平成26年度ライフサイエンスデータベース統合推進事業

ゲノム・メタゲノム情報統合による微生物DBの超高度化推進

国立遺伝学研究所
黒川 顕

主たる共同研究者

国立遺伝学研究所1

黒川　顕：微生物DBにおける研究統括

森　宙史：ゲノム・メタゲノムデータ＆真菌類データの整備

東京工業大学

山田拓司：メタゲノムデータの整備

山本　希：解析Stanza＆オントロジー開発

鈴木真也：高度解析Stanzaの開発、DB自動更新システムの開発

国立遺伝学研究所2

中村保一：藻類データの整備

菅原秀明：MiGAPとの連携強化

神沼英里：MeGAPとの連携強化

藤澤貴智：アクセスレベルの制限システムの開発

基礎生物学研究所

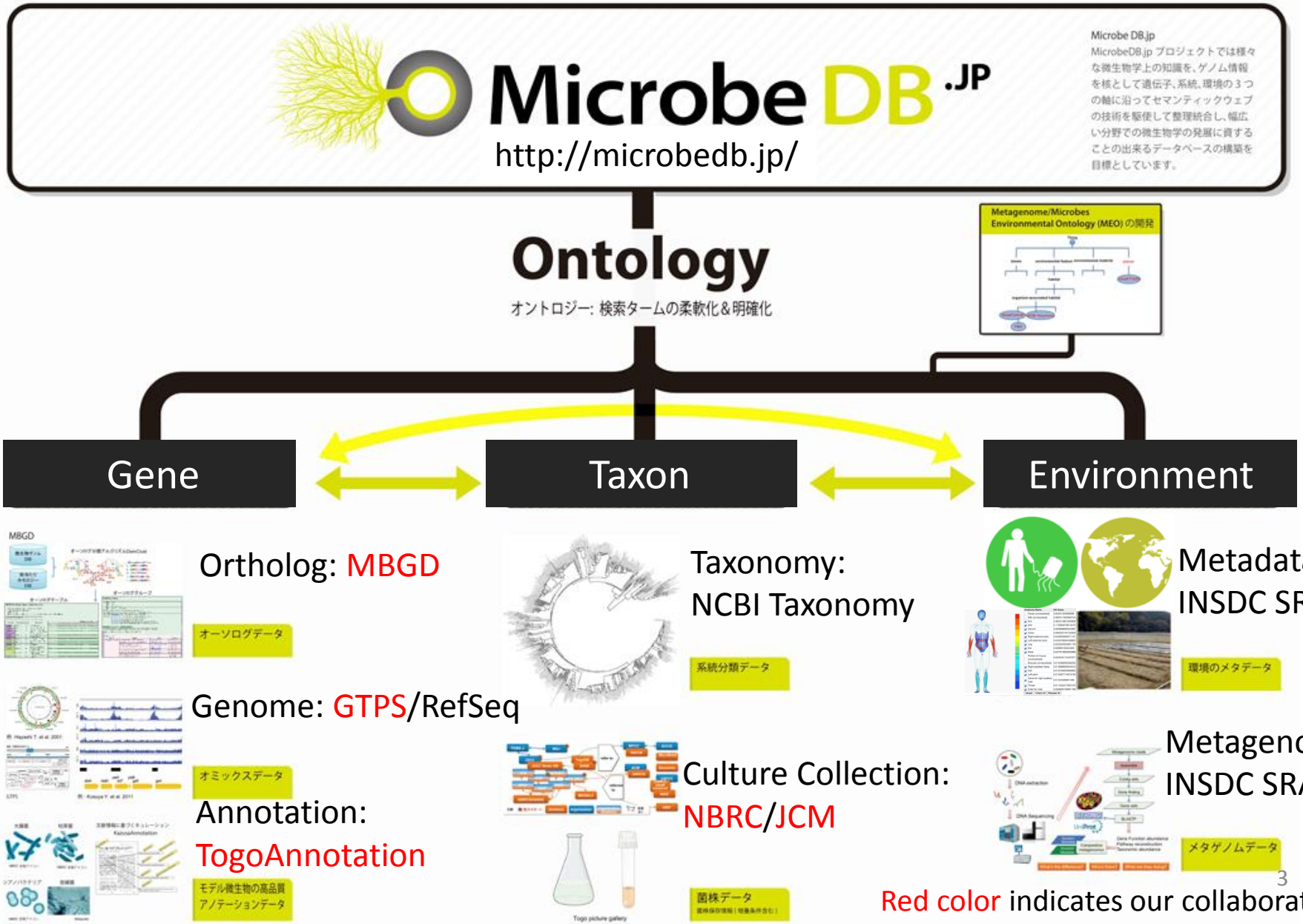
内山郁夫：真核生物に対するオーソログ解析手法の開発

千葉啓和：ドラフトゲノムのオーソログ解析、Stanzaの開発

西出浩世：オーソログを基軸とした各種データ統合の推進

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

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



Red color indicates our collaborators.

http://microbedb.jp/

MicrobeDB

microbedb.jp/MDB

Sign In

Microbe DB.JP

Search

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

JST

NBDC National Bioscience Database Center

DBCLS Database Center for Life Science

MicrobeDB.jpの第1期研究開発で実現したこと

1. 既存のゲノム中の各遺伝子の情報（オーソログ、系統プロファイル、環境プロファイル）
2. 菌株保存機関に存在する菌株の情報（生育培地、表現型情報、遺伝子機能組成）
3. 様々な環境中の細菌群集の情報（系統組成、遺伝子機能組成）
4. 上記の情報をシームレスに統合

問合せ例：

高温環境に多く存在する遺伝子はどのような遺伝子か?その遺伝子は、どの系統が主に持っているのか?

【研究テーマ例】工業排水のメタゲノム解析

1.工業廃水における細菌群集構造は？

→細菌群集構造を明らかにする

2.サンプル間でどのような変化がある？

→サンプル（時系列等）ごとに細菌群集を比較する

3.どのような細菌種、遺伝子が環境因子と強い相関を持つ？

→環境因子と強い相関を持つ遺伝子・種の探索

4.同様の環境ではどのような細菌群集が存在する？

→同様のメタデータを持つサンプルとの比較

5.同様の細菌群集はどのような環境に存在する？

→同様の細菌群集構造を示すサンプルとの比較

第2期研究開発の目標・ねらい

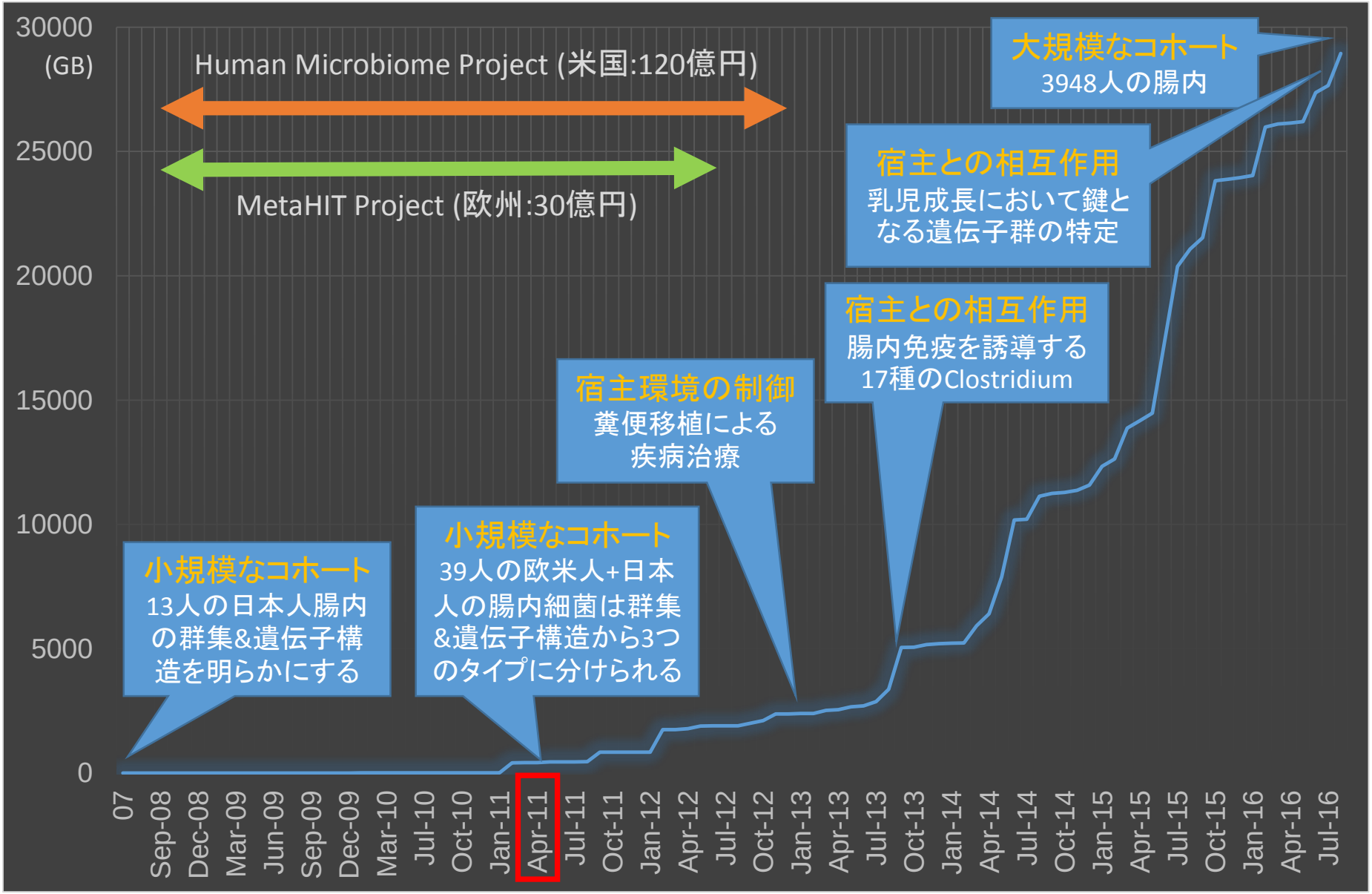
MicrobeDB.jpを

- より広い微生物種を対象として拡張
- データ収集や更新自動化による持続可能なシステム
- 最先端解析プロトコルを実装した解析結果の可視化

研究者コミュニティだけでなく不特定多数の
イノベータを対象とした利用性の向上を徹底する

単なる統計量の提示ではなく、大規模データから新規知識発見を容易に行う事が可能な、今までのDBを超えたDBシステムを構築する事を目標とする。

ヒトメタゲノムデータの蓄積



オーソログ遺伝子解析手順の問題点

- 同種／同属の近縁ゲノムデータが急激に増加しており、総当たりのホモロジー検索の負荷が増大している。
  - 種内・属内ホモロジーと種間・属間ホモロジーの分割管理

- 標準オーソログテーブルには、属ごとに代表生物種1ゲノムのみしか入っていないので、そこに含まれない遺伝子が解析対象とならない。
  - 種内（属内）オーソログループごとに代表遺伝子を選んで構築した種（属）パン・ゲノムを用いたオーソログ解析

MicrobeDB.jp version 2のデータ

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

これらのデータをRDF化した

MicrobeDB.jp version 2のトリプル数

データカテゴリ	Version 1トリプル数	Version 2トリプル数
Reference genomes	747,494,575	4,786,370,037
MBGD Orthologs	291,714,037	2,154,452,832
Metagenomes	16,841,085	1,924,126,117
BRC strains	903,319	983,319
Disease	8,809	253,201
Transcriptome	0	8,231,210
BioSamples	0	140,074,319
Ontologies	18,721,883	25,786,388
Others	440,773	1,704,121
Total	1,076,124,481	9,041,981,544

http://microbedb.jp/MDBdemo

top page | Sign In

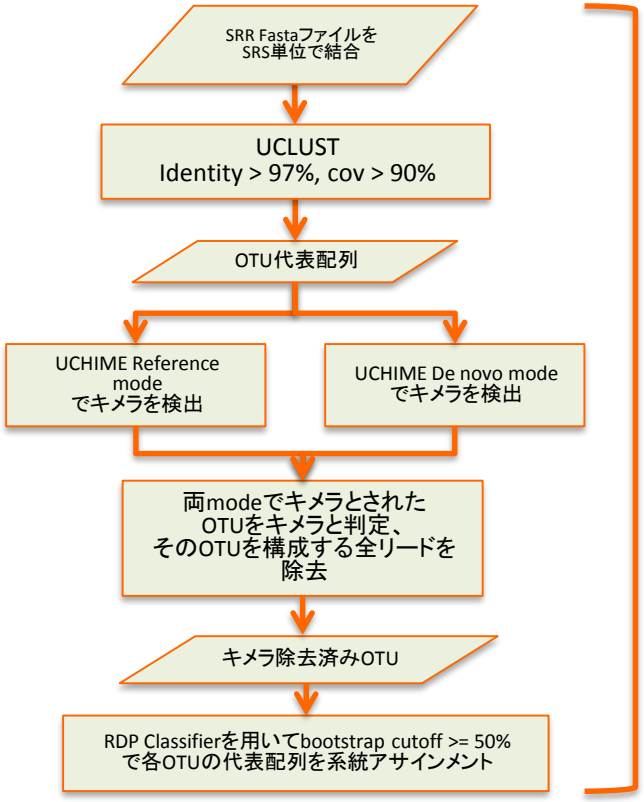
Text Sequence Analysis Statistics

Search

Environment: hot spring
Taxonomy: Enterococcus faecalis
Taxonomy: Streptomyces avermitilis
Gene: psbA
ID: 29

統計Stanza＆解析Stanzaの開発

多種多様な情報が混在しているゲノムやメタゲノム等の複雑なデータから知識発見をするために、比較ゲノム解析や比較メタゲノム解析など様々な解析Stanzaを、TogoStanzaと合わせて約180種類開発し実装した。



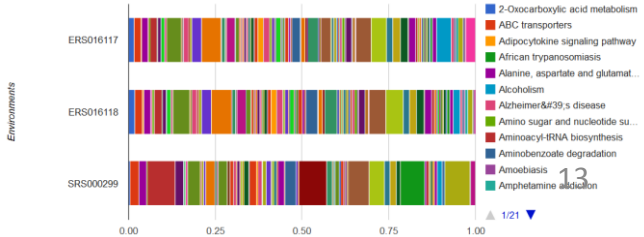
解析Stanzaによる
結果の可視化

各系統と温度との相関係数リスト

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	0.08840924648161 7	29	11.4
http://www.genome.jp/dbget-bin/www_bget?ko00051	Fructose and mannose metabolism	0.08738568321405 2	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	0.05947515368806 5	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	0.01861957147459 1	28	10.7
http://www.genome.jp/dbget-bin/www_bget?ko00061	Fatty acid biosynthesis	0.00919715700696 16	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

環境ごとの系統組成のグラフ



統計Stanza example

・遺伝子アノテーション情報

Feature

[http://purl.obolibrary.org/obo/SO_0000704]

dbxref	http://www.ncbi.nlm.nih.gov/gene/897644
feature_gene	polC
feature_locus_tag	TM0576
location	605923..610026
isPartOf	http://genome.db/uuid/b4d48cd7-00ef-4e03-9adb-fda7de39e078
type	http://purl.obolibrary.org/obo/SO_0000704
label	TM0576

[http://purl.obolibrary.org/obo/SO_0000316]

dbxref	http://www.ncbi.nlm.nih.gov/gene/897644
dbxref	http://www.ncbi.nlm.nih.gov/nuccore/15643342
exons	nodeID://b71582

・オーソログ遺伝子リスト

Ortholog

ID	Genome	Description	Protein	UniProt	GTPS	RefSeq
aac:AACI_1427	aac	DNA polymerase III subunit alpha	YP_003184842.1	C8WWI2	AACI_ACIDOCALDARIUSDSM446:ST2344	NC_013205.1
aar:ACEAR_1599	aar	DNA polymerase III catalytic subunit, PolC type	YP_003828170.1		AARA_DSM5501:ST105	
acl:ACL_0247	acl	DNA polymerase III subunit alpha	YP_001620249.1	A9NEU3	ALAI_PG8A:ST588	NC_010163.1
afl:AFLV_1700	afl	DNA polymerase III PolC	YP_002316046.1	B7GG80	AFLA_WK1:ST2505	NC_011567.1
afn:ACFER_1370	afn	DNA polymerase III subunit alpha	YP_003399045.1		AFER_DSM20731:ST1519	NC_013740.1
amt:AMET_2678	amt	DNA polymerase III subunit alpha	YP_001320489.1	A6TRL2	AMET_QYMF:ST2214	NC_009633.1

・ゲノム基礎情報

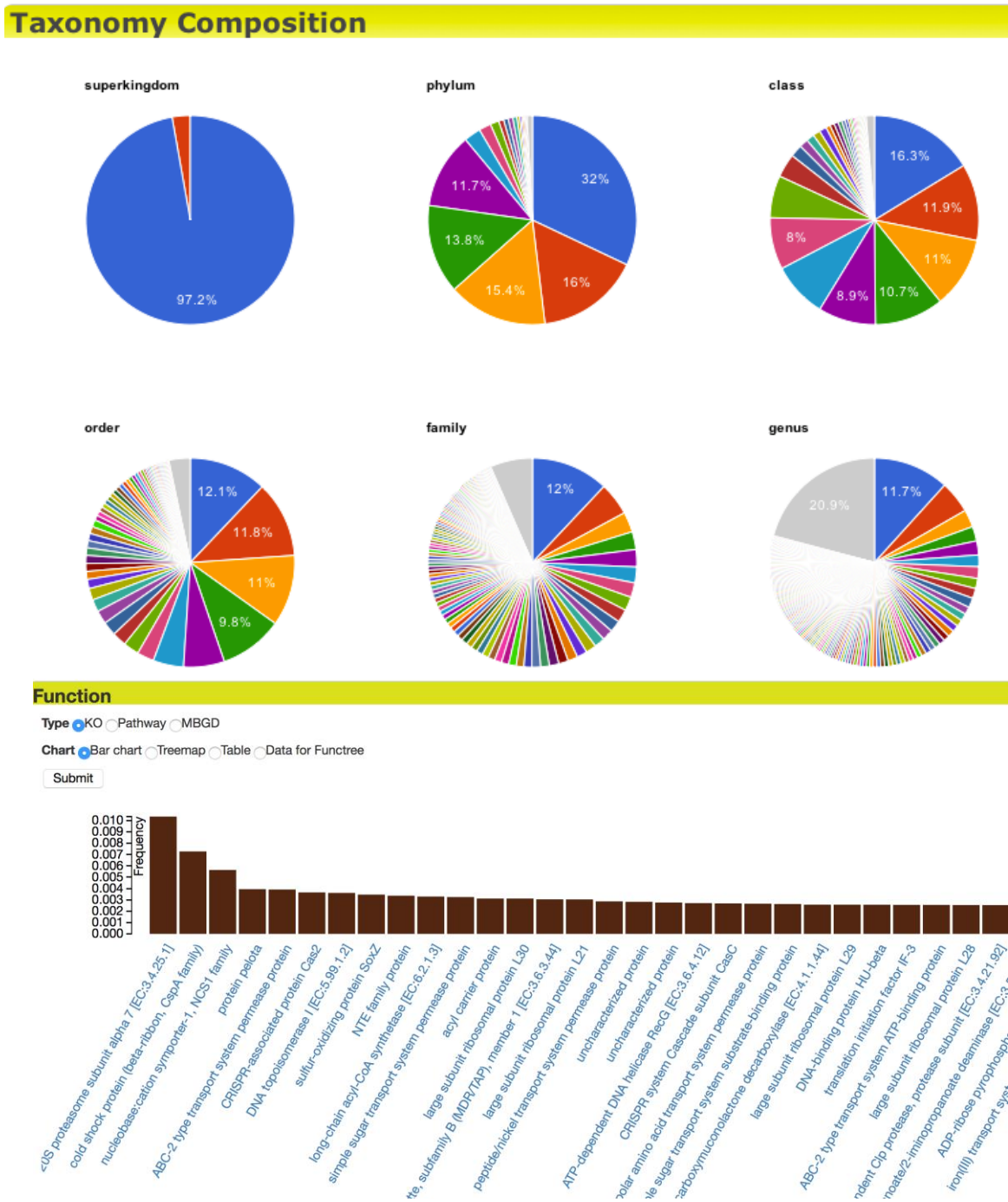
Genome

length	1860725
location	1..1860725
molecularType	genomic DNA
organism	Thermotoga maritima MSB8
sequence	http://genome.db/uuid/b4d48cd7-00ef-4e03-9adb-fda7de39e078.fasta
start	1
stop	1860725
strain	MSB8
version	NC_000853.1
modified	2012-02-13
type	http://purl.obolibrary.org/obo/SO_0000340
type	http://purl.obolibrary.org/obo/SO_0000988
comment	Thermotoga maritima MSB8 chromosome, complete genome.

統計Stanza example

・メタゲノムサンプルにおける
微生物群集構造

・メタゲノムサンプルにおける
遺伝子機能分類



統計＆解析Stanza example

オーソログループに関する多様な情報

Ortholog group (cluster_id=2000)

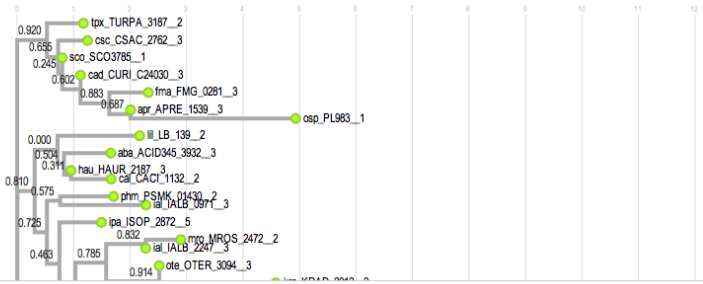
Ortholog description

Name	rsbU
Description	Protein serine/threonine phosphatase
Member count	342

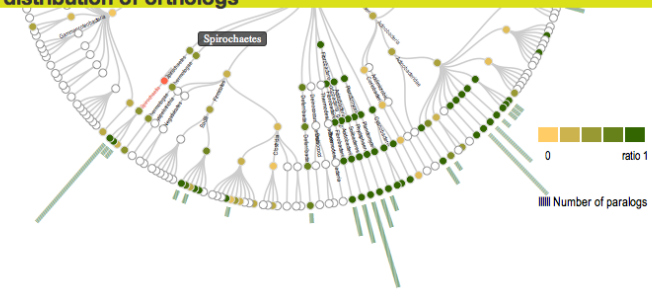
Group members

Superkingdom	Phylum	Organism name	Ortholog group members	Gene description
Archaea	Euryarchaeota	Methanocorpusculum labreanum Z	mla:MLAB_1450(2)	hypothetical protein
		Methanomassiliicoccus intestinalis Issoire-Mx1 Mx1-Issoire	mer:H729_02670(2)	hypothetical protein
		Methanoregula boonei 6A8	mbn:MBOO_0643(2)	stage II sporulation E family protein
			mbn:MBOO_1652(2)	stage II sporulation E family protein
		Methanosaeta thermophila PT	mtp:MTHE_1335(3)	serine phosphatase
		Methanospirillum hungatei JF-1	mhu:MHUN_0060(2)	serine phosphatase
		Acidobacteria bacterium Ellin345	aba:ACID345_0959(3)	serine phosphatase
			aba:ACID345_1335(3)	serine phosphatase
			aba:ACID345_1448(3)	serine phosphatase
			aba:ACID345_2035(3)	serine phosphatase
			aba:ACID345_2718(4)	serine phosphatase

Phylogenetic tree of orthologs



Taxonomic distribution of orthologs

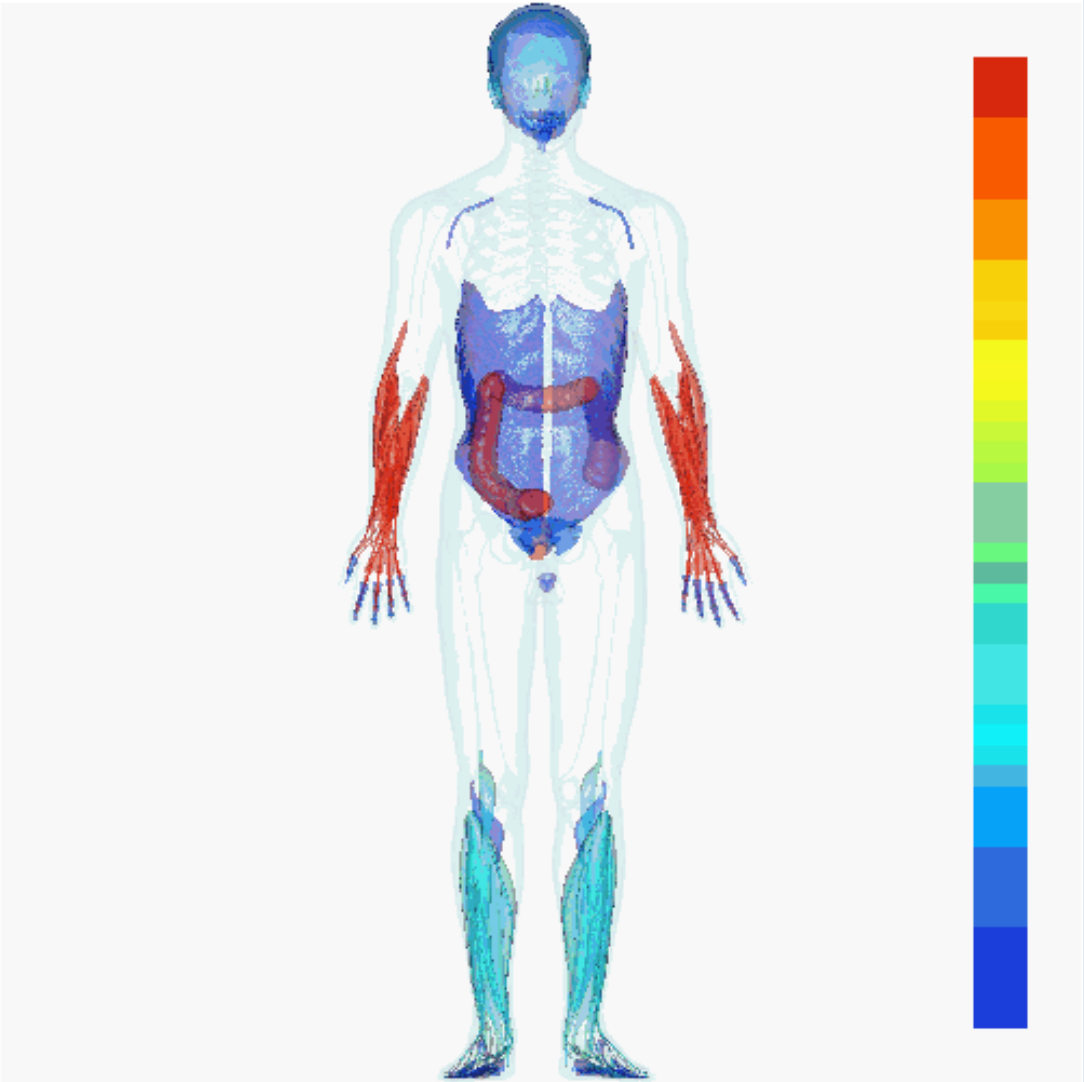


Phenotypes of organisms with the orthologs

Human Meta Body Map Stanza

自分の興味がある系統・遺伝子はヒトの体のどこに多いのか？

Body Mapping



	Anatomy Name	Hit Value
☐	Feces (unvisualized)	0.84051325986689
☐	Milk (unvisualized)	0.585317462682724
☑	Gut	0.362513891694809
☑	Arm	0.110064819613473
☑	Cecum	0.093869656324387
☑	Colon	0.090222145120303
☑	Right external naris	0.045853955671191
☑	Left external naris	0.043376094428822
☑	Leg	0.032520323991776
☑	Ear	0.02869155324353
☑	Nose	0.027815662800968
☐	Portion of mucus (unvisualized)	0.022045172407877
☐	Mucosa (unvisualized)	0.015296295285225
☑	Right popliteal fossa	0.013866629324515
☑	Hair	0.013750255660852
☑	Left palm	0.012587714675729
☑	Canal for right auditory tube	0.01234339691499
☑	Throat	0.011534477907475
☑	External nose	0.009956159861758

reload | Check All | Release All

解析Stanza example

VITCOMIC2スタンザ

ユーザのメタ16S・メタゲノムデータと、類似した系統組成を持った
サンプルおよび環境カテゴリ（MEO ID）を探索し明快に視覚化

VITCOMIC Judge Environment(?)

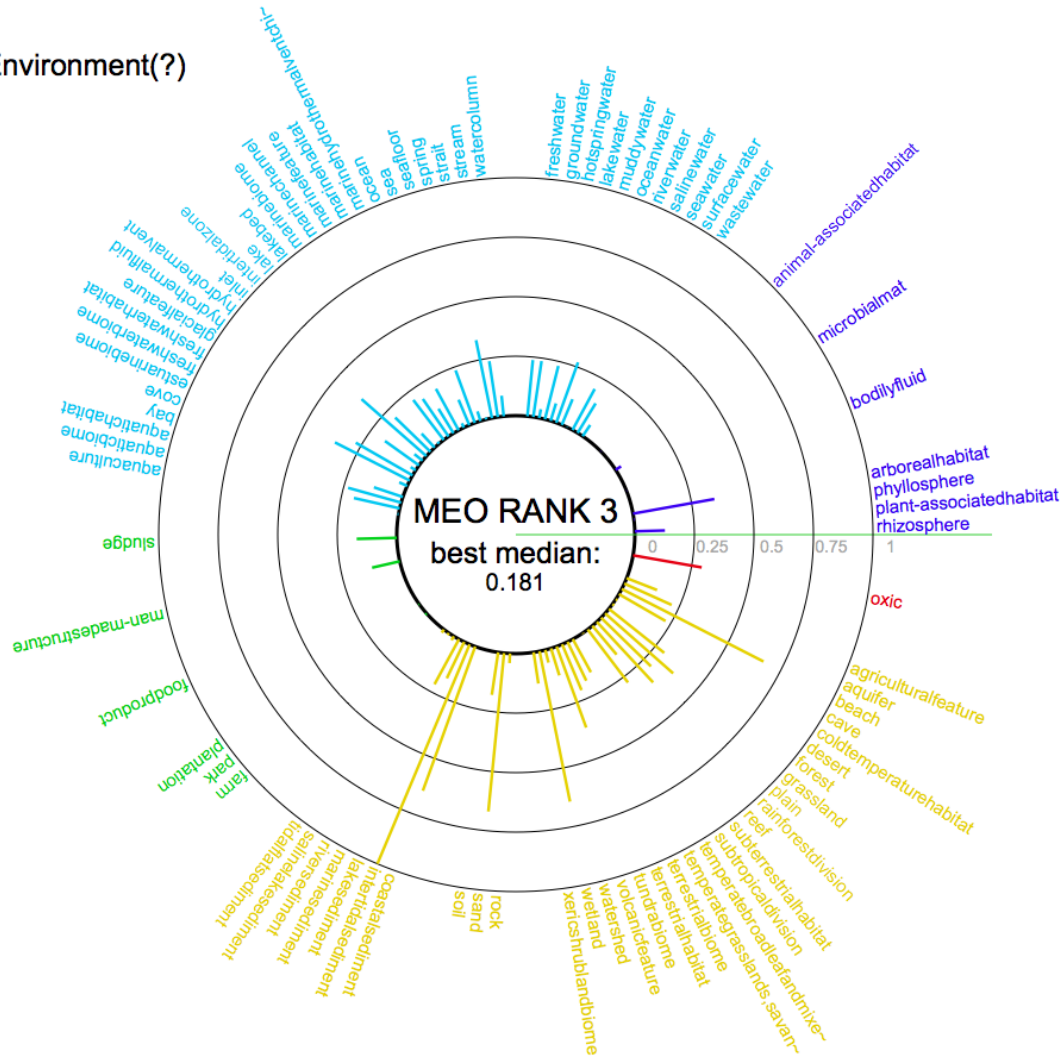
title = mix

MODE : MEDIAN

Rank settings :

MEO Rank 1
MEO Rank 2
MEO Rank 3
MEO Rank 4
MEO Rank 5
MEO Rank 6
MEO Rank 7

atmosphere
geosphere
humanactivityassociation
hydrosphere
organismassociation



解析Stanza example

比較ゲノム解析

top page | Sign In

Microbe DB.JP

Enterococcus faecalis > Enterococcus faecalis TX2141

Category colors: Environment | Taxonomy | Gene [hit column] (hit count) | Phenotype | Other category

This search term has exact match.
Now displaying stanzas in the category: **Taxonomy** . Parameters are tax_id: **1351**

Taxonomies Function Comparison Table

KEGG	Streptobacillus moniliformis DSM 12112	Bacillus amyloliquefaciens LL3
Glycolysis / Gluconeogenesis	+	+
Citrate cycle (TCA cycle)	+	+
Pentose phosphate pathway	+	+
Pentose and glucuronate interconversions	+	+
Fructose and mannose metabolism	+	+
Galactose metabolism	+	+
Ascorbate and aldarate metabolism	+	+
Fatty acid biosynthesis	-	+
Fatty acid degradation	+	+
Synthesis and degradation of ketone bodies	+	+
Secondary bile acid biosynthesis	-	+
Ubiquinone and other terpenoid-quinone biosynthesis	-	+
Oxidative phosphorylation	+	+
Arginine biosynthesis	+	+
Purine metabolism	+	+

解析Stanza example

比較メタゲノム解析

top page | hmori Sign out

Microbe DB.JP

Sample Comparison

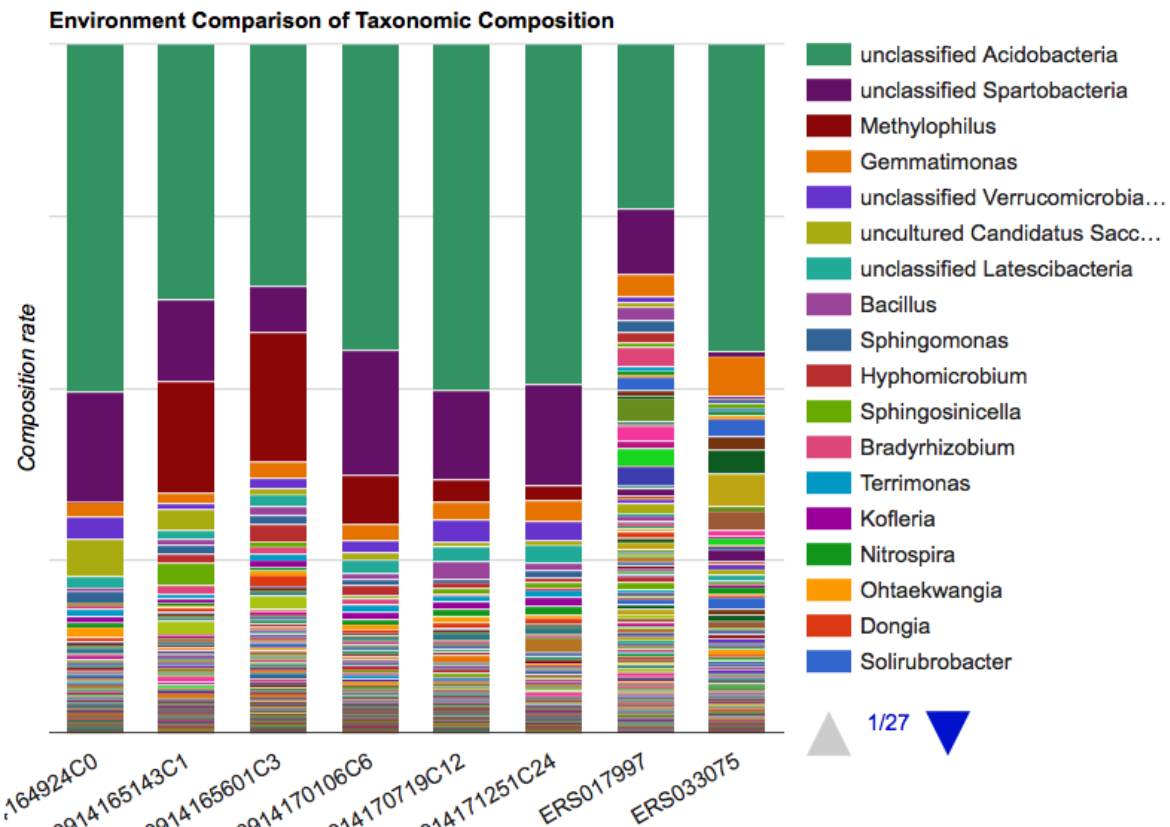
Selected samples:

150914164924C0 | 150914165143C1 | 150914165601C3 | 150914170106C6 | 150914170719C12 | 150914171251C24 | ERS017997 | ERS033075

Taxonomic rank: Genus

Compare samples

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



Stanzaの型(context)を定義

stanza	context
disease_attributes_stanza	Disease
disease_symptom_list_stanza	Disease
disease_taxonomy_list_stanza	Disease
environment_attributes_stanza	Environment
environment_function_table_stanza	Environment
environment_gold_list_stanza	Environment
environment_gold_table_stanza	Environment
environment_linage_info_stanza	Environment
environment_sample_amount_stackcharts_stanza	Environment
environment_strain_list_stanza	Environment
environment_to_brother_function_comparison_stackbar_stanza	Environment
environment_to_brother_taxonomy_comparison_stackbar_stanza	Environment
gold_taxonomy_compositon_stanza	Environment
meo_ontology_viewer_stanza	Environment
meta16s_sample_list_stanza	Environment
meta16s_taxonomy_compositon_stanza	Environment
metagenome_sample_list_stanza	Environment
gene_annotation_list_stanza	Gene
gene_ortholog_list_stanza	Gene
metadata_correlation_analysis_wrapper_stanza	Metadata
metadata_function_correlation_analysis_scatterplot_stanza	Metadata
metadata_function_correlation_analysis_table_stanza	Metadata
metadata_taxonomy_correlation_analysis_scatterplot_stanza	Metadata
metadata_taxonomy_correlation_analysis_table_stanza	Metadata
sample_taxonomy_envtree_stanza	Metagenome
taxonomy_envtree_stanza	Taxonomy
sample_cross_reference_stanza	Metagenome
sample_definition_info_stanza	Metagenome
sample_function_stanza	Metagenome
sample_metadata_info_stanza	Metagenome
sample_spot_map_stanza	Metagenome
sample_taxonomy_composition16s_stanza	Metagenome
sample_taxonomy_similarity_search_stanza	Metagenome
sample_to_environment_function_comparison_stackbar_stanza	Metagenome
sample_to_environment_taxonomy_comparison_stackbar_stanza	Metagenome
sample_to_project_function_comparison_stackbar_stanza	Metagenome
sample_to_project_taxonomy_comparison_stackbar_stanza	Metagenome

environments_function_comparison_stackbar_stanza	Multi Environment
environments_taxonomy_comparison_stackbar_stanza	Multi Environment
samples_function_comparison_heatmap_stanza	Multi Metagenome
samples_function_comparison_stackbar_stanza	Multi Metagenome
samples_metadata_comparison_table_stanza	Multi Metagenome
samples_taxonomy_comparison_heatmap_stanza	Multi Metagenome
samples_taxonomy_comparison_hierarchical_clustering_stanza	Multi Metagenome
samples_taxonomy_comparison_pcoa_stanza	Multi Metagenome
samples_taxonomy_comparison_stackbar_stanza	Multi Metagenome
samples_taxonomy_diversity_index_comparison_plot_stanza	Multi Metagenome
taxonomies_function_comparison_table_stanza	Multi Taxonomy
ortholog_attributes_info_stanza	Ortholog
ortholog_environment_list_stanza	Ortholog
ortholog_environment_profile_stanza	Ortholog
ortholog_group_members_info_stanza	Ortholog
ortholog_included_taxon_stanza	Ortholog
sample_comparison_wrapper_stanza	Other
phenotype_attributes_stanza	Phenotype
refseq_definition_info_stanza	Refseq
refseq_feature_info_stanza	Refseq
refseq_genome_info_stanza	Refseq
gene_list_by_text_stanza	Search
latitude_to_longitude_sample_map_stanza	Statistics
numeric_metadata_histogram_stanza	Statistics
sample_taxonomy_dashboard_stanza	Statistics
taxonomy_phenotype_dashboard_stanza	Statistics
strain_definition_info_stanza	Strain
strain_metadata_info_stanza	Strain
strain_other_collection_numbers_stanza	Strain
strain_reference_info_stanza	Strain
genome_information_stanza	Taxonomy
pathogen_information_stanza	Taxonomy
phenotype_information_stanza	Taxonomy
taxonomy_body_mapping_stanza	Taxonomy
taxonomy_composition_via_meta16s_stanza	Taxonomy
taxonomy_cross_references_stanza	Taxonomy
taxonomy_definition_info_stanza	Taxonomy
taxonomy_phylogenetics_tree_stanza	Taxonomy
taxonomy_strain_list_stanza	Taxonomy
taxonomy_taxonomy_tree_stanza	Taxonomy

検索語の型（context）をオントロジーを用いて推測

環境 = MEO、系統 = NCBI taxonomy、遺伝子機能 = product名等、Phenotype = MPO

Microbe DB.JP

hot spring x

Category colors: Environment Taxonomy Gene [hit column] (hit count) Phenotype Other category

Displaying related keywords . Please press for change to the new term instead of **"hot spring"**

Environment serpentine hot spring calcite hot spring alkaline hot spring hot spring water neutral hot spring acid hot spring artificial hot spring acidic hot spring water

Taxonomy

Gene

Phenotype

This search term has exact match.
Now displaying stanzas in the category: **Environment** . Parameters are meo_id: **MEO_0000029**

Environment attributes

MEO ID	MEO_0000029
Title	hot spring
Synonyms	hot springs, hotspring, spring, thermal feature, thermal spring, thermal springs
Comment	A spring that is produced by the emergence of geothermally-heated groundwater from the Earth's crust.
MEO SuperClass ID	MEO_0000083
MEO SuperClass Title	spring

Environment Hierarchy

ID	Title
MEO_0000817	Environment for microbes

DDBJとの連携

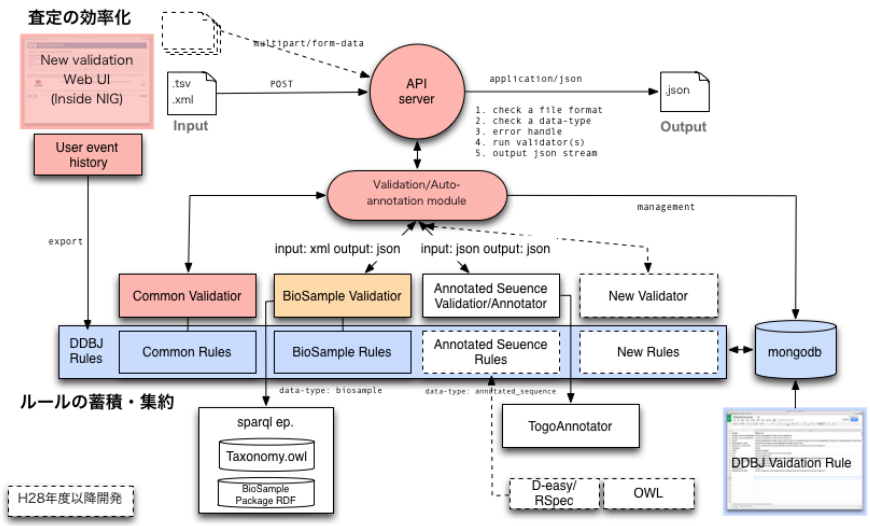
1.DDBJ RDF公開（予定）

OWL　RDF

- DDBJ release 104（239億トリプル、非圧縮2.4TB）
- INSDCオントロジーを利用したGenome RDFと統一モデル
- ftp.ddbj.nig.ac.jp/rdf からの公開（2017 DDBJ DB issue）
- NBDC/RDFポータルからの公開に向けて準備中

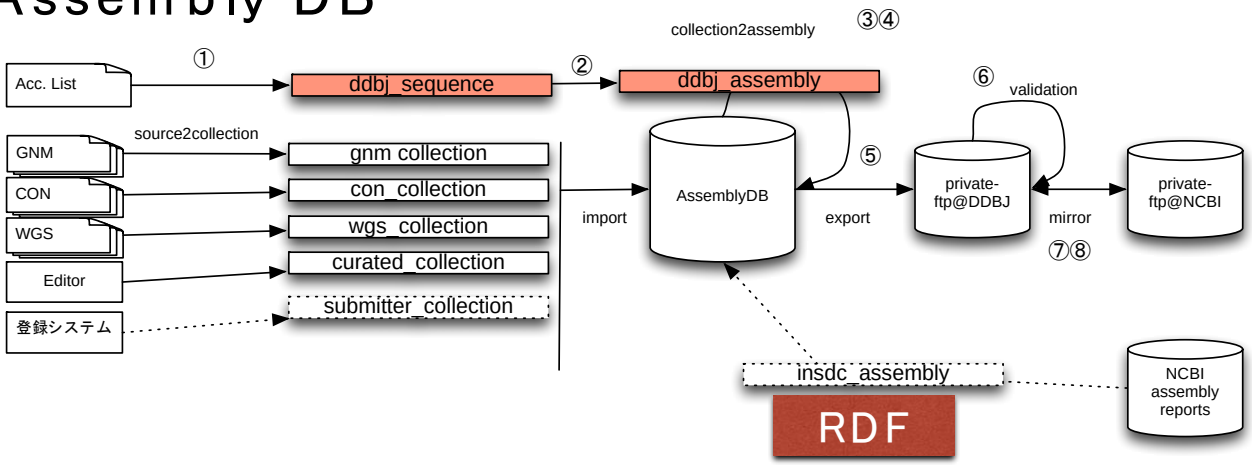
2.DDBJ Validator基盤開発

DDBJ validation module/APIサーバ開発仕様案 2016.05.12



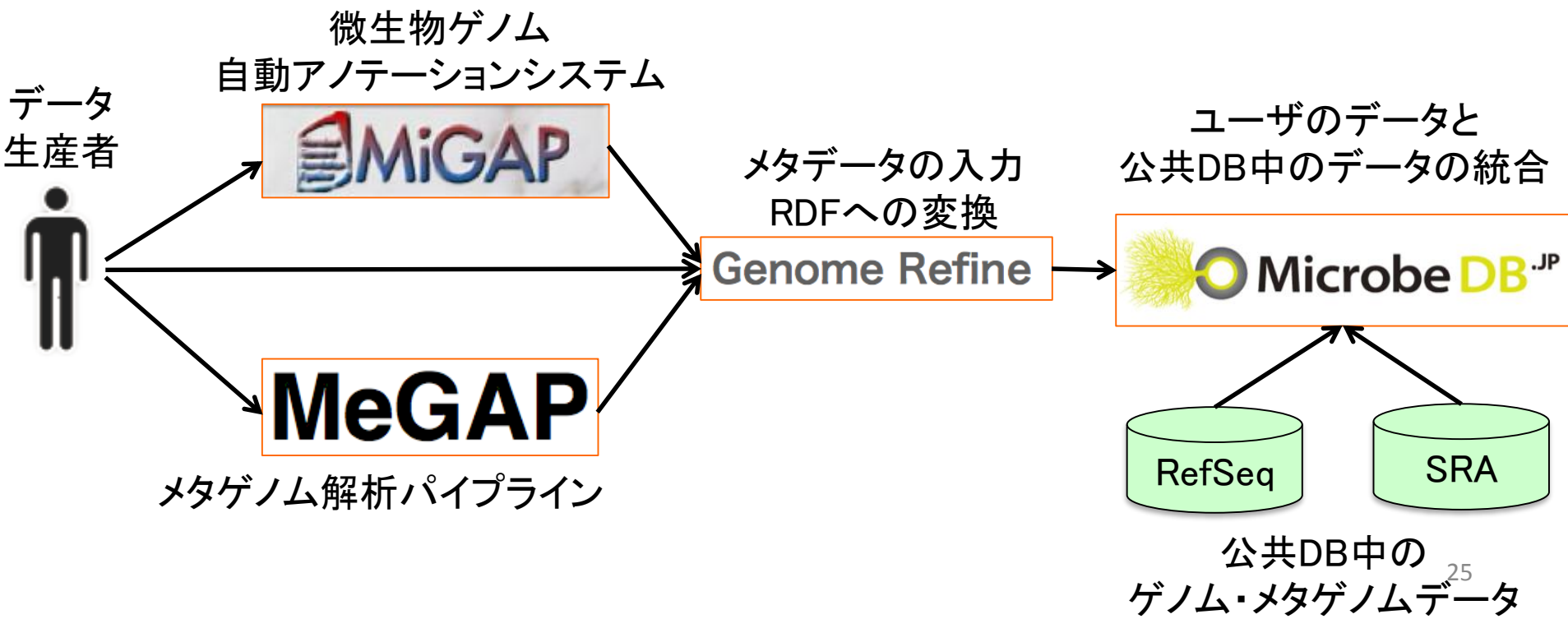
DDBJ Defind BioSample Package/Attributeを定義したOWL開発　OWL

3.DDBJ Assembly DB



データの収集およびクオリティコントロール
更新の自動化など持続可能なシステムの構築

データ生産者から継続的にデータを受け付ける窓口のシステムとして微生物ゲノム自動アノテーションシステム「MiGAP」およびメタゲノム解析パイプライン「MeGAP」を利用し、MicrobeDB.jpと一体運用を実現する。また、これまで手作業で実施してきたDBの更新作業を可能な限り自動化し更新体制を強化する。

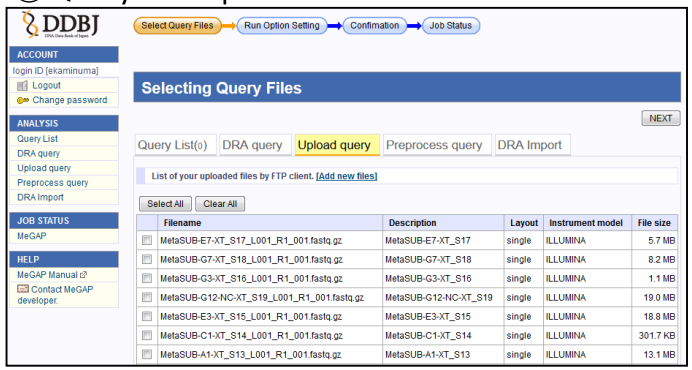


DDBJ Pipeline/MeGAP

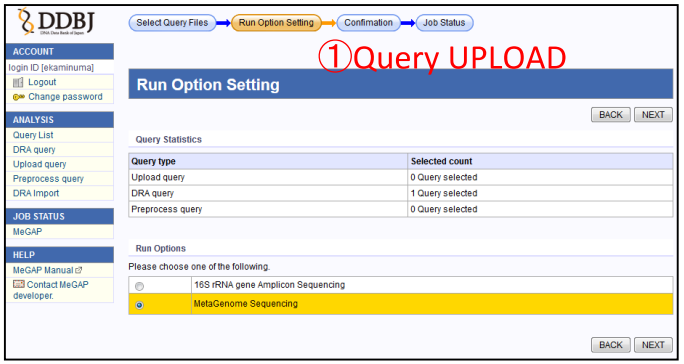
MeGAP powered by DDBJ Pipeline

http://p.ddbj.nig.ac.jp/pipeline/**ext**/Login.doよりログイン
(※公開時TOP画面にLINK作成)

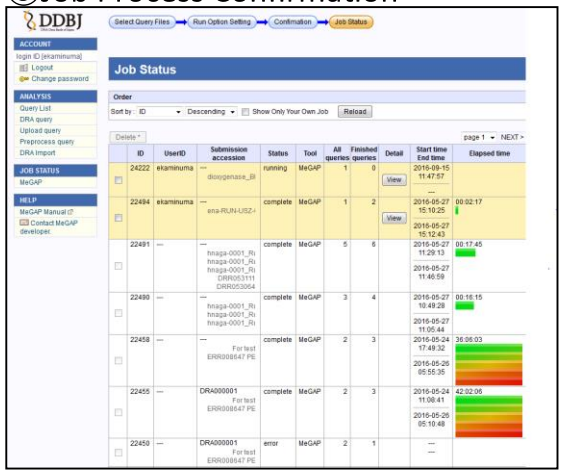
①Query File Upload



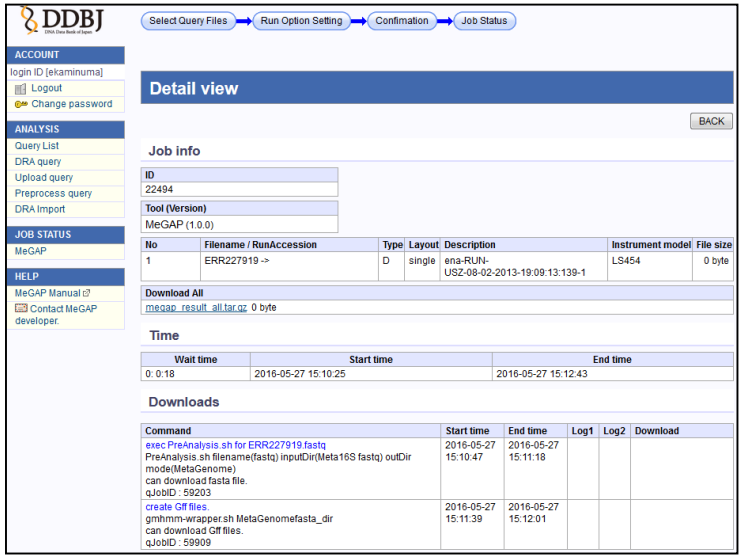
②Branch of 16S/Metagenome workflows



③Job Process Confirmation

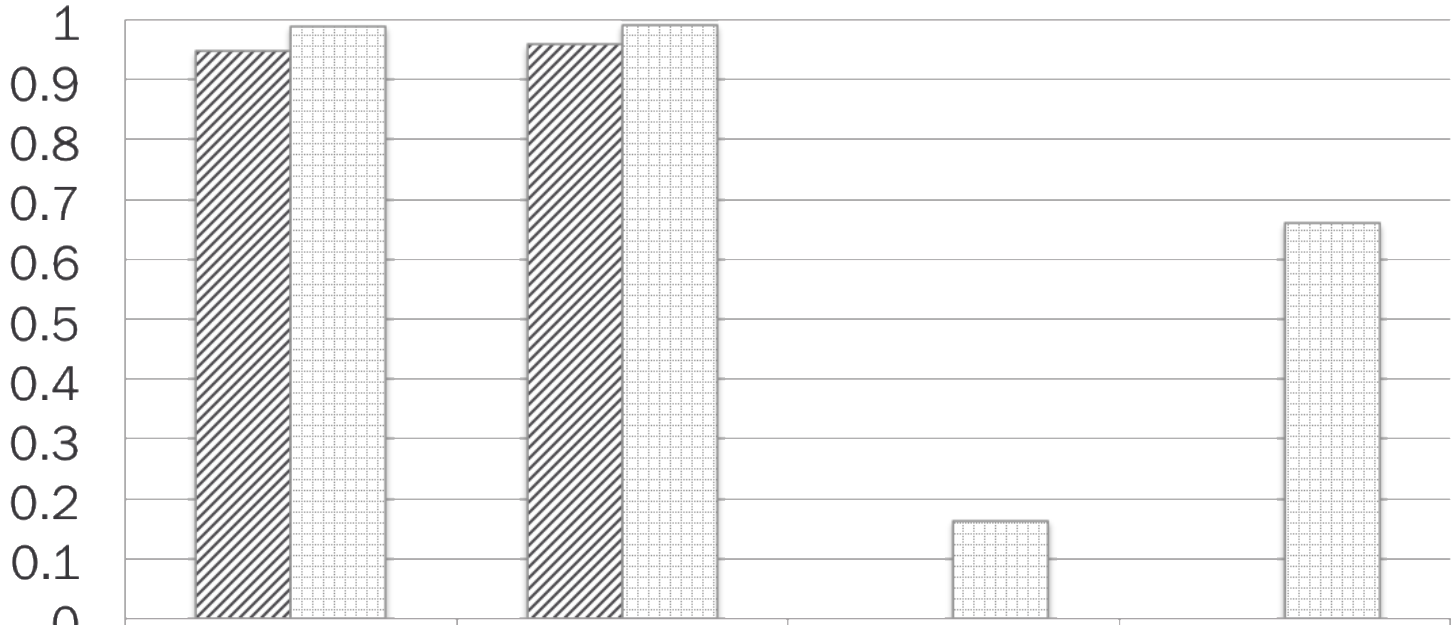


④Result File Download



オントロジーマッピングの自動化

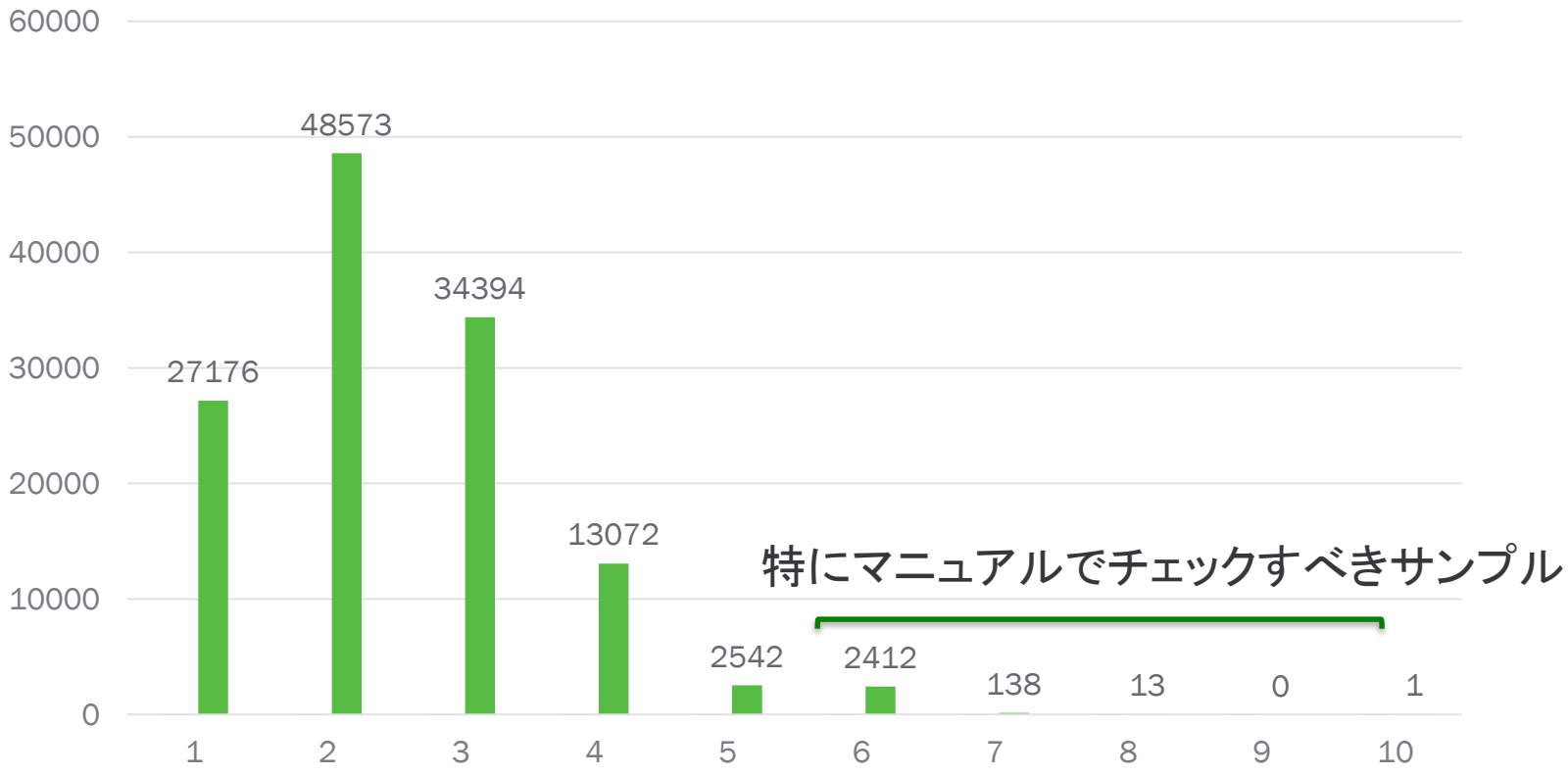
MicrobeDB.jp ver. 1で、SRAのメタゲノムデータをMEOを用いてマニュアルアノテーションした10,755件の由来環境の記述から単語の出現頻度により学習器を構築



	RF	Linear SVM	Poly SVM	Labeled-LDA
Accuracy	0.947	0.958		
F1 measure(micro)2	0.989	0.991	0.164	0.66

MEO auto-annotated count distribution

今回作ったオートアノテーションツールでMicrobeDB.jp ver. 2の
173,359サンプルに対してMEO IDをオートアノテーションした結果



サンプルごとの、オートアノテーションされたMEO IDの数

挑戦的課題

微生物統合DB「MicrobeDB.jp」からの知識発見

さらなる多様なデータの統合化＆超高度化

（統一解析プロトコルによるデータ＆メタデータ連携に基づく高精度データ）

- 環境の物理・化学データ
- リモートセンシングデータ
- 気象データ
- 地質データ
- 河川流域都市の世帯収入、犯罪率、平均寿命
- 抗生物質の使用頻度、院内感染発生率
- 岩石サンプル、ボーリングサンプル

微生物群集を介して「ゆるく」つながっている研究領域や分野において「新たな知識の発見」を促すことができる