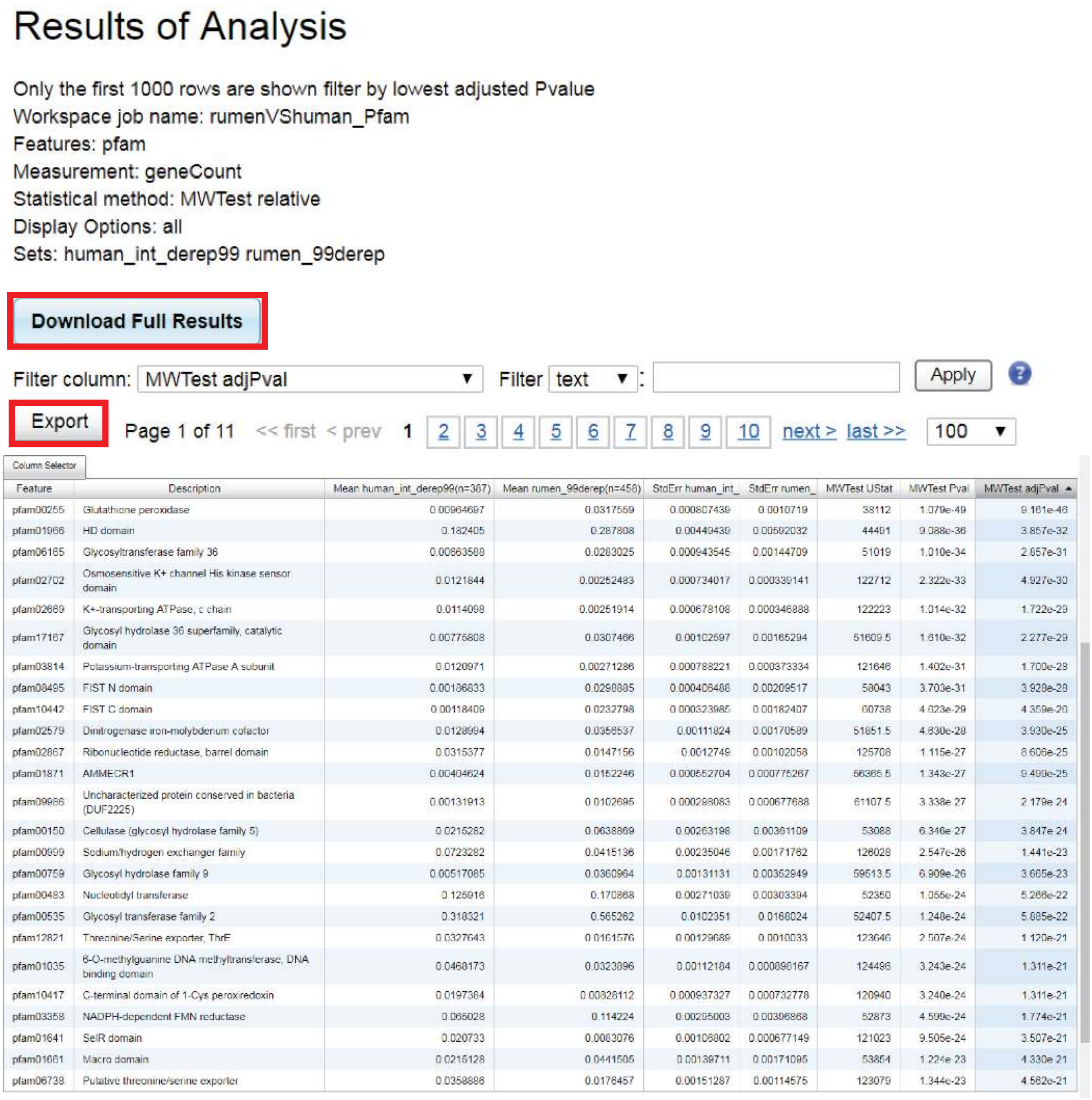
● Show all rows
○ Show only rows with at least one non-zero gene count (filters input data)
○ Show only rows with significant hits (filters display results with adjusted Pvalue ≤ 0.05)

Step 5: Job Name

● Save as a new job with name: Rumen_vs_human

○ Replace the selected job: ▼

Figure 6. The new analysis tools are available in the Statistical Analysis tab of Workspace Genome Sets. In this particular example, the user selects two genome sets to measure gene count by Pfam using default system recommendation. Users can gather more information regarding analysis methods by clicking on the question mark to view a detailed description. The analysis will be run on the background and the result will be saved as a new job. The user can click on the Run Analysis button to submit the analysis request. UI will inform the user which default analysis method has been chosen.



Results of Analysis

Only the first 1000 rows are shown filter by lowest adjusted Pvalue
Workspace job name: rumenVShuman_Pfam
Features: pfam
Measurement: geneCount
Statistical method: MWTest relative
Display Options: all
Sets: human_int_derep99 rumen_99derep

Download Full Results
Filter column: MWTest adjPval ▼
Filter text ▼ :
Apply
?
Export
Page 1 of 11
<< first < prev 1 2 3 4 5 6 7 8 9 10 next > last >>
100 ▼
Column Selector

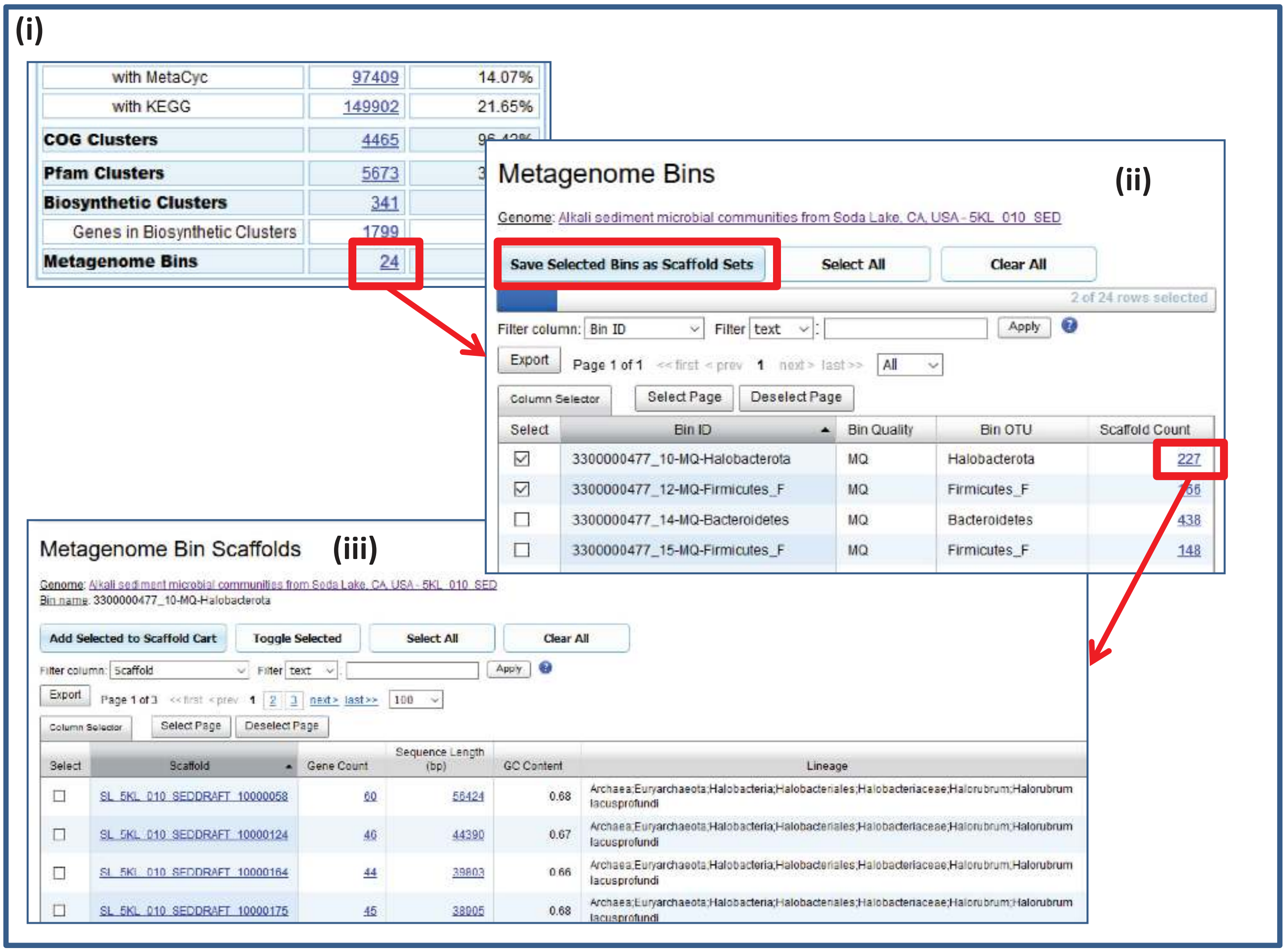
Feature	Description	Mean human_int_derep99(n=387)	Mean rumen_99derep(n=456)	StdErr human_int_	StdErr rumen_	MWTest UStat	MWTest Pval	MWTest adjPval ▲
pfam00255	Glutathione peroxidase	0.00964697	0.0317559	0.000807439	0.0010719	38112	1.079e-49	9.161e-46
pfam01966	HD domain	0.182405	0.287808	0.00449439	0.00592032	44491	9.088e-36	3.857e-32
pfam06165	Glycosyltransferase family 36	0.00663589	0.0283025	0.000943545	0.00144709	51019	1.010e-34	2.857e-31
pfam02702	Osmosensitive K+ channel His kinase sensor domain	0.0121844	0.00252483	0.000734017	0.000339141	122712	2.322e-33	4.927e-30
pfam02669	K+-transporting ATPase, c chain	0.0114098	0.00251914	0.000678108	0.000346888	122223	1.014e-32	1.722e-29
pfam17167	Glycosyl hydrolase 36 superfamily, catalytic domain	0.00775808	0.0307466	0.00102597	0.00165294	51609.5	1.610e-32	2.277e-29
pfam03814	Potassium-transporting ATPase A subunit	0.0120971	0.00271286	0.000788221	0.000373334	121646	1.402e-31	1.700e-28
pfam08495	FIST N domain	0.00136833	0.0298885	0.000406486	0.00209517	58043	3.703e-31	3.928e-28
pfam10442	FIST C domain	0.00118409	0.0232798	0.000323985	0.00182407	60738	4.623e-29	4.359e-26
pfam02579	Dinitrogenase iron-molybdenum cofactor	0.0128994	0.0358537	0.00111824	0.00170589	51851.5	4.830e-28	3.930e-25
pfam02867	Ribonucleotide reductase, barrel domain	0.0315377	0.0147156	0.0012749	0.00102058	125708	1.115e-27	8.606e-25
pfam01871	AMMECR1	0.00404624	0.0152246	0.000552704	0.000775267	56365.5	1.343e-27	9.499e-25
pfam09986	Uncharacterized protein conserved in bacteria (DUF2225)	0.00131913	0.0102695	0.000298083	0.000677688	61107.5	3.338e-27	2.179e-24
pfam00150	Cellulase (glycosyl hydrolase family 5)	0.0215282	0.0638869	0.00263198	0.00361109	53088	6.346e-27	3.847e-24
pfam00999	Sodium/hydrogen exchanger family	0.0723282	0.0415136	0.00235046	0.00171762	126028	2.547e-26	1.441e-23
pfam00759	Glycosyl hydrolase family 9	0.00517085	0.0360964	0.00131131	0.00352949	59513.5	6.909e-26	3.665e-23
pfam00483	Nucleotidyl transferase	0.125916	0.170868	0.00271039	0.00303394	52350	1.055e-24	5.266e-22
pfam00535	Glycosyl transferase family 2	0.318321	0.565262	0.0102351	0.0168024	52407.5	1.240e-24	5.885e-22
pfam12821	Threonine/Serine exporter, ThrE	0.0327643	0.0161576	0.00129689	0.0010033	123646	2.507e-24	1.120e-21
pfam01035	6-O-methylguanine DNA methyltransferase, DNA binding domain	0.0468173	0.0323896	0.00112184	0.000898167	124496	3.243e-24	1.311e-21
pfam10417	C-terminal domain of 1-Cys peroxiredoxin	0.0197384	0.00828112	0.000937327	0.000732778	120940	3.240e-24	1.311e-21
pfam03358	NADPH-dependent FMN reductase	0.065028	0.114224	0.00295003	0.00396868	52873	4.599e-24	1.774e-21
pfam01641	SelR domain	0.020733	0.0083076	0.00106802	0.000677149	121023	9.505e-24	3.507e-21
pfam01661	Macro domain	0.0215128	0.0441505	0.00139711	0.00171095	53854	1.224e-23	4.330e-21
pfam06738	Putative threonine/serine exporter	0.0358898	0.0178457	0.00151287	0.00114575	123079	1.344e-23	4.562e-21

Figure 7. Analysis statuses and results are available from My Jobs in Workspace. A job starts with waiting status. Users will be able to view the analysis result when a job is complete. The result data table can be exported. Users can also select to download a complete report.

have bins automatically assigned. We also plan to backfill metagenome binning information for all existing metagenomes in the IMG database, subject to availability of computational resources.

CONCLUSION

The current version of IMG/M v.5.0 (as of July 2018) contains 77 821 (64 446 public) archaeal, bacterial and eukaryotic genomes and 28 799 (among them 19 942 public) metagenome datasets with over 54 billion (48 billion public) protein coding genes. It continues to grow, as new genomes and metagenomes generated at the JGI and obtained from external sources are being added on a regular basis. Thus, our major challenge remains the same: how to process and store hundreds of thousands of datasets, while efficiently serving the data to the scientific community and providing up-to-date analysis tools in order to support biological discovery (45). There is also a need to develop new analysis and visualization tools to support phylogenetic and functional exploration of available metagenome bins.



with MetaCyc	97409	14.07%
with KEGG	149902	21.65%
COG Clusters	4465	96.42%
Pfam Clusters	5673	3
Biosynthetic Clusters	341	
Genes in Biosynthetic Clusters	1799	
Metagenome Bins	24	

Metagenome Bins

Genome: Alkali sediment microbial communities from Soda Lake, CA, USA - 5KL_010_SED

Select	Bin ID	Bin Quality	Bin OTU	Scaffold Count
☑	3300000477_10-MQ-Halobacterota	MQ	Halobacterota	227
☑	3300000477_12-MQ-Firmicutes_F	MQ	Firmicutes_F	156
□	3300000477_14-MQ-Bacteroidetes	MQ	Bacteroidetes	438
□	3300000477_15-MQ-Firmicutes_F	MQ	Firmicutes_F	148

Metagenome Bin Scaffolds

Genome: Alkali sediment microbial communities from Soda Lake, CA, USA - 5KL_010_SED
Bin name: 3300000477_10-MQ-Halobacterota

Select	Scaffold	Gene Count	Sequence Length (bp)	GC Content	Lineage
□	SL_5KL_010_SEDDRAFT_10000058	60	56424	0.68	Archaea;Euryarchaeota;Halobacteria;Halobacteriales;Halobacteriaceae;Halorubrum;Halorubrum lacusprofundi
□	SL_5KL_010_SEDDRAFT_10000124	46	44390	0.67	Archaea;Euryarchaeota;Halobacteria;Halobacteriales;Halobacteriaceae;Halorubrum;Halorubrum lacusprofundi
□	SL_5KL_010_SEDDRAFT_10000164	44	39803	0.66	Archaea;Euryarchaeota;Halobacteria;Halobacteriales;Halobacteriaceae;Halorubrum;Halorubrum lacusprofundi
□	SL_5KL_010_SEDDRAFT_10000175	45	38905	0.68	Archaea;Euryarchaeota;Halobacteria;Halobacteriales;Halobacteriaceae;Halorubrum;Halorubrum lacusprofundi

Figure 8. New metagenome bins in IMG. (**i**) Metagenome Statistics in the metagenome detail page shows the number of bins. (**ii**) Users can view more detailed information of bins by clicking on the count. Expert Review users can also select one or more bins to save as workspace scaffold sets. (iii) After the user clicks the scaffold count, a new Metagenome Bin Scaffolds page will show up listing all scaffolds in the bin together with more detailed information on each scaffold.

Some of the future developments we envision are based on popular demand and include the possibility of providing quick annotation, summary statistics, and data downloads for unassembled metagenome data. Another line of developments includes upgrading the content of the IMG database by introducing new reference databases and analysis tools, such as adding non-coding RNA prediction in metagenome sequences and assignment of metagenomic proteins to TIGRfams and some InterPro families, and including a newer version of antiSMASH (46) for biosynthetic cluster prediction.

We are also investigating the possibility of providing some of the existing IMG tools as microservices. As part of this development, we have recently implemented a Microbial Species Identifier (MiSI) system, which is the first IMG microservice providing ANI data (22) to IMG users through API access. MiSI will be integrated into IMG through IMG UI in the near future. This effort will decouple ANI computation from the IMG annotation pipeline, thus obviating the need to submit the genomes to IMG in order to obtain ANI results. New microservice is expected to reduce the load on the IMG annotation pipeline, while providing IMG users with more flexible analysis options.

FUNDING

Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy [DE-AC02-05CH11231]; National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy. Funding for open access charge: Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy [DE-AC02-05CH11231].

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrahi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
2. Mukherjee,S., Stamatis,D., Bertsch,J., Ovchinnikova,G., Verezemska,O., Isbandi,M., Thomas,A., Ali,R., Sharma,K.,

Kyrpides,N.C. *et al.* (2017) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res.*, **45**, D446–D456.

3. Huntemann,M., Ivanova,N.N., Mavromatis,K., Tripp,H.J., Paez-Espino,D., Palaniappan,K., Szeto,E., Pillay,M., Chen,I.A., Pati,A. *et al.* (2015) The standard operating procedure of the DOE-JGI microbial genome annotation pipeline (MGAP v. 4). *Stand. Genomic Sci.*, **10**, 86.
4. Field,D., Sterk,P., Kottmann,R., De Smet,J.W., Amaral-Zettler,L., Cochrane,G., Cole,J.R., Davies,N., Dawyndt,P., Garrity,G.M. *et al.* (2014) Genomic standards consortium projects. *Stand Genomic Sci.*, **9**, 599–601.
5. Bland,C., Ramsey,T.L., Sabree,F., Lowe,M., Brown,K., Kyrpides,N.C. and Hugenholtz,P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
6. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935
7. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA familes. *Nucleic Acids Res.*, **46**, D335–D342.
8. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
9. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,FW and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
10. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat. Protoc.*, **2**, 953–971.
11. Moller,S., Croning,M.D.R. and Apweiler,R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.
12. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
13. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
14. Haft,D.H., Selengut,J.D., Richter,R.A., Harkins,D., Basu,M.K. and Bec,K.E. (2013) TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.
15. Finn,R.D., Clements,J., Arndt,W., Benjamin,L.M., Travis,J.W., Fabian,S., Alex,B. and Sean,R.E. (2015) HMMER web server: 2015 update. *Nucleic Acids Res.*, **43**, W30–W38.
16. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
17. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
18. Kielbasa,S.M., Wan,R., Sato,K., Horton,P. and Frith,M.C. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.*, **21**, 487–493.
19. Caspi,R., Billington,R., Ferrer,L., Foerster,H., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.
20. Ivanova,N.N., Anderson,I., Lykidis,A., Mavrommatis,K., Mikhailova,N., Chen,I.A., Szeto,E., Palaniappan,K., Markowitz,V.M. and Kyrpides,N.C. (2007) *Metabolic Reconstruction of Microbial Genomes and Microbial Community Metagenomes*. Technical Report 62292, Lawrence Berkeley National Laboratory.
21. Chen,I.A., Markowitz,V.M., Chu,K., Anderson,I., Mavromatis,K., Kyrpides,N.C. and Ivanova,N.N. (2013) Improving microbial genome annotations in an integrated database context, *PLoS One*, **8**, e54859.
22. Varghese,N.J., Mukherjee,S., Ivanova,N., Konstantinidis,K.T., Mavrommatis,K., Kyrpides,N.C. and Pati,A. (2015) Microbial species delineation using whole genome sequences. *Nucleic Acids Res.*, **43**, 6761–6771.
23. Mavromatis,K., Chu,K., Ivanova,N., Hooper,S.D., Markowitz,V.M. and Kyrpides,N.C. (2009) Gene context analysis in the Integrated Microbial Genomes (IMG) data management system, accepted for publication, *PLoS One*, **4**, e7979.
24. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
25. Hadjithomas,M., Chen,I.A., Chu,K., Huang,J., Ratner,A., Palaniappan,K., Andersen,E., Markowitz,V., Kyrpides,N.C. and Ivanova,N.N. (2017) IMG-ABC: new features for bacterial secondary metabolism analysis and targeted biosynthetic gene cluster discovery in thousands of microbial genomes. *Nucleic Acids Res.*, **45**, D560–D565.
26. Huntemann,M., Ivanova,N.N., Mavromatis,K., Tripp,H.J., Paez-Espino,D., Tennessen,K., Palaniappan,K., Szeto,E., Pillay,M., Chen,I.A. *et al.* (2015) The standard operating procedure of the DOE-JGI metagenome annotation pipeline (MAP v. 4). *Stand. Genomic Sci.*, **11**, 17.
27. Chen,I.A., Markowitz,V.M., Chu,K., Palaniappan,K., Szeto,E., Pillay,M., Ratner,A., Huang,J., Andersen,E., Huntemann,M. *et al.* (2017) IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.*, **45**, D507–D516.
28. Paez-Espino,D., Chen,I.A., Palaniappan,K., Ratner,A., Chu,K., Szeto,E., Pillay,M., Huang,J., Markowitz,V.M., Nielsen,T. *et al.* (2016). IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.*, **45**, D457–D465.
29. Chen,I.A., Markowitz,V.M., Palaniappan,K., Szeto,E., Chu,K., Huang,J., Ratner,A., Pillay,M., Hadjithomas,M., Huntemann,M. *et al.* (2016) Supporting community annotation and user collaboration in the integrated microbial genomes (IMG) system. *BMC Genomics*, **17**, 307.
30. Fisher,R.A. (1956) Mathematics of a lady tasting tea. In: Newman,JR (ed). *The World of Mathematics*. Courier Dover Publications, Vol. **3**, ISBN 978-0-486-41151-4.
31. Mann,H.B. and Whitney,D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.
32. Welch,B.L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, **34**, 28–35.
33. Fisher,R.A. (1921) On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 3–32.
34. Field,A. (2009) *Discovering Statistics using SPSS*. Sage Publications, Inc., ISBN-13: 978–1847879073. ISBN-10: 1847879071.
35. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B (Methodological)*, **57**, 289–300.
36. Odintsova,V., Tyakht,A. and Alexeev,D. (2017). Guidelines to statistical analysis of microbial composition data inferred from metagenomic sequencing. *Curr. Issues Mol. Biol.*, **24**, 17–36.
37. Sangwan,N., Xia,F. and Gilbert,J.A. (2016) Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, **4**, 8.
38. Quince,C., Walker,A.W., Simpson,J.T., Loman,N.J. and Segata,N. (2017) Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, **35**, 833.
39. Parks,D.H., Rinke,C., Chuvochina,M., Chaumeil,P.-A., Woodcroft,B.J., Evans,P.N., Hugenholtz,P. and Tyson,G.W. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.*, **2**, 1533–1542.
40. Tully,B.J., Graham,E.D. and Heidelberg,J.F. (2018) The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, **5**, 170203.
41. Sczyrba,A., Hofmann,P., Belmann,P., Koslicki,D., Janssen,S., Droge,J., Gregor,I., Majda,S., Fiedler,J., Dahms,E *et al.*2017) Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.
42. Bowers,R.M., Kyrpides,N.C., Stepanauskas,R., Harmon-Smith,M., Doud,D., Reddy,T.B.K., Schulz,F., Jarett,J., Rivers,A.R., Eloe-Fadrosh,E.A. *et al.* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, **35**, 725–731.

43. Kang,D.D., Froula,J., Egan,R. and Wang,Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **3**, e1165.
44. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
45. Chen,I.A., Markowitz,V., Szeto,E., Palaniappan,K. and Chu,K. (2014) *Maintaining a Microbial Genome & Metagenome Data Analysis System in an Academic Setting*. SSDBM, 14, July 2014.
46. Blin,K., Wolf,T., Chevrette,M.G., Lu,X., Schwalen,C.J., Kautsar,S.A., Suarez Duran,H.G., de Los Santos,E.L.C., Kim,H.U., Nave,M. *et al.* (2017) antiSMASH 4.0 – improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, **45**, W36–W41.