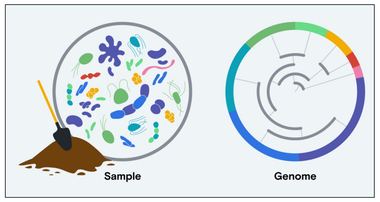
Developing a Pipeline to Improve Pathogen Identification in Metagenomes

Hoekema, S.[1], Pierce-Ward, N.T.[2], Kaur, S.[3], Vinatzer, B.A.[3]

1 Oregon State University, 2 University of California, Davis, 3 Virginia Tech

INTRODUCTION

Metagenomics is the study of the collective genetic material of a microbial community in an environment, such as the microbes, including pathogens, present in a diseased plant. Classification tools like Sourmash and Kraken2 are often used to assign the obtained sequences to species to determine the microbial diversity in the community, possibly including the pathogen that caused a disease in a human, animal, or plant with an unknown disease.

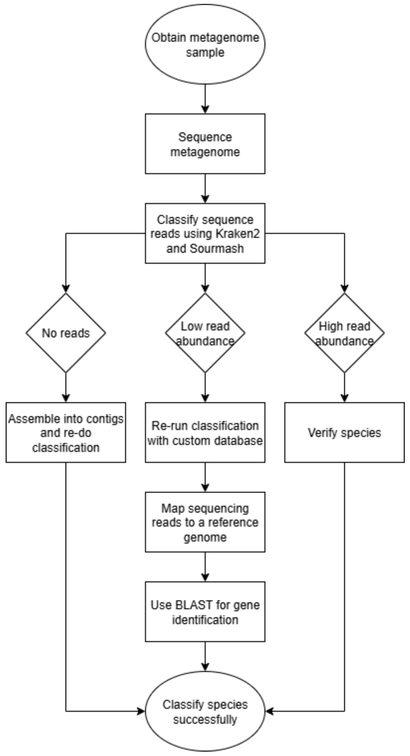


Typically, the lowest level classification of organisms with classification tools is the species, but with tools that utilize a taxonomic ranking system called LINs, more specific strain-level classifications can be made.

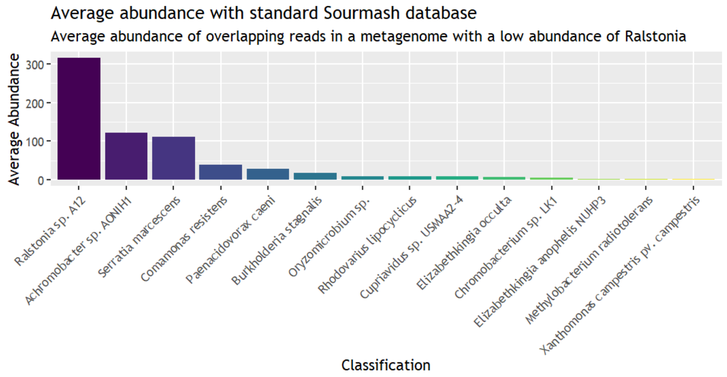
HYPOTHESIS

If methods like mapping reads to a reference, performing BLAST searches, and using custom databases for classification software are used, then low abundance species will be classified resulting in fewer false negatives.

METHODOLOGY

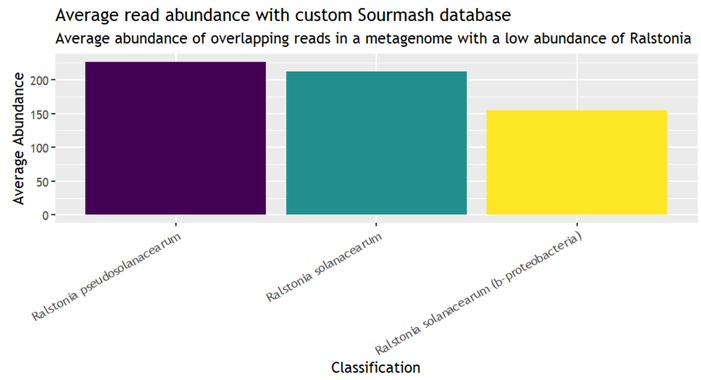


RESULTS



Using a standard Sourmash database with a metagenome with a low abundance of bacteria in the genus *Ralstonia* indicates that there is *Ralstonia* in the metagenome, but many other organisms are identified in the sample as well.

Using a custom *Ralstonia* Sourmash database with a metagenome with a low abundance of *Ralstonia* returns more specific results for Ralstonia, resulting in a higher abundance of overlapping reads identified.



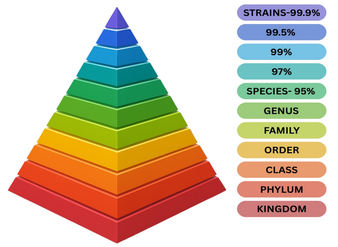
Percent containment of metagenome reads for each Ralstonia species, phylotype, and sequevar

Metagenome	*Ralstonia solanacearum*	*Ralstonia pseudo-solanacearum*	*Ralstonia syzygii*	Phylotype I	Phylotype II	Phylotype IIA	Phylotype IIB	Phylotype IIC	Phylotype III	Phylotype IV	Sequevar 1	Sequevar 2
SRR29654720	0.02	37.36	0.37	36.03	0.02	0	0.02	0	0.01	0.37	0	0
SRR16817055	0	28.67	0	0	0	0	0	0	0.01	0	0	0
SRR29654726	0.07	33.29	0.33	32.7	0.07	0.07	0	0	0.01	0.33	0	0
SRR16817052	0.1	20.56	0	0	0.1	0	0.1	0	0	0	0	0
SRR16817058	0	20.59	0	0	0	0	0	0	0	0	0	0
SRR7340328	0	0.1	0	0	0	0	0	0	0	0	0	0
SRR25013881	0	0.12	0	0.12	0	0	0	0	0	0	0	0
SRR28044133	0	0.12	0	0.12	0	0	0	0	0	0	0	0
SRR15510224	0	0.11	0	0.01	0	0	0	0	0	0	0	0
SRR28044136	0	0.11	0	0.11	0	0	0	0	0	0	0	0
ERR3674559	0.4	1.04	0.15	0	0.4	0	0.4	0	0	0.15	0	0
SRR25388119	0.53	1.32	0.15	0	0.53	0	0.53	0	0	0	0	0
SRR16588138	0.43	1.09	0.19	0	0.43	0.43	0	0	0	0.19	0	0
SRR22387909	1.03	0.72	0.19	0.6	1.03	0.9	0.14	0	0	0	0	0
SRR25388137	0.39	1.22	0.11	0	0.39	0	0.39	0	0	0	0	0

Running Sourmash with the taxonomic ranking system LINs allows more specific taxonomic levels to be identified. For *Ralstonia* metagenomes, the taxonomic levels phylotype and sequevar of the *Ralstonia* species could be obtained when using the custom *Ralstonia* database.

CONCLUSIONS

Using a custom database with classification tools like Sourmash was more successful in classifying expected organisms in a metagenome sample. With the low abundance *Ralstonia* metagenome, there was a higher abundance of reads identified for *Ralstonia* with the custom Sourmash database than with the standard one, making it more reliable for identifying an expected organism. This provides an improved option for identifying a species when there are few reads in a metagenome.



Using a LINs-based approach with classification tools helped identify more specific taxonomic levels in *Ralstonia*, which allows for better identification of the type of bacteria and its characteristics. More specific identification enables people to better handle issues like plant pathogen infections when provided with the specific rankings of a species and strain.

FUTURE WORK

- Use additional bioinformatic tools to confirm Sourmash LIN results.
- Use additional metagenomes for which the microbial composition is known to determine false positive and false negative rates.
- Adapt the Sourmash LIN approach to additional plant, animal, and human pathogens to expand its use from plant pathology to veterinary and human medicine.

FUNDING

Support from the Agriculture and Food Research Initiative program of the National Institute of Food and Agriculture, USDA (grant #2020-68018-30674).