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

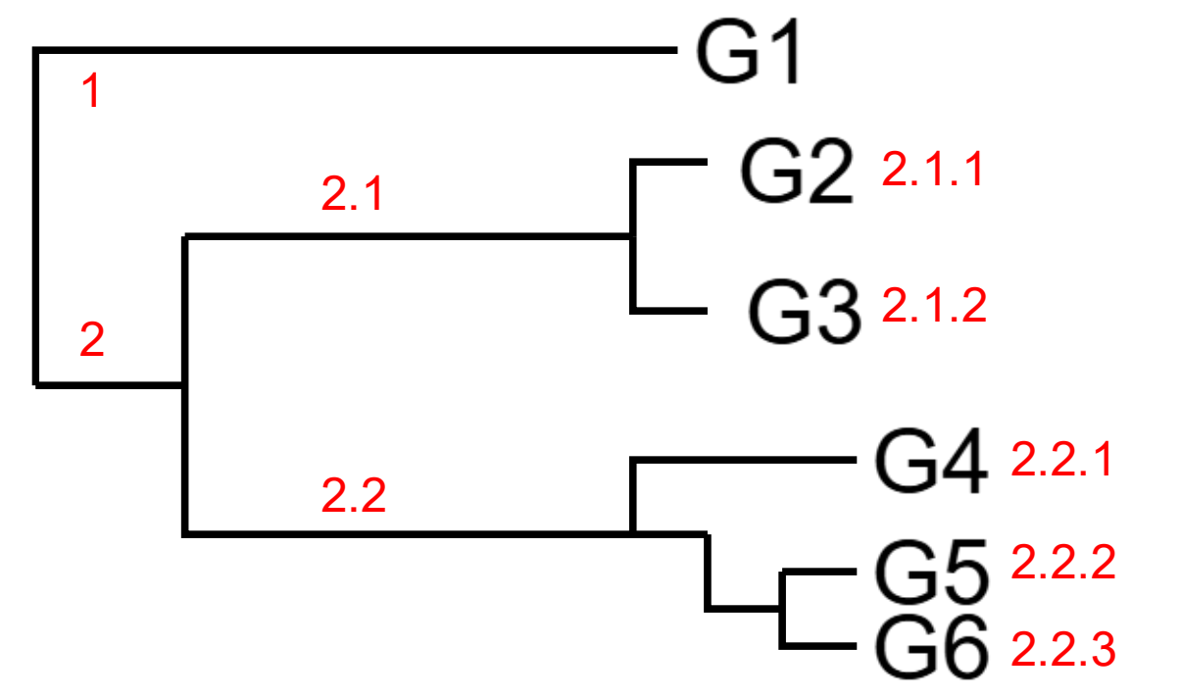
Introduction

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.

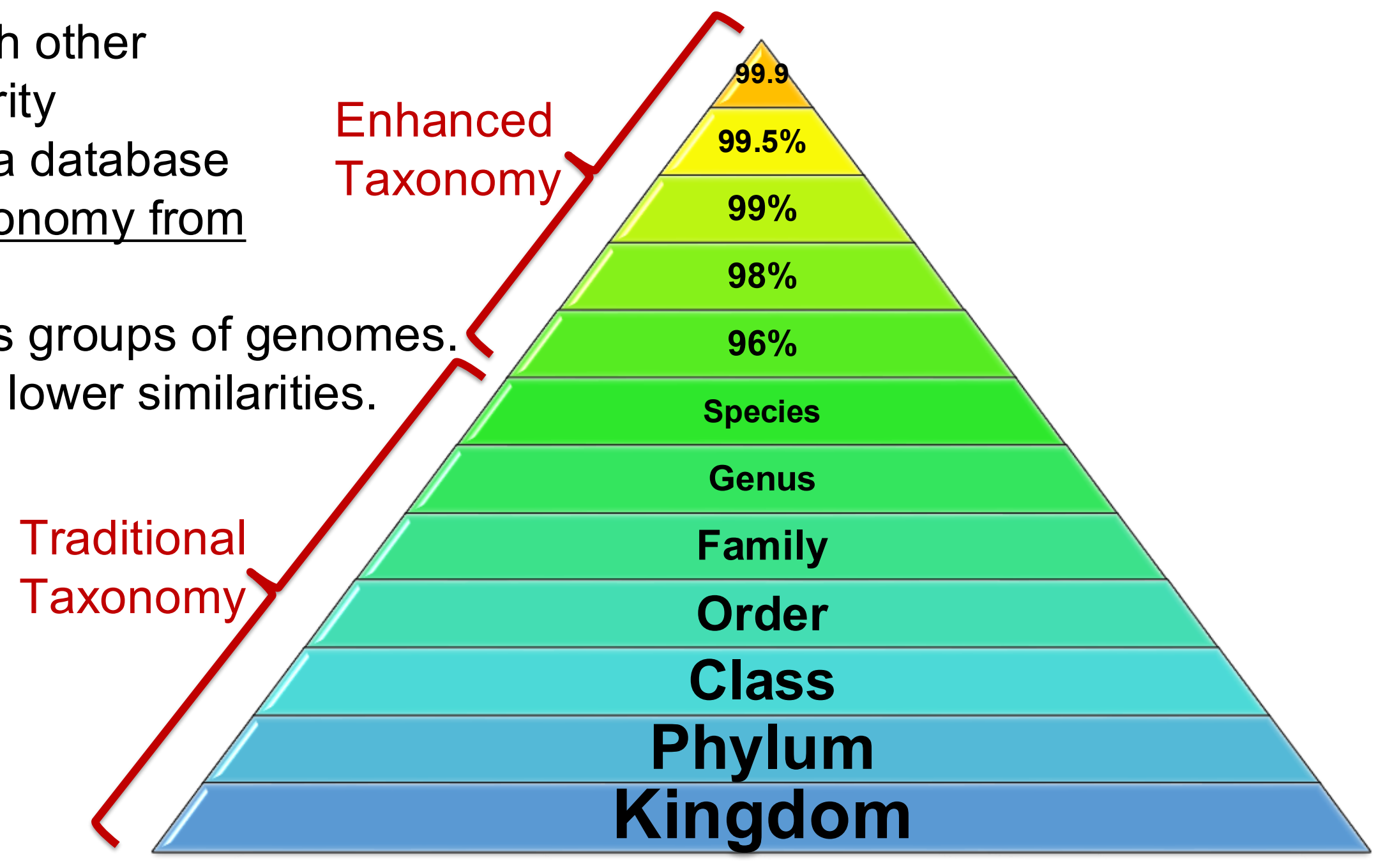
Methods

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



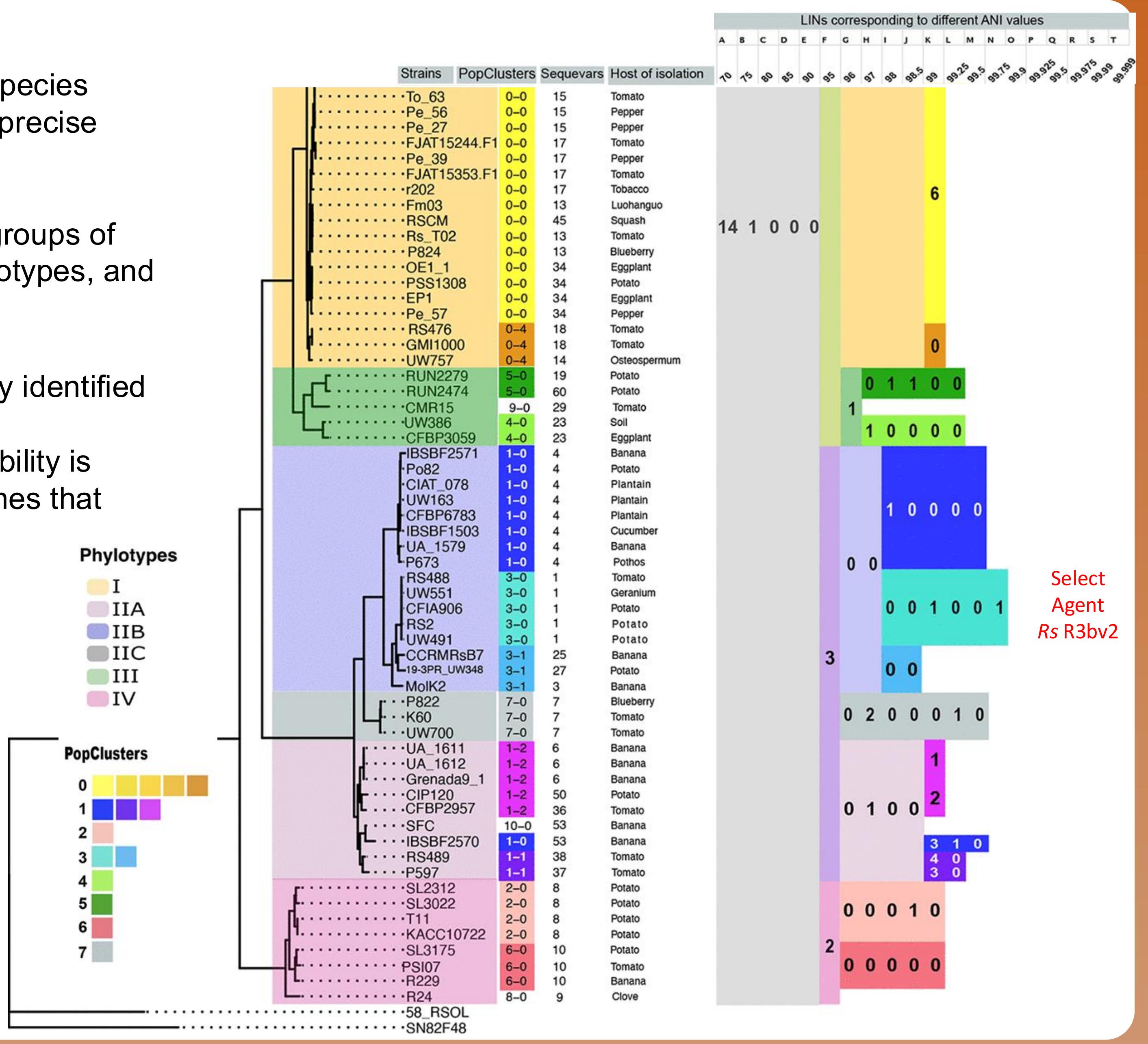
Genus → Species

within Species

%ANI Similarity	70	75	80	85	90	95	96	97	98	98.5	99	99.25	99.5	99.75	99.9	99.925	99.95	99.975	99.99	99.999		
Strain	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	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.



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).

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

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

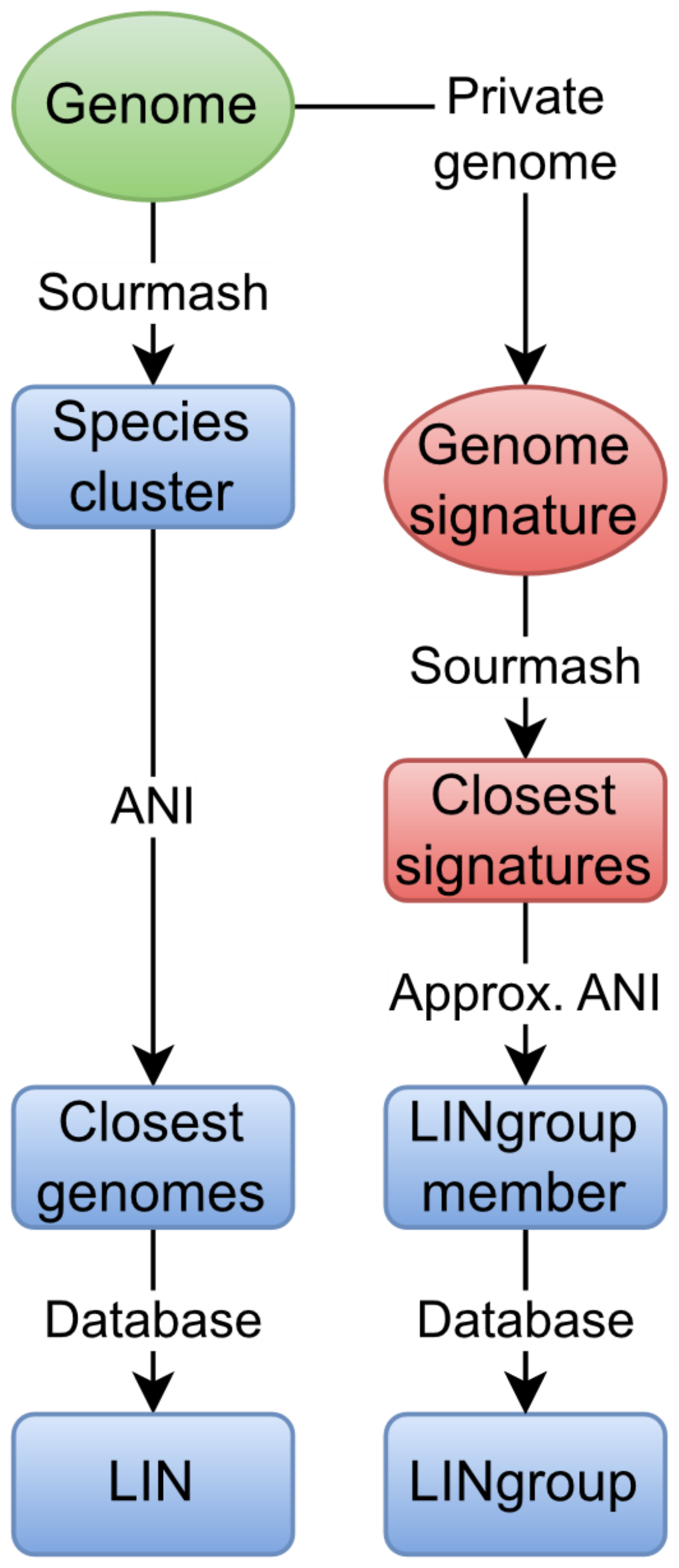
GenomeRxiv
https://www.nsf.gov/awardsearch/showAward?AWD_ID=2018522

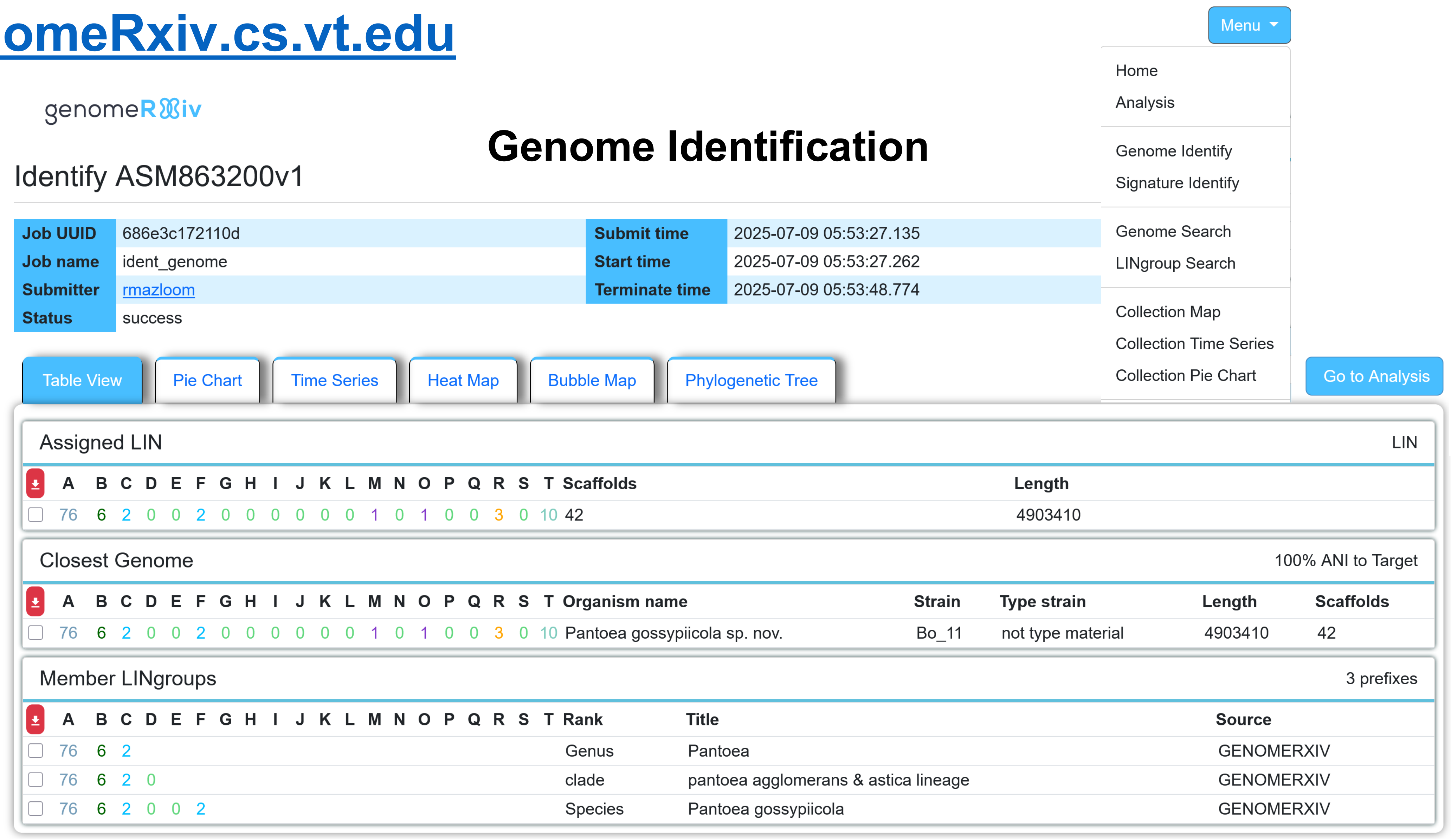
Pantoea project *Crosby, K. C., Rojas, M., Sharma, P., Johnson, M. A., Mazloom, R., Kvitko, B. H., ... & Vinatzer, B. A. (2023). Genomic delineation and description of species and within-species lineages in the genus Pantoea. Frontiers in Microbiology, 14, 1254999.*

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





Genome Identification

Identify ASM863200v1

Job UUID	686e3c172110d	Submit time	2025-07-09 05:53:27.135
Job name	ident_genome	Start time	2025-07-09 05:53:27.262
Submitter	rmazloom	Terminate time	2025-07-09 05:53:48.774
Status	success		

Assigned LIN — LIN

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Scaffolds	Length
□	76	6	2	0	0	2	0	0	0	0	0	0	1	0	1	0	0	3	0	10	42	4903410

Closest Genome — 100% ANI to Target

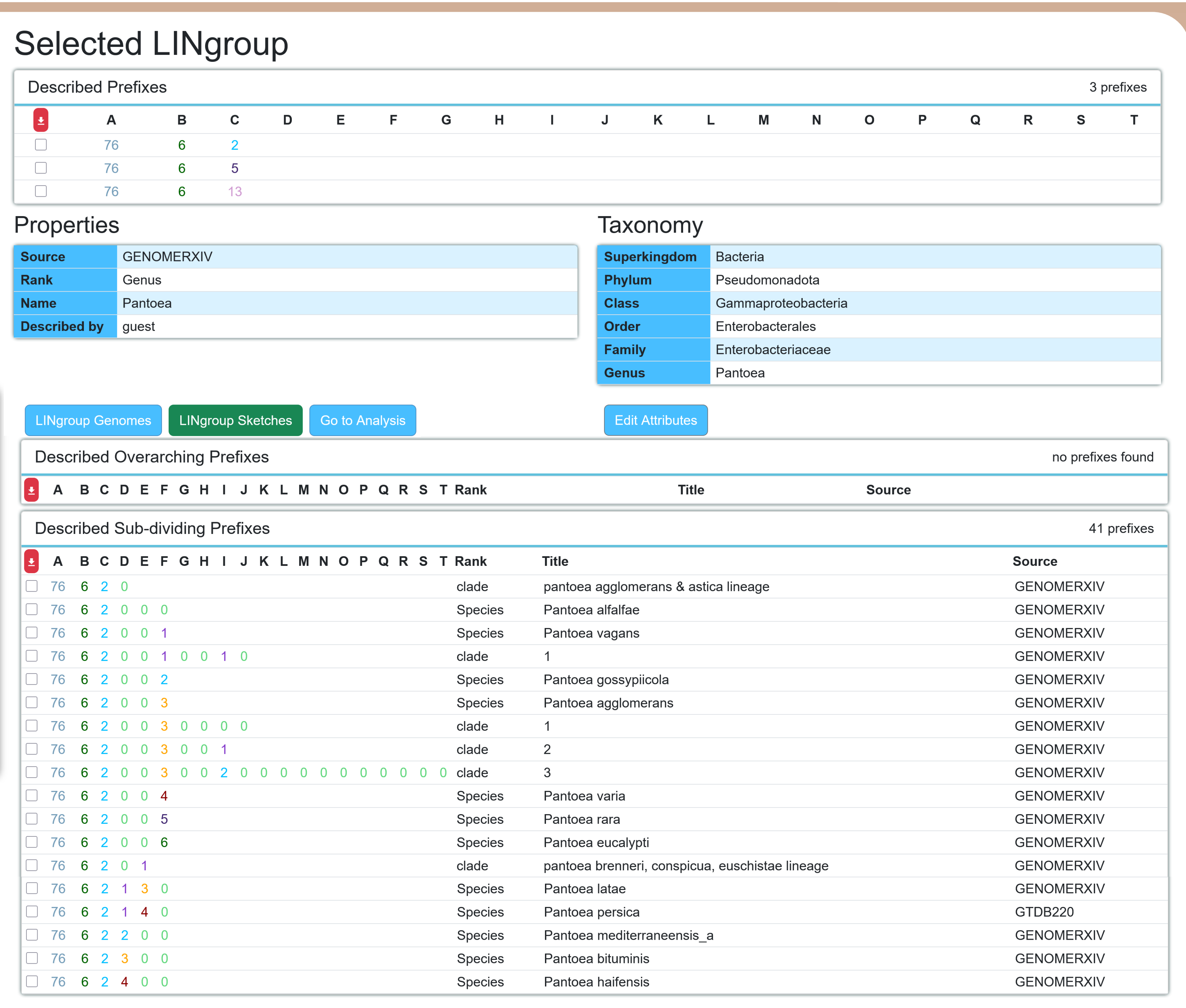
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Organism name	Strain	Type strain	Length	Scaffolds
□	76	6	2	0	0	2	0	0	0	0	0	0	1	0	1	0	0	3	0	10	Pantoea gossypiicola sp. nov.	Bo_11	not type material	4903410	42

Member LINgroups — 3 prefixes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Rank	Title	Source
□	76	6	2																		Genus	Pantoea	GENOMERXIV
□	76	6	2	0																	clade	pantoea agglomerans & astica lineage	GENOMERXIV
□	76	6	2	0	0	2															Species	Pantoea gossypiicola	GENOMERXIV

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



Selected LINgroup

Described Prefixes — 3 prefixes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
□	76	6	2																	
□	76	6	5																	
□	76	6	13																	

Properties

Source	GENOMERXIV
Rank	Genus
Name	Pantoea
Described by	guest

Taxonomy

Superkingdom	Bacteria
Phylum	Pseudomonadota
Class	Gammaproteobacteria
Order	Enterobacterales
Family	Enterobacteriaceae
Genus	Pantoea

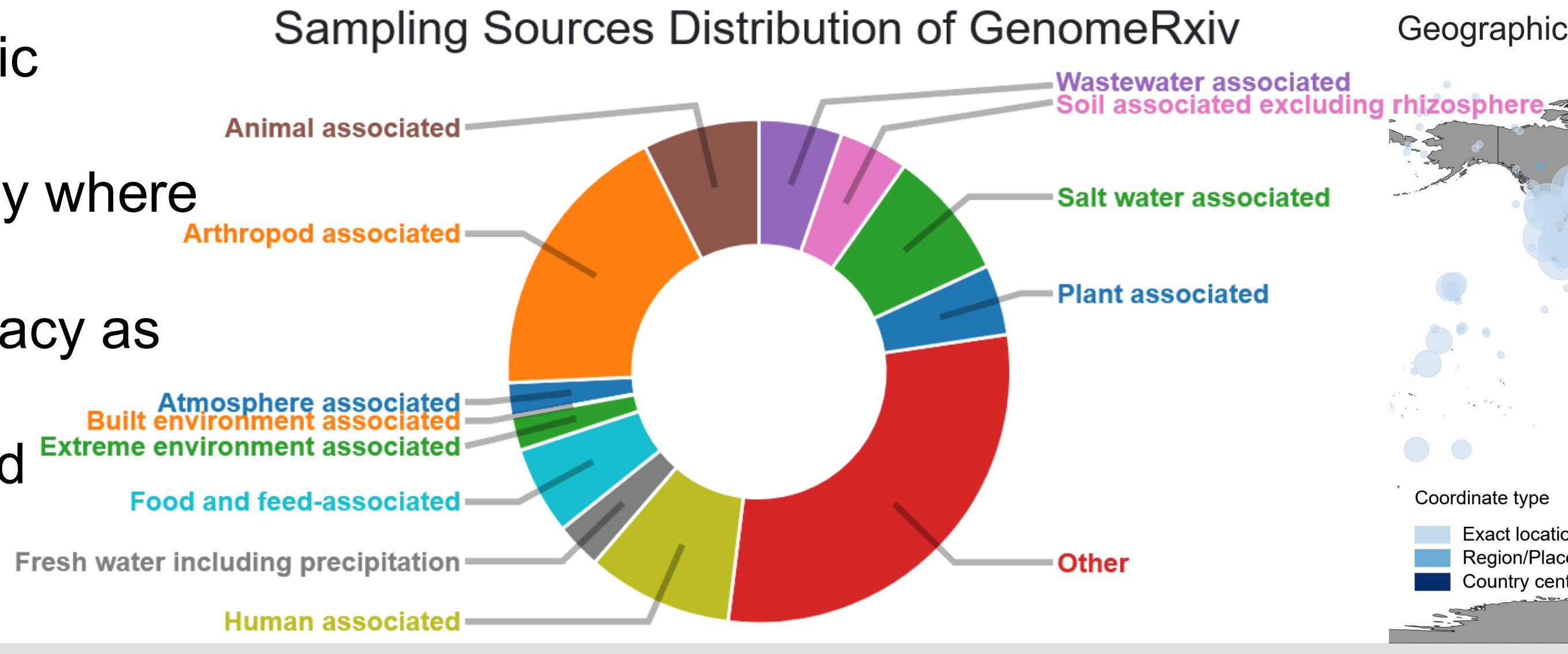
Described Overarching Prefixes — no prefixes found

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Rank	Title	Source

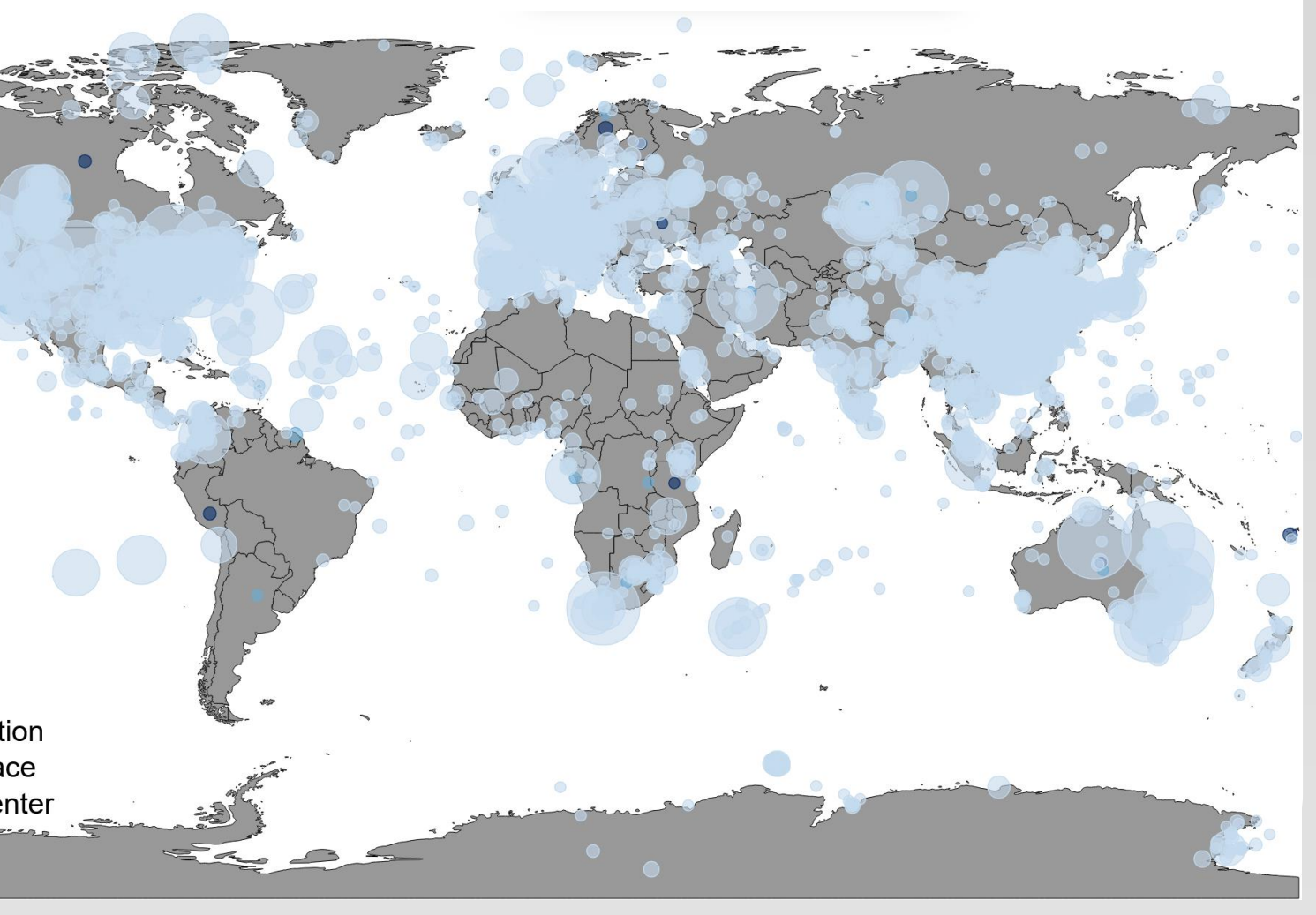
Described Sub-dividing Prefixes — 41 prefixes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Rank	Title	Source
□	76	6	2	0																	clade	pantoea agglomerans & astica lineage	GENOMERXIV
□	76	6	2	0	0	0															Species	Pantoea alfalfae	GENOMERXIV
□	76	6	2	0	0	1															Species	Pantoea vagans	GENOMERXIV
□	76	6	2	0	0	1	0	0	1	0											clade	1	GENOMERXIV
□	76	6	2	0	0	2															Species	Pantoea gossypiicola	GENOMERXIV
□	76	6	2	0	0	3															Species	Pantoea agglomerans	GENOMERXIV
□	76	6	2	0	0	3	0	0	0	0											clade	1	GENOMERXIV
□	76	6	2	0	0	3	0	0	1												clade	2	GENOMERXIV
□	76	6	2	0	0	3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	clade	3	GENOMERXIV
□	76	6	2	0	0	4															Species	Pantoea varia	GENOMERXIV
□	76	6	2	0	0	5															Species	Pantoea rara	GENOMERXIV
□	76	6	2	0	0	6															Species	Pantoea eucalypti	GENOMERXIV
□	76	6	2	0	1																clade	pantoea brenneri, conspicua, euschistae lineage	GENOMERXIV
□	76	6	2	1	3	0															Species	Pantoea latae	GENOMERXIV
□	76	6	2	1	4	0															Species	Pantoea persica	GTDB220
□	76	6	2	2	0	0															Species	Pantoea mediterraneensis_a	GENOMERXIV
□	76	6	2	3	0	0															Species	Pantoea bituminis	GENOMERXIV
□	76	6	2	4	0	0															Species	Pantoea haifensis	GENOMERXIV

Sampling Sources Distribution of GenomeRxiv



Geographic location of genome-sequenced samples in genomeRxiv


Coordinate type
Exact location
Region/Place
Country center