Bigelow | Laboratory for Ocean Sciences

DRAM annotation outputs

with further caveats!

Melody Lindsay & Maria Pachiadaki

SCGC Course Day 2 - PM

DRAM outputs

Now:

Annotation summaries

Let's go through the files you generated from the "annotate" and "distill" DRAM commands

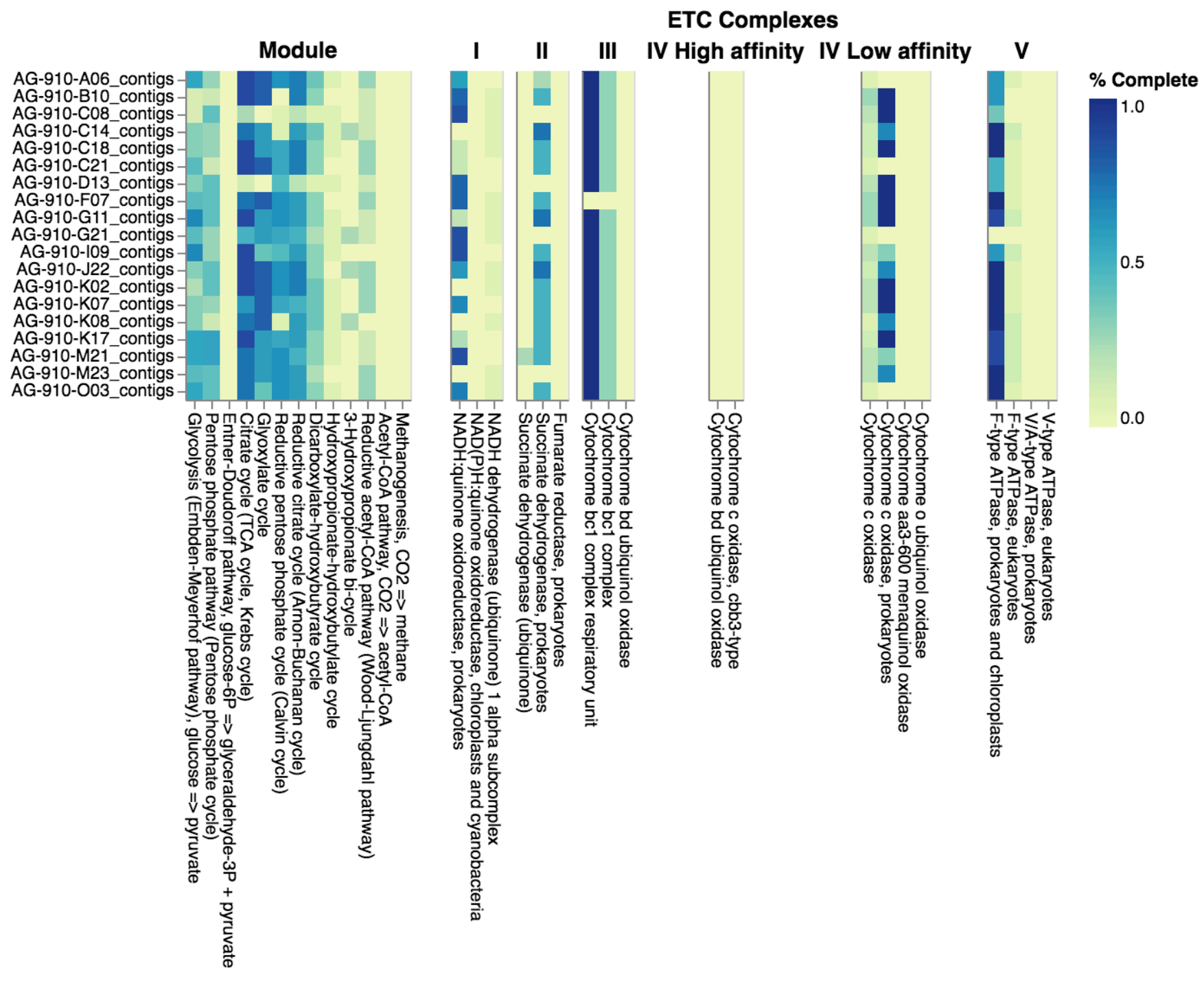
What are they, how can they be used? What will be most useful to you?

Then:

Some examples of misannotations, and ways to avoid

(not covering "missed" annotations, or proteins of unknown function)

Work on examples of DRAM output, manipulate and look through annotations



- One example of visualization (made with DRAM in the “Distill” command) – we won’t do a whole lot with manipulating this data, but will show you how to access what goes into a dataset like this!

SCFA and alcohol conversions
Alcohol production
Propionate, pt 2
Propionate, pt 1
Butyrate, pt 2
Butyrate, pt 1
lactate D
lactate L
acetate, pt 3
acetate, pt 2
acetate, pt 1
pyruvate => acetylCoA f+ formate v3
pyruvate => acetyl CoA v2
pyruvate => acetyl CoA v1
Methanogenesis and methanotrophy
methane => methanol, with oxygen (mmo)
methane => methanol, with oxygen (pmo)
putative but not defining CO2 => methane
monomethylamine => ammonia
dimethylamine => monomethylamine
trimethylamine => dimethylamine
methanol => methane
acetate => methane, pt 3
acetate => methane, pt 2
acetate => methane, pt 1
Key functional gene
Photosynthesis
Photosystem II
Photosystem I
Other Reductases
arsenate reduction, pt 2
arsenate reduction, pt 1
mercury reduction
TMAO reductase
selenate/Chlorate reduction
Sulfur metabolism
tetrathionate => thiosulfate
Thiosulfate oxidation by SOX complex, thiosulfate => sulfate
thiosulfate => sulfite
dissimilatory sulfate reduction (and oxidation) sulfate => sulfide
Nitrogen metabolism
Nitrogen fixation altennative
nitrogen => ammonia
nitrous oxide => nitrogen
nitric oxide => nitrous oxide
Dissimilatory nitrite reduction to ammonia (DNRA)
nitrite => nitric oxide
nitrate => nitrite
nitrite => nitrate
Bacterial (anaerobic-specific) ammonia oxidation
Bacterial (aerobic-specific) ammonia oxidation
Bacterial/Archaeal ammonia oxidation
ammonia => nitrite
CAZy
Arabinose cleavage
Fucose Cleavage
Rhamnose cleavage
Arabinan
Alpha-mannan
Beta-mannan
Xylans
Mixed-Linkage glucans
Beta-galactan (pectic galactan)
Alpha-galactans
Pectin
Chitin
Starch
Xyloglucan
Sulf-Polysachharides
Mucin
Amorphous Cellulose
Crystalline Cellulose
Polyphenolics
AG-910-A06_contigs
AG-910-B10_contigs
AG-910-C08_contigs
AG-910-C14_contigs
AG-910-C18_contigs
AG-910-C21_contigs
AG-910-D13_contigs
AG-910-F07_contigs
AG-910-G11_contigs
AG-910-G21_contigs
AG-910-I09_contigs
AG-910-J22_contigs
AG-910-K02_contigs
AG-910-K07_contigs
AG-910-K08_contigs
AG-910-K17_contigs
AG-910-M21_contigs
AG-910-M23_contigs
AG-910-O03_contigs
Function is Present
true
false

Misannotations, or genes involved in multiple pathways (ascribed to other functions)

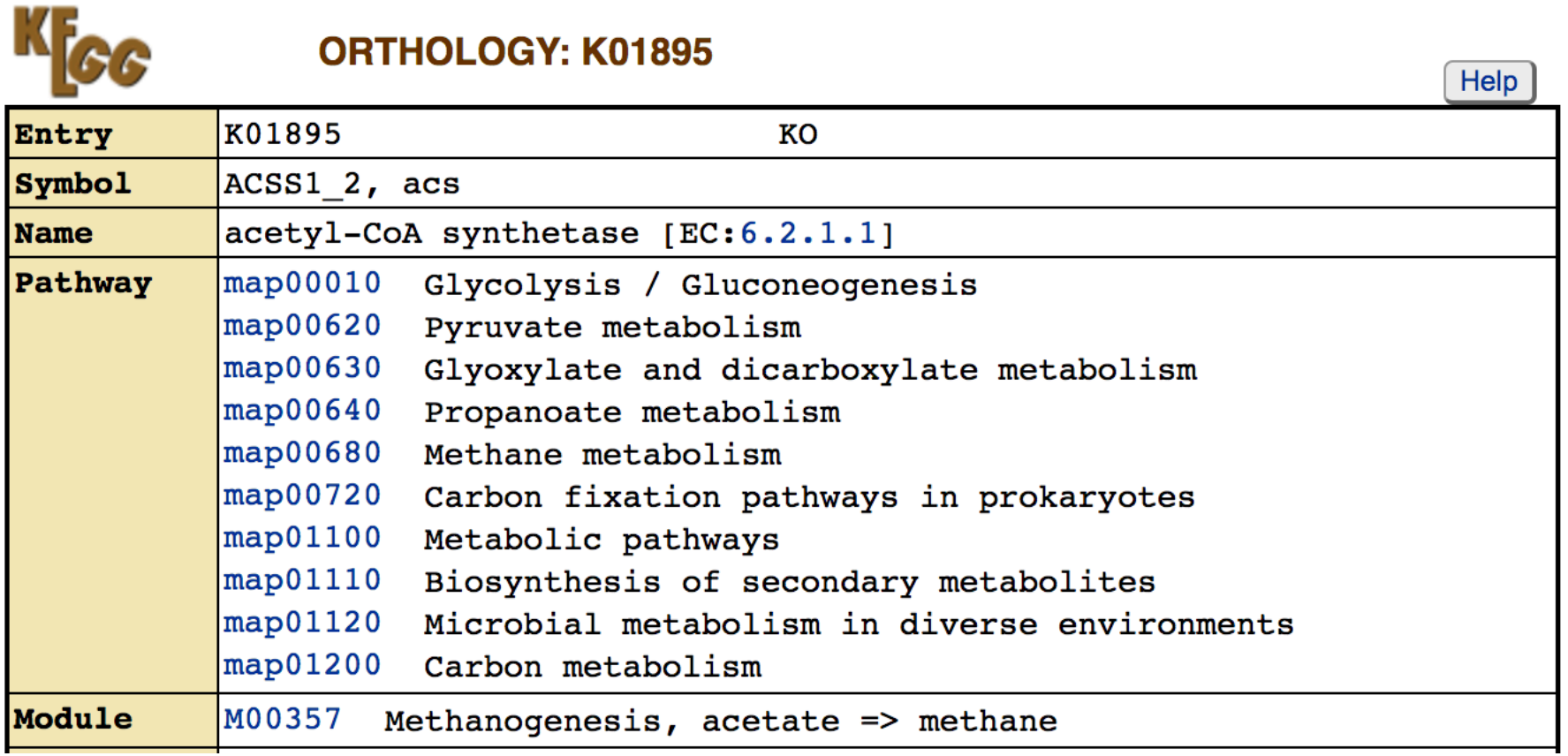
- Rampant in databases, particularly as larger and larger sequence datasets are deposited using automated annotation programs
- How would misannotations become “the norm” for some genes?

- Do you know of any common misannotations in data from organisms you work with/might work with in the future?
- Common genes?
- Some manual curation may be needed in some datasets

Notice anything suspicious?

How would you go about verifying? (check which file??)

Check databases (KEGG)



ORTHOLOGY: K01895

Help

Entry	K01895 KO
Symbol	ACSS1_2, acs
Name	acetyl-CoA synthetase [EC:6.2.1.1]
Pathway	map00010 Glycolysis / Gluconeogenesis map00620 Pyruvate metabolism map00630 Glyoxylate and dicarboxylate metabolism map00640 Propanoate metabolism map00680 Methane metabolism map00720 Carbon fixation pathways in prokaryotes map01100 Metabolic pathways map01110 Biosynthesis of secondary metabolites map01120 Microbial metabolism in diverse environments map01200 Carbon metabolism
Module	M00357 Methanogenesis, acetate => methane

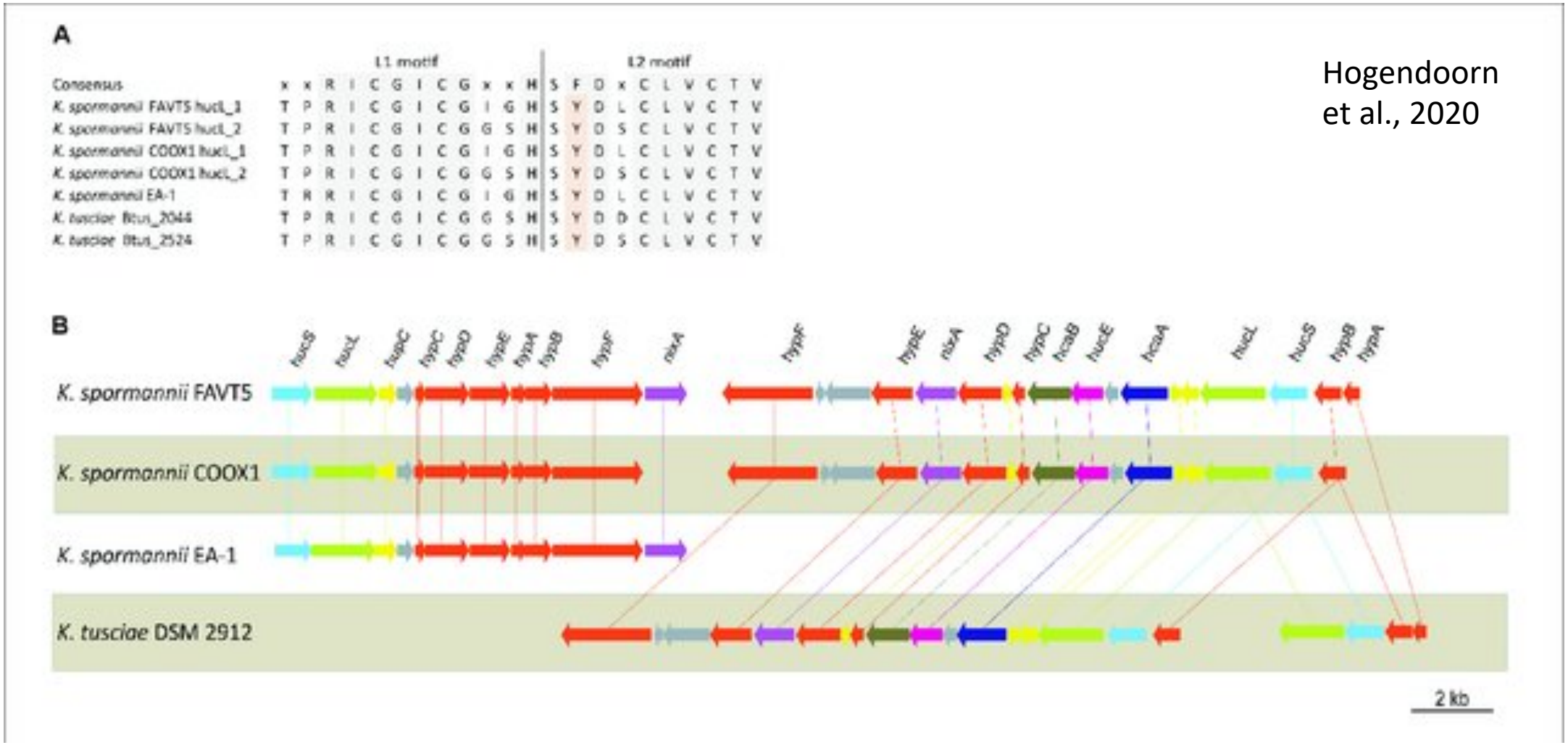
KEGG database

Other ways to check for potential misannotations? Or divergence?

- Active sites for enzymes (do they possess key motifs)?
  - Example: cytochromes
  - Example: hydrogenases

- Specific databases of your choice
  - Examples: hydrogenases/HydDB, cytochrome P450 database, caZY,

- Do you have any commonly misannotated genes that you know of? Common in environments/samples/microbes that you study?



Hogendoorn et al., 2020

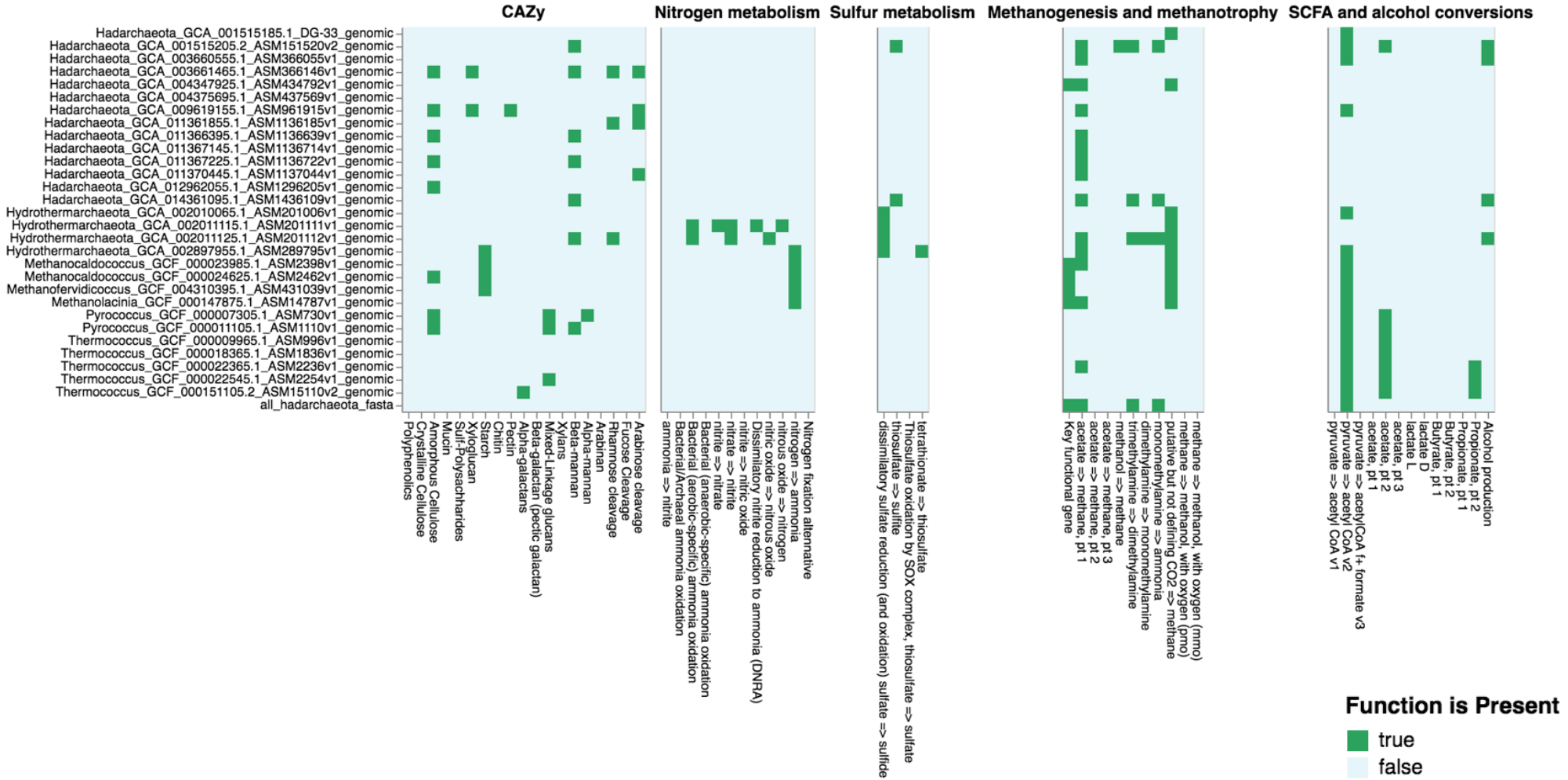
- Example: Cysteine residues key for binding site of hydrogenases (for [NiFe]-hydrogenases, need paired CxxC motifs)

Other Misannotations?

- Carbon fixation – TCA cycle, ATP citrate lyase involved in carbon fixation (appear to be TCA cycle, unless you compare with another ATP citrate lyase). **Challenge: how would you do this?**

- Arsenic reduction misannotation – one of the genes that is a regulation factor (actually in multiple pathways, need to verify that other genes are present in the arsenic reduction pathway!). **Challenge: how would you do this?**

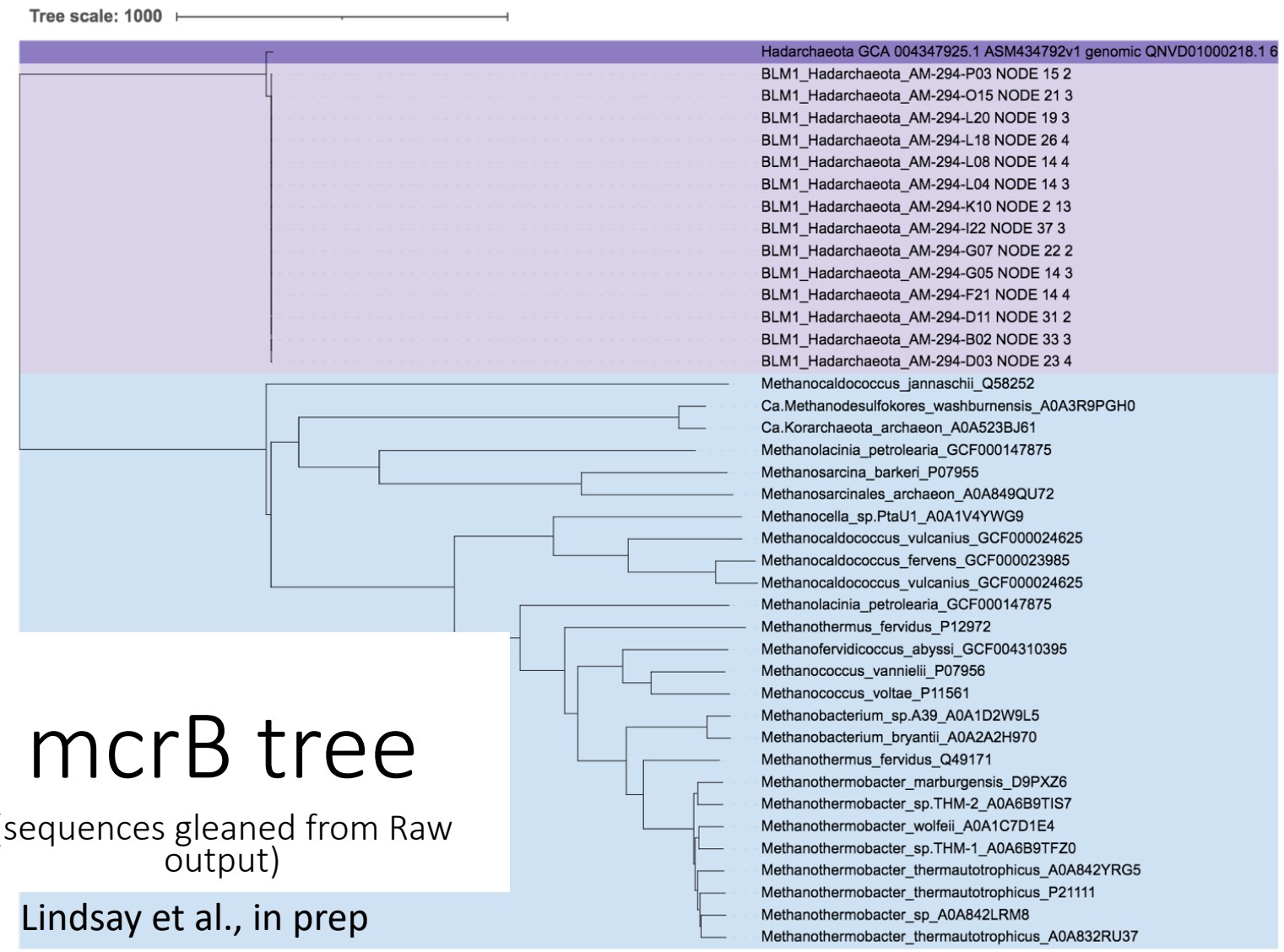
Other ways to check for potential misannotations? Or divergence?



Lindsay et al., in prep

	A	B	C	D	G	H	I	J	K	L	M
1	gene_id	gene_description	module	header	294-B02_co	294-C15_co	294-C19_co	294-C23_co	294-D03_co	294-D11_co	294-D14_co
174	K00320	coenzyme F420-dependent N5,N10-methenyltetrahydromethar	Methanogenesis, CO2 => methane	C1-methane	0	0	0	0	0	0	0
175	K00399	methyl-coenzyme M reductase [EC:2.8.4.1] [RN:R04541]	Methanogenesis, CO2 => methane	C1-methane	0	0	0	0	0	0	0
176	K00401	methyl-coenzyme M reductase [EC:2.8.4.1] [RN:R04541]	Methanogenesis, CO2 => methane	C1-methane	1	0	0	0	1	1	0
177	K00402	methyl-coenzyme M reductase [EC:2.8.4.1] [RN:R04541]	Methanogenesis, CO2 => methane	C1-methane	0	0	0				
178	K00577	tetrahydromethanopterin S-methyltransferase [EC:2.1.1.86] [RN	Methanogenesis, CO2 => methane	C1-methane	0	0	0				

Distillate output



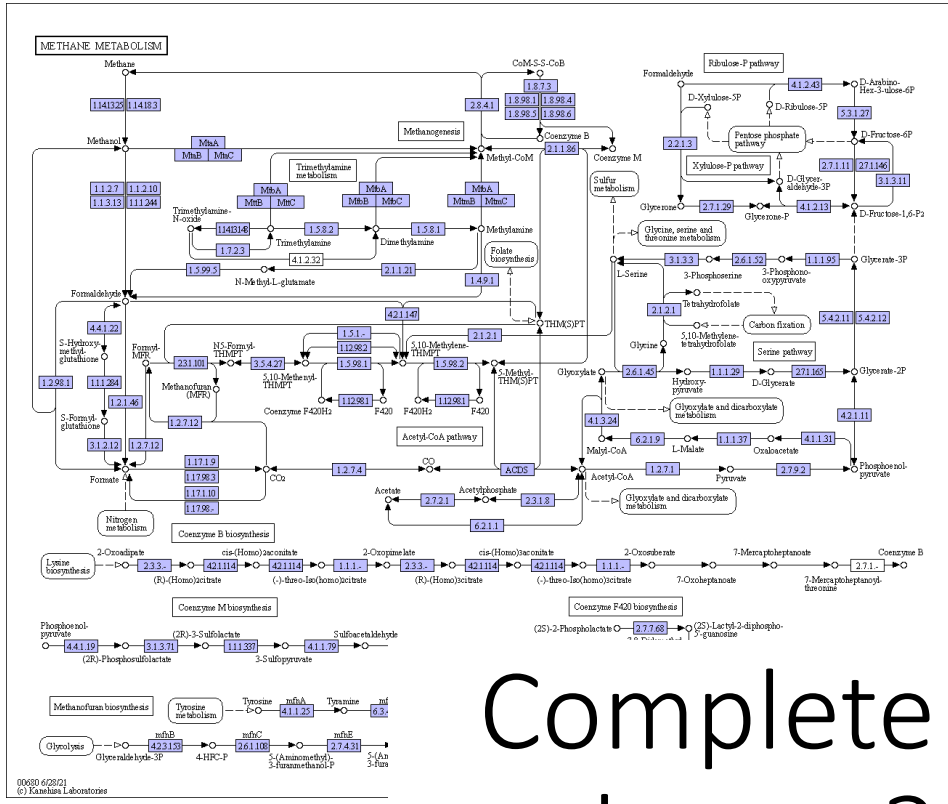
mcrB tree

(sequences gleaned from Raw output)

Lindsay et al., in prep

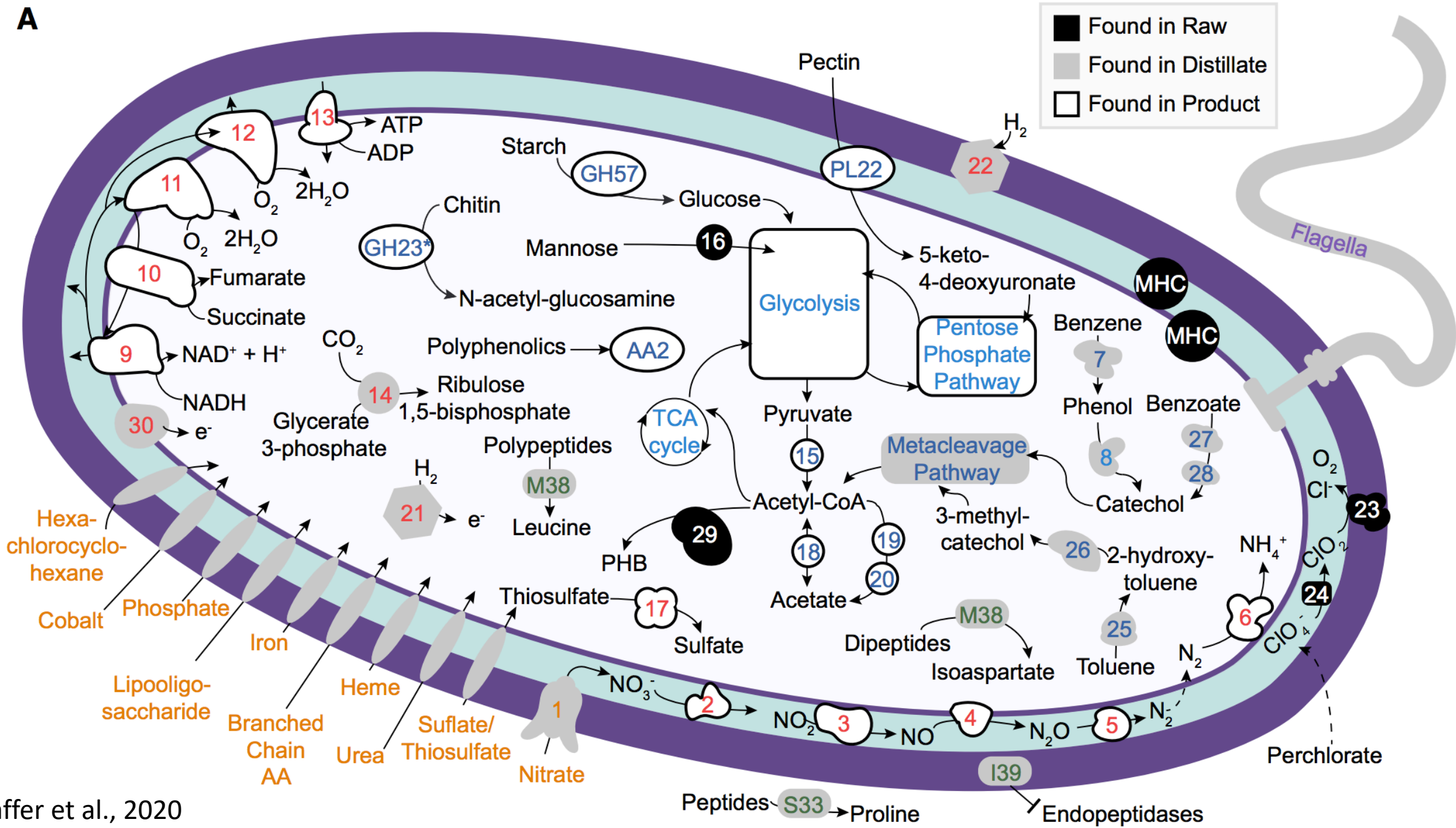
KAAS - KEGG Automatic Annotation Server

for ortholog assignment and pathway mapping



Genome.jp

Complete pathways?



Scaffer et al., 2020

- What other metabolism summaries are you interested in? Try this command with your broad category of choice (as seen in product.html)
- As a reminder - you will need to parse out the excel file sheets again, in case you want to look at different broad categories (other than energy)!

Other exploratory ideas:

How complete are individual pathways in this dataset? (as annotated by the DRAM pipeline?)

1) how complete are cytochrome bc complexes from Pelagibacter SAGs? (where can you find this data?)

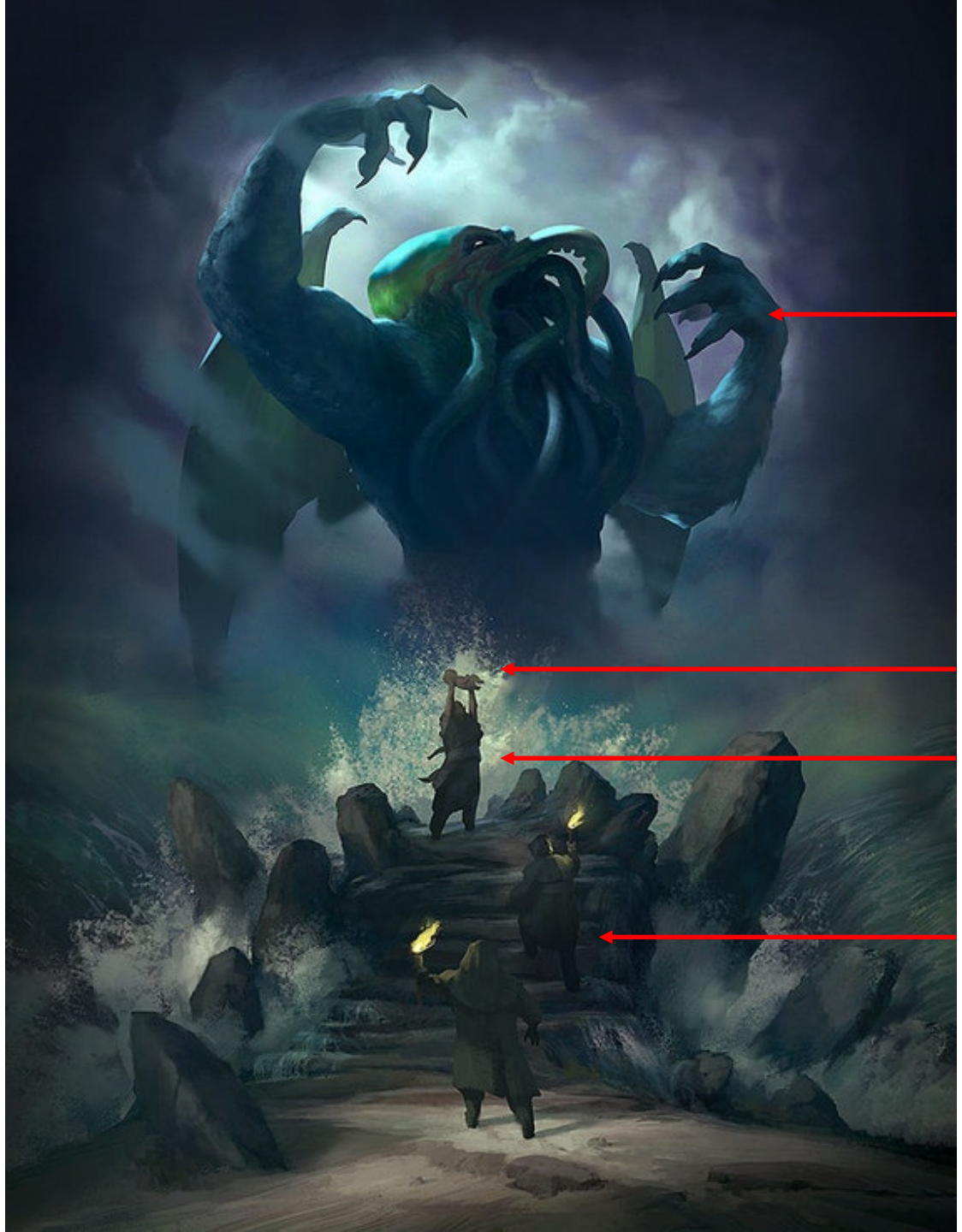
2) how complete are key carbon cycling pathways?

We also have all concatenated SAGs under /storage/lesson_analyses/annotations/AG-910. Can you use the distill command (from DRAM github) to make summary annotations.tsv, product.html & .tsv, metabolism.xlsx, of your choice of taxa?

1) Flavobacteriales?

2) Rhodobacterales?

Keep in mind that DRAM also works with MAGs and other data types.



Functional annotations/metabolic mapping of a confusing archaea? (which is not a methanogen!)

My sanity

Me

My postdoc advisors

https://www.etsy.com/ie/listing/653073086/lovecraft-cthulhu-poster-cthulhu-art