

UC Merced

UC Merced Previously Published Works

Title

Genomes OnLine Database (GOLD) v.10: new features and updates

Permalink

<https://escholarship.org/uc/item/2bp3t0kq>

Authors

Mukherjee, Supratim

Stamatis, Dimitri

Li, Cindy Tianqing

et al.

Publication Date

2024-11-05

DOI








10.1093/nar/gkae1000

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Genomes OnLine Database (GOLD) v.10: new features and updates

Supratim Mukherjee , Dimitri Stamatis , Cindy Tianqing Li , Galina Ovchinnikova , Mahathi Kandimalla, Van Handke, Anuha Reddy, Natalia Ivanova, Tanja Woyke, Emiley A. Elor-Fardosh , I-Min A. Chen , Nikos C. Kyrpides and T.B.K. Reddy *

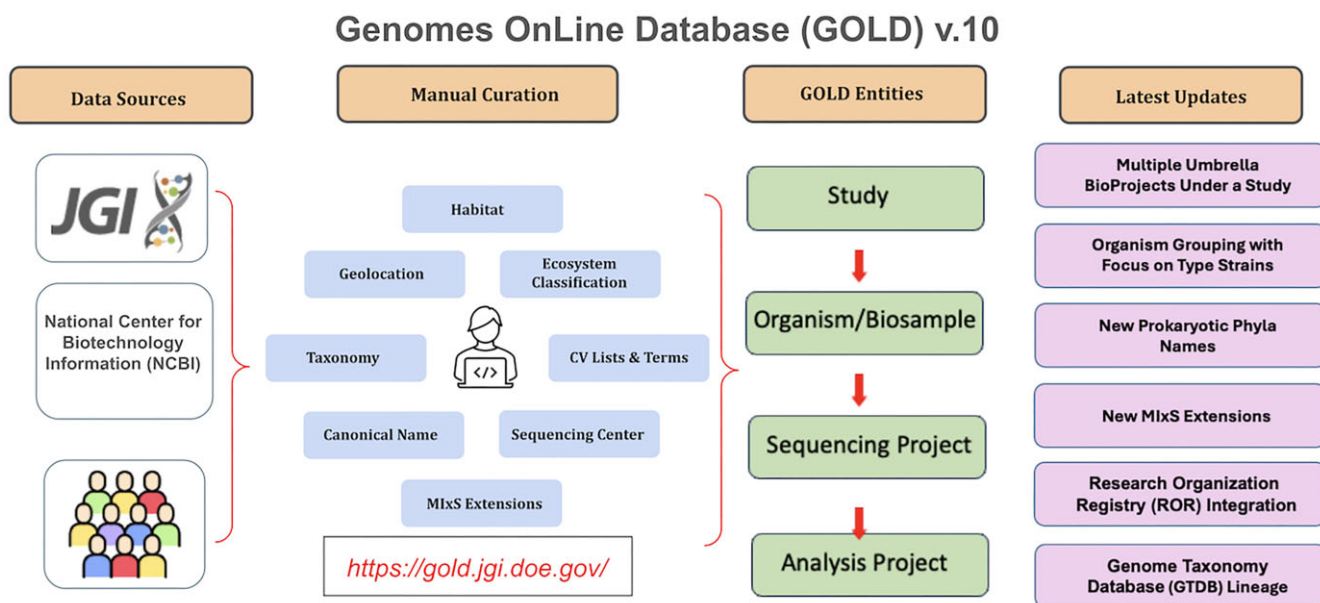
DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

*To whom correspondence should be addressed. Tel: +1 510 495 8400; Email: tbreddy@lbl.gov

Abstract

The Genomes OnLine Database (GOLD; <https://gold.jgi.doe.gov/>) at the Department of Energy Joint Genome Institute is a comprehensive online metadata repository designed to catalog and manage information related to (meta)genomic sequence projects. GOLD provides a centralized platform where researchers can access a wide array of metadata from its four organization levels namely Study, Organism/Biosample, Sequencing Project and Analysis Project. GOLD continues to serve as a valuable resource and has seen significant growth and expansion since its inception in 1997. With its expanded role as a collaborative platform, it not only actively imports data from other primary repositories like National Center for Biotechnology Information but also supports contributions from researchers worldwide. This collaborative approach has enriched the database with diverse datasets, creating a more integrated resource to enhance scientific insights. As genomic research becomes increasingly integral to various scientific disciplines, more researchers and institutions are turning to GOLD for their metadata needs. To meet this growing demand, GOLD has expanded by adding diverse metadata fields, intuitive features, advanced search capabilities and enhanced data visualization tools, making it easier for users to find and interpret relevant information. This manuscript provides an update and highlights the new features introduced over the last 2 years.

Graphical abstract



Introduction

Genomes OnLine Database (GOLD) is a comprehensive metadata management resource designed to facilitate the study and analysis of genomic and metagenomic information across the domains of life. Initially launched to record genome sequences in a tabular format (1), GOLD has matured into a relational

database that provides free access to an expansive list of metadata for researchers in genomics, microbiology and related fields (2). As genomes and metagenomes have increased in volume and complexity, the importance of available and accurate metadata attributes has also grown; without these, genomic data have limited value (3). By offering detailed descriptions

of sequencing projects, organisms, samples and experimental conditions, GOLD's structured metadata provides essential contextual information that enhances the use, reuse and comparative analyses of meta(genomic) data.

One of the primary applications of GOLD is supporting comparative genomics by providing a comprehensive repository of genomic and metagenomic projects, along with advanced search and retrieval tools, standardized metadata and integration with other bioinformatics resources. These features enable researchers to perform detailed comparative analyses across a wide range of organisms and genomic datasets. For example, Hackmann *et al.* (4) used organism-specific metadata from GOLD in conjunction with the Integrated Microbial Genomes and Microbiomes system (IMG/M) (5), BacDive (6), HydBDB (7) and other resources to perform a comprehensive analysis of fermentative prokaryotes, identifying knowledge gaps in our understanding of this complex metabolic process. In another example, Torrence *et al.* (8) used 35 metadata fields from 163 GOLD organisms to study the relationship between bacterial homologous recombination rates and specific genomic or phenotypic traits.

Over the past two years, the GOLD database has significantly expanded its project and metadata coverage and capabilities. Since the last release in November 2022, 86 930 Sequencing Projects (SPs) and 63 311 Analysis Projects (APs) were added representing an increase of 15% and 14.6%, respectively, compared with the previous release. This expansion includes the integration of new parameters related to environmental conditions, experimental methodologies and genomic annotations. GOLD has also substantially enhanced its metadata fields, providing a richer and more detailed context for the (meta)genomic data it houses. Throughout this growth phase, GOLD has continued to adhere to the community-driven standards developed by the Genomic Standards Consortium (GSC) (9). The expansion involved the integration of new MIxS [Minimum Information about any (x) Sequence] (10) extensions that enhance the ability to capture and describe the environmental and contextual information associated with genomic sequences. In addition to the MIxS extensions, GOLD integrated the Registry of Research Organizations (ROR) to standardize the curation of sequencing centers and funding agencies. This new feature not only helps in tracking contributions and collaborations across the global research community but also facilitates better data recording and sharing. These advancements reflect GOLD's commitment to evolving in response to the needs of the scientific community, ensuring that it remains at the forefront of metadata management for genomic and metagenomic projects and analysis. The continuous enhancement of the database underscores its pivotal role in advancing our understanding of (meta)genomes and supporting ongoing research efforts worldwide. The current status of the database along with all of its recent updates and improvements are described in detail in the following sections.

GOLD by the numbers

While the actual number of GOLD entities along with their associated metadata fields continue to increase at a steady rate, the overall organization and structure of the database remain unchanged. As described in more detail in earlier publications (2,11), GOLD is broadly organized into a four-level system namely Study, Biosample/Organism, SP and AP. These four

Table 1. Summary of GOLD entities and their metadata field counts. Biosamples and Organisms share ~500 terms from 11 MIxS extensions

GOLD entity	Counts	No. of metadata fields
Studies	61 296	26
Biosamples	208 629	720
Organisms	515 833	770
SPs	572 093	60
APs	432 186	53

levels contain a growing list of hundreds of metadata fields including >180 controlled vocabularies (CV) with >5600 terms.

A Study represents the broader goals of a research proposal. It describes the overall objectives of the SPs that it contains and is placed at the top of GOLD's four-level classification system. As of August 2024, there are 61 300 Studies in GOLD, ~4200 of which contain metagenome projects. Biosample refers to the physical material collected from the environment (i.e. environmental samples) while living biological material such as bacteria, archaea, fungi, plants or animals is termed as an Organism. GOLD currently has >200 000 Biosamples and 515 000 Organisms containing >700 metadata fields each (Table 1). The sequencing output of a Biosample (for metagenome or metatranscriptome) or Organism (isolate whole genome sequencing or similar) makes up an SP, and the subsequent analysis and data processing methods are detailed in an AP. As of August 2024, there are >570 000 SPs and 430 000 APs in GOLD (Table 1).

This organization structure establishes a logical flow of information from the initial proposal concept to the generated data, ensuring that all aspects of SPs and their related metadata are connected in a logical, coherent and systematic manner.

Updates since the last release

New prokaryotic phyla names

The rank phylum has been widely used and accepted in scientific literature for several decades. However, it was not formally recognized in the International Code of Nomenclature for Prokaryotes (ICNP). In a recent update, the rank 'phylum' was officially included in ICNP (12) and the suffix '-ota' was added to the stem of all formal rank names. To conform with these changes, the names of 42 prokaryotic phyla were updated in the current version of GOLD. This affected >485 000 bacterial and archaeal records in the database, and commonly used terms like *Actinobacteria*, *Firmicutes* and *Proteobacteria* were changed to newly formalized names such as *Actinomycetota*, *Bacillota* and *Pseudomonadota*, respectively. While this update created some confusion to users who were not aware of the new phylum names, it was necessary to make GOLD data FAIR (13) and to follow the established taxonomic standards for organisms.

Short 'how-to?' video tutorials

To educate new users and enhance the experience of existing users, we recently produced a series of concise video tutorials that guide users through several of the database's features and functionalities (Figure 1). These short 'how-to' videos offer clear instructions on how to navigate different parts of the website, from basic search techniques to registering

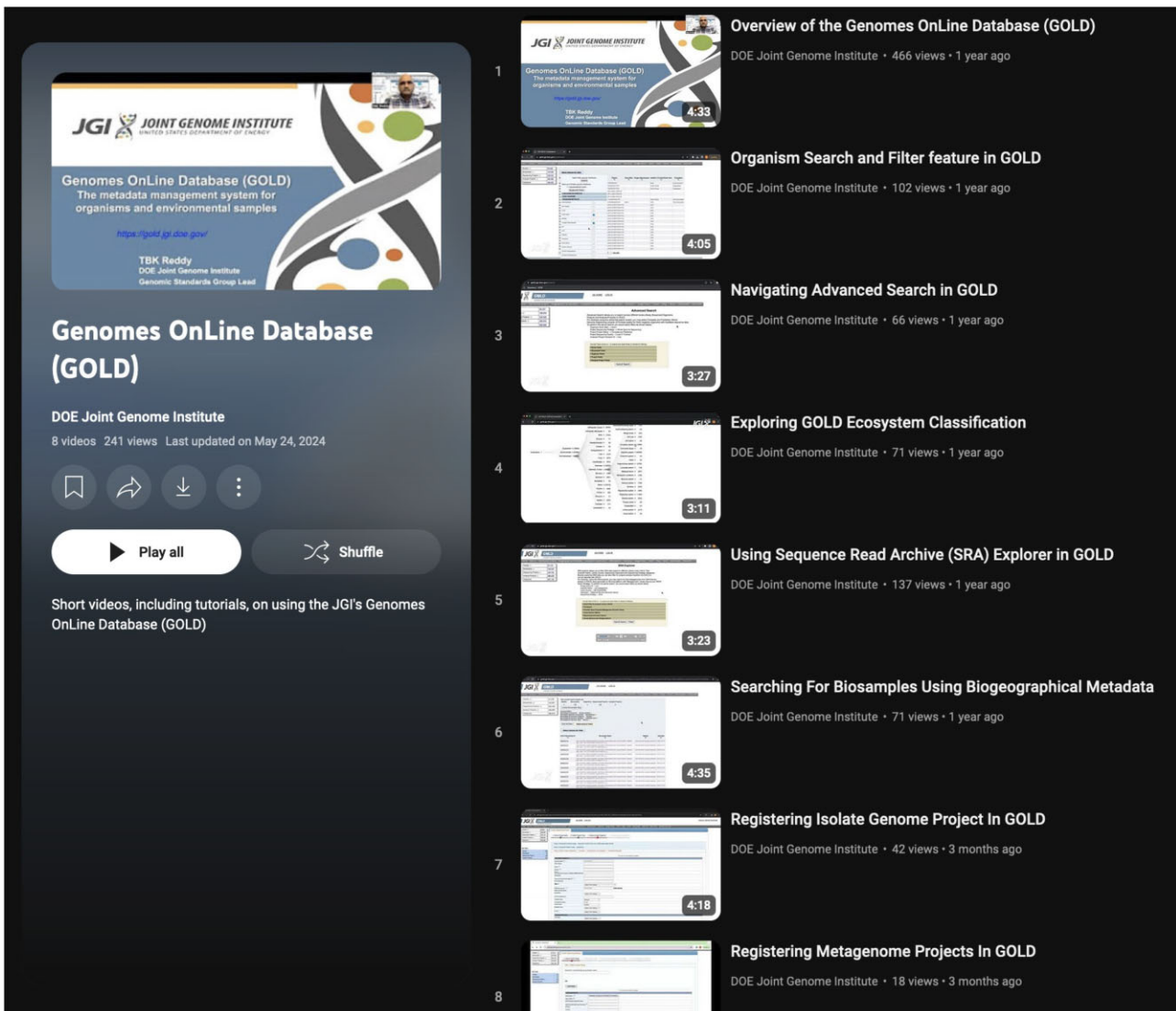


Figure 1. Overview of the landing page of GOLD's YouTube channel containing 8 short video tutorials on how to use its different features.

isolate and metagenome projects. By providing visual demonstrations and easy-to-follow explanations, the tutorials aim to empower users of all skill levels to efficiently access and to utilize the wealth of information available in GOLD. These videos can be accessed in our YouTube channel available at the following URL: <https://www.youtube.com/playlist?list=PLkxZMDuKlaKsMt2m2vYWN7AuggXDjwlt>. This initiative reflects GOLD's commitment to educate and support its growing user base and to ensure that researchers can fully leverage the database's capabilities to advance their work.

Streamlined organism creation process

The recent overhaul of our Organism entry form marks a significant improvement in the ease and efficiency of entering individual organisms and their corresponding projects. The first new feature that users will notice in GOLD v.10 is a screening for genomes and metagenomes that are publicly available in National Center for Biotechnology Information (NCBI). Since we have a standard process to import public data from NCBI that is subsequently annotated in IMG, users are discouraged from manually entering such projects into GOLD.

Detailed guidance on how to search for existing NCBI projects in GOLD is now easily accessible from within the new Organism entry form.

We have also streamlined the process of adding new organisms by incorporating probing questions that determine which metadata fields will be presented and which of these fields will be mandatory. For example, if an Organism is isolated by a user (or their colleagues) who is populating the entry form, they are expected to have specific knowledge about isolation metadata such as collection date, geographic location, latitude, longitude, etc. Consequently, these fields are made mandatory (Figure 2A).

Conversely, if the organism is obtained from a culture collection such as the American Type Culture Collection (14) or Leibniz Institute DSMZ (<https://www.dsmz.de/>), this information may not be readily available to a user, and therefore these fields are not required in the organism entry form (Figure 2B). By simplifying the submission process, the New Organism Creation Wizard not only enhances the accuracy and consistency of data entry but also speeds up the inclusion of newly sequenced genomes into the database. This development reflects GOLD's ongoing commitment to improving user

A

Create Organism

Select the type of Organism:

Cultured

Uncultured

Isolated by me/colleague

Obtained from a culture collection

ISOLATION METADATA

★ Sample Collection Date * Month (mm): Day (dd): Year (yyyy) *

★ Sample Collection Site * *

★ Isolation Country/Ocean * Select from below...

Sample Isolation Comments

Collection Method

Contact Name * Required field

Contact Email * Required field

ENVIRONMENTAL METADATA

Ecosystem * Select from below...

Ecosystem Category *

★ Ecosystem Type *

Ecosystem Subtype

Specific Ecosystem

Ecosystem Suggestion

★ Geographic Location * *

Lat/Long Lookup Get Coords

★ Latitude * e.g. 12.3408

★ Longitude * e.g. -12.578 Reset Map

Lat/Long Information * Select from below...

B

Create Organism

Select the type of Organism:

Cultured

Uncultured

Isolated by me/colleague

Obtained from a culture collection

ISOLATION METADATA

Sample Collection Date Month (mm): Day (dd): Year (yyyy):

Sample Collection Site *

Isolation Country/Ocean * Select from below...

Sample Isolation Comments

Collection Method

Contact Name * Required field

Contact Email * Required field

ENVIRONMENTAL METADATA

Ecosystem * Select from below...

Ecosystem Category *

Ecosystem Type *

Ecosystem Subtype

Specific Ecosystem

Ecosystem Suggestion

Geographic Location *

Lat/Long Lookup Get Coords

Latitude * e.g. 12.3408

Longitude * e.g. -12.578 Reset Map

Lat/Long Information * Select from below...

Figure 2. Different versions of the Organism entry form are used based on the organism's origin. Specific metadata fields (marked with an asterisk) are mandatory if the organism was isolated by an individual (A) rather than obtained from a culture collection (B).

experience and supporting the expanding needs of the global research community.

New MIxS extensions

The GOLD has recently expanded its MIxS extensions (packages) to a total of 11 with the addition of three new MIxS environmental packages: Human Associated, Wastewater and Built Environment. This expansion introduced 246 new fields and 72 new CVs to GOLD's metadata repertoire. As their names suggest, the Human Associated package provides detailed metadata fields for capturing sample information specific to human-related environments, such as host diet, medical history, specific disease status and more facilitating research into human health and disease. The Wastewater package allows for the documenting metadata specific to microbial communities present in wastewater environments, thereby supporting studies on public health and environmental monitoring. Finally, the Built Environment package includes several new metadata fields and CV terms specific to constructed environments, including buildings and infrastructure, that may not be applicable for regular Biosamples or Organisms. These new MIxS packages enhance GOLD's ability to support a wide range of environmental and applied genomic research, providing more precise and contextually relevant data for scientists interested in these specific environments.

Organism grouping with emphasis on type strains

Type strains are specific strains of microorganisms that serve as reference points for the identification and classification of a species. They are the 'standard' or 'typical' representatives of a species, and their characteristics are used as the basis for defining and naming new isolates. Having a type strain or a point of reference is crucial for accurately identifying a novel species. Hundreds of thousands of new prokaryotic species are

described every year (15). When new type strains are identified, the International Committee on Systematics of Prokaryotes (16) mandates that they be deposited in two or more culture collections in different countries. As a result, although they are an individual strain of an organism, they end up receiving multiple culture collection specific identifiers. These are commonly known as equivalent strains, which are often shared and published as independent organisms over time. As the number of distinct organisms continues to grow, so does the number of their corresponding equivalent organisms, highlighting the need to document these equivalent strains belonging to a single organism/type strain. In this current version of the database, we are in the process of grouping all organisms that are part of a synonymous group. While our initial focus is specifically on type strain species, we plan to expand the scope of grouping equivalent strains to all prokaryotic GOLD organisms. In the first phase of this grouping process, 51 440 equivalent type strains have been grouped into 18 217 distinct groups.

Integration of institutes from ROR

In the earlier versions of the GOLD, if a particular sequencing center was not available, a user had to manually send a request to our support team or select an incorrect institute. This was a time-consuming process and often resulted in SPs with incorrect metadata. To avoid this, GOLD v.10 has integrated >110 000 institutes from the ROR (17) into its list of institutions. This represents a significant advancement in expanding the network of sequencing centers, funding agencies and collaborating institutes. ROR provides a standardized and comprehensive ROR, enabling GOLD users to seamlessly incorporate and update information about institutions while adding or updating Sequencing Centers within GOLD's SPs. This integration has increased the number of available

institutions in GOLD by several orders of magnitude and enhanced our ability to catalog and track contributions from a diverse array of global research facilities. By linking genomic data with detailed organizational information, this assimilation of ROR institutes within GOLD facilitates more effective data sharing, ultimately advancing the collective efforts of the scientific community while enriching the resources available within GOLD.

Integrating Genome Taxonomy Database for enhanced prokaryotic classification

With the exponential increase in the number of available genome sequences, there is a growing interest in the prokaryotic community to supplement standard taxonomic nomenclature with genome-level taxonomy (18,19). This is particularly relevant now, since many high-quality novel genome sequences come from uncultured microorganisms, such as metagenome-assembled genomes (MAGs) and single cells. The Genome Taxonomy Database (GTDB) (20,21) is an initiative that infers microbial taxonomy from phylogeny based on conserved single copy marker proteins and rank normalization. GTDB-Tk is an open-source software toolkit (22) that assigns taxonomic classifications to prokaryotic genomes based on the GTDB. We added GTDB taxonomy lineages to APs by matching with the NCBI assembly accessions (when available) from GOLD to GTDB import. For genomes without assembly accessions, we obtain GTDB taxonomy from the IMG system. IMG generates GTDB lineages by using GTDB-Tk. Both NCBI taxonomy (23) lineage and GTDB lineages are displayed on the AP page. Panel (A) in Figure 3 shows an example where both NCBI and GTDB taxonomic lineages are identical, demonstrating consistent classification across both systems. Panel (B) shows an instance where GTDB taxonomy provides more specific classification down to the genus level, whereas the NCBI lineage is limited to the phylum level, highlighting the enhanced granularity of GTDB. In panel (C), we see a case where the NCBI and GTDB taxonomies conflict, showcasing discrepancies between the two systems, allowing users to review and potentially update genome classifications.

Stable isotope probing and low-complexity metagenomes

Limitations of conventional microbiological practices have led to the development of several innovative metagenome techniques that provide a deeper understanding of diversity, functions and interactions of microbial communities. Stable isotope probing (SIP) and low-complexity are two such types of metagenome projects that were recently included in GOLD v.10. SIP metagenome projects are designed to identify substrate-active microorganisms within complex environmental communities by tracking the incorporation of stable isotope-labeled substrates (e.g. carbon-13 or nitrogen-15) into their DNA or RNA (24). GOLD organizes SIP projects into a hierarchical structure consisting of SIP-Parent projects and their associated fractions. The SIP-Parent project represents the overall experimental setup and includes metadata about the environment, the type of stable isotope used, the labeled substrate and general project goals. This Parent project provides a broad overview of the study's objectives and the environmental context from which the samples/fractions were derived. Associated with each SIP-Parent project are multiple SIP-Fraction projects as shown in Figure 4. These frac-

tions correspond to specific density gradients obtained after the DNA or RNA has been separated based on its incorporation of the isotope. Each fraction represents a different level of isotope incorporation, reflecting the different microbial populations or genes that were metabolically active during the experiment. Metadata for SIP-Fractions includes details about the specific density of the fraction, the sequencing methods used and the level of isotopic enrichment. By querying SIP-Parent and fraction data in GOLD, researchers can link specific metabolic activities to specific microbial taxa, facilitating the study of active microbial functions in natural and engineered environments.

This setup in GOLD allows for detailed analysis of microbial roles in biogeochemical cycles, degradation processes and other ecosystem functions, significantly advancing our understanding of microbial ecology. Detailed instructions on how to search for individual SIP metagenome projects and their associated SIP-Fractions can be found in our updated Help section at the following URL: https://gold.jgi.doe.gov/resources/Metagenome_SIP_Help.pdf.

Low-complexity metagenomes represent microbial communities with a limited number of species, typically originating from environments where microbial diversity is naturally low, or from samples that have undergone adaptive selection or efforts to culture axenic strains, resulting in less complex communities. The simplicity of these communities makes them ideal for studying microbial interactions, metabolic pathways and the impact of environmental changes on microbial populations. By focusing on a smaller number of species, researchers can more easily link specific genes and functions to organisms leading to clearer insights into microbial ecology and evolution. Low-complexity metagenome projects in GOLD provide a valuable resource for scientists seeking to explore fundamental biological processes in controlled or extreme environments. Publicly available low-complexity metagenomes can be accessed at <https://gold.jgi.doe.gov/projects?page=1&Project.Sequencing+Strategy=Metagenome++Low+Complexity&count=25>.

Multiple umbrella BioProjects under one study

An umbrella BioProject at the NCBI is a type of BioProject that serves as a top-level, organizational entity for a group of related data-level projects. It acts as a container to aggregate various studies or datasets under a common theme or objective. For instance, the Human Microbiome Project is an umbrella BioProject that groups various sub-projects focused on studying different microbial communities in the human body, such as in the gut, skin or mouth (25). This organizational structure simplifies data sharing and improves collaboration among research groups working on interconnected themes. Similarly, a study in GOLD often contains several different types of projects that share a common goal or research question, functioning much like an NCBI umbrella BioProject. A single GOLD study may consist of several large-scale research initiatives or projects that span multiple organisms, environmental samples and conditions.

When Joint Genome Institute (JGI) users start publishing results for a subset of projects from large GOLD studies, we organize those subsets of data-level BioProjects under umbrella BioProjects at NCBI. Since more than one umbrella BioProject can be set up at NCBI for a given GOLD study to better organize related data and publications, we have implemented

ANALYSIS PROJECT METADATA	
NCBI Taxonomy Lineage	d__Archaea;p__Euryarchaeota;c__Methanomicrobia;o__Methanosarcinales;f__Methanosarcinaceae;g__Methanococoides;s__Methanococoides vulcani
GTDB Taxonomy Lineage	d__Archaea;p__Halobacteriota;c__Methanosarcinia;o__Methanosarcinales;f__Methanosarcinaceae;g__Methanococoides;s__Methanococoides vulcani
GTDB Version ID	220

A (Ga0668190)

ANALYSIS PROJECT METADATA	
NCBI Taxonomy Lineage	d__Bacteria;p__Pseudomonadota;c__o__f__g__s__uncultured Pseudomonadota bacterium
GTDB Taxonomy Lineage	d__Bacteria;p__Pseudomonadota;c__Alphaproteobacteria;o__Rs-D84;f__Rs-D84;g__Enterousia;s__Enterousia sp900313105
GTDB Version ID	220

B (Ga642446)

ANALYSIS PROJECT METADATA	
NCBI Taxonomy Lineage	d__Bacteria;p__Bacillota;c__Bacilli;o__Bacillales;f__Planococcaceae;g__Planococcus;s__Planococcus sp. APC 4015
GTDB Taxonomy Lineage	d__Bacteria;p__Actinomycetota;c__Actinomycetes;o__Actinomycetales;f__Microbacteriaceae;g__Microbacterium;s__Microbacterium sp030407505
GTDB Version ID	220

C (Ga664464)

Figure 3. Illustration of three scenarios comparing NCBI and GTDB taxonomies for prokaryotic genomes. Panel (A) shows an example where both taxonomic lineages are identical. Panel (B) highlights a case where GTDB taxonomy provides more detailed classification down to the genus level. Panel (C) presents a case where NCBI and GTDB taxonomies conflict, offering users an opportunity to review.

PROJECT COMPOSITION				
Study	Soil microbial communities from agricultural fields across the midwestern and eastern USA			
Biosample	Soil microbial communities from an agricultural field in Butler, South Dakota, USA - 87_897_SD_DR_0cm_c3_T1_16O_SIP_20			
Analysis Projects	Soil microbial communities from an agricultural field in Butler, South Dakota, USA - 87_897_SD_DR_0cm_c3_T1_16O_SIP_20			
SIP Projects	SIP Parent Gp0592553 → 19 SIP Fractions			
Number of Analysis Projects (AP)	1			
Number of Derived Analysis Projects (DAP)	0			
GOLD Project ID	Project Name	Project Status	Add Date	Sequencing Strategy
Gp0665695	Soil microbial communities from an agricultural field in Butler, South Dakota, USA - 87_897_SD_DR_0cm_c3_T1_16O_SIP_20	Permanent Draft	2022-07-29	Metagenome - SIP
Gp0665568	Soil microbial communities from an agricultural field in Butler, South Dakota, USA - 87_897_SD_DR_0cm_c3_T1_16O_SIP_5	Permanent Draft	2022-07-29	Metagenome - SIP
Gp0665566	Soil microbial communities from an agricultural field in Butler, South Dakota, USA - 87_897_SD_DR_0cm_c3_T1_16O_SIP_6	Permanent Draft	2022-07-29	Metagenome - SIP
Gp0665565	Soil microbial communities from an agricultural field in Butler, South Dakota, USA - 87_897_SD_DR_0cm_c3_T1_16O_SIP_7	Permanent Draft	2022-07-29	Metagenome - SIP
Gp0665564	Soil microbial communities from an agricultural field in Butler, South Dakota, USA - 87_897_SD_DR_0cm_c3_T1_16O_SIP_8	Permanent Draft	2022-07-29	Metagenome - SIP
Gp0665563	Soil microbial communities from an agricultural field in Butler, South Dakota, USA - 87_897_SD_DR_0cm_c3_T1_16O_SIP_9	Permanent Draft	2022-07-29	Metagenome - SIP
Gp0665557	Soil microbial communities from an agricultural field in Butler, South Dakota, USA - 87_897_SD_DR_0cm_c3_T1_16O_SIP_2	Permanent Draft	2022-07-29	Metagenome - SIP

Figure 4. Overview of a SIP-Parent metagenome project (Gp0592553) composition containing 19 individual SIP-Fractions, 7 of which are shown here.

the capability to associate more than one umbrella BioProject with an individual GOLD study. Figure 5 shows an example of a study with four umbrella BioProjects. The overall goal of the study is to develop a pan-genus genomic platform for wildflower genus *Mimulus* and determine the genomic basis for ecological adaptations under different environmental conditions. Each of the four umbrella BioProjects includes a combination of genome and transcriptome data from different *Mimulus* genotypes, tissues and habitats. For example, *Mimulus guttatus* IM62 is adapted to grow in mine tailings, thermal springs, serpentine soils and high salt environments.

Umbrella BioProject PRJNA112459 contains a single BioProject with *Mimulus guttatus* IM62 genome data and 12 different transcriptome BioProjects with gene expression data from different parts of the plant such as leaf, stem, flowers and more.

Help page updates

GOLD's help page (<https://gold.jgi.doe.gov/help>) provides new and existing users with a wide range of resources and tools to effectively utilize the database's features and max-

STUDY INFORMATION																					
GOLD Study ID	Gs0154496																				
Study Name ⓘ	Genomic Resources for Mimulus, a Powerful Plant System for Analyses of Environmental Adaptations																				
Other Names ⓘ																					
NCBI Umbrella BioProject ⓘ	PRJNA1112459 - Mimulus guttatus cultivar: IM62 genome assembly and annotation PRJNA1112462 - Mimulus nasutus cultivar:SF genome assembly and annotation PRJNA1112461 - Mimulus guttatus cultivar: IM767 genome assembly and annotation PRJNA1112463 - Mimulus tilingii cultivar: LVR genome assembly and annotation																				
SRA Studies	<table border="0"> <tr> <td>SRA Study Id</td> <td>SRP509593 (Link to NCBI)</td> </tr> <tr> <td>Study Title</td> <td>Mimulus nasutus SF Transcriptome - SF_R2</td> </tr> <tr> <td>Study Abstract</td> <td></td> </tr> <tr> <td>SRA Study Id</td> <td>SRP509594 (Link to NCBI)</td> </tr> <tr> <td>Study Title</td> <td>Mimulus nasutus SF Transcriptome - SF_B2</td> </tr> <tr> <td>Study Abstract</td> <td></td> </tr> <tr> <td>SRA Study Id</td> <td>SRP509591 (Link to NCBI)</td> </tr> <tr> <td>Study Title</td> <td>Mimulus nasutus SF Transcriptome - SF_ST1</td> </tr> <tr> <td>Study Abstract</td> <td></td> </tr> <tr> <td colspan="2">Show more...</td> </tr> </table>	SRA Study Id	SRP509593 (Link to NCBI)	Study Title	Mimulus nasutus SF Transcriptome - SF_R2	Study Abstract		SRA Study Id	SRP509594 (Link to NCBI)	Study Title	Mimulus nasutus SF Transcriptome - SF_B2	Study Abstract		SRA Study Id	SRP509591 (Link to NCBI)	Study Title	Mimulus nasutus SF Transcriptome - SF_ST1	Study Abstract		Show more...	
SRA Study Id	SRP509593 (Link to NCBI)																				
Study Title	Mimulus nasutus SF Transcriptome - SF_R2																				
Study Abstract																					
SRA Study Id	SRP509594 (Link to NCBI)																				
Study Title	Mimulus nasutus SF Transcriptome - SF_B2																				
Study Abstract																					
SRA Study Id	SRP509591 (Link to NCBI)																				
Study Title	Mimulus nasutus SF Transcriptome - SF_ST1																				
Study Abstract																					
Show more...																					

Figure 5. GOLD study with four separate NCBI umbrella BioProjects, each containing genome and transcriptome data for different cultivars of Mimulus.

imize its utility. We regularly update the help page to keep users informed about the latest developments. For example, we added new guidance on submitting public NCBI genomes and metagenomes into GOLD and detailed instructions on how to access Metagenome – SIP projects and their associated metadata in the ‘GOLD Documentation’ subsection. Description of new terms were added to the ‘GOLD Terminology’ subsection, including Metagenome – SIP specific terms like ‘SIP-Parent’, ‘SIP-Fraction’ and ‘qSIP’; while new videos and webinars made their way into the ‘How-to Short Videos and Workshops’.

JGI-USCCN

A new partnership was recently launched at the JGI in collaboration with the United States Culture Collection Network (USCCN) (26) allowing users to submit their isolate bacterial or archaeal strains for sequencing at the JGI. Individual users can nominate their organisms by completing a short form accessible from the URL (<https://gold.jgi.doe.gov/usccn>). Participation is open to anyone as long as the organism and the scientific questions are relevant to the Department of Energy mission (<https://jgi.doe.gov/user-programs/program-info/doe-mission-relevance/>).

Future plans

To meet the needs of the scientific community, proper planning is essential for GOLD to remain at the forefront of genomic

research and data management. As genomic science rapidly evolves, GOLD must keep pace and continuously update its infrastructure and features. Below are some of the future development initiatives that we have planned.

Homepage update

GOLD plays a crucial role in organizing and providing access to metadata on genome and metagenome projects. However, as scientific research evolves and the volume of data increases, the need to revamp or refresh GOLD’s homepage becomes essential.

By undertaking home page design, we aim to achieve improved usability, enhanced data accessibility, responsive design and integration with new technologies. Planned updates include an enhanced user interface of our landing page with a new design, cleaner layout and improved navigation elements. Additionally, the new home page will have redesigned menus and navigation bars to simplify access to GOLD’s core features such as search functions, metadata and project listings. Currently, we are gathering feedback from users and developing requirements for homepage redesign. This will be followed by the development and deployment of the redesigned home page.

Revamping the GOLD homepage will not only improve functionality but also improve the overall user experience, making the platform more user-friendly, efficient and capable of meeting the growing needs of the scientific community.

Using ROR IDs as sequencing centers

Currently, GOLD users start entering a sequencing center name to search and find a match in GOLD for the sequencing center of their choice. However, a search and selection by name can often be tedious or lead to mistakes by selecting similar names. Typos in search terms can result in no hits. We currently allow GOLD users to search ROR institutions by name, and we plan to further extend this to let users enter ROR id instead, eliminating the need to enter and search by text. A ROR ID consists of a unique nine-character string that is appended to the ROR domain. For example, the ROR ID for the JGI is '<https://ror.org/04xm1d337>'. The unique strings in ROR identifiers are assigned randomly and do not contain any specific organizational information. This approach will reduce errors that occur when a user selects a sequencing center by name.

Enhanced metadata standardization and curation

Enhanced metadata standardization and curation refers to improving the processes of organizing, maintaining and ensuring the accuracy and consistency of metadata associated with datasets, particularly in large-scale databases like GOLD. The key aspects of enhanced metadata standardization and curation are standardized metadata formats such as canonical metagenome sample names, CVs and ontologies, semi-automated data import processes, interoperability across databases through data exchange like NCBI submissions and APIs for providing programmatic access to power users. We continue to add more curated CV lists, improve Standard Operating Procedures (SOP) for data acquisition and curation, and expand public APIs to include more metadata fields for data exchange and sharing.

Training and outreach programs

We continue to regularly update and add new help documents in GOLD, developing new 'How-to' short videos to help current and new users. We also plan to present about GOLD at various meetings to broaden our outreach.

Importing isolate and MAGs based on the GTDB type material label

After integrating GTDB-Tk taxonomy to isolate and MAG projects in GOLD, we plan to continue the import of isolates and MAGs based on the GTDB type material label. This allows us to efficiently integrate genome projects according to their taxonomic classifications (missing representation in GOLD), using the GTDB which provides standardized taxonomy for bacteria and archaea.

Revamp of statistics and distribution graph pages

Revamping the statistics and distribution graphs in the GOLD is crucial for enhancing the user experience, data accessibility and the overall utility of the platform. As the volume and complexity of genome and metagenome projects grow, modernizing the way GOLD presents statistical insights becomes essential for researchers to effectively interpret and utilize the data. We aim to achieve improved data interpretation through revamped statics and distribution graphs and offer customizable visualizations.

SIP MixS extensions

The need for MixS extensions to SIP data in GOLD is driven by the growing importance of high-quality, standardized metadata for SIP experiments. SIP is a key method used to trace the metabolic pathways of microorganisms in complex environments, and ensuring that metadata from SIP experiments is standardized and easily shareable is crucial for advancing microbial ecology and biogeochemistry research. GSC is currently working on the MixS extensions for the SIP samples, and we plan to implement them once finalized (27).

National Microbiome Data Collaborative collaborations and metadata exchange

The collaboration between the National Microbiome Data Collaborative and the GOLD aims to streamline and enhance the way microbiome data are collected, curated and shared within the scientific community. Both platforms play pivotal roles in cataloging and organizing metadata related to microbial genomics. In particular, we would like to develop processes to capture metadata from NMDC's metadata submission for JGI projects (28).

Data availability

Data are freely available at <https://gold.jgi.doe.gov>.

Acknowledgements

We extend our gratitude to GOLD users and members of the microbiome research community for contributing metadata to the GOLD database. We are also grateful to the JGI project management team, as well as the microbial genomics and metagenomics programs, for their ongoing support and valuable feedback. Additionally, we thank the members of the microbial genomics and metagenomics research and standards communities for their insightful discussions and contributions. The visualizations presented in this manuscript were created using Snagit v. 2024.3.1, Adobe Acrobat Pro v. 2024.002.20687 and MS-Office Suite. This research was carried out by the U.S. Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility. Additionally, it utilized resources from the National Energy Research Scientific Computing Center, which is also supported by the DOE Office of Science.

Funding

U.S. Department of Energy Office of Science [DE-AC02-05CH11231]. Funding for open access charge: U.S. Department of Energy Office of Science.

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. Mukherjee, S., Stamatis, D., Li, C.T., Ovchinnikova, G., Bertsch, J., Sundaramurthi, J.C., Kandimalla, M., Nicolopoulos, P.A., Favognano, A., Chen, I.-M.A., *et al.* (2023) Twenty-five years of

- Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Res.*, **51**, D957–D963.
3. Huttenhower, C., Finn, R.D. and McHardy, A.C. (2023) Challenges and opportunities in sharing microbiome data and analyses. *Nat. Microbiol.*, **8**, 1960–1970.
 4. Hackmann, T.J. and Zhang, B. (2023) The phenotype and genotype of fermentative prokaryotes. *Sci. Adv.*, **9**, eadg8687.
 5. Chen, I.-M.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S.J., Webb, C., Wu, D., *et al.* (2023) The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.*, **51**, D723–D732.
 6. Reimer, L.C., Sardà Carbasse, J., Koblitz, J., Ebeling, C., Podstawka, A. and Overmann, J. (2022) BacDive in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res.*, **50**, D741–D746.
 7. Søndergaard, D., Pedersen, C.N.S. and Greening, C. (2016) HydDB: a web tool for hydrogenase classification and analysis. *Sci. Rep.*, **6**, 34212.
 8. Torrance, E.L., Burton, C., Diop, A. and Bobay, L.-M. (2024) Evolution of homologous recombination rates across bacteria. *Proc. Natl Acad. Sci. U.S.A.*, **121**, e2316302121.
 9. Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J.R., Dawyndt, P., Garrity, G.M., Gilbert, J., Glöckner, F.O., Hirschman, L., Karsch-Mizrachi, I., *et al.* (2011) The Genomic Standards Consortium. *PLoS Biol.*, **9**, e1001088.
 10. 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.
 11. Mukherjee, S., Stamatidis, D., Bertsch, J., Ovchinnikova, G., Sundaramurthi, J.C., Lee, J., Kandimalla, M., Chen, I.-M.A., Kyrpides, N.C. and Reddy, T.B.K. (2021) Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Res.*, **49**, D723–D733.
 12. Oren, A., Arahal, D.R., Rosselló-Móra, R., Sutcliffe, I.C. and Moore, E.R.B. (2021) Emendation of Rules 5b, 8, 15 and 22 of the International Code of Nomenclature of Prokaryotes to include the rank of phylum. *Int. J. Syst. Evol. Microbiol.*, **71**, 004851.
 13. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
 14. Nguyen, S.V., Puthuveetil, N.P., Petrone, J.R., Kirkland, J.L., Gaffney, K., Tabron, C.L., Wax, N., Duncan, J., King, S., Marlow, R., *et al.* (2024) The ATCC genome portal: 3938 authenticated microbial reference genomes. *Microbiol. Resour. Announc.*, **13**, e0104523.
 15. Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hershendorf, A.W., Amano, Y., Ise, K., *et al.* (2016) A new view of the tree of life. *Nat. Microbiol.*, **1**, 16048.
 16. Oren, A., Arahal, D.R., Göker, M., Moore, E.R.B., Rossello-Mora, R. and Sutcliffe, I.C. (2023) International Code of Nomenclature of Prokaryotes. Prokaryotic code (2022 Revision). *Int. J. Syst. Evol. Microbiol.*, **73**, 005585.
 17. Lammey, R. (2020) Solutions for identification problems: a look at the Research Organization Registry. *Sci. Editing*, **7**, 65–69.
 18. Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., Kyrpides, N.C. and Pati, A. (2015) Microbial species delineation using whole genome sequences. *Nucleic Acids Res.*, **43**, 6761–6771.
 19. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T. and Aluru, S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, 5114.
 20. Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J. and Hugenholtz, P. (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.*, **38**, 1079–1086.
 21. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A. and Hugenholtz, P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**, D785–D794.
 22. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P. and Parks, D.H. (2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, **36**, 1925–1927.
 23. Schoch, C.L., Ciuffo, S., Domrachev, M., Hottton, C.L., Kannan, S., Khovanskaya, R., Leippe, D., McVeigh, R., O'Neill, K., Robbertse, B., *et al.* (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*, **2020**, baaa062.
 24. Coyotzi, S., Pratscher, J., Murrell, J.C. and Neufeld, J.D. (2016) Targeted metagenomics of active microbial populations with stable-isotope probing. *Curr. Opin. Biotechnol.*, **41**, 1–8.
 25. 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.
 26. Boundy-Mills, K., Hess, M., Bennett, A.R., Ryan, M., Kang, S., Nobles, D., Eisen, J.A., Inderbitzin, P., Sitepu, I.R., Torok, T., *et al.* (2015) The United States Culture Collection Network (USCCN): enhancing microbial genomics research through living microbe culture collections. *Appl. Environ. Microbiol.*, **81**, 5671–5674.
 27. Simpson, A., Charlson, E.M.W., Smith, M., Koch, B., Beilsmith, K., Kimbrel, J., Kellom, M., Hunter, C., Walls, R.L., Scriml, L.M., *et al.* (2024) MISIP: a data standard for the reuse and reproducibility of any stable isotope probing-derived nucleic acid sequence and experiment. *GigaScience*, **13**, giae071.
 28. Eloë-Fadrosh, E.A., Ahmed, F., Anubhav, Babinski, M., Baumes, J., Borkum, M., Bramer, L., Canon, S., Christianson, D.S., Corilo, Y.E., *et al.* (2022) The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.*, **50**, D828–D836.