Genome-based Taxonomy to Improve the Regulation of Biological Control Organisms

Long Tian[1], Reza Mazloom[2], Lenwood S. Heath[2] and Boris A. Vinatzer[1]

(1) School of Plant and Environmental Sciences, Virginia Tech

(2) Department of Computer Science, Virginia Tech

Acknowledgements

- Haitham El Marakeby
- Alex Weisberg
- Caroline L. Monteil
- Kellye Eversole
- Gwyn Beattie
- Caitilyn Allen
- Mohammad Arif
- Titus C. Brown
- Leighton Pritchard

NSF IOS-1754721

COLLEGE OF AGRICULTURE AND LIFE SCIENCES VIRGINIA TECH

USDA APHIS

COI

Life Identification Number® and LIN® are registered trademarks of This Genomic Life, Inc. Lenwood S. Heath and Boris A. Vinatzer report in accordance with Virginia Tech policies and procedures and their ethical obligation as researchers, that they have a financial interest in the company This Genomic Life, Inc., that may be affected by this presentation. They have disclosed those interests fully to Virginia Tech, and they have in place an approved plan for managing any potential conflicts arising from this relationship.

Outline

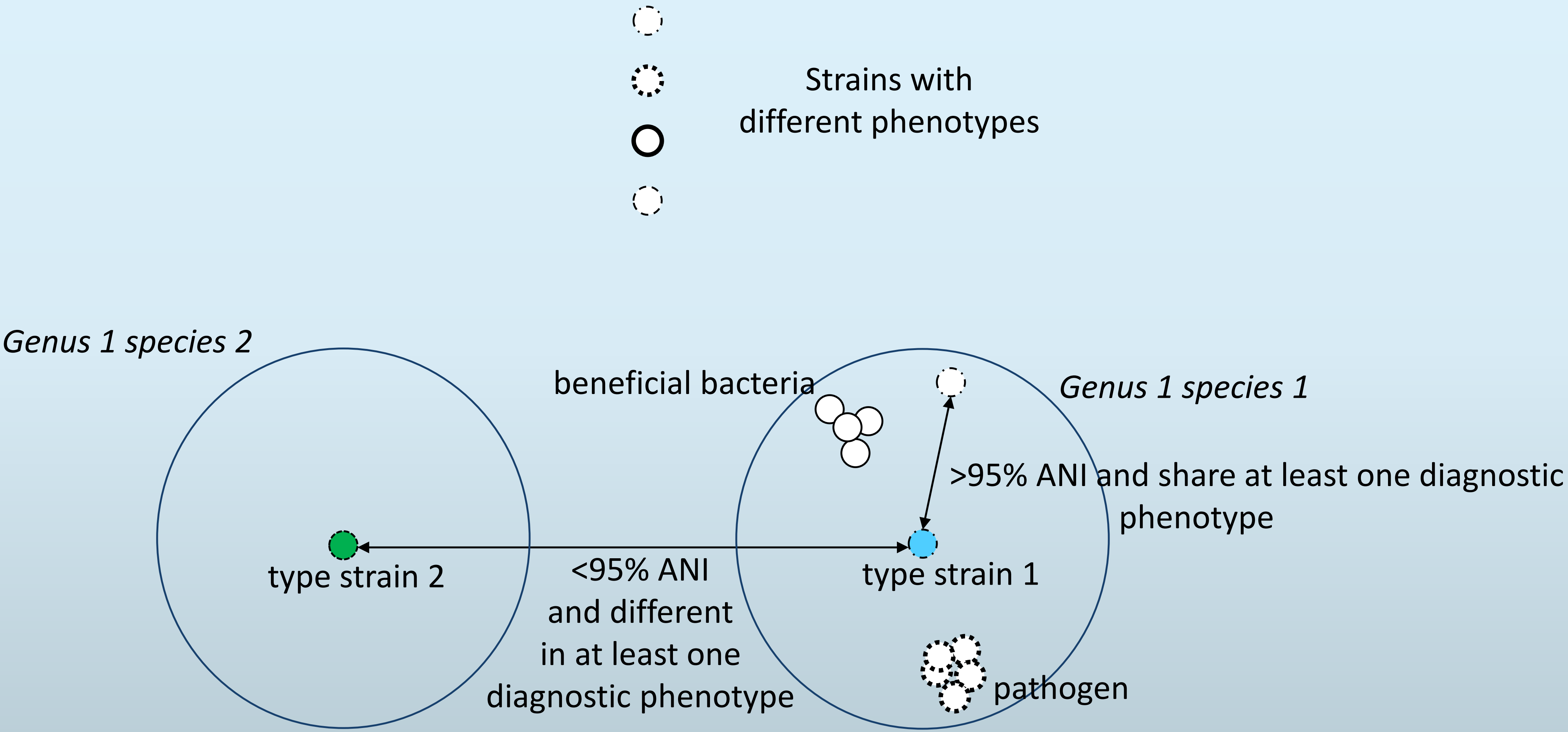
- What is taxonomy really all about?
- The importance of taxonomy for the regulation of plant pathogens and biological control organisms
- The promise of genome-based taxonomy
- The LINbase web server and how it can provide precise classification and identification (as long as phenotypic data/metadata exist)
- Genome-based risk assessment:
  - based on genome phylogeny
  - based on the presence of known/predicted pathogenicity genes

Taxonomy

- The basic “what” definition of taxonomy: the science of naming, describing, and classifying organisms.
- The “why” definition of taxonomy: the science of assigning individual organisms to named groups in a way that being identified as a member of one named group predicts an organism’s characteristics that distinguish the organism from all organisms that are not members of that group.
- A good example: assigning all bacteria that cause the disease anthrax to the named group *Bacillus anthracis* allows us to predict that an unknown bacterium assigned to that group also causes anthrax but an organism assigned to a different group does not cause anthrax.
- A bad example: assigning bacteria that cause different diseases or no disease at all to one group of organisms that share some characteristics that nobody cares about and call that group *Escherichia coli*.

The “*Escherichia coli* problem” is a result of today’s polyphasic taxonomy and the operational species concept

Today’s polyphasic taxonomy – the operational species concept



Strains with
different phenotypes

Genus 1 species 2

beneficial bacteria

Genus 1 species 1

>95% ANI and share at least one diagnostic
phenotype

type strain 2

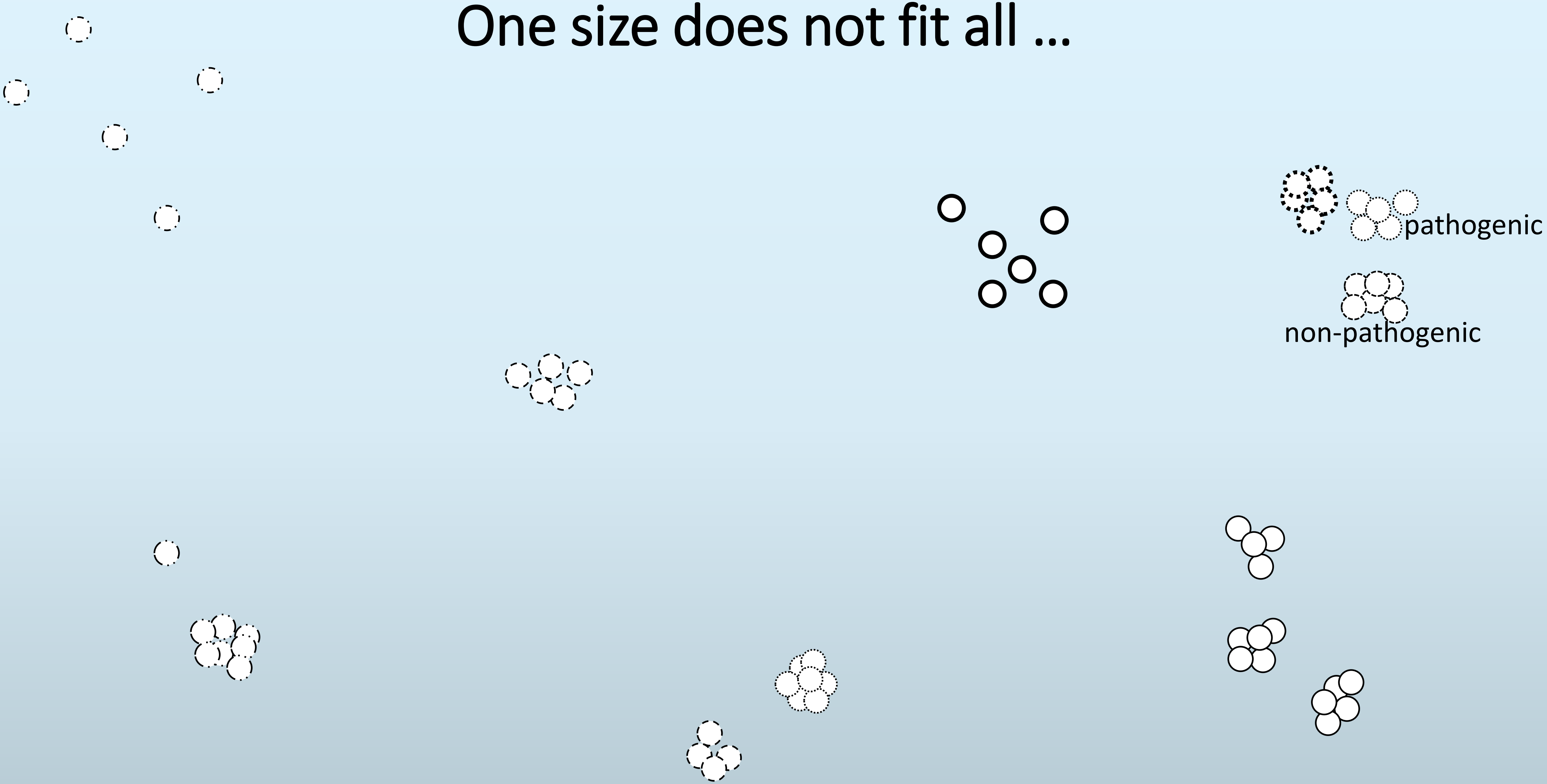
<95% ANI
and different
in at least one
diagnostic phenotype

type strain 1

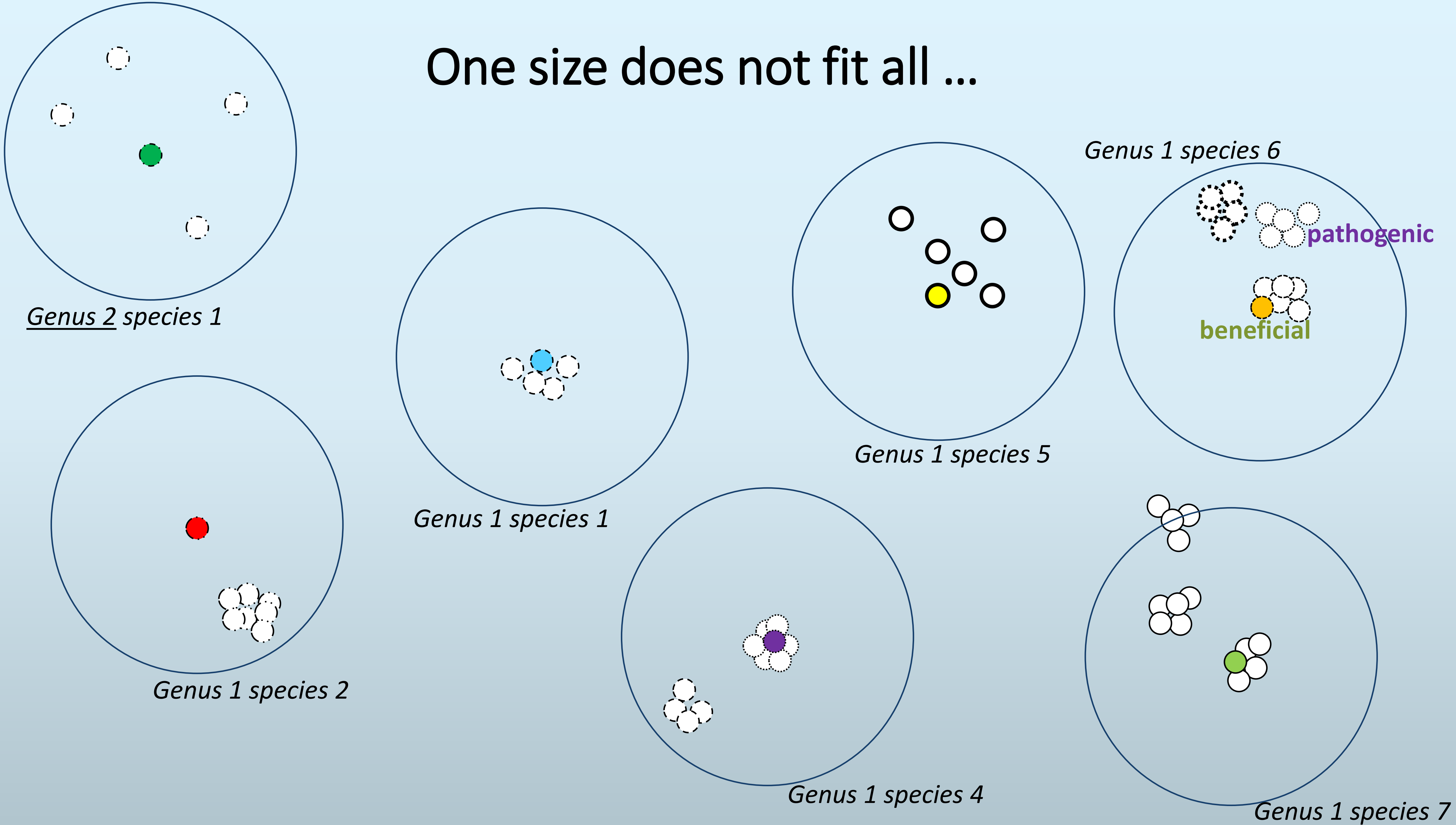
pathogen

ANI: Average Nucleotide Identity

One size does not fit all …

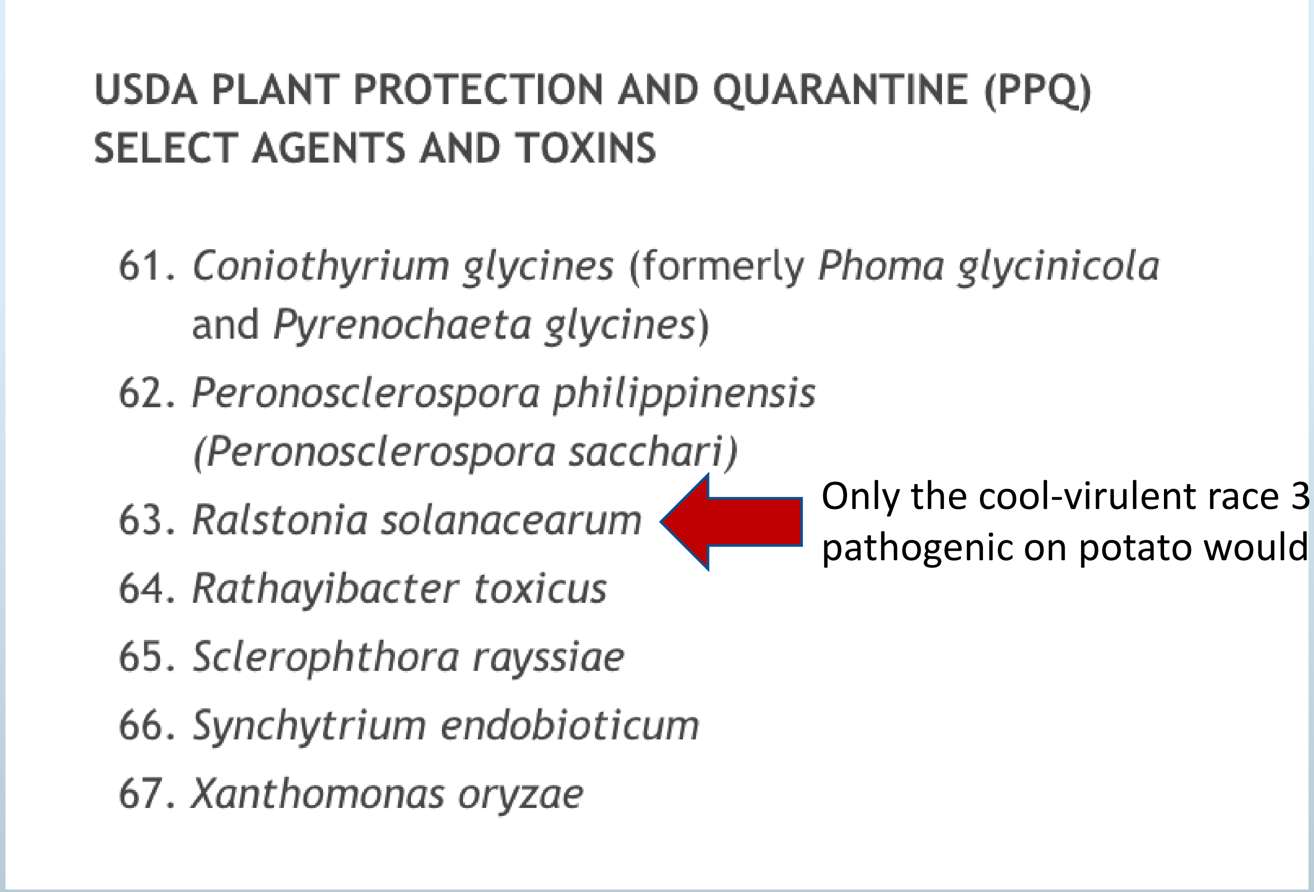


One size does not fit all …



Species and regulatory agencies

- Regulators rely on named species, for example, see the select agent list:

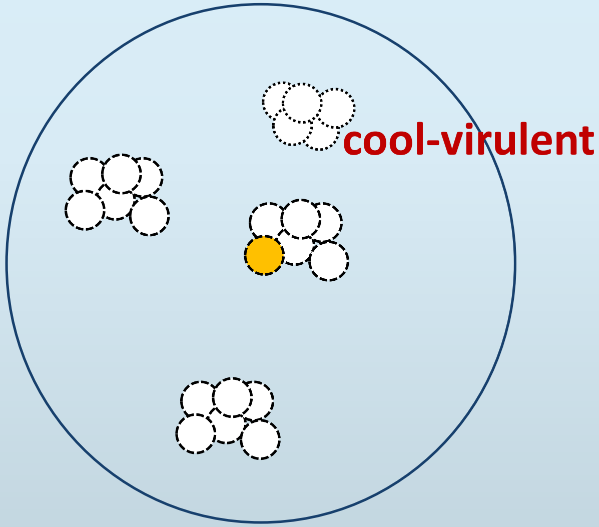


USDA PLANT PROTECTION AND QUARANTINE (PPQ)
SELECT AGENTS AND TOXINS

61. *Coniothyrium glycines* (formerly *Phoma glycinicola* and *Pyrenochaeta glycines*)
62. *Peronosclerospora philippinensis* (*Peronosclerospora sacchari*)
63. *Ralstonia solanacearum*
64. *Rathayibacter toxicus*
65. *Sclerophthora rayssiae*
66. *Synchytrium endobioticum*
67. *Xanthomonas oryzae*

Only the cool-virulent race 3 biovar 2 strain pathogenic on potato would need to be listed!

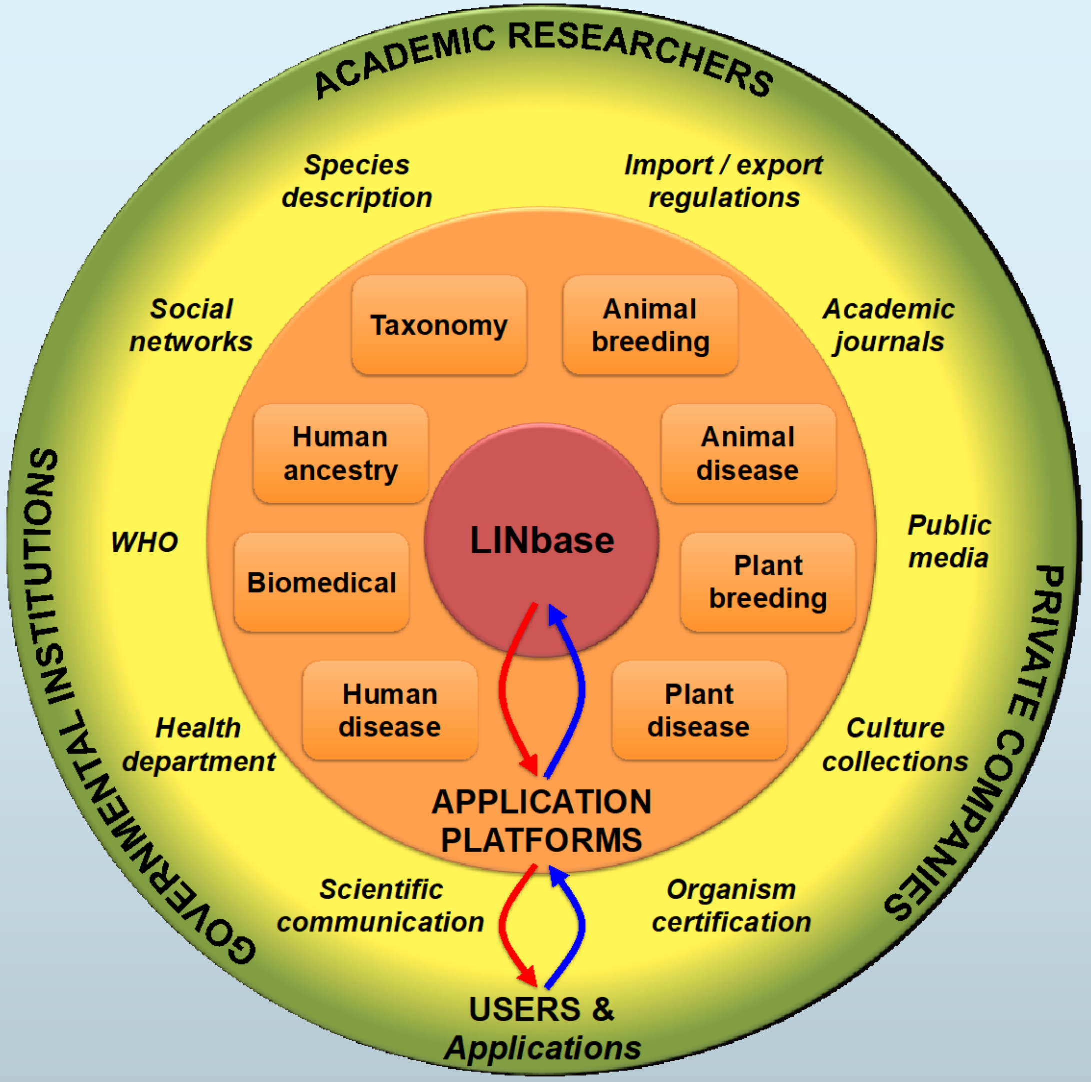
https://www.selectagents.gov/SelectAgentsandToxinsList.html



The challenge with plant-beneficial bacteria

- Plant-beneficial bacteria are sometimes closely related to plant pathogens and even to human pathogens, for example, bacteria in the genera *Burkholderia* and *Bacillus*.
- This is a problem, in particular, when trying to register and commercialize biological control agents.
- Today’s genera and species are not precise enough to develop regulations that reflect risk.
- We need named groups with distinct phenotypes that we care about because they affect human, animal, and plant health.
- Careful **phenotyping** is necessary to do this!

The Life Identification Number® (LIN®) concept



What are LINs?

- Stable and unique codes that are:
    - assigned to individual organisms (for example, bacterial isolates)
    - based on a measure of **genome** similarity, such as average nucleotide identity (ANI)
    - informative of the similarity of an organism’s genome to the genomes of all other organisms.
- Codes consist of a series of positions, each expressing a different threshold of **genome** similarity.
- The more similar the genomes of two organisms are, the more similar the LINs of the two organisms are.
- **Importantly: instead of a single species threshold of 95% ANI, LINs have many ANI thresholds to circumscribe groups of many different breadths!**

LIN DATABASE (LINbase)

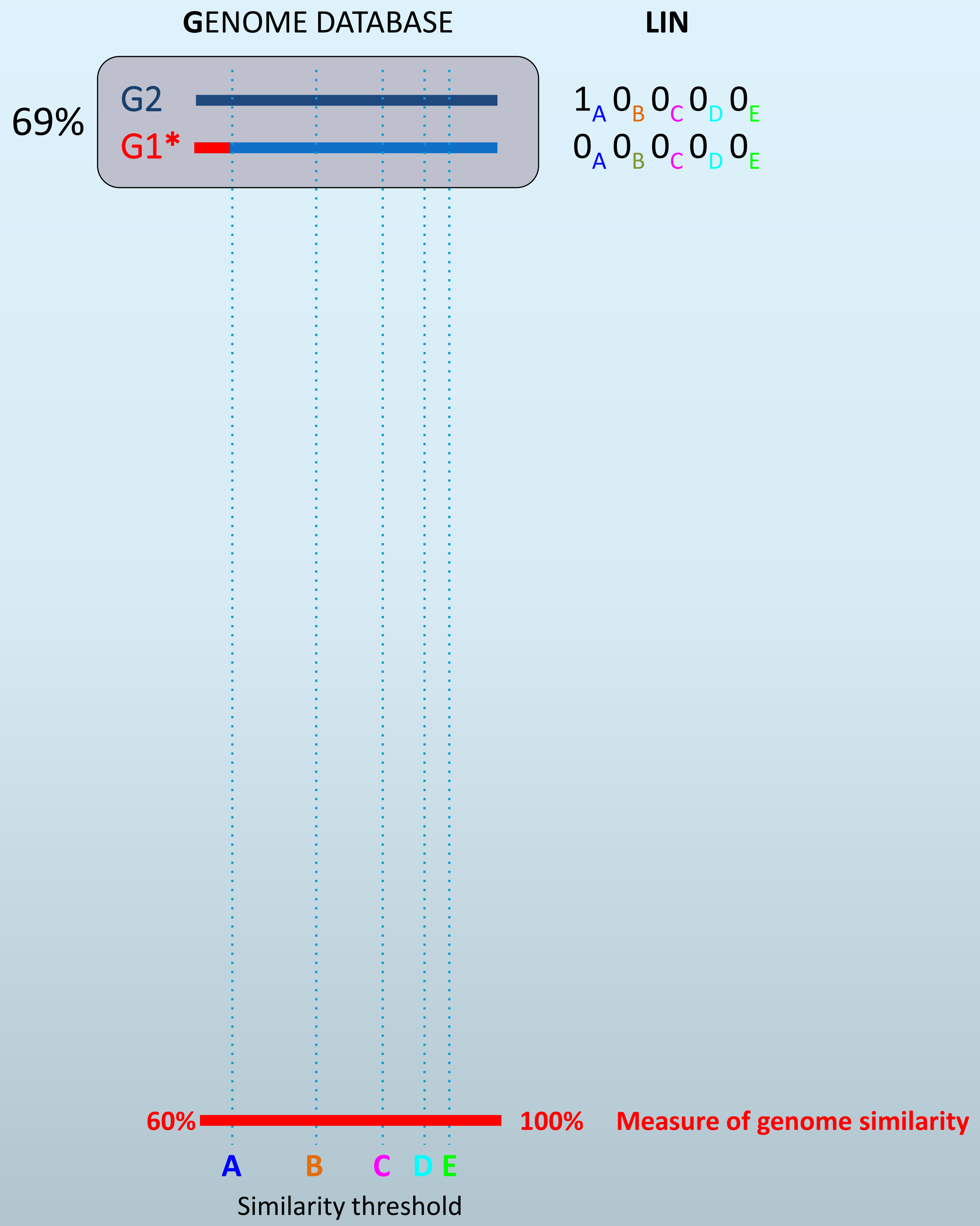
LIN

$0_A\ 0_B\ 0_C\ 0_D\ 0_E$

A	70%
B	80%
C	90%
D	95%
E	96%

Similarity thresholds

- LINs are assigned sequentially starting with any one **genome** in one database.
- “0” is used as a symbol, not as a value.



GENOME DATABASE

LIN

69%

G2 $1_A 0_B 0_C 0_D 0_E$

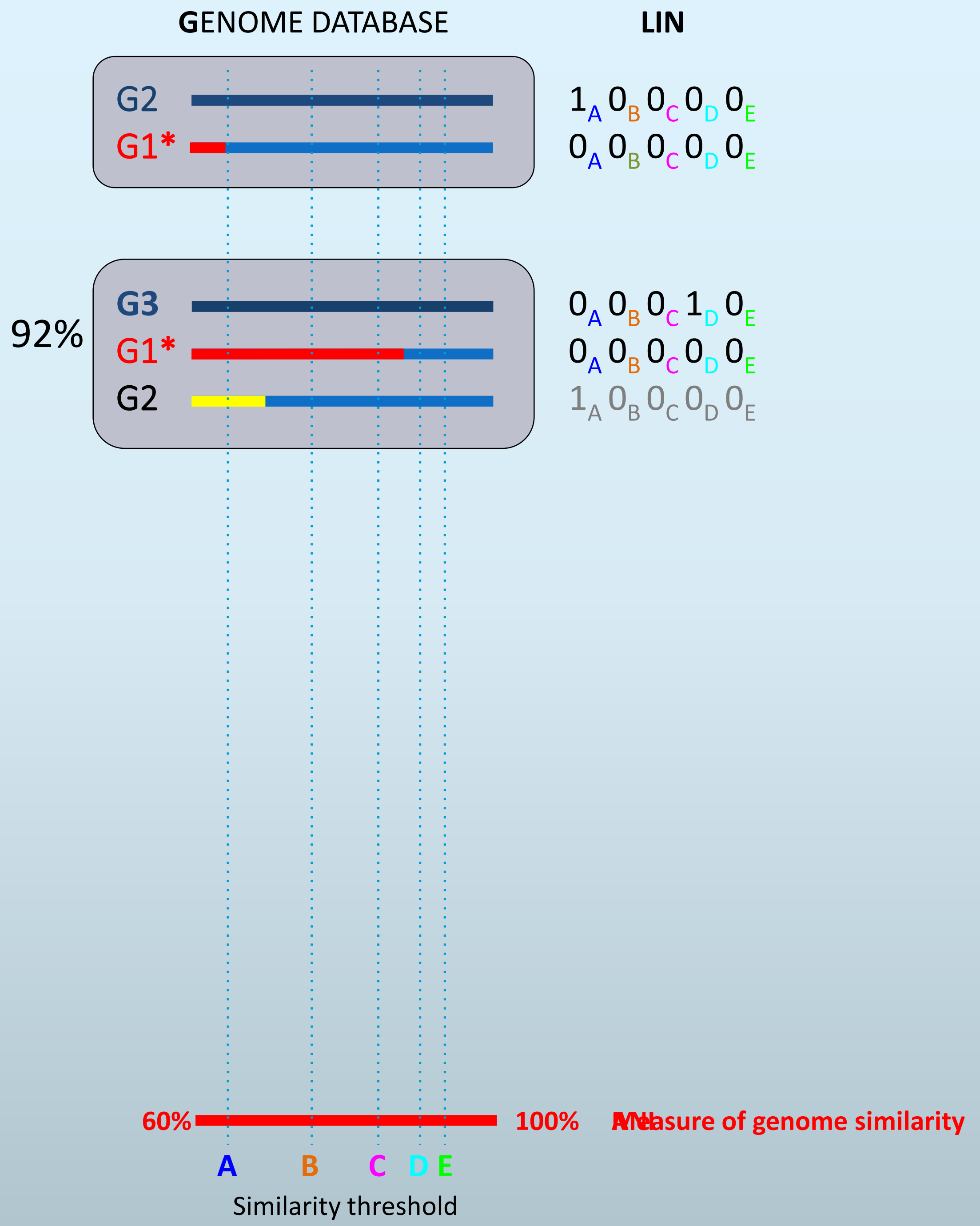
G1* $0_A 0_B 0_C 0_D 0_E$

60% 100% Measure of genome similarity

A B C D E

Similarity threshold

A 70%
B 80%
C 90%
D 95%
E 96%



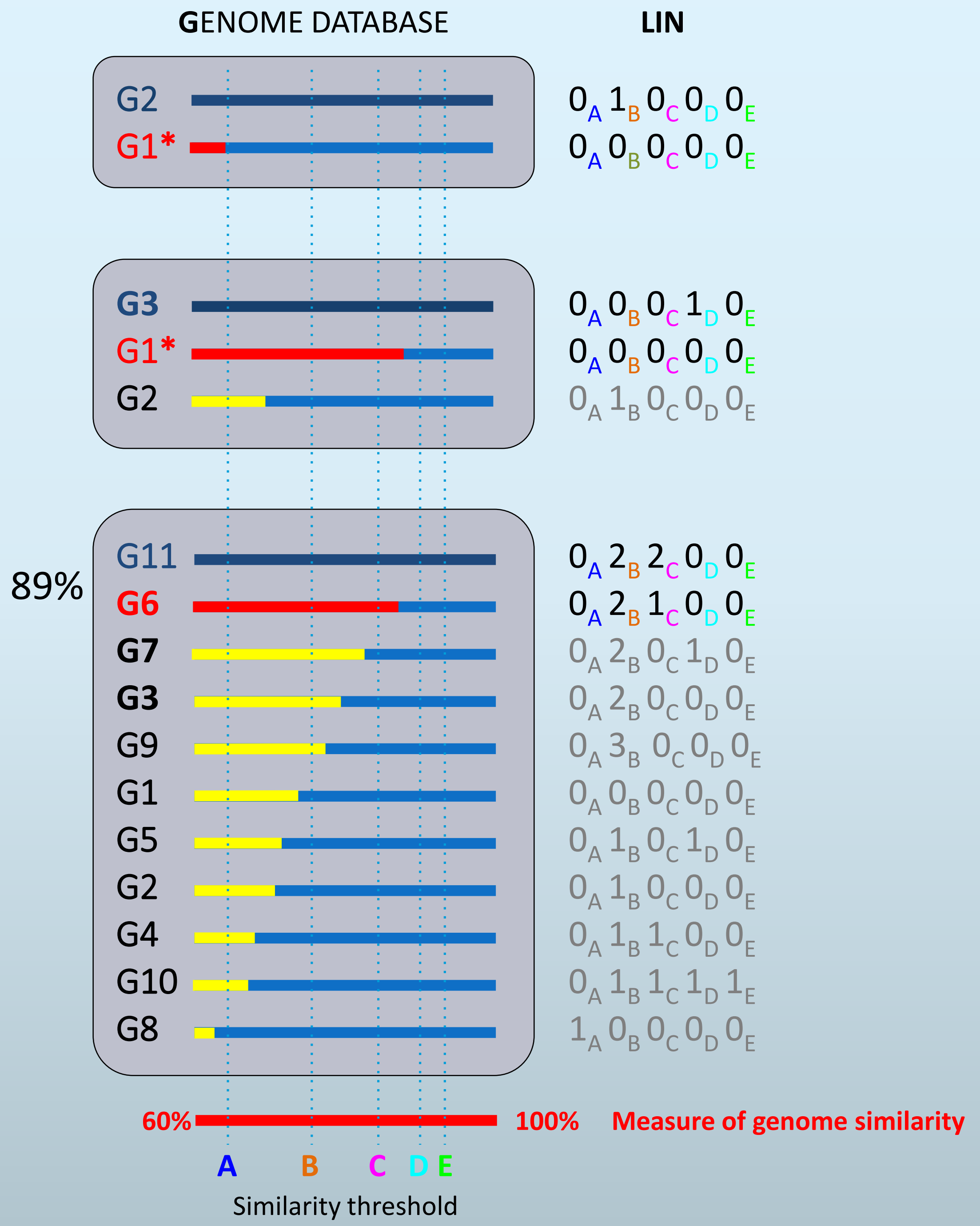
A 70%

B 80%

C 90%

D 95%

E 96%



GENOME DATABASE	LIN
G2	$0_A 1_B 0_C 0_D 0_E$
G1*	$0_A 0_B 0_C 0_D 0_E$

GENOME DATABASE	LIN
G3	$0_A 0_B 0_C 1_D 0_E$
G1*	$0_A 0_B 0_C 0_D 0_E$
G2	$0_A 1_B 0_C 0_D 0_E$

GENOME DATABASE	LIN
G11	$0_A 2_B 2_C 0_D 0_E$
G6	$0_A 2_B 1_C 0_D 0_E$
G7	$0_A 2_B 0_C 1_D 0_E$
G3	$0_A 2_B 0_C 0_D 0_E$
G9	$0_A 3_B 0_C 0_D 0_E$
G1	$0_A 0_B 0_C 0_D 0_E$
G5	$0_A 1_B 0_C 1_D 0_E$
G2	$0_A 1_B 0_C 0_D 0_E$
G4	$0_A 1_B 1_C 0_D 0_E$
G10	$0_A 1_B 1_C 1_D 1_E$
G8	$1_A 0_B 0_C 0_D 0_E$

89%

60% 100% Measure of genome similarity

A B C D E

Similarity threshold

A 70%
B 80%
C 90%
D 95%
E 96%

ANI thresholds used in current LIN implementation

within-species thresholds!

	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
	A	B	C	D	E	F	G	H	I	J	K	L	**M**	N	O	P	Q	R	**S**	T
genome 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
genome 2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
genome 3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
genome 4	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
genome 5	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

- LINs are informative of precisely how similar genomes are to each other.
- LINs are indices that automatically organize individual genomes in a database based on reciprocal similarity (expanding hierarchical taxonomy from the species all the way to the individual).

but how can LINs be used
to describe groups of organisms that need to be regulated?

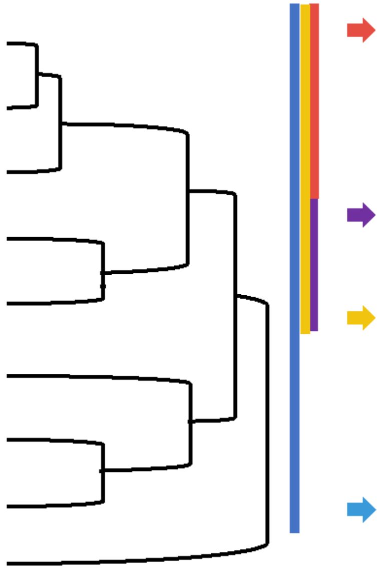
and how can LINs be used
to precisely identify unknown organisms as members of groups (that have phenotypes we care about)?

LINgroups:

any group of related organisms
(that share the same LIN over a number of positions)

LINgroup concept

		70	80	90	95	96	97	98	99	99.9
Genomes	**T***	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
G1	No	0	0	0	0	0	0	0	0	0
G2	No	0	0	0	0	0	0	1	0	0
G3	No	0	0	0	0	0	0	2	0	0
G4	No	0	0	0	0	0	1	0	0	0
G5	No	0	0	0	0	0	1	0	0	0
G6	No	0	0	0	1	0	0	0	0	0
G7	**Yes**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**
G8	No	0	0	0	2	0	0	0	0	0
G9	No	0	0	1	0	0	0	0	0	0



LINgroup: $0_A0_B0_C0_D0_E0_F$
Comment: Chile outbreak 2010 - 2014

LINgroup: $0_A0_B0_C0_D0_E1_F$
Comment: Italy outbreak 2008 - 2010

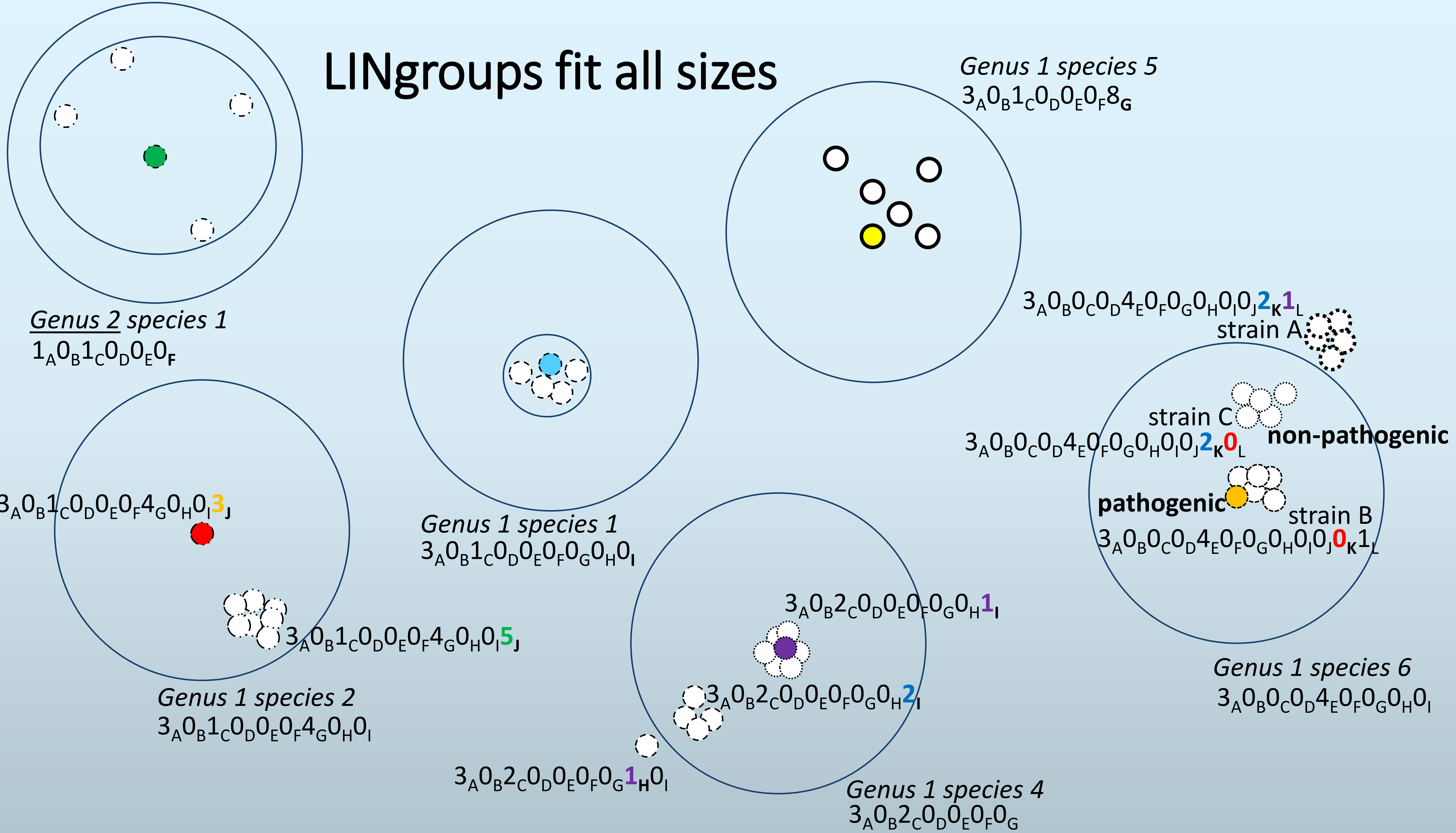
LINgroup: $0_A0_B0_C0_D0_E$
Nickname: pathovar *actinidiae*

LINgroup: $0_A0_B0_C$
Species: *Pseudomonas syringae*

* Type strain

Vinatzer et al 2017

LINgroups fit all sizes



Genus 2 species 1

$1_A0_B1_C0_D0_E0_F$

Genus 1 species 5

$3_A0_B1_C0_D0_E0_F8_G$

Genus 1 species 1

$3_A0_B1_C0_D0_E0_F0_G0_H0_I$

$3_A0_B0_C0_D4_E0_F0_G0_H0_I0_J2_K1_L$

strain A

strain C

$3_A0_B0_C0_D4_E0_F0_G0_H0_I0_J2_K0_L$

non-pathogenic

pathogenic

strain B

$3_A0_B0_C0_D4_E0_F0_G0_H0_I0_J0_K1_L$

Genus 1 species 6

$3_A0_B0_C0_D4_E0_F0_G0_H0_I$

$3_A0_B1_C0_D0_E0_F4_G0_H0_I3_J$

$3_A0_B1_C0_D0_E0_F4_G0_H0_I5_J$

Genus 1 species 2

$3_A0_B1_C0_D0_E0_F4_G0_H0_I$

$3_A0_B2_C0_D0_E0_F0_G0_H1_I$

$3_A0_B2_C0_D0_E0_F0_G0_H2_I$

$3_A0_B2_C0_D0_E0_F0_G1_H0_I$

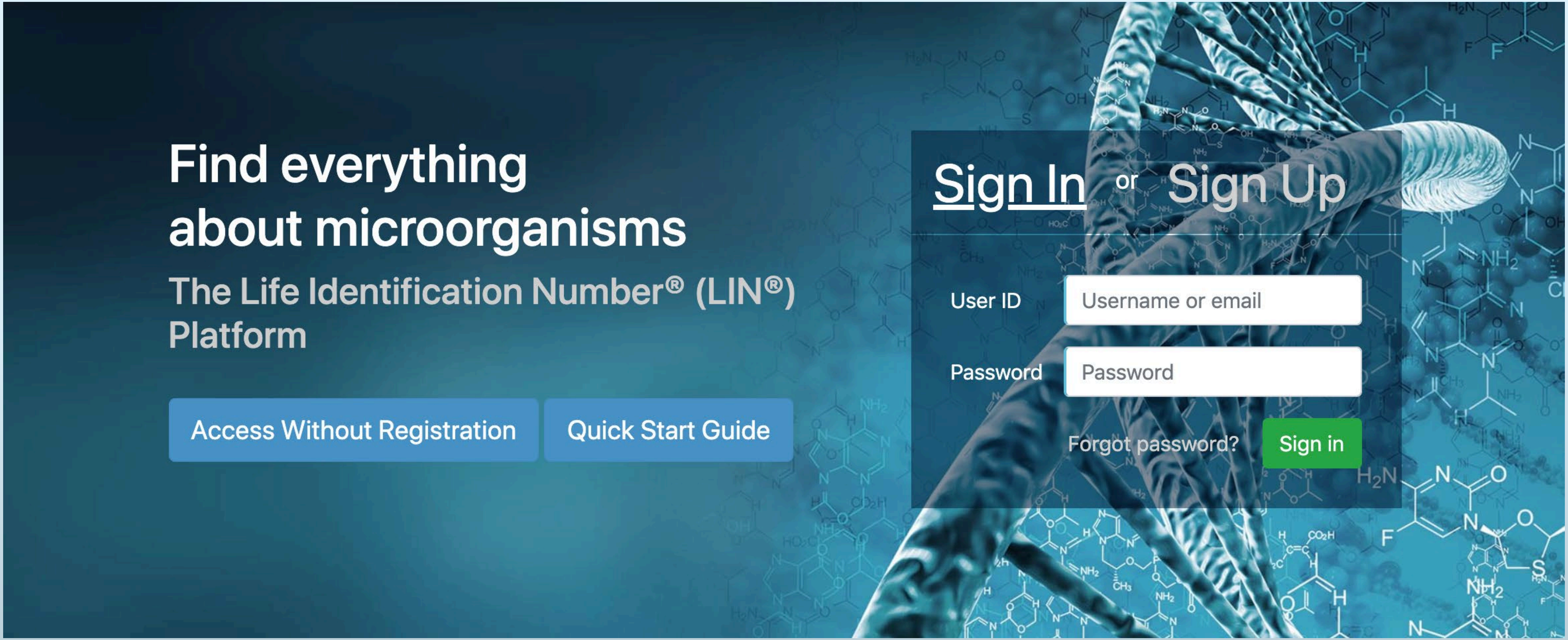
Genus 1 species 4

$3_A0_B2_C0_D0_E0_F0_G$

LINs and LINgroups have been implemented in:

LINbase, a crowdsourcing web server

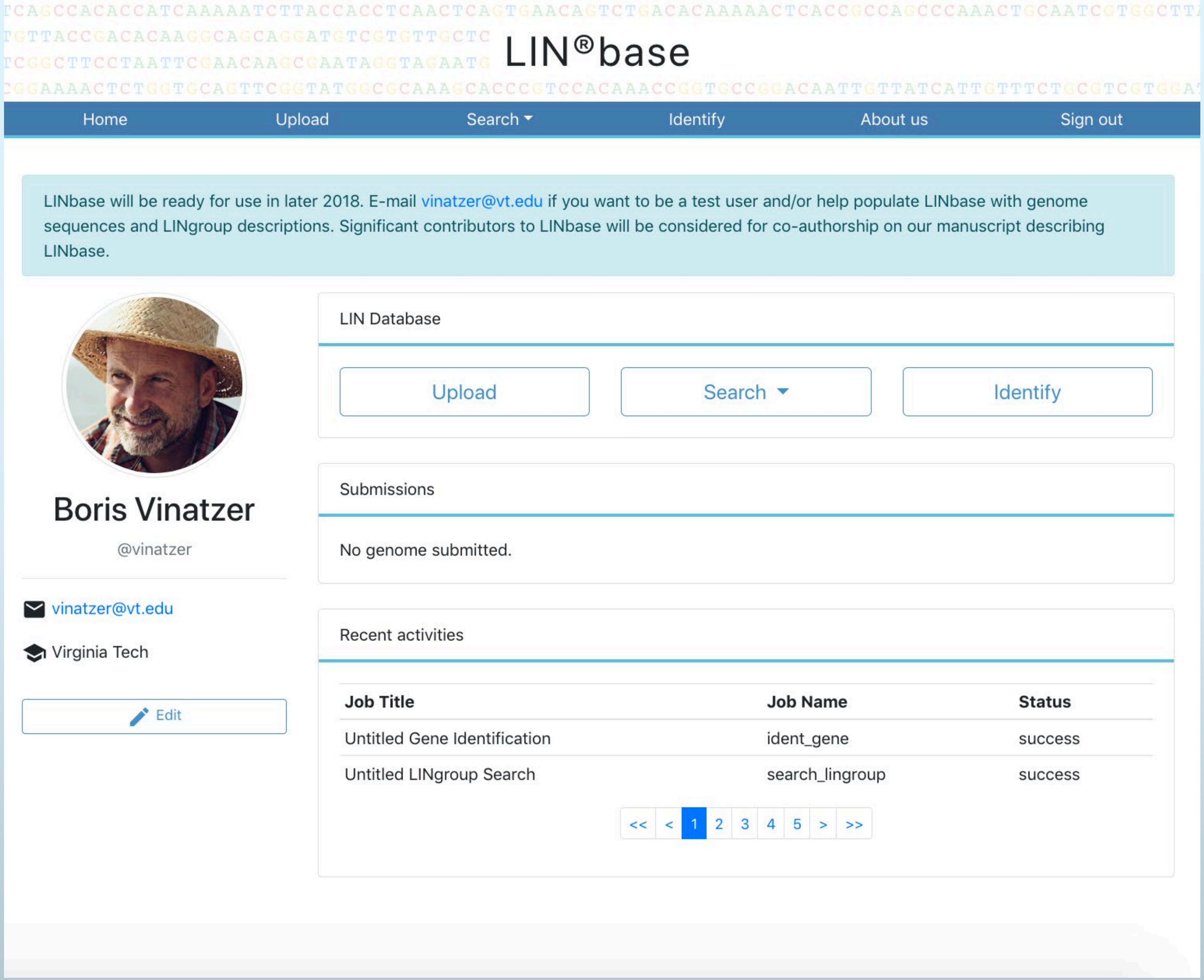
linbase.org



Find everything about microorganisms

The Life Identification Number® (LIN®) Platform

https://www.biorxiv.org/content/10.1101/752212v2



LINbase will be ready for use in later 2018. E-mail vinatzer@vt.edu if you want to be a test user and/or help populate LINbase with genome sequences and LINgroup descriptions. Significant contributors to LINbase will be considered for co-authorship on our manuscript describing LINbase.

Boris Vinatzer

@vinatzer

vinatzer@vt.edu

Virginia Tech

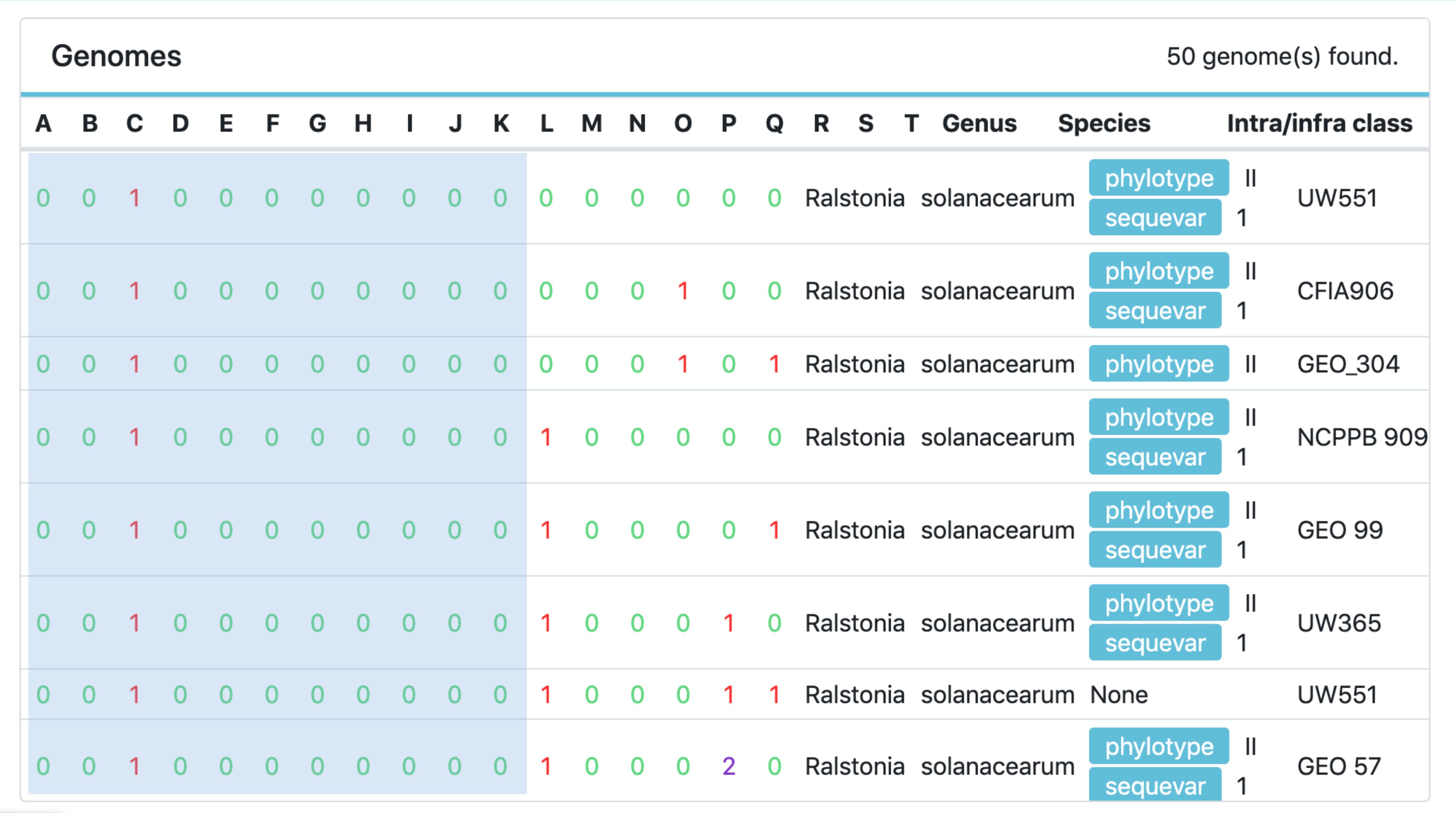
Submissions

No genome submitted.

Recent activities

Job Title	Job Name	Status
Untitled Gene Identification	ident_gene	success
Untitled LINgroup Search	search_lingroup	success

Describing a LINgroup

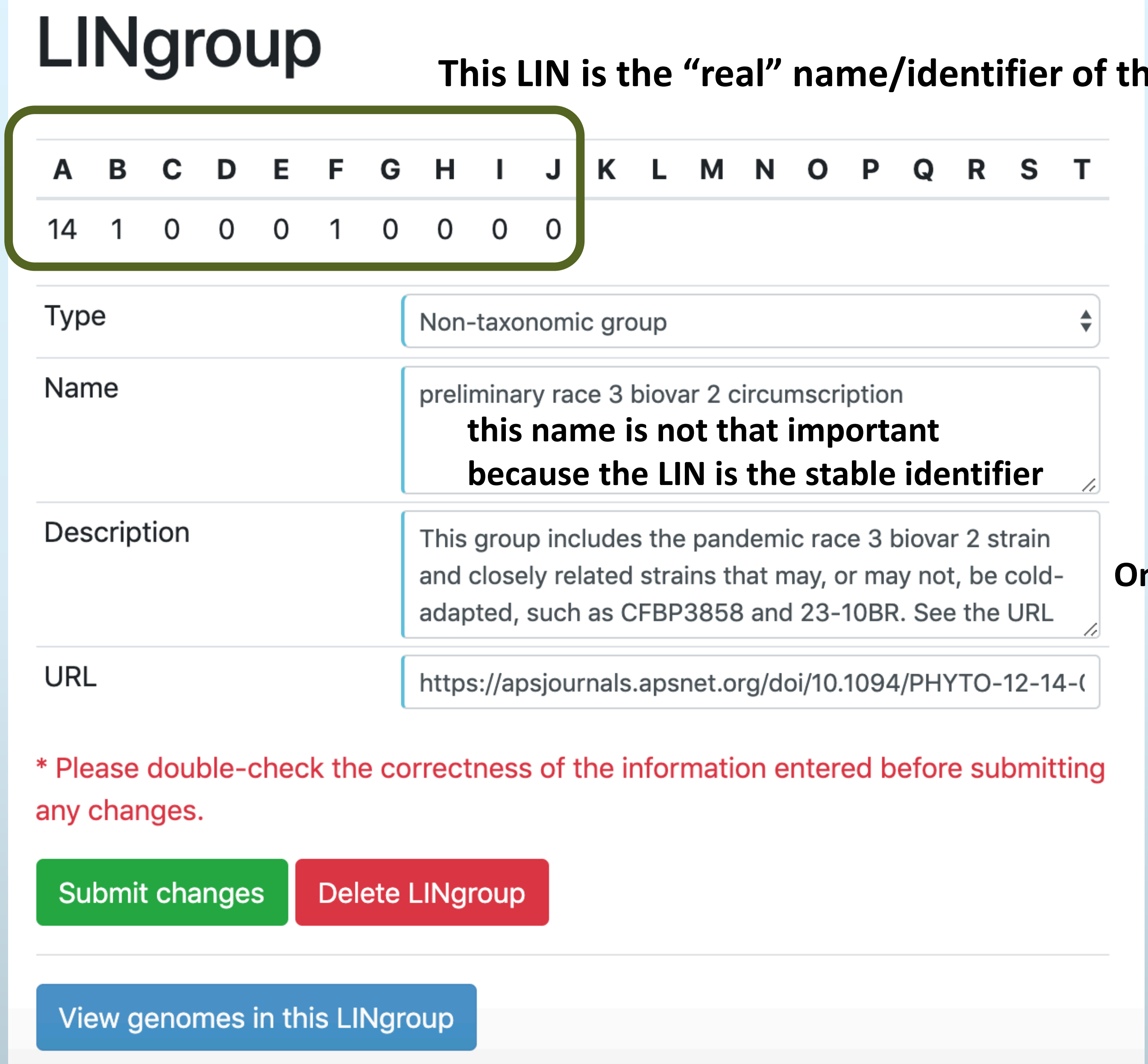


Genomes

50 genome(s) found.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Genus	Species	Intra/infra class
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Ralstonia			solanacearum	phylotype II sequevar 1	UW551
0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	Ralstonia			solanacearum	phylotype II sequevar 1	CFIA906
0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	Ralstonia			solanacearum	phylotype II	GEO_304
0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	Ralstonia			solanacearum	phylotype II sequevar 1	NCPPB 909
0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	Ralstonia			solanacearum	phylotype II sequevar 1	GEO 99
0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	Ralstonia			solanacearum	phylotype II sequevar 1	UW365
0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	1	Ralstonia			solanacearum	None	UW551
0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	2	0	Ralstonia			solanacearum	phylotype II sequevar 1	GEO 57

A user can select the conserved LIN positions for strains that are cool-virulent …



LINgroup

This LIN is the “real” name/identifier of the group

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
14	1	0	0	0	1	0	0	0	0										

Type	Non-taxonomic group
Name	preliminary race 3 biovar 2 circumscription
Description	This group includes the pandemic race 3 biovar 2 strain and closely related strains that may, or may not, be cold-adapted, such as CFBP3858 and 23-10BR. See the URL
URL	https://apsjournals.apsnet.org/doi/10.1094/PHYTO-12-14-(

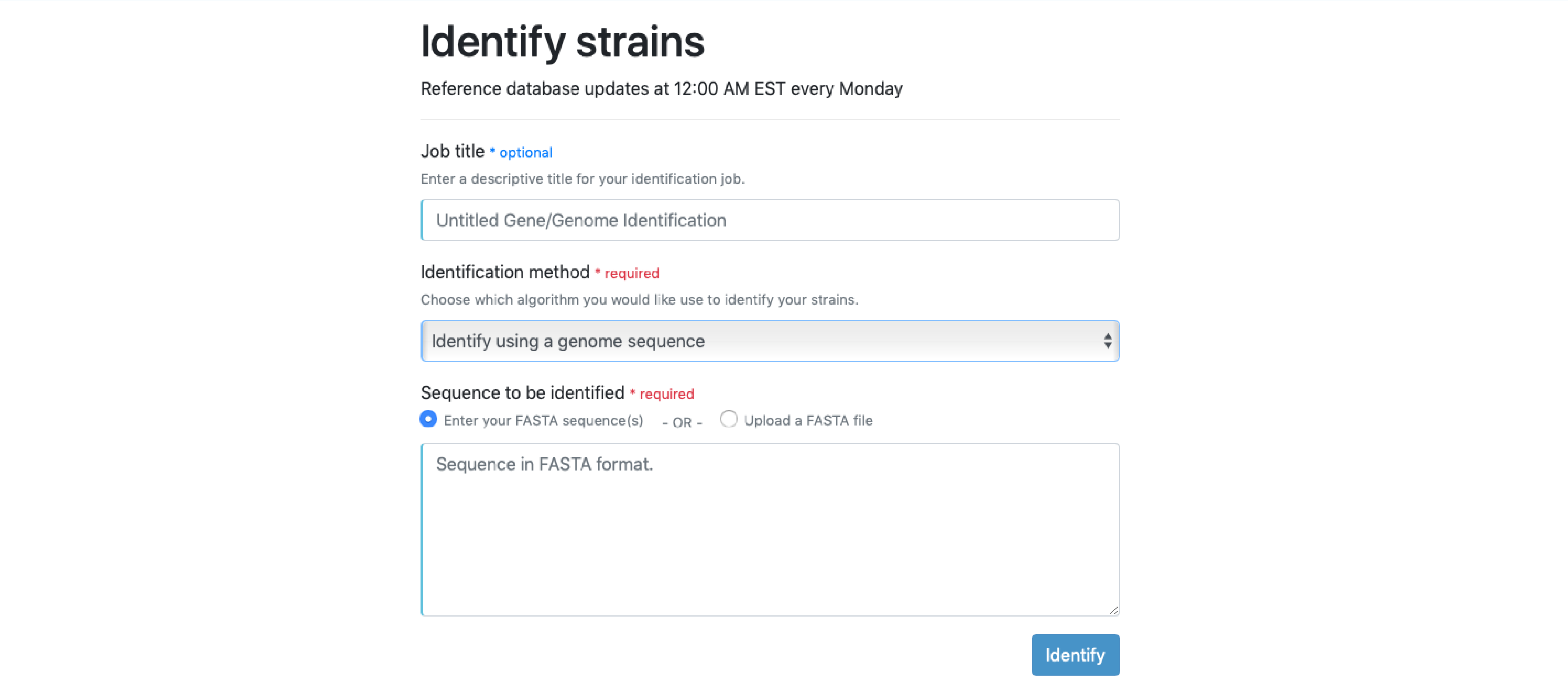
this name is not that important because the LIN is the stable identifier

Only as good as the phenotyping!

* Please double-check the correctness of the information entered before submitting any changes.

The user enters a name, a description,
and possibly the URL of a relevant publication …

Identifying an unknown using a genome sequence



Identify strains

Reference database updates at 12:00 AM EST every Monday

Job title * optional

Enter a descriptive title for your identification job.

Untitled Gene/Genome Identification

Identification method * required

Choose which algorithm you would like use to identify your strains.

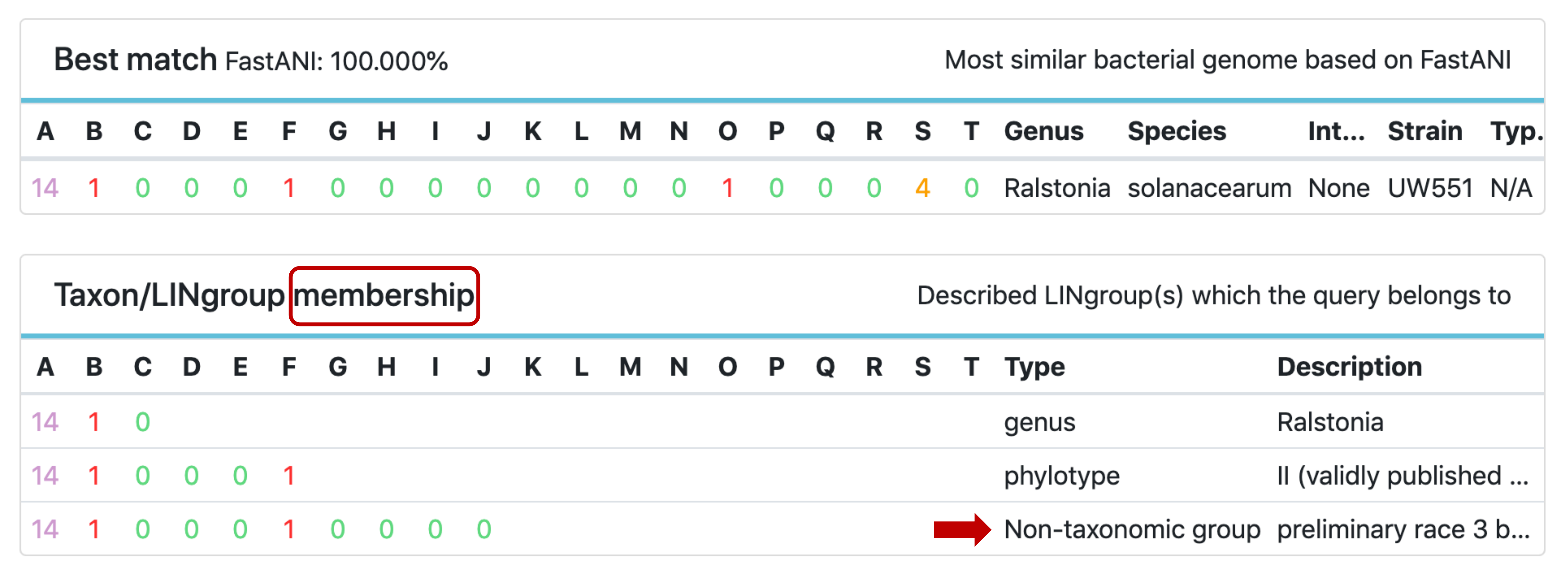
Identify using a genome sequence

Sequence to be identified * required

● Enter your FASTA sequence(s) - OR - ○ Upload a FASTA file

Sequence in FASTA format.

Next time a LINbase user queries LINbase with the genome of an unknown isolate



Best match FastANI: 100.000%

Most similar bacterial genome based on FastANI

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Genus	Species	Int...	Strain	Typ.
14	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	4	0	Ralstonia	solanacearum	None	UW551	N/A

Taxon/LINgroup membership

Described LINgroup(s) which the query belongs to

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Type	Description
14	1	0																		genus	Ralstonia
14	1	0	0	0	1															phylotype	II (validly published ...
14	1	0	0	0	1	0	0	0	0											Non-taxonomic group	preliminary race 3 b...

The user may get the search result that the unknown isolate is a member of the newly described LINgroup.
… NCBI does not give membership; MiGA and GTDB do, but only to the species rank.

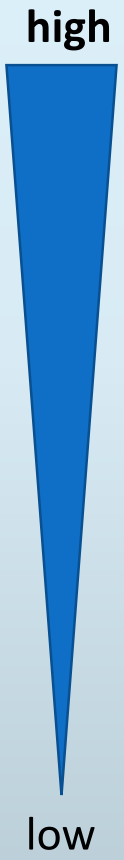
LINbase is in beta testing but the goal is to …

- provide immediate, unique identifiers to communicate about yet un-named and un-characterized organisms, such as emerging pathogens.
- make it very easy and very fast to identify any organism based on its **genome** sequence alone as a member of named groups that have distinctive phenotypes.
- offer a genome-based classification and identification platform to federal agencies to precisely regulate plant pathogens and biological control organisms independently of the 95% ANI threshold of named species (or to protect IP).
- requires metadata and phenotypic data!

Genome-based risk assessment

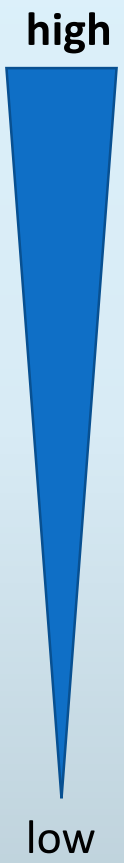
1. Risk based on genome similarity/relatedness to known pathogens of humans, animals, and plants
2. Risk based on the presence of virulence genes

Risk based on genome similarity (using LINs)



- The organism belongs to a LINgroup that is associated with disease (all characterized members of the group cause disease).
- The organism belongs to a LINgroup that is not known to be associated with disease but is closely related to a LINgroup that is associated with disease.
- The organism belongs to a LINgroup that has been shown not to cause disease but the LINgroup is closely related to a LINgroup which members cause disease.
- The organism belongs to a LINgroup which members have been shown not to cause disease and the group is not closely related to any LINgroup which members cause disease.

Risk based on virulence gene content



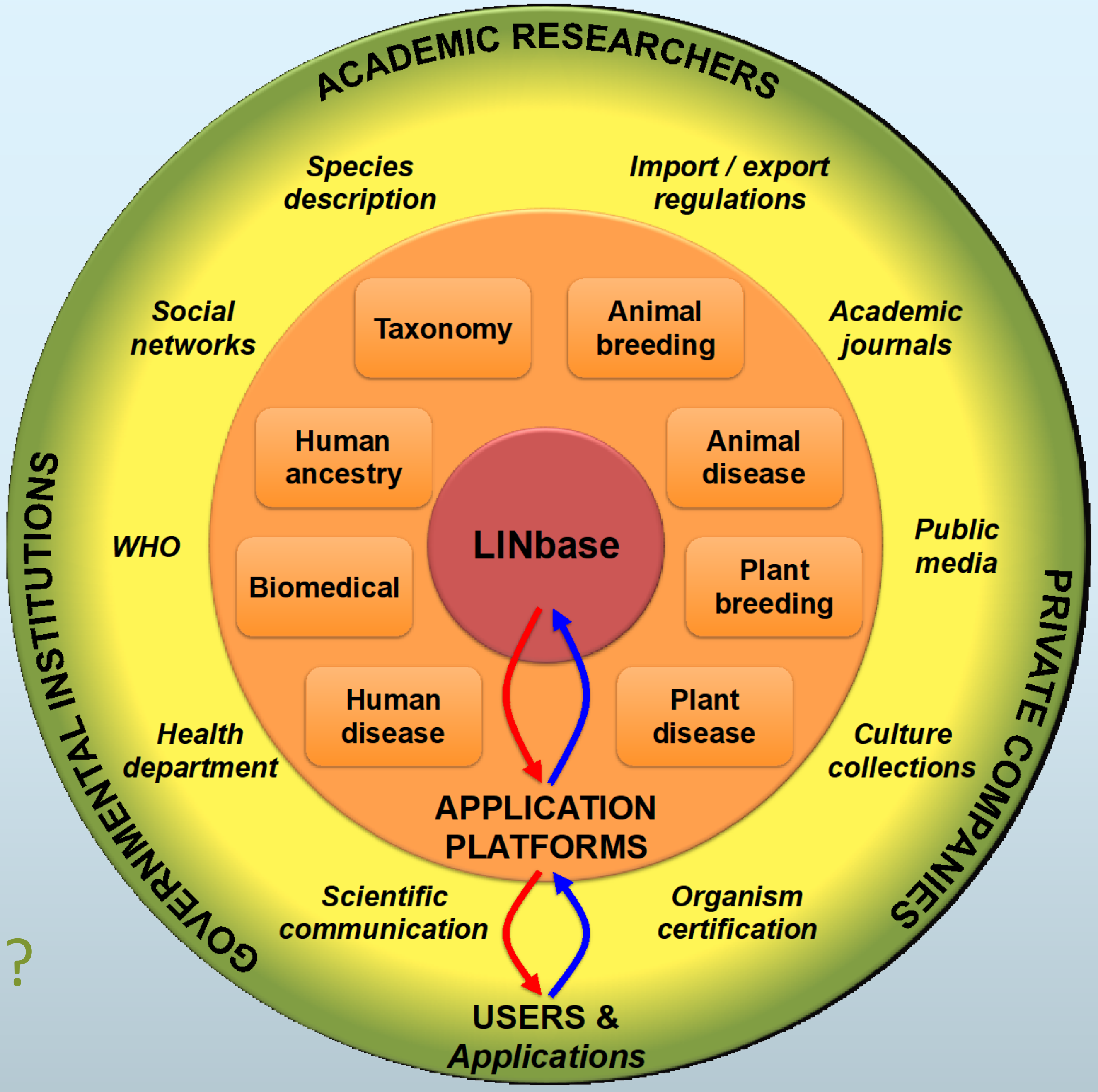
- The organism contains close homologs of pathogenicity or virulence genes typical of human/animal/plant pathogens
- The organisms contains distant homologs of virulence or pathogenicity genes
- The organism contains some genes sometimes associated with virulence or fragments of such genes
- The organism does not contain any genes known to be associated with virulence in any pathogen

In practice, a future version of LINbase could …

- Assign a LIN to a potential biological control organism
- Determine exactly to which LINgroups of non-pathogenic or pathogenic organisms it belongs to and/or how closely related it is to pathogenic or non-pathogenic LINgroups
- Compute a genome similarity-based risk score
- Determine the presence of potential virulence genes
- Compute a virulence gene-based risk score
- Compute a combined genome-based risk score

What we need to get there …

- Circumscribe most species in LINbase (can be automated)
- Enter in LINbase circumscriptions of clades, phylogroups, phylotypes, serogroups, serotypes, pathotypes based on peer-reviewed literature … (a lot of work by many experts …)
- Thorough characterization of species that include both pathogenic as well as non-pathogenic members to circumscribe them as separate groups
- Integrate virulence gene content analysis
- Develop a framework to compute a combined phylogeny and gene-based risk score



ACADEMIC RESEARCHERS

PRIVATE COMPANIES

GOVERNMENTAL INSTITUTIONS

Species description

Import / export regulations

Social networks

Academic journals

WHO

Public media

Health department

Culture collections

Scientific communication

Organism certification

Taxonomy

Animal breeding

Human ancestry

Animal disease

Biomedical

Plant breeding

Human disease

Plant disease

LINbase

APPLICATION PLATFORMS

USERS & *Applications*

Group Discussion

- What are the biggest regulatory hurdles you face today?
- Do you use genome sequencing to characterize biological control organisms?
- What are your needs in regard to genome analysis?
- How do you currently perform risk assessment?