

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  
トーゴの日シンポジウム2013

# ゲノム・メタゲノム情報を基盤とした微生物DBの統合

東京工業大学地球生命研究所  
黒川 顕



©2013 黒川 顕 (東京工業大学) licensed under CC表示2.1日本

# 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

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

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

菅原秀明:微生物ゲノム基盤情報資源の共用化

藤澤貴智:モデル微生物情報の高度化

神沼英里:KazusaAnnotation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

西出浩世: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)

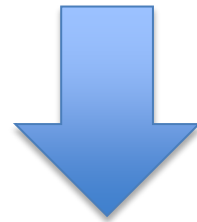
岡本忍, 片山俊明, 川島秀一, 川本祥子, 山本泰智:技術協力

# 微生物研究を取り巻く状況

- 微生物は地球上のあらゆる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には利用困難

# 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**



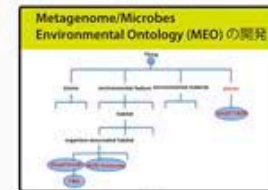
# Microbe DB .JP

<http://microbedb.jp/MDBdemo/>

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使用して整理統合し、幅広い分野での微生物学の実進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

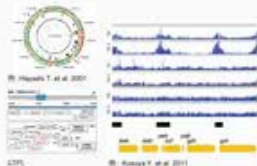
### Taxonomy

### Environment



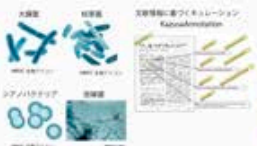
Ortholog: MBGD

オソログデータ



Genome: GTPS/RefSeq

オミックスデータ



Annotation: TogoAnnotation

モデル微生物の製品  
アノテーションデータ



Taxonomy: NCBI Taxonomy

系統分類データ



Strain: NBRC/JCM

菌株データ

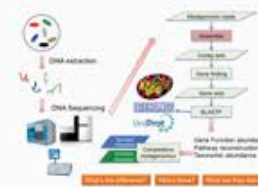


Togo picture gallery



Metadata: GOLD/SRA

環境のメタデータ



Meta-genome: SRA

メタゲノムデータ

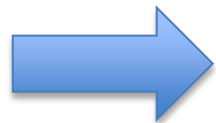
多種多様な細菌の情報をゲノム情報を核として統合し、MicrobeDB.jpから情報を一括検索!

# 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

- 異分野データの統合化のため、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
- Stanzaの開発
- ゲノム自動アノテーションシステムMiGAPとの連携

# MicrobeDB.jpの基盤データ整備

- Bacteria&Archaea Genome (GTPS, RefSeq, GenBank): 2,516株
- Amplicon seq (NCBI SRA): 15,504サンプル
- Metagenome (NCBI SRA): 3,269サンプル
- Meta-transcriptome (NCBI SRA): 196サンプル
- Ortholog groups (MBGD): 340,439グループ
- Strain data (NBRC): 17,367株
- Strain data (JCM): 13,396株



すべてRDFにて記述

# 各種データのトリプル数

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	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が集計結果のデータを作成		440,773
合計			1,075,964,773



# オントロジー・ボキャブラリ

- 異なるDBの統合化のためには、多様な表現を統一する必要がある
- 特に環境情報は単語の定義および語彙の階層的分類・定義が必須
- メタゲノム情報(メタデータ)や菌株情報などは国際的に標準化されていない



各種オントロジーの開発

# ゲノムアノテーション標準のためのオントロジー開発

## ● INSDCオントロジー (w/DBCLS)

- INSDC配列エントリー(GenBank, EMBL, DDBJ 形式)のRDF化を目的としたオントロジー
  - INSDC Feature Table Definition (Version 10.2) をOWL定義(H24年度)
  - Genomic Standards Consortium (GSC15; April 22-24, 2013. NIH, USA) にて発表
  - SPARQL検索性向上を目的としたアップデートを実施(H25年度)

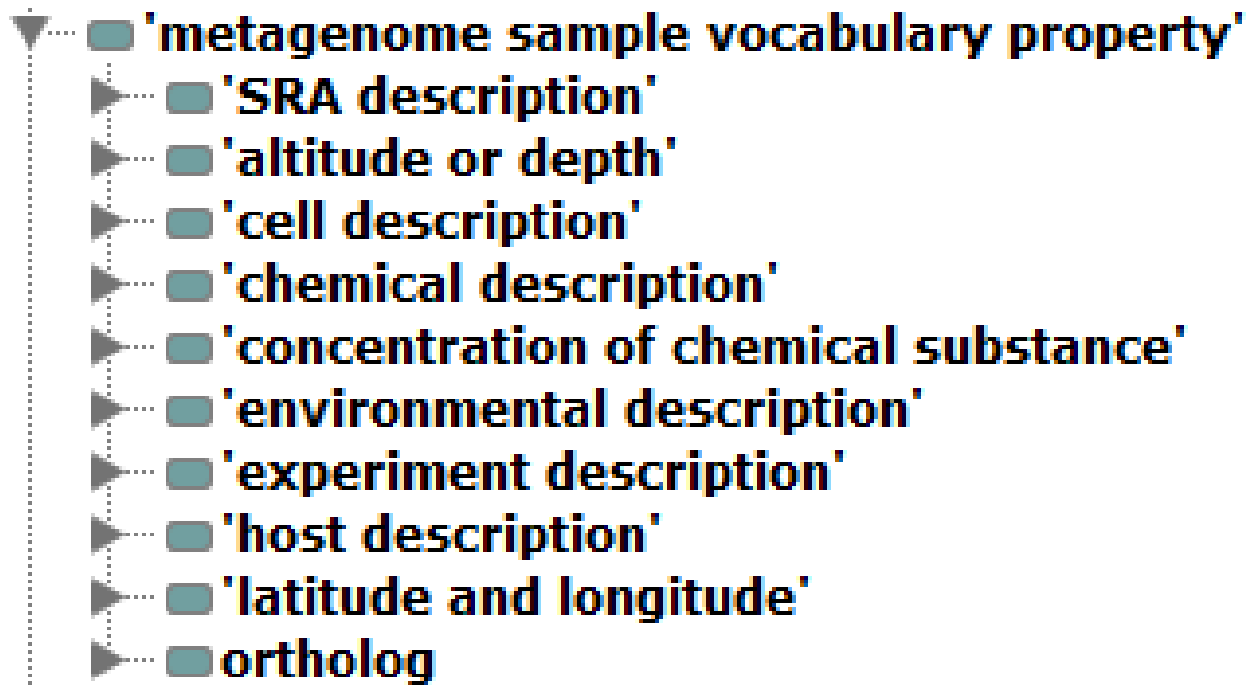
## ● FALDOオントロジー (w/BioHackathon)

- 塩基配列およびタンパク質などbiological featureの複雑な位置情報に関する記述の標準化を目的としたオントロジー
  - BioHackathon2012(2012年9月開催)で開発
  - BioHackathon2013(2013年6月開催)で課題の検討およびアップデート

これらの標準化を目的としたオントロジーを利用することでフェデレーション型の問い合わせでも配列アノテーションの統合が容易となる

# MSV (Metagenome Sample Vocabulary)

- メタゲノムサンプルのメタデータを整理したオントロジー  
同義語と判断できたメタデータカテゴリ名をまとめた
- クラス数 : 306 クラス  
大分類 : 10クラス



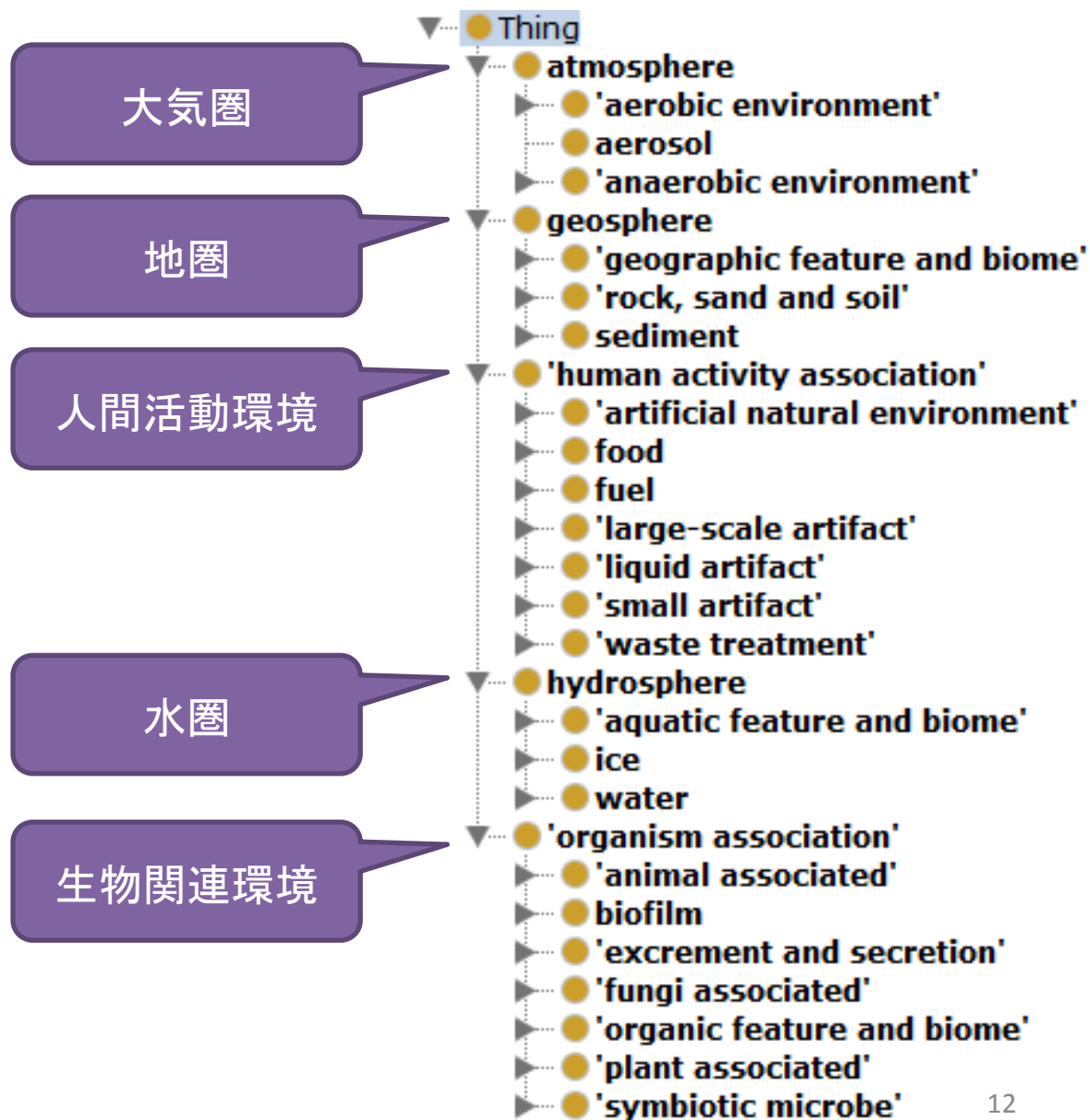
# MEO (Metagenome Environmental Ontology) ver0.6

- 微生物の生息環境  
についてのオントロジー

- ・クラス数: 782
- ・プロパティ数: 19

- クラス構造

- ・5つの大分類
- ・23の中分類



# 疾病関連語句オントロジーの開発:PDO

疾病語句を発生部位により分類

⇒ DO (Pathogenic Disease Ontology)

クラス数: 266

- ▼ ● 'Disease involving body sites'
  - ▶ ● 'Breast disease'
  - ▶ ● 'Cardiovascular disease'
  - ▶ ● 'Digestive system disease' 消化器の疾病
  - ▶ ● 'Immune system disease'
  - ▶ ● 'Musculoskeletal system disease'
  - ▶ ● 'Nervous system disease'
  - ▶ ● 'Reproductive system disease'
  - ▶ ● 'Respiratory system disease'
  - ▶ ● 'Skin disease'
  - ▶ ● 'Systemic disease' 全身性の疾病
  - ▶ ● 'Urinary system disease'
- ▶ ● 'Disease involving unidentified body site'  
感染組織を特定できないもの

SNOMED (国際医療用語集) より同義語を登録  
→ 様々なキーワードで検索可能  
階層を考慮した推論検索も可能

# 疾病関連語句オントロジーの開発: CSSO

症状語句を発生部位により分類

→ **CSSO** (Clinical Signs and Symptoms Ontology)

クラス数: 272

- ▶ ● 'Abdomen and digestive system symptom'
- ▶ ● 'Circulatory system symptom'
- ▶ ● 'Cognition, perception, emotional state and behavior system symptom'
- ▶ ● 'Food and fluid intake system symptom'
- ▶ ● 'General symptom'
- ▶ ● 'Head and neck symptom'
- ▶ ● 'Immune system symptom'
- ▶ ● 'Musculoskeletal system symptom'
- ▶ ● 'Nervous system symptom'
- ▶ ● 'Reproductive system symptom'
- ▶ ● 'Respiratory system symptom'
- ▶ ● 'Skin symptom'
- ▶ ● 'Speech and voice system symptom'
- ▶ ● 'Thoracic symptom'
- ▶ ● 'Urinary system symptom'

SNOMEDより同義語を登録  
→ 様々なキーワードで検索可能

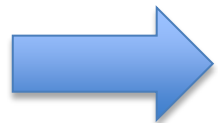
PDO/CSSOの利用により感染症関連語句の  
柔軟な検索が可能

# 開発したオントロジー

- 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 (w/BioHackathon)
- 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 (w/DBCLS)
- 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/CSSO (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 (w/DBCLS)

# オーソログ遺伝子情報

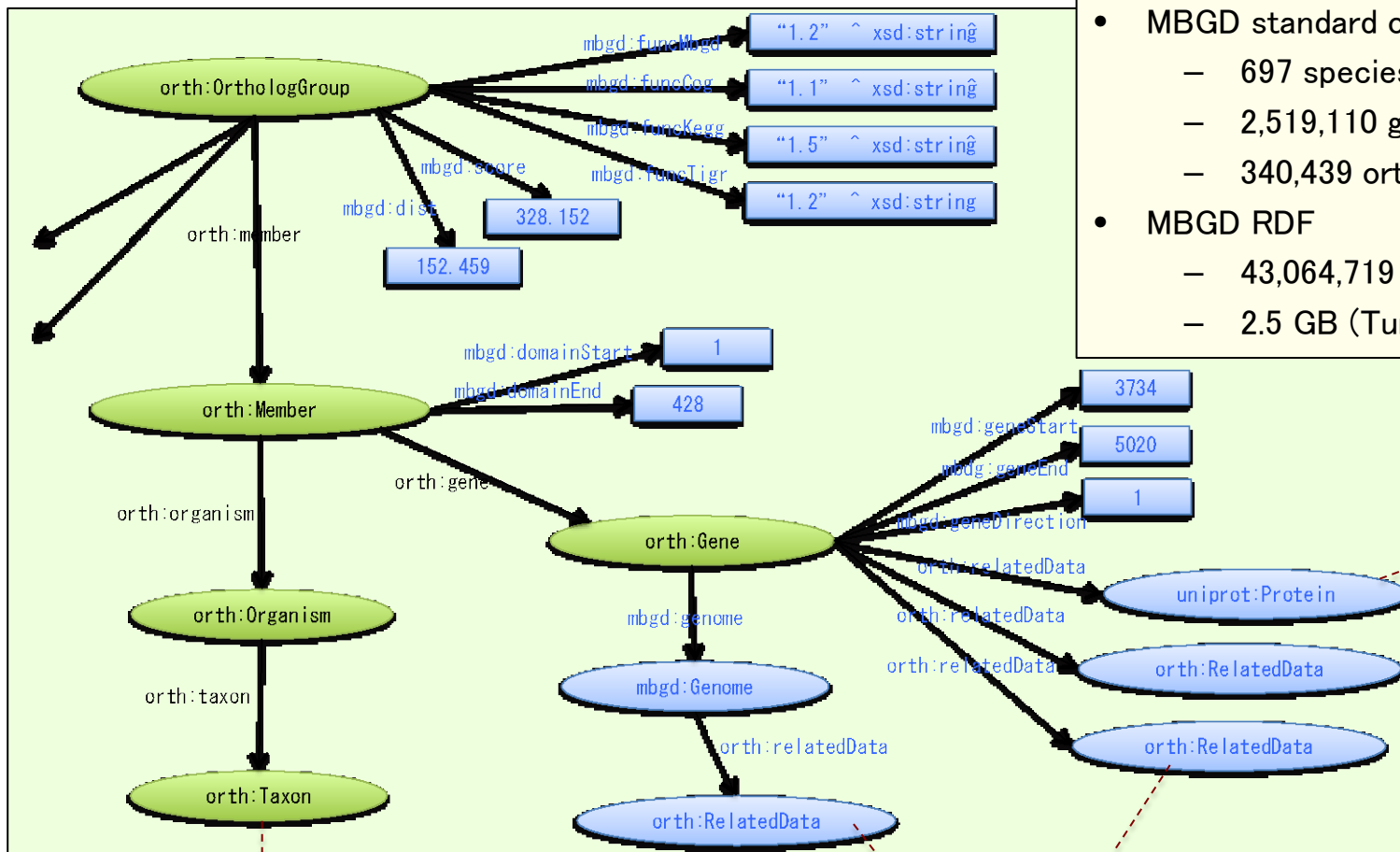
- ゲノム情報を種間で連結するためには、オーソログ遺伝子情報が有効となる
- MBGDオーソログデータベースをRDF化
- オーソロググループのさらなる精密化
- ドラフトゲノムデータにも対応



MBGDのRDF化



# MBGD RDF オースログ情報のRDF化と それに基づく微生物ゲノムデータの統合

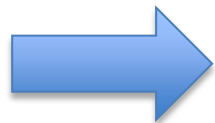


- MBGD standard ortholog groups
  - 697 species
  - 2,519,110 genes
  - 340,439 ortholog groups
- MBGD RDF
  - 43,064,719 triples
  - 2.5 GB (Turtle)



# メタゲノム情報と環境情報の統合

- メタゲノム情報は遺伝子情報と環境情報とをブリッジするユニークなデータ
- メタゲノム情報を通じて異分野との融合が可能となる
- 各メタゲノムデータを解析パイプラインにて統一処理し、遺伝子機能組成データを作成
- 遺伝子機能組成データおよびメタデータ(環境情報)を統合



メタゲノム情報のRDF化

# 公共データベースにデポジットされている メタゲノムメタデータ

	サンプル数	メタデータの カテゴリ 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

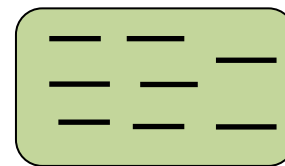
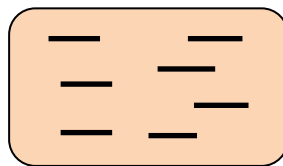
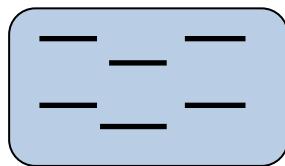
( )内は実際に配列が存在したエントリ

海水

土壌

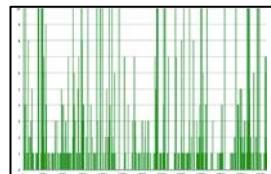
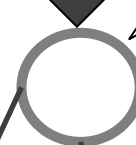
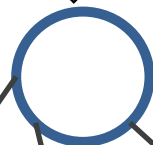
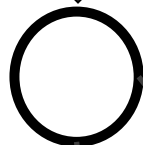
ヒト腸内

メタゲノム



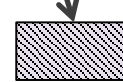
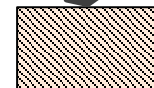
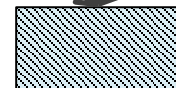
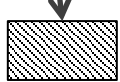
配列相同性検索

ゲノム



遺伝子のクラスタリング

オーソログ



ゲノムを軸に遺伝子・系統・環境の情報を統合化したことで、

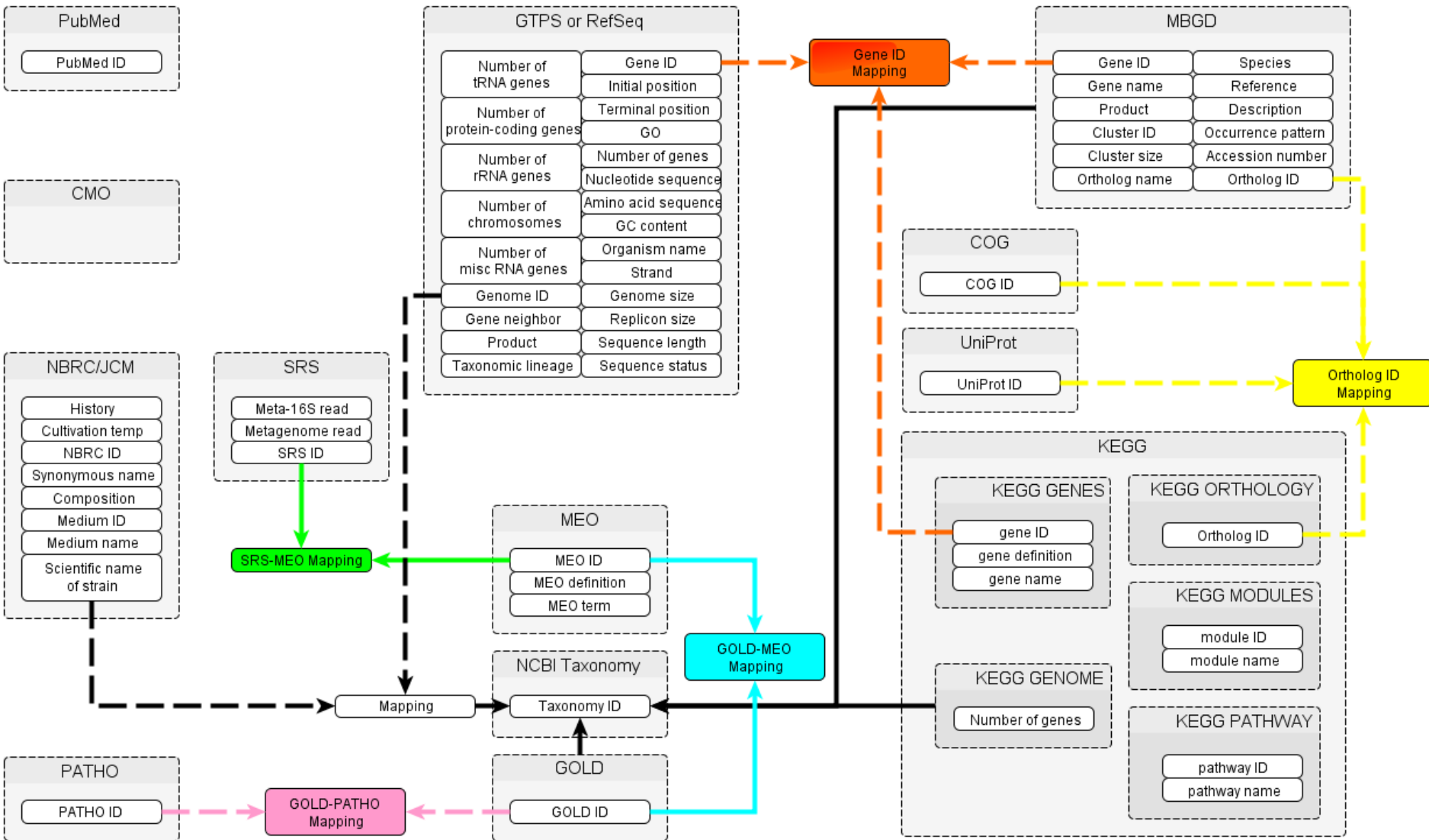
- ・ある環境にどのような遺伝子が多いのか？
- ・それらの遺伝子はどのような機能を持っているのか？
- ・それらの遺伝子はどのような系統群に多いのか？

等の情報が簡単に得られるようになった

# Stanza開発

- データベース利用のためのテンプレート
- 異なるデータベースを横断的に利用するため、各データのID対応関係を整理
- TogoStanzaの仕様をDBCCLSと共に策定
- MicrobeDB.jpにて58種類のStanzaを開発（さらにDBCCLSで開発された34種類の Stanzaを導入）

# 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

# Stanzaの例 (遺伝子・ゲノム)

- Gene Annotation

### Feature

[[http://purl.obolibrary.org/obo/SO\\_0000704](http://purl.obolibrary.org/obo/SO_0000704)]

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

[[http://purl.obolibrary.org/obo/SO\\_0000316](http://purl.obolibrary.org/obo/SO_0000316)]

dbxref	<a href="http://www.ncbi.nlm.nih.gov/gene/897644">http://www.ncbi.nlm.nih.gov/gene/897644</a>
dbxref	<a href="http://www.ncbi.nlm.nih.gov/nucleotide/15643342">http://www.ncbi.nlm.nih.gov/nucleotide/15643342</a>
exons	nodeID://b71582

- Ortholog list

### Ortholog

ID	Genome	Description	Protein	UniProt	GTSPS	RefSeq
<a href="#">aac:AACT_1427</a>	<a href="#">aac</a>	DNA polymerase III subunit alpha	<a href="#">YP_003184842.1</a>	<a href="#">C8WWI2</a>	<a href="#">AACT_ACIDOCALDARIUSDSM446:ST2344</a>	NC_013205.1
<a href="#">aar:ACEAR_1599</a>	<a href="#">aar</a>	DNA polymerase III catalytic subunit, PolC type	<a href="#">YP_003828170.1</a>		<a href="#">AARA_DSM5501:ST105</a>	
<a href="#">acl:ACL_0247</a>	<a href="#">acl</a>	DNA polymerase III subunit alpha	<a href="#">YP_001620249.1</a>	<a href="#">A9NEU3</a>	<a href="#">ALA1_PG8A:ST588</a>	NC_010163.1
<a href="#">afj:AFLY_1700</a>	<a href="#">afj</a>	DNA polymerase III PolC	<a href="#">YP_002316046.1</a>	<a href="#">B7GG80</a>	<a href="#">AFLA_WK1:ST2505</a>	NC_011567.1
<a href="#">afn:ACFER_1370</a>	<a href="#">afn</a>	DNA polymerase III subunit alpha	<a href="#">YP_003399045.1</a>		<a href="#">AFER_DSM20731:ST1519</a>	NC_013740.1
<a href="#">amt:AMET_2678</a>	<a href="#">amt</a>	DNA polymerase III subunit alpha	<a href="#">YP_001320489.1</a>	<a href="#">A6TRL2</a>	<a href="#">AMET_QYME:ST2214</a>	NC_009633.1

- Genome Information

### Genome

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

- Taxonomic Hierarchy

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

- Strain Metadata

Metadata	
Medium	<a href="http://www.nbrc.nite.go.jp/NBRC2/NBRCMediumDetailServlet?NO=227">http://www.nbrc.nite.go.jp/NBRC2/NBRCMediumDetailServlet?NO=227</a>
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 deposit	IFO 12841 <-- SAJ <-- OWU (ISP 5226) <-- Squibb & Sons (F. Arnow, MD 2428, ETH 24234, NIHJ 501)
Taxonomy	<a href="http://purl.uniprot.org/taxonomy/67263">http://purl.uniprot.org/taxonomy/67263</a>
Temperature for growth	28

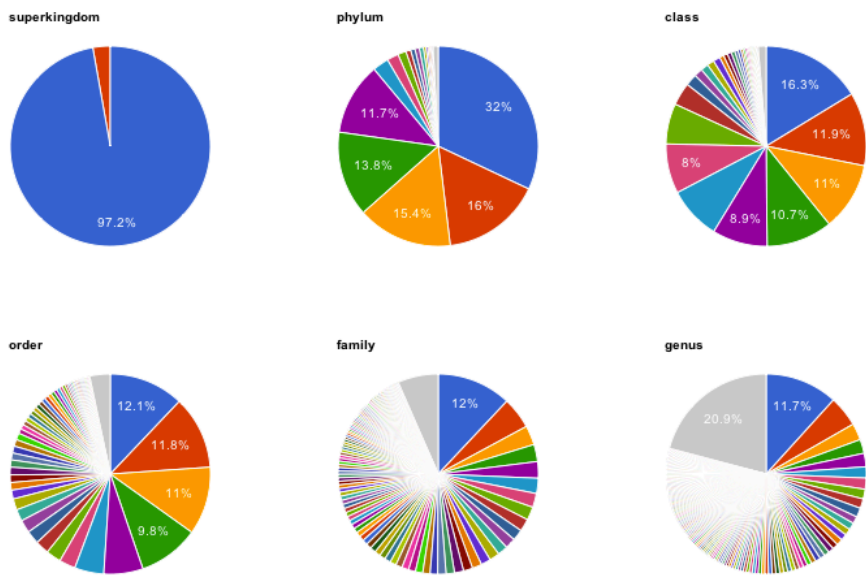
- Related Strain in Other Culture Collection

Other Collection Numbers
AS 4.1693
ATCC 11009
ATCC 23882
<a href="#">BCRC 11815</a>
CBS 662.68
<a href="#">DSM 40226</a>
ISP 5226
<a href="#">JCM 4229</a>
<a href="#">JCM 4623</a>
KCTC 1742
<a href="#">LMG 5967</a>
NCIMB 9625
<a href="#">NRRL B-1806</a>



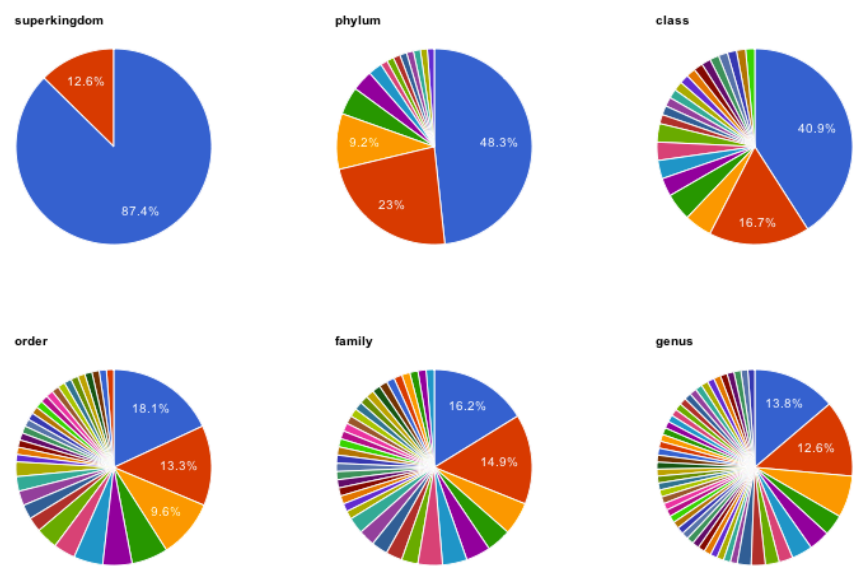
- Taxonomic Composition of the Environment from the Metagenome

Taxonomy Composition



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

Taxonomy Composition via GOLD

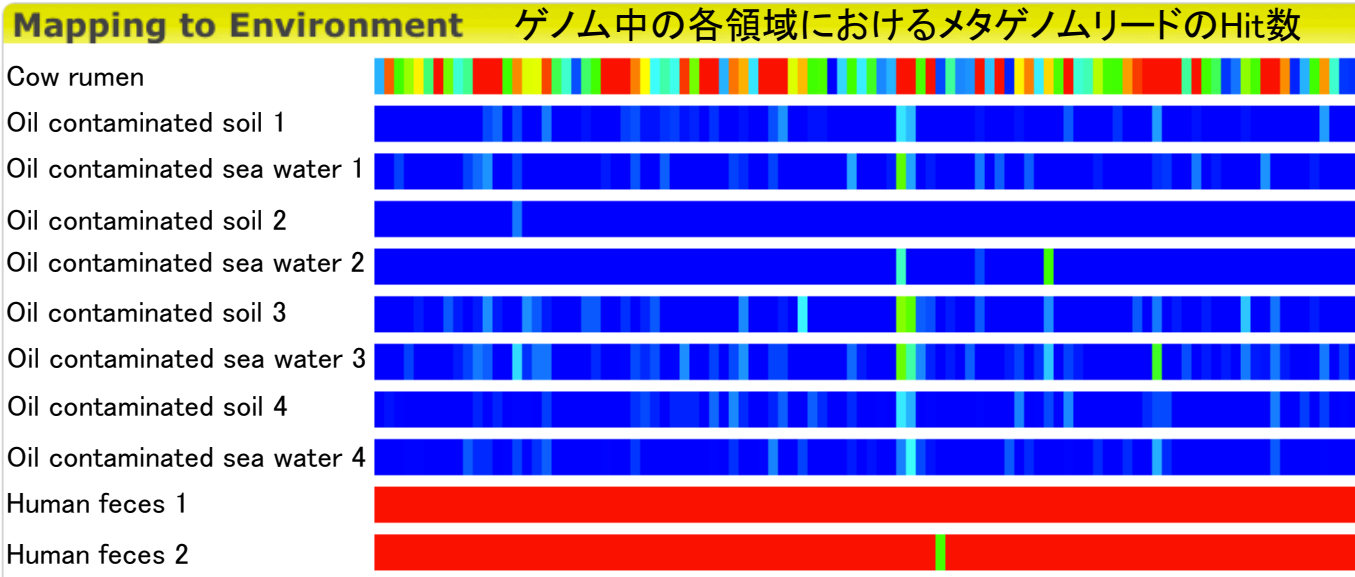


# Stanzaの例 (環境)

## Gene Abundance of the Genome from Several Metagenomes

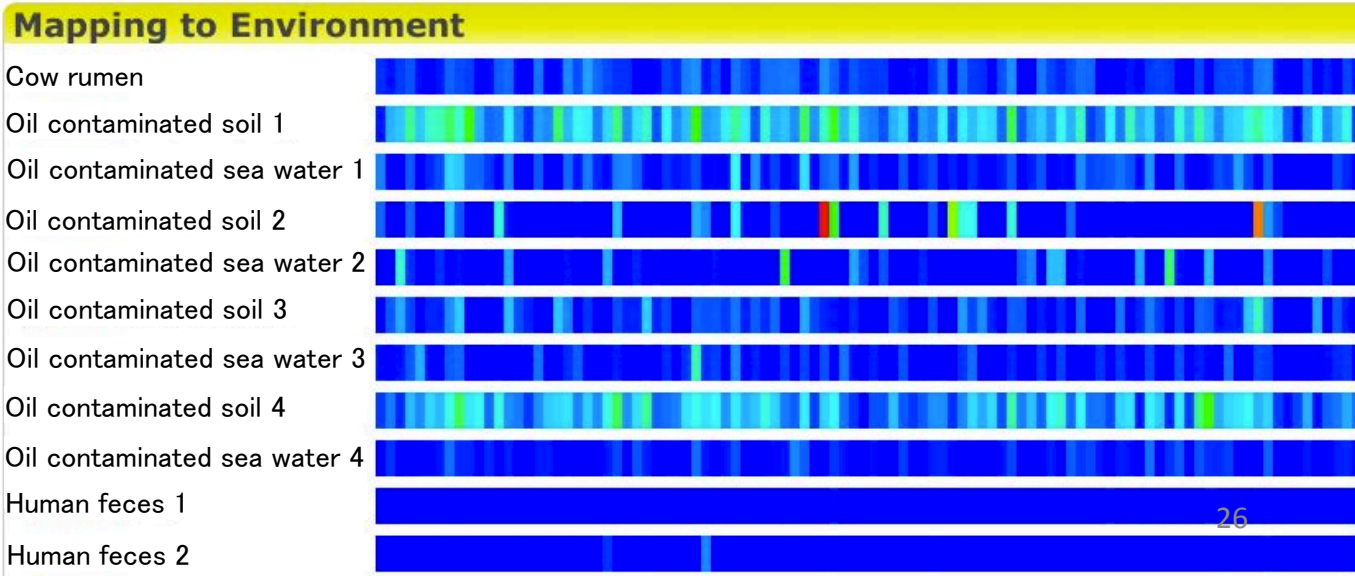
例1: *Bacteroides thetaiotamicron* VPI-5482

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



## Body Mapping




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

http://microbedb.jp/

MicrobeDB

microbedb.jp/MDB

[Sign In](#)



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

28

# 検索ワード: lake

MEO OWL

菌株-MEOマッピングRDF

lake → meo:pond is\_a meo:lake → Strain\_A mccv:isolation\_source meo:pond → Strain\_A

のようなオントロジーを介した推論検索を実行

lake由来の  
ゲノム解読  
済み株

lake由来のメタゲノムサンプル  
で多いオーソログ

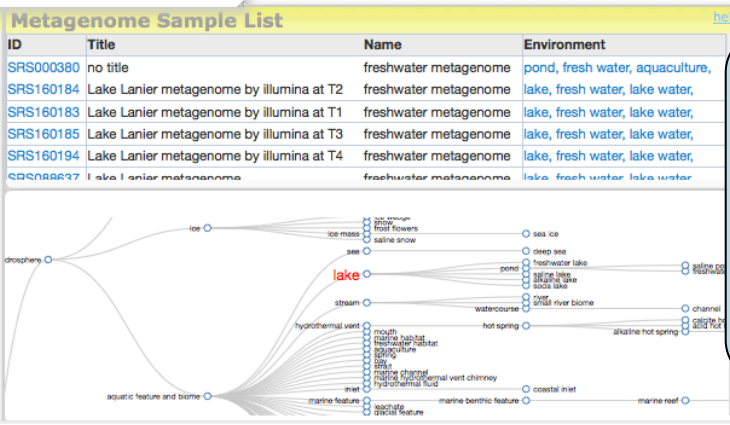
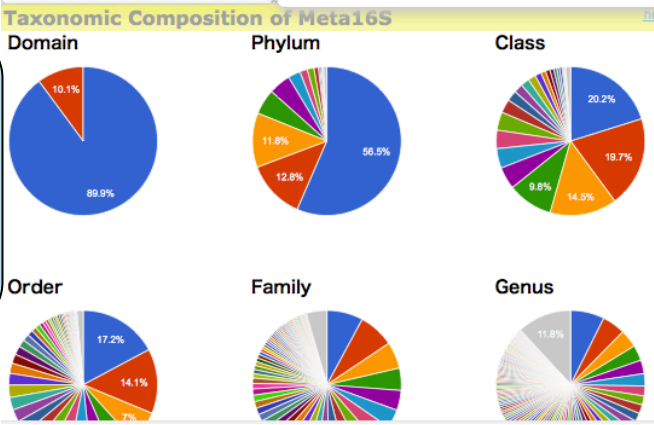
lake由来  
の菌株

MicrobeDB-JP

Sequenced Genome List		Ortholog Abundance in Environment			Strain List					
ID	Description	Ortholog Name	Description	Hit Count	Number	Name	Isolated From	Temperature	Application	
Gi09828	lake	279	groEL	Chaperonin, groel protein	0.505158	JCM 9655	Natronococcus amylolyticus Kanai et al. 1995	Soil from shores of the Lake Magadi, Kenya	37	Production of maltotriose-forming &alpha;-amyl
Gi09831	lake	112	ilvB	Acetolactate synthase large subunit thiamine pyrophosphate	0.250809	JCM 11574	Streptomyces rameus Shibata 1959	Soil, the shore of Lake Inawashiro, Fukushima Pref., Japan	28	Production of streptomycin
Gi09841	lake	151	ilvB	Acetolactate synthase large subunit thiamine pyrophosphate	0.250524	JCM 10172	Delftia acidovorans (den Dooren and de Jona 1926) Wen et al.	Lake Biwa, Shiga Pref., Japan	30	Production of (3-hydroxybu
Gi12111	lake	19342	ydiP	C-5 cytosine-specific DNA methylase protein	0.155533					
Gc01055	freshwater lake	732	dnaK	Molecular chaperone DnaK	0.0695482					
Gc01057	freshwater lake	3522	yciK	Short-chain dehydrogenase/reductase SDR	0.0515464					
Gc01302	freshwater lake	1073	nuoM	Proton-translocating NADH-quinone oxidoreductase subunit M	0.0446604					
Gc01334	freshwater lake	294	glgC	Glucose-1-phosphate thymidyltransferase	0.0440619					
Gc01372	freshwater lake	1782	ureG	Hydrogenase nickel incorporation protein HypB	0.0438833					
Gc01760	freshwater lake									

lake由来のメタゲノムサンプル

lake由来のメタ16Sサンプルの平均的な系統組成



lakeと他の語彙のMEO中での位置関係

検索ワード: *Streptococcus pyogenes* MGAS2096 (ゲノム解読済みの、病原連鎖球菌)



### Definition [help](#)

**Streptococcus pyogenes MGAS2096**

### Taxon Hierarchy [help](#)

Superkingdom	Bacteria
Phylum	Firmicutes
Class	Bacilli
Order	Lactobacillales
Family	Streptococcaceae
Genus	Streptococcus

### Genome Information [help](#)

BioProject:58573

Description	RefSeq	Type	Size	Gene	tRNA	rRNA	Other
Streptococcus pyogenes MGAS2096 chromosome, complete genome.	<a href="#">NC_008023.1</a>	chromosome	1860355	1979	63	18	0

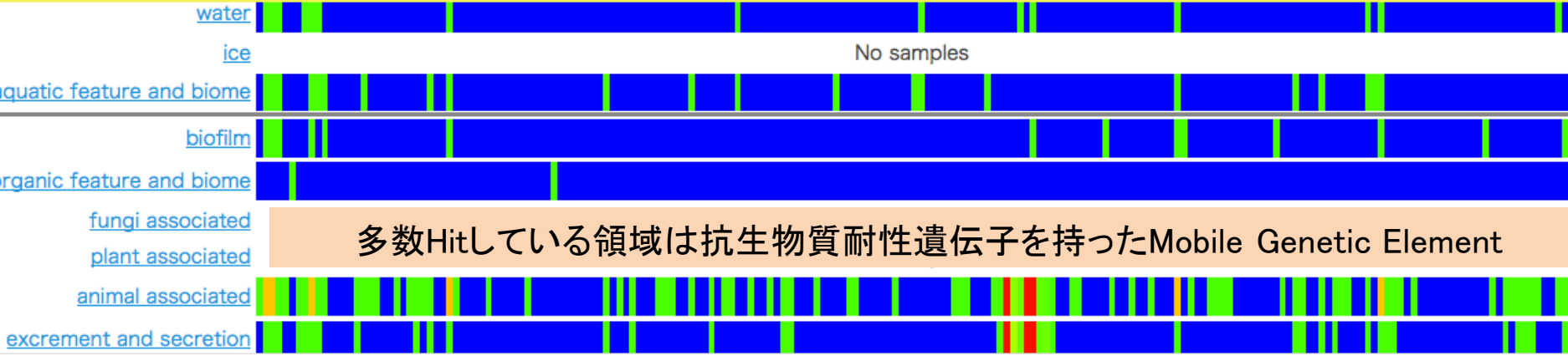
### Pathogen Information [help](#)

Taxonomy Name	Disease Name	Infectious Type	Strain Type
Streptococcus pyogenes MGAS2096	Glomerulonephritis, Necrotizing fasciitis, Streptococcal toxic shock syndrome, Vulvovaginitis, Scarlet fever, Septicemia		

### Phenotype Information [help](#)

Cell shape	Coccus
Oxygen requirement	Facultative anaerobe
Temperature	Mesophile

### Mapping to Environment(Chromosome) [help](#)



多数Hitしている領域は抗生物質耐性遺伝子を持ったMobile Genetic Element

*Streptococcus pyogenes* MGAS2096株は健全な動物の体には多くは居ないが、ゲノム中の、抗生物質耐性遺伝子を多数持ったMobile Genetic Elementは動物の体内の細菌間で広く飛び回っている？

検索ワード: *Clostridium difficile* 630 (ゲノム解読済みの、腸炎の原因菌)

**Microbe DB .JP**

### Definition [help](#)

***Clostridium difficile* 630**

### Taxon Hierarchy [help](#)

Superkingdom	Bacteria
Phylum	Firmicutes
Class	Clostridia
Order	Clostridiales
Family	Peptostreptococcaceae
Species	[ <i>Clostridium</i> ] <i>difficile</i>

### Genome Information [help](#)

BioProject:57679

Description	RefSeq	Type	Size	Gene	tRNA	rRNA	Other
<i>Clostridium difficile</i> 630, complete genome.	NC_009089.1	chromosome	4290252	4008	87	32	102
<i>Clostridium difficile</i> 630 plasmid pCD630, complete sequence.	NC_008226.1	plasmid	7881	11	0	0	6

### Pathogen Information [help](#)

Taxonomy Name	Disease Name	Infectious Type	Strain Type
<i>Clostridium difficile</i> 630	Colitis, Peritonitis, Diarrhea		

### Phenotype Information [help](#)

Cell shape	Rod
Oxygen requirement	Obligate anaerobe
Temperature	Mesophile

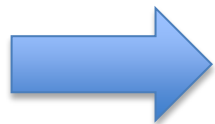
### Mapping to Environment(Chromosome) [help](#)

メタゲノムリードが多数Hitしている3領域は全てMobile Genetic Element

*Clostridium difficile* 630株は健常な動物の体には多くは居ないが、ゲノム中のMobile Genetic Elementは動物の体内の細菌に広く共有されている？

# ドラフトゲノム解析への対応

- 新型シーケンサーの浸透により、誰でもゲノム解読が可能となった
- しかしその多くが完成ゲノム配列ではなく、ドラフトゲノム配列
- シングルセルゲノミクス技術も発達し、ドラフトゲノム配列が爆発的に増加している
- ドラフトゲノム配列を公共DBに登録する際、遺伝子アノテーションは不要
- ドラフトゲノム配列の段階で、自動アノテーションを実施し、同時に新たな知識を発見する事が重要



自動アノテーションシステムMiGAPとの連携



# 微生物ゲノム解析の流れ



パイプライン  
入口



## MiGAP Microbial Genome Annotation Pipeline

ホーム フォーラム FAQ ヘルプ 管理情報

- トップメニュー
  - MiGAPについて
  - お知らせ
  - パイプラインについて
  - MiGAPサーバの運用主体
  - MiGAP引用・関連文献リスト
  - 謝辞
- パイプラインにログイン
- MiGAP
- パイプラインにログイン

公開中です

MiGAP

- フォーラムにログイン

ユーザ名

パスワード

### Home

英国国旗

#### ジョブ投入受付再開しました

2013年 5月 23日(木曜日) 00:00 | 最終更新 2013年 5月 23日(木曜日) 17:03 | 作者: Administrator |

保守のため停止していましたがMiGAPへのジョブ投入受付を再開しました。  
ジョブ投入再開時刻: 2013年5月23日 (木) 16:43

#### 保守に伴うMiGAP投入受付停止のお知らせ

2013年 5月 22日(水曜日) 00:00 | 最終更新 2013年 5月 23日(木曜日) 17:02 | 作者: Administrator |

保守のため、MiGAPへのジョブ投入受付を以下の期間停止しています。  
ジョブ投入受付停止時刻: 2013年5月22日 (水) 16:30  
なお、現在実行中のジョブはそのまま処理されます。また、過去の実行結果の閲覧やダウンロードは投入停止期間中でも可能です。

#### 投入塩基長によるパイプラインへの投入間隔が変更されました

2013年 5月 16日(木曜日) 00:00 | 最終更新 2013年 5月 23日(木曜日) 17:04 | 作者: Administrator |

ジョブ投入間隔の制限設定が変更されました。  
同一利用者が現在投入しているジョブが登録されている場合には、その投入塩基長によって、次のジョブ投入可能になるまでの待ち時間が変わります。

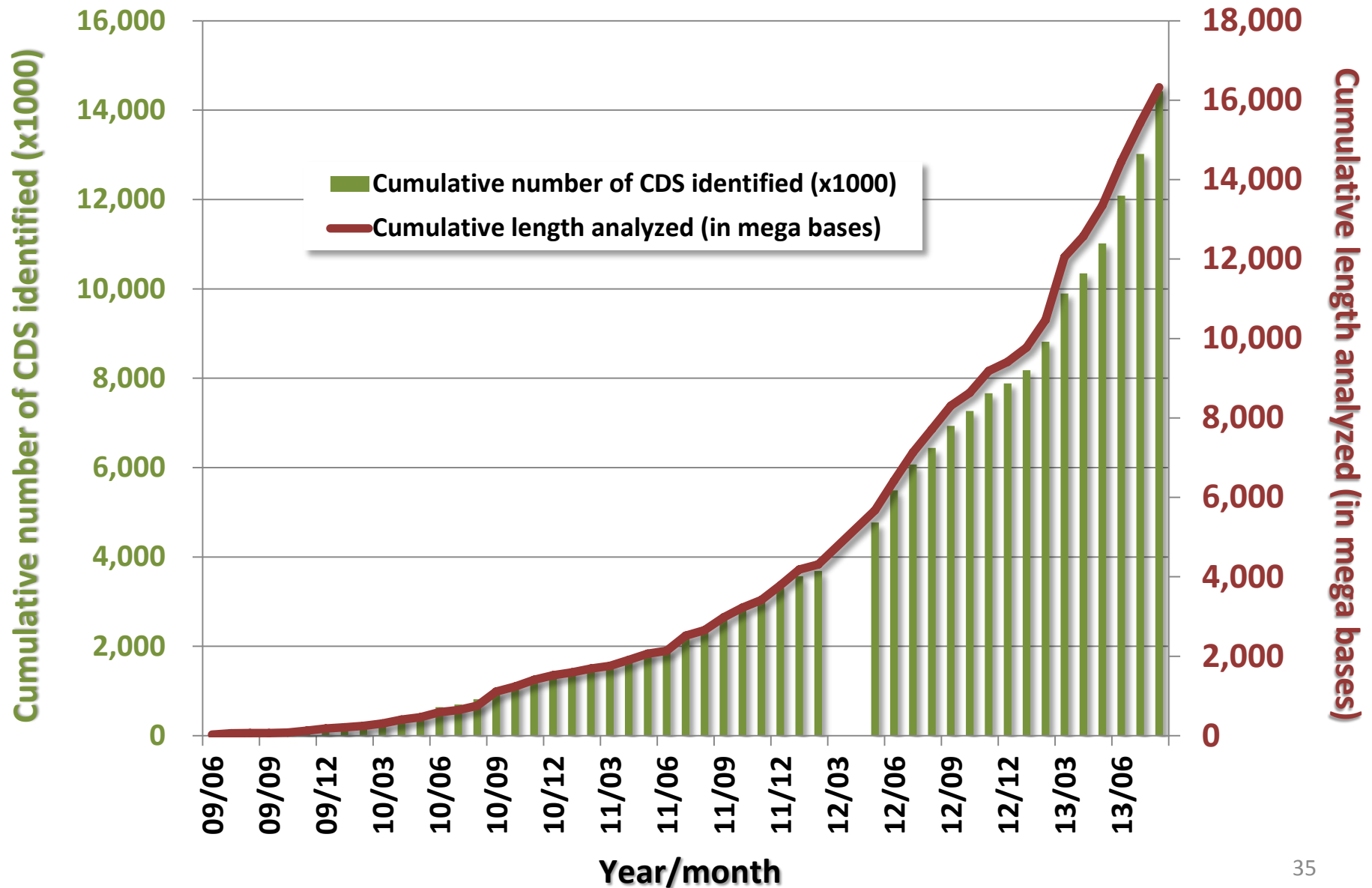
### 最新ニュース

- ジョブ投入受付再開しました
- 保守に伴うMiGAP投入受付停止のお知らせ
- 投入塩基長によるパイプラインへの投入間隔が変更されました
- 遺伝研所内電源工事に伴うMiGAP投入受付停止のお知らせ
- MiGAP 引用・関連文献リスト
- 旧パソコンでのMiGAP実行結果の閲覧とダウンロード再開しました
- 謝辞
- Introduction & Practice (as of May 2012)
- MiGAPにおけるtRNA予測機能の不具合について
- 遺伝研新パソコンでのMiGAPサービスの再開

### 閲覧ランキング

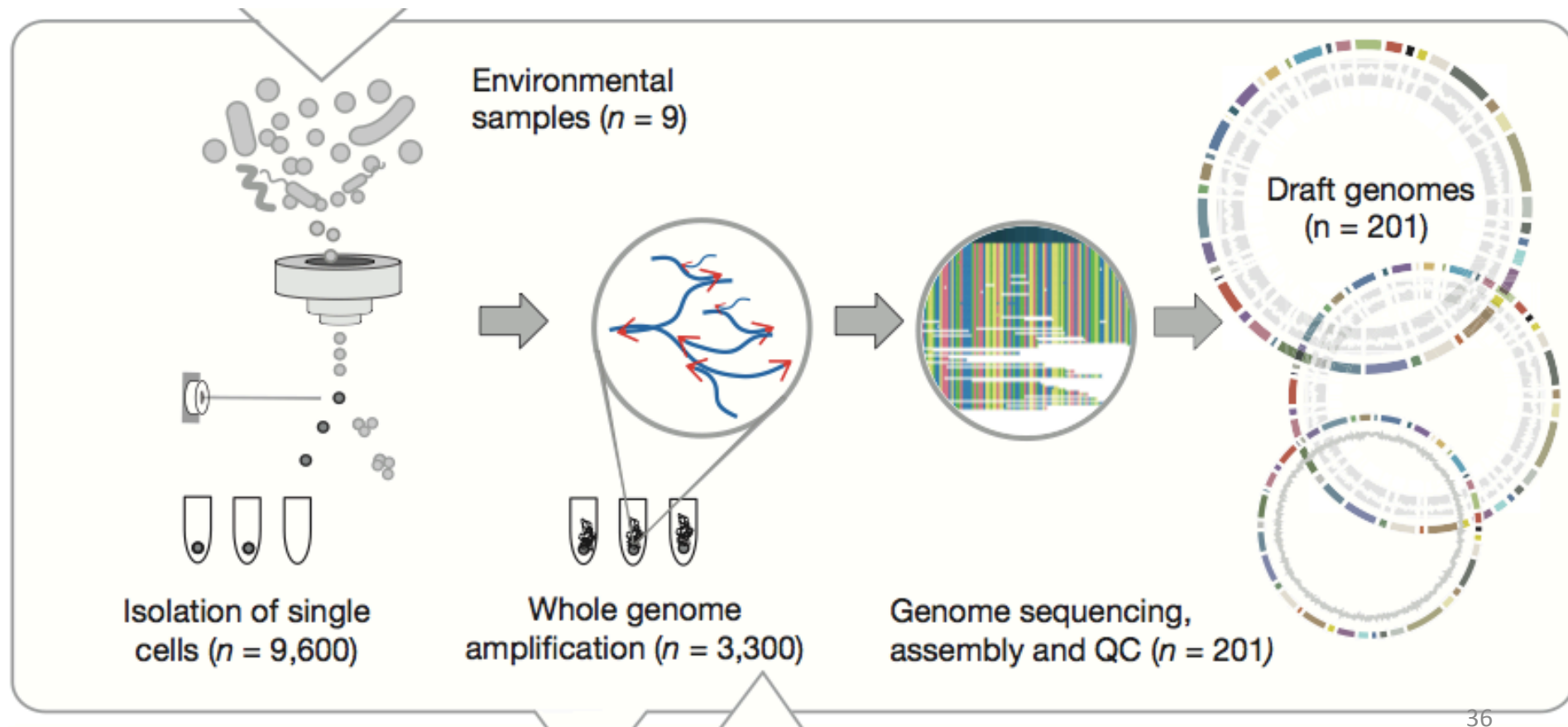
- MiGAPとは?
- MiGAP講習会のお知らせ
- フォーラムヘルプ
- MiGAPサーバの運用主体
- b-MiGAPチュートリアル
- 解析結果をダウンロードす

# MiGAP利用統計

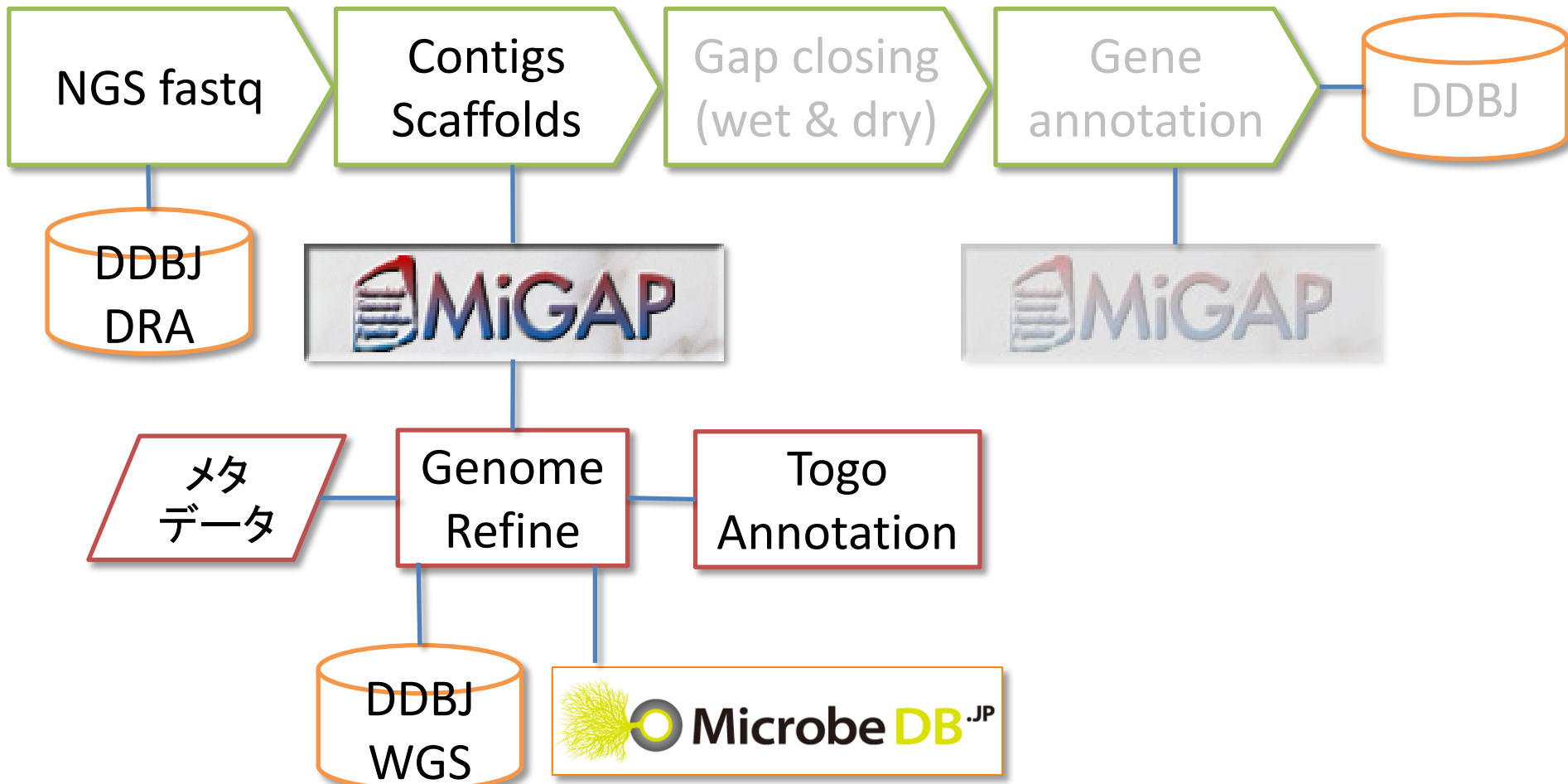


# Insights into the phylogeny and coding potential of microbial dark matter

Christian Rinke<sup>1</sup>, Patrick Schwientek<sup>1</sup>, Alexander Sczyrba<sup>1,2</sup>, Natalia N. Ivanova<sup>1</sup>, Iain J. Anderson<sup>1,†</sup>, Jan-Fang Cheng<sup>1</sup>, Aaron Darling<sup>3,4</sup>, Stephanie Malfatti<sup>1</sup>, Brandon K. Swan<sup>5</sup>, Esther A. Gies<sup>6</sup>, Jeremy A. Dodsworth<sup>7</sup>, Brian P. Hedlund<sup>7</sup>, George Tsiamis<sup>8</sup>, Stefan M. Sievert<sup>9</sup>, Wen-Tso Liu<sup>10</sup>, Jonathan A. Eisen<sup>3</sup>, Steven J. Hallam<sup>6</sup>, Nikos C. Kyrpides<sup>1</sup>, Ramunas Stepanauskas<sup>5</sup>, Edward M. Rubin<sup>1</sup>, Philip Hugenholtz<sup>11</sup> & Tanja Woyke<sup>1</sup>



# ドラフトゲノム解析

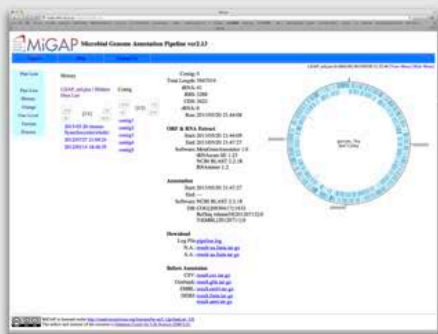


未発表ドラフトゲノムは希望によりアクセスコントロール

# Genome Refineを介したサービス連携

## MiGAP

微生物ゲノム解析において定評あるデータベースと定評あるアルゴリズムを組み合わせたアノテーション実行パイプライン



fasta.csvを入力

## Genome Refine

Genome Refine Sign out

Project Configuration

Genome Project created by MiGAP

Sequence

Annotation

Metadata

organism

locus\_tag\_prefix

bioproject

hold\_date

Assembly Method

Assembly Name

Genome Coverage

Sequence Technology

Isolation Source

Optional Menu

create annotation database

refine gene annotation

Download

[fasta](#) [genbank](#) [gff3](#) [rdf/ttl](#)

未公開genome RDFのデータ共有  
isolation\_sourceに基づく比較解析

## MicrobeDB.jp

微生物に関する多種多様な情報を遺伝子・系統・環境の3つの軸に沿って整理統合し、セマンティックWeb技術を利用して単一の検索ウィンドウからそれらの情報を検索可能な統合DB



## TogoAnnotator

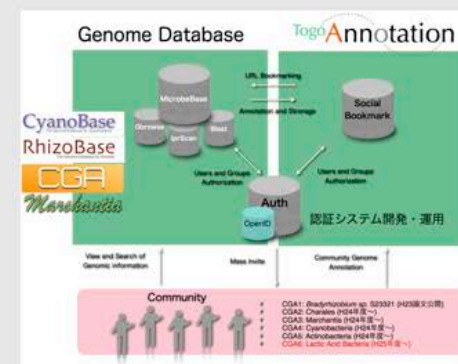
自動機能アノテーションの高精度化

テキストマイニング技術を用いて遺伝子産物名称のアノテーション高精度化ツール。NITE標準アノテーション辞書を利用。

データベースおよびアノテーション環境の提供

CyanoBaseで用いられているゲノムDBとソーシャルブックマークの仕組みでアノテーションを行なうTogoAnnotationを組み合わせ、コミュニティゲノムアノテーション環境を提供。DDBJ登録、データベース公開等のゲノムアノテーション支援も実施。

## ゲノムアノテーション編集環境



# モデル微生物アノテーション高度化

- 研究者コミュニティの規模が大きく、多数の研究成果が期待できるモデル微生物をターゲットとする
- 日々更新されるデータを文献情報から収集するとともに、研究者およびキュレータによる高度アノテーションをデータに付加
- 得られた高度なアノテーション情報を MicrobeDB.jp と統合するため、オミックス情報など全情報をRDF化



TogoAnnotationシステムの開発

# 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

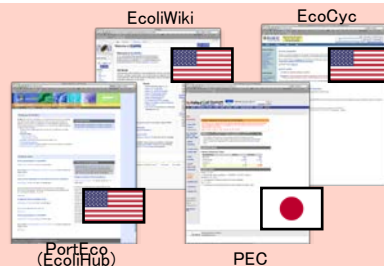
モデル微生物

リファレンス株

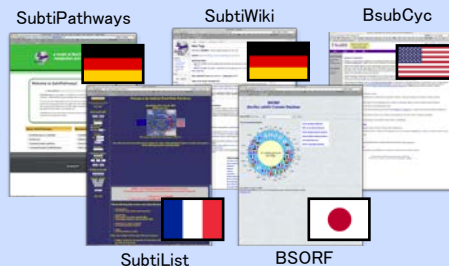
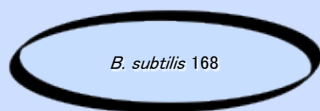
リファレンス株遺伝子の関連文献が  
参照可能なデータベース

国内でゲノム解析された  
病原性/産業有用株

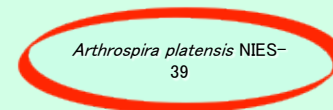
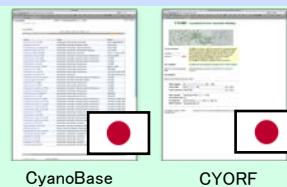
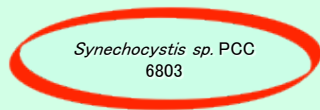
大腸菌



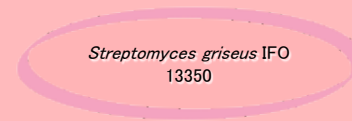
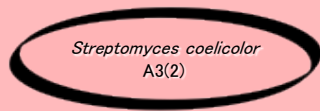
枯草菌



ラン藻



放線菌

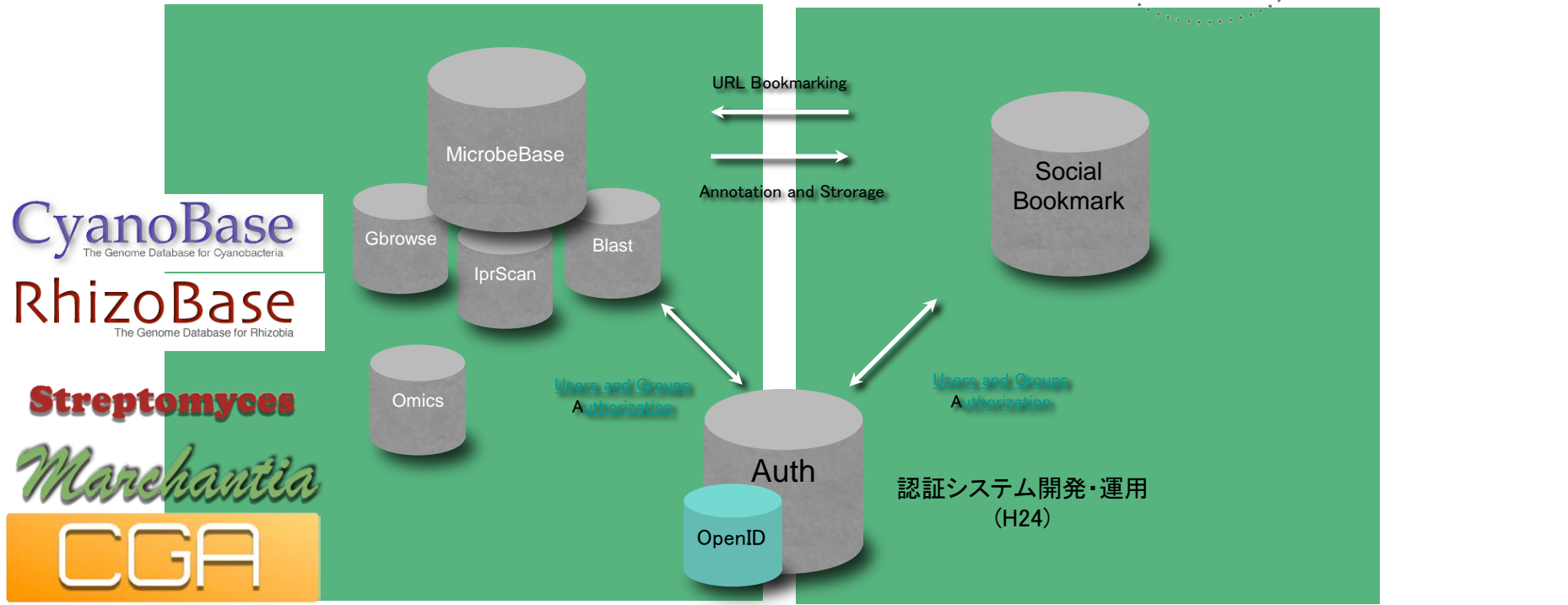




# ゲノムアノテーションプラットフォームによる 研究コミュニティ連携・支援

## Genome Database

## TogoAnnotation



[View and Search of Genomic information](#)

[Mass Invite](#)

[Community Genomic Annotation](#)

CGA1: *Bradyrhizobium* sp. S23321 (H23論文公開)

CGA2: Charales (H24年度～)

CGA3: Marchantia (H24年度～)

CGA4: Cyanobacteria (H24年度～)

CGA5: Actinobacteria (H24年度～)

CGA6: Lactic Acid Bacteria (H25年度～)

Community



# モデル微生物遺伝子の文献リファレンス集積

Phylum	Organism	Literatures	Annotations	Annotated genes	Total genes
cyanobacteria	<i>Synechocystis</i> sp. PCC 6803	2346	80204	3064	3725
	<i>Anabaena</i> sp. PCC 7120	959	29154	2754	6223
	<i>Synechococcus elongatus</i> PCC 7942	815	17060	794	2715
	<i>Thermosynechococcus elongatus</i> BP-1	270	6768	2528	2528
	<i>Synechococcus</i> sp. PCC 7002	264	3999	265	3235
	<i>Nostoc punctiforme</i> ATCC 29133	151	3349	768	6794
	<i>Chlorobium tepidum</i> TLS	143	5532	751	2310
	<i>Anabaena variabilis</i> ATCC 29413	119	1731	258	5724
	<i>Prochlorococcus marinus</i> MED4	64	2155	390	1756
	<i>Gloeobacter violaceus</i> PCC 7421	52	5600	4483	4484
	<i>Prochlorococcus marinus</i> MIT9313	44	919	248	2326
	<i>Prochlorococcus marinus</i> SS120	37	539	135	1928
	<i>Arthrospira platensis</i> NIES-39	9	787	260	6676
	<i>Trichodesmium erythraeum</i> IMS101	5	22	14	4498
	<i>Synechococcus</i> sp. WH8102	5	38	22	2579
	<i>Synechococcus elongatus</i> PCC 6301	2	5	2	2580
actinobacteria	<i>Streptomyces coelicolor</i> A3(2)	161	1185	18703	8300
	<i>Streptomyces griseus</i> NBRC 13350	116	311	12970	7224

2013.08.31現在

シアノバクテリア 5,285 papers, 放線菌 277 papers

# モデル微生物ゲノム情報と オミックスデータのRDFによる統合

シアノバクテリアの例

Data type	The number of resource	RDF
genome project	39	○
gene	138896	○
publication	5285	*
operon*	86	○
protein complex*	68	○
protein-protein interaction	3054	○


\* RDF data model is under development

<http://microbedb.jp/>

MicrobeDB

microbedb.jp/MDB

[Sign In](#)

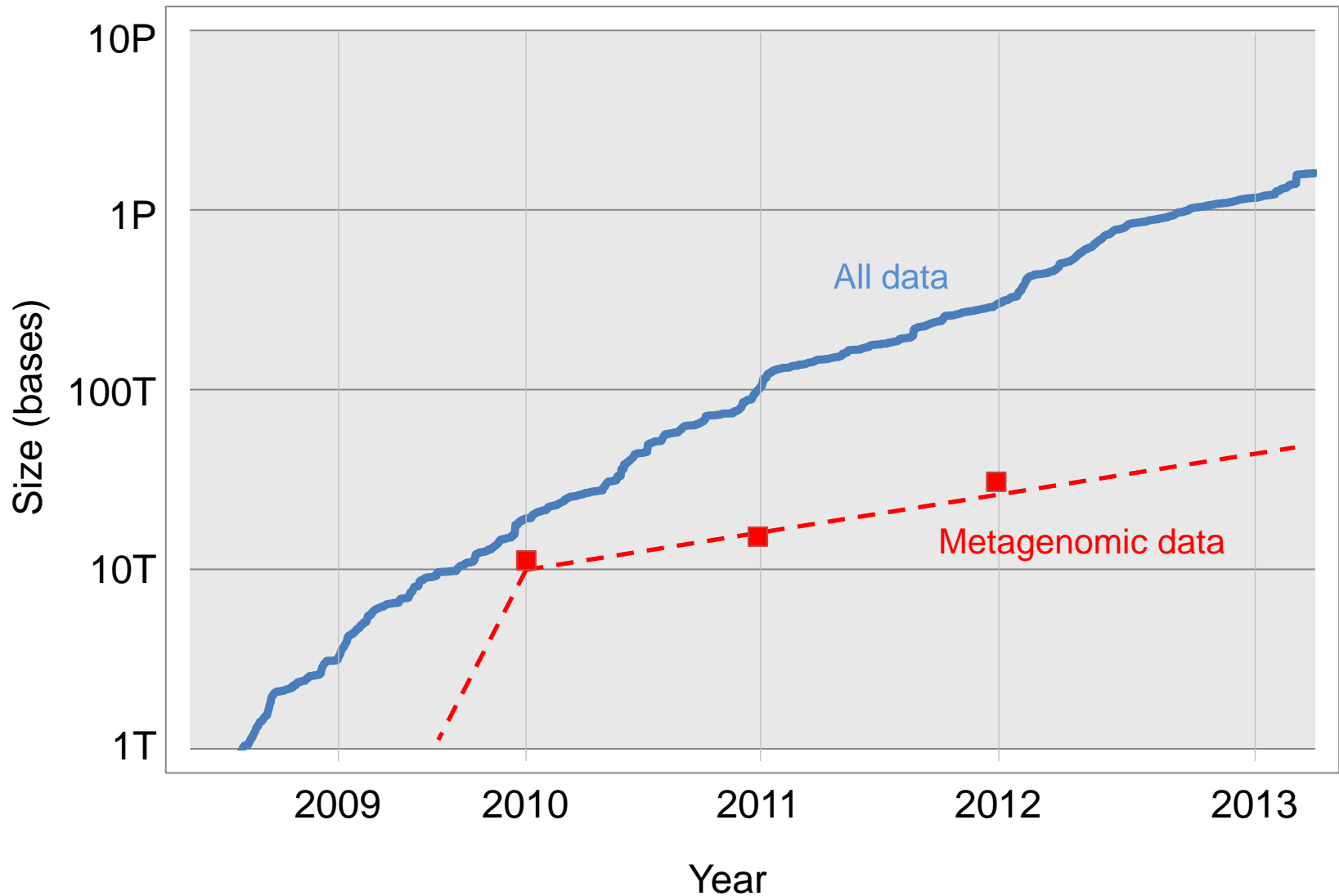


Gene: psbA  
Taxonomy: Streptococcus glycerinaceus  
Mapping: Escherichia coli O157:H7 str. Sakai  
Environment: hot spring  
SRS: rumen  
Strain: Bifidobacterium  
Disease: Cholera  
MiGap: GAF

# MicrobeDB.jpの国際化

- International Council for Science (ICSU) Committee on the Data for Science and Technology (CODATA) Task Group on Advancing Informatics for Microbiology (TG-AIM)
  - 微生物情報の標準化にむけて
- GBIF(DarwinCore) および Genomic Standards Consortium (GSC) (MIxS)
  - 共同でオントロジーを設計
- WFCC-MIRGEN World Data Center for Microorganisms (WDCM)
  - Global Catalogue of Microorganismsにおいてオントロジーの採用を検討中
- Microbial Resource Research Infrastructure (MIRRI@EU)
  - 用語の標準化/オントロジー開発にむけて議論
- Quest for Ortholog MTG
  - オースログの標準形式開発においてRDFを検討

# NCBI SRA DataBase Growth



# Marine Phage Sequencing Project

Marine Phage Sequencing Project: Home

www.broadinstitute.org/annotation/viral/Phage/Home.html

BROAD INSTITUTE

History & Leadership Education Contribute Careers Contact Us

What is Broad News and Publications For the Scientific Community

Science Data Software

Home > Science > Data > Marine Phage Sequencing Project > Home

## Marine Phage, Viruses and Viromes

The Gordon & Betty Moore Foundation's Marine Microbiology Initiative (MMI) aims to generate new knowledge about the composition, function, and ecological role of the microbial communities that serve as the basis of the ocean's food webs and that facilitate the flow of nitrogen, carbon, and energy in the ocean. Phage and viruses play a critical role in shaping microbial diversity, abundance and evolution and in this capacity have a significant impact on atmospheric composition and ecosystem function. In an effort to understand the ecology and evolution of marine phage and viruses and to explore the diversity and ecological roles of entire phage/viral communities through metagenomics, the Broad Institute is collaborating with MMI and research groups whose sequencing priorities were chosen by

**News**

- 2010.12.20: Sequencing, Assembly, and Annotation for all samples completed; all project data available from CAMERA
- Project status updates are [here](#)
- Audio of assembly, annotation, and file format tutorial is available

Marine Phage Sequencing Project

- Home
- Project Info
- Protocols
- Links
- Contact

Genomic Sequencing Center for Infectious Disease

# Earth Microbiome Project

Earth Microbiome Project

www.earthmicrobiome.org

earth microbiome project

Home Defining the Tasks Getting Involved EMP Protocols and Standards Affiliations Publications Meetings EMP Logo No categories

## The Earth Microbiome Project is a systematic attempt to characterize the global taxonomic and functional diversity for the benefit of the planet and mankind

**Constructing the Microbial Biomap for Planet Earth**

The Earth Microbiome Project is a proposed massively multidisciplinary effort to analyze microbial communities across the globe. The general premise is to examine microbial communities from their own perspective. Hence we propose to characterize the Earth by environmental parameter space into different biomes and then explore these using samples currently available from researchers across the globe. We will analyze 200,000 samples from these communities using metagenomics, metatranscriptomics and amplicon sequencing to produce a

SEARCH

Meetings

There are currently no EMP centric meetings planned, however we will update this space as soon as the next meeting is organized.

News

Neil Hall highlights the new-stabilizing price of sequencing and how the EMP and others have

# Home Microbiome Study

Home Microbiome Study

homemicrobiome.com

HOME ANNOUNCEMENTS RESULTS ABOUT

## The Home Microbiome Study

Published February 24, 2012 - No Comments

Most of us are aware of the bacteria on the surfaces we come in contact with. The doorknob for the bathroom, coins and paper currency in our pockets, handrails in subway cars, computer keyboards and mice at the library... the list of built environments in which microbes thrive is nearly on ...

read more

Microbial Biodiversity of Cell Phones and Shoes

Published February 24, 2012 at 4:08 am - No Comments

The results are in! So, what lives on reporters' cell phones and shoes, you ask? As you can see from the graph at left, quite a lot! Each vertical bar is one shoe or cell phone and each color represents a different

The Home Microbiome Study

Published February 24, 2012 at 2:58 pm - No Comments

Most of us are aware of the bacteria on the surfaces we come in contact with. The doorknob for the bathroom, coins and paper currency in our pockets, handrails in subway cars, computer keyboards and mice at the

Latest Articles

Latest Articles from the Blog

- GSC3 Presentation
- Published June 12, 2012 at 2:41 pm
- Microbial Biodiversity of Cell Phones and Shoes
- Published February 24, 2012 at 4:08 am

# Hospital Microbiome Project

Hospital Microbiome Project

hospitalmicrobiome.com

Home Goals Overview Design Timeline Consortium Findings Affiliations

## Hospital Microbiome

This study aims to collect microbial samples from surfaces, air, staff, and patients from the University of Chicago's new hospital pavilion in order to better understand the factors that influence bacterial population development in healthcare environments.

Study Design

47

# MicrobeDB.jpの今後の展開

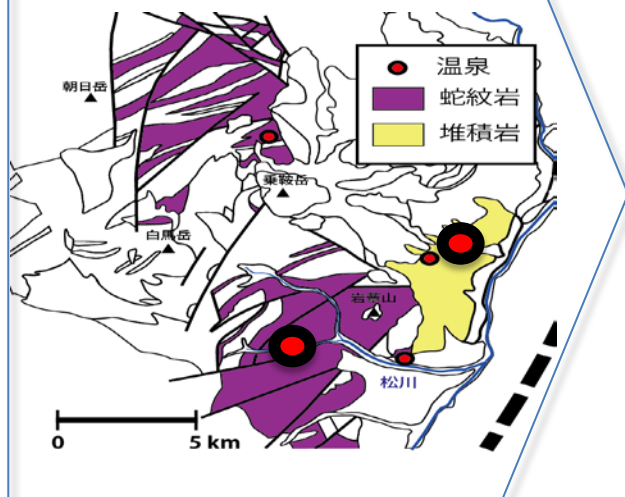
- 上位生物データの統合化（藻類、菌類 etc.）
  - Hologenomicsの到来に備える必要性
- より高度な環境データとの連携の模索（地質学データ、地球化学データ、リモートセンシングデータ etc.）
  - 各分野のオンロジーとの連携





# 蛇紋岩体を貫く熱水環境

## Geological survey



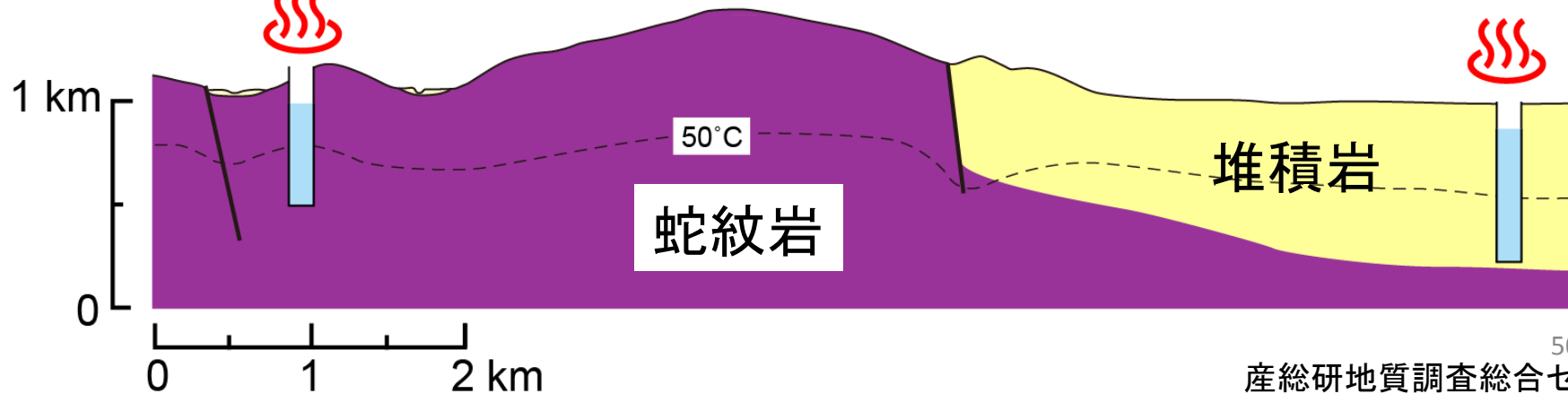
## Sample collection Env. measurements



## Metagenomic sequencing by NGS



## A温泉



## B温泉

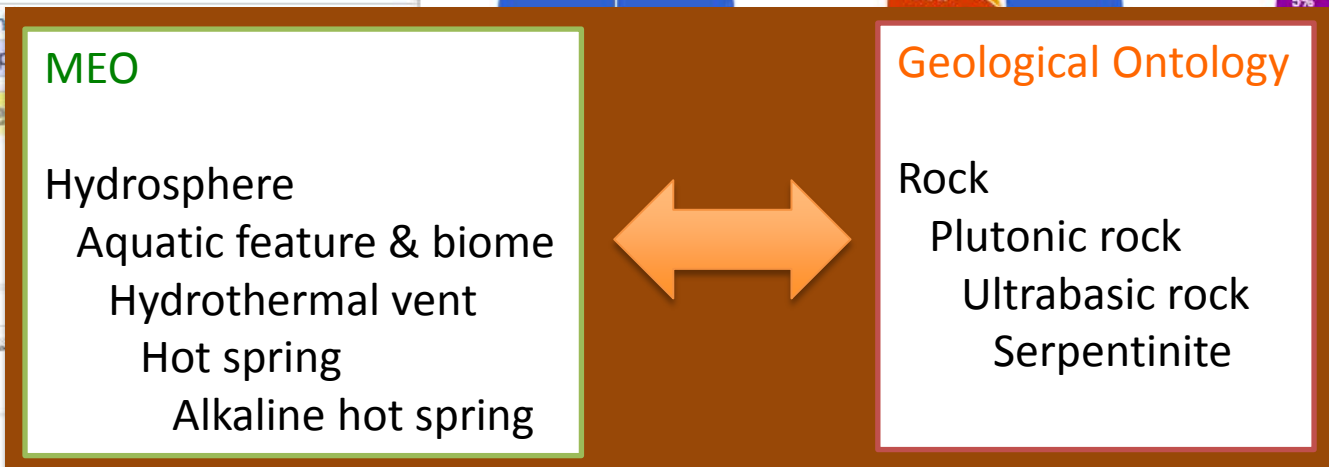
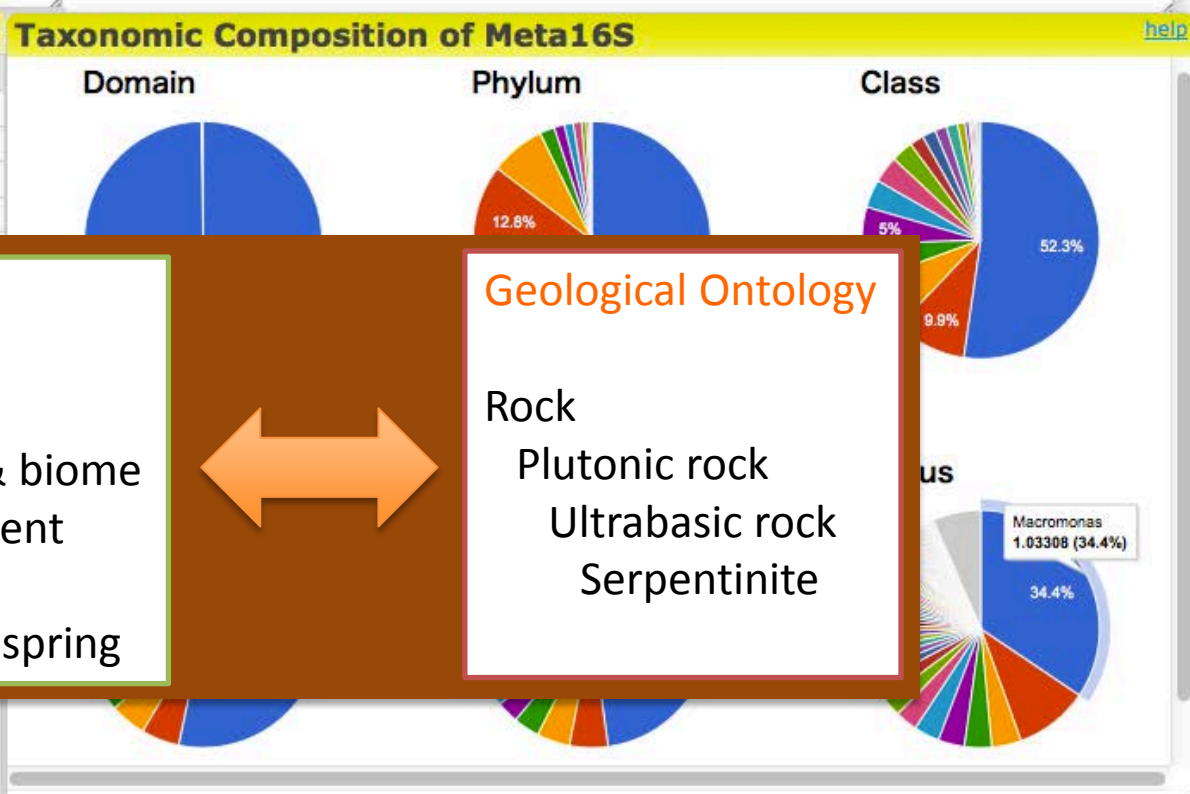


# 検索ワード: serpentine (蛇紋岩)

Definition <a href="#">help</a>	
MEO ID	MEO_0000813
Title	serpentine hot spring
Definition	A hot spring whose water venting from serpentinites.
MEO SuperClass ID	MEO_0000730
MEO SuperClass Title	alkaline hot spring

Meta16S Sample List <a href="#">help</a>			
ID	Title	Name	Environment
<a href="#">SRS417033</a>		groundwater metagenome	serpentine hot spring,
<a href="#">SRS417034</a>		groundwater metagenome	serpentine hot spring,
<a href="#">SRS417032</a>		groundwater metagenome	serpentine hot spring,

MEO Hierarchy <a href="#">help</a>	
ID	Title
MEO_0000004	<a href="#">hydrosphere</a>
MEO_0000425	<a href="#">aquatic feature and biome</a>
MEO_0000031	<a href="#">hydrothermal vent</a>
MEO_0000029	<a href="#">hot spring</a>
MEO_0000730	<a href="#">alkaline hot spring</a>
MEO_0000813	<a href="#">serpentine hot spring</a>



蛇紋岩熱水由来の細菌群集の系統組成等の情報が容易に得られる

# MicrobeDB.jpの今後の展開

- 上位生物データの統合化（藻類、菌類 etc.）
  - Hologenomicsの到来に備える必要性
- より高度な環境データとの連携の模索（地質学データ、地球化学データ、リモートセンシングデータ etc.）
  - 各分野のオンロジーとの連携
- ビッグデータ次世代基盤技術開発との連携
  - 次世代基盤開発に至適な課題

# 謝辞

- NBRC、JCM
- 日本微生物資源学会、日本ゲノム微生物学会、日本微生物生態学会、日本農芸化学会、日本細菌学会
- DBCLS、豊田フェノームG



# 參考資料

# 3年間の活動実績

内部開発会議	30回
SPARQLthon	12回
BioHackathon(国内版含む)	4回
学会等でのブース出展	7回
国内学会発表	23回
国際学会発表	5回



# Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution

Ilana Zilber-Rosenberg<sup>1</sup> & Eugene Rosenberg<sup>2</sup>

<sup>1</sup>Teaching at the Open University of Israel, Raanana, Israel; and <sup>2</sup>Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Ramat Aviv, Israel

FEMS Microbial Rev, 2008

## Hologenomics & Hologenome Sequencing Applications

Hologenome<sup>§</sup>, [*Holo* Greek, from *holos*, whole;] a term borrowed from evolutionary biology is defined as the sum of the genetic information of an organism, or in rather words, the organism and its microbiota. The Hologenome thus signifies a mixed population of genomes. Hologenomics deals with the genomics of a hologenome of mixed populations of genomes, as in the case of interacting populations in host-pathogen and commensals. Hologenome differs from the widely popular term Metagenome, which involves the study of communities of microbes directly in their natural environments.

The availability of next-generation sequencing (NGS) technology has enabled the scale and ease of addressing biological questions on a genomics perspective. The throughput of sequencing enables deep sequencing of nucleic acids, adequate to provide for enough reads of the pathogen, even while the interference of the host genetic material is very high. Metagenomics has been one of the major applications of NGS technology for understanding the composition and dynamics of mixed population of organisms.

### Publications

#### Application of Hologenome Sequencing in Microbiology

Patowary A, Chauhan RK, Singh M, Shamsudheen KV, Periwal V, Kushwaha KP, Sapkal GN, Bondre VP, Gore MM, Sivasubbu S and Scaria V

#### De novo identification of viral pathogens from cell culture hologenomes

*BMC Research Notes* 2012, 5:1