genomeRxiv: a microbial whole-genome database for classification, identification, and data sharing

Leighton Pritchard[1], Bailey Harrington[1], Luiz Irber[2], Reza Mazloom[3], Tessa Pierce[2], Parul Sharma[3], Lenwood Heath[3], C Titus Brown[2], Boris Vinatzer[3]

1. Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, Scotland, UK
2. University of California-Davis, USA
3. Virginia Polytechnic Institute and State University, Blacksburg, USA

1. We need a stable, genome-based classification system for microbes

The mapping of traditional taxonomic nomenclature to the history revealed through genome analysis is not exact, leading to significant challenges:

Genomic disagreement with nomenclature
genome-based classifications do not always agree with published taxonomies [1]

Genome-based classifications resolve novel taxa
genome-based classifications produce highly-resolved taxa at levels that are not represented in prokaryotic taxonomy [2]

Inaccuracies in reference databases
a significant minority of genomes in public databases are misidentified [3]

Our goal is to build ***genomeRxiv***, a "preprint genome server" that provides:

A stable, taxonomy-independent classification scheme
a transparent, quantitative "co-ordinate" scheme in sequence space, with fine-grained resolution (LINs… see right)

Genome-based quantitative identification
precise, secure and confidential taxonomy-independent classification of submitted microbial genomes

Candidate diagnostic markers
practical molecular diagnostic tools targeted at precise groups of microbial genomes

The authors gratefully acknowledge funding support from the Biotechnology and Biological Sciences Research Council (BB/V010417/1) and the National Science Foundation (2018911)

2. *genomeRxiv*

genomeRxiv will provide a service for rapid, quantitative classification of microbial genomes using **Life Identification Numbers (LINs)**, extending the existing **LINbase** service.

LINs work like map co-ordinates in sequence space. Degrees of genome sequence identity are marked with letters (e.g. A-T as in **Figure 1**; example in **Figure 2**), and numeric symbols assigned to indicate a particular grouping of genomes sharing at least that degree of identity with each other.

This string of numeric symbols precisely locates each genome in a region of sequence space. For example, in Figure 1 the LIN $0_A1_B0_C0_D0_E3_F$ circumscribes species *G1 s2*.

			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%
Genus	Species	Strain	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
G1	S1	X1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S2	X2	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S2	X3	0	1	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0
G1	S3	X4	0	1	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S3	X5	0	1	0	0	0	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0
G1	S3	X6	0	1	0	0	0	4	1	0	0	0	0	0	0	0	1	0	0	0	0	0

Figure 1. Each LIN position (A-T) represents an average nucleotide identity (ANI) threshold, ranging from 70% (A) to 99.999% (T). The more similar two genomes are, the further to the right their LINs match.

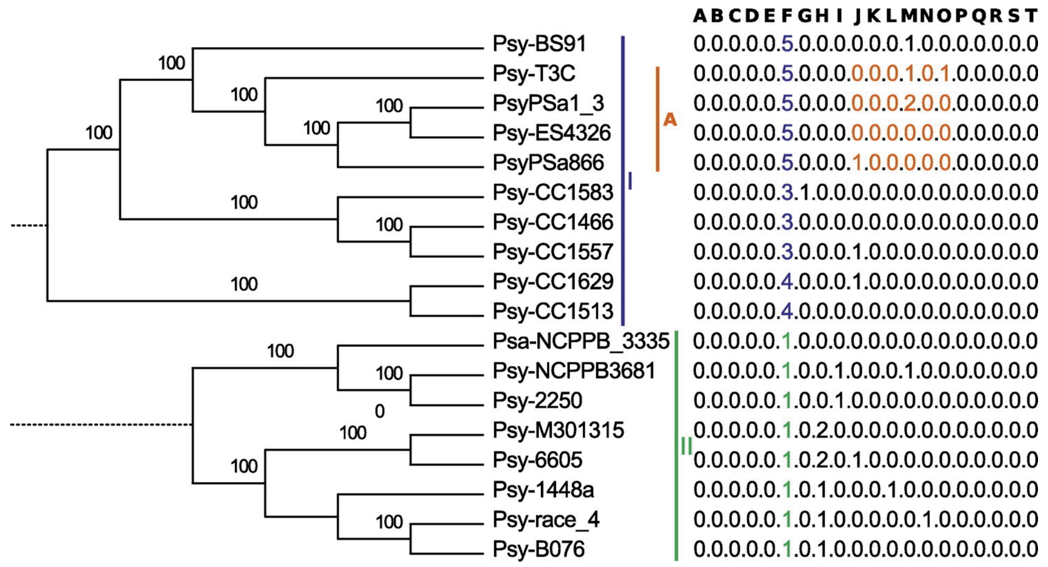
A B C D E F G H I J K L M N O P Q R S T
100
100
100
100
100
100
100
100
100
100
0
100
100
100
Psy-BS91 0.0.0.0.0.5.0.0.0.0.0.1.0.0.0.0.0.0.0.0
Psy-T3C 0.0.0.0.0.5.0.0.0.0.0.0.1.0.1.0.0.0.0.0
PsyPSa1_3 0.0.0.0.0.5.0.0.0.0.0.0.2.0.0.0.0.0.0.0
Psy-ES4326 0.0.0.0.0.5.0.0.0.0.0.0.0.0.0.0.0.0.0.0
PsyPSa866 0.0.0.0.0.5.0.0.0.1.0.0.0.0.0.0.0.0.0.0
Psy-CC1583 0.0.0.0.0.3.1.0.0.0.0.0.0.0.0.0.0.0.0.0
Psy-CC1466 0.0.0.0.0.3.0.0.0.0.0.0.0.0.0.0.0.0.0.0
Psy-CC1557 0.0.0.0.0.3.0.0.0.1.0.0.0.0.0.0.0.0.0.0
Psy-CC1629 0.0.0.0.0.4.0.0.0.1.0.0.0.0.0.0.0.0.0.0
Psy-CC1513 0.0.0.0.0.4.0.0.0.0.0.0.0.0.0.0.0.0.0.0
Psa-NCPPB_3335 0.0.0.0.0.1.0.0.0.0.0.0.0.0.0.0.0.0.0.0
Psy-NCPPB3681 0.0.0.0.0.1.0.0.1.0.0.0.1.0.0.0.0.0.0.0
Psy-2250 0.0.0.0.0.1.0.0.1.0.0.0.0.0.0.0.0.0.0.0
Psy-M301315 0.0.0.0.0.1.0.2.0.0.0.0.0.0.0.0.0.0.0.0
Psy-6605 0.0.0.0.0.1.0.2.0.1.0.0.0.0.0.0.0.0.0.0
Psy-1448a 0.0.0.0.0.1.0.1.0.0.0.1.0.0.0.0.0.0.0.0
Psy-race_4 0.0.0.0.0.1.0.1.0.0.0.0.0.1.0.0.0.0.0.0
Psy-B076 0.0.0.0.0.1.0.1.0.0.0.0.0.0.0.0.0.0.0.0
A
I
II

Figure 2. Two clades of *Pseudomonas syringae sensu lato*, showing assignment of LINs (from Vinatzer et al. (2017))

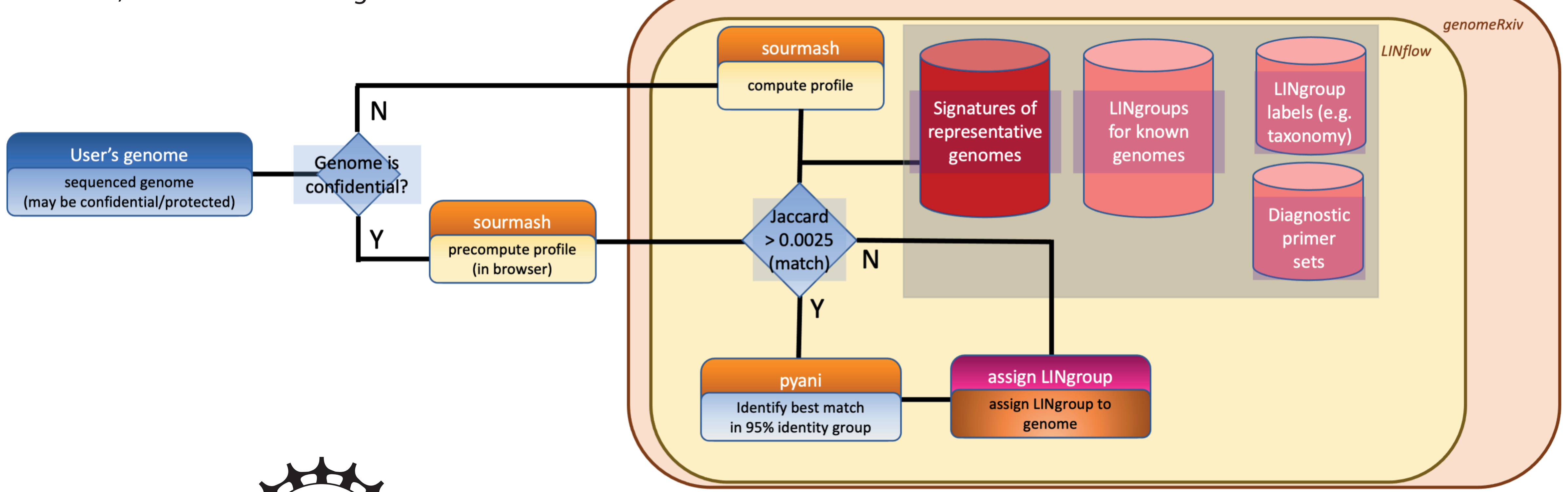

Figure 3. Flowchart of LIN assignment (**LINflow**). The user submits a sequenced genome, which is translated into a **sourmash** profile (in the browser if the genome is confidential). The profile is compared against a set of representative genome profiles. If a match is found, the best-matching genome is selected for ANI (**pyani**) comparison and a new LIN assigned; if not, a new LIN is assigned directly. Adapted from Tian *et al.* (2021)

3. More Information

The genomeRxiv project is at an early stage. We invite you to follow its development and learn more about the underlying technologies at the links below:

Vinatzer *et al.* (2017) *Phytopathology*
https://doi.org/10.1094/phyto-07-16-0252-r
Proposal for LINs

Tian *et al.* (2021) *PeerJ*
https://doi.org/10.1094/phyto-07-16-0252-r
LINflow computational pipeline

https://code.vt.edu/linbaseproject
LINbase repository

https://sourmash.readthedocs.io/en/latest/
sourmash documentation; MinHash-based classification

https://github.com/widdowquinn/pyani
pyani repository; ANI-based classification

https://github.com/widdowquinn/find_differential_primers
pdp repository; diagnostic primer prediction

References

[1] Pritchard *et al.* (2016) *Analytical Methods* doi:10.1039/c5ay02550h

[2] Rodriguez-R *et al.* (2018) *Nuc. Acids Res.* doi:10.1093/nar/gky467

[3] Varghese *et al.* (2015) *Nuc. Acids Res.* doi:10.1093/nar/gkv657