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 ...

Ortholog (e.g., MBGD, eggNOG)

Taxonomy (e.g., NCBI Taxonomy)

Pathogen (e.g., PATRIC)

Genome (e.g., INSDC, Ensembl, GOLD)

Culture Collection (e.g., StrainInfo, WDCM)

Gene Function (e.g., KEGG, MetaCyc, GO)

Metagenome (e.g., EBI-Metagenomics, IMG/M, MG-RAST)

Which DBs should we use (or recommend for beginners)?
integrates lots of data related to microbes. Especially, we integrates the microbial data that can be linked to genomes. since 2011

Ortholog: MBGD
Genome: RefSeq
Annotation: TogoAnnotation
Taxonomy: NCBI Taxonomy
Culture Collection: NBRC/JCM
Metadata: INSDC DRA
Metagenome: INSDC DRA

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**
RDF is a standard data model of Semantic Web technology

RDF (Resource Description Framework)
- Data model which uses Triples (Subject – Predicate – Object)

Gene1 has Function GO:0003700

Genome1 organism Escherichia coli

Organism1 genome Genome1

Organism1 inhabit Lake

URI node can be linked to other nodes

gtps:Gene1 rdfs:label "16S rRNA gene"

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](http://bioportal.bioontology.org/).
## MicrobeDB.jp version 2 data

<table>
<thead>
<tr>
<th>Data categories</th>
<th>Data sources</th>
<th>Ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome</td>
<td>RefSeq Prokaryotes, Fungi, Algae</td>
<td>SO, FALDO, NCBITAX, INSDCO</td>
</tr>
<tr>
<td>Ortholog</td>
<td>MBGD</td>
<td>ORTHO</td>
</tr>
<tr>
<td>Culture collection</td>
<td>JCM, NBRC</td>
<td>MCCV, MPO</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>INSDC DRA</td>
<td>BAO</td>
</tr>
<tr>
<td>Genome &amp; RNA-Seq</td>
<td>INSDC BioSample</td>
<td>MPO, MEO, MSV, PDO, CSSO</td>
</tr>
<tr>
<td>Metadata</td>
<td>INSDC DRA</td>
<td>MEO, MSV</td>
</tr>
</tbody>
</table>
ID Mapping

Ontology Manual/Semiautomatic Annotation

- Culture collection
  - Strain ID
  - Name

- MEO, MSV

- Ortholog
  - Protein ID
  - Ortholog ID

- Metagenome
  - SRS ID

- Genome
  - Protein ID
  - BioSample ID
  - Name
  - Taxonomy ID

- RNA Seq
  - BioSample ID

- NCBI Taxonomy

- BioSample
  - BioSample ID
  - Taxonomy ID
Metagenome (Environment)

Sea water

Genome (Taxon)

Soil

Gene clustering using Sequence similarity search

Ortholog (Gene)

Human gut

Sequence similarity search

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

- Metagenome and 16S rRNA gene amplicon sequencing data from INSDC DRA
  - PhiX and human genome sequences (GRCh38)
    - Bowtie 2
      - cutadapt (Illumina adapter trimming, 3’ quality trimming (QV <17), length filtering (<50 bp))
        - Sequence quality filtering (average QV <20 or contain N)
          - Sequencing strategy inference by the 5’ prefix (20 bp) sequence diversity
  - 16S rRNA gene amplicon sequencing data
    - Taxonomic composition inference by VITCOMIC2
      - Genus composition per sample
    - Functional composition per sample

- Metagenome sequencing data
  - MetaGeneMark
    - MBGD ortholog ID and KEGG orthology ID assignments
      - MBGD
      - KEGG
  - CDS sequences from RefSeq/INSDC draft/complete prokaryote genomes
MicrobeDB.jp portal beta version is available.

Environment: hot spring
Taxonomy: Enterococcus faecalis
Taxonomy: Streptomyces avermitilis
Gene: psbA
ID: 29
Integrating and representing genome, metagenome, taxonomy resources and the analysis datasets with Semantic Web Technologies.

**Database statistics**

- Total number of Metagenomic samples (SRA/SRS):
  - with taxonomic analysis results: 173,359 samples
  - with functional analysis results: 60,551 samples

- Total number of Assembled Genomes (RefSeq/Genbank):
  - 4,048 samples
  - 16,983 taxa

- Total number of Strains (JCM/NBRC):
  - 16,671 strains

- Total number of Environmental terms in ontology (MEO):
  - 2,381 terms
### Metagenomic samples

**3729 results found in 112ms**

<table>
<thead>
<tr>
<th>Select</th>
<th>MDB SampleID</th>
<th>msv:sampleTitle</th>
<th>msv:scientificName</th>
<th>msv:hasTaxonID</th>
<th>msv:hasBioProjectID</th>
<th>msv:hasBioSampleID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remove</td>
<td>SRS551059</td>
<td>Content of the intestine from animals fed one meal of heparinized sheep blood</td>
<td>gut metagenome</td>
<td>749906</td>
<td>PRJNA237098</td>
<td>SAMN02614592</td>
</tr>
<tr>
<td>Remove</td>
<td>SRS551061</td>
<td>Content of the intestine of animals fed two meals of heparinized sheep blood four weeks apart</td>
<td>gut metagenome</td>
<td>749906</td>
<td>PRJNA237098</td>
<td>SAMN02614592</td>
</tr>
<tr>
<td>Remove</td>
<td>SRS452599</td>
<td>Environmental/Metagenome sample for mouse gut metagenome</td>
<td>mouse gut metagenome</td>
<td>410661</td>
<td>PRJNA209582</td>
<td>SAMN02213836</td>
</tr>
<tr>
<td>Add</td>
<td>SRS367344</td>
<td>Pooled 16S rRNA gene sequences</td>
<td>Bacteria</td>
<td>2</td>
<td>PRJNA209582</td>
<td>SAMN01758737</td>
</tr>
<tr>
<td>Add</td>
<td>SRS369251</td>
<td>Pooled 16S rRNA gene sequences</td>
<td></td>
<td>2</td>
<td>PRJNA209582</td>
<td>SAMN01765638</td>
</tr>
</tbody>
</table>
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 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: [Select...]
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: [Enter 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)