skani: fast and robust metagenomic sequence comparison*

Jim Shaw, Yun William Yu

University of Toronto

UNIVERSITY OF TORONTO

*Now accepted to Nature Methods

Introduction

How do we measure sequence similarity for (microbial) genomes?

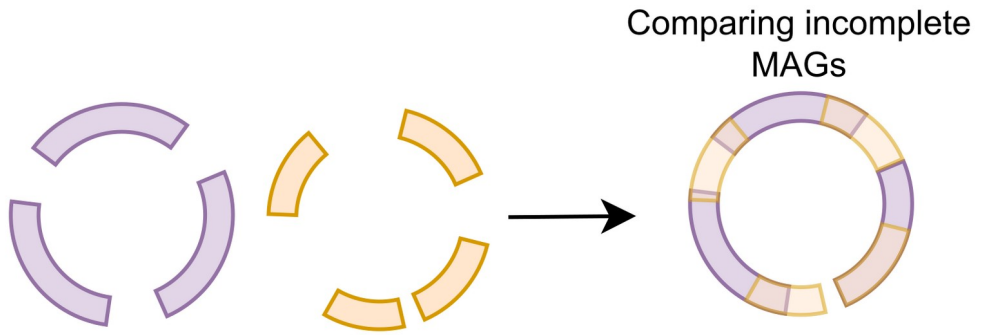
- Average nucleotide identity (ANI) - % of nucleotides shared for orthologous regions
- Common use cases: species delineation (95%), etc

Calculating average nucleotide identity

- **Alignment**: compute alignments + estimate sequence identity
  - ANIm (MUMmer)
  - ANIb/ANIu (BLAST, USEARCH)
- **Sketching:** obtain subset of k-mers, use k-mers to estimate ANI, > 10,000x faster than alignment
  - Mash (MinHash) - Ondov et al. 2016
  - Sourmash (FracMinHash) - Irber et al. 2022
- **Hybrid:** combination of above
  - FastANI (Minimizer MinHash + mapping) - Jain et al. 2018
  - **skani (Sparse k-mer chaining)** - Shaw and Yu 2023

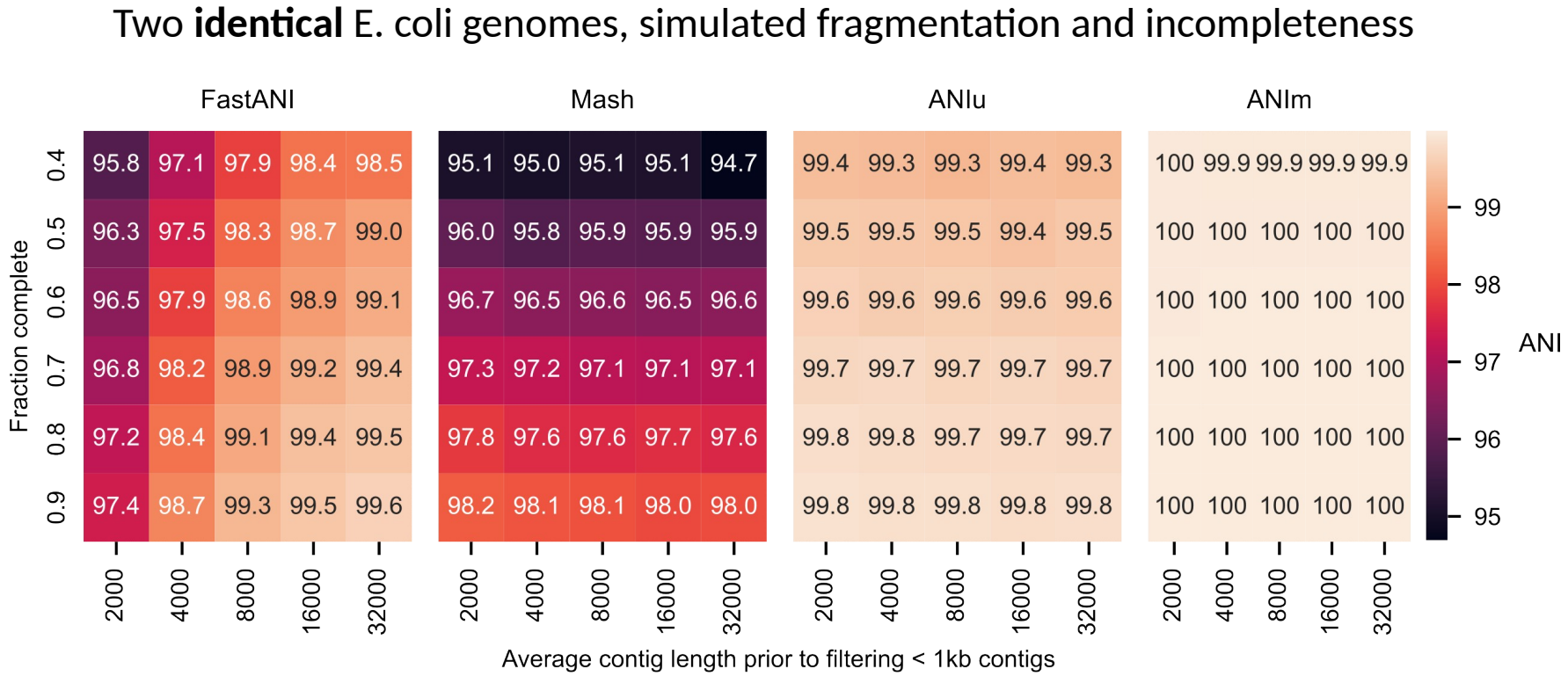
Metagenomic data is **noisy**

- Metagenome-assembled genomes (MAGs) are **noisy**
  - Incomplete (missing sequences)
  - Contaminated (spurious sequences included)
  - Fragmented (small contigs, low N50)



UNIVERSITY OF TORONTO

MAG noise biases ANI calculations



UNIVERSITY OF TORONTO

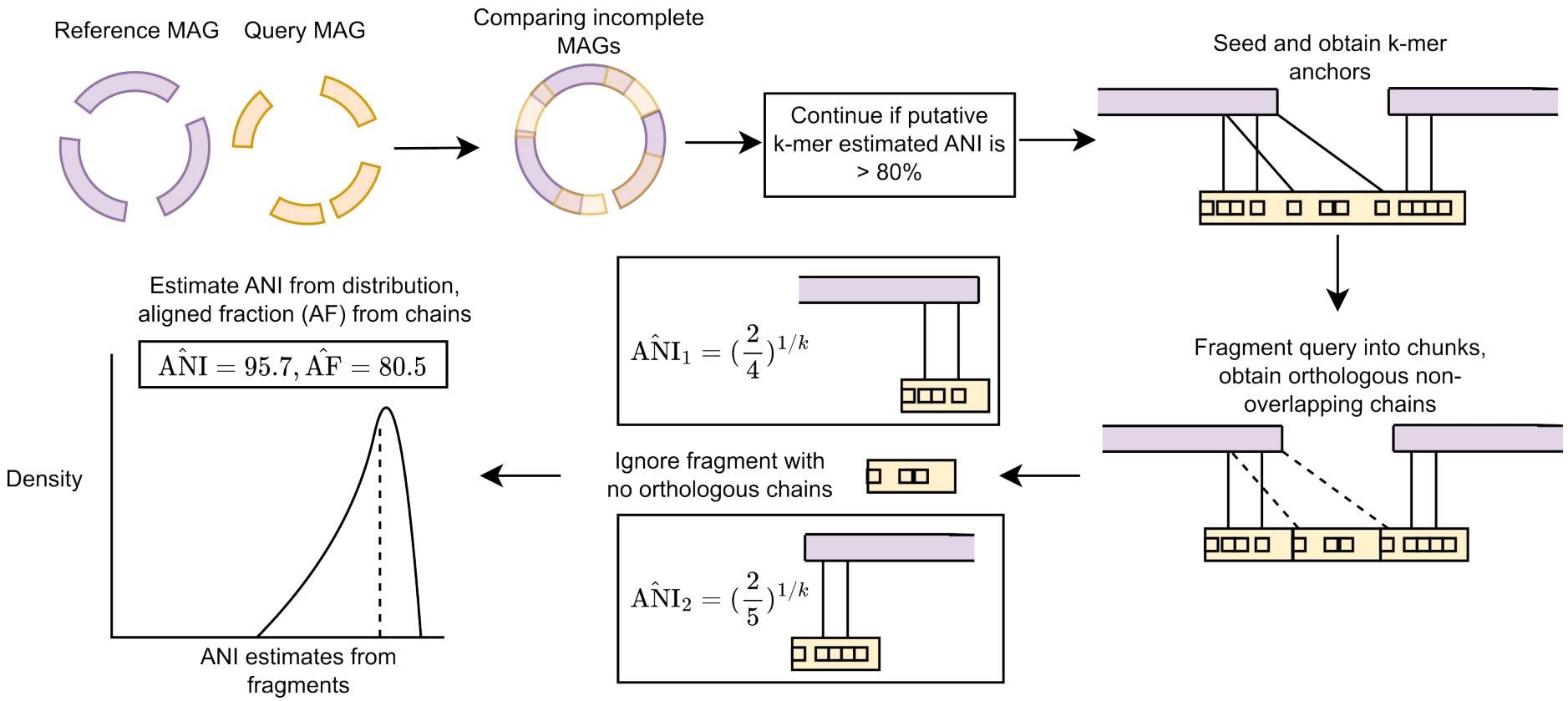
skani: a new tool

- skani is a new **ANI** and **aligned fraction** tool for genomes > 82% ANI
  - Works with MAGs, genomes, eukaryotic MAGs, contigs
- **Database search** is supported by fast k-mer ANI filtering
- Easy install – bioconda, static binary, written in rust with no 3rd party dependencies

skani methods: key points

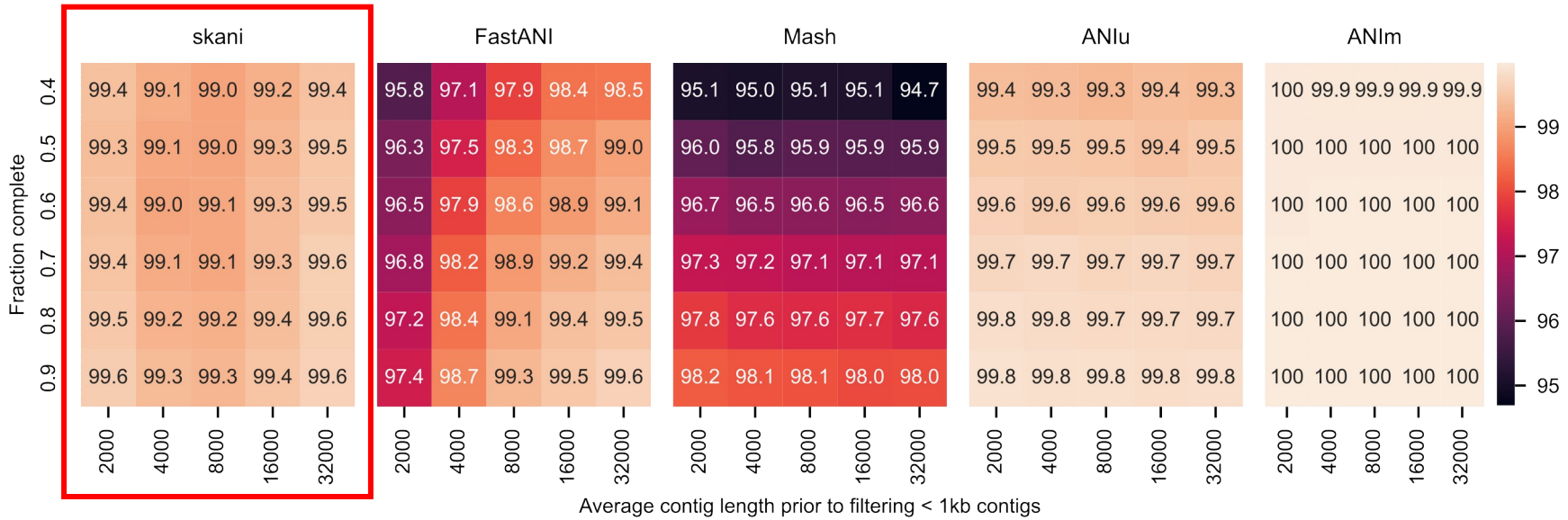
- FracMinHash sketched k-mer filtering > 80% ANI (like sourmash, Irber et al. 2022)
- Sparse k-mer seed-chain-mapping (like minimap2, Li 2018)
- Estimate ANI by using unbiased k-mer statistics only on **orthologous regions**
  - Builds on theoretical work explaining ANI estimation bias in seeding (Belbasi et al. 2022, Hera et al. 2023)

skani - fast, accurate, **robust**



UNIVERSITY OF TORONTO

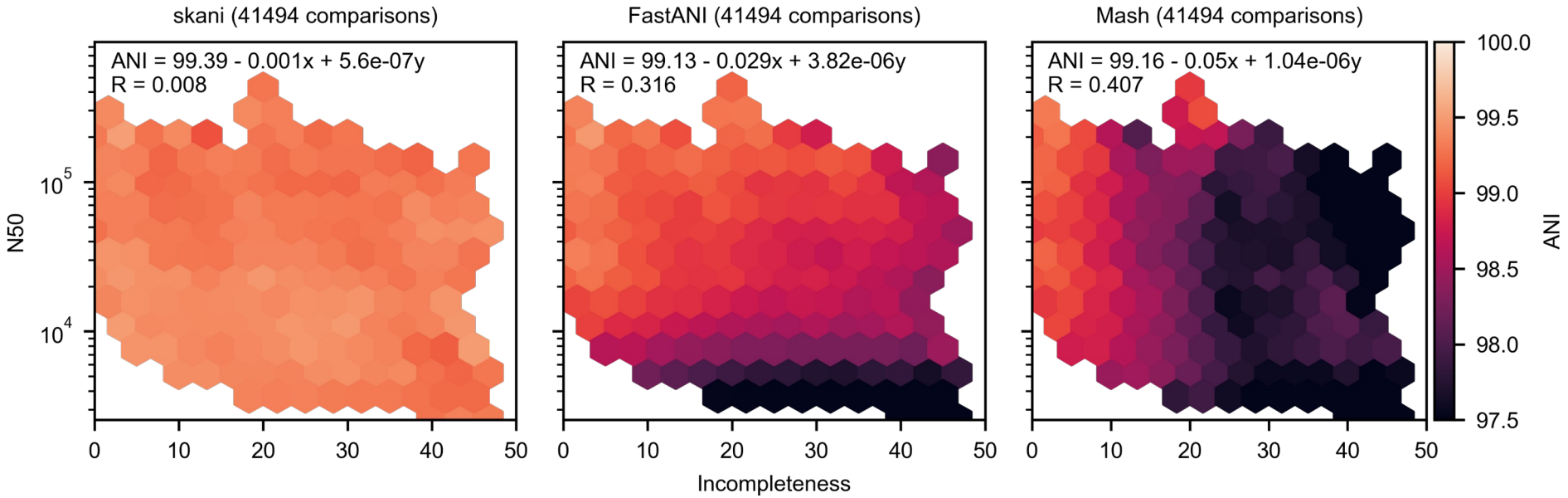
skani is more robust - **simulated**



UNIVERSITY OF TORONTO

skani is more robust - **real**

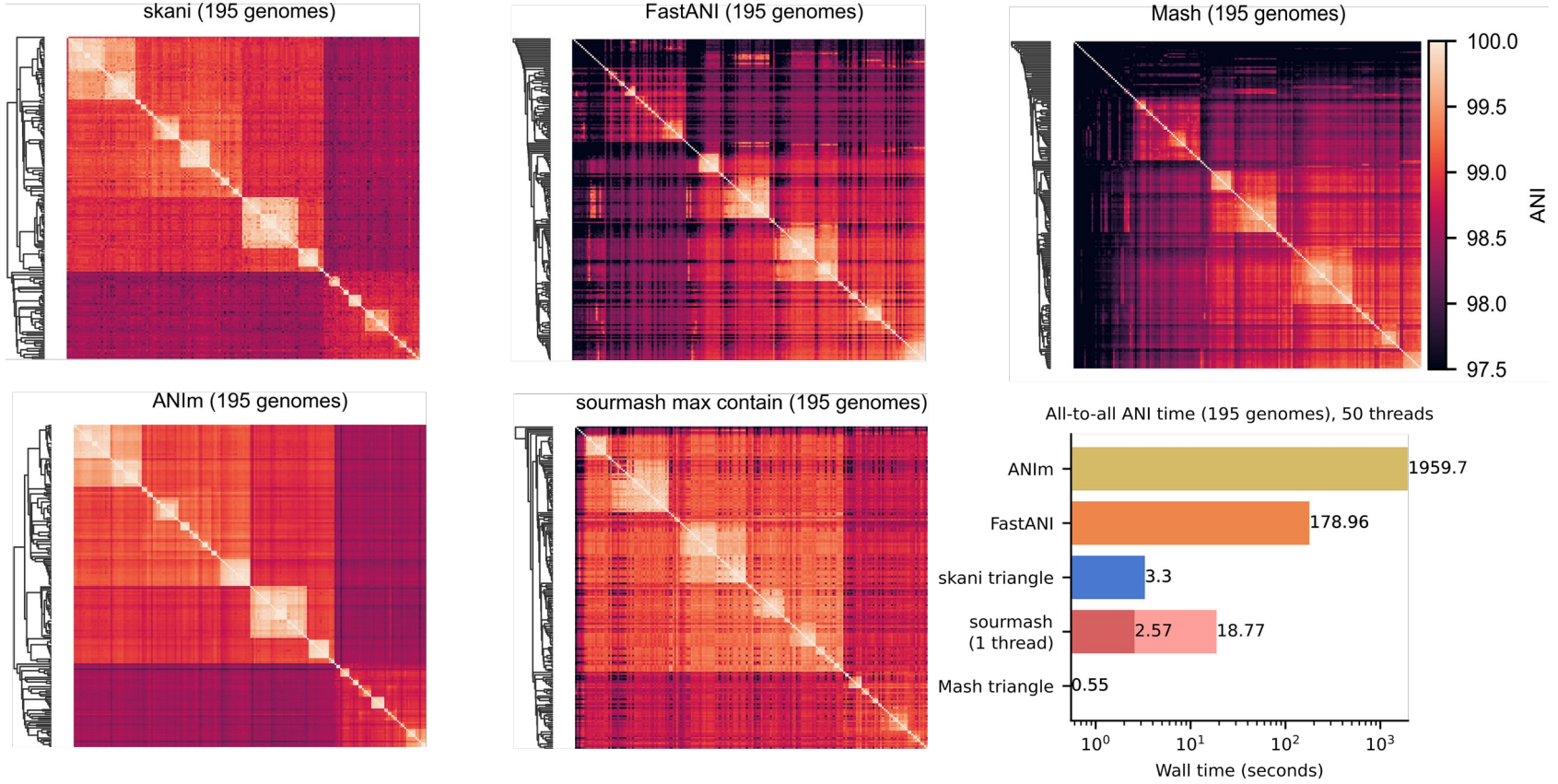
Real MAGs from Pasolli et al (2019) with > 99% ANIm estimate



UNIVERSITY OF TORONTO

skani gives **better** downstream results **faster**

Clustering on 195 *Alistipes ihumii* MAGs from Pasolli et al (2019).



~50x faster than FastANI, > 500x faster than ANIm, slower than Mash

skani can do database search

- Searching **an *E.coli* genome** against GTDB R214 (85,202 genomes)
  - **7 seconds**
  - **6 GB of RAM**
  - **1 thread**
- Searching **an entire long-read assembly (300 Mb contigs)** against GTDB R214 (85,202 genomes)
  - **1.5 minutes**
  - **20 GB of RAM**
  - **10 threads**

UNIVERSITY OF TORONTO

Key takeaways

- **MAGs are noisy** - induces bias in ANI calculation
  - Mash: incompleteness lowers ANI
  - FastANI: low N50 lowers ANI
  - ANIm, ANIb: good but slow
- skani is more **robust against noise – better downstream results**
- skani is **fast** (500x faster than MUMmer method)
- skani can **do database search –** search 85,202 genomes in seconds

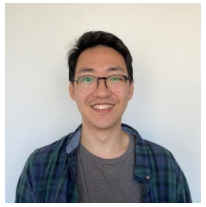
Conclusion

Github

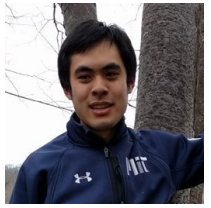
Paper

Fast and robust metagenomic sequence comparison through sparse chaining with skani

- Thanks for the referees for lots of benchmarking advice
- Check our poster out! (B-180)



Jim Shaw
(PhD student)



Yun William Yu
(PhD advisor)

UNIVERSITY OF TORONTO

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

Canada

References

Pasolli, E. et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell 176, 649–662.e20 (2019).

Ondov, B. D. et al. Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biology 17, 132 (2016).

Pierce, N. T., Irber, L., Reiter, T., Brooks, P. & Brown, C. T. Large-scale sequence comparisons with sourmash (2019).

Hera, M. R., Pierce, T. & Koslicki, D. Debiasing FracMinHash and deriving confidence intervals for mutation rates across a wide range of evolutionary distances. bioRxiv 2022.01.11.475870 (2022).

Belbasi, M., Blanca, A., Harris, R. S., Koslicki, D. & Medvedev, P. The minimizer Jaccard estimator is biased and inconsistent. Bioinformatics 38, i169–i176 (2022).

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications 9, 5114 (2018).