November 29, 2019
統合データベース講習会：AJACS世田谷

微生物統合データベース MicrobeDB.jpの活用法

森宙史
Hiroshi Mori
国立遺伝学研究所
情報研究系

MicrobeDB.jp portal beta version is available.

http://microbedb.jp/

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

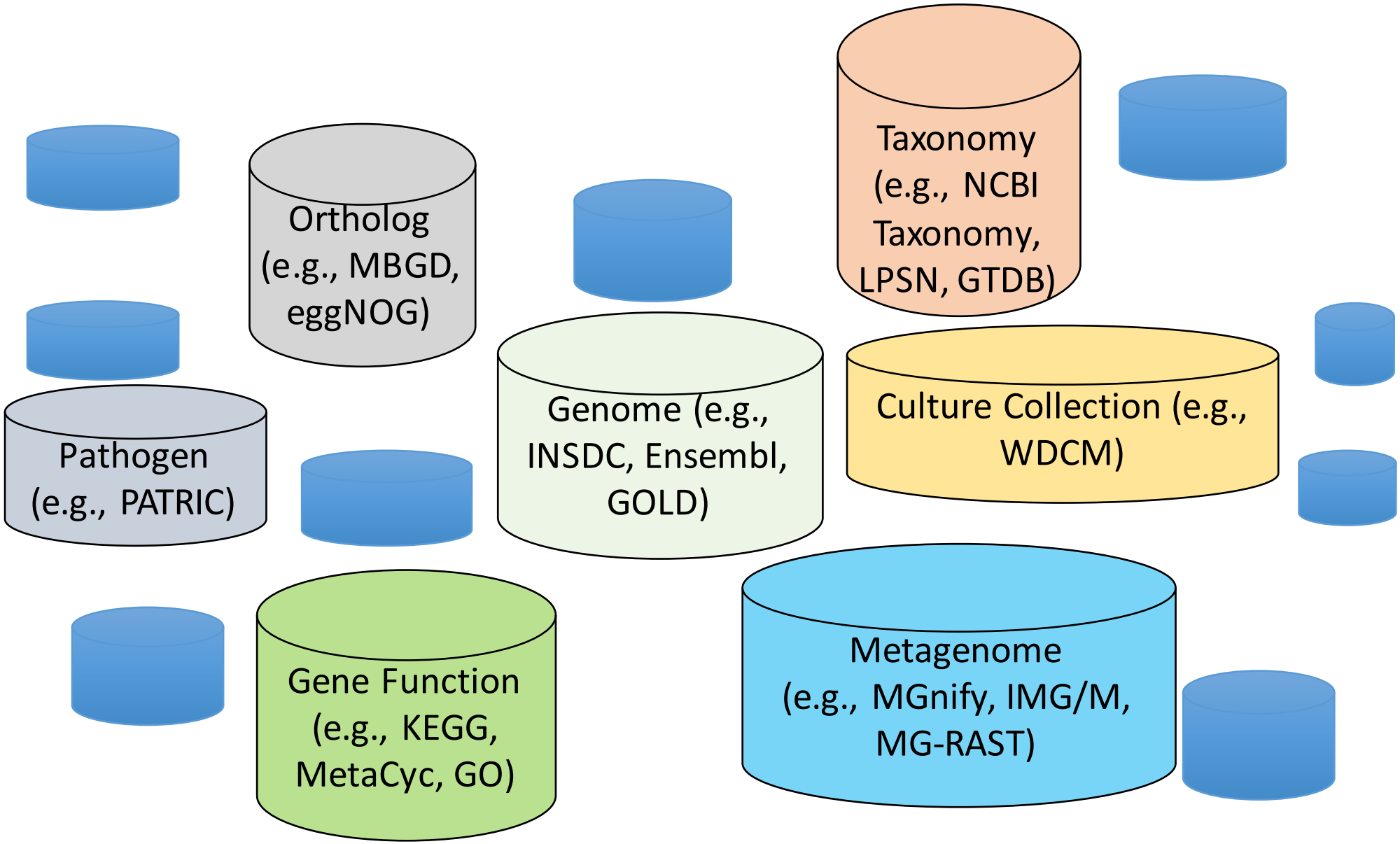
Tokyo Institute of Technology: (Metagenome)
Takuji Yamada, Zenichi Nakagawa

Chiba University: (Fungal & Bacterial culture collection info.)
Hiroki Takahashi, Takashi Yaguchi

Technical adviser:
DBCLS (especially Shuichi Kawashima, Toshiaki Katayama)

Funding JST 科学技術振興機構 NBDC

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

- 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

数%ぐらいの菌しか
培養できない

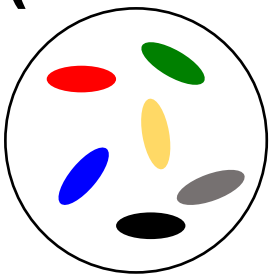
What's metagenomics?

- **Microflora, microbiota, microbial community**: 微生物群集
  Total collection of microorganisms within a community

- **Metagenome**: ある群集の遺伝情報の総体
  Total genomic potential of a community
  [Handelsman et al. 1998, Chem. & Biol.]

- **Microbiome**: マイクロバイオーム
  Microbiota and metagenome in a microbial community

amplicon sequencing analysis
(アンプリコン解析, 16S rRNA遺伝子のアンプリコン解析)



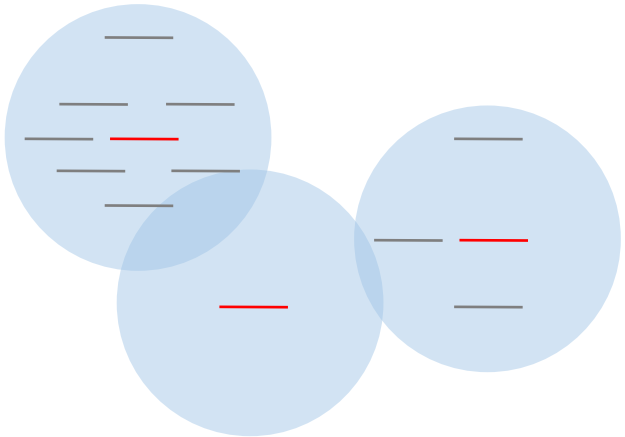
DNA extraction

PCR amplification

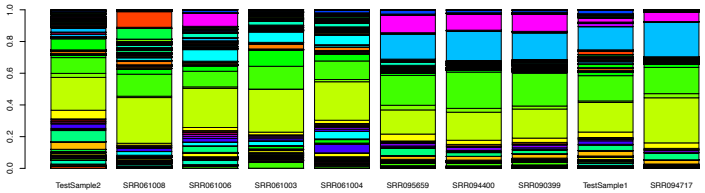
DNA Sequencing

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering or denoising



Taxonomic assignment and
Comparison between samples



Who's there?

Metagenomic sequencing analysis
(メタゲノム解析, ショットガンメタゲノム解析)

DNA extraction

DNA Sequencing

Metagenomic reads

Assemble — MEGAHIT, metaSPAdes, etc.

Contig sets

Gene finding — Prodigal, MetaGeneMark, etc.

Gene sets

Sequence similarity search — BLASTP, DIAMOND, MMseqs2, etc.

KEGG (Kyoto Encyclopedia of Genes and Genomes)

UniProt

Gene Function abundance
Pathway abundance
Pathway reconstruction
Taxonomic abundance

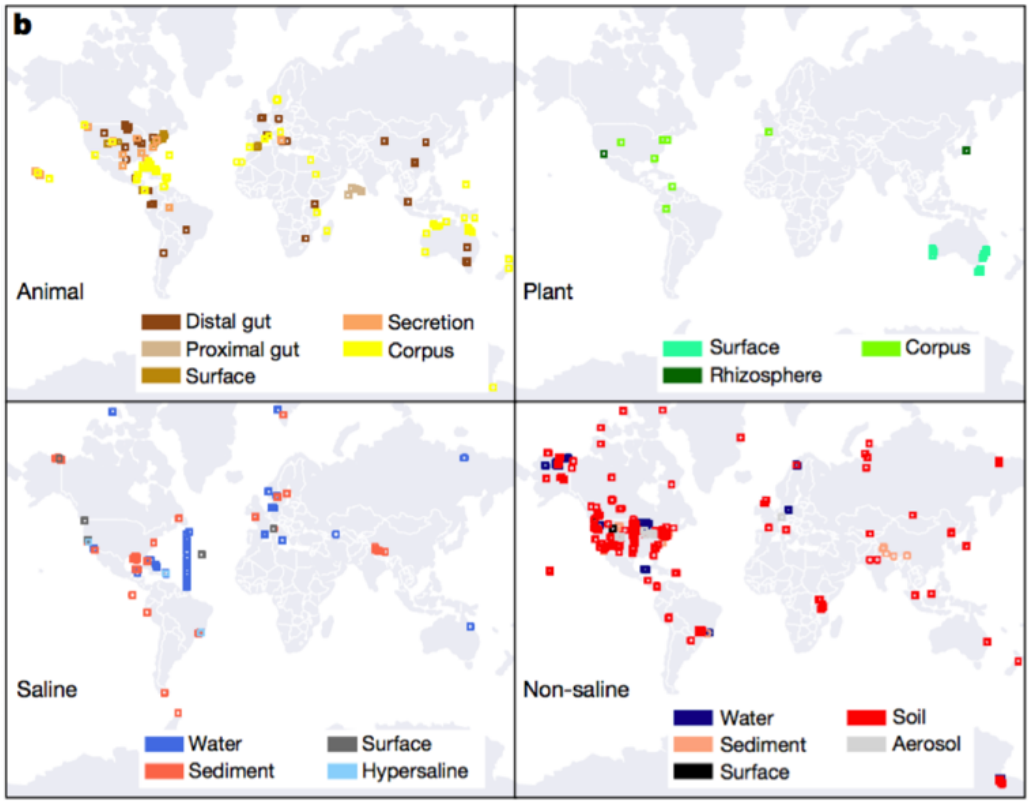
Sample1 Metadata

Sample2 Metadata

Comparative metagenomics

Who's there?

What are they doing?



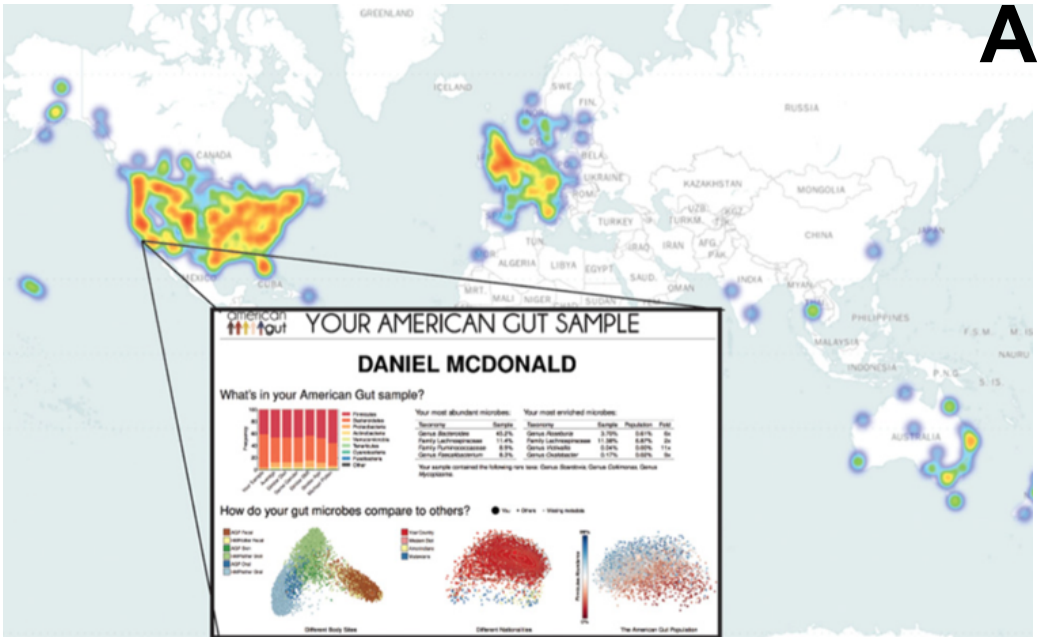
Earth Microbiome Project

Thompson et al. 2017, Nature

Soil, Sea, Pond, Animal, etc.

23,828 samples

Standardized experimental and bioinformatics procedure



American Gut Project

McDonald et al. 2018, mSystems

Human feces

9,511 samples

Standardized experimental and bioinformatics procedure.

Cloud founding

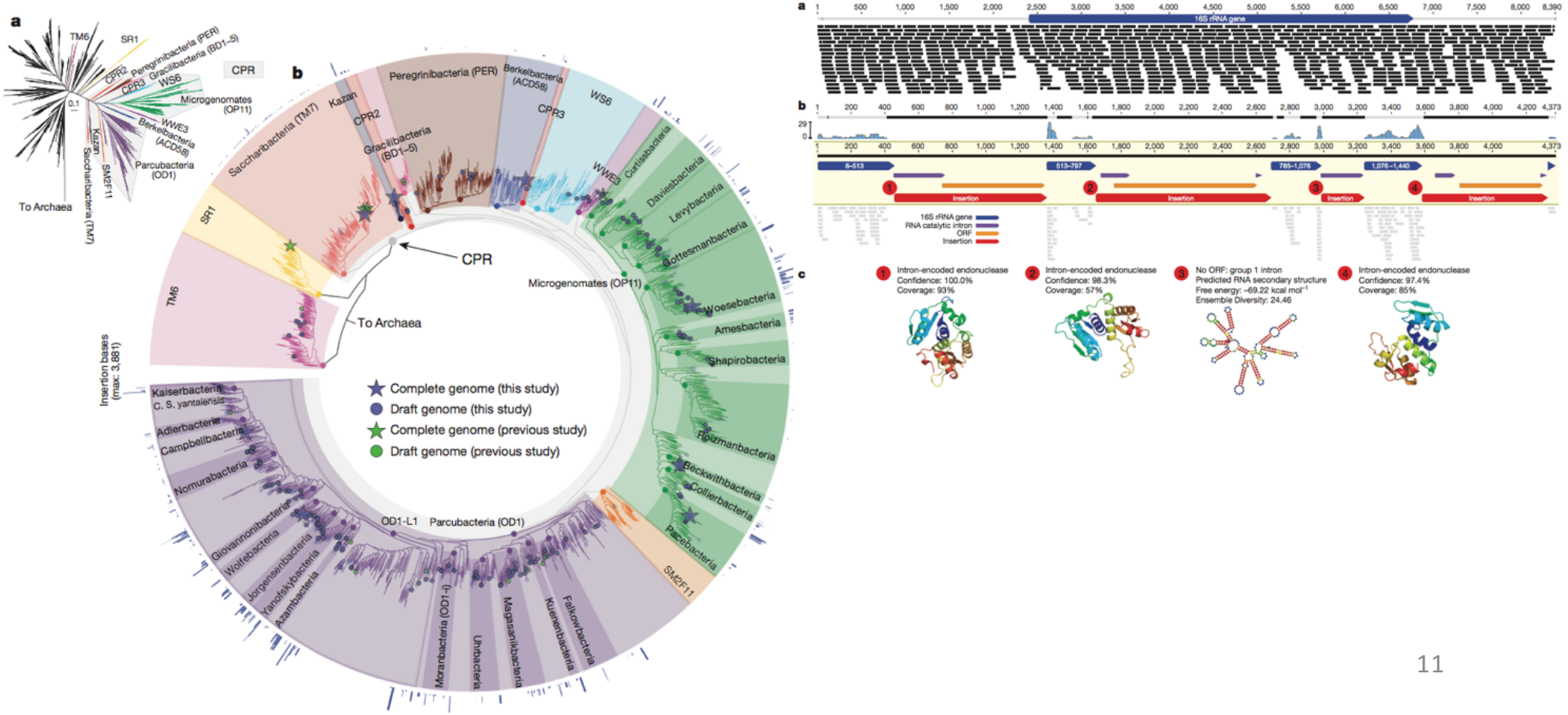
LETTER

Nature, 2015

doi:10.1038/nature14486

Unusual biology across a group comprising more than 15% of domain Bacteria

Christopher T. Brown[1], Laura A. Hug[2], Brian C. Thomas[2], Itai Sharon[2], Cindy J. Castelle[2], Andrea Singh[2], Michael J. Wilkins[3,4], Kelly C. Wrighton[4], Kenneth H. Williams[5] & Jillian F. Banfield[2,5,6]



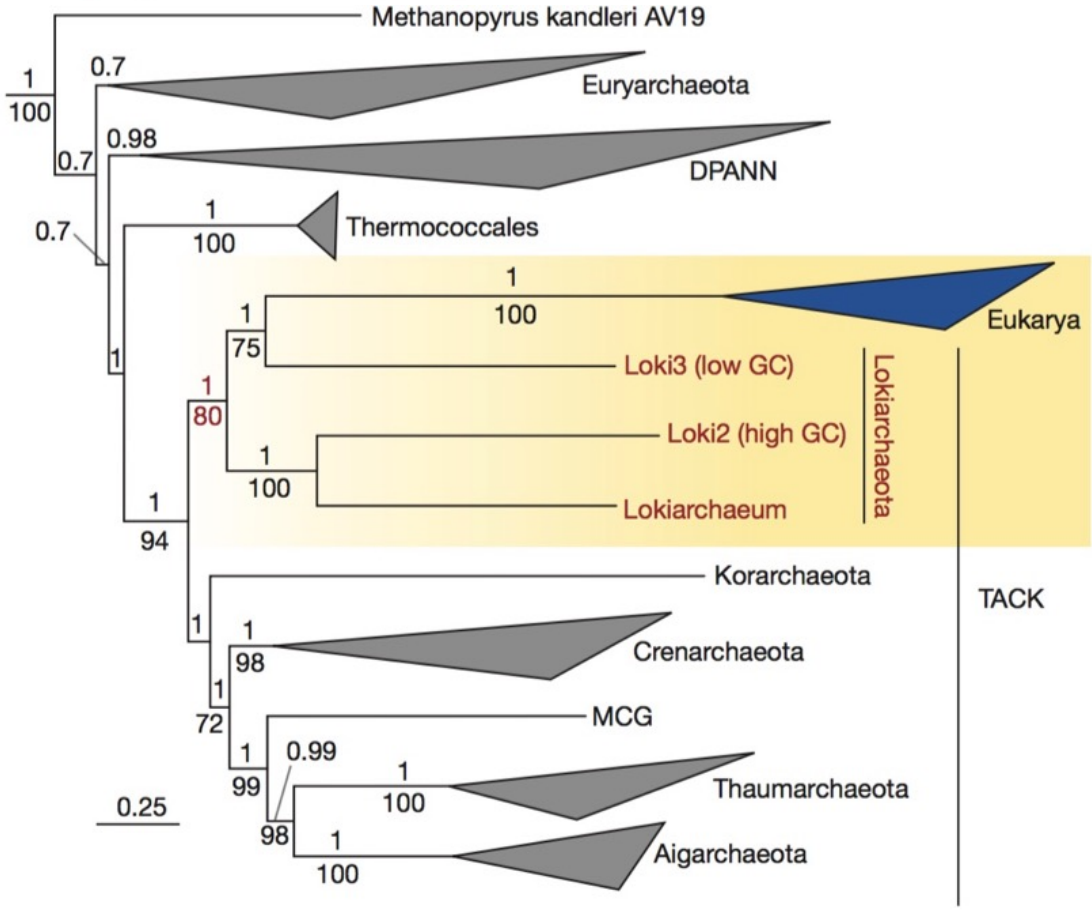
ARTICLE

Nature, 2015

doi:10.1038/nature14447

Complex archaea that bridge the gap between prokaryotes and eukaryotes

Anja Spang[1]*, Jimmy H. Saw[1]*, Steffen L. Jørgensen[2]*, Katarzyna Zaremba-Niedzwiedzka[1]*, Joran Martijn[1], Anders E. Lind[1], Roel van Eijk[1]†, Christa Schleper[2,3], Lionel Guy[1,4] & Thijs J. G. Ettema[1]



メタゲノム解析の情報解析手法については、
https://github.com/AJACS-training/AJACS79

Microbiome data in public DB

	Admin	Database URL	Sequence data	Separate amplicon and shotgun?	Taxa	Function	Number of samples in Sep. 2019
INSDC DRA/ERA/SRA	Japan, EU, USA		○	×	△	×	>1,600,000
GOLD	JGI, USA	https://gold.jgi.doe.gov/	×	×	×	×	50,821
IMG/M	JGI, USA	https://img.jgi.doe.gov/cgi-bin/m/main.cgi	○	×	○	○	16,170
MG-RAST	Chicago U. USA	http://www.mg-rast.org/	○	○	○	○	392,514
MGnify	EBI, EU	https://www.ebi.ac.uk/metagenomics/	×	○	○	○	176,360
MicrobeDB.jp v.2	NIG, Japan	http://microbedb.jp/MDB/	×	○	○	○	60,551

1次データベースであるINSDC DRA/ERA/SRAには
系統組成・遺伝子機能組成のデータが無い。
メタデータもまともに構造化されておらず整理されていない

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

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

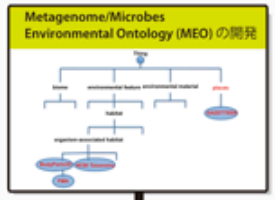
Microbe DB .JP

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



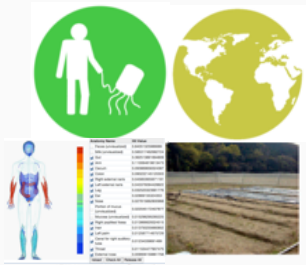
Gene | Taxon | Environment

Ortholog: MBGD

オーソログデータ

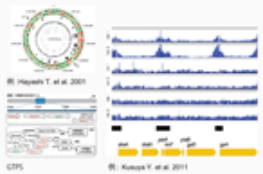
Taxonomy:
NCBI Taxonomy

系統分類データ



Metadata:
INSDC DRA

環境のメタデータ



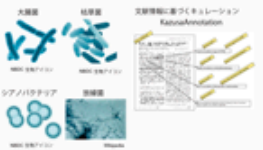
Genome: RefSeq

オミックスデータ

Culture Collection:
NBRC/JCM

Metagenome:
INSDC DRA

メタゲノムデータ

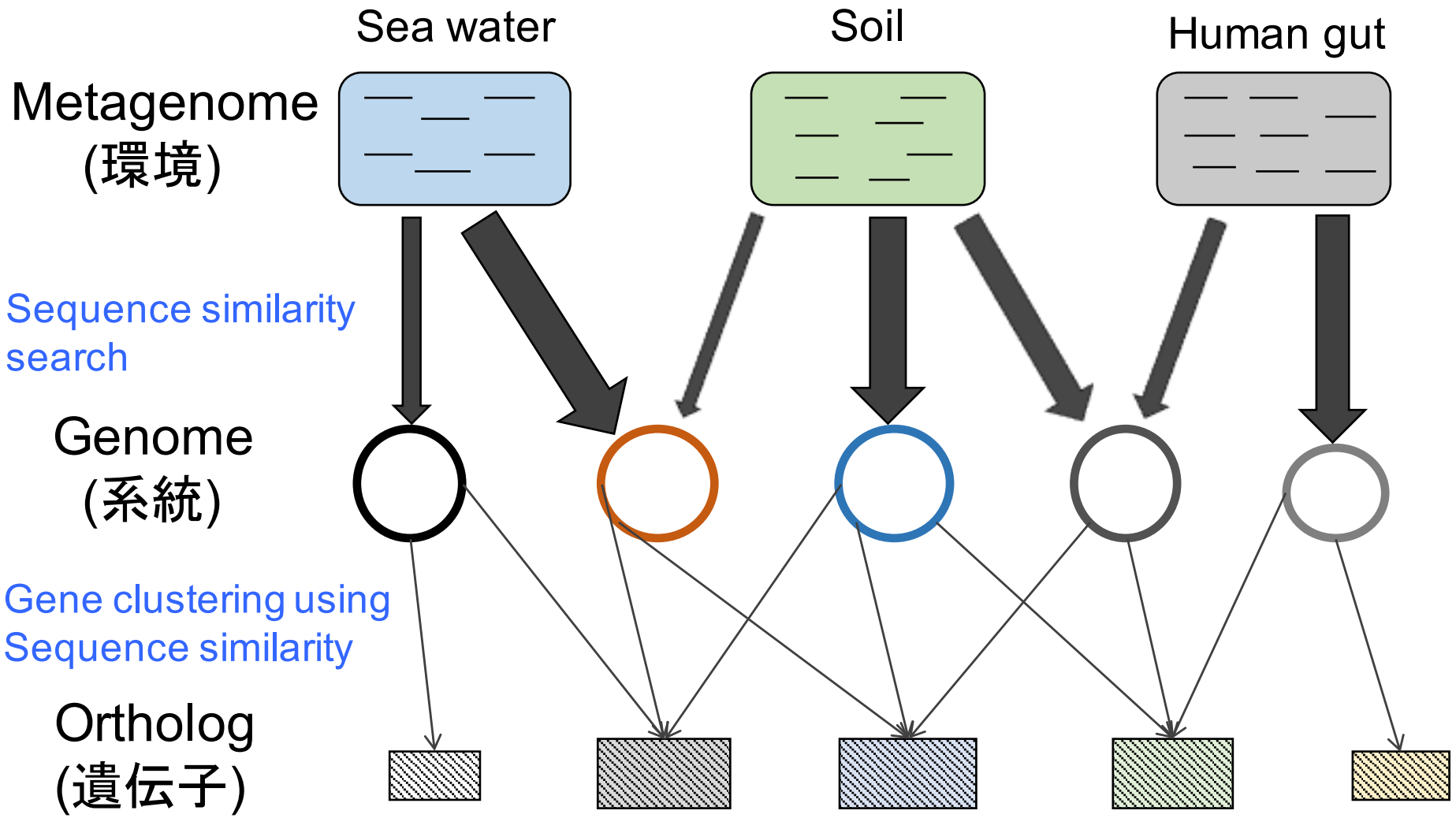


モデル微生物の高品質アノテーションデータ

菌株データ

Red color indicates our collaborators.

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)

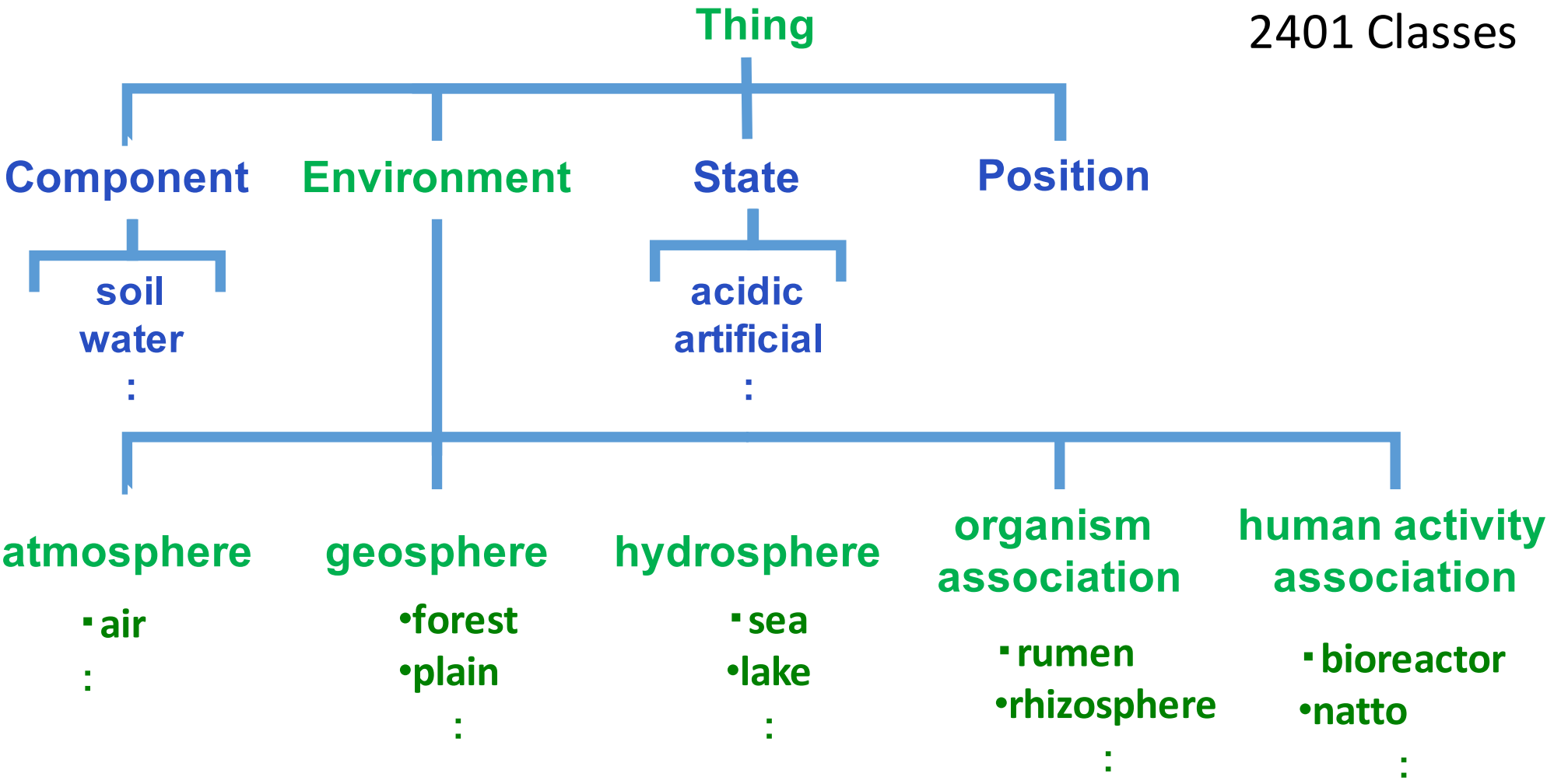
MPO (Microbial Phenotype Ontology)

MBGD Ontology

PDO (Pathogenic Disease Ontology)

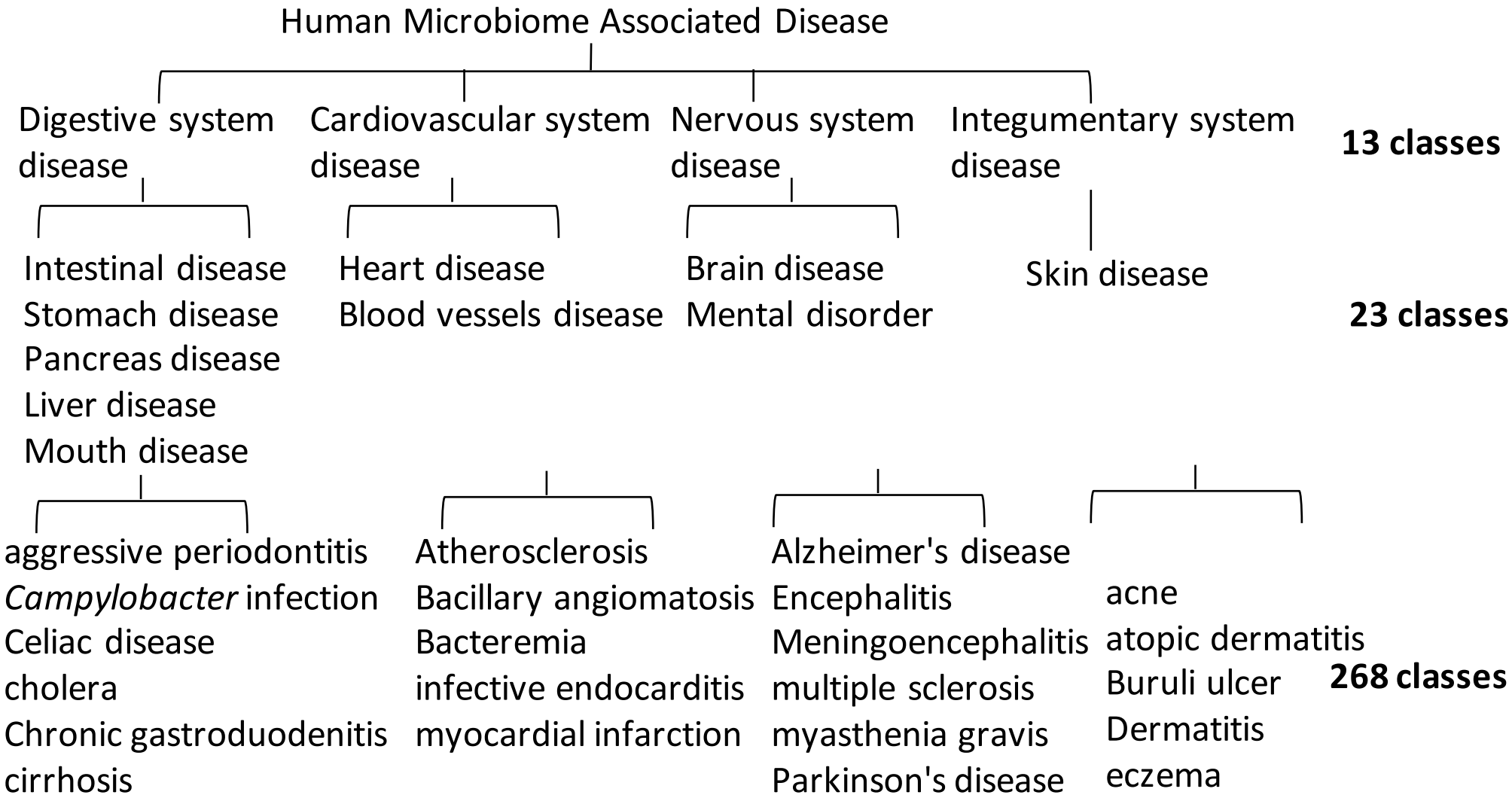
Most of them can be obtained from

Metagenome/Microbes Environmental Ontology (MEO) Ver. 0.9.2



意味・概念をDB内で統一してオントロジーとして語彙を定義

Human Microbiome Associated Disease Ontology (HMADO)



感染症なのか否かなど、微生物群集が関係するヒトの病気の分類

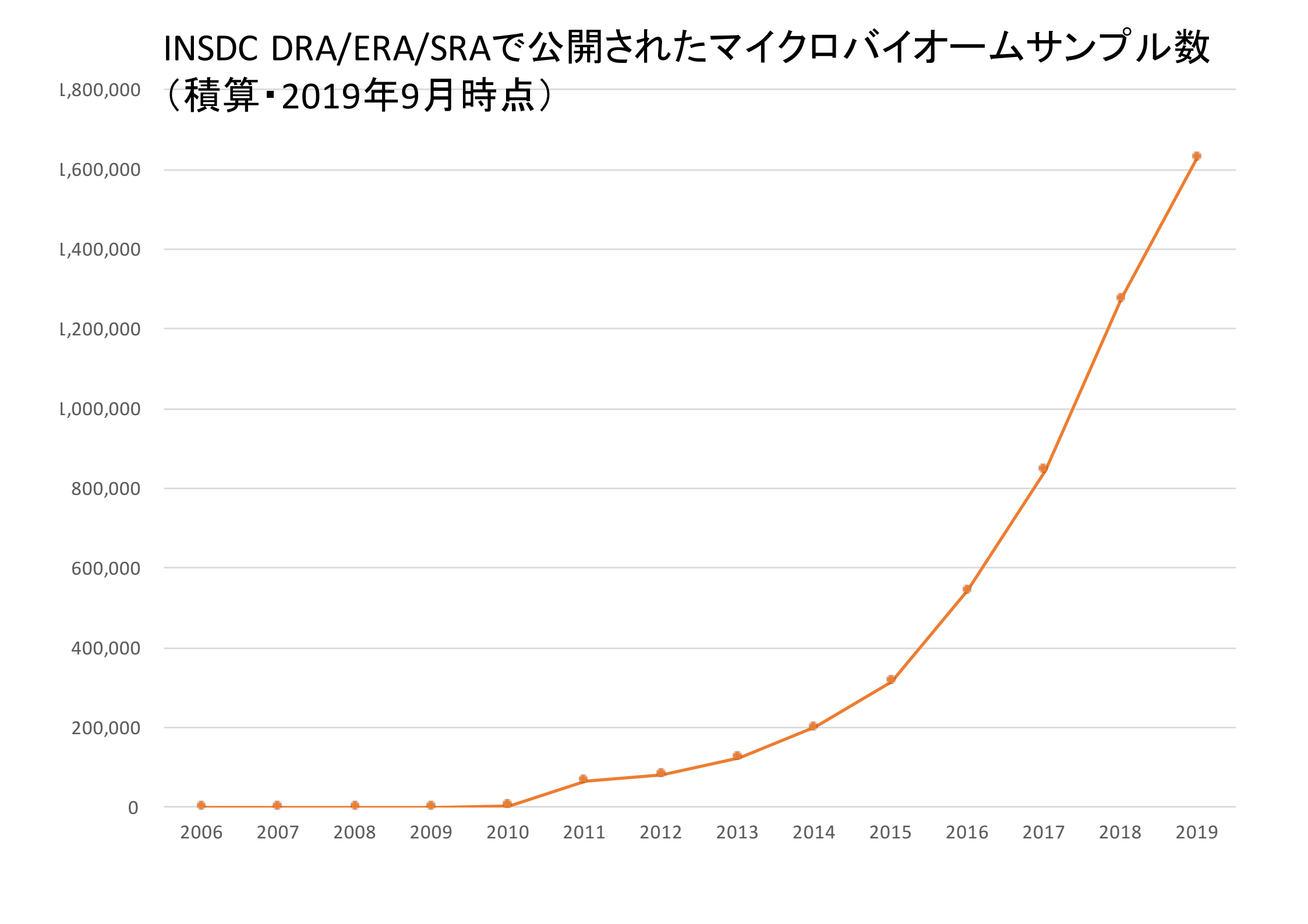
MicrobeDB.jp version 3 data

Data category	Data sources	Ontologies
Genome Metadata	INSDC BioSample, NCBI Assembly report	MEO, HMADO, NCBITAX
Ortholog	MBGD	ORTH
Culture collection	JCM (RDF-Portal)	MCCV, MPO, MEO, PDO, CSSO
Culture collection	NBRC (RDF-Portal)	MCCV, MPO
Metagenome	INSDC DRA	NCBITAX, KEGG Orthology
Metagenome metadata	INSDC BioSample	MEO, HMADO, SIO

MicrobeDB.jp version 3 data

Data category	Number of entry
Genome metadata (from RefSeq)	290,208 genomes
Ortholog cluster data (from MBGD)	375,228 clusters
Culture collection strain data from JCM/NBRC (RDF-Portal)	38,414 strains
Microbiome metadata (from INSDC DRA)	1,631,611 samples
Microbiome taxonomic composition data	96,766 samples
Microbiome functional composition data	4,784 samples

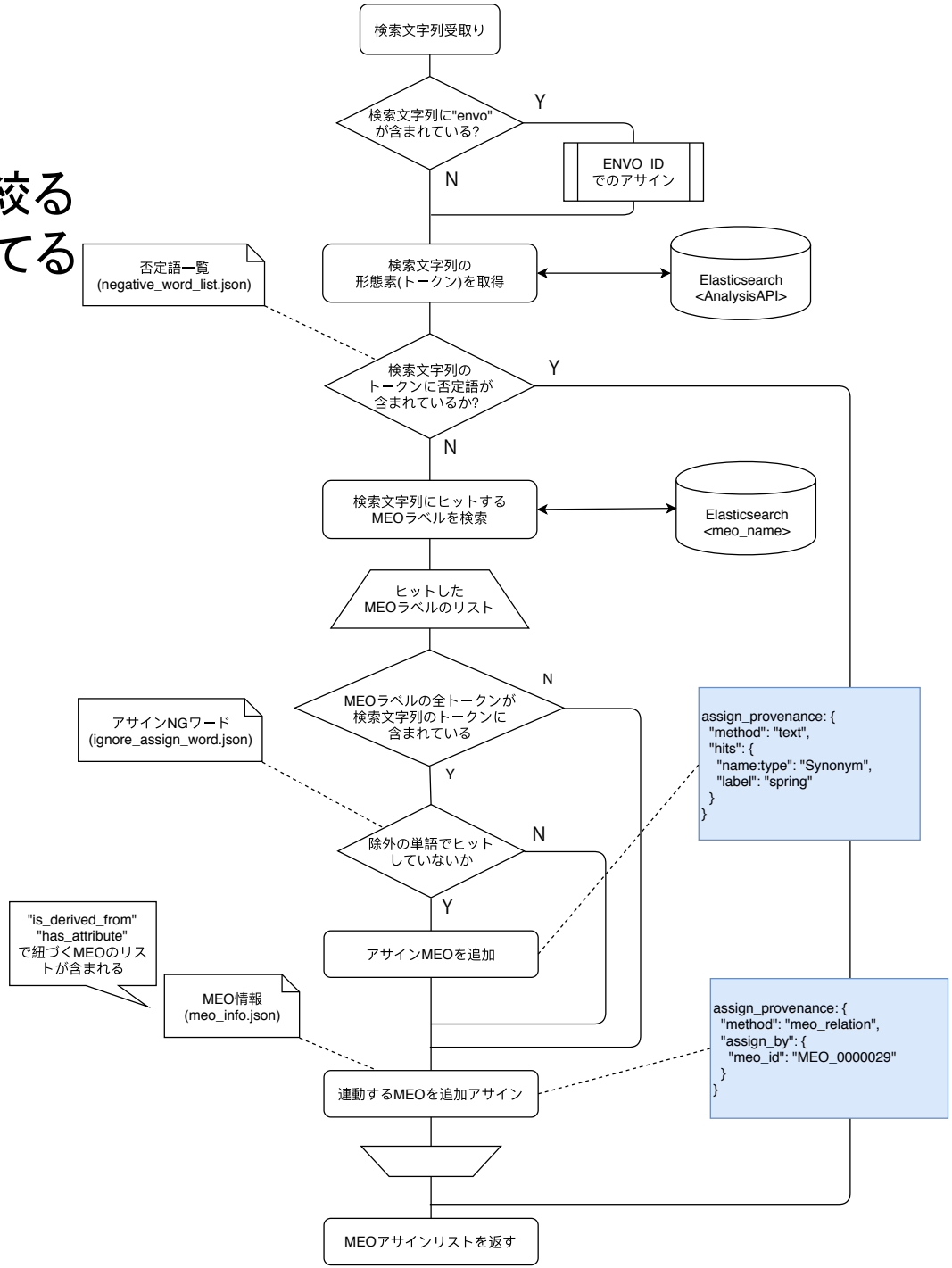
Ver. 2(2014)と比べてゲノムは約20倍、メタゲノムは約10倍の数に


INSDC DRA/ERA/SRAで公開されたマイクロバイオームサンプル数
（積算・2019年9月時点）
1,800,000
1,600,000
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

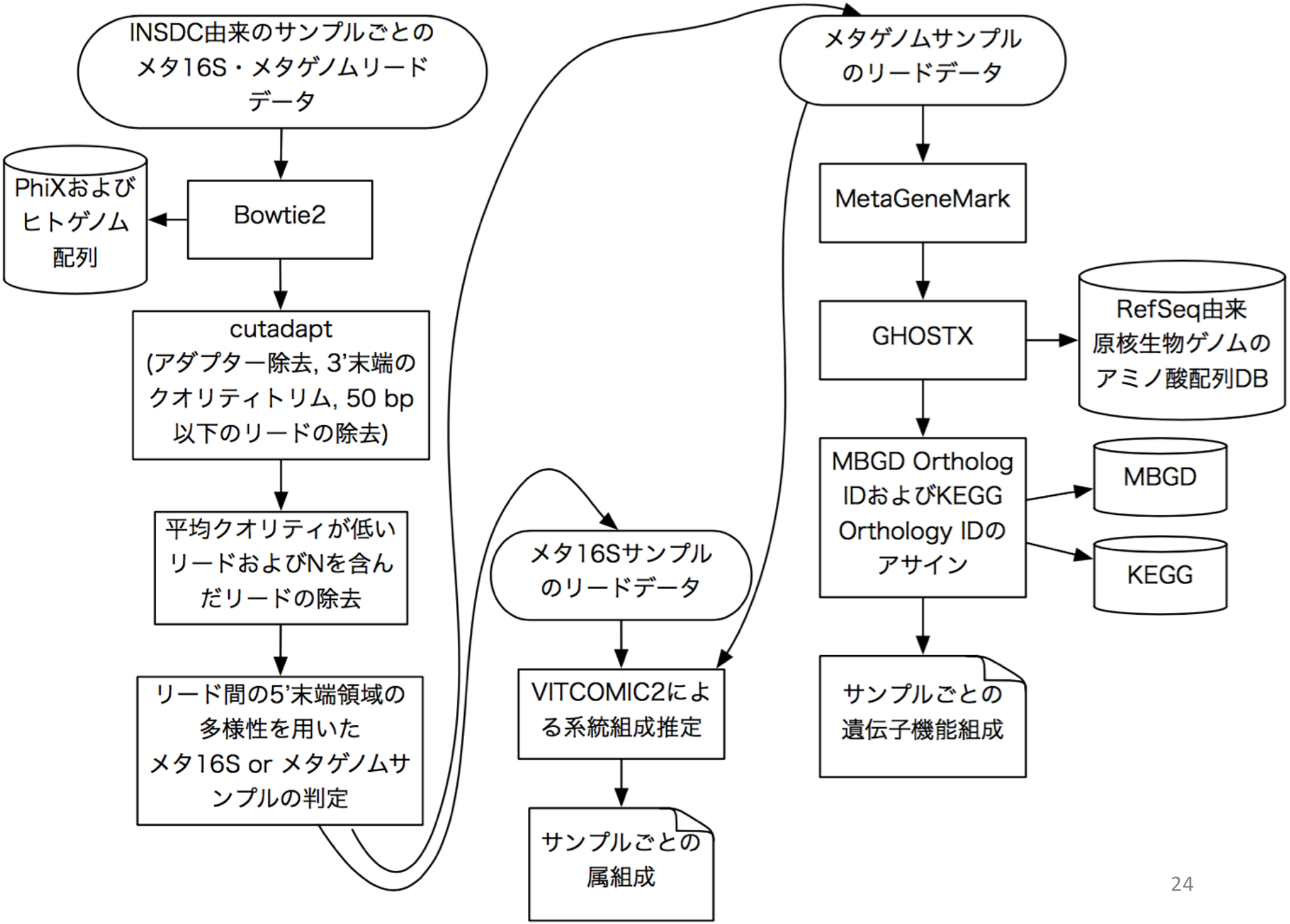
非構造Textデータからの MEOの自動アサイン

- 数的にマイナーなものは諦める
- attribute name(harmonized name)を絞る
- オントロジーのラベル、synonymをあてる
- 否定語対策
- stemmingの有無
- 例外ワードリストの作成
- より下位概念のアサイン

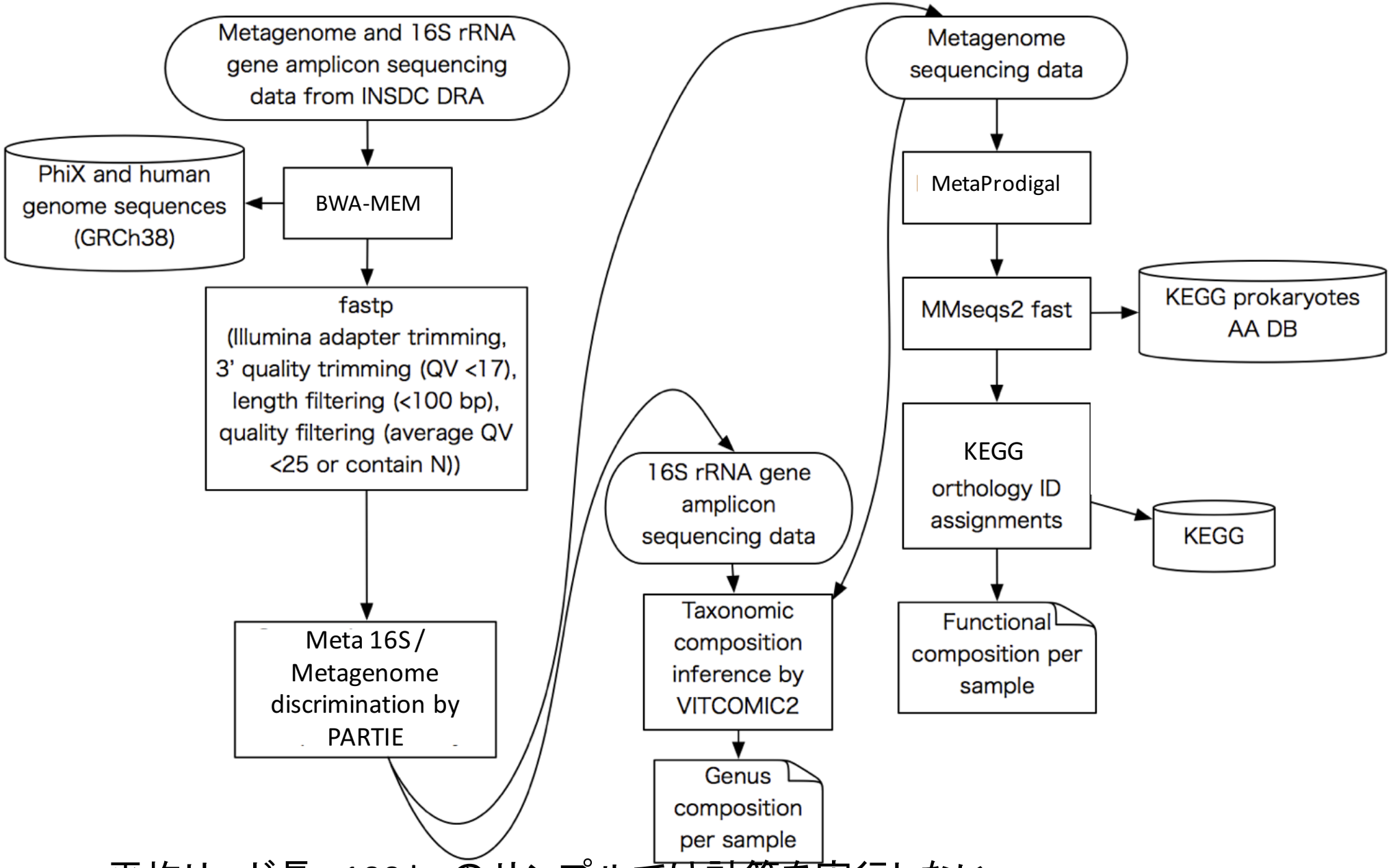
6割のゲノム、
9割のメタゲノムのBioSampleに
MEOがアサインされた



MeGAP2 (MicrobeDB.jp ver. 2のパイプライン)



MeGAP3 (MicrobeDB.jp ver. 3のパイプライン)



平均リード長 <100 bpのサンプルでは計算を実行しない。
16S rRNAアンプリコンについては、結果のファイルサイズでフィルタリング、
メタゲノムについては、アサインされたKOの種類数 <100のサンプルを除外

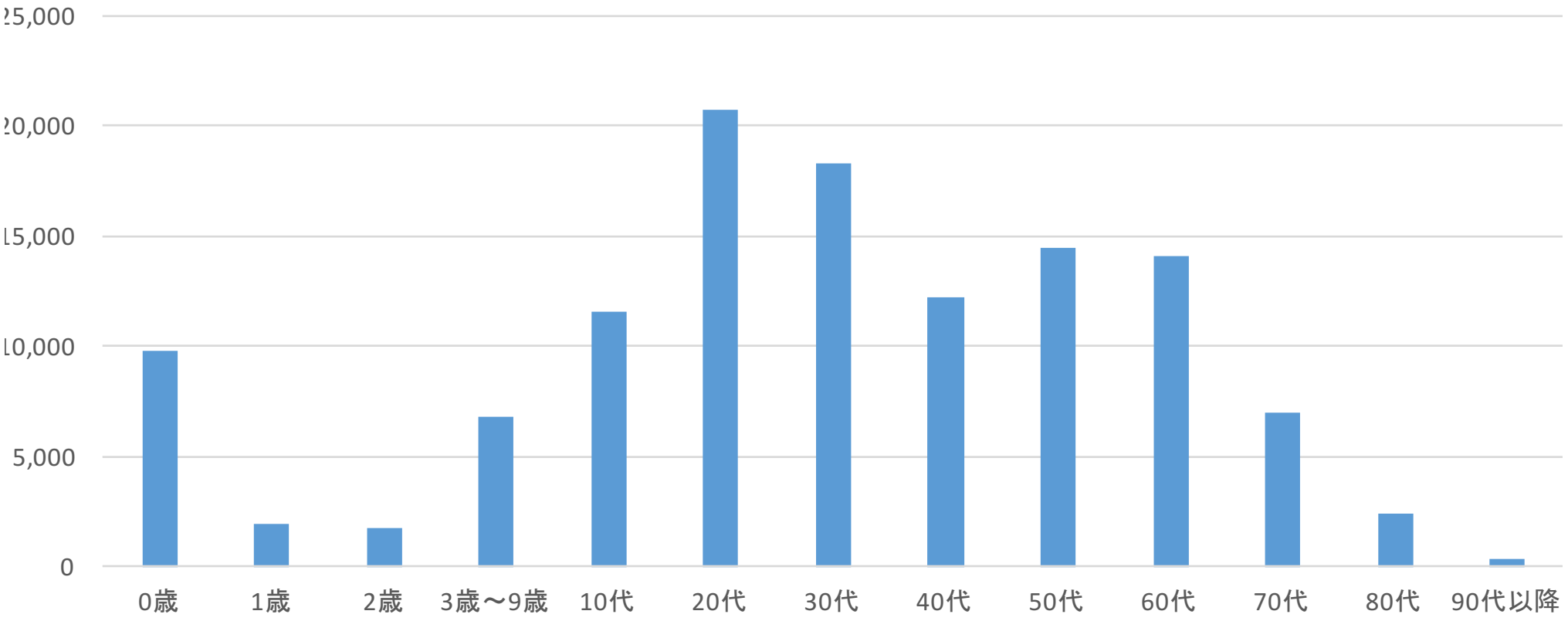
INSDC DRA/ERA/SRAで公開されたマイクロバイオームサンプル数
（2019年9月時点）

	サンプル数
マイクロバイオーム	約　1,600,000
ヒト	約　420,000
マウス	約　95,000
土壌	約　240,000
水環境	約　200,000
人工環境	約　140,000
植物共生	約　120,000

160万サンプルはヒトや土壌等マイクロバイオームサンプルの総数

ヒトマイクロバイオームについて、特に詳細にメタデータをキュレーション

	サンプル数
ヒト	約　420,000
ヒト (年齢情報あり)	約　120,000



ヒトマイクロバイオームについて、特に詳細にメタデータをキュレーション

ヒトの体の部位	サンプル数
gut (feces)	243,958
oral cavity	55,990
respiratory system	45,655
skin	41,381
vagina	21,761

- 抗生物質投与（約2-3万サンプル）
- Probiotics投与（約5千サンプル）
- Ethnicity（約5千サンプルを、Ethnicity ontologyとGAZETTEERでアノテーション）

INSDC DRA/ERA/SRAで公開された
疾患患者のマイクロバイオーム
サンプル数（2019年9月時点）

サンプル数が100以上ある疾患は
約50疾患

サンプル数が1000以上ある疾患が
右表の18疾患
（肥満と未熟児を除く）

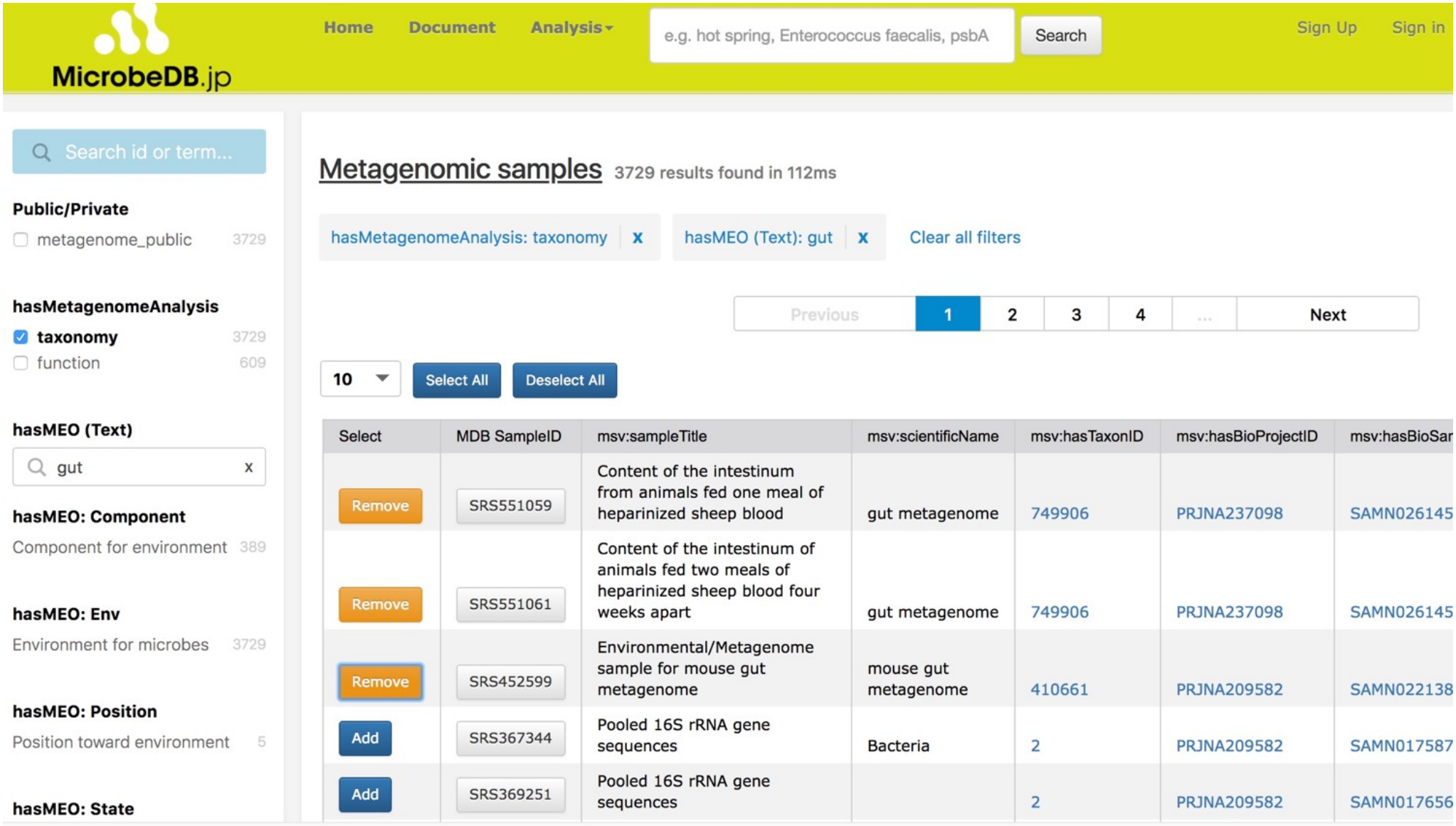
大多数が16S rRNA遺伝子の
アンプリコン解析

疾患名 (HMADO)	サンプルサイズ
IBD	37,094
皮膚炎	13,219
にきび	7,488
喘息	6,178
乾癬	5,151
大腸炎	4,453
嚢胞性線維症	4,040
アレルギー	3,902
下痢	2,684
大腸がん	2,085
細菌性膣炎	2,030
アテローム性動脈硬化	1,968
赤痢	1,918
パーキンソン病	1,760
腺がん	1,742
統合失調症	1,712
歯周炎	1,555
虫歯	1,370

MicrobeDB.jpの使い方

詳しくは、
https://togotv.dbcls.jp/20190902.html


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

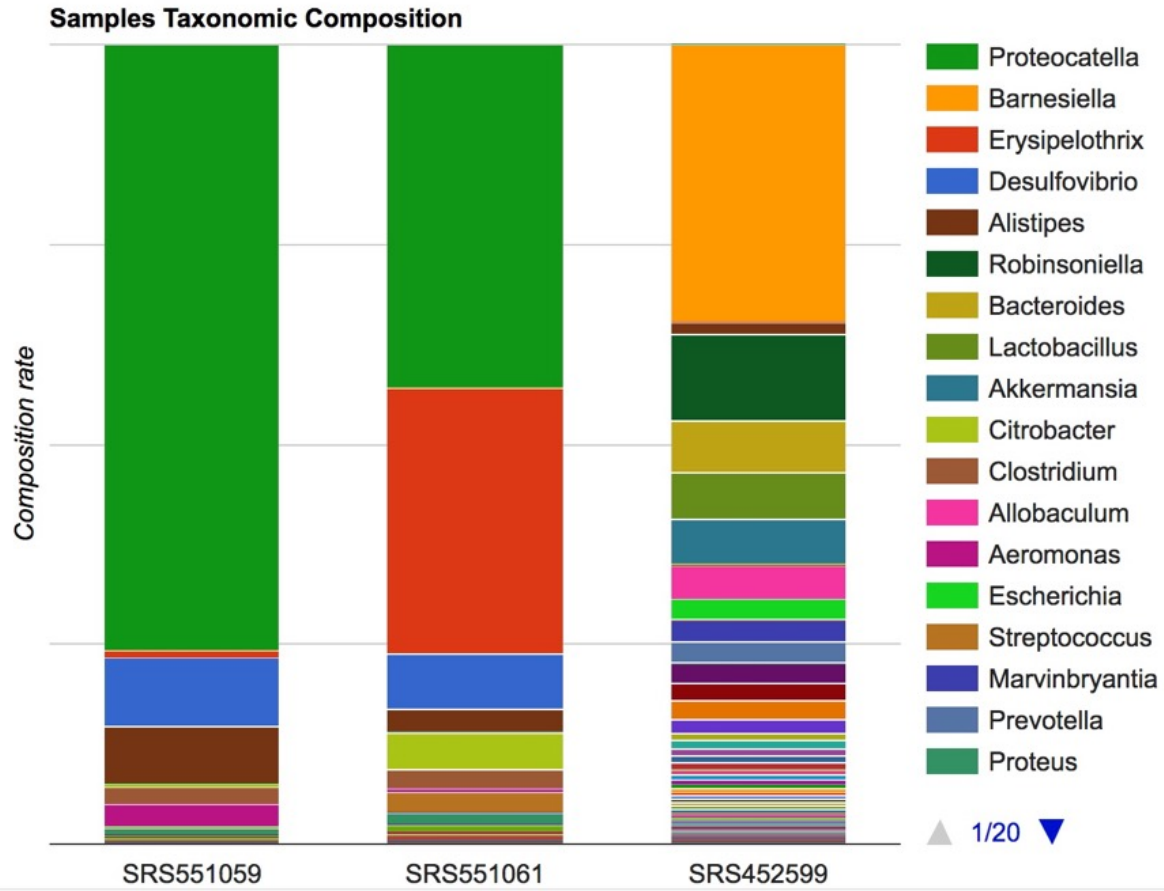

MicrobeDB.jp
Home
Document
Analysis
e.g. hot spring, Enterococcus faecalis, psbA
Search
Sign Up
Sign in
Search id or term...
Public/Private
metagenome_public 3729
hasMetagenomeAnalysis
taxonomy 3729
function 609
hasMEO (Text)
gut
hasMEO: Component
Component for environment 389
hasMEO: Env
Environment for microbes 3729
hasMEO: Position
Position toward environment 5
hasMEO: State
Metagenomic samples 3729 results found in 112ms
hasMetagenomeAnalysis: taxonomy
hasMEO (Text): gut
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

MicrobeDB.jp

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

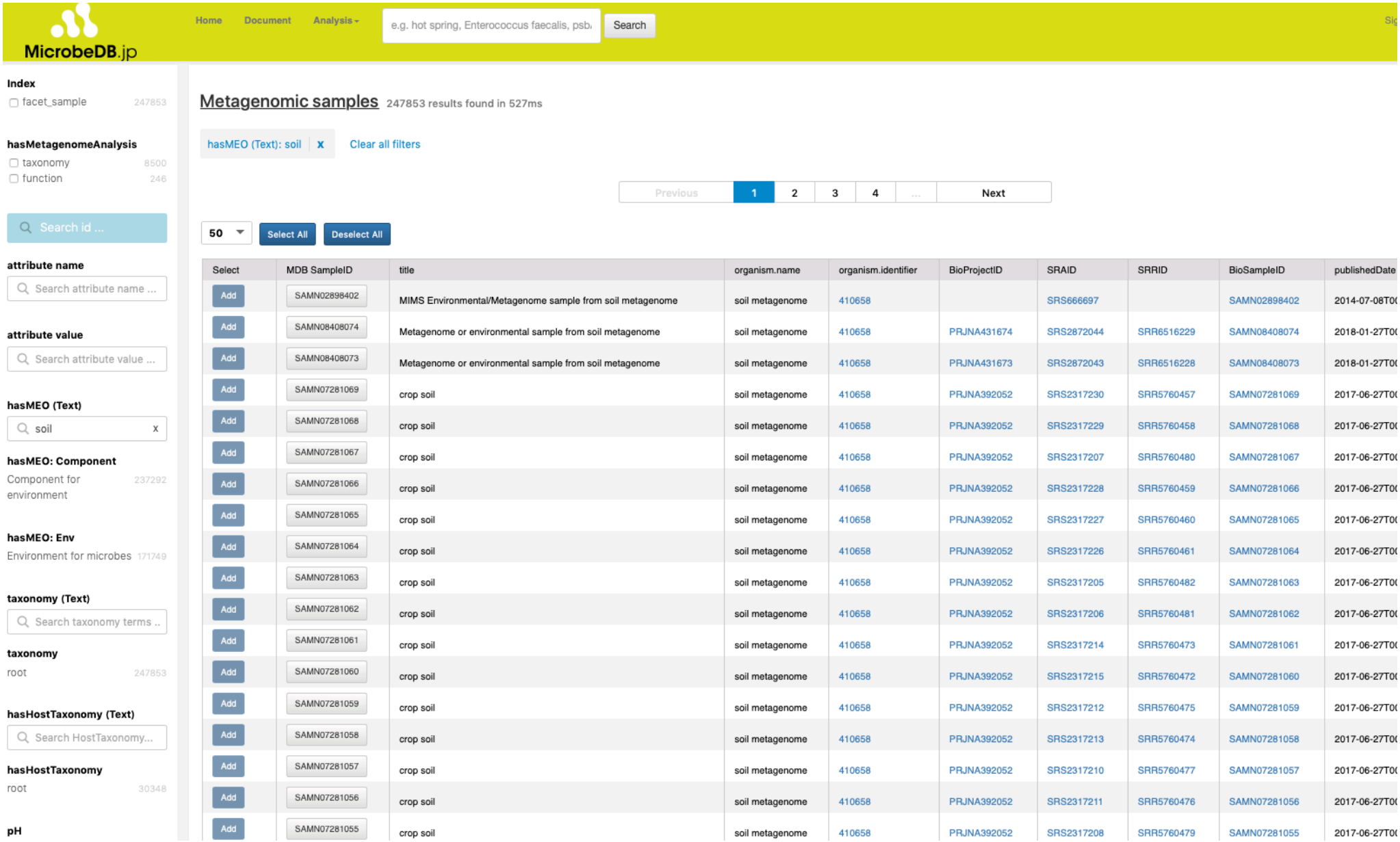
Taxonomic composition (bar) | Taxonomic composition (heatmap) | Diversity index | Hierarchical clustering | PCoA | Functional composition (bar)

Functional composition (heatmap)



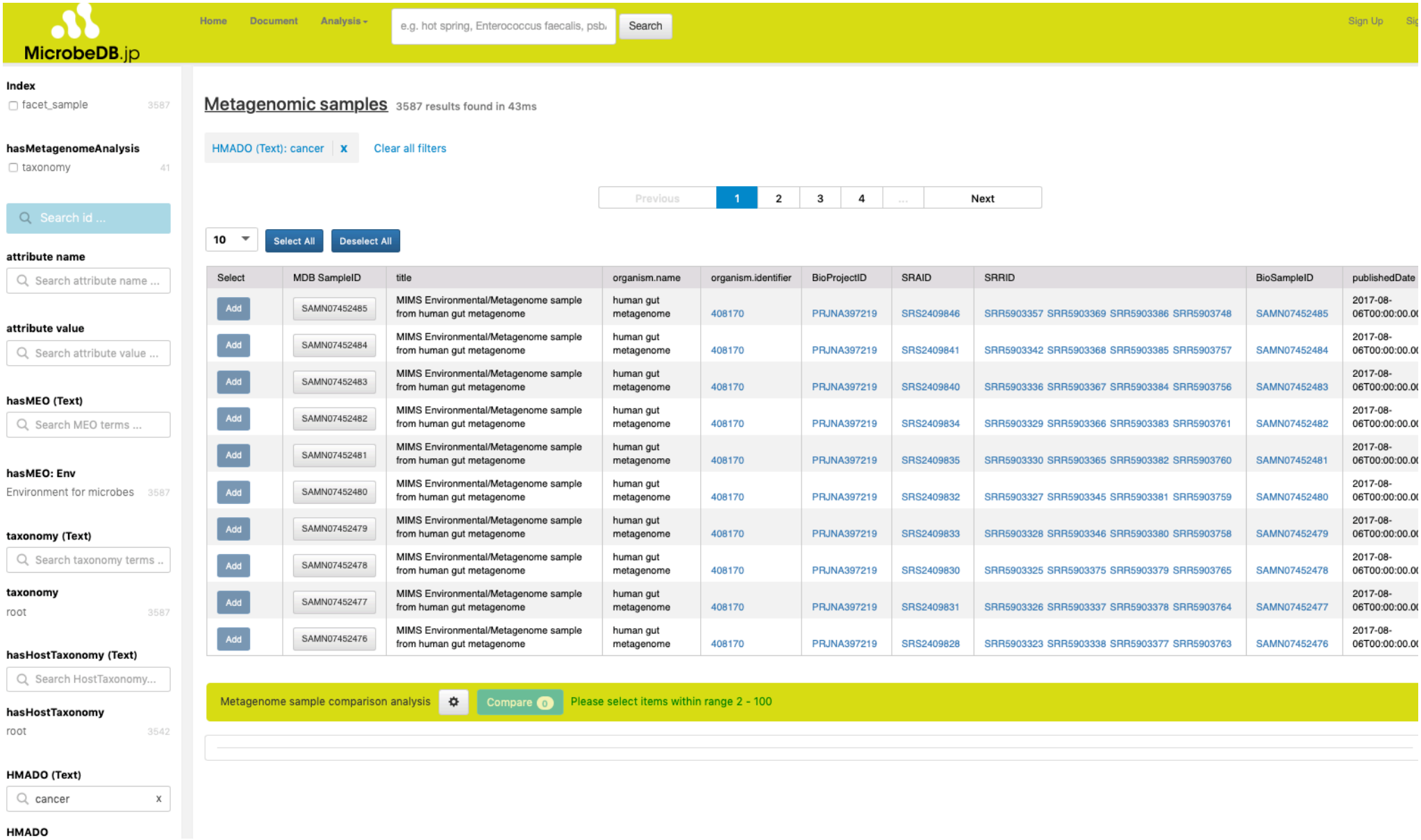
ここからversion 3の新機能
（12月中に切り替わります）

MEO(soil)由来マイクロバイオームサンプルの絞り込み検索 (これはver. 2でもできました)

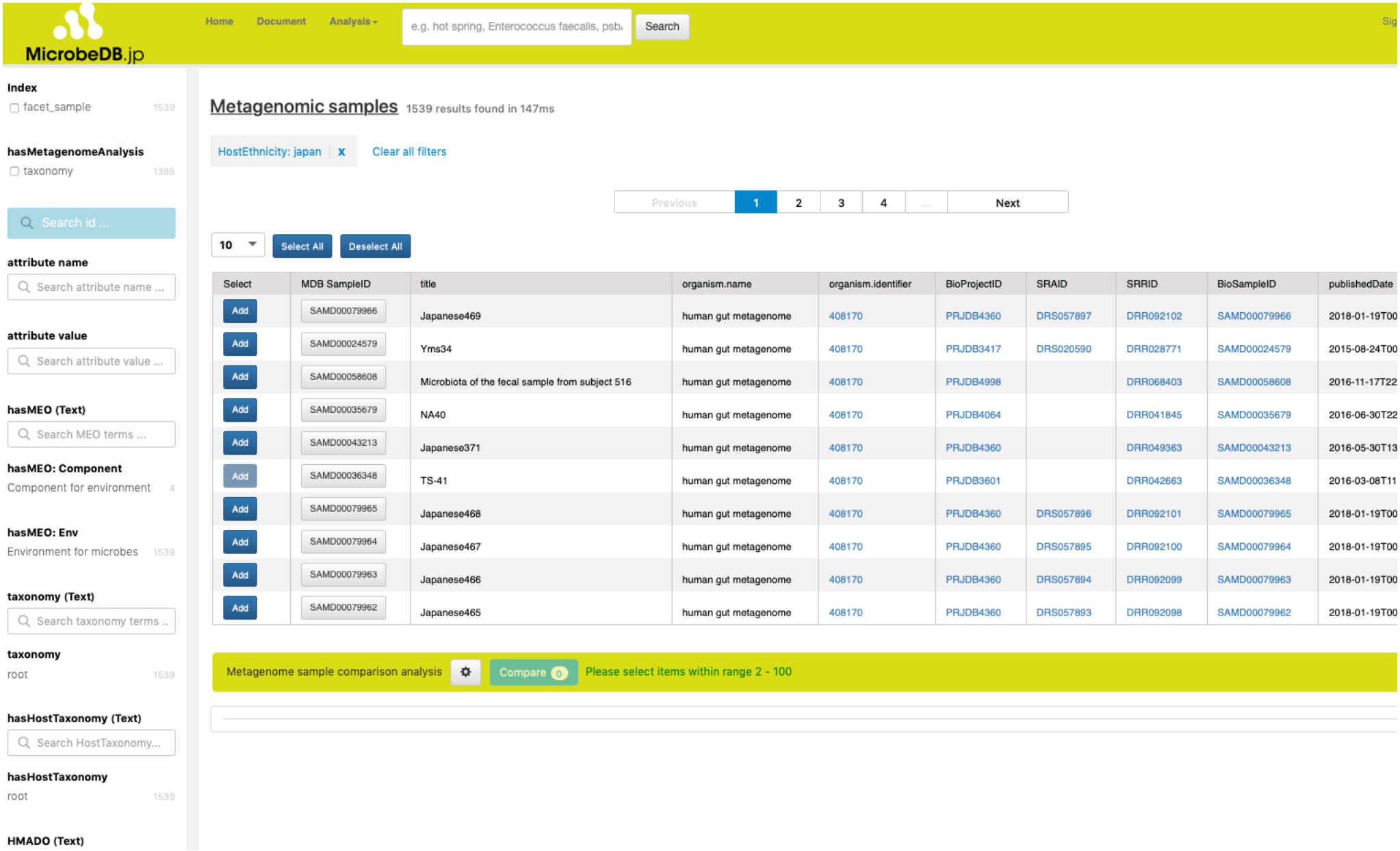


Select	MDB SampleID	title	organism.name	organism.identifier	BioProjectID	SRAID	SRRID	BioSampleID	publishedDate
Add	SAMN02898402	MIMS Environmental/Metagenome sample from soil metagenome	soil metagenome	410658		SRS666697		SAMN02898402	2014-07-08T0(
Add	SAMN08408074	Metagenome or environmental sample from soil metagenome	soil metagenome	410658	PRJNA431674	SRS2872044	SRR6516229	SAMN08408074	2018-01-27T0(
Add	SAMN08408073	Metagenome or environmental sample from soil metagenome	soil metagenome	410658	PRJNA431673	SRS2872043	SRR6516228	SAMN08408073	2018-01-27T0(
Add	SAMN07281069	crop soil	soil metagenome	410658	PRJNA392052	SRS2317230	SRR5760457	SAMN07281069	2017-06-27T0(
Add	SAMN07281068	crop soil	soil metagenome	410658	PRJNA392052	SRS2317229	SRR5760458	SAMN07281068	2017-06-27T0(
Add	SAMN07281067	crop soil	soil metagenome	410658	PRJNA392052	SRS2317207	SRR5760480	SAMN07281067	2017-06-27T0(
Add	SAMN07281066	crop soil	soil metagenome	410658	PRJNA392052	SRS2317228	SRR5760459	SAMN07281066	2017-06-27T0(
Add	SAMN07281065	crop soil	soil metagenome	410658	PRJNA392052	SRS2317227	SRR5760460	SAMN07281065	2017-06-27T0(
Add	SAMN07281064	crop soil	soil metagenome	410658	PRJNA392052	SRS2317226	SRR5760461	SAMN07281064	2017-06-27T0(
Add	SAMN07281063	crop soil	soil metagenome	410658	PRJNA392052	SRS2317205	SRR5760482	SAMN07281063	2017-06-27T0(
Add	SAMN07281062	crop soil	soil metagenome	410658	PRJNA392052	SRS2317206	SRR5760481	SAMN07281062	2017-06-27T0(
Add	SAMN07281061	crop soil	soil metagenome	410658	PRJNA392052	SRS2317214	SRR5760473	SAMN07281061	2017-06-27T0(
Add	SAMN07281060	crop soil	soil metagenome	410658	PRJNA392052	SRS2317215	SRR5760472	SAMN07281060	2017-06-27T0(
Add	SAMN07281059	crop soil	soil metagenome	410658	PRJNA392052	SRS2317212	SRR5760475	SAMN07281059	2017-06-27T0(
Add	SAMN07281058	crop soil	soil metagenome	410658	PRJNA392052	SRS2317213	SRR5760474	SAMN07281058	2017-06-27T0(
Add	SAMN07281057	crop soil	soil metagenome	410658	PRJNA392052	SRS2317210	SRR5760477	SAMN07281057	2017-06-27T0(
Add	SAMN07281056	crop soil	soil metagenome	410658	PRJNA392052	SRS2317211	SRR5760476	SAMN07281056	2017-06-27T0(
Add	SAMN07281055	crop soil	soil metagenome	410658	PRJNA392052	SRS2317208	SRR5760479	SAMN07281055	2017-06-27T0(

HMADO(cancer)由来マイクロバイオームサンプルの絞り込み検索



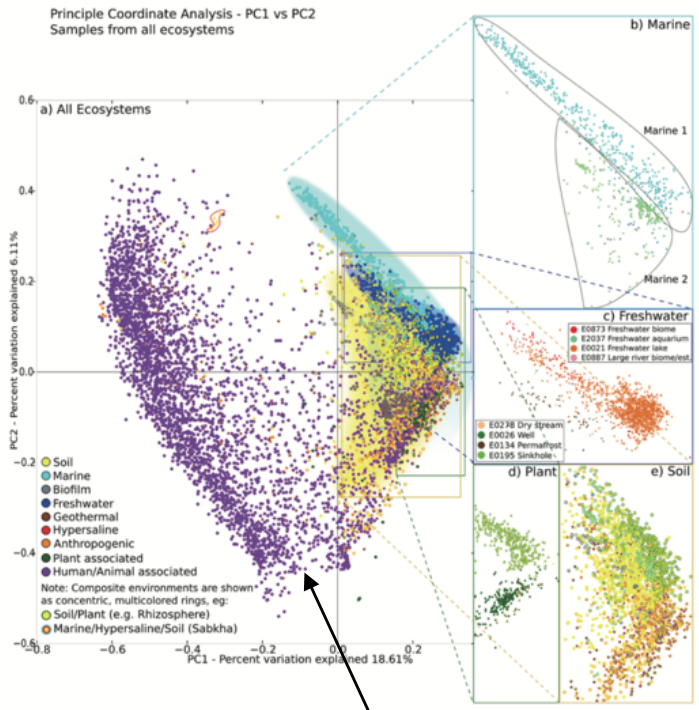
Japanese由来マイクロバイオームサンプルの絞り込み検索



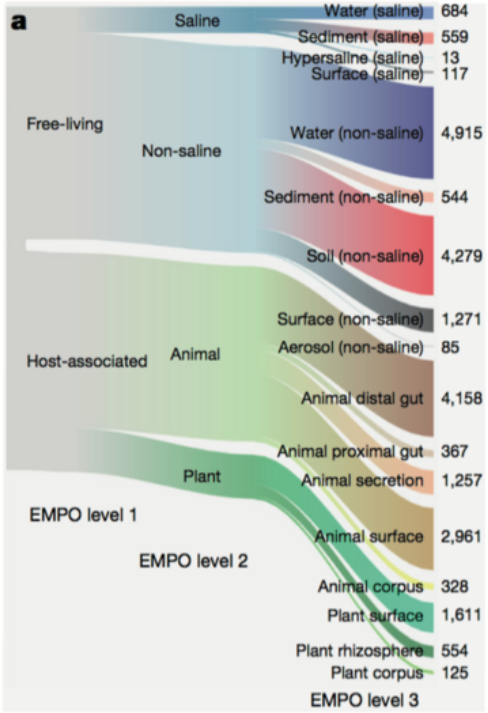
もう一つの検索アプローチ

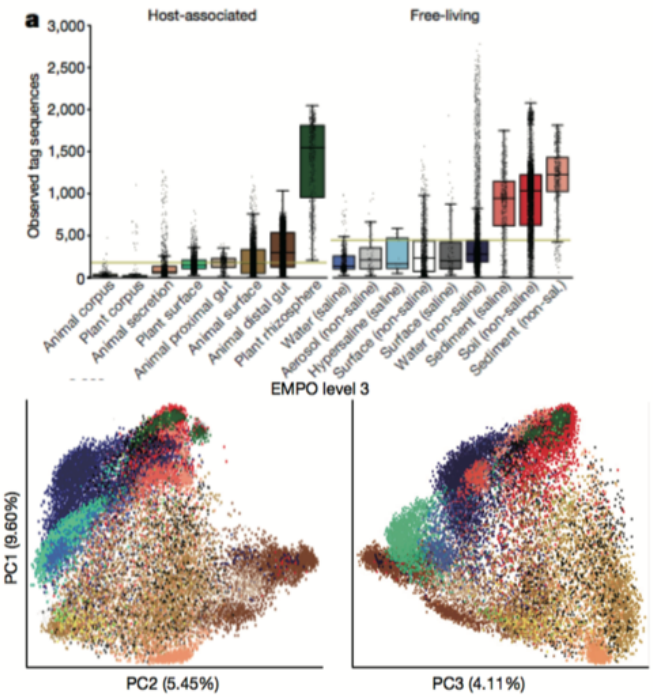
配列データで判断する

Global diversity of microbial community structures

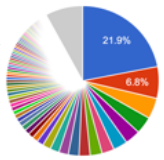


Henschel, A., Anwar, M. Z., & Manohar, V. (2015). PLoS computational biology, 11(10), e1004468.





Thompson, Luke R., et al. Nature 551.7681 (2017).



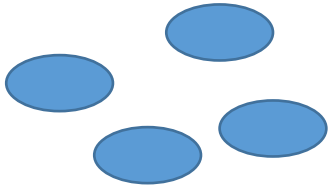
Community structure
(Bacterial & Archaeal taxonomic abundance)

Design "environmental labels" to be compared
=> Assign environmental labels to samples
=> Evaluate similarities or differences between labeled groups
But, can "environment" be discretely categorized?

All samples in DB has metadata (documents described by researchers).
Using these documents and community structure data analyzed by MicrobeDB.jp,
Correspondence between microbes and natural languages can be learned by a hierarchical Bayesian model.

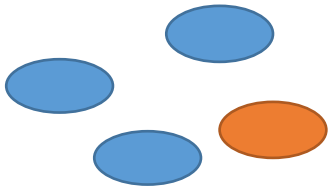
Sample #1

river microbiome.
thames river.
water samples collected from …



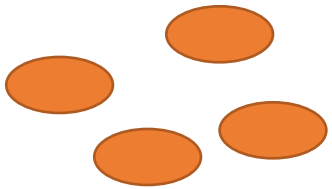
Sample #2

river microbiome.
wastewater contaminated river.



Sample #3

Activated sludge.
wastewater.
collected from the sewage treatment facility

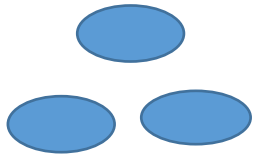


A combination of words and microbes can be extracted as a base variable of whole data set

Machine learning (Topic model)

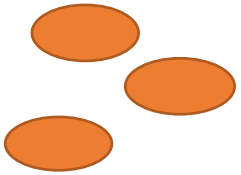
A topic representing "river"

river
water

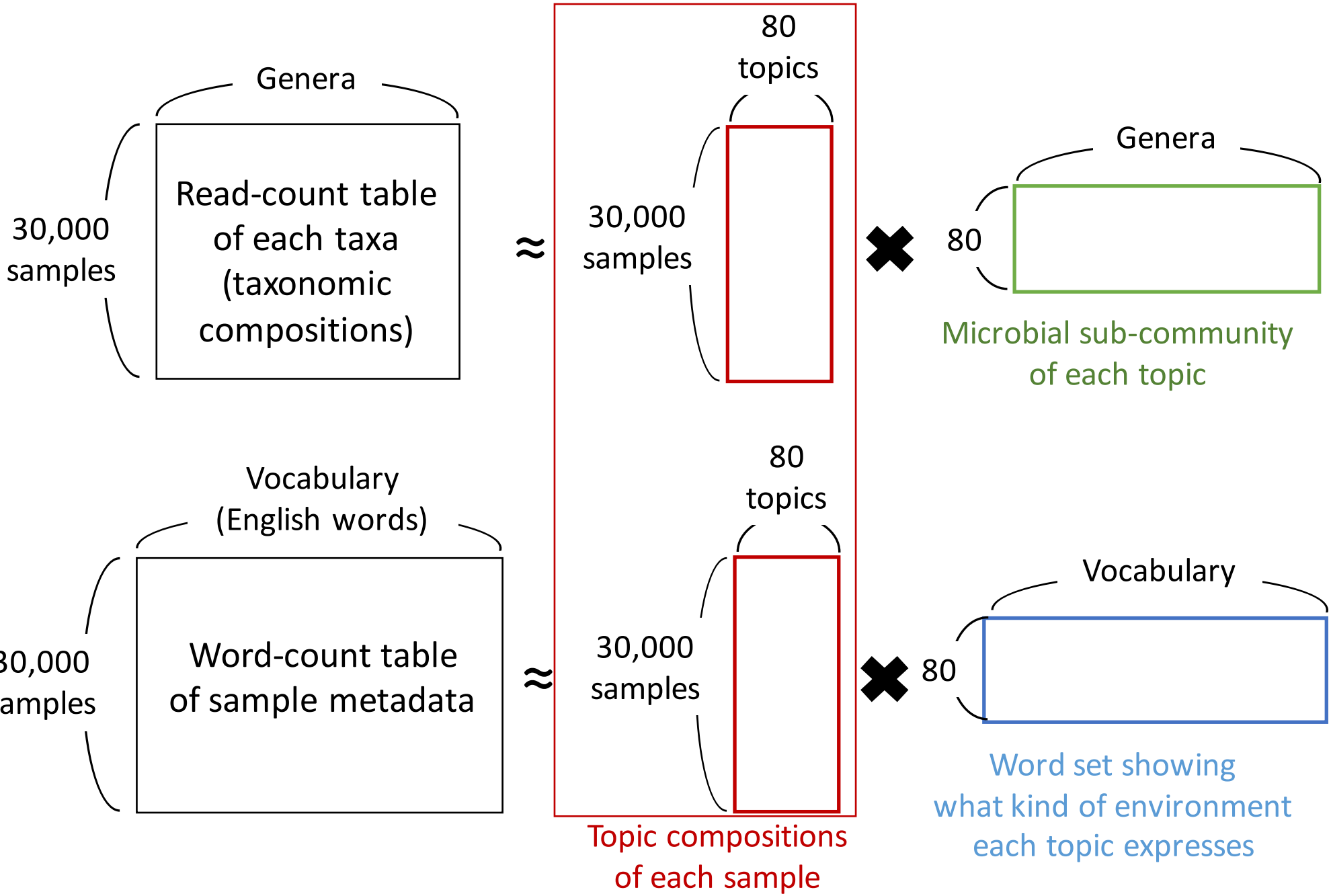


A topic representing "wastewater"

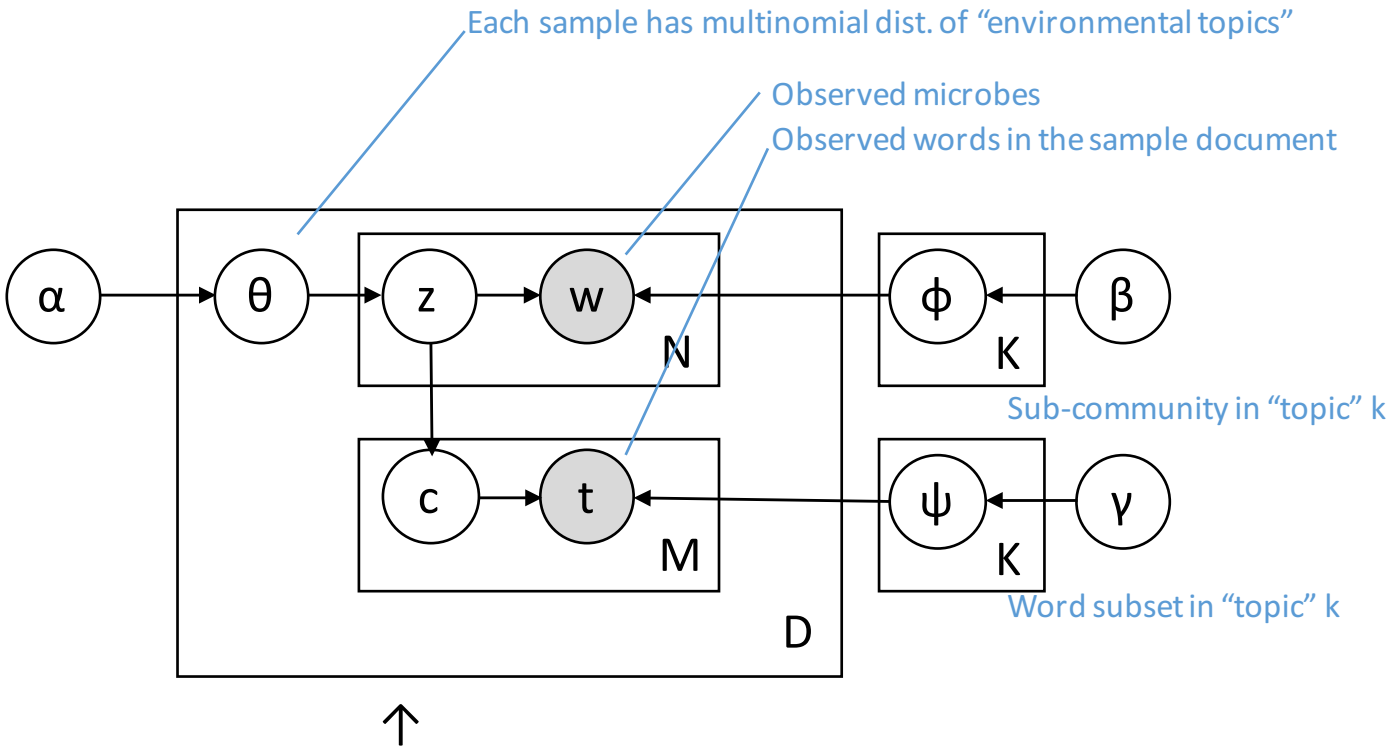
activate
sludge
wastewater
sewage



Matrix factorization for {read | word}-count tables (MicrobeDB.jp ver. 2 data)



Generative model of Latent environment allocation (LEA)
based on the topic model approach (Latent Dirichlet Allocation-model)
(cf. Admixture model of population genetics)



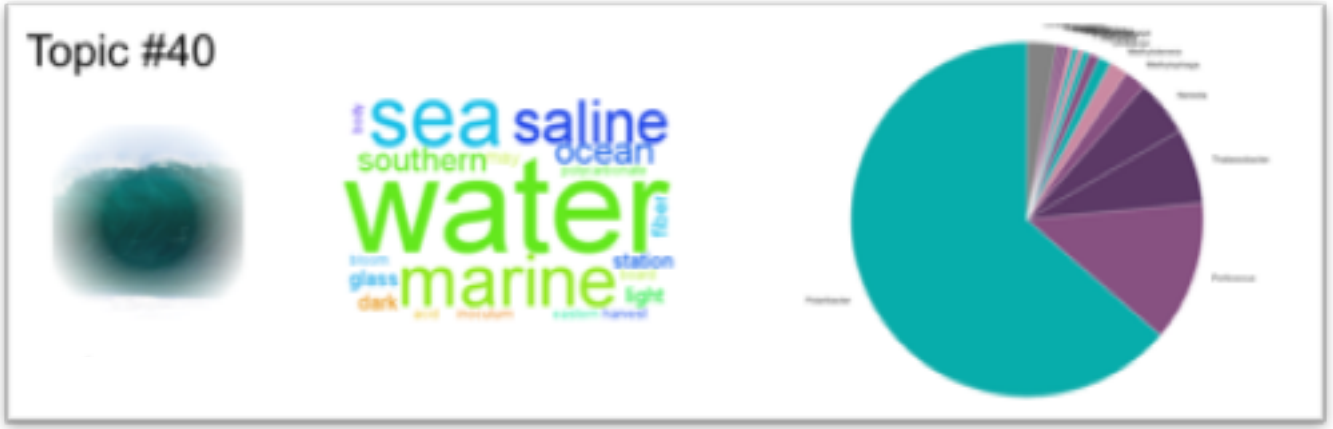
$$P(\mathbf{W},\mathbf{T},\mathbf{Z},\mathbf{C}|\alpha,\beta,\gamma) = P(\mathbf{Z}|\alpha)P(\mathbf{W}|\mathbf{Z},\beta)P(\mathbf{C}|\mathbf{Z})P(\mathbf{T}|\mathbf{C},\gamma)$$

Parameter optimization by Gibbs sampling

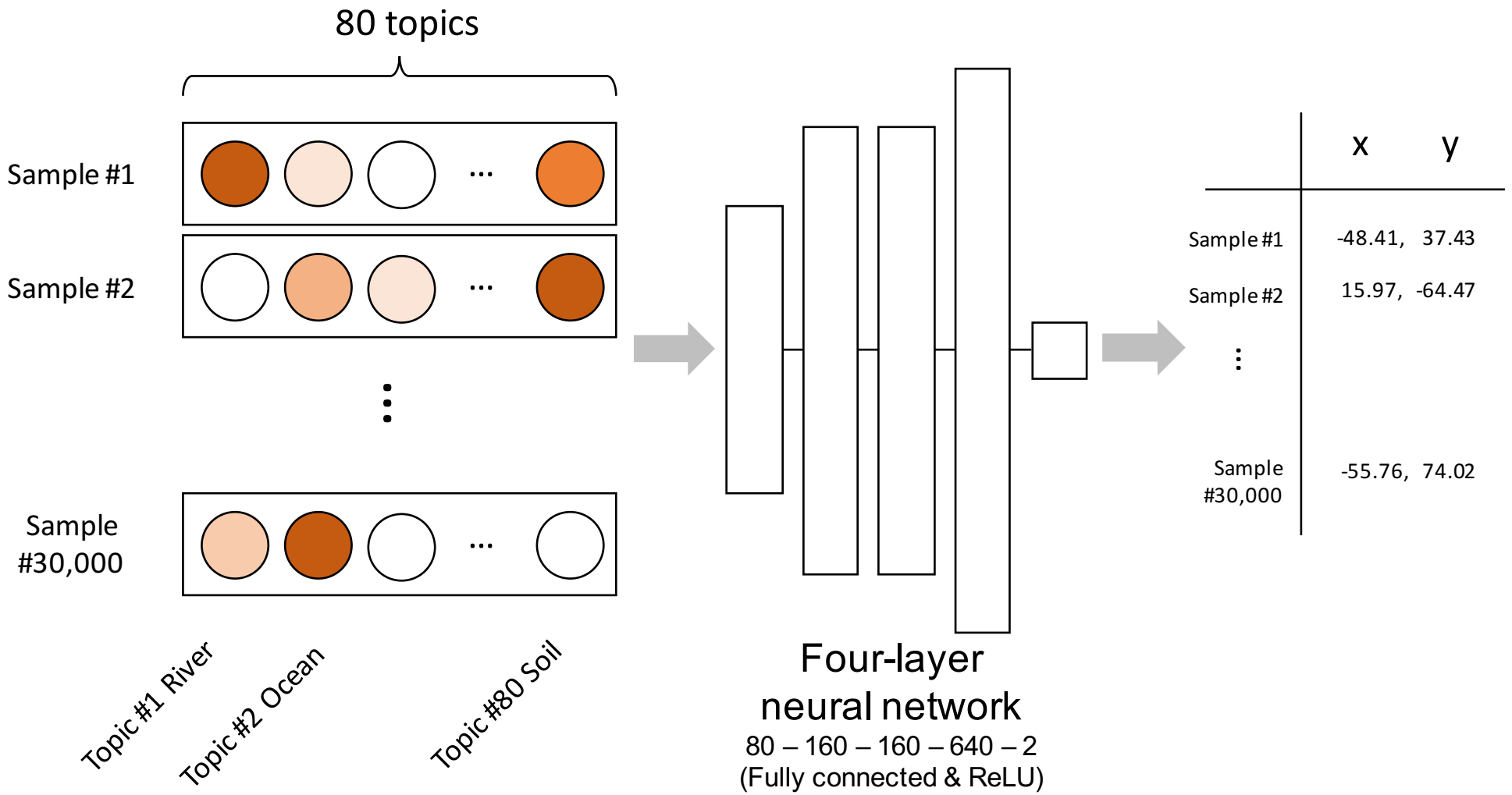
$$P(z_{dn}=k|\mathbf{W},\mathbf{T},\mathbf{Z}_{\backslash dn},\mathbf{C}) \propto (N_{kd\backslash dn}+\alpha_k)\frac{N_{kw_{dn}\backslash dn}+\beta}{N_{k\backslash dn}+\beta W}\left(\frac{N_{kd\backslash dn}+1}{N_{kd\backslash dn}}\right)^{M_{kd}}$$

$$P(c_{dm}=k|\mathbf{W},\mathbf{T},\mathbf{C}_{\backslash dm},\mathbf{Z}) \propto N_{kd}\frac{M_{kt_{dm}\backslash dm}+\gamma}{M_{k\backslash dm}+\gamma T}$$

Examples of estimated core “environmental topics”

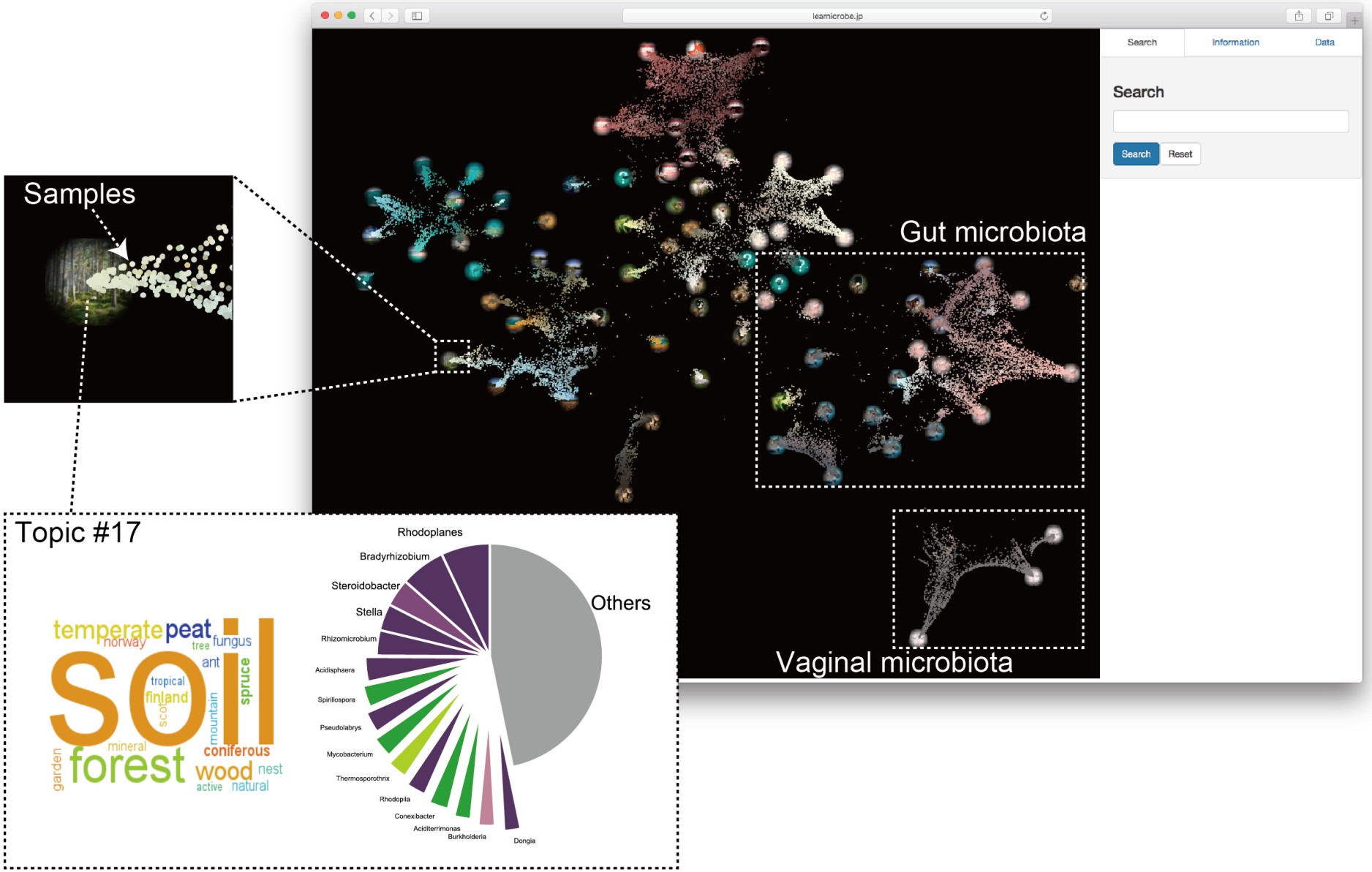


We extracted 80 topics using >30,000 microbiome samples
To visualize topic-based samples,
we applied parametric t-SNE method as dimensionality reduction



LEA (Latent Environment Allocation)

http://leamicrobe.jp/



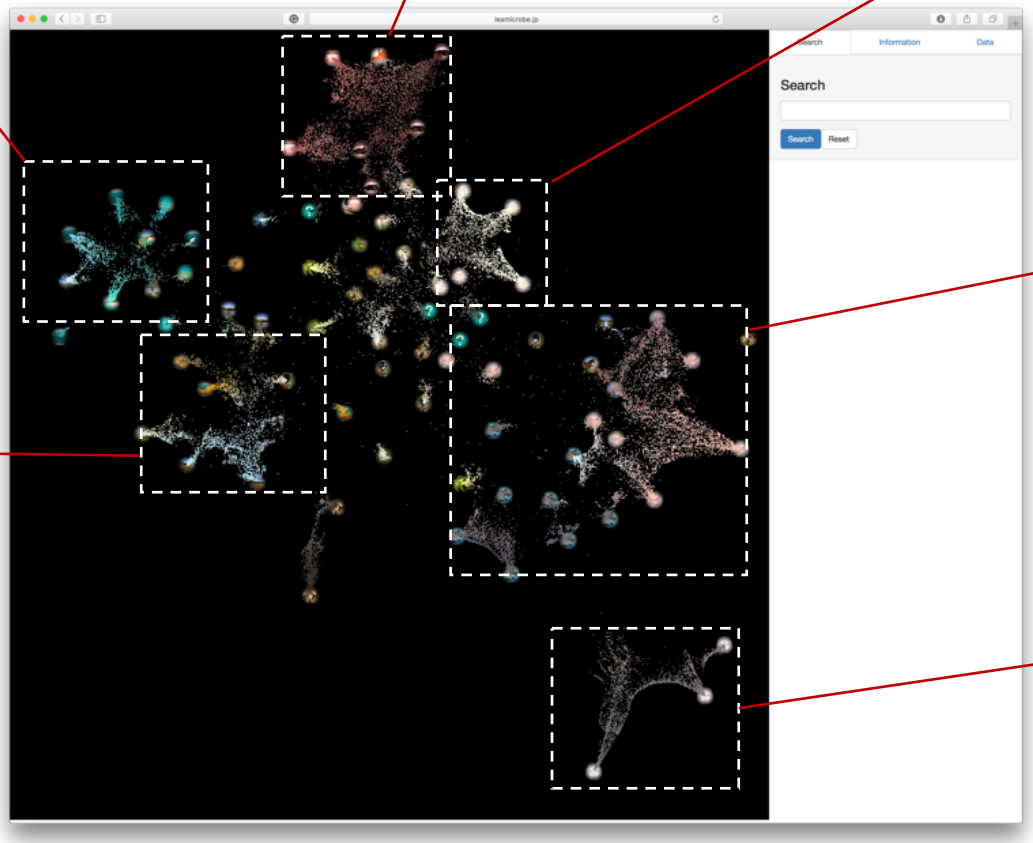
Higashi K *et al*. 2018

6 continents and some isolated islands

“Ocean” area

“Oral cavity” area

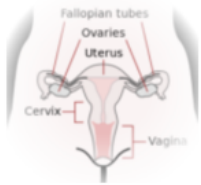
“Skin” area

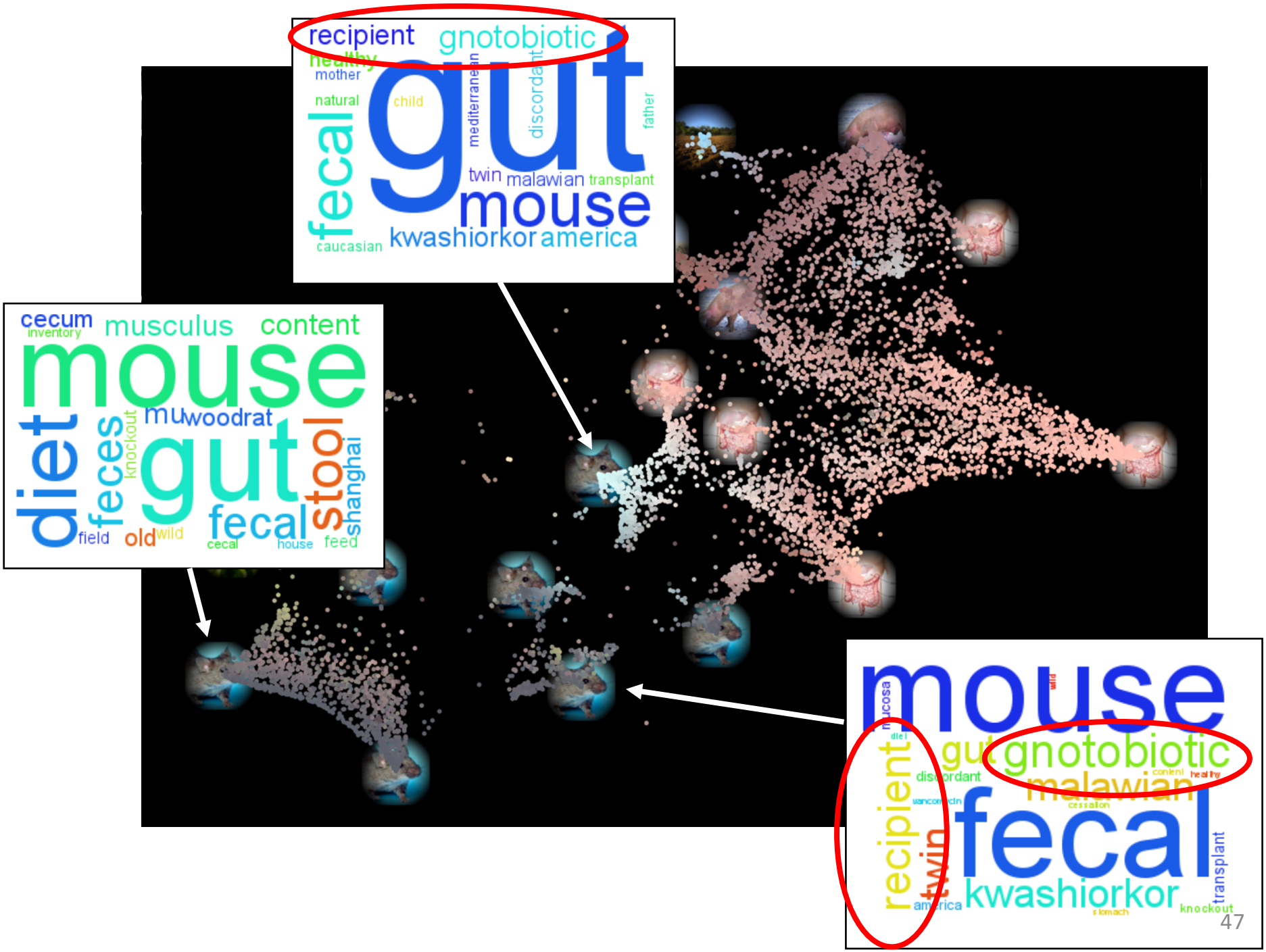


“Gut” area

“Soil” area

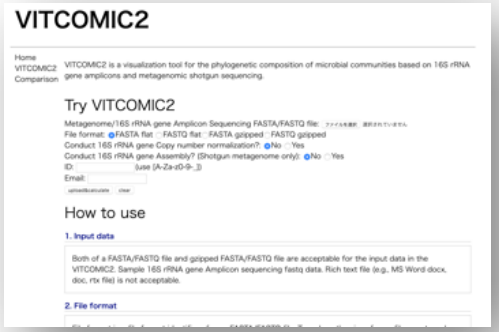
“Vagina” area



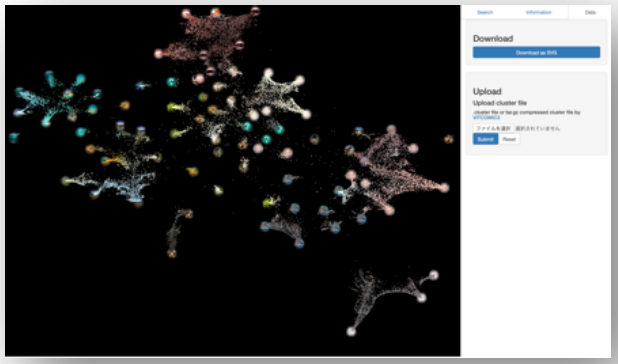

recipient
gnotobiotic
healthy
mother
natural
child
gut
mediterranean
discordant
father
fecal
twin
malawian
transplant
mouse
caucasian
kwashiorkor
america
cecum
inventory
musculus
content
mouse
diet
feces
knockout
muwoodrat
gut
stool
fecal
shanghai
field
old
wild
cecal
house
feed
mouse
mucosa
wild
diet
gut
gnotobiotic
discordant
malawian
content
healthy
vancomycin
cessation
recipient
twin
fecal
america
kwashiorkor
stomach
knockout
transplant

How to map your data on LEA

VITCOMIC2
(http://vitcomic.org/)



LEA
(http://leamicrobe.jp/)



fastq files

~ 6 min. / a sample (100,000 sequences)

cluster files (taxonomic compositions)

~ 0.5 sec. / a sample

MicrobeDB.jp ver. 3の特徴

- かなりマイクロバイオームに特化した統合DB
- 環境や病気のオントロジーはメタデータから自動的にアノテーション
- 160万サンプル以上のマイクロバイオームデータをメタデータ等で検索可能
- MeGAP3による計算が終わったサンプルから、数ヶ月に一度、MicrobeDB.jpで検索できるように
- 現在主にセキュリティ面等をテスト中であり、12月中にはhttp://microbedb.jp/から公開