8/5/2025
P-186

Genome-based Identification of Prokaryotic Plant Pathogens at Within-Species Resolution Using the genomeRxiv Web Server

Mazloom R.[1], Heath L. S.[1], Brown C. T.[2], Pritchard L.[3], Pierce-Ward N. T.[3], Vinatzer B. A.[1]
[1]Virginia Tech, Blacksburg, VA, USA; [2]University of California at Davis; [3]University of Strathclyde, Glasgow,UK

rmazloom@vt.edu
vinatzer@vt.edu

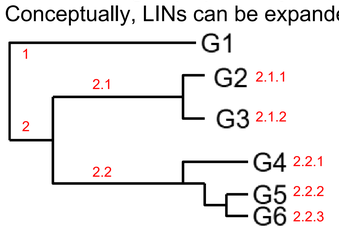
Introduction

New human, animal, and plant pathogens can emerge any time without warning and may start spreading widely before they are detected, making their containment challenging or impossible. The continued exponential surge of complete genome assemblies of pathogens is being assisted by the increase in accuracy, throughput, and accessibility to sequencing tools and technologies. However, this surge is also creating pressure on many bioinformatic processes, such as taxonomic assignment, which are done mostly in curation processes to this day. This curative process not only increases the time needed to obtain information from the genomes but also reduces our ability to focus on fine grained details, such as information on groups within species or outbreaks. We present genomeRxiv as a web service, backed by a database and pipeline, that identifies and assigns taxonomies to genomes from the genus rank towards the strain level.

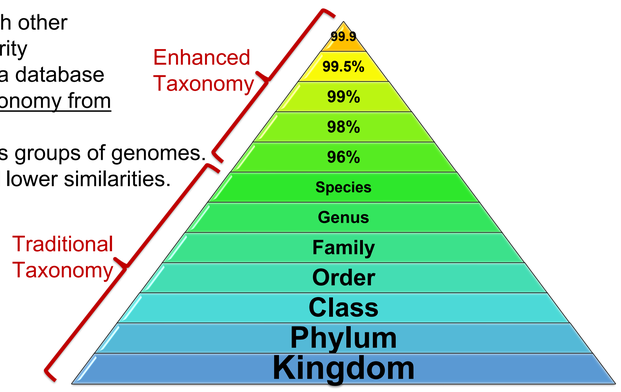
Methods

Life Identification Numbers (LINs)

- LINs represent precisely how similar genomes are to each other
- The current LINs uses ANI as measure of genome similarity
- LINs are indices that automatically organize genomes in a database based on reciprocal similarity (expanding hierarchical taxonomy from the species to almost the individual).
- LINs and LINgroups precisely circumscribe within-species groups of genomes.
- Conceptually, LINs can be expanded to cover higher and lower similarities.



Within-Species Resolution



Genus → Species

within Species

%ANI Similarity

Strain	70 A	75 B	80 C	85 D	90 E	95 F	96 G	97 H	98 I	98.5 J	99 K	99.25 L	99.5 M	99.75 N	99.9 O	99.925 P	99.95 Q	99.975 R	99.99 S	99.999 T	Similarity	Closest
SM116	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	SM116
SM120	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	99.99292	SM116
SM85	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	99.76014	SM116
SM109	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	99.57089	SM116
SM97	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	99.52636	SM109

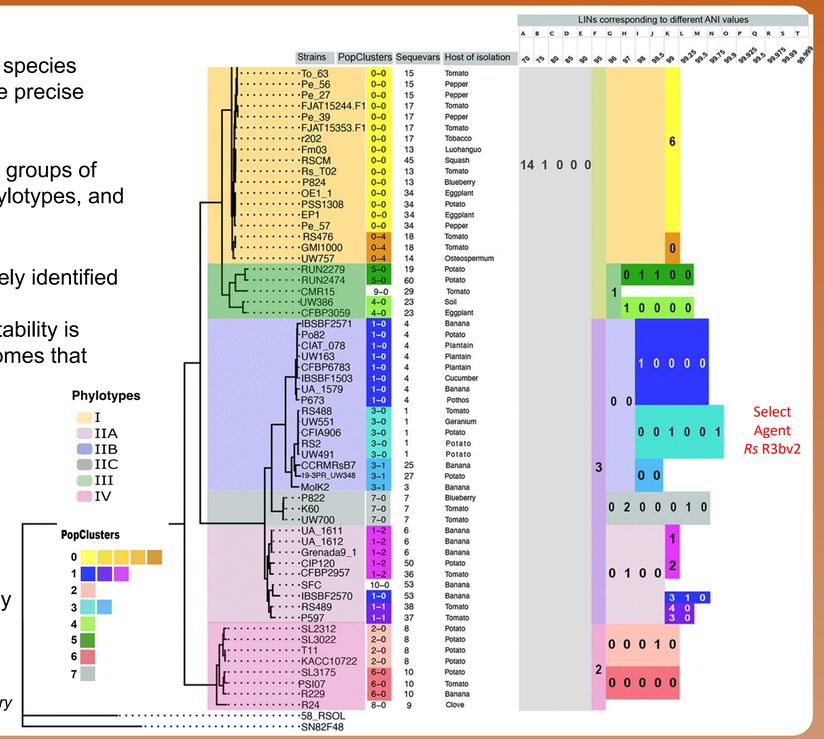
Discussion

- LINs assigned to individual *Ralstonia solanacearum* species complex (RSSC) genomes in genomeRxiv reflect the precise genome similarity among genomes
- LINs reflect core genome phylogenetic relationships
- LINs assigned to genomes are used to circumscribe groups of genome (LINgroups) that correspond to species, phylotypes, and population clusters.
- LINgroups are annotated in genomeRxiv
- Genomes of putative RSSC genomes can be precisely identified as members of the circumscribed LINgroups
- While LINgroup circumscriptions may get updated stability is provided by the LINs assigned to the individual genomes that remain the same.

Data

- 50,000 bacterial species circumscribed on genomeRxiv using LINs
- 310,000 bacterial genomes assigned LINs
- 5.8 million metadata entries linked to genomes
- **Archaea** species integration in progress
- **Fungal** and **Viral** genome LIN assignment underway

Sharma, P., Johnson, M. A., Mazloom, R., Allen, C., Heath, L. S., Lowe-Power, T. M., & Vinatzer, B. A. (2022). Meta-analysis of the Ralstonia solanacearum species complex (RSSC) based on comparative evolutionary genomics and reverse ecology. Microbial Genomics, 8(3).



Conclusion

We expect the high resolution, speed, ease of use, and online availability of genomeRxiv to facilitate genome-based identification for applications in environmental microbiology, biosecurity, and disease surveillance.

References

LINflow python implementation of LINs
https://code.vt.edu/linbaseproject/LINflow
Version1 publication at https://doi.org/10.7717/peerj.10906
BrookLIN https://code.vt.edu/linbaseproject/linbase_api

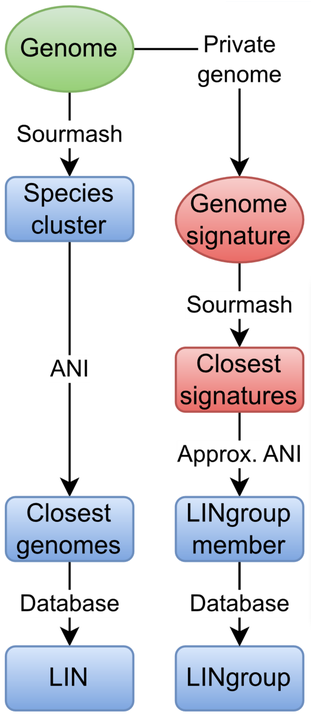
GenomeRxiv
https://www.nsf.gov/awardsearch/showAward?AWD_ID=2018522

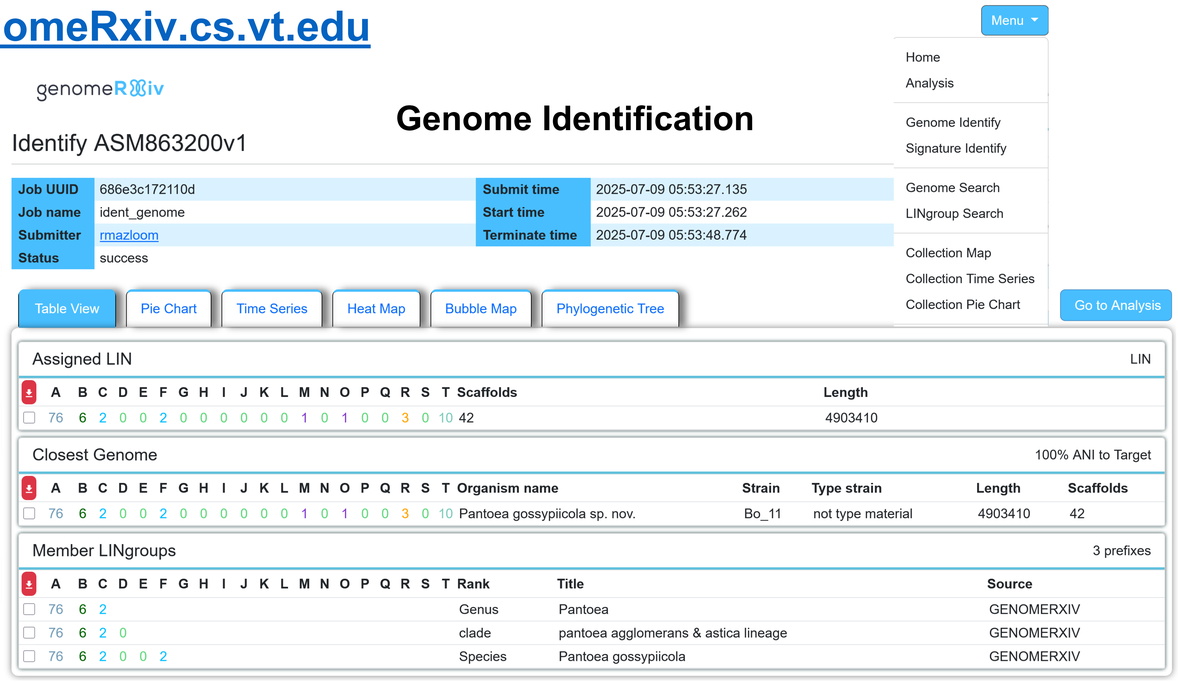
Pantoea project *Crosby, K. C., Rojas, M., Sharma, P., Johnson, M. A., Mazloom, R., Kvitko, B. H., ... & Vinatzer, B. A. (2023). Genomic delineation and description of species and within-species lineages in the genus Pantoea. Frontiers in Microbiology, 14, 1254999.*

Funding

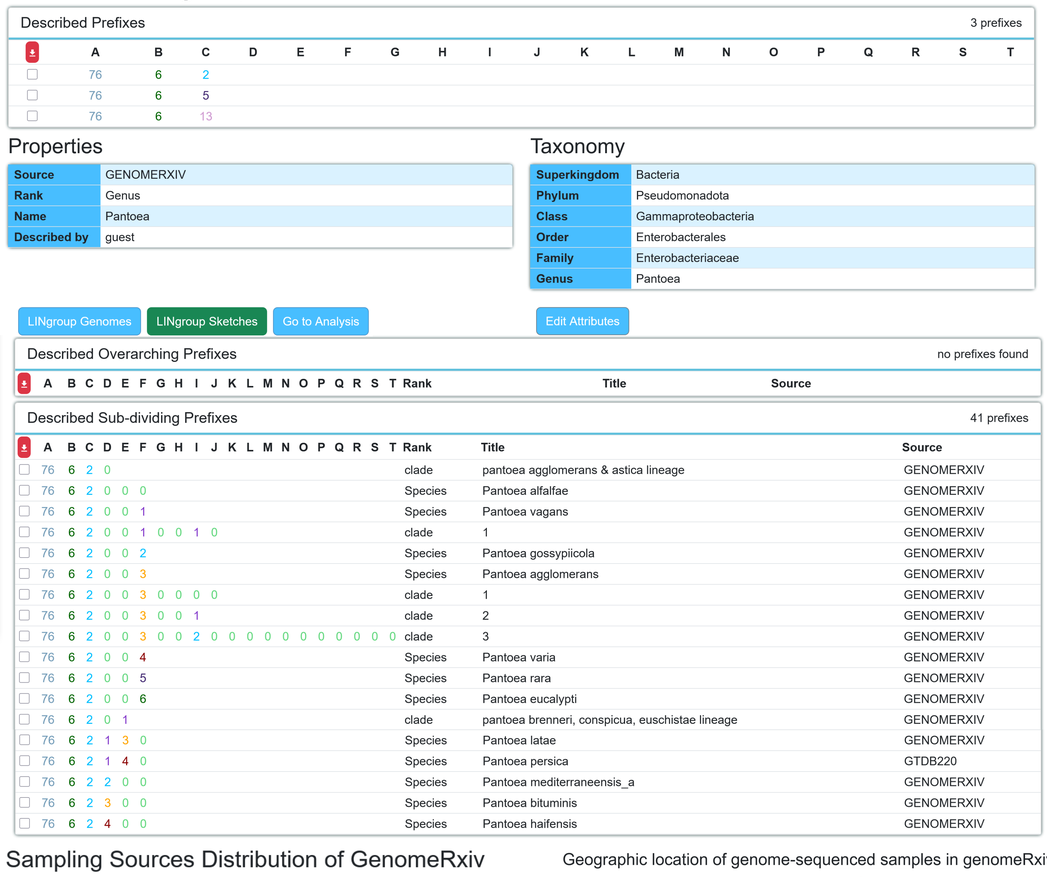
Development of the *genomeRxiv* platform is supported by the US NSF (DBI-2018522) and the UK BBSRC (BB/V010417/1) and the CAIA/Sanghani Center Graduate Scholarship

Results in GenomeRxiv.cs.vt.edu





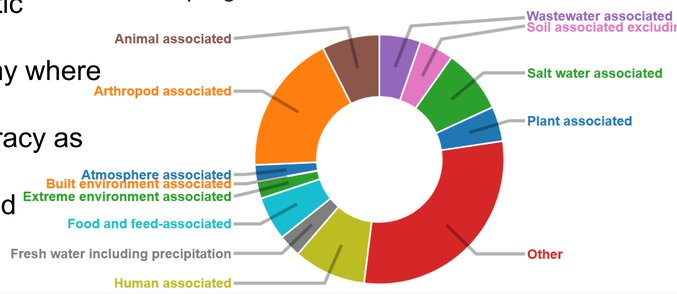
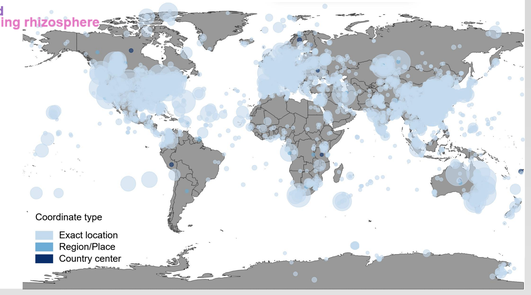
Selected LINgroup



genomeRxiv is a publicly accessible website designed to cluster and identify prokaryotic genomes using their nucleotide sequence or their sourmash signatures (when sharing sequence is not ideal).

- Users can not only identify their genomes, but also contribute them to genomeRxiv and describe their identified LINgroups, increasing the speed, depth and breadth of future runs
- LINs can be used to serve as a bridge between traditional and phylogenetic taxonomy using genome similarity as a unit
- genomeRxiv LINs can be used as a template for similarity-based taxonomy where the similarity criteria could be refined or changed as needed
- Similar to learning models, more users will result in better clustering accuracy as more genomes, signatures, and group descriptions are added
- Encourages and facilitates collaboration between parties working in related organism groups
- Has a command line interface (CLI) called BrookLIN for automation

Sampling Sources Distribution of GenomeRxiv



Geographic location of genome-sequenced samples in genomeRxiv