“Small Bugs, Big Data”:
Developing an integrated Database for Microbes with Semantic Web Technologies

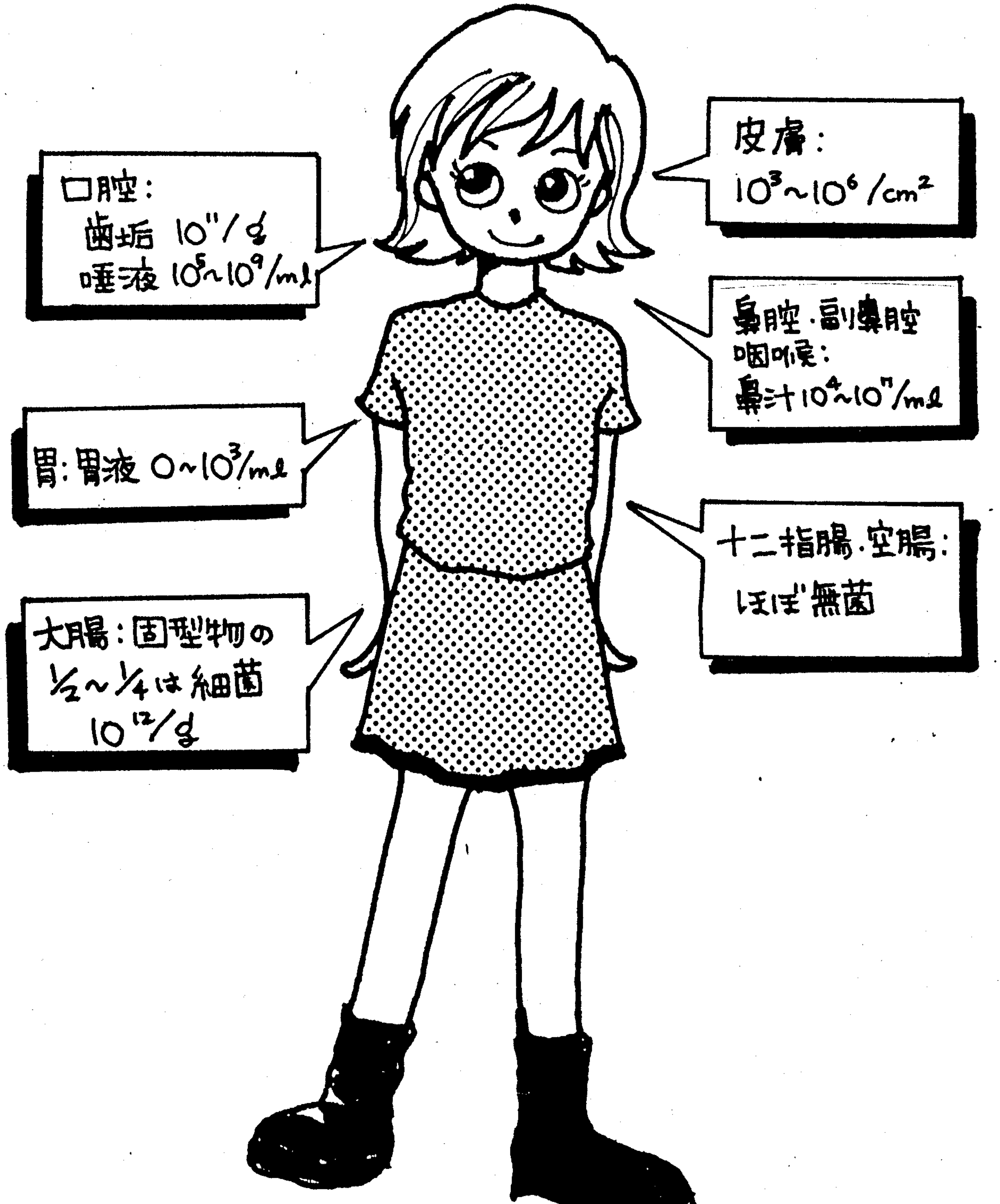
Ken Kurokawa (Earth-Life Science Institute, TITECH)

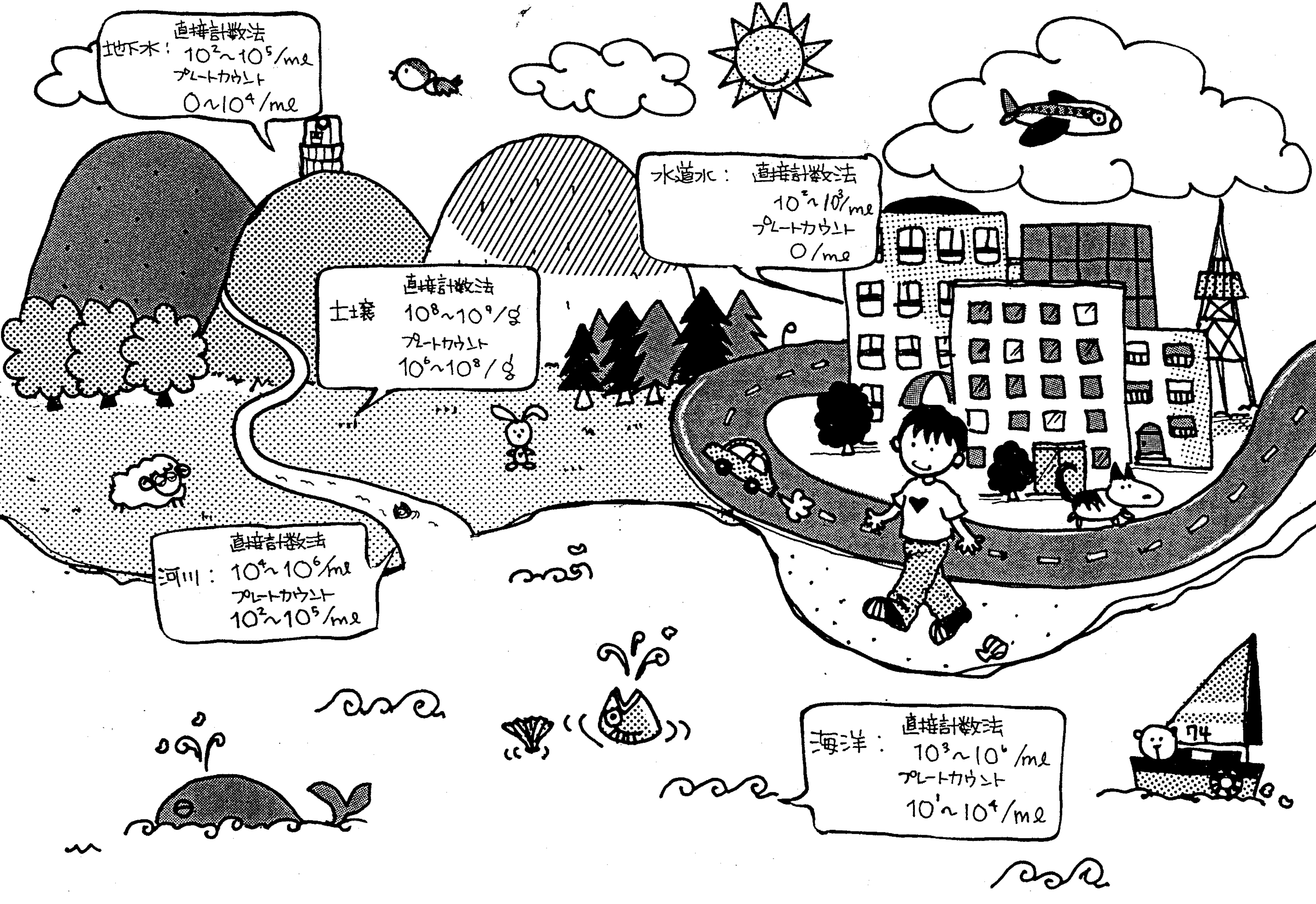
東京工業大学
Tokyo Institute of Technology

wpi
World Premier International
Research Center Initiative

JST

ELSI
EARTH - LIFE SCIENCE INSTITUTE


口腔：
歯垢 10^{11}/g
唾液 10^5~10^9/ml
皮膚：
10^3~10^6/cm^2
鼻腔・副鼻腔
咽喉：
鼻汁 10^4~10^7/ml
胃：胃液 0~10^3/ml
十二指腸・空腸：
ほぼ無菌
大腸：固型物の
1/2~1/4は細菌
10^{12}/g



地下水：直接計数法 10^2~10^5/mℓ
プレートカウント 0~10^4/mℓ
水道水：直接計数法 10^2~10^3/mℓ
プレートカウント 0/mℓ
土壌 直接計数法 10^8~10^9/g
プレートカウント 10^6~10^8/g
河川：直接計数法 10^4~10^6/mℓ
プレートカウント 10^2~10^5/mℓ
海洋：直接計数法 10^3~10^6/mℓ
プレートカウント 10^1~10^4/mℓ
74

The next generation…

Metagenomics

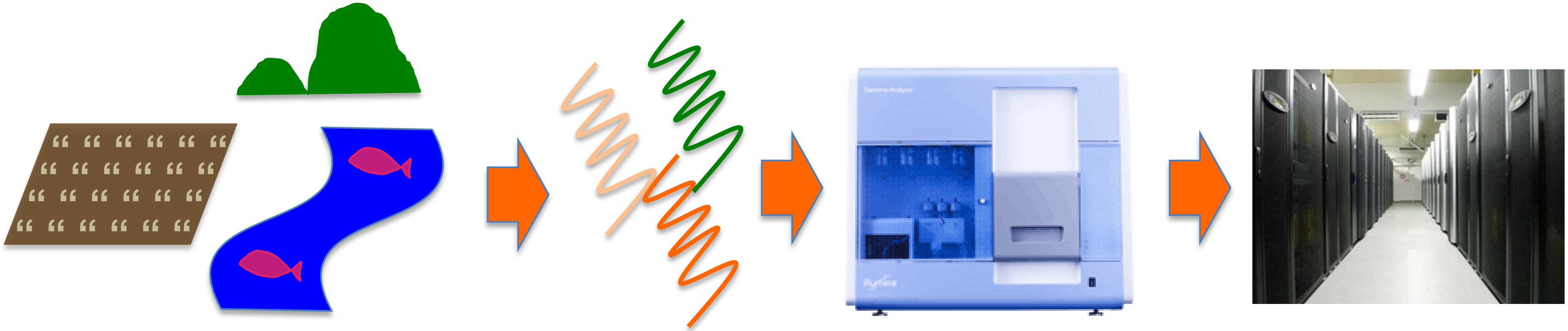
Bacterial community …

Natural environment (marine, river, soil..)
Human, animal (intestine, oral, skin..)

To elucidate the bacterial gene pools in environments, deeply sequencing the entire genomes extracted from bacterial community without cultivation

Genome analysis against “Microbiome”

Metagenomic analysis



Microbiome sampling

DNA extraction

DNA sequencing

Bioinformatics analysis

Both genomic and metagenomic data stored in public Databases

Microbial genome data (NCBI RefSeq DB)

Taxonomic division	Genomes
Archaea	375
Bacteria	24,119

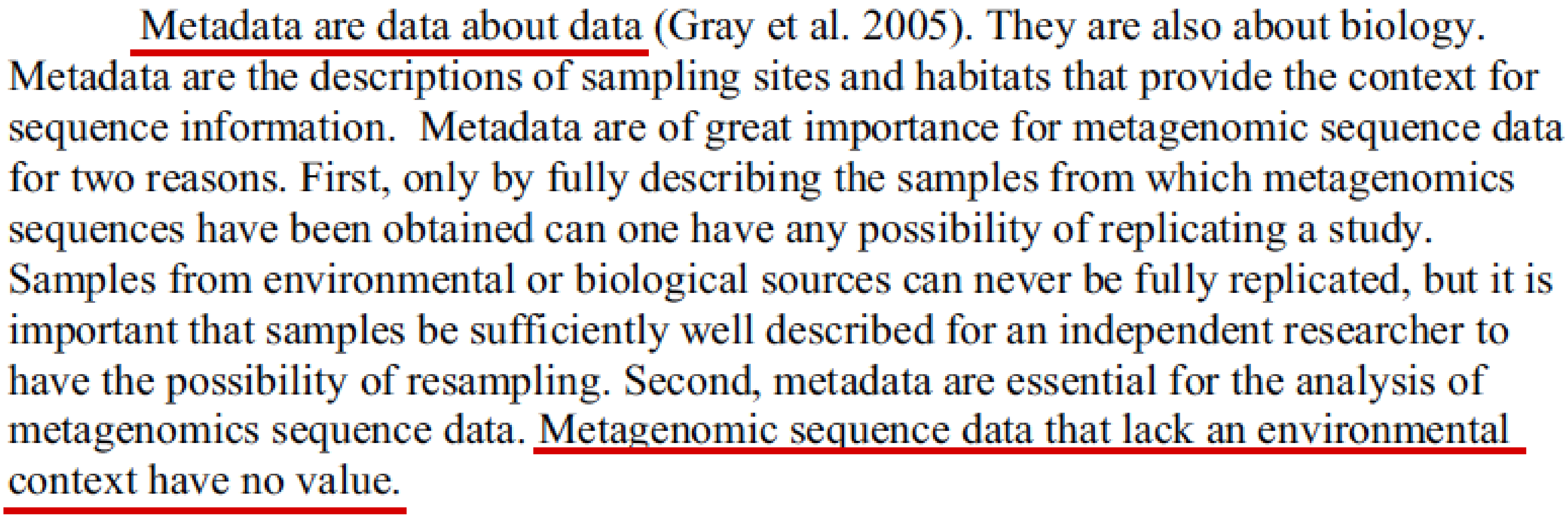
Metagenomic data

DB	Env. metagenomics	Human metagenomics
MG-RAST	14,188	3,291
JGI IMG/M	1,694	840
INSDC SRA	23,214	18,108

Aprirl 2014

The Importance of Metadata

The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet
The National Academy of Science, 2007



Metadata are data about data (Gray et al. 2005). They are also about biology. Metadata are the descriptions of sampling sites and habitats that provide the context for sequence information. Metadata are of great importance for metagenomic sequence data for two reasons. First, only by fully describing the samples from which metagenomics sequences have been obtained can one have any possibility of replicating a study. Samples from environmental or biological sources can never be fully replicated, but it is important that samples be sufficiently well described for an independent researcher to have the possibility of resampling. Second, metadata are essential for the analysis of metagenomics sequence data. Metagenomic sequence data that lack an environmental context have no value.

Metadata are the description of sampling sites and habitats that provide the context for sequence information

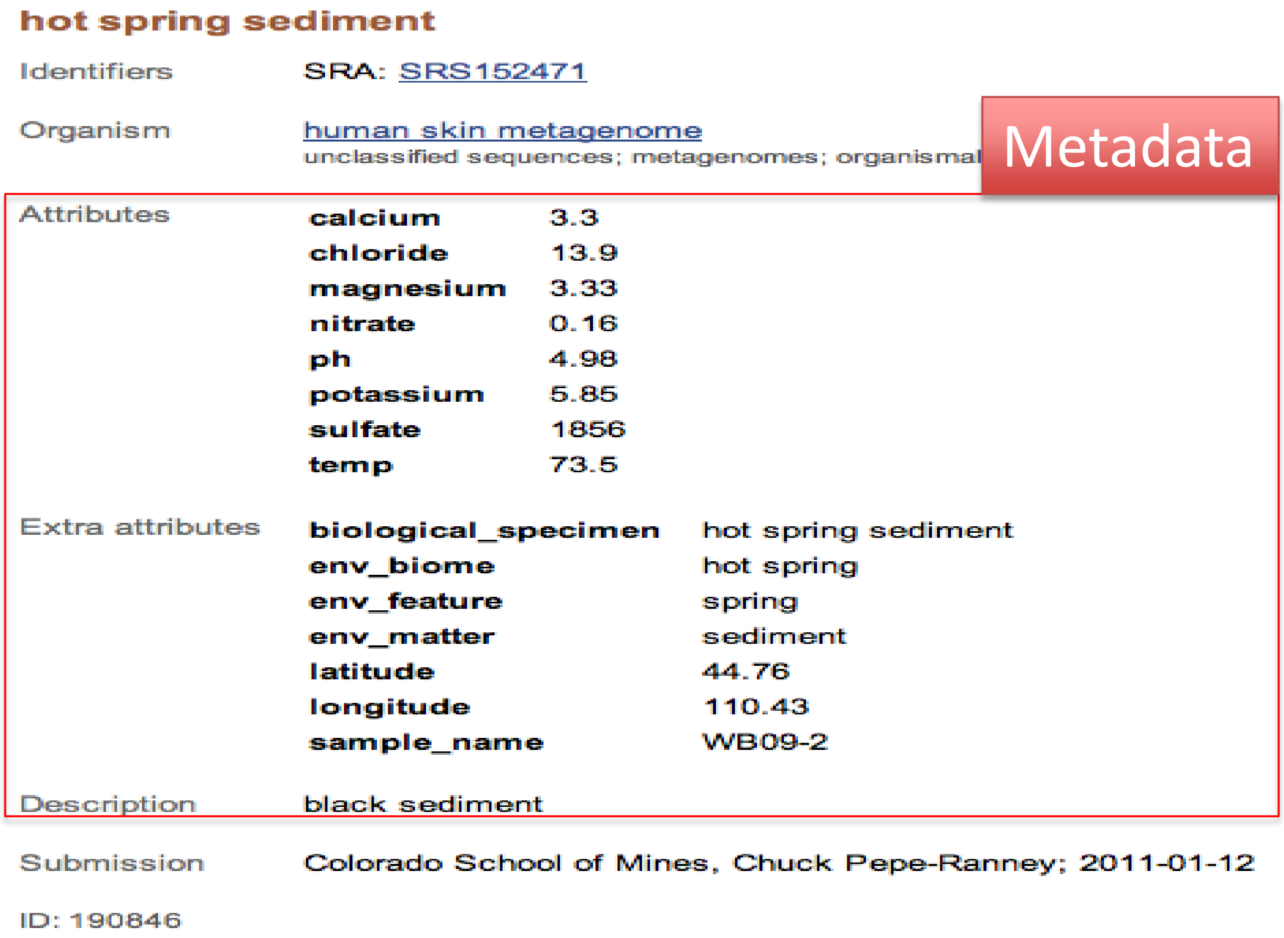
Human microbiome:

Age, Sex, BMI, Body habitat, Country, Diet, Disease stage, Family relationship …

Environmental microbiome:

lat / long, pH, Depth, Dissolved oxygen, Wind speed, Total nitrogen, Temp. …

Metagenomic analysis is performed in a variety of environments, and its massive data is stored in the public databases



hot spring sediment

Identifiers SRA: SRS152471

Organism human skin metagenome
unclassified sequences; metagenomes; organismal

Metadata

Attributes		
	calcium	3.3
	chloride	13.9
	magnesium	3.33
	nitrate	0.16
	ph	4.98
	potassium	5.85
	sulfate	1856
	temp	73.5
Extra attributes	**biological_specimen**	hot spring sediment
	env_biome	hot spring
	env_feature	spring
	env_matter	sediment
	latitude	44.76
	longitude	110.43
	sample_name	WB09-2

Description black sediment

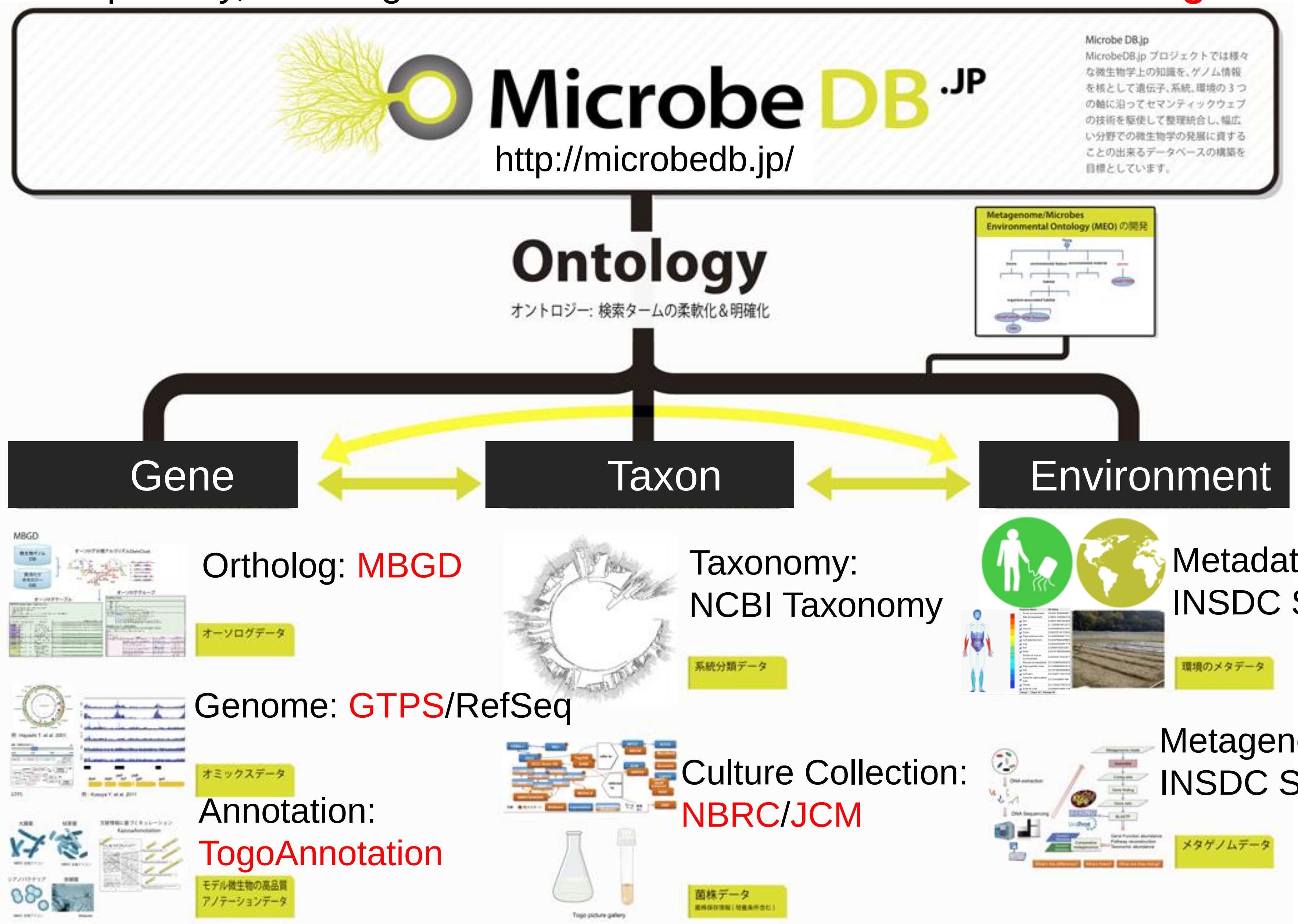
Submission Colorado School of Mines, Chuck Pepe-Ranney; 2011-01-12

ID: 190846

Temperature
pH
Anion & Cation
Long/Lat
:
:

Use of both metagenome sequence data and its metadata will enable the large-scale comparative metagenomic analysis

Microbe DB.JP integrates lots of data related to microbes.
Especially, we integrates the microbial data that can be linked to **genomes**.



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
オントロジー: 検索タームの柔軟化＆明確化

Gene | **Taxon** | **Environment**

Ortholog: MBGD

Genome: GTPS/RefSeq

Annotation:
TogoAnnotation

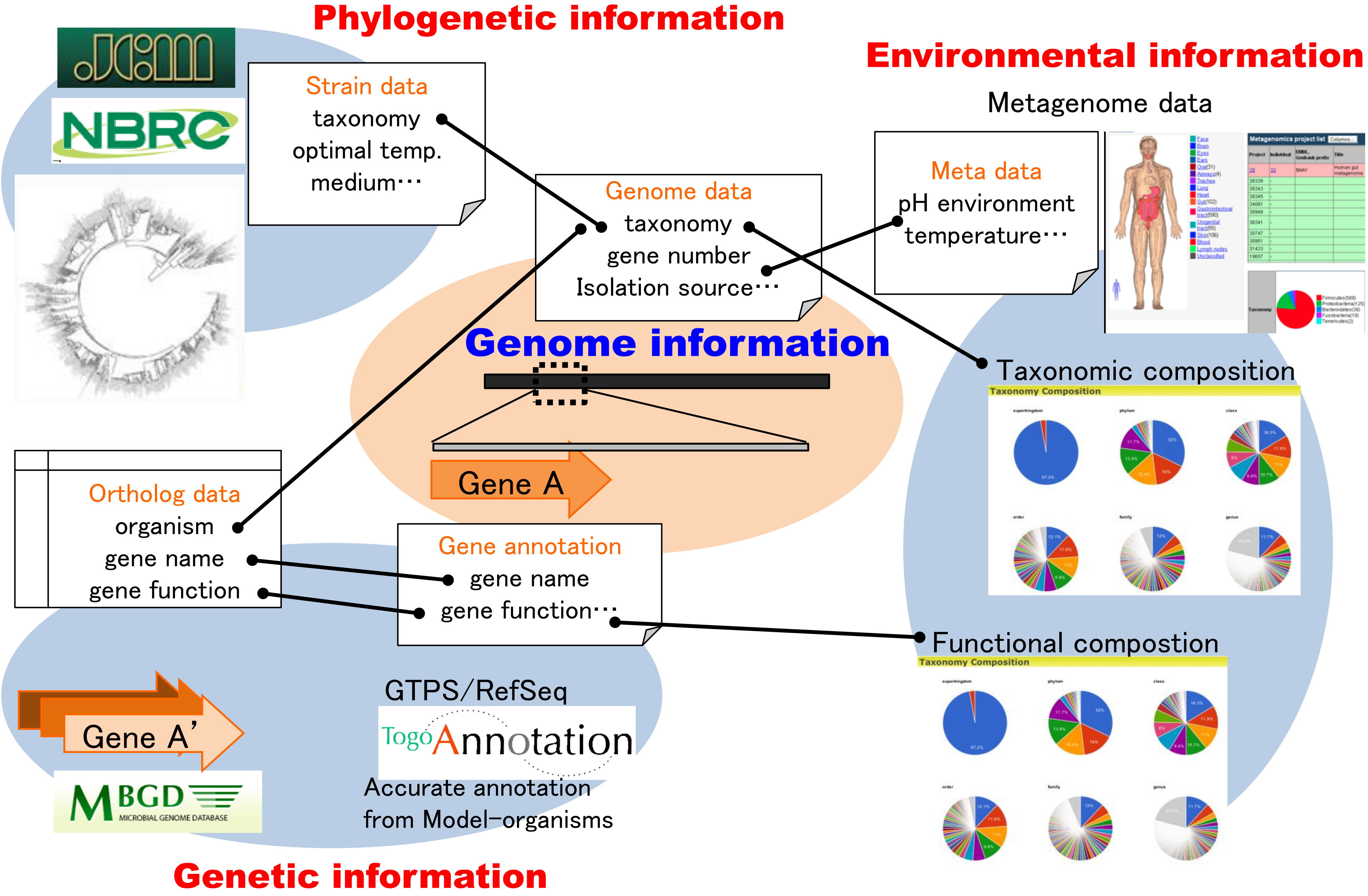
Taxonomy:
NCBI Taxonomy

Culture Collection:
NBRC/JCM

Metadata:
INSDC SRA

Metagenome
INSDC SRA

Integration of microbe’s data centers around genome information



Phylogenetic information

Strain data
taxonomy
optimal temp.
medium･･･

Genome data
taxonomy
gene number
Isolation source･･･

Genome information

Gene A

Environmental information

Metagenome data

Meta data
pH environment
temperature･･･

Taxonomic composition

Ortholog data
organism
gene name
gene function

Gene annotation
gene name
gene function･･･

Functional compostion

Gene A’

GTPS/RefSeq

Accurate annotation
from Model-organisms

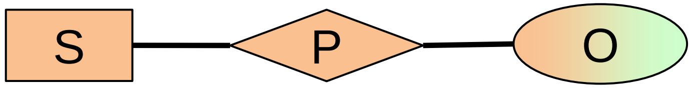
Genetic information

RDF is a standard data model of Semantic Web technology

RDF (Resource Description Framework)

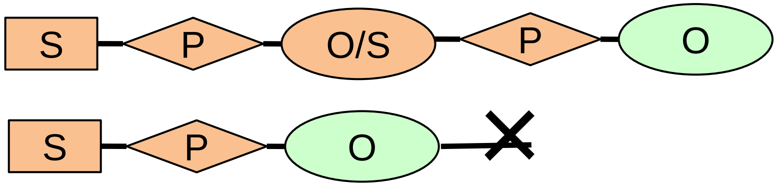
Data model which uses Triples

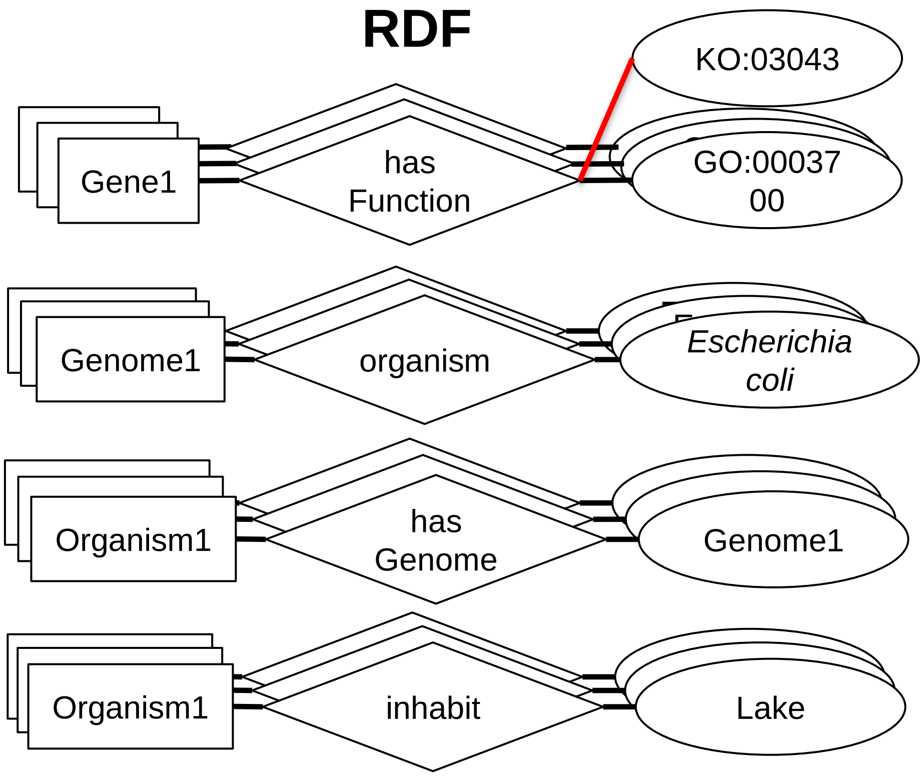
(**S**ubject – **P**redicate – **O**bject)

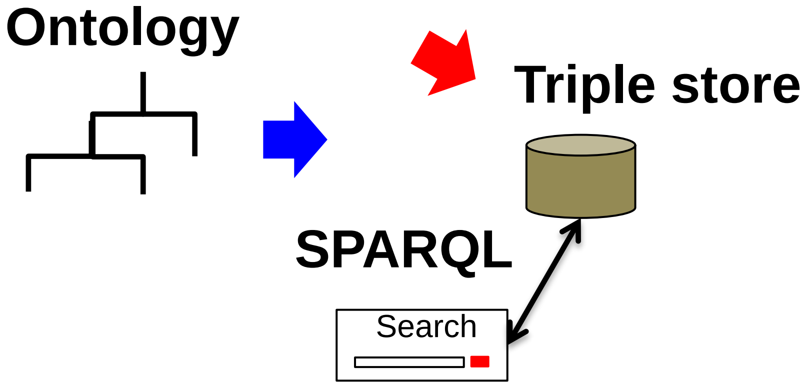


<URI>	<URI>	<URI>/Literal
gtps:Gene1	rdfs:label	“16S rRNA gene”

URI node can be linked to other nodes

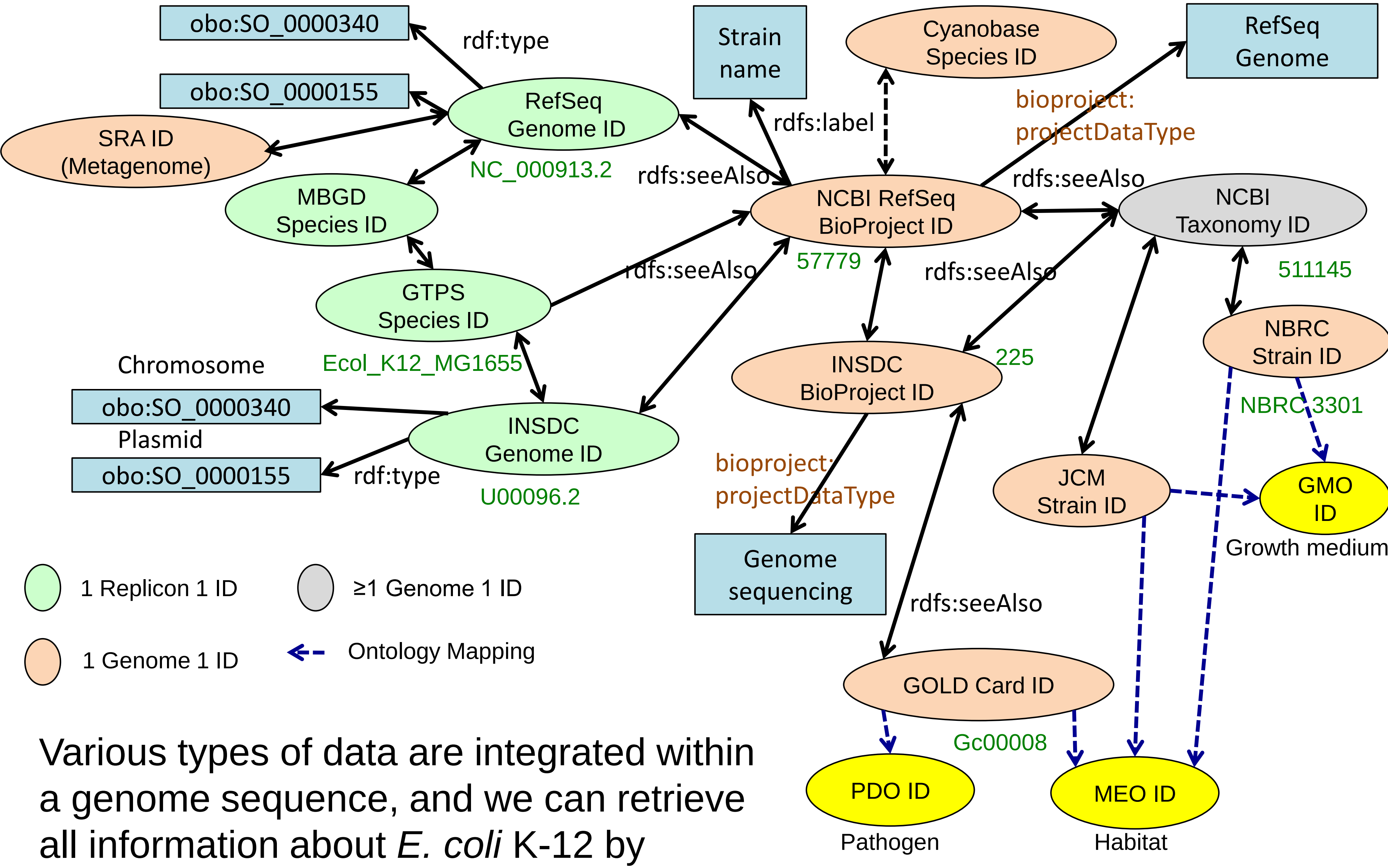






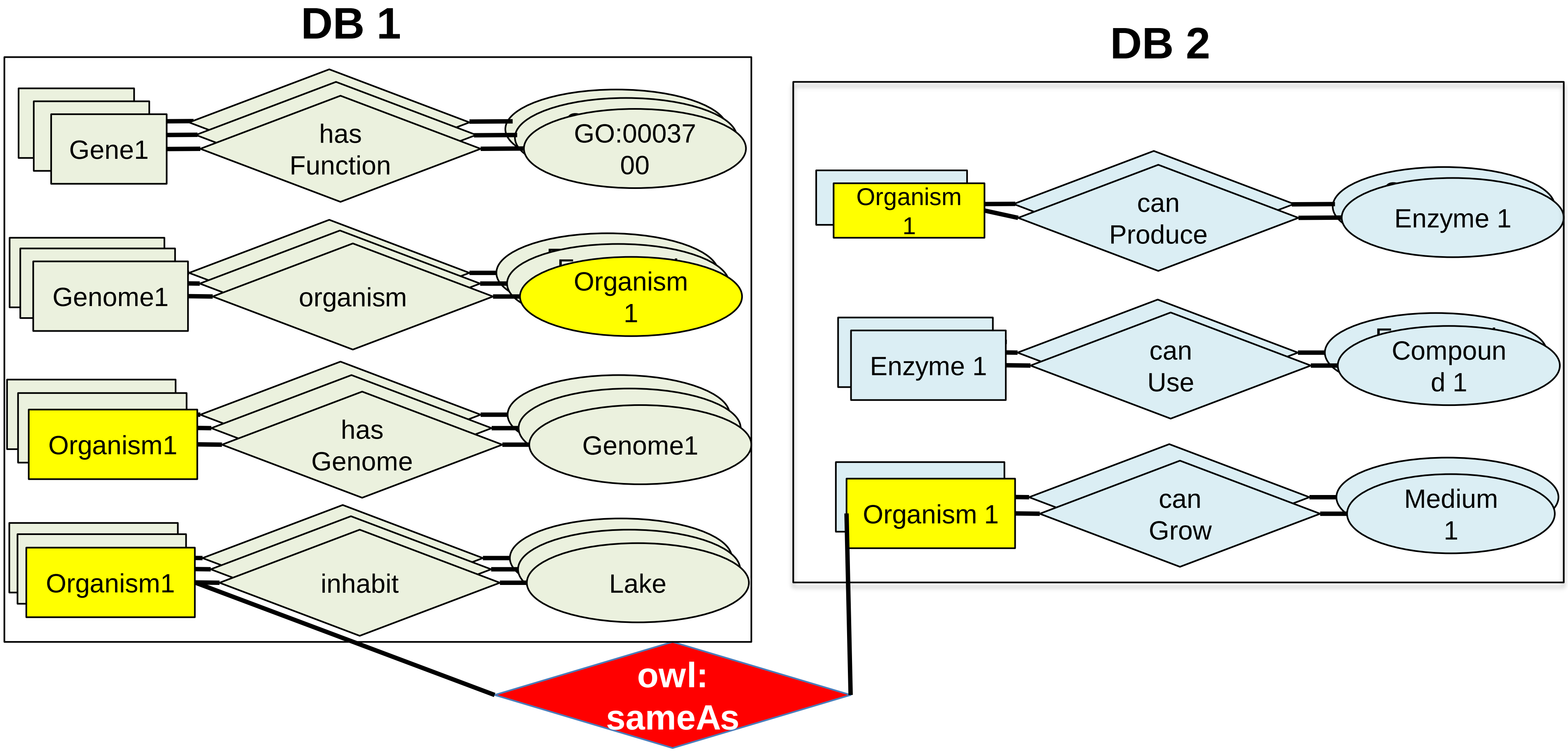
To prepare data in RDF,
the database management system automatically recognize same resources.

An example of RDF relationships for *E. coli* K-12 genome data



Various types of data are integrated within a genome sequence, and we can retrieve all information about *E. coli* K-12 by following these graphs.

How to integrate the data from two different DBs?



1. When two DBs use same URI, already two DB’s data are integrated.
2. If not, you can integrate two DB’s data by adding one Triple (db1:A owl:sameAs db2:B)

You don’t need to place all of these data in one DB management system.

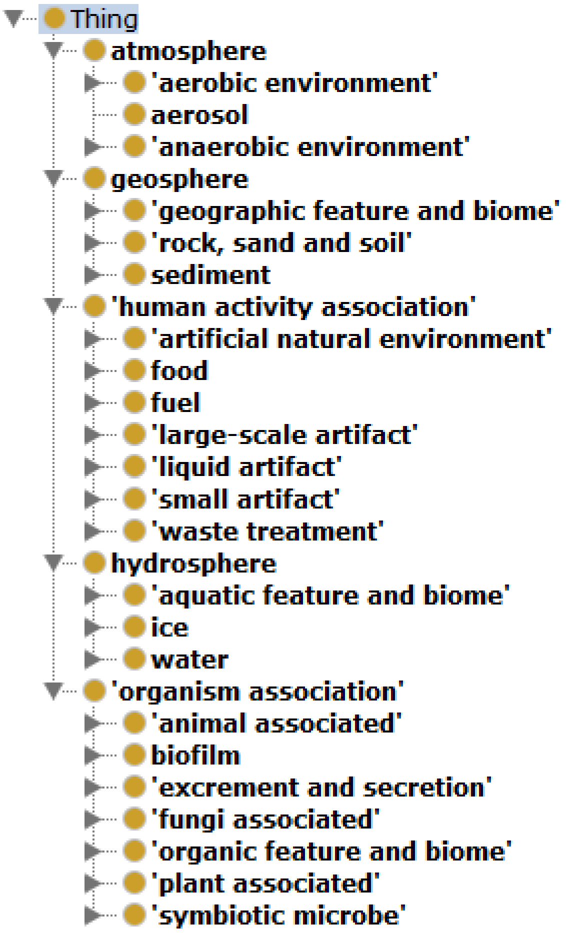
How can we discriminate whether two DB’s resources are same or not?

You should describe your resource by using some Ontologies

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

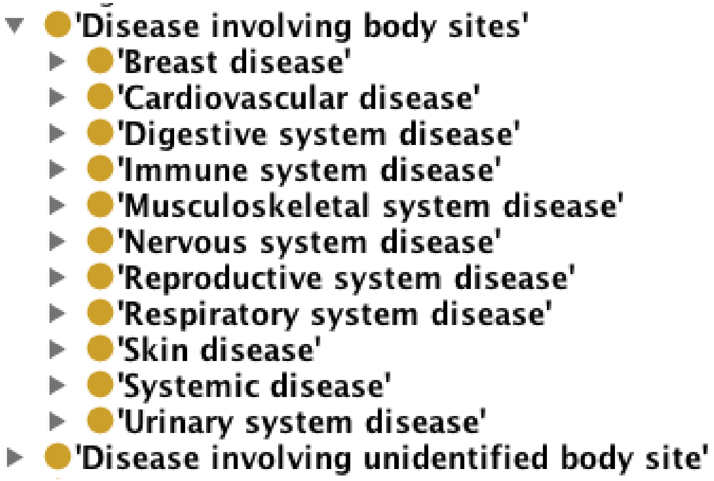
For example, to answer: What is soil? What is a relationship between soil and sand?

MEO (Microbes Environmental Ontology)



- Thing
  - atmosphere
    - 'aerobic environment'
    - aerosol
    - 'anaerobic environment'
  - geosphere
    - 'geographic feature and biome'
    - 'rock, sand and soil'
    - sediment
  - 'human activity association'
    - 'artificial natural environment'
    - food
    - fuel
    - 'large-scale artifact'
    - 'liquid artifact'
    - 'small artifact'
    - 'waste treatment'
  - hydrosphere
    - 'aquatic feature and biome'
    - ice
    - water
  - 'organism association'
    - 'animal associated'
    - biofilm
    - 'excrement and secretion'
    - 'fungi associated'
    - 'organic feature and biome'
    - 'plant associated'
    - 'symbiotic microbe'

PDO (Pathogenic Disease Ontology)



- 'Disease involving body sites'
  - 'Breast disease'
  - 'Cardiovascular disease'
  - 'Digestive system disease'
  - 'Immune system disease'
  - 'Musculoskeletal system disease'
  - 'Nervous system disease'
  - 'Reproductive system disease'
  - 'Respiratory system disease'
  - 'Skin disease'
  - 'Systemic disease'
  - 'Urinary system disease'
- 'Disease involving unidentified body site'

MCCV (Microbial Culture Collection Vocabulary)

MSV (Metagenome Sample Vocabulary)

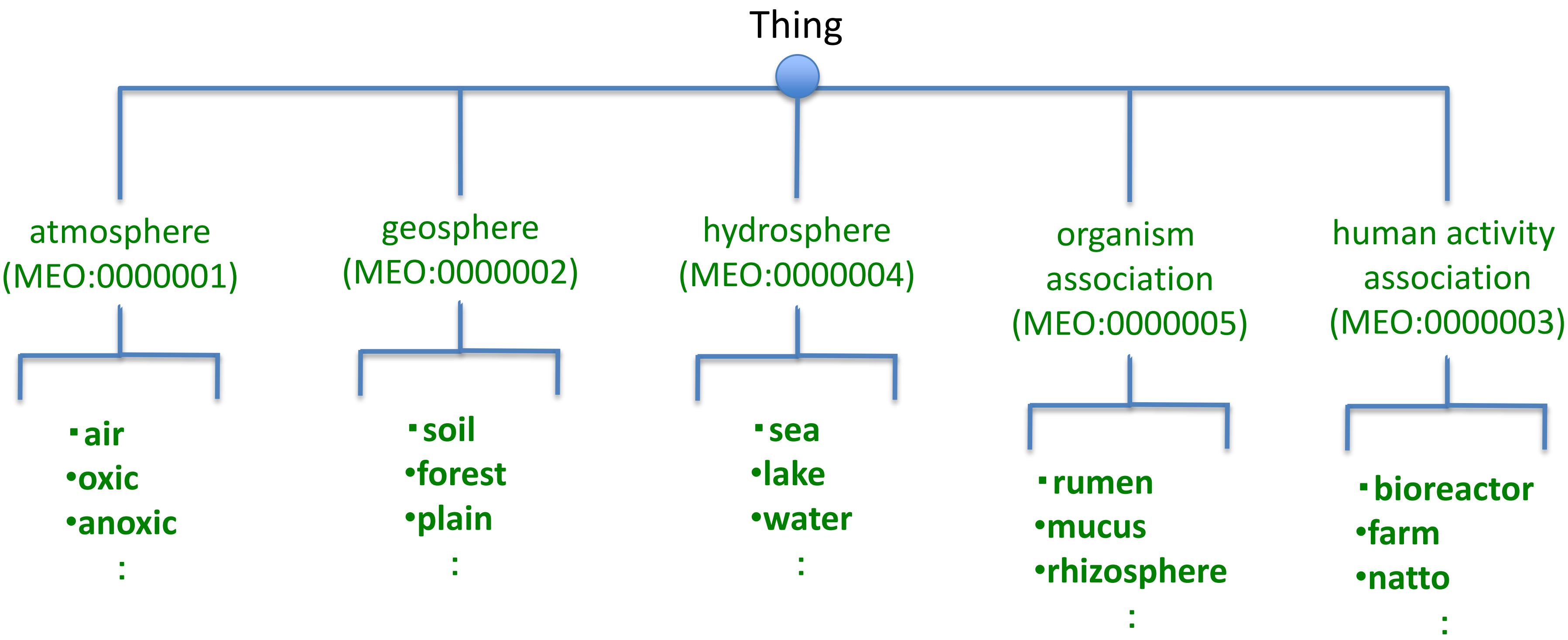
MPO (Microbial Phenotype Ontology)

MBGD Ortholog Ontology

Most of them can be obtained from BioPortal

Metagenome/Microbes Environmental Ontology (MEO) Ver. 0.3

MEO describes the vocabulary related to microbial habitat information



Thing

atmosphere (MEO:0000001)
- air
- oxic
- anoxic

:

geosphere (MEO:0000002)
- soil
- forest
- plain

:

hydrosphere (MEO:0000004)
- sea
- lake
- water

:

organism association (MEO:0000005)
- rumen
- mucus
- rhizosphere

:

human activity association (MEO:0000003)
- bioreactor
- farm
- natto

:

We mapped the MEO terms to the metagenome metadata terms

Example

MetagenomeMetadata: sand → isIncludedto → MEO: soil

Questions: What are the enriched genes in high temperature environment? Who have these genes?

Taxon list

Ortholog list

High temperature
& Gene

MicrobeDB.JP

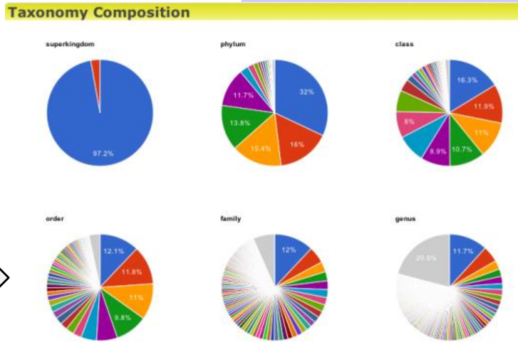
Search

MEO

High temperature
= ?? – ??℃

Environmental information

Functional compostion of
Metagenome data



Meta data
environment
temperature･･･

Phylogenetic information

Taxonomy data
Phylum
Class
Order･･･

Genome information

Organism list

Ortholog list

Gene A

Ortholog data
organism
gene name･･･

Gene A

Gene list

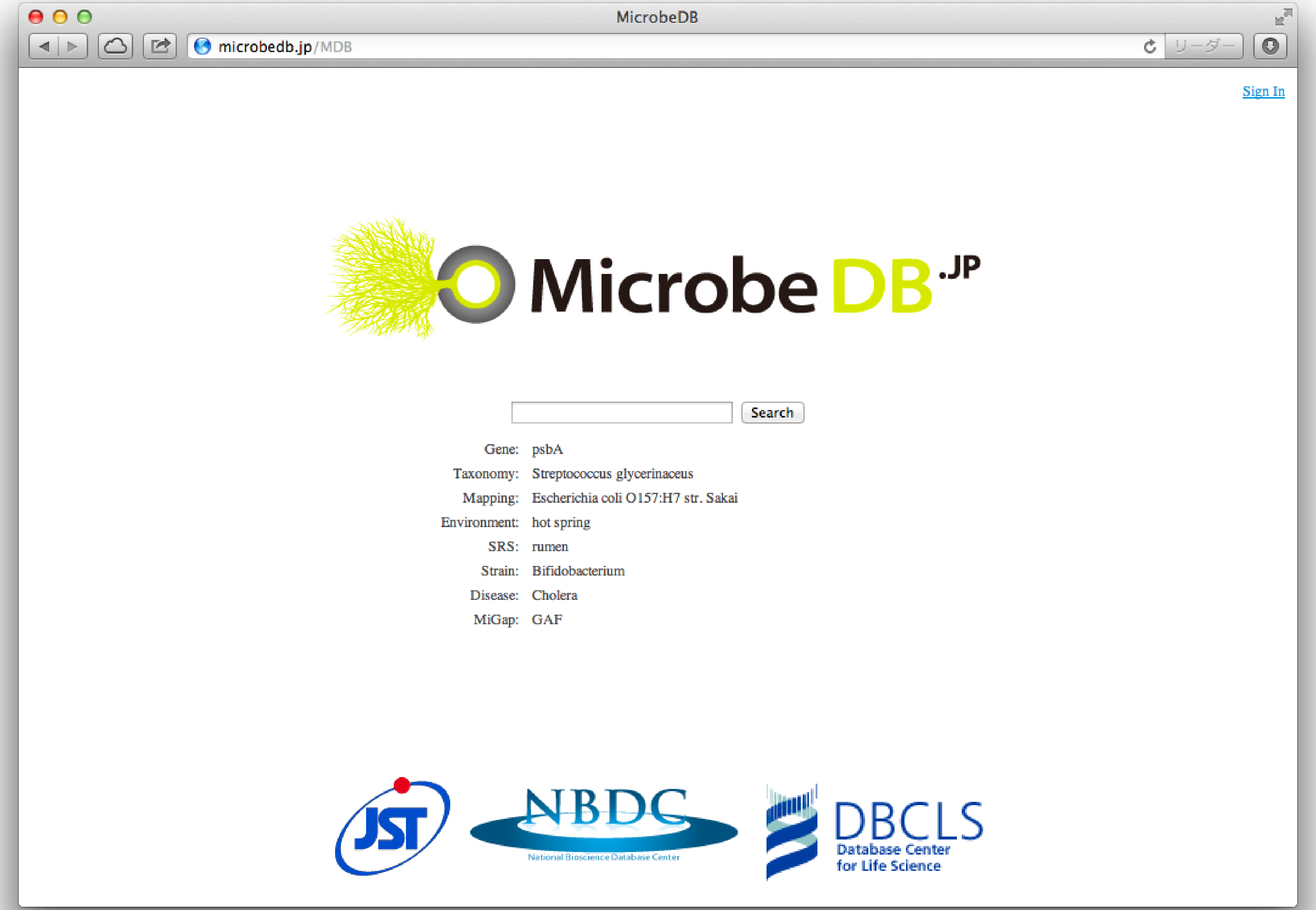
Gene annotation
gene name
gene function･･･

Genetic information

More than 1 billion Triples!

グラフ名	説明	作成元	トリプル数
refseq	RefSeq Prokaryoteゲノムデータ	DBCLS	550,273,744
mbgd	MBGD Orthologデータ	基生研	291,714,037
gtps	GTPSゲノムデータ	遺伝研	197,069,932
taxonomy	SPARQLthonで作成したNCBI Taxonomyオントロジー改良版 版	DBCLS,遺伝研,東工大	10,183,714
meta16S	各SRSメタ16Sの系統組成データ	東工大	9,831,600
gazetteer	地理オントロジー	外部機関	7,062,536
srs_metadata	SRSメタ16S・メタゲノムの様々なメタデータ	東工大	4,982,739
srs_ortholog	各SRSメタゲノムのMBGD Ortholog組成	東工大,基生研	2,026,746
go	Geneオントロジー	外部機関	1,211,571
brc	JCM/NBRC菌株データ with NCBI Taxonomy ID	遺伝研,東工大,DBCLS	903,319
gold	GOLDの個別ゲノムのMEO等へのオントロジーマッピングデータ	東工大,DBCLS	150,899
srs	SRSメタ16S・メタゲノムのMEO等へのオントロジーマッピング ピングデータ	東工大	53,691
so	Sequenceオントロジー	外部機関	43,060
pdo	感染症オントロジー + 症状オントロジー + ゲノムへのオント オントロジーマッピングデータ	東工大	8,809
meo	微生物の生息環境オントロジー	東工大	4,975
msv	SRSメタ16S・メタゲノムのメタデータオントロジー	東工大	1,601
mpo	微生物フェノタイプオントロジー	DBCLS	734
mccv	菌株オントロジー	東工大,DBCLS	293
その他中間データ	いくつかのデータ集計系のSPARQLクエリは遅いため、MSS MSSが集計結果のデータを作成		440,773
合計			1,075,964,773

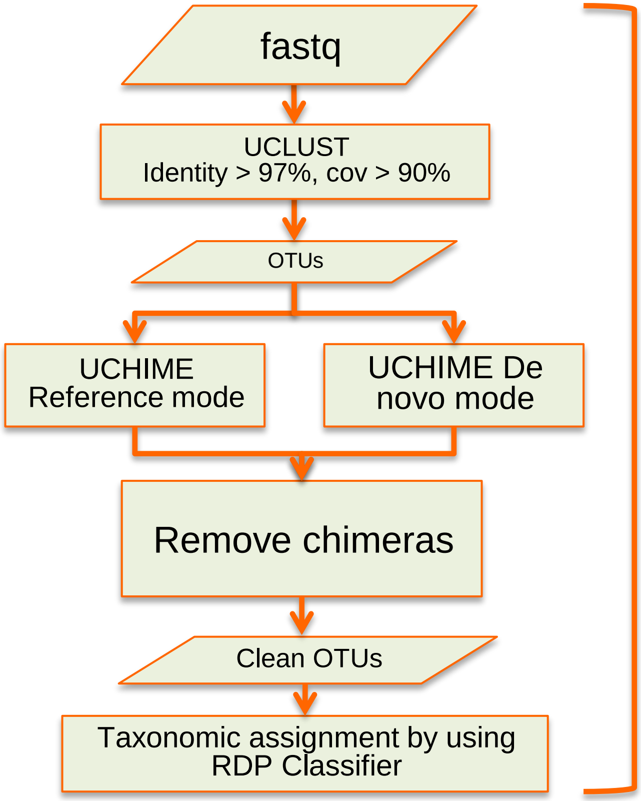
http://microbedb.jp/



Gene:	psbA
Taxonomy:	Streptococcus glycerinaceus
Mapping:	Escherichia coli O157:H7 str. Sakai
Environment:	hot spring
SRS:	rumen
Strain:	Bifidobacterium
Disease:	Cholera
MiGap:	GAF

Stanza Development

To obtain biological knowledge from low data (sequence and metadata), we developed a variety of “Stanza”, which is a compact, modular, and reusable application for data analysis.



fastq

UCLUST
Identity > 97%, cov > 90%

OTUs

UCHIME Reference mode

UCHIME De novo mode

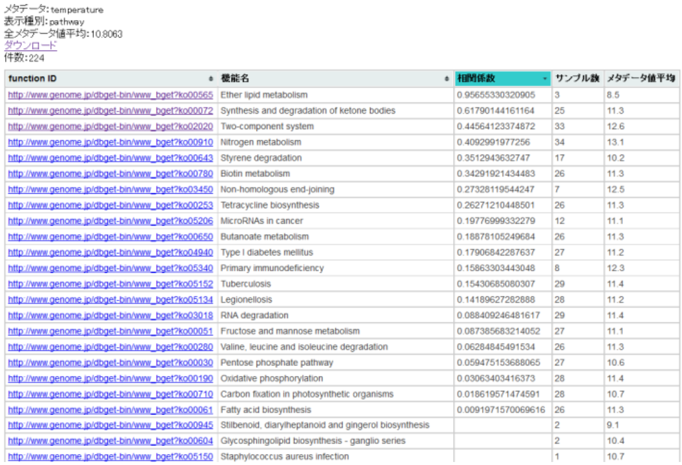
Remove chimeras

Clean OTUs

Taxonomic assignment by using RDP Classifier

Analyze data by using the Stanza

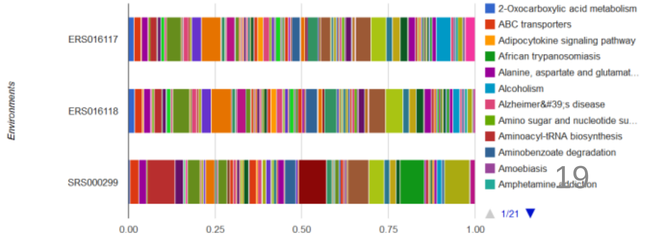
Correlation analysis between gene abundance and metadata



メタデータ: temperature
表示種別: pathway
全メタデータ値平均: 10.8063
ダウンロード
件数: 224

function ID	機能名	相関係数	サンプル数	メタデータ値平均
http://www.genome.jp/dbget-bin/www_bget?ko00565	Ether lipid metabolism	0.95655330320905	3	8.5
http://www.genome.jp/dbget-bin/www_bget?ko00072	Synthesis and degradation of ketone bodies	0.61790144161164	25	11.3
http://www.genome.jp/dbget-bin/www_bget?ko02020	Two-component system	0.44564123374872	33	12.6
http://www.genome.jp/dbget-bin/www_bget?ko00910	Nitrogen metabolism	0.4092991977256	34	13.1
http://www.genome.jp/dbget-bin/www_bget?ko00643	Styrene degradation	0.3512943632747	17	10.2
http://www.genome.jp/dbget-bin/www_bget?ko00780	Biotin metabolism	0.34291921434483	26	11.3
http://www.genome.jp/dbget-bin/www_bget?ko03450	Non-homologous end-joining	0.27328119544247	7	12.5
http://www.genome.jp/dbget-bin/www_bget?ko00253	Tetracycline biosynthesis	0.26271210448501	26	11.3
http://www.genome.jp/dbget-bin/www_bget?ko05206	MicroRNAs in cancer	0.19776999332279	12	11.1
http://www.genome.jp/dbget-bin/www_bget?ko00650	Butanoate metabolism	0.18878105249684	26	11.3
http://www.genome.jp/dbget-bin/www_bget?ko04940	Type I diabetes mellitus	0.17906842287637	27	11.2
http://www.genome.jp/dbget-bin/www_bget?ko05340	Primary immunodeficiency	0.15863303443048	8	12.3
http://www.genome.jp/dbget-bin/www_bget?ko05152	Tuberculosis	0.15430685080307	29	11.4
http://www.genome.jp/dbget-bin/www_bget?ko05134	Legionellosis	0.14189627282888	28	11.2
http://www.genome.jp/dbget-bin/www_bget?ko03018	RNA degradation	0.088409246481617	29	11.4
http://www.genome.jp/dbget-bin/www_bget?ko00051	Fructose and mannose metabolism	0.087385683214052	27	11.1
http://www.genome.jp/dbget-bin/www_bget?ko00280	Valine, leucine and isoleucine degradation	0.06284845491534	26	11.3
http://www.genome.jp/dbget-bin/www_bget?ko00030	Pentose phosphate pathway	0.059475153688065	27	10.6
http://www.genome.jp/dbget-bin/www_bget?ko00190	Oxidative phosphorylation	0.03063403416373	28	11.4
http://www.genome.jp/dbget-bin/www_bget?ko00710	Carbon fixation in photosynthetic organisms	0.018619571474591	28	10.7
http://www.genome.jp/dbget-bin/www_bget?ko00061	Fatty acid biosynthesis	0.0091971570069616	26	11.3
http://www.genome.jp/dbget-bin/www_bget?ko00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis		2	9.1
http://www.genome.jp/dbget-bin/www_bget?ko00604	Glycosphingolipid biosynthesis - ganglio series		2	10.4
http://www.genome.jp/dbget-bin/www_bget?ko05150	Staphylococcus aureus infection		1	10.7

Comparison of taxonomic composition



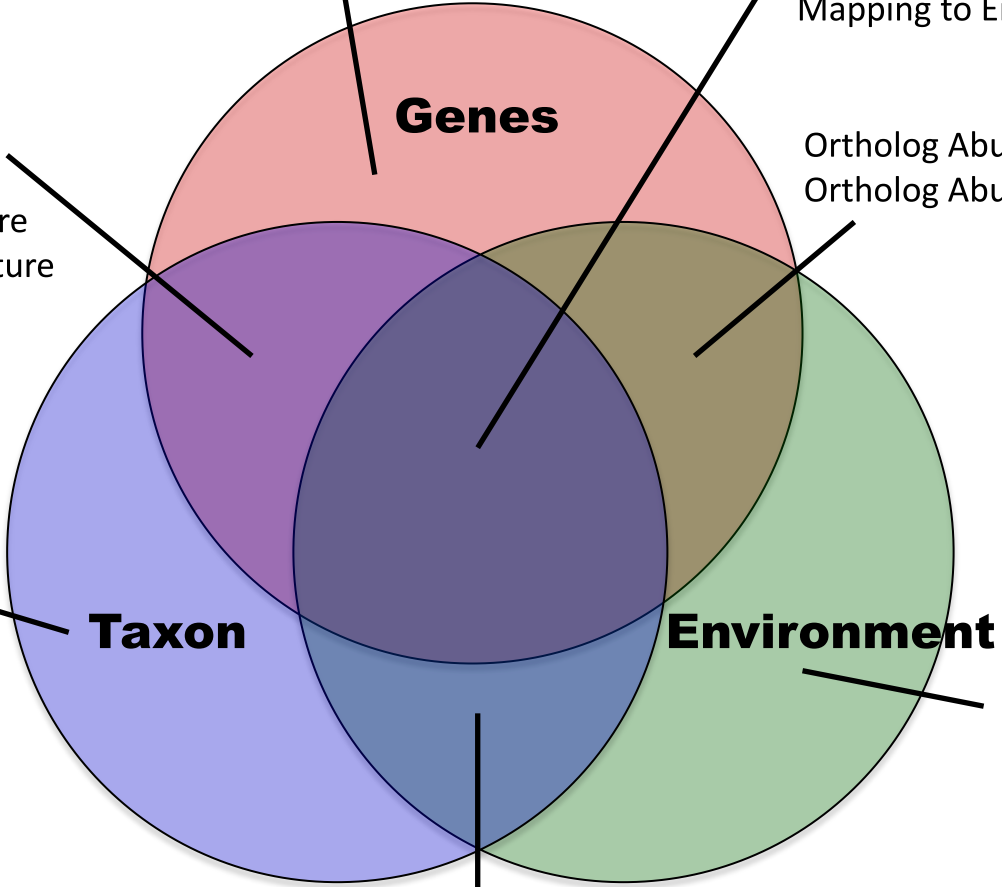
Stanza categories in MicrobeDB.jp

Gene Definition
Gene Publication
Ortholog Definition

Sample Function
Mapping to Environment (Chromosome)
Mapping to Environment (Plasmid)

Gene Annotation
Ortholog Group Members
Ortholog Cluster
Genome Information
GTPS Gene/Genome Feature
RefSeq Gene/Genome Feature

Ortholog Abundance among Environments
Ortholog Abundance in Environment

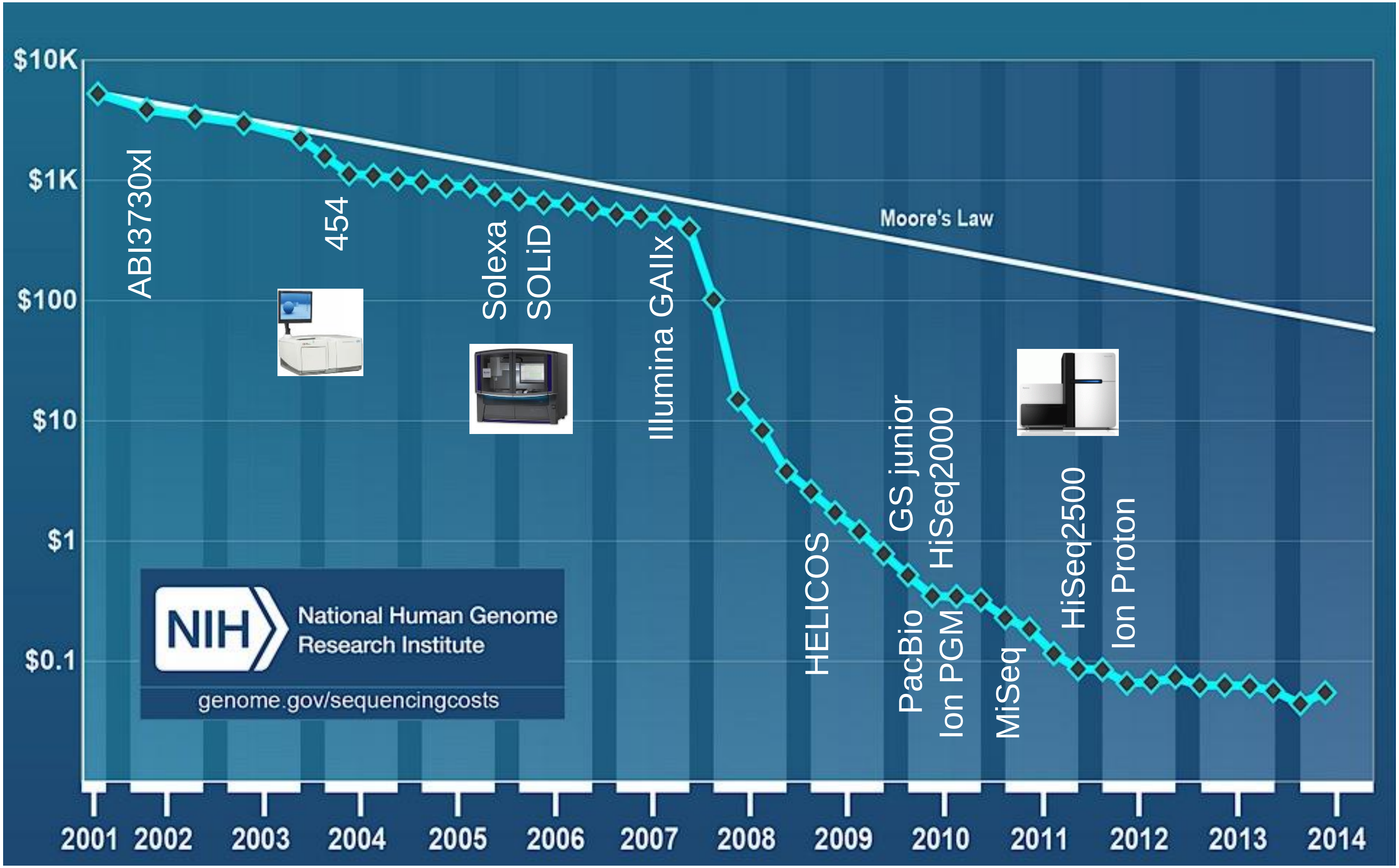


GTPS Genome
GTPS Genome Definition
Other Collection Numbers
Pathogen Information
Phenotype Information
RefSeq Genome
RefSeq Genome Definition
Strain Definition
Strain Genome
Strain Reference
Taxon Definition
Taxon Hierarchy

Disease Definition
Environment Definition
MEO Hierarchy
MEO Ontology View
Meta16S Sample List
Metagenome Sample List
Numeric Metadata Histogram
Sample Definition
Sample Metadata
SRS Cross Reference
Symptom List

Genome-Sequenced Strains
Sequenced Genome List
Strain List
Taxonomic Composition of Genomes
Taxonomic Composition of Meta 16S
Human Meta Body Mapping
Strain Metadata

Cost per Raw Megabase of DNA Sequences



Data from the NHGRI Genome Sequencing Program (GSP)

http://www.genome.gov/sequencingcosts/

Microbe DB .JP

MicrobeDB.jp Project Team

- Ken Kurokawa (Tokyo Institute of Technology)
  Hiroshi Mori, Junichi Takehara, Koji Yoshino, Shinya, Suzuki,
  Nozomi Yamamoto, Takuji Yakada

- Yasukazu Nakamura (National Institute of Genetics, DDBJ)
  Takatomo Fujisawa, Eri Kaminuma, Hideaki Sugawara

- Ikuo Uchiyama (National Institute for Basic Biology)
  Hirokazu Chiba, Hiroyo Nishide

Advisor (DataBase Center for Life Science)
Shinobu Okamoto, Shuichi Kawashima, Toshiaki Katayama,
Yasunori Yamamoto, Shoko Kawamoto

Funding

NBDC
National Bioscience Database Center

JST