Published online 24 October 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D649–D659***
doi: 10.1093/nar/gky977

Genomes OnLine database (GOLD) v.7: updates and new features

Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Hema Y. Katta, Alejandro Mojica, I-Min A. Chen, Nikos C. Kyrpides[*] and T.B.K. Reddy[*]

Prokaryotic Super Program, DOE Joint Genome Institute, Walnut Creek, CA 94598, USA

Received September 14, 2018; Revised October 04, 2018; Editorial Decision October 05, 2018; Accepted October 08, 2018

ABSTRACT

The Genomes Online Database (GOLD) (https://gold.jgi.doe.gov) is an open online resource, which maintains an up-to-date catalog of genome and metagenome projects in the context of a comprehensive list of associated metadata. Information in GOLD is organized into four levels: Study, Biosample/Organism, Sequencing Project and Analysis Project. Currently GOLD hosts information on 33 415 Studies, 49 826 Biosamples, 313 324 Organisms, 215 881 Sequencing Projects and 174 454 Analysis Projects with a total of 541 metadata fields, of which 80 are based on controlled vocabulary (CV) terms. GOLD provides a user-friendly web interface to browse sequencing projects and launch advanced search tools across four classification levels. Users submit metadata on a wide range of Sequencing and Analysis Projects in GOLD before depositing sequence data to the Integrated Microbial Genomes (IMG) system for analysis. GOLD conforms with and supports the rules set by the Genomic Standards Consortium (GSC) Minimum Information standards. The current version of GOLD (v.7) has seen the number of projects and associated metadata increase exponentially over the years. This paper provides an update on the current status of GOLD and highlights the new features added over the last two years.

INTRODUCTION

Genomes OnLine Database (GOLD) is a freely available information rich resource of sequencing projects and their associated metadata (1). GOLD serves as a major resource that catalogs and monitors genome and metagenome projects from around the world. Since its inception in 1997 (2), GOLD has grown exponentially, keeping pace with the growth in number of sequencing projects. DNA sequencing recently celebrated its 40th anniversary (3,4). There have been several technological revolutions in this relatively short period of time. A reduction in sequencing cost as well as advancements in sequencing technologies and bioinformatics analyses have led to an increase in both the number and diversity of genomes that were sequenced. Large-scale, multi-institute projects such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA), Genome 10k Project, Earth BioGenome Project are few of the several recent initiatives to sequence thousands or hundreds of thousands of isolate organisms (5–7). This exponential growth in genomics, compared to other data science fields, has promoted calls for substituting the term 'Astronomical' with 'Genomical' to describe the vast quantities of genomic data being generated (8).

Sequencing of cultured isolate microorganisms is important as they serve as references for related, less known organisms. At the same time, a large fraction of the prokaryotic diversity on Earth remains uncultured. Our knowledge about these organisms stems from the analyses of single cells, environmental DNA and metagenome-assembled genomes (MAGs). An example of a large-scale project targeting the uncultured diaspora is the Earth Microbiome Project (EMP) (9), which aims to create a global catalogue of the Earth's uncultured microorganisms. In fact, cultivation-independent sequencing is predicted to outrank the rate at which cultured isolates are being sequenced (10).

Sequence data facilitates comparative analysis and leads to discoveries only when it is accompanied with accurate metadata. This is precisely where GOLD, with its collection of rich and carefully curated metadata, comes in. The data in GOLD is freely available through an easy-to-use web interface. A user can browse through the entire collection of public genome and metagenome projects, study the metadata and download statistics and figures for use in publications and presentations. GOLD Metadata is used in research projects carried out by individual researchers as well as other resources. GOLD geolocation metadata powered BioAtlas, that provides geographic profiles of 16S rRNA sequences from metagenomes, whether they came from geographical and/or host-oriented locations (11). MAR databases that are part of Marine Metagenomics Portal use

[*] 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

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

curated data from GOLD (12). Several publications have also used GOLD metadata. For example, Sibbald *et al.* (13) called for more expansive sequencing of protists after surveying the sequencing status metadata in GOLD. The ecosystem metadata from GOLD was used in a publication describing the ecological and biological factors influencing viral coinfection (14).

There are three different sources for projects in GOLD: internal projects carried out at the Department of Energy Joint Genome Institute (DOE-JGI), projects submitted by external users and projects sourced from public databases such as NCBI, EBI and others. GOLD serves as a doorkeeper for sequencing projects submitted to IMG system (15) for analysis. Before sequence data can be submitted to IMG for annotation, it must be defined in GOLD (Sequencing Project and Analysis Project) along with all the relevant metadata. GOLD follows standards mandated by the Genomics Standards Consortium (16) and is compliant with its Minimum Information about any (x) Sequence (MIxS) standards (17). This ensures uniform standards applied to all entities in the database and helps in comparative analysis.

This is a very exciting time in the field of genomics. The growth in number and variety of sequencing projects and analysis strategies are promising in terms of providing better insights into research questions and hypothesis testing. Simultaneously it demands an agile database platform that can quickly adopt by adding new features and capabilities to organize, store and track associated metadata efficiently. This manuscript provides a status update and description of new features of GOLD that were implemented over the last two years in response to the growing demands of the genomics community. At a time when database errors (18,19) are easily propagated leading to incorrect and misleading research, the value of a reliable genomic data management system such as GOLD becomes more important.

OVERVIEW AND ORGANIZATION OF GOLD

The current version of GOLD is organized by a four-level classification system namely Study, Biosample/Organism, Sequencing Project and Analysis Project. A GOLD Study lies at the top of this hierarchical classification system and broadly describes the overall objective of the sequencing projects that it contains. The physical material collected from the environment is called a Biosample, while living biological material such as bacteria, fungus, plant or animal is termed as an Organism in GOLD. The sequencing output of a GOLD Biosample or Organism makes up a Sequencing Project (SP) and the subsequent analysis and data processing methods are described in an Analysis Project (AP). This organization structure ensures that the different aspects of sequencing projects and their related metadata are connected to each other in a coherent manner.

Study

A Study summarizes the overall goal of a research initiative and outlines the key objective of its underlying projects. A GOLD Study may have only one Sequencing and Analysis Project, where sequencing a single organism fulfills a research objective. A user may need multiple Sequencing Projects with a wide range of sequencing strategies such as isolate genome, metagenome, metatranscriptome and single-cell sequencing to tackle a research problem. In such a scenario, all related projects will be grouped under a single GOLD Study. There is no limit to the number of Sequencing or Analysis Projects that can constitute a Study.

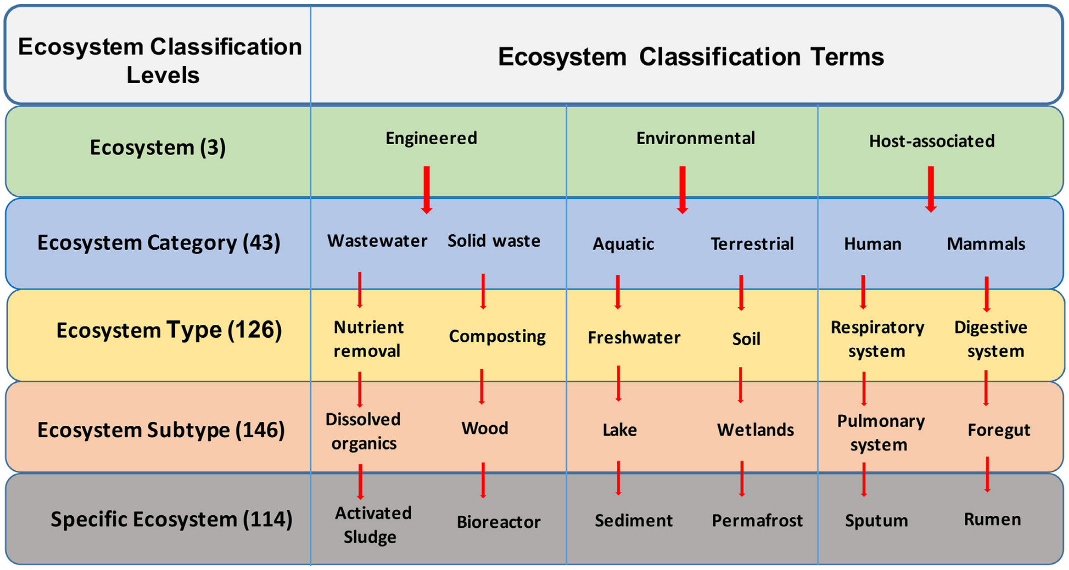
Biosample/Organism

A Biosample or Organism in GOLD is used to describe the material that is being sequenced. An environmental sample containing genetic material from multiple individuals is commonly referred to as a Biosample. Examples of GOLD Biosamples include soil from a farm, lake sediment, fecal sample from a diseased animal etc. A Biosample is described with metadata such as habitat, ecosystem, geographical location, latitude and longitude, among others and is required for creating a Metagenome or Metatranscriptome Sequencing Project. A GOLD Organism on the other hand is used to describe an individual entity such as a bacterium, fungus, plant, animal or a virus. It can be a cultured isolate of a pure strain of bacterium or an uncultured single-cell isolated using cell sorting. Metagenome-Assembled Genomes (MAGs) associate with a new type of uncultured, non-living Organism in GOLD. All Organisms are required to have basic taxonomic information such as genus, species, strain, NCBI taxonomy ID, phylum etc. Defining an Organism entity is essential to create GOLD Sequencing Projects with sequencing strategies such as Whole genome sequencing, Transcriptome and others.

Environmental features of a Biosample or an Organism is described in a five-level ecosystem classification (Figure 1) (20). At the top of this classification system is Ecosystem, which consists of Engineered, Environmental and Host-associated terms to describe the broader environment where a Biosample or Organism came from. As shown in Figure 1 these top three terms are further classified into four classification levels namely: Ecosystem Category, Ecosystem Type, Ecosystem Subtype and Specific Ecosystem. These terms provide a detailed description of the source environment of a Biosample or Organism. For example, a GOLD Biosample from lake sediment will have the following ecosystem classification: Ecosystem: Environmental, Ecosystem Category: Aquatic, Ecosystem Type: Freshwater, Ecosystem Subtype: Lake and Specific Ecosystem: Sediment. Similarly, an Organism isolated from human sputum will have Ecosystem: Host-associated, Ecosystem Category: Human, Ecosystem Type: Respiratory system, Ecosystem Subtype: Pulmonary system and Specific Ecosystem: Sputum. All GOLD Biosamples and a large number of Organisms have at least the top two-level classifications populated in GOLD.

Sequencing Project

The methods and processes used to generate sequencing output from a Biosample or Organism are described in a Sequencing Project. Details about the sequencing strategy, type of nucleic acid used, sequencing center, sequencing technology etc. are included in a Sequencing Project. Isolate



Ecosystem Classification Levels	Ecosystem Classification Terms					
Ecosystem (3)	Engineered		Environmental		Host-associated	
Ecosystem Category (43)	Wastewater	Solid waste	Aquatic	Terrestrial	Human	Mammals
Ecosystem Type (126)	Nutrient removal	Composting	Freshwater	Soil	Respiratory system	Digestive system
Ecosystem Subtype (146)	Dissolved organics	Wood	Lake	Wetlands	Pulmonary system	Foregut
Specific Ecosystem (114)	Activated Sludge	Bioreactor	Sediment	Permafrost	Sputum	Rumen

Figure 1. Ecosystem classification in GOLD. The five ecosystem classification levels are displayed in the left column with the number of unique terms at each level in parenthesis. Select terms from each classification level are shown in three right columns, with arrows showing possible ecosystem classification paths.

Whole Genome Sequencing (WGS), transcriptomes, targeted gene surveys, metagenomes, metatranscriptomes are some of the several different types of Sequencing Projects in GOLD. Information from the NCBI database such as BioProject Accession, BioSample Accession are also available. A Biosample or Organism in GOLD may be linked to several Sequencing Projects with different sequencing strategies. For example, DNA and RNA from the same soil sample can be used in separate metagenome and metatranscriptome projects respectively.

Analysis Project

A GOLD Analysis Project describes the assembly and annotation processes that are performed on a Sequencing Project. A user must create a GOLD Analysis Project in order to submit sequence data to IMG for analysis (21). The different types of Analysis Projects in GOLD are Genome, Metagenome, Metatranscriptome, Single Cell (Unscreened), Single Cell (Screened), Genome from Metagenome, Transcriptome and Combined Assembly Analysis Projects. A single Sequencing Project may have multiple Analysis Projects. A user has the option to choose which of these Analysis Projects is the primary one for analysis. The remaining Analysis Projects from the same Sequencing Project become designated as reanalysis. Assembly method, gene calling method, sequencing depth, estimated genome size are some of the key metadata fields in an Analysis Project. Submitted data sets may take between 2 and 4 weeks for processing and loading into IMG database, depending upon the number of submissions in queue.

CURRENT GOLD STATUS

Studies

As of August 2018, there are 33 408 Studies in GOLD out of which 1521 are metagenomic Studies, with at least one metagenome or metatranscriptome project. The current number of Studies represents an increase of 7291 (28%) since the 2016 release and 14 208 (74%) since the 2014 release of the database (Figure 2A). Compared to the relative increase of Sequencing Projects and Analysis Projects (discussed later) during the last two GOLD releases, the number of Studies has not increased significantly. With improvements in sequencing capabilities and reduction in cost, individual Studies now contain a larger number of Sequencing and Analysis Projects compared to earlier years. We expect this trend to continue in the near future.

Biosample/Organism

As of August 2018, there are 49 821 Biosamples in GOLD. Out of these 4603 are from Engineered, 22 150 are from Environmental and 23 068 are from Host-associated (23 068) ecosystems. There are 563 unique ecosystem classification paths to describe these biosamples. There are 313 378 Organisms from 263 different phyla and candidate phyla. These numbers represent an increase of 33 934 (214%) and 74 278 (31%) for Biosamples and Organisms respectively, compared to the last release. The jump in the number of Biosamples is both due to the fact of increasing metagenome projects submitted by users as well as our data import efforts.

Sequencing Project

We currently have 215 881 Sequencing Projects in GOLD. This represents an increase of 118,669 (122%) compared to the last release and 159 881 (285%) since the 2014 release of the database (Figure 2A). WGS projects make up the bulk (65%) of Sequencing Projects followed by metagenome (20%) and transcriptome (9%) (Figure 2B). Majority of the WGS projects are comprised of bacteria (78.7%) followed

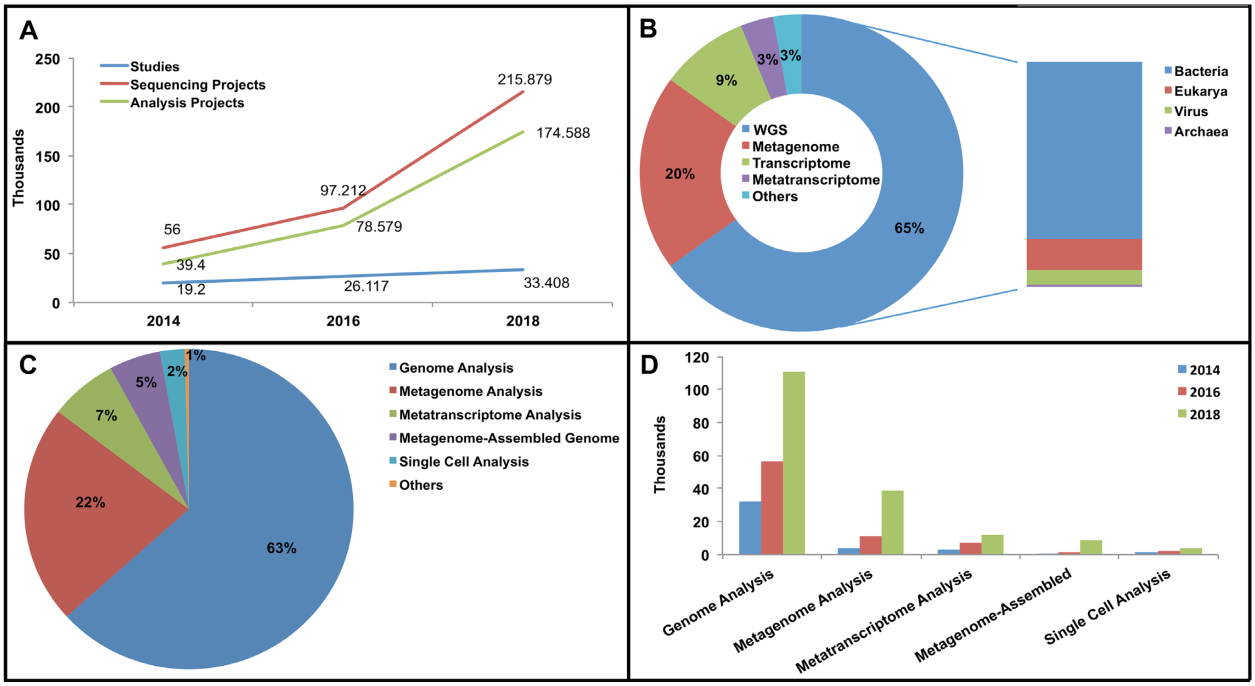


Figure 2. GOLD by the numbers. (**A**) Growth in Studies, Sequencing Projects and Analysis Projects during the last three releases of GOLD database. (**B**) Pie diagram showing distribution of Sequencing Project types in the current version. Vertical panel on the right displaying domain level distribution of organisms for WGS. (**C**) Different types of Analysis Projects and their percentage breakdown in the current release of GOLD. (**D**) Growth of different Analysis Project types during the last three releases.

by eukarya (13.8%), viruses (6.5%) and archaea (1%) (Figure 2B).

Analysis Project

As of August 2018, there are 174 454 Analysis Projects in GOLD. This is an increase of 95 875 (122%) since the 2016 release and 135 054 (343%) since the 2014 release of the database (Figure 2A). Of the current GOLD Analysis Projects, 85% are comprised of Genome Analysis (63%) and Metagenome Analysis (22%) followed by Metatranscriptome, MAGs, Single cells and others such as Transcriptome and Combined Assembly Analysis Projects (Figure 2C). While the number of isolate genome and Single Cell Analysis Projects essentially doubled compared to 2016 and increased ∼4-fold compared to 2014, metagenome and MAG Analysis Projects increased by 2.5-fold and 4-fold from 2016 and around 10-fold and 23-fold from the 2014 numbers respectively (Figure 2D). This can be largely attributed to the increasing interest in sequencing of metagenome samples as well as several recent large-scale projects to identify novel MAGs from environmental samples (22,23).

EXPLORING GOLD

Users have unrestricted access to projects and associated metadata for research and comparative analysis. The GOLD homepage provides an option to download the complete list of public projects as an Excel file. Users can get a summary of the different types of GOLD Studies, Organisms, Biosamples, Sequencing Projects and Analysis Projects along with their respective counts directly on the homepage. These counts are updated daily and are presented as clickable links. For example, a user can click on the number next to 'Sequencing Projects' on the top left corner of the homepage and go to a page with complete list of available Sequencing Projects. This list is sortable and searchable. Using 'Select Columns for Table' option, one can configure the list to display more columns of interest such as library method, GC percent, NCBI BioProject Accession, NCBI BioSample Accession etc. Using the magnifying lens on top of each column one can further filter the list based on project status, sequencing strategy, project name etc. Different menu tabs on the homepage are described below:

Search

The advanced search feature in GOLD can be used to seamlessly query across different levels of GOLD like Study, Biosample, Organism, Sequencing Project and Analysis Project. A combination of free text fields and CV-based drop-down menus can be customized to search across a total of 74 metadata fields. Using the advanced search feature, a user can perform a simple query like search for metagenomes from a freshwater environment or a relatively complex one, using multiple search filters, such as look for WGS projects from phylum Actinobacteria with a public GenBank ID, whose sequencing was completed after the year 2015.

Distribution graphs

Distribution Graphs tab on GOLD User Interface (UI) presents several graphical and tabular views of Sequencing

Project types, sequencing status, phylogenetic breakdown of Organisms, ecosystem classification of Biosamples etc. For example, the Phylogenetic Table subsection displays precomputed pie-charts of the distribution of major phyla from the three domains of life for Sequencing Projects that are in GOLD. Through expandable tabs a user can navigate down taxonomic lineage to see the distribution of Sequencing Projects at each level.

Biogeographical metadata

Interactive Biosample and Organism distribution maps are presented under biogeographical metadata section. Biosample and Organism geographic isolation site information is used to generate these maps. On Distribution maps individual Biosamples and Organisms are shown as placeholders on a google map based on their geographic location. The map can be zoomed in or out to focus on a particular location. A user can switch between satellite and map views. By clicking on individual placeholders, one can go to the respective Organism or Biosample pages.

Statistics

The statistics tab on the GOLD homepage contains several pre-computed graphs and charts summarizing different metadata fields from Sequencing Projects. For example, one of the charts show the growth and distribution of metagenome projects and WGS projects (by organism domain) over the years. The growth of metagenome Sequencing Projects, especially in the last few years can be easily interpreted from a quick glance at this chart. The precomputed graphs and pie-charts in the Statistics page is updated on a weekly basis.

GOLD METADATA FIELDS

GOLD currently has 541 metadata fields out of which 80 are based on Controlled Vocabularies (CV). This represents an increase of 73% in the total number of fields and an increase of 38% for CV based fields respectively compared to GOLD v.6. Most of this increase came with the addition of Soil and Water environmental packages. Metadata fields and CV terms were also added to accommodate new sequencing technologies or to better describe samples from a specific environment. For example, CV terms to represent new sequencing instruments such as PacBio Sequel and Illumina HiSeq X Ten were added. Another example is the newly added subsurface depth field, which helps to differentiate between multiple samples collected at different subsurface depths (Table 1).

Table 1. Number of metadata and cv based fields in GOLD

GOLD classification Level	No. of fields	No. of CV based fields
Study	27	6
Biosample	195	21
Organism	230	43
Sequencing Project	45	8
Analysis Project	44	2

NEW FEATURES SINCE LAST RELEASE

UI improvements

GOLD users access metadata through an interactive web interface. Grouping related information and presenting it in an easily navigable fashion is essential for a positive user experience as well as making metadata discoverable to users. Page layouts, search and sort features as well as page loading times are constantly optimized as per growing data and changing needs. We also implemented updates in response to user feedback. One such update is the selective displaying of relevant metadata tabs for Biosamples based on ecosystem classification. For example, the 'Host Metadata' tab, with details of host name, host taxonomy ID, host body site etc. are displayed only when the Ecosystem classification of a Biosample is set to 'Host-associated'. Similarly, the host metadata fields are displayed under the Isolation Metadata tab for 'Host-associated' Organisms only and were not displayed for Organisms that are from Engineered or Environmental ecosystems.

Several metadata fields in the GOLD UI have tooltips in order to explain what the field is and to assist users in populating such fields. Tooltips come handy during project entry and also while updating an existing GOLD entry. Based on user comments and feedback new tooltips have been added in the current release of the database. Several tooltips were also updated by adding specific examples. For example, the tooltip for 'Assembly Method' in the Analysis Project page has examples like Velvet v. 0.7.57, Newbler v. 2.3 etc. to indicate what value to add and in what format. We also implemented checks to make sure users enter data in expected format. We require assembly method to include version number starting with 'v.'. Values that fail to meet this format can't be saved and users will see a prompt asking to enter in expected format.

Environmental packages

Associated metadata for a genome or a metagenome project varies widely based on where the metagenome sample collected, or organism has been isolated from. For example, a metagenome sample may come from the bottom of a deep ocean and an organism might have been isolated from a hot spring. These two distinct environments require different set of metadata fields to accurately record data related to the collection and/or isolation of metagenome and organisms. Previous versions of GOLD offered a standard set of metadata fields for all biosamples and organisms. On a need basis we expanded metadata fields in GOLD over the years to accommodate expanding variety of metagenome samples and diverse environments from which organisms are being isolated. In our last release we included metadata fields specific to ocean and biogas reactor ecosystems. As more and more genomic data become publicly available for comparative analysis, researchers are realizing the need for standards and minimum set of metadata to describe organisms and metagenome samples. The Genomic Standards Consortium is coordinating efforts among the research community to come up with standard metadata packages to describe specific samples. In the current version we incorporated Soil and Water packages. Inclusion of these two packages in GOLD further expands the list of available metadata fields by 112, out of which 7 are controlled vocabularies comprising of 66 new terms. Figure 3 shows a select list of metadata fields used in soil package. For example, the soil package has a metadata field 'Fire' which records the historical and/or physical evidence of fire in a particular soil sample. While this field may not be applicable to most of the soil samples, it is a critical piece of metadata for a project, which studies the effect of long-term prescribed fire on soil bacterial community (24). Similarly, the 'Petroleum hydrocarbon' field in the water package is key to differentiate between samples studying methane cycling in hydrocarbons in groundwater from Pennsylvania, USA (25).

NCBI SRA explorer

Unassembled read data is available at SRA both for genome and metagenome projects. There is rich metadata associated with SRA data both about the underlying organisms or samples and instrument, sequencing technology and library strategies used. We import both genome and metagenome projects from NCBI. For genome projects, we check BioProject, BioSample entries and look at NCBI genome reports for assembly information. We import prokaryote genomes with assemblies and assembled eukaryotic genomes (fungi and protists) with CDS information. However, for metagenomes mostly users have to rely on unassembled read data at SRA and deal with myriad of sequencing instruments, library strategies to proceed with assemblies on their own. At GOLD we currently importing metagenomes sequenced on Illumina platform using WGS sequencing strategy that we can locally assemble and annotate in IMG. However, we want to provide an easy to use SRA data exploration tool to filter SRA data based on different criteria and check which SRA data is already in GOLD. This will serve as a bridge between a public repository like SRA and a curated metadata resource like GOLD plus IMG annotation platform. SRA explorer will let users filter (Figure 4A) on Library Source, Sequencing Instrument, Sequencing Strategy, Organism, various text search options and if see the data is in GOLD or not. As shown in Figure 4B, the applied filters of Scientific Name: soil metagenome, Library Strategy: WGS, Library Instrument: Illumina HiSeq 2000, Library Source: metagenomic and Projects in GOLD: True yielded 277 results, a portion of which is shown in Figure 4C. Users can also edit their search results by using the 'Refine SRA Search Filters' button (Figure 4B) and by checking/unchecking the radio buttons.

NCBI import tracker

GOLD routinely imports isolate genome and metagenome projects from NCBI's GenBank and SRA through a semi-automated process. Assembled whole genome sequencing projects are imported from GenBank and metagenome and metatranscriptome projects are imported from SRA database. In order to provide our users with a visual comparison between the number of projects that are available in NCBI/SRA and GOLD and subsequently in IMG, we have developed an NCBI Import Tracker (Figure 5).

Type strain tracker

Type strain is the representative unit of a microbial species that is defined when a species name is first established as per the International Code of Nomenclature of Prokaryotes (26). Type strains have well characterized phenotypes and established taxonomic information. Type strain genome sequence provides key insights into the phylogeny and systematics of well characterized representative microorganisms and are used as a reference for other organisms. Additionally, type strains are maintained and widely distributed by culture collection centers, which further promotes follow-up studies. To provide users an overview of prokaryotic type strains with and without publicly available genome sequences, we developed a Type Strain tracker that is available on GOLD homepage. Type Strain tracker displays the number of type strains that have been sequenced as well as those that are yet to be sequenced. There are several large and small-scale type strain sequencing initiatives completed or underway (5,27,28). By clicking on 'Type Strain Projects' link on GOLD home page, users can access the full list of type strain projects in GOLD.

External batch load Excels

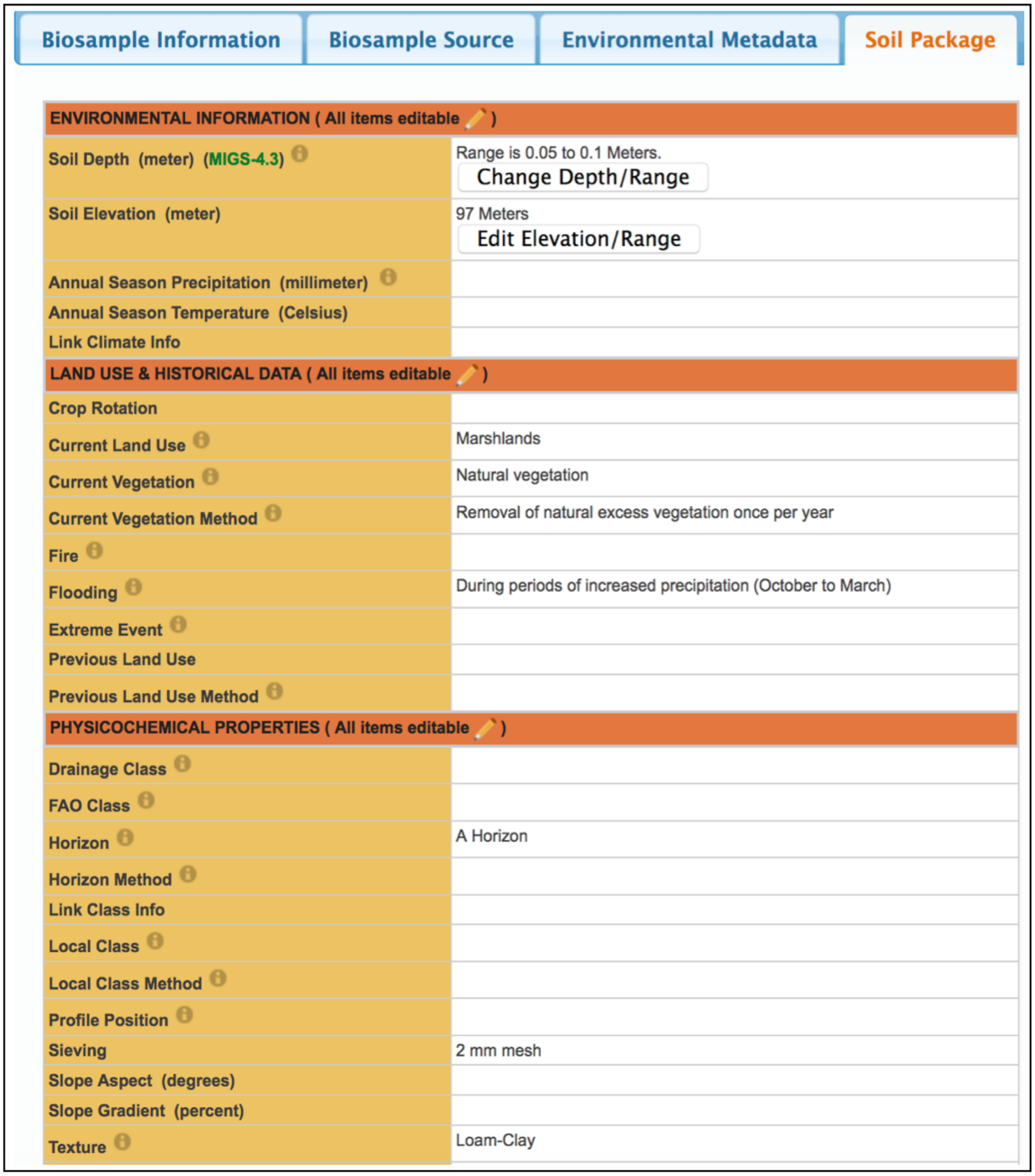
Researchers from around the world submit projects to GOLD. Individual researchers and consortia carrying out large number of projects typically use GOLD's web interface to enter new projects one at a time. It was not an issue in the past as the number of projects each individual or group entering are manageable and in cases where they have large number of projects they entered those over a period of time. With the ease of sequencing and increase in the number projects users are typically entering dozens of projects at a time. It is cumbersome to enter each project one-by-one. So, to help users to enter multiple projects in one go, we implemented a batch load option for external users. Typically, users contact us with such requests and we provide Excel formatted batch load file to suit their needs and facilitate loading of their projects in a batch.

Analysis Projects on public Sequencing Projects

Another feature we added in the current release is the option to create Analysis Projects on public Sequencing Projects. For example, a user may want to generate a new assembly on a public genome or metagenome and want to get it annotated in IMG. Accordingly, now a user can create Analysis Project on any public Sequencing Project and submit data to IMG. Users can add these newly created APs to their existing studies or if that does not apply, a new Study can be created for these APs.

Downloadable biosample geolocation map

In response to user feedback, we developed a new feature that takes results from advanced search to create figures showing the location of selected Biosamples on a world map. For example, a user can select Ecosystem filters of 'Environmental', 'Aquatic', 'Freshwater' and 'Lake' on Biosample section of advanced search, to get a matching list of Biosamples. By clicking on 'Create Biosample Map' a



ENVIRONMENTAL INFORMATION (All items editable)	
Soil Depth (meter) (MIGS-4.3)	Range is 0.05 to 0.1 Meters. Change Depth/Range
Soil Elevation (meter)	97 Meters Edit Elevation/Range
Annual Season Precipitation (millimeter)	
Annual Season Temperature (Celsius)	
Link Climate Info	
LAND USE & HISTORICAL DATA (All items editable)	
Crop Rotation	
Current Land Use	Marshlands
Current Vegetation	Natural vegetation
Current Vegetation Method	Removal of natural excess vegetation once per year
Fire	
Flooding	During periods of increased precipitation (October to March)
Extreme Event	
Previous Land Use	
Previous Land Use Method	
PHYSICOCHEMICAL PROPERTIES (All items editable)	
Drainage Class	
FAO Class	
Horizon	A Horizon
Horizon Method	
Link Class Info	
Local Class	
Local Class Method	
Profile Position	
Sieving	2 mm mesh
Slope Aspect (degrees)	
Slope Gradient (percent)	
Texture	Loam-Clay

Figure 3. Soil metadata package in GOLD. GOLD Biosample using Soil package. Representative metadata fields from the soil package are displayed here.

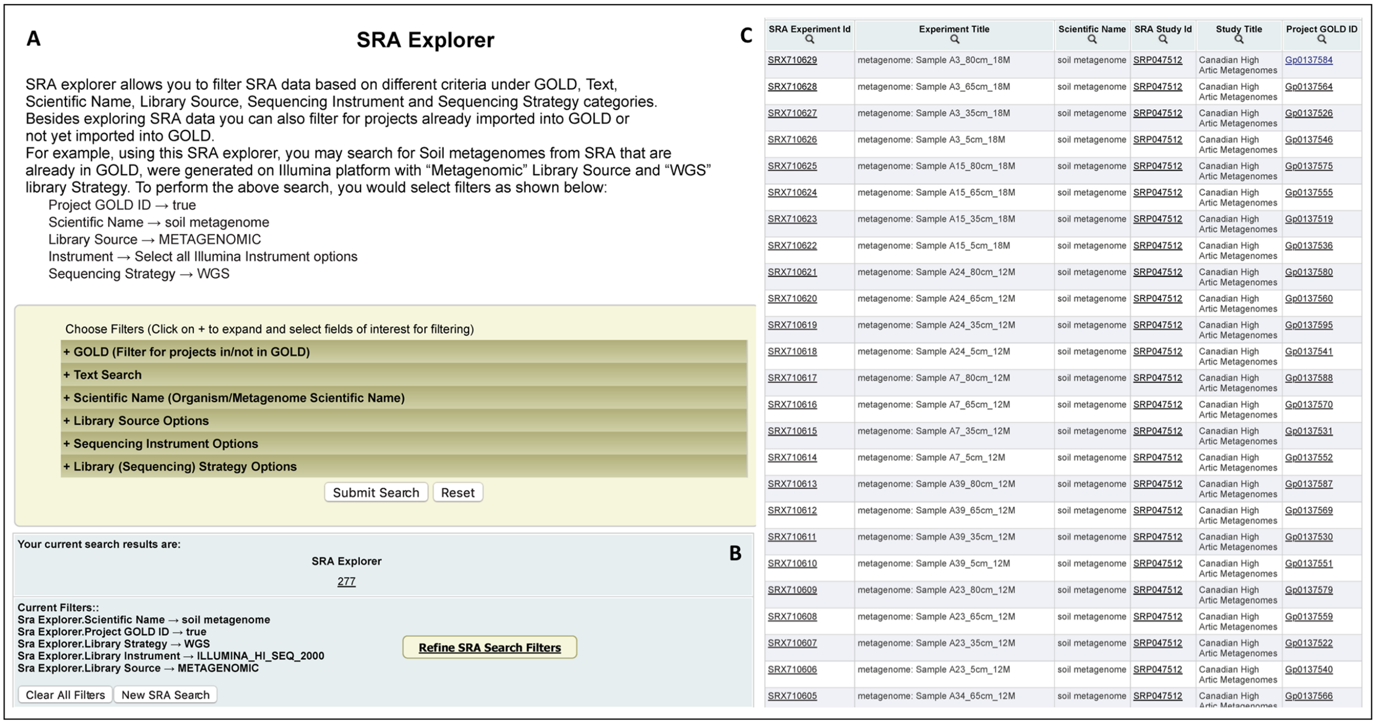
downloadable image of the location of freshwater Biosamples in GOLD (Figure 6) can be generated. Additional filters can be applied at Study, Biosample, Sequencing Project and Analysis Project level to get a list of interested biosamples for plotting on map. These images can be downloaded and saved for use in presentations or publications.

Downloadable genome report tables

Researchers regularly use metadata from GOLD to prepare manuscripts for journals such as Standards in Genomic Sciences (SIGS) and Genome Announcements. This involves gathering metadata from multiple fields. This can be both time consuming and can lead to errors in copy pasting. To address these issues, we implemented a new feature to automatically generate genome report tables and make them available for download. This feature compiles a pre-defined list of metadata fields from Organism/Biosample, Sequencing Project and Analysis Project to create customized tables for genome/metagenome reports. Genome report tables can be accessed by clicking on ‘View Genome/Metagenome Report Tables’ on GOLD Analysis Projects.

Dynamic map view and geographic coordinate lookup

Latitude and longitude values are required fields for Biosamples and are highly recommended for GOLD Organism entries. To help users enter the correct values and


A
B
C
Choose Filters (Click on + to expand and select fields of interest for filtering)
+ GOLD (Filter for projects in/not in GOLD)
+ Text Search
+ Scientific Name (Organism/Metagenome Scientific Name)
+ Library Source Options
+ Sequencing Instrument Options
+ Library (Sequencing) Strategy Options
Submit Search
Reset
Refine SRA Search Filters
Clear All Filters
New SRA Search

SRA Explorer

SRA explorer allows you to filter SRA data based on different criteria under GOLD, Text, Scientific Name, Library Source, Sequencing Instrument and Sequencing Strategy categories. Besides exploring SRA data you can also filter for projects already imported into GOLD or not yet imported into GOLD.

For example, using this SRA explorer, you may search for Soil metagenomes from SRA that are already in GOLD, were generated on Illumina platform with "Metagenomic" Library Source and "WGS" library Strategy. To perform the above search, you would select filters as shown below:

Project GOLD ID → true
Scientific Name → soil metagenome
Library Source → METAGENOMIC
Instrument → Select all Illumina Instrument options
Sequencing Strategy → WGS

Your current search results are:

SRA Explorer

277

Current Filters::
Sra Explorer.Scientific Name → soil metagenome
Sra Explorer.Project GOLD ID → true
Sra Explorer.Library Strategy → WGS
Sra Explorer.Library Instrument → ILLUMINA_HI_SEQ_2000
Sra Explorer.Library Source → METAGENOMIC

SRA Experiment Id	Experiment Title	Scientific Name	SRA Study Id	Study Title	Project GOLD ID
SRX710629	metagenome: Sample A3_80cm_18M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137584
SRX710628	metagenome: Sample A3_65cm_18M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137564
SRX710627	metagenome: Sample A3_35cm_18M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137526
SRX710626	metagenome: Sample A3_5cm_18M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137546
SRX710625	metagenome: Sample A15_80cm_18M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137575
SRX710624	metagenome: Sample A15_65cm_18M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137555
SRX710623	metagenome: Sample A15_35cm_18M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137519
SRX710622	metagenome: Sample A15_5cm_18M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137536
SRX710621	metagenome: Sample A24_80cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137580
SRX710620	metagenome: Sample A24_65cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137560
SRX710619	metagenome: Sample A24_35cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137595
SRX710618	metagenome: Sample A24_5cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137541
SRX710617	metagenome: Sample A7_80cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137588
SRX710616	metagenome: Sample A7_65cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137570
SRX710615	metagenome: Sample A7_35cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137531
SRX710614	metagenome: Sample A7_5cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137552
SRX710613	metagenome: Sample A39_80cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137587
SRX710612	metagenome: Sample A39_65cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137569
SRX710611	metagenome: Sample A39_35cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137530
SRX710610	metagenome: Sample A39_5cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137551
SRX710609	metagenome: Sample A23_80cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137579
SRX710608	metagenome: Sample A23_65cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137559
SRX710607	metagenome: Sample A23_35cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137522
SRX710606	metagenome: Sample A23_5cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137540
SRX710605	metagenome: Sample A34_65cm_12M	soil metagenome	SRP047512	Canadian High Artic Metagenomes	Gp0137566

Figure 4. SRA Explorer. (**A**) SRA Explorer launch page with a description of how to launch search. (**B**) SRA Explorer search parameters and results. (**C**) List of SRA Explorer search results obtained along with GOLD project IDs.

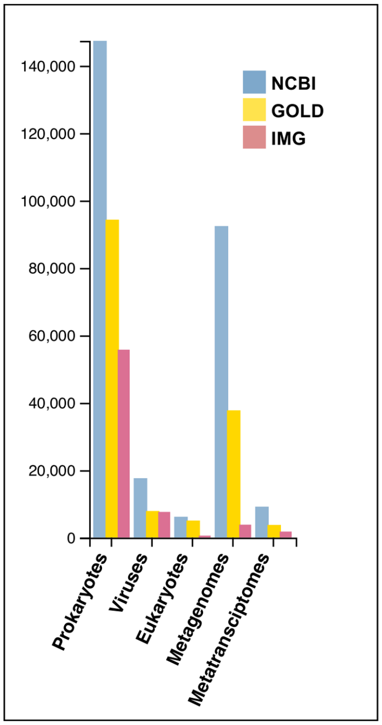


Figure 5. NCBI import tracker displaying the number of projects publicly available at NCBI and GOLD/IMG databases. Metagenome and metatranscriptome projects are displayed separately in the tracker, while WGS projects are displayed by domain: prokaryotes, eukaryotes and viruses.

prevent mistakes we created a dynamic map view during the Organism/Biosample creation process. When latitude and longitude fields are populated, a map is displayed below with a pin placed on the location corresponding to the values entered. This helps users to view and confirm the geolocation and correct errors, if any. Additionally, if the exact geographic coordinates are not available, we have designed a latitude/longitude lookup feature where a user can enter the name of the location and click on the 'Get Coords' button. This auto populates the latitude and longitude fields and provides an updated map view for the user to confirm.

Organism DOIs

Organism taxonomy information is important to accurately identify the organism being sequenced. Given an organism may be known by different vernacular names, synonyms in literature and across different labs, it is often leads to confusion about taxonomy information associated with a genome project. In order to address this potential ambiguity arising out of historical practices associated with conventional taxonomy and the use of different names in lab and literature, we worked with Names4Life (www.namesforlife.com) to obtain Digital Object Identifier (DOI) for organisms. Organism DOIs are stable and use of these identifiers in literature and public database repositories will help in accurate cross-referencing and comparative analysis. As a first step we obtained DOIs for type strains from Names4Life. For each type strain, we have Name DOI, Taxon DOI and Exemplar DOI. For example, GOLD organism (Go0000013)

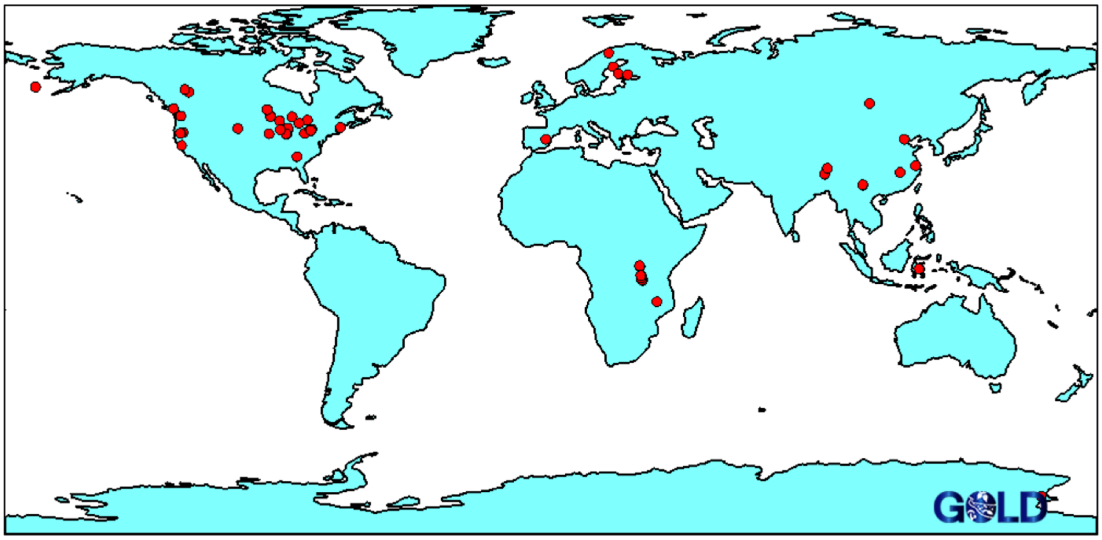


Figure 6. Geographic location map of freshwater Biosamples. Advanced search for Biosamples from freshwater environment was used and those results are plotted onto a downloadable geographic location map.

Corynebacterium kroppenstedtii DSM 44385 lists the following values on Organism detail page. Name DOI: 10.1601/nm.6233, Taxon DOI:10.1601/tx.6233 and Exemplar DOI: 10.1601/ex.6233.

ENTERING PROJECTS IN GOLD

GOLD is a login free open resource. However, in order to create projects in GOLD user registration is required. Registered users can access project entry interface from the homepage. At the beginning one need to specify if the material being sequenced is an Organism or Biome. After this step, a user is asked to create a GOLD Study, or choose from one of his existing Studies. The next step, before creating a Sequencing Project, is to create a GOLD Organism or Biosample. For example, an isolate genome Sequencing Project requires an Organism, which can either be selected from a list of over 300 000 existing GOLD Organisms or a new one entered using the Organism entry form. For a new GOLD Organism, taxonomy-specific metadata such as phylum, genus, species, strain and the top two Ecosystem Classification levels are some of the required fields. Similarly, for a metagenome Sequencing project, a Biosample is required where habitat, geographic location, latitude, longitude, sample collection site and the top three Ecosystem Classification levels are some of the required fields. For new Organisms and Biosamples, users have the option to select the Soil or Water package or go with the default list of metadata fields. After an Organism or Biosample is defined the next steps involve creating a Sequencing Project followed by an Analysis Project. Detailed instructions on how to enter different types of Sequencing and Analysis Projects, along with screenshots, is available in the GOLD help document at https://gold.jgi.doe.gov/resources/project_help_doc.pdf. Additionally, GOLD metagenomes follow a canonical naming system, details of which can be accessed in the metagenome naming guide at https://gold-dev.jgi.doe.gov/resources/Standardized_Metagenome_Naming.pdf.

FUTURE PLANS

Handling the increase in genome and metagenome projects from around the world, with the need to accurately record metadata and disseminate information to research community is a balancing act between the genomical growth, implementing standards for metadata and user experience. In an effort to better manage this we plan to undertake the following future development plans.

Semi-automated metadata curation

Manual curation steps are the rate limiting steps in importing projects from public resources. For genome projects, taxonomy information which accurately identifies underlying organism being sequenced and for metagenome projects, habitat, location and communities sequenced (bacterial, viral) are critical aspects of metadata that require manual curation. Projects/samples come with a variety of metadata fields often with synonymous names. Users submit metadata under different existing fields at NCBI or under newly defined attributes for existing fields with a case(upper/lower) difference or underscore in field names *etc*. Sifting through this nuanced metadata attributes/fields and mapping/parsing them to standardized GOLD metadata fields is a time consuming, but very important step for accurately recording metadata. Automating some of the steps in this metadata curation process will go a long way in addressing these rate limiting steps of metadata curation. We aim to explore and implement semi-automated processes in metadata curation using text parsing augmented with semantic dictionary and/or machine learning approaches.

Downloadable search results

Rich, manually curated metadata from GOLD is widely used by users who are studying few genomes at a time to those who are conducting comparative analysis of hundreds of genomes or metagenomes. At present users access individual records one at a time or contact GOLD for customized metadata exports. This diverts our limited curation resources to on demand metadata exports. To ease this burden and streamline user experience, we plan to implement an option to download search results directly from GOLD webpage. This should eliminate the need to access one record at a time on web or contact GOLD for bulk data exports.

Additional packages

As described above, we implemented water and soil packages in the current version of GOLD. We will be promoting the use of these packages and also seeking user feedback. This will drive our plans to include additional packages from MixS environmental package group such as the sediment, wastewater and host-associated packages in future.

Batch metadata updater

Users enter projects either online or via a batch load process. As additional metadata becomes available or existing data need to be corrected, users need to update individual projects and/or contact GOLD curators for bulk updates. Providing an option to submit additional metadata and/or updates in a batch file will ease the burden of updating individual entries on web and will streamline metadata acquisition process.

ACKNOWLEDGMENTS

We are thankful to the research community for submitting detailed metadata to GOLD or via their submissions to other public resources. We thank members of the microbial genomics and metagenomics research communities for their constant support, feedback and helpful discussions. Visualizations were generated using MS-Office Suite, GNU Image Manipulation Program (GIMP) and Adobe Acrobat Professional.

FUNDING

This work was conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility [DE-AC02-05CH11231]. Funding for open access charge: Office of Science of the U.S. Department of Energy [DE-AC02-05CH11231].

Conflict of interest statement. None declared.

REFERENCES

1. Mukherjee,S., Stamatis,D., Bertsch,J., Ovchinnikova,G., Verezemska,O., Isbandi,M., Thomas,A.D., Ali,R., Sharma,K., Kyrpides,N.C. *et al.* (2017) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res.*, **45**, D446–D456.
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. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5463–5467.
4. Shendure,J., Balasubramanian,S., Church,G.M., Gilbert,W., Rogers,J., Schloss,J.A. and Waterston,R.H. (2017) DNA sequencing at 40: past, present and future. *Nature*, **550**, 345–353.
5. Mukherjee,S., Seshadri,R., Varghese,N.J., Eloe-Fadrosh,E.A., Meier-Kolthoff,J.P., Göker,M., Coates,R.C., Hadjithomas,M., Pavlopoulos,G.A., Paez-Espino,D. *et al.* (2017) 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.*, **35**, 676–683.
6. Koepfli,K.-P., Paten,B. and Genome 10K Community of ScientistsGenome 10K Community of Scientists and O'Brien,S.J. (2015) The genome 10K project: a way forward. *Annu. Rev. Anim. Biosci.*, **3**, 57–111.
7. Lewin,H.A., Robinson,G.E., Kress,W.J., Baker,W.J., Coddington,J., Crandall,K.A., Durbin,R., Edwards,S.V., Forest,F., Gilbert,M.T.P. *et al.* (2018) Earth BioGenome project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 4325–4333.
8. 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.
9. Gilbert,J.A., Jansson,J.K. and Knight,R. (2014) The Earth Microbiome project: successes and aspirations. *BMC Biol.*, **12**, 69.
10. Anantharaman,K., Brown,C.T., Hug,L.A., Sharon,I., Castelle,C.J., Probst,A.J., Thomas,B.C., Singh,A., Wilkins,M.J., Karaoz,U. *et al.* (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.*, **7**, 13219.
11. Lund,J.B., List,M. and Baumbach,J. (2017) Interactive microbial distribution analysis using BioAtlas. *Nucleic Acids Res.*, **45**, W509–W513.
12. Klemetsen,T., Raknes,I.A., Fu,J., Agafonov,A., Balasundaram,S.V., Tartari,G., Robertsen,E. and Willassen,N.P. (2018) The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.*, **46**, D692–D699.
13. Sibbald,S.J. and Archibald,J.M. (2017) More protist genomes needed. *Nat. Ecol. Evol.*, **1**, 145.
14. Díaz-Muñoz,S.L. (2017) Viral coinfection is shaped by host ecology and virus–virus interactions across diverse microbial taxa and environments. *Virus Evol.*, **3**, vex011.
15. Chen,I.-M.A., Markowitz,V.M., Chu,K., Palaniappan,K., Szeto,E., Pillay,M., Ratner,A., Huang,J., Andersen,E., Huntemann,M. *et al.* (2017) IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.*, **45**, D507–D516.
16. 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.
17. 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 (x) sequence (MIxS) specifications. *Nat. Biotechnol.*, **29**, 415–420.
18. 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.
19. Langdon,W.B. (2014) Mycoplasma contamination in the 1000 genomes project. *BioData Min.*, **7**, 3.
20. Ivanova,N., Tringe,S.G., Liolios,K., Liu,W.-T., Morrison,N., Hugenholtz,P. and Kyrpides,N.C. (2010) A call for standardized classification of metagenome projects. *Environ. Microbiol.*, **12**, 1803–1805.
21. 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.
22. Parks,D.H., Rinke,C., Chuvochina,M., Chaumeil,P.-A., Woodcroft,B.J., Evans,P.N., Hugenholtz,P. and Tyson,G.W. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.*, **2**, 1533–1542.
23. Brown,C.T., Hug,L.A., Thomas,B.C., Sharon,I., Castelle,C.J., Singh,A., Wilkins,M.J., Wrighton,K.C., Williams,K.H. and Banfield,J.F. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, **523**, 208–211.

24. Shen,J., Chen,C.R. and Lewis,T. (2016) Long term repeated fire disturbance alters soil bacterial diversity but not the abundance in an Australian wet sclerophyll forest. *Sci. Rep.*, **6**, 19639.
25. Vigneron,A., Bishop,A., Alsop,E.B., Hull,K., Rhodes,I., Hendricks,R., Head,I.M. and Tsesmetzis,N. (2017) Microbial and isotopic evidence for methane cycling in Hydrocarbon-Containing groundwater from the pennsylvania region. *Front. Microbiol.*, **8**, 593.
26. 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.
27. Whitman,W.B., Woyke,T., Klenk,H.-P., Zhou,Y., Lilburn,T.G., Beck,B.J., De Vos,P., Vandamme,P., Eisen,J.A., Garrity,G. *et al.* (2015) Genomic encyclopedia of bacterial and archaeal type strains, phase III: the genomes of soil and plant-associated and newly described type strains. *Stand. Genomic Sci.*, **10**, 26.
28. 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.