

# AJACS MAG (Metagenome-Assembled Genome) を 知って・学んで・使う」

2024年7月25日

## メタゲノム・MAGデータを 検索する

森 宙史, Ph.D.

(Hiroshi Mori)

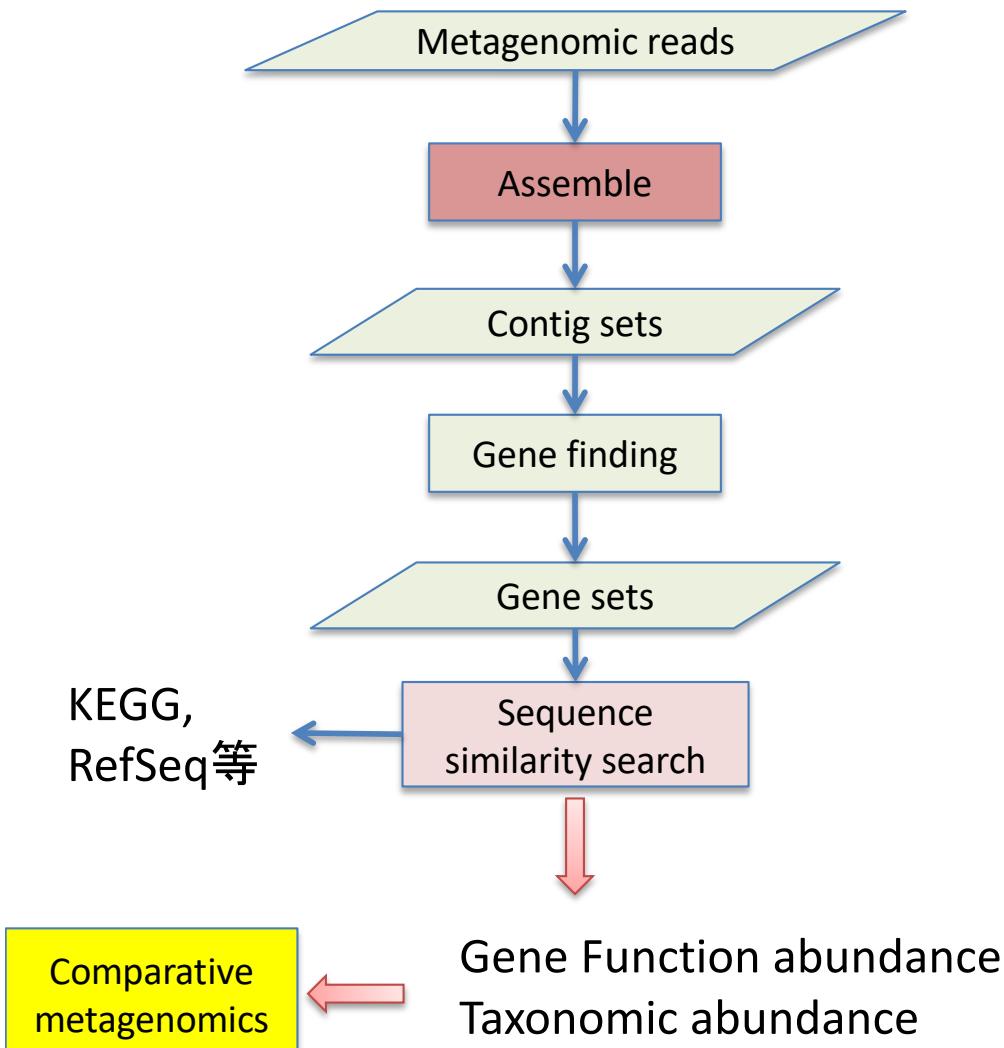
国立遺伝学研究所

先端ゲノミクス推進センター

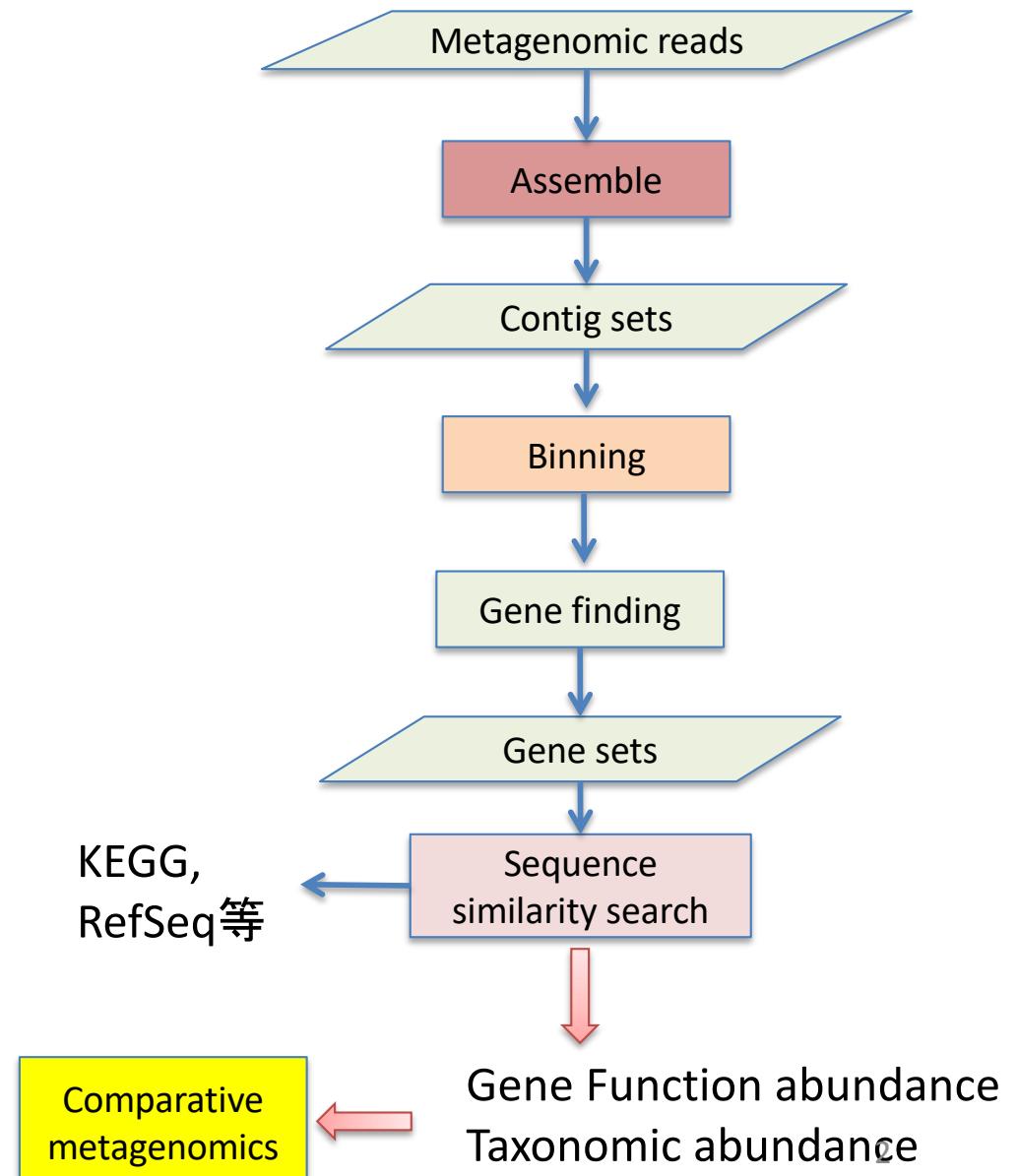
[hmori@nig.ac.jp](mailto:hmori@nig.ac.jp)

# 遺伝子組成解析とMAG解析

## Assembly approach



## Assembly + Binning approach



# 遺伝子機能組成解析

## 利点

- ・群集全体でどの系統由来のどのような遺伝子機能が多いかがわかる
- ・群間での統計的仮設検定による比較解析を行いやすい
- ・同じ種類の機能遺伝子を持つ系統の多様性を解析可能

## 欠点

- ・存在量の多い遺伝子の大半は必須遺伝子等全系統が持つ機能遺伝子
- ・マイナー系統が担う機能については統計的に有意な差は出にくい
- ・数百万遺伝子の解析なので情報解析が複雑になりがち

# MAG解析

## 利点

- ・単一種由来のドラフトゲノム配列を対象にするので群集中での頻度情報を考慮する必要が無い
- ・近縁種の既存のゲノム・MAGデータと比較ゲノム解析が可能
- ・数千遺伝子の解析なので情報解析が比較的単純

## 欠点

- ・rRNA遺伝子や水平伝播アイランド等MAGに含まれにくい遺伝子の存在
- ・キメラやコンタミの問題

# MAG関連のデータベース(DB)

様々なMAGをメタデータで検索し、配列データを取得する

The GTDB homepage features a large, detailed photograph of a tree's root system in the background. Overlaid on the top right is the logo of the Australian Centre for Ecogenomics, which consists of a circular arrangement of stylized DNA helixes.

**BACTERIA (394,932)**

Taxonomic Rank	Count
SPECIES	80,789
GENUS	19,153
FAMILY	4,264
ORDER	1,624
CLASS	488
PHYLUM	161

\*\*\* GTDB Release 214 is now available download files \*\*\*

\*\*\* GTDB-Tk for R214 will follow next week \*\*\*

Species distribution chart showing the count of genomes across taxonomic ranks:

- 20 PHYLUM
- 60 CLASS
- 148 ORDER
- 508 FAMILY
- 1,586 GENUS
- 4,416 SPECIES

Welcome to GTDB

# GENOME TAXONOMY DATABASE

402,709 genomes  
Release 08-RS214 (28th April 2023)

Archaea distribution chart showing the count of genomes across taxonomic ranks:

- 20 PHYLUM
- 60 CLASS
- 148 ORDER
- 508 FAMILY
- 1,586 GENUS
- 4,416 SPECIES

ARCHAEA (7,777)

**THE UNIVERSITY OF QUEENSLAND AUSTRALIA**

The GTDB homepage features a large background image of a tree trunk and branches. Overlaid on the top left is a summary of bacterial taxonomy counts: BACTERIA (584,382), SPECIES (107,235), GENUS (23,112), FAMILY (4,870), ORDER (1,840), CLASS (538), and PHYLUM (175). A green banner at the top provides release information: "\*\*\* GTDB Release 220 is now available [download files](#) \*\*\*" and "\*\*\* GTDB-Tk has been updated to use the R220 taxonomy from v2.4.0 \*\*\*". On the right side, there is a logo for the "Australian Centre for Ecogenomics" featuring a circular emblem with stylized DNA helixes. A sidebar on the right lists updates: "GTDB-Tkも Sourmashも CheckMも 变わるので、更新されたら MAG解析はやり直し". At the bottom, there is a summary of archaeal taxonomy counts: ARCHAEA (12,477), PHYLUM (19), CLASS (64), ORDER (166), FAMILY (564), GENUS (1,847), and SPECIES (5,869). Logos for The University of Queensland Australia and the Australian Centre for Ecogenomics are also present.

BACTERIA (584,382)

SPECIES ■ 107,235  
GENUS ■ 23,112  
FAMILY ■ 4,870  
ORDER ■ 1,840  
CLASS ■ 538  
PHYLUM ■ 175

\*\*\* GTDB Release 220 is now available [download files](#) \*\*\*

\*\*\* GTDB-Tk has been updated to use the R220 taxonomy from v2.4.0 \*\*\*

Australian Centre for Ecogenomics

GTDB-Tkも  
Sourmashも  
CheckMも  
变わるので、  
更新されたら  
MAG解析は  
やり直し

Welcome to GTDB

# GENOME TAXONOMY DATABASE

596,859 genomes  
Release 09-RS220 (24th April 2024)

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

19 ■ PHYLUM  
64 ■ CLASS  
166 ■ ORDER  
564 ■ FAMILY  
1,847 ■ GENUS  
5,869 ■ SPECIES

ARCHAEA (12,477)

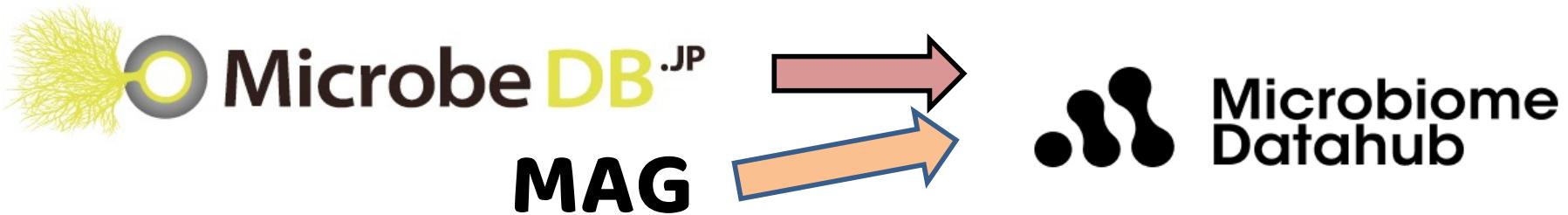
# 代表的なMAG DB間の比較

	Admin	Database URL	DNA sequence data	Protein sequence data	Number of MAGs in July 2024
NCBI Datasets	NCBI, USA	<a href="https://www.ncbi.nlm.nih.gov/datasets/genome/">https://www.ncbi.nlm.nih.gov/datasets/genome/</a>	○	○	443,763
IMG/M	JGI, USA	<a href="https://img.jgi.doe.gov/cgi-bin/m/main.cgi">https://img.jgi.doe.gov/cgi-bin/m/main.cgi</a>	○	○	25,507
SPIRE	EMBL, EU	<a href="http://spire.embl.de/">http://spire.embl.de/</a>	○	○	1,160,000
MGnify	EBI, EU	<a href="https://www.ebi.ac.uk/metagenomics">https://www.ebi.ac.uk/metagenomics</a>	○	○	478,810
Microbiome Datahub	NIG, Japan	<a href="https://mdatahub.org/">https://mdatahub.org/</a>	○	○	218,653

データベースによってMAGのクオリティはバラバラ  
残りの時間で、各DBの特徴と使い方について解説する

# JST-NBDC 統合化推進プログラム

爆発的な勢いで増加するマイクロバイオームデータをいち早く収録し、検索・解析可能な統合DBとして、マイクロバイオーム研究の国際的なデータハブを構築し発展させることを目標とする。



<https://microbedb.jp> → <https://mdatahub.org>

**単離菌ゲノムとMAGを基盤とした  
マイクロバイオームの統合データベース**

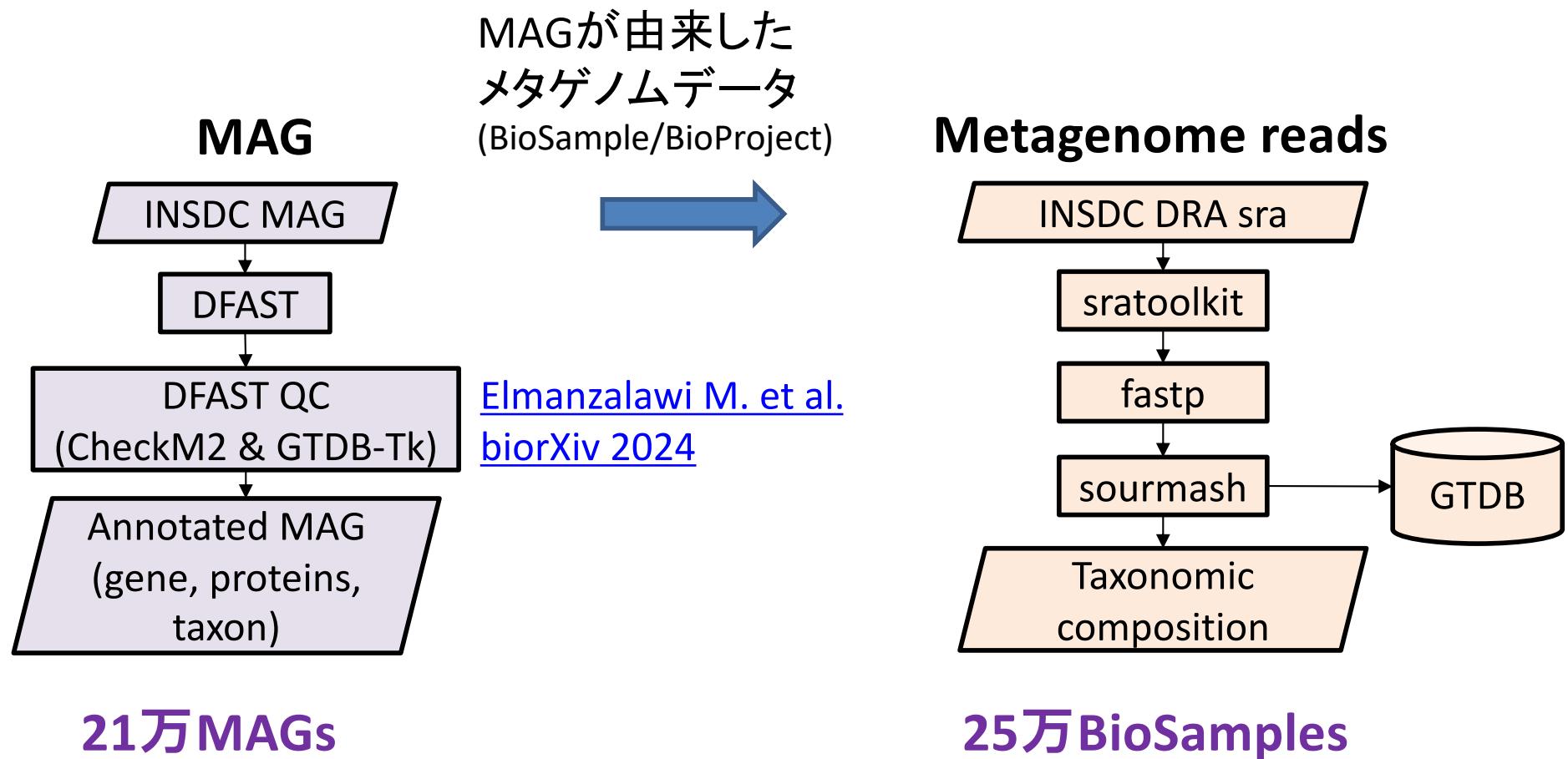
# Microbiome Datahubの今のUI

The screenshot shows the current user interface of the Microbiome Datahub. On the left, there is a sidebar with various filters: Environment (soil, marine, freshwater, hot spring, sediment, air, gut, oral, skin, reproductive system, human activity related), Genome taxon, MAG completeness (50), Host taxon, and Host disease. A search bar at the top of the sidebar contains the placeholder "Search Keyword". The main area has tabs for "PROJECT" and "GENOME", with "GENOME" currently selected. The genome list shows the following entries:

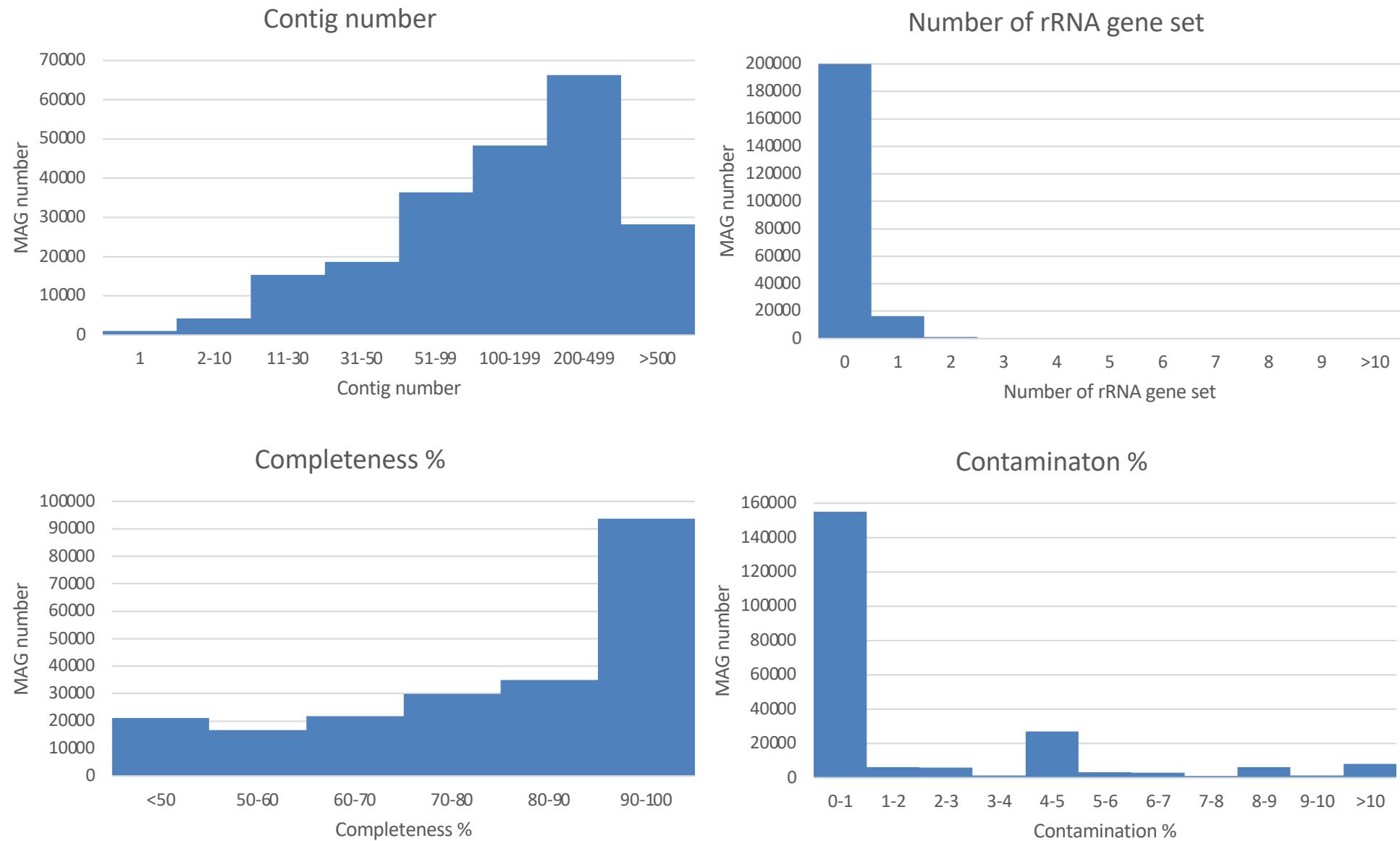
- Candidatus Thermoplasmatota archaeon** GCA\_029762495.1
- Methanocalculus sp.** GCA\_029762515.1
- Candidatus Thermoplasmatota archaeon** GCA\_029762525.1
- Candidatus Thermoplasmatota archaeon** GCA\_029762535.1
- Candidatus Thermoplasmatota archaeon** GCA\_029762575.1

Each entry includes a checkbox, environment information, host taxon, bio samples, data size, and a date created (2023-04-16). There are also "Select" and "Download" buttons, and a "order by Date Created" dropdown. The URL <https://mdatahub.org> is displayed prominently on the right.

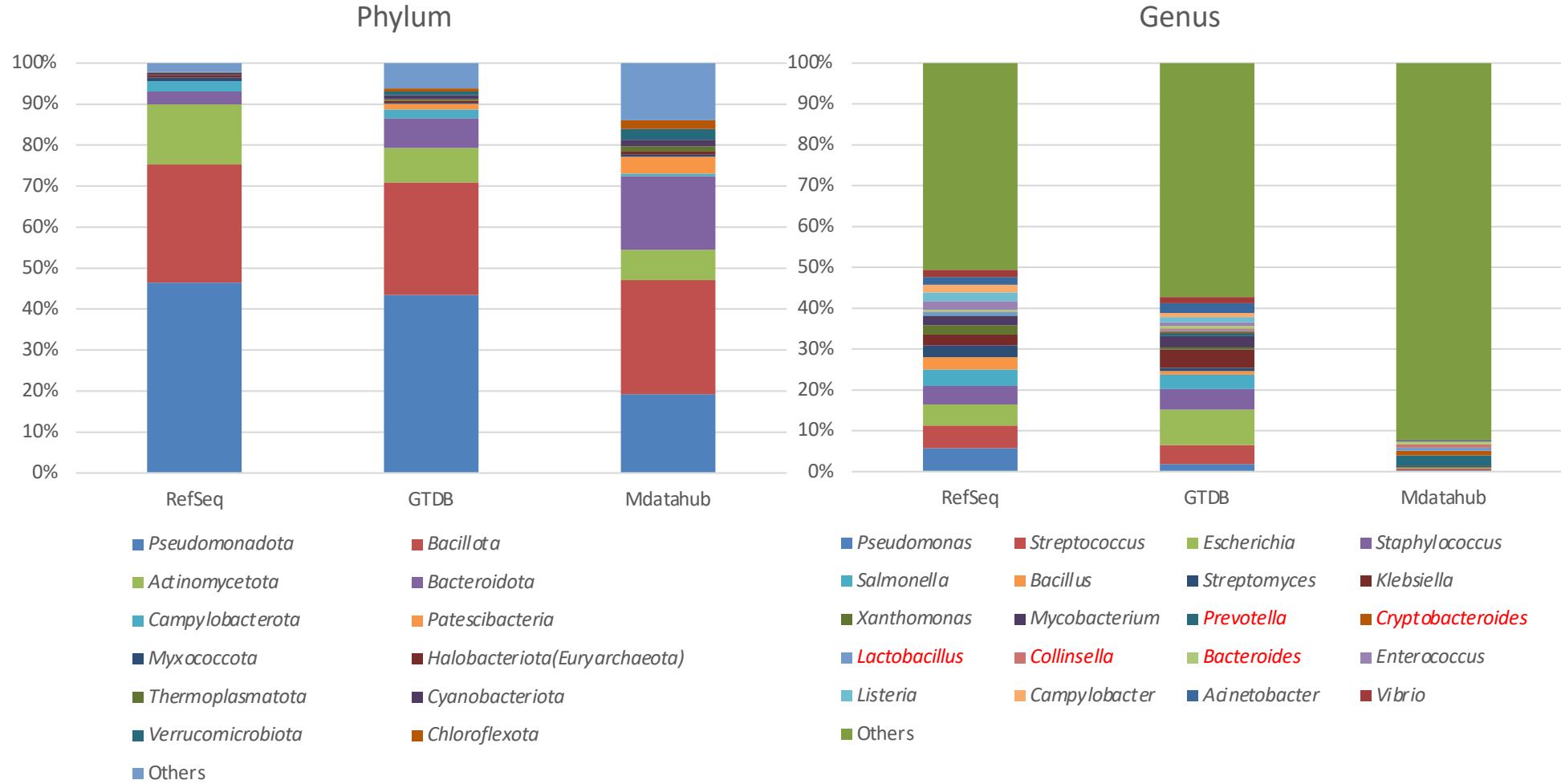
# Microbiome DatahubにおけるMAGの収集とアノテーション



# Microbiome Datahubにおける21万MAGの内訳



# Microbiome Datahubにおける21万MAGの内訳と比較



## RefSeq単離菌31万ゲノムデータの内訳

*Escherichia*: 3.5万、*Klebsiella*: 2万、*Staphylococcus*: 2万、

*Streptococcus*: 1.8万、*Pseudomonas*: 1.5万、*Salmonella*: 1.2万、

*Acinetobacter*: 1万      RefSeqは病原菌が全体の半数以上を占める

# Microbiome Datahubでできること

- Projectのメタデータ、サンプルの系統組成の検索
  - 環境情報を用いてショットガンメタゲノムのINSDC BioProjectを検索、メタデータをTSV形式で取得
  - 環境情報を用いてショットガンメタゲノムのINSDC BioProjectを検索、サンプルの系統組成データをTSV形式で取得
- MAGのメタデータによる検索
  - 系統名やメタデータを用いてMAGデータを検索、メタデータや配列データを取得

# Microbiome DatahubのMAG検索

<https://mdatahub.org>

Microbiome Datahub

Document

Search Keyword

Environment

soil  
marine  
freshwater  
hot spring  
sediment  
air  
gut  
oral  
skin  
reproductive system  
human activity related

Genome taxon

MAG completeness   
50

Host taxon

Host disease

PROJECT GENOME

10 / 218653 order by Date Created

**Candidatus Thermoplasmatota archaeon** GCA\_029762495.1

Environment Candidatus Thermoplasmatota archaeon Host taxon BioSamples 1 Data size (GB) 0.53 MB

Date Created 2023-04-16

**Methanocalculus sp.** GCA\_029762515.1

Environment Methanocalculus sp. Host taxon BioSamples 1 Data size (GB) 0.67 MB

Date Created 2023-04-16

**Candidatus Thermoplasmatota archaeon** GCA\_029762525.1

Environment Candidatus Thermoplasmatota archaeon Host taxon BioSamples 1 Data size (GB) 0.515 MB

Date Created 2023-04-16

**Candidatus Thermoplasmatota archaeon** GCA\_029762535.1

Environment Candidatus Thermoplasmatota archaeon Host taxon BioSamples 1 Data size (GB) 0.847 MB

Date Created 2023-04-16

**Candidatus Thermoplasmatota archaeon** GCA\_029762575.1

Environment Candidatus Thermoplasmatota archaeon Host taxon BioSamples 1 Data size (GB) 0.453 MB

Date Created 2023-04-16

# Microbiome DatahubのMAG検索

<https://mdatahub.org>

The screenshot shows the Microbiome Datahub search interface. On the left, there is a sidebar with filters for Environment (soil, marine, freshwater, hot spring, sediment, air, gut, oral, skin, reproductive system, human activity related), Genome taxon, MAG completeness (87), Host taxon, and Host disease. The main search bar at the top contains the query "Prevotella". The results are displayed under the "GENOME" tab, showing 10 out of 4223 results. Each result row includes the species name, GCA identifier, date created, environment, host taxon, BioSamples count, and data size. A dropdown menu for "Download" is open over the first result, listing options for metadata, genome sequence, gene sequence, and protein sequence.

Result	GCA Identifier	Date Created	Environment	Host Taxon	BioSamples	Data Size (GB)
1	GCA_029050585.1	2023-03-09	Environment	Prevotella sp.	1	0.458 MB
2	GCA_029052985.1	2023-03-08	Environment	Prevotella sp.	1	0.784 MB
3	GCA_029054825.1	2023-03-08	Environment	Prevotella sp.	1	0.747 MB
4	GCA_029054915.1	2023-03-08	Environment	Prevotella sp.	1	0.66 MB

# Microbiome DatahubのMAG検索のダウンロード結果例

## MAG metadata例 (tsv形式)

DFAST結果

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	identifier	organism	quality	bioproject	biosample	species_taxid	san	san	sample_host_organism	sample_host_location	sample_host_location	Total Sequence Length (bp)	Number of Sequences	Longest N50	GCcontent (%)	Number of CDSs					
2	GCA_029050585.1	Prevotella sp.	2	PRJNA725899	SAMN19956079	59823	Pre	##	Microtus arvalis	Lithuania: Vilnius	{ "l": "Lithuania: Vilnius" }	1572930	703	#### ####	51.3	813					
3	GCA_029052985.1	Prevotella sp.	3	PRJNA725899	SAMN19955964	59823	Pre	##	Alexandromys oeconomus	Lithuania: Vilnius	{ "l": "Lithuania: Vilnius" }	2726451	158	#### ####	46.8	2085					
4	GCA_029054825.1	Prevotella sp.	3	PRJNA725899	SAMN19955874	59823	Pre	##	Apodemus agrarius	Lithuania: Vilnius	{ "l": "Lithuania: Vilnius" }	2610678	250	#### ####	48.7	1875					
5	GCA_029054915.1	Prevotella sp.	3	PRJNA725899	SAMN19955868	59823	Pre	##	Apodemus agrarius	Lithuania: Vilnius	{ "l": "Lithuania: Vilnius" }	2283067	724	#### ####	52	1254					
6	GCA_029071365.1	Prevotella sp.	3	PRJNA725899	SAMN19956450	59823	Pre	##	Alexandromys oeconomus	Lithuania: Vilnius	{ "l": "Lithuania: Vilnius" }	2595038	74	#### ####	51	2081					

## Genome sequence, gene sequence, protein sequence例 (fasta形式)

Gene sequence

Category	File Type	File Name	Last Modified	Size	Description	
Gene sequence	sequence_cds	GCA_029050585.1/cds.fna	今日 6:44	728 KB	書類	
	sequence_cds	GCA_029052985.1/cds.fna	今日 6:44	2.5 MB	書類	
	sequence_cds	GCA_029054825.1/cds.fna	今日 6:44	2.1 MB	書類	
	sequence_cds	GCA_029054915.1/cds.fna	今日 6:44	1.3 MB	書類	
	sequence_cds	GCA_029071365.1/cds.fna	今日 6:44	2.4 MB	書類	
	sequence_genome	GCA_029050585.1/genome.fna	今日 6:42	1.6 MB	書類	
Genome sequence	sequence_genome	GCA_029052985.1/genome.fna	今日 6:42	2.7 MB	書類	
	sequence_genome	GCA_029054825.1/genome.fna	今日 6:42	2.6 MB	書類	
	sequence_genome	GCA_029054915.1/genome.fna	今日 6:42	2.3 MB	書類	
	sequence_genome	GCA_029071365.1/genome.fna	今日 6:42	2.6 MB	書類	
	sequence_protein	GCA_029050585.1/protein.faa	今日 6:44	270 KB	書類	
	sequence_protein	GCA_029052985.1/protein.faa	今日 6:44	894 KB	書類	
Protein sequence	sequence_protein	GCA_029054825.1/protein.faa	今日 6:44	756 KB	書類	
	sequence_protein	GCA_029054915.1/protein.faa	今日 6:44	464 KB	書類	
	sequence_protein	GCA_029071365.1/protein.faa	今日 6:44	878 KB	書類	
	sequence_protein					
	sequence_protein					
	sequence_protein					

 National Library of Medicine  
National Center for Biotechnology Information

Search NCBI ... Log in

NCBI Datasets Taxonomy Genome Gene Command-line tools Documentation

## Genome

Search by taxonomic name or ID, Assembly name, BioProject, BioSample, WGS or Nucleotide accession

Search term  Search

Try examples: [Homo sapiens](#) [GCF\\_000001405.40](#) [PRJNA489243](#) [SAMN15960293](#) [WFKY01](#) [GRCh38.p14](#) [NC\\_000913.3](#)

### Genomic data available from NCBI Datasets

Click below to learn more about the genomic data available from NCBI Datasets.



Eukaryota



Archaea



Bacteria



Viruses

### All Genomes

2.42M 36.73K 1.99M

Total Reference Annotated

National Library of Medicine  
National Center for Biotechnology Information

Search NCBI ... Log in

NCBI Datasets Taxonomy Genome Gene Command-line tools Documentation

## Genome

Search by taxonomic name or ID, Assembly name, BioProject, BioSample, WGS or Nucleotide accession

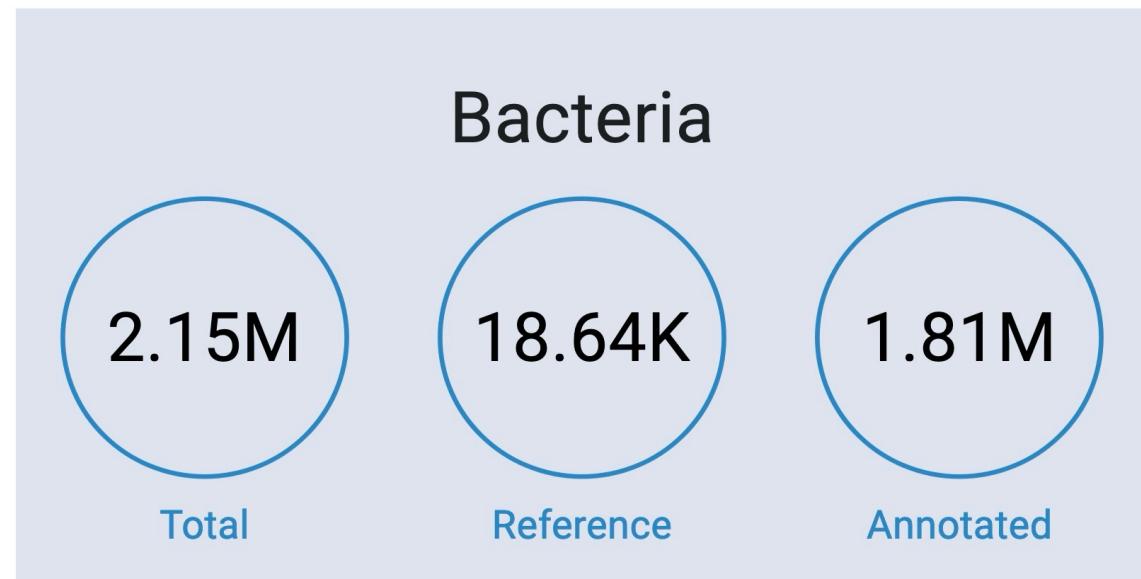
Search term

Search

Try examples: [Homo sapiens](#) [GCF\\_000001405.40](#) [PRJNA489243](#) [SAMN15960293](#) [WFKY01](#) [GRCh38.p14](#) [NC\\_000913.3](#)

## Genomic data available from NCBI Datasets

Click below to learn more about the genomic data available from NCBI Datasets.



National Library of Medicine  
National Center for Biotechnology Information

Search NCBI ... Log in

NCBI Datasets Taxonomy Genome Gene Command-line tools Documentation

## Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa

Bacteria  Enter one or more taxonomic names

Filters

<input type="checkbox"/> Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
<input type="checkbox"/> ASM694v2	GCA_000006945.2	GCF_000006945.2	Salmonella enterica subsp. ent...	LT2 (strain)	NCBI RefSeq Submitter	<input type="button" value="⋮"/>
<input type="checkbox"/> ASM19595v2	GCA_000195955.2	GCF_000195955.2	Mycobacterium tuberculosis H...	H37Rv (strain)	NCBI RefSeq Submitter	<input type="button" value="⋮"/>
<input type="checkbox"/> ASM904v1	GCA_000009045.1	GCF_000009045.1	Bacillus subtilis subsp. subtilis ...	168 (strain)	NCBI RefSeq Submitter	<input type="button" value="⋮"/>
<input type="checkbox"/> ASM584v2	GCA_000005845.2	GCF_000005845.2	Escherichia coli str. K-12 substr...	K-12 substr. MG165...	NCBI RefSeq Submitter	<input type="button" value="⋮"/>
<input type="checkbox"/> ASM886v2	GCA_000008865.2	GCF_000008865.2	Escherichia coli O157:H7 str. S...	Sakai substr. RIMD ...	NCBI RefSeq Submitter	<input type="button" value="⋮"/>
<input type="checkbox"/> ASM24018v2	GCA_000240185.2	GCF_000240185.1	Klebsiella pneumoniae subsp. ...	HS11286 (strain)	NCBI RefSeq Submitter	<input type="button" value="⋮"/>
<input type="checkbox"/> ASM676v1	GCA_000006765.1	GCF_000006765.1	Pseudomonas aeruginosa PAO1	PAO1 (strain)	NCBI RefSeq Submitter	<input type="button" value="⋮"/>

19

NCBI Datasets [https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=2&mags\\_only=true](https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=2&mags_only=true)

## Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa

Bacteria  Enter one or more taxonomic names

**INCLUDE ONLY**

Reference genomes  
 Annotated genomes  
 Annotated by NCBI RefSeq  
 Annotated by GenBank submitter  
 Metagenome-assembled genomes (MAGs)  
 From type material  
 From ICTV exemplar

**SEARCH WITHIN RESULTS**

**ASSEMBLY LEVEL**

contig scaffold chromosome complete

**YEAR RELEASED**

1980 2024

		424,066 Genomes		Rows per page	20	1-20 of 424,066	<	>
<input type="checkbox"/>	Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action	
<input type="checkbox"/>	piMetPara1.Trichococcus_floc...	GCA_963662535.1	GCF_963662535.1	Trichococcus flocculiformis	659e37ff-3402-402...	NCBI RefSeq	<input type="button" value="::"/>	20
<input type="checkbox"/>	ASM1318141v1	GCA_013181415.1	GCF_013181415.1	Ferrovum myxofaciens	S2.4 (isolate)	NCBI RefSeq	<input type="button" value="::"/>	

NCBI Datasets [https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=838&mags\\_only=true](https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=838&mags_only=true)

## Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa

Prevotella x Enter one or more taxonomic names

Filters MAGs x

**INCLUDE ONLY**

Reference genomes  
 Annotated genomes  
 Annotated by NCBI RefSeq  
 Annotated by GenBank submitter  
 Metagenome-assembled genomes (MAGs)  
 From type material  
 From ICTV exemplar

**SEARCH WITHIN RESULTS**

**ASSEMBLY LEVEL**

  
contig scaffold chromosome complete

**YEAR RELEASED**

  
1980 2024

**EXCLUDE**

Atypical genomes  
 Metagenome-assembled genomes (MAGs)  
 Genomes from large multi-isolate projects

Download Select columns 5,640 Genomes Rows per page 20 1-20 of 5,640 < >

<input type="checkbox"/> Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
<input type="checkbox"/> ASM1526260v1 <span style="color: green;">✓</span>	GCA_015262605.1	GCF_015262605.1	Prevotella aurantiaca	JCVI_44_bin.5 (isol...	NCBI RefSeq Submitter	...
<input type="checkbox"/> 6755_HET_CTRL_contig_661...	GCA_963931425.1		Prevotella sp. CAG:1058	6755_HET_CTRL_c...		...

NCBI Datasets [https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=838&mags\\_only=true&assembly\\_level=2:3](https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=838&mags_only=true&assembly_level=2:3)

## Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa

Prevotella × Enter one or more taxonomic names

Filters MAGs × chromosome+ ×

**INCLUDE ONLY**

Reference genomes  
 Annotated genomes  
 Annotated by NCBI RefSeq  
 Annotated by GenBank submitter  
 Metagenome-assembled genomes (MAGs)  
 From type material  
 From ICTV exemplar

**SEARCH WITHIN RESULTS**

**ASSEMBLY LEVEL**

contig scaffold chromosome complete

**YEAR RELEASED**

1980 2024

**EXCLUDE**

Atypical genomes  
 Metagenome-assembled genomes (MAGs)  
 Genomes from large multi-isolate projects

Download Select columns 3 Genomes Rows per page 20 ▾ 1-3 of 3 < >

<input type="checkbox"/> Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
<input type="checkbox"/> 6755_HET_CTRL_contig_661...	GCA_963931425.1		Prevotella sp. CAG:1058	6755_HET_CTRL_c...		⋮
<input type="checkbox"/> metabat2_pooled_hifiasm.683...	GCA_963931275.1		uncultured Prevotella sp.	metabat2_pooled_h...		⋮

# NCBI Datasets [https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=838&mags\\_only=true](https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=838&mags_only=true)

## Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa  
Prevotella  Enter one or more taxonomic names

**Filters** MAGs

**INCLUDE ONLY**

Reference genomes

Annotated genomes

Annotated by NCBI RefSeq

Annotated by GenBank submitter

Metagenome-assembled genomes (MAGs)

From type material

From ICTV exemplar

**SEARCH WITHIN RESULTS**

**ASSEMBLY LEVEL**

contig scaffold chromosome complete

**YEAR RELEASED**

1980 2024

Download <input type="button" value="v"/>		Select columns		5,640 Genomes 5,640 selected		Rows per page	20 <input type="button" value="▼"/>	1-20 of 5,640	< >
Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action			
<input checked="" type="checkbox"/>	ASM1526260v1 	GCA_015262605.1	GCF_015262605.1	Prevotella aurantiaca	JCVI_44_bin.5 (isol...	<b>NCBI RefSeq</b> Submitter	<input type="button" value="::"/>		
<input checked="" type="checkbox"/>	6755_HET_CTRL_contig_661...	GCA_963931425.1		Prevotella sp. CAG:1058	6755_HET_CTRL_c...		<input type="button" value="::"/>		

# Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa  
Prevotella  Enter one or more taxonomic names

Filters MAGs

INCLUDE ONLY

Reference genomes  
 Annotated genomes  
 Annotated by NCBI RefSeq  
 Annotated by GenBank submitter  
 Metagenome-assembled genomes  
 From type material  
 From ICTV exemplar

EXCLUDE

Atypical genomes  
 Metagenome-assembled genomes  
 Genomes from large multi-isolate samples

Download Package

5,640 genomes selected for download

Select file source

All  
 RefSeq only  
 GenBank only

Select file types

Genome sequences (FASTA)  
 Annotation features (GTF)  
 Annotation features (GFF)  
 Sequence and annotation (GBFF)  
 Transcripts (FASTA)  
 Genomic coding sequences (FASTA)  
 Protein (FASTA)  
 Sequence report (JSONL)  
 Assembly data report (JSONL)

Your selected data will be downloaded as a ZIP archive  
Estimated file size is 977 MB

Name your file  
ncbi\_dataset.zip

Cancel Download

20 ▾ 1-20 of 5,640 < >

Action	Annotation	Modifier	Scientific name	RefSeq	GenBank	Download Table	Download Package
⋮	NCBI RefSeq Submitter	JCVI_44_bin.5 (isol...)	Prevotella aurantiaca	GCF_015262605.1	GCA_015262605.1	<input type="checkbox"/> ASMT1526260V1	<input checked="" type="checkbox"/>
⋮	6755_HET_CTRL_contig_661...	6755_HET_CTRL_c...	Prevotella sp. CAG:1058	GCA_963931425.1	6755_HET_CTRL_...	<input type="checkbox"/> 6755_HET_CTRL_contig_661...	<input checked="" type="checkbox"/>

The screenshot shows the MGNify homepage. At the top, there's a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, a search icon, and the EMBL-EBI logo. The main header features the MGNify logo and the tagline "Submit, analyse, discover and compare microbiome data". Below this is a search bar with placeholder text "Search MGNify" and a magnifying glass icon. A "Example searches" section lists "Tara oceans", "MGYS00000410", and "Human Gut". The background features a 3D illustration of a diverse microbiome community.

Overview    Submit data ▾    Text search ▾    Sequence search ▾    Browse data ▾    About    Help ▾    Login

---

### Search by

**Text search →**  
Name, biome, or keyword

**Sequence search →**  
Sequence search

### Or by data type

**Analysis types**

480962	amplicon
57629	assemblies
2050	metabarcoding
39920	metagenomes
2581	metatranscriptomics
2	long reads assemblies

**Public data**

5004	studies
597736	analyses
478810	genomes in 11 MAG catalogues

### Request analysis of

**Submit and/or Request →**  
Your data

**Request →**  
A public dataset

### Latest studies

**EMG produced TPA metagenomics assembly of the Gut microbial dysbiosis in young adults with obesity (...)**  
The Gut microbial dysbiosis in young adults with obesity Third Party Annotation (TPA) assembly was derived from the primary whole genome shotgun (WGS) data set: PRJEB12123. This project includes samples from the following biomes: Host-associated, Hu...

**Oil spill dispersant strategies and bioremediation efficiency**  
After accidental oil spills to seawater from production or transport facilities the oil will be spread on the water surface and in the water column, and a number of weathering processes will appear. Typical operational methods to remove the oil from ...

**nifH gene in KH-11-10 and KH-13-7 cruises**

# MGNify <https://www.ebi.ac.uk/metagenomics/>

Search by

**Text search** →

Name, biome, or keyword

**Sequence search** →

Sequence search

Or by data type

xx Analysis types

480962 amplicon

57629 assemblies

2050 metabarcoding

39920 metagenomes

2581 metatranscriptomics

2 long reads assemblies

Public data

5004 studies

597736 analyses

478810 genomes in 11 MAG catalogues

Or by selected biomes



Human  
(213874)



Digestive system  
(111024)



Aquatic  
(51540)



Marine  
(38036)



Digestive system  
(35543)



Plants  
(28859)



Soil  
(25893)



Skin  
(11527)

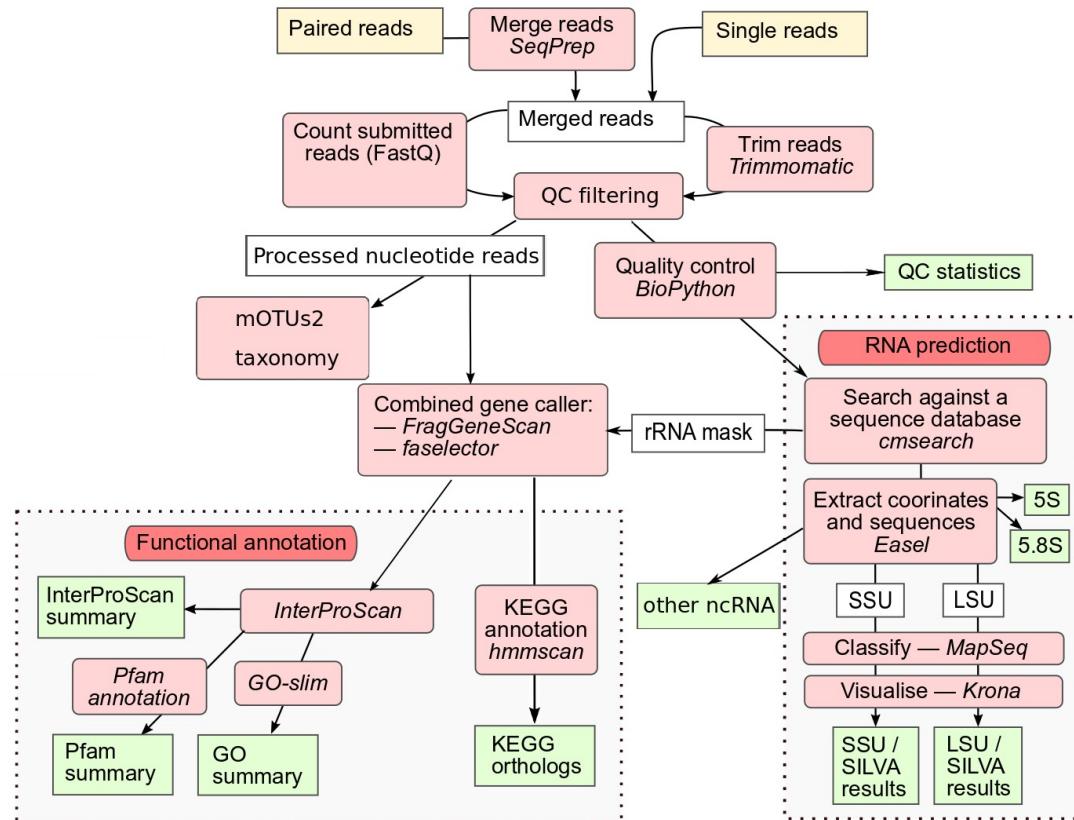


Wastewater  
(4358)



Food production  
(2923)

**View all biomes**



<https://emg-docs.readthedocs.io/en/latest/analysis.html>



Overview Search Results Help Contact

phmmer hmmsearch

## Search our non-redundant protein database using HMMER

Protein database version: 2024\_04

Paste a Sequence | Upload a File

Paste in your sequence or use the example ?

Submit Reset

### ▼ Sequence subset ?

Sequence type:  All sequences  Full length sequences  Partial sequences

Current database selection:

Full length sequences

### ▼ Cut-Offs ?

E-value  Bit score

Sequence	Hit
Significance E-values: 0.001	0.001

Sequence	Hit
Report E-values: 0.1	0.1

# MGnify <https://www.ebi.ac.uk/metagenomics/browse/genomes>

Home > Browse > Genomes

## Browse MGnify

Super Studies   Studies   Samples   Publications   **Genomes**   Biomes

Genome catalogues are biome-specific collections of metagenomic-assembled and isolate genomes. The latest version of each catalogue is shown on this website. Data for current and previous versions are available on the [FTP server](#).

If you use the MGnify Genomes resource, please cite:

**MGnify Genomes: a resource for biome-specific microbial genome catalogues** *Journal of Molecular Biology* (2023) doi:[10.1016/j.jmb.2023.168016](https://doi.org/10.1016/j.jmb.2023.168016)  
Gurbich TA, Almeida A, Beracochea M, Burdett T, Burgin T, Cochrane G, Raj S, Richardson LJ, Rogers AB, Sakharova E, Salazar GA and Finn RD.

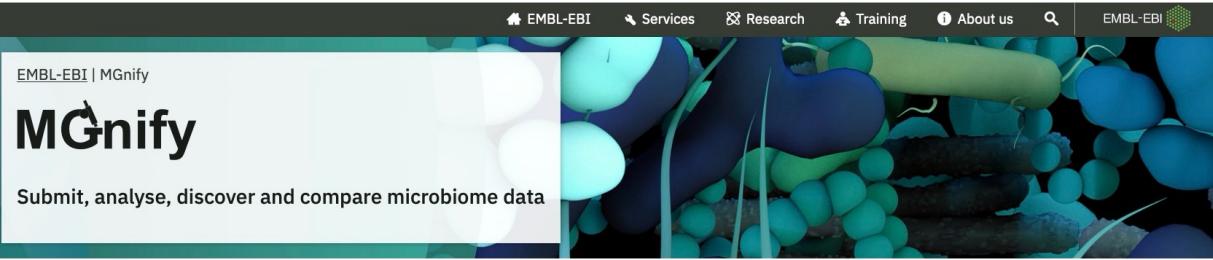
[Catalogues list](#)   All genomes   Gene search   MAG search

Select a catalogue in the table to browse or search its genomes.

Filter biome

 Download

Biome	Catalogue ID	Catalogue name	Catalogue version	Species count	Total genomes count	Last Updated
	chicken-gut-v1-0-1	Chicken Gut v1.0.1	1.0.1	1322	13386	2024/1/23
	cow-rumen-v1-0-1	Cow Rumen v1.0.1	1.0.1	2729	5578	2024/1/23
	honeybee-gut-v1-0-1	Honeybee gut v1.0.1	1.0.1	154	627	2024/1/23
	human-gut-v2-0-2	Unified Human Gastrointestinal Genome (UHGG) v2.0.2	2.0.2	4744	289232	2024/1/23
	human-oral-v1-0-1	Human Oral v1.0.1	1.0.1	452	1225	2024/1/23
	human-vaginal-v1-0	Human Vaginal v1.0	1.0	280	618	2023/8/15
	marine-v2-0	Marine v2.0	2.0	13223	50866	2024/5/12
	mouse-gut-v1-0	Mouse Gut v1.0	1.0	2847	112951	2023/12/8
	non-model-fish-gut-v2-0	Non-model Fish Gut v2.0	2.0	178	254	2023/4/26
	pig-gut-v1-0	Pig Gut v1.0	1.0	1376	3972	2022/11/29
	zebrafish-fecal-v1-0	Zebrafish Fecal v1.0	1.0	79	101	2022/11/28



EMBL-EBI | MGnify

# MGnify

Submit, analyse, discover and compare microbiome data

Overview Submit data ▾ Text search ▾ Sequence search ▾ Browse data ▾ About Help ▾ Login

Home > Browse > Genomes

## Browse MGnify

Super Studies Studies Samples Publications **Genomes** Biomes

Genome catalogues are biome-specific collections of metagenomic-assembled and isolate genomes. The latest version of each catalogue is shown on this website. Data for current and previous versions are available on the [FTP server](#).

If you use the MGnify Genomes resource, please cite:

**MGnify Genomes: a resource for biome-specific microbial genome catalogues** *Journal of Molecular Biology* (2023) doi:[10.1016/j.jmb.2023.168016](https://doi.org/10.1016/j.jmb.2023.168016)  
Gurbich TA, Almeida A, Beracocha M, Burdett T, Burgin T, Cochrane G, Raj S, Richardson LJ, Rogers AB, Sakharova E, Salazar GA and Finn RD.

Catalogues list [All genomes](#) Gene search MAG search

Search for genomes across all catalogues.

Genomes (552)

Biome	Accession	Catalogue	Type	Taxonomy
	MGYG000411028	mouse-gut-v1-0	MAG	Prevotella rodentium
	MGYG000410844	mouse-gut-v1-0	MAG	Paraprevotella sp910586505
	MGYG000409046	mouse-gut-v1-0	MAG	Paraprevotella sp910584955
	MGYG000348300	mouse-gut-v1-0	MAG	Prevotella sp002933775
	MGYG000331390	mouse-gut-v1-0	MAG	Prevotella sp002265625
	MGYG000331388	mouse-gut-v1-0	MAG	Alloprevotella sp900542795
	MGYG000331371	mouse-gut-v1-0	MAG	Prevotella sp000436035
	MGYG000331314	mouse-gut-v1-0	MAG	Prevotella sp900553595
	MGYG000331239	mouse-gut-v1-0	MAG	Prevotella pectinovora
	MGYG000330961	mouse-gut-v1-0	MAG	Prevotella sp900552675

Filter  [Download](#)

## MGNify <https://www.ebi.ac.uk/metagenomics/browse/genomes?browse-by=gene-search>

The screenshot shows the MGnify homepage with a banner featuring a 3D visualization of microbial cells. The main navigation bar includes links for Overview, Submit data, Text search, Sequence search, Browse data, About, Help, and Login. Below the navigation is a breadcrumb trail: Home > Browse > Genomes. A sub-navigation bar below the breadcrumb includes Super Studies, Studies, Samples, Publications, Genomes (which is underlined), and Biomes. A text box provides citation information for MGnify Genomes. The central search area is titled "Search DNA fragments across catalogues" and includes a "Gene search" tab. A list of selected catalogues includes: Chicken Gut v1.0.1, Cow Rumen v1.0.1, Honeybee gut v1.0.1, Unified Human Gastrointestinal Genome (UHGG) v2.0.2, Human Oral v1.0.1, Human Vaginal v1.0, Marine v2.0, Mouse Gut v1.0, Non-model Fish Gut v2.0, Pig Gut v1.0, and Zebrafish Fecal v1.0. Below this is a sequence input field with three submission options: Paste a sequence, Browse for file..., and Use the example. A threshold selector is set to 0.4. A yellow warning box at the bottom states: "The query can't be submitted because: • The sequence has to have at least 50 nucleotides".

The screenshot shows the MGnify homepage with a banner featuring a 3D visualization of gut microbiome bacteria. The navigation bar includes links for EMBL-EBI, Services, Research, Training, About us, a search icon, and the MGnify logo. Below the banner, the MGnify logo and tagline "Submit, analyse, discover and compare microbiome data" are displayed. A secondary navigation bar below the main one includes Overview, Submit data, Text search, Sequence search, Browse data, About, Help, and Login.

The main content area is titled "Browse MGnify" and includes a sub-navigation bar with links for Super Studies, Studies, Samples, Publications, Genomes (which is highlighted in blue), and Biomes.

A text block explains that genome catalogues are biome-specific collections of metagenomic-assembled and isolate genomes, with the latest version shown on the website. It also mentions the availability of data for current and previous versions via an FTP server.

An information box provides citation details: "If you use the MGnify Genomes resource, please cite: MGnify Genomes: a resource for biome-specific microbial genome catalogues. Journal of Molecular Biology (2023) doi:10.1016/j.jmb.2023.168016 Gurbich TA, Almeida A, Beracochea M, Burdett T, Burgin T, Cochrane G, Raj S, Richardson LJ, Rogers AB, Sakharova E, Salazar GA and Finn RD."

The interface features a "Catalogues list" tab and a "MAG search" tab, with the latter being active. The "Search MAG files across catalogues" section includes a "Powered by Sourmash" logo. A note states that users can compare their MAG file or collection against MGnify's catalogues to see if they are novel.

Instructions for file upload are provided, noting that users can upload a single FastA file, multiple files using [ctrl] or [shift], or a whole directory using directory mode. It is mentioned that the tool processes all FastA files in the selected directory but does not descend into subdirectories.

A note clarifies that files are not uploaded to servers; instead, Sourmash generates a signature of the file(s) in the browser and compares it against the MAG catalogue.

Successful searches result in a CSV file per signature, which is compiled into a TGZ for download. These results are stored for 30 days.

The "Select catalogues to search against" section lists available catalogues: Chicken Gut v1.0.1, Cow Rumen v1.0.1, Honeybee gut v1.0.1, Unified Human Gastrointestinal Genome (UHGG) v2.0.2, Human Oral v1.0.1, Human Vaginal v1.0, Marine v2.0, Mouse Gut v1.0, Non-model Fish Gut v2.0, Pig Gut v1.0, and Zebrafish Fecal v1.0. A dropdown menu is shown next to the list.

The "Select your files" section includes a "Select nucleotides FastA files:" input field with a "Choose Files..." button, a "Browse" button, and tabs for "Directory" and "Files".

At the bottom right, there are "Search" and "Clear" buttons.

The screenshot shows the MGnify homepage with a banner for the Unified Human Gastrointestinal Genome (UHGG) v2.0.2. Key statistics displayed include 289232 total genomes and 4744 species-level clusters. There are links to an FTP site, a pipeline workflow, and a genome list. A search bar is present at the top, and a detailed table below lists species-level cluster representatives with columns for Biome, Accession, Length, Num. of genomes, Completeness, Contamination, Type, Taxonomy, and Last Updated.

Biome	Accession	Length	Num. of genomes	Completeness	Contamination	Type	Taxonomy	Last Updated
	MGYG000000001	3221717	4	98.59	0.7	Isolate	GCA-900066495 sp902362365 ⓘ	2024/1/23
	MGYG000000002	4441003	359	99.37	0	Isolate	Blautia_A faecis ⓘ	2024/1/23
	MGYG000000003	3234232	1181	100	0	Isolate	Alistipes shahii ⓘ	2024/1/23
	MGYG000000004	3707250	25	98.66	0.22	Isolate	Anaerotruncus colihominis ⓘ	2024/1/23
	MGYG000000005	3932530	2	99.3	0	Isolate	Terrisporobacter glycolicus_A ⓘ	2024/1/23
	MGYG000000006	2823639	1	99.26	1.39	Isolate	Staphylococcus xylosus ⓘ	2024/1/23
	MGYG000000007	2017471	1	98.94	0	Isolate	Lactobacillus intestinalis ⓘ	2024/1/23
	MGYG000000008	1972012	24	99.22	0.78	Isolate	Lactobacillus johnsonii ⓘ	2024/1/23
	MGYG000000009	2221226	2	99.21	0	Isolate	Ligilactobacillus murinus ⓘ	2024/1/23

The screenshot shows the MGnify homepage with a banner featuring a 3D visualization of gut microbiome bacteria. Below the banner, a navigation bar includes links for Overview, Submit data, Text search, Sequence search, Browse data, About, Help, and Login.

The main content area displays statistics: 289232 Total genomes and 4744 Species-level clusters. It also features links for the FTP Site, Pipeline v1.2.1, and a Download full catalogue.

A note below the stats states: "This is an update to the UHGG1.0 catalogue described by Almeida et al., Nature Biotechnol (2021)."

The central feature is a Taxonomy tree. The tree starts with the Domain level, branching into Archaea (28), Bacteria (4716), Firmicutes\_A (1906), Bacteroidota (619), and Bacteroidia (619). The Bacteroidia branch further into Bacteroidales (609), Rikenellaceae (61), Bacteroidaceae (260), Bacteroides (53), Phocaeicola (36), and Prevotella (113). The Prevotella branch lists numerous species and strains, many of which are highlighted in green.

Below the tree, there are tabs for Genome list, Taxonomy tree (which is active), Protein catalogue, Search by Gene, and Search by MAG.

The screenshot shows the MGnify genome overview page for MGYG000000526. The top navigation bar includes links for EMBL-EBI, Services, Research, Training, About us, a search bar, and the MGnify logo. Below the header is a banner with the text "Submit, analyse, discover and compare microbiome data". The main content area displays the genome's taxonomic lineage (Bacteria > Bacteroidota > Bacteroidia > Bacteroidales > Bacteroidaceae > Prevotella > Prevotella sp900770395), its type (MAG), and its length (3510035 bp). It also provides links for Overview, Browse genome, COG analysis, KEGG class analysis, KEGG module analysis, and Downloads. The page is divided into several sections: Genome statistics, Genome annotations, Pan-genome statistics, Genome RNA coverage, Geographic metadata, and External links. Each section contains detailed data tables.

EMBL-EBI | MGnify

# MGnify

Submit, analyse, discover and compare microbiome data

Overview Submit data ▾ Text search ▾ Sequence search ▾ Browse data ▾ About Help ▾ Login

Home > Genomes > human-gut-v2-0-2 > MGYG000000526

## Genome MGYG000000526

Type: MAG

**Taxonomic lineage:** Bacteria > Bacteroidota > Bacteroidia > Bacteroidales > Bacteroidaceae > Prevotella > Prevotella sp900770395

[Overview](#) [Browse genome](#) [COG analysis](#) [KEGG class analysis](#) [KEGG module analysis](#) [Downloads](#)

**▼ Genome statistics**

Type:	MAG
Length (bp):	3510035
Contamination:	0.19%
Completeness:	97.78%
Num. of contigs:	38
Total number of genomes in species:	4
Number of proteins:	2880
GC content:	47.78%
Taxonomic lineage:	Bacteria > Bacteroidota > Bacteroidia > Bacteroidales > Bacteroidaceae > Prevotella > Prevotella sp900770395
N50:	182536

**▼ Genome annotations**

InterPro coverage:	87.6%
EggNog coverage:	91.11%

**▼ Pan-genome statistics**

Pan-genome size:	3119
Pan-genome core size:	1696
Pan-genome accessory size:	1423

**▼ Genome RNA coverage**

rRNA 5s total gene length coverage:	0%
rRNA 16s total gene length coverage:	0%
rRNA 23s total gene length coverage:	0%
tRNAs:	19
ncRNA:	17

**▼ Geographic metadata**

Origin of representative genome:	Oceania
Geographic range of pan-genome:	Europe, Oceania

**▼ External links**

ENA sample accession:	<a href="#">SRS476306</a>
ENA study accession:	<a href="#">SRP029441</a>

MGnify <https://www.ebi.ac.uk/metagenomics/genomes/MGYG000000526>

EMBL-EBI | MGnify

# MGnify

Submit, analyse, discover and compare microbiome data

Overview Submit data ▾ Text search ▾ Sequence search ▾ Browse data ▾ About Help ▾ Login

[Home](#) > [Genomes](#) > [human-gut-v2-0-2](#) > MGYG000000526

## Genome MGYG000000526

Type: MAG

**Taxonomic lineage:** Bacteria > Bacteroidota > Bacteroidia > Bacteroidales > Bacteroidaceae > Prevotella > Prevotella sp900770395

Overview **Browse genome** COG analysis KEGG class analysis KEGG module analysis Downloads

IGV MGYG000000526\_36:1-6,408 Q 6,408 bp Cursor Guide Center Line Track Labels Save SVG - +

0 kb 1 kb 2 kb 3 kb 4 kb 5 kb 6 kb

Functional annotation MGYG000000526\_02912 MGYG000000526\_02913 MGYG000000526\_02914 MGYG000000526\_02915 MGYG000000526\_02916

Functional annotation track colour Feature

Feature type ID MGYG000000526\_02914

Product hypothetical protein

Functional annotation

Pfam PF02775 PF01855

KEGG ko:K00179

eggNOG 1410608.JNKKX0100005\_gene1063

ELIXIR Core Data Resource

MGnify is part of the ELIXIR infrastructure  
MGnify is an ELIXIR Core Data Resource

EMBL-EBI is the home for big data in biology. We help scientists exploit data to make discoveries that benefit humankind.

SERVICES RESEARCH ABOUT

MGnify <https://www.ebi.ac.uk/metagenomics/genomes/MGYG000000526#kegg-module-analysis>

The screenshot shows the MGnify genome analysis interface for genome MGYG000000526. The top navigation bar includes links to EMBL-EBI, Services, Research, Training, About us, a search bar, and the MGnify logo. Below the header is a banner with the text "Submit, analyse, discover and compare microbiome data". The main content area displays the genome's taxonomic lineage: Bacteria > Bacteroidota > Bacteroidia > Bacteroidales > Bacteroidaceae > Prevotella > Prevotella sp900770395. A navigation menu at the top includes Overview, Submit data, Text search, Sequence search, Browse data, About, Help, and Login. The "KEGG module analysis" tab is currently selected. A bar chart titled "Top 10 KEGG module categories" shows the number of matches for various modules, with M00178 having the highest count (approx. 55). Below the chart is a table titled "All 190 KEGG modules" listing the top 10 modules along with their descriptions, genome counts, and pan-genome counts. The bottom of the page includes a page size dropdown set to "Show 10", a navigation bar with page numbers 1 through 19, and a "Next" button.

EMBL-EBI | MGnify

# MGnify

Submit, analyse, discover and compare microbiome data

Overview Submit data ▾ Text search ▾ Sequence search ▾ Browse data ▾ About Help ▾ Login

Home > Genomes > human-gut-v2-0-2 > MGYG000000526

## Genome MGYG000000526

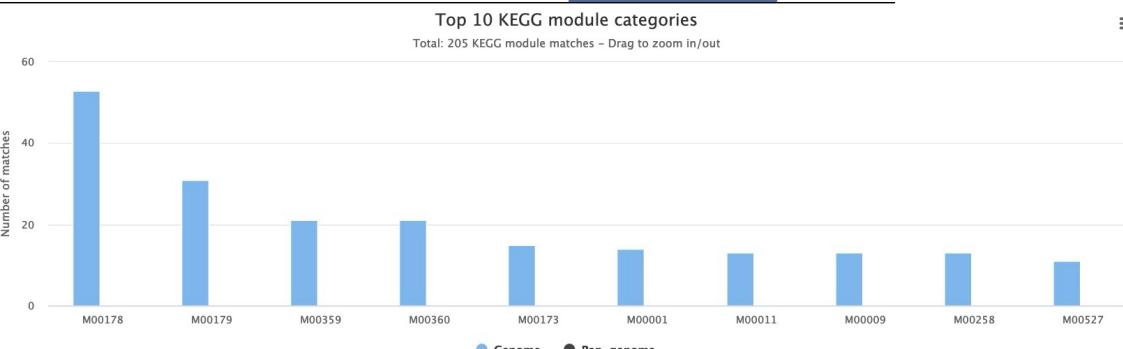
Type: MAG

**Taxonomic lineage:** Bacteria > Bacteroidota > Bacteroidia > Bacteroidales > Bacteroidaceae > Prevotella > Prevotella sp900770395

Overview Browse genome COG analysis KEGG class analysis **KEGG module analysis** Downloads

### Top 10 KEGG module categories

Total: 205 KEGG module matches – Drag to zoom in/out



Module ID	Description	Genome Count	Pan-genome count
M00178	Ribosome, bacteria	53	
M00179	Ribosome, archaea	31	
M00359		21	
M00360		21	
M00173	Reductive citrate cycle (Arnon-Buchanan cycle)	15	
M00001	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	14	
M00011	Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate	13	
M00009	Citrate cycle (TCA cycle, Krebs cycle)	13	
M00258		13	
M00527	Lysine biosynthesis, DAP aminotransferase pathway, aspartate => lysine	11	

Page Size: Show 10 ▾ Previous 1 2 3 4 5 6 7 ... 18 19 Next

## MGnify <https://www.ebi.ac.uk/metagenomics/genomes/MGYG000000526>

EMBL-EBI | MGnify

# MGnify

Submit, analyse, discover and compare microbiome data

Overview    Submit data ▾    Text search ▾    Sequence search ▾    Browse data ▾    About    Help ▾    Login

Home > Genomes > [human-gut-v2-0-2](#) > MGYG000000526

## Genome MGYG000000526

Type: MAG

Taxonomic lineage: Bacteria > Bacteroidota > Bacteroidia > Bacteroidales > Bacteroidaceae > Prevotella > Prevotella sp900770395

Overview    Browse genome    COG analysis    KEGG class analysis    KEGG module analysis    **Downloads**

### Pan-genome analysis

Name	Compression	Format	Action
List of core genes in the entire pangenome	-	TAB	<a href="#">Download</a>
Presence/absence binary matrix of the pan-genome across all conspecific genomes	-	TSV	<a href="#">Download</a>
Tree generated from the pairwise Mash distances of conspecific genomes	-	Newick format	<a href="#">Download</a>
DNA sequence FASTA file of the pangenome	-	FASTA	<a href="#">Download</a>

### Genome analysis

Name	Compression	Format	Action
All predicted CDS	-	FASTA	<a href="#">Download</a>
DNA sequence FASTA file of the genome assembly of the species representative	-	FASTA	<a href="#">Download</a>
DNA sequence FASTA file index of the genome assembly of the species representative	-	FAI	<a href="#">Download</a>
Genome GFF file with various sequence annotations	-	GFF	<a href="#">Download</a>
eggNOG annotations of the protein coding sequences	-	TSV	<a href="#">Download</a>
InterProScan annotation of the protein coding sequences	-	TSV	<a href="#">Download</a>

# IMG/M <https://img.jgi.doe.gov/cgi-bin/m/main.cgi>


**IMG/M** 

INTEGRATED MICROBIAL GENOMES & MICROBIOMES

My Analysis Carts: 0 Genomes | 0 Scaffolds | 0 Functions | 0 Genes | 0 Genome Search History | 0 Gene Search History | 0 Scaffold Search History | 0 Bin Search

---

- [Home](#)
- [IMG/M](#)
- [Find Genomes](#)
- [Find Genes](#)
- [Find Functions](#)
- [Compare Genomes](#)
- [OMICS](#)
- [My IMG](#)
- [Collaborations](#)
- [Help](#)

Quick Genome Search:

[Login into IMG/MER](#)

---

[Home](#)
[IMG Survey](#)

**IMG Content**

Datasets	JGI	All
Bacteria	<a href="#">21131</a>	<a href="#">137535</a>
Archaea	<a href="#">786</a>	<a href="#">2968</a>
Eukarya	<a href="#">370</a>	<a href="#">591</a>
Viruses	<a href="#">13</a>	<a href="#">19529</a>
Metagenome	<a href="#">19982</a>	<a href="#">34779</a>
Cell Enrichments	<a href="#">2799</a>	<a href="#">2849</a>
Single Particle Sorts	<a href="#">7368</a>	<a href="#">7758</a>
Metatranscriptome	<a href="#">6219</a>	<a href="#">8722</a>
Combined Assembly	<a href="#">275</a>	<a href="#">694</a>
Total Datasets		<a href="#">217387</a>

Last Datasets Added On:

Genome	<a href="#">2024-06-13</a>
Metagenome	<a href="#">2024-07-14</a>

 [Project Maps](#):  
[Genomes](#) [Metagenomes](#)



 [IMG Webinar YouTube Playlist](#)

The **Integrated Microbial Genomes (IMG)** system serves as a community resource for analysis and annotation of genome and metagenome datasets in a comprehensive comparative context. The **IMG data warehouse** integrates genome and metagenome datasets provided by IMG users with a comprehensive set of publicly available genome and metagenome datasets.

IMG provides users with tools for analyzing publicly available genome datasets and metagenome datasets ([Nucleic Acids Research, January 2019](#)).

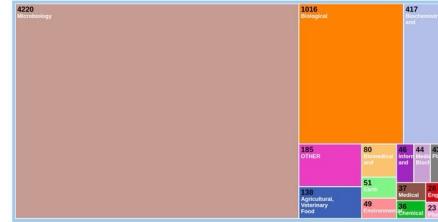
If you use IMG web resources or data to assist in research publications or proposals. Please cite: **IMG - Chen et al., 2022 (Nucleic Acids Research, gkac976).**  
**GOLD v9 - Mukherjee et al., 2022 (Nucleic Acids Research, gkac974).**

[IMG Statistics](#)
[Data Usage Policy](#)

Sequenced at:	Isolates		SAGs		MAGs	
	JGI	All	JGI	All	JGI	All
Bacteria	<a href="#">13458</a>	<a href="#">113595</a>	<a href="#">2282</a>	<a href="#">9447</a>	<a href="#">5357</a>	<a href="#">14457</a>
Archaea	<a href="#">220</a>	<a href="#">1365</a>	<a href="#">318</a>	<a href="#">601</a>	<a href="#">248</a>	<a href="#">1002</a>
Eukarya	<a href="#">370</a>	<a href="#">591</a>	0	0	0	0
Viruses	<a href="#">13</a>	<a href="#">16510</a>	0	<a href="#">75</a>	0	<a href="#">2918</a>

(Only data sets with GOLD metadata were counted.)

**View Citations**



4220 Combined assembly  
4016 Industrial  
417 Environmental  
89 Medical and Veterinary  
51 Agricultural, Forestry, Food  
49 Biotechnology  
31 Engineering  
23 Other

Combined assembly data sets were excluded in the following metagenome and metatranscriptome table statistics.

[Metagenome](#)
[Metatranscriptome](#)

Engineered	JGI	All	Environmental	JGI	All	Host-associated	JGI	All
Artificial ecosystem	<a href="#">173</a>	<a href="#">203</a>	Air	<a href="#">60</a>	<a href="#">116</a>	Algae	<a href="#">53</a>	<a href="#">168</a>
Bioreactor	<a href="#">902</a>	<a href="#">1074</a>	Aquatic	<a href="#">16111</a>	<a href="#">20696</a>	Annelida	<a href="#">138</a>	<a href="#">152</a>
Bioremediation	<a href="#">25</a>	<a href="#">61</a>	Terrestrial	<a href="#">9184</a>	<a href="#">12514</a>	Arthropoda	0	1
Biotransformation	<a href="#">3</a>	<a href="#">8</a>				Arthropoda: Crustaceans	0	7
Built environment	<a href="#">297</a>	<a href="#">1428</a>				Arthropoda: Insects	<a href="#">123</a>	<a href="#">324</a>
Food production	0	<a href="#">61</a>				Birds	<a href="#">17</a>	<a href="#">48</a>
Industrial production	<a href="#">20</a>	<a href="#">43</a>				Cephalochordata	0	4
Lab culture	<a href="#">4</a>	<a href="#">4</a>				Fish	0	<a href="#">29</a>
Lab enrichment	<a href="#">5</a>	<a href="#">69</a>				Fungi	<a href="#">145</a>	<a href="#">146</a>
Modeled	<a href="#">11</a>	<a href="#">91</a>				Human	<a href="#">2</a>	<a href="#">1001</a>

38

**JGI** **IMG/M** INTEGRATED MICROBIAL GENOMES & MICROBIOMES

Quick Genome Search:  Go Login into IMG/MER

My Analysis Carts: 0 Genomes | 0 Scaffolds | 0 Functions | 0 Genes | 0 Genome Search History | 0 Gene Search History | 0 Scaffold Search History | 0 Bin Search History

Home IMG/M Find Genomes Find Genes Find Functions Compare Genomes OMICS My IMG Collaborations Help

Home > Find Genomes 14457 Loaded  [IMG Survey](#)

## Bacteria - Metagenome-Assembled Genome

- Table Configuration

**hint:** Data Statistics with \* [assembled, unassembled, both] means metagenomes counts or percentages only use assembled data or unassembled data or both (assembled data and unassembled data) for its calculations. This does not apply to isolates. The [assembled, unassembled, both] pick list is available under the Table Configuration Data Statistics. The default is assembled data.

Add Selected to Genome Cart Display Selected on Map View Phylogenetically

Showing 1 to 10 of 14,457 entries First Previous 1 2 3 4 5 6 7 8 9 10 ... 1,446 Next Last Export Select All Clear All Select - page Deselect - page Toggle Column Selector Show 10

Domain	Sequencing Status	Study Name	Genome Name / Sample Name	Sequencing Center	IMG Genome ID	Genome Size * assembled	Gene Count * assembled
Bacteria	Permanent Draft	Candidatus Caldatrivacterium californiense OP9-cSCG	Candidatus Caldatrivacterium californiense OP9-cSCG	Unknown	2527291510	2080528	2534
Bacteria	Permanent Draft	Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage	SAR86 cluster bacterium SAR86A	Unknown	2597489919	1245342	1339
Bacteria	Permanent Draft	23 microbial genomes reconstructed from anaerobic bioreactors	Sulfurovum sp. G1	Beijing Genomics Institute (BGI)	2600254903	1781553	1864
Bacteria	Permanent Draft	Candidatus Cloacamonas acidaminovorans	Candidatus Cloacamonas acidaminovorans Evry	Unknown	642555115	2246820	1868
Bacteria	Permanent Draft	Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage	SAR86 cluster bacterium SAR86B	Unknown	2597489920	1679540	1890
Bacteria	Permanent Draft	CSP_777133	Candidatus Endomicrobia bacterium trichonymphae	Unknown	2014613004	3375854	4350
Bacteria	Permanent Draft	Uncultured BPS1	Uncultured Chlorobi bacterium Chlorobi5	Unknown	2061766010	343942	738
Bacteria	Permanent Draft	Uncultured Spirochaete Spiro3	Uncultured Spirochaete Spiro3	Unknown	2061766011	512975	1181
Bacteria	Permanent Draft	Red algae associated microbial communities from Porphyra blades in Schoodic Point, Maine, United States	Planctomycetes bacterium P1	DOE Joint Genome Institute (JGI)	2509276000	8468922	7084

Data downloadには  
アカウント登録が必要

SPIRE

FAQs Downloads Contribute Search



Explore Environments Explore Taxonomy

# SPIRE

The microbial world at your fingertips.

1M+  
MAGs

700+  
Studies

35B+  
Genes

100+  
Countries

## FAQs

### What is SPIRE?



SPIRE is short for Searchable Planetary-scale mIcrobiome REsource: a one stop shop for microbial data, integrated and consistently processed across habitats and phylogeny, at global scales. SPIRE holds five main types of data, derived from ~100k metagenomic samples in 739 studies encompassing ~500Tbp of publicly available raw metagenomes:

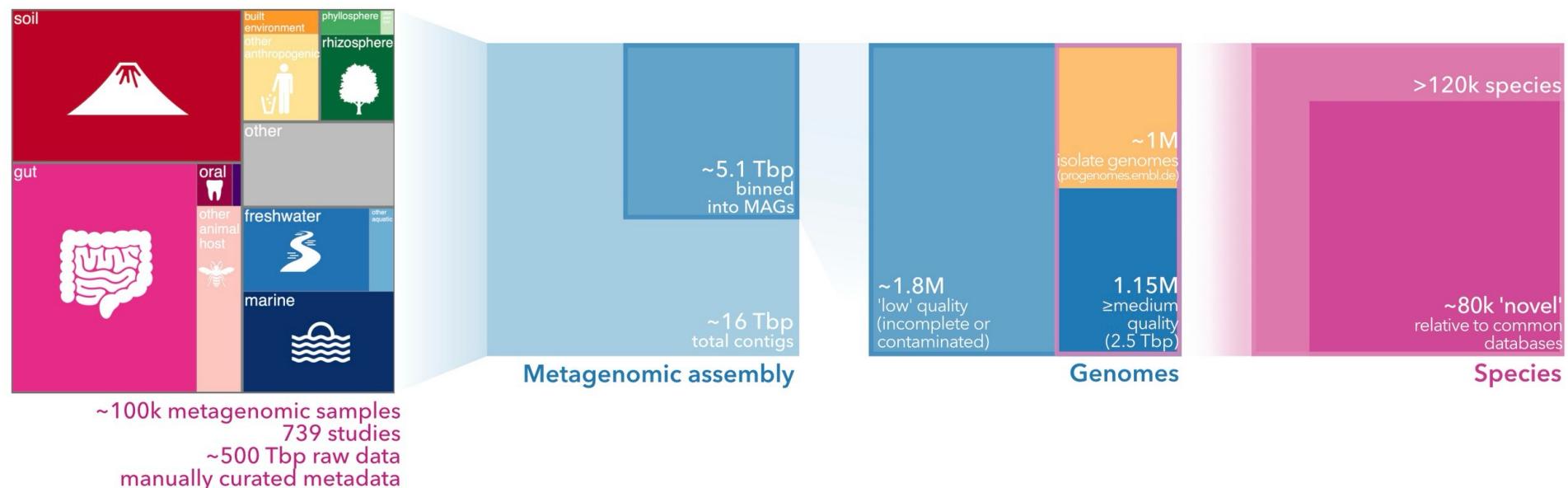
manually curated contextual data, including annotations against a custom microntology (See FAQ: What is microntology?)

~16Tbp worth of metagenomic assemblies (contigs)

~35 billion metagenomic open reading frames (ORFs) called from these contigs, functionally annotated in various ways 1.16 million metagenome-assembled genomes (MAGs) of medium or high quality, co-clustered with isolate genomes into ~120 thousand species-level clusters

a custom mOTU database [mOTUs](#) to allow taxonomic profiling against these species-level clusters, as well as pre-computed taxonomic profiles (coming soon...)

The aim of SPIRE is to provide an integrated view of microbial diversity at various taxonomic granularities, at genome and at gene level, across known microbial habitats.



x fresh water

## Select Filters:

- Habitat Terms
  - + age group
  - air
  - ancient microbiome
  - + anthropogenic
  - aquatic
    - benthic
    - + brackish water
    - briny water
    - fresh water
      - fluvial
      - lacustrine
      - lake bed
      - stream
      - stream bed
    - groundwater
    - ice
    - lentic
    - littoral
    - lotic
    - + marine
    - paludal
    - pelagic
    - saline water
    - sediment
    - + spring
  - + birth term
  - contaminated
  - + host lifestyle
  - + host-associated
  - + oxygen level
  - + ph

Studies: 83

Samples: 2913

Type a keyword...

Sample ID	Study ID
<a href="#">SAMEA103890758</a>	<a href="#">PRJEB19833_trout</a>
<a href="#">SAMEA103890776</a>	<a href="#">PRJEB19833_trout</a>
<a href="#">SAMEA103890767</a>	<a href="#">PRJEB19833_trout</a>
<a href="#">SAMEA103890785</a>	<a href="#">PRJEB19833_trout</a>
<a href="#">SAMEA103890789</a>	<a href="#">PRJEB19833_trout</a>
<a href="#">SAMEA103890781</a>	<a href="#">PRJEB19833_trout</a>
<a href="#">SAMEA103890772</a>	<a href="#">PRJEB19833_trout</a>
<a href="#">SAMEA103890778</a>	<a href="#">PRJEB19833_trout</a>
<a href="#">SAMEA103890769</a>	<a href="#">PRJEB19833_trout</a>
<a href="#">SAMEA103890787</a>	<a href="#">PRJEB19833_trout</a>

Showing 1 to 10 of 2913 results

Previous 1 2 3 ... 292 Next

# SPIRE <http://spire.embl.de/sample/SAMEA103890758>

SPIRE

FAQs Downloads Contribute Search

Search

## SAMEA103890758

ENA Accession: [SAMEA103890758](#)

Study: [PRJEB19833\\_trout](#)

Microntology Tags:

[fresh water](#) [aquatic](#) [host-associated](#) [animal host](#) [digestive tract](#) [intestine](#)

SPIRE Cluster Count: 1

SPIRE Genome Count: 2

## MAGs

Type a keyword...

Spire ID	Genome ...	Num Con...	Contig ...	Complete...	Contamina...	GUNC ...	GUNC ...	Gene C...	Spire Cluster	Downlo
<a href="#">spire_mag_0024440_0</a>	644814	18	70291	88.99	0.01	0.25	0.38	608	<a href="#">spire_v1_095_00021366_1</a>	<a href="#">↓</a>
<a href="#">spire_mag_0024440_1</a>	204712	41	5414	3.53	0.0	0.0	0.09	165	None	<a href="#">↓</a>

An error happened while fetching the data

[Previous](#) [Next](#)

[Download Table Data](#)

## Sample Downloads

File Type	Tool	Download
Assembly	megahit/1.2.9	<a href="#">↓</a>
EggNOG Annotation	eggNOG-mapper/2.1.3	<a href="#">↓</a>
tRNA	trnascan/2.0.9	<a href="#">↓</a>
CRISPR Sequences	minced/0.4.2	<a href="#">↓</a>
Antibiotic Resistance Annotations	deepARG/1.0	<a href="#">↓</a>

# 代表的なMAG DB間の比較

	Admin	Database URL	環境の多さ	プロジェクトの多さ	Number of MAGs in July 2024
NCBI Datasets	NCBI, USA	<a href="https://www.ncbi.nlm.nih.gov/datasets/genome/">https://www.ncbi.nlm.nih.gov/datasets/genome/</a>	未整理	大	443,763
IMG/M	JGI, USA	<a href="https://img.jgi.doe.gov/cgi-bin/m/main.cgi">https://img.jgi.doe.gov/cgi-bin/m/main.cgi</a>	小	小	25,507
SPIRE	EMBL, EU	<a href="http://spire.embl.de/">http://spire.embl.de/</a>	中	中	1,160,000
MGnify	EBI, EU	<a href="https://www.ebi.ac.uk/metagenomics">https://www.ebi.ac.uk/metagenomics</a>	小	大	478,810
Microbiome Datahub	NIG, Japan	<a href="https://mdatahub.org/">https://mdatahub.org/</a>	大(整理中)	大	218,653

環境情報をいかに整理するかが各MAG DBの今後の課題