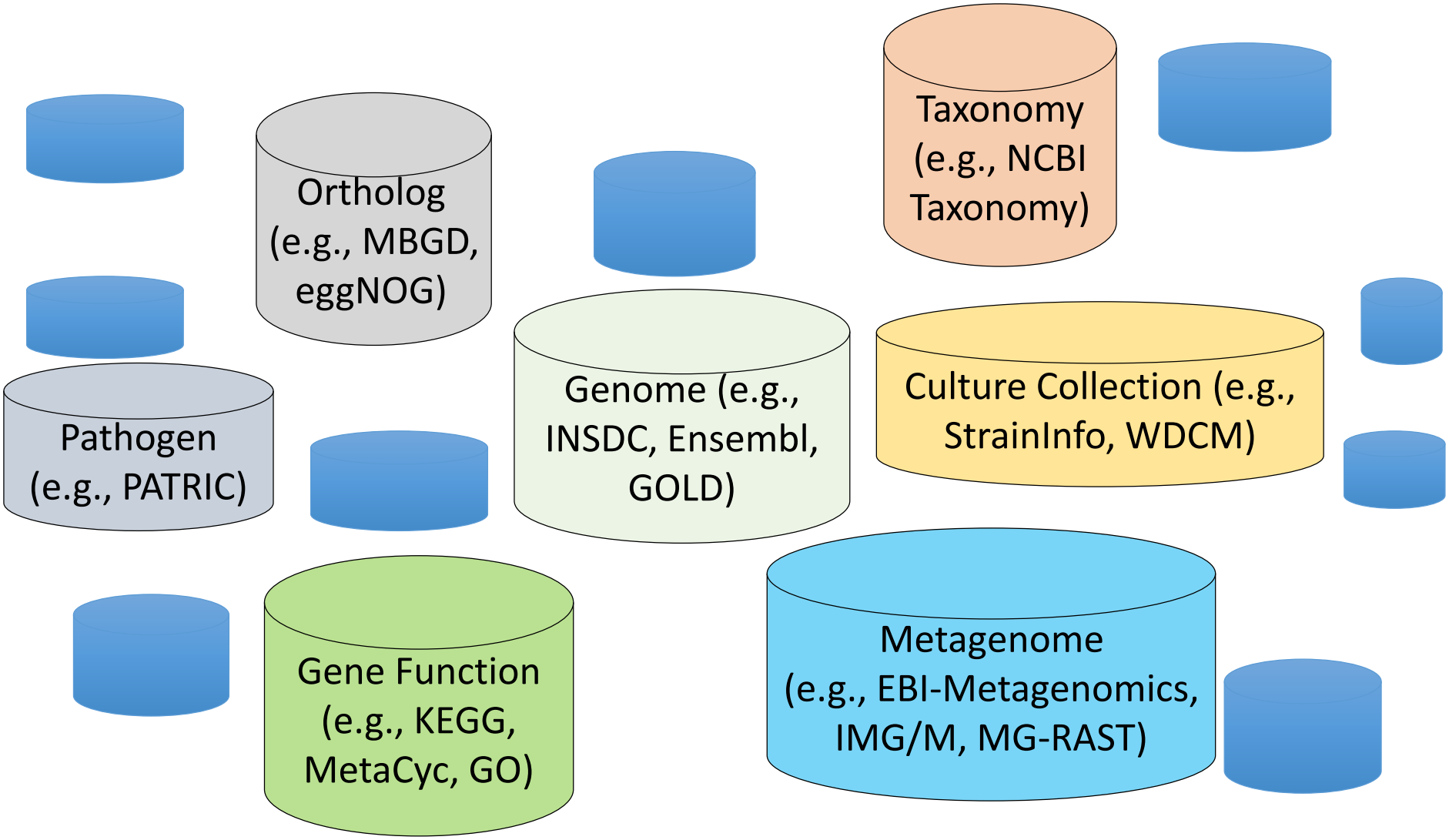
Nov. 28, 2018
[1PW2-09] 生命科学のデータベース活用法2018

微生物統合データベース MicrobeDB.jpの活用法

森宙史, 黒川顕
国立遺伝学研究所
生命情報研究センター

Many microbial databases (DBs) exist …



Which DBs should we use (or recommend for beginners)?

Microbe DB .JP integrates lots of data related to microbes.

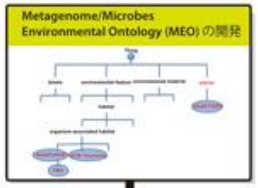
Especially, we integrates the microbial data that can be linked to **genomes**. since 2011

Microbe DB .JP

http://microbedb.jp/

Microbe DB.jp
MicrobeDB.jp プロジェクトでは様々な微生物学上の知識を、ゲノム情報を核として遺伝子、系統、環境の３つの軸に沿ってセマンティックウェブの技術を駆使して整理統合し、幅広い分野での微生物学の発展に資することの出来るデータベースの構築を目標としています。

Ontology



Gene | Taxon | Environment

Ortholog: MBGD

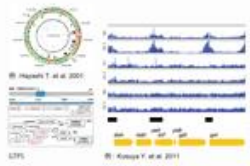
オーソログデータ

Taxonomy:
NCBI Taxonomy

系統分類データ

Metadata:
INSDC DRA

環境のメタデータ



Genome: RefSeq

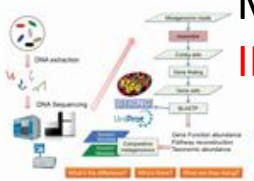
オミックスデータ

Annotation:
TogoAnnotation

モデル微生物の高品質アノテーションデータ

Culture Collection:
NBRC/JCM

菌株データ



Metagenome:
INSDC DRA

メタゲノムデータ

Red color indicates our collaborators.

MicrobeDB.jp v.3 project members

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

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

National Institute of Basic Biology: (Ortholog)

Ikuo Uchiyama, Hirokazu Chiba (DBCLS), Hiroyo Nishide

Tokyo Institute of Technology: (Metagenome)

Takuji Yamada

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

Hiroki Takahashi, Takashi Yaguchi

Technical adviser:

DBCLS (especially Shuichi Kawashima, Toshiaki Katayama)

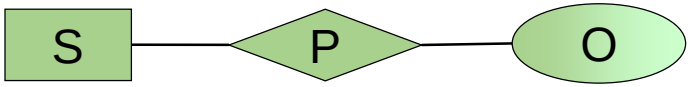
Funding

JST

NBDC
National Bioscience Database Center

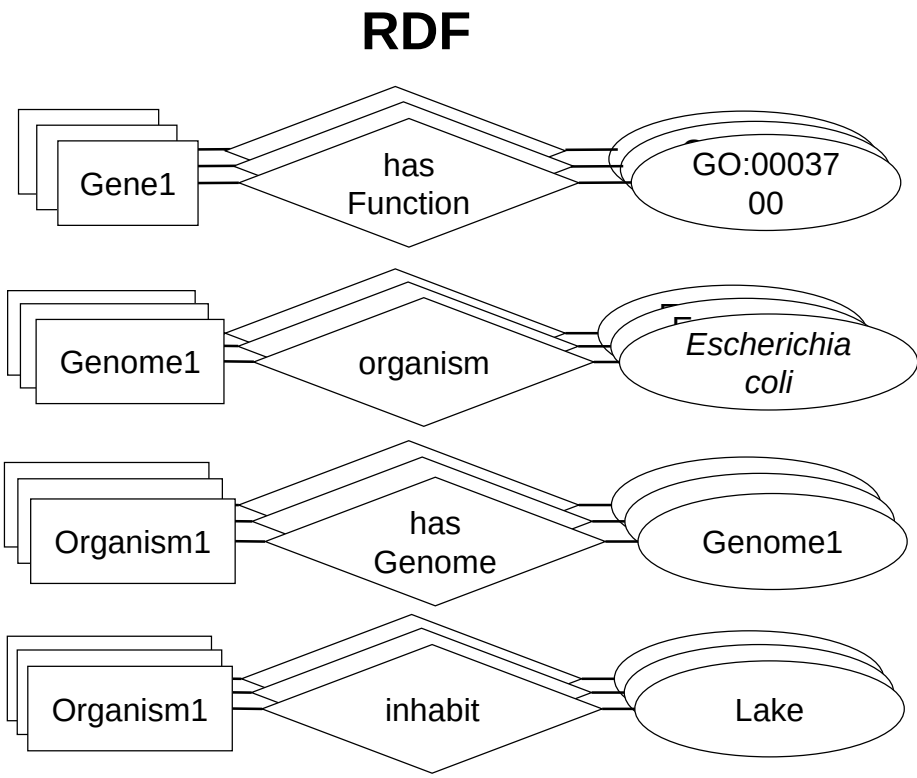
RDF is a standard data model of Semantic Web technology

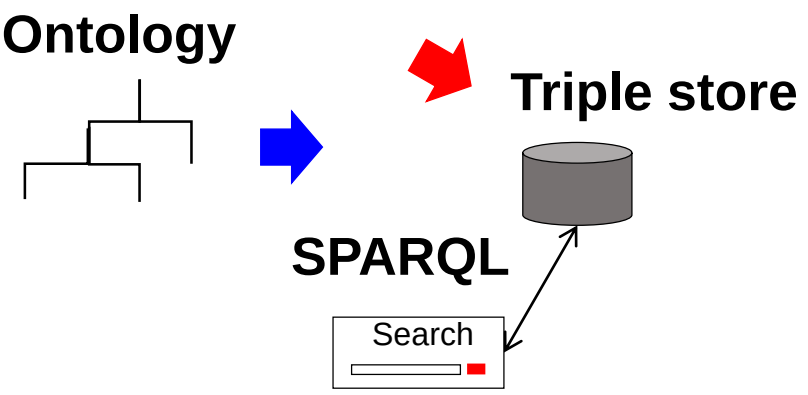
RDF (Resource Description Framework)
Data model which uses Triples
(**S**ubject – **P**redicate – **O**bject)



<URI>	<URI>	<URI>/Literal
gtps:Gene1	rdfs:label	“16S rRNA gene”

URI node can be linked to other nodes





To prepare data in RDF,
the database management system automatically recognize same resources.

You should describe your resource by using some Ontologies

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

For example, to answer: What is soil? What is a relationship between soil and sand?

MEO (Microbes/Metagenomes Environmental Ontology)

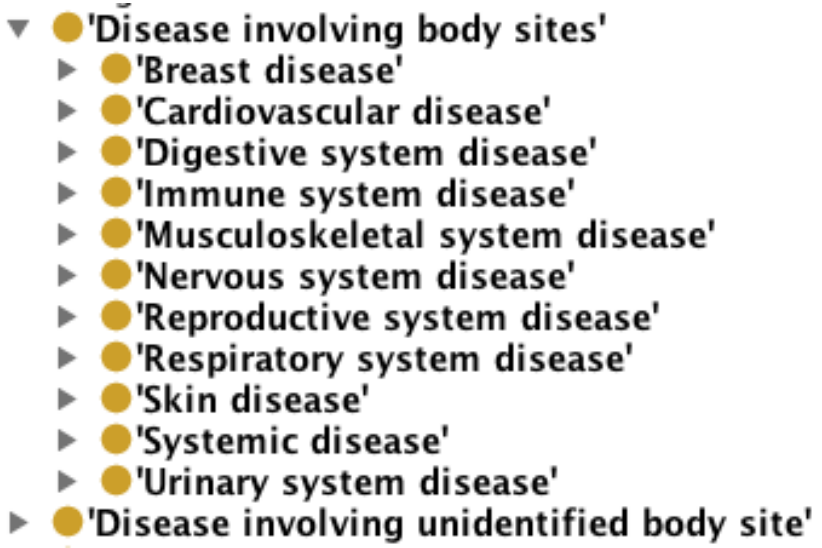
MSV (Metagenome Sample Vocabulary)

MCCV (Microbial Culture Collection Vocabulary)

MPO (Microbial Phenotype Ontology)

MBGD Ontology

PDO (Pathogenic Disease Ontology)



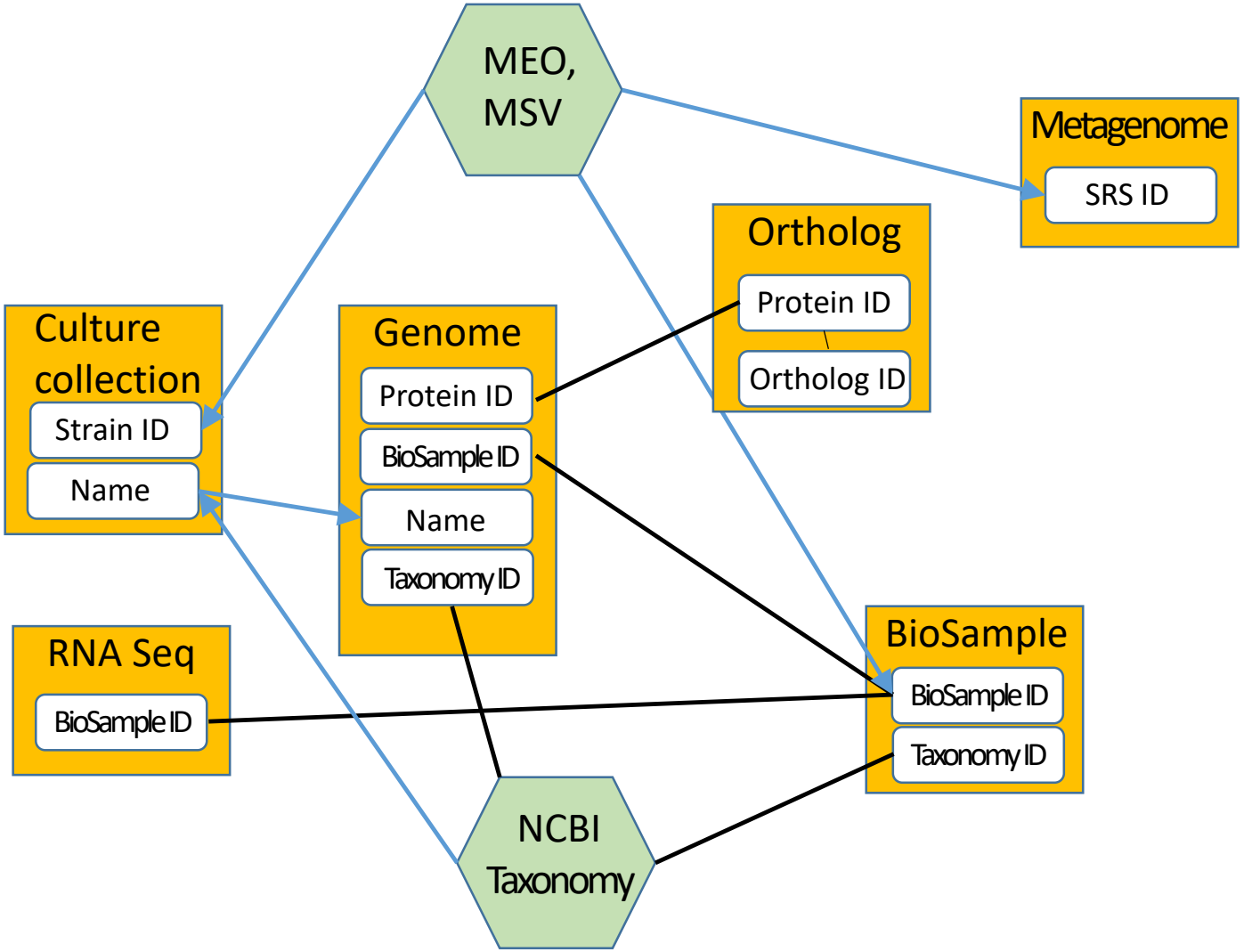
Most of them can be obtained from BioPortal

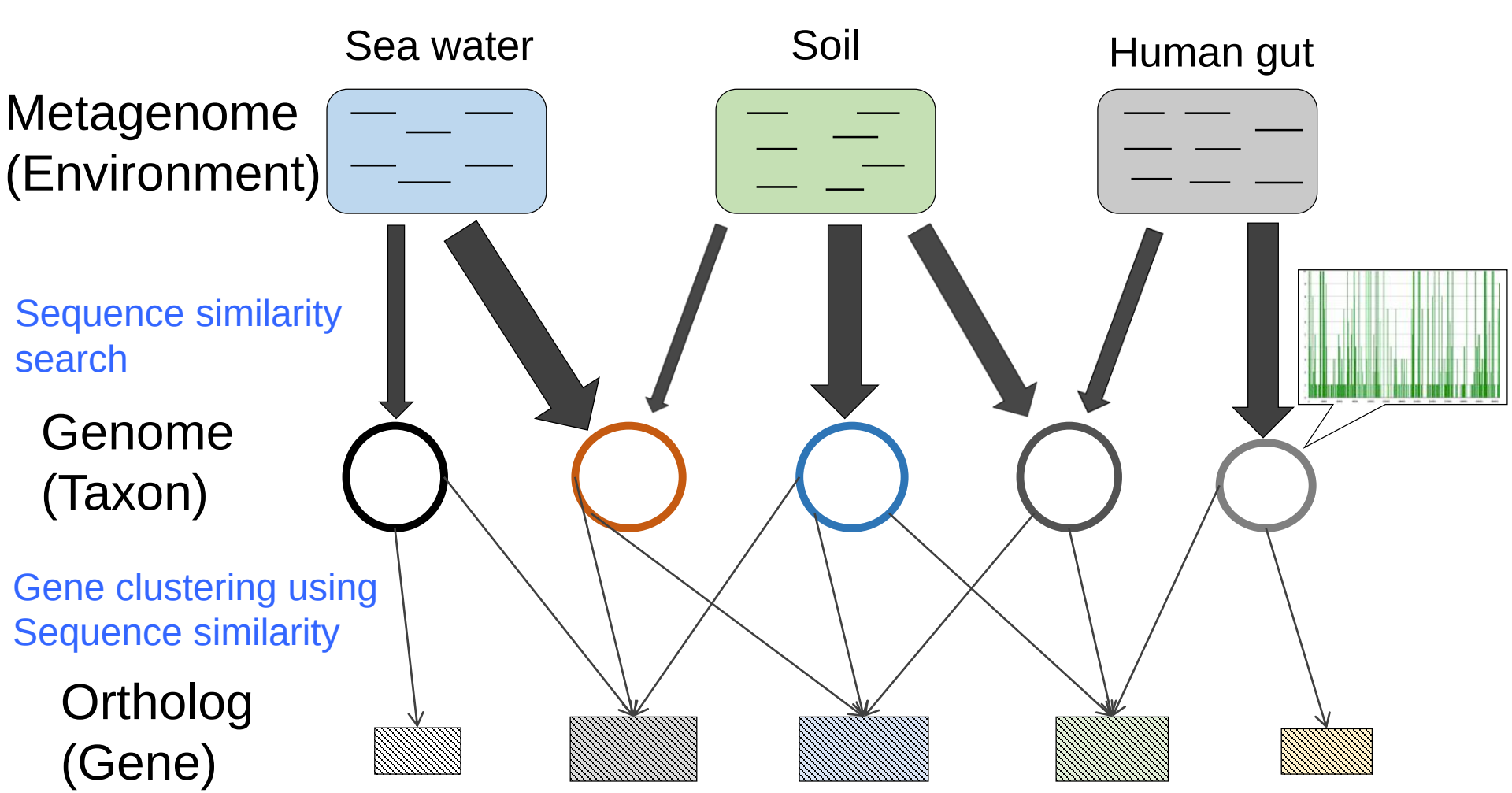
MicrobeDB.jp version 2 data

Data categories	Data sources	Ontologies
Genome	RefSeq Prokaryotes, Fungi, Algae	SO, FALDO, NCBITAX, INSDCO
Ortholog	MBGD	ORTHO
Culture collection	JCM, NBRC	MCCV, MPO
RNA-Seq	INSDC DRA	BAO
Genome & RNA-Seq Metadata	INSDC BioSample	MPO, MEO, MSV, PDO, CSSO
Metagenome	INSDC DRA	MEO, MSV

ID Mapping

Ontology Manual/Semiautomatic Annotation





- Ontology-based integration
- Sequence-based integration

Microbiome data in DRA/ERA/SRA

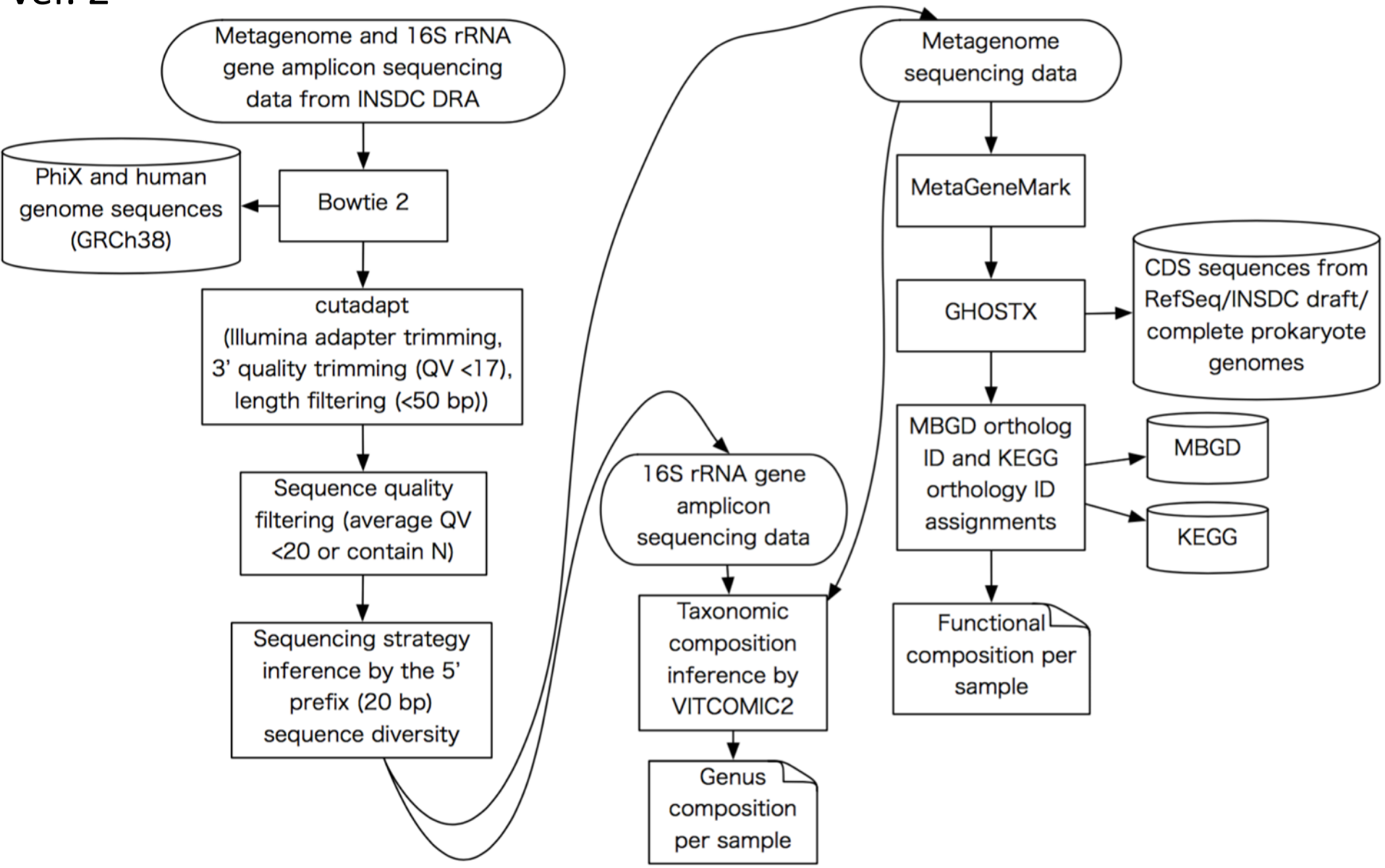
- 2014
  - Microbiomes: 173,359 (samples)

    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



http://microbedb.jp/

https://beta.microbedb.jp

MicrobeDB.jp

Home Document Analysis

e.g. hot spring, Enterococcus faecalis, psbA Search

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

MicrobeDB.jp

Home Document Analysis e.g. hot spring, Enterococcus faecalis, psbA Search Sign Up Sign in

Search id or term...

Public/Private

- metagenome_public 3729

hasMetagenomeAnalysis

- **taxonomy** 3729
- function 609

hasMEO (Text)

gut x

hasMEO: Component

Component for environment 389

hasMEO: Env

Environment for microbes 3729

hasMEO: Position

Position toward environment 5

hasMEO: State

Metagenomic samples

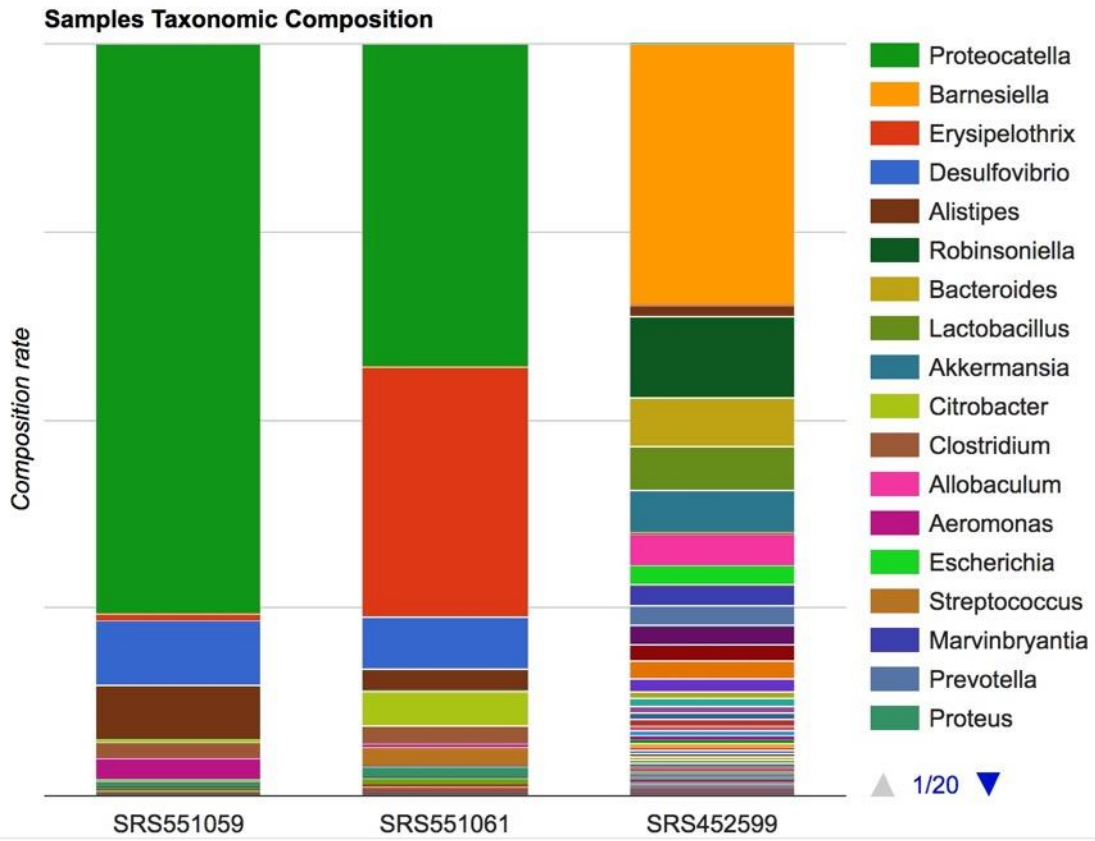
3729 results found in 112ms

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
Remove	SRS551059	Content of the intestinum from animals fed one meal of heparinized sheep blood	gut metagenome	749906	PRJNA237098	SAMN026145
Remove	SRS551061	Content of the intestinum of animals fed two meals of heparinized sheep blood four weeks apart	gut metagenome	749906	PRJNA237098	SAMN026145
Remove	SRS452599	Environmental/Metagenome sample for mouse gut metagenome	mouse gut metagenome	410661	PRJNA209582	SAMN022138
Add	SRS367344	Pooled 16S rRNA gene sequences	Bacteria	2	PRJNA209582	SAMN017587
Add	SRS369251	Pooled 16S rRNA gene sequences		2	PRJNA209582	SAMN017656


Samples Taxonomic Composition
Composition rate
SRS551059
SRS551061
SRS452599
Proteocatella
Barnesiella
Erysipelothrix
Desulfovibrio
Alistipes
Robinsoniella
Bacteroides
Lactobacillus
Akkermansia
Citrobacter
Clostridium
Allobaculum
Aeromonas
Escherichia
Streptococcus
Marvinbryantia
Prevotella
Proteus
1/20

VITCOMIC2

http://vitcomic.org

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: 選択... 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:
upload&calculate clear

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

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（微生物版GPS）。

(Higashi K et al. 2018, PLoS Comp Biol)