15

MBGDのオーソログ推定機能を用いた新規ゲノム配列の機能推定

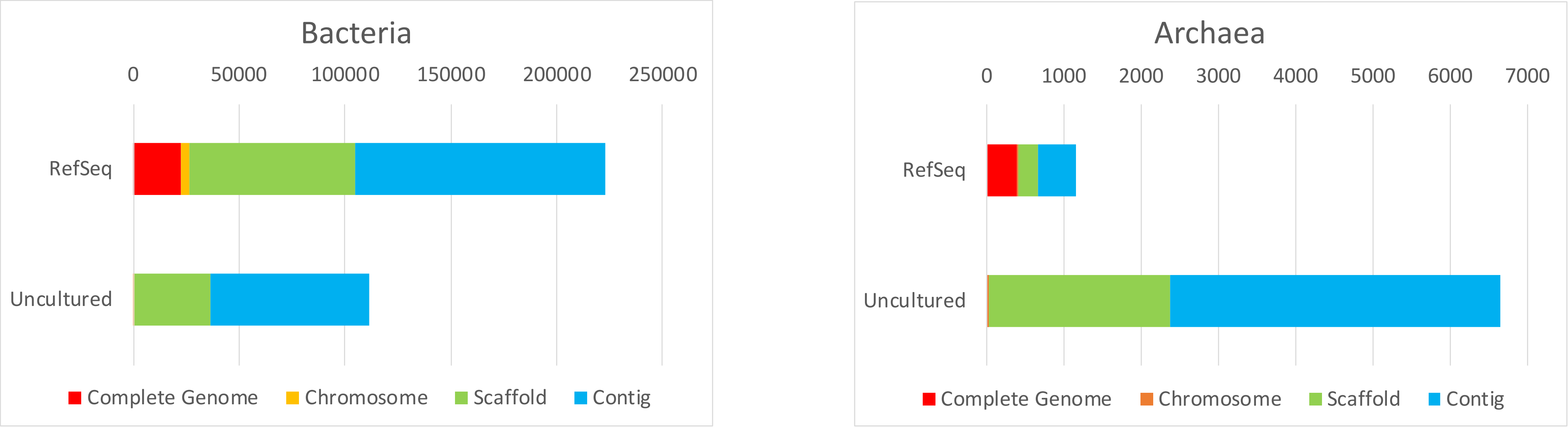
〇内山郁夫1)、三原基広2)、西出浩世1)、千葉啓和3)、高柳正彦4)、髙見英人5)

1) 基礎生物学研究所、2) 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) ウェブブレイン、5)東大大気海洋研究所

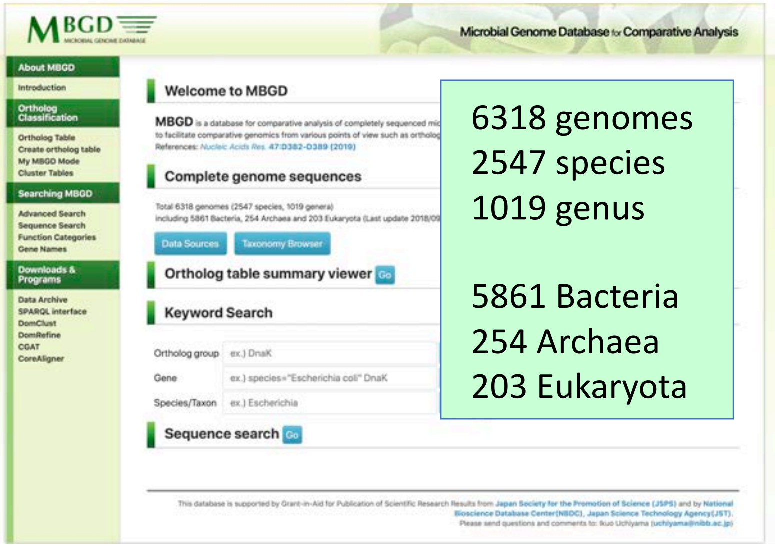
モチベーション：未培養菌ゲノムデータの活用

- メタゲノムからゲノム再構築などの手法で、未培養菌のゲノム解読が進み、環境中の未知微生物の生態をゲノム配列から推定することが重要になりつつある。
- MBGDのオーソログデータを用いて、新規ゲノムのアノテーションおよびゲノム機能推定に活用する。

The number of genomes released from NCBI (from Assembly Reports)



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

- About MBGD
  - Introduction
- Ortholog Classification
  - Ortholog Table
  - Create ortholog table
  - My MBGD Mode
  - Cluster Tables
- Searching MBGD
  - Advanced Search
  - Sequence Search
  - Function Categories
  - Gene Names
- Downloads & Programs
  - Data Archive
  - SPARQL interface
  - DomClust
  - DomRefine
  - CGAT
  - CoreAligner

Welcome to MBGD

MBGD is a database for comparative analysis of completely sequenced mic… to facilitate comparative genomics from various points of view such as ortholog… References: Nucleic Acids Res. 47:D382-D389 (2019)

Complete genome sequences

Total 6318 genomes (2547 species, 1019 genera) including 5861 Bacteria, 254 Archaea and 203 Eukaryota (Last update 2018/09…

Data Sources | Taxonomy Browser

Ortholog table summary viewer Go

Keyword Search

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

Sequence search Go

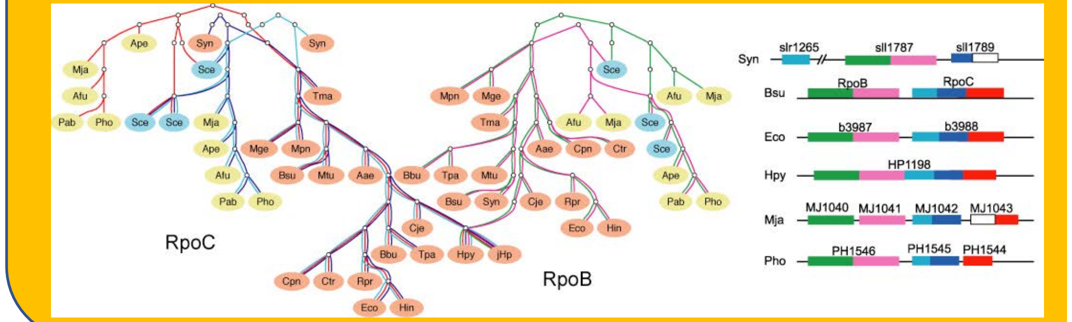
This database is supported by Grant-in-Aid for Publication of Scientific Research Results from Japan Society for the Promotion of Science (JSPS) and by National Bioscience Database Center(NBDC), Japan Science Technology Agency(JST). Please send questions and comments to: Ikuo Uchiyama (uchiyama@nibb.ac.jp)

6318 genomes
2547 species
1019 genus

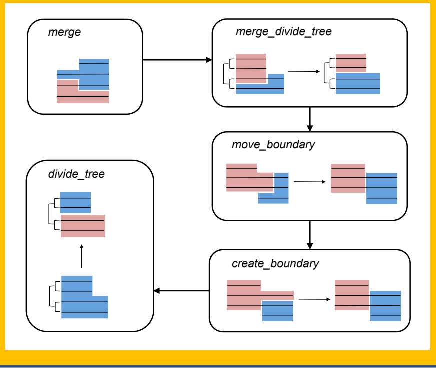
5861 Bacteria
254 Archaea
203 Eukaryota

DomClust

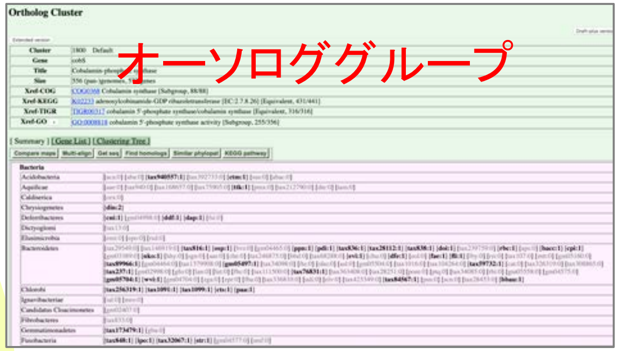
ドメイン単位のオーソロググルーピング



DomRefine



オーソロググループ



Ortholog Cluster

Extended version

Draft-plus version

Cluster	1800 Default
Gene	cobS
Title	Cobalamin-phosphate synthase
Size	556 (pan-genomes, … genes
Xref-COG	COG0368 Cobalamin synthase [Subgroup, 88/88]
Xref-KEGG	K02233 adenosylcobinamide-GDP ribazoletransferase [EC:2.7.8.26] [Equivalent, 431/441]
Xref-TIGR	TIGR00317 cobalamin 5'-phosphate synthase/cobalamin synthase [Equivalent, 316/316]
Xref-GO	GO:0008818 cobalamin 5'-phosphate synthase activity [Subgroup, 255/356]

[Summary] [Gene List] [Clustering Tree]

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

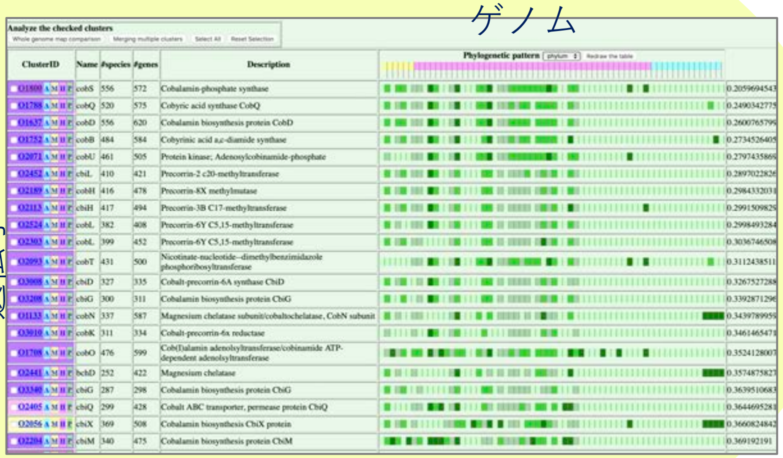
Bacteria

Acidobacteria	[illegible]
Aquificae	[illegible]
Caldiserica	[illegible]
Chrysiogenetes	[din:2]
Deferribacteres	[cni:1] [illegible] [ddf:1] [dap:1] [illegible]
Dictyoglomi	[illegible]
Elusimicrobia	[illegible]
Bacteroidetes	[illegible]
Chlorobi	[tax256319:1] [tax1091:1] [tax1099:1] [cts:1] [paa:1]
Ignavibacteriae	[illegible]
Candidatus Cloacimonetes	[illegible]
Fibrobacteres	[illegible]
Gemmatimonadetes	[tax173479:1] [illegible]
Fusobacteria	[tax848:1] [ipo:1] [tax32067:1] [str:1] [illegible]

オーソログテーブル

ゲノム

遺伝子



Analyze the checked clusters

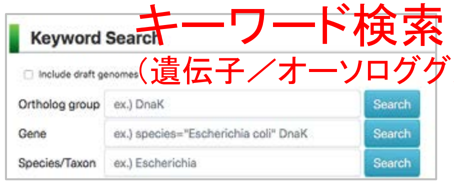
Whole genome map comparison | Merging multiple clusters | Select All | Reset Selection

Phylogenetic pattern phylum Redraw the table

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

キーワード検索

（遺伝子／オーソロググループ／生物種）

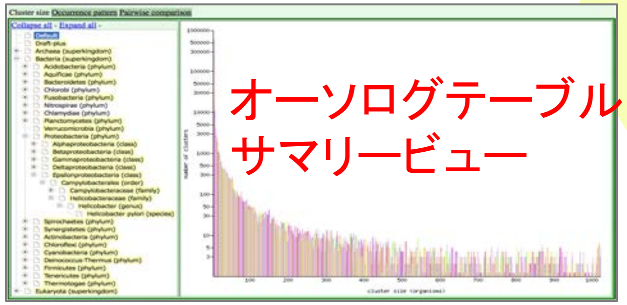


Keyword Search

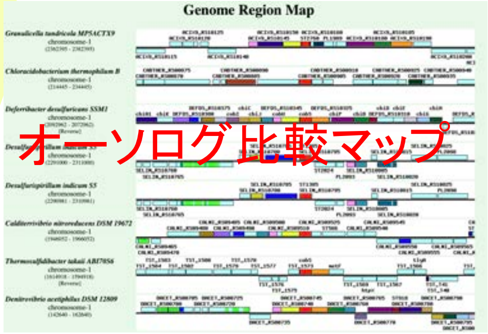
□ Include draft genomes

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

オーソログテーブル サマリービュー

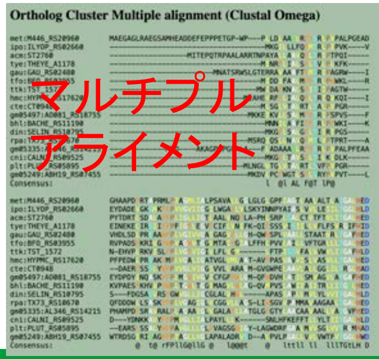


オーソログ比較マップ

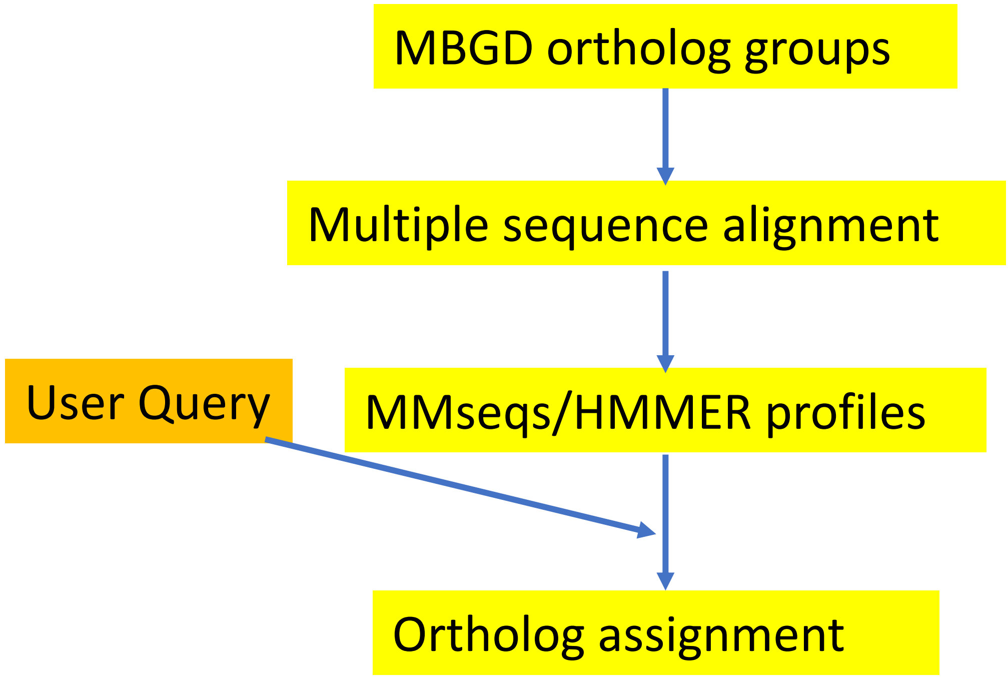


Genome Region Map
Granulicella tundricola MP5ACTX9
Chloracidobacterium thermophilum B
Deferribacter desulfuricans SSM1
Desulfurispirillum indicum S5
Caldisericum exile
Thermosulfidibacter takaii ABI70S6
Denitrovibrio acetiphilus DSM 12809
chromosome-1

マルチプルアライメント

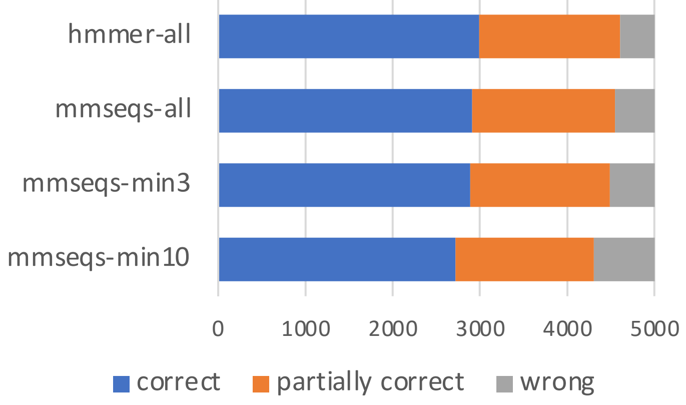


プロファイル検索によるオーソログ推定



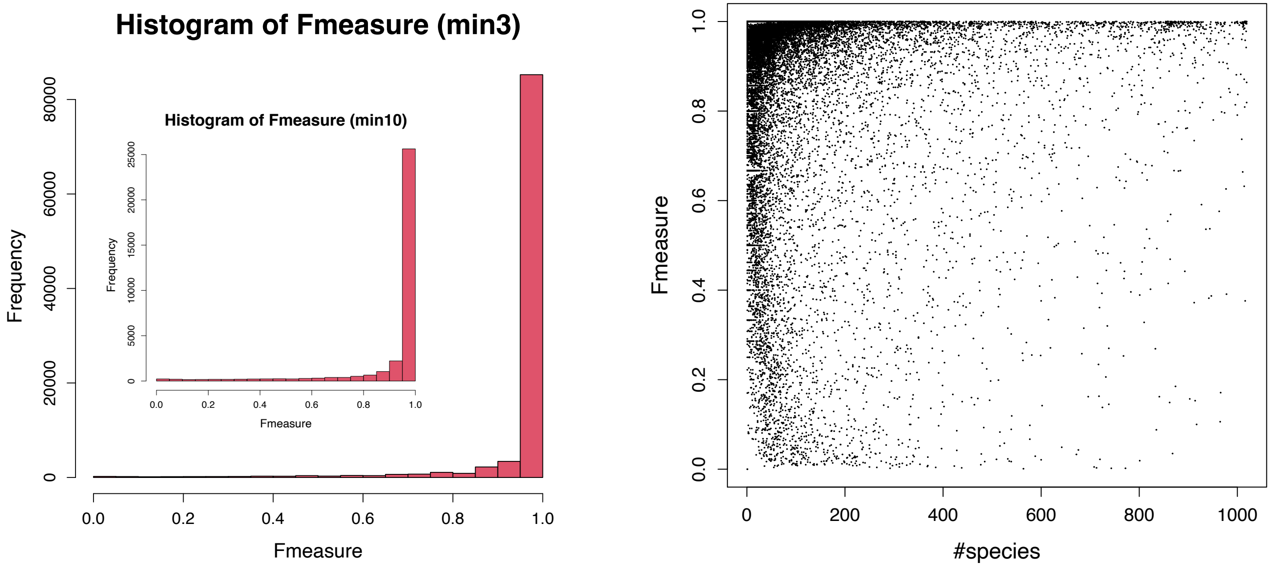
	# clusters
all	755852
min3 (members ≥ 3)	97661
min10 (members ≥10)	38446

Ortholog group assignment test (5000 queries)

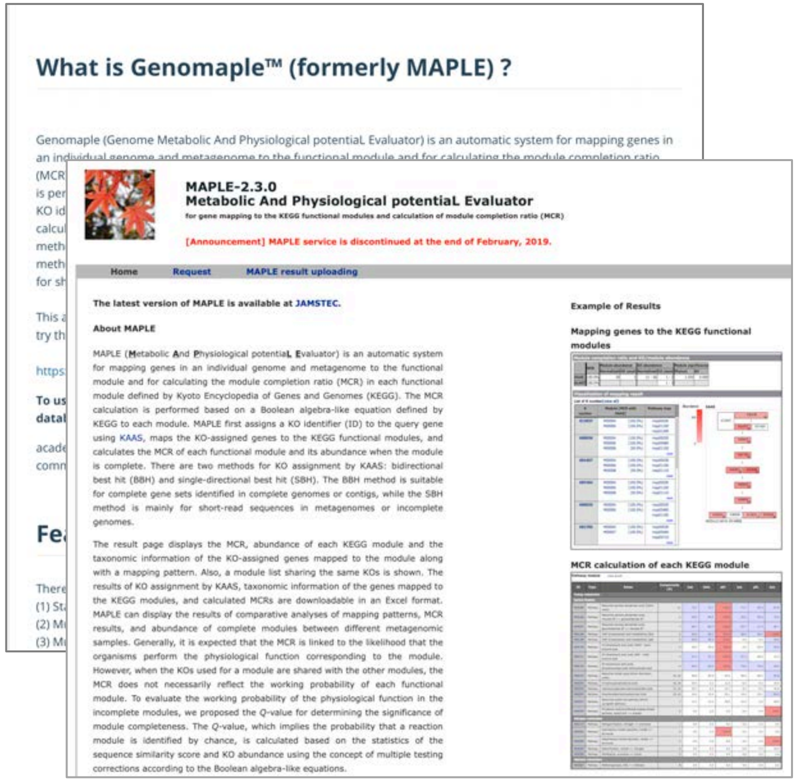


	SearchTime(s)
hmmer-all	20298.6
mmseqs-all	4632.0
mmseqs-min3	1758.0
mmseqs-min10	1105.6

Result of profile quality test



Genomapleを用いた機能モジュール充足判定



What is Genomaple™ (formerly MAPLE) ?

Genomaple (Genome Metabolic And Physiological potentiaL Evaluator) is an automatic system for mapping genes in an individual genome and metagenome to the functional module and for calculating the module completion ratio (MCR

is per

KO id

calcu

meth

meth

for sh

This a

try th

https

To us

datal

acade

comn

Fe

There

(1) St

(2) M

(3) M

MAPLE-2.3.0
Metabolic And Physiological potentiaL Evaluator
for gene mapping to the KEGG functional modules and calculation of module completion ratio (MCR)

[Announcement] MAPLE service is discontinued at the end of February, 2019.

The latest version of MAPLE is available at JAMSTEC.

About MAPLE

MAPLE (**M**etabolic **A**nd **P**hysiological potentia**L** **E**valuator) is an automatic system for mapping genes in an individual genome and metagenome to the functional module and for calculating the module completion ratio (MCR) in each functional module defined by Kyoto Encyclopedia of Genes and Genomes (KEGG). The MCR calculation is performed based on a Boolean algebra-like equation defined by KEGG to each module. MAPLE first assigns a KO identifier (ID) to the query gene using KAAS, maps the KO-assigned genes to the KEGG functional modules, and calculates the MCR of each functional module and its abundance when the module is complete. There are two methods for KO assignment by KAAS: bidirectional best hit (BBH) and single-directional best hit (SBH). The BBH method is suitable for complete gene sets identified in complete genomes or contigs, while the SBH method is mainly for short-read sequences in metagenomes or incomplete genomes.

The result page displays the MCR, abundance of each KEGG module and the taxonomic information of the KO-assigned genes mapped to the module along with a mapping pattern. Also, a module list sharing the same KOs is shown. The results of KO assignment by KAAS, taxonomic information of the genes mapped to the KEGG modules, and calculated MCRs are downloadable in an Excel format. MAPLE can display the results of comparative analyses of mapping patterns, MCR results, and abundance of complete modules between different metagenomic samples. Generally, it is expected that the MCR is linked to the likelihood that the organisms perform the physiological function corresponding to the module. However, when the KOs used for a module are shared with the other modules, the MCR does not necessarily reflect the working probability of each functional module. To evaluate the working probability of the physiological function in the incomplete modules, we proposed the *Q*-value for determining the significance of module completeness. The *Q*-value, which implies the probability that a reaction module is identified by chance, is calculated based on the statistics of the sequence similarity score and KO abundance using the concept of multiple testing corrections according to the Boolean algebra-like equations.

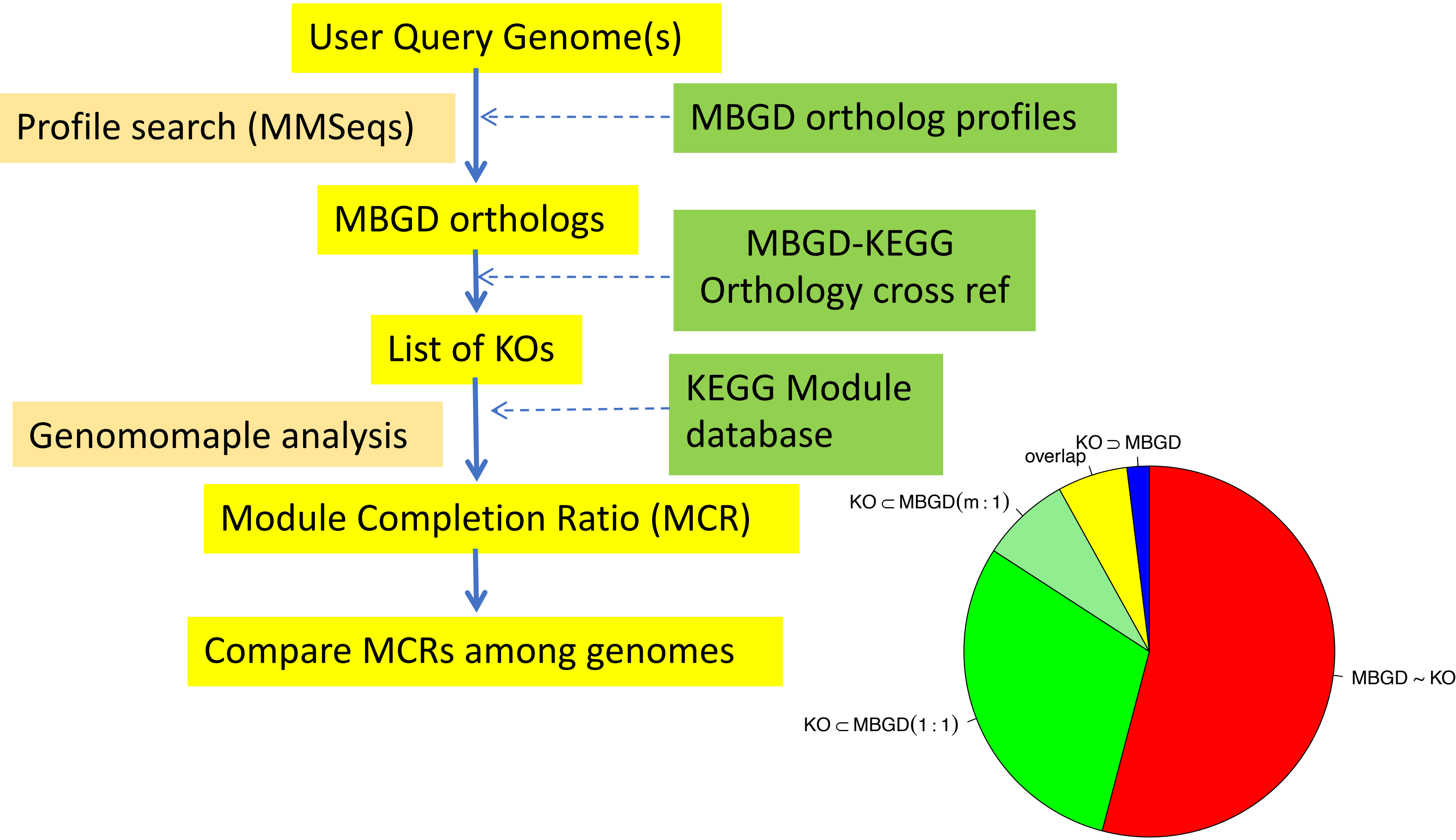
Example of Results

Mapping genes to the KEGG functional modules

MCR calculation of each KEGG module

MAPLE (Takami et al. 2016):
ゲノム中に同定されたKOに基づいて KEGG Module の充足率を計算するツール

MBGD-Genomaple pipeline



User Query Genome(s)

Profile search (MMSeqs) ← MBGD ortholog profiles

MBGD orthologs

MBGD-KEGG Orthology cross ref

List of KOs

KEGG Module database

Genomomaple analysis

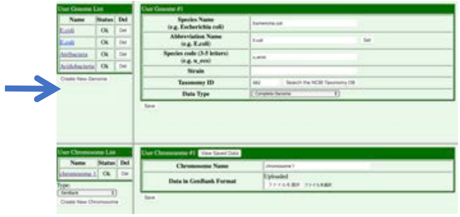
Module Completion Ratio (MCR)

Compare MCRs among genomes

MBGD-Genomaple解析のための新規MyMBGDインターフェイス

MyMBGD: upload user genomes

User Genomes


User Genome List

Name	Status	Del
E.coli	Ok	Del
E.coli	Ok	Del
Atribacteria	Ok	Del
Acidobacteria	Ok	Del

Create New Genome

User Genome #1

Species Name (e.g. Escherichia coli)	Escherichia coli
Abbreviation Name (e.g. E.coli)	E.coli
Species code (3-5 letters) (e.g. u_eco)	u_eco
Strain	
Taxonomy ID	562
Data Type	Complete Genome

Save

User Chromosome List

Create New Chromosome

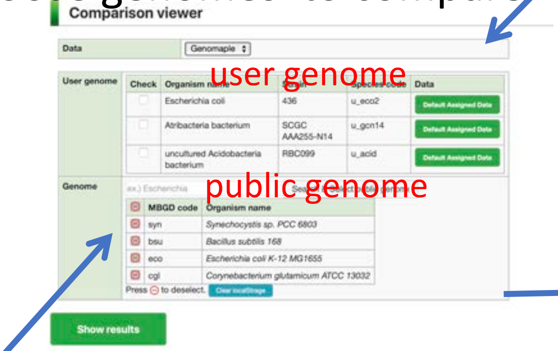
User Chromosome #1

Chromosome Name	chromosome 1
Data in GenBank Format	Uploaded

Save

MBGD-Genomaple pipeline

Choose genomes to compare


Comparison viewer

Data: Genomaple

user genome

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

public genome

Genome

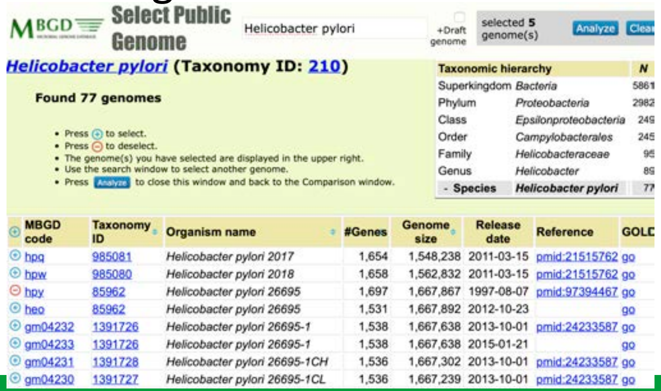
ex.) Escherichia — Search & Select public genome

⊖	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 localStorage

Show results

Public genome selection


Select Public Genome — Helicobacter pylori — □ +Draft genome — selected 5 genome(s) — Analyze — Clear

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

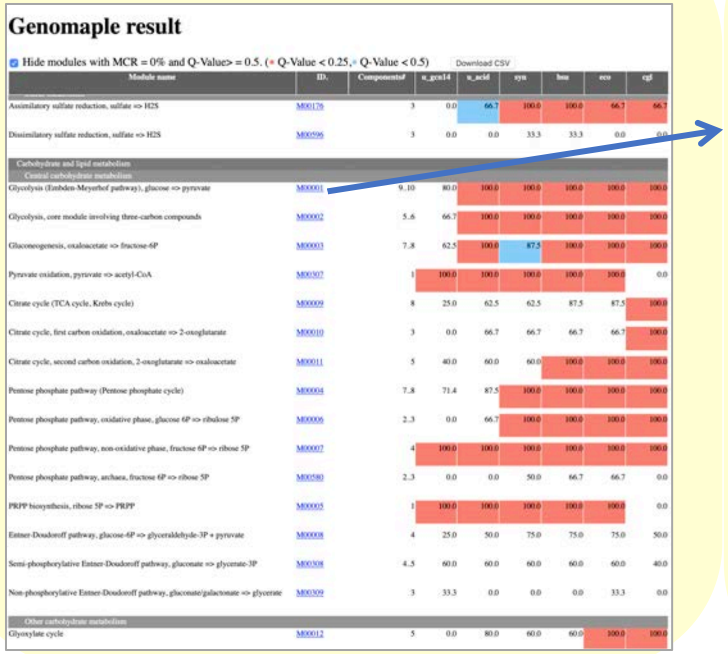
Found 77 genomes

- 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

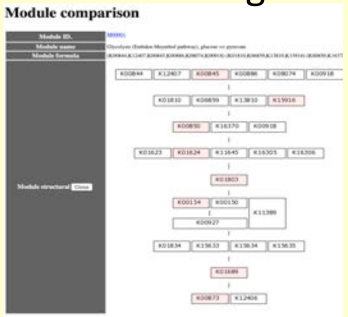
Module comparison viewer


Genomaple result

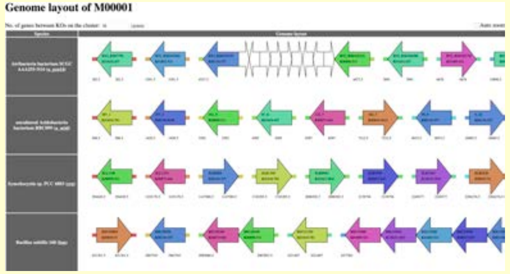
☑ Hide modules with MCR = 0% and Q-Value> = 0.5. (• Q-Value < 0.25, • Q-Value < 0.5) Download CSV

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

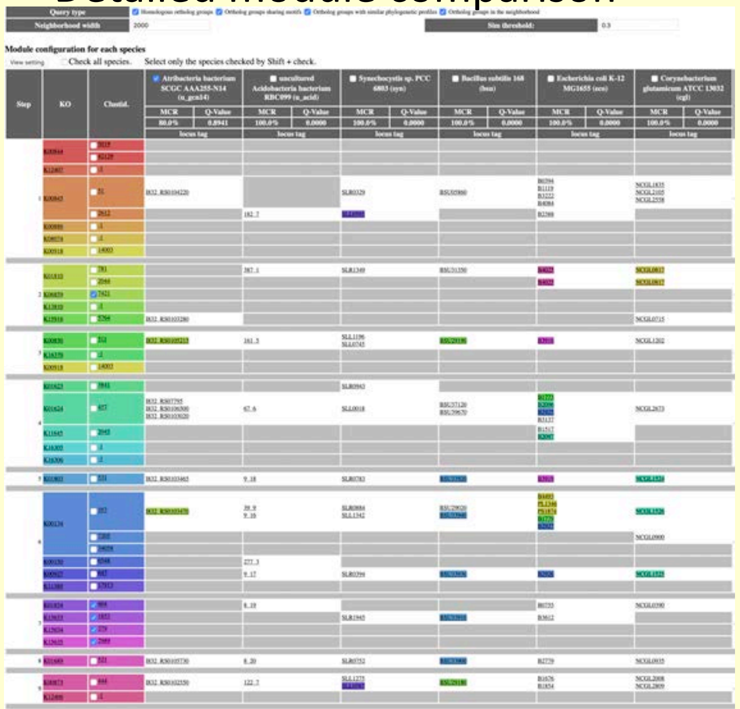
Module diagram


Module comparison

Gene arrangement comparison viewer

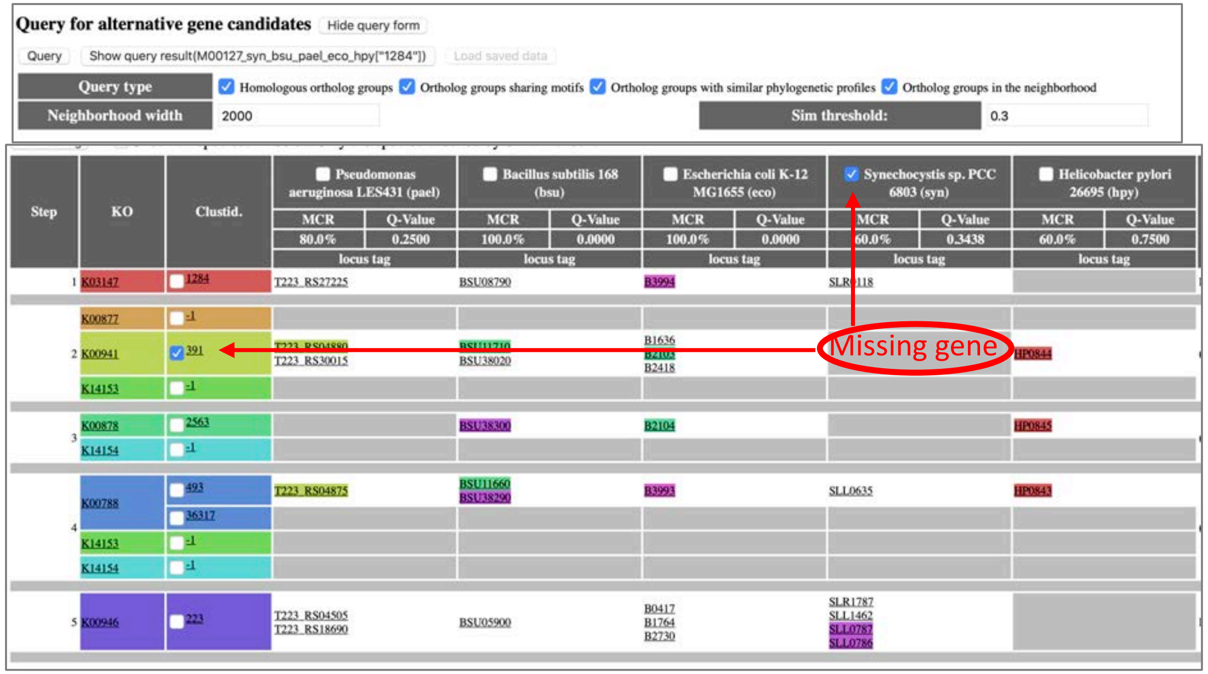

Genome layout of M00001

Detailed module comparison


Module configuration for each species

欠損遺伝子に対する代替候補遺伝子の検索

Query for searching alternative genes



Query for alternative gene candidates Hide query form

Query | Show query result(M00127_syn_bsu_pael_eco_hpy["1284"]) | Load saved data

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

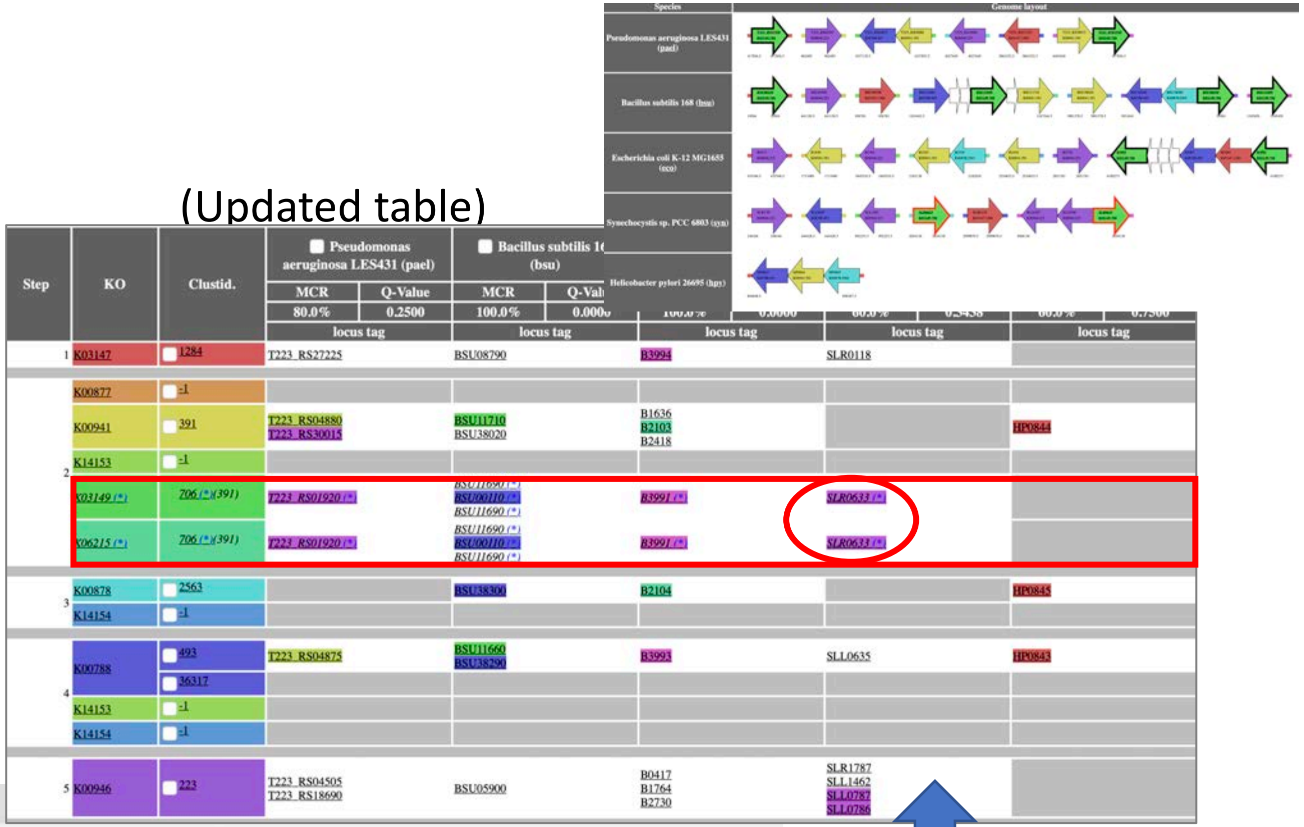
Step	KO	Clustid.	☐ Pseudomonas aeruginosa LES431 (pael)	☐ Bacillus subtilis 168 (bsu)	☐ Escherichia coli K-12 MG1655 (eco)	☑ Synechocystis sp. PCC 6803 (syn)	☐ Helicobacter pylori 26695 (hpy)
			MCR 80.0% Q-Value 0.2500	MCR 100.0% Q-Value 0.0000	MCR 100.0% Q-Value 0.0000	MCR 60.0% Q-Value 0.3438	MCR 60.0% Q-Value 0.7500
			locus tag	locus tag	locus tag	locus tag	locus tag
1	K03147	☐ 1284	T223_RS27225	BSU08790	B3994	SLR0118	
2	K00877	☐ -1					
	K00941	☑ 391	T223_RS04880 T223_RS30015	BSU11710 BSU38020	B1636 B2103 B2418		HP0844
	K14153	☐ -1					
3	K00878	☐ 2563		BSU38300	B2104		HP0845
	K14154	☐ -1					
4	K00788	☐ 493	T223_RS04875	BSU11660 BSU38290	B3993	SLL0635	HP0843
		☐ 36317					
	K14153	☐ -1					
	K14154	☐ -1					
5	K00946	☐ 223	T223_RS04505 T223_RS18690	BSU05900	B0417 B1764 B2730	SLR1787 SLL1462 SLL0787 SLL0786	

Missing gene

Searching candidate genes

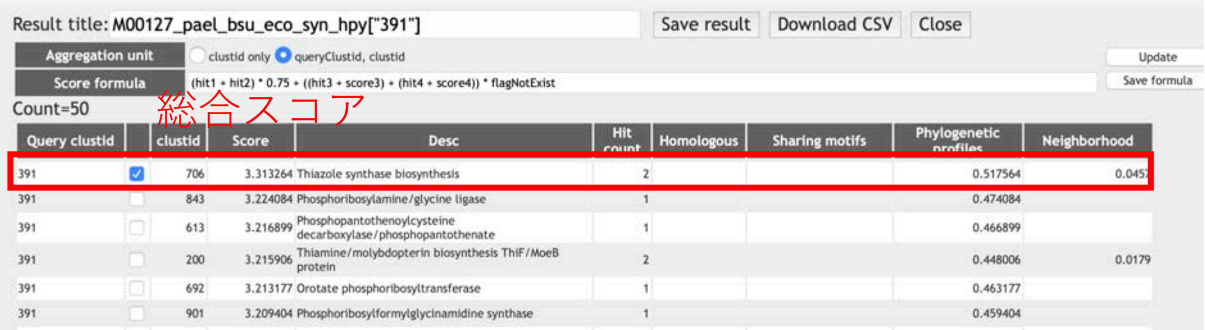
- Homologous cluster
- Sharing motifs
- Phylogenetic profiles
- Neighborhood orthologs

(Updated table)


Species
Genome layout
Pseudomonas aeruginosa LES431 (pael)
Bacillus subtilis 168 (bsu)
Escherichia coli K-12 MG1655 (eco)
Synechocystis sp. PCC 6803 (syn)
Helicobacter pylori 26695 (hpy)

Step	KO	Clustid.	☐ Pseudomonas aeruginosa LES431 (pael)	☐ Bacillus subtilis 16 (bsu)			
			MCR 80.0% Q-Value 0.2500	MCR 100.0% Q-Val 0.000	100.0% 0.0000	60.0% 0.3438	60.0% 0.7500
			locus tag	locus tag	locus tag	locus tag	locus tag
1	K03147	☐ 1284	T223_RS27225	BSU08790	B3994	SLR0118	
2	K00877	☐ -1					
	K00941	☐ 391	T223_RS04880 T223_RS30015	BSU11710 BSU38020	B1636 B2103 B2418		HP0844
	K14153	☐ -1					
	K03149 (*)	706 (*)(391)	T223_RS01920 (*)	BSU11690 (*) BSU00110 (*) BSU11690 (*)	B3991 (*)	SLR0633 (*)	
	K06215 (*)	706 (*)(391)	T223_RS01920 (*)	BSU11690 (*) BSU00110 (*) BSU11690 (*)	B3991 (*)	SLR0633 (*)	
3	K00878	☐ 2563		BSU38300	B2104		HP0845
	K14154	☐ -1					
4	K00788	☐ 493	T223_RS04875	BSU11660 BSU38290	B3993	SLL0635	HP0843
		☐ 36317					
	K14153	☐ -1					
	K14154	☐ -1					
5	K00946	☐ 223	T223_RS04505 T223_RS18690	BSU05900	B0417 B1764 B2730	SLR1787 SLL1462 SLL0787 SLL0786	

Query result



Result title: M00127_pael_bsu_eco_syn_hpy["391"]

Aggregation unit: ○ clustid only ● queryClustid, clustid

Score formula: (hit1 + hit2) * 0.75 + ((hit3 + score3) + (hit4 + score4)) * flagNotExist

Count=50

総合スコア

Query clustid		clustid	Score	Desc	Hit count	Homologous	Sharing motifs	Phylogenetic profiles	Neighborhood
391	☑	706	3.313264	Thiazole synthase biosynthesis	2			0.517564	0.0457
391	☐	843	3.224084	Phosphoribosylamine/glycine ligase	1			0.474084	
391	☐	613	3.216899	Phosphopantothenoylcysteine decarboxylase/phosphopantothenate	1			0.466899	
391	☐	200	3.215906	Thiamine/molybdopterin biosynthesis ThiF/MoeB protein	2			0.448006	0.0179
391	☐	692	3.213177	Orotate phosphoribosyltransferase	1			0.463177	
391	☐	901	3.209404	Phosphoribosylformylglycinamidine synthase	1			0.459404	

Adding selected orthologs