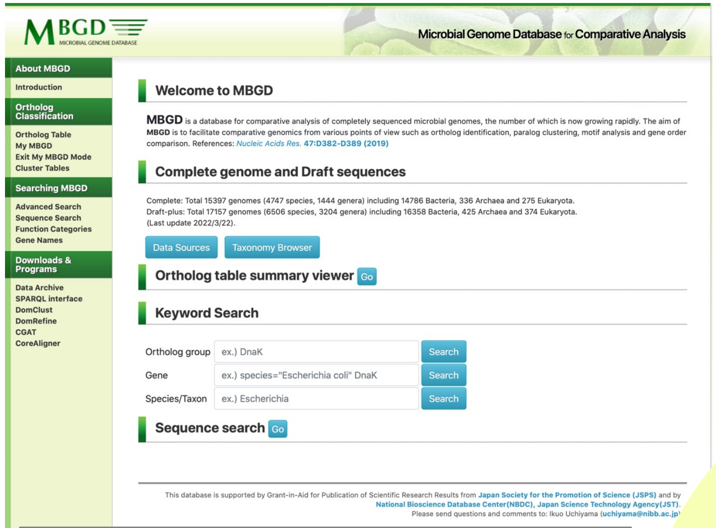
12

MBGD 2022:　大規模な微生物オーソログデータベースの効果的な活用に向けて

〇内山郁夫[1]、三原基広[2]、西出浩世[1]、千葉啓和[3]、高柳正彦[4]、髙見英人[5]

1. 基礎生物学研究所、2. （株）ダイナコム、3. ライフサイエンス統合データベースセンター、
4. （株）ウェブブレイン、5. 東京大学大気海洋研究所

微生物比較ゲノムデータベースMBGD



MBGD MICROBIAL GENOME DATABASE

Microbial Genome Database for Comparative Analysis

Welcome to MBGD

MBGD is a database for comparative analysis of completely sequenced microbial genomes, the number of which is now growing rapidly. The aim of **MBGD** is to facilitate comparative genomics from various points of view such as ortholog identification, paralog clustering, motif analysis and gene order comparison. References: *Nucleic Acids Res.* **47:D382-D389 (2019)**

Complete genome and Draft sequences

Complete: Total 15397 genomes (4747 species, 1444 genera) including 14786 Bacteria, 336 Archaea and 275 Eukaryota.
Draft-plus: Total 17157 genomes (6506 species, 3204 genera) including 16358 Bacteria, 425 Archaea and 374 Eukaryota.
(Last update 2022/3/22).

Ortholog table summary viewer

Keyword Search

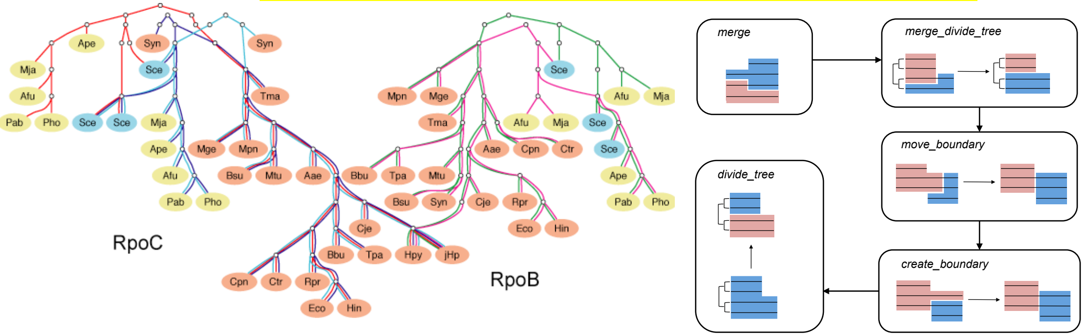
Ortholog group: ex.) DnaK
Gene: ex.) species="Escherichia coli" DnaK
Species/Taxon: ex.) Escherichia

Sequence search

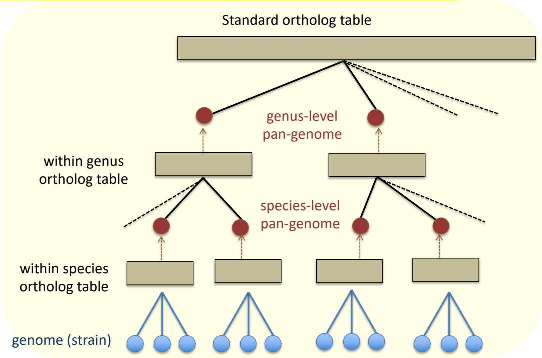
This database is supported by Grant-in-Aid for Publication of Scientific Research Results from Japan Society for the Promotion of Science (JSPS) and by National Bioscience Database Center(NBDC), Japan Science Technology Agency(JST).
Please send questions and comments to: Ikuo Uchiyama (uchiyama@nibb.ac.jp)

微生物ゲノムデータベースＭＢＧＤは、オーソログ解析に基づいて微生物ゲノムの比較解析を行うためのデータベースである。全ゲノムが決定されて公開された微生物ゲノムに対して全生物種および系統群毎のオーソログテーブルを作成て提供している。

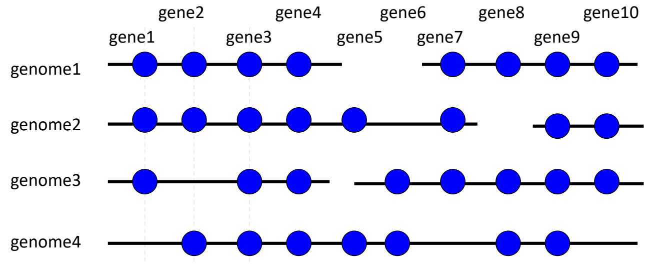
ドメイン単位のオーソロググルーピング



階層的オーソロググループ



コアゲノム解析



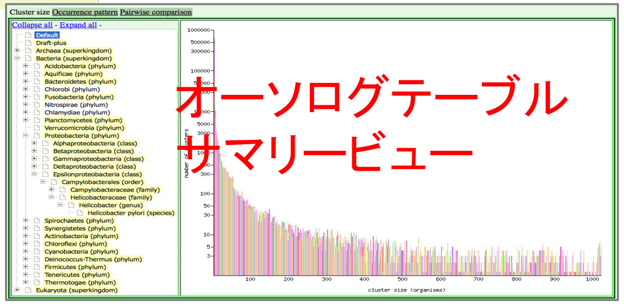
オーソログテーブル

ゲノム

遺伝子

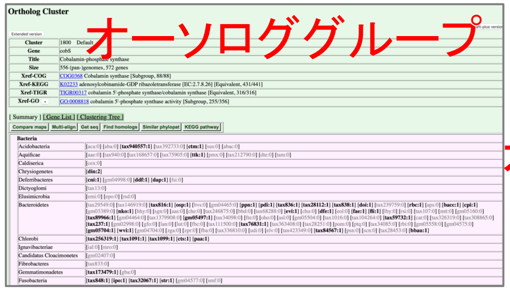
Analyze the checked clusters

ClusterID	Name	#species	#genes	Description	Phylogenetic pattern	
O1800 A M H P	cobS	556	572	Cobalamin-phosphate synthase		0.2059694543
O1788 A M H P	cobQ	520	575	Cobyric acid synthase CobQ		0.2490342775
O1637 A M H P	cobD	556	620	Cobalamin biosynthesis protein CobD		0.2600765799
O1752 A M H P	cobB	484	584	Cobyrinic acid a,c-diamide synthase		0.2734526405
O2071 A M H P	cobU	461	505	Protein kinase; Adenosylcobinamide-phosphate		0.2797435869
O2452 A M H P	cbiL	410	421	Precorrin-2 c20-methyltransferase		0.2897022826
O2189 A M H P	cobH	416	478	Precorrin-8X methylmutase		0.2984332031
O2113 A M H P	cbiH	417	494	Precorrin-3B C17-methyltransferase		0.2991509829
O2524 A M H P	cobL	382	408	Precorrin-6Y C5,15-methyltransferase		0.2998493284
O2303 A M H P	cobL	399	452	Precorrin-6Y C5,15-methyltransferase		0.3036746508
O2093 A M H P	cobT	431	500	Nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase		0.3112438511
O3008 A M H P	cbiD	327	335	Cobalt-precorrin-6A synthase CbiD		0.3267527288
O3208 A M H P	cbiG	300	311	Cobalamin biosynthesis protein CbiG		0.3392871296
O1133 A M H P	cobN	337	587	Magnesium chelatase subunit/cobaltochelatase, CobN subunit		0.3439789959
O3010 A M H P	cobK	311	334	Cobalt-precorrin-6x reductase		0.3461465471
O1708 A M H P	cobO	476	599	Cob(I)alamin adenolsyltransferase/cobinamide ATP-dependent adenolsyltransferase		0.3524128007
O2441 A M H P	bchD	252	422	Magnesium chelatase		0.3574875827
O3340 A M H P	cbiG	287	298	Cobalamin biosynthesis protein CbiG		0.3639510683
O2405 A M H P	cbiQ	299	428	Cobalt ABC transporter, permease protein CbiQ		0.3644695281
O2056 A M H P	cbiX	369	508	Cobalamin biosynthesis CbiX protein		0.3660824842
O2204 A M H P	cbiM	340	475	Cobalamin biosynthesis protein CbiM		0.369192191



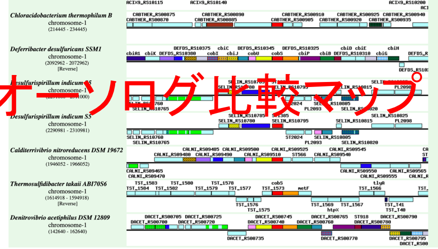
Cluster size Occurrence pattern Pairwise comparison

オーソログテーブル サマリービュー



Ortholog Cluster

オーソロググループ

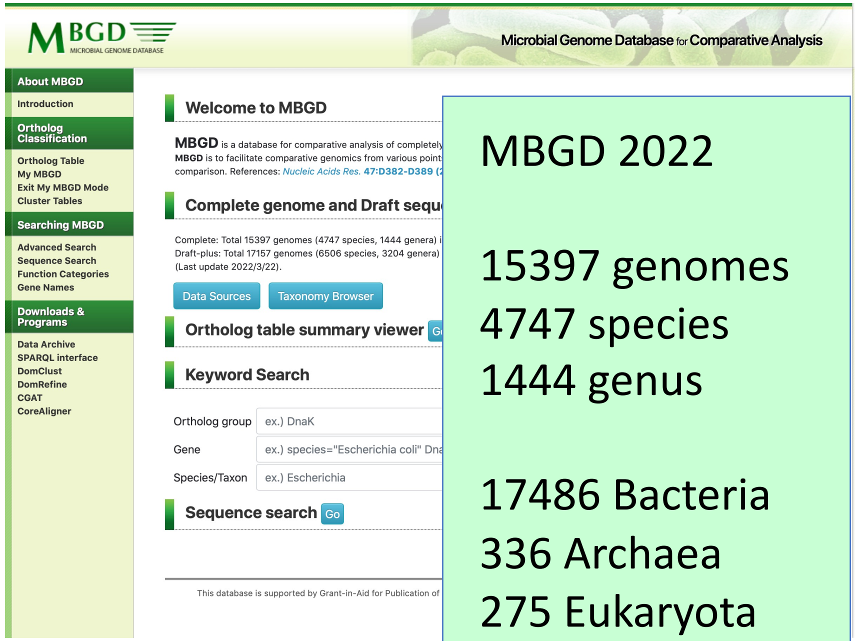


Genome R

オーソログ比較マップ

マルチプル アライメント

MBGD登録ゲノム数の推移



Welcome to MBGD

MBGD is a database for comparative analysis of completely
MBGD is to facilitate comparative genomics from various point
comparison. References: *Nucleic Acids Res.* **47:D382-D389**

Complete genome and Draft sequ

Complete: Total 15397 genomes (4747 species, 1444 genera) i
Draft-plus: Total 17157 genomes (6506 species, 3204 genera)
(Last update 2022/3/22).

Ortholog table summary viewer

Keyword Search

Ortholog group ex.) DnaK
Gene ex.) species="Escherichia coli" Dna
Species/Taxon ex.) Escherichia

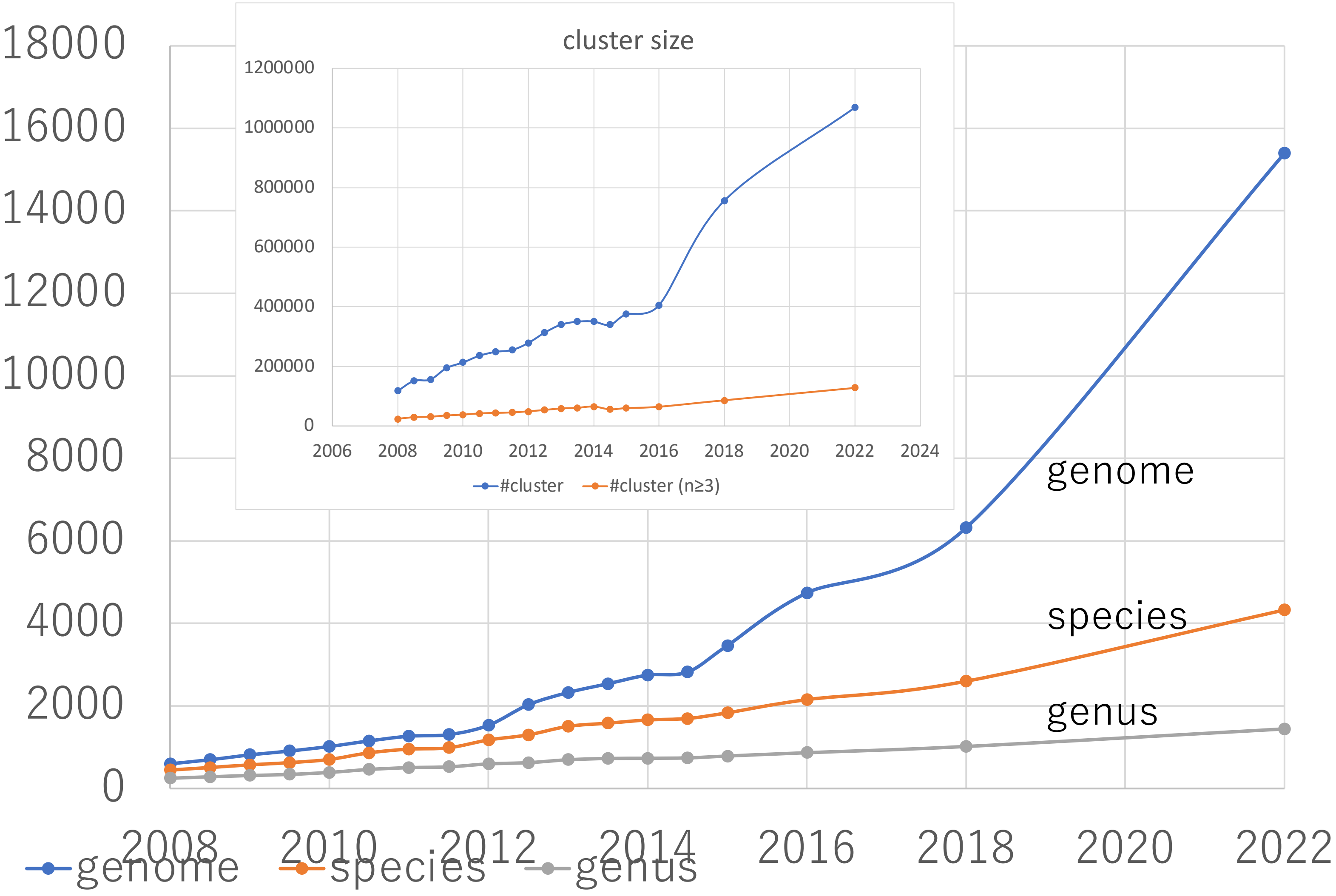
Sequence search

This database is supported by Grant-in-Aid for Publication of

MBGD 2022

15397 genomes
4747 species
1444 genus

17486 Bacteria
336 Archaea
275 Eukaryota



MyMBGD mode: ユーザゲノムの解析

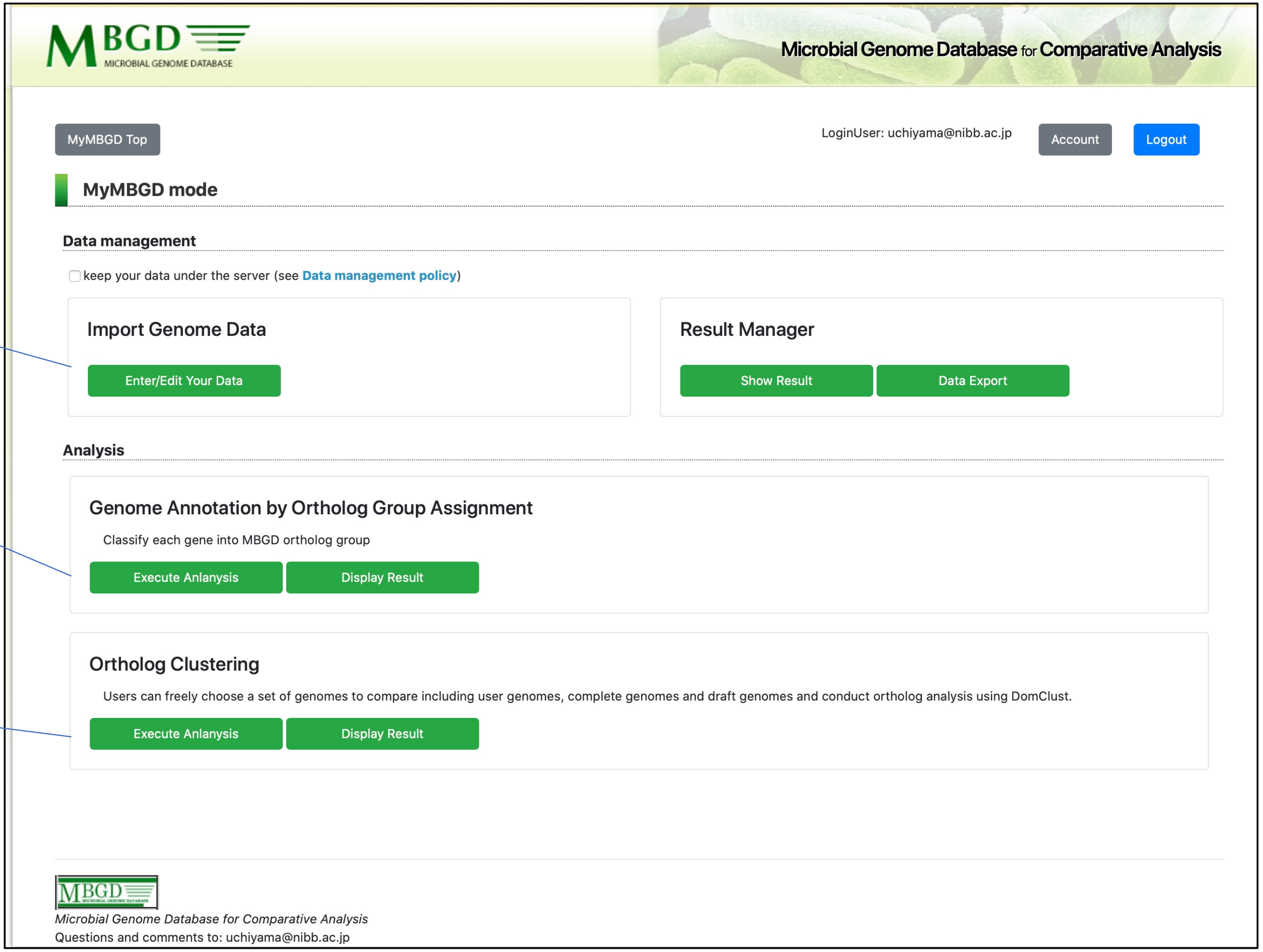
ユーザゲノムの登録

アサインメントモード

ユーザゲノムの各遺伝子を既存オーソロググループに分類する

クラスタリングモード

ユーザゲノムとMBGDに登録された公開ゲノムを合わせた新規クラスタリングによるオーソログ解析



Microbial Genome Database for Comparative Analysis

LoginUser: uchiyama@nibb.ac.jp

MyMBGD mode

Data management

☐ keep your data under the server (see **Data management policy**)

Import Genome Data

Result Manager

Analysis

Genome Annotation by Ortholog Group Assignment

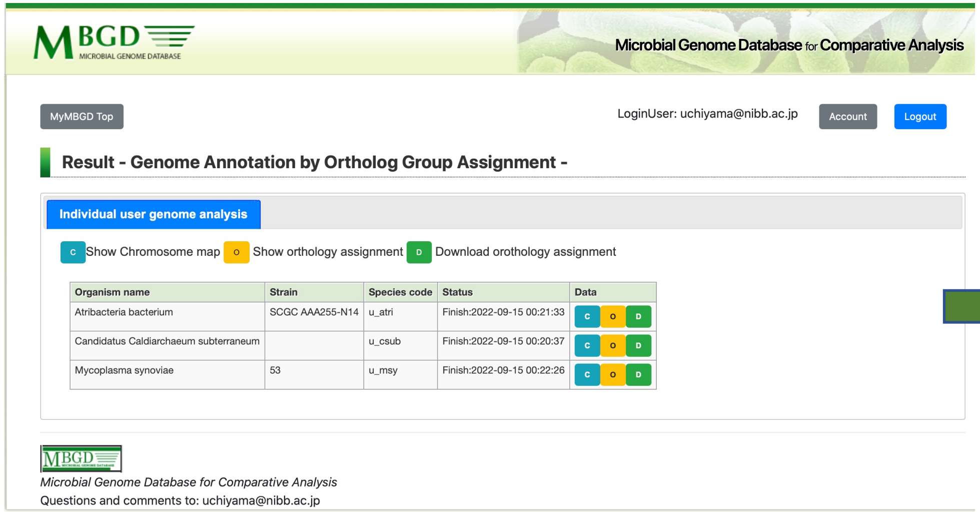
Classify each gene into MBGD ortholog group

Ortholog Clustering

Users can freely choose a set of genomes to compare including user genomes, complete genomes and draft genomes and conduct ortholog analysis using DomClust.

Microbial Genome Database for Comparative Analysis
Questions and comments to: uchiyama@nibb.ac.jp

MyMBGD mode: アサインメントモード



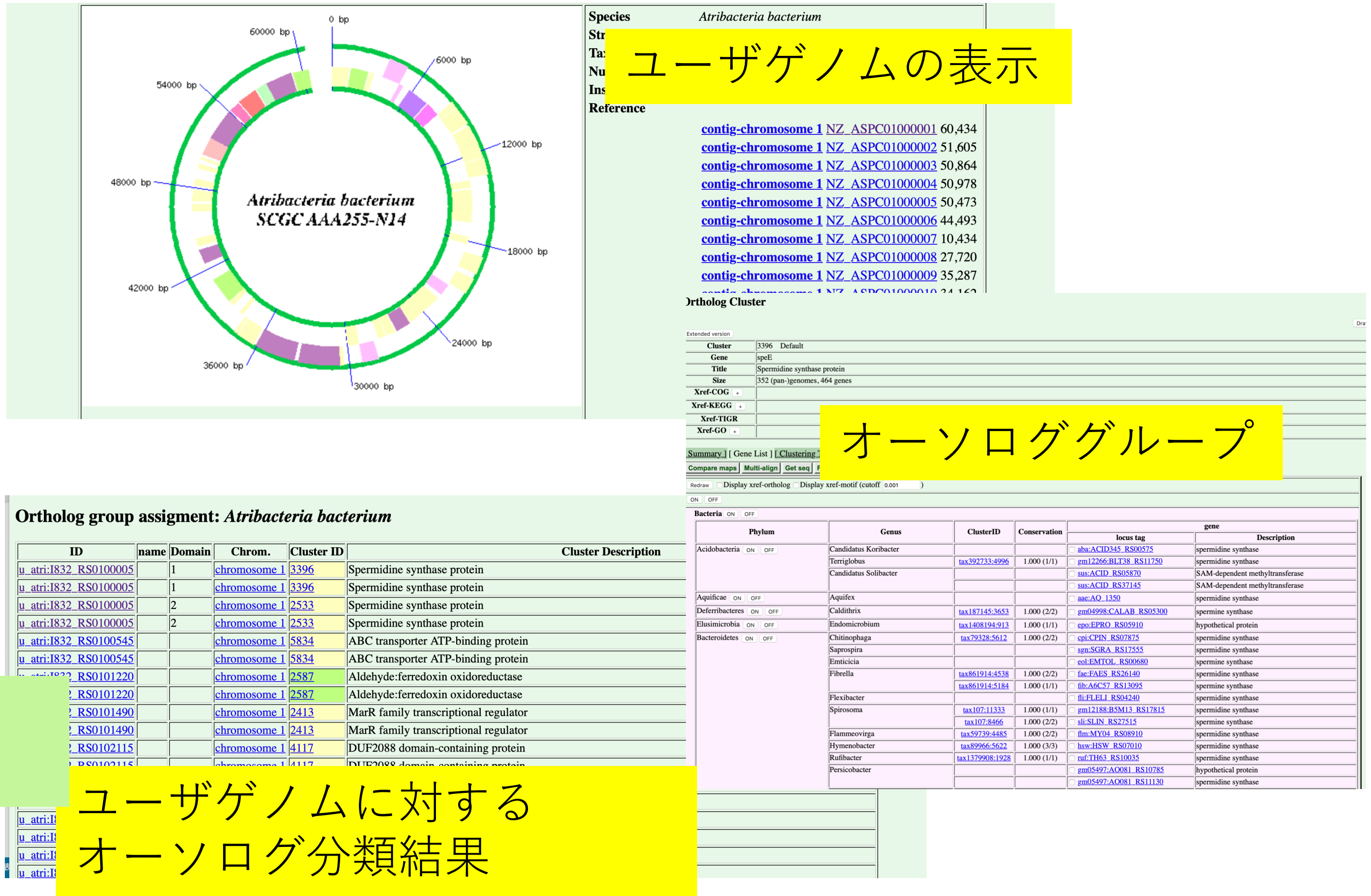
LoginUser: uchiyama@nibb.ac.jp

Result - Genome Annotation by Ortholog Group Assignment -

Organism name	Strain	Species code	Status	Data
Atribacteria bacterium	SCGC AAA255-N14	u_atri	Finish:2022-09-15 00:21:33	C O D
Candidatus Caldiarchaeum subterraneum		u_csub	Finish:2022-09-15 00:20:37	C O D
Mycoplasma synoviae	53	u_msy	Finish:2022-09-15 00:22:26	C O D

Microbial Genome Database for Comparative Analysis
Questions and comments to: uchiyama@nibb.ac.jp

MMSeqsによるオーソログ
DBのプロファイル検索



ユーザゲノムの表示

Species *Atribacteria bacterium*
Str
Ta
Nu
Ins
Reference

contig-chromosome 1	NZ_ASPC01000001	60,434
contig-chromosome 1	NZ_ASPC01000002	51,605
contig-chromosome 1	NZ_ASPC01000003	50,864
contig-chromosome 1	NZ_ASPC01000004	50,978
contig-chromosome 1	NZ_ASPC01000005	50,473
contig-chromosome 1	NZ_ASPC01000006	44,493
contig-chromosome 1	NZ_ASPC01000007	10,434
contig-chromosome 1	NZ_ASPC01000008	27,720
contig-chromosome 1	NZ_ASPC01000009	35,287

rtholog Cluster

Cluster	3396 Default
Gene	speE
Title	Spermidine synthase protein
Size	352 (pan-)genomes, 464 genes
Xref-COG +	
Xref-KEGG +	
Xref-TIGR	
Xref-GO +	

オーソロググループ

☐ Display xref-ortholog ☐ Display xref-motif (cutoff 0.001)

Bacteria ON OFF

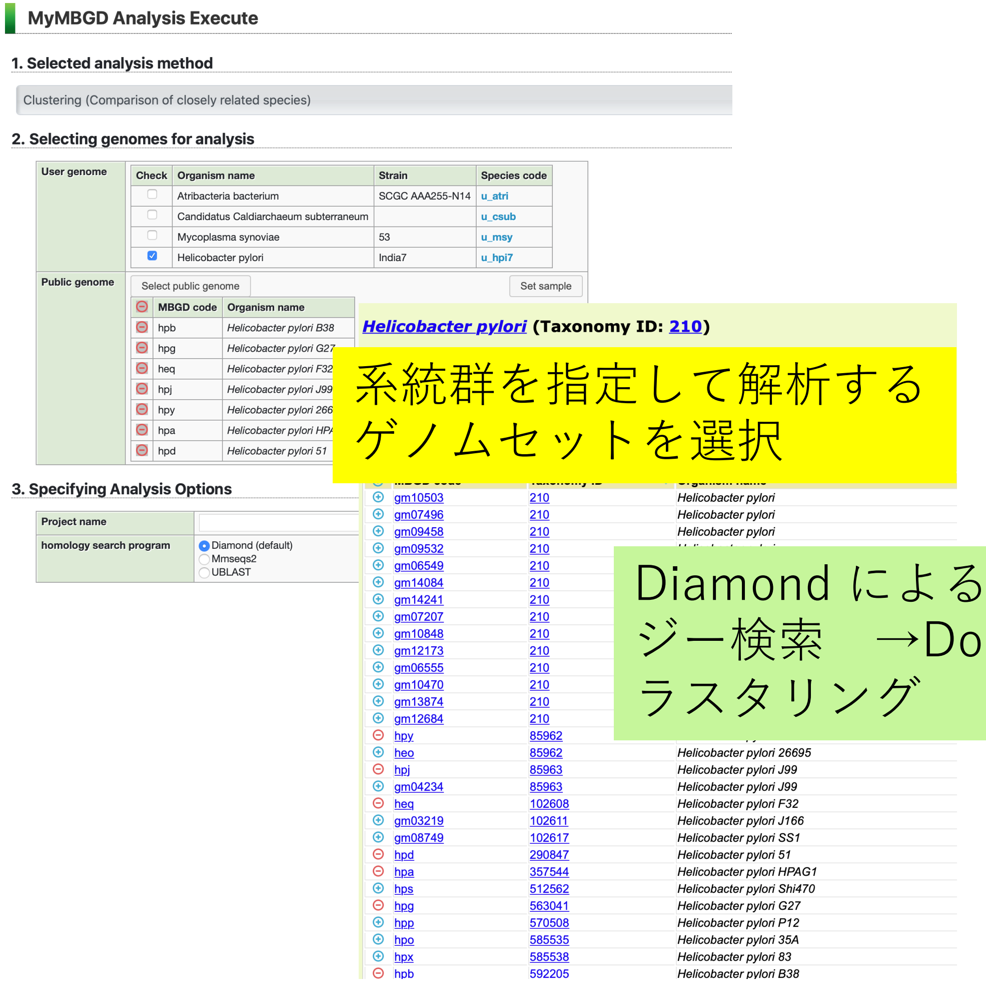
Phylum	Genus	ClusterID	Conservation	gene	
				locus tag	Description
Acidobacteria ON OFF	Candidatus Koribacter			☐ aba:ACID345_RS00575	spermidine synthase
	Terriglobus	tax392733:4996	1.000 (1/1)	☐ gm12266:BLT38_RS11750	spermidine synthase
	Candidatus Solibacter			☐ sus:ACID_RS05870	SAM-dependent methyltransferase
				☐ sus:ACID_RS37145	SAM-dependent methyltransferase
Aquificae ON OFF	Aquifex			☐ aae:AQ_1350	spermidine synthase
Deferribacteres ON OFF	Caldithrix	tax187145:3653	1.000 (2/2)	☐ gm04998:CALAB_RS05300	spermine synthase
Elusimicrobia ON OFF	Endomicrobium	tax1408194:913	1.000 (1/1)	☐ epo:EPRO_RS05910	hypothetical protein
Bacteroidetes ON OFF	Chitinophaga	tax79328:5612	1.000 (2/2)	☐ cpi:CPIN_RS07875	spermidine synthase
	Saprospira			☐ sgn:SGRA_RS17555	spermidine synthase
	Emticicia			☐ eol:EMTOL_RS00680	spermine synthase
	Fibrella	tax861914:4538	1.000 (2/2)	☐ fae:FAES_RS26140	spermidine synthase
		tax861914:5184	1.000 (1/1)	☐ fib:A6C57_RS13095	spermine synthase
	Flexibacter			☐ fli:FLELI_RS04240	spermidine synthase
	Spirosoma	tax107:11333	1.000 (1/1)	☐ gm12188:B5M13_RS17815	spermidine synthase
		tax107:8466	1.000 (2/2)	☐ sli:SLIN_RS27515	spermine synthase
	Flammeovirga	tax59739:4485	1.000 (2/2)	☐ flm:MY04_RS08910	spermidine synthase
	Hymenobacter	tax89966:5622	1.000 (3/3)	☐ hsw:HSW_RS07010	spermidine synthase
	Rufibacter	tax1379908:1928	1.000 (1/1)	☐ ruf:TH63_RS10035	spermidine synthase
	Persicobacter			☐ gm05497:AO081_RS10785	hypothetical protein
				☐ gm05497:AO081_RS11130	spermidine synthase

Ortholog group assigment: *Atribacteria bacterium*

ID	name	Domain	Chrom.	Cluster ID	Cluster Description
u_atri:I832_RS0100005		1	chromosome 1	3396	Spermidine synthase protein
u_atri:I832_RS0100005		1	chromosome 1	3396	Spermidine synthase protein
u_atri:I832_RS0100005		2	chromosome 1	2533	Spermidine synthase protein
u_atri:I832_RS0100005		2	chromosome 1	2533	Spermidine synthase protein
u_atri:I832_RS0100545			chromosome 1	5834	ABC transporter ATP-binding protein
u_atri:I832_RS0100545			chromosome 1	5834	ABC transporter ATP-binding protein
_RS0101220			chromosome 1	2587	Aldehyde:ferredoxin oxidoreductase
_RS0101220			chromosome 1	2587	Aldehyde:ferredoxin oxidoreductase
_RS0101490			chromosome 1	2413	MarR family transcriptional regulator
_RS0101490			chromosome 1	2413	MarR family transcriptional regulator
_RS0102115			chromosome 1	4117	DUF2088 domain-containing protein
u_atri:I					
u_atri:I					
u_atri:I					
u_atri:I					

ユーザゲノムに対する
オーソログ分類結果

MyMBGD mode: クラスタリングモード



MyMBGD Analysis Execute

1. Selected analysis method

Clustering (Comparison of closely related species)

2. Selecting genomes for analysis

User genome

Check	Organism name	Strain	Species code
☐	Atribacteria bacterium	SCGC AAA255-N14	u_atri
☐	Candidatus Caldiarchaeum subterraneum		u_csub
☐	Mycoplasma synoviae	53	u_msy
☑	Helicobacter pylori	India7	u_hpi7

Public genome

	MBGD code	Organism name
⊖	hpb	*Helicobacter pylori B38*
⊖	hpg	*Helicobacter pylori G27*
⊖	heq	*Helicobacter pylori F32*
⊖	hpj	*Helicobacter pylori J99*
⊖	hpy	*Helicobacter pylori 266*
⊖	hpa	*Helicobacter pylori HP*
⊖	hpd	*Helicobacter pylori 51*

3. Specifying Analysis Options

Project name:

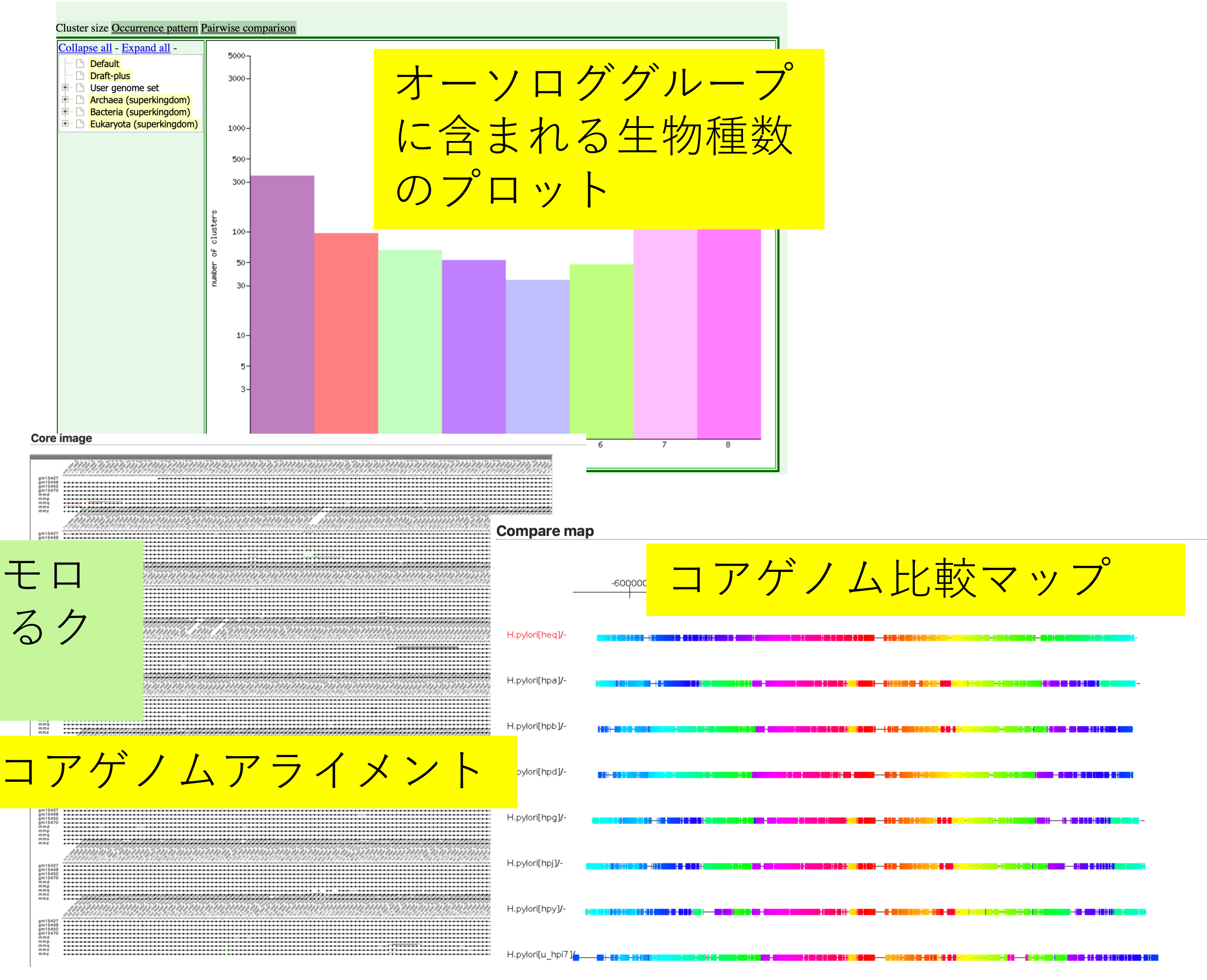
homology search program: ● Diamond (default) ○ Mmseqs2 ○ UBLAST

Helicobacter pylori **(Taxonomy ID: 210)**

	MBGD code	Taxonomy ID	Organism name
⊕	gm10503	210	*Helicobacter pylori*
⊕	gm07496	210	*Helicobacter pylori*
⊕	gm09458	210	*Helicobacter pylori*
⊕	gm09532	210	
⊕	gm06549	210	
⊕	gm14084	210	
⊕	gm14241	210	
⊕	gm07207	210	
⊕	gm10848	210	
⊕	gm12173	210	
⊕	gm06555	210	
⊕	gm10470	210	
⊕	gm13874	210	
⊕	gm12684	210	
⊖	hpy	85962	
⊕	heo	85962	*Helicobacter pylori 26695*
⊖	hpj	85963	*Helicobacter pylori J99*
⊕	gm04234	85963	*Helicobacter pylori J99*
⊖	heq	102608	*Helicobacter pylori F32*
⊕	gm03219	102611	*Helicobacter pylori J166*
⊕	gm08749	102617	*Helicobacter pylori SS1*
⊖	hpd	290847	*Helicobacter pylori 51*
⊖	hpa	357544	*Helicobacter pylori HPAG1*
⊕	hps	512562	*Helicobacter pylori Shi470*
⊖	hpg	563041	*Helicobacter pylori G27*
⊕	hpp	570508	*Helicobacter pylori P12*
⊕	hpo	585535	*Helicobacter pylori 35A*
⊕	hpx	585538	*Helicobacter pylori 83*
⊖	hpb	592205	*Helicobacter pylori B38*

系統群を指定して解析する
ゲノムセットを選択

Diamond による総当たりホモロジー検索　→DomClustによるクラスタリング



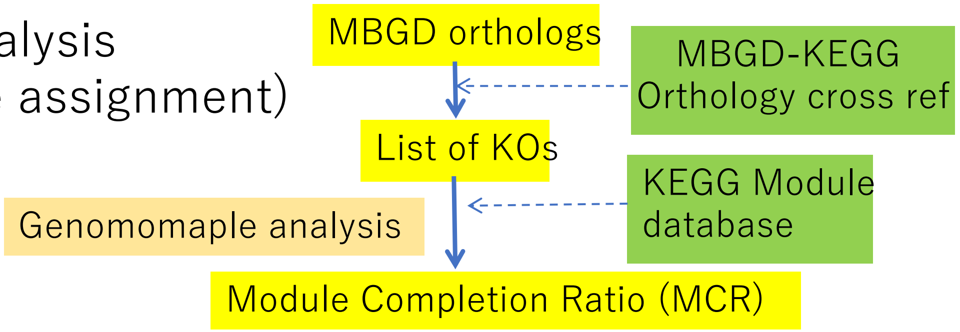
オーソログループ
に含まれる生物種数
のプロット

コアゲノムアライメント

コアゲノム比較マップ

MBGD-Genomemaple (upcoming) :Module単位のアノテーション

Genomaple analysis
(KEGG Module assignment)


MBGD orthologs

MBGD-KEGG Orthology cross ref

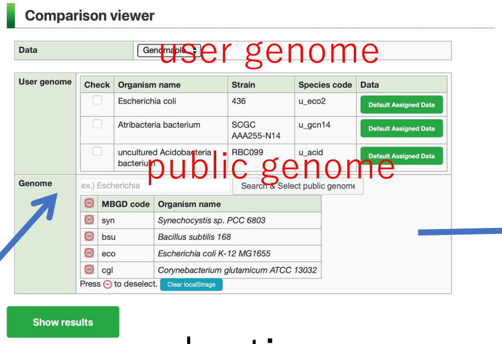
List of KOs

KEGG Module database

Genomomaple analysis

Module Completion Ratio (MCR)

Genome selection



Comparison viewer

Data	Genomaple

User genome

Check	Organism name	Strain	Species code	Data
☐	Escherichia coli	436	u_eco2	Default Assigned Data
☐	Atribacteria bacterium	SCGC AAA255-N14	u_gcn14	Default Assigned Data
☐	uncultured Acidobacteria bacterium	RBC099	u_acid	Default Assigned Data

Genome

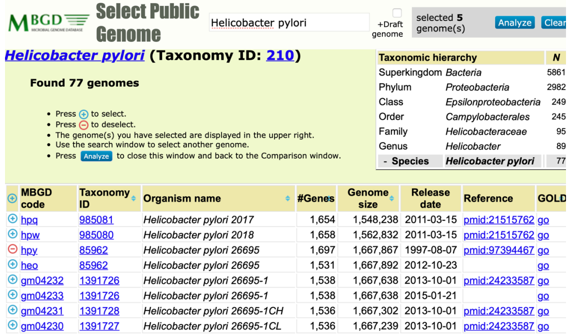
ex.) Escherichia | Search & Select public genome

⊖	MBGD code	Organism name
⊖	syn	*Synechocystis sp. PCC 6803*
⊖	bsu	*Bacillus subtilis 168*
⊖	eco	*Escherichia coli K-12 MG1655*
⊖	cgl	*Corynebacterium glutamicum ATCC 13032*

Press ⊖ to deselect. Clear localStrage

Show results

Public genome selection



Helicobacter pylori

☐ +Draft genome

selected **5** genome(s)

Helicobacter pylori **(Taxonomy ID: 210)**

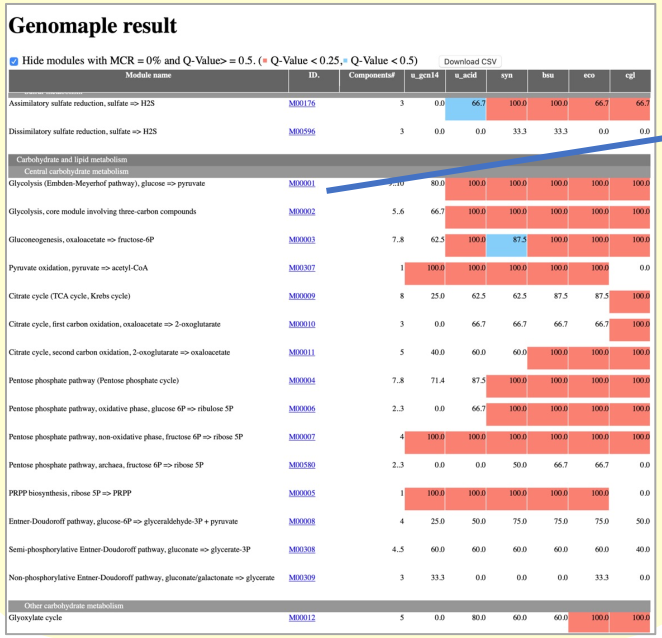
Found 77 genomes

- Press ⊕ to select.
- Press ⊖ to deselect.
- The genome(s) you have selected are displayed in the upper right.
- Use the search window to select another genome.
- Press Analyze to close this window and back to the Comparison window.

Taxonomic hierarchy		*N*
Superkingdom	*Bacteria*	5861
Phylum	*Proteobacteria*	2982
Class	*Epsilonproteobacteria*	249
Order	*Campylobacterales*	245
Family	*Helicobacteraceae*	95
Genus	*Helicobacter*	89
- Species	***Helicobacter pylori***	77

⊕	MBGD code	Taxonomy ID	Organism name	#Genes	Genome size	Release date	Reference	GOLD
⊕	hpq	985081	*Helicobacter pylori 2017*	1,654	1,548,238	2011-03-15	pmid:21515762	go
⊕	hpw	985080	*Helicobacter pylori 2018*	1,658	1,562,832	2011-03-15	pmid:21515762	go
⊖	hpy	85962	*Helicobacter pylori 26695*	1,697	1,667,867	1997-08-07	pmid:97394467	go
⊕	heo	85962	*Helicobacter pylori 26695*	1,531	1,667,892	2012-10-23		go
⊕	gm04232	1391726	*Helicobacter pylori 26695-1*	1,538	1,667,638	2013-10-01	pmid:24233587	go
⊕	gm04233	1391726	*Helicobacter pylori 26695-1*	1,538	1,667,638	2015-01-21		go
⊕	gm04231	1391728	*Helicobacter pylori 26695-1CH*	1,536	1,667,302	2013-10-01	pmid:24233587	go
⊕	gm04230	1391727	*Helicobacter pylori 26695-1CL*	1,536	1,667,239	2013-10-01	pmid:24233587	go

Module comparison viewer



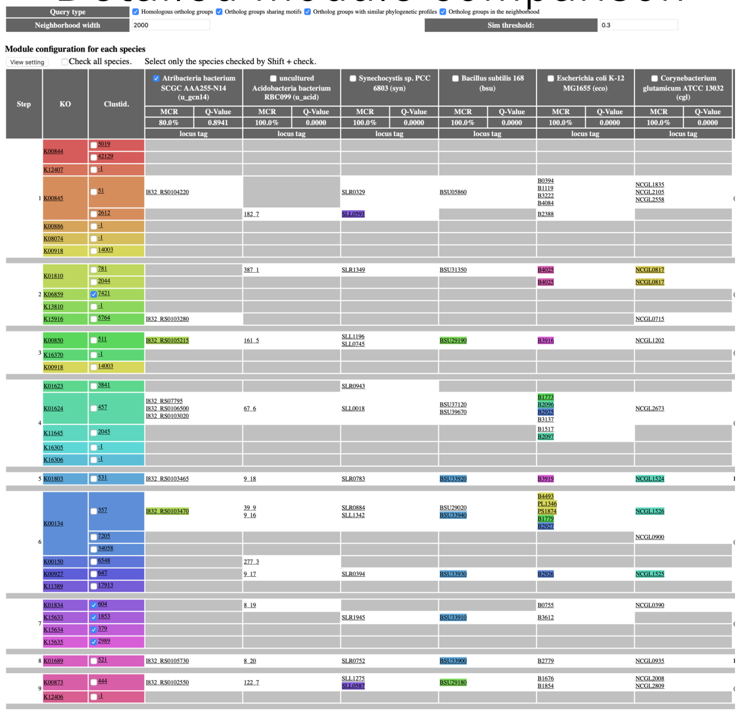
Genomaple result

☑ Hide modules with MCR = 0% and Q-Value> = 0.5. (🟥 Q-Value < 0.25, 🟦 Q-Value < 0.5)

Download CSV

Module name	ID.	Components#	u_gcn14	u_acid	syn	bsu	eco	cgl
Assimilatory sulfate reduction, sulfate => H2S	M00176	3	0.0	66.7	100.0	100.0	66.7	66.7
Dissimilatory sulfate reduction, sulfate => H2S	M00596	3	0.0	0.0	33.3	33.3	0.0	0.0
Carbohydrate and lipid metabolism								
Central carbohydrate metabolism								
Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	M00001	9..10	80.0	100.0	100.0	100.0	100.0	100.0
Glycolysis, core module involving three-carbon compounds	M00002	5..6	66.7	100.0	100.0	100.0	100.0	100.0
Gluconeogenesis, oxaloacetate => fructose-6P	M00003	7..8	62.5	100.0	87.5	100.0	100.0	100.0
Pyruvate oxidation, pyruvate => acetyl-CoA	M00307	1	100.0	100.0	100.0	100.0	100.0	0.0
Citrate cycle (TCA cycle, Krebs cycle)	M00009	8	25.0	62.5	62.5	87.5	87.5	100.0
Citrate cycle, first carbon oxidation, oxaloacetate => 2-oxoglutarate	M00010	3	0.0	66.7	66.7	66.7	66.7	100.0
Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate	M00011	5	40.0	60.0	60.0	100.0	100.0	100.0
Pentose phosphate pathway (Pentose phosphate cycle)	M00004	7..8	71.4	87.5	100.0	100.0	100.0	100.0
Pentose phosphate pathway, oxidative phase, glucose 6P => ribulose 5P	M00006	2..3	0.0	66.7	100.0	100.0	100.0	100.0
Pentose phosphate pathway, non-oxidative phase, fructose 6P => ribose 5P	M00007	4	100.0	100.0	100.0	100.0	100.0	100.0
Pentose phosphate pathway, archaea, fructose 6P => ribose 5P	M00580	2..3	0.0	0.0	50.0	66.7	66.7	0.0
PRPP biosynthesis, ribose 5P => PRPP	M00005	1	100.0	100.0	100.0	100.0	100.0	0.0
Entner-Doudoroff pathway, glucose-6P => glyceraldehyde-3P + pyruvate	M00008	4	25.0	50.0	75.0	75.0	75.0	50.0
Semi-phosphorylative Entner-Doudoroff pathway, gluconate => glycerate-3P	M00308	4..5	60.0	60.0	60.0	60.0	60.0	40.0
Non-phosphorylative Entner-Doudoroff pathway, gluconate/galactonate => glycerate	M00309	3	33.3	0.0	0.0	0.0	33.3	0.0
Other carbohydrate metabolism								
Glyoxylate cycle	M00012	5	0.0	80.0	60.0	60.0	100.0	100.0

Detailed module comparison



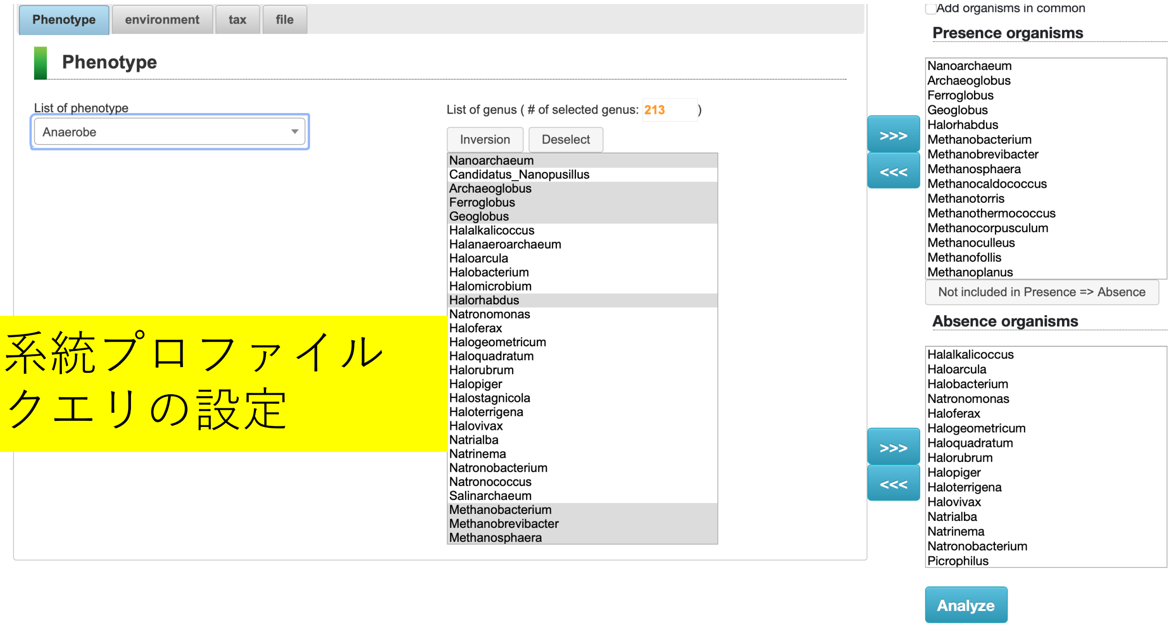
Query type	☑ Homologous ortholog groups ☑ Ortholog groups sharing motifs ☑ Ortholog groups with similar phylogenetic profiles ☑ Ortholog groups in the neighborhood		
Neighborhood width	2000	Sim threshold:	0.3

Module configuration for each species

View setting ☐ Check all species. Select only the species checked by Shift + check.

Step	KO	Clustid.	☑ Atribacteria bacterium SCGC AAA255-N14 (u_gcn14)		☐ uncultured Acidobacteria bacterium RBC099 (u_acid)		☐ Synechocystis sp. PCC 6803 (syn)		☐ Bacillus subtilis 168 (bsu)		☐ Escherichia coli K-12 MG1655 (eco)		☐ Corynebacterium glutamicum ATCC 13032 (cgl)	
			MCR	Q-Value	MCR	Q-Value	MCR	Q-Value	MCR	Q-Value	MCR	Q-Value	MCR	Q-Value
			80.0%	0.8941	100.0%	0.0000	100.0%	0.0000	100.0%	0.0000	100.0%	0.0000	100.0%	0.0000
			locus tag		locus tag		locus tag		locus tag		locus tag		locus tag	

系統プロファイル解析



Phenotype

List of phenotype: Anaerobe

List of genus (# of selected genus: 213)

Nanoarchaeum
Candidatus_Nanopusillus
Archaeoglobus
Ferroglobus
Geoglobus
Halalkalicoccus
Halanaeroarchaeum
Haloarcula
Halobacterium
Halomicrobium
Halorhabdus
Natronomonas
Haloferax
Halogeometricum
Haloquadratum
Halorubrum
Halopiger
Halostagnicola
Haloterrigena
Halovivax
Natrialba
Natrinema
Natronobacterium
Natronococcus
Salinarchaeum
Methanobacterium
Methanobrevibacter
Methanosphaera

☐ Add organisms in common

Presence organisms

Nanoarchaeum
Archaeoglobus
Ferroglobus
Geoglobus
Halorhabdus
Methanobacterium
Methanobrevibacter
Methanosphaera
Methanocaldococcus
Methanotorris
Methanothermococcus
Methanocorpusculum
Methanoculleus
Methanofollis
Methanoplanus

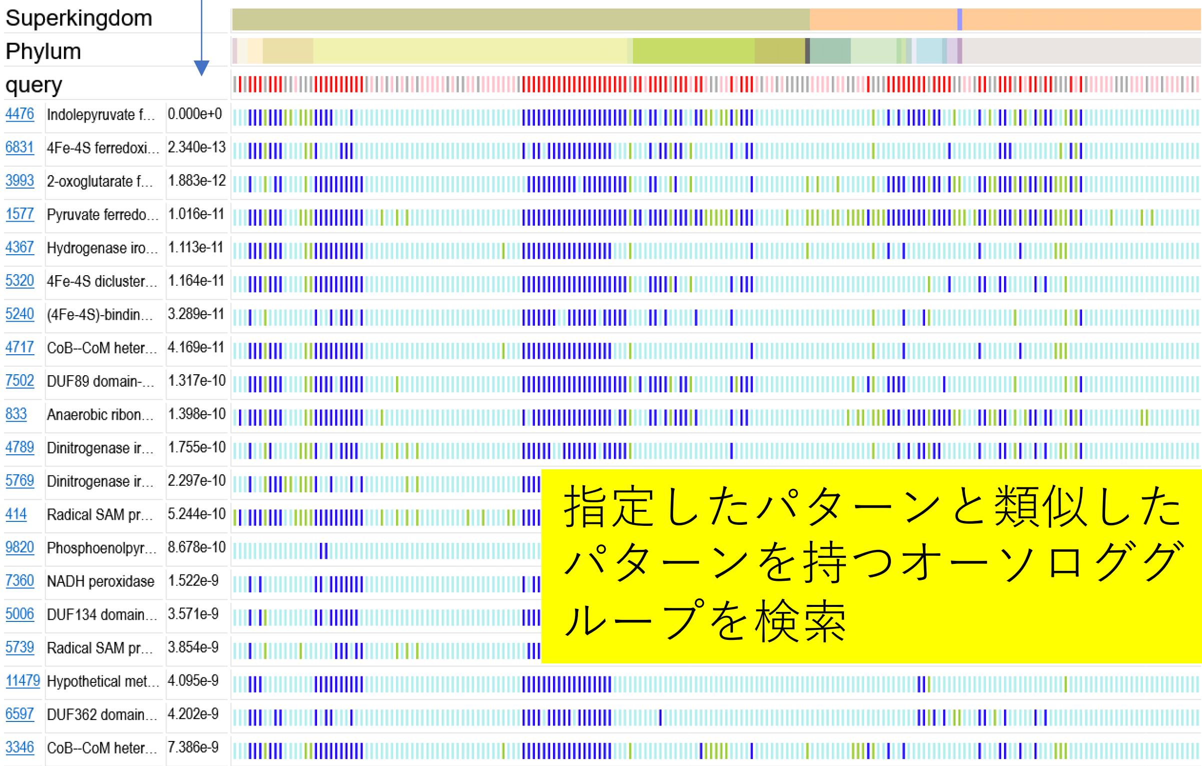
Absence organisms

Halalkalicoccus
Haloarcula
Halobacterium
Natronomonas
Haloferax
Halogeometricum
Haloquadratum
Halorubrum
Halopiger
Haloterrigena
Halovivax
Natrialba
Natrinema
Natronobacterium
Picrophilus

系統プロファイル
クエリの設定

presence/absenceの条件とする生物種セットをPhenotype/Environment/Taxonomyを用いて指定できる

Dollo-parsimony compress profiles(DPCP)法(Ruano-Rubio et al 2009)による、生物系統の偏りを排除した統計評価



4476	Indolepyruvate f...	0.000e+0
6831	4Fe-4S ferredoxi...	2.340e-13
3993	2-oxoglutarate f...	1.883e-12
1577	Pyruvate ferredo...	1.016e-11
4367	Hydrogenase iro...	1.113e-11
5320	4Fe-4S dicluster...	1.164e-11
5240	(4Fe-4S)-bindin...	3.289e-11
4717	CoB--CoM heter...	4.169e-11
7502	DUF89 domain-...	1.317e-10
833	Anaerobic ribon...	1.398e-10
4789	Dinitrogenase ir...	1.755e-10
5769	Dinitrogenase ir...	2.297e-10
414	Radical SAM pr...	5.244e-10
9820	Phosphoenolpyr...	8.678e-10
7360	NADH peroxidase	1.522e-9
5006	DUF134 domain...	3.571e-9
5739	Radical SAM pr...	3.854e-9
11479	Hypothetical met...	4.095e-9
6597	DUF362 domain...	4.202e-9
3346	CoB--CoM heter...	7.386e-9

指定したパターンと類似したパターンを持つオーソロググループを検索

まとめと今後の展開

- MBGD2022は前バージョン(MBGD2018)と比べてゲノム数2.4倍、genus数としては1.4倍に増加した。オーソロググループ数もgenus数と同程度に単調に増加している。
- このデータをユーザゲノムの解析に利用できるよう、MyMBGDインターフェイスを改良し、使い勝手を向上させた。
- アサインメントモードではGenomaple, クラスタリングモードではCoreAligner を用いた解析機能を提供。
- 系統プロファイル検索の統計評価法を改良し、精度の高い検索を可能にした。
- 今後、Metagenome Assembled Genomeのアノテーションに用いることを想定して、さらなる改良を行う。