D446–D456 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 27 October 2016*
doi: 10.1093/nar/gkw992

Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements

Supratim Mukherjee[1], Dimitri Stamatis[1], Jon Bertsch[1], Galina Ovchinnikova[1], Olena Verezemska[1], Michelle Isbandi[1], Alex D. Thomas[1], Rida Ali[1], Kaushal Sharma[1], Nikos C. Kyrpides[1,2,*] and T. B. K. Reddy[1,*]

[1]Prokaryotic Super Program, DOE Joint Genome Institute, Walnut Creek, 94598 CA, USA and [2]Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

Received September 20, 2016; Revised October 11, 2016; Editorial Decision October 12, 2016; Accepted October 19, 2016

ABSTRACT

The Genomes Online Database (GOLD) (https://gold.jgi.doe.gov) is a manually curated data management system that catalogs sequencing projects with associated metadata from around the world. In the current version of GOLD (v.6), all projects are organized based on a four level classification system in the form of a Study, Organism (for isolates) or Biosample (for environmental samples), Sequencing Project and Analysis Project. Currently, GOLD provides information for 26 117 Studies, 239 100 Organisms, 15 887 Biosamples, 97 212 Sequencing Projects and 78 579 Analysis Projects. These are integrated with over 312 metadata fields from which 58 are controlled vocabularies with 2067 terms. The web interface facilitates submission of a diverse range of Sequencing Projects (such as isolate genome, single-cell genome, metagenome, metatranscriptome) and complex Analysis Projects (such as genome from metagenome, or combined assembly from multiple Sequencing Projects). GOLD provides a seamless interface with the Integrated Microbial Genomes (IMG) system and supports and promotes the Genomic Standards Consortium (GSC) Minimum Information standards. This paper describes the data updates and additional features added during the last two years.

INTRODUCTION

The Genomes OnLine Database (GOLD) is a data management system for the curation and visualization of sequencing projects pursued around the world. Ever since its first release (1) and subsequent updates (2–6), GOLD has been a pioneering centralized public resource for monitoring sequencing projects and their associated metadata, promoting comparative analyses and groundbreaking discoveries through biological translation of sequence data (7–9). An important component in the analysis and interpretation of sequence data is the availability of high quality and accurate metadata. With the increasing amounts of sequence data released in the public domain, without an accurate account of metadata any comparative analysis will be less meaningful and prone to misinterpretations. GOLD carries the critical role in providing manually curated metadata from the literature and various other resources, enabling more efficient comparative analysis of sequence data. The data are provided to the community through a login free, user-friendly web interface. Thus, GOLD serves as the curated catalogue of world wide sequencing projects as well as a central resource of curated metadata records.

The decreasing sequencing costs coupled with continuous improvements in sequencing and longer read technologies are driving the continuation of doubling the amount of data produced every seven months over the past 10 years (10). These technological developments have enabled several large scale sequencing efforts including the Human Microbiome Project (HMP) (11), 1000 Fungal Genomes (12), Genomic Encyclopedia of Bacteria and Archaea (13–15) and others. More recently, single cell genomics from environmental samples (16), i.e. sequencing the genome from a single cell, and genomes reconstructed from metagenomes have significantly increased our ability to sequence phylogenetically diverse and hitherto uncultured organisms. A growing number of Sequencing Projects in GOLD during the last few years are from genomes of uncultured organisms with these approaches leading to characterizing the genome of several new phyla (17–20).

GOLD serves as the entry point for all the projects submitted for analysis to the Integrated Microbial Genomes (IMG) data management systems (21,22) and ensures that projects are correctly defined along with their necessary

*To whom correspondence should be addressed. Tel: +1 925 296 5768; Fax: +1 925 296 5850; Email: tbreddy@lbl.gov
Correspondence may also be addressed to Nikos C. Kyrpides. Tel: +1 925 296 5718; Fax: +1 925 296 5666; Email: nckyrpides@lbl.gov
Present address: Alex D. Thomas, Department of Environmental Science, Policy, & Management, University of California Berkeley, Berkeley, CA 94720, USA.

metadata before being passed on to the IMG pipelines for annotation (23,24). GOLD also supports the International community-driven standards of the Genomics Standards Consortium (25) and is fully compliant with its recommendations for Minimum Information about any (x) Sequence (MIxS) standards (26). Documenting and organizing metadata in a centralized database that serves both as a worldwide catalogue and as an entry point for annotation and comparative analysis, has been shown to be very convenient for the users (27). Documented metadata in GOLD can be readily accessed to create genome reports for journals such as Standards in Genomic Sciences (28).

The increase in the number of sequencing projects worldwide and the diversity of research studies coupled with novel and sophisticated analysis approaches users are applying for their data, is driving the need for a more flexible project and metadata management system. In addition, there is a constant need for new metadata fields, intuitive search mechanisms and new approaches to data analysis. These are some of the main requirements that have driven the development of GOLD since its last major update two years ago. An Advanced Search feature, custom metadata package for biogas reactor and support for NCBI's data imports (29) are few of the major updates described in this paper.

GOLD OVERVIEW AND CURRENT STATUS

GOLD data structure

GOLD is based on a four level classification system to clearly distinguish and organize different entities for better tracking and metadata management. The four levels are Study, Biosample or Organism, Sequencing Project (SP) and Analysis Project (AP). Each level holds a unique set of metadata fields and is connected to one or more levels in a hierarchical fashion.

GOLD Study

A Study represents the top level in GOLD's four level organization scheme (Figure 1). Studies broadly represent the umbrella project or the overall goal of a research proposal that a researcher sets out to explore. A GOLD Study can consist of any number of genome Sequencing Projects, e.g. the HMP under which several hundred genome projects were completed (11). While the majority of the Studies in GOLD involve either isolate genome or metagenome SPs, there are several cases where multiple sequencing strategies (such as isolate genome, single-cell genome, transcriptome, metagenome, metatranscriptome and others) are pursued under a single Study. Currently, 26 117 Studies are reported in GOLD. Since the last update, the number of Studies has increased by approximately 7000.

GOLD Biosample

The GOLD Biosample corresponds to the physical material collected from the environment, and by effect represent the descriptor of the metadata that is associated with an environmental sample. GOLD's Biosample allows the connection of multiple Sequencing Projects to a single physical sample (e.g. a metagenome, a metatranscriptome and several single cell genome projects may be originating from the same environmental material). Metadata associated with GOLD Biosamples include data such as the description of the ecosystem, habitat, place of isolation etc. Rich metadata facilitates comparative analysis as well as helping to drive new discoveries through the availability of specific and accurate metadata. For example, having fine-grained metadata was instrumental in mapping the biogeography of marine viral sequences to different ecological regions of the ocean such as estuaries, coastal waters, coastal sediments and to different depths like surface water, deep ocean, hydrothermal vents and more (7). GOLD's definition of Biosample is conceptually different from the NCBI's BioSample that encompasses both organism and environmental samples. While a GOLD Biosample may be associated with more than one Sequencing Project, a separate BioSample is required for each sequencing project submitted to NCBI. As an example, the chromosome and the plasmid of a single organism may be under a single NCBI BioProject (e.g. PRJNA48991) but under two different NCBI BioSample IDs. Overall, 174 of the GOLD's Biosamples are associated with more than one Sequencing Project, connecting different sequencing strategies to the same original sample. Currently there are 15 887 Biosamples in GOLD distributed across Environmental (47%), Host-associated (35.7%) and Engineered (17.3%) ecosystems.

GOLD Organism

An Organism in GOLD corresponds to any living biological material (virus, bacteria, fungus, plant or animal) that is associated to a Sequencing Project. A GOLD Organism may be cultured or uncultured (such as single cells) and can be linked to more than one Sequencing Project. For example, one organism may be sequenced by different research groups to address similar or different research questions. There are two main sources for new Organism entries in GOLD. One is through the regular addition of a new Sequencing Project, where a new Organism has to be entered (if not already available in the system). The second is a mass import of cultured organisms from StrainInfo (30) most of which are not yet associated with a Sequencing Project. These organisms are readily available for researchers to choose from while creating a new Sequencing Project in GOLD. Currently, there are 239 100 Organisms in GOLD from which 76 759 are associated with 81 289 Sequencing Projects. Using the strain mapping information provided from StrainInfo (30), equivalent strains from different culture collections are mapped to a single Organism in GOLD.

One important metadata field associated with the Organism in GOLD is the information on whether an Organism represents a type strain (31). A type strain is the strain used when the species was first described. Authors reporting a new species usually also designate the type strain of the species. Type strains are maintained in at least two independent culture collections and serve as reference point for a species. As per 'International Code of Nomenclature of Prokaryotes' (32) these are referred to as the 'nomenclatural type of the species'. GOLD acquires type strain infor-

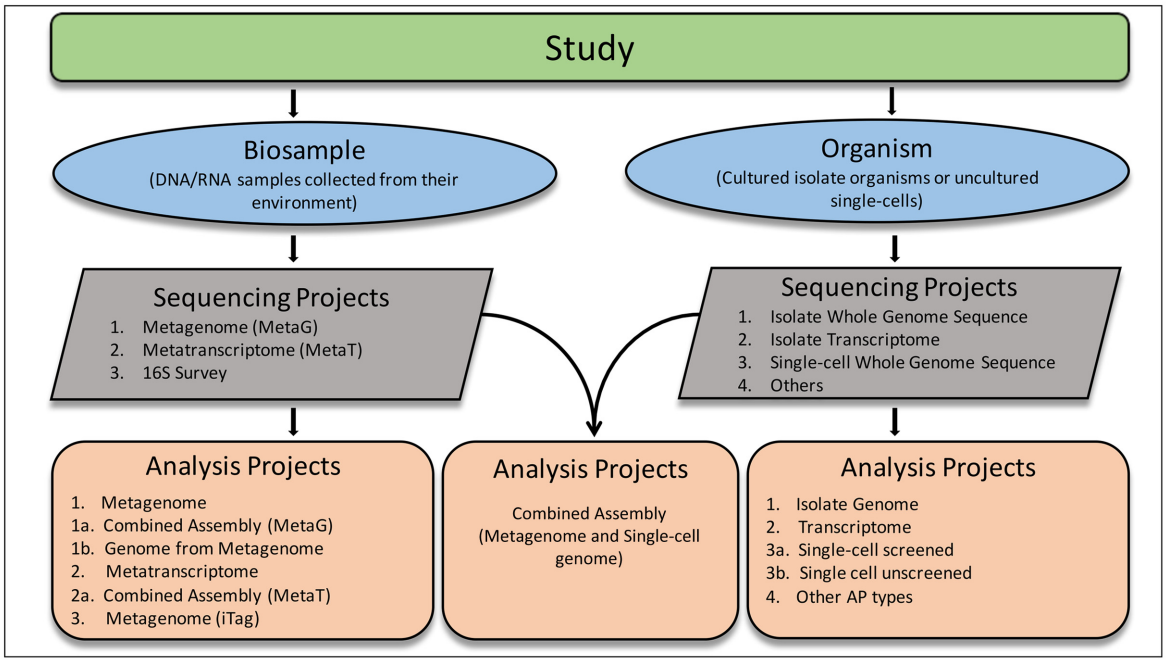

Figure 1. Four level classification system of the Genomes OnLine Database (GOLD) database. A Study lies at the helm of the project classification system in GOLD and is comprised of either Biosamples or Organisms, which in turn form their respective Sequencing Projects. The assembly and analysis of GOLD Sequencing Projects culminate into Analysis Projects, which are passed on to the Integrated Microbial Genomes (IMG) data management and analysis system.

mation through a collaboration with NamesforLife (www.namesforlife.com), publicly available information at culture collections and the literature. GOLD currently has 11 096 type strains with Sequencing Projects associated with 3321. A total of 186 type strains have more than one Sequencing Project. A total of 10 809 of the types strains in GOLD also have a digital object identifier (DOI), which uniquely identifies each GOLD Organism and can be used as a direct reference in publications or online platforms.

GOLD's Organism classification conform to NCBI's taxonomy conventions (33). Several taxonomy specific fields like genus, species, strain, NCBI taxonomy id and phylogeny are mandatory for registering a new organism in GOLD. Additional organism-specific information such as type strain, culture collection ID, Gram stain, phenotype, motility, oxygen requirement, biotic relationship and others are also available at the Organism level, along with other environmental metadata. Figure 2 shows the geographic distribution of GOLD's Biosamples and Organisms that were collected from different parts of the globe. A total of 73% of Biosamples and 10% of the Organisms that are associated with sequencing projects have geographic location information in GOLD.

GOLD Sequencing Project

A GOLD Sequencing Project represents the sequencing output from an individual Organism or Biosample. Recent developments in sequencing technologies have resulted in a wide array of sequencing strategies that can be applied to a biological or environmental sample. As such, several different types of Sequencing Projects are available in GOLD, ranging from isolate WGS, single cell sequencing, targeted gene surveys, transcriptomes, metagenomes, metatranscriptomes and more (Table 1). Currently GOLD has 97 212 SPs with 71 295 WGS projects spread across bacteria (81.3%), eukaryotes (10.5%), virus (6.5%) and archaea (1.7%) followed by metagenome and metatranscriptome projects. An interesting observation comparing the metadata fields from GOLD Sequencing Projects is shown in Figure 3. In terms of the total number of Sequencing Projects, Broad Institute leads the way; however, over the years, the Joint Genome Institute (JGI) has sequenced a significantly diverse selection of organisms (in terms of unique genus and species) than any other sequencing center.

GOLD Analysis Project

Analysis Project represents the data processing and analysis methods applied to individual Sequencing Projects, specifically detailing the assembly and annotation approaches. A GOLD AP is required for submitting a data set to IMG for analysis. Each Sequencing Project in GOLD can have one or more APs associated with it. For example, a user can apply multiple assembly techniques to the same raw sequence data (i.e. same Sequencing Project) and have them annotated in IMG. However, each AP can drive a single submission to IMG, so that a one-to-one relation is preserved between a GOLD AP and an IMG Taxon OID (i.e. data set). Only one annotated AP can be part of IMG's reference data set and is designated as primary AP. The primary AP denotes the default assembly and annotation of a Sequencing Project. A reanalysis AP is created when a user has reassembled or reannotated a data set and would like

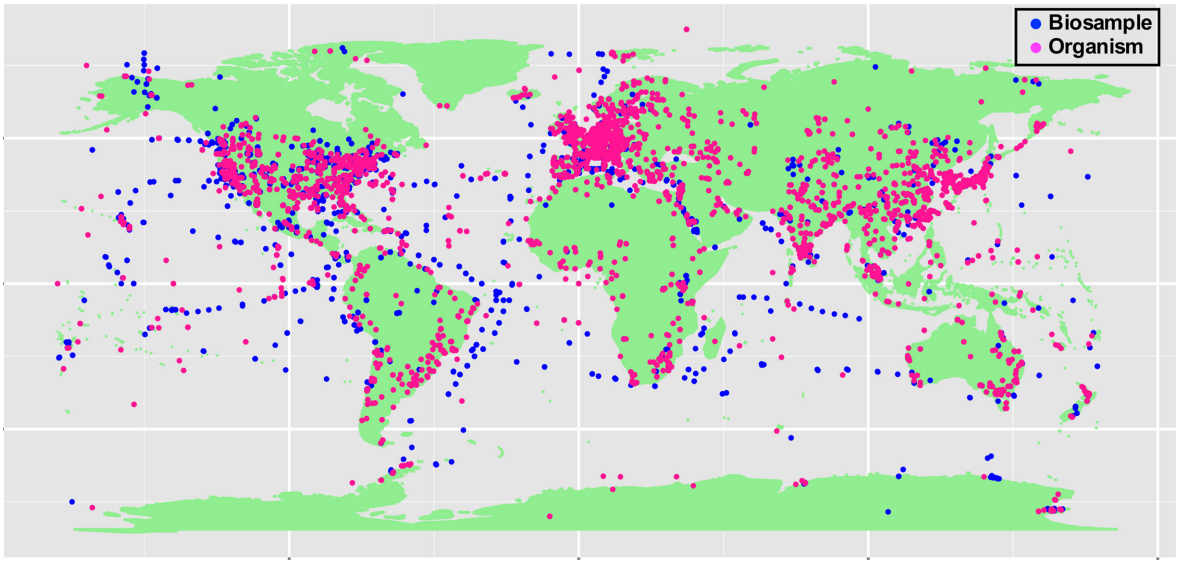

Figure 2. Geographic Distribution of GOLD Biosamples and Organisms. Organism location of isolation is marked in pink while Biosample location of collection is denoted with blue dots.

Table 1. Sequencing Project types in GOLD

Sequencing Strategy	No. of SPs
Whole Genome Sequencing	78 246
Metagenome	13 417
Metatranscriptome	2320
Transcriptome	1595
Genome fragments	1185
Targeted Gene Survey	198
Methylation	66
Transposon Mutagenesis	60
Chloroplast	52
Others	69

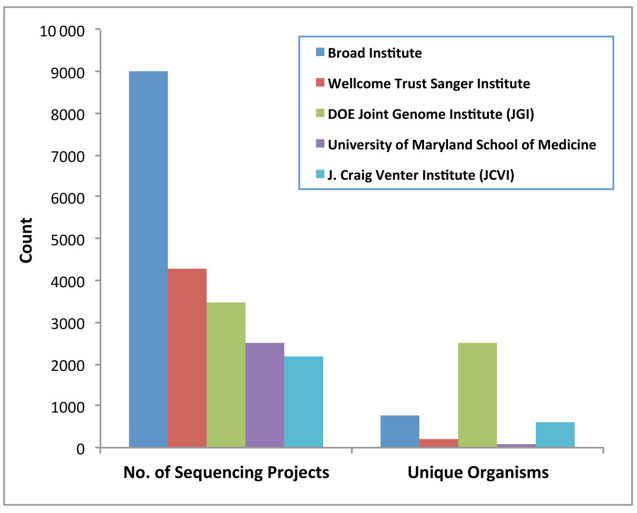

Figure 3. Sequencing projects across top sequencing centers. Comparison of the total number of GOLD Sequencing Projects and corresponding unique Organisms (in terms of genus and species names) per sequencing center. Color of the bars represent each sequencing center as shown in the legend. Unique Organisms are defined as unique species names.

to compare the results with those of the primary AP, which already exists in IMG. There is no limit on the number of reanalysis APs that can be issued from a user. The prerequisite for creating a reanalysis AP is that a primary AP must already exist. A user can also convert a reanalysis AP to a primary AP. Different metadata fields of an AP gather information about the data processing methods that differentiate one AP from another. Currently there are 78 579 Analysis Projects in GOLD, which is more than twice the number of APs since our last release. A total of 68% of the APs have been submitted to IMG and have an IMG Taxon OID. Over 56 000 Analysis Projects are for individual genomes, 92% of which have a GenBank ID (34).

Table 2 lists the different Analysis Project types in GOLD. Driven by the absence of appropriate culturing techniques and improvement in bioinformatics methods to assemble environmental sequences, there has been a recent increase in the number of partial or near-complete reconstruction of genomes from metagenomes (GFMs) (35). Accordingly, GOLD has observed a marked increase in the number of GFM APs. Since GFMs are not direct product of sequencing an individual organism (either an isolate or a single cell), but rather computationally derived from a metagenome, they are not directly connected to an SP. Instead, they are connected to an AP of a metagenome SP. Single-cell ge-

nomics is another example where uncultured microbes were isolated from environmental samples (36). While sequence contamination is common in isolate genomes (37), single amplified genome extraction, being a nascent technology, is equally prone to contamination and often requires extensive decontamination procedures (38). Thus, to differentiate APs that have gone through a thorough contamination check from those that have not, GOLD has two different kinds of single-cell APs, namely, single cell analysis (screened) and single-cell analysis (unscreened). Transcriptome, metatranscriptome, 16S based targeted metagenome assembly and an expanded range of combined assembly APs (discussed later) make up the remaining different types of Analysis Projects in the current version of GOLD.

GOLD DATA SOURCES

Data in GOLD are imported from three main sources: (i) projects deposited by users, (ii) projects imported from public resources like NCBI's BioProject and BioSample databases (39) and (iii) projects sequenced at JGI. User entered data are regularly monitored for data accuracy and consistency. The later two are imported into GOLD using semi-automatic import processes after manual checks. Out of the total 97 212 public Sequencing Projects in GOLD, 13 140 were entered by users, 24 923 are JGI projects and 59 149 were imported from external resources.

GOLD METADATA STATISTICS

The four project levels of GOLD have a total of 312 metadata fields out of which 58 are represented by controlled vocabularies (CV) and the remaining are free text fields (Table 3). The 58 CVs comprise a total of 2067 CV terms. At all four levels of GOLD around 45 metadata fields are mandatory fields. The most well populated fields across metagenome projects/biosamples are ecosystem classification, habitat, geographic location, latitude, longitude, etc. Among isolate Organism based Sequencing Projects, Organism specific fields such as taxonomy information (genus, species, strain, NCBI taxonomy id, phylogeny) and Organism specific metadata such as Gram stain, cell shape, color, isolation site and habitat are commonly populated fields. Organisms identified as type strains tend to posses more metadata in GOLD. Organisms associated with specific Studies list metadata relevant to that initiative. For example, HMP project associated Organisms often list host name, host body site, subsite, body product and disease.

GOLD FEATURE AND DATA UPDATES SINCE LAST RELEASE

Change has always been constant in GOLD as it continues to develop and evolve over the years to keep up with the growing demands of the larger scientific community. Since the last release (6) there were several key updates to the database. New features were added for better data organization, increased efficiency and to make it more intuitive and user-friendly. GOLD also grew significantly with respect to the volume of data that was incorporated over the last couple of years. Below we list some of the major updates both in terms of new features and data since the last release.

New features

A select list of new features added to GOLD since our last release are Bifurcation of Organism and Biosample, Advanced Search, Metadata Packages and New Combined Assembly Analysis Project Types.

Bifurcation of Organism and Biosample

As described earlier, a GOLD Biosample refers to a physical sample from which genetic material (DNA or RNA) is isolated for subsequent Sequencing Projects. In the previous version of GOLD, a Biosample entity was defined/created for environmental samples as well as organisms including isolate and uncultured single cell organisms. Traditionally environmental samples were pursued for metagenome and metatranscriptome projects. In some cases, single cells were isolated from environmental samples for genome sequencing. Having a Biosample entity both for environmental samples and organisms created some confusion among our users, with a question why a separate Biosample entity in GOLD is required if all the metadata for a particular organism can be captured and organized at the Organism level itself. Also it puts undue burden on users who enter projects manually. Users were previously required to enter both a Biosample and an Organism if it was not already present in GOLD. To clearly distinguish between environmental samples and organisms, better organize metadata as well as to reduce the data entry burden on our users, we decided to bifurcate the Biosample, as defined in earlier versions of GOLD, into Biosample and Organism entities. As shown in Figure 1, GOLD Biosamples now specifically refer to environmental samples. Organisms will not have a Biosample entry, instead all the metadata is now stored at the Organism level. As a way to support our users and reduce their data entry burden we have added a large number of Organisms from the StrainInfo database (30) to GOLD.

Advanced Search

We implemented the Advanced Search feature to allow users to explore GOLD's different project levels such as Study, Biosample/Organism, Sequencing Projects and Analysis Projects. In earlier versions, one had to perform several iterations of the individual search feature and track those results offline from one search to another. Our current implementation of the Advanced Search feature (Figure 4A) is designed to eliminate those shortcomings. Now a user can apply multiple metadata filters across different levels to explore GOLD. For example, the current advanced search feature enables the search for a list of finished whole-genome sequencing projects with GenBank IDs from Gram positive, aerobic bacteria. As shown in Figure 4B, this advanced search allows searching GOLD by applying six different metadata filtering criteria across three different levels. Search results are organized and presented with hits at all levels, with a clickable link on the number of results. By clicking on the number, a list of corresponding GOLD entries filtered by the complex search criteria outlined above are retrieved. For instance, clicking on Analysis Projects, a list of Analysis Projects from Advanced

Table 2. Types of different Analysis Projects in GOLD

Type of Analysis Project	AP count
Genome Analysis	56 386
Metagenome Analysis	10 814
Metatranscriptome mapping	5827
Genome from Metagenome	1713
Metatranscriptome Analysis	1684
Single Cell Analysis (screened)	1185
Single Cell Analysis (unscreened)	840
Combined Assembly	109
Transcriptome Analysis	12
Targeted Gene Survey	9

Table 3. Number of metadata and CV fields in GOLD

GOLD Classification Level	No. of fields	No. of CV based fields
Study	26	6
Biosample	83	11
Organism	124	31
Sequencing Project	44	8
Analysis Project	35	2

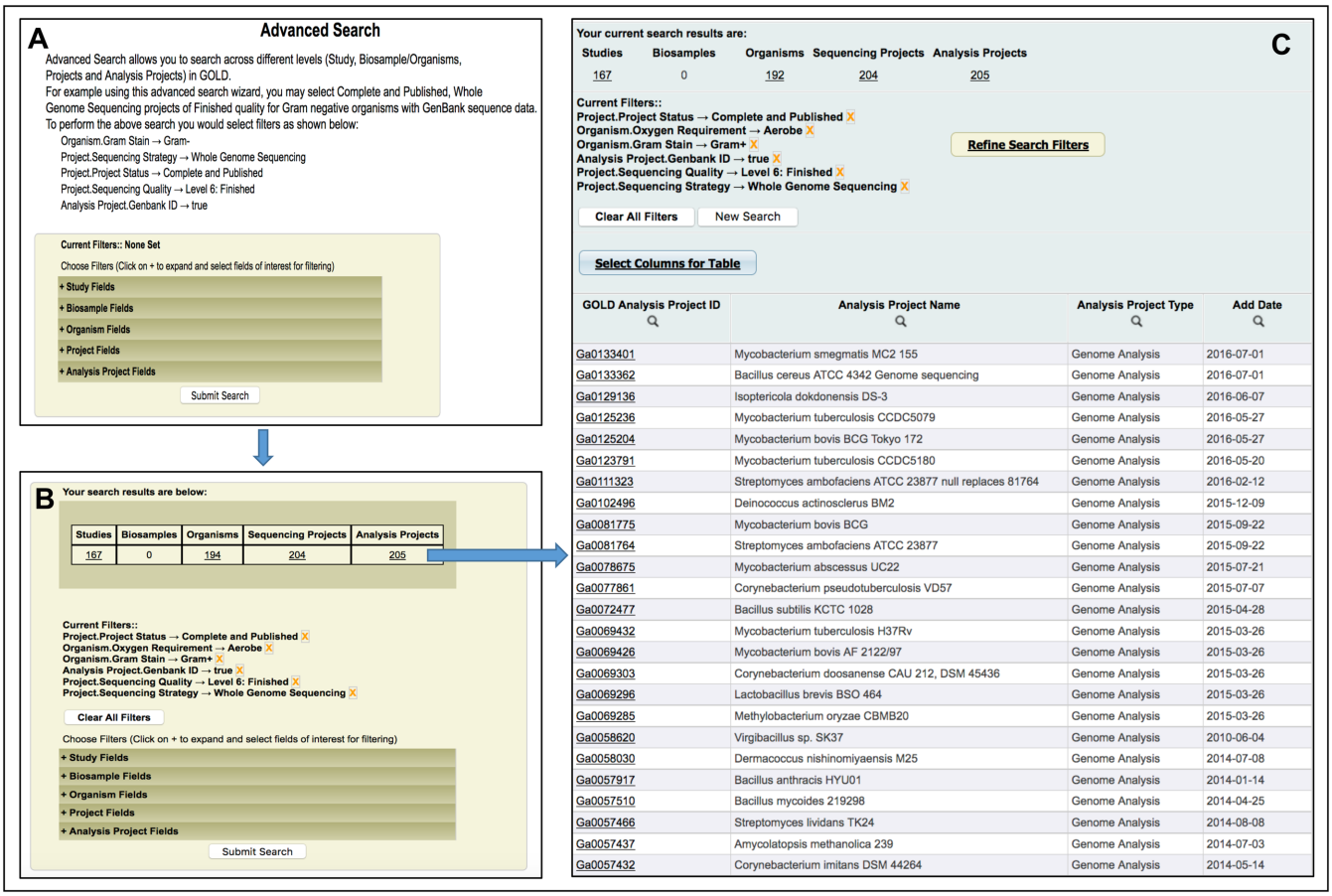

Figure 4. Advanced Search feature in GOLD. (**A**) Advanced Search launch page in GOLD with a brief explanation of how to conduct an advanced search. (**B**) Advanced Search results after applying six different search filters across three GOLD levels. (**C**) List of GOLD Analysis Projects obtained from the Advanced Search.

Search results page are displayed (Figure 4C). The Analysis Projects list/table can be explored as previously by selecting/adding new columns for display and filtering on those columns. At the top of the results page there are several options for exploring advanced search results. These include: (i) remove one or more of the already applied filters; (ii) refine current filters by adding new filters or removing already applied filters and (iii) launch a new search. In another example of the Advanced Search feature, if a user is interested in metagenome projects from Thermal springs whose analysis was completed after January 2014 the following filtering criteria will be applied:

Biosample.Ecosystem → Environmental, Biosample.Ecosystem Category → Aquatic, Biosample.Ecosystem Type → Thermal springs, Project.Sequencing Strategy → Metagenome, Analysis Project.Completion Date → >01-01-2014.

Metadata packages

For each of the four project levels, a defined set of metadata fields allows users to describe their entries in GOLD. Metadata fields are being constantly expanded with new entries to accommodate specific needs of the user. For example, in the current version, GOLD Organisms contain metadata fields specific to ocean ecosystems (http://www.nodc.noaa.gov/OC5/woa13/) such as Longhurst Code, World Ocean Atlas (WOA) Temperature, WOA Salinity etc. that capture metadata related to marine cyanobacteria and their phages. However, occasionally, Biosamples or Organisms may be submitted with a specific set of metadata that are not part of GOLD's standard set of metadata fields. In these cases GOLD cannot capture these specific metadata. To address this shortfall and to promote extended metadata acquisition and curation efforts, GOLD now supports metadata packages. We implemented a custom Biogas/Reactor metadata package to capture specific metadata applicable for samples coming from biogas reactors. As shown in Figure 5, Biogas/Reactor package supports close to twenty specific metadata fields that are unique to samples from Biogas reactors. These include biogas plant substrate, retention time, yield, total organic carbon, methane percentage etc.

New Combined Assembly Analysis Project types

Frequently, raw sequencing data from multiple Sequencing Projects (typically metagenomes, but often single cells as well) are co-assembled in order to generate better assemblies. In order to capture this information in GOLD, a Combined Assembly AP is created that is connected to multiple SPs. A combined assembly generally results in a higher number of well-characterized contigs, leading to a better taxonomic and functional annotation of sequence data. For example, a combination of combined assembly and genome binning of high-throughput metagenome sequences of microbial communities (from GOLD study Gs0095506) led to the identification of previously unknown bacterial species from biogas plants in Germany (40). The previous version of GOLD supported combined assemblies among metagenome projects only. The current version supports creation of combined assemblies consisting of the following types of Sequencing Projects: (i) Metagenome SPs, (ii) Metatranscriptome SPs, (iii) Single-cell SPs and (iv) Metagenomic project with Single-Cells. As shown in Table 2, GOLD currently has 109 APs that are defined as Combined Assemblies.

Data updates since last release

Major data updates to GOLD since our last release include the addition of Public Organisms, Sequence Read Archive (SRA) based metagenomes and support for NCBI Multi-isolate Project imports.

Import of public Organisms into GOLD

A new Organism can be created by a user while entering a SP or as part of GOLD's public Sequencing Projects import pipeline from an external resource such as NCBI. When a new Organism is entered by a user there is always a possibility of creating a duplicate entry in GOLD. Potential errors can also creep in if the genus, species, strain or other phylogeny fields of the new Organism are not accurately recorded. Additionally, Organisms that are imported from multiple external sources often require additional curation due to inconsistent quality control standards at other resources. To address these problems, GOLD imported over 150 000 publicly available organisms from the StrainInfo database (30). These Organisms entered are in accordance with standard taxonomic conventions. This expanded set of new Organisms is available for the user to select from when creating a new Project. The availability of these Organisms in GOLD is expected to speed up the Project creation process and also help to reduce manual errors in the process, at least for the Organisms already described.

Import of SRA based metagenomes and associated metadata

The NCBI SRA database (41) stores large volumes of raw sequence data for metagenomic samples. Earlier versions of GOLD did not import metagenome BioProjects or their associated SRA information from NCBI although some select studies were manually entered by GOLD users. The current release supports the systematic import of metagenome projects from NCBI's SRA database. As part of this import process, GOLD has incorporated information from a number of non-amplicon, Illumina-based SRA Runs. Currently GOLD has information for 858 SRA Studies corresponding to 11 914 SRA Experiments and 19 645 Runs. Data from these Projects are subsequently passed on to the IMG assembly and annotation pipeline and are eventually integrated into the IMG system and released to the public.

Incorporation of NCBI Multi-isolate projects

GOLD regularly imports projects from external resources. NCBI BioProject/GenBank is a major source for our external imports. Previously NCBI used to have separate BioProjects for each genome sequencing project. When these projects were imported into GOLD, each Project was associated to a unique NCBI BioProject ID. Recently NCBI introduced the concept of multi-isolate BioProjects where multiple isolate genomes are grouped under a single BioProject ID. To accommodate for this change, GOLD revamped

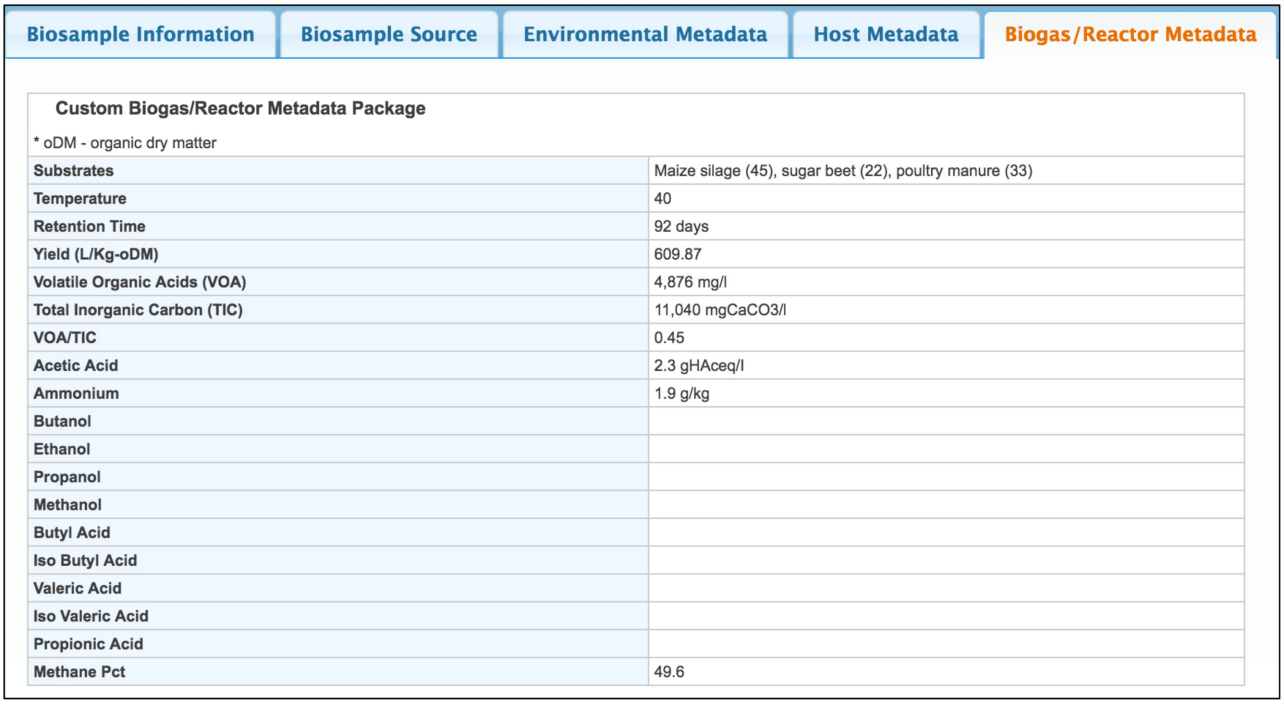

Figure 5. Description of a GOLD Metadata Package. Biosample populated using the Biogas/Reactor metadata package. All the different metadata categories that are unique to bioreactor samples are listed here.

its project import process. The current version of the GOLD database includes over 11 500 multi-isolate Projects. NCBI multi-isolate projects are now a regular component of GOLD's semi-automatic genome import pipeline and as a result GOLD SPs currently have a one-to-one analogy with a NCBI BioSample, in order to account for the inclusion of multi-isolate projects.

NAVIGATING GOLD

GOLD provides login free access to all of its publicly available data**.** The total number and different types of Studies, Biosamples, Organisms, SPs and APs are computed on a daily basis and presented in a table with hyperlinks on the GOLD home page. A brief summary of the different menu tabs in the GOLD web user interface is provided below:

Search

The search option enables a user to query the GOLD database within its multi-level project classification system and different metadata categories. The search drop-down menu is categorized into (i) Advanced Search that is designed to query GOLD across a suite of multiple project features and metadata fields, all at the same time and (ii) Metadata Search that allows the user to search GOLD using metadata identifiers and provides a graphical as well as tabular output of the results.

Distribution Graphs

Data summary of different types of Sequencing Projects, sequencing status, Organism phylogenetic classification, Biosample ecosystem classifications, etc. are provided as pre-computed pie charts and tables in the 'Distribution Graphs' section of the GOLD UI.

Biogeographical Metadata

The Biogeographical Metadata section displays the geographic location of GOLD Biosamples and Organisms using the map and terrain components of Google map. The interactive maps in this segment can be zoomed in or out to focus on a specific geo-location to search for specific Biosamples/Organisms from that region.

Statistics

The statistics component of the GOLD UI consists of graphs and charts encompassing several different metadata categories from Sequencing Projects. A user can access the summary statistics of the growth of genome Sequencing Projects, as they were added in GOLD over the years and also look at their breakdown by sequencing status or project completeness. Pre-computed pie-charts displaying the distribution of projects by relevance or by sequencing centers are also available in the GOLD statistics page.

CREATING SEQUENCING PROJECTS IN GOLD

GOLD continuously imports publicly available genome and metagenome projects from other resources. If a public sequencing project is not yet in GOLD or a user has a private genome project, which they want to define in GOLD and annotate at IMG, they can use the project entry interface to do that. Each isolate genome Sequencing Project re-

quires an Organism entry in GOLD. Typically a user defines an Organism during project entry process or selects an existing Organism. Part of our manual curation effort is to ensure that all Organisms in GOLD are unique, so it is important not to create duplicate Organism entries. Since GOLD now contains over 230 000 public Organisms, the chances of a user requiring to enter a new Organism is greatly diminished. To facilitate project entry, we have put together a help document (https://gold.jgi.doe.gov/resources/project_help_doc.pdf) listing step-by-step instructions with screenshots, showing how to define different Sequencing and Analysis Projects.

GOLD USERS AND USAGE STATISTICS

GOLD has 14 000 registered users. A GOLD user account is required to submit private data to GOLD. All public data can be accessed without a user account. In the last twelve months 75 000 unique users visited GOLD from around the world. Majority of GOLD users come from North America. Besides individual users various other database resources source GOLD metadata. They are the Data Analysis and Coordination Center (DAAC) of HMP (http://hmpdacc.org/), The Pathosystems Resource Integration Center (PATRIC) (42), World Data Center for Microorganism (WDCM) (http://www.wdcm.org/), the EBI Metagenomics (43) etc. We also exchange metadata between external collaborators and provide custom database reports to users as per their research needs.

FUTURE DEVELOPMENT PLANS

GOLD's future development plans can be broadly classified into the following six categories. They are (i) Data acquisition, (ii) Expanding metadata fields, (iii) Metadata packages, (iv) Scalable metadata curation (v) User interface and search enhancements and (vi) implementing unique identifiers.

Data acquisition

We will continue to import genome and metagenomic projects from external resources like NCBI's GenBank and SRA into GOLD. This is an ongoing process with ever increasing data in public domain with more and more complex Studies and associated metadata. It is a constant challenge to fine-tune our semi-automatic import scripts that generate data for manual checks. Our future efforts will be focused on gaining efficiencies on the overall import process as well as on projects that we can process through IMG pipeline.

Expanding metadata fields

We are constantly adding new metadata fields and/or reorganizing existing fields to best suit the needs of emerging research projects. As newer and cheaper technologies make it possible to pursue studies with diverse aims and scope, it necessitates to expand metadata fields. Studies like built environment metagenomes, deep ocean samples, upper atmospheric samples etc. are few diverse examples that require specific set of metadata fields. GOLD currently supports such diverse Studies by accommodating new metadata fields.

Metadata packages

Specific Studies require a unique set of metadata fields that in general may not be applicable across all Biosamples or Organisms in GOLD. In such cases there is a need to implement specific metadata packages. For example biogas reactor Biosamples require a set of unique metadata fields as shown in Figure 5**.** We plan to expand across similar metadata packages in the near future.

Scalable metadata curation

GOLD's current metadata quality and consistency is due to manual curation. However, it is understandable that manual curation cannot scale at the level of data growth. Much of the future operations in this direction will concentrate on developing automatic or semi-automatic Quality Control (QC) checks for metadata, as well as developing more accurate text mining and natural language processing approaches that would parse the existing wealth of metadata available in the literature (44–46). Crowdsourcing could be another mechanism to maintain curation quality that will be explored (47,48).

User interface (UI) and search enhancements

GOLD users interact with our database through UI both to enter new Projects and search GOLD for public Projects. The new Advanced Search feature we described in this paper is aimed at our user needs to explore GOLD's different levels seamlessly. We will continue to develop the Advanced Search feature to include more metadata fields. It is certainly tedious to enter multiple samples with more or less similar metadata. Also some of the critical metadata for environmental samples such as geo-location, latitude, longitude, altitude, collection date, etc. are now captured by researchers in the field using portable devices like smartphones. Because of these changes in how information is captured, we will explore the implementation of a smartphone app to capture metadata at the time of sample collection in field. We also plan to develop and support the option of loading multiple projects using a batch loading process.

Digital object identifier

We plan to obtain DOIs for organisms and APs in GOLD. DOIs are persistent identifiers used to uniquely identify objects and will help our users in referring to GOLD/IMG data in their publications as well as on any digital platforms.

ACKNOWLEDGEMENTS

The authors are thankful to researchers who take time to accurately document and provide metadata directly to GOLD or via other public resources. The authors thank Alexander Sczyrba from Bielefeld University for help in incorporating biogas reactor specific metadata package. We also value

constant community feedback in improving and in maintaining accurate information in GOLD. The authors thank members of the microbial genomics and metagenomics programs at the Joint Genome Institute (JGI) for their constant support, feedback and helpful discussions. Visualizations were generated using the maps and ggplot2 packages in R.

FUNDING

This work was conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under contract number DE-AC02-05CH11231. Funding for open access charge: Office of Science of the U.S. Department of Energy [contract DE-AC02-05CH11231].
Conflict of interest statement. None declared.

REFERENCES

1. Kyrpides,N.C. (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, **15**, 773–774.
2. Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.*, **29**, 126–127.
3. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.*, **34**, D332–D334.
4. Liolios,K., Mavromatis,K., Tavernarakis,N. and Kyrpides,N.C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **36**, D475–D479.
5. Liolios,K., Chen,I.-M.A., Mavromatis,K., Tavernarakis,N., Hugenholtz,P., Markowitz,V.M. and Kyrpides,N.C. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **38**, D346–D354.
6. Reddy,T.B.K., Thomas,A.D., Stamatis,D., Bertsch,J., Isbandi,M., Jansson,J., Mallajosyula,J., Pagani,I., Lobos,E.A. and Kyrpides,N.C. (2015) 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.
7. Paez-Espino,D., Eloe-Fadrosh,E.A., Pavlopoulos,G.A., Thomas,A.D., Huntemann,M., Mikhailova,N., Rubin,E., Ivanova,N.N. and Kyrpides,N.C. (2016) Uncovering Earth's virome. *Nature*, **536**, 425–430.
8. Teeling,H., Fuchs,B.M., Bennke,C.M., Krüger,K., Chafee,M., Kappelmann,L., Reintjes,G., Waldmann,J., Quast,C., Glöckner,F.O. *et al.* (2016) Recuring patterns in bacterioplankton dynamics during coastal spring algae blooms. *eLife*, **5**, e11888.
9. Seshadri,R., Reeve,W.G., Ardley,J.K., Tennessen,K., Woyke,T., Kyrpides,N.C. and Ivanova,N.N. (2015) Discovery of novel plant interaction determinants from the genomes of 163 root nodule bacteria. *Sci. Rep.*, **5**, 16825.
10. Stephens,Z.D., Lee,S.Y., Faghri,F., Campbell,R.H., Zhai,C., Efron,M.J., Iyer,R., Schatz,M.C., Sinha,S. and Robinson,G.E. (2015) Big Data: Astronomical or Genomical? *PLoS Biol.*, **13**, e1002195.
11. Human Microbiome Jumpstart Reference Strains Consortium, Nelson,K.E., Weinstock,G.M., Highlander,S.K., Worley,K.C., Creasy,H.H., Wortman,J.R., Rusch,D.B., Mitreva,M., Sodergren,E. *et al.* (2010) A catalog of reference genomes from the human microbiome. *Science*, **328**, 994–999.
12. Grigoriev,I.V., Nikitin,R., Haridas,S., Kuo,A., Ohm,R., Otillar,R., Riley,R., Salamov,A., Zhao,X., Korzeniewski,F. *et al.* (2014) MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.*, **42**, D699–D704.
13. Wu,D., Hugenholtz,P., Mavromatis,K., Pukall,R., Dalin,E., Ivanova,N.N., Kunin,V., Goodwin,L., Wu,M., Tindall,B.J. *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, **462**, 1056–1060.
14. Kyrpides,N.C., Hugenholtz,P., Eisen,J.A., Woyke,T., Göker,M., Parker,C.T., Amann,R., Beck,B.J., Chain,P.S.G., Chun,J. *et al.* (2014) Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol.*, **12**, e1001920.
15. Kyrpides,N.C., Woyke,T., Eisen,J.A., Garrity,G., Lilburn,T.G., Beck,B.J., Whitman,W.B., Hugenholtz,P. and Klenk,H.-P. (2014) Genomic Encyclopedia of Type Strains, Phase I: The one thousand microbial genomes (KMG-I) project. *Stand. Genomic Sci.*, **9**, 1278–1284.
16. Ishoey,T., Woyke,T., Stepanauskas,R., Novotny,M. and Lasken,R.S. (2008) Genomic sequencing of single microbial cells from environmental samples. *Curr. Opin. Microbiol.*, **11**, 198–204.
17. Rinke,C., Schwientek,P., Sczyrba,A., Ivanova,N.N., Anderson,I.J., Cheng,J.-F., Darling,A., Malfatti,S., Swan,B.K., Gies,E.A. *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.
18. Tsementzi,D., Wu,J., Deutsch,S., Nath,S., Rodriguez-R,L.M., Burns,A.S., Ranjan,P., Sarode,N., Malmstrom,R.R., Padilla,C.C. *et al.* (2016) SAR11 bacteria linked to ocean anoxia and nitrogen loss. *Nature*, **536**, 179–183.
19. Eloe-Fadrosh,E.A., Paez-Espino,D., Jarett,J., Dunfield,P.F., Hedlund,B.P., Dekas,A.E., Grasby,S.E., Brady,A.L., Dong,H., Briggs,B.R. *et al.* (2016) Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat. Commun.*, **7**, 10476.
20. Hedlund,B.P., Dodsworth,J.A., Murugapiran,S.K., Rinke,C. and Woyke,T. (2014) Impact of single-cell genomics and metagenomics on the emerging view of extremophile 'microbial dark matter'. *Extremophiles*, **18**, 865–875.
21. Markowitz,V.M., Chen,I.-M.A., Palaniappan,K., Chu,K., Szeto,E., Pillay,M., Ratner,A., Huang,J., Woyke,T., Huntemann,M. *et al.* (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.*, **42**, D560–D567.
22. Markowitz,V.M., Chen,I.-M.A., Chu,K., Szeto,E., Palaniappan,K., Pillay,M., Ratner,A., Huang,J., Pagani,I., Tringe,S. *et al.* (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.*, **42**, D568–D573.
23. Huntemann,M., Ivanova,N.N., Mavromatis,K., Tripp,H.J., Paez-Espino,D., Palaniappan,K., Szeto,E., Pillay,M., Chen,I.-M.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.
24. Huntemann,M., Ivanova,N.N., Mavromatis,K., Tripp,H.J., Paez-Espino,D., Tennessen,K., Palaniappan,K., Szeto,E., Pillay,M., Chen,I.-M.A. *et al.* (2016) The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Stand. Genomic Sci.*, **11**, 17.
25. 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.
26. Yilmaz,P., Kottmann,R., Field,D., Knight,R., Cole,J.R., Amaral-Zettler,L., Gilbert,J.A., Karsch-Mizrachi,I., Johnston,A., Cochrane,G. *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any sequence (MIxS) specifications. *Nat. Biotechnol.*, **29**, 415–420.
27. Bischof,J., Harrison,T., Paczian,T., Glass,E., Wilke,A. and Meyer,F. (2014) Metazen - metadata capture for metagenomes. *Stand. Genomic Sci.*, **9**, 18.
28. Garrity,G.M. (2011) The state of standards in genomic sciences. *Stand. Genomic Sci.*, **5**, 262–268.
29. NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7–D19.
30. Verslyppe,B., De Smet,W., De Baets,B., De Vos,P. and Dawyndt,P. (2014) StrainInfo introduces electronic passports for microorganisms. *Syst. Appl. Microbiol.*, **37**, 42–50.
31. Krieg,N.R. and Garrity,G.M. (2012) On using the Manual. In: Goodfellow,M, Kämpfer,P, Busse,H-J, Trujillo,ME, Suzuki,K, Ludwig,W and Whitman,WB (eds). *Bergey's Manual® of Systematic Bacteriology*. Springer, NY, Vol. **5**, pp. 23–24.
32. Parker,C.T., Tindall,B.J. and Garrity,G.M. (2015) International code of nomenclature of prokaryotes. *Int. J. Syst. Evol. Microbiol.*, doi:10.1099/ijsem.0.000778.
33. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.

34. Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.
35. Hug,L.A., Baker,B.J., Anantharaman,K., Brown,C.T., Probst,A.J., Castelle,C.J., Butterfield,C.N., Hernsdorf,A.W., Amano,Y., Ise,K. *et al.* (2016) A new view of the tree of life. *Nat. Microbiol.*, **1**, 16048.
36. Rinke,C., Lee,J., Nath,N., Goudeau,D., Thompson,B., Poulton,N., Dmitrieff,E., Malmstrom,R., Stepanauskas,R. and Woyke,T. (2014) Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat. Protoc.*, **9**, 1038–1048.
37. Mukherjee,S., Huntemann,M., Ivanova,N., Kyrpides,N.C. and Pati,A. (2015) Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand. Genomic Sci.*, **10**, 18.
38. Tennessen,K., Andersen,E., Clingenpeel,S., Rinke,C., Lundberg,D.S., Han,J., Dangl,J.L., Ivanova,N., Woyke,T., Kyrpides,N. *et al.* (2016) ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J.*, **10**, 269–272.
39. Federhen,S., Clark,K., Barrett,T., Parkinson,H., Ostell,J., Kodama,Y., Mashima,J., Nakamura,Y., Cochrane,G. and Karsch-Mizrachi,I. (2014) Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Stand. Genomic Sci.*, **9**, 1275–1277.
40. Stolze,Y., Bremges,A., Rumming,M., Henke,C., Maus,I., Pühler,A., Sczyrba,A. and Schlüter,A. (2016) Identification and genome reconstruction of abundant distinct taxa in microbiomes from one thermophilic and three mesophilic production-scale biogas plants. *Biotechnol. Biofuels*, **9**, 156.
41. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database Collaboration (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
42. Wattam,A.R., Abraham,D., Dalay,O., Disz,T.L., Driscoll,T., Gabbard,J.L., Gillespie,J.J., Gough,R., Hix,D., Kenyon,R. *et al.* (2013) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D591.
43. Mitchell,A., Bucchini,F., Cochrane,G., Denise,H., ten Hoopen,P., Fraser,M., Pesseat,S., Potter,S., Scheremetjew,M., Sterk,P. *et al.* (2016) EBI metagenomics in 2016–an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, **44**, D595–D603.
44. Hirschman,L., Burns,G.A.P.C., Krallinger,M., Arighi,C., Cohen,K.B., Valencia,A., Wu,C.H., Chatr-Aryamontri,A., Dowell,K.G., Huala,E. *et al.* (2012) Text mining for the biocuration workflow. *Database J. Biol. Databases Curation*, **2012**, bas020.
45. Pafilis,E., Buttigieg,P.L., Ferrell,B., Pereira,E., Schnetzer,J., Arvanitidis,C. and Jensen,L.J. (2016) EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. *Database J. Biol. Databases Curation*, **2016**, baw005.
46. Papanikolaou,N., Pavlopoulos,G.A., Pafilis,E., Theodosiou,T., Schneider,R., Satagopam,V.P., Ouzounis,C.A., Eliopoulos,A.G., Promponas,V.J. and Iliopoulos,I. (2015) BioTextQuest+: a knowledge integration platform for literature mining and concept discovery. *Bioinformatics*, **30**, 3249–3256.
47. Hirschman,L., Fort,K., Boué,S., Kyrpides,N., Islamaj Doğan,R. and Cohen,K.B. (2016) Crowdsourcing and curation: perspectives from biology and natural language processing. *Database J. Biol. Databases Curation*, **2016**, baw115.
48. McQuilton,P., Gonzalez-Beltran,A., Rocca-Serra,P., Thurston,M., Lister,A., Maguire,E. and Sansone,S.-A. (2016) BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database J. Biol. Databases Curation*, **2016**, baw075.