

# 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)

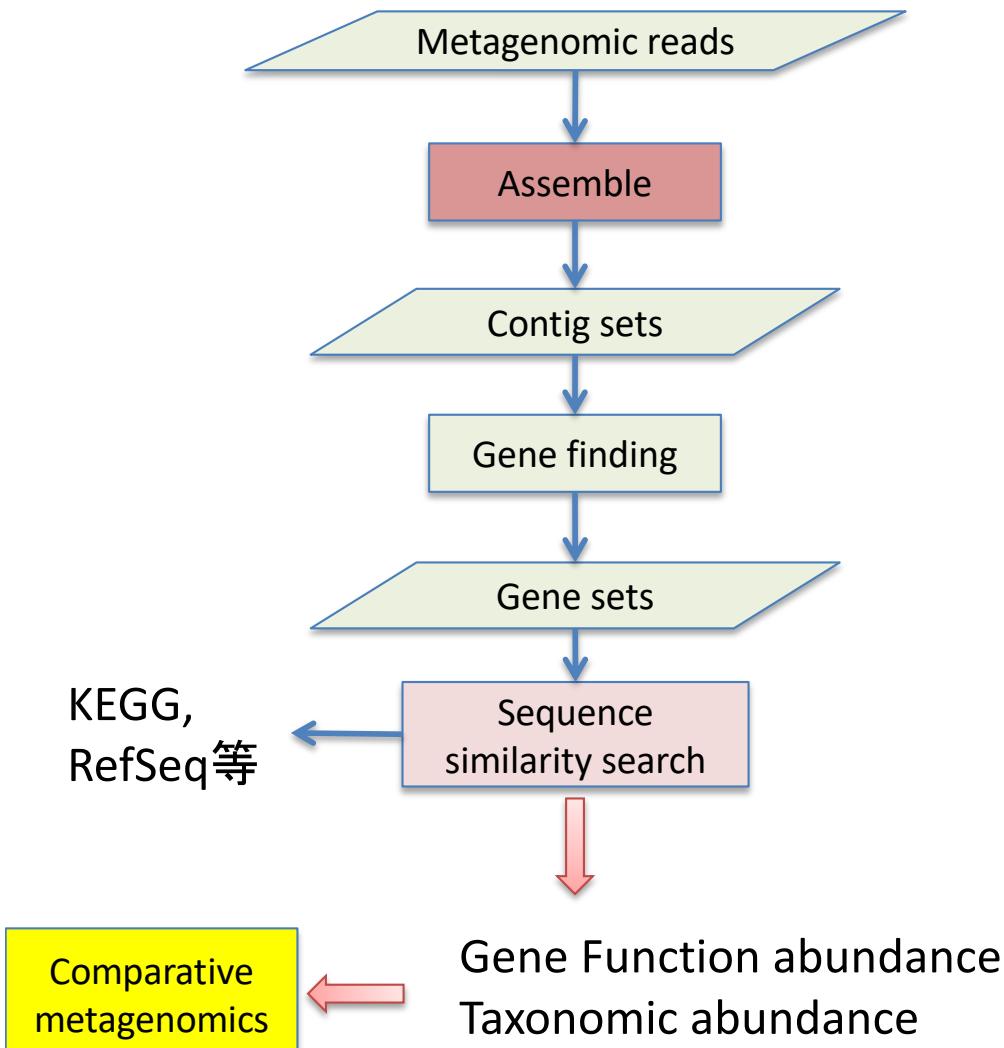
国立遺伝学研究所

先端ゲノミクス推進センター

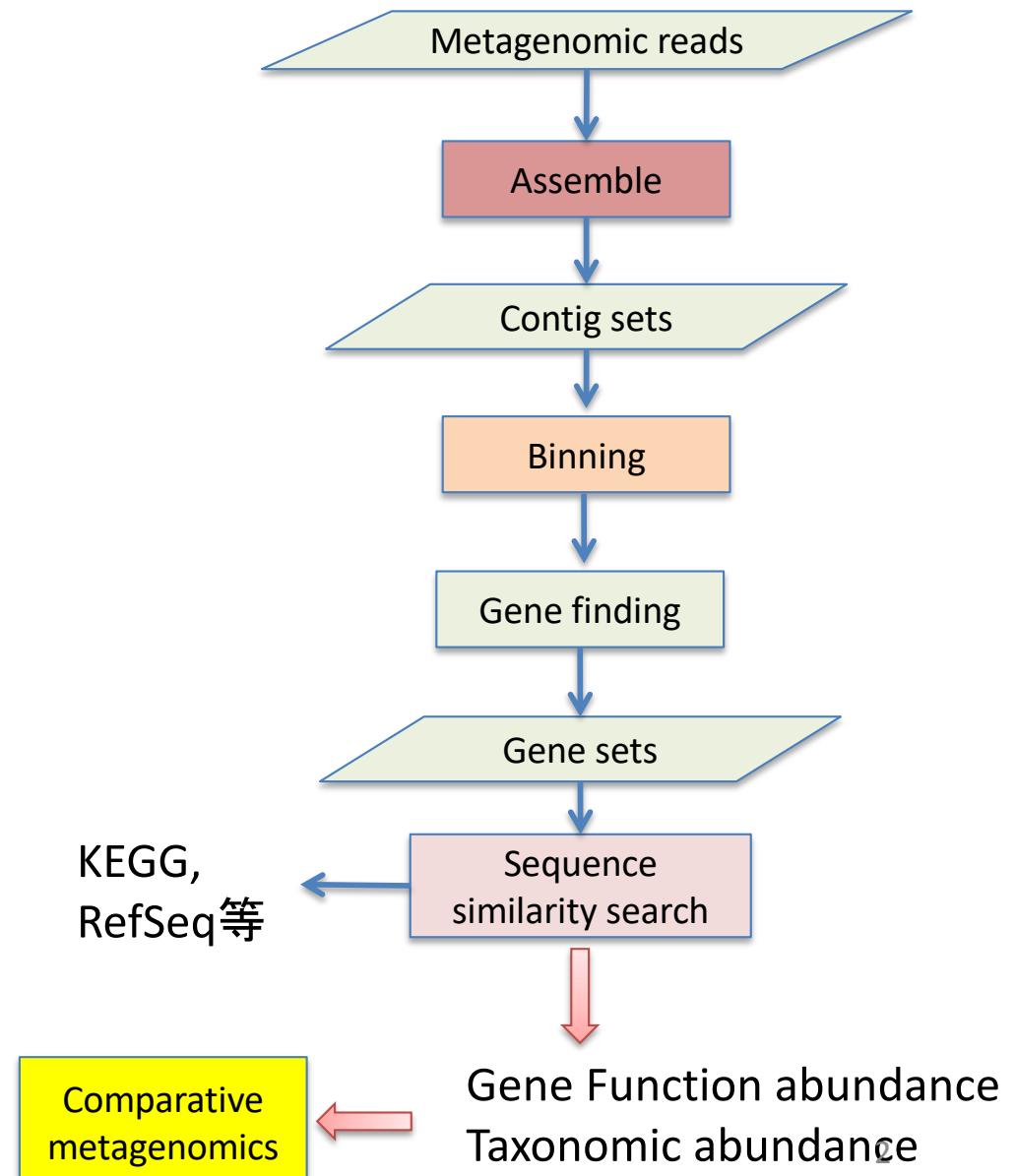
[hmori@nig.ac.jp](mailto:hmori@nig.ac.jp)

# 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)

The GTDB homepage features a large, detailed photograph of a tree trunk and branches 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 counts for Phylum, Class, Order, Family, Genus, and Species.

Concatenated protein phylogenies

毎年4月更新

NCBI Assembly DBからゲノム配列は取得

Isolate+MAG

Welcome to GTDB

# GENOME TAXONOMY DATABASE

402,709 genomes

Release 08-RS214 (28th April 2023)

Archaea distribution chart showing counts for Phylum, Class, Order, Family, Genus, and Species.

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. The main title "GENOME TAXONOMY DATABASE" is prominently displayed in large white letters. Below it, the text "Welcome to GTDB", "596,859 genomes", and "Release 09-RS220 (24th April 2024)" are visible. A detailed breakdown of the taxonomic levels is shown in a bar chart at the bottom right: PHYLUM (19), CLASS (64), ORDER (166), FAMILY (564), GENUS (1,847), and SPECIES (5,869). The total count for ARCHAEA is listed as (12,477). The bottom left corner includes the logo of The University of Queensland Australia.

- Concatenated protein phylogenies
- 毎年4月更新
- NCBI Assembly DBからゲノム配列は取得
- Isolate+MAG

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

- 昨年のversionと比べてゲノムが約20万増加
- Phylumは13增加、Genusは4千属、Speciesは2.8万種増加

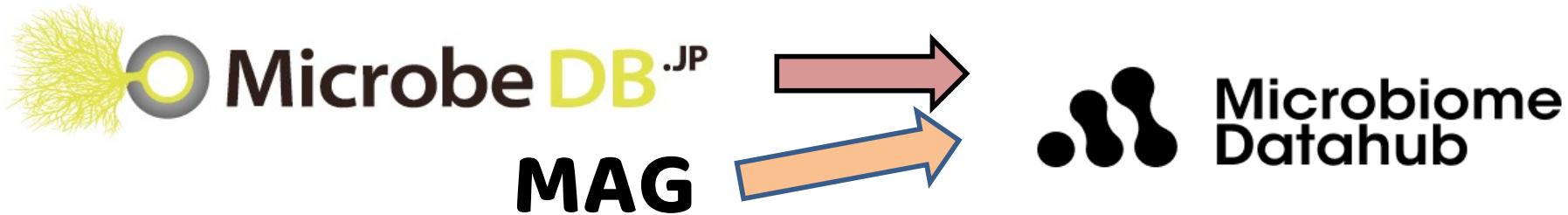
# 代表的なMAG DB間の比較

	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

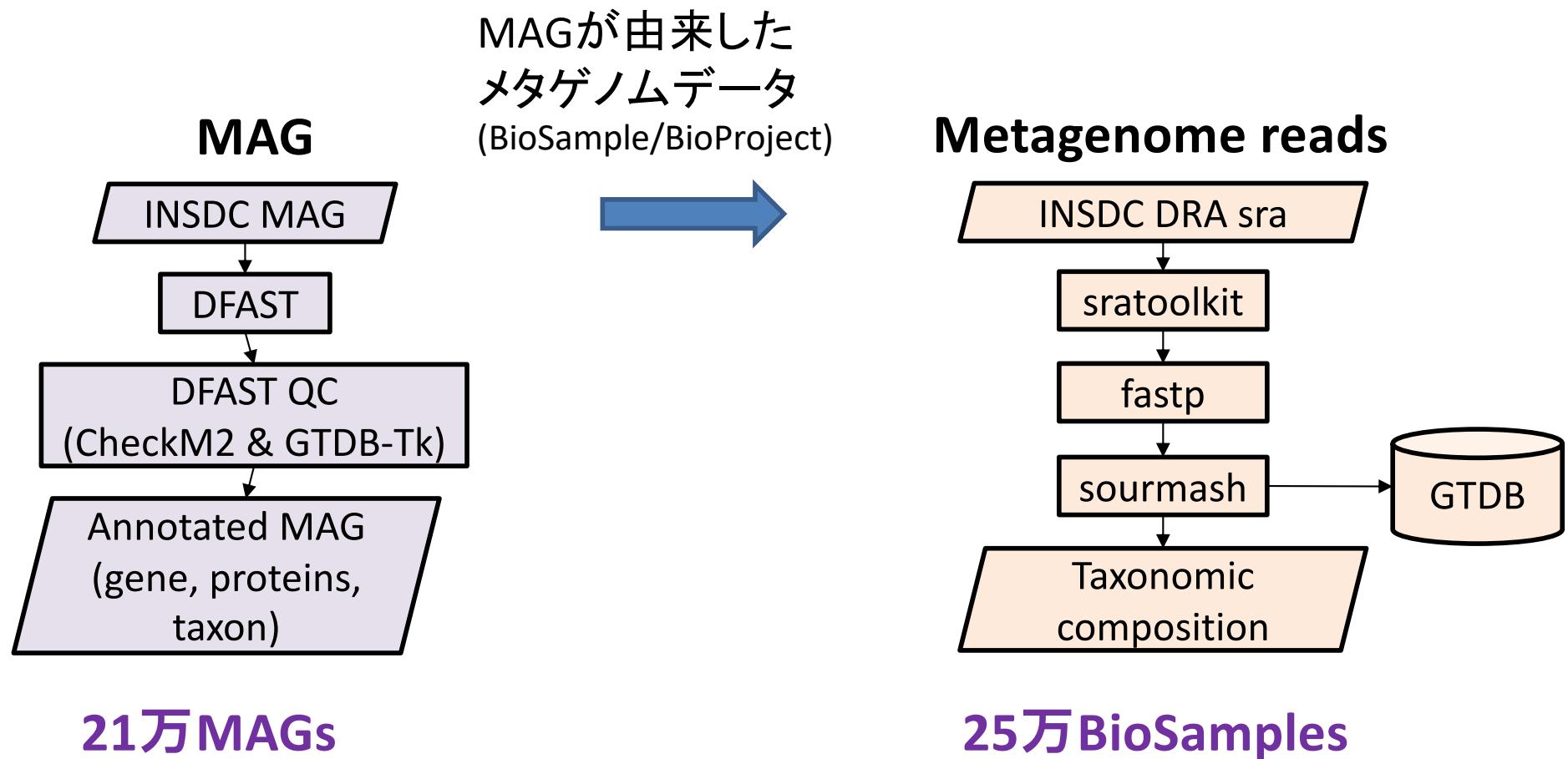
The screenshot shows the current user interface of the Microbiome Datahub. At the top, there is a yellow header bar with the "Microbiome Datahub" logo and a "Document" link. Below the header is a search bar with the placeholder "Search Keyword" and a magnifying glass icon. To the right of the search bar is a close button (X). On the left side, there is a sidebar with several filter sections, each with a toggle switch:

- Environment**: Toggled on. Options include soil, marine, freshwater, hot spring, sediment, air, gut, oral, skin, reproductive system, and human activity related.
- Genome taxon**: Toggled on. This section is currently empty.
- Genome Category**: Toggled on. Options include Isolate complete, Isolate draft, MAG high quality, and MAG low quality.
- MAG source**: Toggled on. Option: INSDC.

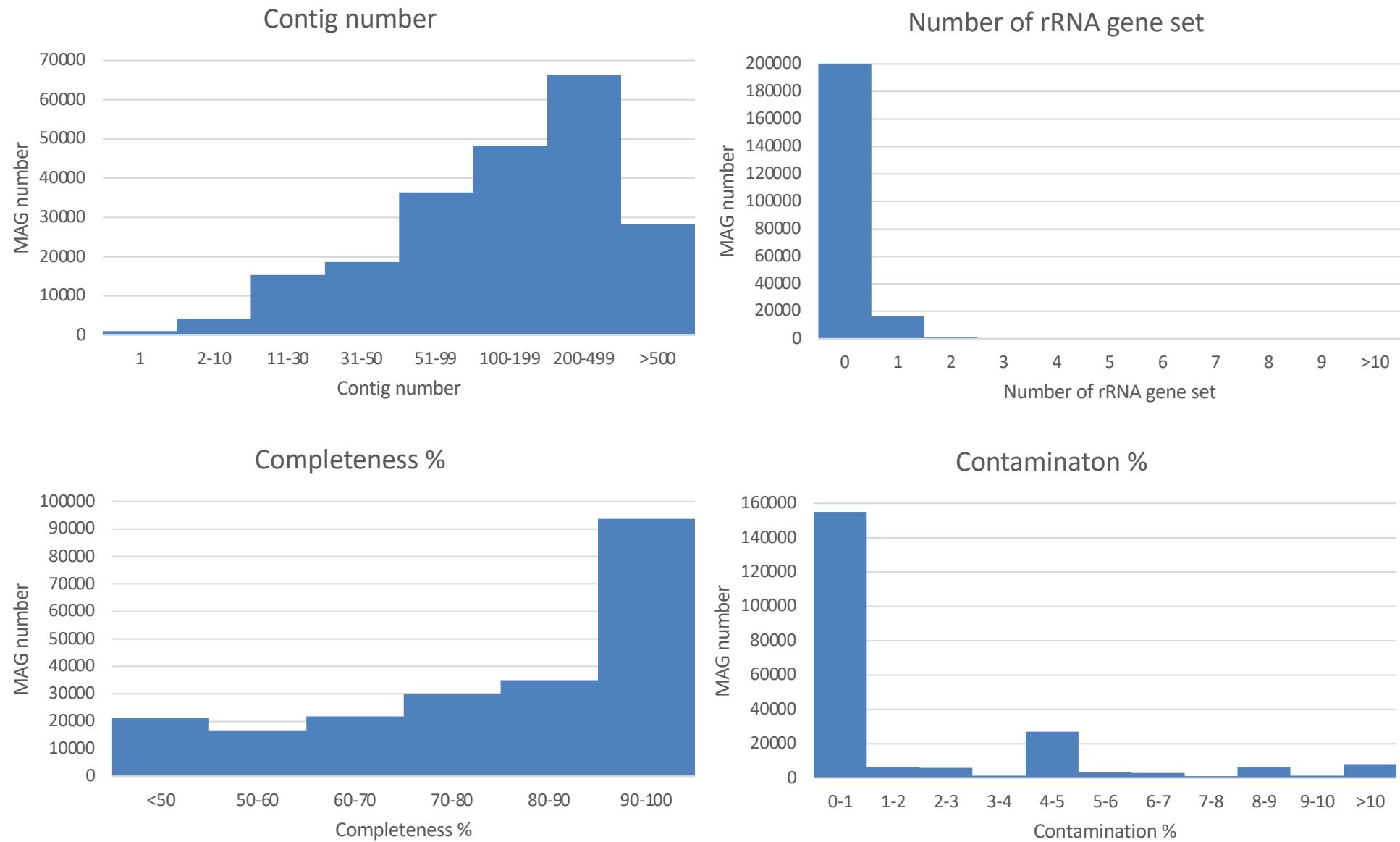
On the right side, there are two tabs: "PROJECT" and "GENOME". The "GENOME" tab is selected, indicated by an orange underline. Below the tabs, there is a search result summary: "10 / 218653" followed by "order by Date Created" with up and down arrows. The main content area displays a list of genome entries:

Genome Name	Accession
<b>Candidatus Thermoplasmatota archaeon</b>	GCA_029762495.1
Environment Host taxon BioSamples 0 Data size (GB) 0.0 GB Date Created 2023-04-16	
<b>Methanocalculus sp.</b>	GCA_029762515.1
Environment Host taxon BioSamples 0 Data size (GB) 0.0 GB Date Created 2023-04-16	
<b>Candidatus Thermoplasmatota archaeon</b>	GCA_029762525.1
Environment Host taxon BioSamples 0 Data size (GB) 0.0 GB Date Created 2023-04-16	
<b>Candidatus Thermoplasmatota archaeon</b>	GCA_029762535.1
Environment Host taxon BioSamples 0 Data size (GB) 0.0 GB Date Created 2023-04-16	
<b>Candidatus Thermoplasmatota archaeon</b>	GCA_029762575.1
Environment Host taxon BioSamples 0 Data size (GB) 0.0 GB Date Created 2023-04-16	
<b>Methanolobus sp.</b>	GCA_029762595.1
Environment Host taxon BioSamples 0 Data size (GB) 0.0 GB Date Created 2023-04-16	
<b>Methanomicrobiaceae archaeon</b>	GCA_029762625.1

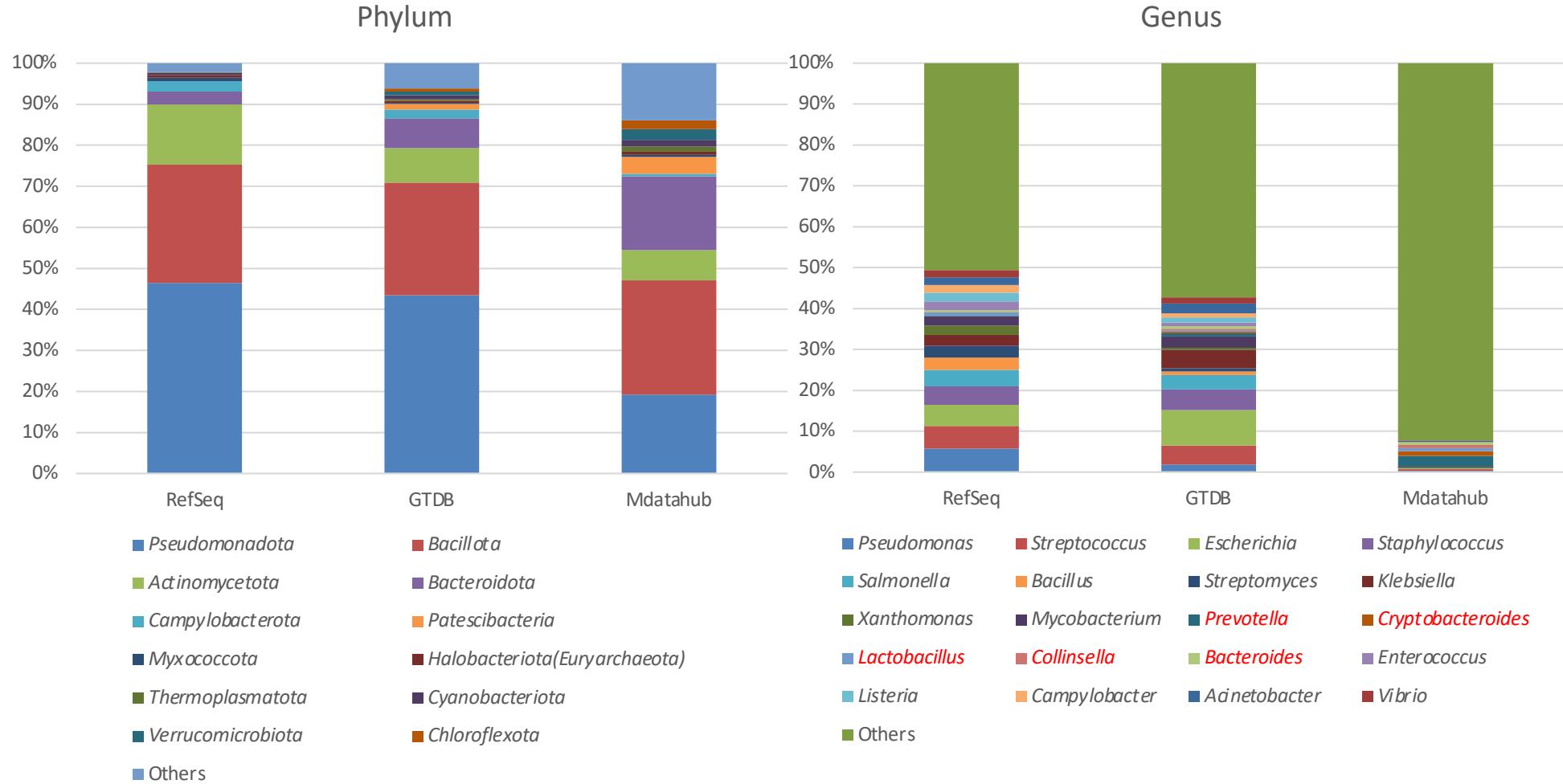
# Microbiome DatahubにおけるMAGの収集とアノテーション



# 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

# 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)

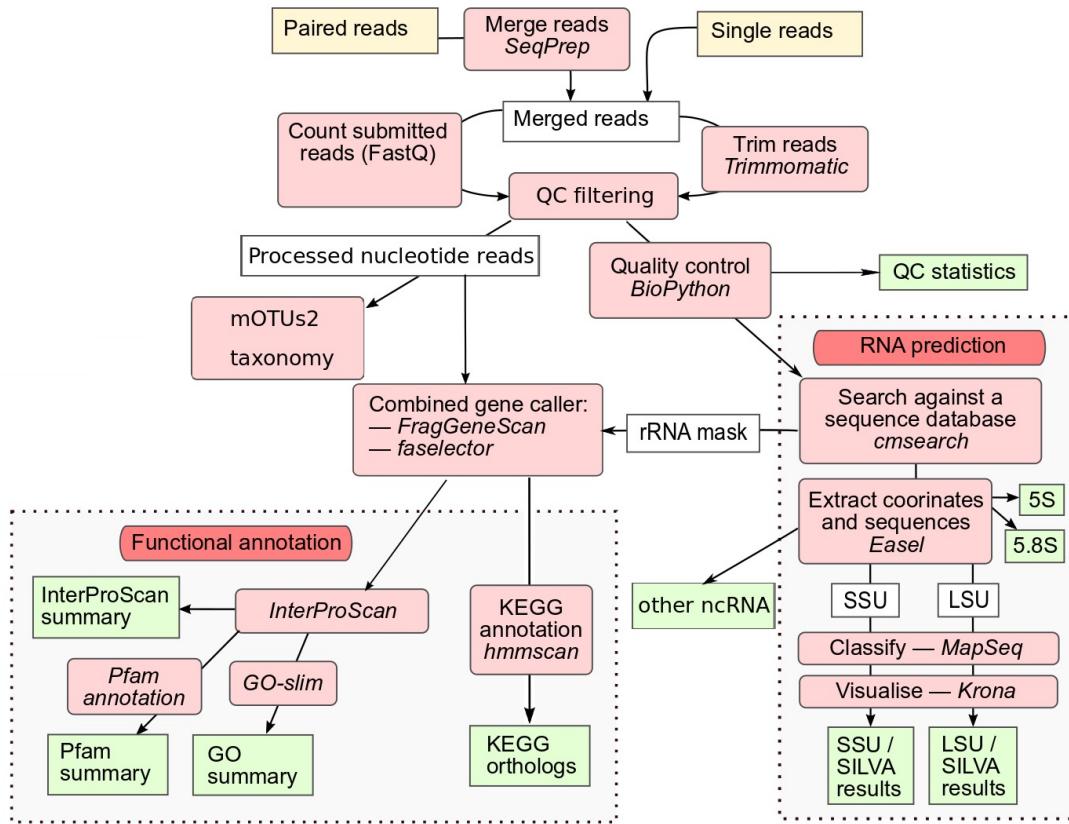


Wastewater  
(4358)



Food production  
(2923)

**View all biomes**



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

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](#)

## Latest eukaryotic RefSeq annotations

From the [NCBI Eukaryotic Genome Annotation Pipeline \(EGAP\)](#)

### Annotation in progress

Species	Assembly	Release	Freeze date
<a href="#">Branchiostoma lanceolatum</a> (amphioxus)	<a href="#">kIBraLanc5.hap2</a>	<a href="#">GCF_035083965.1-RS_2024_07</a>	Jul 12, 2024
<a href="#">Euwallacea similis</a> (beetles)	<a href="#">ESF131.1</a>	<a href="#">GCF_039881205.1-RS_2024_07</a>	Jul 12, 2024

### Recently completed

Species	Assembly	Release	Release date	Action


**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](#)

[Home](#)

[IMG Survey](#)

---

**IMG Content**

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

Last Datasets Added On:

Genome	2024-06-13
Metagenome	2024-07-14

[!\[\]\(2a0a660e8dbc7fc983c587cf59e6cd54\_img.jpg\) Project Maps: Genomes Metagenomes](#)

[!\[\]\(7c27c570ccf2748ecdd8300aed0365eb\_img.jpg\) JGI's Integrated... ::](#)

[!\[\]\(ddabc0c6b79bc12602a6f1c747b91e8e\_img.jpg\) \*\*IMG Webinar YouTube Playlist\*\*](#)

[Download Isolate Genomes](#)

[Download Metagenome BINs](#)

Quick Genome Search:

[Login into IMG/MER](#)

---

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](#)).

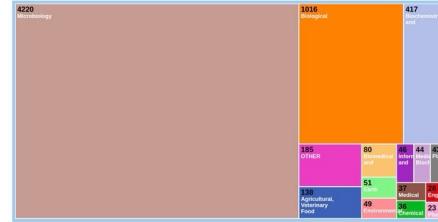
If you use IMG web resources or data to assist in research publications or proposals. Please cite: **IMG - Chen et al., 2022 (Nucleic Acids Research, gkac976).**  
**GOLD v9 - Mukherjee et al., 2022 (Nucleic Acids Research, gkac974).**

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

Sequenced at:	Isolates		SAGs		MAGs	
	JGI	All	JGI	All	JGI	All
Bacteria	13458	113595	2282	9447	5357	14457
Archaea	220	1365	318	601	248	1002
Eukarya	370	591	0	0	0	0
Viruses	13	16510	0	75	0	2918

(Only data sets with GOLD metadata were counted.)

[View Citations](#)



A treemap visualization showing the distribution of citations across various fields of study. The largest category is Microbiology (4220), followed by Chemical and Environmental Engineering (4116), and Chemistry (417). Other categories include Earth and Planetary Sciences, Agricultural and Veterinary Medicine, Medical and Engineered Systems, and Biochemistry.

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	173	203	Air	60	116	Algae	53	168
Bioreactor	902	1074	Aquatic	16111	20696	Annelida	138	152
Bioremediation	25	61	Terrestrial	9184	12514	Arthropoda	0	1
Biotransformation	3	8				Arthropoda: Crustaceans	0	7
Built environment	297	1428				Arthropoda: Insects	123	324
Food production	0	61				Birds	17	48
Industrial production	20	43				Cephalochordata	0	4
Lab culture	4	4				Fish	0	29
Lab enrichment	5	69				Fungi	145	146
Modeled	11	91				Human	2	1001

15

SPIRE

FAQs Downloads Contribute Search



**SPIRE**

Explore Environments Explore Taxonomy

The microbial world at your fingertips.

1M+	700+	35B+	100+
MAGs	Studies	Genes	Countries

Privacy Policy Bork Group