メタゲノム由来ゲノムを収集・整理した統合データベース

「Microbiome Datahub」を開発

—21 万ゲノム以上の MAG 配列と環境・機能情報を統合し、微生物研究を加速—

■ 概要

情報・システム研究機構　国立遺伝学研究所の森宙史准教授、自然科学研究機構　基礎生物学研究所の内山郁夫准教授、東京科学大学　生命理工学院　山田拓司教授、京都大学　化学研究所　松井求助教を中心とする共同研究グループ（国立遺伝学研究所、基礎生物学研究所、東京科学大学、京都大学、東京大学）は、環境中の微生物を解析したメタゲノム由来のゲノム配列（MAG: Metagenome-Assembled Genomes）を公共の塩基配列リポジトリから網羅的に収集し、環境や系統・遺伝子機能等、様々な情報を付加した統合データベース「Microbiome Datahub」を開発・公開しました。本データベースは、公共の塩基配列リポジトリに蓄積された 21 万件以上の MAG 配列に対し、統一された遺伝子予測、系統分類、遺伝子機能アノテーション、表現型予測および環境メタデータを付与したものであり、データ駆動型の微生物学や未知の有用タンパク質探索の基盤として貢献することが期待されます。本研究成果は、科学誌「Microbiome」に 2026 年 3 月 16 日に速報版が掲載されました。

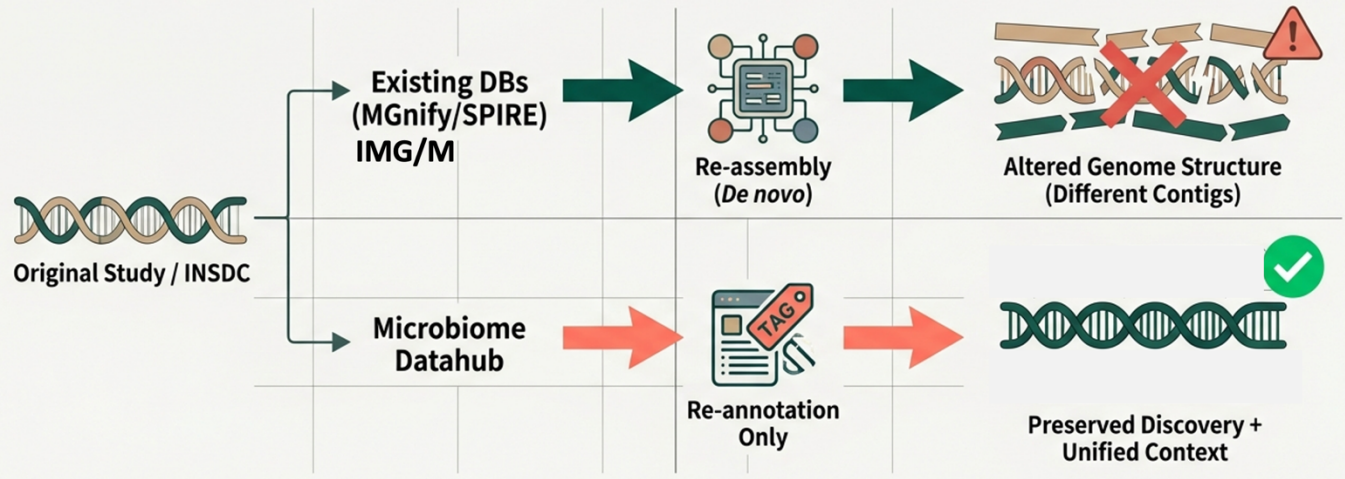
■ 成果掲載誌

- 雑誌名: Microbiome
- 論文タイトル: Microbiome Datahub: an open-access platform integrating environmental metadata, taxonomy, and functional annotation for comprehensive metagenome-assembled genome datasets
- 著者: Hiroshi Mori, Takatomo Fujisawa, Koichi Higashi, Yasuhiro Tanizawa, Zenichi Nakagawa, Hiroyo Nishide, Masaki Fujiyoshi, Yasukazu Nakamura, Ikuo Uchiyama, Motomu Matsui, Takuji Yamada
- **DOI:** https://doi.org/10.1186/s40168-026-02385-x
- **URL:** https://link.springer.com/article/10.1186/s40168-026-02385-x
- **Microbiome Datahub URL:** https://mdatahub.org/
- 全ゲノム配列、メタデータ、機能アノテーションデータは一括ダウンロード（Zenodo 等の web サイト）でも提供されています。

■ 研究の詳細

【研究の背景】環境中の微生物群集から丸ごと DNA 配列を解読するメタゲノム解析技術の発展により、培養困難な微生物のメタゲノム由来ゲノム情報（MAG）[1] が爆発的に増加しています。これらのデータは公共の塩基配列リポジトリ（INSD）[2] に登録されていますが、「品質のばらつき」「環境メタデータ（生息場所）の未整理」「遺伝子情報の欠如や分類体系の不統一」といった課題があり、そのままではデータの検索や横断的な比較解析が困難でした。また、既存の MAG の二次データベース（MGnify、IMG/M、SPIRE 等）は、登録されたリード配列から独自のパイプラインで「再アセンブリ（再構築）」を行っているため、原著論文で報告された MAG とは配列が変わってしまい、原著論文の研究成果を正しく参照・評価できないという問題（再現性の喪失）がありました（図 1）。

図 1. Microbiome Datahub と他の大規模 MAG データベースにおける MAG の違い

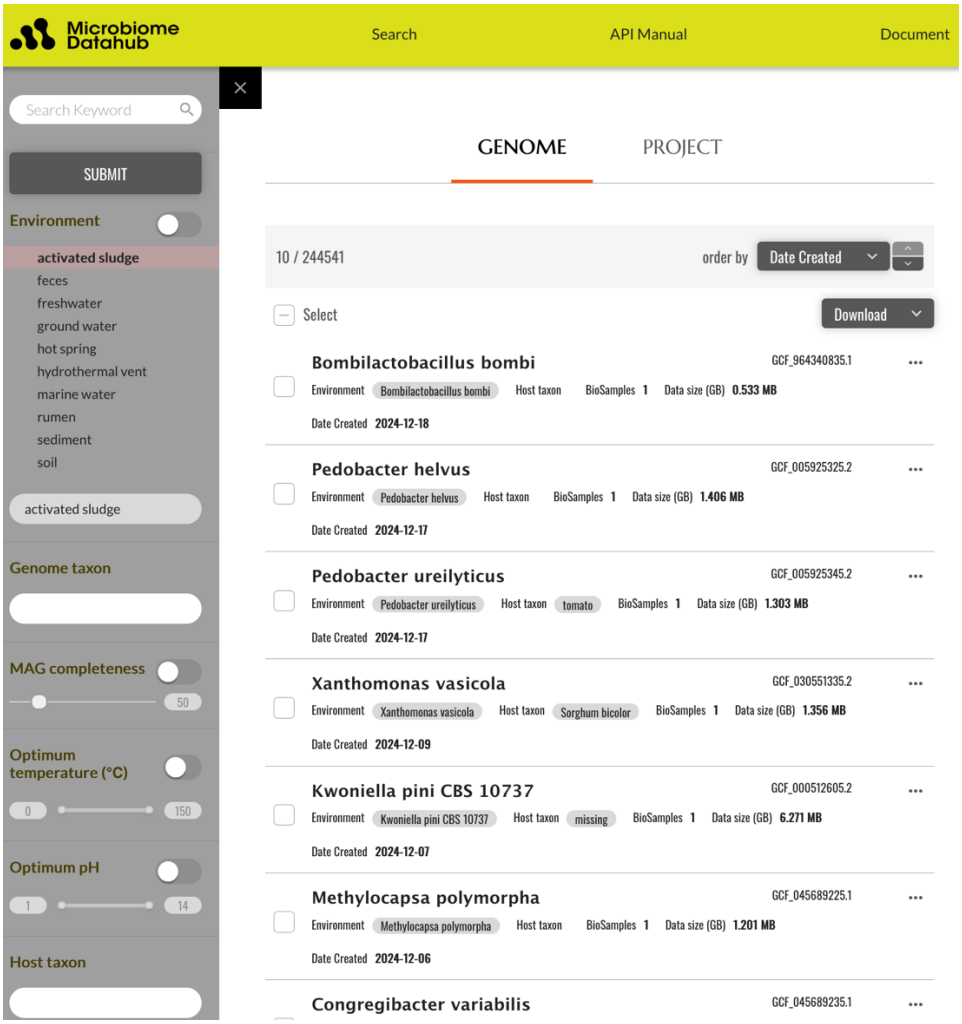


【研究の成果】本研究グループは、これらの課題を解決するため、公共リポジトリ中の MAG の配列はそのままに、メタデータとアノテーションのみを統一・高度化するアプローチを採用した MAG の統合データベース「Microbiome Datahub」を開発しました（図 2）。
主な特徴と成果は以下の通りです：

- **21 万件超の大規模かつ高品質な MAG データセットの構築**：INSD から 214,427 件の MAG を収集しました。そのうち約 17 万件が「Completeness（完全性）>60% かつ Contamination（汚染度）<10%」[3]という高い品質基準を満たしており、公共リポジトリ中の大多数の MAG は高品質であることが確認されました。
- **独自のオントロジーを用いた環境情報の統一**：本研究グループで開発し公開している、微生物の生息環境を体系的に記述するオントロジー[4]（MEO: Metagenome and Microbes Environmental Ontology）[5]を用い、表記揺れの激しい環境メタデータを人手によるキュレーションも交えて「糞便(feces)」「活性汚泥(activated sludge)」「土壌(soil)」等 123 種類の環境に統一・整理しました。
- **27 種類の表現型を予測**：系統名や 16S rRNA 遺伝子配列を用いた原核生物の表現型予測ツール「Bac2Feature」[6]を用いて、系統名から増殖速度、至適温度、至適 pH 等、全 MAG に対して 27 の表現型を予測し、異なる環境間での微生物の生存戦略の違い（例：宿主関連環境の微生物は増殖速度が平均的に速い等）を示唆しました。
- **多数の新規性の高いタンパク質配列の発見**：超高速配列類似性検索ツール「PZLAST」[7]等を用いた解析により、MAG に含まれるタンパク質配列の約 19%が既存のオーソログデータベース（MBGD）[8]に相同性が無い新規性の高いタンパク質配列であることが明らかになりました。

【今後の展望】Microbiome Datahub は、ウェブ上での高速検索や API アクセス、一括ダウンロードに対応しており、基礎的な微生物学研究から、タンパク質構造予測や有用酵素の探索等の応用研究まで、幅広く利用されることが期待されます。本プロジェクトは、今後も急増が予想される公共 MAG データを収載し整理・公開するデータベースとして、継続的な更新と拡張を予定しています。

図 2. Microbiome Datahub のユーザインターフェース。21 万件以上の MAG データについて、生息環境や系統名、MAG の品質等多様なメタデータを用いて絞り込み検索が可能です。



■ 用語解説

1. **MAG**　Metagenome Assembled Genome の略称であり、メタゲノム配列をアセンブルして得られた長い配列(コンティグ)から、配列の連続塩基組成や配列の相対存在量等の情報をもとにコンティグをクラスタリング(binning)して得られる、仮想的なゲノム配列を指します。
2. **INSD**　International Nucleotide Sequence Database の略称であり、様々な生物の塩基配列データを蓄積し公開している公共の塩基配列リポジトリを指す。現在は遺伝研の DDBJ、ヨーロッパにある EMBL-EBI の ENA、アメリカの NCBI の 3 機関で運営されています。
3. **Completeness と Contamination**　原核生物のゲノムや MAG において、配列の品質評価に使用される指標。原核生物のゲノム中に通常 1 コピーのみ存在する必須遺伝子セットを用いて、どの程度セット中の遺伝子が揃っているか（Completeness）、およびどの程度重複してそれらの遺伝子を所持しているか（Contamination）を計算して%で表します。Completeness は 100%に近いほど良く、Contamination は 0%に近いほど良い統計量です。これらの統計量は、CheckM や CheckM2 等のツールを用いてゲノムや MAG ごとに計算することが可能です。
4. **オントロジー**　ある分野の概念（ものや事柄）と、それらの関係性を体系的に定義した知識の構造。遺伝子の機能についてのオントロジー Gene Ontology 等が生物学分野では代表的なオントロジーです。

5. **MEO**　Metagenome and Microbes Environmental Ontology の略称であり、生物の環境を記述した Environmental Ontology の一部を利用した上で、微生物の生息環境に特化した語彙を追加し、クラス構造を大きく改変した、微生物特化の環境オントロジーです。遺伝研の森グループが中心になって開発し、生命科学分野のオントロジーの国際的なポータルサイト、BioPortal にて公開されています。MEO の URL https://bioportal.bioontology.org/ontologies/MEO
6. **Bac2Feature**　標準化された微生物の特性や形質に関するデータセットに様々な予測手法を適用し、16S rRNA 遺伝子配列や系統名から、細胞形態、グラム染色性、胞子形成能や運動性の有無、ゲノムの特徴、生育条件等等、その微生物の様々な表現型を予測するツールです。京都大学の松井グループが開発し公開しています。 Bac2Feature の URL https://www.genome.jp/bac2feature/
7. **PZLAST**　メタゲノム由来の大量のタンパク質のアミノ酸配列に対して、専用のスーパーコンピュータを用いて超高速にアミノ酸対アミノ酸の配列類似性検索を行うことが可能な検索ツールです。森グループが中心になって開発し公開しています。PZLAST の URL https://pzlast.nig.ac.jp/pzlast/meta
8. **MBGD**　主に微生物を対象として、遺伝子のオーソログ関係を、タンパク質をより詳細な機能単位であるドメインに分割した上でクラスタリングを行って推定し、オーソログクラスタとして整理したデータベースです。MBGD は基礎生物学研究所の内山グループが構築・運営しています。MBGD の URL https://mbgd.nibb.ac.jp/

■ 研究体制と支援

本研究開発は、科学技術振興機構（JST）ライフサイエンスデータベース統合事業(統合化推進プログラム)（課題番号：JPMJND2206）の支援を受けて実施されました。