2024年10月5日
トーゴーの日シンポジウム2024

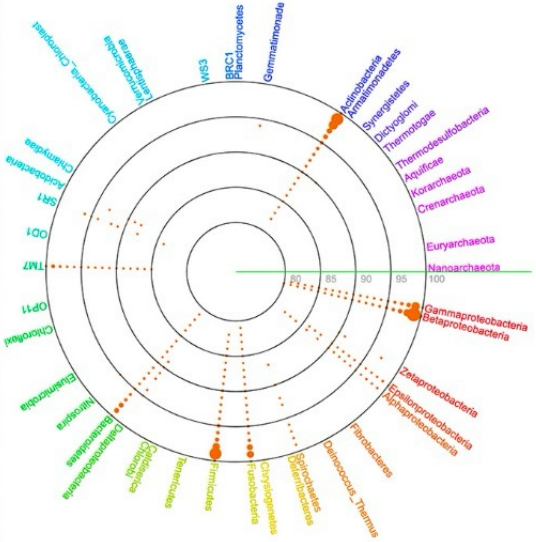
マイクロバイオームのDB開発における既存知識の抽出・整理のためのLLM活用

国立遺伝学研究所
情報研究系
森　宙史
Hiroshi Mori

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



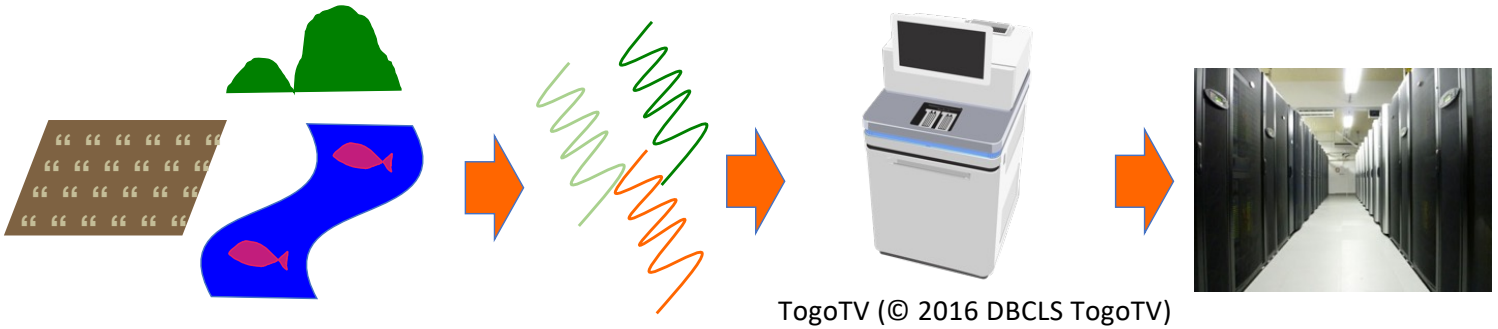
本研究室では、バイオインフォマティクス技術を用いて微生物などが持つゲノムの多様性を解明する研究に取り組んでいます。メタゲノム解析技術の進展によって、培養が難しい微生物も含めてゲノム解読が可能になり、また、メタゲノムデータの一種であるAncient DNAデータを用いることで、数万年以上前に絶滅した生物のゲノム解析も可能になりました。我々は森が兼任する遺伝研の先端ゲノミクス推進センターと強固に連携し、最先端のゲノム解析技術とバイオインフォマティクス解析技術を武器に未だ未知な部分が多い生物のゲノムの多様性に関する幅広い研究を進めております。

- 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

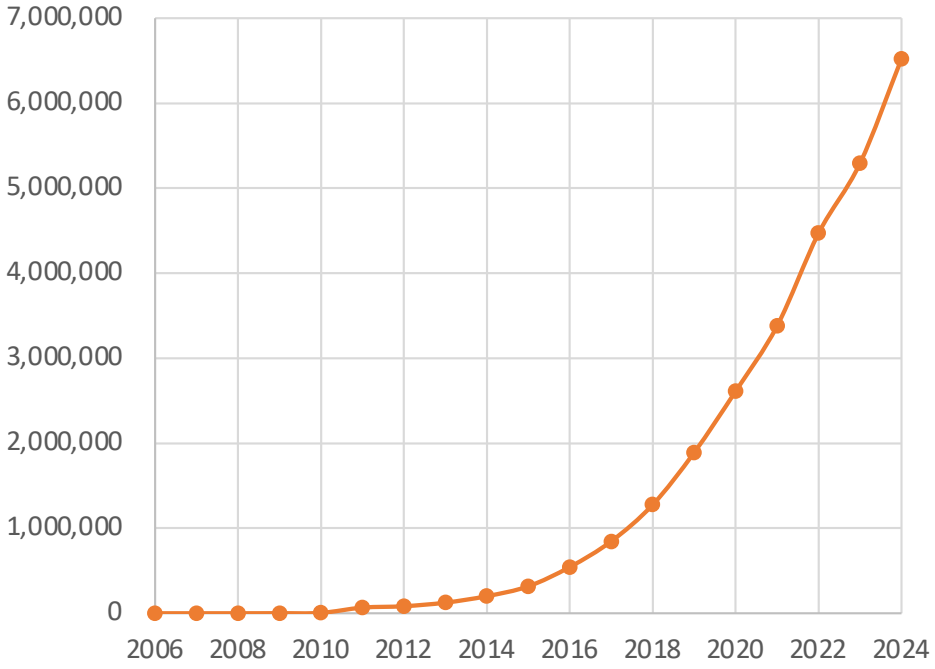
研究室webページ https://www.genome.id

Genome analysis against “Microbial community” to know member compositions and functions



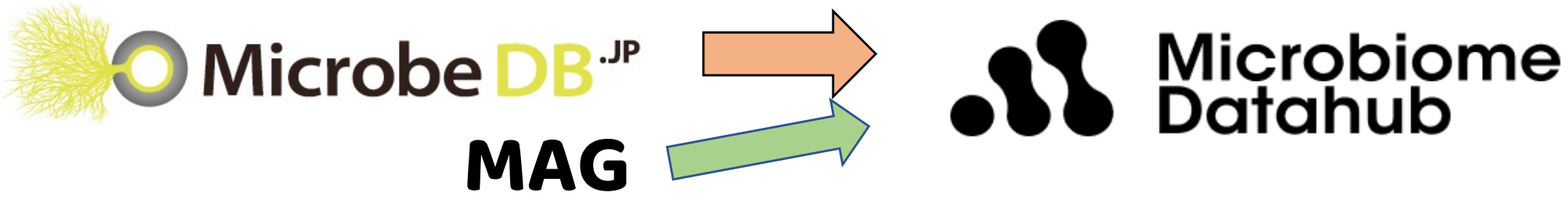
Microbiota sampling → DNA extraction from microbes → DNA sequencing → Bioinformatics analysis

公共のマイクロバイオームサンプルの総数と内訳
(2024年8月時点)



環境区分	プロジェクト数	サンプル数
自然環境	**73,006**	**2,357,828**
土壌	20,967	833,450
海水	8,077	357,054
宿主共生	**41,699**	**3,506,984**
ヒト	9,072	1,618,420
マウス	4,609	295,610

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



https://mdatahub.org

NIG

Hiroshi Mori, Takatomo Fujisawa, Koichi Higashi, Yasuhiro Tanizawa, Yasukazu Nakamura

NIBB

Ikuo Uchiyama, Hirokazu Chiba, Hiroyo Nishide

Institute of Science Tokyo

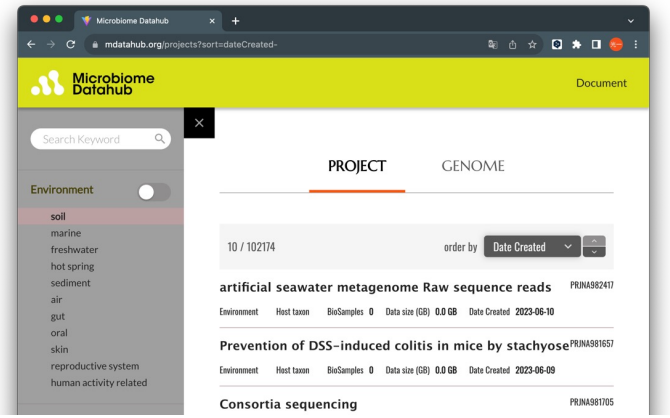
Takuji Yamada, Zenichi Nakagawa

Kyoto Univ.

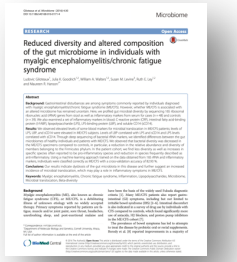
Motomu Matsui, Takao Suzuki, Yuki Nishimura

マイクロバイオーム解析に活かすためのサンプルの環境情報の整理

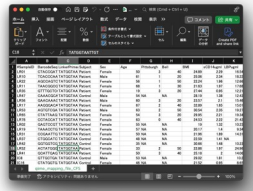
Microbiome Datahub



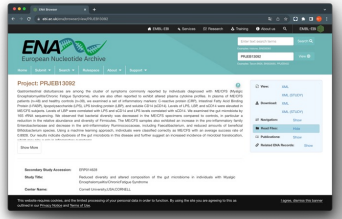
Paper (main text)



Supp tables

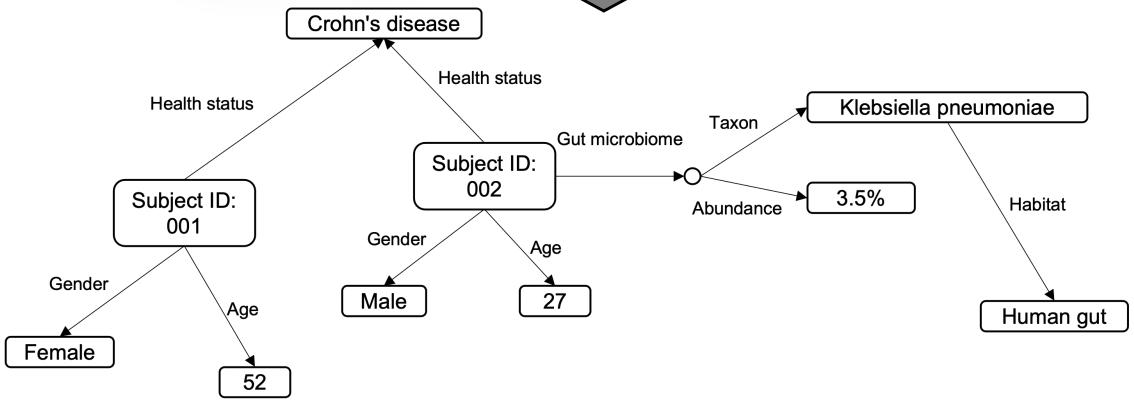


Public DB Registered Info (BioProject & BioSample)



マイクロバイオーム サンプル・ゲノムごとに整理

環境情報



論文からの環境情報の手動抽出は非常に手間がかかる　->　LLMを用いて効率化

EMBERS Project

Encompassing Microbiome-Bibliome Extraction and Retrieval System

ChatGPT APIを利用

Higashi K. et al. in preparation

メタデータの論文からの収集
EMBERS-MINE

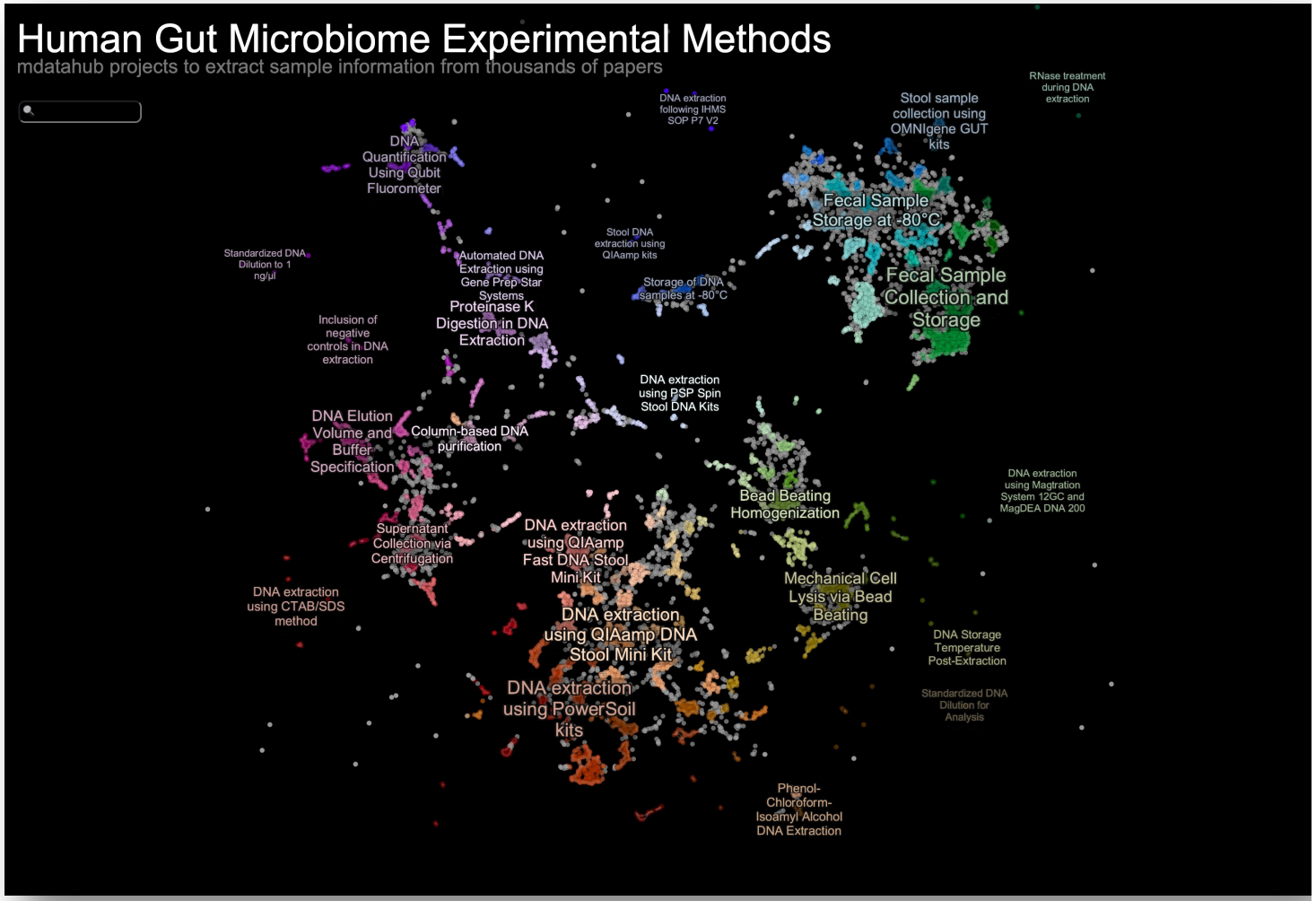
- 26,435 papers from PubMed Central
- Filter papers by analyzing the article text.
- Summarize the project and methods
  Convert them into embedding
- Extract accession number
- Parse the article and supplement tables
  Collect info related to subject metadata.
- Generate descriptions for the newly obtained metadata “keys” from the article.

メタデータの名寄せ
EMBERS-FUSE

- Metadata keys with descriptions
- EMBEDDING
  - Convert “key-description” texts into embedding vectors
- CLUSTERING
  - Compute hierarchical clustering of embedding vectors.
  - Determine a cut-off for each metadata key based on “semantic homogeneity.”
- ALIGNMENT
  - Convert value descriptions to a common format.
- Integrated metadata
  (Standardized keys and values)

EMBERS-MINE

- 26,435 papers from PubMed Central
- Filter papers by analyzing the article text.
- Summarize the project and methods
  Convert them into embedding
- Extract accession number
- Parse the article and supplement tables
  Collect info related to subject metadata.
- Generate descriptions for the newly obtained metadata “keys” from the article.



“Experimental methods” embeddings and their clusters

・科学には論文出版の文化があり大量に論文が蓄積

・大量の論文からの知識抽出とDBへの反映

・論文の要約・知識抽出はLLM技術を活用すれば可能

生成AI時代に、何のため誰のためにDBを作るのか