

August 7, 2019
統合データベース講習会: AJACS番町3

【How to use1】 MicrobeDB.jpの使い方

森宙史

Hiroshi Mori

国立遺伝学研究所
情報研究系

<http://microbedb.jp/>



Text

Analysis

Statistics

Search

Environment: hot spring

Taxonomy: Enterococcus faecalis

Taxonomy: Streptomyces avermitilis

Gene: psbA

ID: 29

2011年から公開している
原核生物を主とした微生物の統合データベース (DB)

MicrobeDB.jp v.3 project members

National Institute of Genetics: (Genome, Metagenome, Ontology)

Ken Kurokawa, Yasukazu Nakamura, Hiroshi Mori,
Takatomo Fujisawa, Eli Kaminuma (TMDU), Koichi Higashi

National Institute of Basic Biology: (Ortholog)

Ikuo Uchiyama, Hirokazu Chiba (DBCLS), Hiroyo Nishide

Tokyo Institute of Technology: (Metagenome)

Takuji Yamada, Zenichi Nakagawa

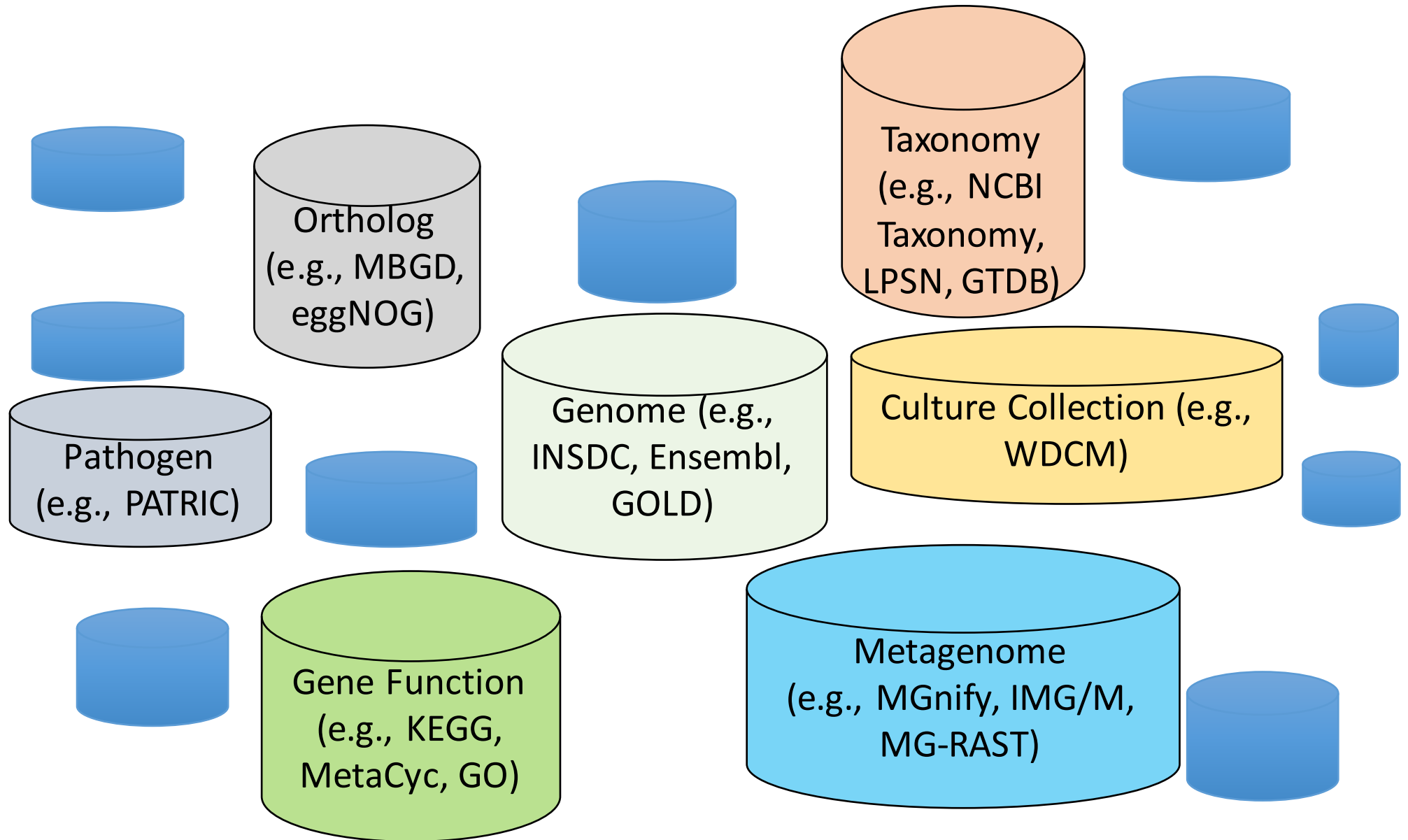
Chiba University: (Fungal & Bacterial culture collection info.)

Hiroki Takahashi, Takashi Yaguchi

Technical adviser:

DBCLS (especially Shuichi Kawashima, Toshiaki Katayama)

微生物のDBは既に良い特化型のDBが多数存在



特化型のDBでは、異なる種類のデータ(ゲノムとメタゲノム等)を
関連付けて検索することは困難 → **統合DB**の出番

微生物統合DB MicrobeDB.jpの基本的な開発方針

微生物統合DB ≠ 微生物の百科事典

統合DBは、何らかの軸で多様なデータを**繋いで**整理したDB

異なる種類のデータ間を関連付けることを重視している

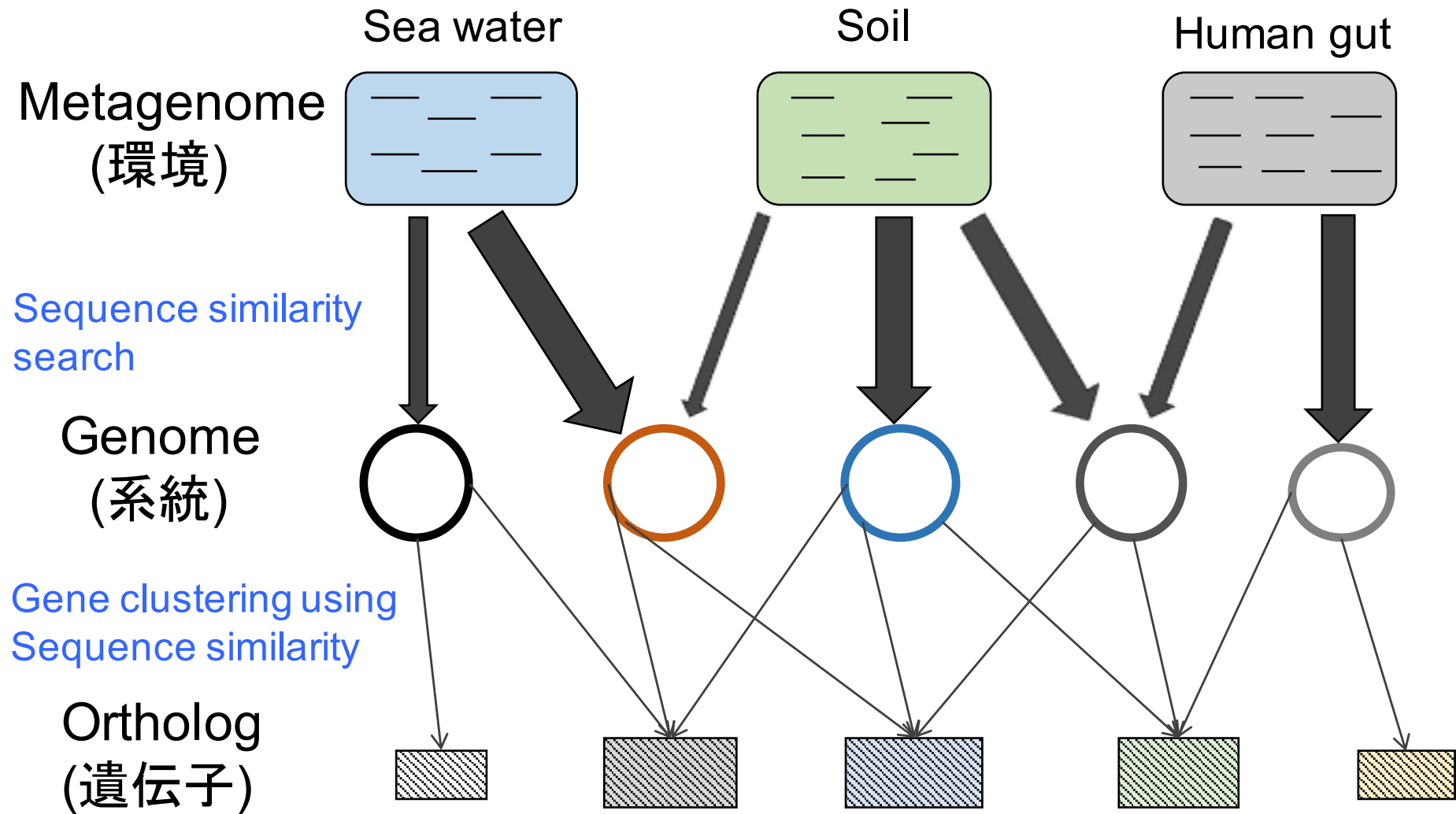
MicrobeDB.jpは、主に原核生物の遺伝子・系統・環境に関する様々なデータを、ゲノム配列を軸に統合化したDB

微生物統合DB MicrobeDB.jpの基本的な開発方針

- 遺伝子と系統のリンク
 - 系統が持つ遺伝子機能的な特徴を推定
- 遺伝子と表現型のリンク
 - 表現型を担う遺伝子を推定
- 遺伝子と環境のリンク
 - その環境で生息するために必要な遺伝子機能を推定
- 系統と環境のリンク
 - その環境に適応した系統を推定

既存のデータにこれらのリンクを付与し、リンクから微生物に関する新たな知見を得られるDBを目指す₆

配列類似性 (進化的類縁性) を基にしたデータの統合化



概念 (Ontology) を基にしたデータの統合化

Ontology is a structured controlled vocabulary to describe properties and types of resources.

例: 森とは何か? 林と何が違うのか?

データの由来が異なれば、同じ語彙でも意味が異なる場合がある
それをOntologyを用いて統一する

MEO (Microbes/Metagenomes Environmental Ontology)

MSV (Metagenome Sample Vocabulary)

MCCV (Microbial Culture Collection Vocabulary)

MPO (Microbial Phenotype Ontology)

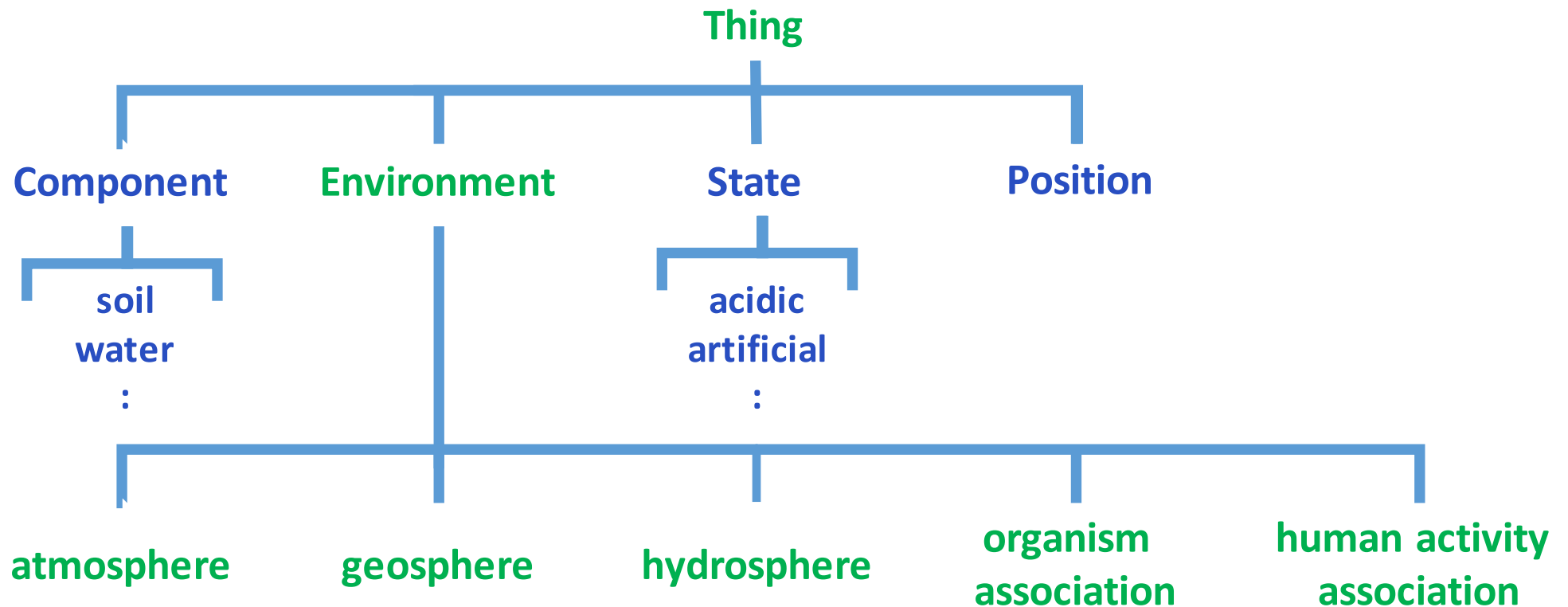
MBGD Ontology

PDO (Pathogenic Disease Ontology)

Most of them can be obtained from



Microbes/Metagenomes Environmental Ontology (MEO) ver. 0.9



Gather microbial habitat related terms from INSDC DRA/BioSample, Japanese Culture Collections (JCM & NBRC).

Class:

2,381

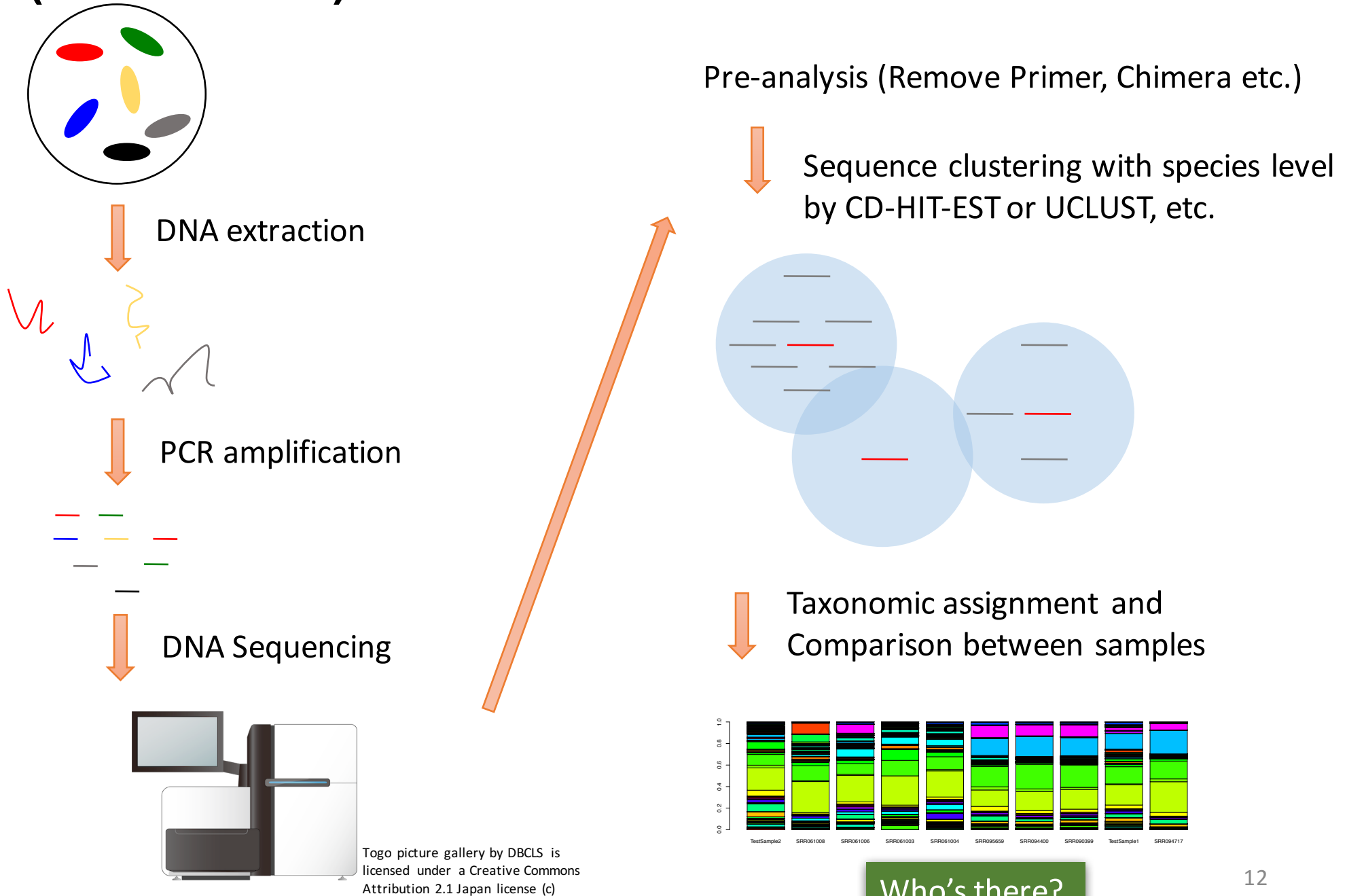
MicrobeDB.jp version 2 data

Data category	Data sources	Ontologies
Genome	RefSeq Prokaryotes	SO, FALDO, NCBITAX, INSDCO
Genome Metadata	INSDC BioSample	MPO, MEO, MSV, PDO, CSSO
Ortholog	MBGD	ORTHO
Culture collection	JCM, NBRC	MCCV, MPO
Metagenome	INSDC DRA	MEO, MSV

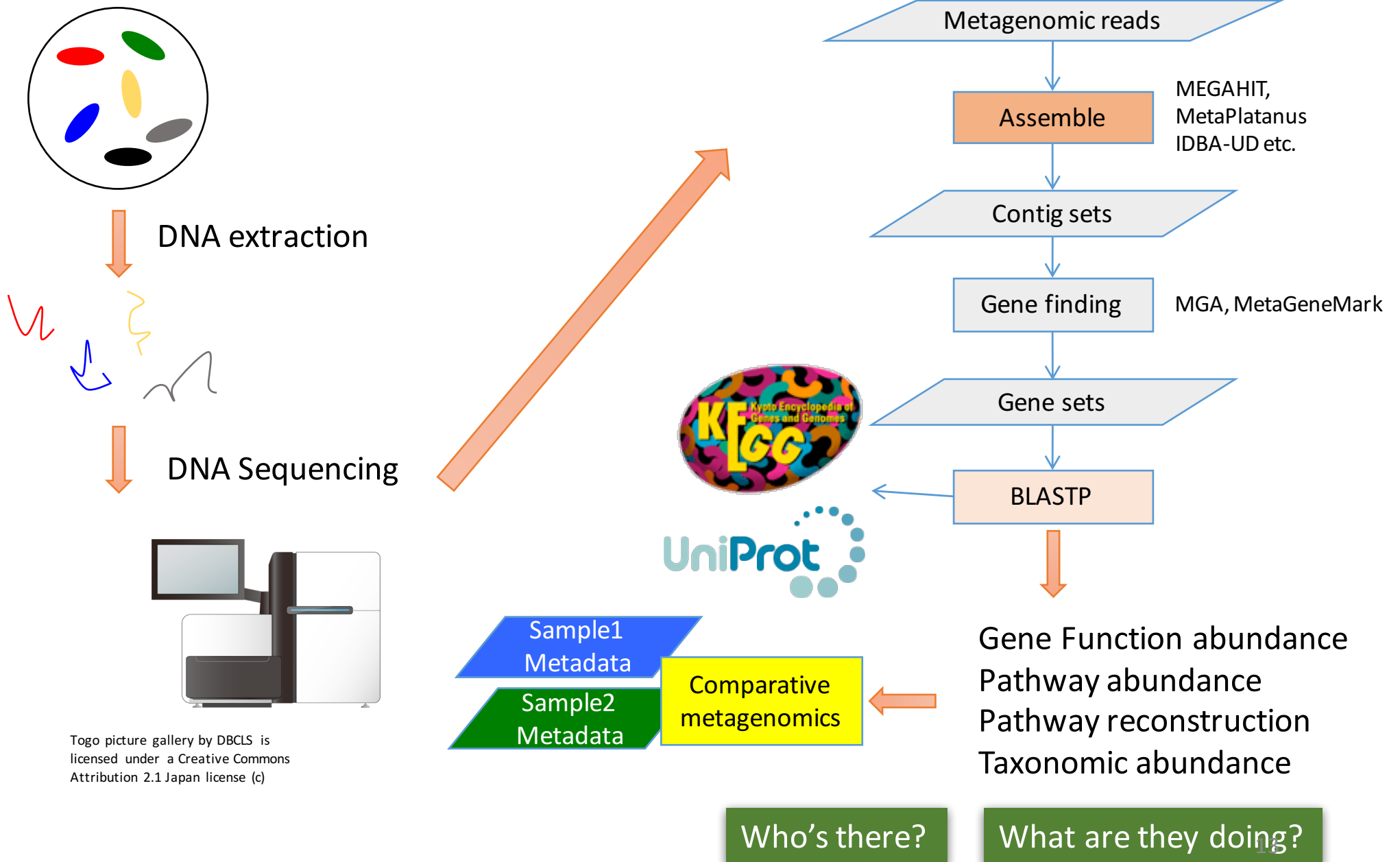
MicrobeDB.jp version 2 data

Data category	Number of entry
Genome data	16,983 taxa
Culture collection strain data	16,671 strains
Microbiome metadata	173,359 samples
Microbiome taxonomic composition data	60,551 samples
Microbiome functional composition data	4,048 samples

16S rRNA gene **amplicon sequencing** analysis (メタ16S解析)



Metagenomic sequencing analysis (メタゲノム解析)



Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

メタ16S解析

利点

- ・安価かつ少量のDNAから系統組成が得られる
- ・reference配列に依存しない解析も可能
- ・マシンパワーは少なくて済み、解析ツールも普及 (QIIME・mothur等)

欠点

- ・PCRバイアスの存在
- ・種以下は分解能に問題あり
- ・個々の系統の機能が不明

メタゲノム解析

利点

- ・系統組成と遺伝子機能組成が得られる
- ・実験によるバイアスが比較的少ない
- ・優占系統のドラフトゲノムの構築 (条件が良ければ可能)

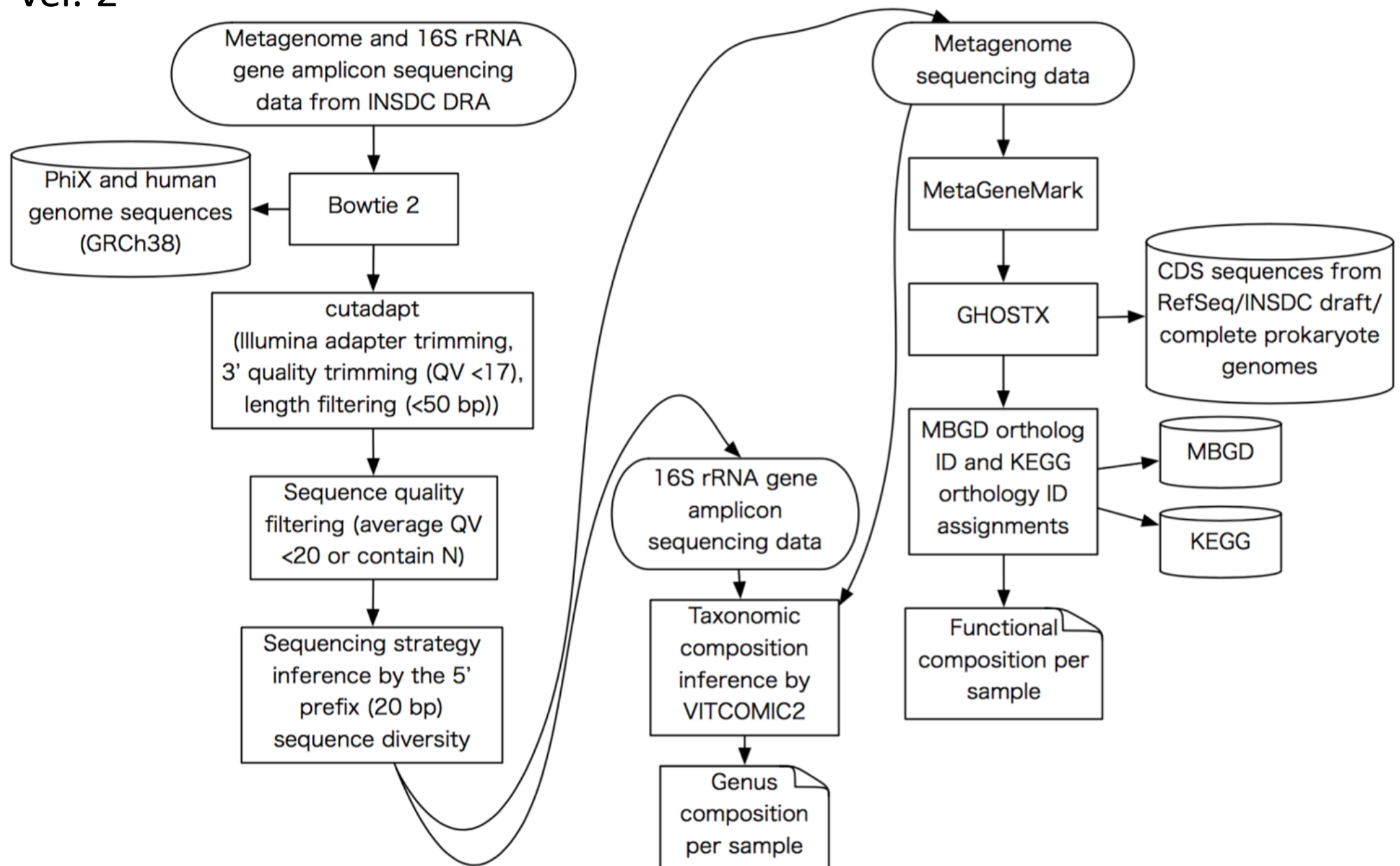
欠点

- ・reference配列に依存した解析
- ・目的依存で解析手法が変化し、マシンパワーも必要

Microbiome data in DRA/ERA/SRA

- 2014 (MicrobeDB.jp ver.2)
 - Microbiomes: 173,359 (samples)
amplicon:metagenome = 7:1
- 2018
 - Microbiomes: 1,117,378
 - ecological microbiomes: 433,491
 - air: 4,109
 - marine: 67,393
 - soil: 146,784
 - host-associated microbiomes: 618,575
 - human: 318,000
 - mouse: 55,600

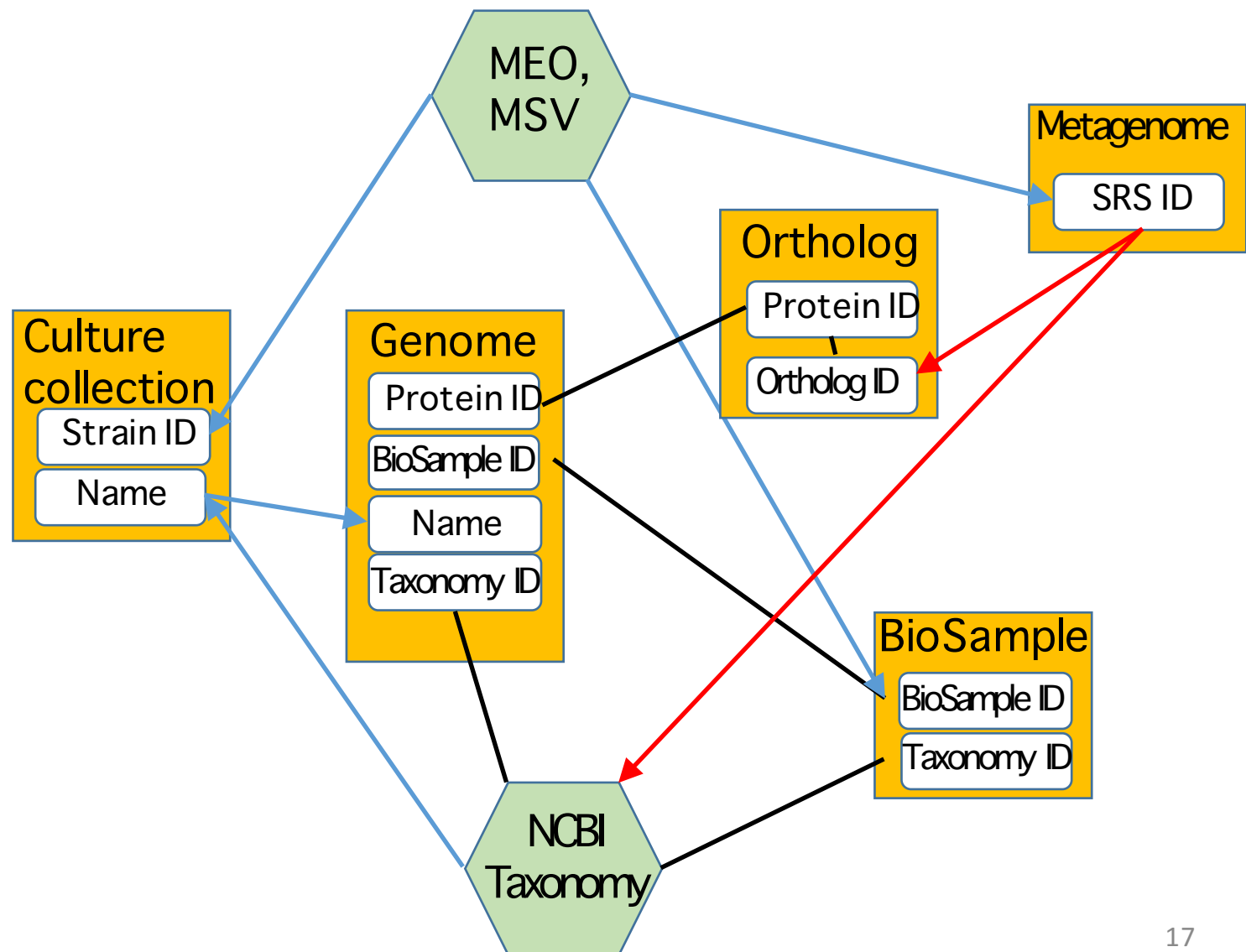
MicrobeDB.jp's 16S rRNA gene amplicon/Metagenome analysis workflow ver. 2



ID Mapping

Ontology Manual/Semiautomatic Annotation

Sequence-based analyses

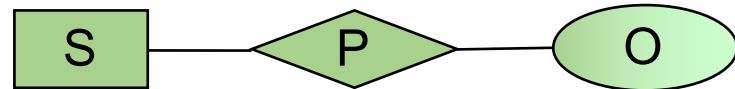


Semantic Web技術を用いた、データの持つ意味の共通性を基にした統合

RDF (Resource Description Framework)

Data model which uses Triples

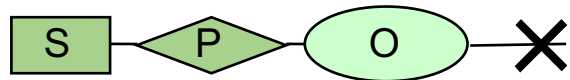
(**S**ubject – **P**redicate – **O**bject)



<URI> <URI> <URI>/Literal

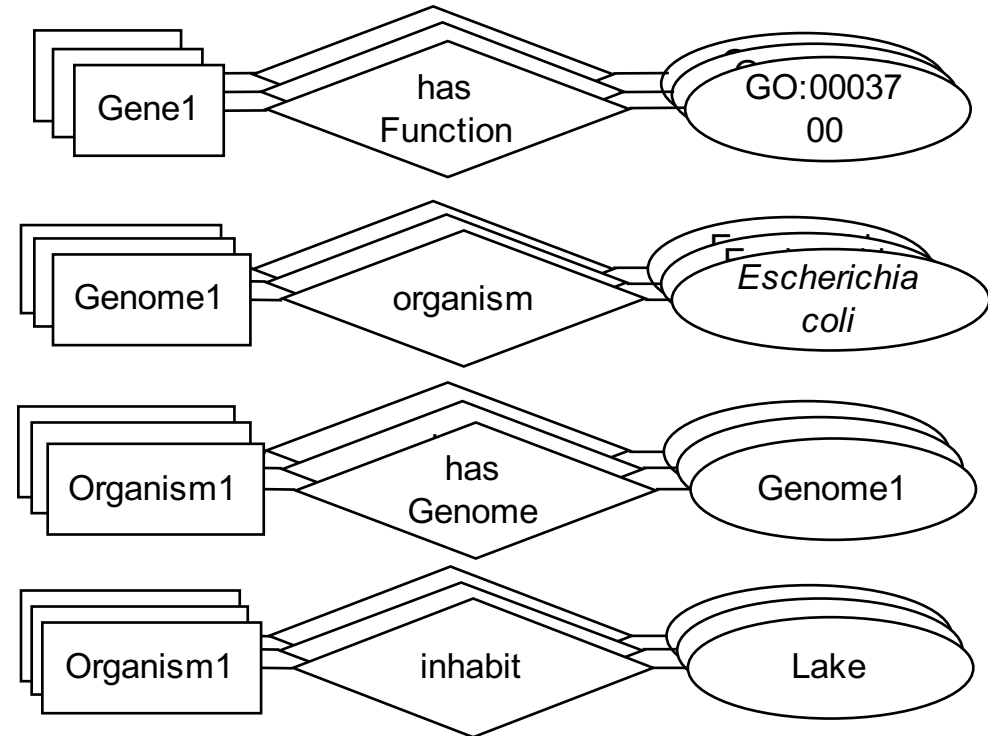
gtps:Gene1 rdfs:label “16S rRNA gene”

URI node can be linked to other nodes

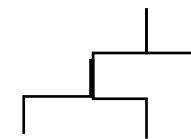


- ・ どの分野のデータもファイル形式が同じ
- ・ IDがURIなので、Web上で一意
- ・ 後から追加も容易
- ・ データが持つ意味やデータ間の関係性を記述可能

RDF

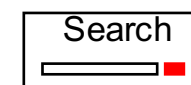


Ontology



Triple store

SPARQL



https://github.com/AJACS-training/AJACS77/tree/master/02_mori




MicrobeDB.jp

Integrating and representing genome, metagenome, taxonomy resources and the analysis datasets with Semantic Web Technologies.

Database statistics

Total number of Metagenomic samples (SRA/SRS):	173,359 samples
- with taxonomic analysis results:	60,551 samples
- with functional analysis results:	4,048 samples
Total number of Assembled Genomes (RefSeq/Genbank):	16,983 taxa
Total number of Strains (JCM/NBRC):	16,671 strains
Total number of Environmental terms in ontology (MEO):	2,381 terms

Show graph



[Home](#)
[Document](#)
[Analysis ▾](#)

[Sign Up](#)
[Sign in](#)

Public/Private

☐ metagenome_public 3729

hasMetagenomeAnalysis

☒ taxonomy 3729
 ☐ function 609

hasMEO (Text)

x

hasMEO: Component

Component for environment 389

hasMEO: Env

Environment for microbes 3729

hasMEO: Position

Position toward environment 5

hasMEO: State

Metagenomic samples

3729 results found in 112ms

hasMetagenomeAnalysis: taxonomy x

hasMEO (Text): gut x

Clear all filters

Previous

1

2

3

4

...

Next

10 ▾

Select All

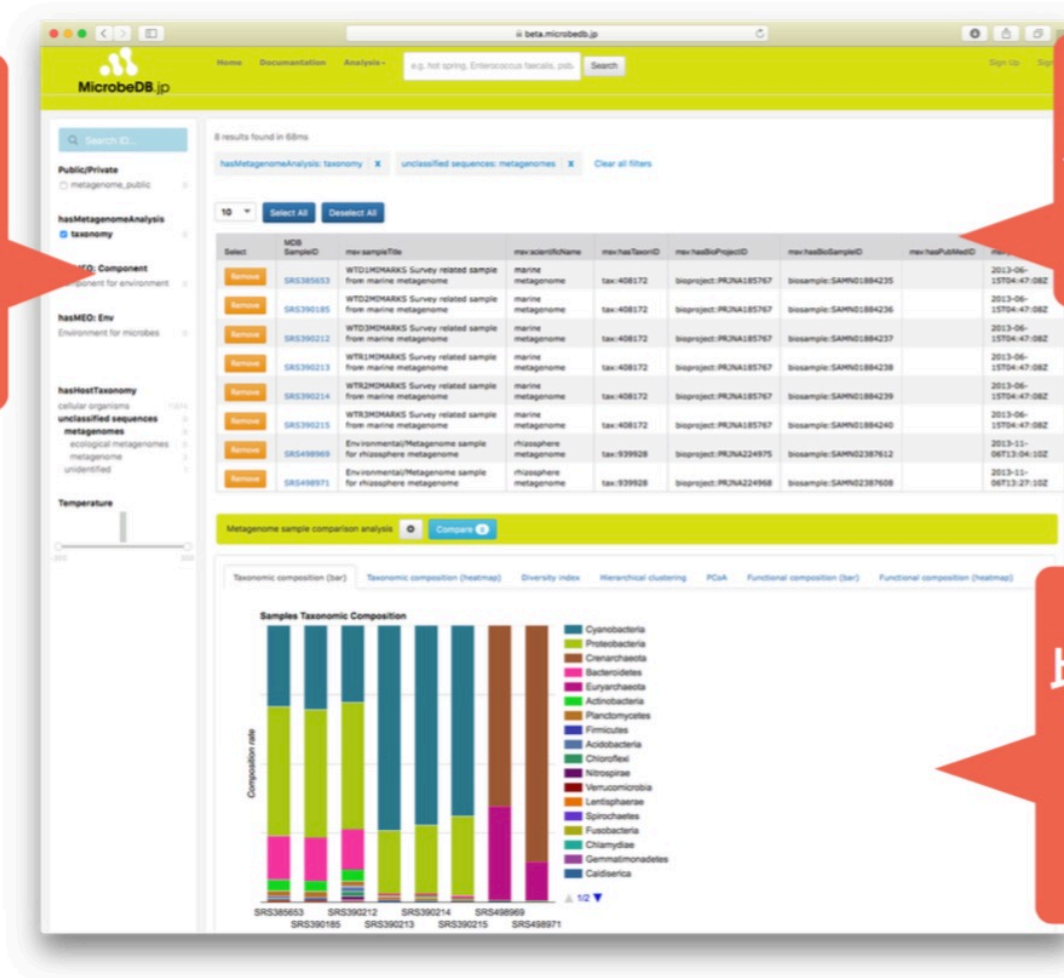
Deselect All

Select	MDB SampleID	msv:sampleTitle	msv:scientificName	msv:hasTaxonID	msv:hasBioProjectID	msv:hasBioSar
<input type="button" value="Remove"/>	SRS551059	Content of the intestinum from animals fed one meal of heparinized sheep blood	gut metagenome	749906	PRJNA237098	SAMN026145
<input type="button" value="Remove"/>	SRS551061	Content of the intestinum of animals fed two meals of heparinized sheep blood four weeks apart	gut metagenome	749906	PRJNA237098	SAMN026145
<input type="button" value="Remove"/>	SRS452599	Environmental/Metagenome sample for mouse gut metagenome	mouse gut metagenome	410661	PRJNA209582	SAMN022138
<input type="button" value="Add"/>	SRS367344	Pooled 16S rRNA gene sequences	Bacteria	2	PRJNA209582	SAMN017587
<input type="button" value="Add"/>	SRS369251	Pooled 16S rRNA gene sequences		2	PRJNA209582	SAMN017656

MicrobeDB.jp ポータル

例) メタゲノムサンプルのファセット検索から比較解析

環境情報、pH、温度、Host情報によるファセット検索



メタゲノム解析済み公開メタもしくは自分の解析データを選択

比較結果をStanzaという可視化コンポーネントにより各種提示

VITCOMIC2 is a visualization tool for the phylogenetic composition of microbial communities based on 16S rRNA gene amplicons and metagenomic shotgun sequencing.

Try VITCOMIC2

Metagenome/16S rRNA gene Amplicon Sequencing FASTA/FASTQ file: ファイルが選択されていません。

File format: ☒ FASTA flat ☐ FASTQ flat ☐ FASTA gzipped ☐ FASTQ gzipped

Conduct 16S rRNA gene Copy number normalization?: ☒ No ☐ Yes

Conduct 16S rRNA gene Assembly? (Shotgun metagenome only): ☒ No ☐ Yes

ID: (use [A-Za-z0-9-_])

Email:

How to use

1. Input data

Both of a FASTA/FASTQ file and gzipped FASTA/FASTQ file are acceptable for the input data in the VITCOMIC2. Sample 16S rRNA gene Amplicon sequencing fastq data.

2. File format

File format is a file format identifier of your FASTA/FASTQ file. To reduce the size of your file, we strongly recommend that you compress your file with gzip. If you don't compress your file, please choose "flat file".

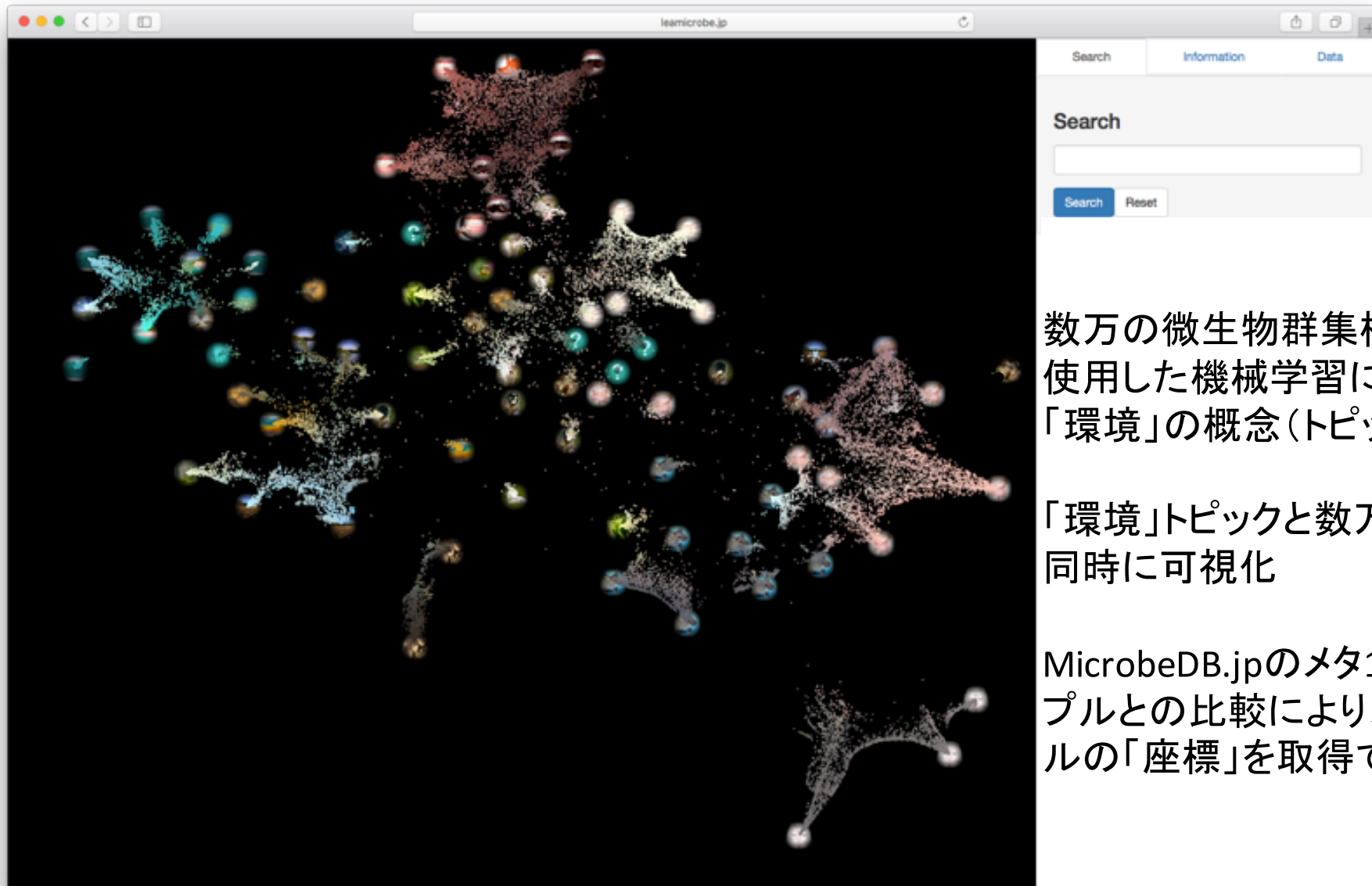
(Mori H et al. 2018, BMC Syst Biol)

Microbiome sequencing data → Genus composition

LEA

<http://leamicrobe.jp/>

Visualize microbiome composition data



数万の微生物群集構造データを使用した機械学習によって、「環境」の概念(トピック)を抽出

「環境」トピックと数万サンプルを同時に可視化

MicrobeDB.jpのメタ16S数万サンプルとの比較により、新規サンプルの「座標」を取得できる