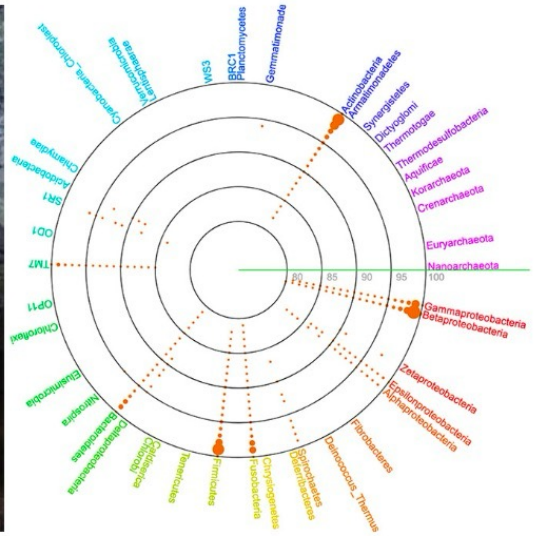
2026年1月30日
第7回バイオ生成AI研究会

学習データについて
-Microbiome Datahub-

国立遺伝学研究所
情報研究系
森　宙史
Hiroshi Mori
hmori@nig.ac.jp

HOME
Members
Research
Publications
DB&Tools
Resources
Link
Access&Contact

国立遺伝学研究所 ゲノム多様性研究室



本研究室では、バイオインフォマティクス技術を用いて微生物などが持つゲノムの多様性を解明する研究に取り組んでいます。メタゲノム解析技術の進展によって、培養が難しい微生物も含めてゲノム解読が可能になり、また、メタゲノムデータの一種であるAncient DNAデータを用いることで、数万年以上前に絶滅した生物のゲノム解析も可能になりました。我々は森が兼任する遺伝研の先端ゲノミクス推進センターと強固に連携し、最先端のゲノム解析技術とバイオインフォマティクス解析技術を武器に未だ未知な部分が多い生物のゲノムの多様性に関する幅広い研究を進めております。

Since 2021-

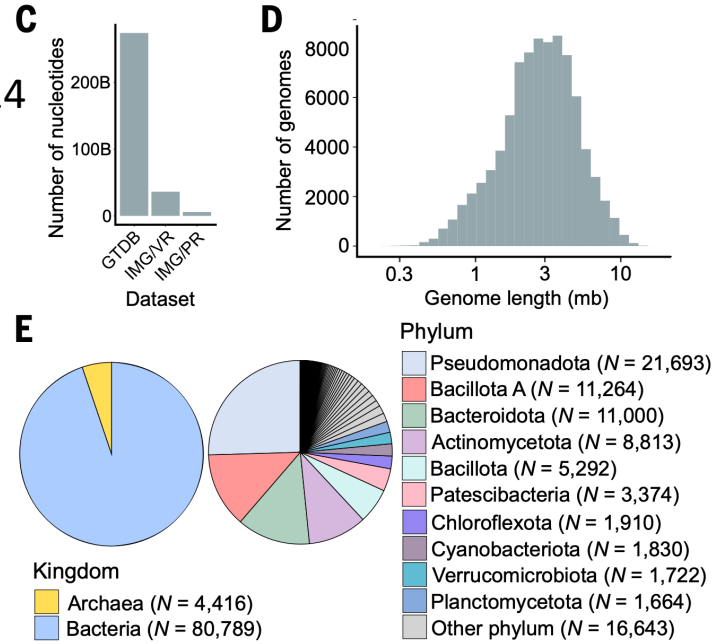
- 微生物のゲノム解析
- 様々な環境のメタゲノム解析
- Ancient DNA解析

- これらに関わる情報解析ツール・DBの開発

研究室webページ https://www.genome.id

様々なGLMの学習データ

Nguyen E. et al. Science 2024



- Evo & TrinityDNA, etc.

  OpenGenome dataset

  GTDB v214.1 原核生物の培養株のゲノムとMAG
  IMG/VR　原核生物に感染するphage
  IMG/PR　原核生物が持つplasmid

  85,205 prokaryotes genomes

- GenomeOcean　Wu W. et al. bioRxiv 2025 (JGI)

  6 Metagenomes (HMP, Tara Ocean, Forest soil, soil, Antarctic Lake, Lake)

- Evo2　Brixi G. et al. bioRxiv 2025

  OpenGenome2 dataset

  OpenGenome dataset +
  GTDB v220 (2024.4) added new species genomes (28,174)

  113,379 prokaryotes genomes

  NCBI Eukaryote genomes (2024.5) clustering by Mash

  15,032 genomes

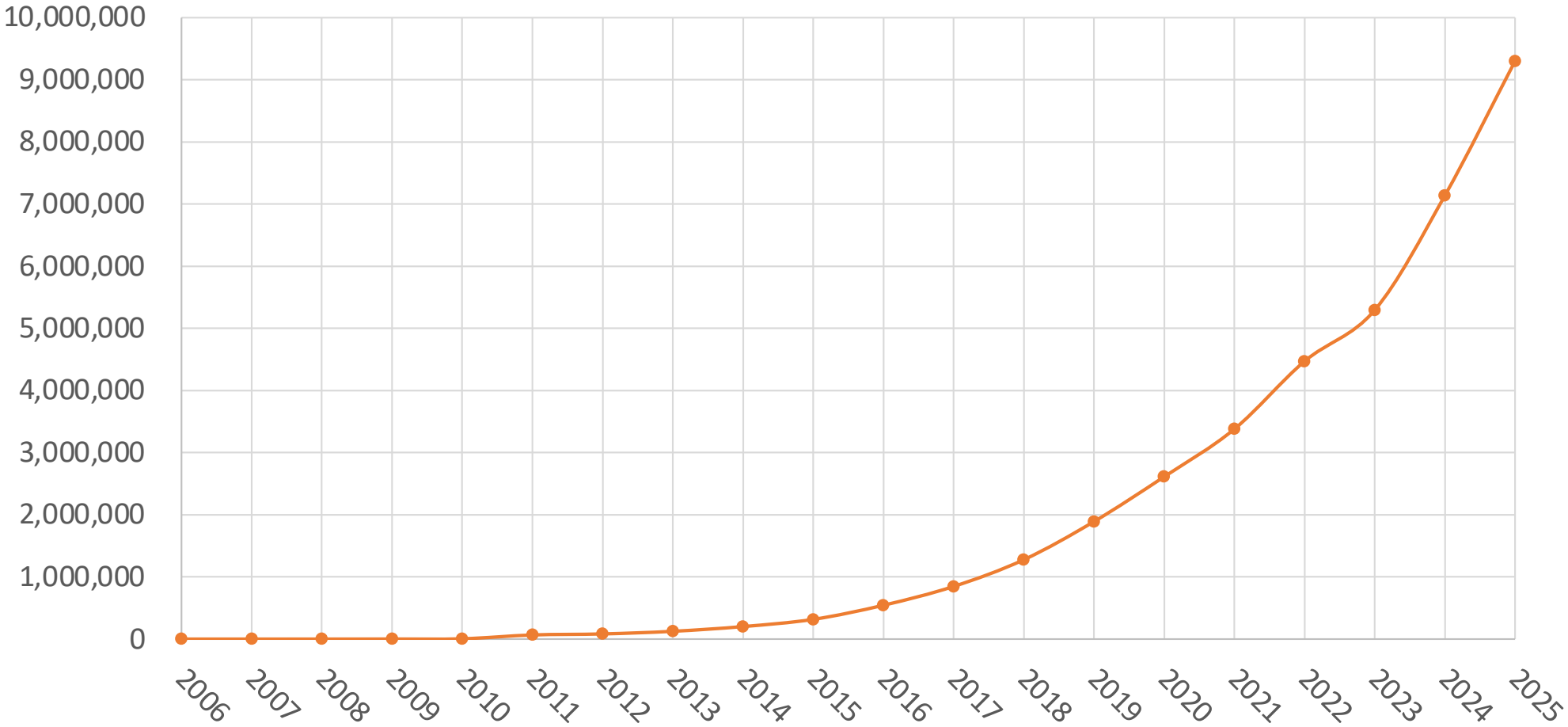
  Metagenomes and MAGs (IMG/M, MGnify, MG-RAST, Tara Ocean, Animal Gut)

  41,253 metagenomes

  NCBI Organelle

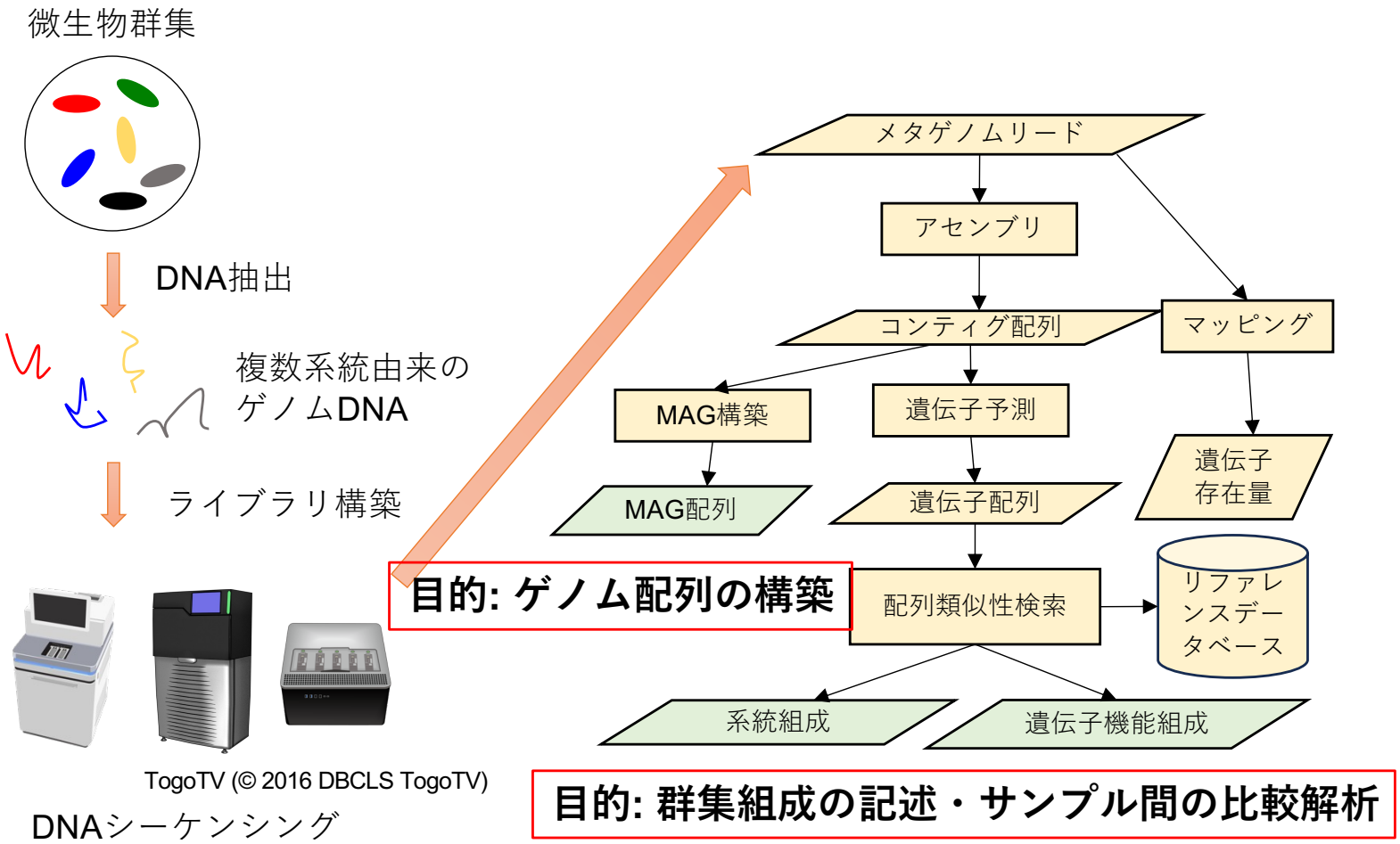
  33,457 organelle genomes

公共のマイクロバイオームサンプルの総数(2025年12月時点)



この2年ほどは年間200万サンプルほど増加
ショットガンメタゲノムに絞ると上記の約1/10

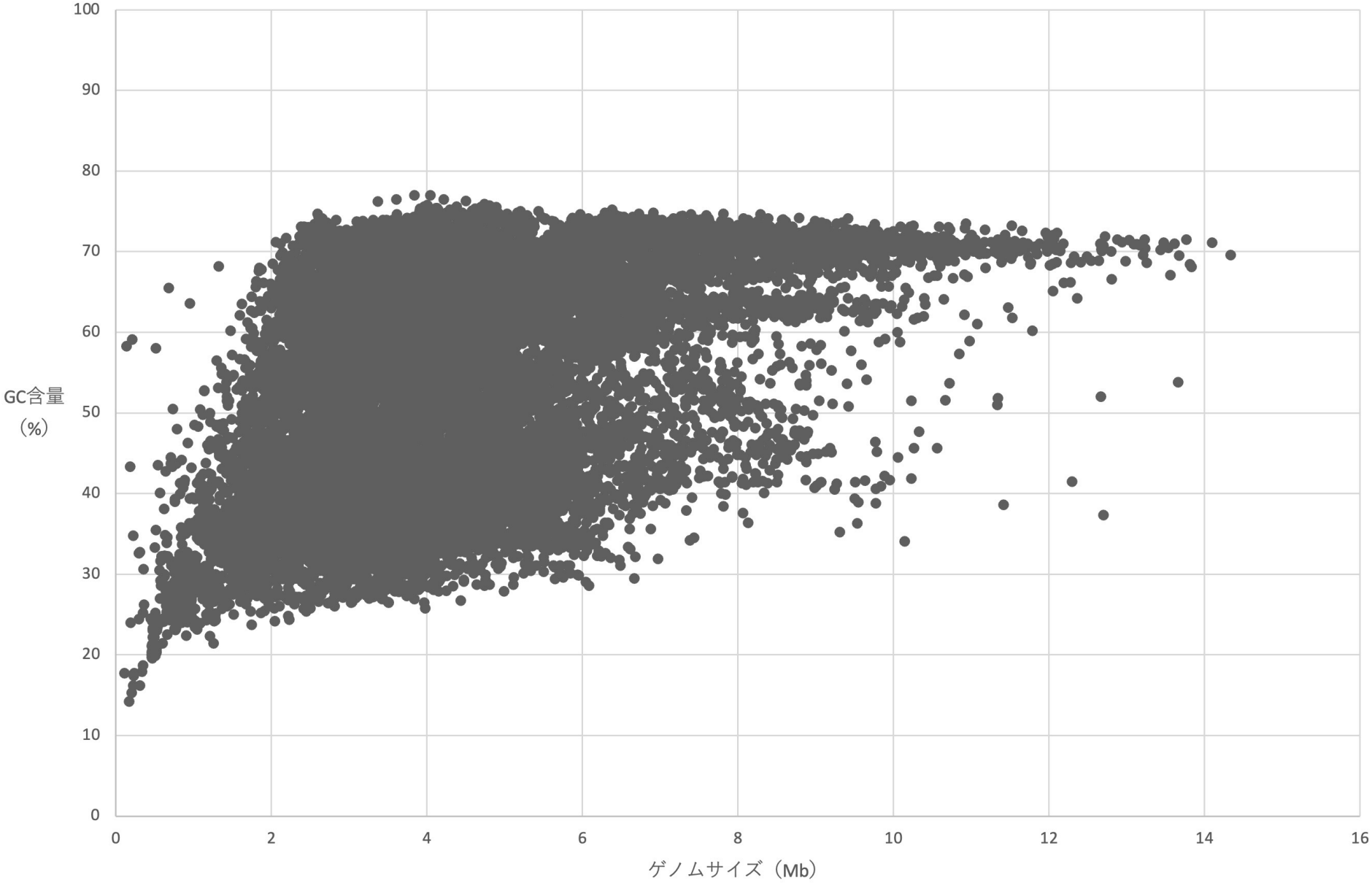
Metagenomic sequencing analysis
(メタゲノム解析, ショットガンメタゲノム解析)



Metagenome Assembled Genome (MAG)

バクテリアゲノムのGC含量の多様性

森,「原核生物のゲノム」, 理論生物学事典, 2025



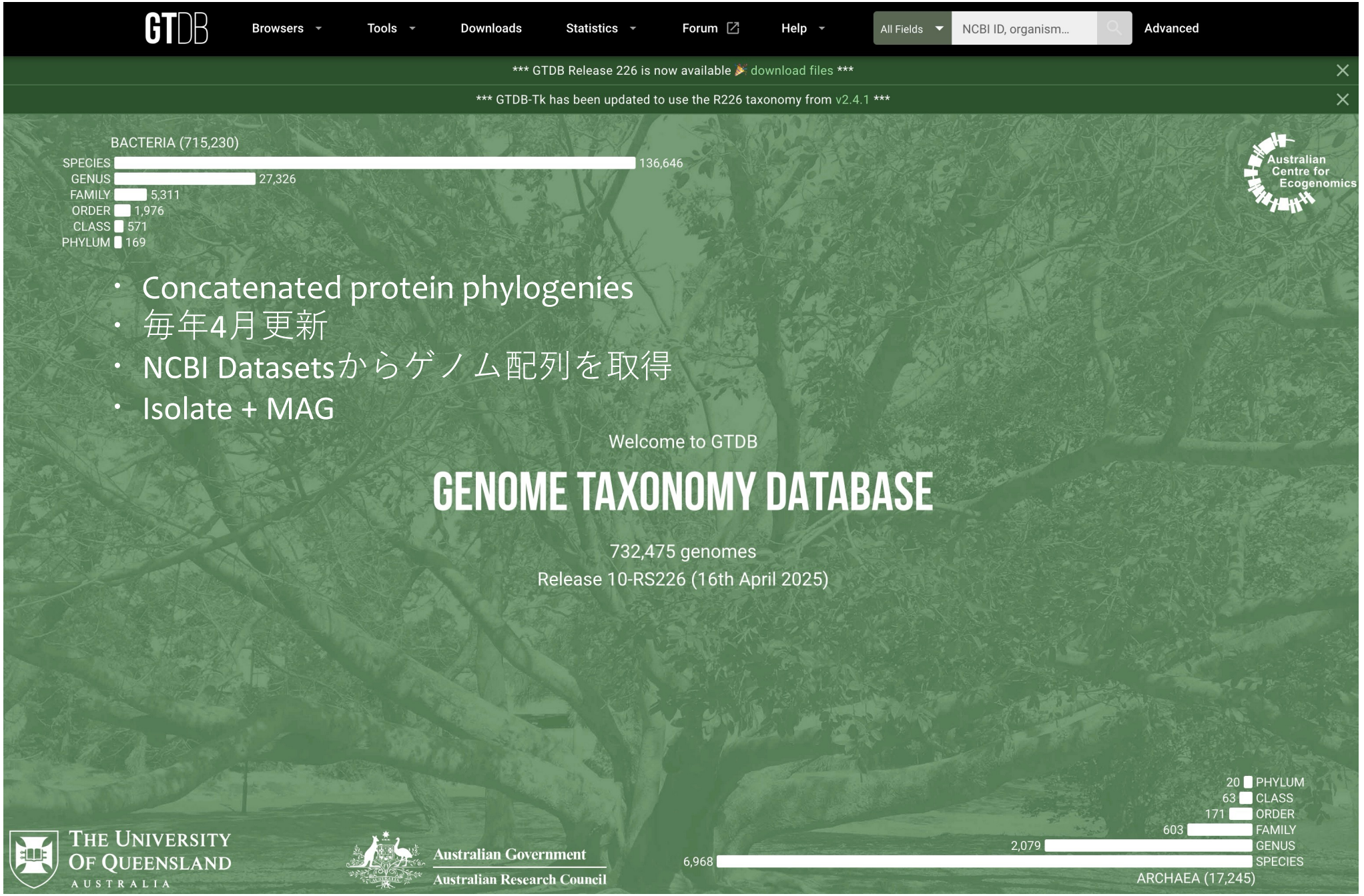
NCBI Datasetsより原核生物18600ゲノムの統計情報をダウンロードして作成

k-mer組成やcoverageを用いてメタゲノムContigを分ける(binning)ツールの例

ツール名	特徴的な統計手法	使用する特徴量	論文
CONCOCT	ガウス混合モデル（GMM）	Coverage, tetranucleotide頻度	Alneberg J. et al. Nature Methods. 2014
MaxBin2	期待値最大化（EM）アルゴリズム	Coverage, tetranucleotide頻度	Wu Y.W. et al. Bioinformatics. 2015
MetaBAT2	グラフクラスタリング	Coverage, tetranucleotide頻度	Kang D.D. et al. PeerJ. 2019
SemiBin2	自己教師あり学習 + DL	Coverage, tetranucleotide頻度, taxonomic information	Pan S. et all. Bioinformatics. 2023
COMEBin	**対照学習** (Contrastive learning)	Coverage, tetranucleotide頻度, Contig内k-mer共起情報	Wang Z. et al. Nature Commun. 2024

メタゲノムデータの全体像を議論するのではなく、優占系統のドラフトゲノム配列を抽出して各ゲノムが持つ機能について議論する

https://gtdb.ecogenomic.org/



- Concatenated protein phylogenies
- 毎年4月更新
- NCBI Datasetsからゲノム配列を取得
- Isolate + MAG

培養を経ずにメタゲノム配列データから
機能や進化的な類縁関係が推定された系統が多数存在

In December 2025

Taxonomic rank	Cultured and validly described taxa	Genome-based putative taxa including uncultured taxa
Phylum	46	189
Class	124	634
Order	253	2,147
Family	654	5,914
Genus	3,908	29,405
Species	21,672	143,614

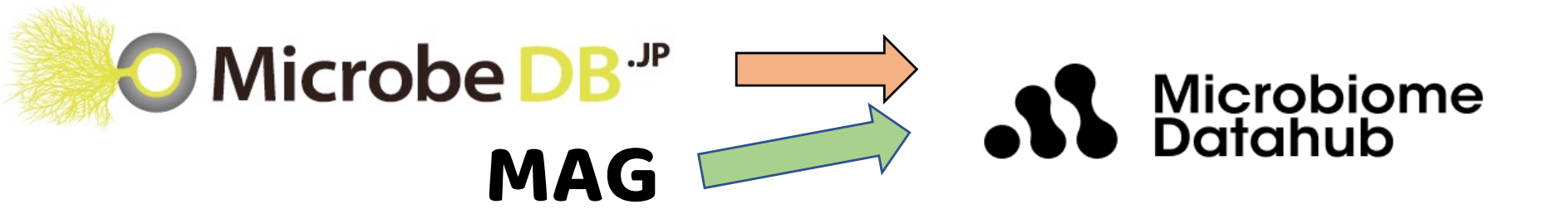
List of prokaryotic names with standing in nomenclature (https://lpsn.dsmz.de/text/numbers)
Genome Taxonomy Database R226 (https://gtdb.ecogenomic.org/stats)

正式に記載された種の約9割が、4 phyla由来

Pseudomonadota, Actinomycetota, Bacillota, Bacteroidota

本プロジェクトの研究開発の目標

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、微生物エンサイクロペディアMicrobeDB.jpをマイクロバイオーム研究の国際的なデータハブMicrobiome Datahubへ発展させることを目標とする。



https://mdatahub.org

国立研究開発法人
科学技術振興機構
Japan Science and Technology Agency

統合化推進プログラム
(DICP)

研究開発実施体制 & 謝辞

国立研究開発法人 科学技術振興機構 Japan Science and Technology Agency
統合化推進プログラム (DICP)

国立遺伝学研究所

森宙史, 藤澤貴智, 東光一, 谷澤靖洋, 飯塚朋代, 中村保一
MAGを中心としたMicrobiome Datahubの開発について主体的に研究開発を行い、各分担グループと連携して研究開発を進める

基礎生物学研究所

内山郁夫, 千葉啓和, 西出浩世, 河合幹彦
MAGのオーソログアサインメントについて主体的に研究開発を行う

東京科学大学

山田拓司, 中川善一
マイクロバイオーム論文からのマニュアル・自動でのメタデータ抽出、ヒトマイクロバイオーム関連のキラーデータセットの開発について主体的に研究開発を行う

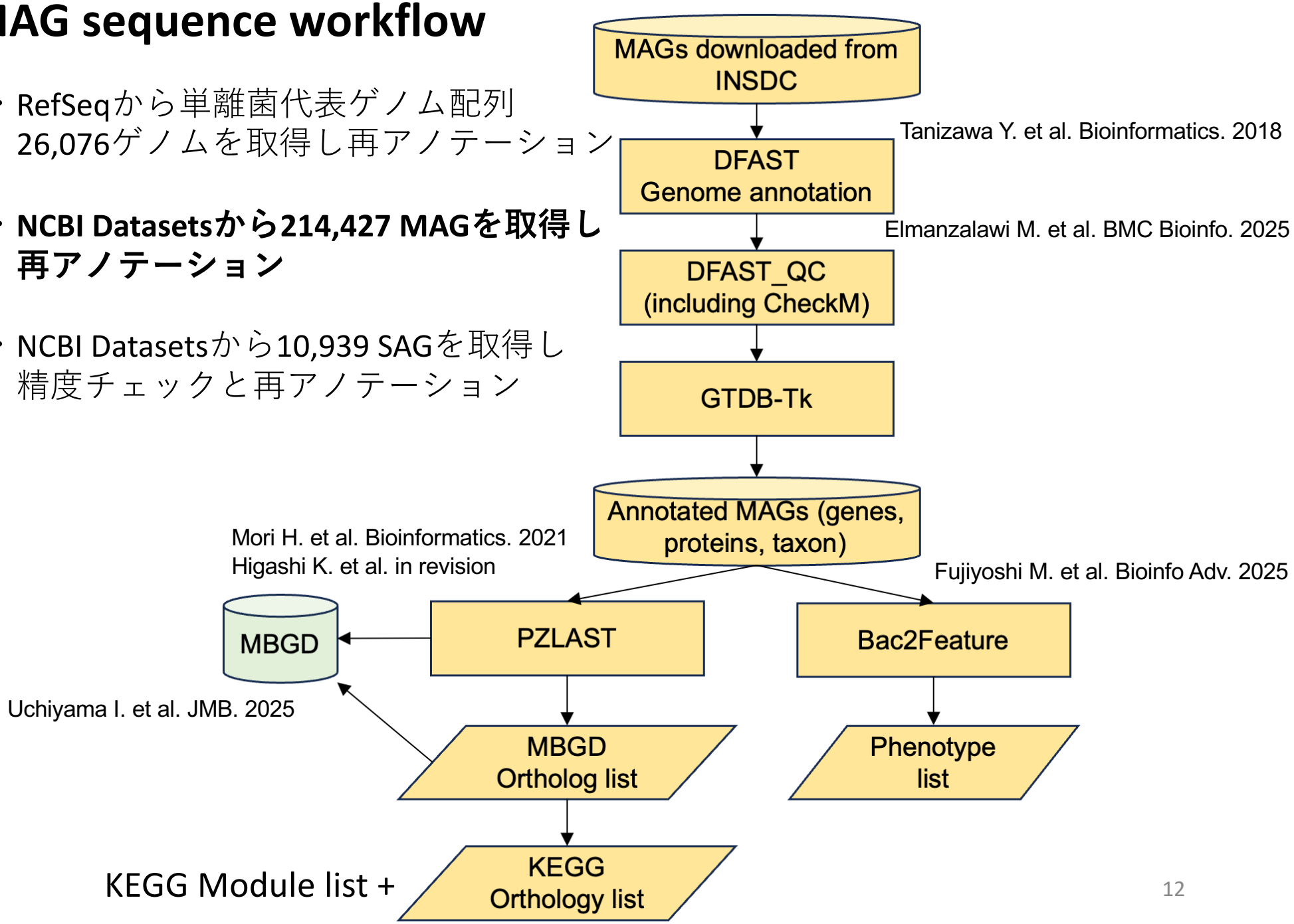
京都大学

松井求, 西村祐貴(東大), 藤吉真生(東大), 鈴木誉保(東大)
MAGのPhenotype推定について主体的に研究開発を行う

Microbiome Datahub
MAG sequence workflow

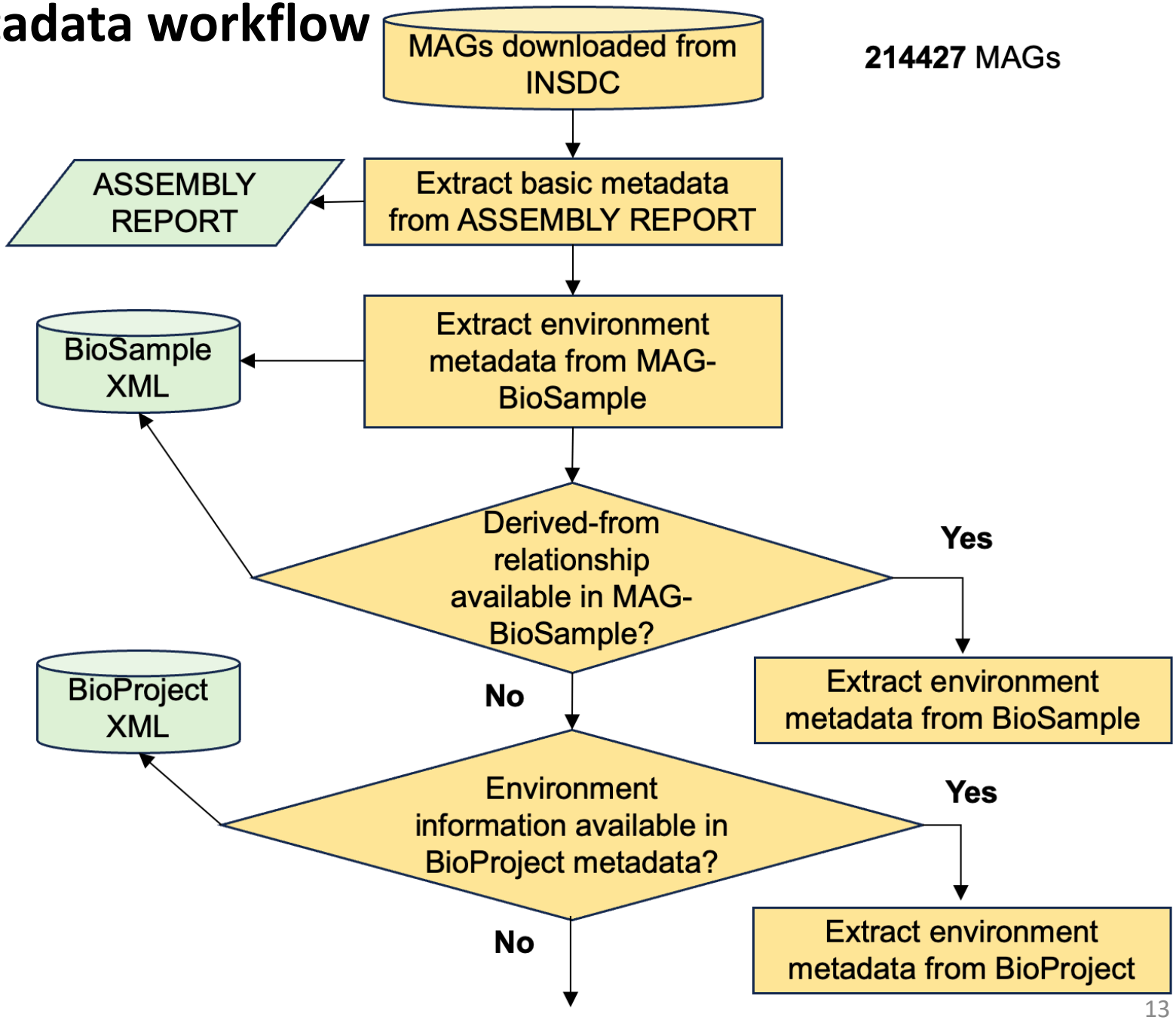
Mori H. et al. in revision

- RefSeqから単離菌代表ゲノム配列26,076ゲノムを取得し再アノテーション
- **NCBI Datasetsから214,427 MAGを取得し再アノテーション**
- NCBI Datasetsから10,939 SAGを取得し精度チェックと再アノテーション



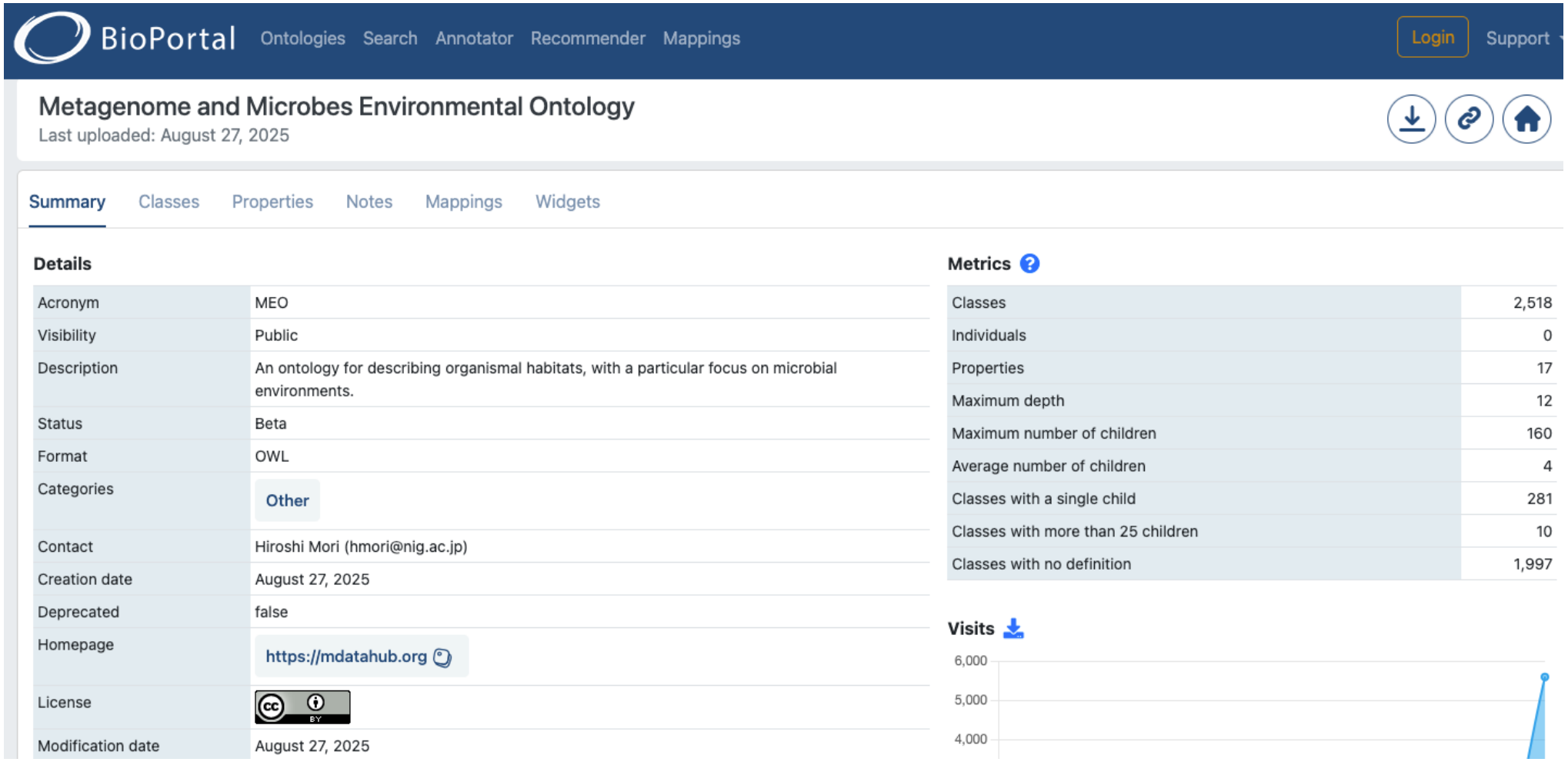
Microbiome Datahub
MAG metadata workflow

Mori H. et al. in revision



Microbiome Datahubにおける環境アノテーション

https://bioportal.bioontology.org/ontologies/MEO



MEO version 1.0をBioPortalで2025年8月に公開

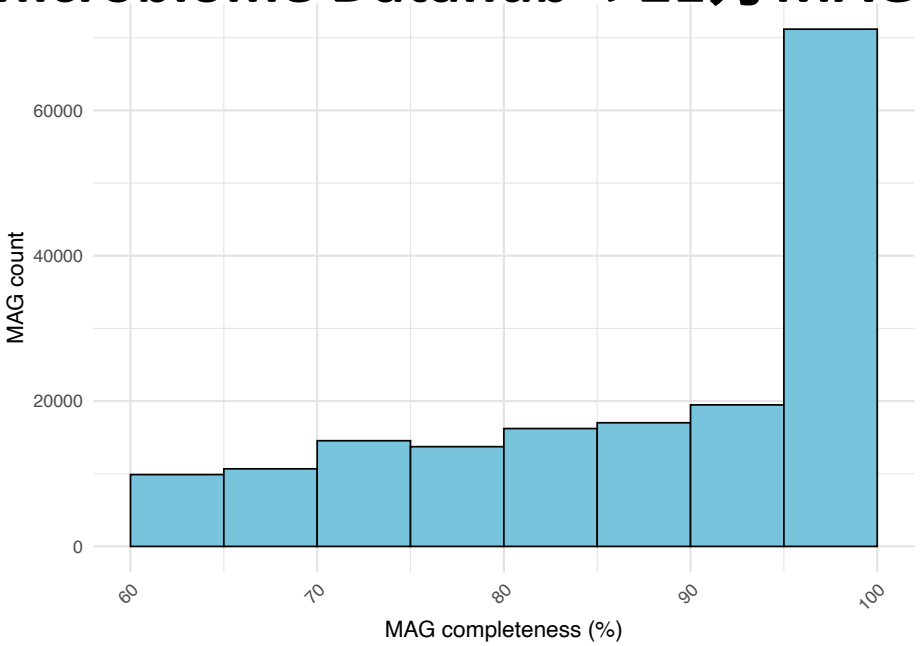
自動アノテーションと手動アノテーションを使い、

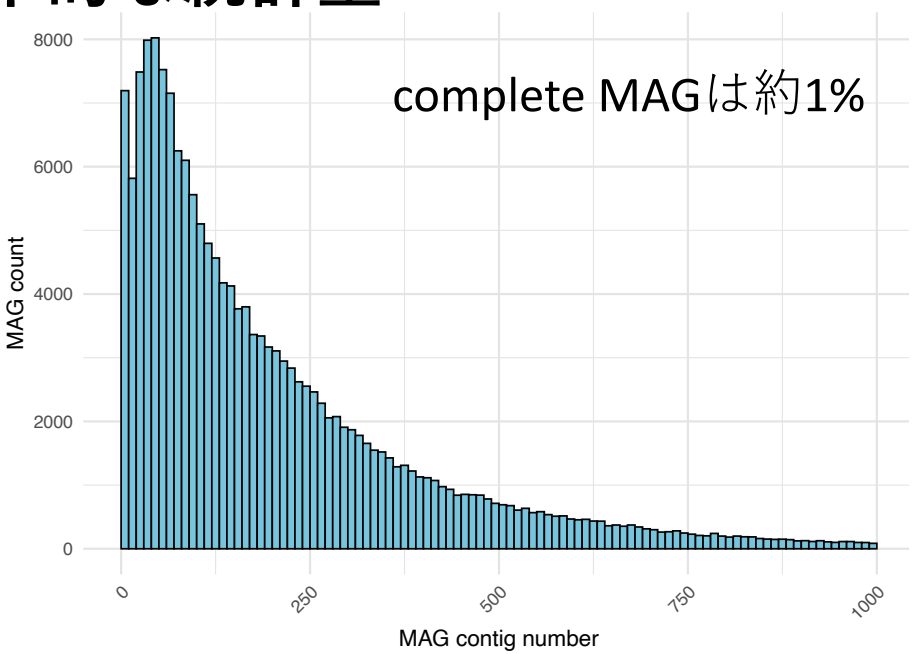
99%のMAGに最低一つはMEO classをアノテーション

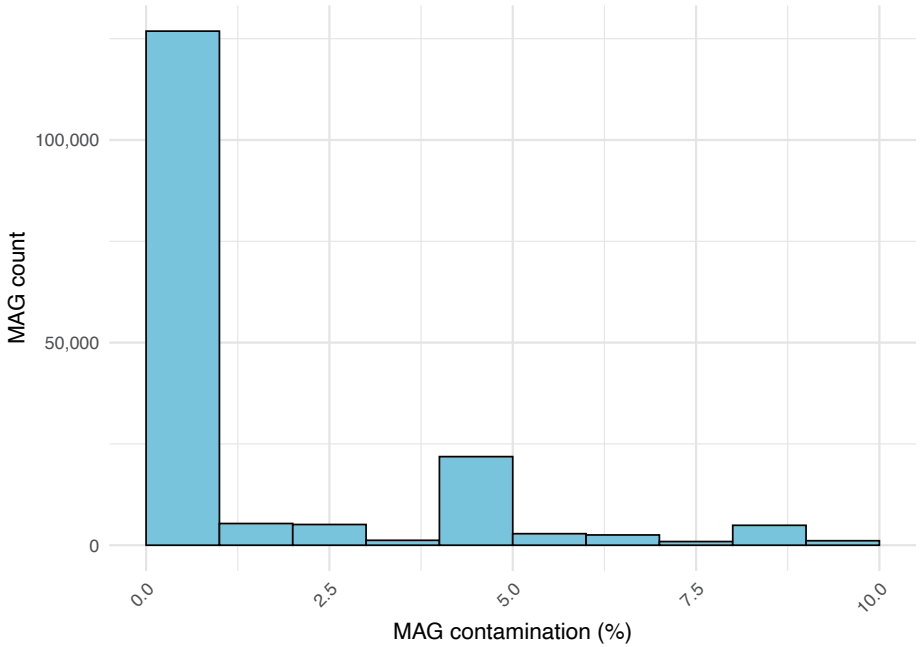
feces, soil, marine water等

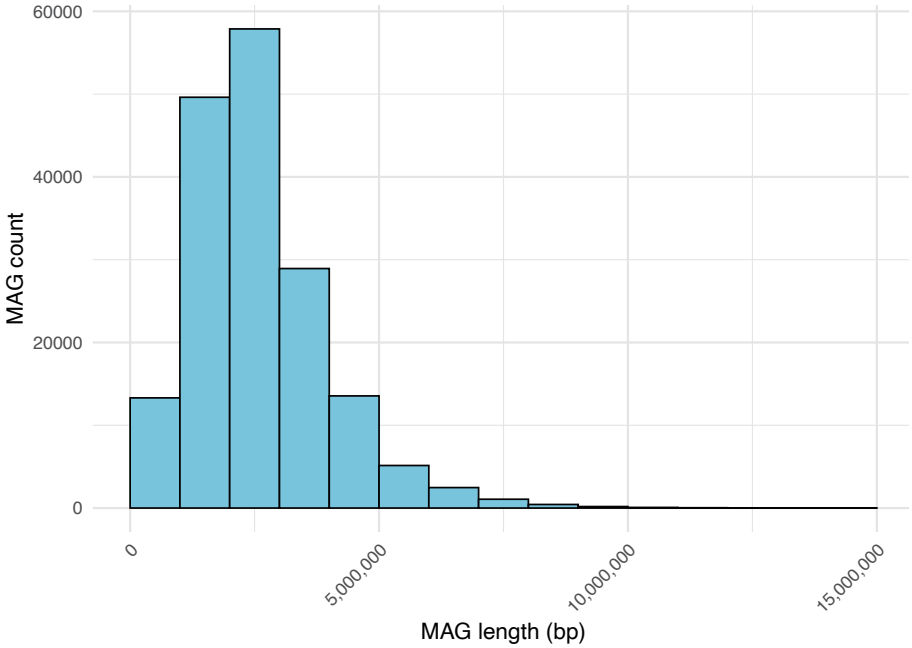
Microbiome Datahubの21万MAGの基本的な統計量

Mori H. et al. in revision







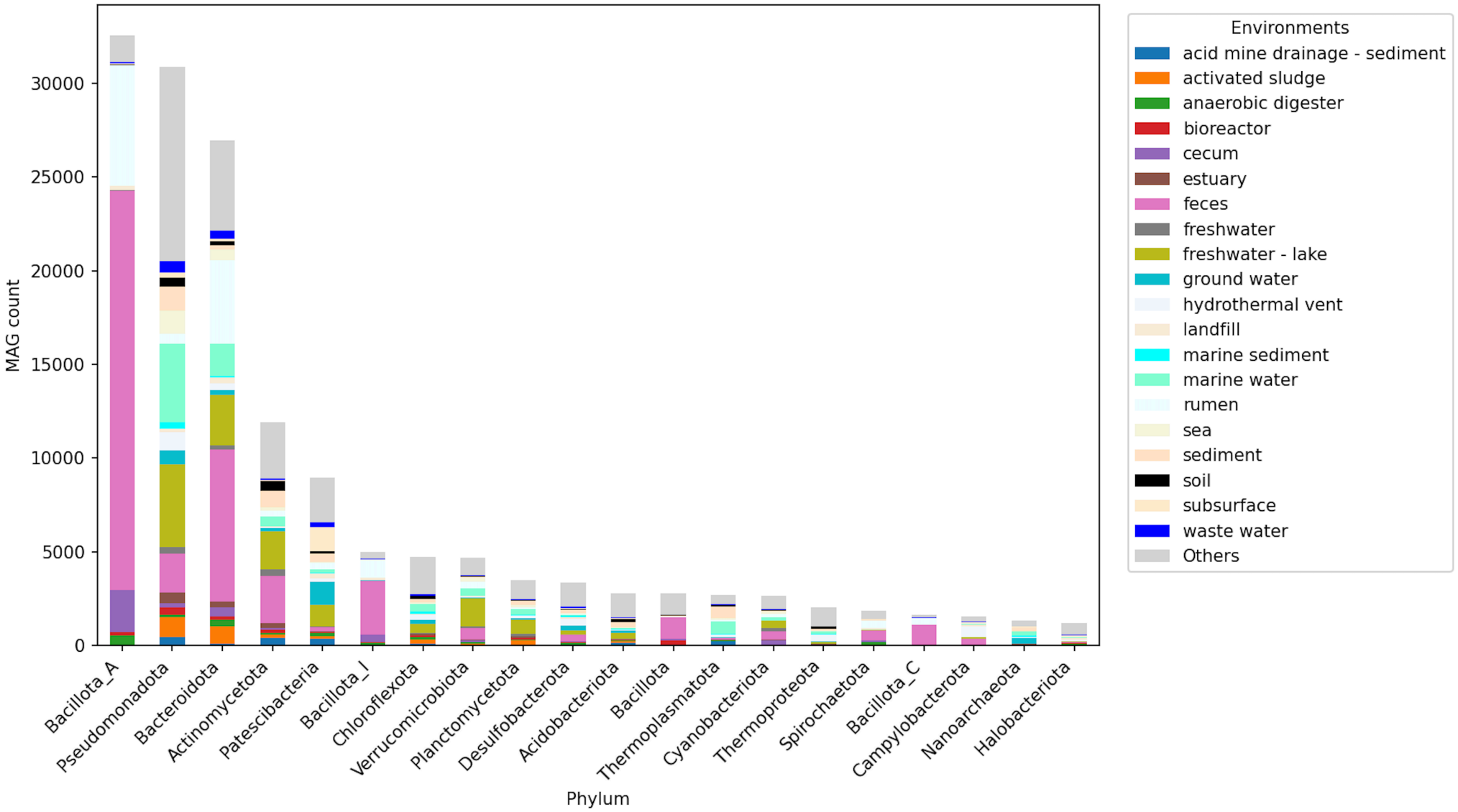


17万MAGはCompleteness > 60% & Contamination < 10%

種の重複を無くすと、77753 MAG

MAGの主要Phylumの環境分布

Mori H. et al. in revision



Microbiome Datahub and other MAG DBs

Mori H. et al. in revision

	Microbiome Datahub	MGnify	IMG/M	SPIRE
Developer	NIG	EMBL	JGI	EMBL
Number of MAGs	214,427	518,533	268,973	**1,158,553**
Number of environments	123	18	178	111
Number of phyla	**213**	182	197	182
Number of proteins	454,799,231	**2,455,939,992**	791,399,378	**2,452,036,556**
Total nucleotides (Gb)	533 Gb	1,291 Gb	784 Gb	**2,627 Gb**
Number of projects	1,759	5,189	8,929	739
Average completeness (%)	80.59	87.07	76.89	81
Average contamination (%)	1.83	1.03	2.10	1.77
Average MAGs per project	121.9	99.9	30.1	1567.7
Minimum completeness (%)	N.A.	50	50	N.A.
Minimum contamination (%)	N.A.	5	10	N.A.
Last data update date	May, 2023	**Nov, 2025**	**Dec, 2025**	Sep, 2023
Bulk download?	**Yes**	**Yes**	No	No

GLM構築用のデータセット

Microbiome DatahubのMAGデータについて、

以下の2種類の学習用データを構築
1. 高精度版
Completeness > 60%
Contamination < 10%
Contig number < 1500
Minimum contig length > 1,000 bp の条件を満たしたHigh quality
164,641 MAGs (509GB)

2. 簡易版
上記データのうち、1種1MAGにした、
77,753 MAGs (194GB)

全MAGに環境情報・系統情報・遺伝子配列・遺伝子機能組成情報が存在

単離菌ゲノムデータと組み合わせた場合とそうでない場合等テスト可能