METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks

C-DEBI Virtual Meeting Series

Feb 4, 2022

Goals for creating METABOLIC

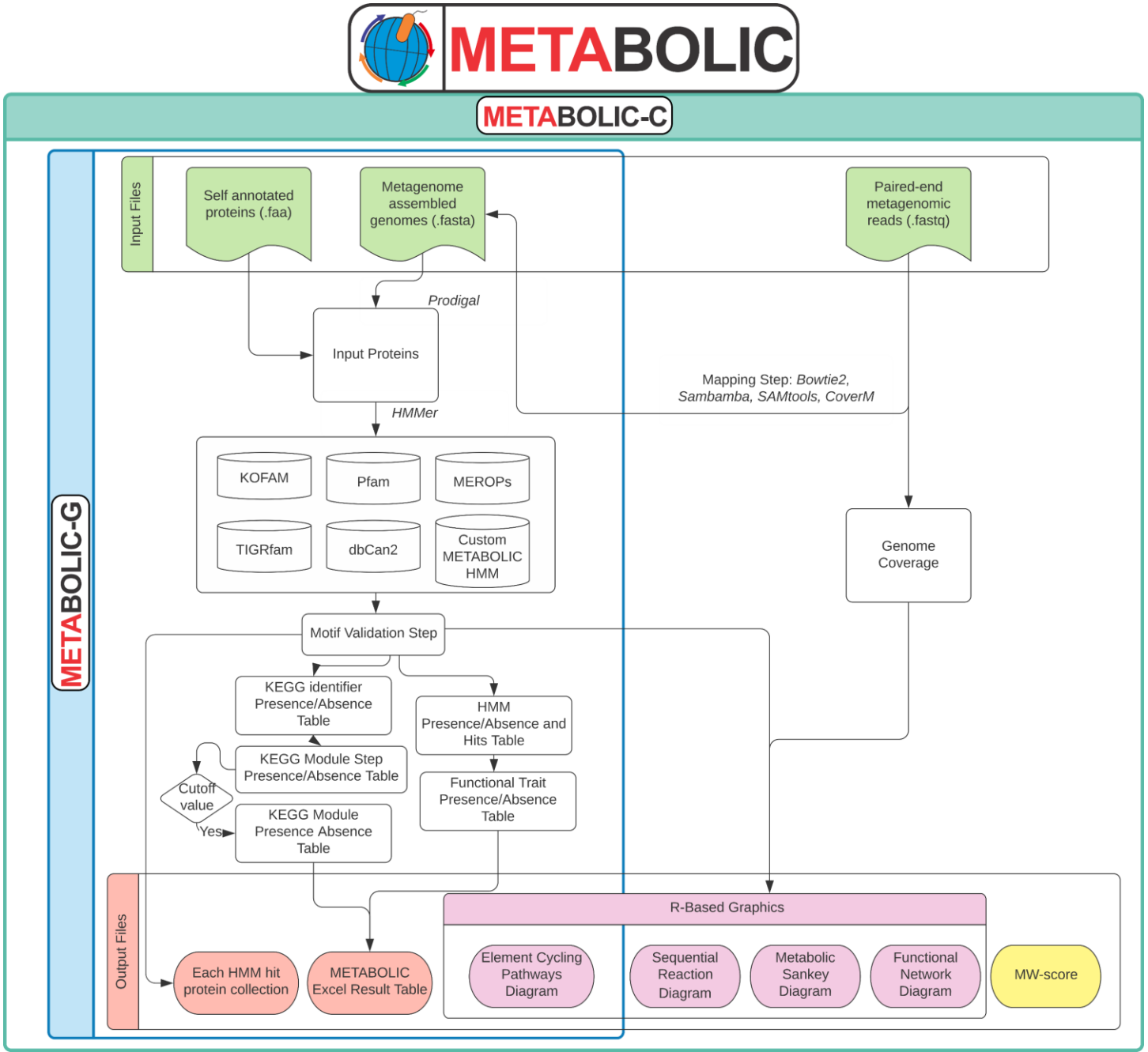
- Enable metabolic and biogeochemical analyses for genomes and microbial communities
- Enable visualization of biogeochemical cycling potential and community-scale functional networks.

Software workflow

METABOLIC-G:
The genome version

METABOLIC-C:
The community version

Include metagenomic reads for community analysis and visualization



Software workflow

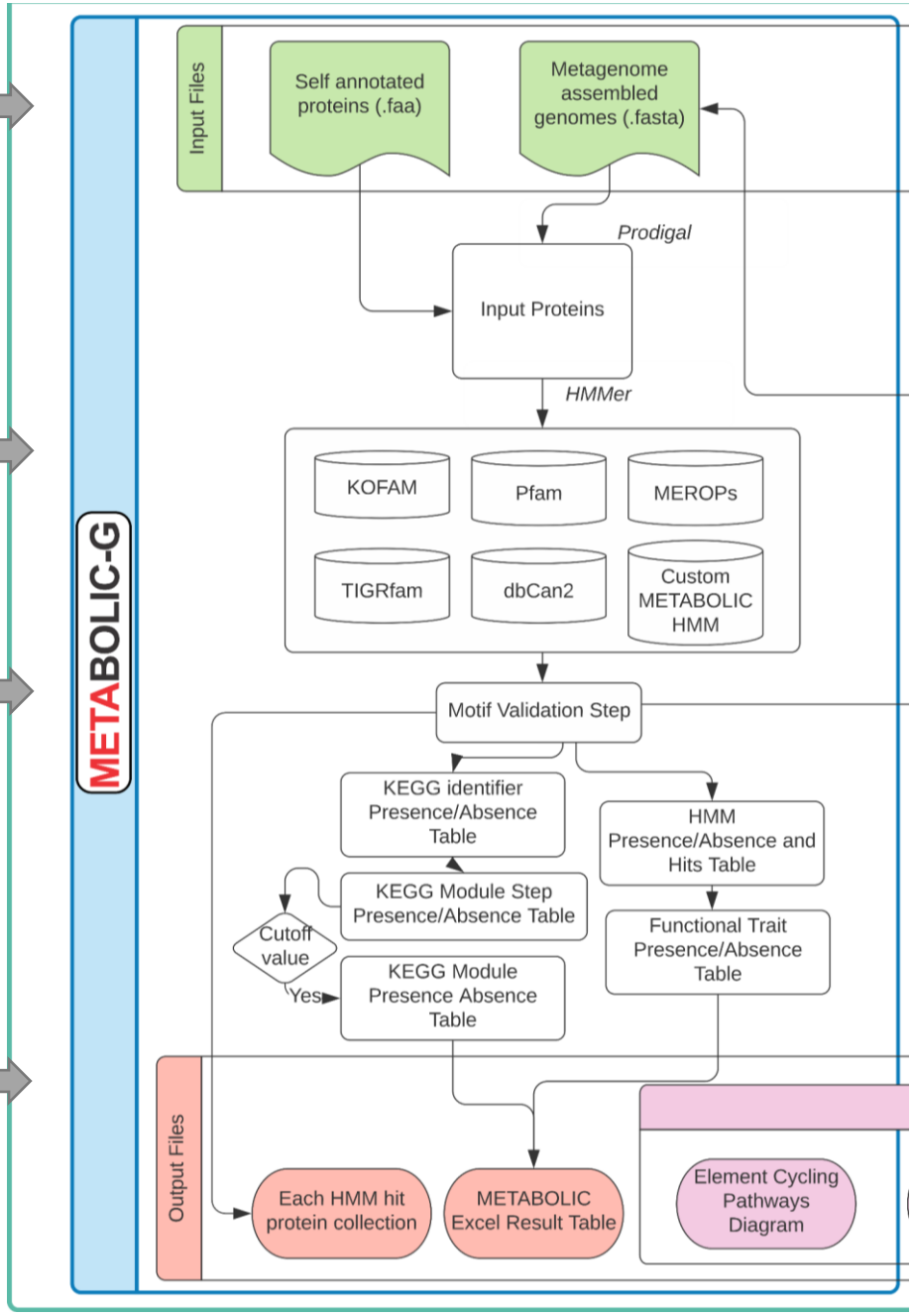
Intake microbial genomes (both *.faa and *.fasta) of: genomes of microbial isolates, MAGs, SAGs

Annotate using a comprehensive set of database: KEGG, TIGRfam, Pfam, custom HMM databases, dbCAN2 (for CAZymes), and MEROPS (for peptidases/inhibitors); manually curated cutoff scores for several HMMs to increase accuracy

Motif validation step by comparing protein motifs against a manually curated set of highly conserved residues in important proteins

Increase the annotation confidence on functionally important proteins with high sequence similarity

Provide user-friendly outputs in the form of tables and figures



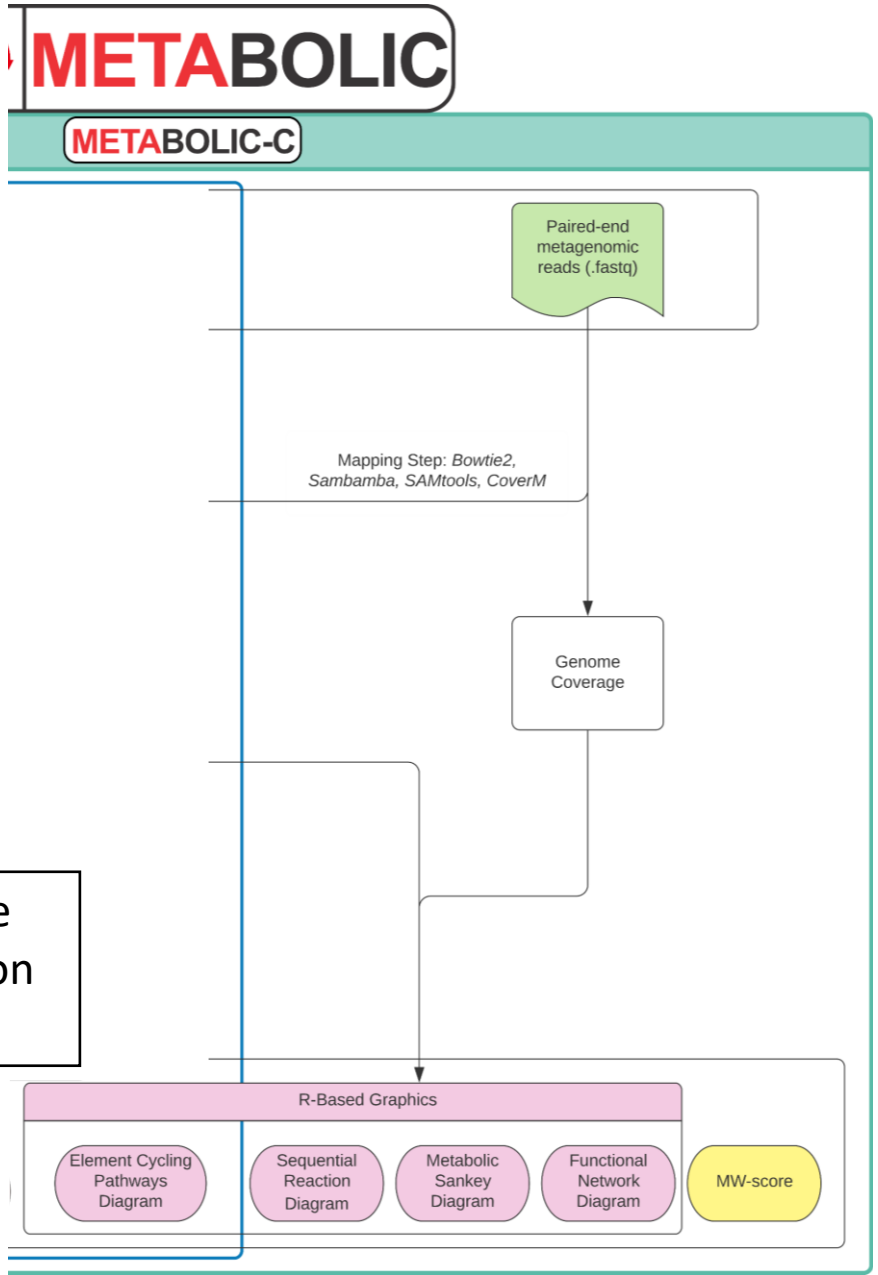
Software workflow

Program Name	Program Description
METABOLIC-G.pl	Allows for classification of the metabolic capabilities of input genomes.
METABOLIC-C.pl	Allows for classification of the metabolic capabilities of input genomes, calculation of genome coverage, creation of biogeochemical cycling diagrams, and visualization of community metabolic interactions and contribution to biogeochemical processes by each microbial group.

Metagenomic read mapping to individual genomes

Obtain genome coverage information

Genome annotation result



METABOLIC outputs

METABOLIC output table

Output File/Folder	File Description	Generated by METABOLIC-C	Generated by METABOLIC-G
All_gene_collections_mapped.depth.txt	The gene depth of all input genes	X	
Each_HMM_Amino_Acid_Sequence/	The faa collection for each hmm file	X	X
intermediate_files/	The hmmsearch, peptides (MEROPS), CAZymes (dbCAN2), and GTDB-Tk (only for METABOLIC-C) running intermediate files	X	X
KEGG_identifier_result/	The hit and result of each genome by Kofam database	X	X
METABOLIC_Figures/	All figures output from the running of METABOLIC	X	X
METABOLIC_Figures_Input/	All input files for R-generated diagrams		
METABOLIC_result_each_spreadsheet/	TSV files representing each sheet of the created METABOLIC_result.xlsx file		
MW-score_result/	The resulted table for MW-score		
METABOLIC_result.xlsx	The resulting excel file of METABOLIC		

1) HMMHit: records the marker HMM hits – including presence/absence state, number of hits, and hit names
2) FunctionHit: records the combined HMM hit results to represent each function presence/absence
3) KEGG module Hit
4) KEGG module step hits to provide the details for each step
5) dbCAN2 result for CAZymes
6) MEROPS hit for peptidases/inhibitors

METABOLIC outputs

Purpose: To address the lack of quantitative and reproducible measures to represent potential metabolic interactions in microbial communities, we developed a new metric that we termed **MW-score (Metabolic-Weight score)**

MW-score (Metabolic-Weight score) resolves metabolic capacity and abundance quantitatively in the co-sharing functional networks

Metrics

MW-scores provide quantitative metrics for:

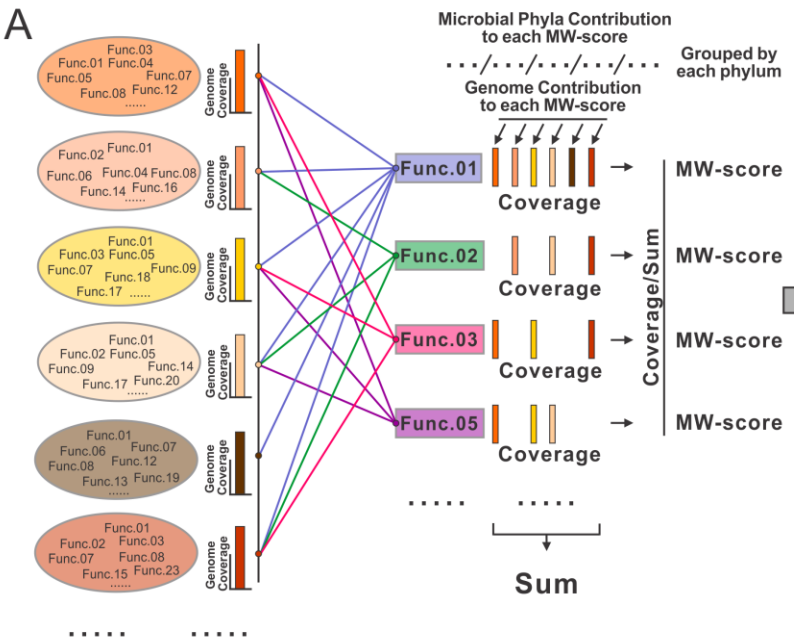
Advantage

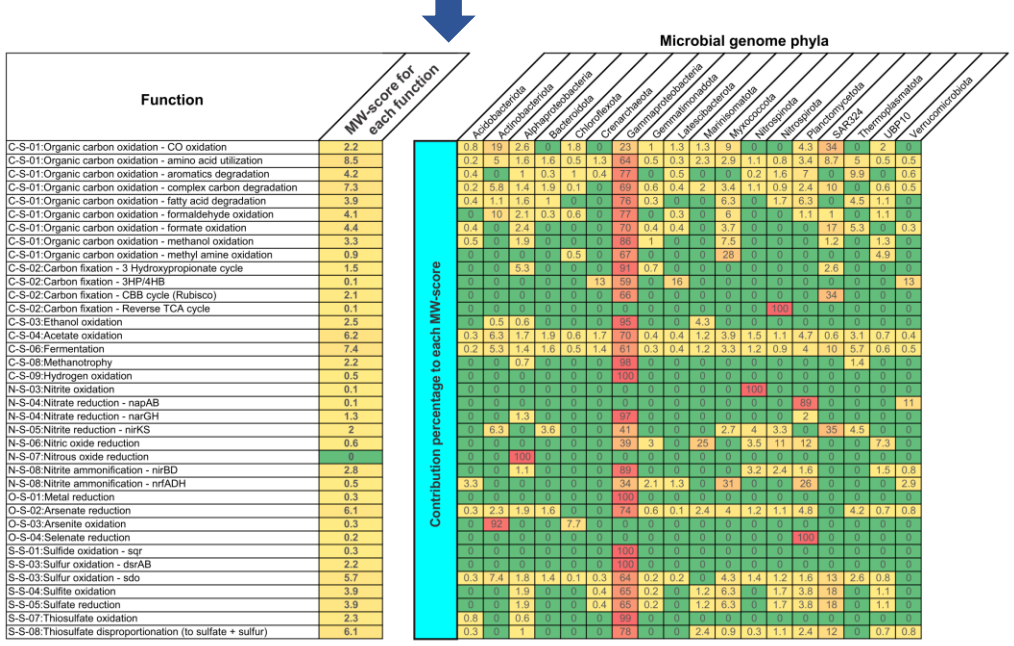
- Quantitative
- Universal for comparison

Graphic pattern of the network

more edges connecting two nodes indicates two steps are more shared, thicker edges indicate higher gene abundance for the metabolic steps

Also provide each microbial group's contribution to the MW-score of a specific function within the community





METABOLIC performance demonstration

Test the prediction of 3HP and 3HP/4HB cycles in three groups of microorganisms

- METABOLIC result has the same annotation compared to that in KEGG genome pathway

				METABOLIC result		KEGG genome pathways	
				Carbon fixation		Carbon fixation	
Accession ID	Organism	KEGG Organism Code	Group	3 Hydroxypropionate cycle (3HP)	3-hydroxypropionate/4-hydroxybutyrate cycle (3HP/4HB)	3 Hydroxypropionate cycle (3HP)	3-hydroxypropionate/4-hydroxybutyrate cycle (3HP/4HB)
GCA_000011905.1	Dehalococcoides mccartyi 195	det	Chloroflexi	Absent	Absent	Absent	Absent
GCA_000017805.1	Roseiflexus castenholzii DSM 1	rca	Chloroflexi	Present	Absent	Present	Absent
GCA_000018865.1	Chloroflexus aurantiacus J-10-fl	cau	Chloroflexi	Present	Absent	Present	Absent
GCA_000021685.1	Thermomicrobium roseum DSM	tro	Chloroflexi	Absent	Absent	Absent	Absent
GCA_000021945.1	Chloroflexus aggregans DSM 94	cag	Chloroflexi	Present	Absent	Present	Absent
GCA_000299395.1	Nitrosopumilus sediminis AR2	nir	Thaumarchaeota	Absent	Present	Absent	Present
GCA_000698785.1	Nitrososphaera viennensis EN76	nvn	Thaumarchaeota	Absent	Present	Absent	Present
GCA_000875775.1	Nitrosopumilus piranensis D3C	nid	Thaumarchaeota	Absent	Present	Absent	Present
GCA_000812185.1	Nitrosopelagicus brevis CN25	nbv	Thaumarchaeota	Absent	Present	Absent	Present
GCA_900696045.1	Nitrosocosmicus franklandus NF	nfn	Thaumarchaeota	Absent	Present	Absent	Present
GCA_000015145.1	Hyperthermus butylicus DSM 54	hbu	Crenarchaeota	Absent	Absent	Absent	Absent
GCA_000017945.1	Caldisphaera lagunensis DSM 1	clg	Crenarchaeota	Absent	Present	Absent	Present
GCA_000148385.1	Vulcanisaeta distributa DSM 144	vdi	Crenarchaeota	Absent	Absent	Absent	Absent
GCA_000193375.1	Thermoproteus uzoniensis 768-2	tuz	Crenarchaeota	Absent	Present	Absent	Present
GCA_003431325.1	Acidilobus sp. 7A	acia	Crenarchaeota	Absent	Absent	Absent	Absent

Remarks: Only Chloroflexii contain 3HP. 3HP/4HB could only be detected in Crenarchaeota and Thaumarchaeota

The reference of biochemical evidence of the existence of corresponding carbon fixation pathways

	3HP	3HP/4HB
Choloflexi	10.1073/pnas.1710798114	
Crenarchaeota		10.1126/science.1149976
Thaumarchaeota		10.1016/j.mib.2011.04.007

- METABOLIC result annotation is also consistent to the reference of biochemical evidence

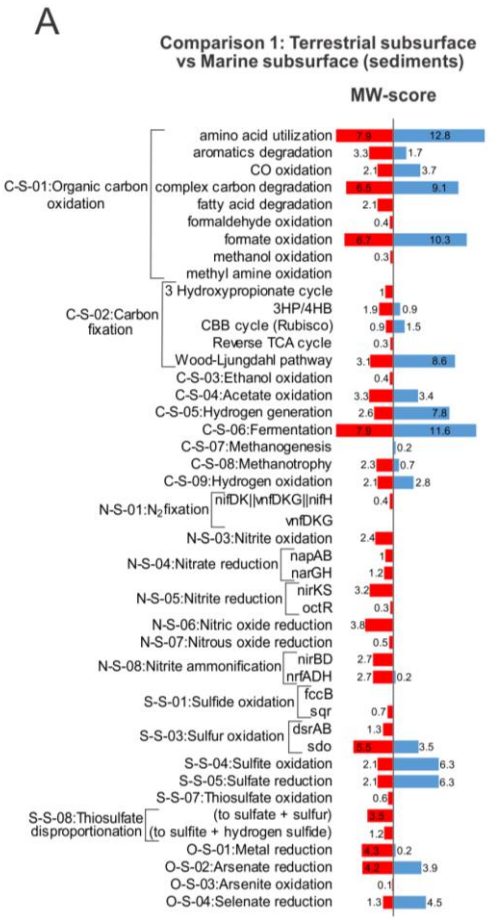
METABOLIC performance demonstration

- **Wide application and good performance of METABOLIC in eight different samples:**

  marine subsurface,
  terrestrial subsurface,
  deep-sea hydrothermal vent,
  freshwater lake

  gut microbiome from patients with colorectal cancer,
  gut microbiome from healthy control,
  meadow soil,
  wastewater treatment facility

- **Community metabolism comparison based on MW-scores**



Terrestrial subsurface contains more abundant metabolic functions related to nitrogen cycling compared to the marine subsurface

Consistency:

Consistent with the previous characterization of these two environments

Anantharaman K et al., *Nat Commun.* 2016;7:13219;
Glass JB et al., *Environ Microbiol*. 2021 Aug;23(8):4646-4660

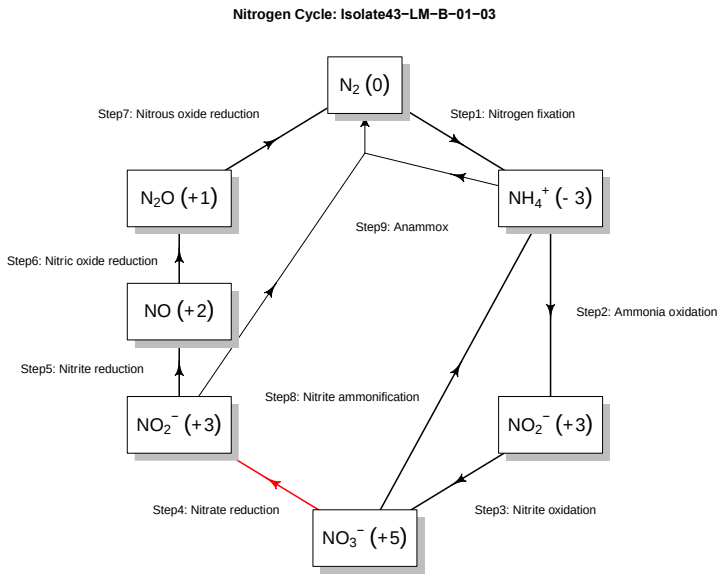
Visualizations:
Four types of diagrams

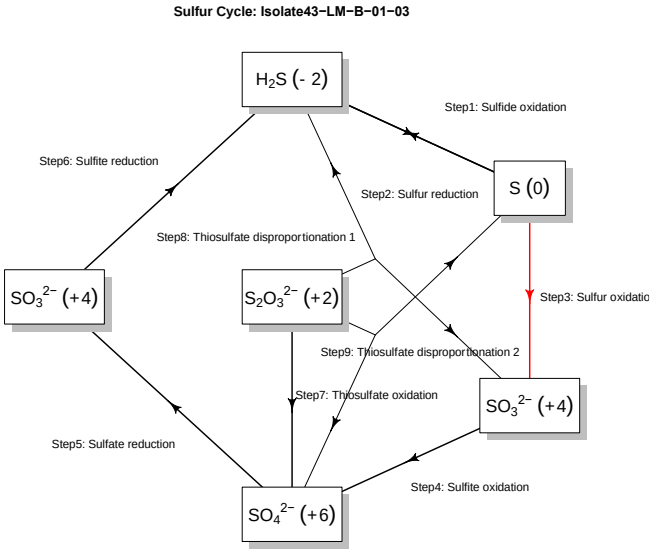
1. Nutrient Cycle Diagram
    1. Organismal-level
    2. Community-level
3. Sequential Metabolic Transformations
4. Functional Networks
5. Metabolic Alluvial Diagram

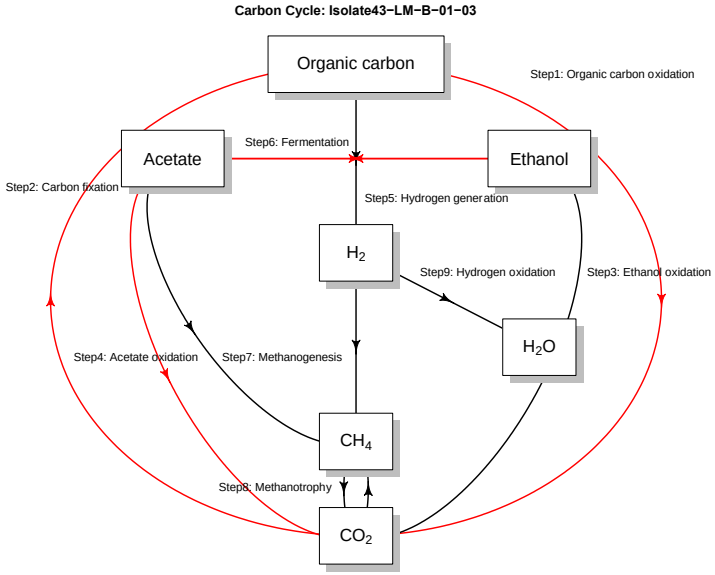
All visualizations are created as part of the default pipeline, but can also be ran individually using an Rscript.

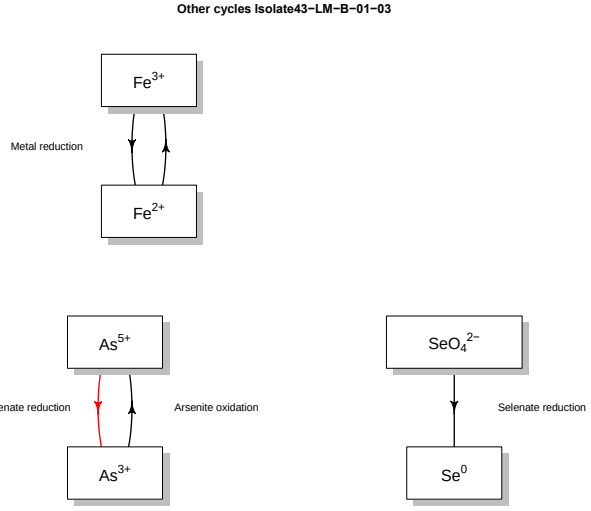
Nutrient cycling diagrams: Organismal-level

Red arrows = present
Black arrows = not present

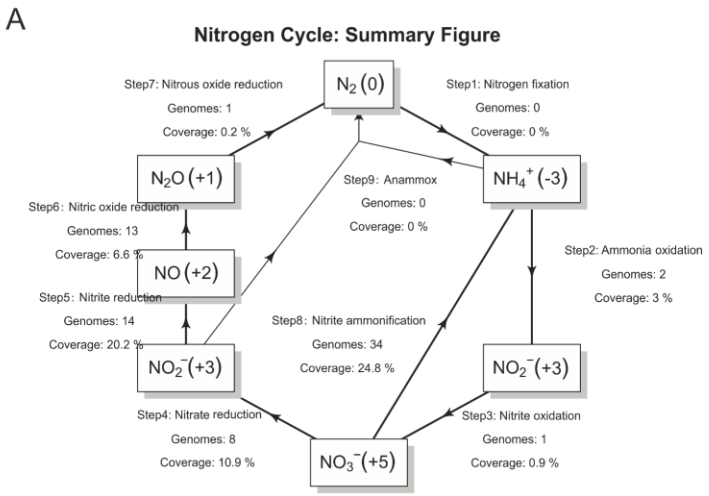


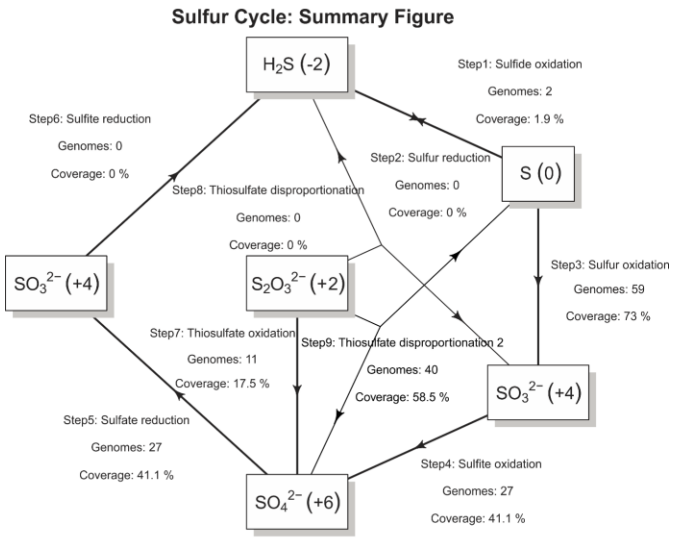


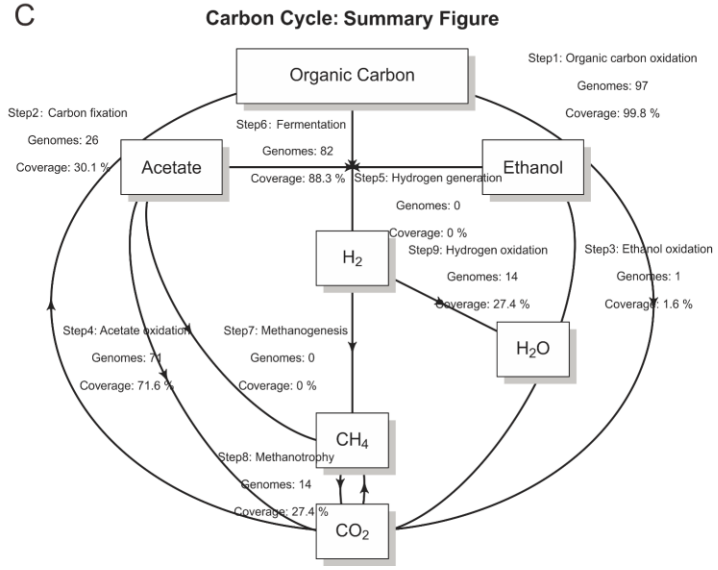


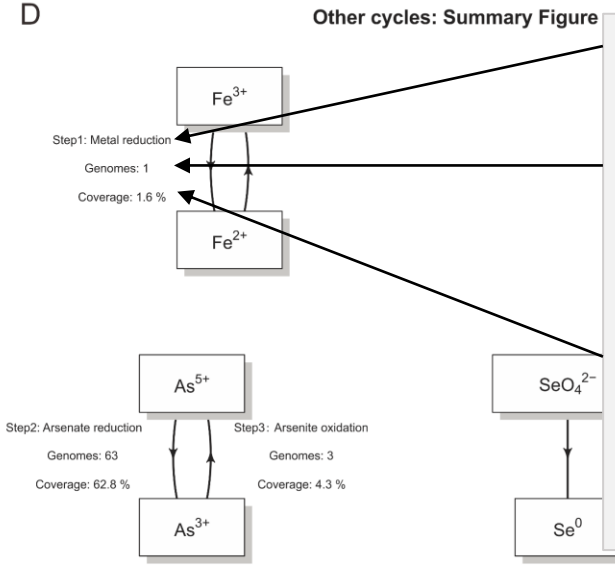


Nutrient cycling diagrams: Community-level









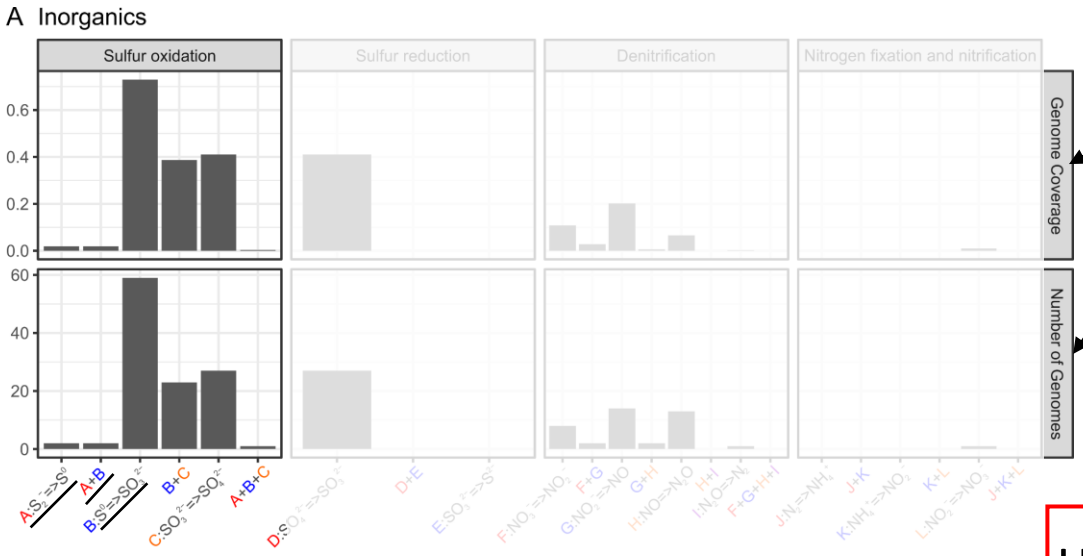
step number and reaction

number of genomes that can conduct these reactions

metagenomic coverage of genomes (represented as a percentage within the community) that can conduct these reactions

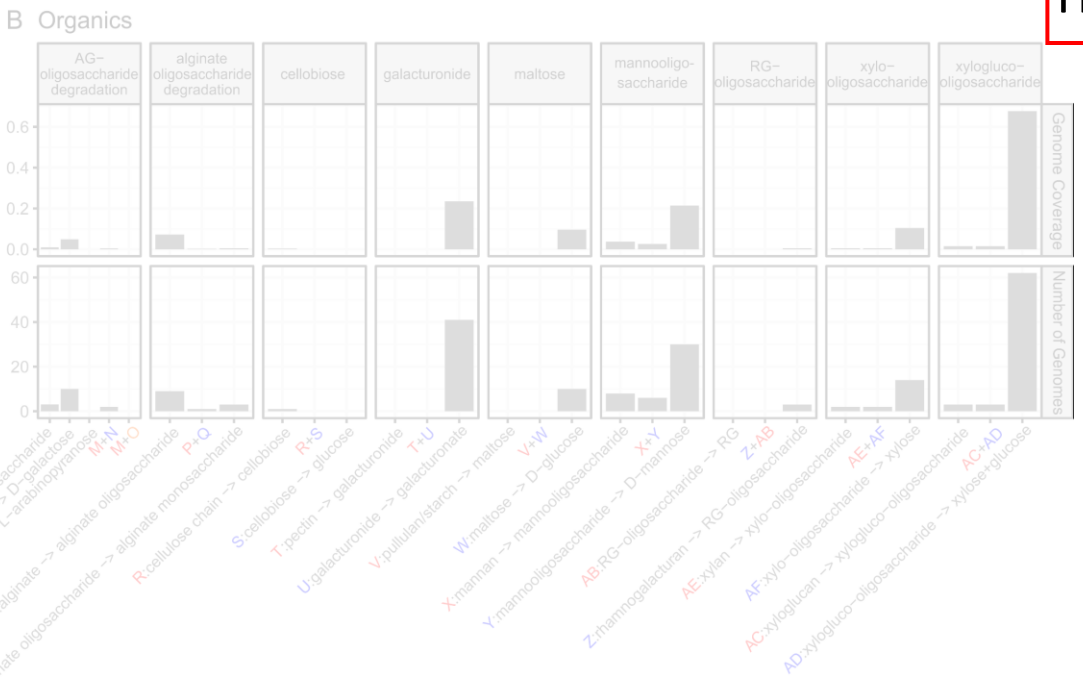
Sequential metabolic transformations:

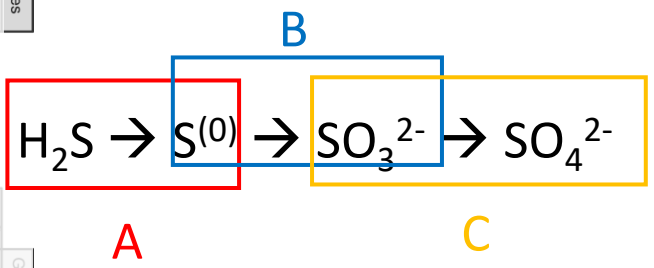
To show how *whole* biogeochemical pathways are *broken up*



The genome coverage

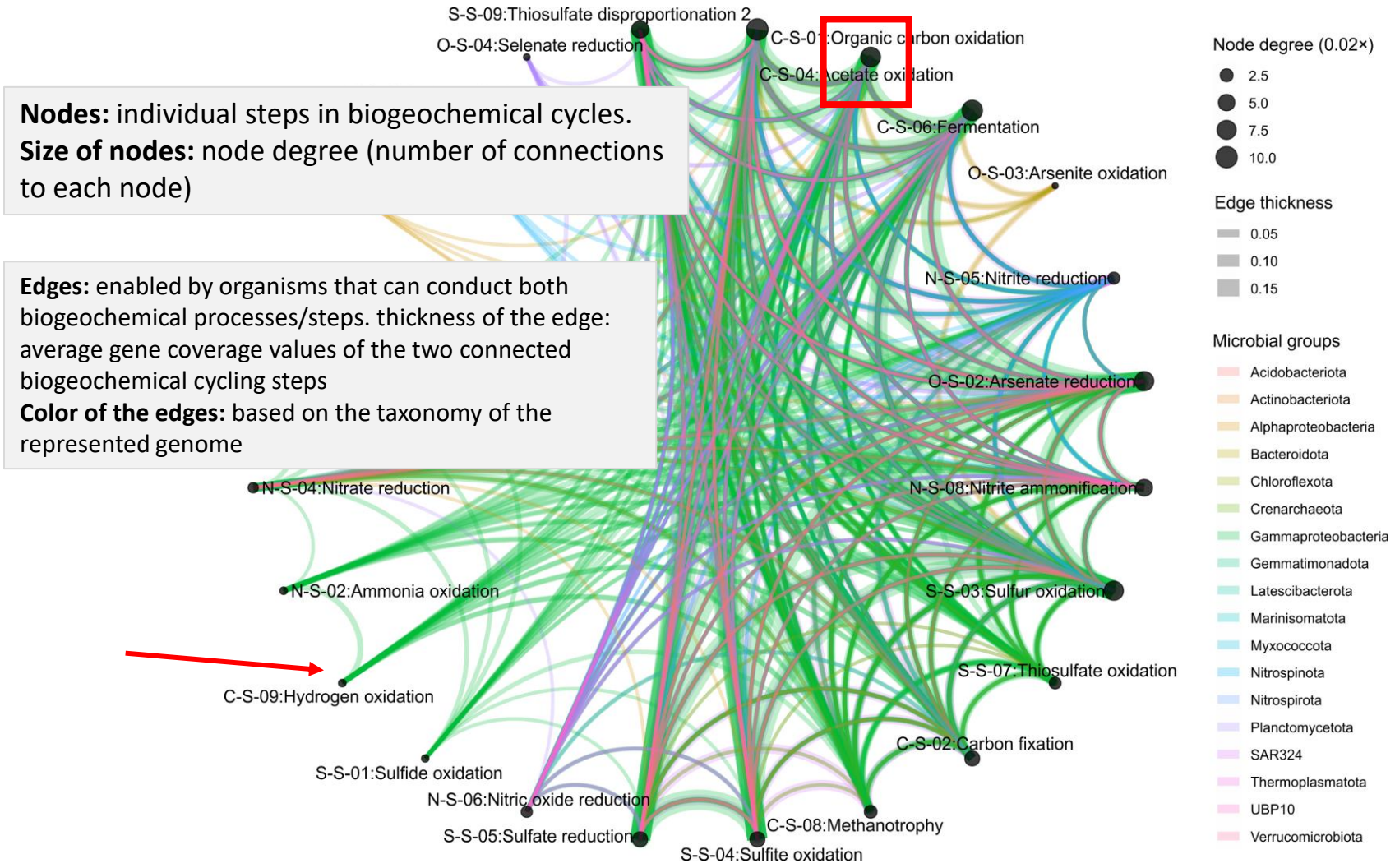
The number of genomes of organisms that are involved in certain sequential metabolic transformations





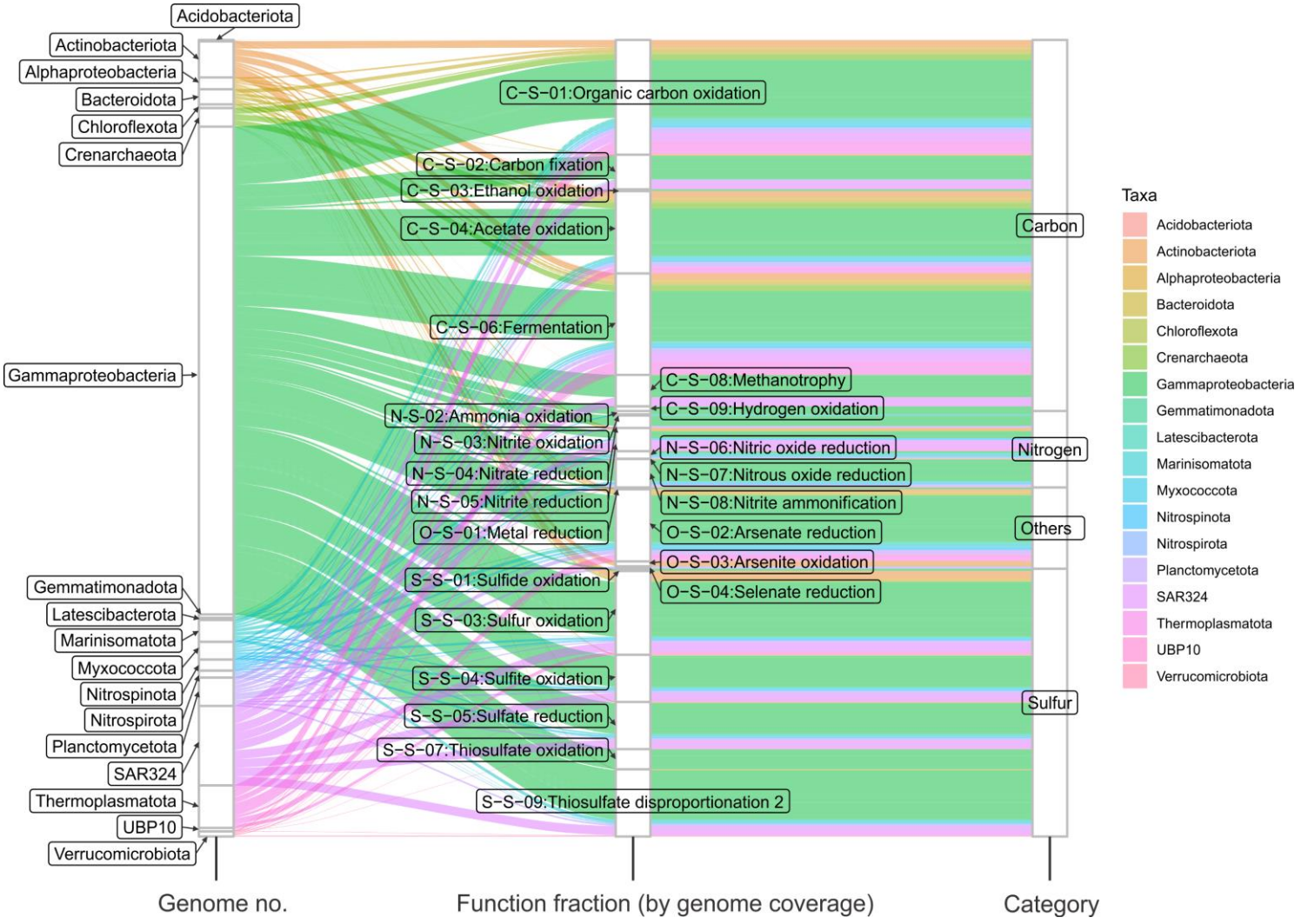
Functional networks:

To visualize *functional networks* and which pathways are more often associated with each other than others



Metabolic Alluvial diagram:

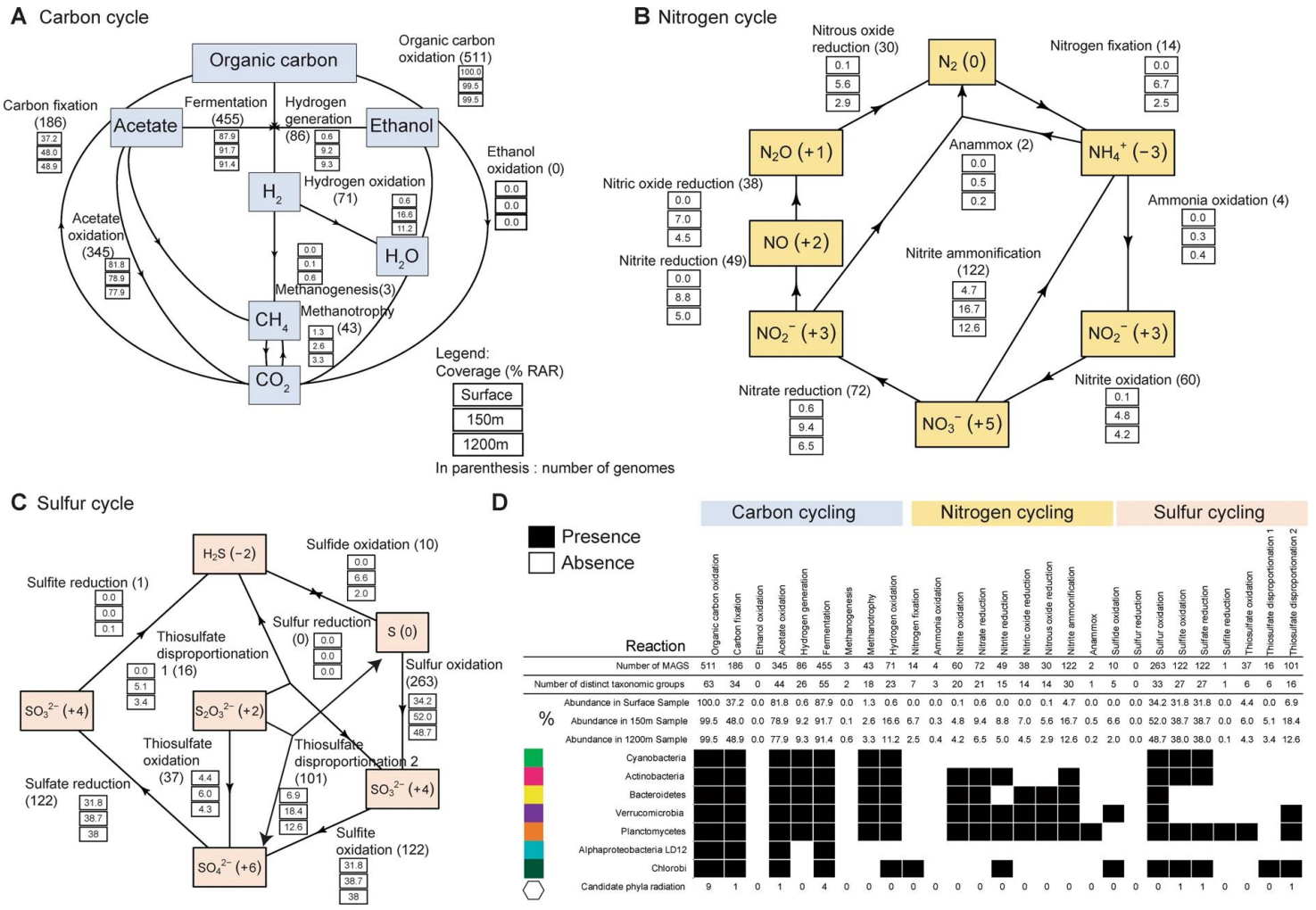
To see how *groups of taxa* are associated with *which biogeochemical reactions/categories*



Width of curved line: from a specific microbial group to a given functional trait indicates their corresponding proportional contribution to a specific metabolism

Applications of METABOLIC to different datasets: Freshwater

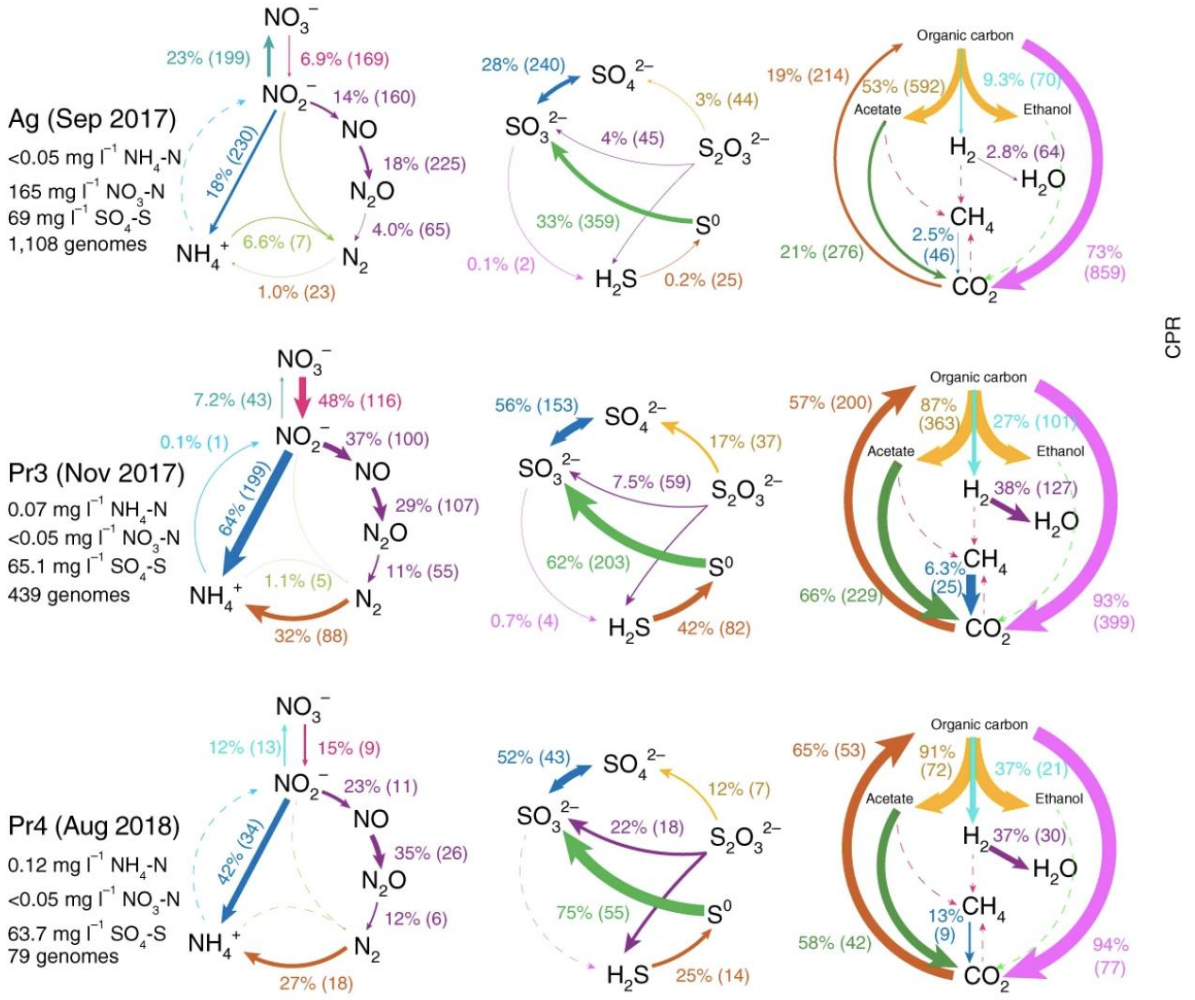
To compare across spatial depths: C common throughout, S and N under the oxycline



Tran, P.Q., Bachand, S.C., McIntyre, P.B. *et al.* Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. *ISME J* **15,** 1971–1986 (2021).

Applications of METABOLIC to different datasets: Groundwater

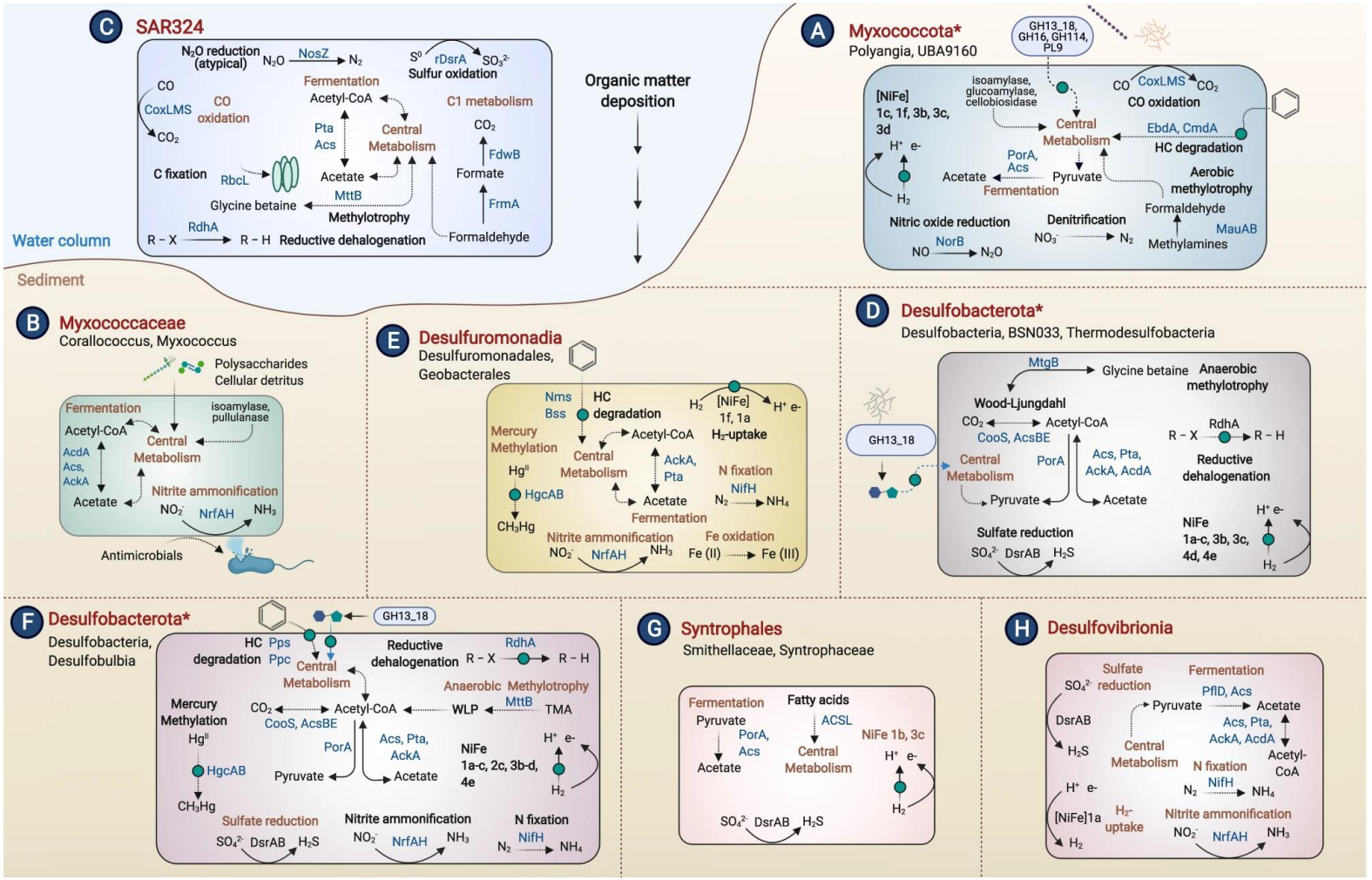
To compare across 2 environmental types (agricultural & pristine), over time



He, C., Keren, R., Whittaker, M.L. *et al.* Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat Microbiol* **6,** 354–365 (2021).

Applications of METABOLIC to different datasets: Phylum-level comparative genomics

To leverage the textual output to create their own user-made cellular maps



Langwig, M.V., De Anda, V., Dombrowski, N. *et al.* Large-scale protein level comparison of Deltaproteobacteria reveals cohesive metabolic groups. *ISME J* **16,** 307–320 (2022).

Summary

- METABOLIC is an open-source, free tool that aims to facilitate the annotations of genomes, or MAGS, with a focus on genes involved in biogeochemical cycling
- Focused on added annotation validation steps, multiple ways to explore the output data, and visualizations
- Developed an M-W score which can be used for standardized, reproducible comparisons between communities & samples
- Can be applied to diverse environments and utilized at different scales of analysis (genome or community).

METABOLIC

https://github.com/AnantharamanLab/METABOLIC

Appendix

Update on the carbon fixation pathway

3HP/4HB

As being not necessarily used in the pathway or being members of "generic" superfamilies (potential dual functionality, having been discovered involved with the fermentation of 4-aminobutyrate by *Clostridium aminobutyricum*), we deleted 4-hydroxybutyryl-CoA dehydratase gene as the marker for 3HP/4HB pathway, but used the 4-hydroxybutyrate---CoA ligase.

- K14466 and K18861 (4-hydroxybutyrate---CoA ligase (AMP-forming) [EC:6.2.1.40]) were used as the marker for 3HP/4HB pathway
- K18861 and 4hbl (K14467)(4-hydroxybutyrate---CoA ligase (AMP-forming) [EC:6.2.1.40]) were used as the marker for DC/4-HB pathway
- The updates have been made to both resulted excel file ("METABOLIC_result.xlsx"), Nutrient cycling diagrams, and MW-scores

3HP

Two genes are currently used as the marker for 3HP pathway:
propionyl-CoA synthase (K14469)
mcr, malonyl-CoA reductase / 3-hydroxypropionate dehydrogenase (NADP+) (K14468)

Feb 4, 2022

METABOLIC

https://github.com/AnantharamanLab/METABOLIC