sylph: metagenome profiling by statistical k-mer sketching

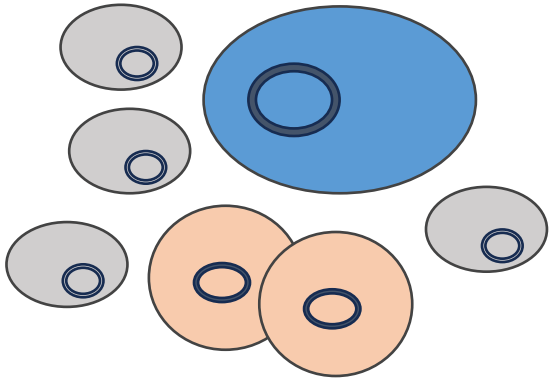
Jim Shaw

University of Toronto

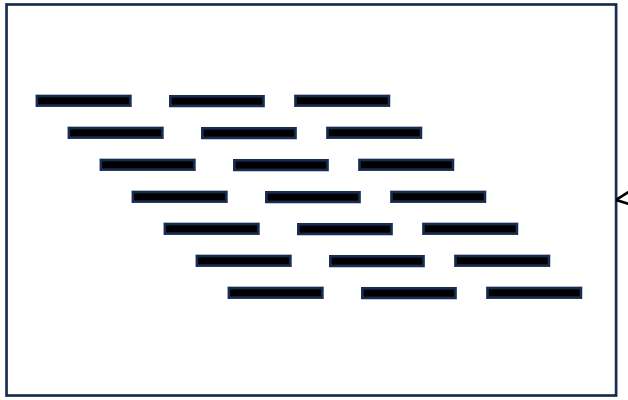
UNIVERSITY OF TORONTO

Metagenomic workflow (shotgun, not 16S)

Microbiome sample



Reads



Profiling

Use database -> infer taxa abundance/presence

Assembly

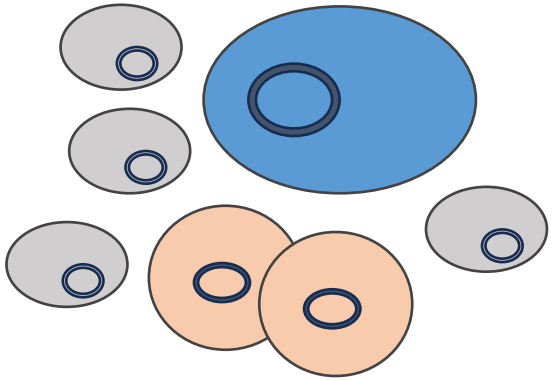
De novo assembly of sequences

What is in my sample?

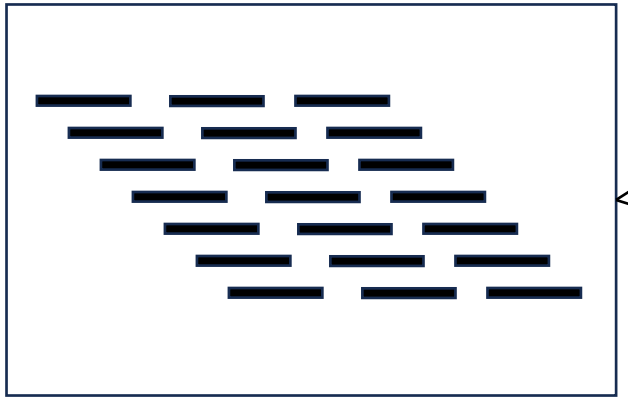
UNIVERSITY OF TORONTO

This work: **metagenomic** profiling

Microbiome sample



Reads



Profiling

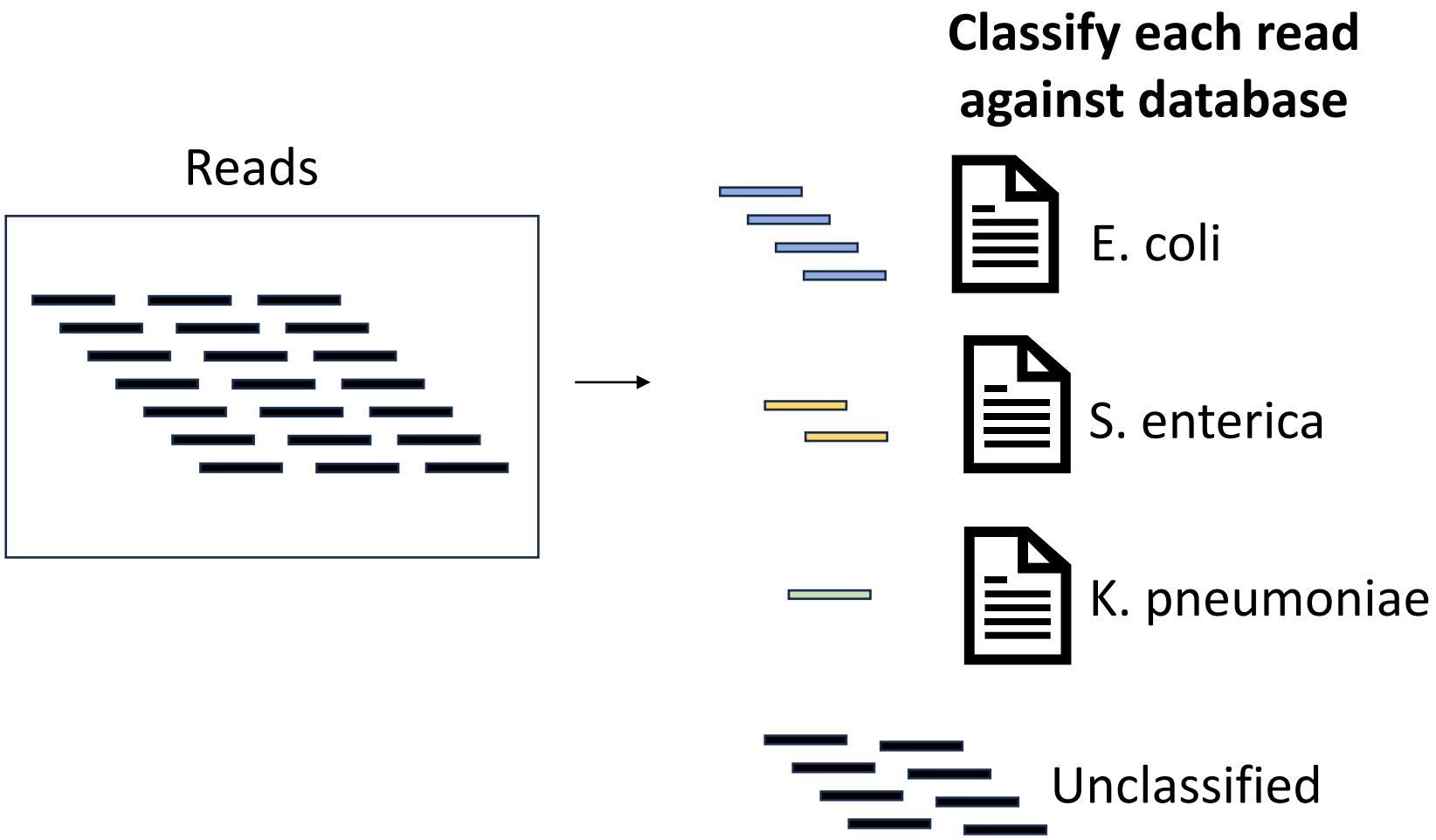
Use database -> infer taxa abundance/presence

Assembly

De novo assembly of sequences

UNIVERSITY OF TORONTO

Standard idea – map reads to genomes + calculate relative abundance



Why build new metagenome profilers?

Problem: classifying reads is hard!

- Indexing 100,000 genomes + mapping **is expensive**
- **False positives** are inevitable (ambiguous reads)

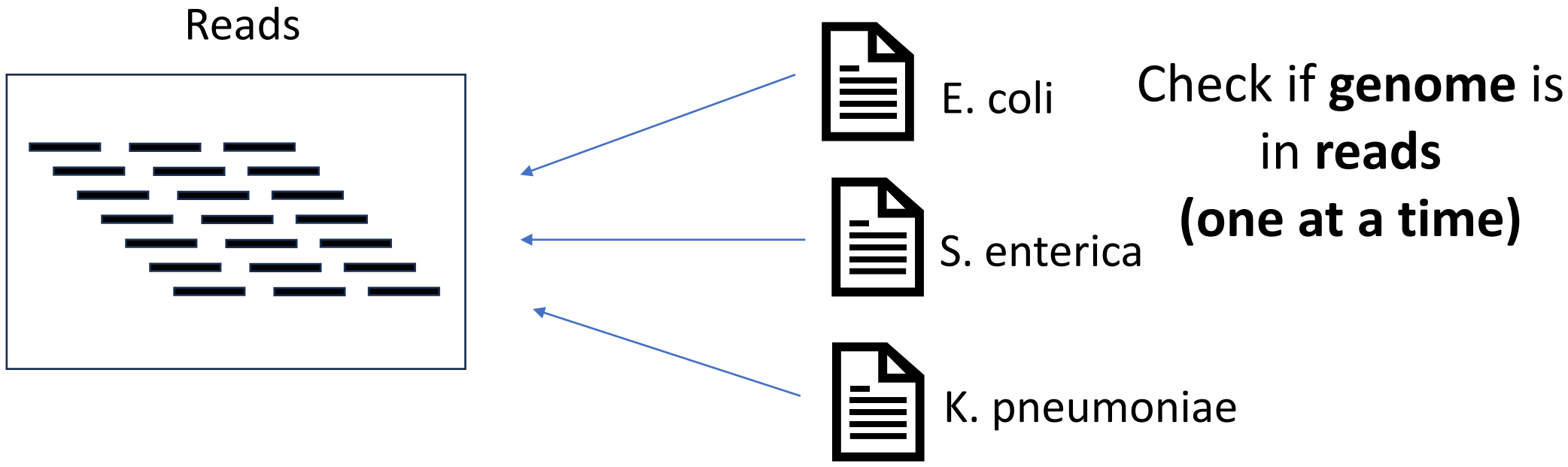
UNIVERSITY OF TORONTO

sylph: metagenome profiling by k-mer containment

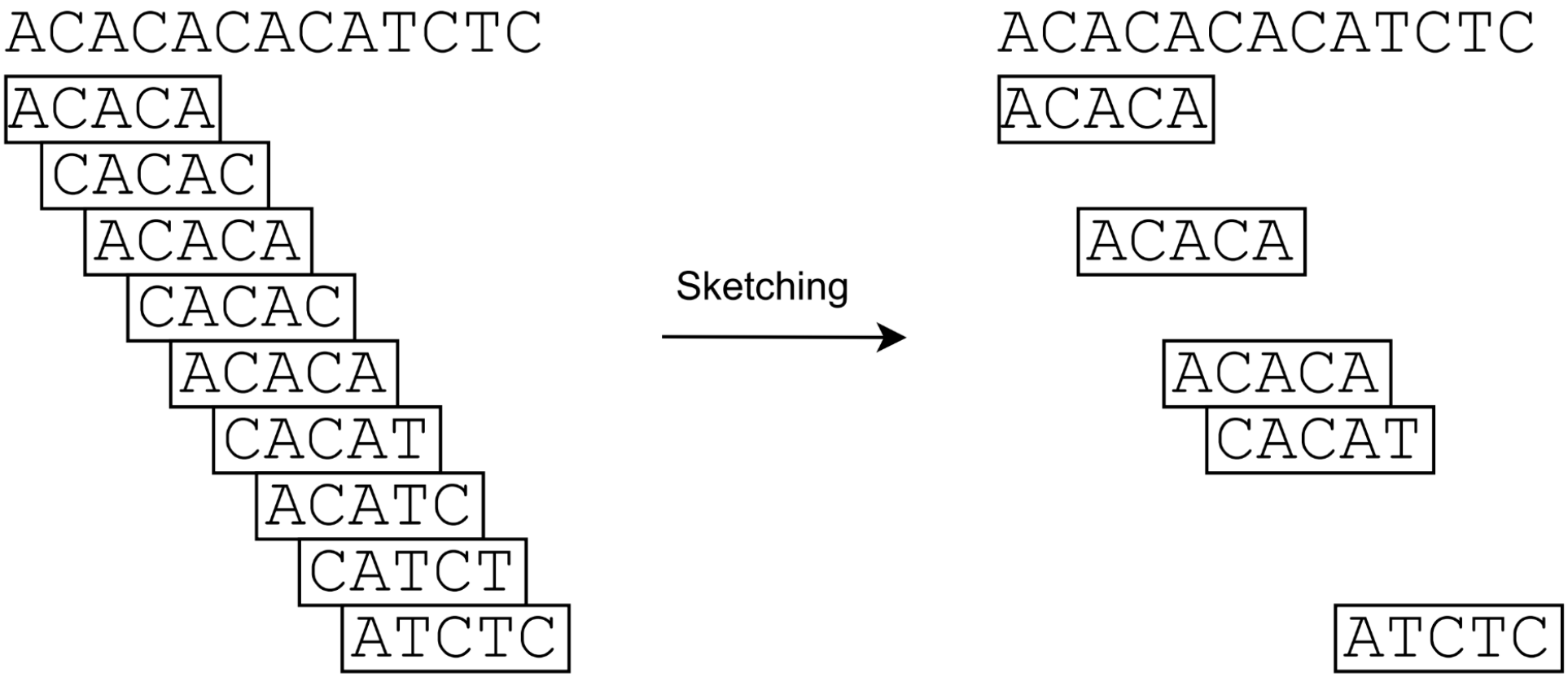
UNIVERSITY OF TORONTO

Sylph (Shaw and Yu, 2023, bioRxiv)

- Classify **genomes** against **reads** instead



How sylph works (1): k-mer sketching

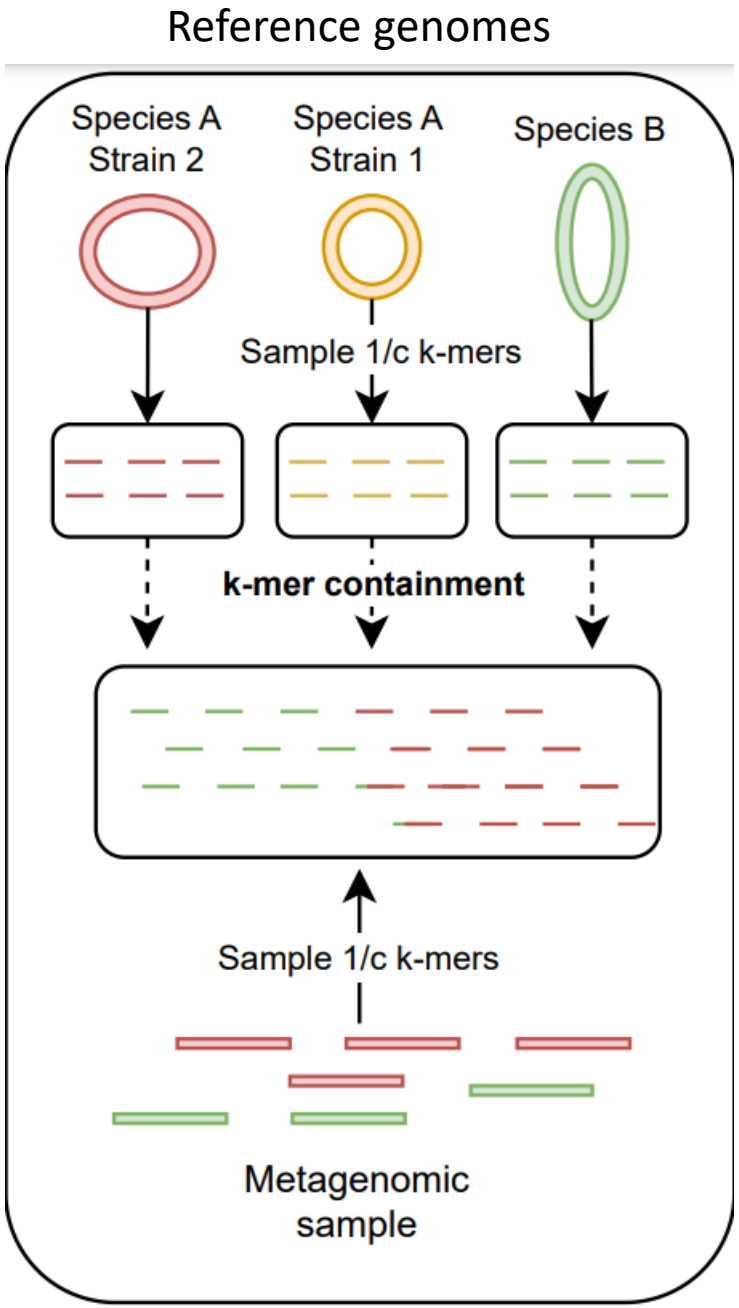


UNIVERSITY OF TORONTO

Step 1: k-mer sketching

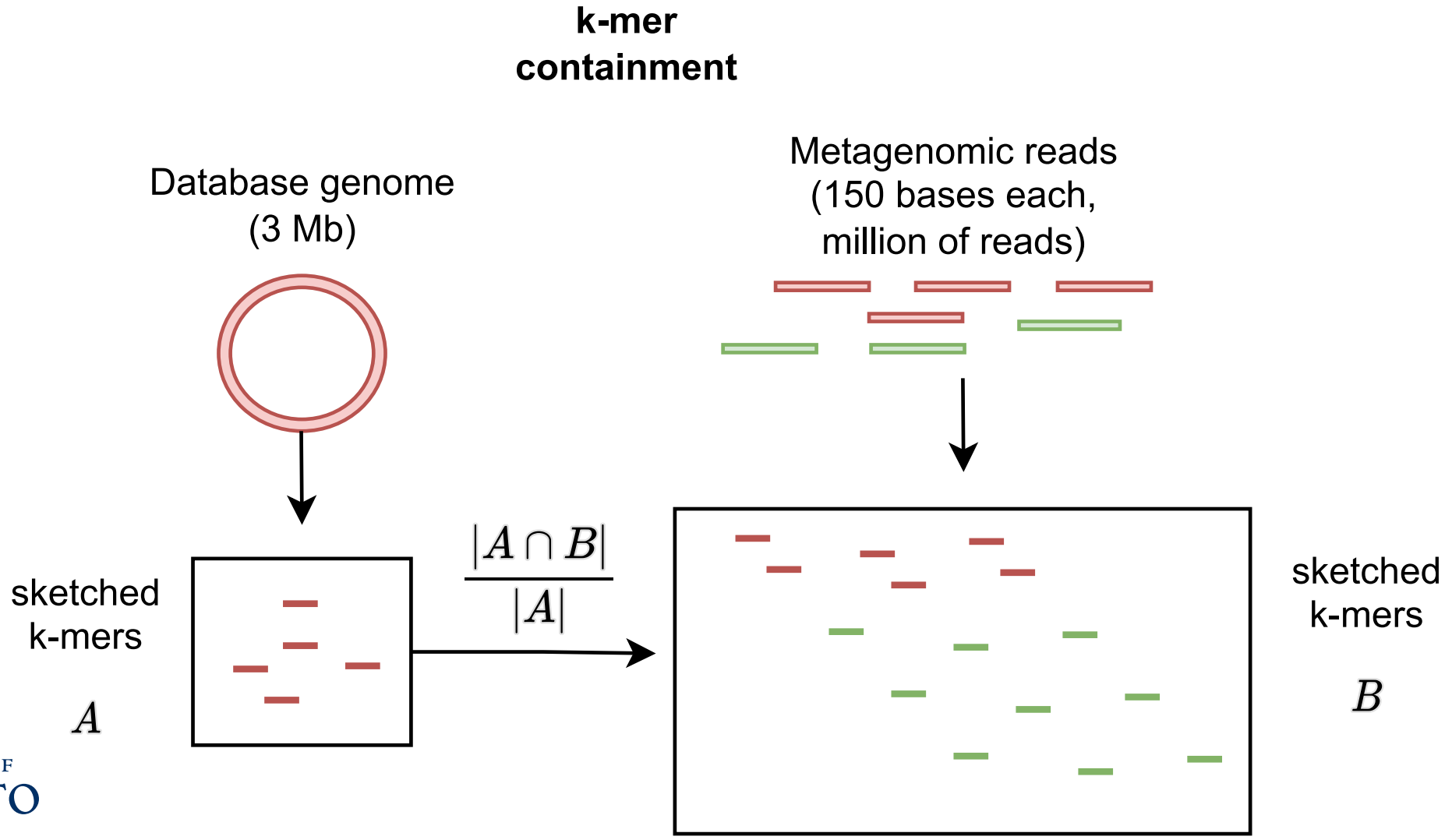
Subsample k-mers using **FracMinHash** (similar to **minimizers**)

Sample 1/200 k-mers by default



UNIVERSITY OF TORONTO

Step 2: k-mer containment



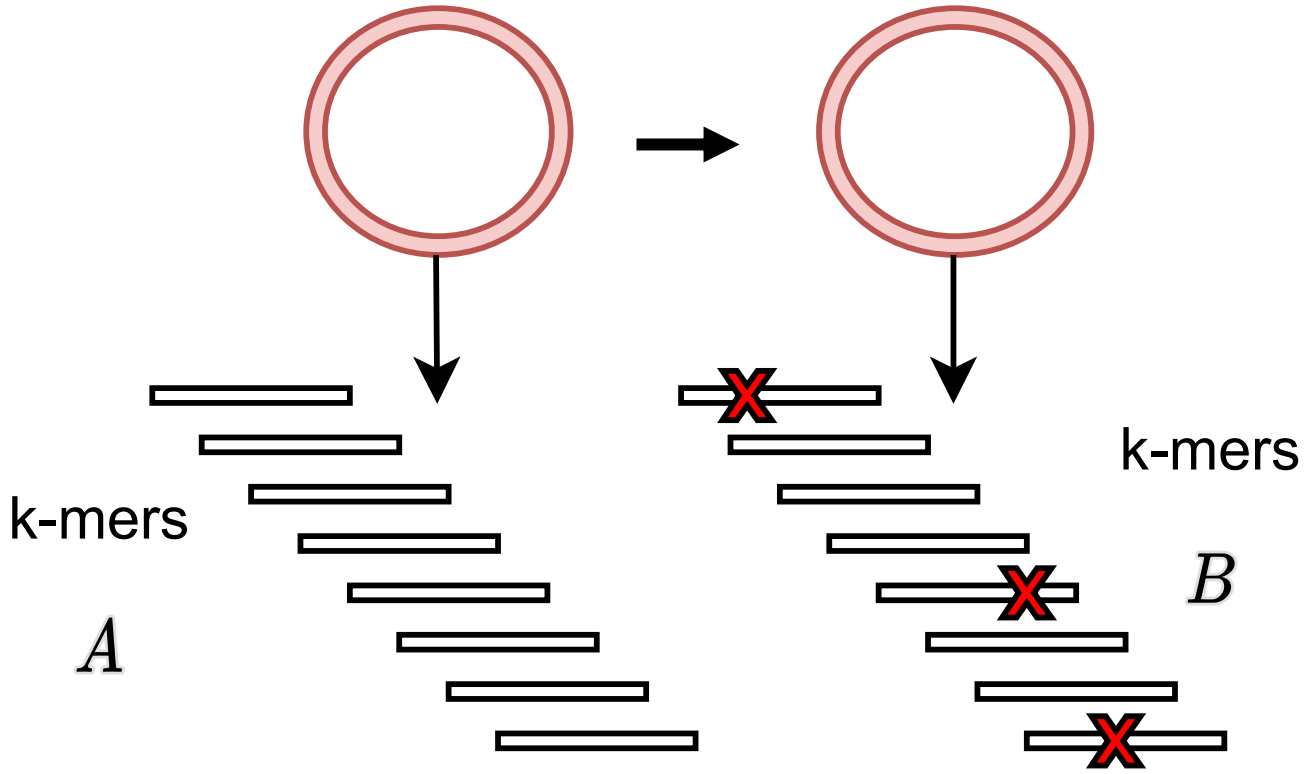
Connection to **average nucleotide identity** (ANI)

Average nucleotide identity: 99% similar strains

Estimate ANI by counting k-mers:

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

(k ≈ 20-32)

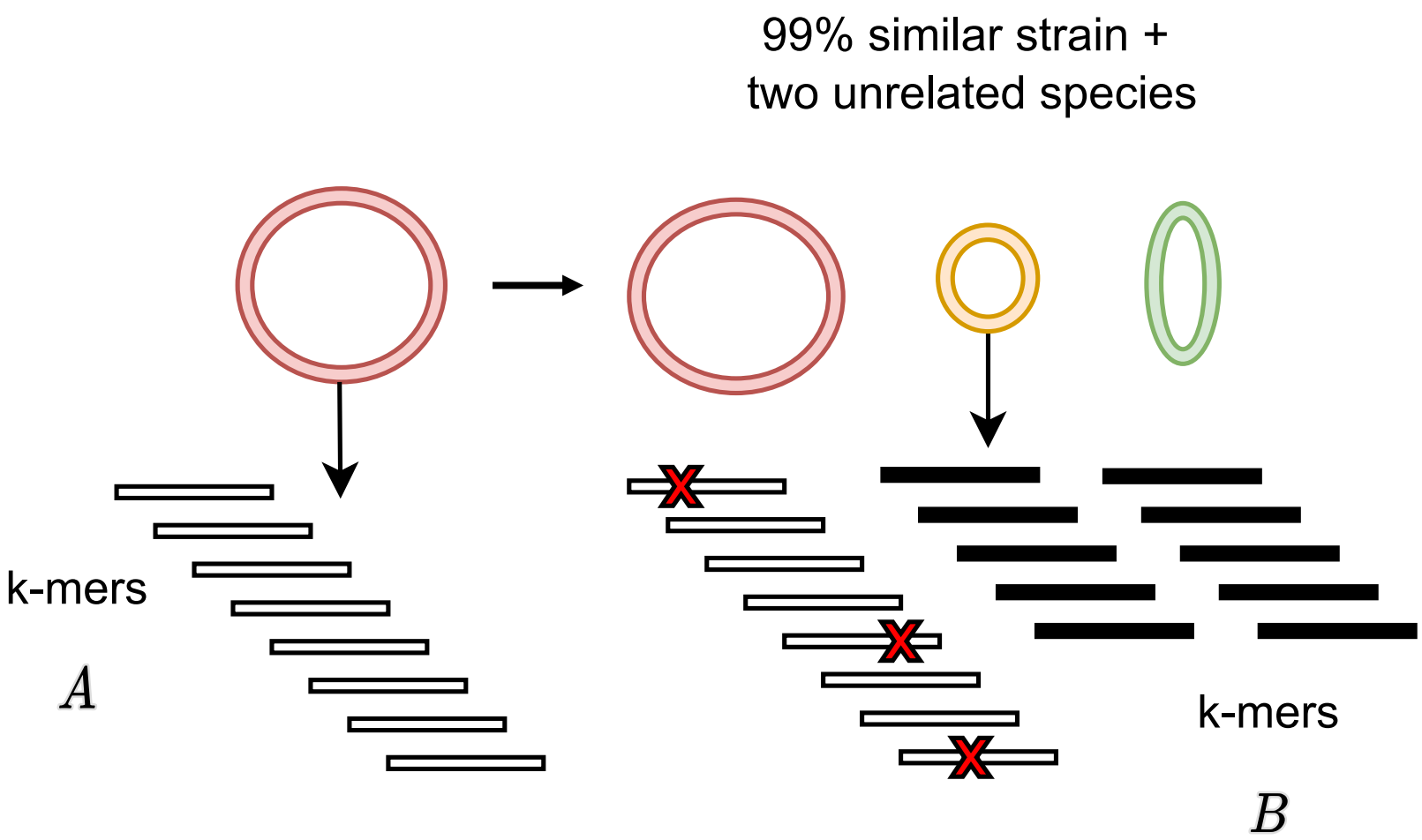


UNIVERSITY OF TORONTO

Containment ANI – metagenomes

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

Extends to metagenomes



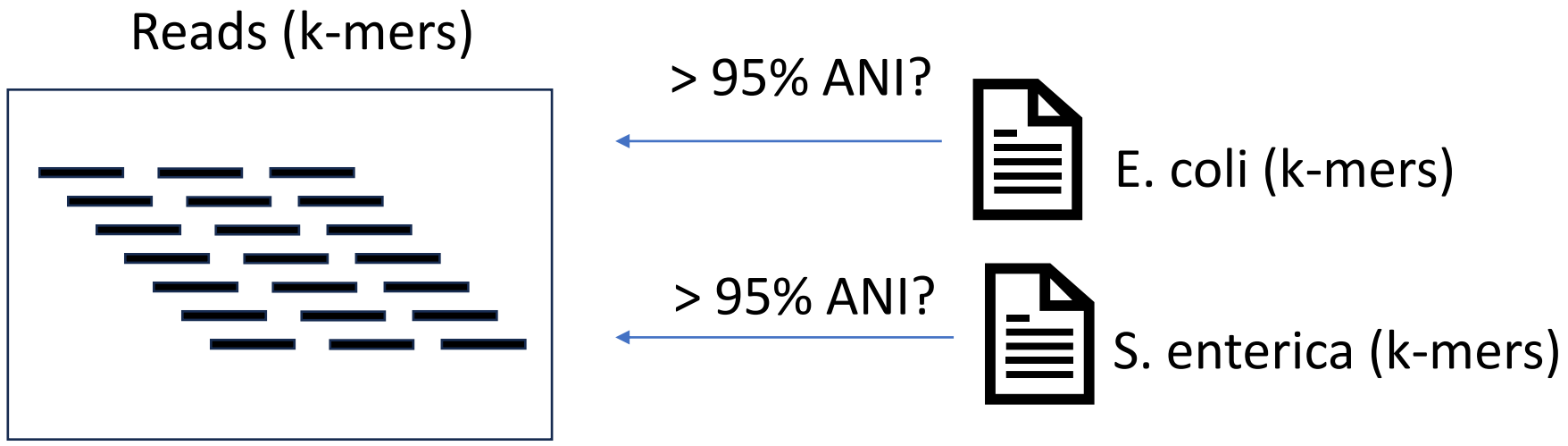
UNIVERSITY OF TORONTO

Why ANI? **Species defining!**

Two (microbial) genomes > 95% ANI ⇒ same species*

Sylph: calculate **metagenome containment ANI**

- **> 95%** ⇒ **present** (at species level)



UNIVERSITY OF TORONTO

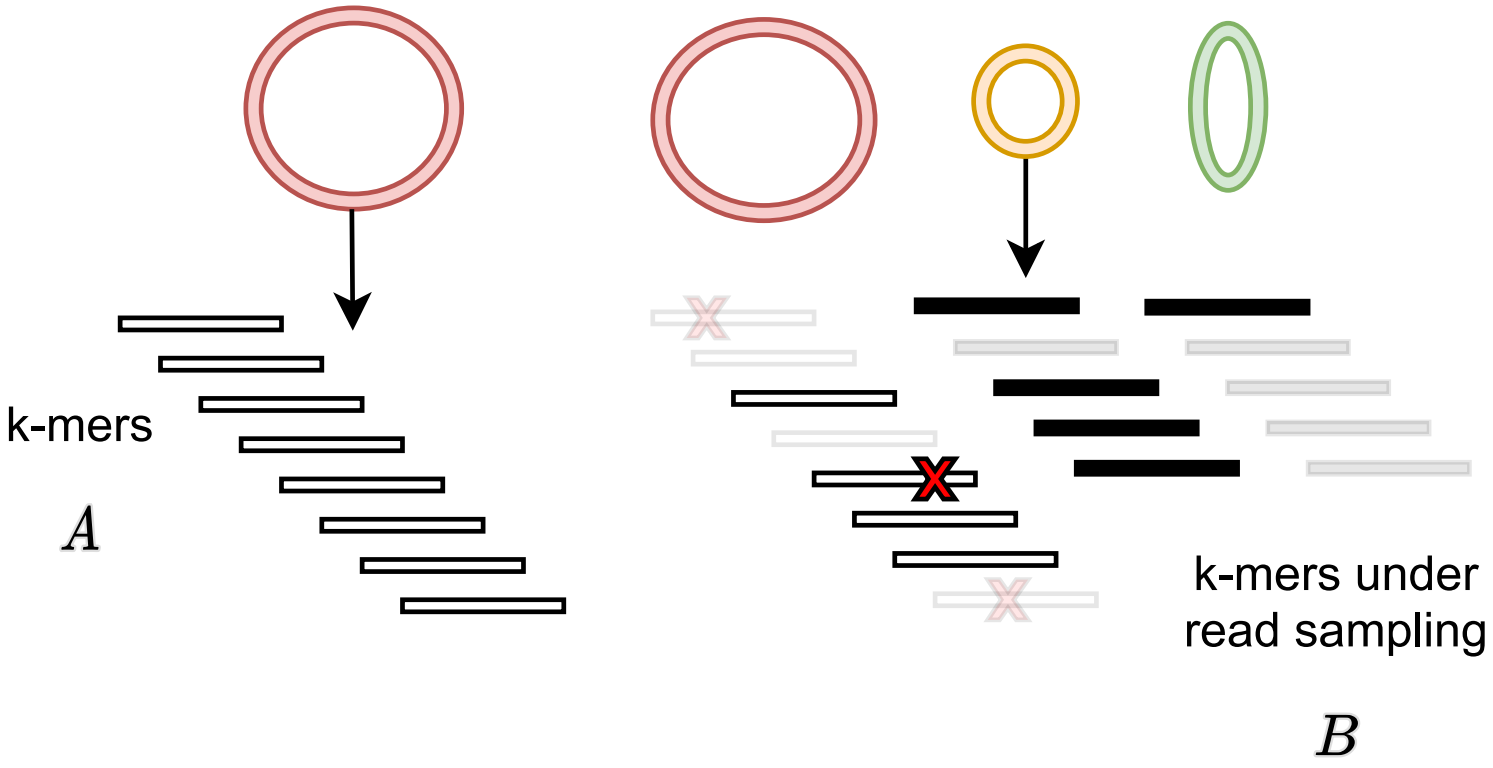
*Jain et al., 2018, Nature Communications

The **low-coverage problem**: sylph's **innovation**

Reads do not cover all k-mers (low-coverage)

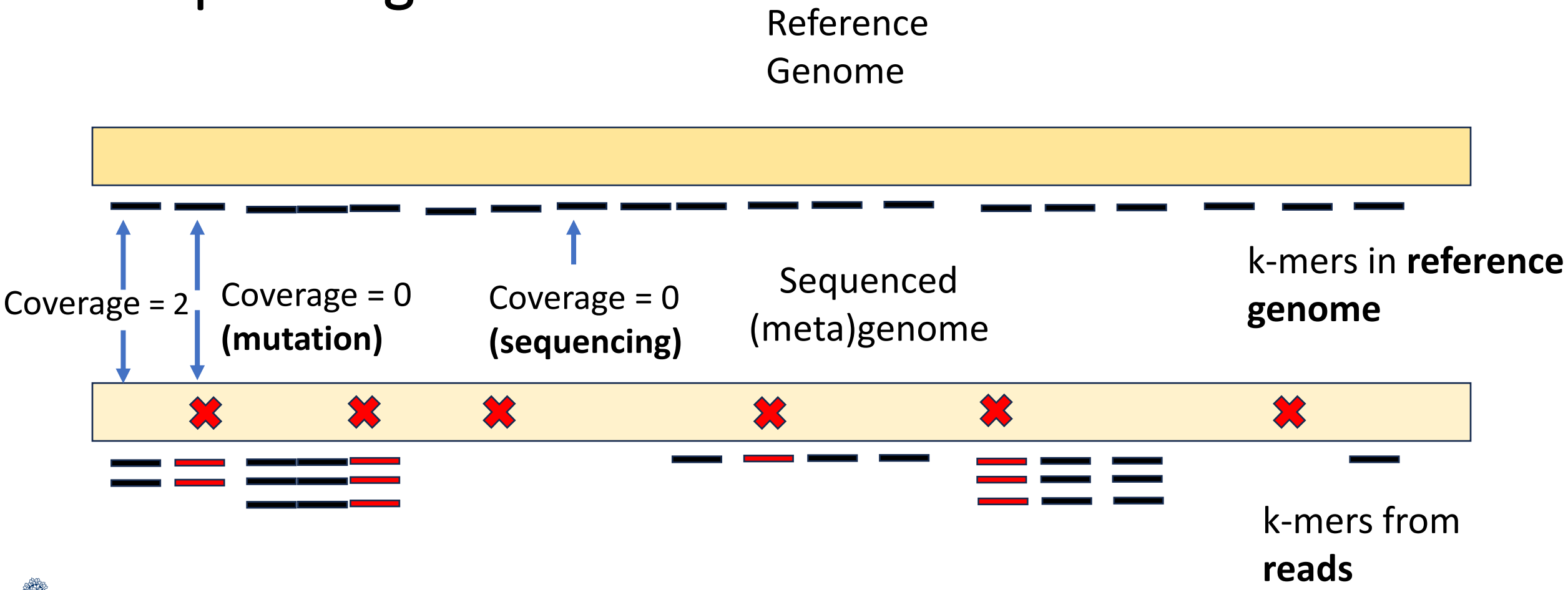
$$ANI \neq \left(\frac{|A \cap B|}{|A|}\right)^{1/k}$$

ANI inference fails because B is **under sequenced**



UNIVERSITY OF TORONTO

K-mer coverage “dropouts” due to **mutation AND sequencing**

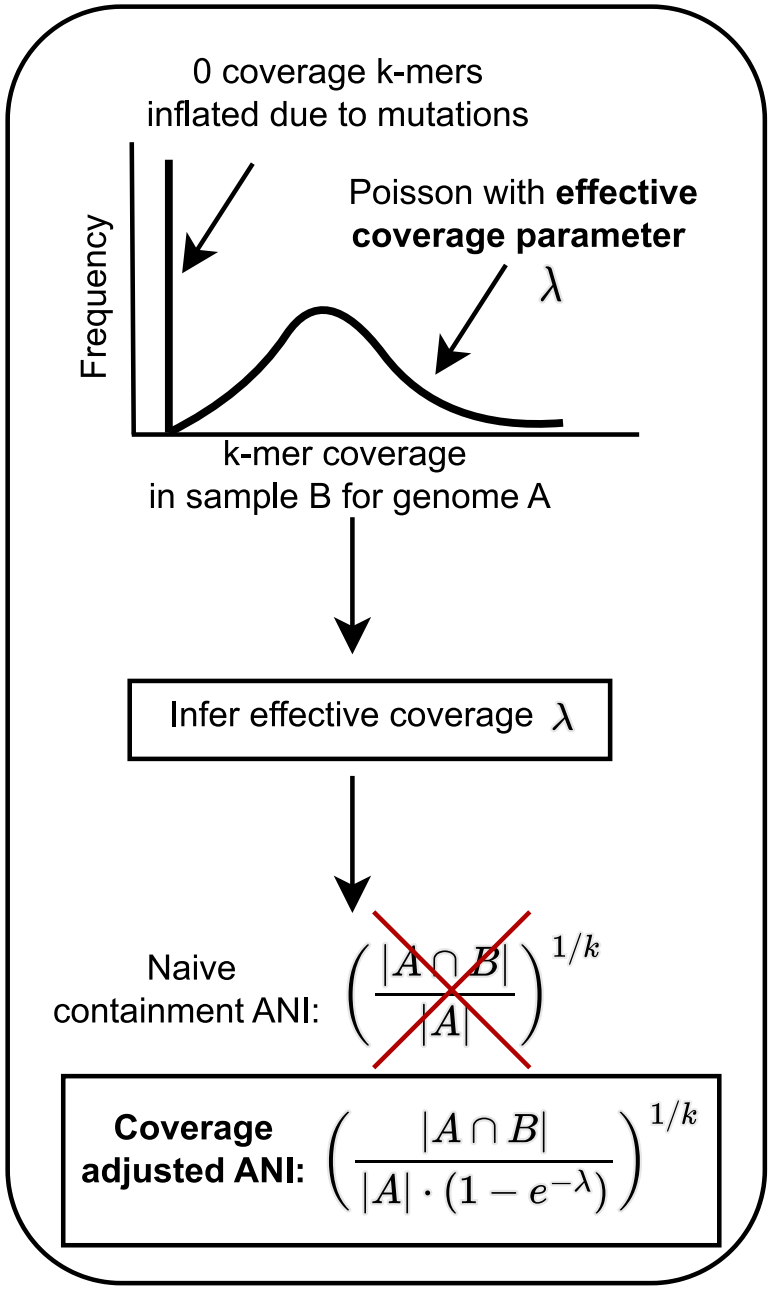


UNIVERSITY OF TORONTO

Sylph:

statistical adjustment for **low-coverage** sequencing

(2) k-mer ANI with coverage adjustment



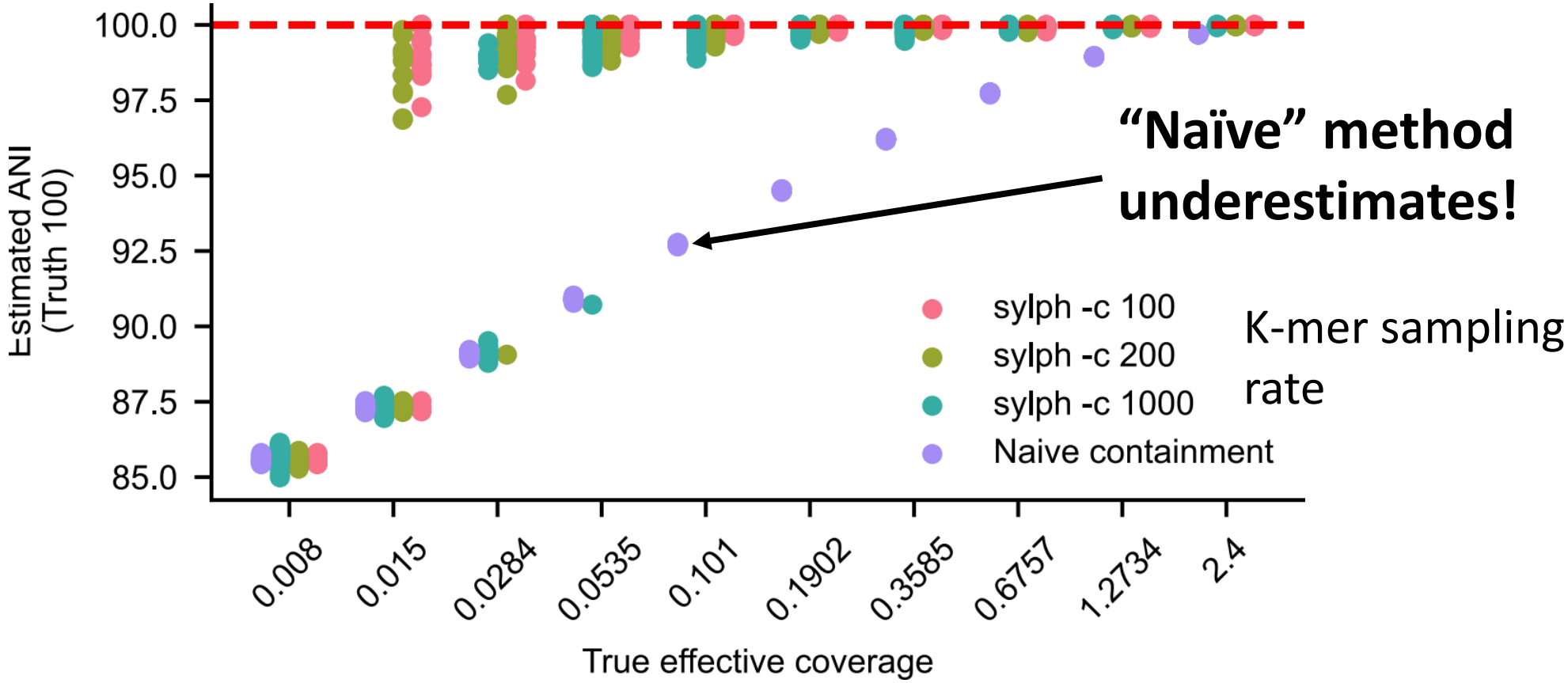
Intuition

1. Some 0s are zero inflation (**mutation**)
2. But some 0s are Poisson (**sequencing**)
3. **ANI: which 0s** are due to **mutation**?
4. Sylph: **infer Poisson** + **re-adjust** containment for ANI

UNIVERSITY OF TORONTO

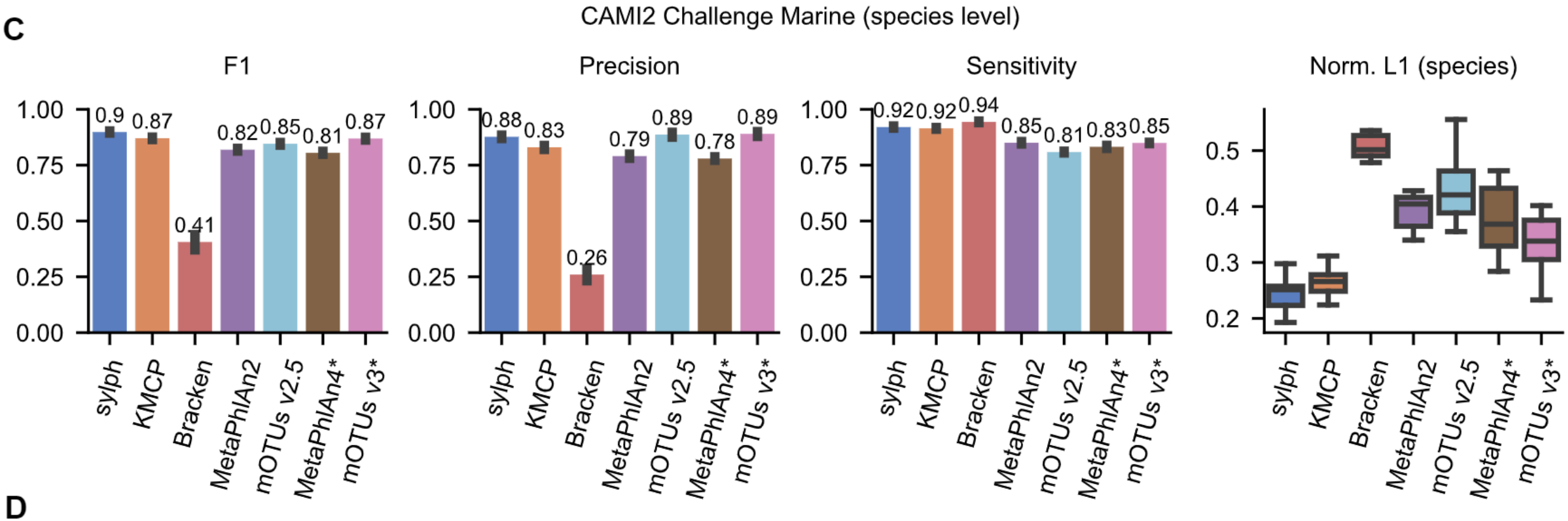
Sylph **results**

Sylph corrects ANI for simulated reads at low coverage



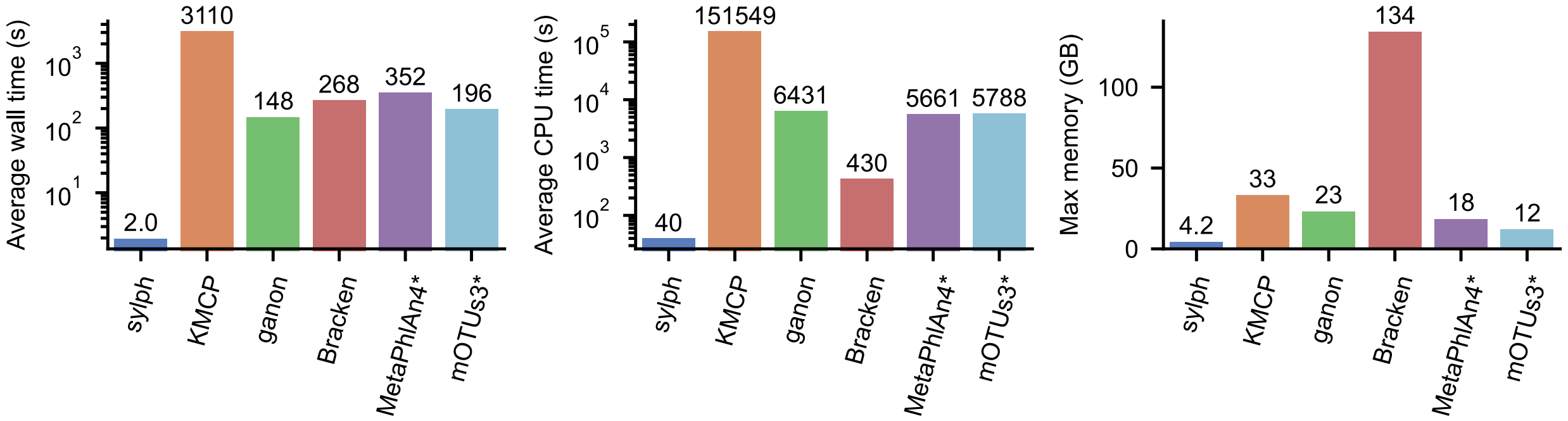
UNIVERSITY OF TORONTO

Sylph is effective for species-level profiling



UNIVERSITY OF TORONTO

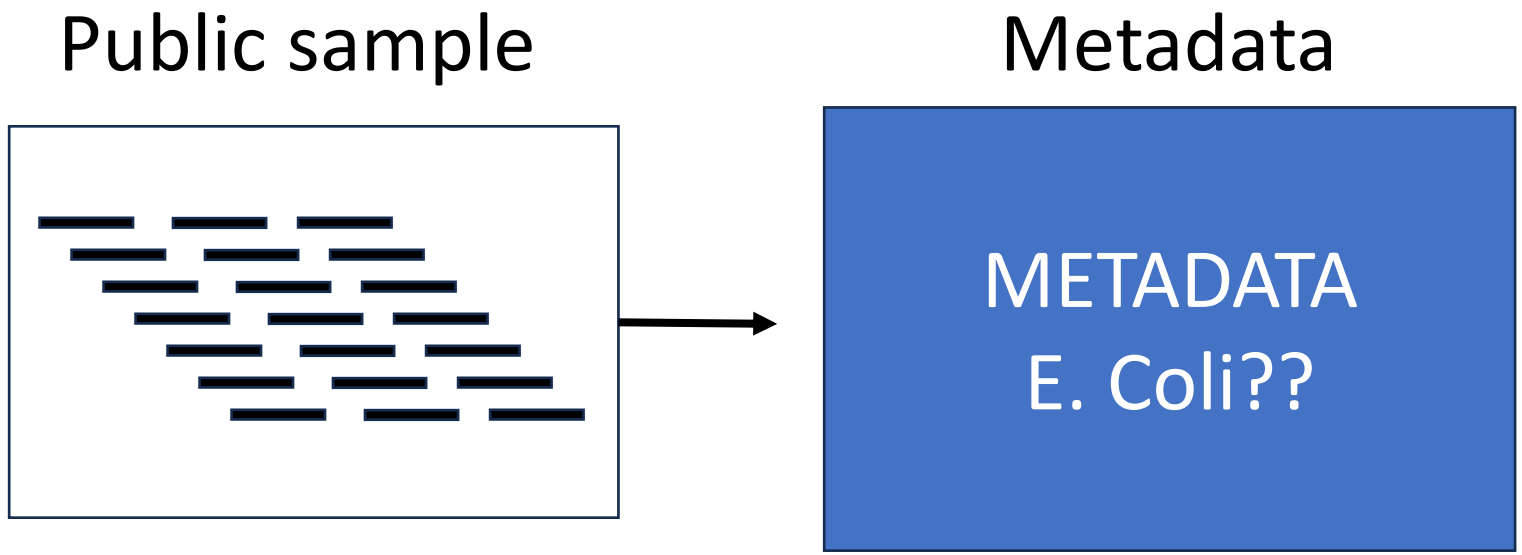
Sylph is extremely fast and efficient (multi-sample profiling)



UNIVERSITY OF TORONTO

Massive contamination detection

- **Contamination/bad metadata** in public data
- Solution: metagenome profile to detect contamination



UNIVERSITY OF TORONTO

AllTheBacteria - all bacterial genomes assembled, available and searchable

Martin Hunt, Leandro Lima, Wei Shen, John Lees, Zamin Iqbal

doi: https://doi.org/10.1101/2024.03.08.584059

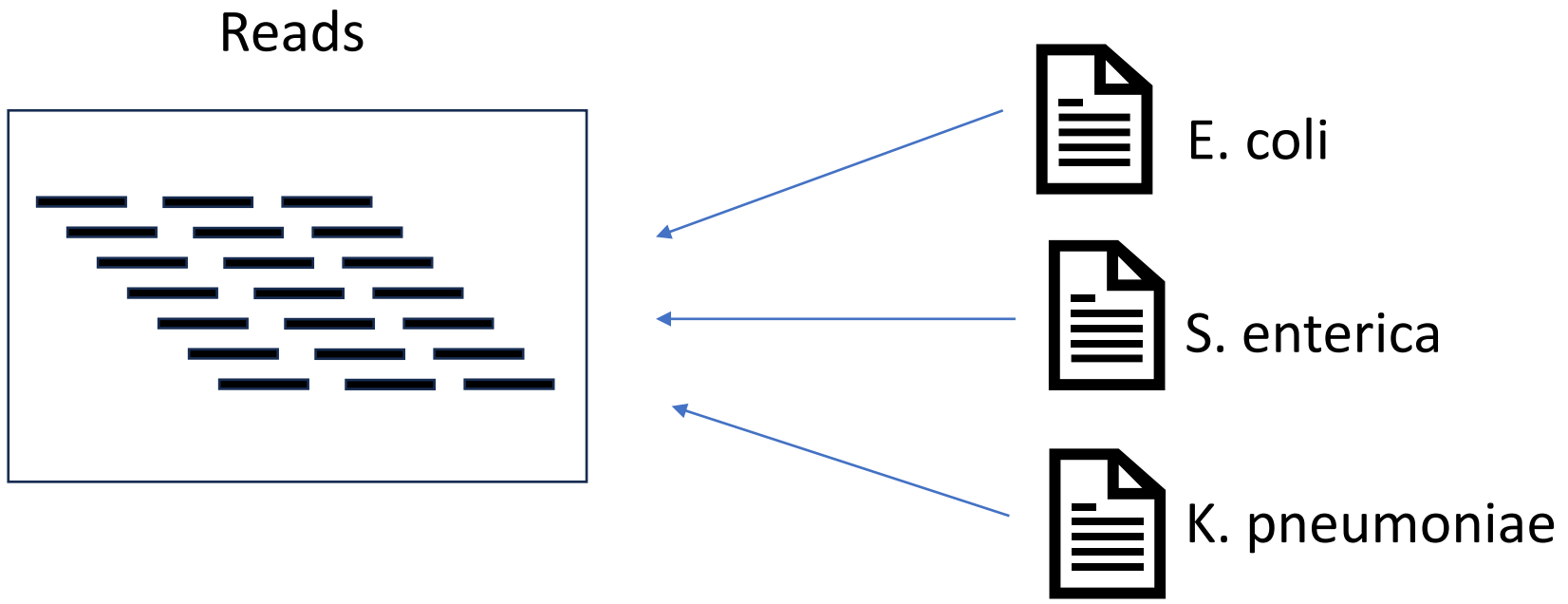
- Contamination check: **every single SRA dataset** (bacterial isolate Illumina WGS)
- Analyzed ≈ 2 **million datasets** with **sylph**

*"… sylph was more **accurate**, **faster** (~1 minute per sample) and required **less RAM** (10Gb of RAM for [85,000 genomes]) than previous tools"*

UNIVERSITY OF TORONTO

Recap: sylph metagenome profiling

1. Classify **genomes** against **reads**
2. k-mer sampling + **coverage-aware ANI statistics**



UNIVERSITY OF TORONTO

Drawbacks

- **Sylph can not classify reads**
- Some tasks: **require** classifying reads (e.g. very low coverage)
- Requires species-level representatives (but can use large databases + MAGs)

Conclusion

- **Jim Shaw** – 5th year PhD student (University of Toronto)
- Yun William Yu – PhD advisor (Assistant Prof. at Carnegie Mellon University)

Metagenome profiling and containment estimation through abundance-corrected k-mer sketching with sylph

by Jim Shaw and Yun William Yu available on **bioRxiv**

UNIVERSITY OF TORONTO

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

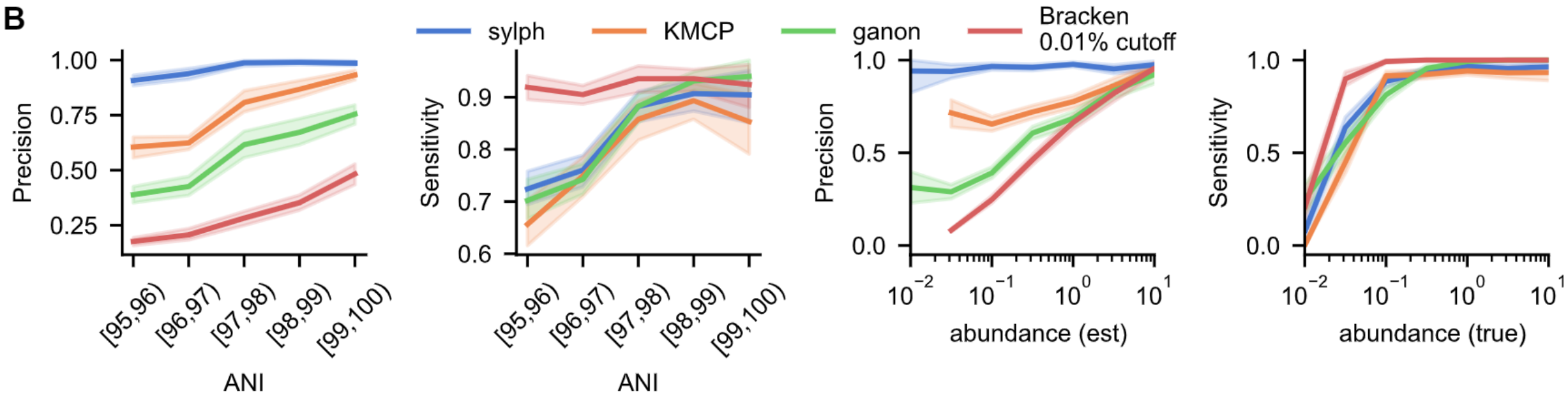
Supported by NSERC CGS-D

Math

- $\hat{\lambda} = \frac{\#kmers\ with\ cov = 2}{\#kmers\ with\ cov = 1} \cdot 2$
- Comes from Poisson PMF:

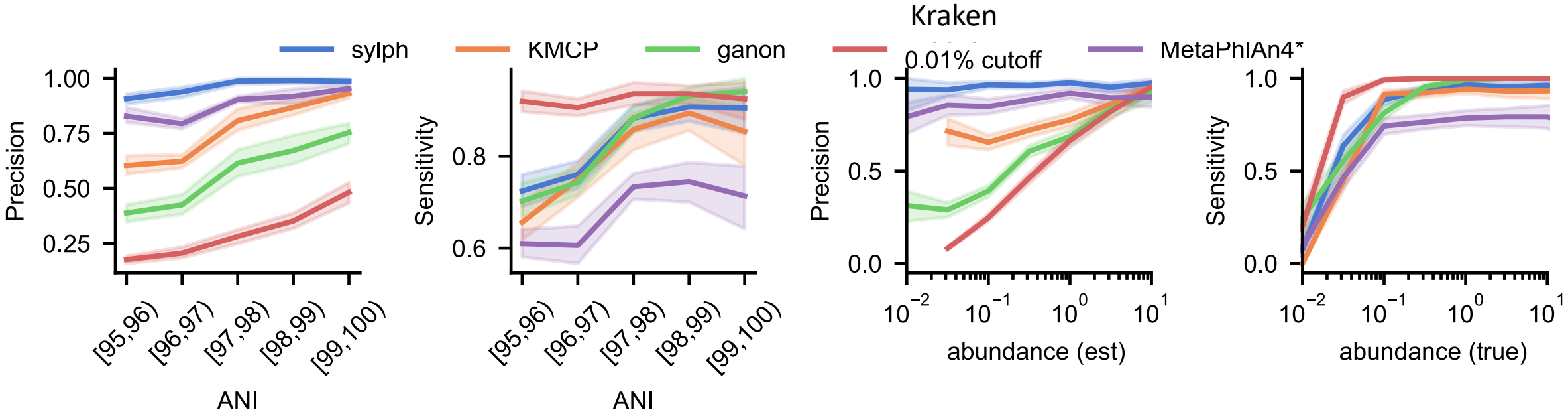
$$\frac{\Pr(Pois{=}2)}{\Pr(Pois{=}1)} = \frac{e^{-\lambda}\lambda^2}{2!} \Big/ \frac{e^{-\lambda}\lambda^1}{1!} = \frac{\lambda}{2}$$

Sylph is precise for **divergent** and **low-abundance** species



UNIVERSITY OF TORONTO

Sylph is precise for low ANI and coverage



UNIVERSITY OF TORONTO