sylph: metagenomic profiling and containment by statistical k-mer sketching

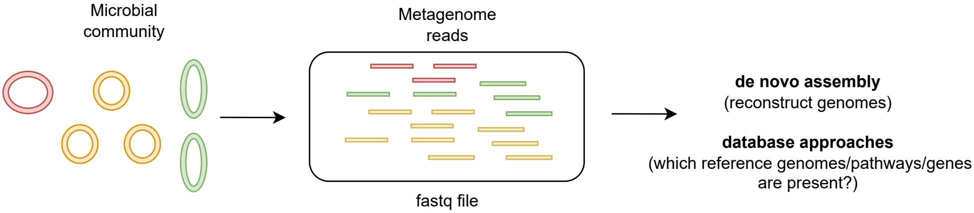
Jim Shaw[1] and Yun William Yu[2]

[1]University of Toronto [2]Carnegie Mellon University

Genome Informatics 2023

Shotgun sequencing of sample

Metagenomics - analyzing shotgun sequences of an environmental sample



Microbial community

Metagenome reads

fastq file

de novo assembly
(reconstruct genomes)

database approaches
(which reference genomes/pathways/genes are present?)

Focus of the talk

- **Database approach**

Focus of the talk

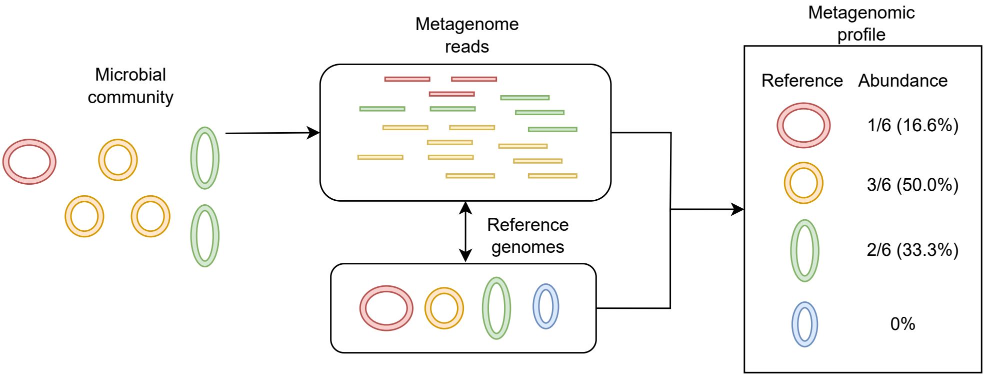
- **Database approach**
- What genomes are in my sample?

Focus of the talk

- **Database approach**
- What genomes are in my sample?
- Two distinct but similar approaches: **profiling** and **containment**

What is metagenomic profiling?

Profiling: **What taxa** are in the community and **how abundant** are they**?**



Microbial community

Metagenome reads

Reference genomes

Metagenomic profile

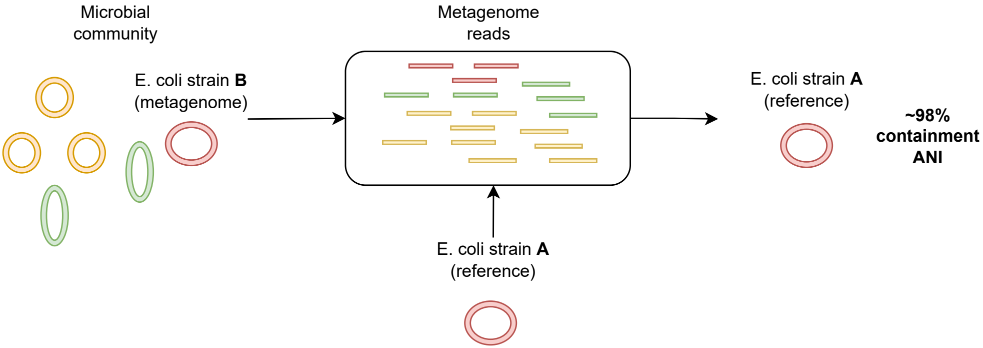
Reference	Abundance
🔴	1/6 (16.6%)
🟡	3/6 (50.0%)
🟢	2/6 (33.3%)
🔵	0%

What is containment estimation?

Average nucleotide identity (ANI): genome-to-genome similarity
containment ANI: genome-to-metagenome similarity (nearest neighbor)

Containment is a continuous measure

Containment: How **similar** is a genome to the genomes in the community?



Microbial community

E. coli strain **B** (metagenome)

Metagenome reads

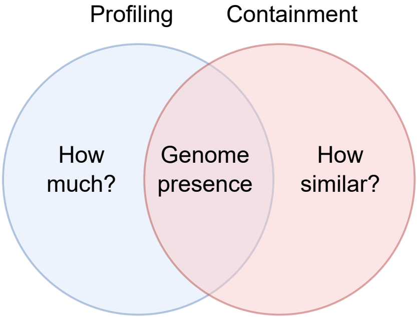
E. coli strain **A** (reference)

E. coli strain **A** (reference)

~98% containment ANI

Profiling vs containment

1. **Profiling** - relative abundances of genomes/taxa
2. **Containment** - nucleotide similarity of genome within metagenome



sylph - containment and profiling

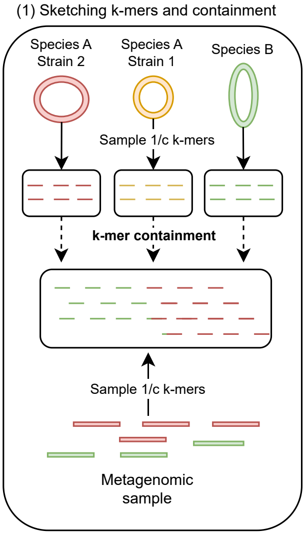
We present **sylph**, a new *profiler* with *containment* capabilities

sylph part 1 - k-mer sketching

Given a genome or a metagenome:

- Take all k-mers, sample only $1/c$ of them using FracMinHash (Irber et al., 2022). Default $c = 200$.

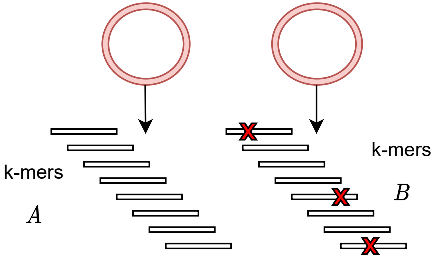
sylph sketching



Estimating containment ANI

k-mers differ between different genomes

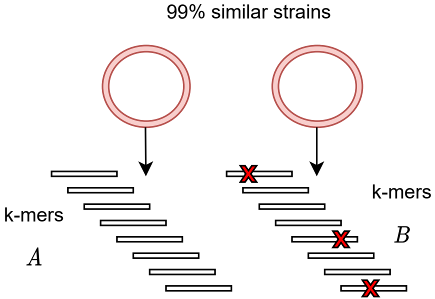
99% similar strains



Estimating containment ANI

Fact:

$$ANI \approx \left(\frac{|A \cap B|}{|A|}\right)^{1/k}$$



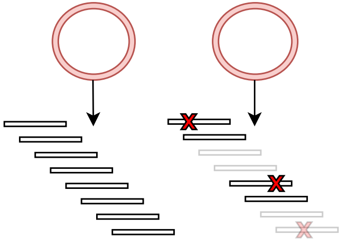
Low abundance genomes

Problem: this model had no **read sampling**. k-mers are missing in low abundance genomes due to read sampling

Low abundance genomes

$|A \cap B|/|A|$ **underestimates** ANI when k-mers are missing!

99% similar strains



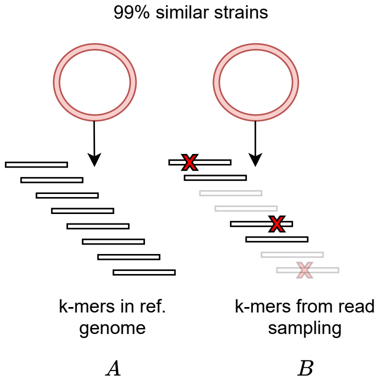
k-mers in ref. genome

k-mers from read sampling

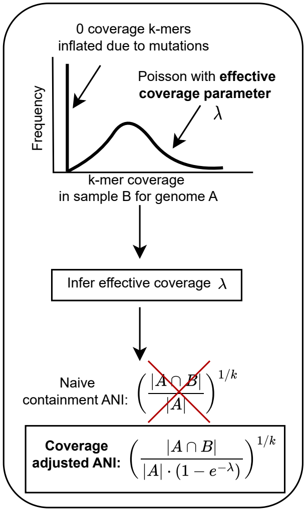
A

B

Zero-inflated Poisson model



(2) k-mer ANI with coverage adjustment



0 coverage k-mers inflated due to mutations
Frequency
Poisson with **effective coverage parameter** λ
k-mer coverage in sample B for genome A
Infer effective coverage λ
Naive containment ANI:

$$\cancel{\left(\frac{|A \cap B|}{|A|}\right)}^{1/k}$$

Coverage adjusted ANI:

$$\left(\frac{|A \cap B|}{|A| \cdot (1 - e^{-\lambda})}\right)^{1/k}$$

Coverage adjustment

1. Estimate the true coverage parameter λ

Coverage adjustment

1. Estimate the true coverage parameter λ
2. $\hat{\lambda} = \frac{\text{\# k-mers with multiplicity a+1}}{\text{\# k-mers with multiplicity a}} \cdot (a+1)$ (similar to Skmer, Sarmashghi et al., 2019)

Coverage adjustment

1. Estimate the true coverage parameter λ
2. $\hat{\lambda} = \frac{\text{\# k-mers with multiplicity a+1}}{\text{\# k-mers with multiplicity a}} \cdot (a+1)$ (similar to Skmer, Sarmashghi et al., 2019)
3. Coverage adjusted ANI:

$$\left(\frac{|A \cap B|}{|A| \cdot (1 - e^{-\hat{\lambda}})} \right)^{1/k}$$

Coverage adjustment

1. Estimate the true coverage parameter λ
2. $\hat{\lambda} = \frac{\text{\# k-mers with multiplicity a+1}}{\text{\# k-mers with multiplicity a}} \cdot (a+1)$ (similar to Skmer, Sarmashghi et al., 2019)
3. Coverage adjusted ANI:

$$\left(\frac{|A \cap B|}{|A| \cdot (1 - e^{-\hat{\lambda}})} \right)^{1/k}$$

4. Intuition: small coverage $\implies$ denominator is small, pushes ANI upwards

Coverage adjusted ANI - synthetic experiment

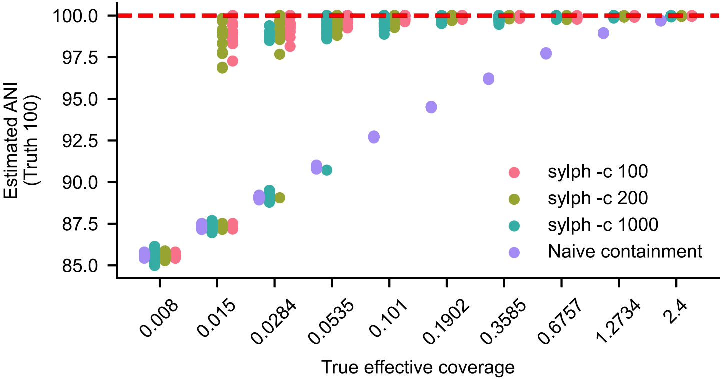


Figure: Containment ANI against synthetic reads from a *Klebsiella pneumoniae* genome.

Coverage adjusted ANI - real experiment

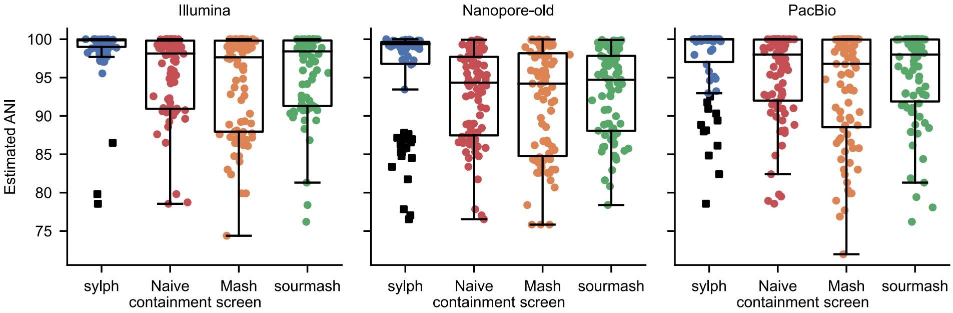


Figure: **Real** reads for mock metagenome from Meslier et al. (2022) with known references; black squares have uncorrected ANI.

sylph can do containment... but profiling?

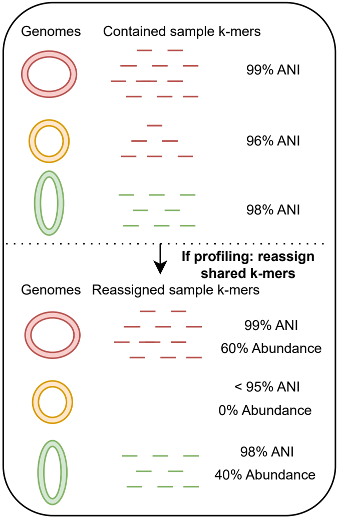
- containment ANI - doesn't say **how abundant** a microbe is

sylph can do containment… but profiling?

- containment ANI - doesn't say **how abundant** a microbe is
- **Problem**: k-mers are shared between genomes… which genome does a k-mer belong to?

Reassigning k-mers for profiling

(3) Top - ANI querying (sylph **query**)
Bot - Taxonomic profiling (sylph **profile**)



Genomes | Contained sample k-mers

99% ANI

96% ANI

98% ANI

If profiling: reassign shared k-mers

Genomes | Reassigned sample k-mers

99% ANI
60% Abundance

< 95% ANI
0% Abundance

98% ANI
40% Abundance

sylph: metagenomic profiling and containment by statistical k-mer sketching

Jim Shaw[1] and Yun William Yu[2]

Introduction

k-mer sketching

Statistical adjustment by ZIP model

Profiling

ANI-based MWAS

Flexible database profiling

Conclusion

Supp. Figs

Synthetic metagenome - CAMI2 Marine

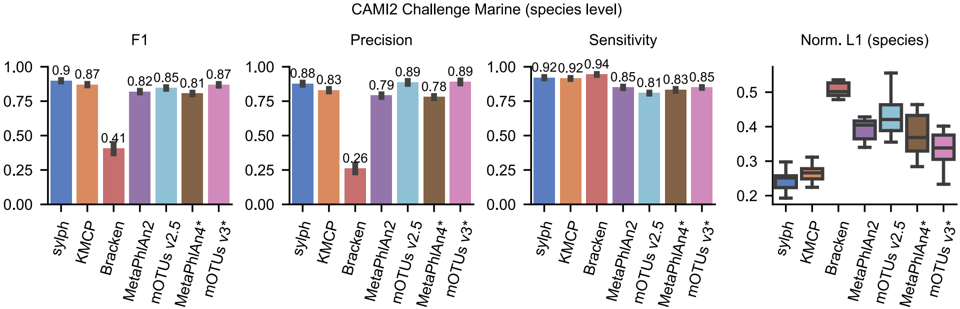


Figure: CAMI2 marine metagenome profiling challenge.

Synthetic metagenome - varying ANI

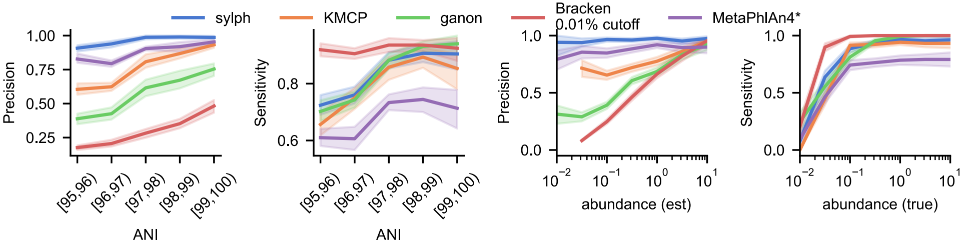


Figure: sylph retains high precision for lower ANI and abundance microbes (species level classification).

Fast and efficient

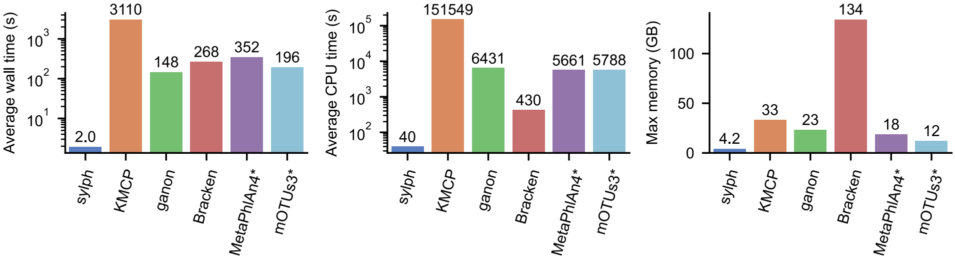


Figure: Runtime/memory on 200 genome synthetic community (3 Gbp, 2x150bp). 50 threads.

Why is it so fast?

- sylph does not classify reads (unlike Kraken et al.)
- sylph does not align reads (unlike MetaPhlAn, mOTUs)
- sylph shares one database for multi-sample profiling (unlike Kraken)
- Engineering (uses AVX2 instructions for k-mer sketching, etc)

Applications

Applications

Wallen et al. Parkinson’s Disease Metagenomics

- Wallen et al. (2022) performed gut metagenome wide association study (MWAS) for 490 Parkinson’s Disease (PD) and 234 controls

Wallen et al. Parkinson's Disease Metagenomics

- Wallen et al. (2022) performed gut metagenome wide association study (MWAS) for 490 Parkinson's Disease (PD) and 234 controls
- They used differential abundance testing → p-value

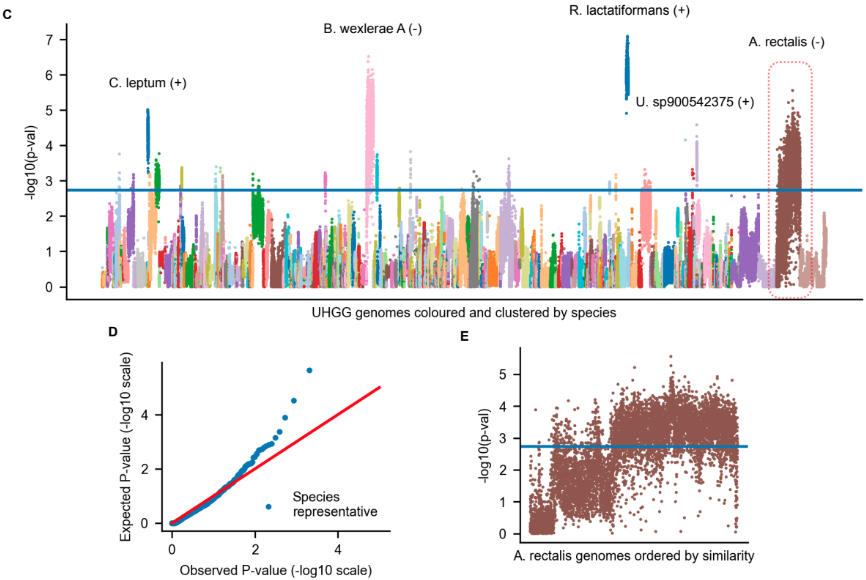
Wallen et al. Parkinson's Disease Metagenomics

- Wallen et al. (2022) performed gut metagenome wide association study (MWAS) for 490 Parkinson's Disease (PD) and 234 controls
- They used differential abundance testing → p-value
- **What we did**: differential containment ANI → p-value

Wallen et al. Parkinson's Disease Metagenomics

- Wallen et al. (2022) performed gut metagenome wide association study (MWAS) for 490 Parkinson's Disease (PD) and 234 controls
- They used differential abundance testing → p-value
- **What we did**: differential containment ANI → p-value
- Queried *289,232* genomes (UHGG) against 5.5 tb of reads; took a $\approx$ 3 hours with 40 threads

Containment ANI MWAS



Results

- Results are concordant with Wallen et al.

Results

- Results are concordant with Wallen et al.
- Butyrate-producing bacteria (*F. prausnitzii*, *A. rectalis*) depleted in PD

Results

- Results are concordant with Wallen et al.
- Butyrate-producing bacteria (*F. prausnitzii*, *A. rectalis*) depleted in PD
- Previous study (Becker et al. 2022) showed:

$$\frac{\text{B. wexlerae abundance}}{\text{R. lactatiformans abundance}} \propto \text{fecal butyrate concentration}$$

Flexible database profiling

- sylph allows for flexible database choice

Flexible database profiling

- sylph allows for flexible database choice
- Viruses and eukaryotes can be profiled too

Virome profiling and customized databases vs RefSeq

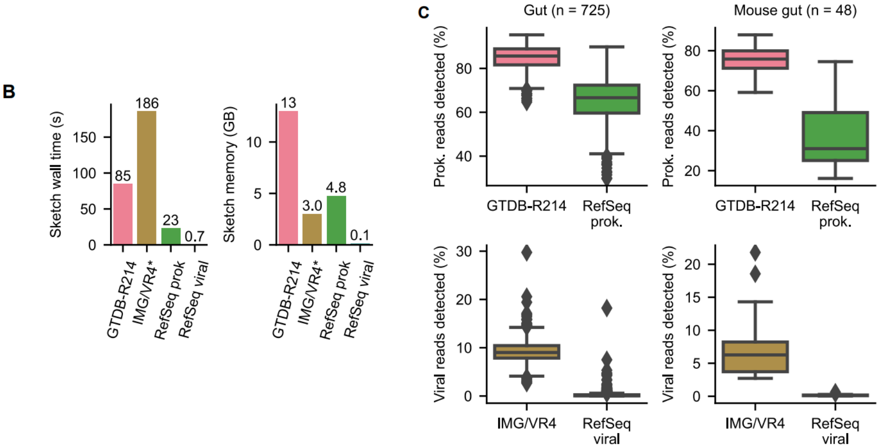
Comprehensive databases:

- GTDB-R214: 85,205 prokaryotic species genomes
- IMG/VR4: 2,917,521 species genomes

RefSeq:

- RefSeq representative prokaryotic: 18,325 genomes
- RefSeq viral: 14,993 genomes

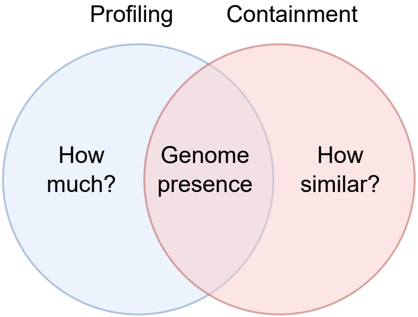
Virome profiling and customized databases vs RefSeq



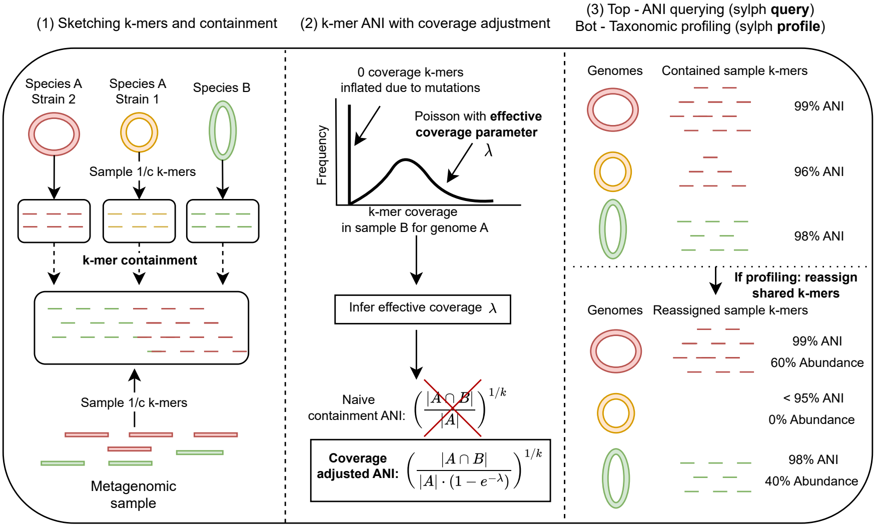
Conclusion

Presented **sylph**

- Statistical k-mer sketching approach for **containment** and **profiling**



Algorithm recap


(1) Sketching k-mers and containment
(2) k-mer ANI with coverage adjustment
(3) Top - ANI querying (sylph **query**)
Bot - Taxonomic profiling (sylph **profile**)
Species A Strain 2
Species A Strain 1
Species B
Sample 1/c k-mers
k-mer containment
Sample 1/c k-mers
Metagenomic sample
0 coverage k-mers inflated due to mutations
Frequency
Poisson with **effective coverage parameter** λ
k-mer coverage in sample B for genome A
Infer effective coverage λ
Naive containment ANI: $\cancel{\left(\frac{|A \cap B|}{|A|}\right)}^{1/k}$
Coverage adjusted ANI: $\left(\frac{|A \cap B|}{|A| \cdot (1 - e^{-\lambda})}\right)^{1/k}$
Genomes
Contained sample k-mers
99% ANI
96% ANI
98% ANI
If profiling: reassign shared k-mers
Genomes
Reassigned sample k-mers
99% ANI
60% Abundance
< 95% ANI
0% Abundance
98% ANI
40% Abundance

Funding and acknowledgements

Metagenome profiling and containment estimation through abundance-corrected k-mer sketching with sylph now available on bioRxiv.

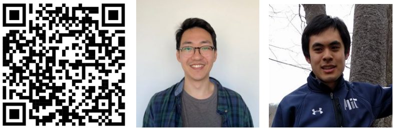


Figure: Github QR code, Jim Shaw (PhD student), Yun William Yu (Advisor)

Natural Sciences and Engineering Research Council of Canada
Conseil de recherches en sciences naturelles et en génie du Canada
Canada

UNIVERSITY OF TORONTO

Carnegie Mellon University

Synthetic metagenome - CAMI2 Marine

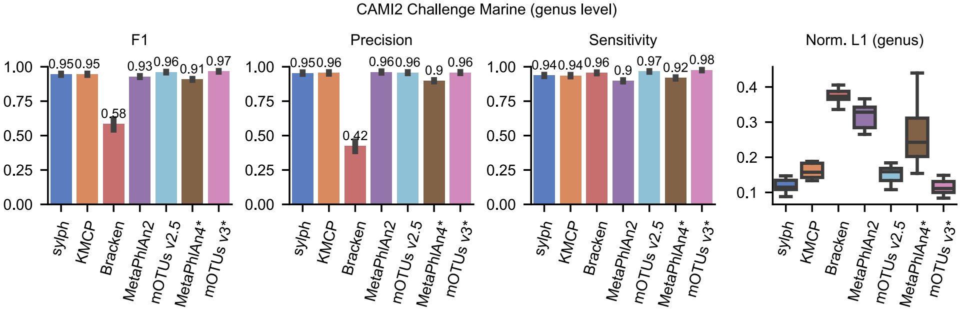


Figure: CAMI2 marine metagenome challenge profiling.

Synthetic metagenome - CAMI2 Strain Madness

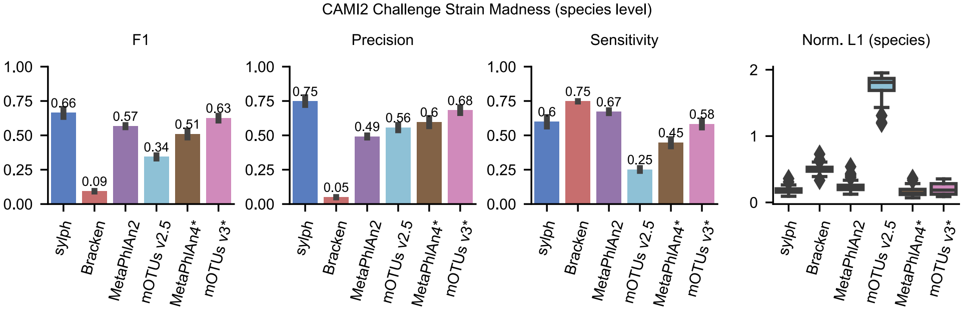


Figure: CAMI2 strain madness challenge.

Synthetic metagenome - CAMI2 Strain Madness

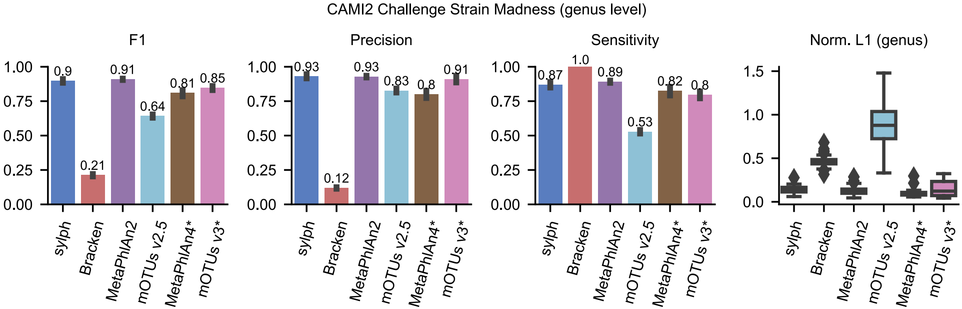


Figure: CAMI2 strain madness metagenome challenge profiling.

Profiling real reads

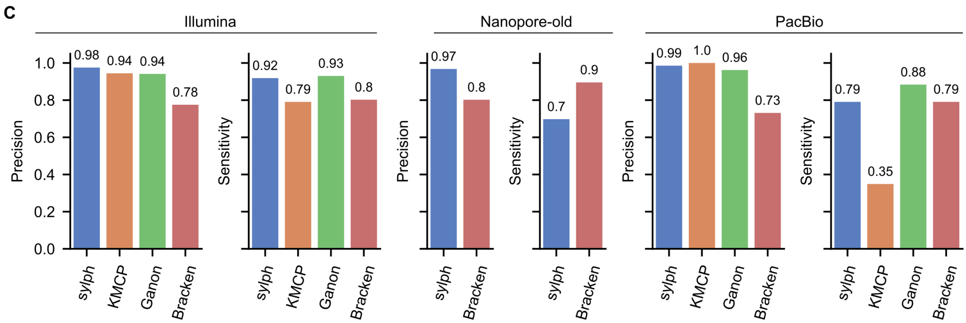


Figure: Mock community profiling from Meslier et al. (2022)