Integrated microbial database leading microbiome research

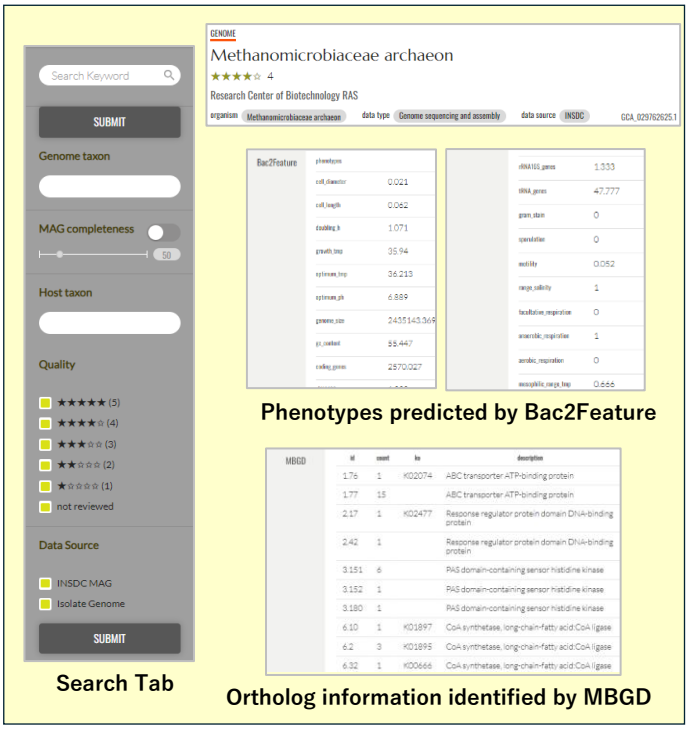
Microbiome Datahub

https://mdatahub.org/

Microbiome Datahub contains approximately 210,000 MAG data sets derived from INSDC and approximately 26,000 reference genomes from isolated bacterial strains in RefSeq, with related BioProject information. It allows users to refine MAG searches by phylogeny, genome quality, host phylogeny, etc., and to download sequence data, annotation information, etc. via web pages and APIs.

MAGs have received attention in recent years for exploring microbial diversity, as they can reveal the genomes of uncultured microorganisms. Microbiome Datahub collects MAGs published by INSDC, performs quality control, and provides 210,000 high-quality MAG data sets, including

(1) gene annotation information and sequence data analyzed by the microbial genome annotation tool DFAST, (2) metadata on isolation source of sample and strain characteristics, (3) ortholog information identified using the microbial comparative genome analysis database MBGD, and (4) predicted phenotypes such as cellular morphology, Gram staining properties, spore formation ability, motility, genomic features, and growth conditions predicted using Bac2Feature. Users can also use the API to obtain this information in bulk.



The most distinctive feature of Microbiome Datahub is the diversity of its data collection. While several existing databases archive MAGs, most focus predominantly on microbiomes derived from human-associated environments, such as the gut or skin. In contrast, Microbiome Datahub targets a wide range of environmental samples, including soil, seawater, freshwater, hot springs, sediments, and air, storing MAG data from a diverse microbial taxa.

When used in conjunction with Microbiome Datahub, PZLAST-MAG enables high-speed homology searches of approximately 450 million protein sequences derived from MAGs archived in Microbiome Datahub.

＜Number of data included in DB＞ (June 5, 2025)

Genome derived from isolated bacteria	26,076 data
Metagenome-Assembled Genome (MAG)	218,653 data

Metagenome-Assembled Genome (MAG): Genome sequences comprehensively analyzed without isolating and culturing microorganisms from soil, environmental water, feces, skin tissue, etc. are called “metagenomes.” Metagenomes contain fragmented genome sequences derived from a variety of microbial species that cannot be cultured in a laboratory. By assembling and clustering these fragmented sequences in a computer to reconstruct the original microbial genome, the resulting virtual sequence is referred to as a “Metagenome-Assembled Genome (MAG).” MAGs are expected to reveal previously unknown microorganisms, as well as contain useful information for industrial applications, such as novel enzyme genes possessed by these microorganisms.

Microbiome Datahub is developed in a project of JST Database Integration Coordination Program (DICP) "Development of an integrated microbiome data hub for microbiome research (PI: MORI Hiroshi, Associate Professor, National Institute of Genetics, Research Organization of Information and Systems)"

Reference

Mori H. *et al.*, Microbiome Datahub: an open-access platform integrating environmental metadata, taxonomy, and functional annotation for comprehensive metagenome-assembled genome datasets. Microbiome (2026)
DOI: 10.1186/s40168-026-02385-x

Japan Science and Technology Agency（JST）
Department of Information Services, Office of NBDC Program
Tel: 03-5214-8491　E-mail: nbdc-dicp@jst.go.jp

(Apr 2026)

Microbiome Datahub quick manual

Microbiome Datahub

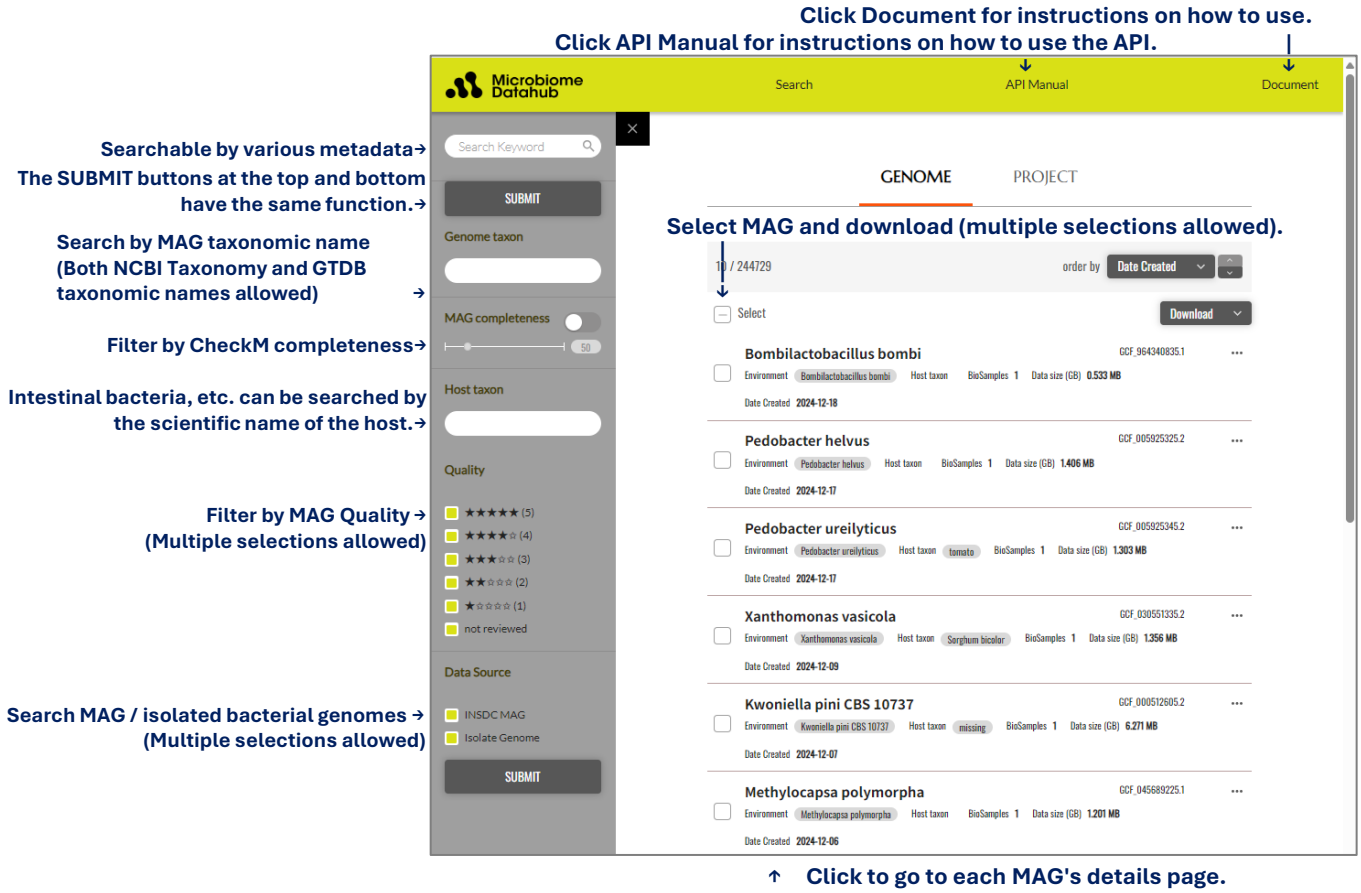
https://mdatahub.org/

Data available from Microbiome Datahub

- MAG metadata TSV
- MAG nucleotide sequence fasta
- MAG gene sequence fasta
- MAG gene amino acid sequence fasta
- MAG gene assigned MBGD Ortholog / KEGG Orthology ID list TSV

How to obtain data

1) Download by specifying the MAG ID in the URL using the API (see the API manual for details).
2) Download from a web browser (see below).



↑ Click to go to each MAG's details page.

Information available on the MAG details page

① **MAG metadata:** BioProject, BioSample, genome ID, taxonomic name, and other metadata in INSDCMAG
② **MBGD Orthology / KEGG Ortholog composition:** MAG gene MBGD Orthology ID and KO ID assigned by proprietary sequence similarity search (PZLAST-MAG) (*)
③ **MAG DFAST:** Statistics from the genome annotation tool
④ **DFASTDFAST_QC:** Results from the quality check tool DFAST_QC (CheckM)
⑤ **Bac2Feature results:** Results from the phenotype estimation tool Bac2Feature

* Individual MAG MBGD Orthology ID / KO ID composition TSVs can be obtained from the download API.
For example, if the MAG ID is GCA_029762515.1, it can be obtained from the following URL:
https://mdatahub.org/api/genome/mbgd/GCA_029762515.1

Japan Science and Technology Agency（JST）
Department of Information Services, Office of NBDC Program
Tel: 03-5214-8491　E-mail: nbdc-dicp@jst.go.jp

(Apr 2026)