Pore Water Microbiome Characterization using Ribosome Profiling

Yochen Zhong

Outline

- Introduction
- Taxonomic classification
- Functional analysis
- Conclusions

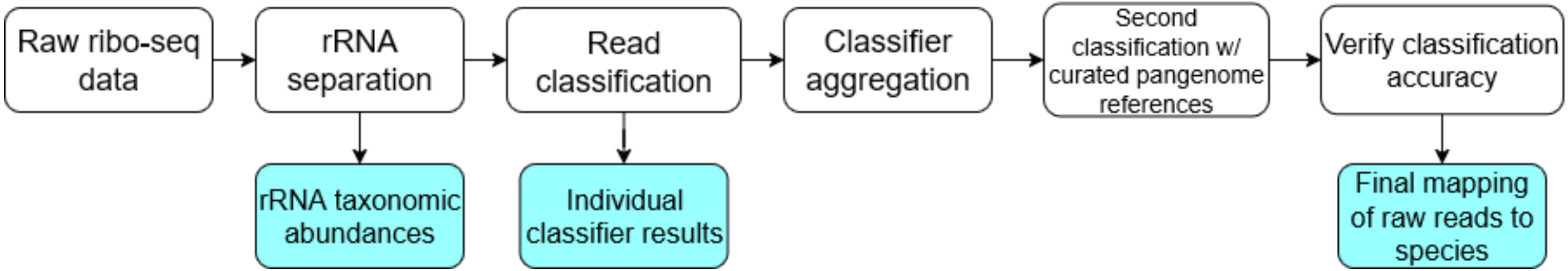
Background

- Ribosome profiling: using short, ribosome-protected mRNA fragments to create a snapshot of protein production.
  - More direct measure of cell activity than RNA-seq
- This technique has rarely been applied to environmental microbiome samples.

Objectives

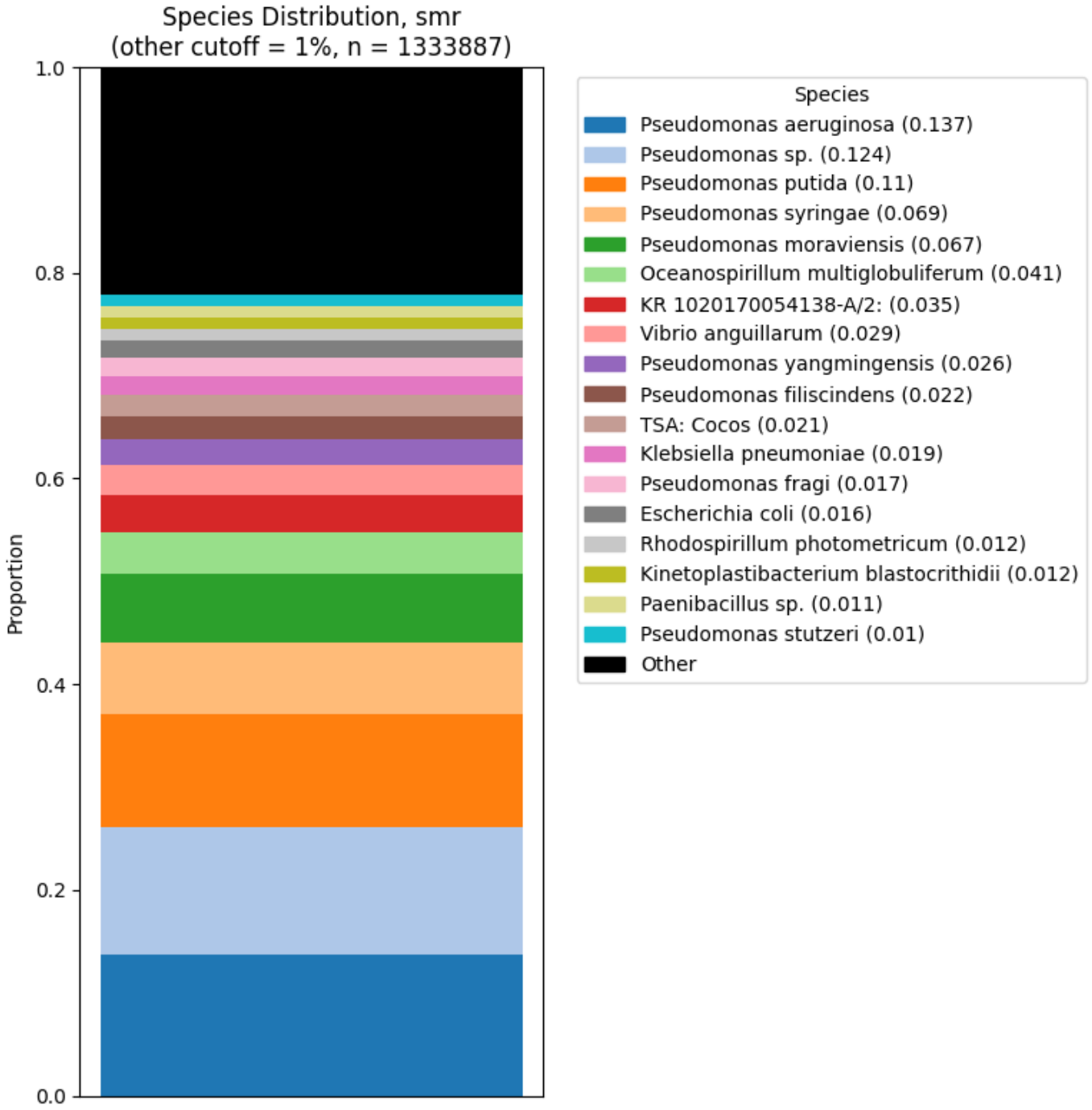
1) Assign and validate taxonomic identity of individual Ribo-ITP reads from pore water sample
2) Identify functional capabilities of the microbiome

Part 1: Taxonomic Classification



rRNA separation

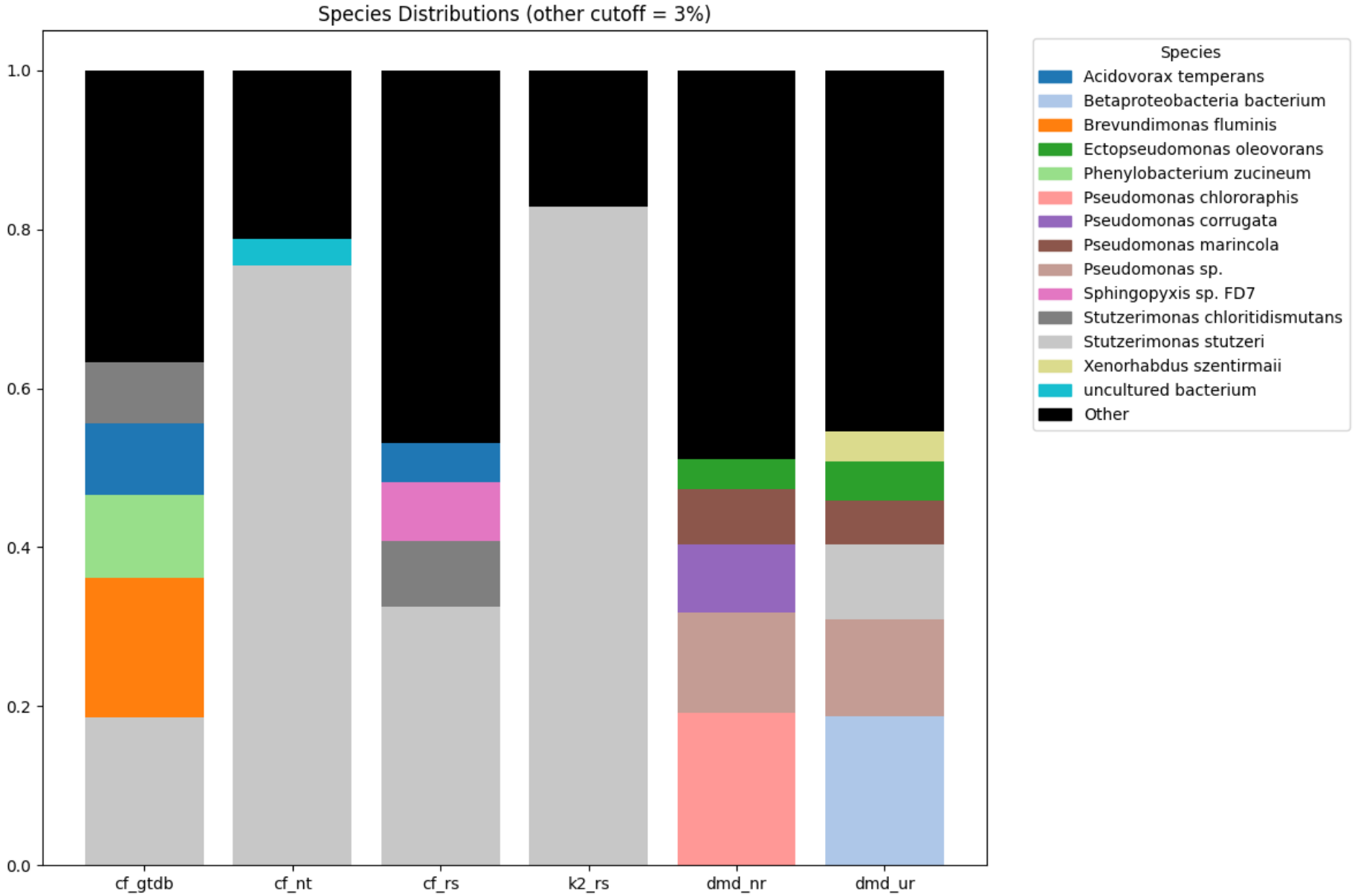
- SortMeRNA (Kopylova et al., 2012)
- 7219457 raw reads:
  - 1333887 (18.48%) classified as rRNA
  - 5885570 (81.52%) retained for downstream classification



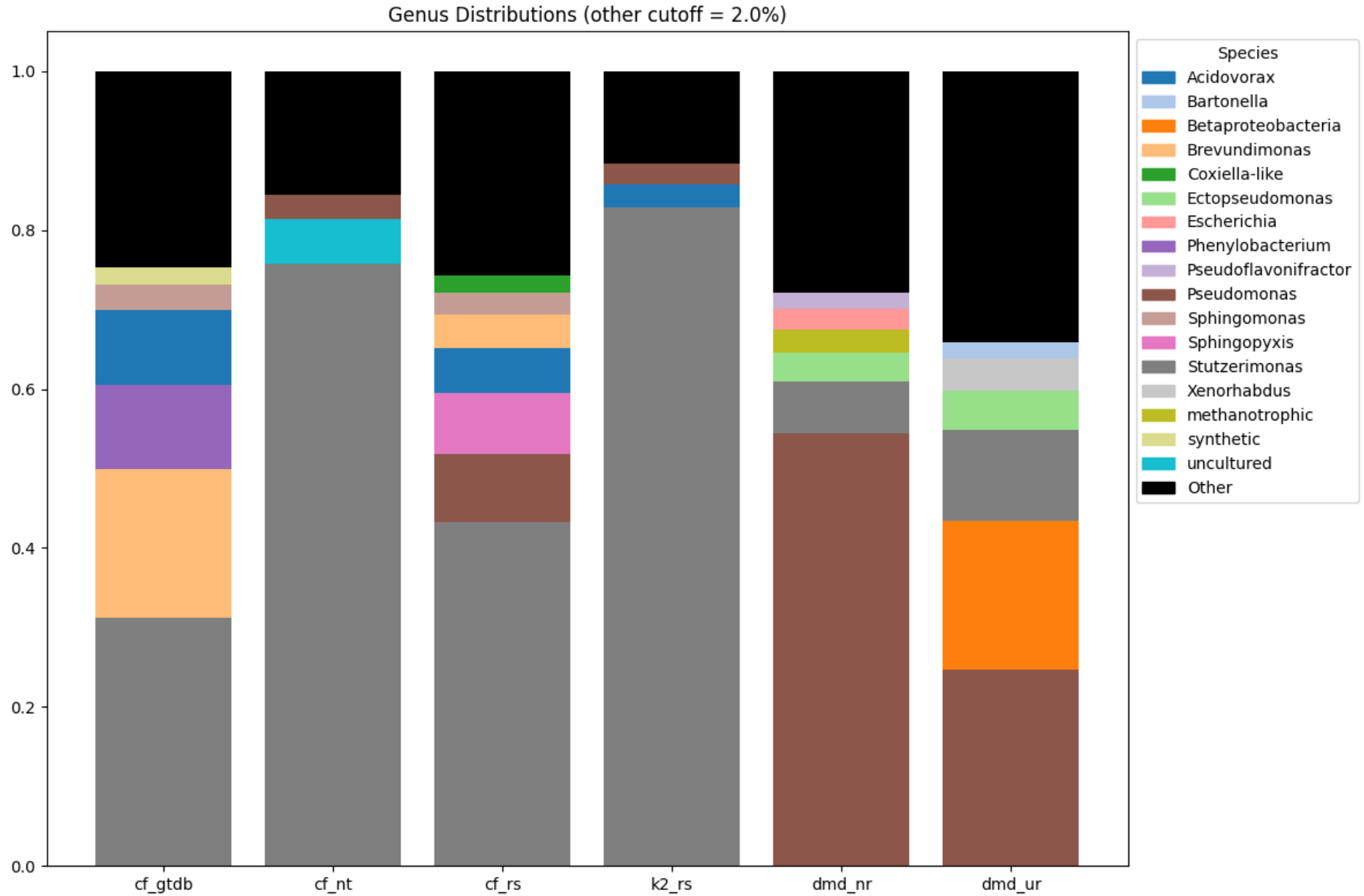
Classifiers and databases used for initial read assignments

- Nucleotide matching
  - Kraken2 (Wood et al., 2014)
    - NCBI RefSeq Bacteria/Human complete genomes (O'Leary et al., 2016)
  - Centrifuger (Song & Langmead, 2024)
    - NCBI RefSeq Bacteria/Human/Virus/Archea, GenBank SARS-CoV2
    - GTDB r226 + RefSeq Human/Virus/Fungi/Contaminants
    - NCBI nt
- Protein matching
  - DIAMOND (Buchfink et al. 2021)
    - NCBI nr
    - UniRef100

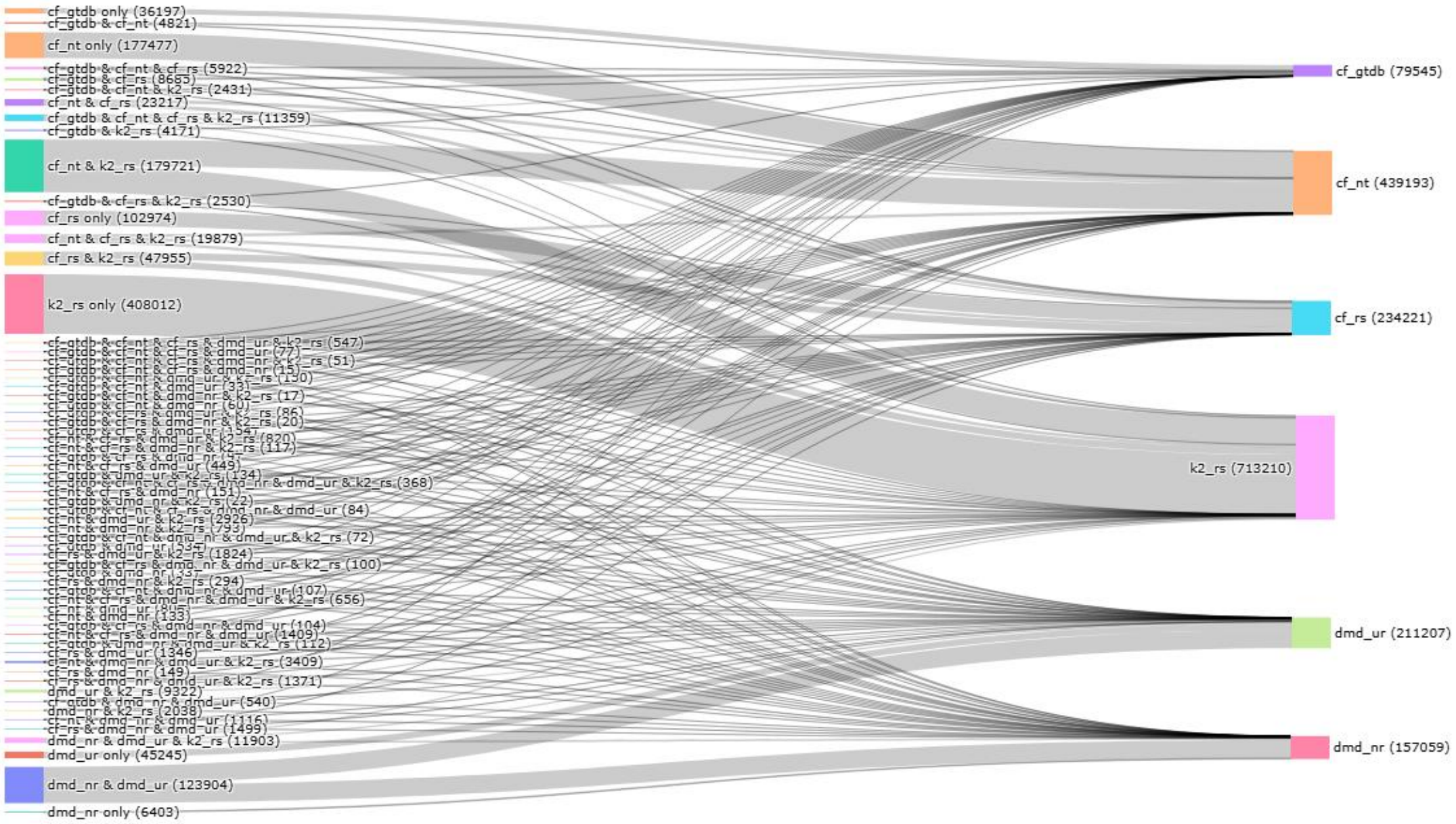
Initial classifier results, species



Initial classifier results, genera

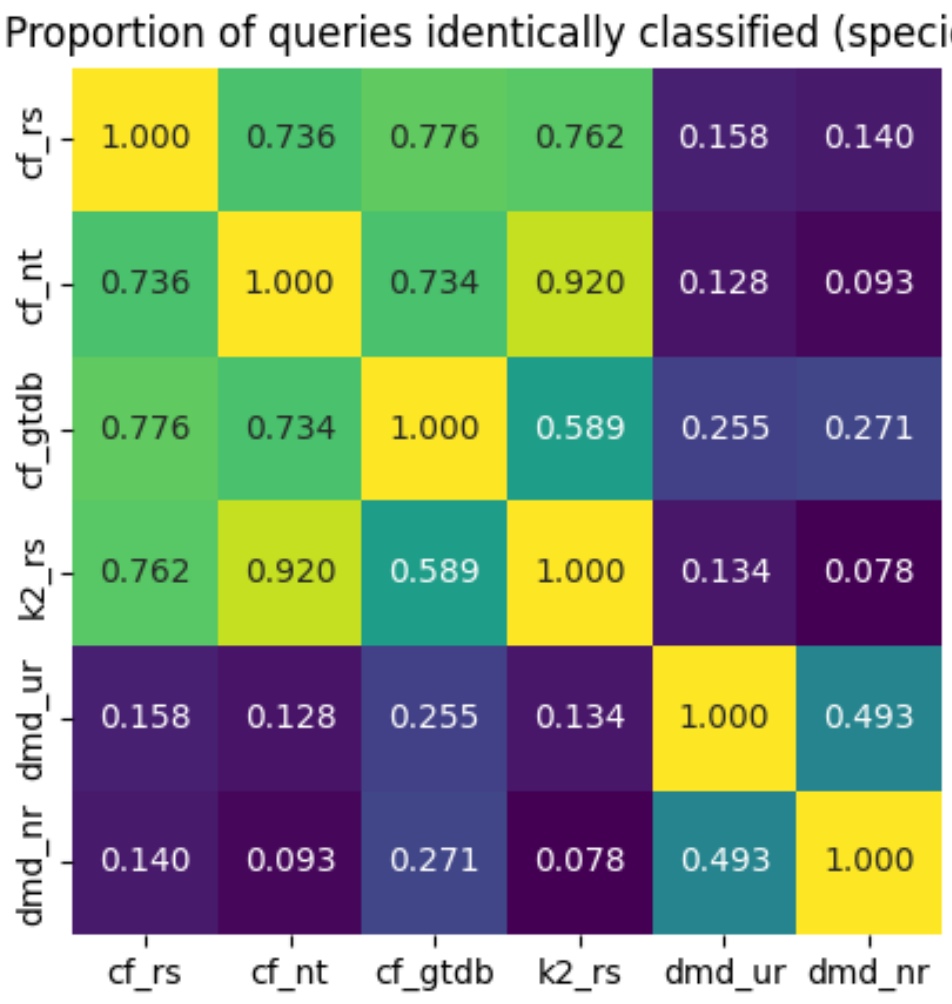


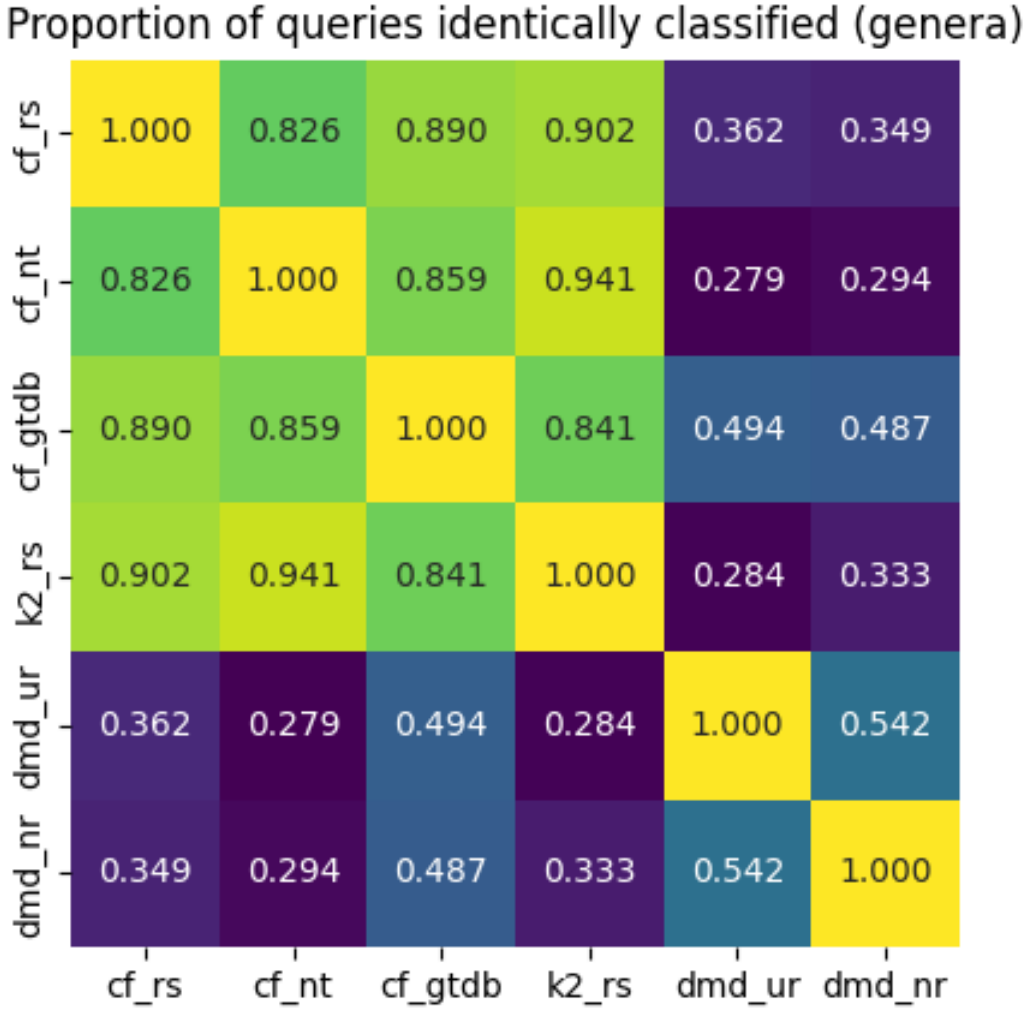
1256863 total reads classified



Classifiers show similar classifications

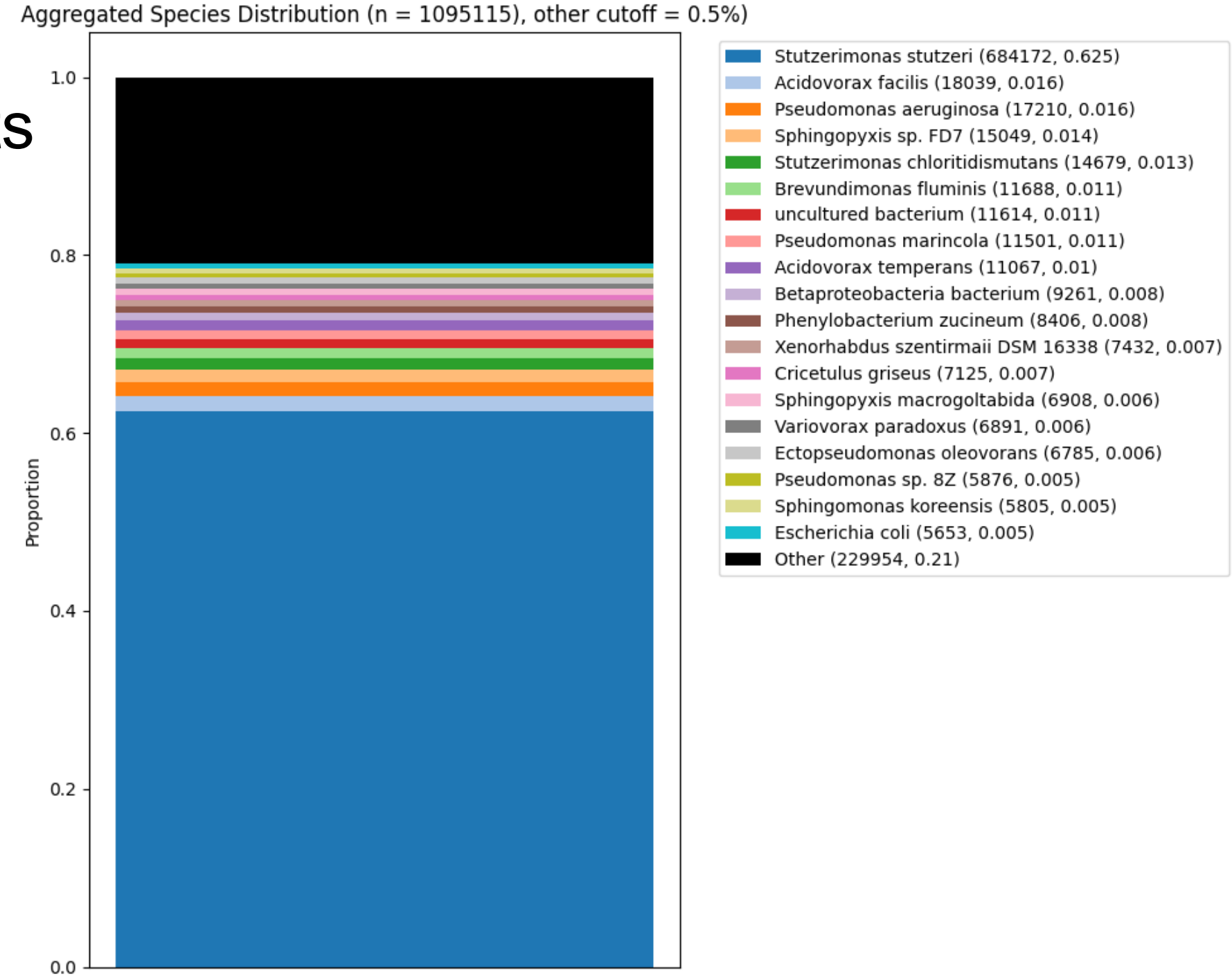
- Out of 480555 shared reads, 321598 (67%) agreed on species
- 354239 (74%) agreed on genus

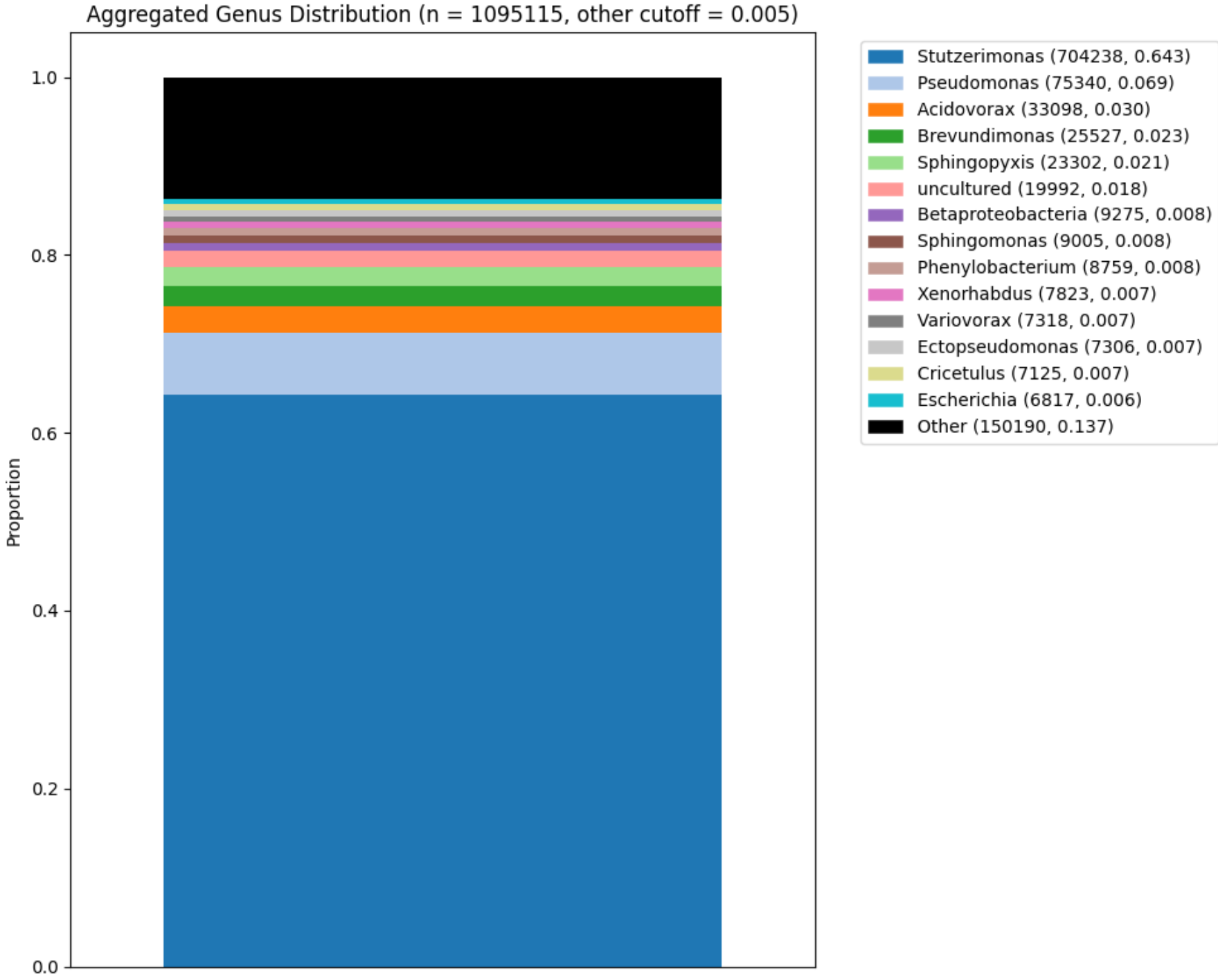




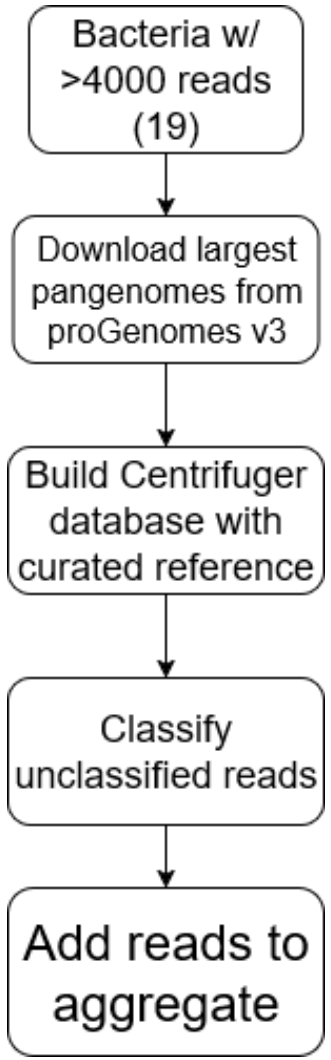
Aggregated classifier results

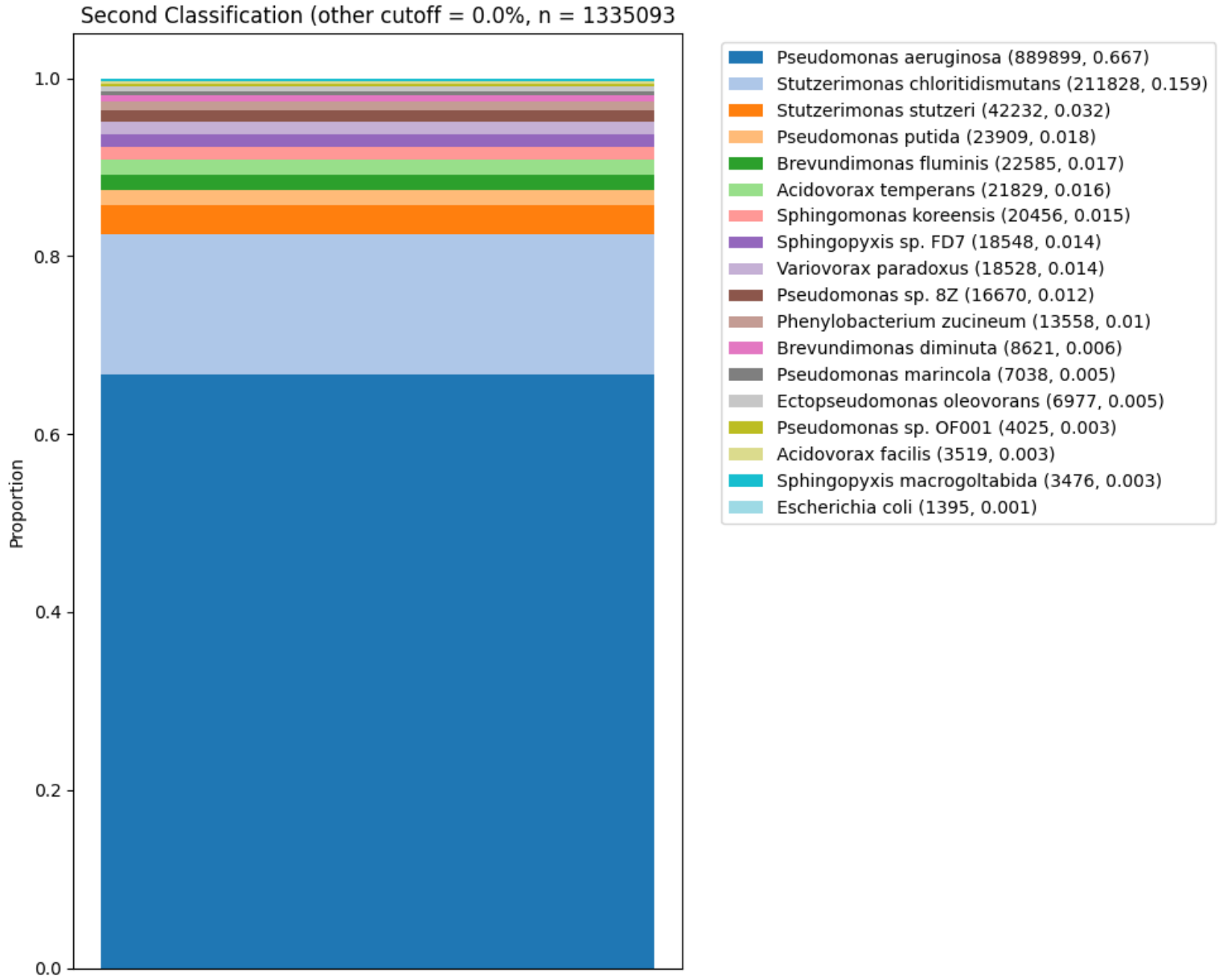
- Aggregated classifier information by combining single hits and unanimous hits
- 1095115/5885570 (18.6%) of raw reads classified




Aggregated Genus Distribution (n = 1095115, other cutoff = 0.005)
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
Stutzerimonas (704238, 0.643)
Pseudomonas (75340, 0.069)
Acidovorax (33098, 0.030)
Brevundimonas (25527, 0.023)
Sphingopyxis (23302, 0.021)
uncultured (19992, 0.018)
Betaproteobacteria (9275, 0.008)
Sphingomonas (9005, 0.008)
Phenylobacterium (8759, 0.008)
Xenorhabdus (7823, 0.007)
Variovorax (7318, 0.007)
Ectopseudomonas (7306, 0.007)
Cricetulus (7125, 0.007)
Escherichia (6817, 0.006)
Other (150190, 0.137)

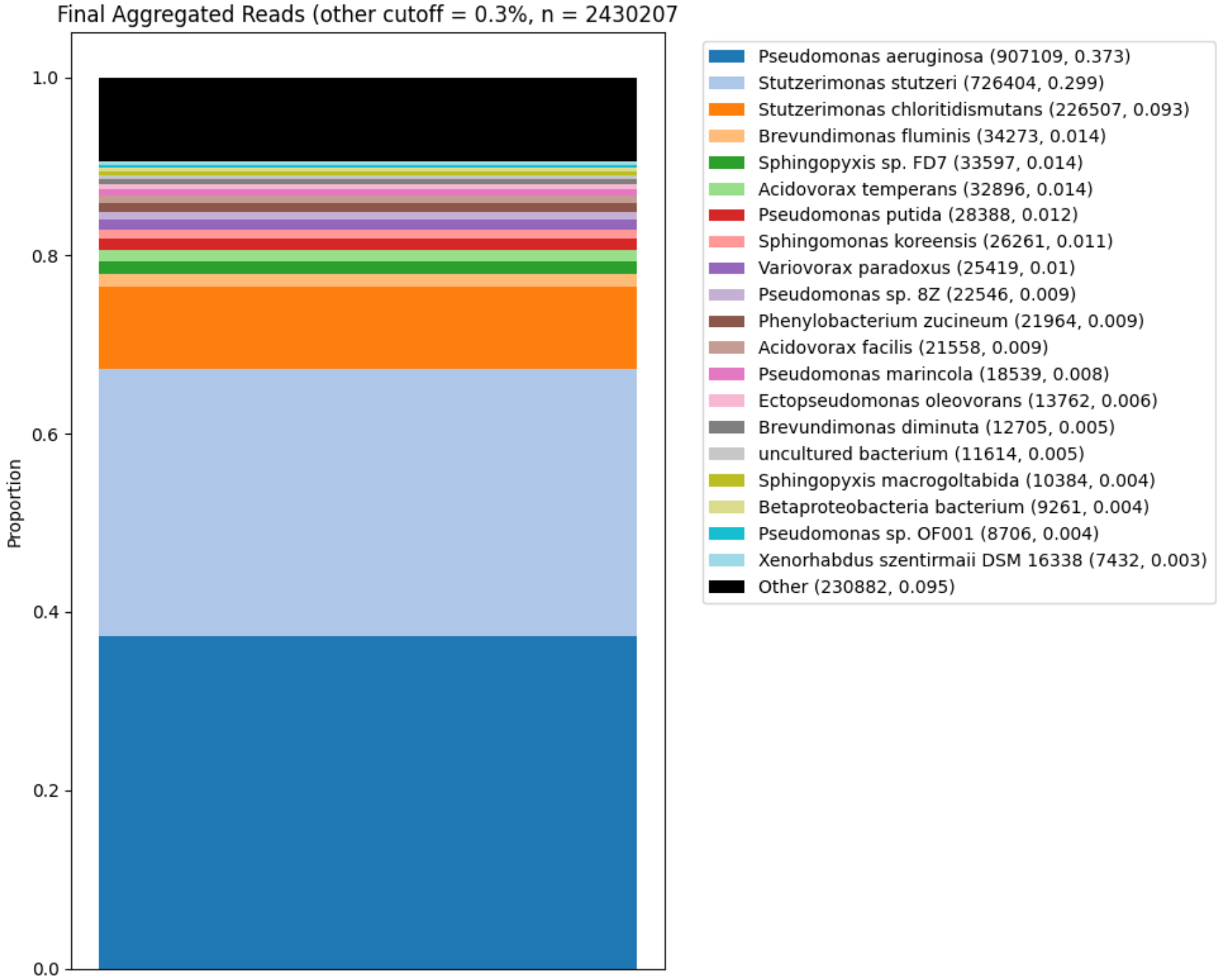
Second classification round



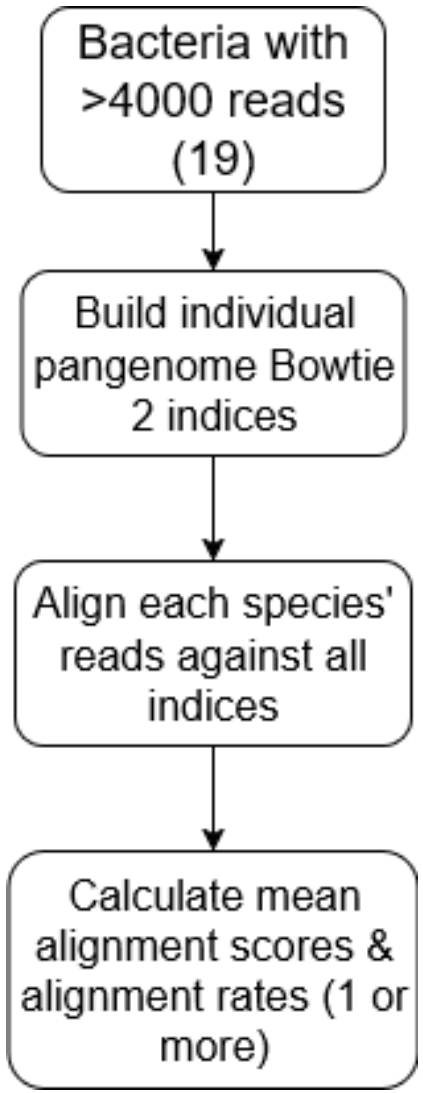


Final taxonomic results

- 2430207/5885570 (41.3%) of raw reads classified

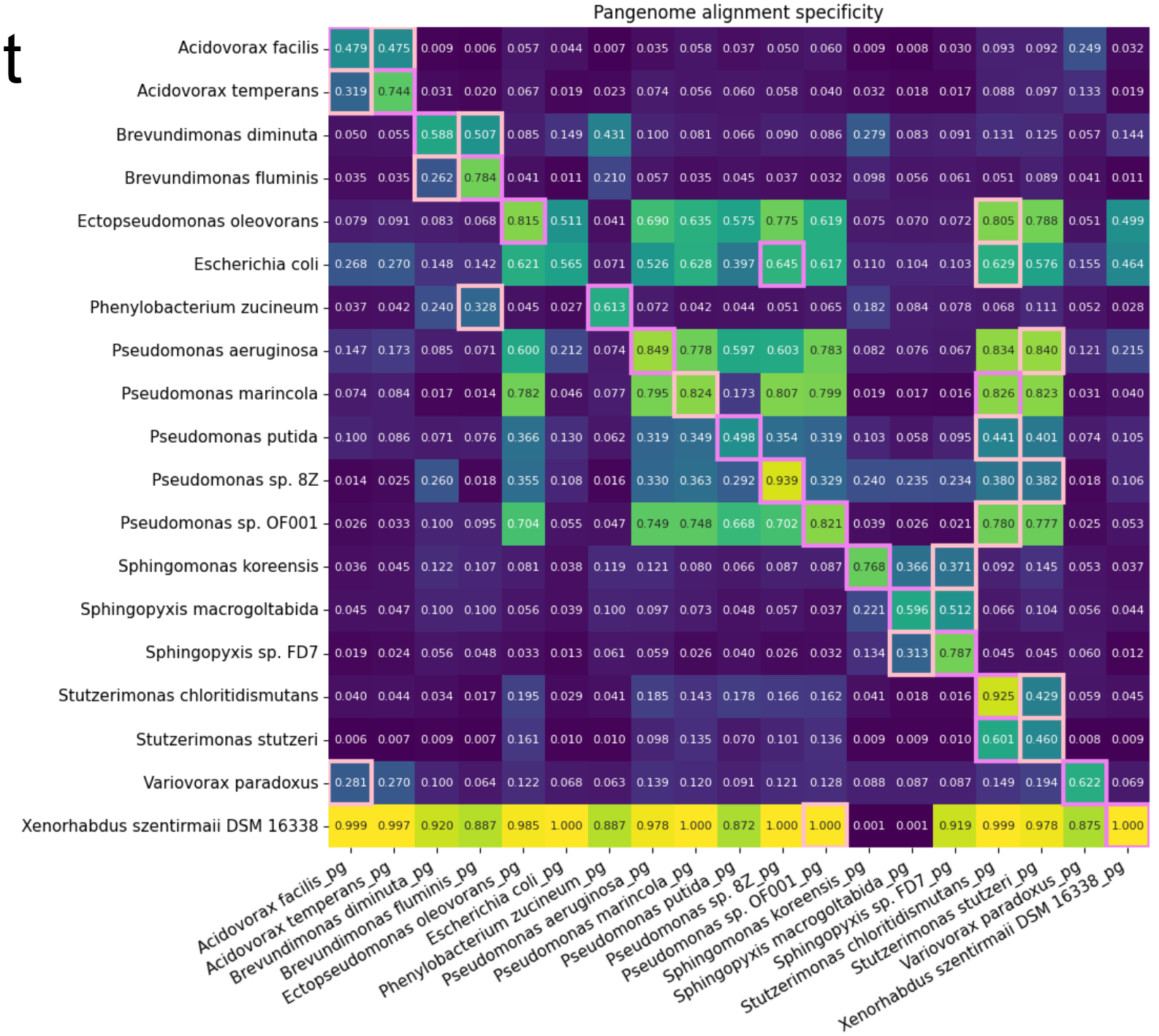


Classification validation



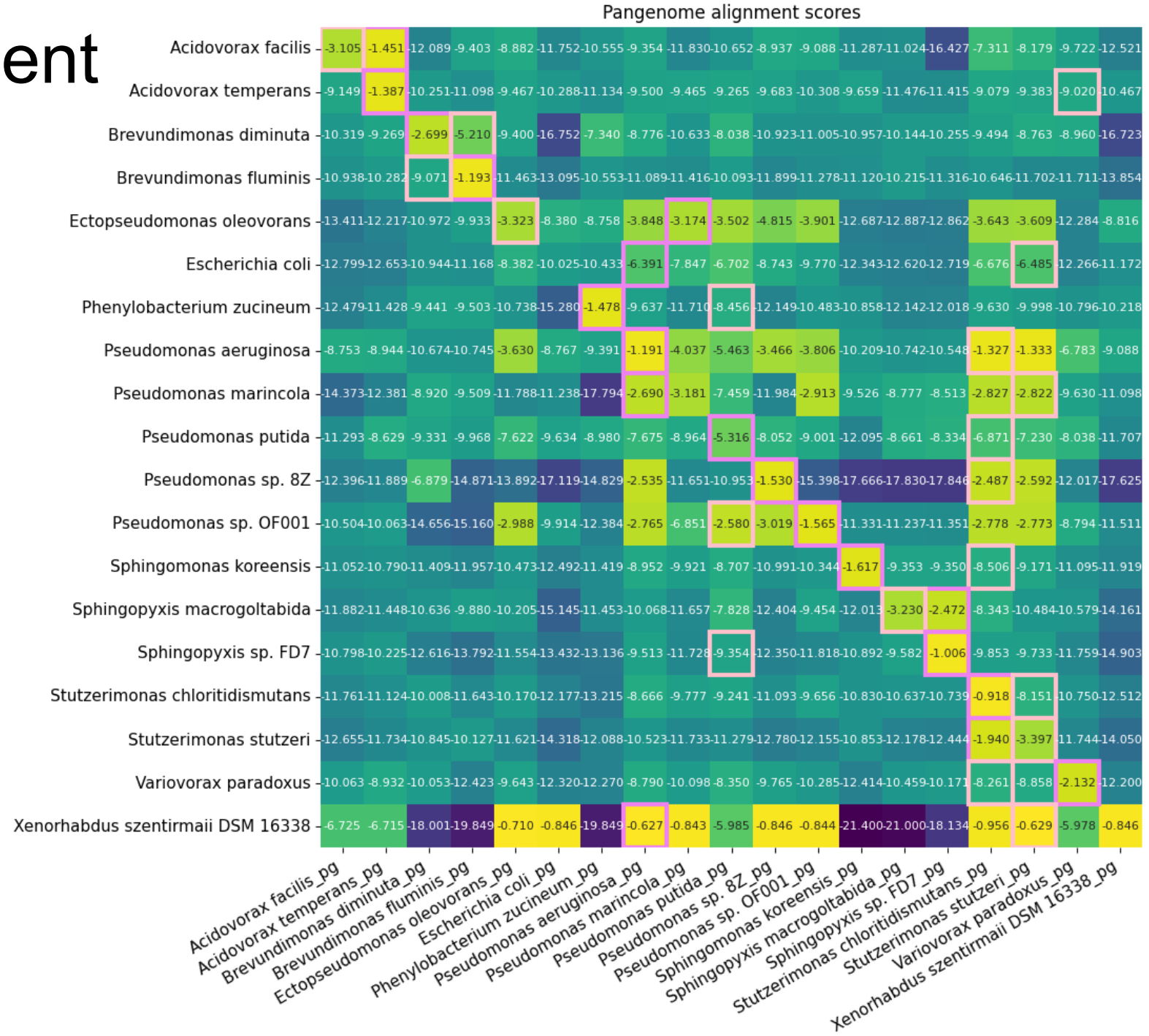
Specificity: Alignment Rate

- Classified reads generally have more alignments to their respective pangenomes



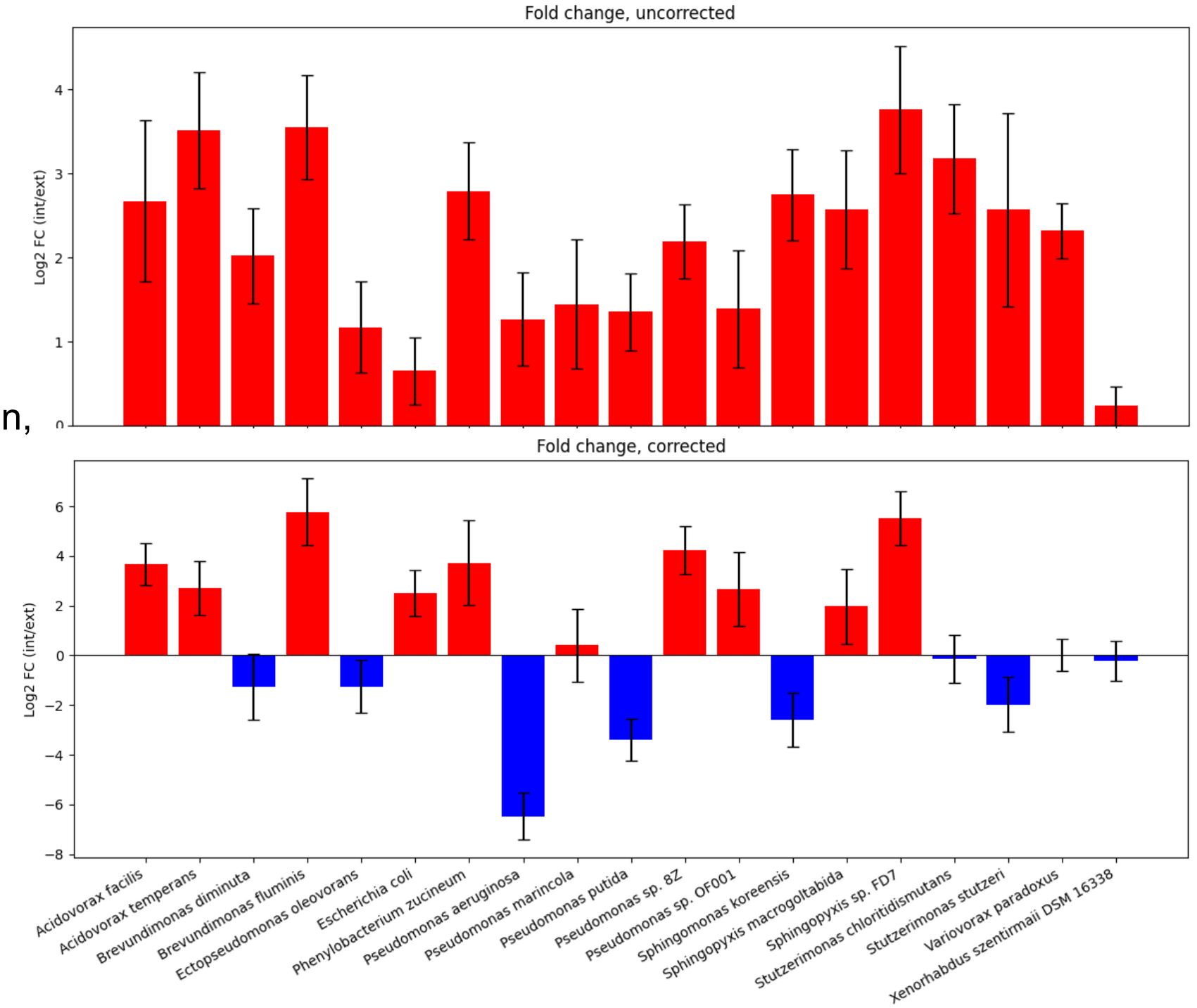
Specificity: Alignment Quality

- Classified reads generally align better to their respective pangenomes

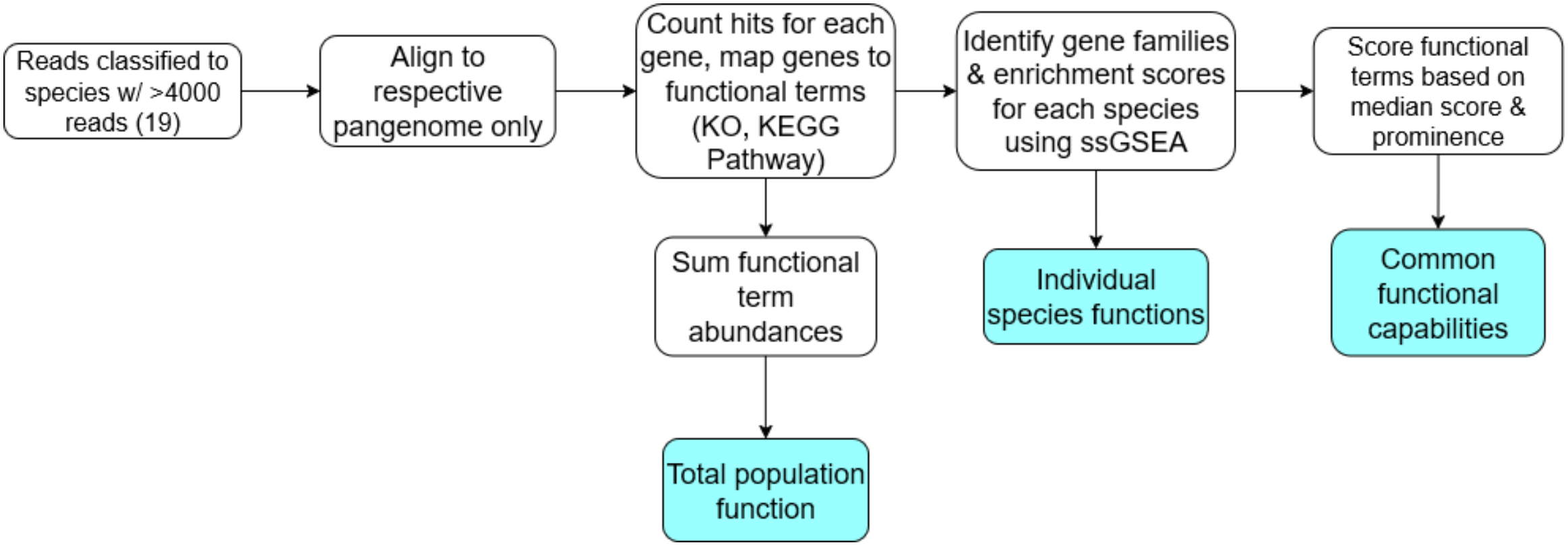


Specificity scores

- Classified reads generally align better to their respective pangenomes
- After pangenome size correction, species with more ambiguous assignments drop in score

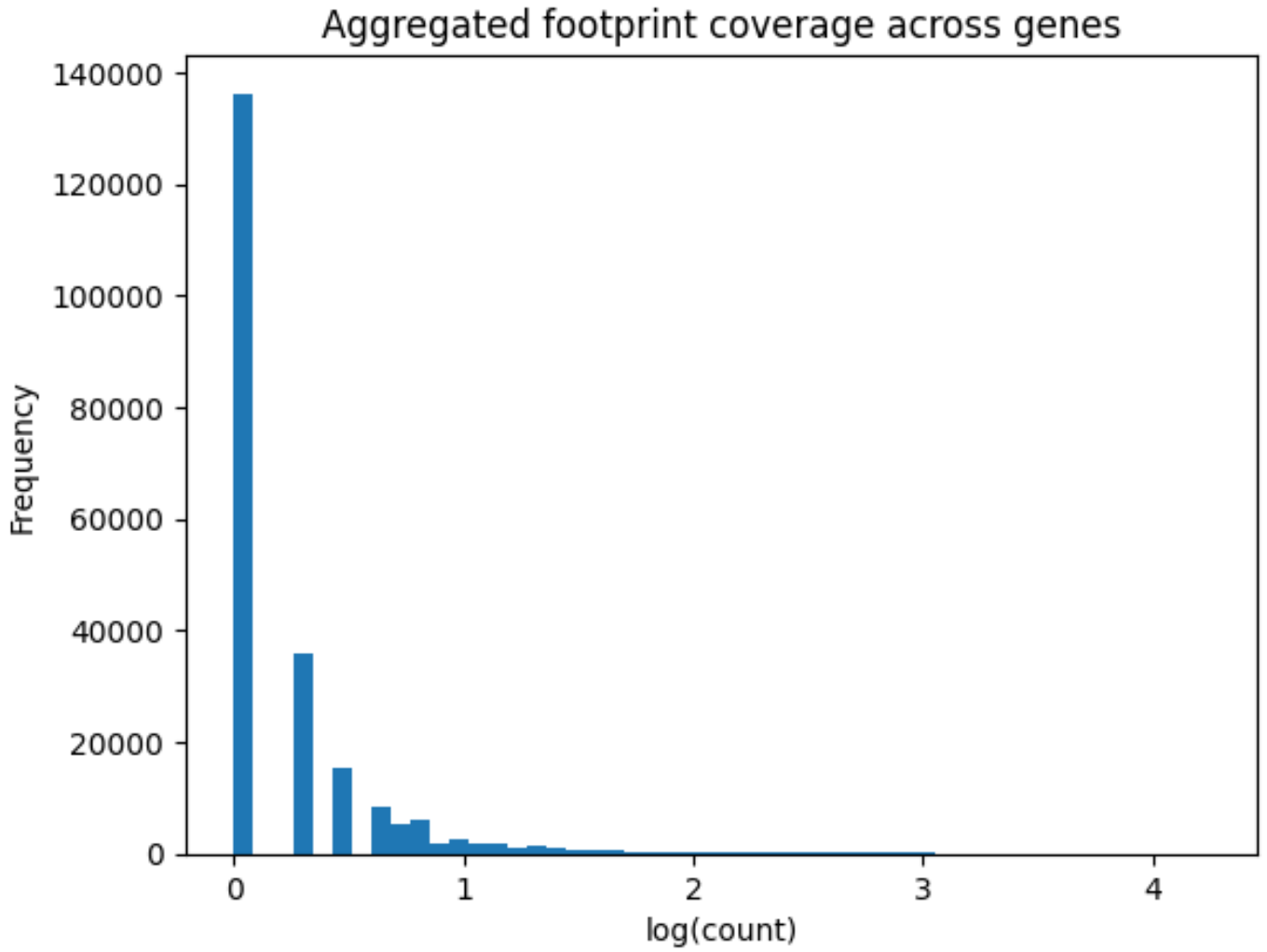


Part 2: Functional Analysis

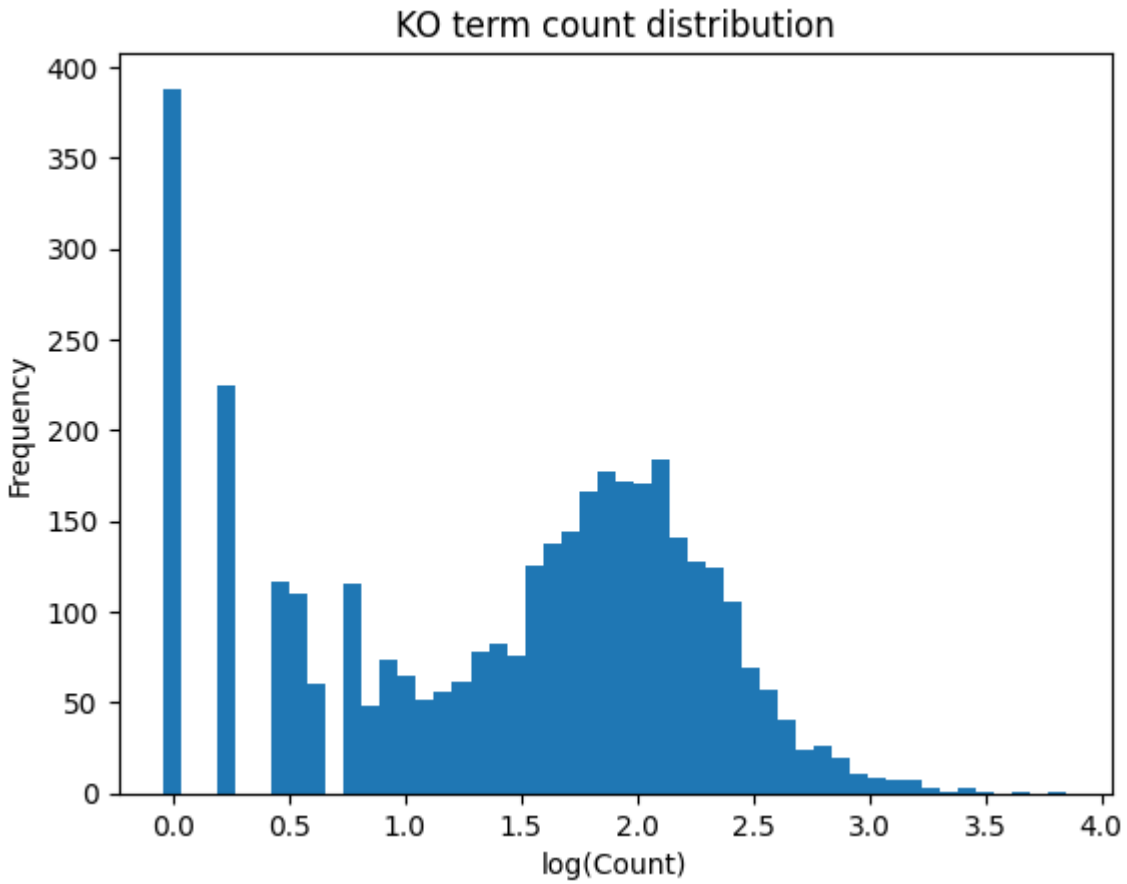


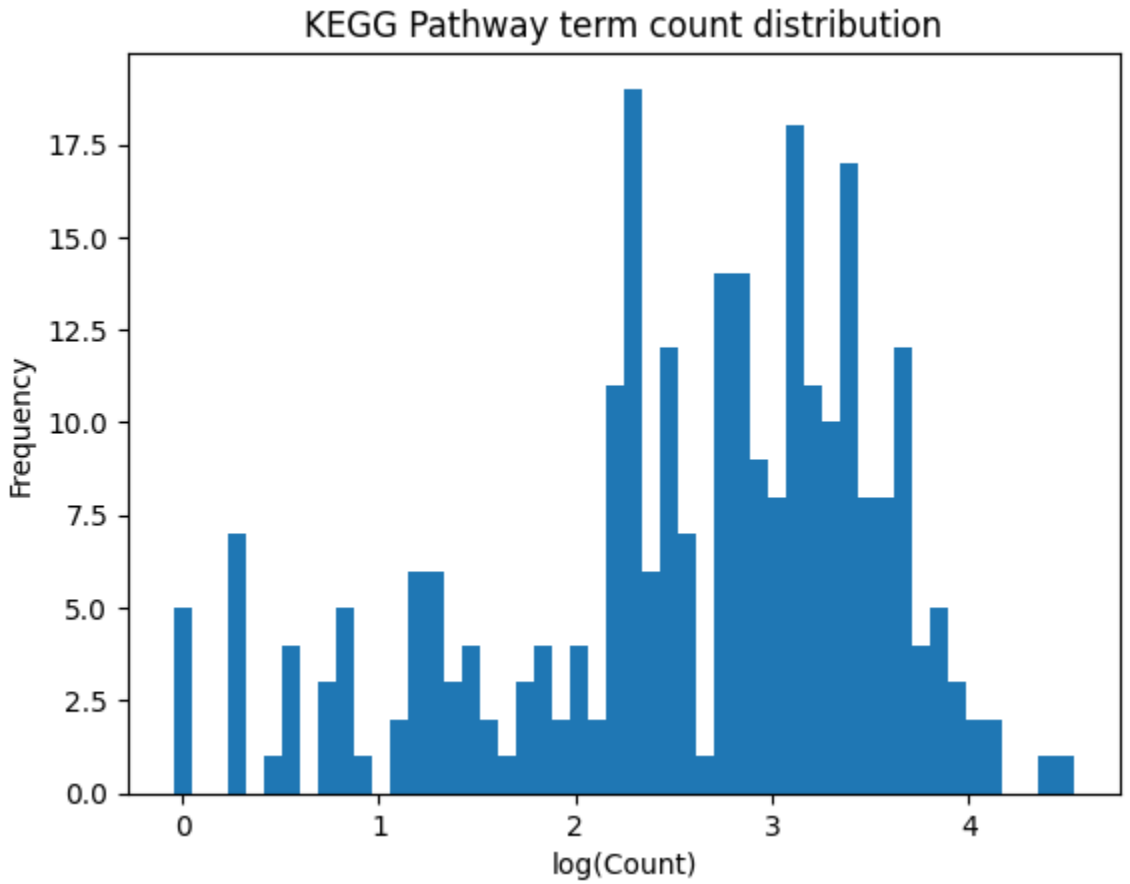
Gene Alignment Quality

- Reads follow expected distribution



Functional Term Quality

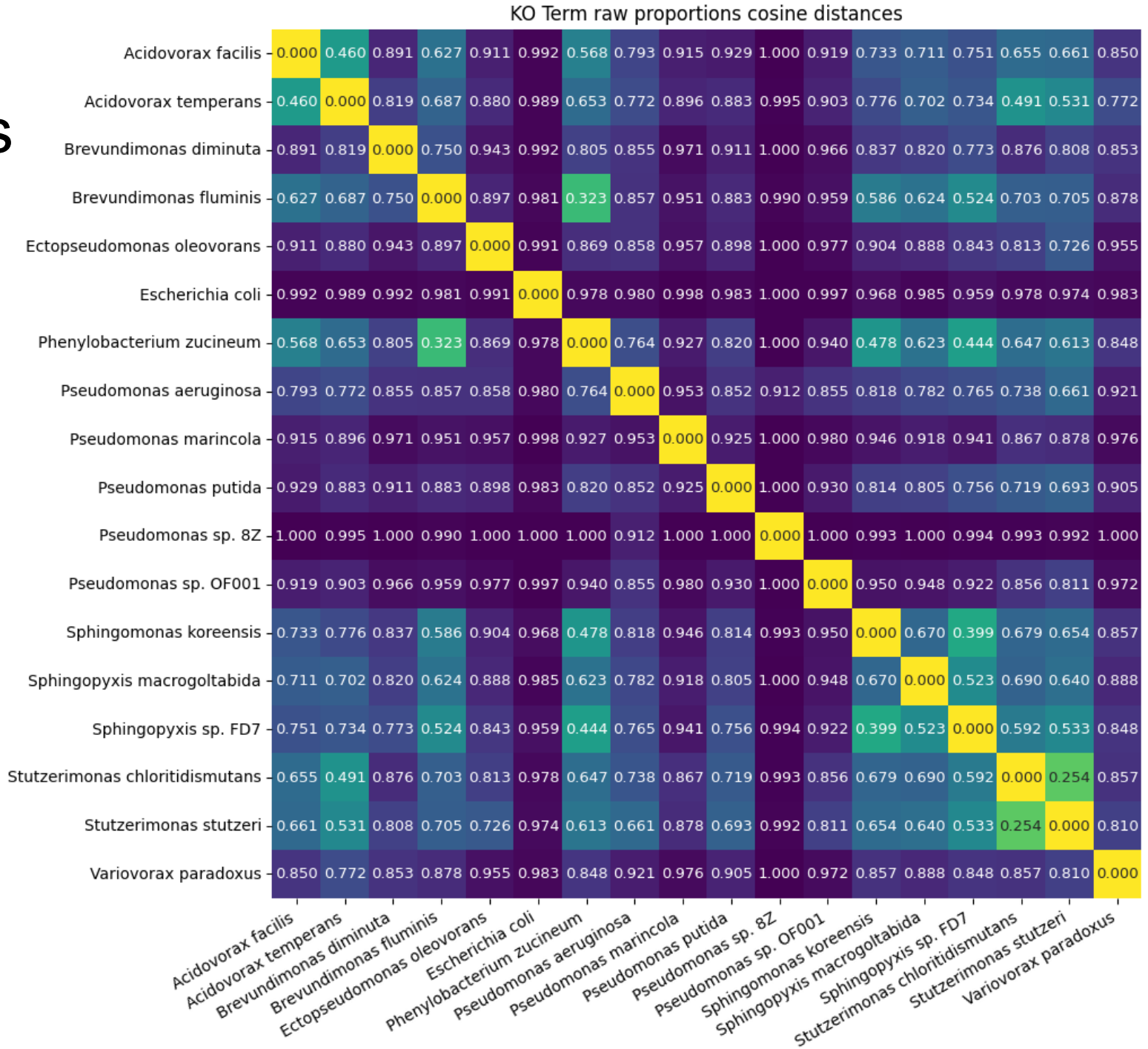




- Term counts follow expected distribution
- *X. szentirmaii* reads returned zero functional terms; dropped from analysis

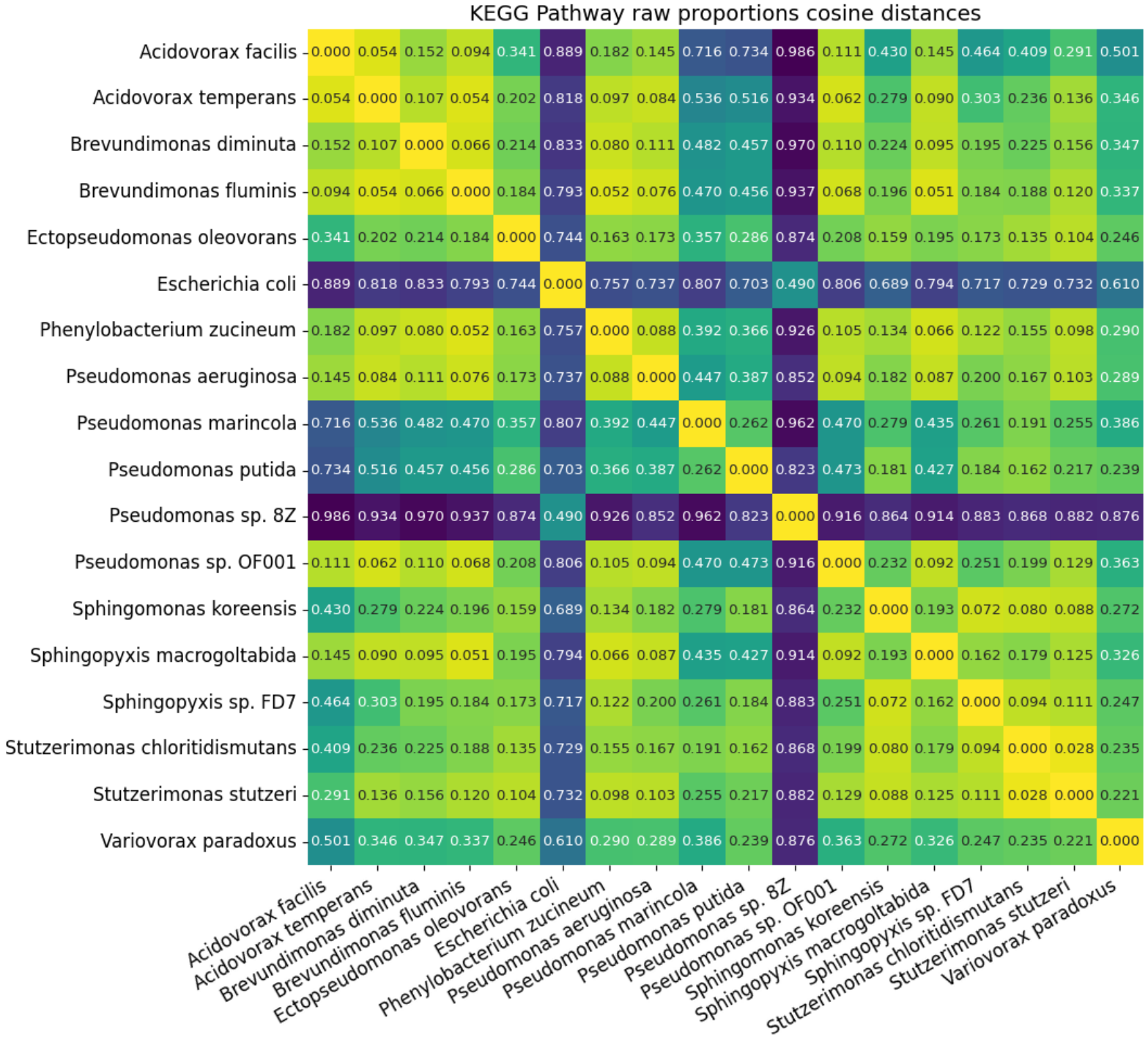
Functional Relatedness: KO Terms

- Species show diverse functional roles

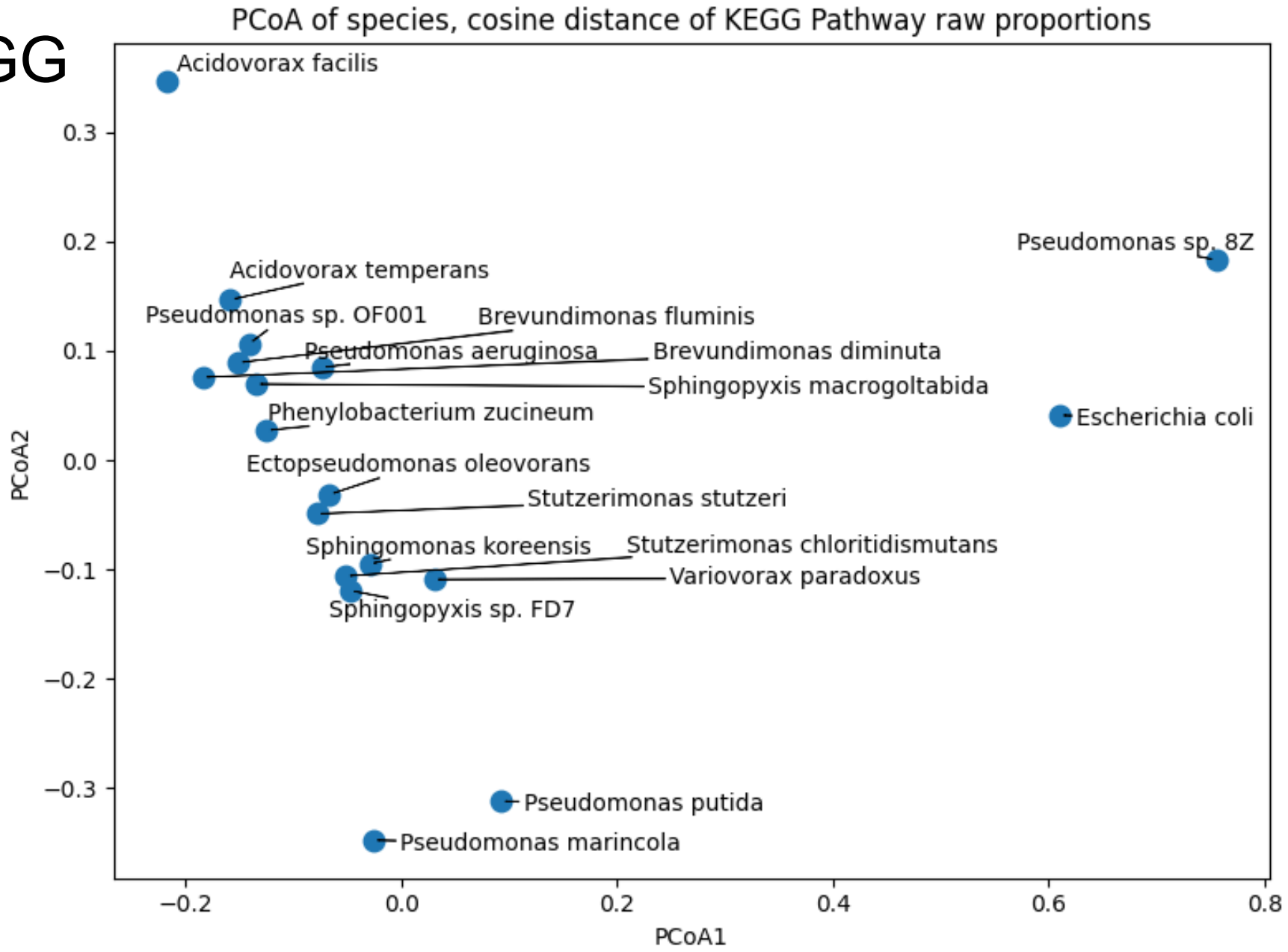


Functional Relatedness: KEGG Pathways

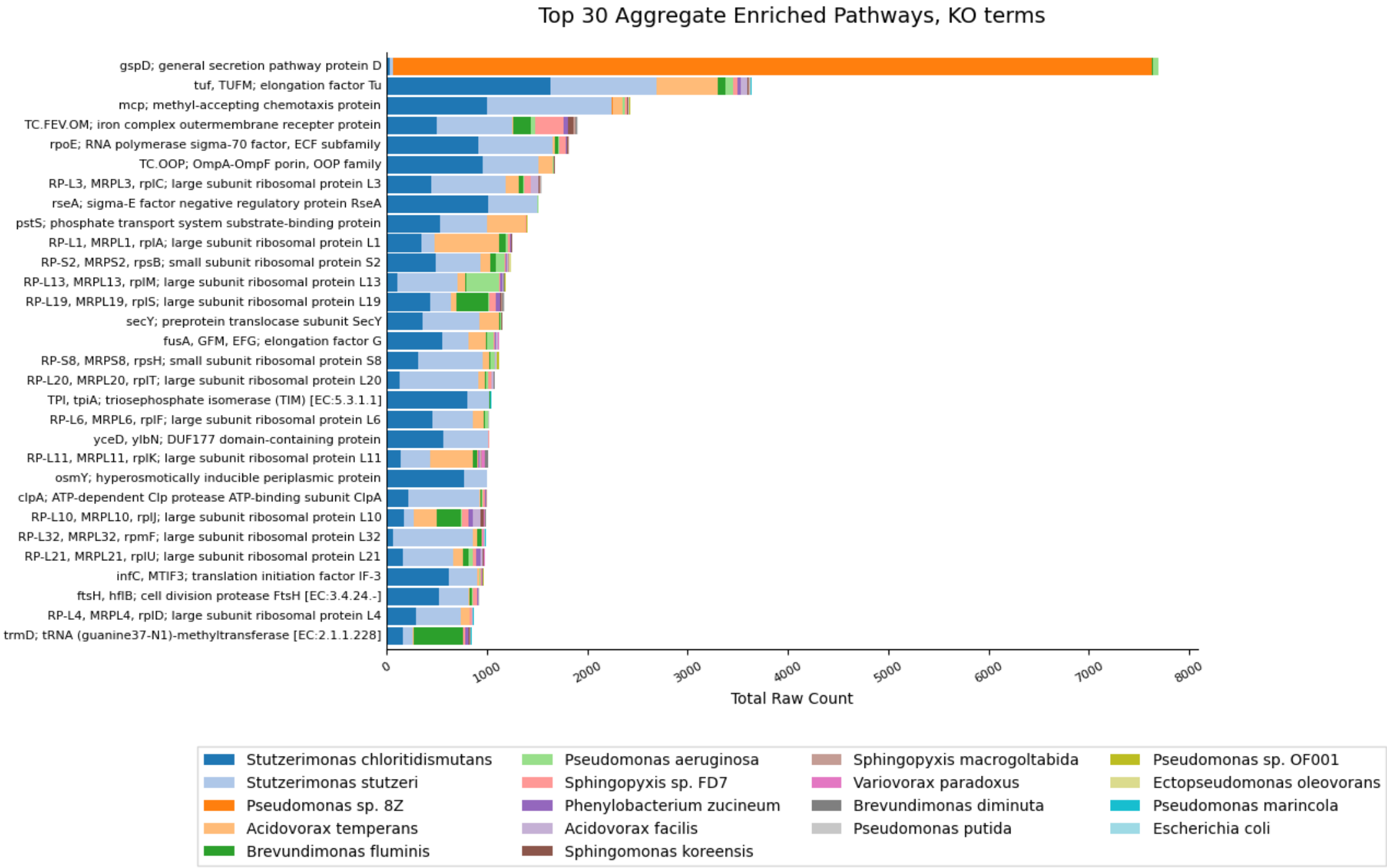
- Related species show similar functional roles
- *X. szentirmaii* reads returned zero functional terms; dropped from analysis



Functional Relatedness: KEGG Pathways

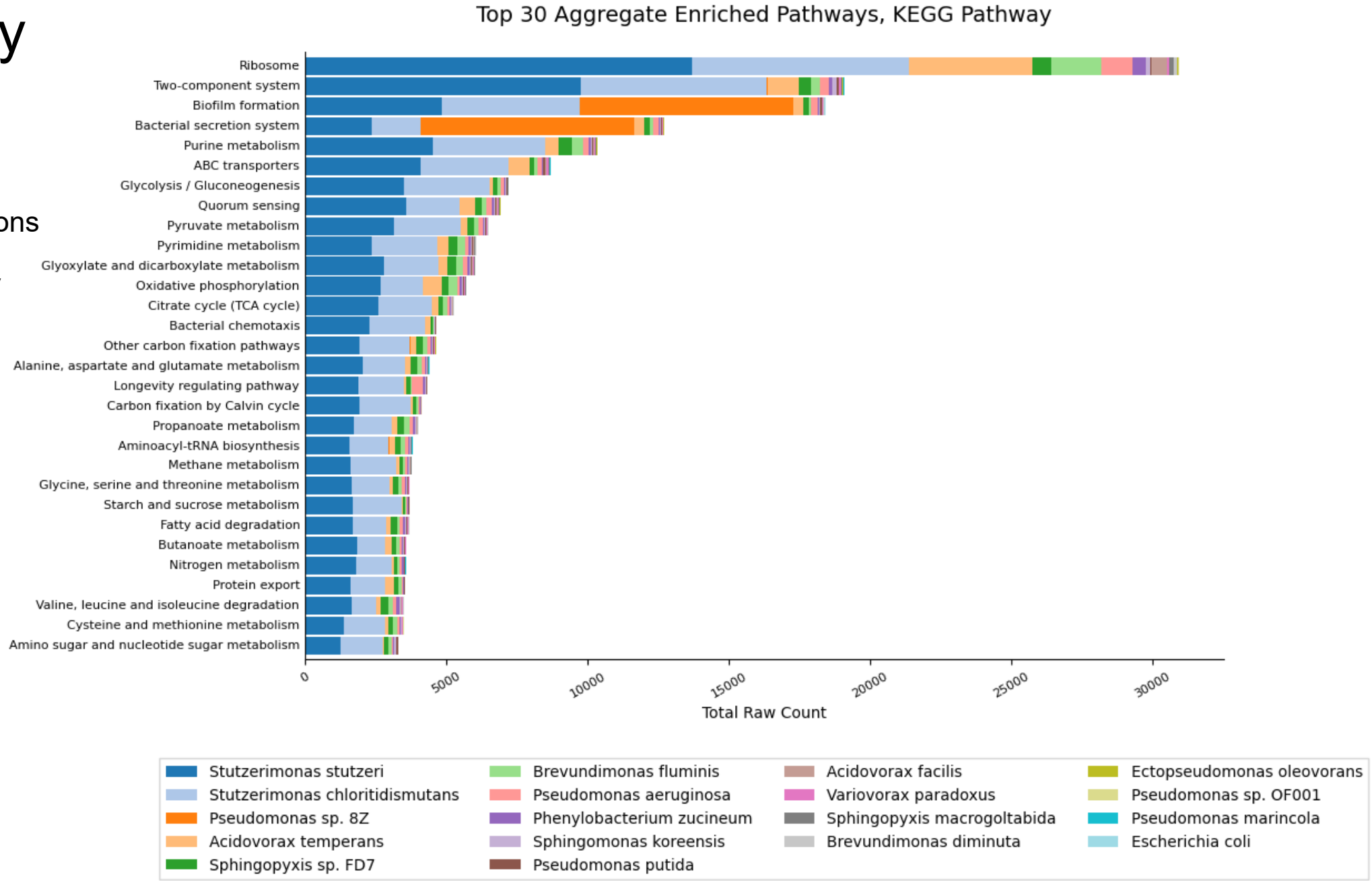


- Essential functions populate top KO terms

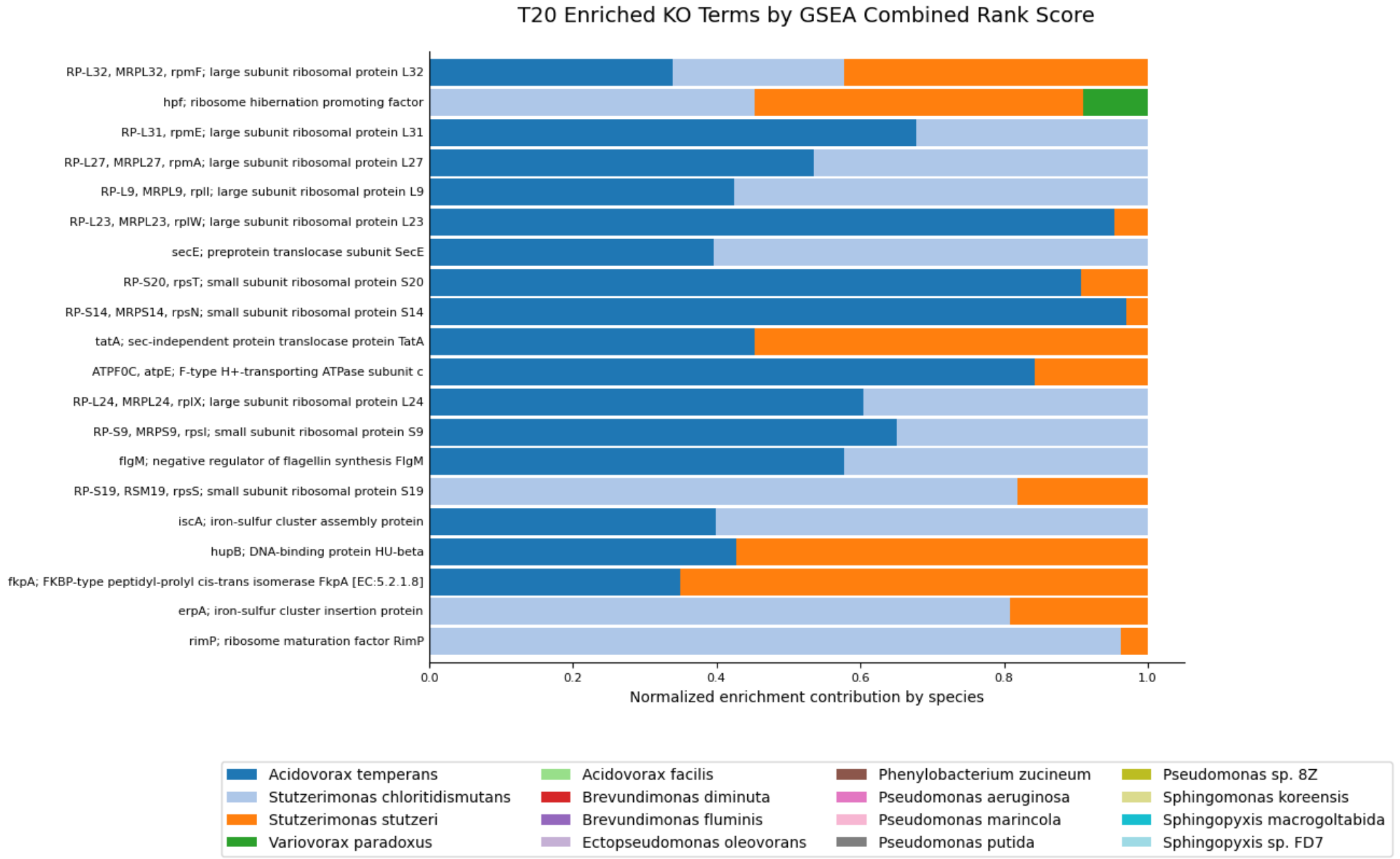


Aggregate community functions

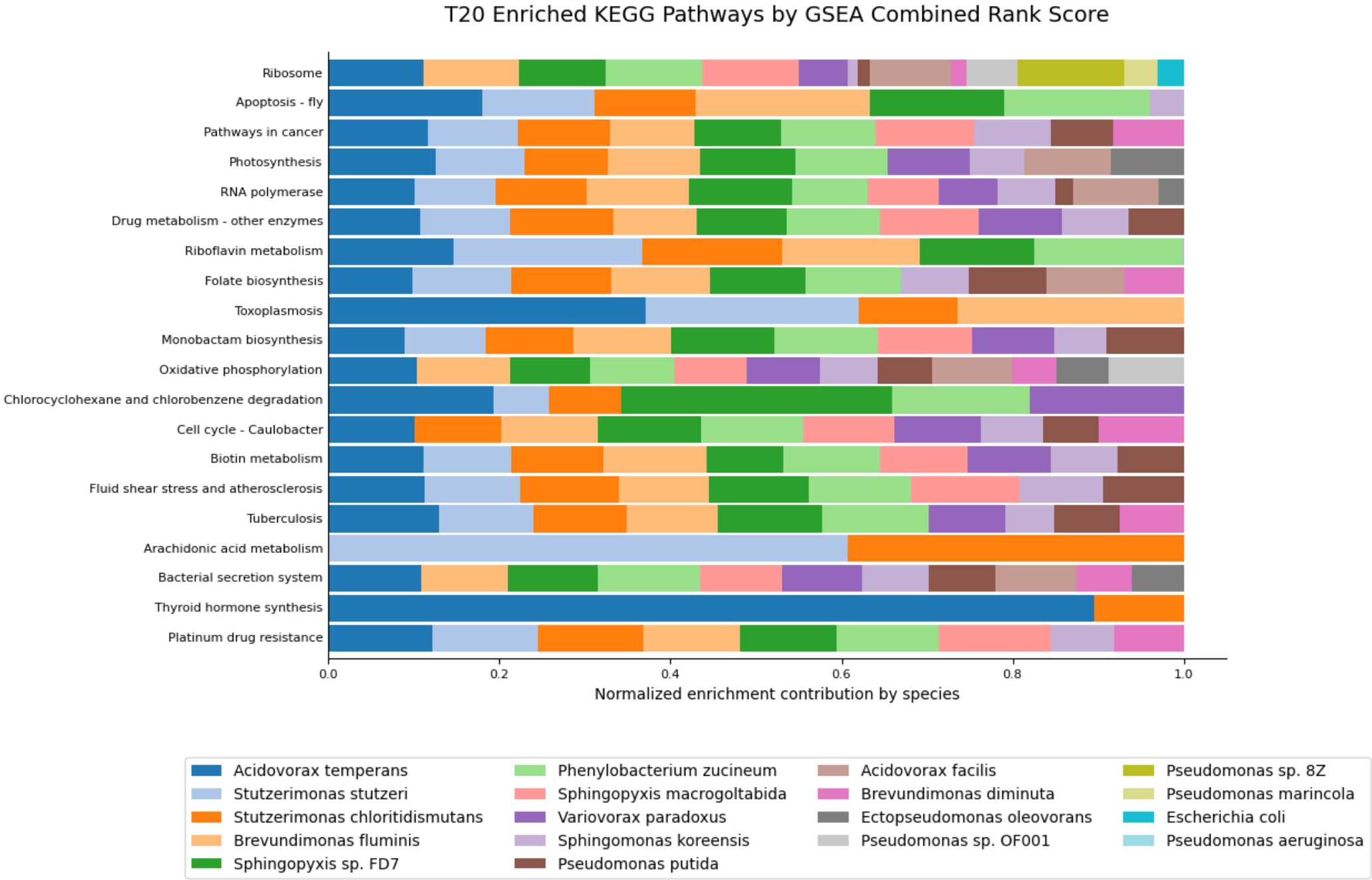
- Essential functions populate top KEGG Pathway terms



- KO Terms enrichment is usually not shared



- KEGG Pathway enrichment shows more similarity



Summary

- A multi-classifier approach is effective in characterizing environmental microbiome ribosome profiling data
- The microbial community translational activity is dominated by a few bacteria, particularly *S. stutzeri*, *P. aeruginosa*, and *S. chloriditismutans*
- Key community functions can be identified from ribosome profiling data

Limitations/Next Steps

- Ribo-seq is not suited for taxonomic abundance estimation
  - Integration with metagenomics data
- Ribo-seq only describes a portion of bacterial activity
  - Integration with metatranscriptomics data
- Many unmapped reads; only proceeded with taxa >4000 raw aggregate reads
- Pangenomes only partially describe a species
- Comparison is needed for significant conclusions
  - Differential analysis
  - Correlation with geochemical/environmental data

Questions

- Environmental sample results or expectations?
- Any specific bacterial processes of interest?

References

- Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
- Fullam, A. *et al.* proGenomes3: approaching one million accurately and consistently annotated prokaryotic genomes. *Nucleic Acids Res.* **51**, D760–D766 (2023).
- Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
- O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
- Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
- Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **52**, D10–D19 (2024).
- Song, L. & Langmead, B. Centrifuger: lossless compression of microbial genomes for efficient and accurate metagenomic sequence classification. *Genome Biol.* **25**, 106 (2024).
- Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).