ライフサイエンスデータベース統合推進事業
統合化推進プログラム
平成24年度　進捗報告会

ゲノム・メタゲノム情報を基盤とした微生物DBの統合

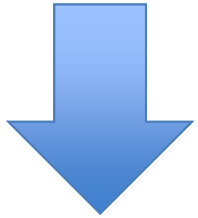
東京工業大学大学院生命理工学研究科
黒川　顕

微生物研究を取り巻く状況

- 微生物は環境と密接に関与し存在する
- 微生物研究はバイオ分野のみならず，他の多くの分野と連携可能
- 微生物研究分野には多様なDBが多数存在する
- 環境との関連性を記述しているDBは未だ存在しない
- ほとんどのDBは専門知識を持っていないバイオ分野以外の人は利用困難

研究開発の目標・ねらい

ゲノム情報を核として様々な微生物学上の知識を統合し、幅広い分野での微生物学の発展に資することのできる「微生物エンサイクロペディア」の構築を目標とする。



微生物学分野のオミックス研究の発展に寄与
データ駆動型研究による新しい仮説の提唱

研究開発メンバー

東京工業大学

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DB、メタデータの構築、オントロジー構築
山田拓司：メタゲノムDBの構築
山本希：オントロジーの構築
吉野弘二，竹原潤一：メタデータDBの構築、オントロジー構築
小西史一：スパコンにおける解析システムの開発および実装

国立遺伝学研究所

中村保一：微生物アノテーションリファレンスの整備と共用化
藤澤貴智：TogoAnnotation構築、モデル微生物情報の高度化
菅原秀明：微生物ゲノム基盤情報資源の共用化
神沼英里：TogoAnnotationの拡張

基礎生物学研究所

内山郁夫：比較ゲノム解析に立脚した微生物ゲノム情報の統合化
千葉啓和：MBGDの統合化

統合データベースセンター（技術アドバイザー）

岡本忍，片山俊明，川島秀一，川本祥子，山本泰智：技術協力

H24年度の当初計画

- 保存菌株情報（NBRC，JCM）のRDF化
- 培地情報オントロジーCMOの整備（w/ DBCLS）
- GTPSのRDF化および各オミックスデータの統合
- MBGDオーソログ遺伝子情報のRDF化
- 放線菌以外のモデル微生物ゲノムアノテーション高度化
- メタゲノムデータのRDF化およびGTPS、RefSeq等との統合
- 各種オントロジーの開発
- 各種アプリケーション、結果表示要素「Stanza」の標準化および開発（w/ DBCLS）

微生物における各種DBを統合化し，環境情報との連携を徹底的に記述した新たなDBの構築

H24年度の実施状況

- セマンティックウェブの技術を活用
- データ間をリンクするためのゲノム情報、オーソログ遺伝子情報、メタゲノム情報の整備
- 全データのRDF化、各データID間のリンク構築
- 各種オントロジーの開発、各データにマッピング
- アノテーション高度化システムの開発
- ユーザ認証システム
- データテンプレート「Stanza」の開発

RDF/OWLファイルを用いてFull RDFなDB「MicrobeDB.jp」の構築

NBRC/JCMの菌株データのRDF化

National Institute of Technology and Evaluation
独立行政法人
製品評価技術基盤機構
Biotechnology Center
NBRC
NITE Biological Resource Center

NBRC No.	NBRC 102086
Scientific Name of this Strain	***Burkholderia multivorans*** Vandamme et al. 1997
Synonymous Name	
Type Strain	
History	NITE <- Meijo Univ. (S. Ichihara, 25)
Other Culture Collection No.	
Other No.	NITE 02208=25
Rehydration Fluid	702
Medium	802
Cultivation Temp.	25 C
Source of Isolation	
Locality of Source	
Country of Origin	Japan
Biosafety Level	
Applications	Phenylacetic acid;degradation
Mating Type	
Genetic Marker	
Plant Quarantine No.	
Animal Quarantine No.	
Herbarium No.	
Restriction	
Comment	
References	
Sequences	16S rDNA

株数：約16,000株
単離元：1,627
培地情報：432種類

ゲノムデータおよびメタゲノムデータ等と統合を目標にRDF化

JCMも対象とする（約14,000株）

NBRCおよびJCMデータ項目とその内容の概観

1. Strain_Number（株番号）
2. Other_Collection_Numbers（他機関での株番号）
3. Name（学名）
4. Organism_Type（生物分類）
5. History_of_Deposit（来歴）
6. Date_of_Isolation（分離日）
7. Isolated_from（分離源）
8. Geographic_Origin（地理）
9. Status（ステータス）
10. Optimum_Temperature_for_Growth（至適培養温度）
11. Maximum_Temperature_for_Growth（最高培養温度）
12. Minimum_Temperature_for_Growth（最低培養温度）
13. Medium（培地）
14. Application（利用法）
15. Literature（文献）

NBRC/JCMにおける総トリプル数

	NBRC株	JCM株
株番号	17,367	13,396
他機関での株番号	62,914	122,286
学名	48,791	51,769
生物分類	45,771	35,350
来歴	59,773	58,442
分離日	0	0
分離源	57,240	46,275
地理	15,338	0
ステータス	4,090	6,004
至適培養温度	16,611	13,396
最高培養温度	746	0
最低培養温度	746	0
培地	18,429	18,787
利用法	1,470	1,316
文献	21,600	15,646

菌株IDとNCBI Taxonomy IDの名寄せ

National Institute of Technology and Evaluation
独立行政法人
製品評価技術基盤機構

NBRC
NITE Biological Resource Center
Biotechnology Center

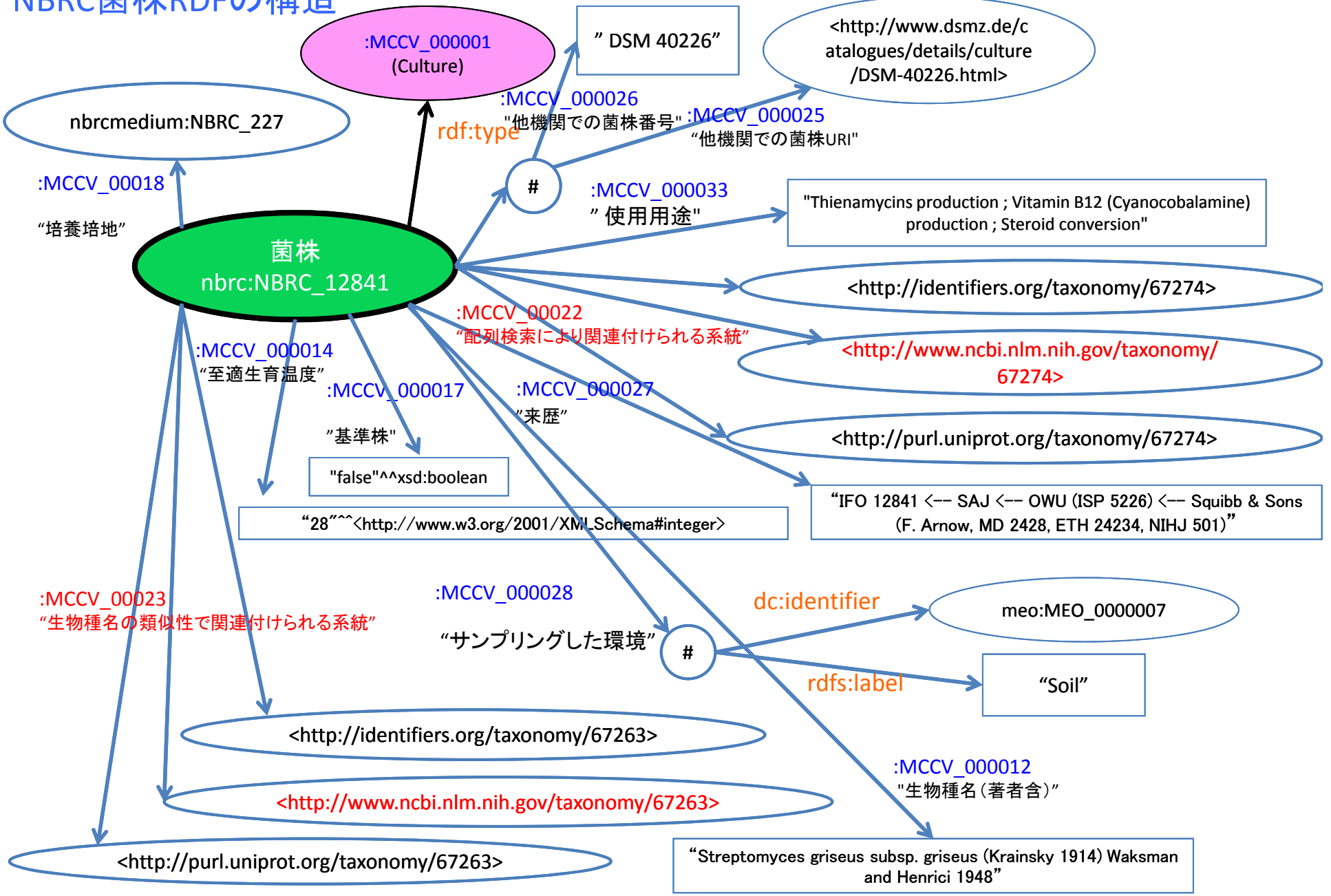
NBRC No.	NBRC 12841
Scientific Name of this Strain	***Streptomyces griseus*** subsp. ***griseus*** (Krainsky 1914) Waksman and Henrici 1948
Synonymous Name	Synonym: *Streptomyces griseus*
Type Strain	
History	IFO 12841 <- SAJ <- OWU (ISP 5226) <- Squibb & Sons (F. Arnow, MD 2428, ETH 24234, NIHJ 501)
Other Culture Collection No.	ATCC 11009=ATCC 23882=CBS 662.68=RIA 1168=ISP 5226=AS 4.1693=BCRC 11815=DSM 40226=JCM 4229=JCM 4623=KCTC 1742=LMG 5967=NCIMB 9625=NRRL B-1806=VKM Ac-747
Other No.	IMET 43659

現状のNBRC/JCMの菌株データは、配列データとは独立しており、
両者の「糊しろ」となるNCBI Taxonomy IDも菌株データには付加されていない

菌株データと配列データを統合するため、
NBRC/JCM菌株IDとNCBI Taxonomy ID間の名寄せが必要

名寄せ結果を含む菌株RDFで用いる様々な語彙のオントロジー、MCCV
(Microbial Culture Collection Vocabulary)を設計し、これを用いて菌株RDFを作成

NBRC菌株RDFの構造



複数の異なる方法による菌株IDとNCBI Taxonomy ID間の名寄せ結果を、predicateで方法ごとに区別可能

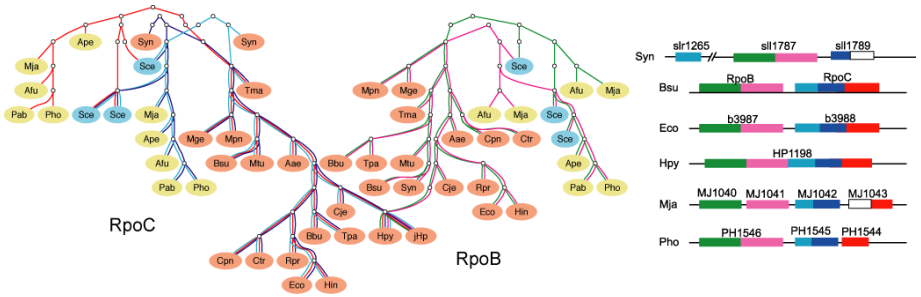
MBGDオーソログ遺伝子情報のRDF化

微生物比較ゲノムデータベースMBGD

微生物ゲノム DB

総当たり ホモロジー DB

オーソログ分類アルゴリズムDomClust



オーソログテーブル

MBGD Ortholog Cluster Table Overview

A Multiple sequence alignment (⊙ ClustalW ○ MAP ○ MAFFT)
M Multiple genome map comparison
H Find homologous clusters
P Similar phylogenetic pattern search (⊙ Correlation coefficient ○ Hamming distance ○ Mutual information)

Analyze the checked clusters
Whole genome map comparison | Merging multiple clusters | Select All | Reset Selection

ClusterID	Name	#species	#genes	Description	Phylogenetic pattern
O1149	folE	328	377	GTP cyclohydrolase I	
O1279	folB	320	352	Dihydroneopterin aldolase	
O1033	ubiE	367	397	Ubiquinone/menaquinone biosynthesis methyltransferase	
O990	folK	371	406	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	
O952	aroB	374	412	3-dehydroquinate synthase	
O1077	atpH	382	388	F0F1 ATP synthase subunit delta	
O1399	hemH	324	332	Ferrochelatase	
O662	sdhB	370	485	Succinate dehydrogenase / fumarate reductase protein, iron-sulfur subunit	
O985	yggS	386	405	Alanine racemase domain-containing protein	
O315	fabF	388	691	3-oxoacyl-(acyl carrier protein) synthase II	

オーソロググループ

Ortholog Cluster

Cluster	1149
Gene	folE
Title	GTP cyclohydrolase I
Size	328 species, 377 genes
Xref-COG	COG0302 GTP cyclohydrolase I [Equivalent, 34/35]
Xref-KEGG	K01495 GTP cyclohydrolase I [EC:3.5.4.16] [Equivalent, 249/258]
Xref-TIGR	TIGR00063 GTP cyclohydrolase I [Subgroup, 172/177]
Xref-GO	GO:0003934 GTP cyclohydrolase I activity [Supergroup, 305/762] GO:0016787 hydrolase activity [Supergroup, 303/173030] GO:0005737 cytoplasm [Supergroup, 287/174310] GO:0046654 tetrahydrofolate biosynthetic process [Supergroup, 287/604] GO:0006730 one-carbon compound metabolic process [Supergroup, 268/4398]

[Summary] [Gene List] [Clustering Tree]

Compare maps | Multi-align | Get seq | Find homologs | Similar phylopat | KEGG pathway

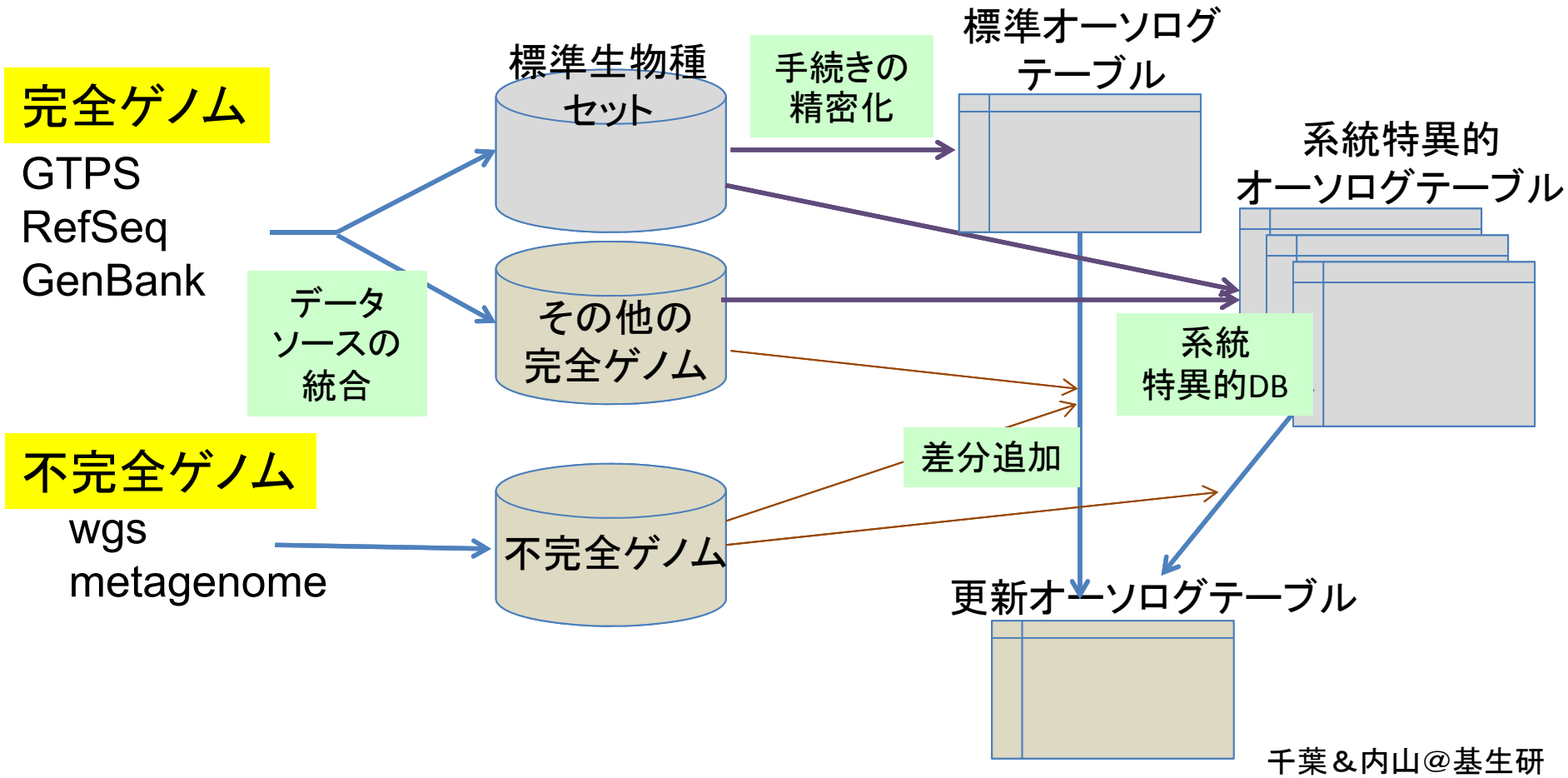
Redraw ☐ Display xref-ortholog ☐ Display xref-motif (cutoff 0.001)
ON OFF

Bacteria ON OFF

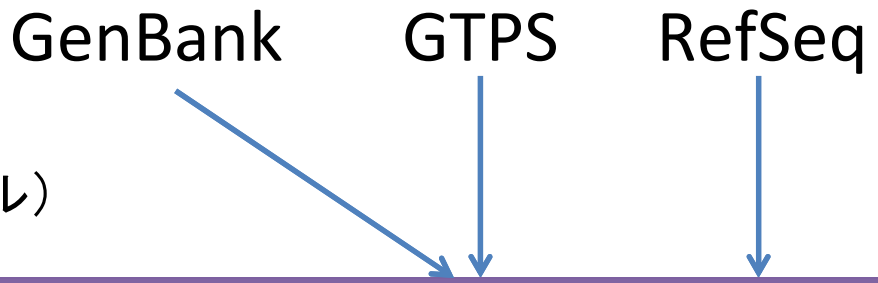
Phylum	Species	Orf ID	Description
Actinobacteria ON OFF	*Acidimicrobium ferrooxidans* DSM 10331	afo:AFER_1912	GTP cyclohydrolase I
	Arcanobacterium haemolyticum DSM 20595	ahe:ARCH_0252	GTP cyclohydrolase I
	Mobiluncus curtisii ATCC 43063	mcu:HMPREF0573_10367	GTP cyclohydrolase I
	Catenulispora acidiphila DSM 44928	cai:CACI_2639	GTP cyclohydrolase I
		cai:CACI_8425	GTP cyclohydrolase I
	Corynebacterium glutamicum ATCC 13032	cgl:NCGL2602	GTP cyclohydrolase I
	Gordonia bronchialis DSM 43247	gbr:GBRO_4016	GTP cyclohydrolase I
	Mycobacterium tuberculosis H37Rv	mtu:RV3609C	GTP cyclohydrolase I
	Nocardia farcinica IFM 10152		

千葉＆内山@基生研

対象ゲノムデータの拡大と効率的なオーソログ解析



データソースの統合による拡張



対応テーブル（染色体レベル）

sp_id	sp_code	GenBank	RefSeq
gm00001	hin	L42023.1	NC_000907.1
gm19792	gm1979	CP003122.1	-

対応テーブル（遺伝子レベル）

name	from	to	GTPS	GenBank	RefSeq
hin:HI0001	2	1021	HINF_KW20:ST73	AAC21680.1	NP_438174.1
hin:HI0002	1214	3013	HINF_KW20:ST725	AAC21681.1	NP_438175.1
hin:PL6	5707	6273	HINF_KW20:PL6	-	-

千葉＆内山@基生研

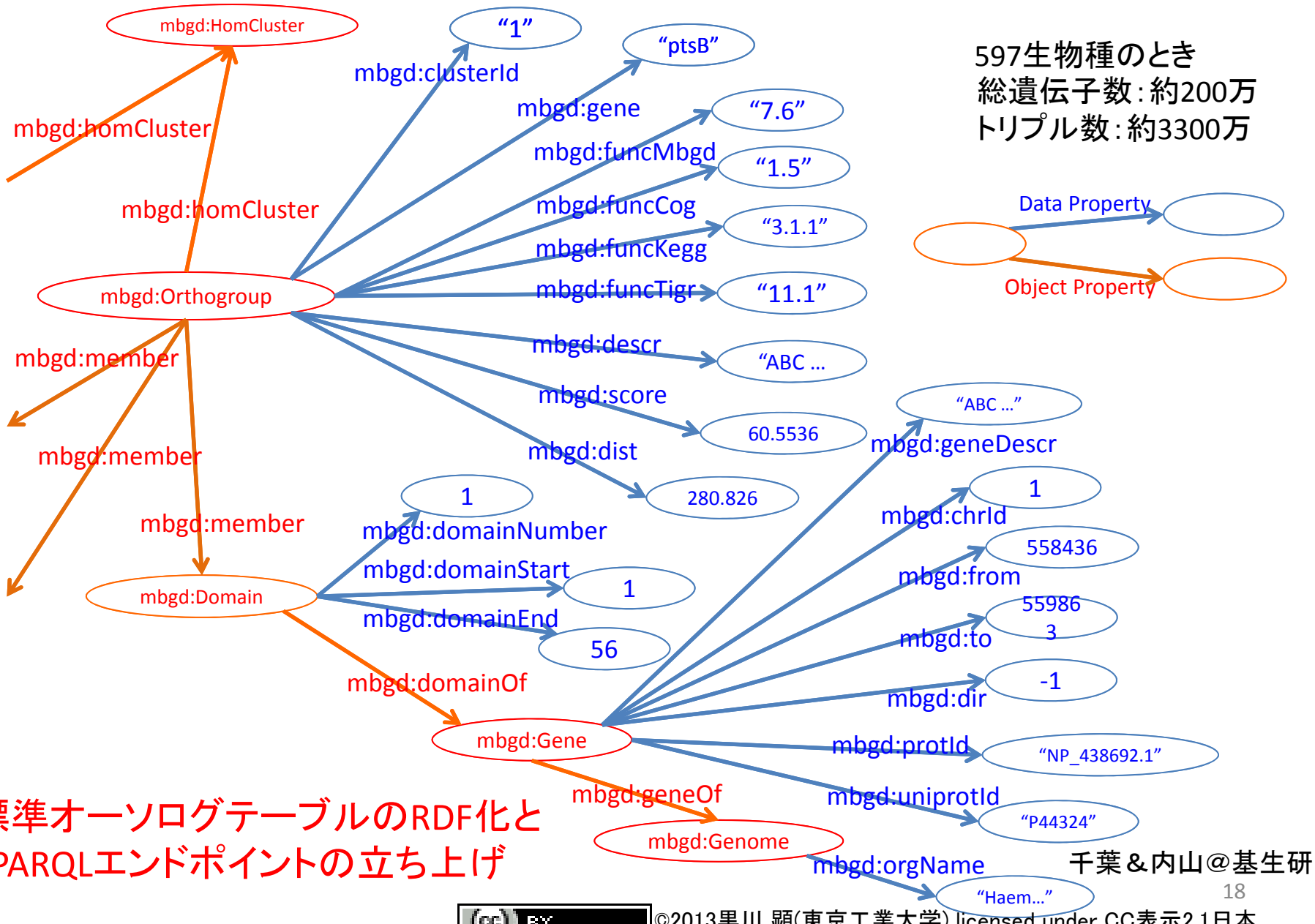
系統特異的オーソログテーブルの作成

Target rank	Rank for representative selection	CoreAlign is available	Num. of taxa
all (default)	genus	No	1
all (extended)	genome	No	1
superkingdom	genus	No	2
phylum	genus	No	13
class	genus	No	18
order	species	No	50
family	species	Yes	58
genus	species	Yes	31
species	genome	Yes	21

代表生物種を6種以上含むタクサの数

千葉＆内山@基生研

MBGDのRDF化



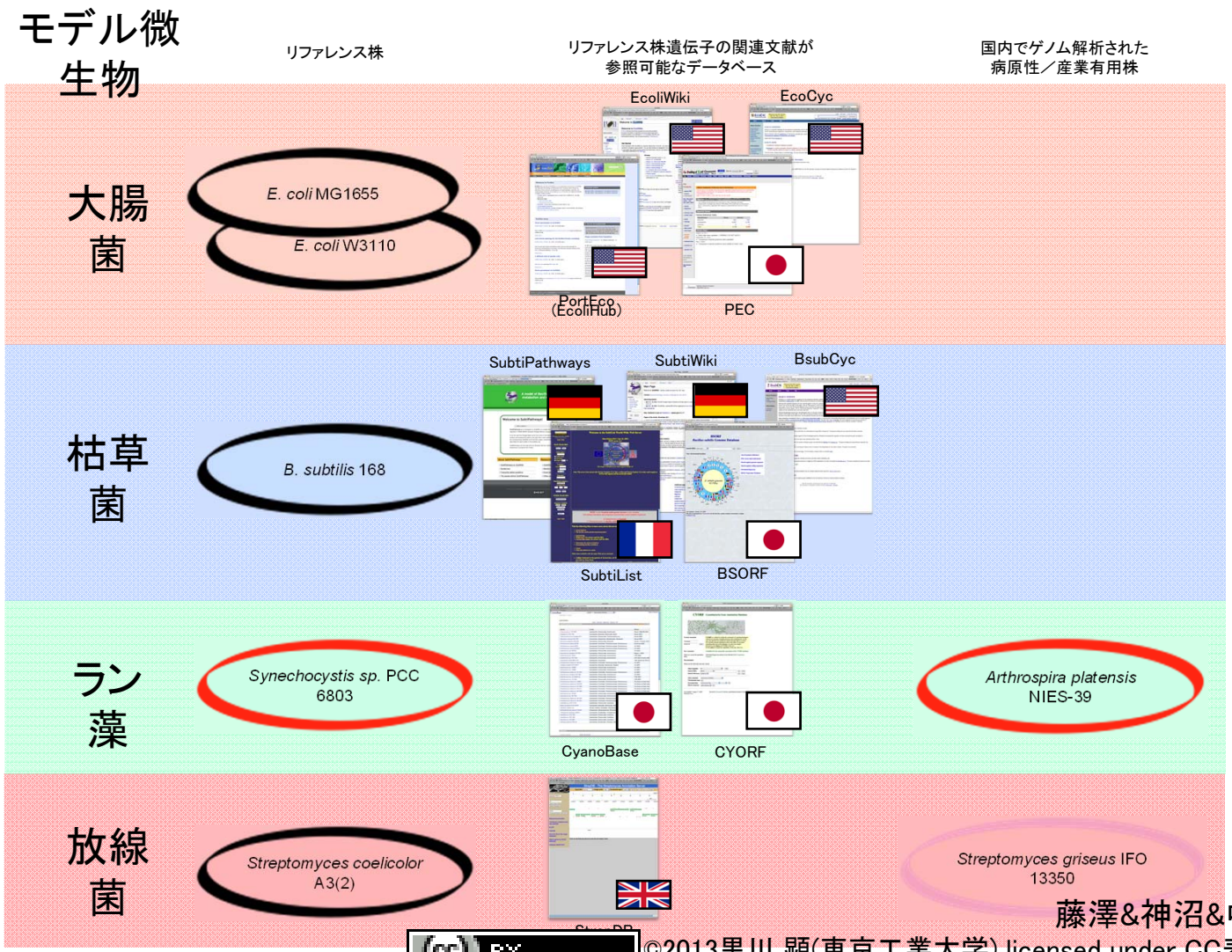
597生物種のとき
総遺伝子数：約200万
トリプル数：約3300万

標準オーソログテーブルのRDF化と
SPARQLエンドポイントの立ち上げ

千葉＆内山@基生研

モデル微生物アノテーション高度化

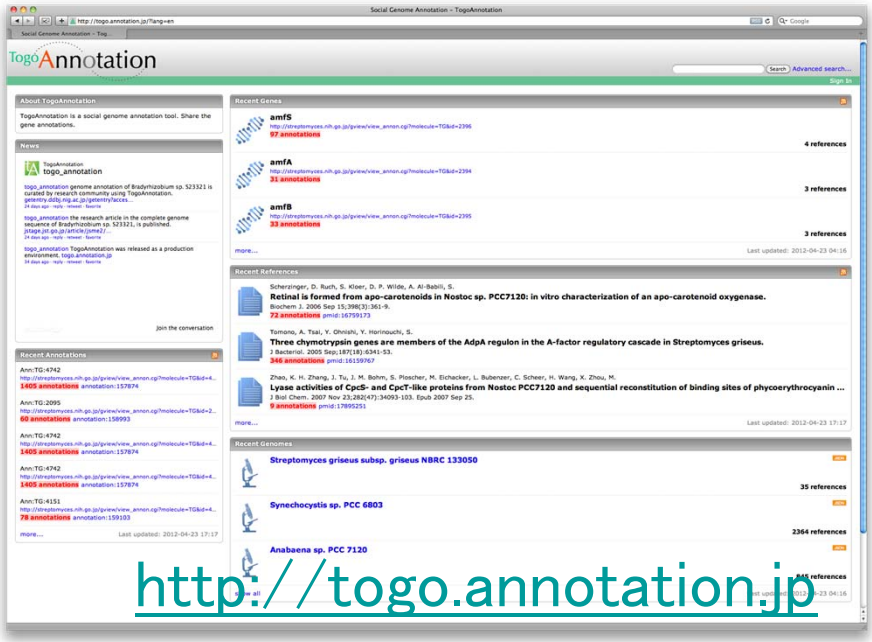
文献情報に基づくモデル微生物ゲノムデータベースの現状



藤澤&神沼&中村@遺伝研

TogoAnnotationの拡張整備

- TogoAnnotation
  ソーシャルブックマークシステムを基盤技術として構成様々なタイプのエンティティを容易にアノテーションすることが可能



KazusaAnnotationから改名（H24.3）

- ☑URL, サービス名称の変更（H23）
- ☑ゲノム、遺伝子、文献単位での再利用を目的としたAPI開発（H23）
- ☑モデル微生物情報（シアノバクテリア、放線菌）のマニュアルキュレーションによる文献リファレンス蓄積（H23,H24）
- ☑OpenID認証と連携したユーザ／グループ認証サービス開発・運用（H24）
- ☑モデル微生物情報（大腸菌、枯草菌）の文献リファレンス蓄積のための整備（H24）
- ☑TogoAnnotationのRDF化（H24）
- ☑モデル微生物情報のRDFを介した情報統合のユースケースプロトタイプ（H24）

モデル微生物のマニュアルキュレーションによる文献リファレンス蓄積：遺伝子情報

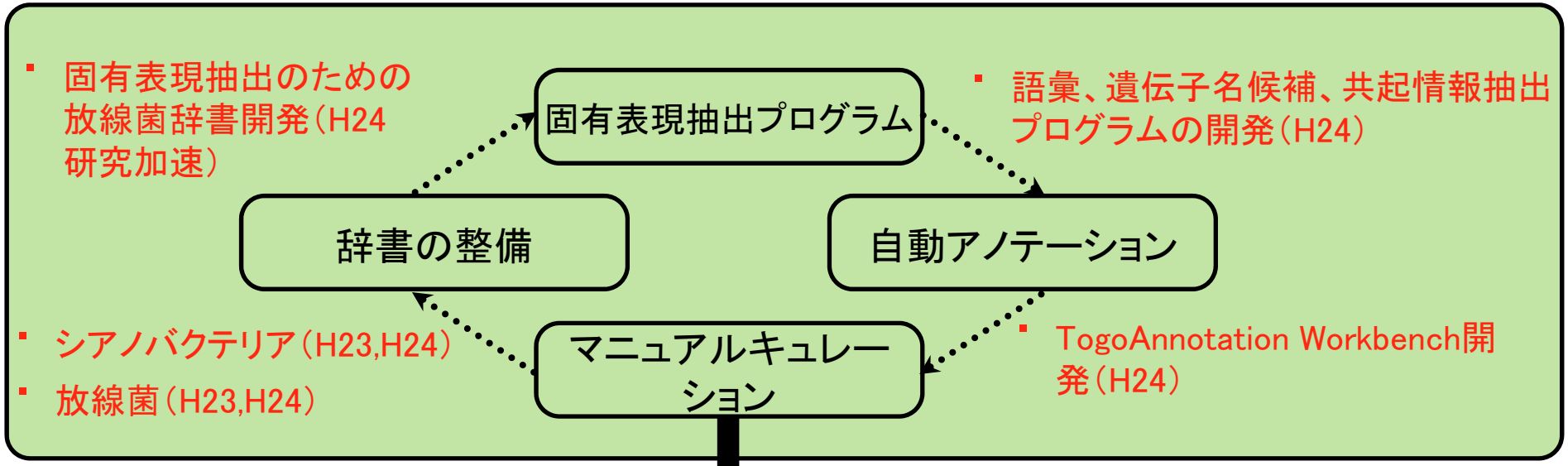
ver. 2013.01.15

モデル微生物	生物種	対象文献数	アノテーション文献数	アノテーション遺伝子数	アノテーションブックマーク数	アノテーションブックマーク数（H23-）
大腸菌	*Escherichia coli* K-12					
枯草菌	*Bacillus subtillis* 168					
ラン藻	*Synechocystis sp.* PCC 6803	2524	2244	3003	74363	7543
	Anabaena sp. PCC 7120	974	929	2685	28031	8196
	Synechococcus elongatus PCC 7942	900	787	763	16356	2105
	Synechococcus sp. PCC 7002	299	251	251	3943	236
	Thermosynechococcus elongatus BP-1	329	252	2534	6659	304
	Chlorobium tepidum TLS	180	143	751	5532	11
	Nostoc punctiforme ATCC 29133	138	143	751	3242	870
	Anabaena variabilis ATCC 29413	130	113	246	1712	121
	Prochlorococcus marinus MED4	66	59	383	2119	335
	Gloeobacter violaceus PCC 7421	49	46	4486	5588	4754
	Prochlorococcus marinus MIT9313	41	40	235	885	166
	Prochlorococcus marinus SS120	40	34	127	527	70
	Arthrospira platensis NIES-39	8	7	258	785	391
	Synechococcus sp. WH8102	1	2	3	7	1
	Synechococcus elongatus PCC 6301	1	1	1	4	4
	Trichodesmium erythraeum IMS101	1	1	1	2	2
放線菌	*Streptomyces coelicolor* A3(2)	2143	21	266	2958	2958
	Streptomyces griseus IFO 13350	169	112	299	12814	12814

TogoAnnotationへモデル微生物17043遺伝子に対して合計165,527アノテーション（H23-24: 8,869遺伝子40,881アノテーション）のブックマークを集積した。H24において、新規に放線菌リファレンスゲノムである *Streptomyces coelicolor* A3(2) のマニュアルキュレーションを開始した。

藤澤&神沼&中村@遺伝研

モデル微生物アノテーションリファレンス集積加速のための整備

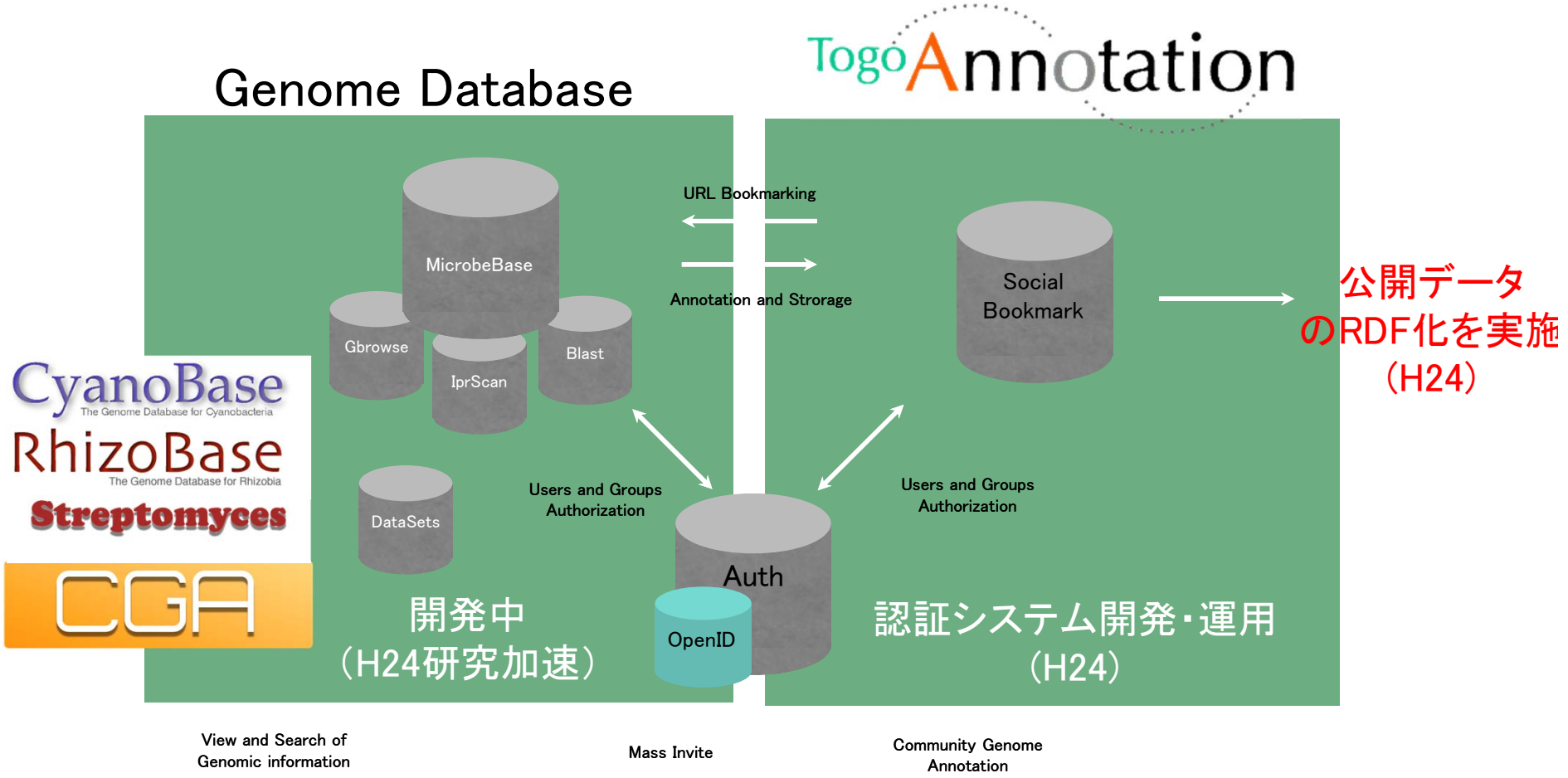


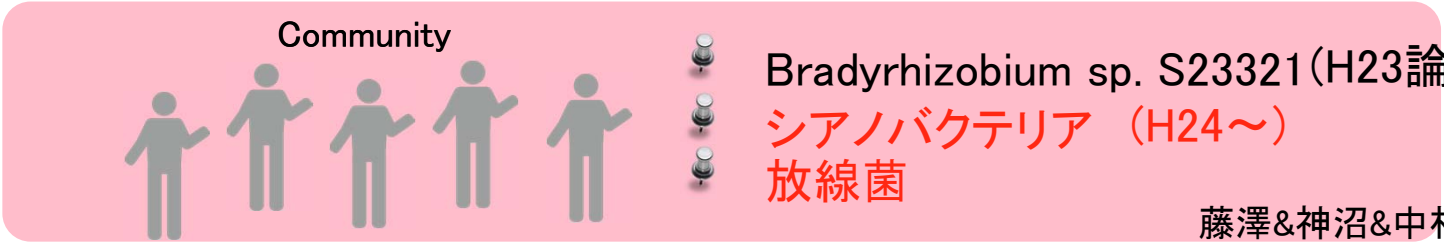
TogoAnnotationへのモデル微生物
アノテーションリファレンス集積

ゲノムデータベースへのマッピング

H24までに各処理系の開発を実施した。H25に実データでの運用を開始する。

ゲノムアノテーションプラットフォームの整備



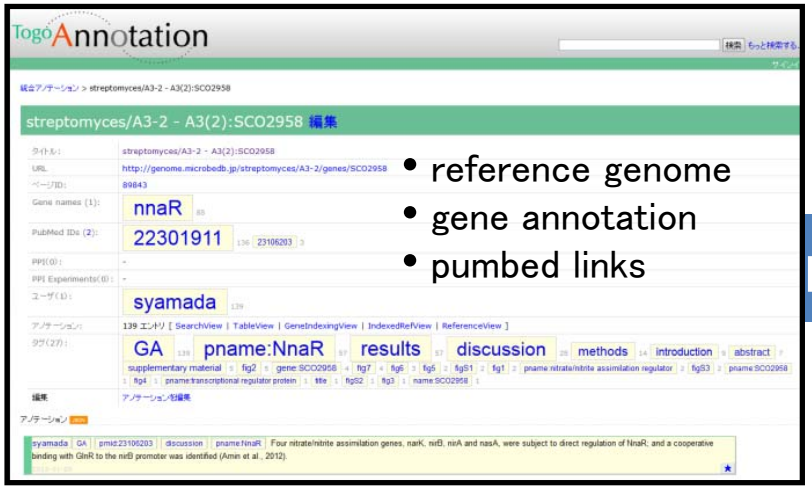


Bradyrhizobium sp. S23321（H23論文公開）
シアノバクテリア （H24～）
放線菌

藤澤&神沼&中村@遺伝研

TogoAnnotationキュレーション情報とNGS多型解析情報の統合

放線菌ゲノム情報統合の事例



- reference genome
- gene annotation
- pumbed links

RDF統合

Sequence Read Archive + DDBJ pipeline

Example: SRA002840

S.coelicolor A3(2)[REF]	All genes	7824
S.lividans TK24	Unmapped genes	599
	Mapped genes with variants(SNP/INDEL)	5795

S.lividans TK24	SNP	27630
	INS	402
	DEL	377

- target strain
- new variants

ゲノム情報　　キュレーション情報　　多型解析情報

gene_id	gene_product	gene_symbol	annotation_id	curated_gene_symbol	reference_count	annotation_count	valiation_id	valiation_type	seqid	location	allele
SCO0613	arginine deiminase		http://togo.annotation.jp/annotations/193815	arcA	1	1	SITK24_000002495	SNP	chr	652849	C
SCO0674	endo-1,4-beta-xylanase		http://togo.annotation.jp/annotations/197580	xysA	1	4	SITK24_000002703	SNP	chr	713507	C
SCO0674	endo-1,4-beta-xylanase		http://togo.annotation.jp/annotations/197580	xysA	1	4	SITK24_000002702	SNP	chr	712924	T
SCO0674	endo-1,4-beta-xylanase		http://togo.annotation.jp/annotations/197580	xysA	1	4	SITK24_000002704	SNP	chr	713822	G
SCO0713	lipase		http://togo.annotation.jp/annotations/191409	lipA	1	1	SITK24_000002874	SNP	chr	756500	G
SCO0713	lipase		http://togo.annotation.jp/annotations/191409	lipA	1	1	SITK24_000002875	SNP	chr	756716	A
SCO0713	lipase		http://togo.annotation.jp/annotations/191409	lipA	1	1	SITK24_000002876	SNP	chr	756797	C
SCO0713	lipase		http://togo.annotation.jp/annotations/191409	lipA	1	1	SITK24_000002877	SNP	chr	757109	C
SCO1483	carbamoyl phosphate synthase large subunit	carB	http://togo.annotation.jp/annotations/193838	pyrA	1	5	SITK24_000005845	SNP	chr	1584905	G
SCO1483	carbamoyl phosphate synthase large subunit	carB	http://togo.annotation.jp/annotations/193838	pyrA	1	5	SITK24_000005847	SNP	chr	1586465	C
SCO1483	carbamoyl phosphate synthase large subunit	carB	http://togo.annotation.jp/annotations/193838	pyrA	1	5	SITK24_000005841	SNP	chr	1584152	A
SCO1483	carbamoyl phosphate synthase large subunit	carB	http://togo.annotation.jp/annotations/193838	pyrA	1	5	SITK24_000005842	SNP	chr	1584191	G
SCO1483	carbamoyl phosphate synthase large subunit	carB	http://togo.annotation.jp/annotations/193838	pyrA	1	5	SITK24_000005843	SNP	chr	1584359	A
SCO1483	carbamoyl phosphate synthase large subunit	carB	http://togo.annotation.jp/annotations/193838	pyrA	1	5	SITK24_000005844	SNP	chr	1584890	A
SCO1483	carbamoyl phosphate synthase large subunit	carB	http://togo.annotation.jp/annotations/193838	pyrA	1	5	SITK24_000005846	SNP	chr	1586213	G
SCO1484	carbamoyl phosphate synthase small subunit		http://togo.annotation.jp/annotations/193839	pyrAa	1	2	SITK24_000005848	SNP	chr	1588303	T
SCO1486	dihydroorotase	pyrC	http://togo.annotation.jp/annotations/193846	pyrC	1	1	SITK24_000005849	SNP	chr	1589698	G
SCO1486	dihydroorotase	pyrC	http://togo.annotation.jp/annotations/193846	pyrC	1	1	SITK24_000005850	SNP	chr	1589974	G
SCO1486	dihydroorotase	pyrC	http://togo.annotation.jp/annotations/193846	pyrC	1	1	SITK24_000005851	SNP	chr	1590196	C
SCO1487	aspartate carbamoyltransferase catalytic subunit	pyrB	http://togo.annotation.jp/annotations/193841	pyrB	1	8	SITK24_000005852	SNP	chr	1590420	G
SCO1487	aspartate carbamoyltransferase catalytic subunit	pyrB	http://togo.annotation.jp/annotations/193841	pyrB	1	8	SITK24_000005853	SNP	chr	1590854	A
SCO1513	GTP pyrophosphokinase		http://togo.annotation.jp/annotations/194993	relA	1	3	SITK24_000005907	SNP	chr	1618230	G
SCO1513	GTP pyrophosphokinase		http://togo.annot		1	3	SITK24_000005908	SNP	chr	1618512	C
SCO1513	GTP pyrophosphokinase		http://togo.annot		1	3	SITK24_000005909	SNP	chr	1618563	G

オミックスデータとゲノム情報との統合イメージ

研究コミュニティ提供の *S. griseus* オミックスデータ

H24研究加速により実施

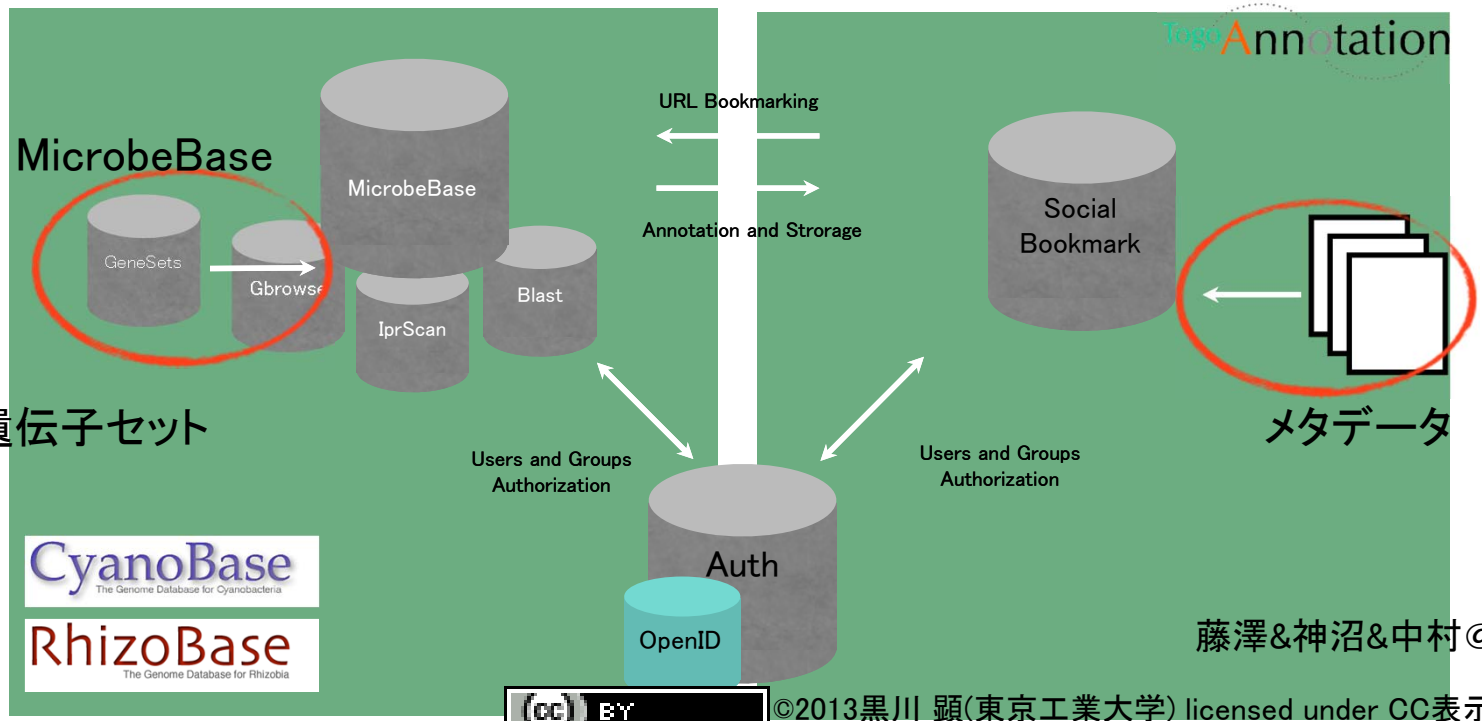
マイクロアレイ	Gbrowse
ChIP-seq解析	新規
プロテオーム解析 (MS)	新規
プロテオーム解析 (二次元電気泳動)	KazusaWiki
RNA-seq	新規
転写解析	新規
遺伝子破壊株	TogoAnnotation
small RNA	Gbrowse
ゲルシフトアッセイとフットプリント	新規
文献情報(遺伝子グループ)	新規

→

- 新規データの格納方法
- サンプル情報などのメタデータ
- 遺伝子セットデータ (.gmt形式)

↓

ゲノムデータベースへのマッピングを可能に



藤澤&神沼&中村@遺伝研

メタゲノムデータのRDF化および GTPS、RefSeq等との統合

メタゲノムメタデータの集計結果

	サンプル数	メタデータのカテゴリー数	メタデータカテゴリーの例
ヒト共生細菌群集	72,236 (18,224)	85	Age , Sex ,Disease stage , Country , Body Habitat , Diet 等
環境細菌群集	6,356 (5,827)	627	pH , Temperature , Wind Speed , Dissolved Oxygen 等

( )内は実際に配列が存在したエントリ （2012年6月5日時点）

	Age	Body Habitat	Body Site	Collection Date	Country	Disease Stage
Sample 1	22	Feces		2008		Obese
Sample 2					Japan	
Sample 3			Scalp			
Sample 4			Skin		USA	
Sample 5	1years		Gut	2011/8/8		Healthy

登録されているカテゴリーや値の語彙は登録者によってバラバラ

微生物の生息環境についてのオントロジーMEOを利用して、メタデータの記述を行った

配列取得可能なSRS IDの内訳

ヒト共生細菌群集 (計 18,224)

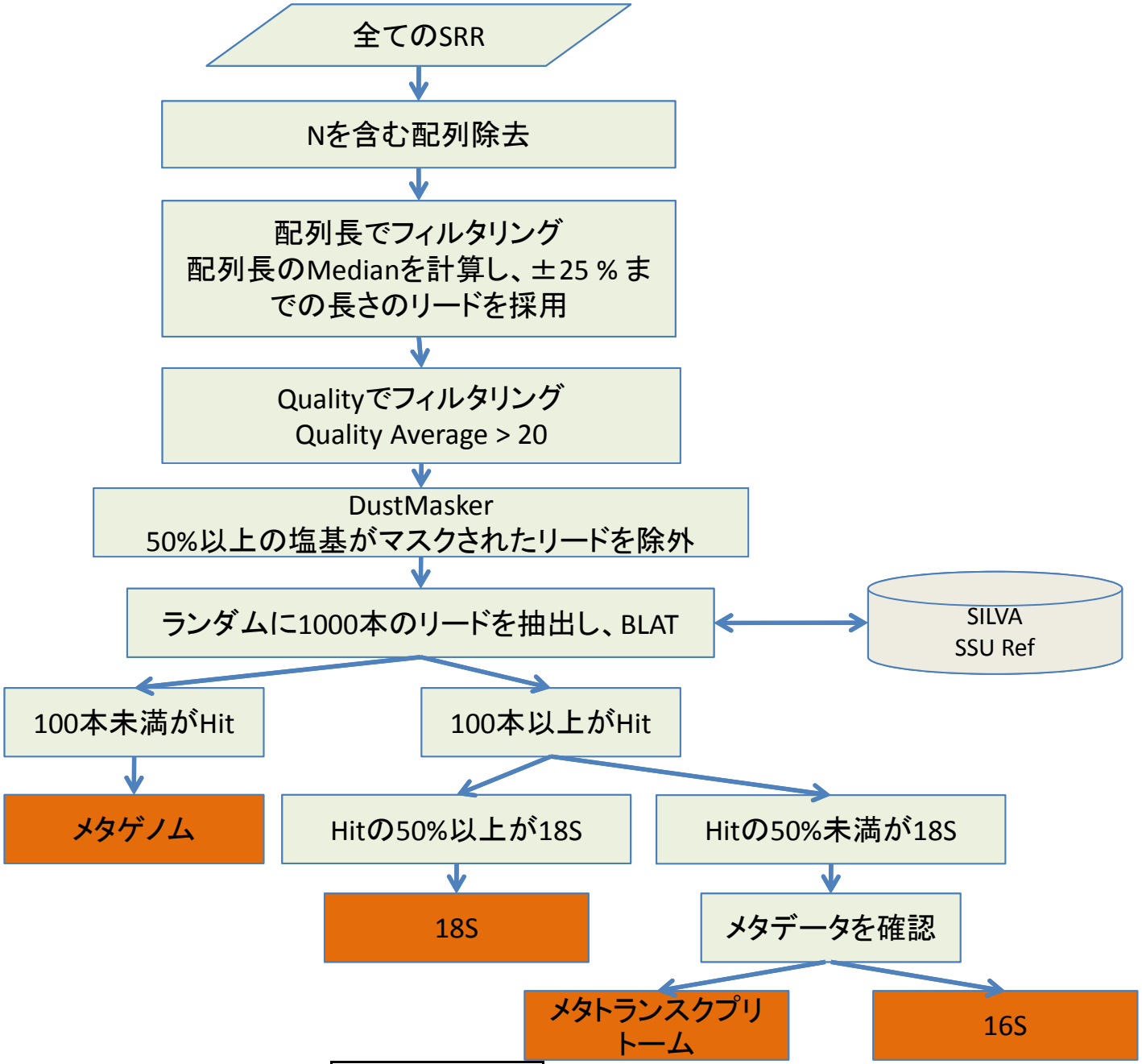
サンプルの種類	SRS IDの数
human metagenome	16,185
human gut metagenome	766
human oral metagenome	504
Homo sapiens	405
human skin metagenome	363
human lung metagenome	1
合計	18,224

配列取得可能なSRS IDの内訳

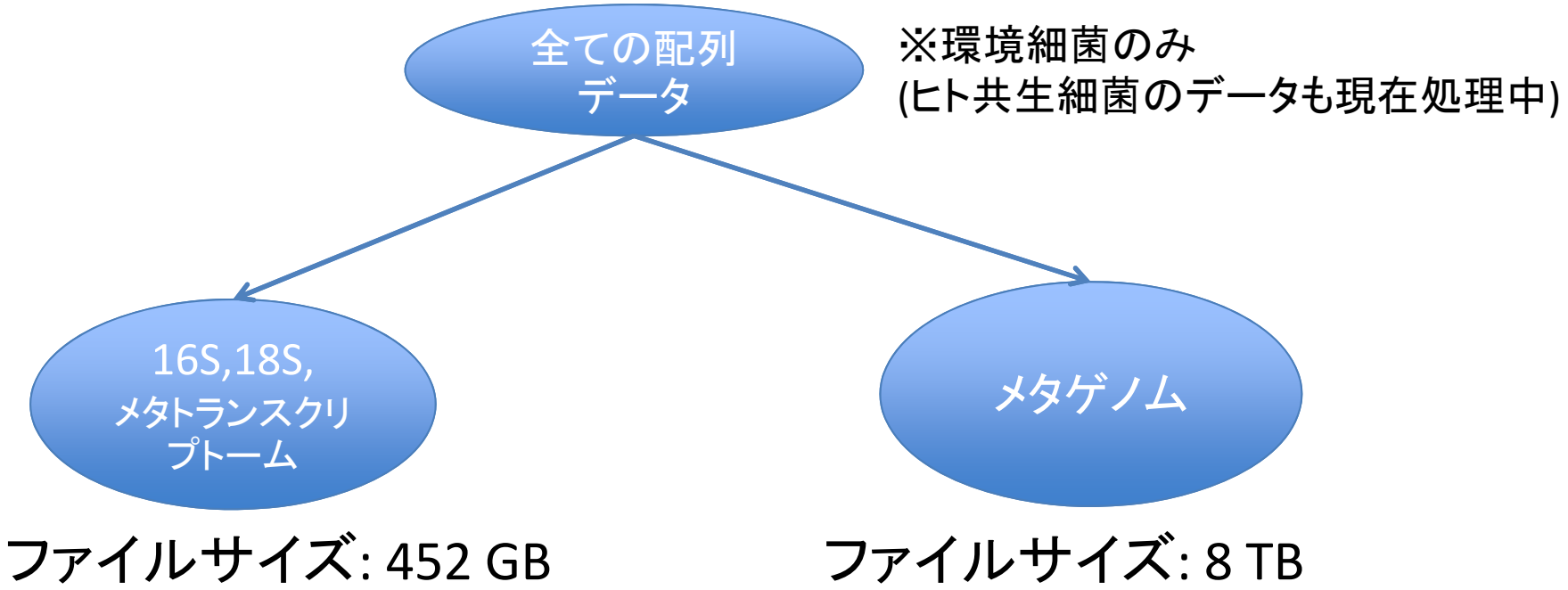
環境細菌群集 (計 5,827)

サンプルの種類	SRS IDの数
marine metagenome	1,212
mouse gut metagenome	985
hydrocarbon metagenome	480
soil metagenome	469
freshwater metagenome	419
gut metagenome	395
organismal metagenomes	210
metagenome sequence	198
bioreactor sludge metagenome	132
marine sediment metagenome	98
sediment metagenome	79
:	:
合計	5,827

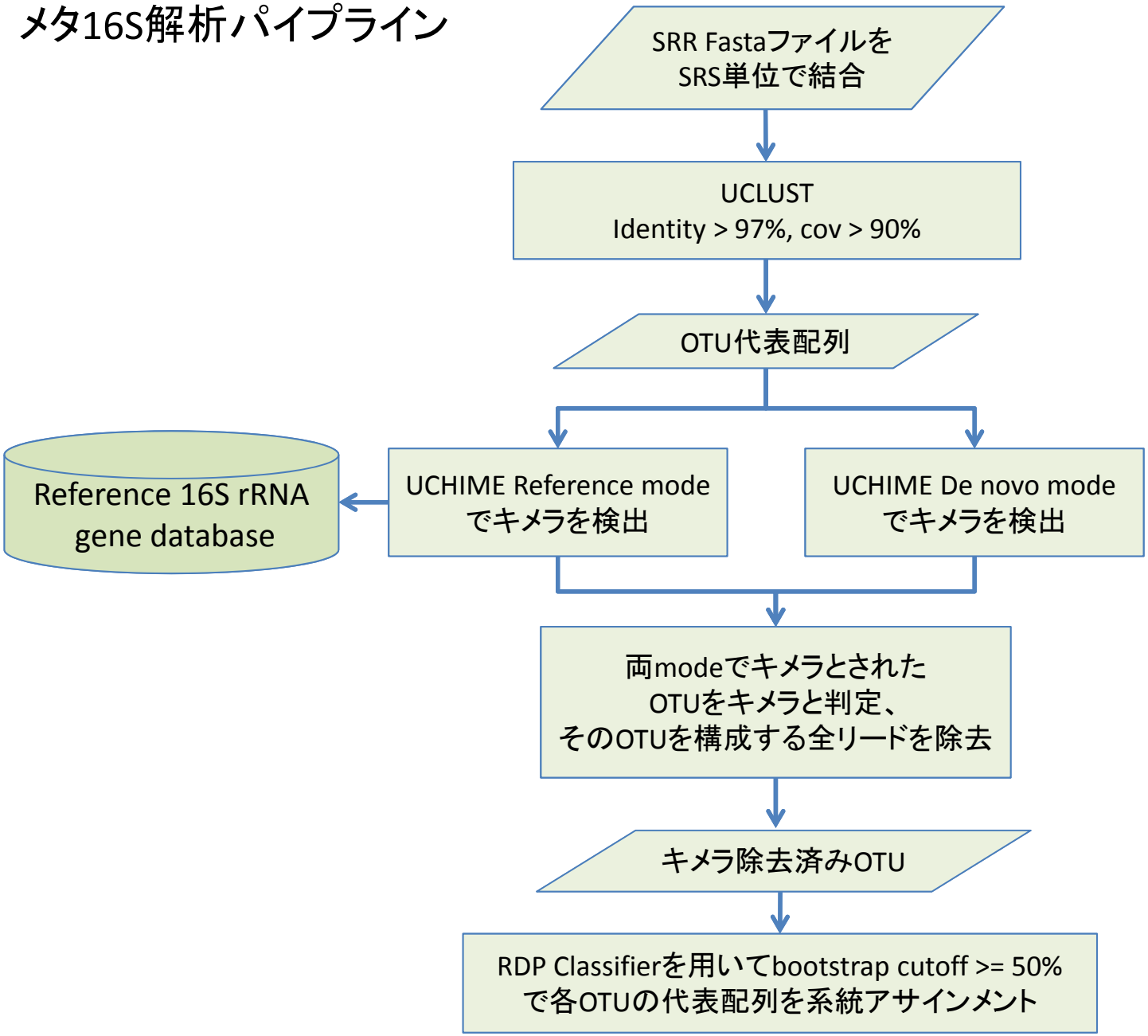
配列の前処理 (目的: メタゲノムとメタ16Sデータの区別 & 高精度配列データの抽出)



前処理によって区別された配列データの内訳

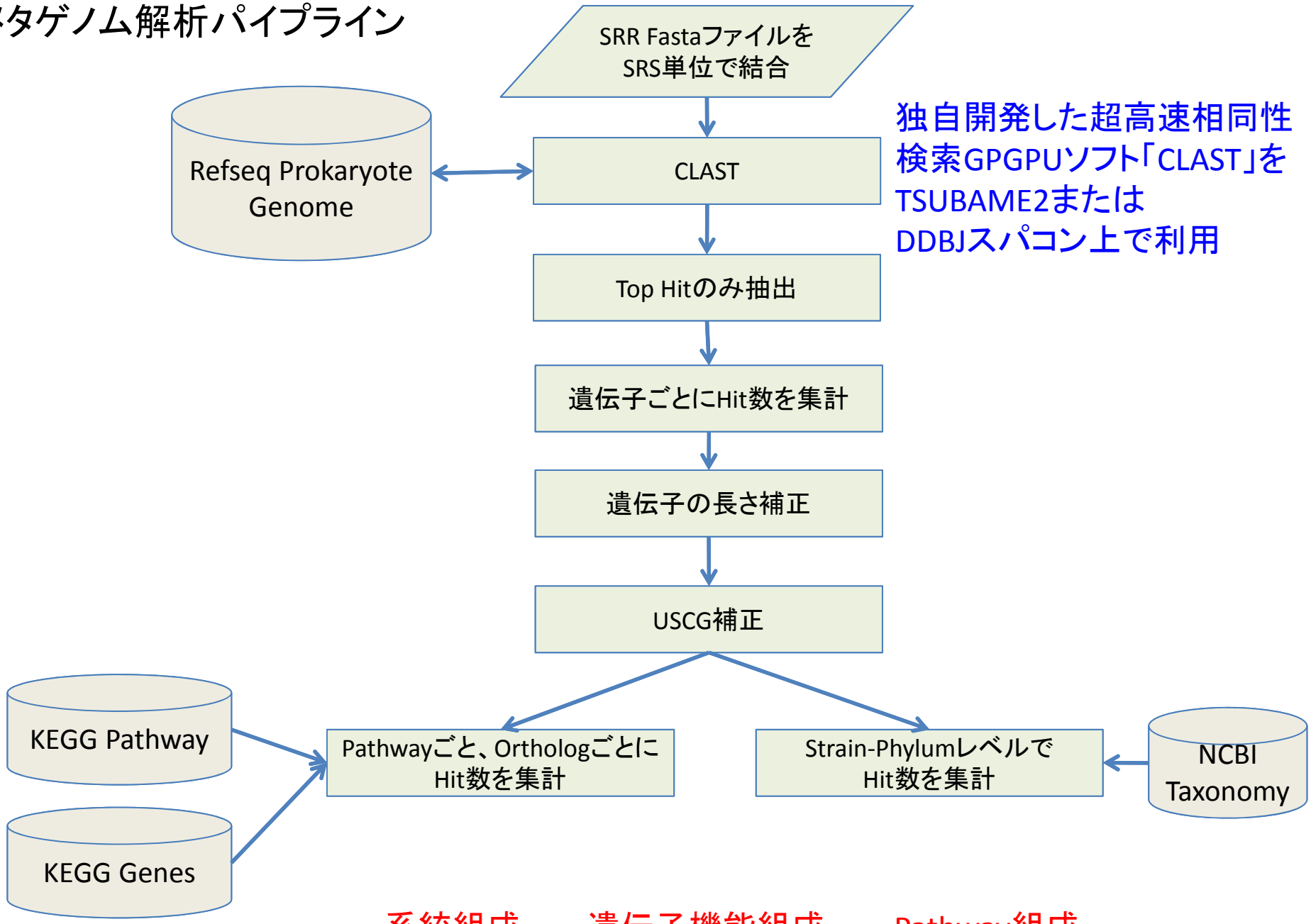


メタ16S解析パイプライン



Genus-Phylumレベルの系統組成

メタゲノム解析パイプライン



独自開発した超高速相同性検索GPGPUソフト「CLAST」をTSUBAME2またはDDBJスパコン上で利用

系統組成　遺伝子機能組成　Pathway組成

文献からの環境メタデータの抽出

SRAへの配列登録時、大部分のメタデータの記述は任意であるため、論文には記述されているが、SRAのXMLには記述されていないメタデータが大量に存在

生息環境の情報と微生物の遺伝子および系統組成の情報を統合して解析し、より多くの知見を得るためには、論文に記述されているメタデータの追加が必要

まず、論文のpubmed IDとSRS IDの対応付けを行う必要がある

2つの方法でpubmed IDとSRS IDの対応付けを行った

①SRAs (http://sra.dbcls.jp/cgi-bin/publication.cgi)の利用
②NCBIのftpサイトからダウンロードしたSRAのXMLファイルの利用

文献からの環境メタデータの抽出

環境細菌群集のメタゲノム・メタ16S解析についての113報の論文を取得
SRAの2,286/5,827サンプル (39.2%)と紐付けられた。

取得した論文の内訳

	論文数
marine metagenome	29
soil metagenome	11
freshwater metagenome	6
mouse gut metagenome	5
sediment metagenome	4
marine sediment metagenome	4
:	:

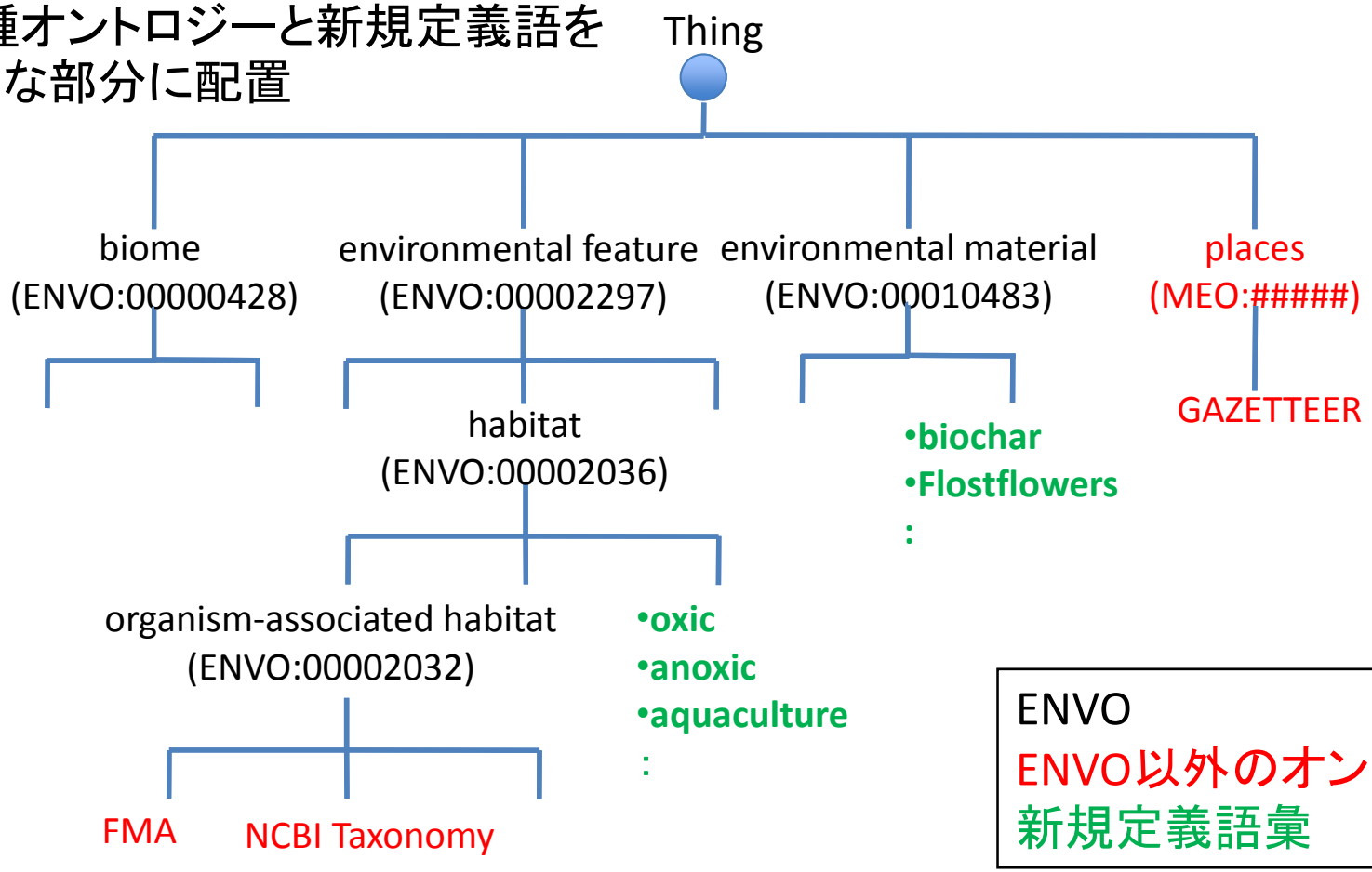
ヒト共生細菌群集のメタゲノム・メタ16S解析についての30報の論文を取得
SRAの12,365/18,224サンプル (67.9%)と紐付けられた。

各種オントロジーの開発

Metagenome/Microbes Environmental Ontology (MEO)

微生物の生息環境メタデータのオントロジー

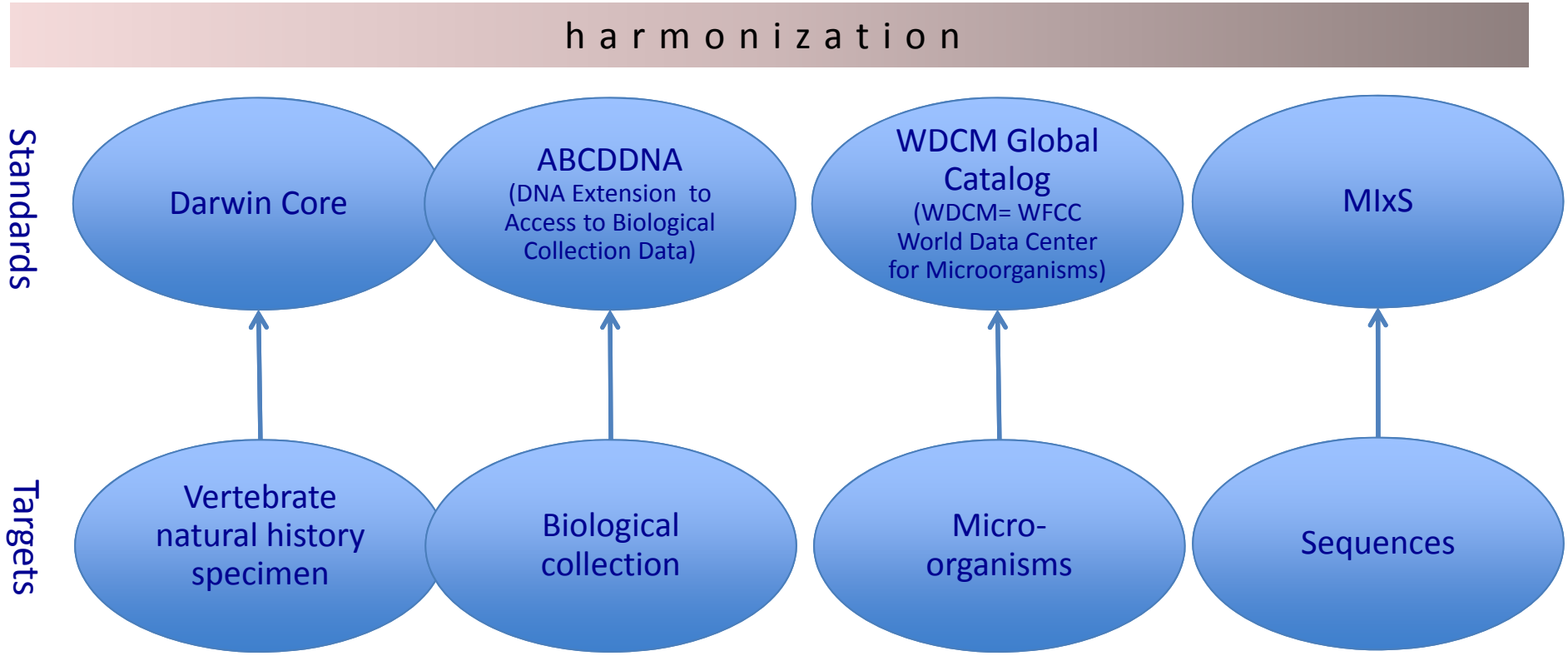
- ENVOを基礎とした構造
- 各種オントロジーと新規定義語を適切な部分に配置



ENVO
ENVO以外のオントロジー
新規定義語彙

BioPortal http://bioportal.bioontology.org/ 及び
プロジェクトページ http://mdb.bio.titech.ac.jp/meo/ でMEOを公開

Further Extension of GBIF standards

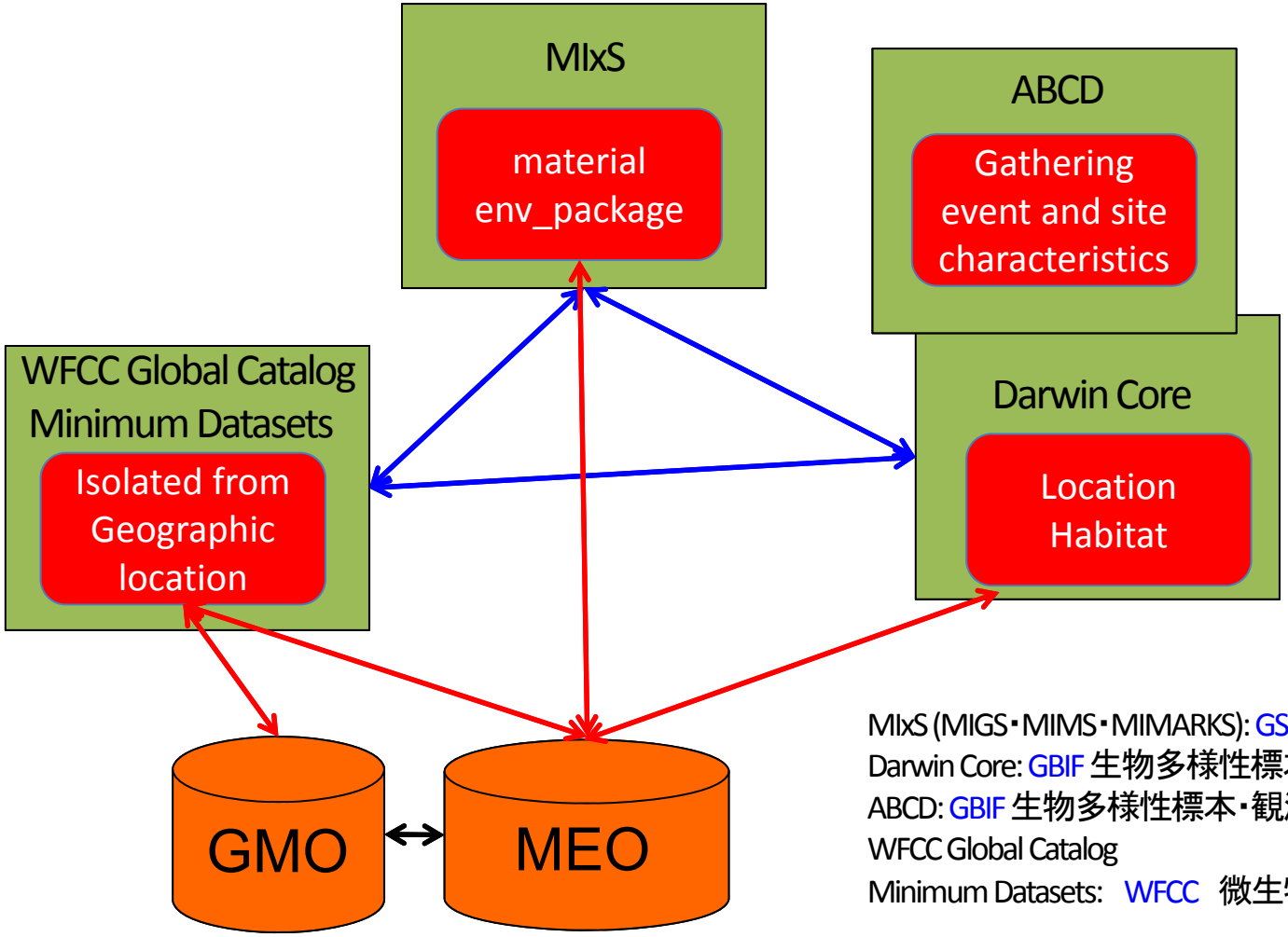


Sequences, sequences and sequences, e.g.,

--- Each GridION node and cartridge is initially designed to deliver tens of Gb of sequence data per 24 hour period, - snip -- Nanopore technology and the GridION system are well suited to miniaturisation of the sensing device. The MinION is a small., portable, single-molecule analysis device that is compatible with Oxford Nanopore's DNA sequencing and protein analysis techniques.--- --- Oxford Nanopore intends to commercialise GridION and MinION directly to customers within 2012 ---

生物資料のゲノム情報 = その生物自体のゲノム + 共生細菌のメタゲノム

国際データ標準化グループとの連携



MIxS (MIGS・MIMS・MIMARKS): GSC ゲノム・メタゲノム
Darwin Core: GBIF 生物多様性標本・観測
ABCD: GBIF 生物多様性標本・観測
WFCC Global Catalog
Minimum Datasets: WFCC 微生物保存株

DwC* - **MIxS**** alignment workshop

27 – 29 February 2012

Oxford e-Research Centre, 7 Keble Road, Oxford, OX1 3QG

*DwC = Darwin Core

**MIxS = Minimum Information of {Genome|Metagenome|Marker gene} Sequence

Standards in Genomic Sciences

An Open Access Journal of the Genomic Standards Consortium

Standards in Genomic Sciences (2012) 7:166-170 DOI:10.4056/sigs.3166513

Meeting Report: Hackathon-Workshop on Darwin Core and MIxS Standards Alignment (February 2012)

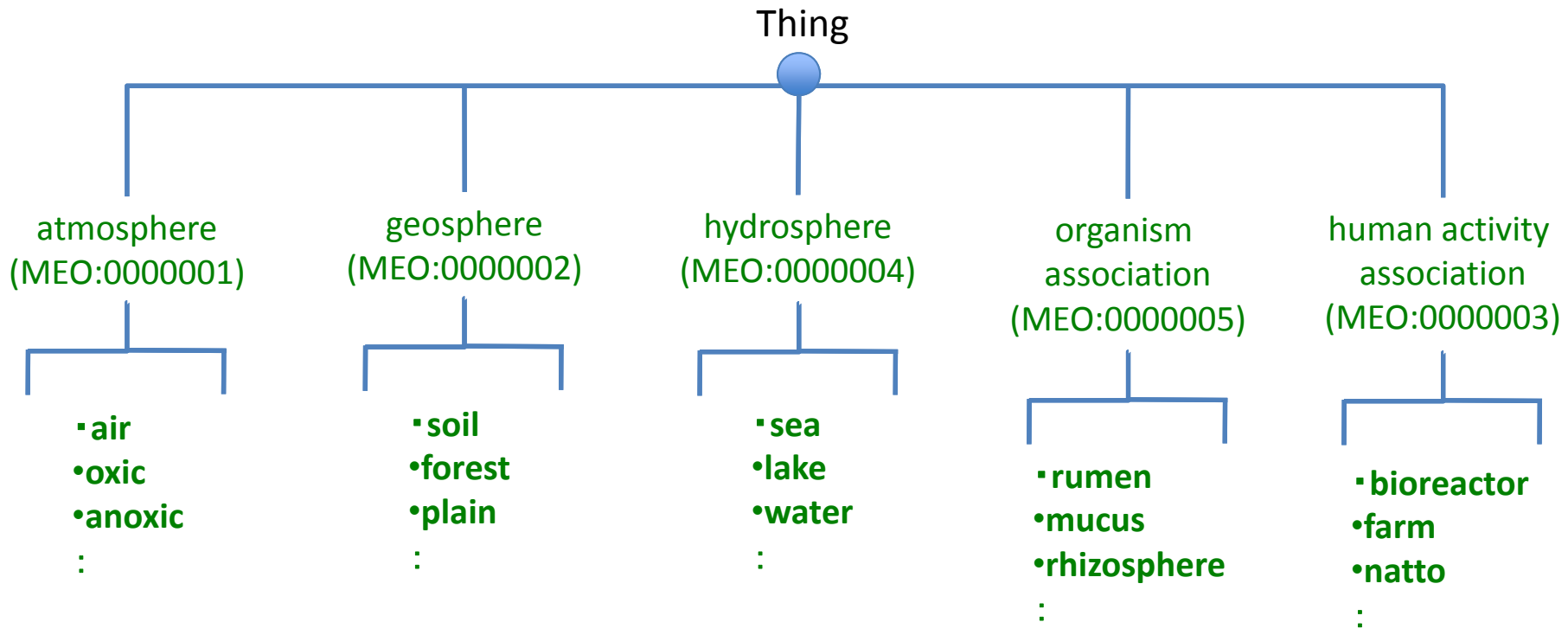
Éamonn Ó Tuama[1], John Deck[2], Gabriel Dröge[3], Markus Döring[4], Dawn Field[5], Renzo Kottmann[6], Juncai Ma[7], Hiroshi Mori[8], Norman Morrison[9,10], Peter Sterk[11], Hideaki Sugawara[12], John Wieczorek[13], Linhuan Wu[14], Pelin Yilmaz[15]

5. The MicrobeDB.jp project including MEO (http://mdb.bio.titech.ac.jp/meo/about_meo) coordinates microbial genomic and metagenomic data, and is supported by NBDC (National Bioscience Database Center, http://biosciencedbc.jp/nbdc.cgi?lng=en), Japan.

http://standardsingenomics.org/index.php/sigen/article/view/sigs.3166513/779

Metagenome/Microbes Environmental Ontology (MEO) Ver. 0.3

MEO is an application ontology for MicrobeDB.jp, not a domain ontology (BioHackathon 2012)



177 terms

MEO version 0.3 contains terms of GOLD complete genome metadata and SRA metagenome metadata by manually picking up terms related to microbial environment

微生物が引き起こす感染症・症状についての オントロジー PDO-SYMP の構築

感染症：病原体（主に微生物）の侵入により引き起こされる病気

既知のゲノム・メタゲノムデータには感染症関連のデータが多数存在

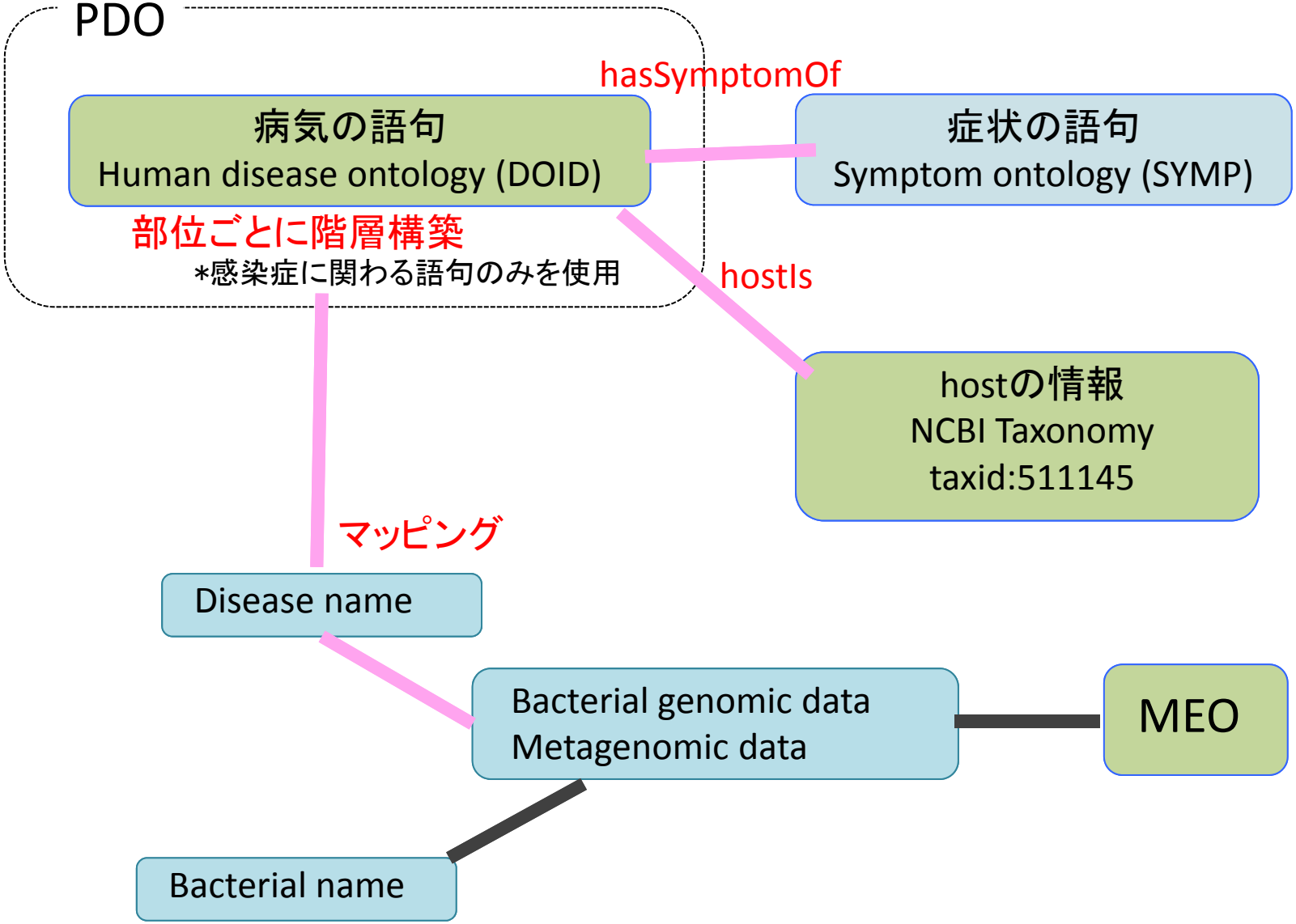
- 原因不明疾患の原因菌推定、微生物・ヒトor動植物間相互作用の解明等の目的にMicrobeDB.jpを利用できるようにしたい
- メタゲノムデータにも対応させたい

MicrobeDB.jpを感染症研究に活用可能にするためには…

感染症に関連する語句（原因菌、病名、症状）
の意味的関連性を明確にする→オントロジーの構築

オントロジー構造（全体）

名称：Pathogenic Disease Ontology (PDO)



ゲノムアノテーション標準のためのINSDCオントロジー開発

- INSDC配列エントリー（GenBank, ENA, DDBJ 形式）のRDF化を目的としたオントロジー開発
- 国内版BH12.12においてDBCLSと連携してINSDC Feature Table Definition (version 10.2)をOWLで定義
- Genomic Standards Consortium (GSC15; April 22-24, 2013. NIH, USA) にて発表予定

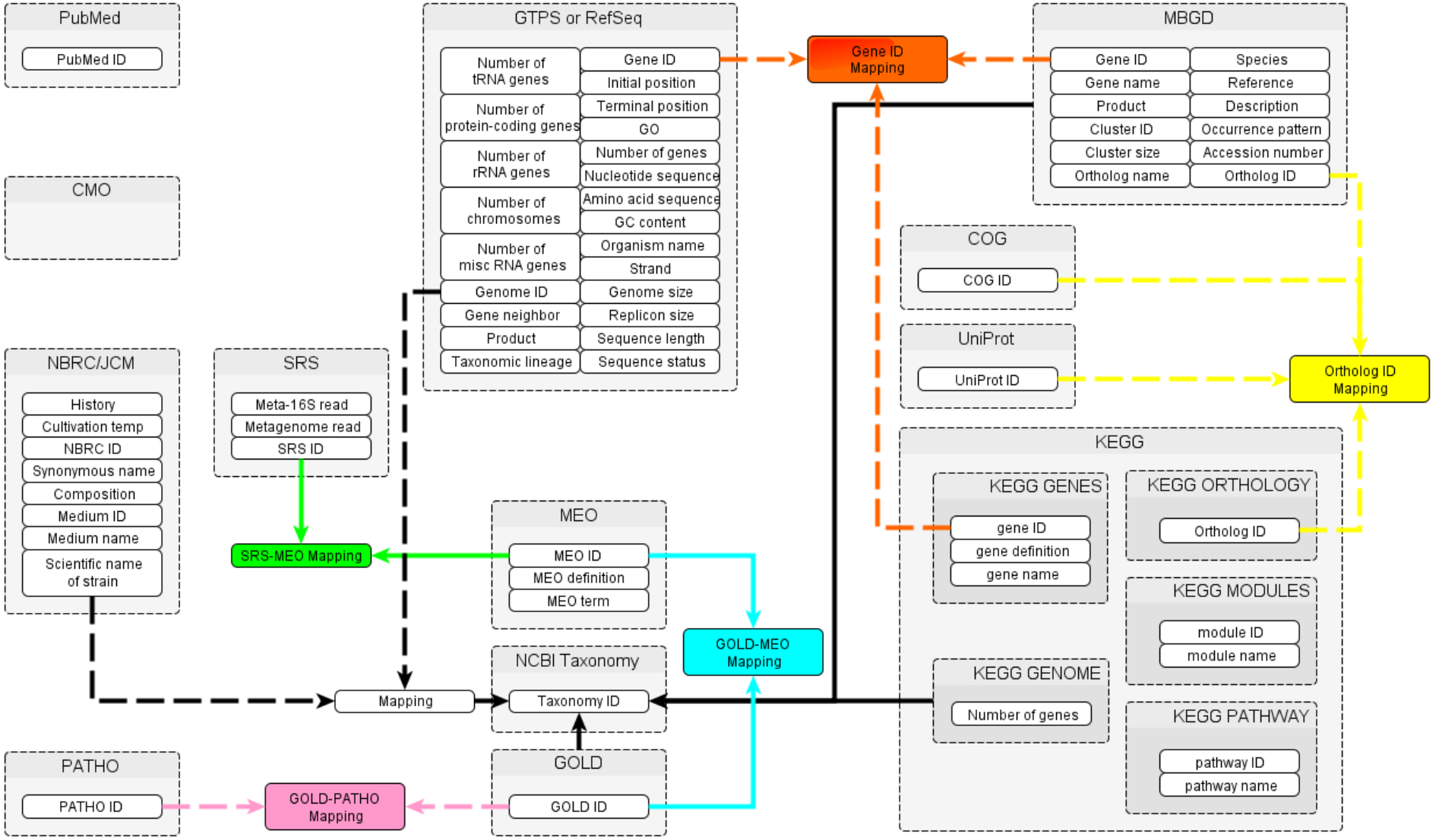
藤澤&神沼&中村@遺伝研

構築したオントロジー

- FALDO (Feature Annotation Location Description Ontology)
  - ゲノム中の各featureの位置情報を記述するためのオントロジー（DBCLS）
- INSDC Ontology
  - INSDCエントリのfeatureとqualifierのターム記述のためのオントロジー
- MCCV (Microbial Culture Collection Vocabulary)
  - 菌株データを記述するためのオントロジー
- MEO (Metagenome/Microbe Environmental Ontology)
  - 細菌の生息環境を記述するためのオントロジー
- PDO-SYMP (Pathogenic Disease Ontology with Symptom)
  - 細菌が引き起こす感染症の情報および感染症の症状を連結したオントロジー
- GMO (Growth Media Ontology)
  - 細菌の培地情報を記述するためのオントロジー（DBCLS）

MicrobeDB.jpにおけるDB間のID統合およびStanzaの開発

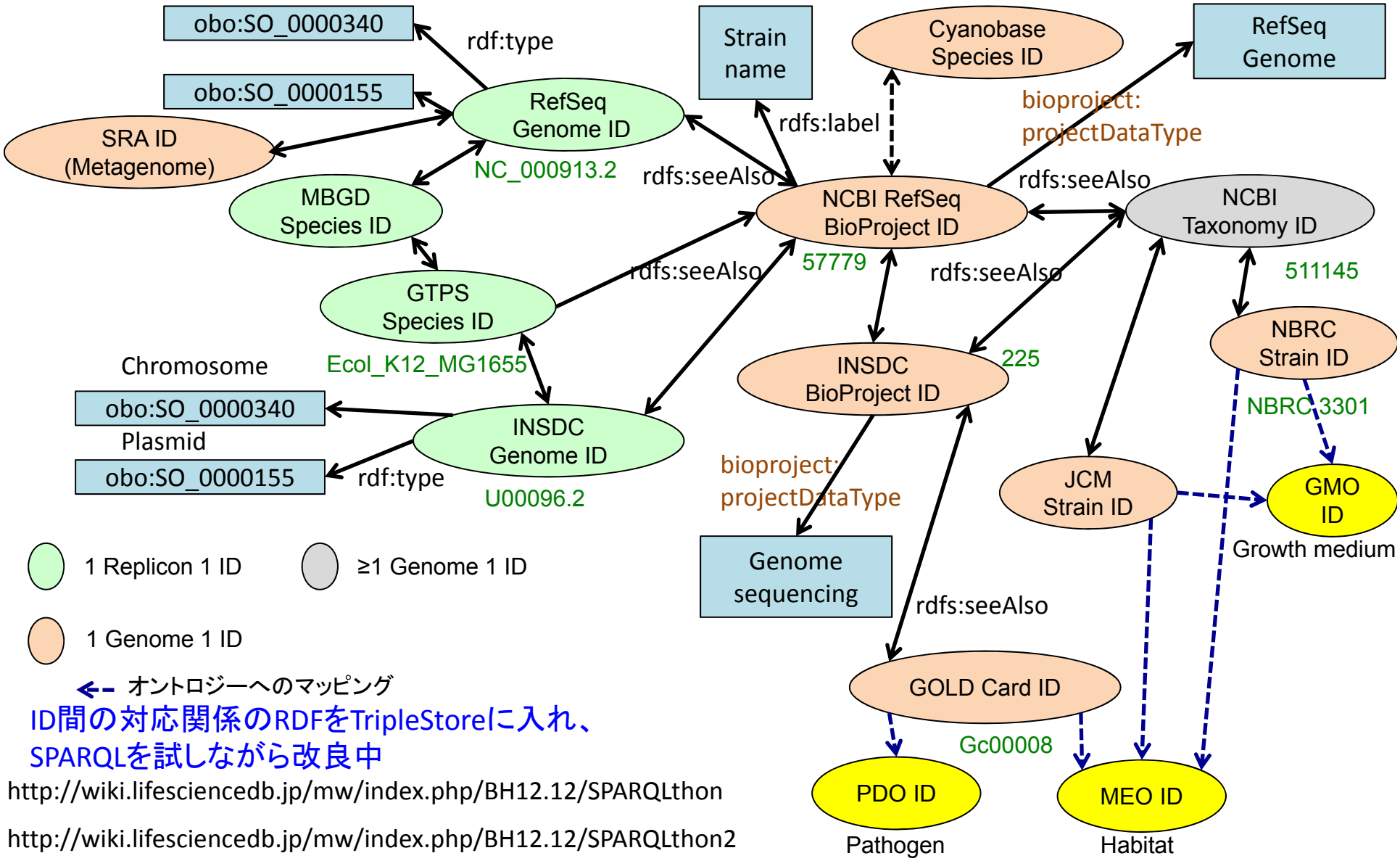
単一DB or 複数DB由来のStanza



複数のDB由来のデータからStanzaを作るには、DB間のID対応関係を整理する必要がある

Integration of DB resources in MicrobeDB.jp

Genomes (example:*Escherichia coli* str. K-12 substr. MG1655)



ID間の対応関係のRDFをTripleStoreに入れ、
SPARQLを試しながら改良中

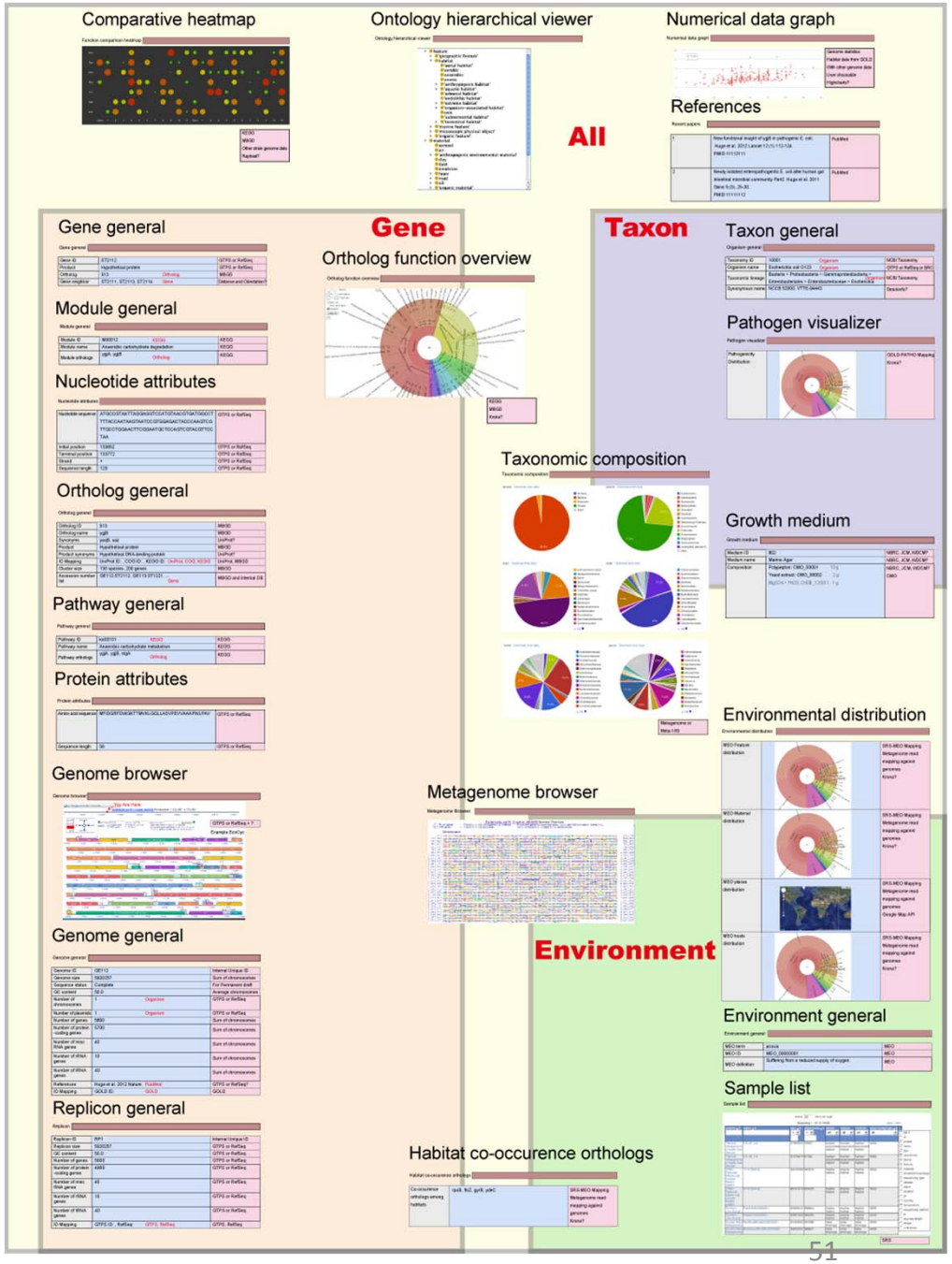
http://wiki.lifesciencedb.jp/mw/index.php/BH12.12/SPARQLthon

http://wiki.lifesciencedb.jp/mw/index.php/BH12.12/SPARQLthon2

http://wiki.lifesciencedb.jp/mw/index.php/BH12.12/SPARQLthon3

Stanza:
data templates for a genome DB

- reusable
- high modularity
- easy to integrate



- TogoStanazaの仕様をDBCLSと共に策定
- MicrobeDBとCyanobaseをユースケースとして開発
- 重複開発を避ける必要あり
- H24年度までに約15種類のStanzaを開発(ゲノム基盤部分はDBCLSが開発)

Stanzaの例 (遺伝子・ゲノム)

- Gene Annotation Stanza
- Ortholog list Stanza
- Genome Information Stanza

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

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

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.

Stanzaの例 (系統・菌株)

・Taxonomic Hierarchy Stanza

Taxonomy Tree

superkingdom	Bacteria
phylum	Firmicutes
class	Bacilli
order	Lactobacillales
family	Enterococcaceae
genus	Enterococcus
species	Enterococcus faecalis

・ Strain Metadata Stanza

Metadata

Medium	http://www.nbrc.nite.go.jp/NBRC2/NBRCMediumDetailServlet?NO=227
Strain number	NBRC 12841
Application	Thienamycins production ; Vitamin B12 (Cyanocobalamine) production ; Steroid conversion
Isolated from	Soil
Strain name	Streptomyces griseus subsp. griseus (Krainsky 1914) Waksman and Henrici 1948
History of deeposit	IFO 12841 <-- SAJ <-- OWU (ISP 5226) <-- Squibb & Sons (F. Arnow, MD 2428, ETH 24234, NIHJ 501)
Taxonomy	http://purl.uniprot.org/taxonomy/67263
Temperature for growth	28

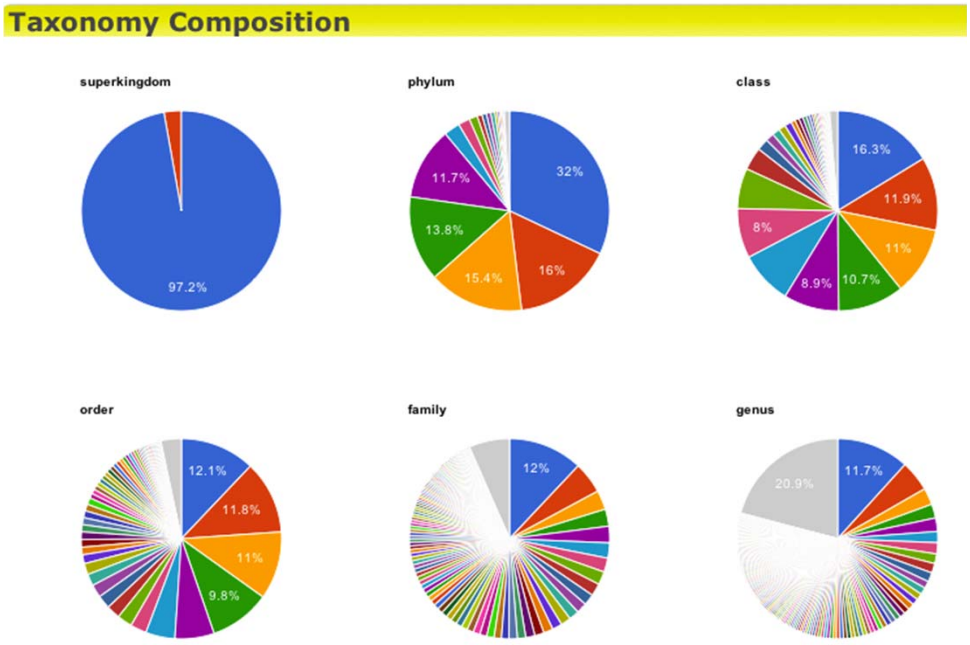
・Related Strain in Other Culture Collection Stanza

Other Collection Numbers

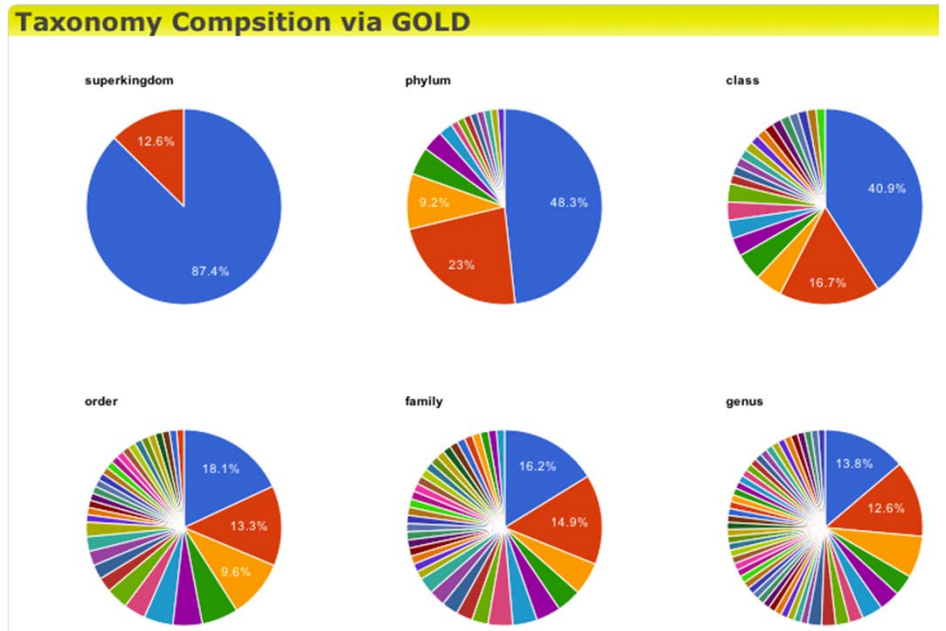
AS 4.1693
ATCC 11009
ATCC 23882
BCRC 11815
CBS 662.68
DSM 40226
ISP 5226
JCM 4229
JCM 4623
KCTC 1742
LMG 5967
NCIMB 9625
NRRL B-1806

Stanzaの例 (環境)

- Taxonomic Composition of the Environment from the Metagenome Stanza



- Taxonomic Composition of the Environment from the Genome-sequenced Strains Stanza



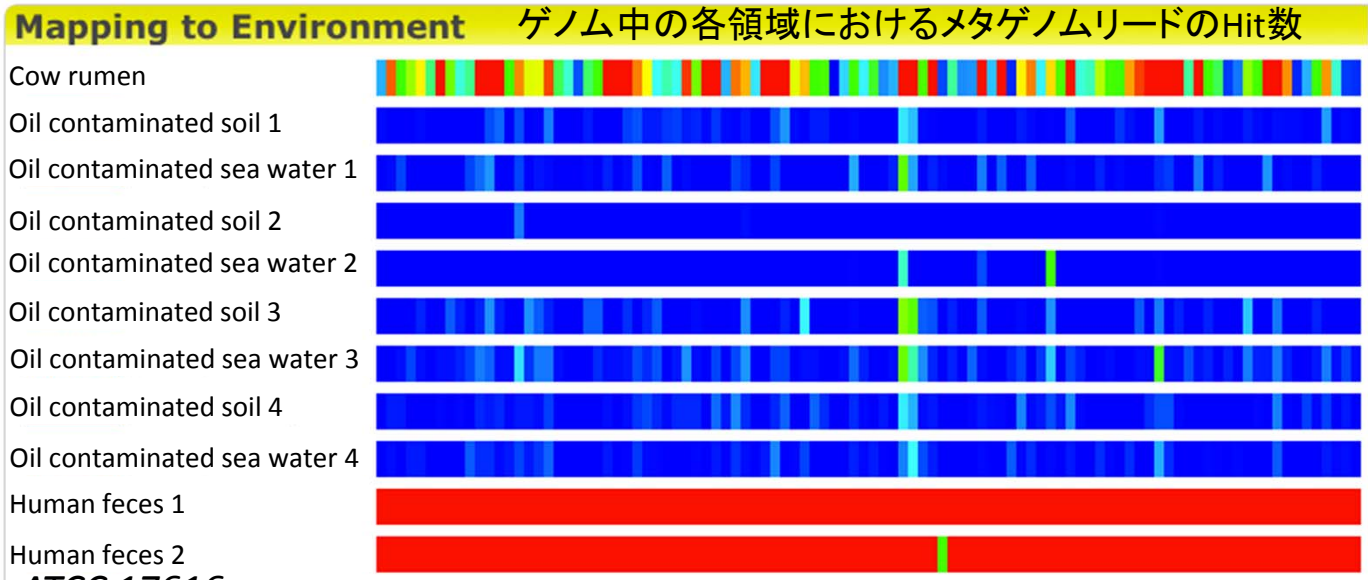
DBCLSが開発中のTogoStanzaと仕様をあわせて特に微生物特有のStanzaを中心に開発中

Stanzaの例 (環境)

・Gene Abundance of the Genome from Several Metagenomes Stanza

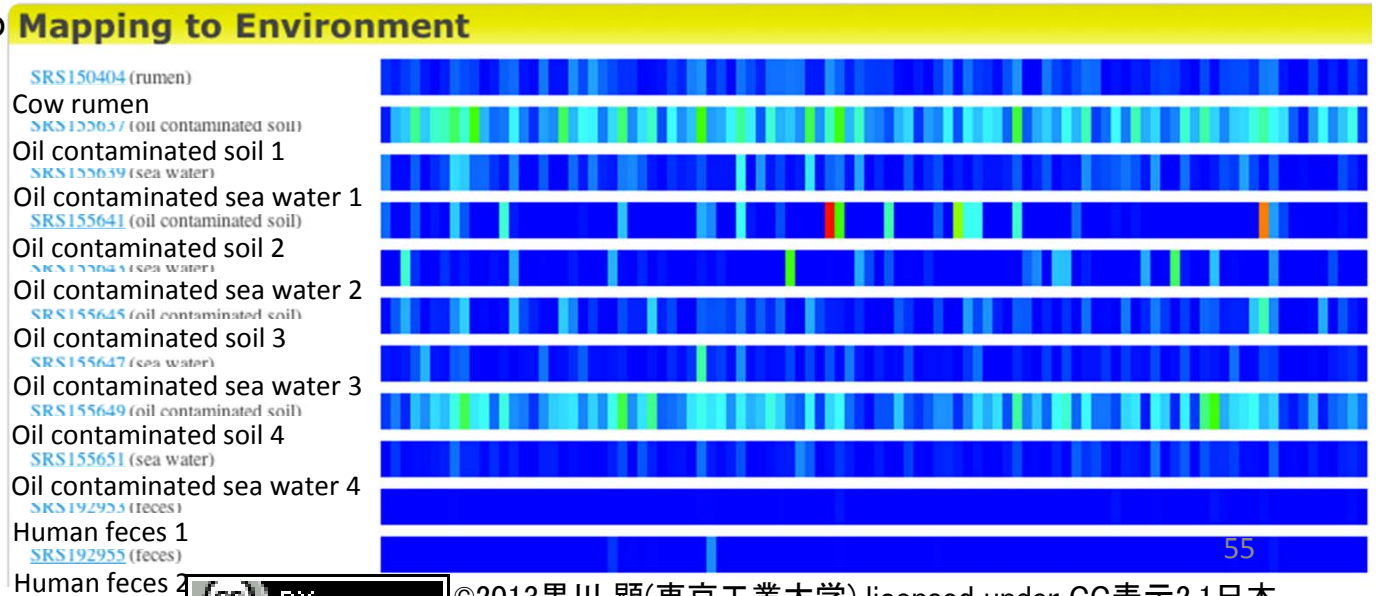
例1: *Bacteroides thetaiotamicron* VPI-5482

ヒト腸内における優占種



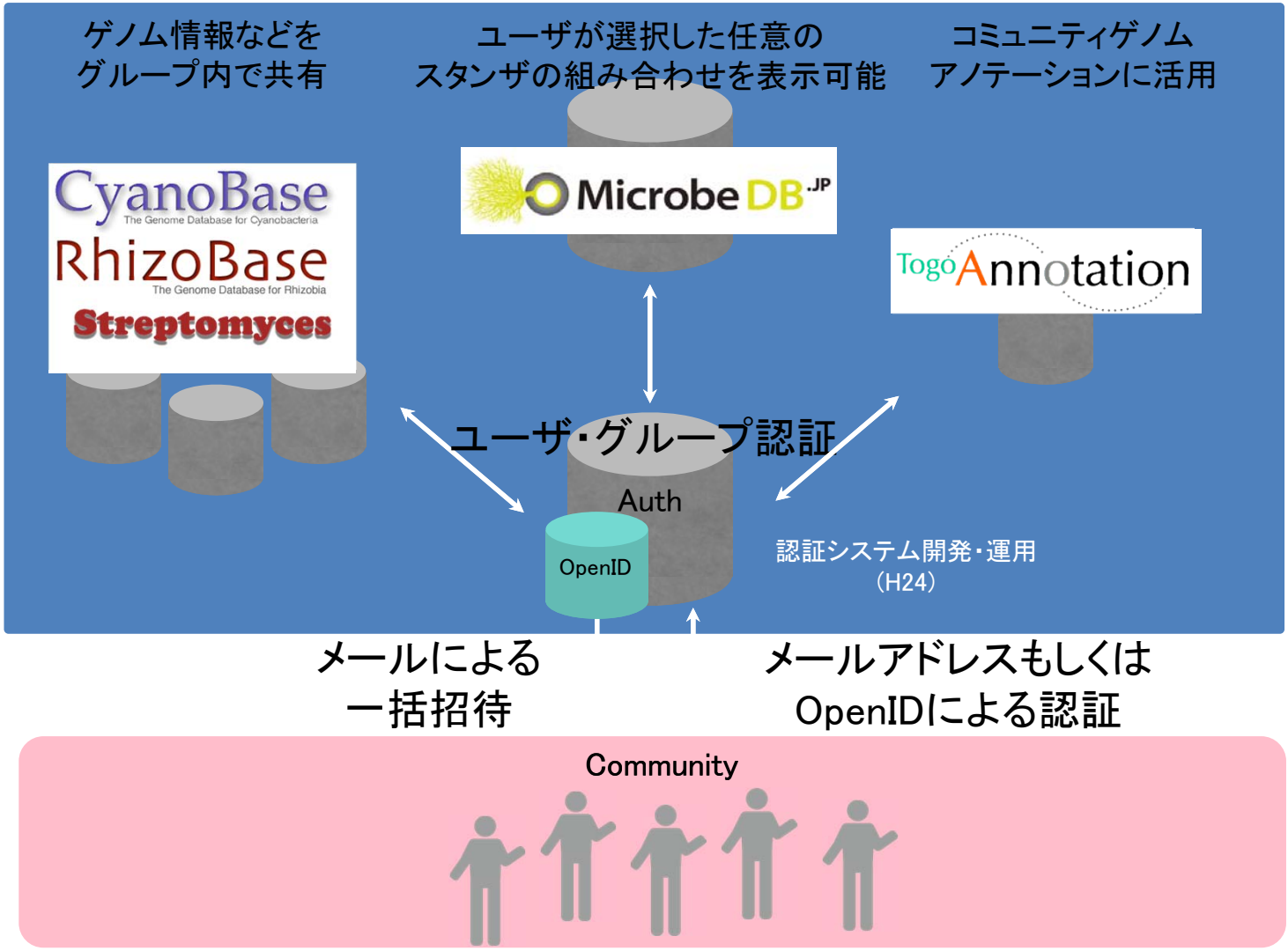
例2: *Burkholderia multivorans ATCC 17616*

芳香族分解関連遺伝子を多数持つ



OpenID-TAuthによる MicrobeDB.jpのユーザ認証

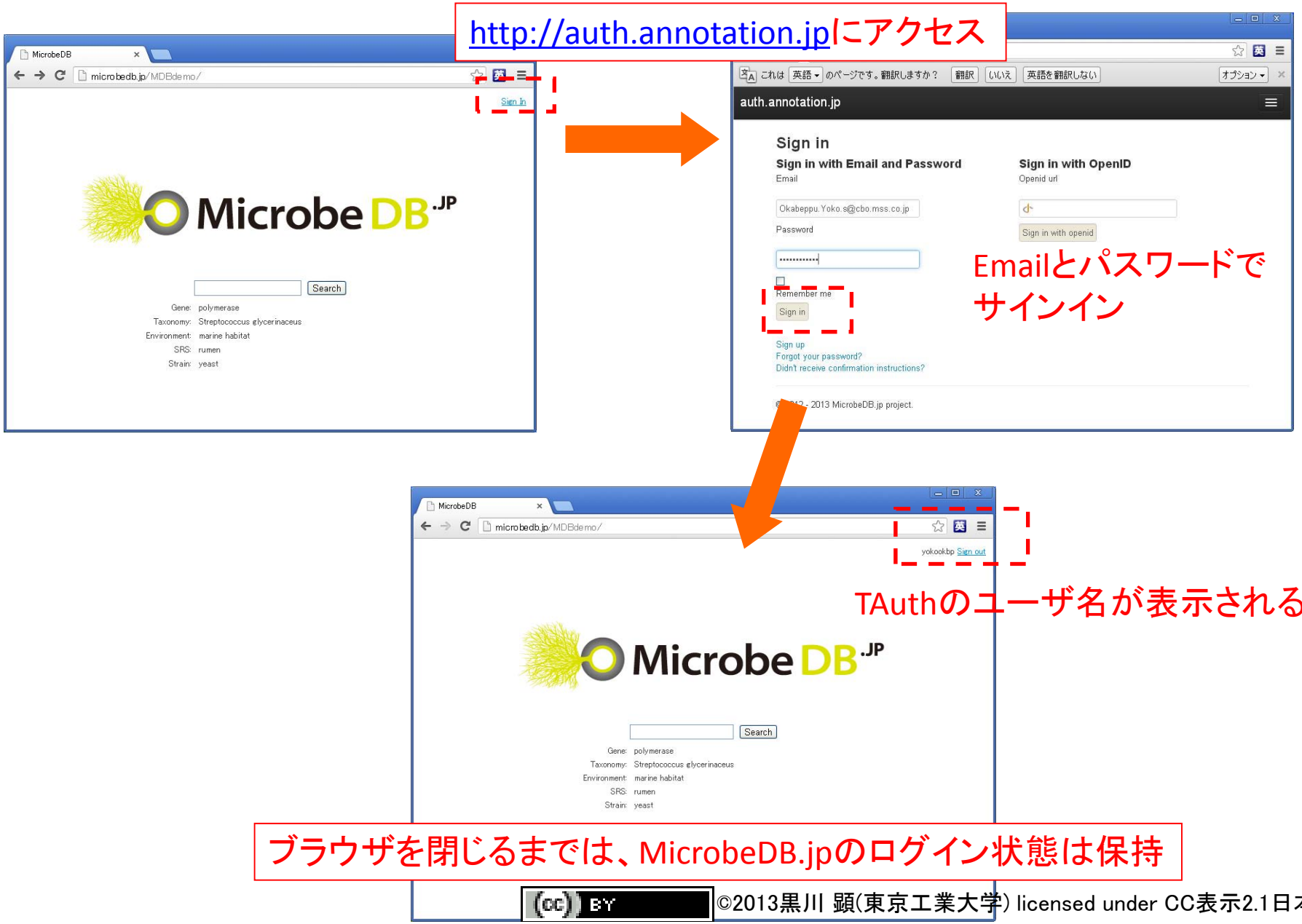
データベース連携に向けたユーザ認証サービス開発・運用



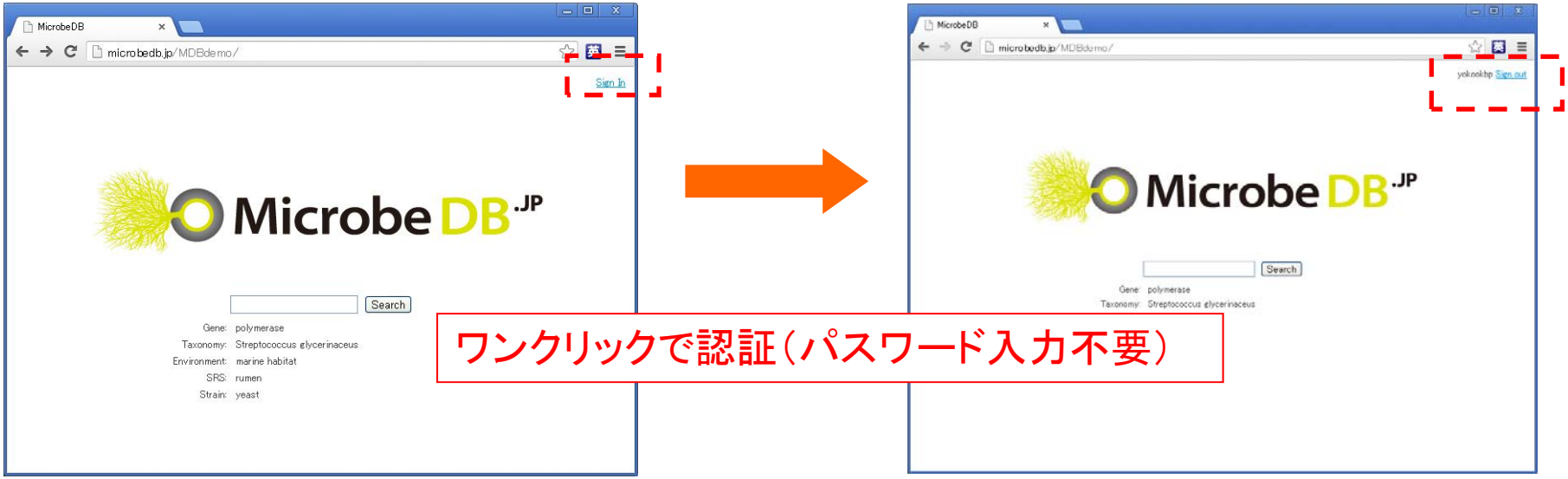
H24年度にOpenID認証と連携したグループ情報を含むユーザ認証サービスを開発し運用を開始した。この機能により、同一認証サービスを利用したMicrobeDB.jpプロジェクト運用データベース間でデータ共有等の連携が可能となった。

藤澤&神沼&中村@遺伝研

①TAuthにログインしていない状態で、ログイン処理を行う場合



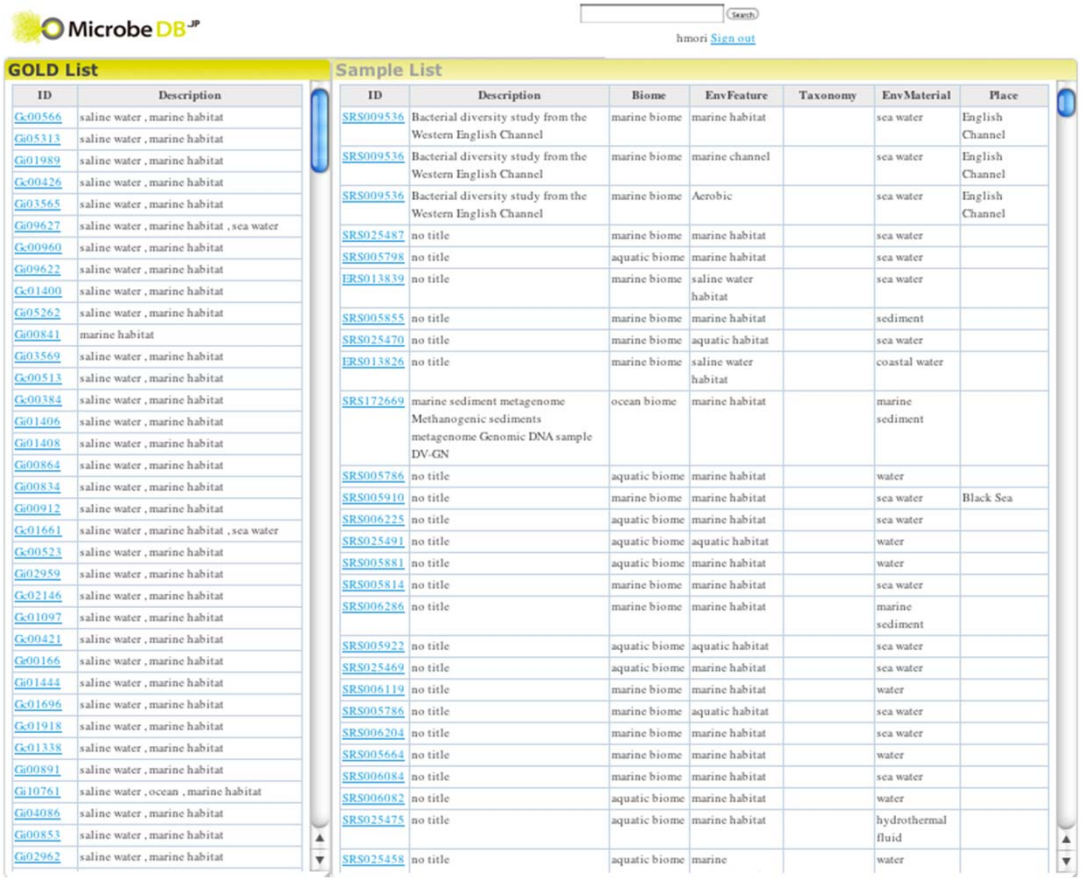
②TAuthにログインしている状態で、ログイン処理を行う場合



ユーザごとのStanzaの表示位置のカスタマイズ

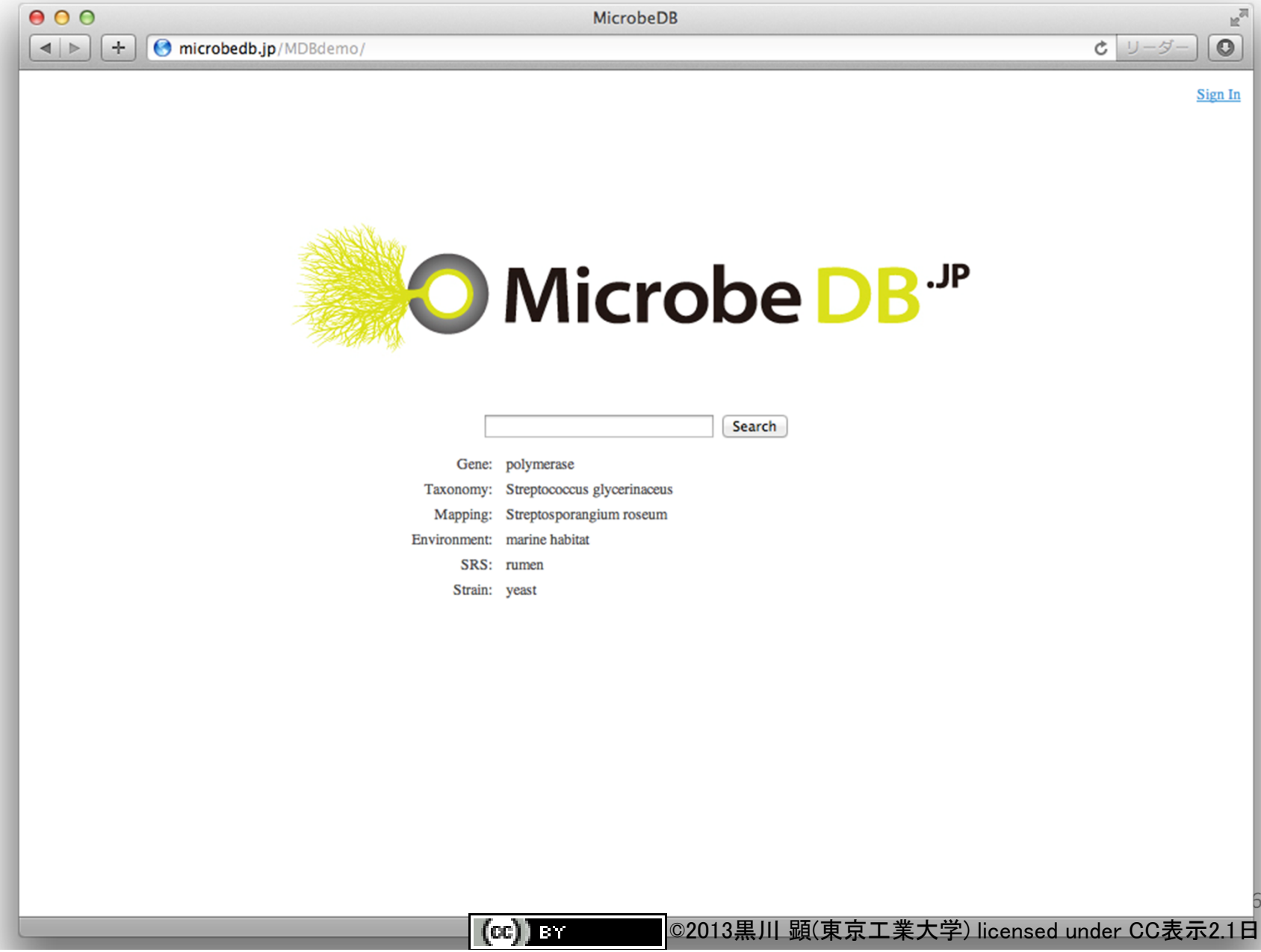
生息環境についてのDefaultページ例

生息環境について関連するゲノム・メタゲノムサンプルの情報のリスト表示に特化するようカスタマイズしたページ



ユーザがStanzaの表示位置をカスタマイズした場合、
ログアウト後もそのユーザのStanza表示位置は記憶され、次のログイン後に再設定は必要無い

http://microbedb.jp/MDBdemo/



H25年度の実施計画

- H23、H24年度に引続き、各DBの高度化および「MicrobeDB.jp」の開発
- GTPSのTogoGenomeを用いたRDF化
- TogoStanzaの仕様確定とMicrobeDB.jpで用いる各種Stanzaの開発（w/ DBCLS）
- TogoAnnotationの各種オミックスデータへの対応
- MBGDの高度化およびバージョンアップ問題への対応
- メタゲノム解析パイプラインをTSUBAME&DDBJスパコンへの対応
- 生息環境以外の様々なメタデータを記述するオントロジーの構築
- オントロジーマッピング、文献情報抽出の自動化
- MicrobeDB.jpにおけるユーザ管理システムの構築
- 培地オントロジー「GMO」の開発（w/ DBCLS）
- 各種オミックスデータの統合化
- 微生物自動アノテーションプログラム「MiGAP」との連携

H24年度 主な活動状況

- 微生物統合DB全体会議（4回開催）
- RDF会議（4回開催）
- オントロジー会議（6回開催）
- 菌株会議（1回開催）

- 微生物データベースの将来に関するフォーラム (5/28@東京）
- BioHackathon 2012（9/2-7@富山）
- The 14th workshop of the GSC（9/17-21@Oxford）
- トーゴーの日シンポジウム 2012（10/5@東京）
- BioJapan 2012（10/10-12@横浜）
- SPARQLthon（10月,11月,12月,1月,計4回@DBCLS）
- 情報・システム研究機構シンポジウム 2012（11/9@東京）
- 分子生物学会（12/11-14@福岡）
- BioHackathon 12.12（12/19-21@東京）
- ゲノム微生物学会（3/8-10@長浜）予定
- 農芸化学会（3/24-28@仙台）予定