

IMG Hands-On EXERCISES

MGM-23 Workshop: Functional Annotation & Comparative Genomics for Gene Discovery 09/27/2016

Rekha Seshadri, DoE Joint Genome Institute, 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](#)

Find Genes

- Gene Search
- Cassette Search
- BLAST
- Phylogenetic Profilers
- Single Genes
- Gene Cassettes

IMG/ER Statistics

Data Submission Site

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

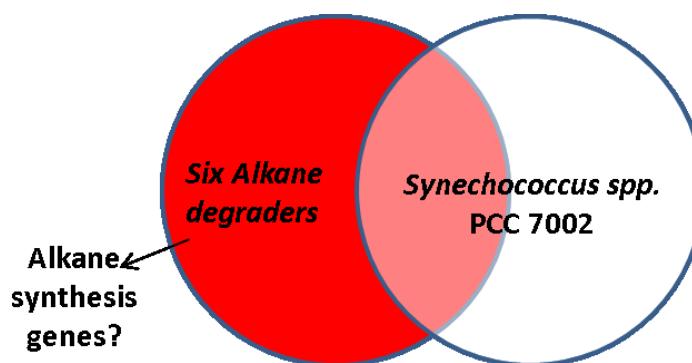
EXERCISE 1

Observation: Alkanes, the major constituents of gasoline and other fuels, are naturally produced by diverse species from metabolites of the fatty acid pathway. Biochemical studies have revealed decarbonylation of fatty aldehydes as the primary mechanism, however, no genes have been discovered. In the microbial world, cyanobacteria are known alkane producers and different cyanobacterial species were evaluated for alkane (particularly heptadecane) production with the following results:

Cyanobacterium	Genome size (Mb)	Alkanes observed
<i>Synechococcus elongatus</i> PCC7942	2.7	Heptadecane, pentadecane
<i>Prochlorococcus marinus</i> <i>pastoris</i> CCMP 1986	1.7	Pentadecane
<i>Anabaena variabilis</i> ATCC 29413	6.4	Heptadecane
<i>Gloeobacter violaceus</i> PCC 7421	4.6	Heptadecane
<i>Nostoc</i> sp. PCC 7120	6.2	Heptadecane
<i>Cyanothece</i> sp. PCC 7425	5.7	Heptadecane
<i>Synechococcus</i> sp. PCC 7002	3.0	No alkanes observed

Table adapted from Schirmer, A. et. al., "Microbial biosynthesis of alkanes" Science 2010, 329(5991):559-562

Objectives: Discover putative conserved genes for alkane synthesis (specific candidates from *Synechococcus elongatus* PCC7942) using a subtractive genome comparison approach against *Synechococcus* sp. PCC 7002



Workflow: Add the above 7 genomes to your genome cart using the "Quick Genome Search" box at the top of the page.



Next, Use **Find Genes > Phylogenetic Profilers > Single Genes**, to perform comparisons

This screenshot shows the "Find Genes" section of the JGI IMG/ER website. The top navigation bar includes links for Home, Find Genomes, Find Genes (which is the active tab), Find Functions, Compare Genomes, and OMICS. Below the navigation, a message states "My Analysis Carts**: 7 Genomes | 0". A sidebar on the left indicates "7 genome(s) in cart". The main content area features several search and analysis tools: Gene Search, Cassette Search, BLAST, Phylogenetic Profilers (with a red box around the "Single Genes" link), and Gene Cassettes. A note at the bottom of the page mentions a "Centerwide Outage" occurring between June 11 and June 12, 2016.

Use *Synechococcus elongatus PCC 7942* as query genome, and find genes that are conserved in the other alkane producers, but NOT in *Synechococcus sp. PCC 7002*. Adjust advanced options as you wish.

Phylogenetic Profiler for Single Genes

Sequencing Status

All Finished, Permanent Draft and Draft

List Tree Selected: 5

Search for: <enter a genome name to search>

Anabaena variabilis ATCC 29413 (B) [F]
Cyanothec sp. PCC 7425 (B) [F]
Gloeobacter violaceus PCC 7421 (B) [F]
Nostoc sp. PCC 7120 (B) [F]
Prochlorococcus marinus pastoris CCMP 1986 (B) [F]
Synechococcus elongatus PCC 7942 (B) [F]
Synechococcus sp. PCC 7002 (B) [F]

Selected Genomes

Find Genes In
Synechococcus elongatus PCC 7942 (B) [F]

With Homologs In 5

Anabaena variabilis ATCC 29413 (B) [F]
Cyanothec sp. PCC 7425 (B) [F]
Gloeobacter violaceus PCC 7421 (B) [F]
Nostoc sp. PCC 7120 (B) [F]
Prochlorococcus marinus pastoris CCMP 1986 (B) [F]

Without Homologs In 1

Synechococcus sp. PCC 7002 (B) [F]

Advance Options

Similarity Cutoffs

Max. E-value	1e-5 <input style="border: none; border-bottom: 1px solid black; width: 15px; height: 15px;" type="button" value="▼"/>
Min. Percent Identity	30 <input style="border: none; border-bottom: 1px solid black; width: 15px; height: 15px;" type="button" value="▼"/>
Exclude Pseudo Genes	10 20 30 <input style="border: none; border-bottom: 1px solid black; width: 15px; height: 15px;" type="button" value="▼"/>
Algorithm	Absent/Absent Homologs <input style="border: none; border-bottom: 1px solid black; width: 15px; height: 15px;" type="button" value="▼"/>
Min. Taxon Percent With Homologs	40 <input style="border: none; border-bottom: 1px solid black; width: 15px; height: 15px;" type="button" value="▼"/>
Min. Taxon Percent Without Homologs	50 60 70 80 90 100 <input style="border: none; border-bottom: 1px solid black; width: 15px; height: 15px;" type="button" value="▼"/>

Function Display Options

Examine the shortlist of genes PCC 7942 genes that were retrieved – which ones look like good candidates for alkane synthesis? Recall that biochemical evidence supports decarbonylation as a key mechanism. Check gene neighborhood – are any candidates in a potential operon?

Result: Out of the ~40 genes that were found to be conserved in alkane producers, but absent in *Synechococcus sp. PCC 7002* (which does not produce alkanes), Synpcc7942_1593 and Synpcc7942_1594 encode a fatty acid decarbonylase and an acyl-ACP reductase, enzymes involved in fatty acid metabolism. Results from Schirmer, A. et. Al. (“Microbial biosynthesis of alkanes” Science, 2010, 329(5991):559-562) show heterologous expression of these two cyanobacterial genes in *E. coli* results in alkane production.

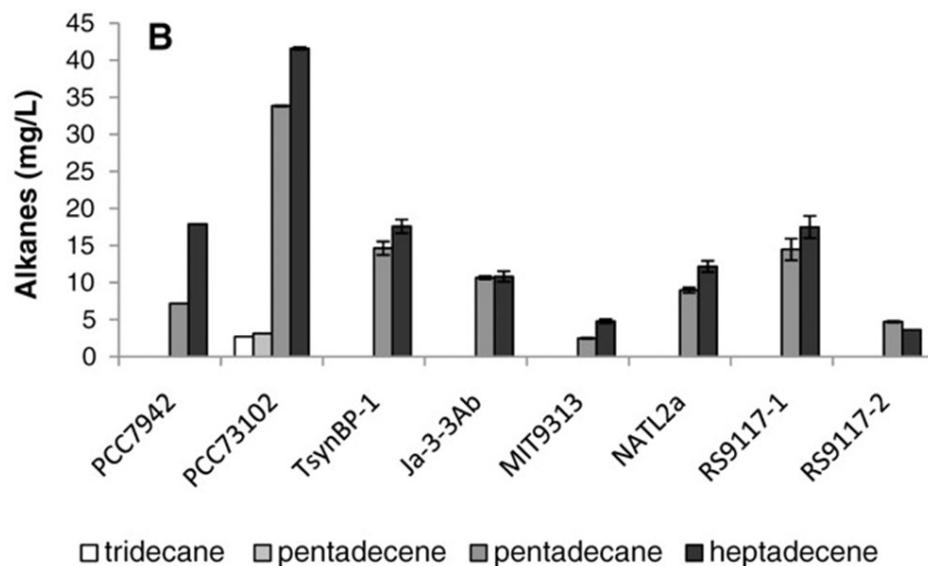


Image from Schirmer, A. et. al., Microbial biosynthesis of alkanes” Science 2010, 329(5991):559-562)

BONUS QUESTION: Examine various annotations on the gene details pages for both these candidates (Synpcc7942_1593 and Synpcc7942_1594). Are the various annotation evidences largely in agreement? Are some annotations less informative? Are some function annotations absent? Are some function annotations contradictory?

EXERCISE 2

Background: Organohalides (e.g., tetrachloroethylene, polychlorinated biphenyls, dioxins) constitute a large proportion of environmental pollutants, and reductive dehalogenases encoded by obligate organohalide-respiring bacteria (OHRB), catalyze their dehalogenation to non-toxic metabolites. Corrinoid cobalamin (B12) is an essential cofactor of reductive dehalogenases. De novo biosynthesis of corrinoid is one of the most complicated metabolic pathways in nature, and OHRB have developed different strategies to cope with their need for this cofactor.

- (1) What is the TIGRFAM ID and/or Pfam ID for “reductive dehalogenase” (Hint: Use “Find Functions > Tigrfam > Tigrfam list”, and use the filter column feature to find the appropriate ID)

The screenshot shows the IMG/MER homepage with a navigation bar at the top. The 'Find Functions' button is highlighted with a red box. A dropdown menu for 'Tigrfam' is also highlighted with a red box. The menu includes options like Function Search, Search Pathways, Secondary Metabolism, COG, KOG, and Pfam. Under Pfam, there are links for TIGRfam, KEGG, and IMG Networks. The 'TIGRfam' link is specifically highlighted with a red box.

[Home](#) > Find Functions

Pfam Families

Add Selected to Function Cart Select All Clear All

Filter column: Pfam Name ▾ Filter text ▾ : reductive dehalogenase

Export Pfam ID First < prev 1 next > last >> 100

Column Selection: All Columns Select Page Deselect Page

Select	Pfam ID	Pfam Name
<input type="checkbox"/>	pfam13486	Dehalogenase - Reductive dehalogenase subunit

Export Page 1 of 1 << first < prev 1 next > last >> 100

Add Selected to Function Cart Select All Clear All

Result: pfam13486 or TIGR02486

(2) How many **FINISHED BACTERIAL** genomes in IMG have reductive dehalogenase genes?

How many genes were found? (Hint: Use "Find Functions > Function Search" and pull down to select "Finished" and "Bacteria", strongly suggest using Pfam ID (pfam13486) or Tigrfam ID (TIGR02486) to search since assigned gene names are not reliably consistent)

JGI **IMG/ER** INTEGRATED MICROBIAL GENOMES / EXPERT REVIEW Quick Genome Search: Go Hi Rekha Seshadri | Logout (JGI SSO) 60728

My Analysis Carts**: [4 Genomes](#) | [0 Scaffolds](#) | [0 Functions](#) | [0 C](#)

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

Sat. June 11 to Sun. June 12 NERSC System Maintenance. This may impact the availability and/or performance of IMG during this time. Sorry for the inconvenience.

Function Search

[Search Pathways](#)
The International Pathway Database provides a central location for pathway analysis and comparison across multiple organisms. It integrates data from various databases and resources, making it easier to compare pathways across different species.

[Secondary Metabolism](#)
The Secondary Metabolism Resource (SMR) is a database that integrates information about secondary metabolites and their biosynthetic pathways. It provides a comprehensive resource for researchers interested in the study of secondary metabolism in microorganisms.

[COG](#)
The Comprehensive Microbial Resource (CMR) is a database that integrates genome-scale data for a wide range of microorganisms. It provides a comprehensive set of publicly available genomic and metagenomic data for research and analysis.

Function Search

Function search allows you to find functions in selected genomes by keyword.

hint: Search term marked by * indicates that it supports metagenomes.
You must add your selections into Selected Genomes.

Keyword: TIGR02486

Filters: Gene Product Name (inexact) *

Search for: COG (list) *
KOG (list)
Pfam (list) *

Sequencing Status: All Finished
 List
All Finishes
Enzyme (list) *
Transporter Classification (list)
KEGG Pathway Enzymes
KEGG Orthology ID (list) *

Search for: Dehalobac KEGG Orthology Name *
Dehalococ KEGG Orthology Definition *
Dehalococ InterPro (list)
Dehalogen MetaCyc (list)
IMG Compound (list)
IMG Term and Synonyms
IMG Pathways
IMG Parts List
All function names (slow, Gene Product Name not included)

MER-FS Metagenome: Assembled

Go **Reset**

Sequencing Status: Finished

List Tree Show

Domain: Bacteria
Archaea
Bacteria
Eukaryota
Metagenome
Plasmid
All (Slow)
Genome Cart

Selected Genomes: 4428 selected

Acaryochloris marina MBIC11017 (B) [F]
ACD20 (ACD20correct) (B) [F]
Acetobacter pasteurianus 386B (B) [F]
Acetobacter pasteurianus Ab3 (B) [F]
Acetobacter pasteurianus IFO 3283-01 (B) [F]
Acetobacter pasteurianus IFO 3283-01-42C (B) [F]
Acetobacter pasteurianus IFO 3283-03 (B) [F]
Acetobacter pasteurianus IFO 3283-07 (B) [F]
Acetobacter pasteurianus IFO 3283-12 (B) [F]
Acetobacter pasteurianus IFO 3283-22 (B) [F]

MER-FS Metagenome: Assembled

Go **Reset**

Result: 41 genomes, 420 genes (this result is for Tigrfam). **BONUS QUESTION:** do you get a different result if you used Pfam, why do you think that might be?)

From this above list, add the following 5 FINISHED genomes to your cart:

Genome	Phylum	Genome Size (Mbp)
<i>Dehalobacter restrictus</i> DSM 9455	Firmicute	2.94
<i>Dehalobacter</i> sp. CF	Firmicute	3.09
<i>Dehalobacter</i> sp. 11DCA	Firmicute	3.06
<i>Dehalogenimonas lykanthroporepellens</i> BL-DC-9	Chloroflexi	1.68
<i>Dehalococcoides mccartyi</i> 195	Chloroflexi	1.46

(3) Recall that the two *Dehalococcoides mccartyi* strains' genomes shared extensive synteny. Is there similar conservation of gene order between any of the *Dehalobacter* genomes? What can you possibly infer about the relationship of the 3 strains? (Hint: Compare Genomes > Synteny Viewers > Dot Plot; Compare Genomes > Average Nucleotide Identity > Pairwise ANI)

DotPlot

Dot Plot employs [Mummer](#) to generate dotplot diagrams between two genomes. It uses input DNA sequences directly for comparing genomes with similar sequences ([NUCmer](#)). It uses the six frame amino acid translation of the DNA input sequences ([PROmer](#)) for comparing genomes with dissimilar sequences (because the DNA sequence is not as highly conserved as the amino acid translation).

Please select 2 genomes.

Sequencing Status Domain

List Tree Show  Selected: 2

Search for: <enter a genome name to search>

- Dehalobacter restrictus DSM 9455 (B) [F]
- Dehalobacter sp. 11DCA (B) [F]
- Dehalobacter sp. CF (B) [F]
- Dehalococcoides mccartyi 195 (B) [F]
- Dehalogenimonas lykanthroporepellens BL-DC-9 (B) [F]

Selected Genomes

Please select 2 genomes: 2 selected

- Dehalobacter sp. 11DCA (B) [F]
- Dehalobacter sp. CF (B) [F]

Algorithm: Nucleotide sequence based comparisons Protein sequence based comparisons

Reference: Use 1 as reference Use 2 as reference

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:

Quick Search: <enter a genome name to search>

Sequencing Status

All Finished, Permanent Draft and Draft

Domain

Genome Cart

List Tree

Show



Selected: 1

Search for: <enter a genome name to search>

Dehalobacter restrictus DSM 9455 (B) [F]

Dehalobacter sp. 11DCA (B) [F]

Dehalobacter sp. CF (B) [F]

Dehalococcoides mccartyi 195 (B) [F]

Dehalogenimonas lykanthroporepellens BL-DC-9 (B) [F]

Selected Genomes

Pairwise 1: 1 selected

Dehalobacter sp. 11DCA (B) [F]

Add

Upload Sets

Remove

Upload File

Pairwise 2: 1

Dehalobacter sp. CF (B) [F]

Add

Upload Sets

Remove

ANI

Result: The two unnamed *Dehalobacter* species, CF and 11DCA, are clearly strains of the same species, sharing high nucleotide sequence identity across nearly their entire genomes (ANI/AF of 99.8%/0.93). There is very extensive synteny as well. *Dehalobacter restrictus* is closely related, possibly a distinct species with ~96%/.75 ANI/AF, but nucmer-based dot plot displays significant translocations, inversions and gaps.

(4) Generate a heatmap contrasting gene counts (using Tigrfam/Pfam) between all 5 genomes. What are the top contrasting functions between the Chlorflexi versus the Firmicute dehalorespirers? (Hint: Compare Genomes > Abundance Profile > Overview (All functions) > Normalize and Select Function type)

Abundance Profile Overview

Display Options:

Output Type Normalization Method [?](#)

Heat Map None
 Scale for genome size

OR

Matrix Gene count Include all rows, including those without hits
 Estimated gene copies [?](#) functions per page
(Slower)

Enter matching text for highlighting clusters/rows (E.g., "kinase")

Function:

- COG
- Enzyme
- KO
- Pfam
- TIGRfam

Genomes [1](#):

MER-FS Metagenome:

Please select 1 to 100 genomes.

Sequencing Status

Finished

Domain

Genome Cart

List Tree



Search for: <click show, then enter a genome name to search>

Abundance Profile Overview Results (Gene Count)

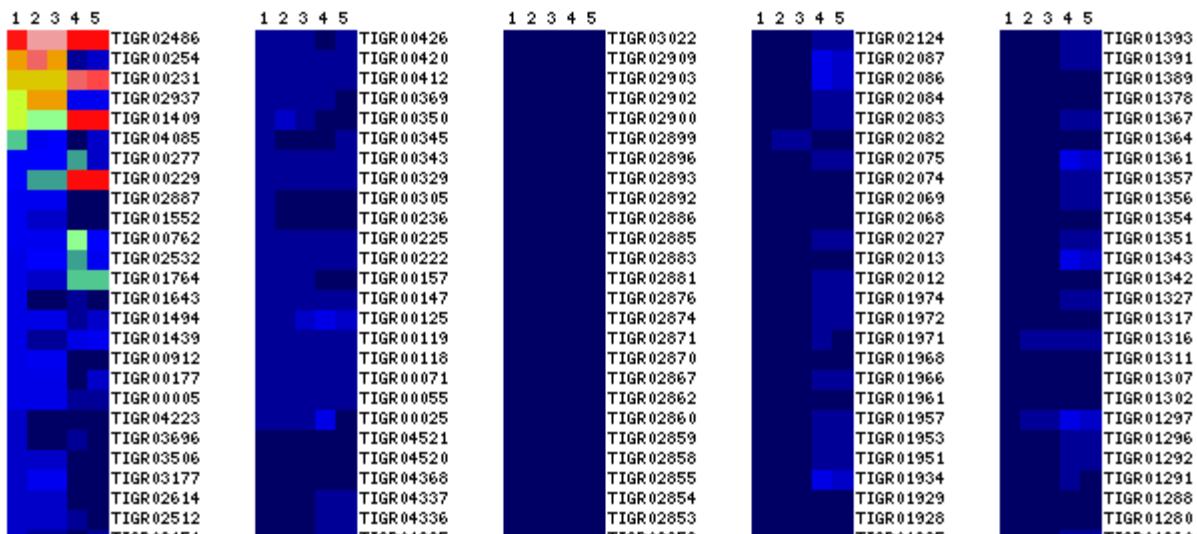
Mouse over labels to see additional information.

Clicking on the column number will sort rows for that column in descending gene count order.

Clicking on row cluster ID will add the cluster to the appropriate analysis cart (if cart is supported).

Mouse over heat map to see gene counts. Clicking in the heat map will take you to the gene list.

- 1 - [Dehalobacter restrictus DSM 9455](#)
- 2 - [Dehalobacter sp. 11DCA](#)
- 3 - [Dehalobacter sp. CF](#)
- 4 - [Dehalococcoides mccartyi 195](#)
- 5 - [Dehalogenimonas lykanthroporepellens BL-DC-9](#)



Result: Normalized counts reveal that certain sensory or regulatory functions (like diguanylate cyclases, sigma factors, PAS folds) are strikingly different. Other notable differences are a membrane lipid synthesis acyltransferase in the Chlorflexi (known to have unusual membrane lipids) or a group II intron reverse transcriptase in the *Dehalobacter* spp.

(5) As mentioned above, reductive dehalogenase function is dependent on a corrinoid cofactor. Synthesis of cobalamin cofactor is highly complex involving >30 enzymatic steps. **Are any of these dehalorespirers (5 genomes in your cart) capable of de novo cobalamin biosynthesis? Do they have partial pathways? Are they all auxotrophs? What is your conjecture based on**

your findings? What is the obvious difference between the Firmicutes and the Chloroflexi genomes? (Hint: Find Functions > KEGG > Pathways via KO terms > Select Porphyrin and chlorophyll metabolism under Metabolism of cofactors and vitamins > Switch Tab to “view map for selected genomes> View maps individually for each genome> *On the KEGG map, focus on the anaerobic portion of the pathway leading from precorrin to Vitamin B12*)

The screenshot shows the IMG/ER web interface with the 'Find Functions' tab selected. A message at the top left states: "Sat. June 11 to Sun. June 12 NERSC S... during this time. Sorry for the inconvenience." On the left, there's an 'IMG/ER Content' sidebar with various dataset counts. Below it are links for 'Project Map', 'Metagenome Projects Map', and 'System Requirements'. A 'Hands on training available at the...' link with a thumbnail image is also present. The main content area shows a hierarchical menu under 'Find Functions':

- Function Search
- Search Pathways
- Secondary Metabolism
- COG
- KOG
- Pfam
- TIGRfam
- KEGG
- IMG Networks
- Enzyme
- MetaCyc
- Biotrans
- Phenotypes
- InterPro Browser
- Protein Family Comparison

The 'KEGG' and 'Pathways via KO Terms' options are highlighted with dark blue boxes.

KEGG Orthology (KO) Terms and Pathways

[KEGG Orthology \(KO\) Terms](#) Based on [BRITE Hierarchy](#)
[KEGG Pathways via KO Terms](#)

KEGG Pathways via KO Terms

01 Metabolism

02 Global and overview maps

03 [Metabolic pathways](#) (2632)

03 [Biosynthesis of secondary metabolites](#) (949)

03 [Microbial metabolism in diverse environments](#) (978)

03 [Biosynthesis of antibiotics](#) (738)

03 [Carbon](#)

03 [2-Oxoc](#)

03 [Fatty ac](#)

03 [Biosynt](#)

03 [Degrad](#)

Scroll down

02 Metabolism of cofactors and vitamins

03 [Thiamine metabolism](#) (24)

03 [Riboflavin metabolism](#) (26)

03 [Vitamin B6 metabolism](#) (23)

03 [Nicotinate and nicotinamide metabolism](#) (61)

03 [Pantothenate and CoA biosynthesis](#) (36)

03 [Biotin metabolism](#) (23)

03 [Lipoic acid metabolism](#) (4)

03 [Folate biosynthesis](#) (47)

03 [One carbon pool by folate](#) (33)

03 [Retinol metabolism](#) (48)

03 [Porphyrin and chlorophyll metabolism](#) (107)

03 [Ubiquinone and other terpenoid-quinone biosynthesis](#) (55)

02 Metabolism of terpenoids and polyketides

03 [Terpenoid backbone biosynthesis](#) (53)

Tab over to "View Map for Selected Genomes from the KEGG Pathway details page:

KEGG Pathway Details

Details for Pathway: *Porphyrin and chlorophyll metabolism*

*Showing counts for genomes in genome cart only

[KO Terms in Pathway](#) [Save to My Workspace](#) [View Map for Selected Genomes](#)

KEGG Orthology (KO) Terms in Pathway

Add Selected to Function Cart Select All Clear All

Filter column: KO Term ID Filter text : Apply ?

Export Page 1 of 2 << first < prev 1 2 next > last >> 100

Column Selector Select Page Deselect Page

Select	KO Term ID	KO Name	Definition	KO Module ID	KO Module Name
<input type="checkbox"/>	KO:K00214	BLVRA, bvdR	biliverdin reductase [EC:1.3.1.24]		
<input type="checkbox"/>	KO:K00218	E1.3.1.33, por	protochlorophyllide reductase [EC:1.3.1.33]		

View map for all 5 genomes INDIVIDUALLY:

KEGG Pathway Details

Details for Pathway: *Porphyrin and chlorophyll metabolism*
*Showing counts for genomes in genome cart only

KO Terms in Pathway Save to My Workspace View Map for Selected Genomes

KEGG Map for Selected Genomes

Sequencing Status Domain

All Finished, Permanent Draft and Draft Genome Cart

List Tree Show ? Selected: 1

Search for: <enter a genome name to search>

Dehalobacter restrictus DSM 9455 (B) [F]
Dehalobacter sp. 11DCA (B) [F]
Dehalobacter sp. CF (B) [F]
Dehalococcoides mccartyi 195 (B) [F]
Dehalogenimonas lykanthroporepellens BL-DC-9 (B) [F]

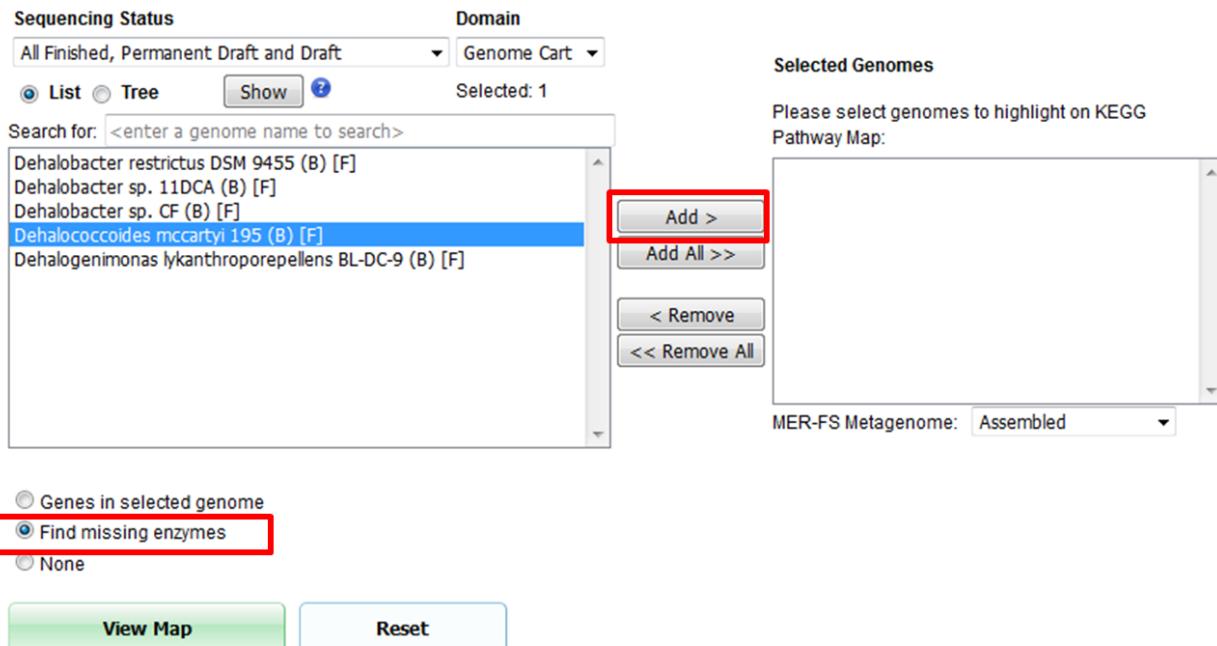
Add > Add All >>
< Remove << Remove All

Please select genomes to highlight on KEGG Pathway Map:

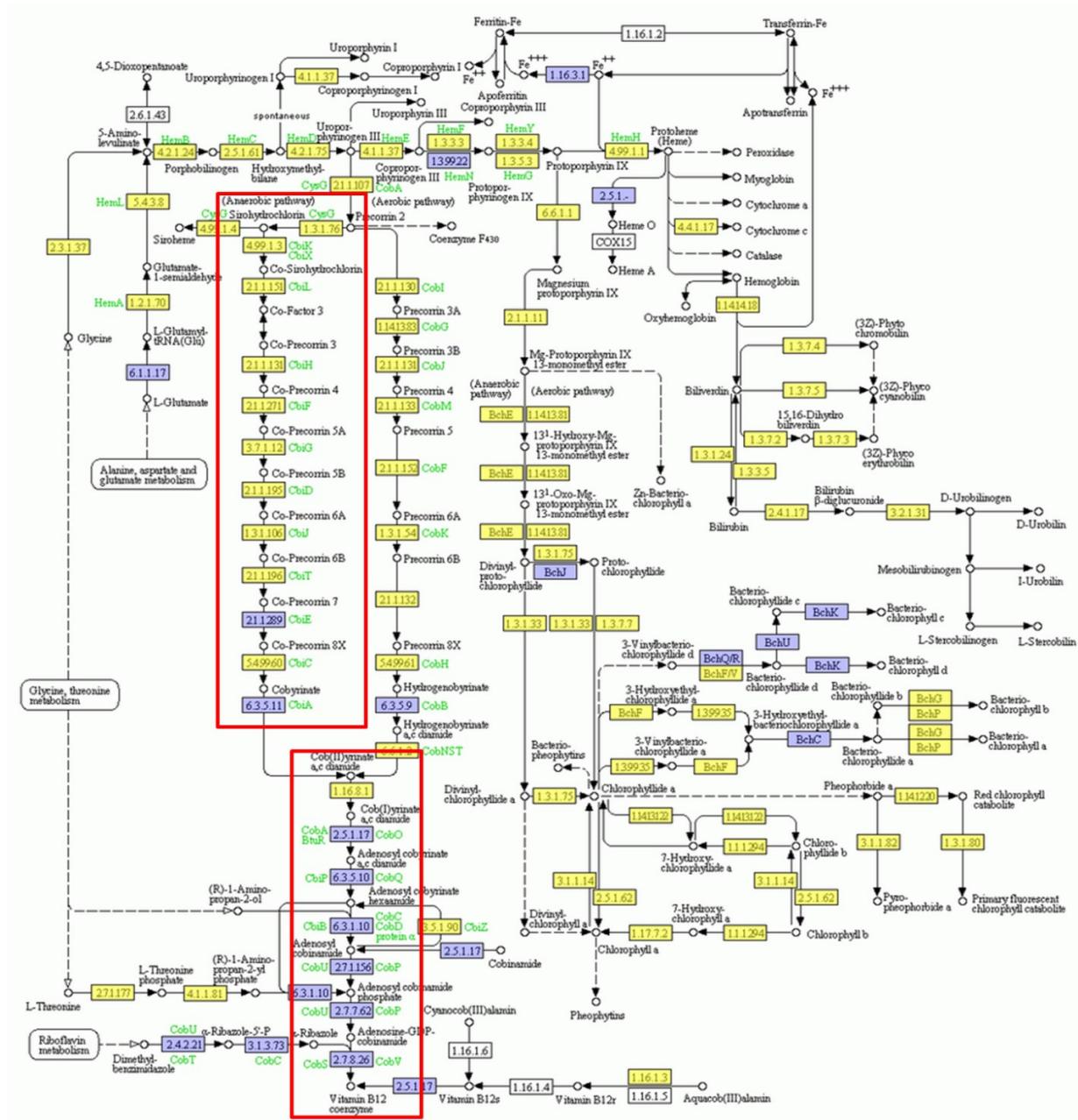
MER-FS Metagenome: Assembled

Genes in selected genome
 Find missing enzymes
 None

View Map **Reset**



On the KEGG map displayed for "Porphyrin and chlorophyll metabolism, examine the "Anaerobic Pathway" leading from Precorrin to Vitamin B12:



Result: Chlorflexi genomes may salvage B12 from a cobinamide intermediate based on the presence of the lower half of the pathway. By contrast, the Firmicutes *Dehalobacter* genomes have a near-complete pathway, but some intermediate steps are not identified – this would require further investigation into whether CDS were not correctly identified or annotated before any conclusions are drawn. Alternately, non-canonical enzymes not identified by existing conserved gene models may be involved. Given the results as is, none of the 5 strains are capable of de novo B12 synthesis.