From Species to Strains: Developing Metagenomics for Fast, Accurate, and Precise Pathogen Identification

Sehgeet Kaur (sehgeetk@vt.edu)
Boris Vinatzer (vinatzer@vt.edu)

Sehgeet Kaur[1,2], Sarah Hoekema[3] Parul Sharma[4], Eric Newberry[5], Tiffany Lowe-Power[6], N. Tessa Pierce-Ward[7], Reza Mazloom[8], Lenwood S. Heath[8], Boris A. Vinatzer[1]

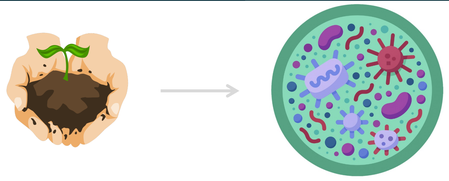
1 School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, USA, 2 Genetics, Bioinformatics, and Computational Biology Graduate Program, Virginia Tech, Blacksburg, VA, USA, 3 Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, US, 4 School of Medicine, Emory University, Atlanta, GA, USA, 5 Science and Technology, Animal and Plant Health Inspection Service, USDA, 6 Department of Plant Pathology, University of California, Davis, CA, USA, 7 School of Veterinary Medicine, University of California, Davis, CA, USA 8 Department of Computer Science, Virginia Tech, Blacksburg, VA, US

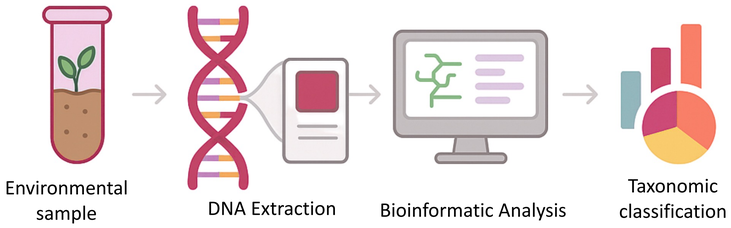
Abstract

Fast, accurate, and precise pathogen identification is critical for plant **disease surveillance** and **biosecurity**. **Metagenomics**, *i.e.*, sequencing all DNA in a sample and identifying all organisms in it, is a promising approach to address this need. However, most metagenomic classification software tools rely on species-rank taxonomy, failing to distinguish between strains of the same species that have different host ranges, differ in other phenotypes, or have different geographic distributions. Therefore, we explored how metagenomic classification can be improved by combining two different taxonomic profilers (**Kraken2 and Sourmash**) with different taxonomies including **NCBI**, **GTDB**, and custom taxonomy based on **Average nucleotide identity (ANI)** known as **Life Identification Numbers (LINs)**. To evaluate this approach, we used metagenomes of plant samples that were infected with ***Ralstonia solanacearum* species complex (RSSC)** strains**,** using short read datasets. We built reference databases using representative genomes at multiple ANI thresholds (95%, 99.5%, and 99.975%) to explore how resolution impacts the accuracy and specificity of profiling. Our initial results show that NCBI taxonomy can lead to incorrect assignments, due to **mislabeled genomes**. In contrast, LIN taxonomy provided accurate species and **sequevar-level assignments**. We are currently working to integrate GTDB taxonomy to complete the comparison. This work lays the foundation for building flexible, genome resolution aware tools for metagenomic analysis.

What is metagenomics?

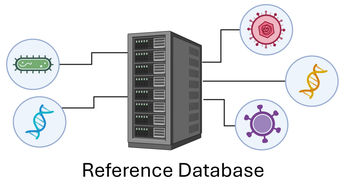
Branch of genomics that involves the study of genetic material recovered directly from environmental samples





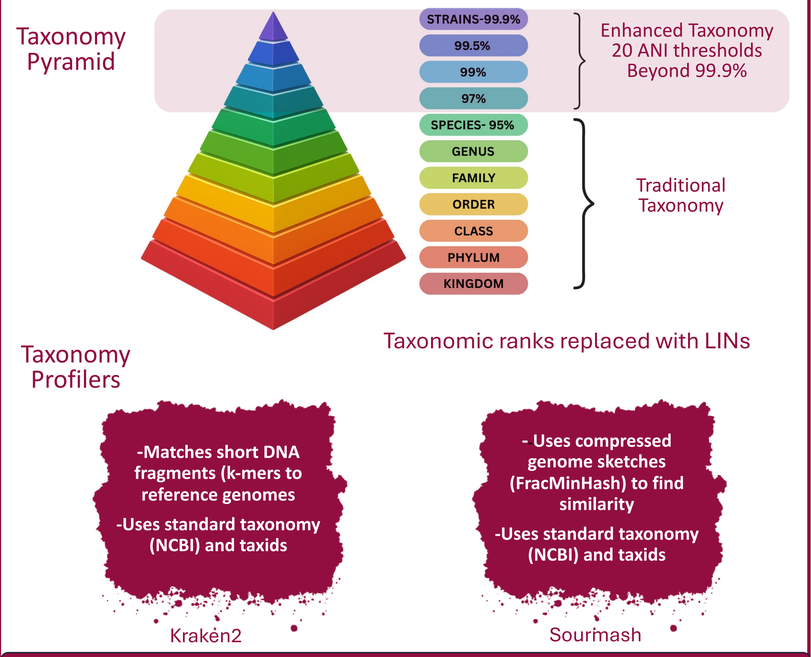
Critical challenge of Metagenomics: Difficulty of Precise Pathogen Identification in Diverse Metagenome

GTDB NCBI

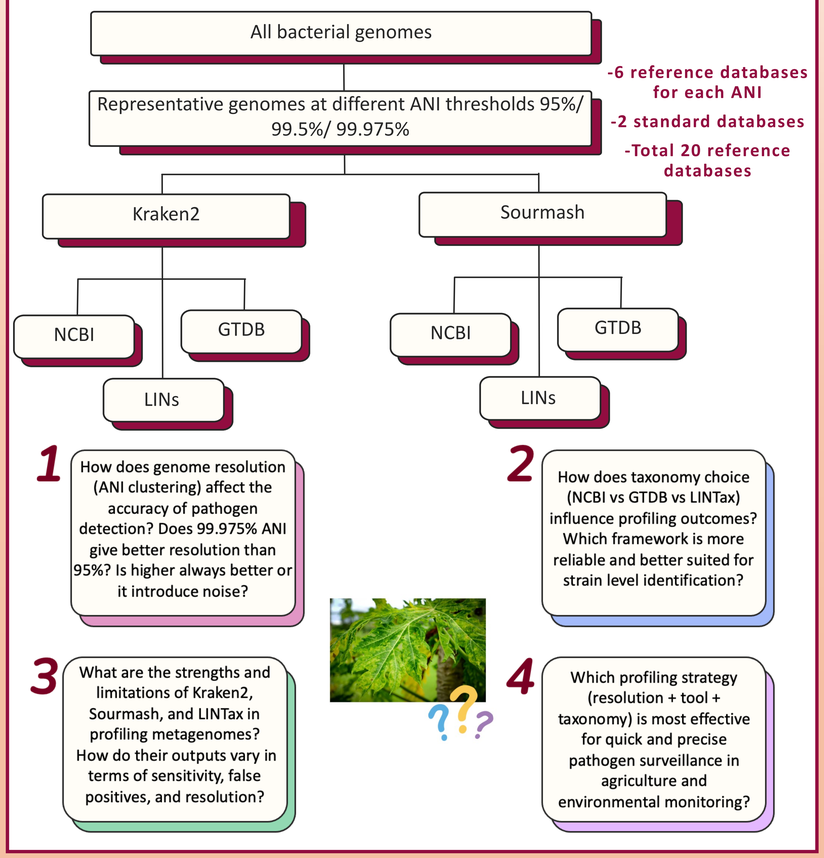


- Taxonomic assignments are inconsistent, missing, inaccurate.
- Taxonomic assignments are frequently updated by scientists, but changes are not always reflected in public databases.
- Neither NBCI or GTDB assign taxid at within-species level.

Enhanced Taxonomy and Taxonomic Profilers

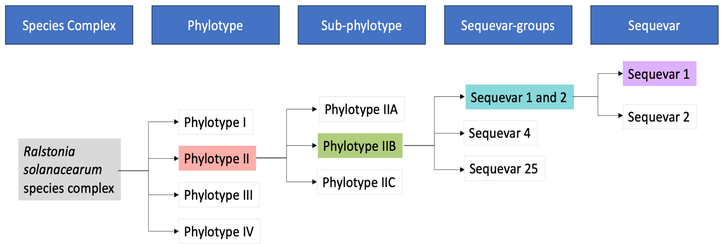


Comparative framework for metagenomic profiling



Results

Ralstonia solanacearum species complex &
select agent "Race 3 Biovar 2 (R3Bv2)" within-species lineage (sequevars 1 and 2)



Example classification results of metagenome with R3Bv2

	Standard Kraken2 (reads)	LINtax Kraken2 (reads)	Standard Sourmash (average abundance)	LINtax Sourmash (% containment)
RSSC	3,189,445	3,225,037	NA	NA
R. solanacearum	3,170,954	3,130,609	138	4.99
R. pseudosolanacearum	589	5,798	0	0
R. syzygii	231	998	0	0
Phylotype I	NA	104	NA	0
Phylotype II	NA	3,130,609	NA	4.99
Phylotype IIA	NA	12,813	NA	0
Phylotype IIB	NA	2,441,812	NA	4.99
Phylotype IIC	NA	2,703	NA	0
Phylotype III	NA	191	NA	0
Phylotype IV	NA	48	NA	0
Sequevar 1	NA	60,428	NA	4.99
Sequevar 2	NA	857	NA	0

Example classification results of metagenome with *R. pseudosolanacearum*

	Standard Kraken2	LINtax Kraken2	95% ANI Kraken2	99.975% ANI Kraken2
RSSC	14,331,337	14,911,822	13,279,335	7,956,163
R. solanacearum	155,582	168,246	13,049,118	300,799
R. pseudosolanacearum	8,750,675	14,128,326	0	133,684
R. syzygii	27,094	54,583	186,086	10,059

Conclusion and Future Directions

- Both Kraken2 and Sourmash can identify sequevars when replacing NCBI taxonomy with LINs.
- Classification using LINtax Kraken2 resulted in some false positives, likely due to limited reference genomes, adding representative genomes from other genera is expected to improve the accuracy.
- The use of NCBI taxonomy led to misclassification, highlighting its limitations.
- Current analysis is based on short reads; evaluation of long reads and GTDB taxonomy is ongoing.
- This study highlights the need for high-resolution taxonomies and diverse reference databases for accurate pathogen detection.

Acknowledgement

USDA APHIS
COLLEGE OF AGRICULTURE AND LIFE SCIENCES
SCHOOL OF PLANT AND ENVIRONMENTAL SCIENCES
VIRGINIA TECH
AP20PPQS &T00C055
TPSC TRANSLATIONAL PLANT SCIENCES CENTER
GBCB
NSF
DBI-2018522