

September 27, 2016

**MGM-23 Workshop Presentation:** Comparative Genomics for Gene Discovery Using Integrated Microbial Genomes (IMG)

**Presenter:** Rekha Seshadri, Joint Genome Institute ([rshadri@lbl.gov](mailto:rshadri@lbl.gov))

**URL:** <https://img.jgi.doe.gov/cgi-bin/mer/main.cgi>

JGI  **IMG/ER** INTEGRATED MICROBIAL GENOMES / EXPERT REVIEW

Quick Genome Search:  Go

Hi Rekha Seshadri | Logout (JGI SSO) 19732

My Analysis Carts: 0 Genomes | 0 Scaffolds | 0 Functions | 0 Data

Home Find Genomes Find Genes Find Functions Compare Genomes OMICS Workspace My IMG Data

**IMG/ER Content**

Datasets

Bacteria	38900
Archaea	1023
Eukarya	257
Plasmids	1221
Viruses	5178
Genome Fragments	1199
Metagenome	8632
Total Datasets	56410
My Private Datasets	10638

Last Datasets Added On:

Genome	2016-01-30
Metagenome	2016-01-31

[Project Map](#) [Metagenome Projects Map](#) [System Requirements](#)

 Hands on training available at the

[Microbial Genomics & Metagenomics Workshop](#)

**Gene Search**

Cassette Search

BLAST

Phylogenetic Profilers

Single Genes

Gene Cassettes

IMG/ER Statistics

The **Microbial Genomes (IMG)** system serves as a community resource for analysis of genome and metagenome datasets in a comprehensive manner. The **IMG data warehouse** integrates genome and metagenome data from various sources and provides tools for analyzing their private (password-protected) datasets ([JGI JLM](#)) or public datasets ([JGI JLM](#)) in the context of all public (free) datasets ([JGI JLM](#)). The system also includes a **Gene Search** feature for finding specific genes across different datasets.

IMG/ER contains 238 public studies, 4766 public metagenome datasets (4403 unique samples) distributed as follows:

Engineered	534	Environmental	2915	Host-associated	1317
Bioreactor	16	Air	31	Algae	29
Bioremediation	47	Aquatic	1790	Animal	1
Biotransformation	28	Terrestrial	1092	Annelida	52
Food production	3	Unclassified	2	Arthropoda	81
Lab enrichment	92			Birds	10
Solid waste	29			Cnidaria	2
Unclassified	22			Human	876
Wastewater	297			Mammals	30
				Microbial	12
				Mollusca	10
				Plants	197
				Porifera	9
				Tunicates	8

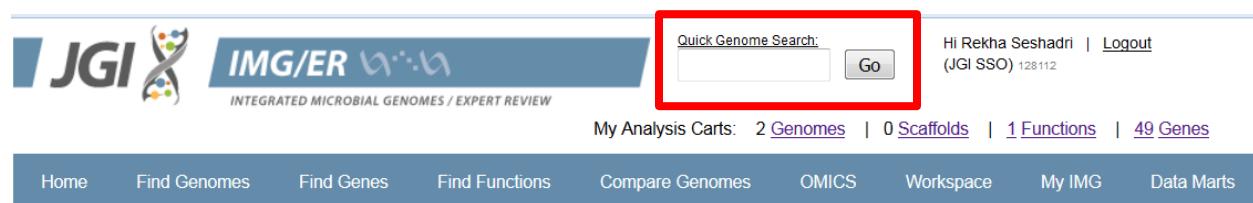
2016-02-01-08.00.00

**CASE STUDY:** Explore and compare two strains of *Dehalococcoides mccartyi* to discover putative gene(s) responsible for complete dechlorination of PCE to nontoxic end product, ethylene.

Strain 195 <sup>1</sup>	Strain CBDB1 <sup>2</sup>
1.467 Mbp	1.395 Mbp
Chloroethylenes, etc	Chlorobenzenes, etc
PCE->TCE->DCE-> <b>VC</b> ->Eth	PCE->TCE->DCE
17(+2) RDs	32 RDs

**Question 1:** Given that these are designated as strains of the same species based on 16S rRNA identity, what is the extent of conservation of gene order or “synteny”?

- Add the following 2 genomes to your genome cart using the “Quick genome Search” box at the top of the page: [Dehalococcoides mccartyi CBDB1](#) and [Dehalococcoides mccartyi 195](#)



The screenshot shows the JGI/IMG/ER website interface. At the top, there is a navigation bar with links for Home, Find Genomes, Find Genes, Find Functions, Compare Genomes, OMICS, Workspace, My IMG, and Data Marts. Below the navigation bar is a search bar labeled "Quick Genome Search:" with a red box highlighting it. To the right of the search bar is a "Go" button. Further to the right, it shows the user's name "Hi Rekha Seshadri" and a link to "Logout". Below the search bar, it says "My Analysis Carts: 2 Genomes | 0 Scaffolds | 1 Functions | 49 Genes". The main content area is currently empty.

## All Fields Genome Search Results

**hint:** Go to [Preferences](#) to show or hide plasmids, GFragment and viruses.  
Go to home page statistics under [IMG Genomes](#) to select individual phylogenetic domains or all genomes.

Domains(D): \* = Microbiome,  
B = Bacteria, A = Archaea, E = Eukarya, P = Plasmids, G = GFragment, V = Viruses.  
Genome Completion(C): F = Finished, P = Permanent Draft, D = Draft.

- o [Table Configuration](#)
- o [Save to Workspace](#)

**hint:** **Data Statistics with \*** [assembled, unassembled, both] means metagenomes counts or percentages only use assembled data or unassembled data or both (assembled data and unassembled data) for its calculations. This does not apply to isolates. The [assembled, unassembled, both] pick list is available under the Table Configuration Data Statistics. **The default is assembled data.**

Add Selected to Genome Cart	Select All	Clear All					
Filter column: Domain	Filter text: :	Apply					
Export	Page 1 of 1 << first < prev 1 next > last >> All						
Column Selector	Select Page	Deselect Page					
Select	Domain	Status	Study Name	Genome Name / Sample Name	Sequencing Center	IMG Genome ID	GOLD Analysis F
<input type="checkbox"/>	B	F	Dehalococcoides ethenogenes 195	<a href="#">Dehalococcoides mccartyi</a> <a href="#">195</a>	J. Craig Venter Institute (JCVI)	637000089	Genome Analysis

- Use Compare Genomes > Synteny Viewers > Dot Plot

Home	Find Genomes	Find Genes	Find Functions	Compare Genomes	OMICS	Workspace	My IMG	Data Marts	
Home > Find Genomes				Genome Statistics	1 Loaded				
<b>Fri. May. 6, 2016.</b> NERSC will be replacing IMG's web server. <b>Tue. May. 24, 2016.</b> NERSC and JGI quarterly maintenance. A				Synteny Viewers	VISTA	inconvenience.			
<b>All Fields Genome Search Results</b>				Abundance Profiles	Dot Plot				
<b>hint:</b> Go to <a href="#">Preferences</a> to show or hide plasmids, GFragment and viruses. Go to home page statistics under <a href="#">IMG Genomes</a> to select individual phylogenetic domains or all genomes.				Phylogenetic Dist.	Artemis ACT				
Domains(D): * = Microbiome, B = Bacteria, A = Archaea, E = Eukarya, P = Plasmids, G = GFragment, V = Viruses. Genome Completion(C): F = Finished, P = Permanent Draft, D = Draft.				Avg Nucleotide Ident.					
				Distance Tree					
				Function Profile					
				Genome Clustering	s only use assembled data or unassembled data for its calculations. The [assembled, unassembled, both] pick list is available under the Table Configuration Data Statistics. <b>The default is assembled data.</b>				
				Genome Gene Best Hmigs					
				Phylo. Marker COGs					
				Filter column: Domain	Filter text: :	Apply	?		

- Select and Add both genomes and click “Dotplot”

**DotPlot**   

**Dot Plot** employs [Mummer](#) to generate dotplot diagrams between two genomes. It uses input DNA sequences directly for comparing genomes with similar sequences and uses the six frame amino acid translation of the DNA input sequences ([PROmer](#)) for comparing genomes with dissimilar sequences (because the DNA sequence is conserved as the amino acid translation).

Please select 2 genomes.

**Sequencing Status**

All Finished, Permanent Draft and Draft Genome Cart

List  Tree Show  Selected: 2

Search for: <enter a genome name to search>

Dehalococcoides mccartyi 195 (B) [F]  
Dehalococcoides mccartyi CBDB1 (B) [F]

**Selected Genomes**

Please select 2 genomes: 2 selected

Dehalococcoides mccartyi 195 (B) [F]  
Dehalococcoides mccartyi CBDB1 (B) [F]

**Add >** **Add All >>**

< Remove  
<< Remove All

**Algorithm:**

Nucleotide sequence based comparisons  
 Protein sequence based comparisons

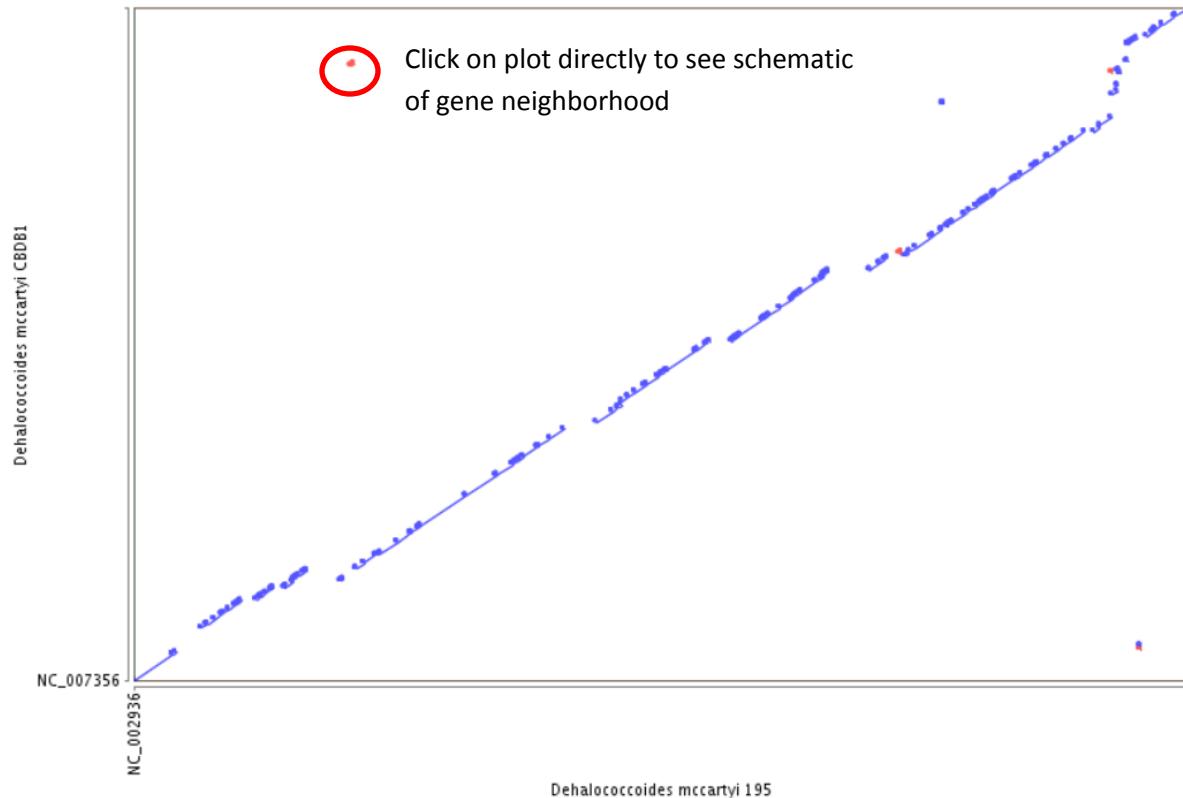
**Reference:**

Use 1 as reference  
 Use 2 as reference

**Dotplot** Reset

Dehalococcoides mccartyi 195 vs.  
Dehalococcoides mccartyi CBDB1

Using [nucmer](#) to compare genomes:



**Answer 1:** Dot plot reveals extensive synteny with some gaps, rearrangement, transposition or inversion appears to involve reductive dehalogenase genes, key enzymes for their hallmark property.

## Question 2: What proportion of genes or proteins is conserved?

- Use Compare Genomes > Average Nucleotide Identity > Pairwise ANI

Fri. May. 6, 2016. NERSC will be replacing IMG's web server.  
Tue. May. 24, 2016. NERSC and JGI quarterly maintenance. A

Datasets	JGI	All
Bacteria	6341	43427
Archaea	373	1174
Eukarya	31	257
Plasmids	1	1220
Viruses	5185	
Genome Fragments	1201	
Metagenome	4747	10147
metatranscriptome	837	1408
Total Datasets	62611	
My Private Datasets	12530	

Last Datasets Added On:  
Genome 2016-05-18  
Metagenome 2016-05-22

[Project Map](#)  
[Metagenome Projects Map](#)  
[System Requirements](#)

 Hands on training available at the

The **Integrated Microbial Genomes** (IMG) database provides a comparative analysis and annotation of genomic datasets in a comprehensive and user-friendly interface. The **IMG** database contains over 60,000 public datasets provided by **IMG** users and the scientific community, including bacterial, archaeal, eukaryotic, and metagenomic datasets.

**IMG/ER** provides users with tools to analyze and compare (with protected access) genome datasets and/or metagenome datasets ([here](#)). **IMG/ER** also provides users with tools to analyze and compare (with protected access) genome and metagenome datasets.

**IMG/ER Statistics**

Sequenced at:	Engine
Metagenome	<a href="#">JGI</a>
Metatranscriptome	<a href="#">104</a>

**Metagenome (excluding metatranscriptome)**

Associated	All
Genome Gene Best HmLgs	<a href="#">1</a> <a href="#">3189</a>
Phylo. Marker COGs	<a href="#">0</a> <a href="#">295</a>

**Distance Tree** **Same Species Plot**

**Function Profile** **ANI Cliques**

**Genome Clustering**

**IMG/ER** contains [246](#) public studies, 5180 public metagenome datasets ([4780](#) unique samples) distributed across [11](#) public projects.

## Pairwise ANI

BBHs between a genome pair are computed as pairwise bidirectional best nSimScan hits of genes having 70% or more identity and at least 70% coverage of the shorter gene. You may either select genome(s) from IMG or you may upload a nucleotide sequence in FASTA format (using the [Upload File](#) button) to compute ANI to selected genome(s) in IMG.

Please select up to 100 genomes:

**Sequencing Status**  
 All Finished, Permanent Draft and Draft

List

Tree
[Show](#)
[?](#)

Selected: 1
[Add](#)
[Upload Sets](#)
[Remove](#)
[Upload File](#)

Search for: <enter a genome name to search>  
 Dehalococcoides mccartyi 195 (B) [F]  
**Dehalococcoides mccartyi CBDB1 (B) [F]**

**Selected Genomes** [?](#)  
 Pairwise 1: 1 selected  
 Dehalococcoides mccartyi 195 (B) [F]

[Add](#)
[Upload Sets](#)
[Remove](#)
[Upload File](#)

Pairwise 2: 1  
 Dehalococcoides mccartyi CBDB1 (B) [F]

[Add](#)
[Upload Sets](#)
[Remove](#)

[ANI](#)

## Pairwise ANI

Filter column:		Genome1 ID	Filter	text	Apply	?				
		Export	Page 1 of 1 << first < prev 1 next > last >>		All					
Column Selector		Genome1 Name	Genome2 ID	Genome2 Name	ANI1->2	ANI2->1	AF1->2	AF2->1	Total BBH	Precomputed ?
637000089	<a href="#">Dehalococcoides mccartyi 195</a>	637000090	<a href="#">Dehalococcoides mccartyi CBDB1</a>		86.35	86.34	0.810	0.830	1251	Yes
Export		Page 1 of 1 << first < prev 1 next > last >>		All						

- To compare proteins, use Compare Genomes > Genome Genes Best Hmlgs

## Genome Gene Best Homologs

Percent Identity

### Sequencing status

All Finished, Permanent Draft and Draft

### Domain

Genome Cart

List  Tree

Show 

Selected: 1

Search for: <enter a genome name to search>

Dehalococcoides mccartyi 195 (B) [F]  
Dehalococcoides mccartyi CBDB1 (B) [F]

### Selected Genomes

#### Reference Genome

Dehalococcoides mccartyi 195 (B) [F]



#### Query Genomes (50 max) 1

Dehalococcoides mccartyi CBDB1 (B) [F]

#### Excluded Genomes

Dehalococcoides mccartyi CBDB1 (B) [F]

**Answer 2:** Both nucleotide and amino acid sequence based comparisons suggest about 75-80% of the genes and proteins are conserved based on best reciprocal hits.

**Question 3:** How many genes or what functions appear to be unique to Strain 195?

- Use Find Genes > Phylogenetic Profilers > Single Genes

## Phylogenetic Profiler for Single Genes

**Sequencing Status**      **Domain**

All Finished, Permanent Draft and Draft      Genome Cart

List  Tree Show ?      Selected: 1

Search for: <enter a genome name to search>

Dehalococcoides mccartyi 195 (B) [F]  
Dehalococcoides mccartyi CBDB1 (B) [F]

**Selected Genomes**

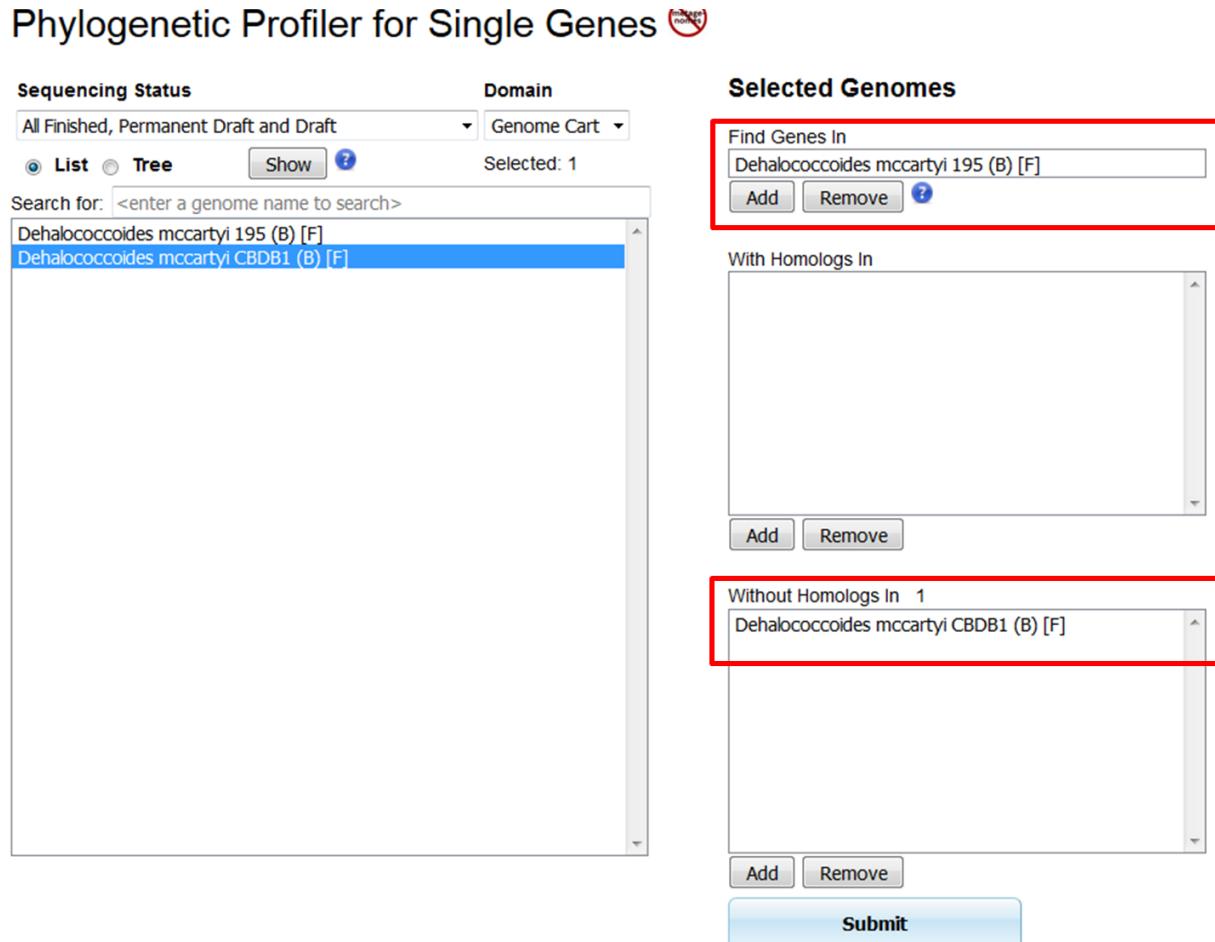
Find Genes In  
Dehalococcoides mccartyi 195 (B) [F]  
Add Remove ?

With Homologs In

Add Remove

Without Homologs In 1  
Dehalococcoides mccartyi CBDB1 (B) [F]  
Add Remove

**Submit**



## Advance Options

### Similarity Cutoffs

Max. E-value	1e-10
Min. Percent Identity	1e-2
Exclude Pseudo Genes	1e-5
Algorithm	By Present/Absent Homologs
Min. Taxon Percent With Homologs	100
Min. Taxon Percent Without Homologs	100

- Results table filtered by Gene Name for “nitrogenase”

Select	Result ▾	Gene Object ID	Locus Tag	Gene Name	Length	KEGG Map Name	KEGG Module Name
<input type="checkbox"/>	231	<a href="#">637120711</a>	DET1148	<b>nitrogenase</b> cofactor biosynthesis protein NifB, putative	276	-	-
<input type="checkbox"/>	234	<a href="#">637120714</a>	DET1151	<b>dinitrogenase</b> iron-molybdenum cofactor NifB/Y/X family protein	134	-	-
<input type="checkbox"/>	235	<a href="#">637120715</a>	DET1152	<b>nitrogenase</b> molybdenum-iron protein, beta subunit, putative	451	-	-
<input type="checkbox"/>	236	<a href="#">637120716</a>	DET1153	<b>nitrogenase</b> MoFe cofactor biosynthesis protein NifE	454	-	-
<input type="checkbox"/>	237	<a href="#">637120717</a>	DET1154	Mo- <b>nitrogenase</b> MoFe protein subunit NifK (EC 1.18.6.1)	461	Microbial metabolism in diverse environments Metabolic pathways Chloroalkane and chloroalkene degradation Nitrogen metabolism	Nitrogen fixation, nitrogen => ammonia Nitrogen fixation, nitrogen => ammonia Nitrogen fixation, nitrogen => ammonia Nitrogen fixation, nitrogen => ammonia
<input type="checkbox"/>	238	<a href="#">637120718</a>	DET1155	<b>nitrogenase</b> molybdenum-iron protein alpha chain	539	Microbial metabolism in diverse environments Chloroalkane and chloroalkene degradation Nitrogen metabolism Metabolic pathways	Nitrogen fixation, nitrogen => ammonia Nitrogen fixation, nitrogen => ammonia Nitrogen fixation, nitrogen => ammonia Nitrogen fixation, nitrogen => ammonia
<input type="checkbox"/>	239	<a href="#">637120721</a>	DET1158	Mo- <b>nitrogenase</b> iron protein subunit NifH (EC 1.18.6.1)	274	Nitrogen metabolism Metabolic pathways Chloroalkane and chloroalkene degradation Microbial metabolism in diverse environments	Nitrogen fixation, nitrogen => ammonia Nitrogen fixation, nitrogen => ammonia Nitrogen fixation, nitrogen => ammonia Nitrogen fixation, nitrogen => ammonia

**Answer 3:** About 300 genes appear to be unique to Strain 195 – these include a nitrogenase operon for nitrogen fixation, and a small number of reductive dehalogenases, that might correspond to observed differences in organohalide substrate specificities.

**Question 4:** The two strains have different dehalogenation profiles (i.e., preferences in halogenated substrate specificities), in particular, *D. mccartyi* strain195 is capable of COMPLETE dechlorination of PCE to ethylene, unlike Strain CBDB1. Can we identify the reductive dehalogenase(s) responsible for the terminal steps?

- Use *Find Genes > Gene Search*

The screenshot shows the JGI/IMG/MER homepage. At the top, there is a logo for JGI and another for IMG/MER. Below the logos, there is a search bar labeled "Quick Genome Search:" and a link "My Analysis Carts\*\*". The main navigation menu includes "Home", "Find Genomes", "Find Genes" (which is highlighted with a red box), "Find Functions", "Compare Genomes", and "OMIC". A dropdown menu titled "IMG/ER Content" is open under the "Find Genes" button. This dropdown menu has a sub-menu titled "Gene Search" which is also highlighted with a red box. Other options in the dropdown include "Cassette Search", "BLAST", and "Phylogenetic Profilers". To the right of the dropdown, there is a brief description of the system: "The Integrated Microbial Genomes (IMG) system serves as a comprehensive resource for genome and metagenome datasets in a community-based environment. The **IMG data warehouse** integrates genome and metagenome datasets with tools (IMG/ER UI Map) for analyzing their genomic features. Access genome and metagenome datasets (<http://nar.oxfordjournals.org/content/42/1/13204>) or access genome and metagenome datasets in IMG." Below the dropdown, there are three buttons: "IMG/ER Statistics", "Data Usage Policy", and "Data".

Datasets	JGI	All
Bacteria	7111	48607
Archaea	402	1445
Eukarya	31	259
Plasmids	1	1223
Viruses	5185	
Genome Fragments	1196	
Metagenome & Metatranscriptome	5569	11399
Total Datasets	69314	
My Private Datasets	13204	

- On the next page, select both genomes from your genome cart, use the pfam ID to search for reductive dehalogenases (pfam13486) (why not use the default gene product name search instead? Names can be inconsistent)

## Gene Search

Find genes in selected genomes by keyword. It's required to add selections into "Selected Genomes" unless blocked.  
\*MER-FS Metagenome supported search filters.

The screenshot shows the Gene Search interface. In the 'Keyword' field, 'pfam13486' is entered. In the 'Filters' dropdown, 'Pfam Domain Search (list) \*' is selected. The 'List' radio button is selected under 'Search for'. The results list includes various search filters like 'Gene Product Name (inexact)\*', 'MyIMG Annotation (inexact)', etc. At the bottom right, it says 'MER-FS Metagenome: Assembled'. Below the search bar are 'Go' and 'Reset' buttons.

- Nest, from Gene Cart page, use the Sequence Alignment Tab and “Do alignment”

## Gene Cart

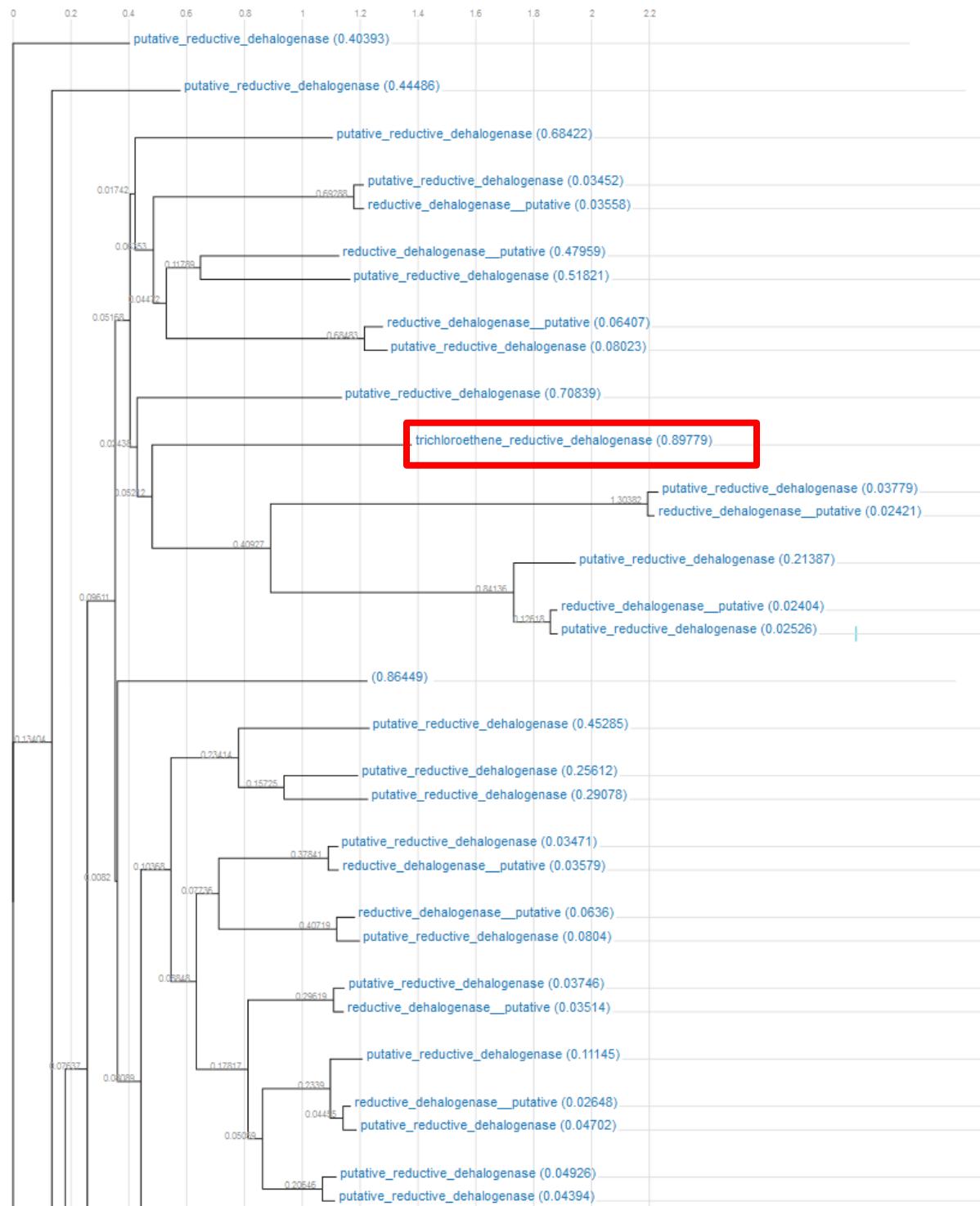
Only a maximum of 20000 genes can be in cart.

49 gene(s) in cart

The screenshot shows the Gene Cart interface. At the top, there are tabs: 'Genes in Cart', 'Functions', 'Upload & Export & Save', 'Chromosome Map', 'Sequence Alignment' (which is highlighted with a red box), 'Gene Neighborhoods', 'Profile & Alignment', and 'Edit'. Below these are buttons for 'Add Genomes of Selected Genes to Cart' and 'Add Scaffolds of Selected Genes to Cart'. Further down are buttons for 'Toggle Selected', 'Select All', 'Clear All', and 'Remove Selected'. A message '49 of 49 rows selected' is displayed. At the bottom, there are buttons for 'Export', 'Page 1 of 1', 'Filter column: Gene Product Name', 'Filter text: :', 'Apply', and a help icon. The main area is a table with columns: Select, Gene ID, Locus Tag, Gene Product Name, Genome ID, Genome Name, Batch<sup>1</sup>, and Amino Acid Sequence Length (aa). The table lists 49 genes, all of which have 'Selected' checked in the first column.

Select	Gene ID	Locus Tag	Gene Product Name	Genome ID	Genome Name	Batch <sup>1</sup>	Amino Acid Sequence Length (aa)
<input checked="" type="checkbox"/>	<a href="#">640733214</a>	DET0162		637000089	<a href="#">Dehalococcoides mccartyi</a> <u>195</u>	3	486
<input checked="" type="checkbox"/>	<a href="#">637703851</a>	cbdb_A1491	putative reductive dehalogenase	637000090	<a href="#">Dehalococcoides mccartyi</a> <u>CBDB1</u>	2	482
<input checked="" type="checkbox"/>	<a href="#">637702640</a>	cbdb_A84	putative reductive dehalogenase	637000090	<a href="#">Dehalococcoides mccartyi</a> <u>CBDB1</u>	2	488
<input checked="" type="checkbox"/>	<a href="#">637703982</a>	cbdb_A1624	putative reductive dehalogenase	637000090	<a href="#">Dehalococcoides mccartyi</a> <u>CBDB1</u>	2	495
<input checked="" type="checkbox"/>	<a href="#">637703944</a>	cbdb_A1582	putative reductive dehalogenase	637000090	<a href="#">Dehalococcoides mccartyi</a> <u>CBDB1</u>	2	491

- From the alignment results page, select “Rectangular Phylogram” tab to view a tree



**Answer 4:** From the tree, it is possible to discern a handful of divergent candidates in strain 195 relative to CBDB1—the gene described as “trichloroethene reductive dehalogenase” is experimentally validated to dehalogenate DCE and VC to ethylene<sup>4</sup>

**CASE STUDY:** Discovering novel determinants of symbiosis through comparative genomics of 100+ root nodule bacteria against a control genomes set of non-plant associated bacteria. The objective is to showcase the metadata available in IMG courtesy of the GOLD effort<sup>5</sup>.

The screenshot shows the JGI/IMG/ER homepage. At the top, there is a navigation bar with links for Home, Find Genomes, Find Genes, Find Functions, Compare Genomes, and OMICs. A "Quick Genome Search" input field is also present. Below the navigation bar, there is a message about a scheduled centerwide outage. On the left, there is a sidebar with categories for Bacteria, Archaea, Eukarya, Plasmids, Viruses, Genome Fragments, Metagenome & Metatranscriptome, Total Datasets, and My Private Datasets. The "Genome Browser" link under the "Find Genomes" section is highlighted with a red box. A tooltip for this link states: "ERSC Scheduled Centerwide Outage. This may impact convenience." To the right of the tooltip, there is a detailed description of the Integrated Microbial Genomes (IMG) system. At the bottom right, there are links for "IMG/ER Statistics" and "Data Submission Site".

Quick Genome Search:

My Analysis Carts\*\*: 0 [Genomes](#)

Home Find Genomes Find Genes Find Functions Compare Genomes OMICs

**Sat. Jun 10, 2017** during this time.

**Genome Browser** ERSC Scheduled Centerwide Outage. This may impact convenience.

**Genome Search**

**Scaffold Search**

**Deleted Genomes**

Bacteria 3907

Archaea 1192

Eukarya 4886

Plasmids 5606

Viruses

Genome Fragments

Metagenome & Metatranscriptome

Total Datasets 50468

My Private Datasets 0

The **Integrated Microbial Genomes (IMG)** system serves as a comprehensive analysis and annotation of genome and metagenome datasets in a comparative context. The **IMG data warehouse** integrates genome and metagenome datasets provided by IMG users with a comprehensive set of publicly available datasets.

IMG/ER provides users with tools ([IMG/ER UI Map](#)) for analyzing their protected access genome datasets (<http://nar.oxfordjournals.org/content/42>) and/or metagenome datasets (<http://nar.oxfordjournals.org/content/42>) genome and metagenome datasets in IMG.

[IMG/ER Statistics](#) [Data Submission Site](#)

[Home](#) > Find Genomes

 **Sat. June 11 to Sun. June 12 NERSC Scheduled Center**  
during this time. Sorry for the inconvenience.

## Genome Browser

[Add Selected to Genome Cart](#)

[View Alphabetically](#)

Only visible selected genomes will be saved.

- ◆ Green plus to select
- Red minus to clear

[Open All](#)

[Close All](#)

--- ▾ --- ▾

01 ► ◆ - \*Microbiome (5606)

01 ► ◆ - Archaea (763)

**01 ► ◆ - Bacteria (37588)**

01 ► ◆ - Eukaryota (220)

01 ► ◆ - GFragment (1192)

01 ► ◆ - Plasmid (1192)

01 ► ◆ - Viruses (3907)

[Click on arrow to EXPAND a](#)

[lineage , in this instance](#)

[EXPAND Bacteria >](#)

[Proteobacteria >](#)

[Alphaproteobacteria >](#)

[Add Selected to Genome Cart](#)

[View Alphabetically](#)

02▼ • - Proteobacteria (16879)  
 03► • - Acidithiobacillia (11)  
 03▼ • - Alphaproteobacteria (2389)  
 04► • - Caulobacterales (69)  
 04► • - Kiloniellales (6)  
 04► • - Kordiimonadales (3)  
 04► • - Magnetococcales (1)  
 04► • - Parvularculales (2)  
 04► • - Pelagibacterales (30)  
**04▼** • - Rhizobiales (1241) **Scroll down, and click on Green Dot to select all 1241 genomes under "Rhizobiales"**  
 05► • - Aurantimonadaceae (—, —)  
 05► • - Bartonellaceae (81)  
 05► • - Beijerinckiaceae (10)  
 05► • - Bradyrhizobiaceae (167)  
 05► • - Brucellaceae (369)  
 05► • - Cohaesibacteraceae (2)  
 05► • - Hyphomicrobiaceae (49)  
 05► • - Methylobacteriaceae (87)  
 05► • - Methylocystaceae (17)  
 05► • - Phyllobacteriaceae (78)  
 05► • - Rhizobiaceae (323)  
 05► • - Rhodobiaceae (9)  
 05► • - Xanthobacteraceae (13)  
 05► • - unclassified (13)

- When the next screen loads, add these 1241 selected genomes to Genome Cart:

## Genome Browser

**Add Selected to Genome Cart**

**View Alphabetically**

Only visible selected genomes will be saved.

- Green plus to select
- Red minus to clear

- 1241 genomes are now loaded in your genome cart. On this page, scroll down to “Table configuration” and make your “metadata “selections and Redisplay the table:

## Table Configuration

**Redisplay**

Genome Field	Metadata (Updated Jun 7 2016)	Data Statistics
<input type="button" value="All"/> <input type="button" value="Clear"/> <input checked="" type="checkbox"/> Domain <input checked="" type="checkbox"/> Status <input checked="" type="checkbox"/> Study Name <input checked="" type="checkbox"/> Genome Name / Sample Name <input checked="" type="checkbox"/> Sequencing Center <input checked="" type="checkbox"/> IMG Genome ID (IMG Taxon ID) <input type="checkbox"/> Phylum <input type="checkbox"/> Class <input type="checkbox"/> Order <input type="checkbox"/> Family <input type="checkbox"/> Genus <input type="checkbox"/> Species <input type="checkbox"/> NCBI Taxon ID <input type="checkbox"/> RefSeq Project ID <input type="checkbox"/> NCBI Project ID <input type="checkbox"/> IMG Submission ID <input type="checkbox"/> JGI Project ID / ITS PID <input type="checkbox"/> GOLD Study ID <input type="checkbox"/> GOLD Project ID <input type="checkbox"/> GOLD Analysis Project ID <input type="checkbox"/> GOLD Analysis Project Type <input type="checkbox"/> Gene Model QC <input type="checkbox"/> Submission Type <input type="checkbox"/> Assembly Method <input type="checkbox"/> Strain <input type="checkbox"/> Funding Agency <input type="checkbox"/> Is Public <input type="checkbox"/> Comments <input type="checkbox"/> IMG Release <input type="checkbox"/> IMG Product Assignment <input type="checkbox"/> High Quality <input type="checkbox"/> Add Date <input type="checkbox"/> Release Date <input type="checkbox"/> Distance Matrix Calc. Date	<input type="button" value="All"/> <input type="button" value="Clear"/> <input type="checkbox"/> Alt. Contact Email <input type="checkbox"/> Alt. Contact Name <input type="checkbox"/> Alt2. Contact Emails (GOLD) <input type="checkbox"/> Alt2. Contact Names (GOLD) <input type="checkbox"/> Altitude <input type="checkbox"/> Bioproject Accession <input type="checkbox"/> Biosample Accession <input type="checkbox"/> Biotic Relationships <input type="checkbox"/> Cell Arrangement <input type="checkbox"/> Cell Shape <input type="checkbox"/> Clade <input type="checkbox"/> Contact Email <input type="checkbox"/> Contact Name <input type="checkbox"/> Culture Type <input type="checkbox"/> Cultured <input type="checkbox"/> Depth <input type="checkbox"/> Diseases <input checked="" type="checkbox"/> Ecosystem <input checked="" type="checkbox"/> Ecosystem Category <input checked="" type="checkbox"/> Ecosystem Subtype <input checked="" type="checkbox"/> Ecosystem Type <input type="checkbox"/> Ecotype <input type="checkbox"/> Energy Source <input type="checkbox"/> Funding Program <input type="checkbox"/> GOLD Sequencing Depth <input type="checkbox"/> GOLD Sequencing Quality <input type="checkbox"/> GOLD Sequencing Strategy <input type="checkbox"/> GPTS Proposal Id <input type="checkbox"/> Geographic Location <input type="checkbox"/> Gram Staining <input type="checkbox"/> HMP ID <input checked="" type="checkbox"/> Habitat <input type="checkbox"/> Host Gender <input type="checkbox"/> Host Name	<input type="button" value="All"/> <input type="button" value="Clear"/> <input type="button" value="Select Counts"/> <input type="button" value="Select Percentage"/> * Assembled (Metagenomes) <input checked="" type="checkbox"/> * Genome Size (Number of total bases) <input checked="" type="checkbox"/> * Gene Count (Number of total Genes) <input checked="" type="checkbox"/> * Scaffold Count (Number of scaffolds) <input type="checkbox"/> * CRISPR Count (Number of CRISPRs) <input type="checkbox"/> * GC Count (Number of GC) <input type="checkbox"/> * GC (GC % in fraction) <input type="checkbox"/> Coding Base Count (Total number of coding bases) <input type="checkbox"/> Coding Base Count % (Percentage of Total number of coding bases) <input type="checkbox"/> Coding Base Count NP (Total number of coding bases no pseudogenes) <input type="checkbox"/> Coding Base Count NP % (Percentage of Total number of coding bases no pseudogenes) <input checked="" type="checkbox"/> * CDS Count (Number of CDS genes) <input type="checkbox"/> * CDS % (Percentage of CDS genes) <input type="checkbox"/> * RNA Count (Number of RNA genes) <input type="checkbox"/> * RNA % <input type="checkbox"/> * rRNA Count (Number of rRNA genes) <input type="checkbox"/> * 5S rRNA Count (Number of 5S rRNAs) <input type="checkbox"/> * 16S rRNA Count (Number of 16S rRNAs) <input type="checkbox"/> * 18S rRNA Count (Number of 18S rRNAs) <input type="checkbox"/> * 23S rRNA Count (Number of 23S rRNAs) <input type="checkbox"/> * 28S rRNA Count (Number of 28S rRNAs) <input type="checkbox"/> * tRNA Count (Number of tRNA genes) <input type="checkbox"/> * Other RNA Count (Number of other unclassified RNA genes) <input type="checkbox"/> Pseudo Genes Count (Number of pseudo genes) <input type="checkbox"/> Pseudo Genes % (Percentage of pseudo genes) <input type="checkbox"/> Unchar Count (Number of uncharacterized genes) <input type="checkbox"/> Unchar % (Percentage of uncharacterized genes) <input type="checkbox"/> Dubious Count <input type="checkbox"/> Dubious % <input type="checkbox"/> * w/ Func Pred Count (Number of genes with predicted protein product) <input type="checkbox"/> * w/ Func Pred % (Percentage of genes with predicted protein product) <input type="checkbox"/> * w/o function prediction <input type="checkbox"/> * w/o function prediction %

- Curate your list of “Control” genomes (genomes from within the Order Rhizobiales with NO PLANT ASSOCIATION) using these “metadata”
- Next, retrieve gene counts of all Pfams for all Control genomes. First use Find Functions > Pfam > Pfam list > Select All 16,295 pfams and Add to Function Cart

Home Find Genomes Find Genes Find Functions Compare Genomes OMICS Workspace My IMG Data Marts Help

[Home](#) > Analysis Cart

**Sat. June 11 to Sun. June 12 NERSC S** during this time. Sorry for the inconvenience.

## Genome Cart

Only a maximum of 20000 genomes can be in cart.  
1241 genome(s) in cart

[Genomes in Cart](#) [Upload & Export & Save](#)

[Group Genome Cart by Phyla](#)

**hint:** Scaffolds will not be added into cart for Only scaffolds (assembled data only) or

**Add Scaffolds of Selected Genomes to Cart**

Filter column: Domain Export Page 1 of 13 << first < prev 1 2 Select Page Deselect

Select	Domain	Status	Study Name	Name	Sequencing Center	IMG Genome ID	Genome Size * assembled	Gene Count * assembled
<input type="checkbox"/>	B	D	Comparative genomics of bacterial root endophytes of switchgrass collected from prairies over two seasons	DOE Joint Genome Institute (JGI)	2600254933	4750527	4541	
<input type="checkbox"/>	B	P	Brucella melitae bv. 1 str. M5	IM5	Chinese Academy of Science, Institute of Microbiology	2562617108	3277964	3188

## Pfam Families

[Add Selected to Function Cart](#) [Select All](#) [Clear All](#)

16295 of 16295 rows selected

Filter column: Pfam ID Export Page 1 of 163 << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> 100

Column Selector Select Page Deselect Page

Select	Pfam ID	Pfam Name
<input checked="" type="checkbox"/>	<a href="#">pfam00001</a>	7tm_1 - 7 transmembrane receptor (rhodopsin family)
<input checked="" type="checkbox"/>	<a href="#">pfam00002</a>	7tm_2 - 7 transmembrane receptor (Secretin family)
<input checked="" type="checkbox"/>	<a href="#">pfam00003</a>	7tm_3 - 7 transmembrane sweet-taste receptor of 3 GPCR
<input checked="" type="checkbox"/>	<a href="#">pfam00004</a>	AAA - ATPase family associated with various cellular activities (AAA)
<input checked="" type="checkbox"/>	<a href="#">pfam00005</a>	ABC_tran - ABC transporter
<input checked="" type="checkbox"/>	<a href="#">pfam00006</a>	ATP-synt_ab - ATP synthase alpha/beta family, nucleotide-binding domain
<input checked="" type="checkbox"/>	<a href="#">pfam00007</a>	Cys_knot - Cystine-knot domain

- Tab over to “Profile & Alignment” on the Function Cart page, populate with genomes from “Genome Cart” and retrieve gene counts for all genomes using “View Genomes vs. Functions”

## Function Cart

Only a maximum of 20000 functions can be in cart.

16295 function(s) in cart

Functions in Cart	Upload & Export & Save	Profile & Alignment	KEGG Pathways	Analysis
-------------------	------------------------	---------------------	---------------	----------

### Profile and Alignment Tools

**hint:**

- Hold down the control key (or command key in the case of the Mac) to select multiple genomes.
- Drag down list to select all genomes.
- More genome and function selections result in slower query.

Sequencing Status	Domain
All Finished, Permanent Draft and Draft	Genome Cart
Finished	Selected: 1
Permanent Draft	
Draft	
All Finished, Permanent Draft and Draft	
Afifella pfennigii DSM 17143 (B) [P]	
Afipia birgiae 34632 (B) [P]	
Afipia broomeae ATCC 49717 (B) [P]	
Afipia clevelandensis ATCC 49720 (B) [P]	
Afipia felis felis ATCC 53690 (B) [P]	
Afipia felis genospecies A, 76713 (B) [P]	
Afipia massiliensis LC387 (B) [P]	
Afipia sp. 1NLS2 (B) [P]	
Afipia sp. AAP120 (B) [P]	

MER-FS Metagenome: Assembled

## Function Profile

View selected function(s) against selected genomes. GO, Interpro and IMG Network are not supported.

MER-FS metagenomes only support COG, EC, Pfam, TIGRFam, KO and MetaCyc.

Use the Genome Filter above to select 1 to 500 genome(s).

View Functions vs. Genomes	View Genomes vs. Functions
----------------------------	----------------------------

- Export the Function Profile table for further analyses.

## Function Profile

16295 functions and 172 genomes are selected.

**hint:** Mouse over genome abbreviation to see genome name.  
Cell coloring is based on gene count: white = 0, bisque = 1-4, yellow >= 5.

Show Slim View

Filter column: Function ID Filter text: Apply ?

Export Page 1 of 163 << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> 100

Column Selector Save selected rows from this table as an Excel spreadsheet or tab delimited file

Function ID	Name	Agr sp 33 650716007	Agr tum Ac5 2645727797	Agr tum 813 2654588028	Agr tum UW <sub>h</sub> 639279302	Agr tum Ac5 2558860256	Agr tum WR1 2554235003	Agr vit S44 64334850
pfam00001	7 transmembrane receptor (rhodopsin family)	0	0	0	0	0	0	0
pfam00002	7 transmembrane receptor (Secretin family)	0	0	0	0	0	0	0
pfam00003	7 transmembrane sweet-taste receptor of 3 GCPR	0	0	0	0	0	0	0
pfam00004	ATPase family associated with various cellular activities (AAA)	7	7	7	8	7	9	7
pfam00005	ABC transporter	198	179	213	209	213	204	201
pfam00006	ATP synthase alpha/beta family, nucleotide-binding domain	4	4	4	4	4	4	4
pfam00007	Cystine-knot domain	0	0	0	0	0	0	0
pfam00008	EGF-like domain	0	0	0	0	0	0	0
pfam00009	Elongation factor Tu GTP binding domain	9	8	8	9	8	8	8

**Results:** Results from this exercise, overall strategy, and more, are described in the following publication<sup>6</sup>: <http://www.nature.com/articles/srep16825>

## REFERENCES

1. Genome sequence of the PCE-dechlorinating bacterium *Dehalococcoides ethenogenes*. Seshadri R, et al. Science, 2005 Jan 7. PMID 15637277
2. Genome sequence of the chlorinated compound-respiring bacterium *Dehalococcoides* species strain CBDB1. Kube M et. Al. Nat Biotechnol. 2005 Oct;23(10):1269-73.

3. Microbial species delineation using whole genome sequences, Varghese, N. et. al. Nucl. Acids Res. (2015)
4. Trichloroethene reductive dehalogenase from Dehalococcoides ethenogenes: sequence of tceA and substrate range characterization. Magnuson JK, et al. Appl Environ Microbiol, 2000 Dec. PMID 11097881
5. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Reddy TBK, et. al., Nucl. Acids Res. (2014)
6. Discovery of Novel Plant Interaction Determinants from the Genomes of 163 Root Nodule Bacteria. Seshadri, R. et.al. Nat. Sci. Rep. 2015. 5:16825.