IMG Hands-On EXERCISES

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

Rekha Seshadri, DoE Joint Genome Institute, rseshadri@lbl.gov

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

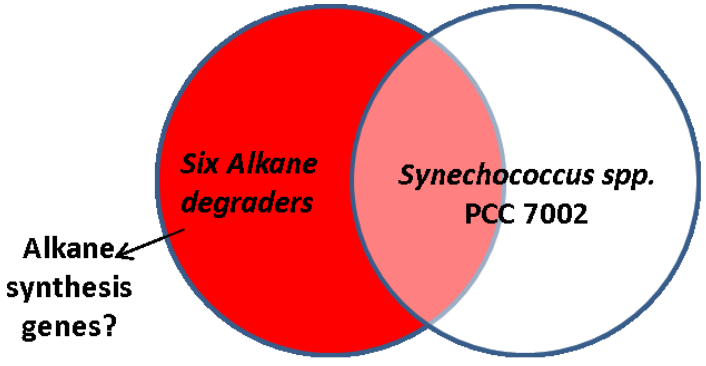
EXERCISE 1

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
Synechococcus elongatus PCC7942	2.7	Heptadecane, pentadecane
Prochlorococcus marinus pastoris CCMP 1986	1.7	Pentadecane
Anabaena variabilis ATCC 29413	6.4	Heptadecane
Gloeobacter violaceus PCC 7421	4.6	Heptadecane
Nostoc sp. PCC 7120	6.2	Heptadecane
Cyanothece sp. PCC 7425	5.7	Heptadecane
Synechococcus sp. PCC 7002	**3.0**	**No alkanes observed**

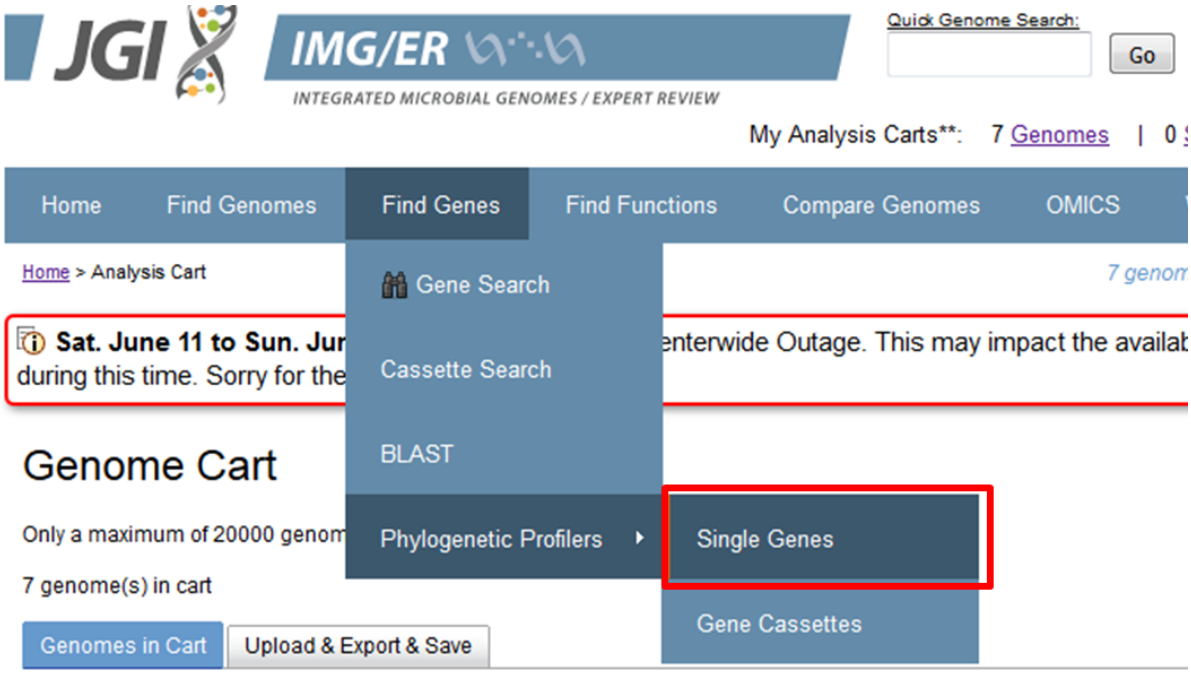
Table adapted from Schirmer, A. et. al. ,Microbial biosynthesis of alkanes" Science 2010, 329(5991):559-562)

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

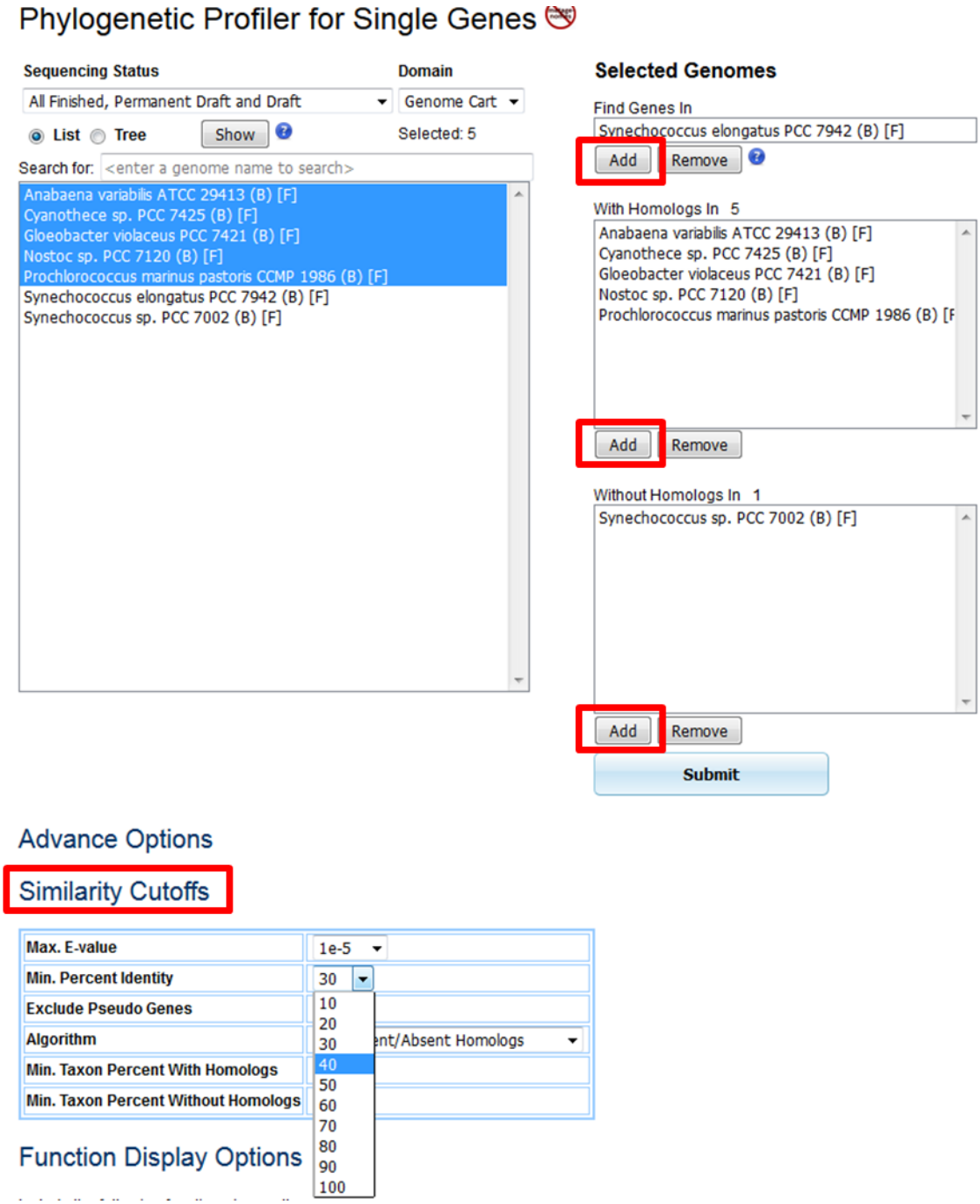


Workflow: *Add the above 7 genomes to your genome cart using the "Quick Genome Search" box at the top of the page.*

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



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.*



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.

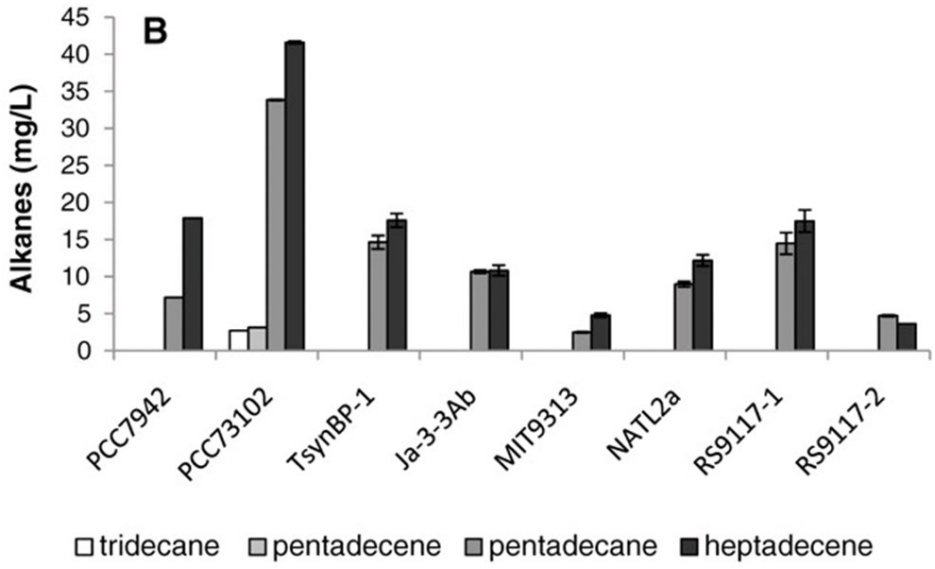


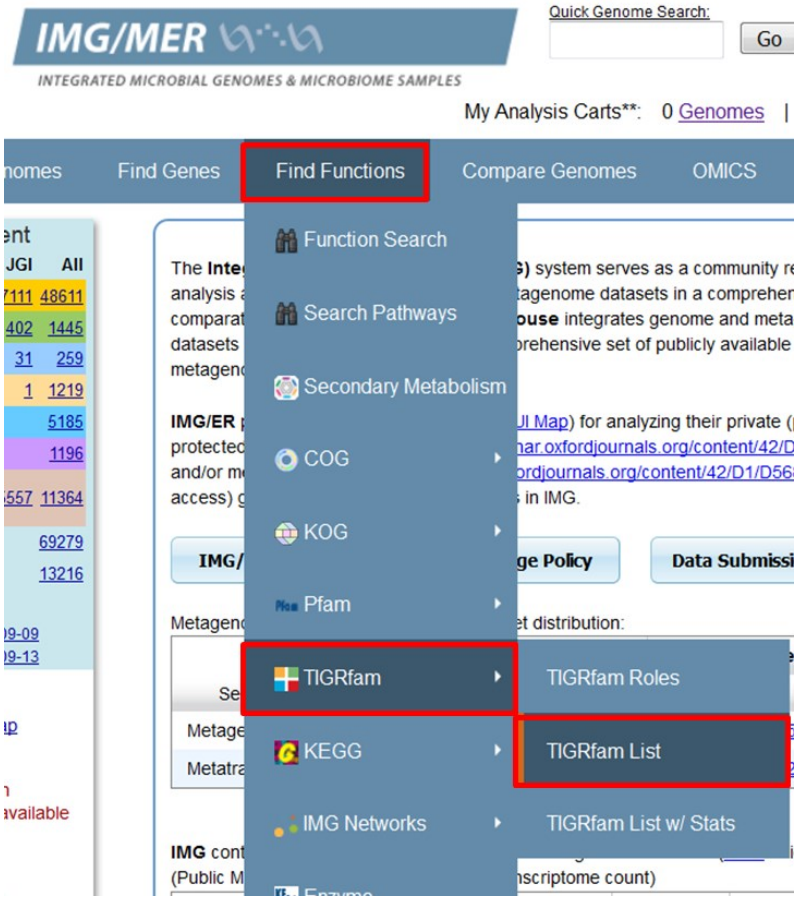
Image from Schirmer, A. et. al., Microbial biosynthesis of alkanes" Science 2010, 329(5991):559-562)

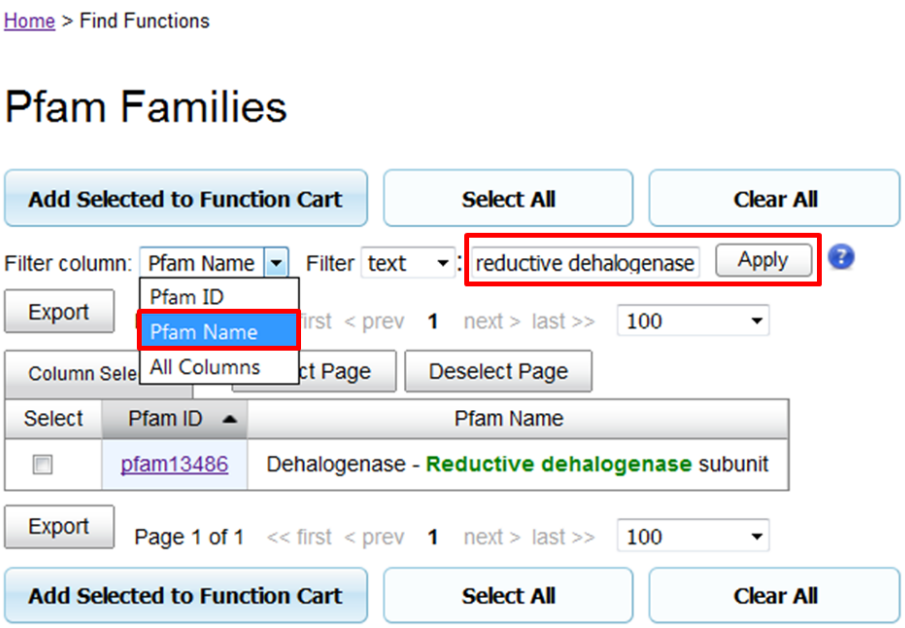
BONSUS QUESTION: Examine various annotations on the gene details pages for both these candidates (Synpcc7942_1593 and Synpcc7942_1594). Are the various annotation evidences largely in agreement? Are some annotations less informative? Are some function annotations absent? Are some function annotations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

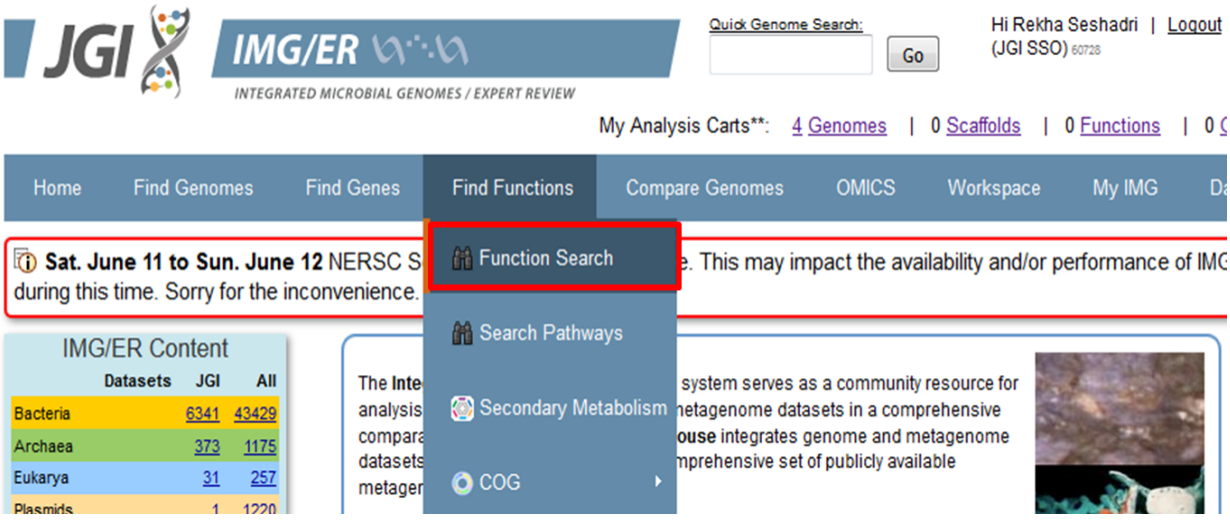
(1) **What is the TIGRFAM ID and/or Pfam ID for "reductive dehalogenase"** *(Hint: Use "Find Functions > Tigrfam > Tigrfam list", and use the filter column feature to find the appropriate ID)*

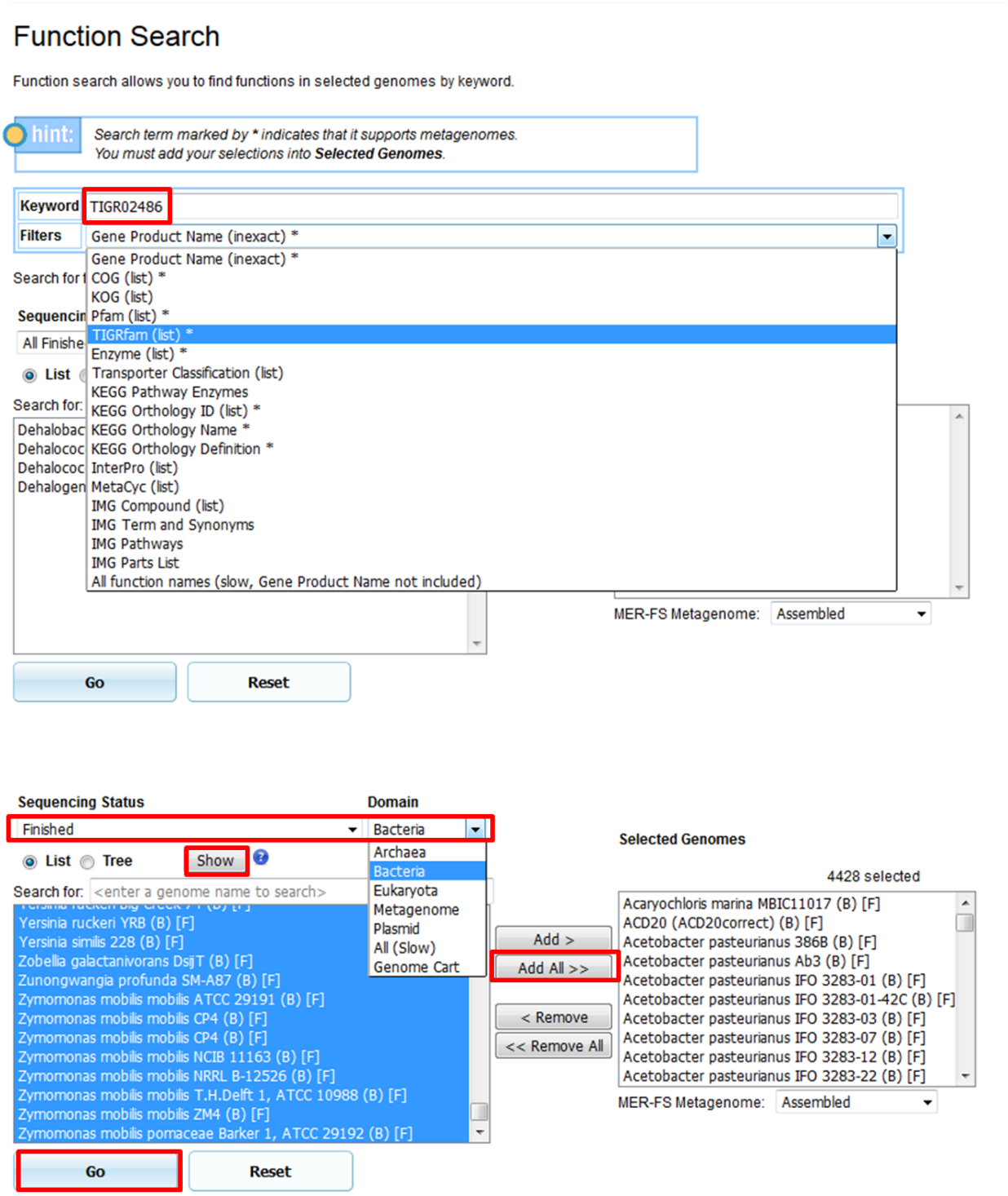




Result: pfam13486 or TIGR02486

(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)*



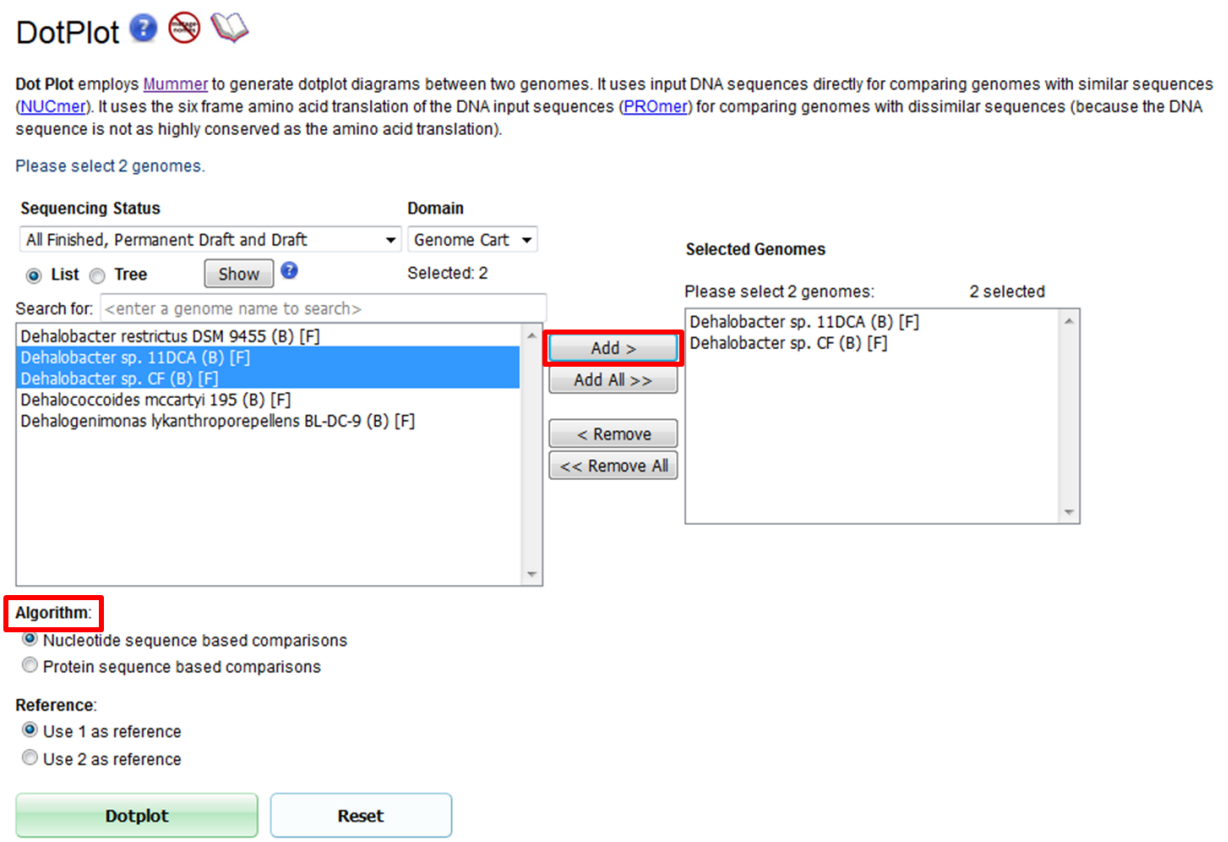


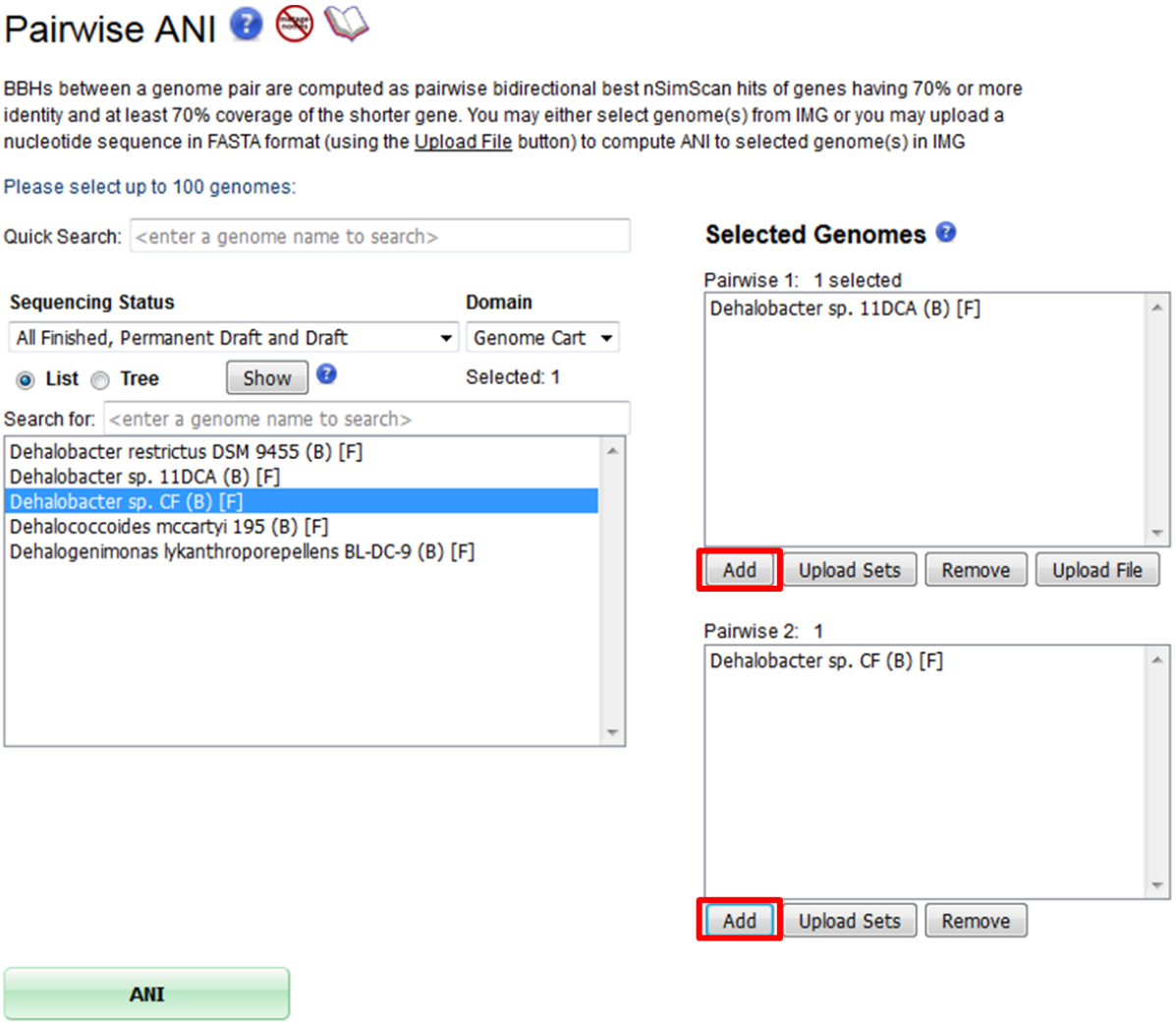
Result: 41 genomes, 420 genes *(this result is for Tigrfam). BONUS QUESTION: do you get a different result if you used Pfam, why do you think that might be?)*

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

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





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** [3]

- (•) Heat Map
  - ( ) None
  - (•) Scale for genome size

OR

- ( ) Matrix
  - (•) Gene count
  - ( ) Estimated gene copies [2] *(Slower)*
  - [ ] Include all rows, including those without hits
  - 500 functions per page

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

Function:

- ( ) COG
- ( ) Enzyme
- ( ) KO
- ( ) Pfam
- (•) TIGRfam

Genomes [1]:

MER-FS Metagenome: Assembled

Please select 1 to 100 genomes.

Sequencing Status **Domain**

Finished | Genome Cart

(•) **List** ( ) **Tree** Show

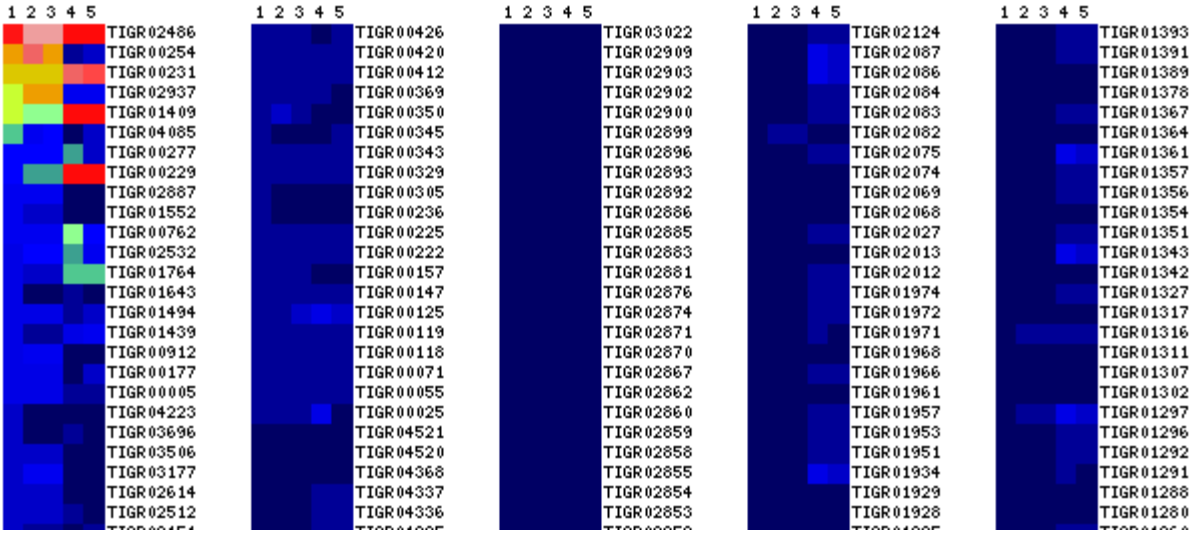
Search for: <click show, then enter a genome name to search>

Go

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.

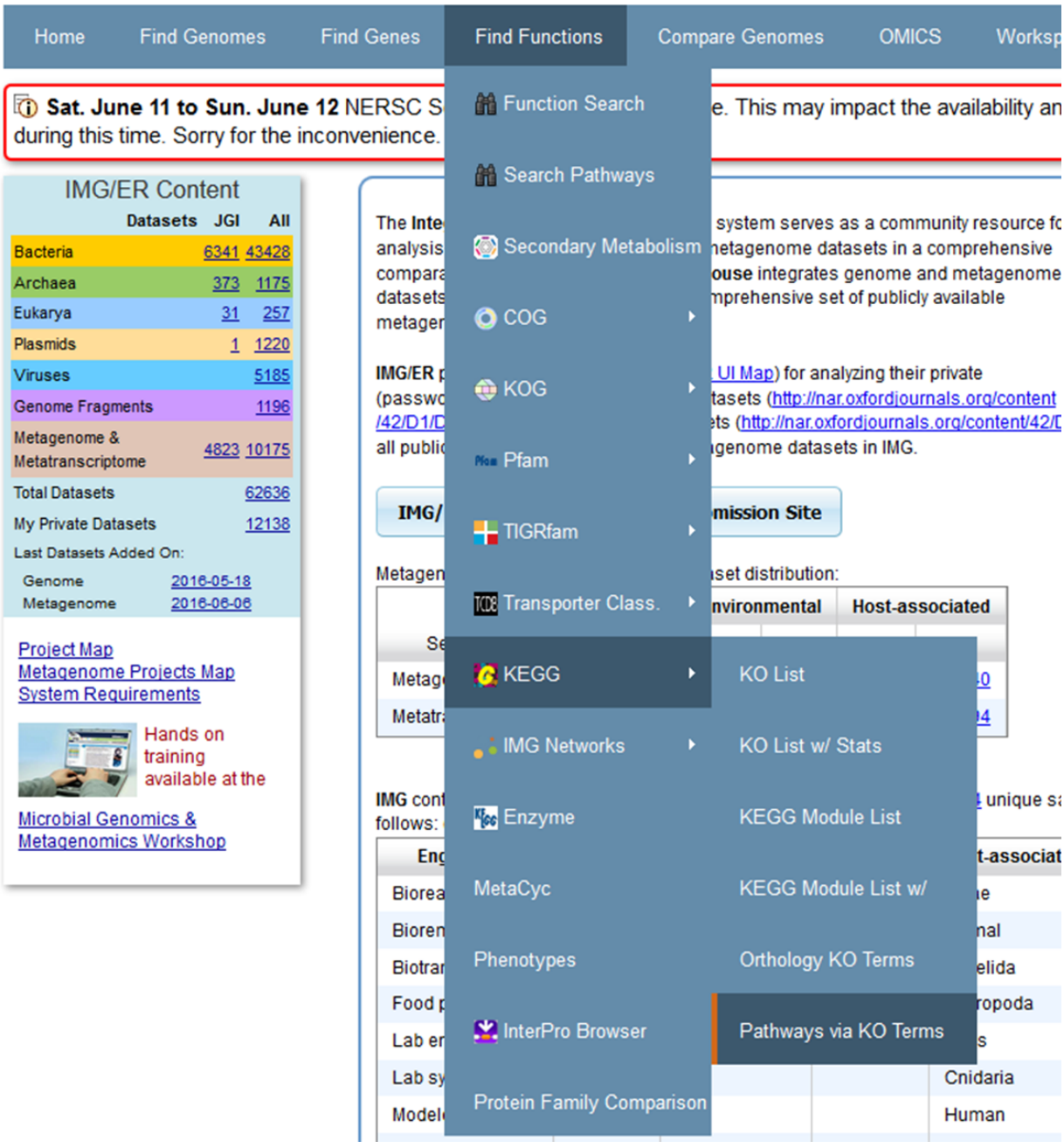
1 - Dehalobacter restrictus DSM 9455
2 - Dehalobacter sp. 11DCA
3 - Dehalobacter sp. CF
4 - Dehalococcoides mccartyi 195
5 - Dehalogenimonas lykanthroporepellens BL-DC-9

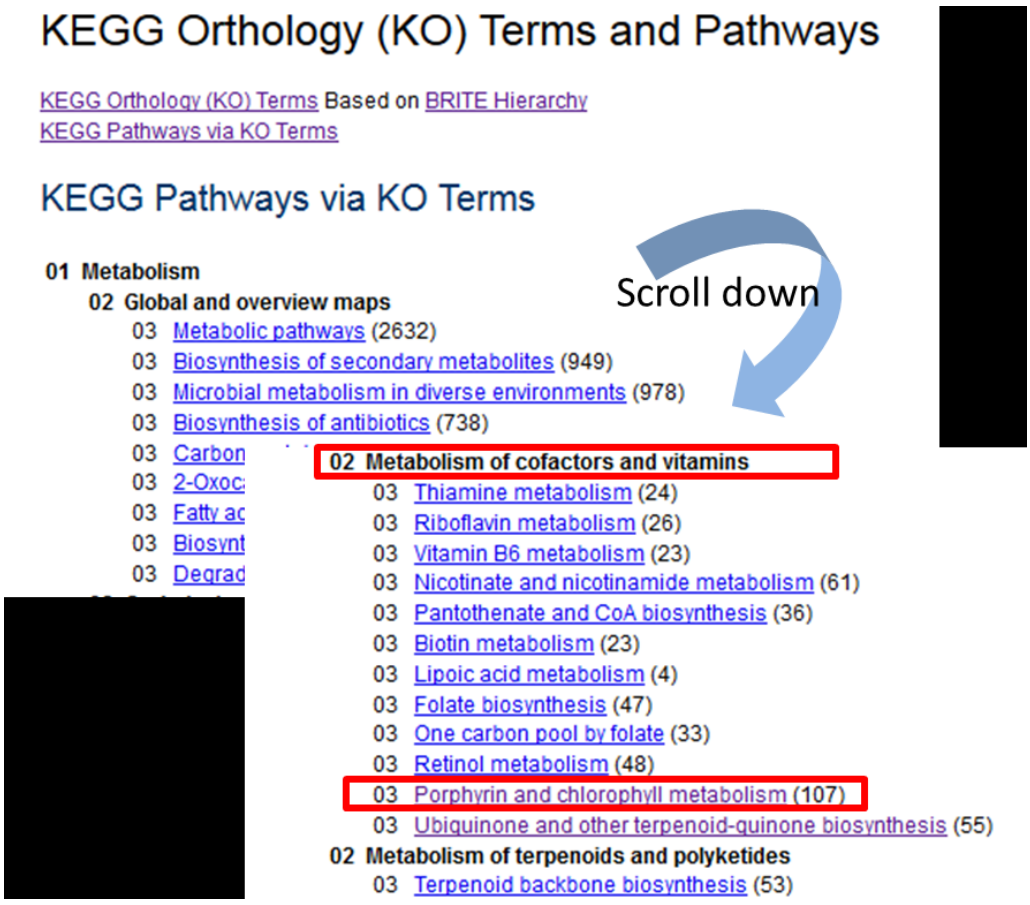


Result: Normalized counts reveal that certain sensory or regulatory functions (like diguanylate cyclases, sigma factors, PAS folds) are strikingly different. Other notable differences are a membrane lipid synthesis acyltransferase in the Chlorflexi (known to have unusual membrane lipids) or a group II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

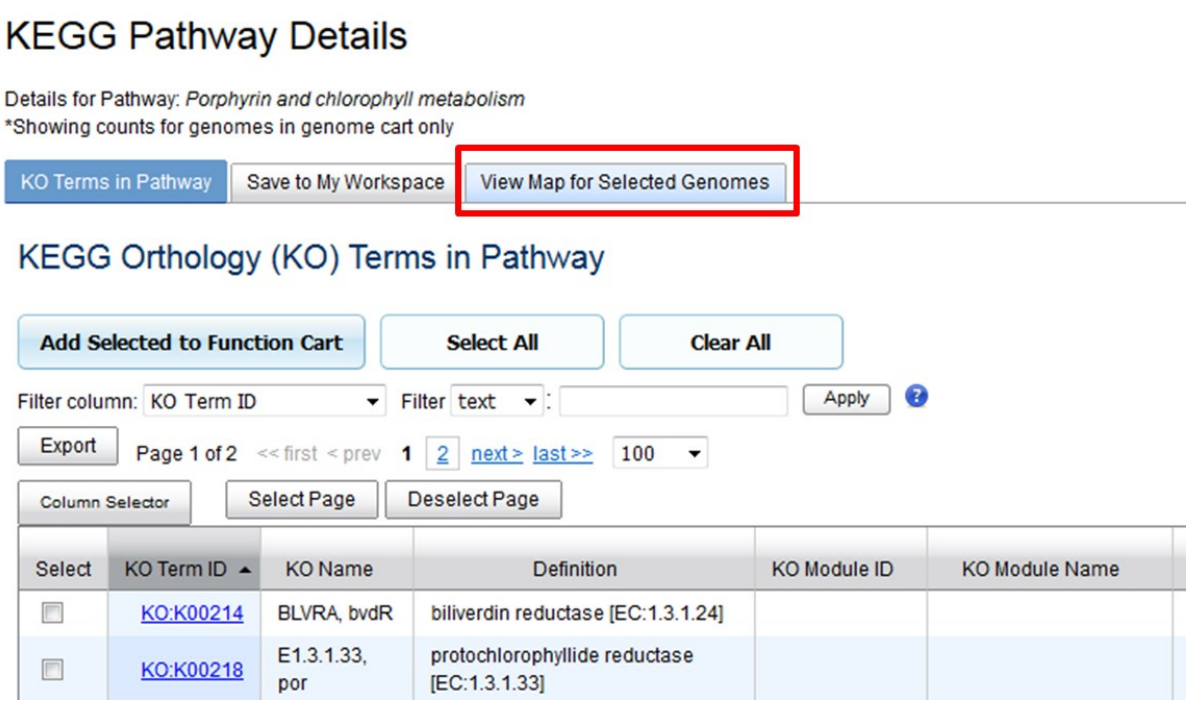
(5) As mentioned above, reductive dehalogenase function is dependent on a corrinoid cofactor. Synthesis of cobalamin cofactor is highly complex involving >30 enzymatic steps. **Are any of these dehalorespirers (5 genomes in your cart) capable of de novo cobalamin biosynthesis? Do they have partial pathways?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

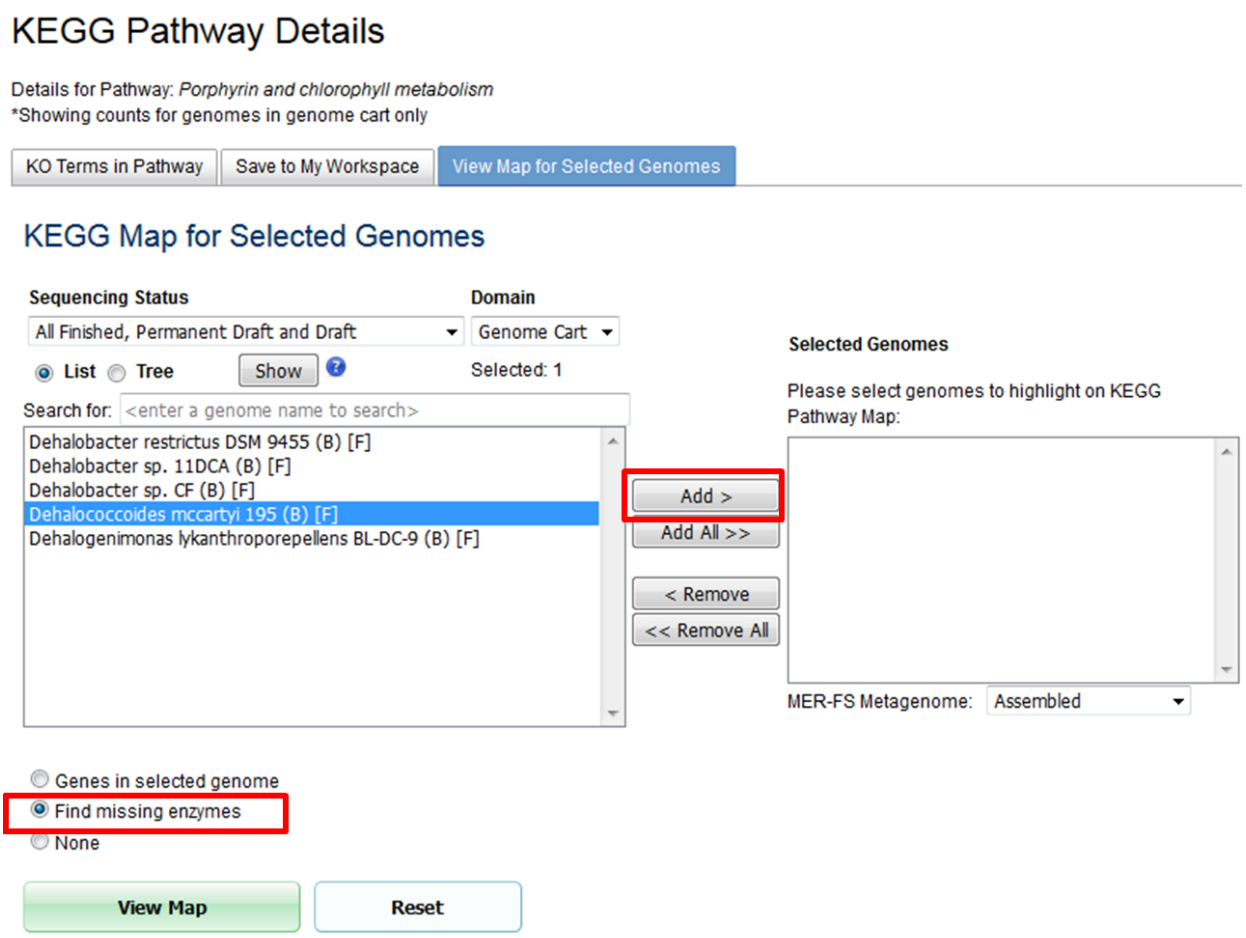




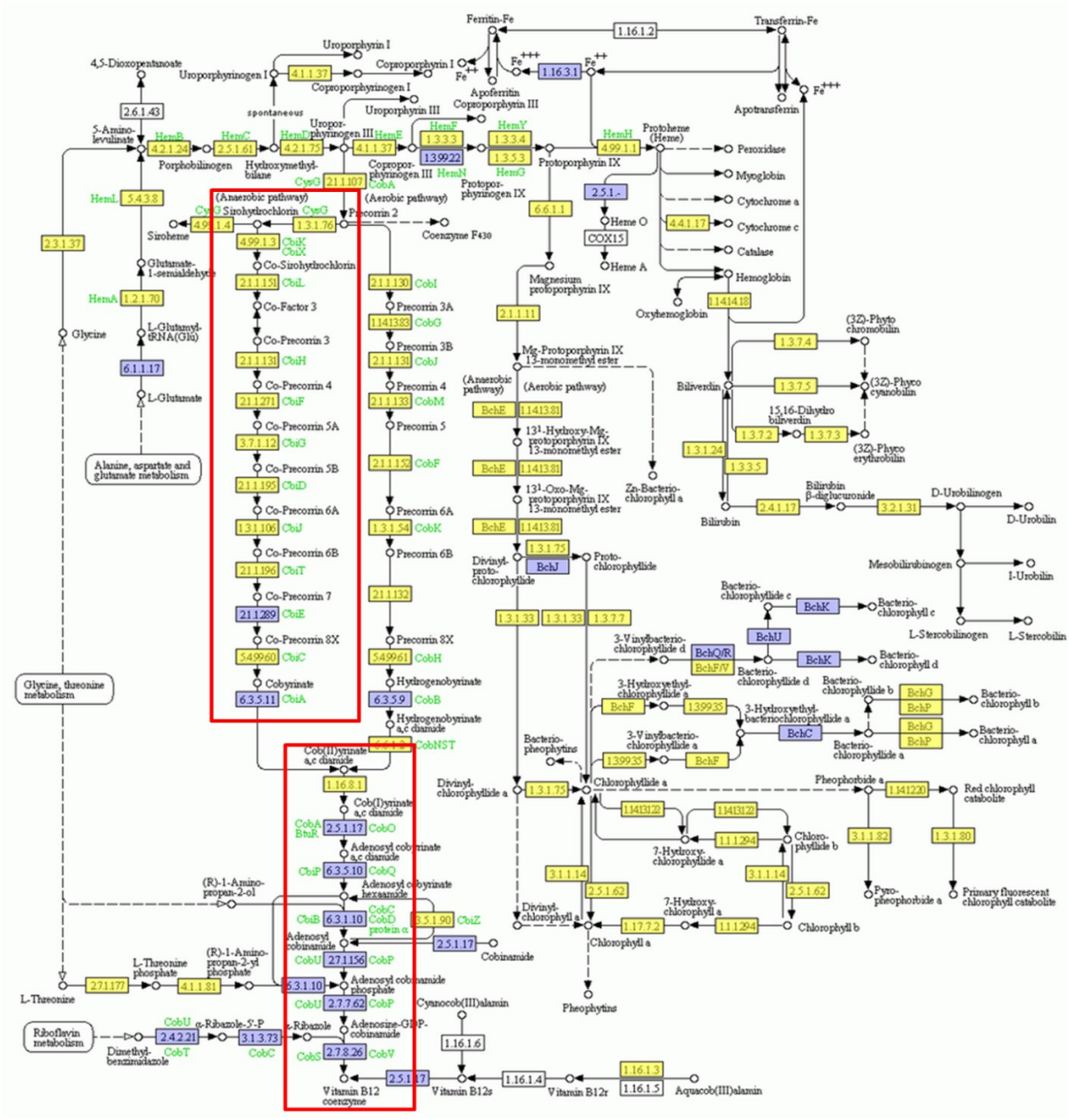
Tab over to "View Map for Selected Genomes from the KEGG Pathway details page:



View map for all 5 genomes INDIVIDUALLY:



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.