Published online 13 October 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D507–D516***
doi: 10.1093/nar/gkw929

IMG/M: integrated genome and metagenome comparative data analysis system

I-Min A. Chen[1,*], Victor M. Markowitz[1], Ken Chu[1], Krishna Palaniappan[1], Ernest Szeto[1], Manoj Pillay[1], Anna Ratner[1], Jinghua Huang[1], Evan Andersen[1], Marcel Huntemann[2], Neha Varghese[2], Michalis Hadjithomas[2], Kristin Tennessen[2], Torben Nielsen[2], Natalia N. Ivanova[2] and Nikos C. Kyrpides[2,*]

[1]Biosciences Computing Group, Computational Science Department, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA and [2]Microbial Genome and Metagenome Program, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

Received September 16, 2016; Accepted October 05, 2016

ABSTRACT

The Integrated Microbial Genomes with Microbiome Samples (IMG/M: https://img.jgi.doe.gov/m/) system contains annotated DNA and RNA sequence data of (i) archaeal, bacterial, eukaryotic and viral genomes from cultured organisms, (ii) single cell genomes (SCG) and genomes from metagenomes (GFM) from uncultured archaea, bacteria and viruses and (iii) metagenomes from environmental, host associated and engineered microbiome samples. Sequence data are generated by DOE's Joint Genome Institute (JGI), submitted by individual scientists, or collected from public sequence data archives. Structural and functional annotation is carried out by JGI's genome and metagenome annotation pipelines. A variety of analytical and visualization tools provide support for examining and comparing IMG/M's datasets. IMG/M allows open access interactive analysis of publicly available datasets, while manual curation, submission and access to private datasets and computationally intensive workspace-based analysis require login/password access to its expert review (ER) companion system (IMG/M ER: https://img.jgi.doe.gov/mer/). Since the last report published in the 2014 NAR Database Issue, IMG/M's dataset content has tripled in terms of number of datasets and overall protein coding genes, while its analysis tools have been extended to cope with the rapid growth in the number and size of datasets handled by the system.

DATA SOURCES AND PROCESSING

The Integrated Microbial Genomes with Microbiome Samples (IMG/M: https://img.jgi.doe.gov/m/) includes archaea, bacteria, eukarya, plasmids, viruses, genome fragments (partially sequenced genomes), as well as metagenomes and metatranscriptome datasets. Since 2014, the two separate IMG systems for isolate genomes and metagenomes have been merged into a single one.

NCBI still serves as IMG's major source of genome data for cultured and uncultured organisms. However, starting from November 2012, all new sequence data are downloaded from GenBank (1) rather than RefSeq (2) using GOLD v.5 (3) metadata. These sequence data files go through IMG Submission system (https://img.jgi.doe.gov/submit/) and IMG annotation pipeline (4) before being integrated into the IMG data warehouse.

IMG continues to support external submissions of assembled genome data generated with any sequencing technology. Each genome or metagenome submission must be associated with a GOLD v.5 analysis project definition rather than a sequencing project. The new GOLD v.5 supports more extensive metadata definition and better provenance for output of complex analysis pipelines, such as combined assembly of metagenomes and single cell genomes or genomes extracted from metagenomes (3).

All isolate genomes submitted for annotation and integration are processed through JGI's Microbial Genome Annotation Pipeline (MGAP v. 4) (4). IMG supports two types of isolate genome submissions: GenBank files with predicted features, and unannotated sequence files in FASTA format. For the latter, IMG performs feature prediction including identification of protein-coding genes, non-coding RNAs and regulatory RNA features, as well as CRISPR

*To whom correspondence should be addressed. Tel: +1 925 296 5697; Fax: +1 925 296 5666; Email: IMAChen@lbl.gov
Correspondence may also be addressed to Nikos C. Kyrpides. Email: nckyrpides@lbl.gov
Present address: I-Min A. Chen, Biosciences Computing Group, Computational Science Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

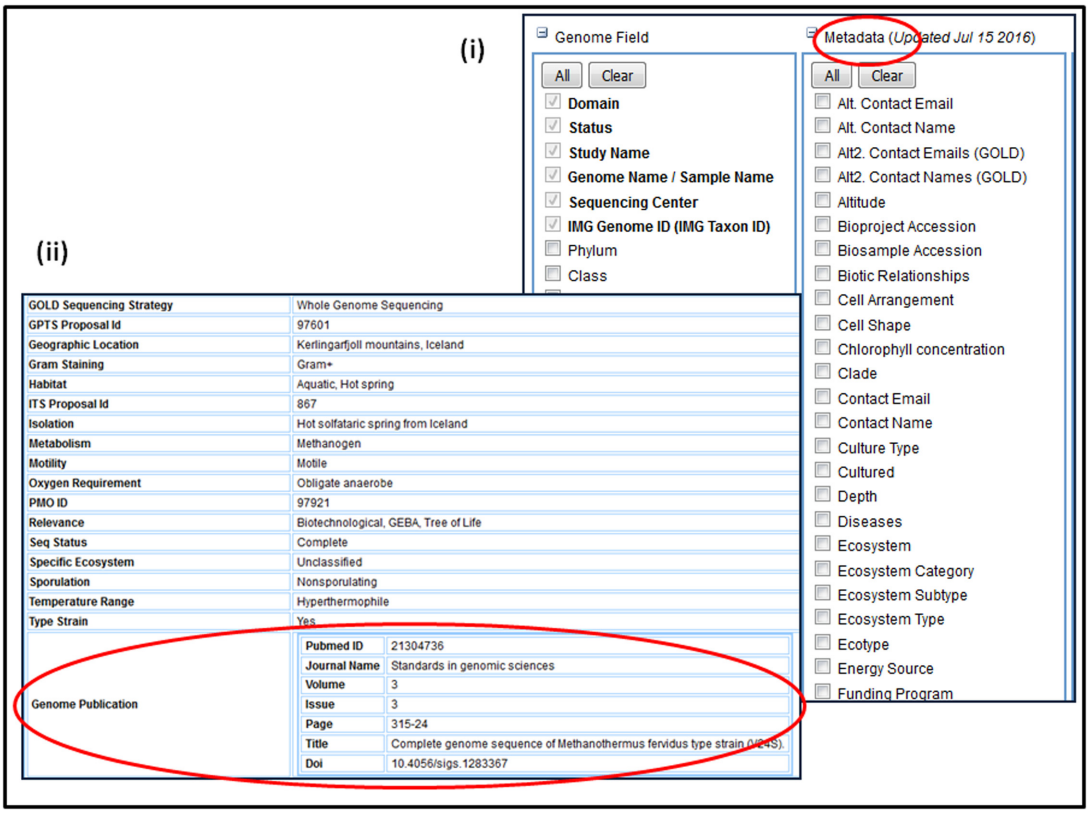

Figure 1. New Metadata from GOLD v.5. (**i**) IMG now provides additional metadata field selection obtained from GOLD v.5. (**ii**) Genome publication list in the Genome Detail page shows researchers the publication reference.

elements. Briefly, feature prediction involves detection of CRISPR elements using a modified CRT (5), tRNAs using cmsearch from the Infernal 1.1 package (6), Ribosomal RNA genes (5S, 16S, 23S) using hmmsearch tool from the pack-age HMMER 3.1b2, other non-coding RNAs (ncRNA) using Rfam 10.1 (7) models and Infernal 1.0 (8), protein-coding genes using Prodigal v2.6.2 (9).

After feature prediction submissions undergo prediction of signal peptides using SignalP (10), transmembrane helices using TMHMM (11), as well as protein family assignments and functional annotation steps. These involve comparing predicted proteins to COG (12) position-specific scoring matrices using RPS-BLAST, comparing to Pfam-A (13) and TIGRfam (14) Hidden Markov Models using HMMER 3.0b2 and HMMER 3.0, respectively (15), comparing to InterPro models using a customized version of InterproScan5 (16) and associating proteins with KEGG Ortholog (KO) terms (17) using LAST (18). Proteomes are associated with KEGG pathways based on KO term assignments to genes, and are associated with MetaCyc pathways (19) based on gene annotations with Enzyme Commission numbers derived from KO terms. In addition, the new genes also go through an IMG term assignment step to be associated with IMG pathways (20). Pathway assertions will then lead to phenotype predictions for genomes (21). In addition, for each isolate proteome, genome vs. genome Bidirectional Best Hits (BBH) and best hits against IMG reference isolates (high quality public genomes) are computed using LAST (18). The latter 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), fused genes (24), biosynthetic clusters (25) and putatively horizontally transferred genes as previously described.

On the metagenome side, while IMG processes both assembled and unassembled sequences for the JGI-generated datasets, only assembled data submissions in FASTA format are accepted for the data generated outside the JGI. Since early 2016, unassembled 454 reads are no longer accepted. Metagenome feature prediction and functional annotation are similar to the process for isolate genomes described above. The differences include the use of hmmsearch for assigning COGs to metaproteomes, the omission of ncRNAs annotations, and several functional annotation steps, such as IMG term assignment and pathway assertions. The detailed standard operating procedure for IMG metagenomes can be found in (26).

Metadata for genomes and metagenomes provided by GOLD v.5 can be accessed through the IMG Genome Browser metadata field selection and in the Genome Detail page as illustrated in Figure 1(i) and 1(ii), respectively. It is worth mentioning that Genome Detail pages now also list related genome publications as shown in Figure 1(ii).

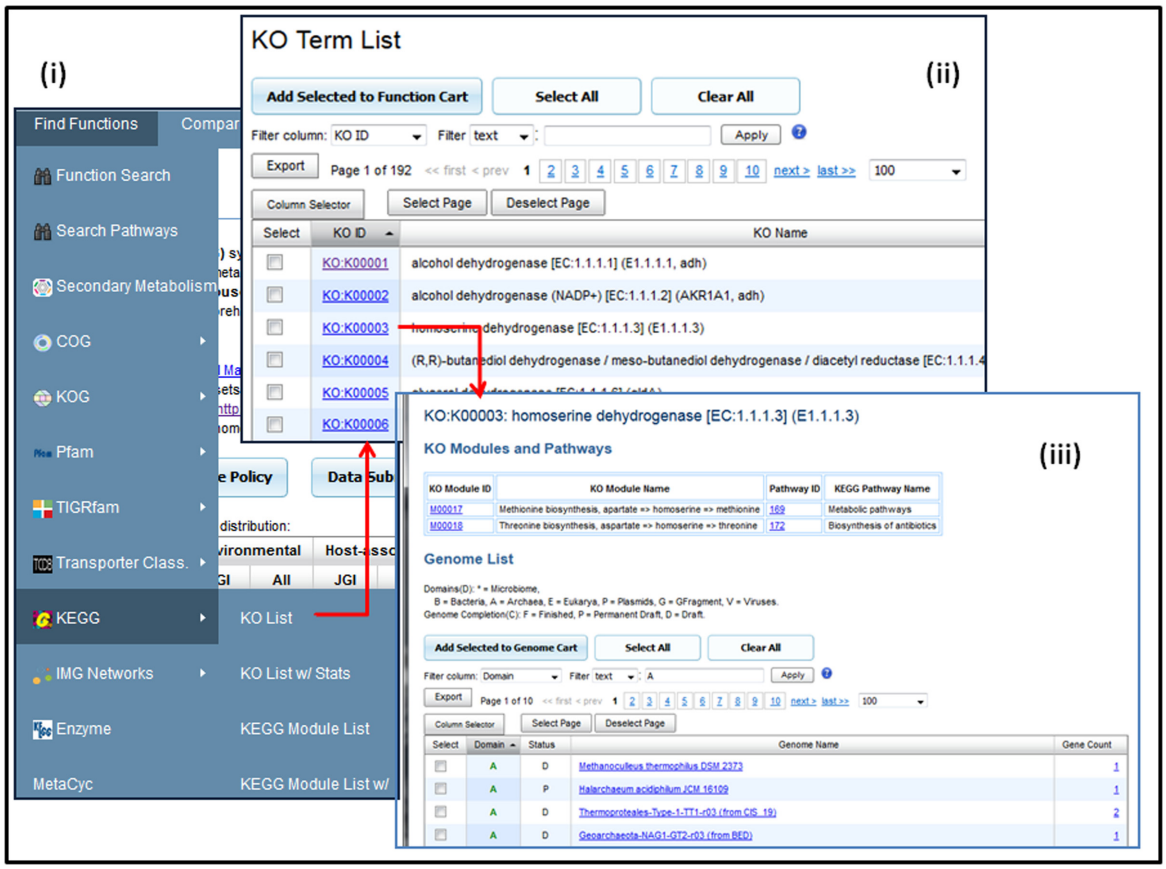

Figure 2. Find Functions with KO List. (**i**) Newly added KEGG functions are shown: KO List, KO List w/ Stats, KEGG Module List and KEGG Module List w/ Stats. (**ii**) KO List shows all KO terms in the IMG database. (**iii**) KO Term detail page for K00003 *homoserine dehydrogenase [EC:1.1.1.3]* shows associated KO modules (M00017, M00018) and pathways (169 172), as well as all genomes and metagenomes with genes annotated with this KO term.

DATA CONTENT

Genomics data and Microbiome samples

Ever since the first release of IMG in March 2005, its data warehouse content has continuously experienced exponential growth. Most of the IMG genomes and metagenomes are publicly available (https://img.jgi.doe.gov/m/). However, ∼15% of the isolate genomes and 48% of metagenomes remain private and password protected until the PIs and submitters can publish their research results. These private genomes are only accessible through the 'Expert Review' version of IMG/M ER (https://img.jgi.doe.gov/mer/). The current version of the system (as of July 2016) contains a total of 47 516 (among them 40 894 public) archaeal, bacterial and eukaryotic genomes, which represents an over 300% increase since September 2013 (27). In addition, IMG also includes 5185 (3907 public) viral genomes, 1220 (1192 public) plasmids and 1196 (1192 public) genome fragments, bringing its total content to 55 117 (47 185 public) genome datasets with more than 173 million (153 million public) protein coding genes. IMG content reflects an increasing interest in genomes from uncultured organisms: there are 3189 (1454 public) single cell genomes compared to only 1341 in September 2013 and 2649 (1557 public) genomes extracted from metagenome datasets.

Metagenome datasets in IMG also increased substantially. As of July 2016, there are 11 004 (among them 5735 public) metagenome datasets from 544 (250 public) metagenome studies with over 45.7 billion (35.9 billion public) protein coding genes in IMG compared to only 3328 metagenomes from 460 studies with 19.5 billion genes back in September 2013. Many new metagenome datasets were conducted within existing studies, which explains why the metagenome study numbers did not increase dramatically. About 60% of metagenome datasets in IMG are derived from environmental samples, with another 32% classified as host-associated and ∼8% coming from engineered environments.

Experimentally validated and predicted biosynthetic clusters in IMG are associated with various pathways including KEGG and Metacyc, as well as GOLD metadata about secondary metabolites produced by them. We have seen increased community interest in biosynthetic clusters in the past couple years with many new IMG users claiming this to be their main purpose of using IMG. There is a special IMG datamart (https://img.jgi.doe.gov/abc/) called IMG-ABC (25), dedicated to biosynthetic clusters.

We continue supporting external submissions to IMG through the IMG Submission system. There are currently more than 9000 total external isolate genome submissions. Among these, more than 5000 were submitted be-

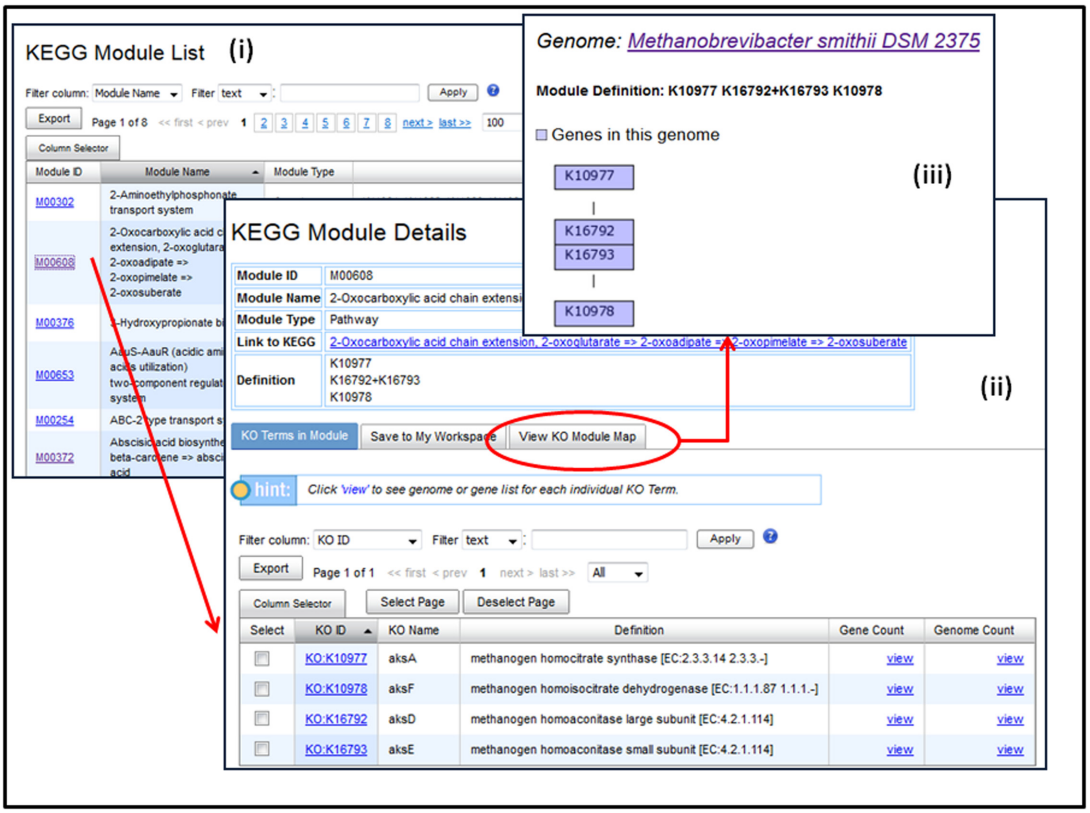

Figure 3. KEGG Module Viewer. **(i)** 'KEGG Module List' in the Find Functions menu shows a list of all KEGG modules in IMG. **(ii)** KEGG Module detail for M00608 *2-Oxocarboxylic acid chain extension, 2-oxoglutarate = > 2-oxoadipate = > 2-oxopimelate = > 2-oxosuberate* shows the module definition and all KO terms in the module. **(iii)** The 'View KO Module Map' feature shows that selected genome *Methanobrevibacter smithii DSM 2375* has genes annotated with all KO terms in this module.

tween September 2013 and July 2016. On the metagenome side, there are around 4800 total external submissions, and among those around 4000 were submitted in that time period. These external metagenome submissions account for around 50% of all metagenome submissions made between September 2013 and July 2016.

Omics data

IMG started to include proteomics datasets back in 2009. The organization and analysis of proteomic data in IMG has been described in (28). There are now 10 (nine public) transcriptome studies with 92 (82 public) datasets in IMG (as of July 2016):

- Transcription Profiling of the Model Cyanobacterium Synechococcus sp. Strain PCC 7002 study with 11 datasets.
- Meta-omics analysis of microbial carbon cycling responses to altered rainfall inputs in native prairie soils study with 19 datasets.
- Sequencing the transcriptome of cyanobacteria study with 30 datasets.
- Thermo-Regulation of Genes Mediating Motility and Plant Interactions in Pseudomonas syringae study with 12 datasets.
- Corynebacterium glutamicum codon depletion transcriptome study with six datasets.
- Two bacteroides fragilis 610 transcriptome studies with one dataset each.
- Two bacteroides fragilis 638R transcriptome studies with one dataset each.

IMG also started to include metatranscriptomic datasets back in 2012. There are now 5156 (4185 public) datasets across 52 (39 public) RNASeq studies as of July 2016, compared with only 16 such studies in 2013. IMG/M contains metatranscriptomic datasets from a very rich variety of samples including marine, estuarine and freshwater microbial communities, communities from forest soil, peatland and rhizosphere, microbial communities from different bioreactors, etc.

DATA ANALYSIS

Data querying, visualization and comparative analysis can be performed using the IMG User Interface (UI) (https://img.jgi.doe.gov/m/). Most IMG UI features have remained unchanged since September 2013 and are summarized only briefly, while more detailed description is provided to new functions below.

The UI **Main** page has a summary of IMG's content listing the counts of genomes for bacteria, archaea, eukarya,

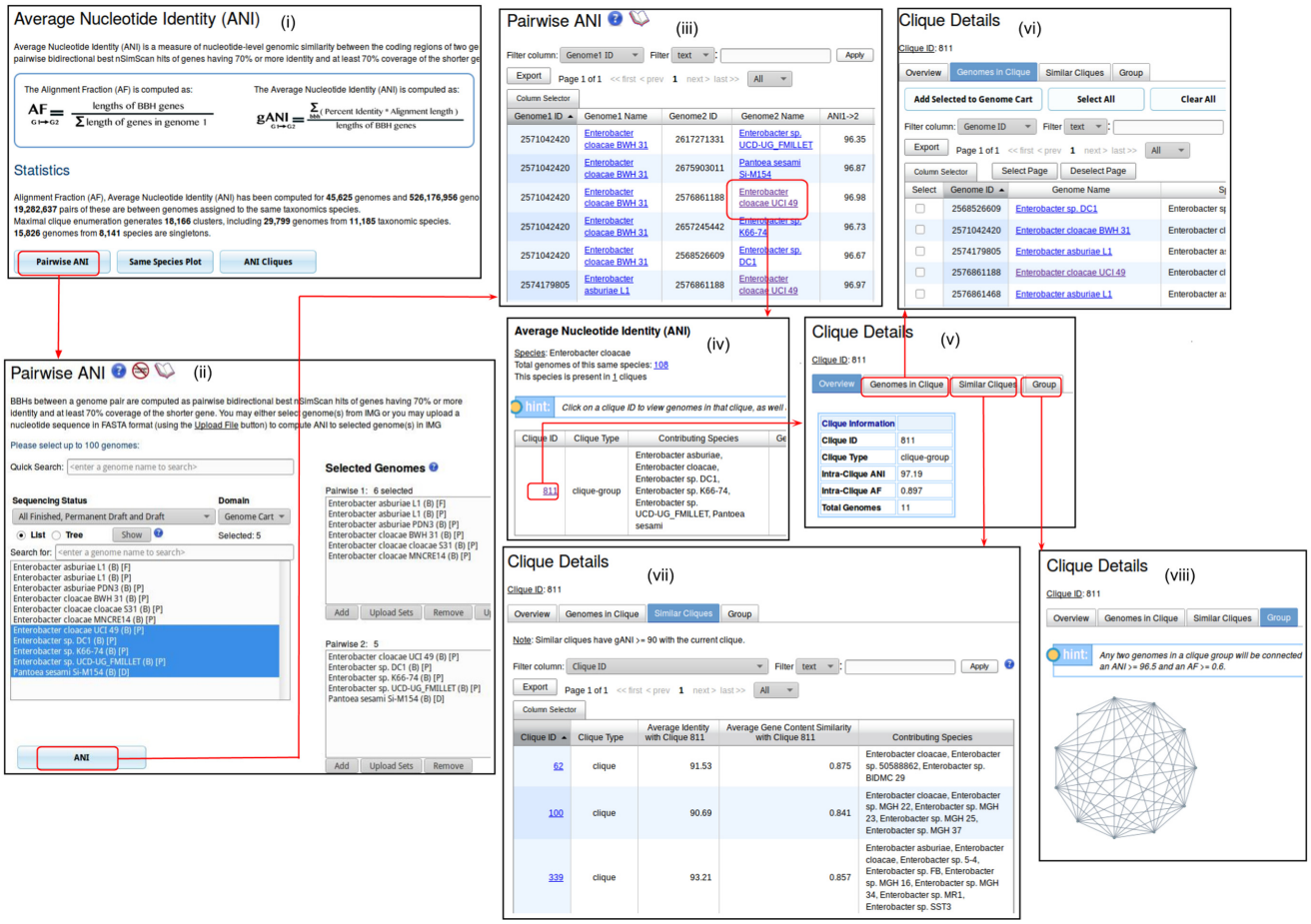

Figure 4. Pairwise ANI tool. **(i)** ANI's landing page in IMG is found under menu item Compare Genomes -> Avg Nucleotide Ident. **(ii)** Pairwise ANI genome selection page, only isolate genomes can be selected. **(iii)** The results of the selected genomes pairwise comparisons. **(iv)** The ANI details on the genome's detail page. **(v)** A clique details. **(vi)** All the genomes that belong to the given clique. **(vii)** A list of similar cliques to the given clique. **(viii)** A graphical representation of the clique group.

plasmids, viruses, genome fragments and metagenomes. More details can be found from **IMG Statistics**. Users who are interested in using IMG content in their publications should check **Data Usage Policy**.

The **Find Genomes** menu allows users to browse or search existing genomes in the IMG database, while **Deleted Genomes** function lists deprecated genomes and metagenomes that are no longer available for analysis in IMG. The **Find Genes** menu enables gene search and gene content-based comparison of genomes. The 'Phylogenetic Profiler for Single Genes' allows gene search in a query genome with and/or without homologs in other genomes of interest, and 'Phylogenetic Profiler for Gene Cassettes' identifies genes co-located in a query genome and in other genomes of interest (23,29).

The **Find Functions** menu provides searching and browsing capabilities for protein families such as COGs, Pfams and TIGRfams, and functional families such as enzymes across isolate genomes and metagenomes. We have recently extended Find Functions tool to enable browsing of KO terms and KEGG modules, as illustrated in Figure 2(i). Select 'KO List' option will lead to a list of KO terms in the IMG database as shown in Figure 2(ii), while 'KO List w/ Stats' will lead to the same list but with additional genome and metagenome count information. Clicking on KO ID will lead to the detailed description of this KO term with associated KO module and KEGG pathway information as well as the list of genomes with genes annotated with this KO term (see Figure 2(iii)).

KEGG Module List function (in Figure 3(i)) lists all KEGG modules in the IMG database as illustrated in Figure 3(i). Clicking on any Module IDs will lead to the detailed description of particular modules as shown in Figure 3(ii). The new 'View KO Module Map' option provides a view of the KEGG module map and enables analysis of the distribution of genes from selected genome(s) across the module (see Figure 3(iii)).

We have also extended the KEGG Module Viewer to biosynthetic clusters listing KEGG Modules associated with the genes in biosynthetic cluster and displaying gene count for each of these modules. For example, predicted biosynthetic cluster 160349372 of *Methanobrevibacter smithii DSM 2375* has two genes connected to KEGG Module M00028 as shown in the KEGG detail panel.

Figure 5. ANI Same Species Plot. **(i)** Tool found under menu Compare Genomes -> Avg Nucleotide Ident. -> Same Species Plot. (ii) Example of comparing some Enterobacter and Pantoea.

The function 'Find missing enzyme' introduced in 2009 (30) has been extended to narrow down the potential genome selections, and to enable search for potentially missing enzymes in more than 400 KEGG pathways. More detailed description of the new features with examples can be found in (31).

The **Compare Genomes** menu provides many comparative analysis tools including Genome Statistics, Synteny Viewers, Phylogenetic Distribution, Distance Tree, Function Profile, Genome Clustering, Best Homologs and Phylogenetic Marker COGs (28,30), as well as newly added Average Nucleotide Identity (ANI) function (22).

ANI is a measure of nucleotide level similarity between the coding regions of two genomes, which can be used to infer species relationship and divergence (22). To compute the ANI and Alignment Fraction (AF) between two genomes, the nucleotide sequences of protein-coding genes of genome A and genome B are compared using the high performance similarity search tool, NSimScan (http://www.scidm.org/). The results are then filtered to retain bidirectional best hits (BBHs) with at least 70% sequence identity over at least 70% of the length of the shorter sequence in each BBH pair. The ANI of genome A to genome B is defined as the sum of the percent identity multiplied by alignment length for all BBH's, divided by the sum of the lengths of the BBH genes. The AF is computed by dividing the sum of the lengths of all BBH genes by the sum of the length of all the genes in genome A. This computation is performed separately in both directions: from Genome A to Genome B and from Genome B to Genome A, between multiple IMG genomes, or against an external genome of choice using the 'Pairwise ANI' tool (Figure 4).

The ANI,AF values are precomputed in IMG between all high quality genomes. As of July 2016 precomputed values are available for 45 625 genomes and 526 176 956 genome pairs. Since ANI,AF values are a reflection of the phylogenetic distance between genomes, the genomes within a taxonomic species are expected to exhibit high ANI and AF pairwise values. The 'Same Species' tool (Figure 5) allows to plot the ANI,AF values computed between all genomes pairs of a particular species to visually explore the relationship between these genomes and identity those that might be misidentified or contaminated. It also allows overlaying the plotted values of several species to provide the ability to compare the extent of genomic conservation among their members.

IMG also uses ANI to provide the ability to explore how genomes group based solely on their genomic relatedness, irrespective of their existing taxonomy. Maximal Clique Enumeration (MCE), a form of complete linkage clustering (32) is applied to genome pairs linked with species-level AF and ANI (22). MCE generates two types of clusters: (i) 'cliques' which are complete graphs, where each vertex represents a genome and every vertex is linked to every other vertex and (ii) 'clique-groups' which are formed when multiple maximal clique configurations are possible, resulting in cliques having genomes in common. Those genomes that do not display the required ANI and AF values to any other genome are denoted as singletons. The 'ANI Cliques' tool (Figure 6) enables exploration of precomputed cliques, clique-groups and singletons by clusters, species or taxonomic hierarchy and also provides visual representation of the clique-groups.

The **OMICS** menu provides three types of 'omics' data in the IMG data warehouse: Protein, RNASeq and Methyla-

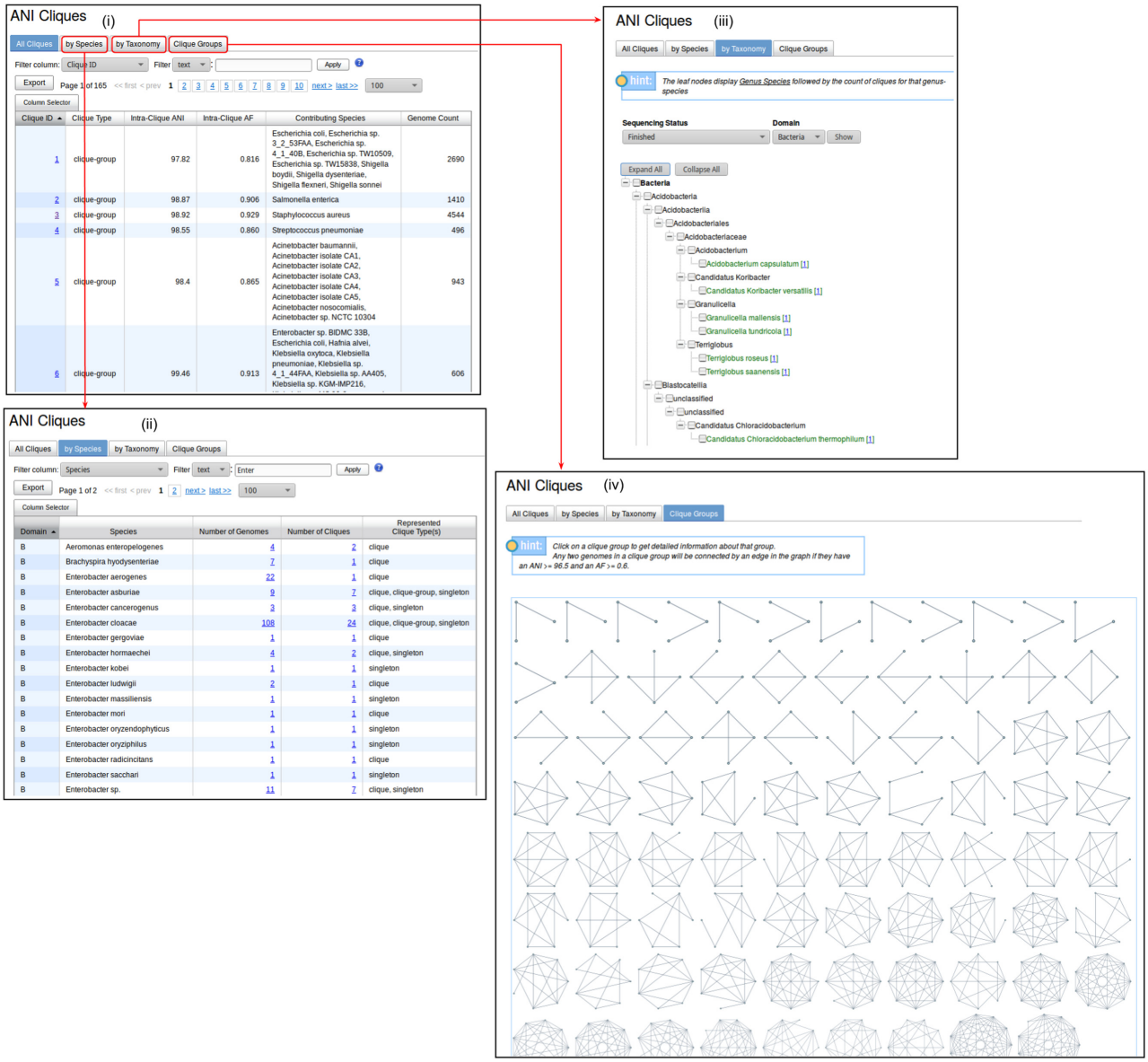

Figure 6. ANI Cliques. **(i)** List of all clique types in IMG. **(ii)** All cliques grouped by species. **(iii)** All cliques grouped by taxonomy. **(iv)** Cliques groups.

tion. Though no new functions have been added here, there are more public datasets available for analysis.

The **MyIMG** menu in the public version of IMG only allows users to set browser preference. However, in the 'Expert Review' version (https://img.jgi.doe.gov/mer/) of IMG it provides additional functions such as the listing of MyIMG gene annotations (30,31) and **MyJob** tracking of jobs submitted for computation on demand (33). In addition, MyIMG Home includes the new IMG Group feature, which enables IMG users to form user groups for sharing genomes and workspace datasets among group members. This new effort to facilitate user collaboration and community annotation is described in more detail in (31).

The **Data Marts** menu provides a quick link to various IMG data marts such as IMG ABC (25) and IMG HMP (34), with associated reading materials under the **Help** menu. IMG continues to provide Analysis Cart functions, but the feature has been moved to the top of the window above the menu bar.

Due to the dramatic increase in the number and size of metagenome datasets, it is no longer feasible to perform certain on-the-fly operations such as listing of all the genes assigned to protein families in large sets of metagenomes. Therefore, our recent work on UI development has focused mainly on performance improvement rather than adding more new analysis functions. As part of this development, many analysis features rely on pre-computations of summary statistics (e.g. lists of protein families with counts of genes in isolate genomes and metagenomes in Find Functions menu). Moreover, many metagenomic analysis have been re-routed to rely on the 'computation on demand' features in the IMG Workspace, which enables submission of long computational jobs. Such features are only available in the Expert Review version of IMG, since workspace

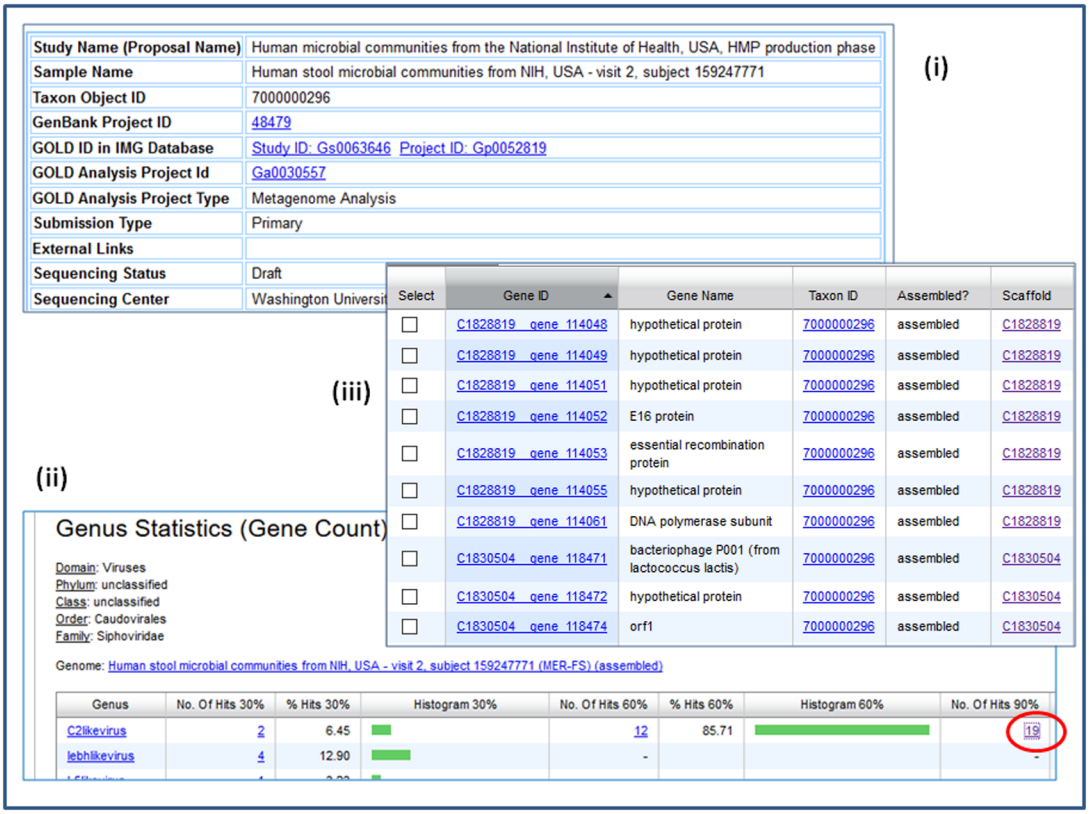

Figure 7. Metagenome binning. (**i**) From a metagenome detail page, a user can scroll down to find the Phylogenetic Distrubution of Genes function. (**ii**) The phylogenetic distribution result shows that the metagenome has 19 genes with more than 90% hits to *C2likevirus* (genus). (**iii**) The 19 genes all came from two scaffolds C1828819 and C1830504, which can be selected and saved to individual workspace scaffold datasets for further analysis.

datasets, computation requests and computation results are defined by individual users and are password protected.

Metagenome binning and genomes from Microbiome samples

During the last two years there is increasing interest in metagenome binning and studies of genomes extracted from metagenome datasets (GFMs). Metagenome bins can be stored in IMG as individual workspace scaffold datasets, and analyzed using many tools, such as function profiles, histograms and phylogenetic distributions. For example, a user can use the Phylogenetic Distribution of Genes function from the metagenome detail page of HMP genome *Human stool microbial communities from NIH, USA - visit 2, subject 159247771* (IMG OID: 7000000296) to identify scaffolds of interests (see Figure 7(i)). From the phylogenetic distribution result table, the user can select *Virus* unclassified by clicking the link, and then can follow the link to unclassified class, *Caudovirales* (order) and *Siphoviridae* (family). The result shows that there are 19 potential genes with more than 90% hits to *C2likevirus* (genus) (Figure 7(ii)). These 19 genes came from two (viral) scaffolds C1828819 and C1830504 (Figure 7(iii)). The user can select to store the two scaffolds as separate workspace scaffold datasets for further analysis.

Most workspace functions now include 'computation on demand' feature allowing submission of computationally intensive analysis jobs, such as computing abundances of protein families across large sets of scaffolds from multiple metagenomes.

IMG users can export nucleotide sequence of selected scaffold set(s) in FASTA format using the **Data Export** function provided in the Workspace Scaffold Sets. Submission of these files for selected scaffold sets as new genomes or metagenomes is also available using the IMG Submission system. In the future additional functionality will enable users to publish their workspace scaffold sets as 'public metagenome bins,' thus making them available to all other IMG users.

CONCLUDING REMARKS AND FUTURE PLANS

The current version of IMG/M (as of 15 July 2016) contains 11 004 (among them 5735 public) metagenome datasets from 544 (250 public) metagenome studies with over 45.7 billion (35.9 billion public) protein coding genes, while new metagenomes both from JGI and from external sources keep being submitted into IMG. During the last year we have initiated an effort to assemble and annotate selected metagenomics and metatranscriptomic datasets available at NCBI's SRA system in order to further increase the diversity of projects integrated into IMG (35). While IMG will continue to provide and further enhance its tools for comparative analysis of unassembled reads, the main emphasis will remain on the analysis of assembled metagenomics data. In this direction, IMG does provide a unique environ-

ment that has been proven extremely effective in a variety of research efforts requiring high quality metagenomic assembly, such as the identification and genomic reconstruction of novel phylogenetic lineages (36), discovery of novel biosynthetic gene clusters (25), identification of alternative genetic codes (37), uncovering gaps in amplicon-based detection of microbial diversity (38) and discovery of novel viruses (39).

We are working on consolidation of our feature prediction and functional annotation pipelines for both isolate genomes and metagenomes. This effort will allow for better data consistency across genomes and metagenomes as well as more streamlined operations. Moving forward, we expect that the sheer size of the data will further reduce the scope and scale of comparative analyses available on the fly. Our work in the past few years have been focused on exploring new data management techniques and effective data analysis methods (40). We expect the efforts to be continued into the future.

FUNDING

Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, US Department of Energy [DE-AC02-05CH11231]; National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy [DE-AC02-05CH11231]; US National Institutes of Health Data Analysis and Coordination Center [U01-HG004866 to IMG/M-HMP]. Funding for open access charge: Joint Genome Institute and Lawrence Berkeley National Laboratory.

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. O'Leary,N.A., Wright,M.W., Brister,J.R. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
3. Reddy,T.B.K., Thomas,A., Stamatis,D., Bertsch,J., Isbandi,M., Jansson,J., Mallajosyula,J., Pagani,I., Lobos,E. and Kyrpides,N.C. (2014) The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.*, **43**, D1099–D1106.
4. 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.
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. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2015) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
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. *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 Beck,E. (2013) TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.
15. Finn,R.D., Clements,J., Arndt,W. *et al.* (2015) HMMER web server: 2015 update. *Nucleic Acids Res.*, **43**, W30–W38.
16. Jones,P., Binns,D., Chang,H-Y *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. *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., Ratner,A., Palaniappan,K., Szeto,E., Huang,J., Reddy,T.B.K., Cimermanc̆ic̆,P., Fischbach,M.A. *et al.* (2015) IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *mBio*, **6**, doi:10.1128/mBio.00932-15.
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. Markowitz,V.M., Chen,I.A., Palaniappan,K., Chu,K., Szeto,E., Pillay,M., Ratner,A., Huang,J., Woyke,T., Huntemann,M. *et al.* (2013) IMG 4 version of the integrated microbial genomes comparative analysis system, *Nucleic Acids Res.*, **42**, D560–D567.
28. Markowitz,V.M., Chen,I.A., Palaniappan,K., Chu,K., Szeto,E., Grechkin,Y., Ratner,A., Biju,J., Huang,J., Williams,P. *et al.* (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.*, **40**, D115–D122.
29. Romosan,A., Shoshani,A., Wu,K., Markowitz,V.M. and Mavromatis,K. (2013) Accelerating gene context analysis using bitmaps. *Proc. of the 25th Int. Conference on Scientific and Statistical Database Management (SSDBM 2013)*.
30. Markowitz,V.M., Mavromatis,K., Ivanova,N.N., Chen,I.A., Chu,K. and Kyrpides,N.C. (2009) IMG ER: a system for microbial annotation expert review and curation. *Bioinformatics*, **25**, 2271–2278.

31. 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.
32. Bron,C. and Kerbosch,J. (1973) Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, **16**, 575–577.
33. Markowitz,V.M., Chen,I.A., Chu,K., Szeto,E., Palaniappan,k, Pillay,M., Ratner,A., Huang,J., Pagani,I., Tringe,S. *et al.* (2013) IMG 4 version of the integrated metagenomes comparative analysis system, *Nucleic Acids Res.*, **42**, D568–D573.
34. Markowitz,V.M., Chen,I.M., Chu,K., Szeto,E., Palaniappan,K., Jacob,B., Ratner,A., Liolios,K., Pagani,I., Hunteman,M. *et al.* (2012) IMG/M-HMP: a metagenome comparative analysis system for the human microbiome project. *PLoS ONE*, **7**, e40151.
35. Kyrpides,N.C., Eloe-Fadrosh,E.A. and Ivanova,N.N. (2016) Microbiome Data Science: understanding our microbial planet. *Trends Microbiol.*, **24**, 425–427.
36. Eloe-Fadrosh,E.A., Paez-Espino,D., Jarett,J., Dunfield,P.F., Hedlund,B.P., Grasby,S.E., Brady,A.L., Dong,H., Briggs,B.R., Li,W-J *et al.*2016) Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat. Commun.*, **7**, 10476.
37. Ivanova,N., Schwientek,P., Tripp,J.J., Rinke,C., Pati,A., Huntemann,M., Visel,A., Woyke,T., Kyrpides,N.C. and Rubin,E.M. (2014) Stop codon reassignments in the wild. *Science*, **344**, 909–913.
38. Eloe-Fadrosh,E.A., Ivanova,N.N., Woyke,T. and Kyrpides,N.C. (2016) Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat. Microbiol.*, **1**, 15032.
39. Paez-Espino,D., Eloe-Fadrosh,E.A., Pavlopoulos,G., Thomas,A.D., Huntemann,M., Pati,A., Rubin,E., Ivanova,N.N. and Kyrpides,N.C. (2016) Uncovering earth's virome. *Nature*, **536**, 425–430.
40. 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.