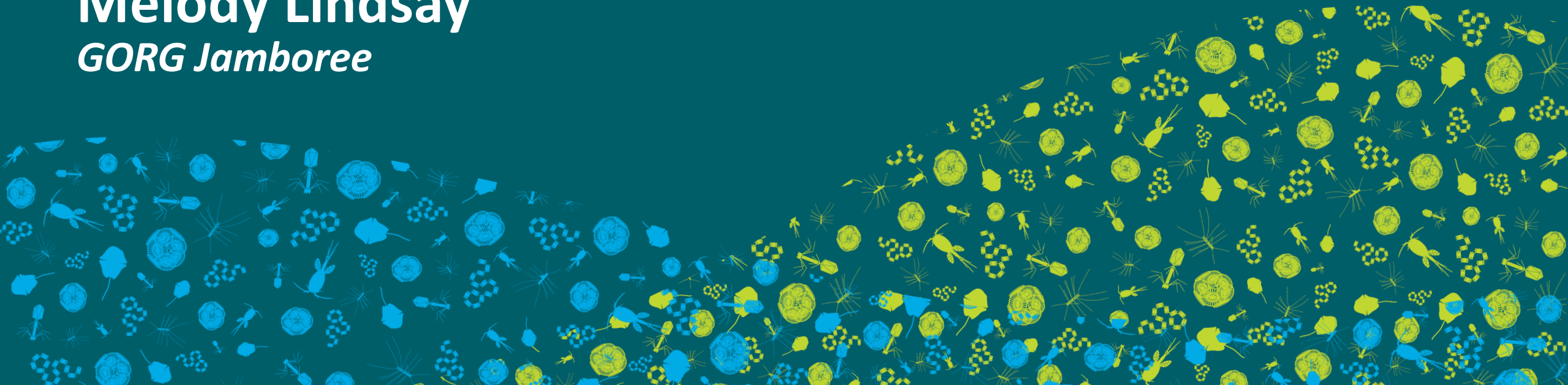
Bigelow | Laboratory for Ocean Sciences

Genome Annotation

"from sequence to biology"

Melody Lindsay

GORG Jamboree



Annotation

The process of identifying and associated function to various segments of the assembled DNA sequence. It consists of 3 main stages:

- Gene prediction
- Gene identification (protein coding, ribosomal RNA genes, transfer RNA genes)
- Identification of other genomic elements (such as: promoters, enhancers, transcription factor binding sites)

Some commonly used annotation pipelines include Prokka (Seeman, 2014) and DRAM (Shaffer et al., 2020)

And now: GORG Classifier! Incorporates annotations of database SAGs!

But there are so many: general, specific, depends on what your question is.

Question: What do you look for in annotations?

Prokka

- First command line software tool to fully annotate bacterial, archaeal, and viral genomes
- Relies on external feature prediction tools to ID the coordinates of genomic features within contigs

Table 1.

Feature prediction tools used by Prokka

Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen *et al.* , 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen *et al.* , 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA

Prokka

- For the protein coding genes, the function is infered by comparison to databases of known proteins in a hierarchical matter.
- Briefly, prokka first searches, using BLAST+, three (small but well curated) databases, ISfinder, NCBI Bacterial Antimicrobial Resistance Reference Gene Database, UniProtKB (SwissProt).
- Then HMMER3 is used for similarity searching against protein family profiles (mostly bacterial specific).

Prokka

- SCGC deliverable – let's see what this file looks like

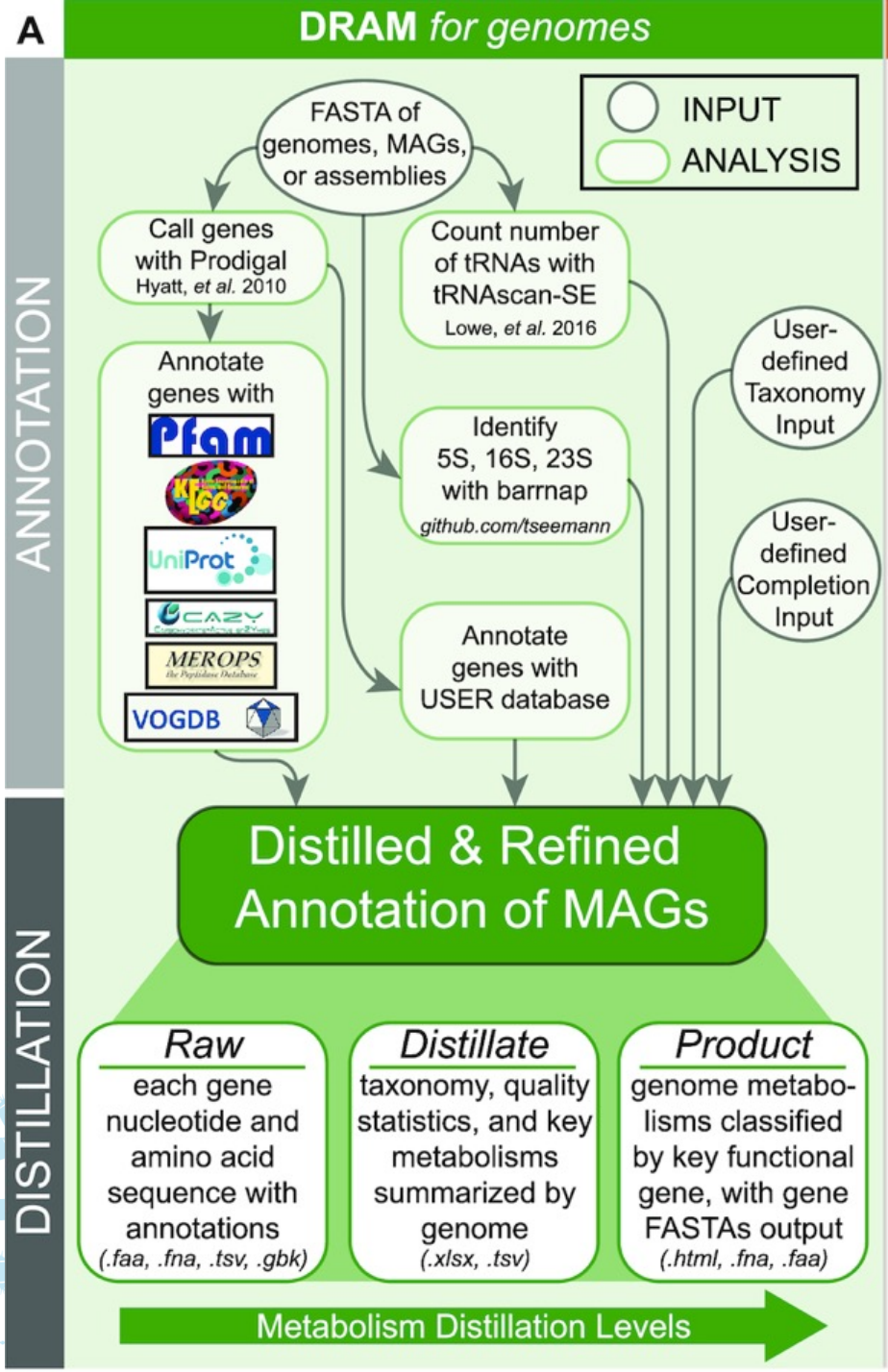
/storage/data/sag-data/gorg-dark/v1_deliverables/

```
>AG-538-A01_contigs_00001 hypothetical protein
MCTVKEAILPSVEYDKFFIFNFLSKSKLIFVFKSALPLISLKKSIFKPNKEDILLISEKL
NFPFVRKKLFIFVLFSSRFKLTLLSVNSKSNKKPFSITAVNPSFSIPLP
>AG-538-A01_contigs_00002 hypothetical protein
MAPSVPNGKINSPCKFPLANIFLSCLLTILGRKILVILVKSLISLFLKAISALFNKKSEL
KNDLSDKNKSNFDKSISLFLTKTFGSCSFTRSFSPLISRCIFVF
>AG-538-A01_contigs_00003 product::Peptide chain release factor 2 (RF-2)^^eggNO
MLGGIFEKNDIKGKVRTFDKKILENNFWKDKLSAQKILKEKKFFENIINKFNSVVNELEN
FEELLKLASQENDTVVIKECEEKISLLFNQIKKVEVECFLSGENDFLDTYLEIHAGAGGT
ESQDWAQMLRRMYSRWIEKKKCKFEIISEHKGDEAGIKSSVLKVTGPYMYGWLKSESGVH
RLVRISPFDSGARRHTSFASVWVYPVVNDSIKIDIKEKDIRIDTYRSSGAGGQHVNTTDS
AVRITHLPTNIVVQCQNERSQHKNKETCFNMLKARLYRYEIQRKEEVSENLAKSKKDIGW
GHQIRSYVLQPYQLVKDLRNNLENTDPEDVLNGEIDSFLEAALYQNQIEEKSK
>AG-538-A01_contigs_00004 product::Penicillin-binding protein 1A (PBP-1a) (PBP1
transglycosylase^^eggNOG::ENOG4108JQC,COG5009^^KO::K05366^^Pfam::PF17092,PF0091
MFVFSTLWYFSSDLPDYKILSNYKPPVSSRVHSGEGQLIAEYALQKRLFIPYDYIPKNVI
YSFLSAEDKNFFSHPGVDVKSITRATIKNIKNIFSTKRLEGASTITQQVAKNFLLTNEIS
LKRKIKEVILAFRIERAYSKERIMELYLNQIYLGEGTYGVAAASLEYFDKSVRNLNYEEA
ALLAALRAPSKYNPYKSKKQATIRRNFVLKNLYFNSYTNKTFYFNFKKKFIKTKKRKIK
```

Extension	Description
.gff	This is the master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV.
.gbk	This is a standard Genbank file derived from the master .gff. If the input to prokka was a multi-FASTA, then this will be a multi-Genbank, with one record for each sequence.
.fna	Nucleotide FASTA file of the input contig sequences.
.faa	Protein FASTA file of the translated CDS sequences.
.ffn	Nucleotide FASTA file of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)
.sqn	An ASN1 format "Sequin" file for submission to Genbank. It needs to be edited to set the correct taxonomy, authors, related publication etc.
.fsa	Nucleotide FASTA file of the input contig sequences, used by "tbl2asn" to create the .sqn file. It is mostly the same as the .fna file, but with extra Sequin tags in the sequence description lines.
.tbl	Feature Table file, used by "tbl2asn" to create the .sqn file.
.err	Unacceptable annotations - the NCBI discrepancy report.
.log	Contains all the output that Prokka produced during its run. This is a record of what settings you used, even if the --quiet option was enabled.
.txt	Statistics relating to the annotated features found.
.tsv	Tab-separated file of all features: locus_tag,ftype,len_bp,gene,EC_number,COG,product

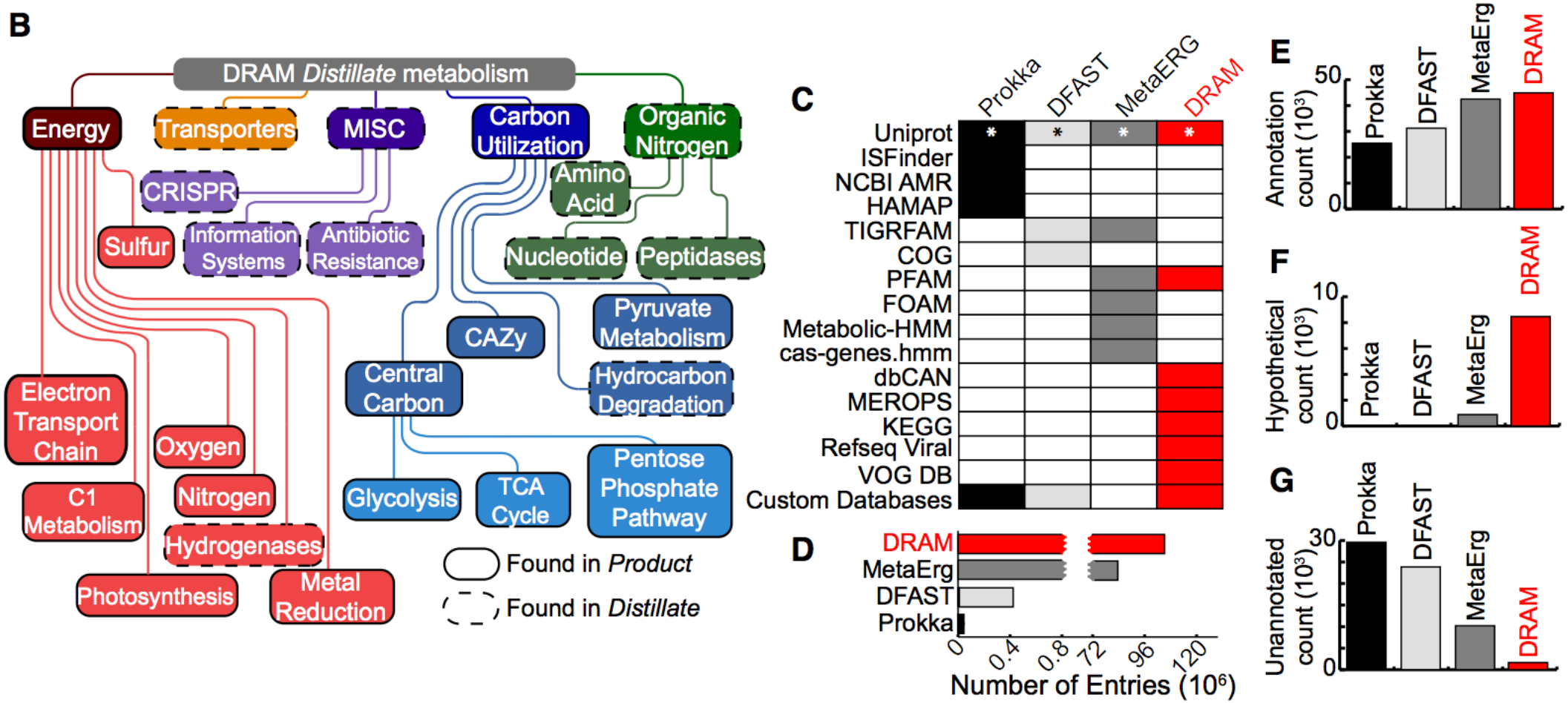
DRAM

- DRAM has been recently developed for the annotation of genomes and viral contigs. Besides the identification of individual genes, DRAM attempts to put information into context: to assign them to metabolic pathways.
- Predicted amino acid sequences are searched with: KEGG, Pfam, UniRef90 and MEROPS using M Mseqs2.
- Additionally, HHMER3 is used for HMM profile searches of dbCAN and VOGDB.



Shaffer et al., 2020

DRAM



Shaffer et al., 2020

Running DRAM for analysis…

- First, identify 3-4 SAGs of interest – you can choose based on GORG annotations, or lineage identity
- Then, we will run the DRAM commands so that they can run during the lecture ☺
- Post-lecture, will check in to see if DRAM worked okay

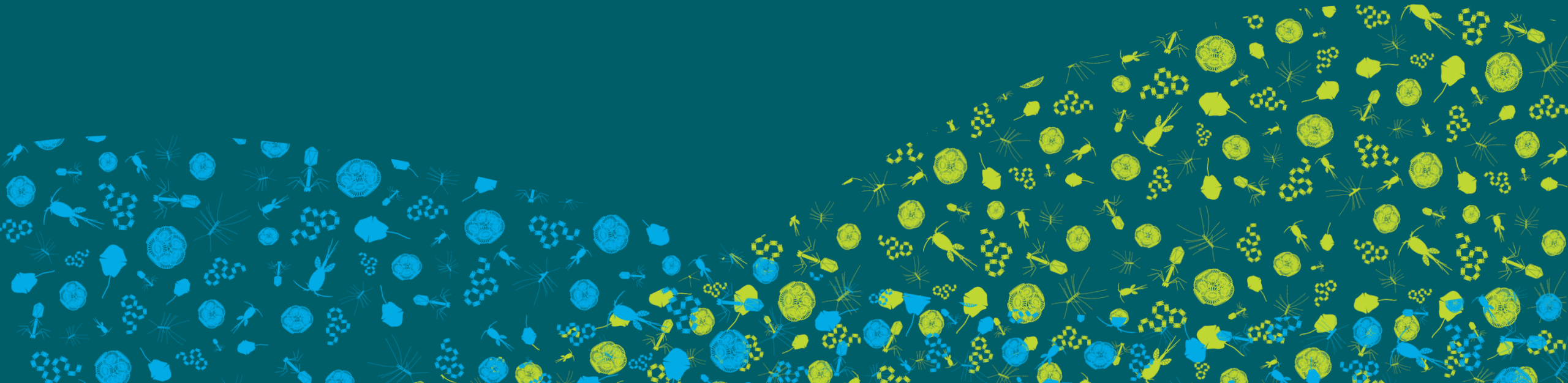
-then coffee break-

- Analysis of outputs, how we can manipulate it at a larger scale, and a deeper dive into misannotations

Bigelow | Laboratory for Ocean Sciences

DRAM annotation outputs

with further caveats!



DRAM outputs

Now:

Annotation summaries

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

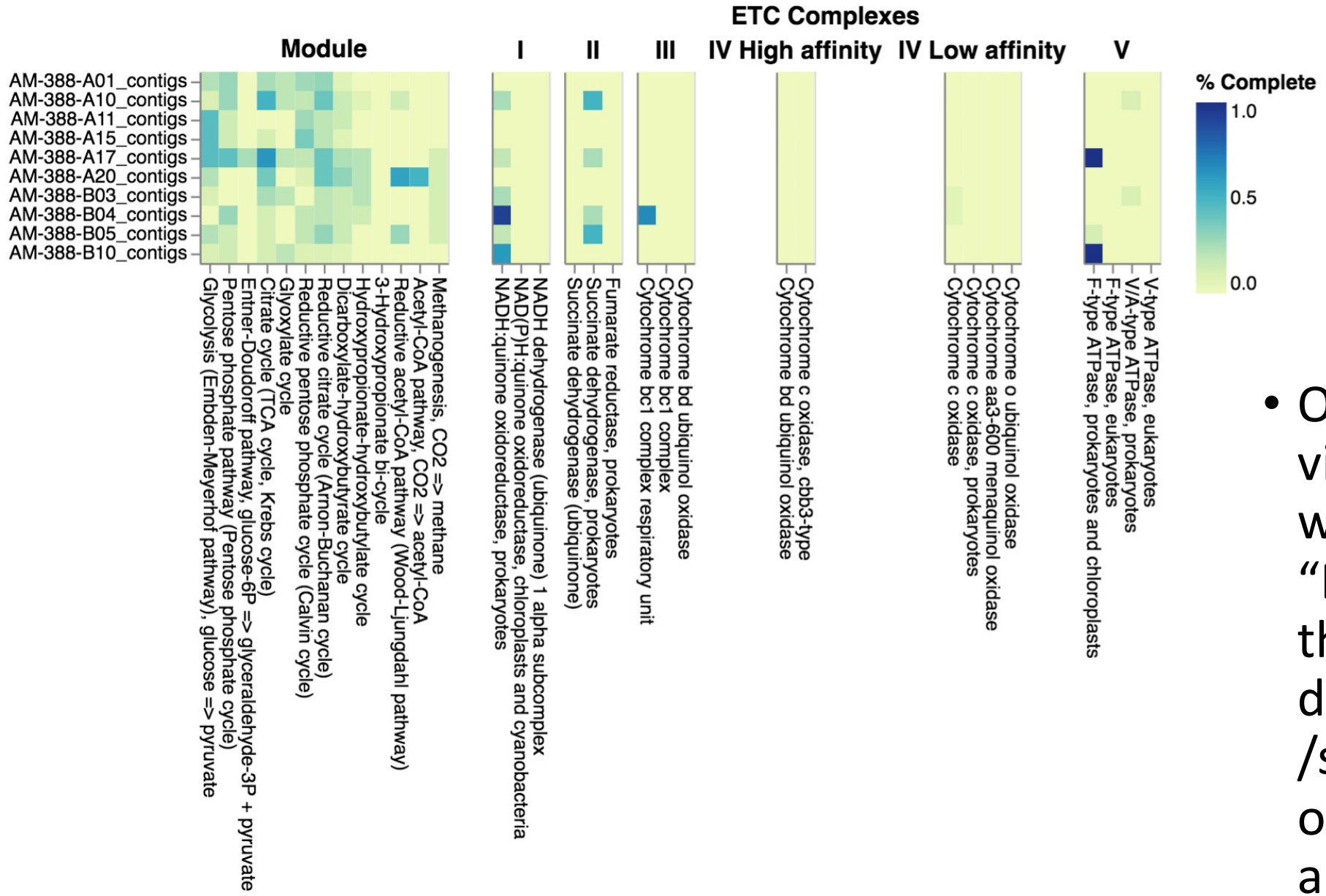
What are they, how can they be used? What will be most useful to you?

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) – this includes the dataset available in: /storage/tutorials/jamboree-day4-annotation/BulkSAGs_DRAMoutput/

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 => acetyl CoA 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 alternative
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
AM-388-A01_contigs
AM-388-A10_contigs
AM-388-A11_contigs
AM-388-A15_contigs
AM-388-A17_contigs
AM-388-A20_contigs
AM-388-B03_contigs
AM-388-B04_contigs
AM-388-B05_contigs
AM-388-B10_contigs
Function is Present
true
false

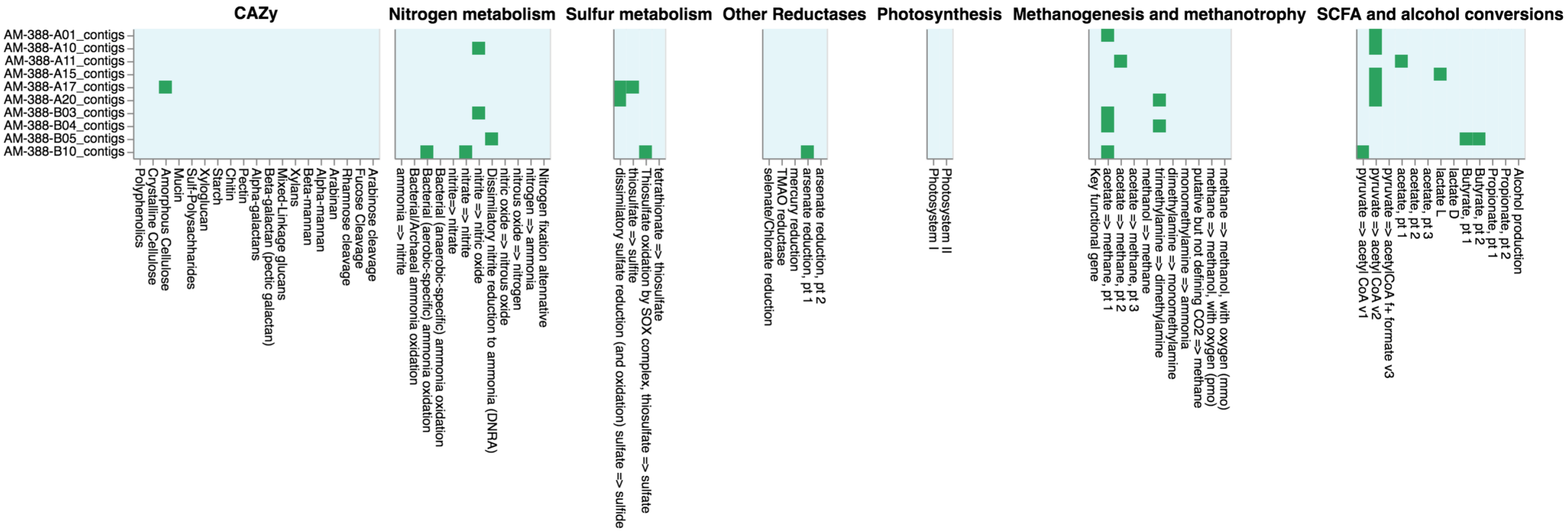
KEGG Visualization

- https://www.genome.jp/kaas-bin/kaas_main?prog=GHOSTX&way=s

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, or as databases include contamination
- 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

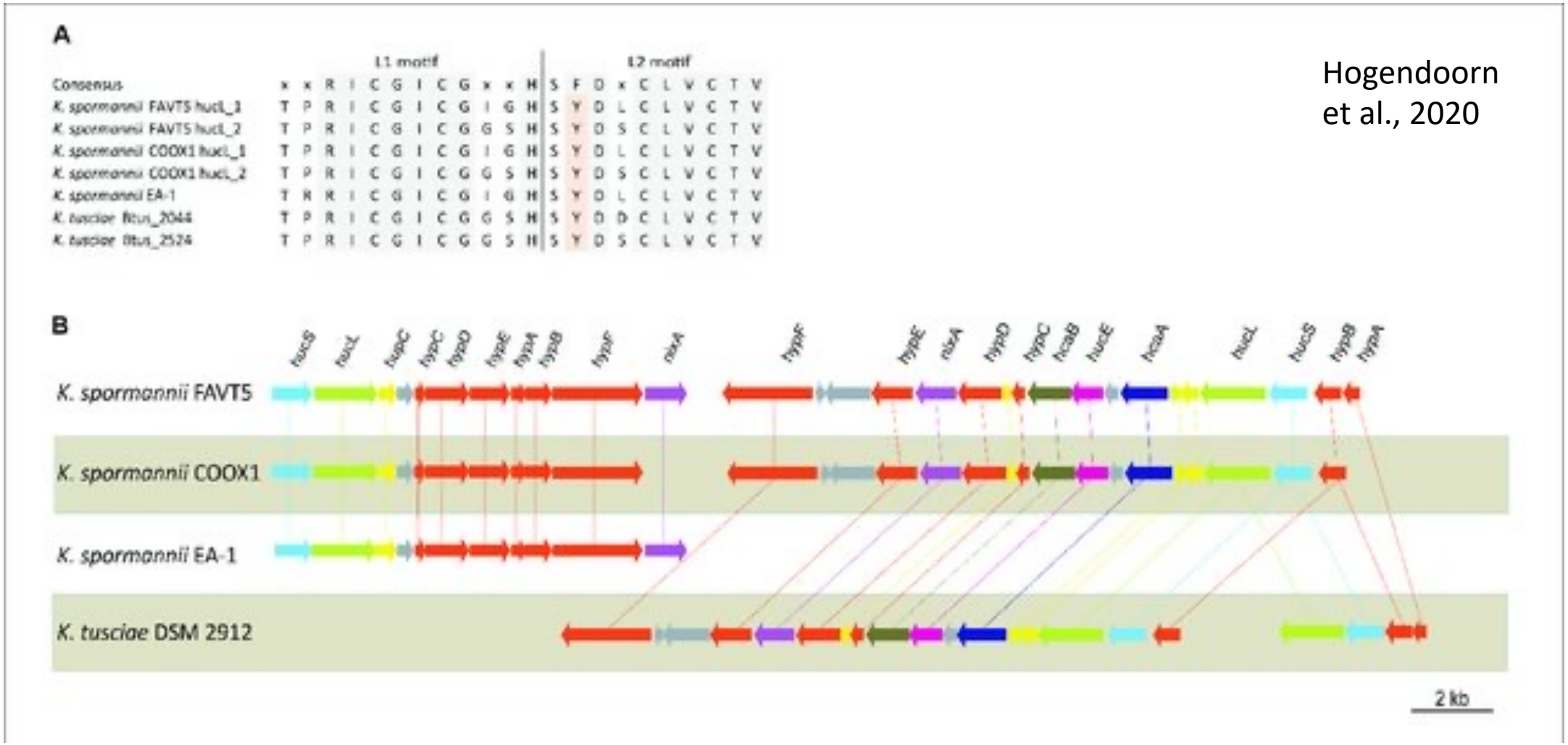
KEGG database

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

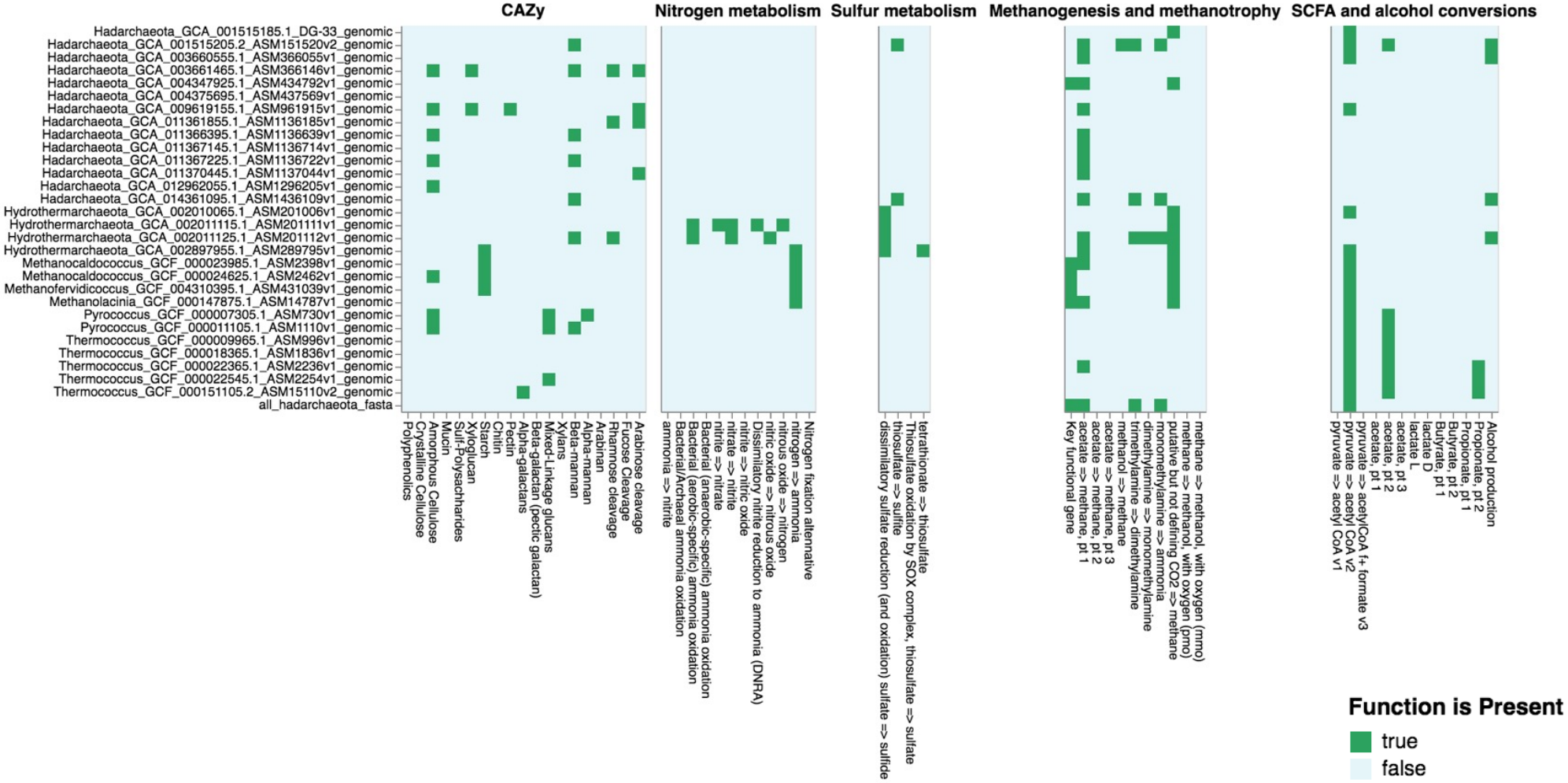
- Example: Cysteine residues key for binding site of hydrogenases (for [NiFe]-hydrogenases, need paired CxxC motifs)

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?**

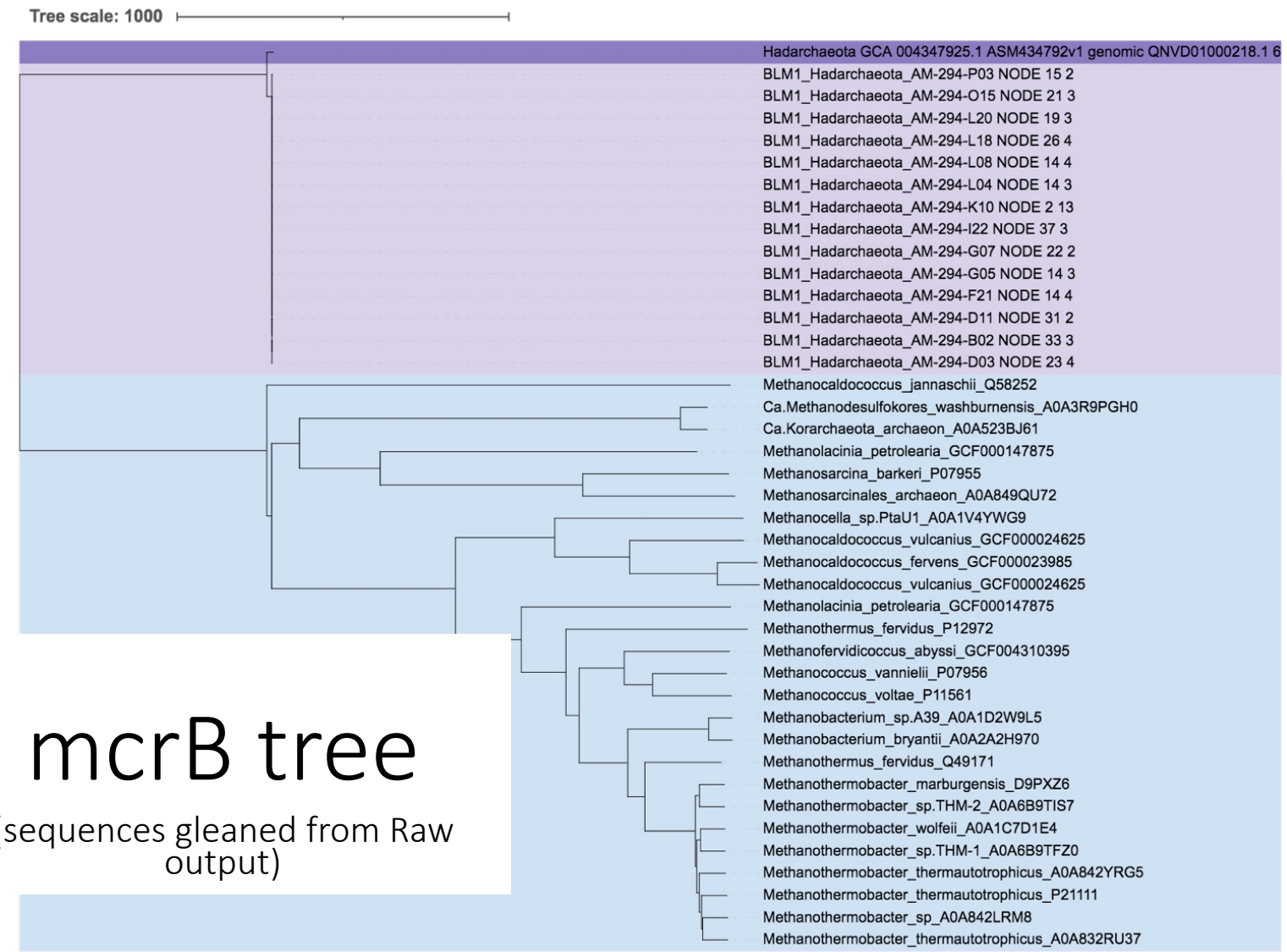
Other ways to check for potential misannotations? Or divergence?



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

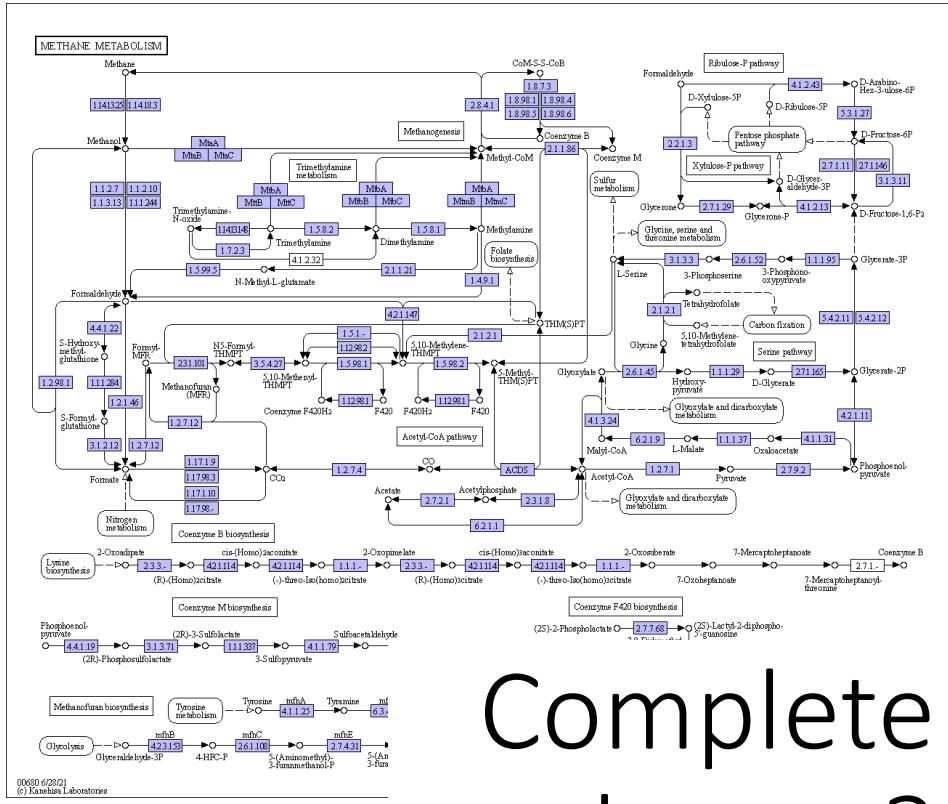


mcrB tree

(sequences gleaned from Raw output)

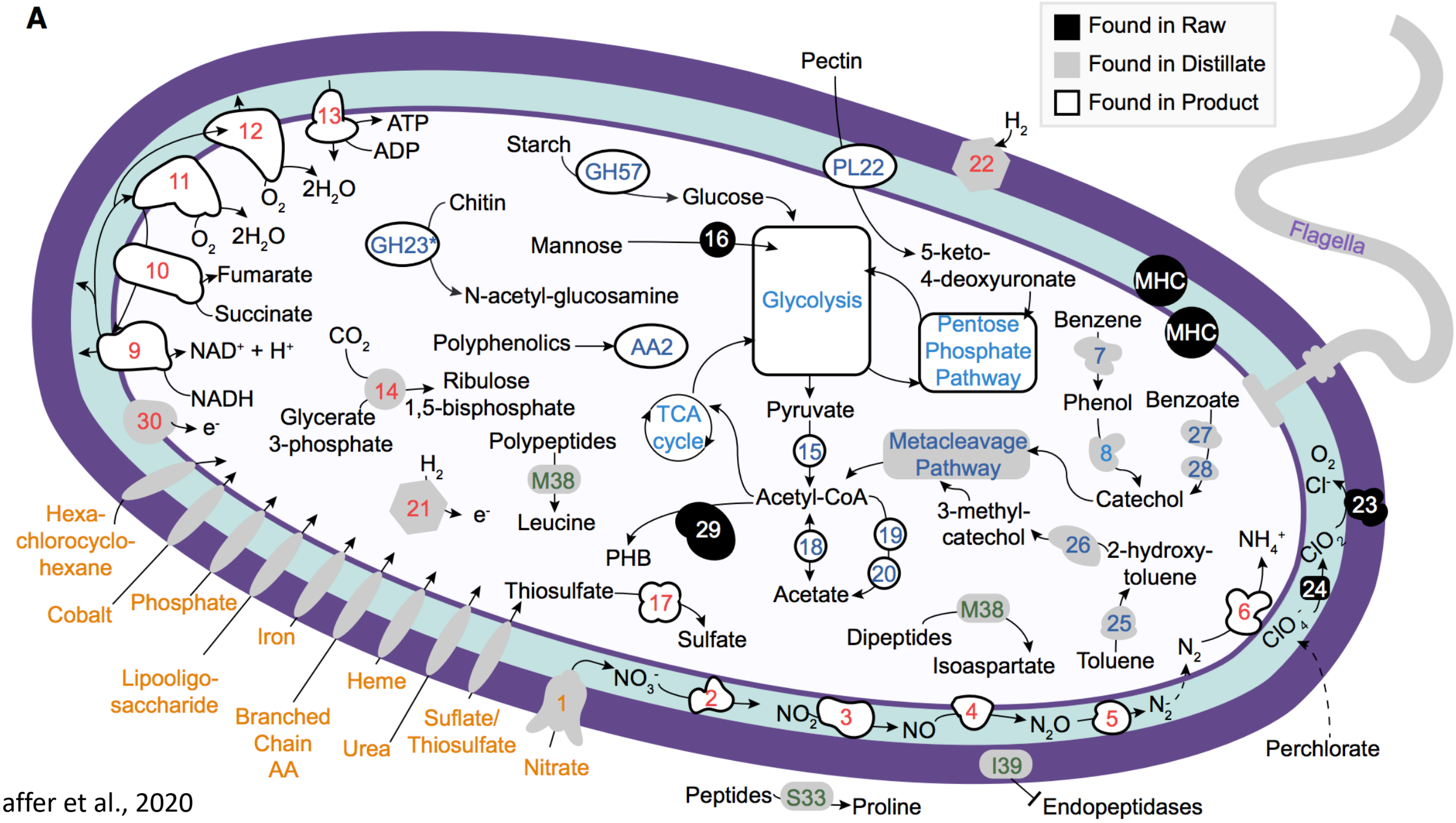
KAAS - KEGG Automatic Annotation Server

for ortholog assignment and pathway mapping



Genome.jp

Complete pathways?



Shaffer 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 your SAGs of choice? (where can you find this data?)

2) how complete are key carbon cycling pathways?

We also have SAG data under /storage/lesson_analyses/annotations/. Can you use the annotate and distill commands (from DRAM) to make summary annotations.tsv, product.html & .tsv, metabolism.xlsx, of your choice of taxa?

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

Guided questions for exploration:

1) What metabolic capabilities are present in cells belonging to the most abundant genus in GORG-Dark?
2) Are any cells capable of {insert interesting metabolism here}?
   What might be a good first step in locating a metabolism of interest?
3) If you have a cell capable of carbon fixation, how complete is the pathway? Are ‘key genes’ present?