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

I-Min A. Chen1,*, Ken Chu1, Krishna Palaniappan1, Manoj Pillay1, Anna Ratner1, Jinghua Huang1, Marcel Huntemann1, Neha Varghese1, James R. White2, Rekha Seshadri1, Tatyana Smirnova1, Edward Kirton1, Sean P. Jungbluth1, Tanja Woyke1, Emiley A. Eloe-Fadrosh1, Natalia N. Ivanova1,* and Nikos C. Kyrpides1,*

1Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA and 2Resphera 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-
Nucleic Acids Research, 2019, Vol. 47, Database issue

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.

Fusion 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 hmmscan (15) for assignment of COGs to metaproteomes, and omission of several computationally intensive steps, such as non-coding RNA pre-
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, 9,674 (8,388 public) viruses and 1,215 (1,190 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, 4,605 (2,423 public) metatranscriptome datasets, and two new types of reduced complexity ‘microbiomes’ - cell enrichments (1,333 including 801 public), and single particle sorts (3,954 including 3,486 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 6,500 genomes and 6,200 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
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-
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, TIGIfam, 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.’ An 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
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 (4,425 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
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.
have bins automatically assigned. We also plan to back-fill 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.
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-Mizrachi, 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., Verezenska, O., Isbandi, M., Thomas, A., Ali, R., Sharma, K.,
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., Schwelen, 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.