How Genome-Based Classification can Improve Regulation of Pathogens and Beneficials

Boris A. Vinatzer

School of Plant and Environmental Sciences

VIRGINIA TECH™

Contributors, Collaborators, Funding & COI

Virginia Tech

- Lenwood S. Heath
- Reza Mazloom
- Parul Sharma

Collaborators (*R. solanacearum*)

- Caitilyn Allen (UW-Madison)
- Kellye Eversole (Phytobiomes)
- Gwyn Beattie (ISU)

Collaborators (genomeRxiv)

- C. Titus Brown (UC Davis)
- Leighton Pritchard (U Strathclyde)

DBI-2018522

USDA APHIS

AP19PPQS&T00C083

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

- The **Why** of microbial taxonomy
- The importance of taxonomy for the regulation of plant pathogens and biological control organisms
- The promise of genome-based taxonomy
- The LINbase/genomeRxiv web server to implement genome-based taxonomy

Taxonomy

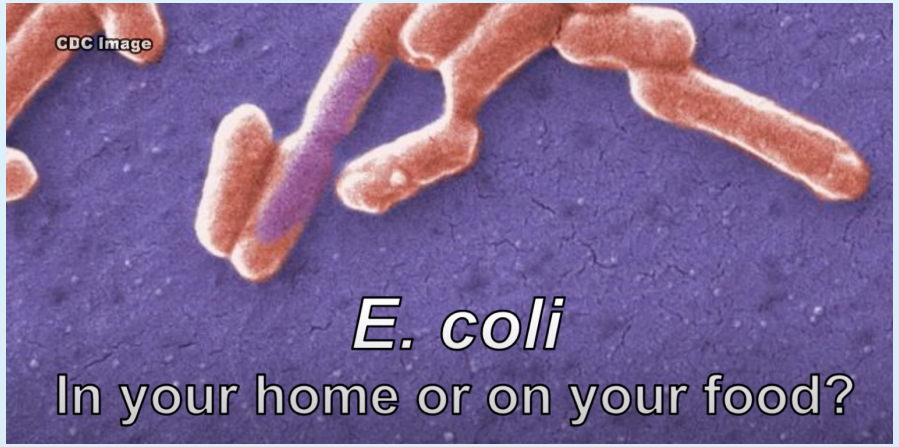
- The most common **what** definition of taxonomy: the science of (discovering), describing, classifying, and naming organisms (life forms).
- But **why** do we need taxonomy?
- … clear communication about organisms

www.britannica.com

Example 1

- discovering that some bacteria cause the highly fatal disease anthrax
- describing bacteria that either cause or do not cause anthrax
- group the bacteria that cause anthrax together in one group/class/taxon
- name that group *Bacillus anthracis*
- When I publish a scientific article about *Bacillus anthracis,* other scientists have no doubt about what organism I am talking about.

Example 2
sarcasm alert!



- discovering that some bacteria live in the intestin
- describing bacteria that either live in the intestin or not
- group the bacteria that live in the intestin as one group/taxon
- name that group *Escherichia coli.*
- When I publish a scientific article about *Escherichia coli,* other scientists have no doubt about what organism I am talking about.

What I really need to know:

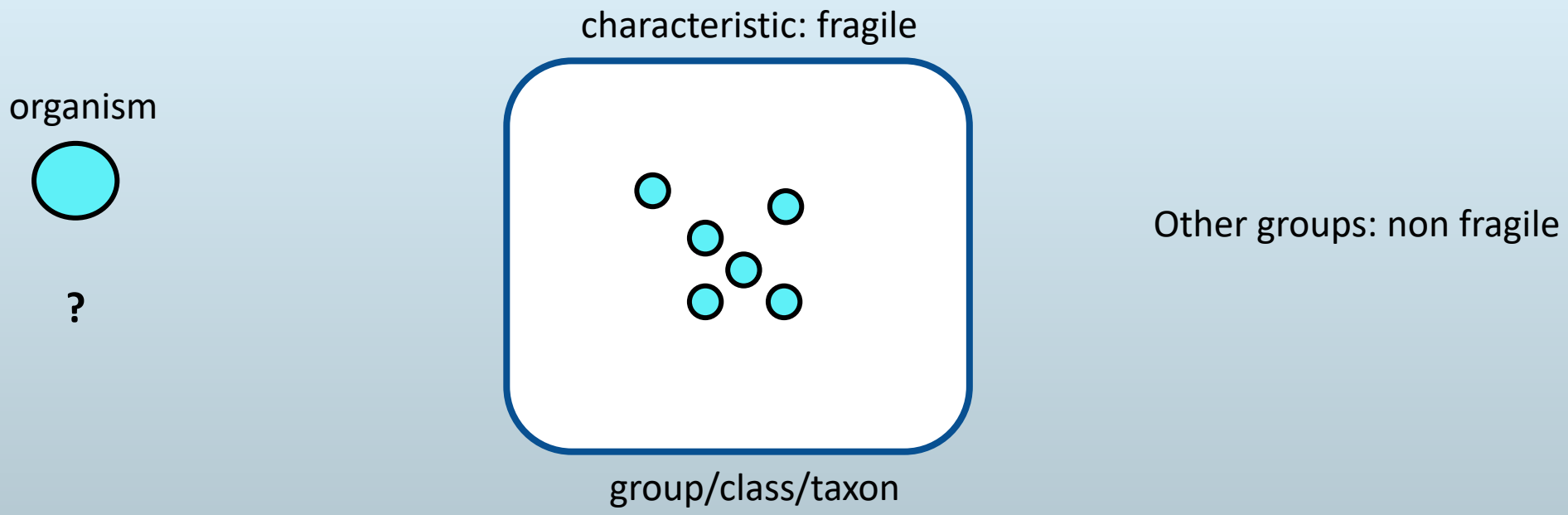
- Is it pathogenic?
- If yes, what disease does it cause?
- Urinary tract infection? Hemorrhagic diarrhea?
- What is the treatment?
- How can I contain an outbreak?

- Or, is it beneficial?
- What is the benefit?
- Is it safe?

Shouldn’t taxonomy define classes for which the answer to these questions is clear?

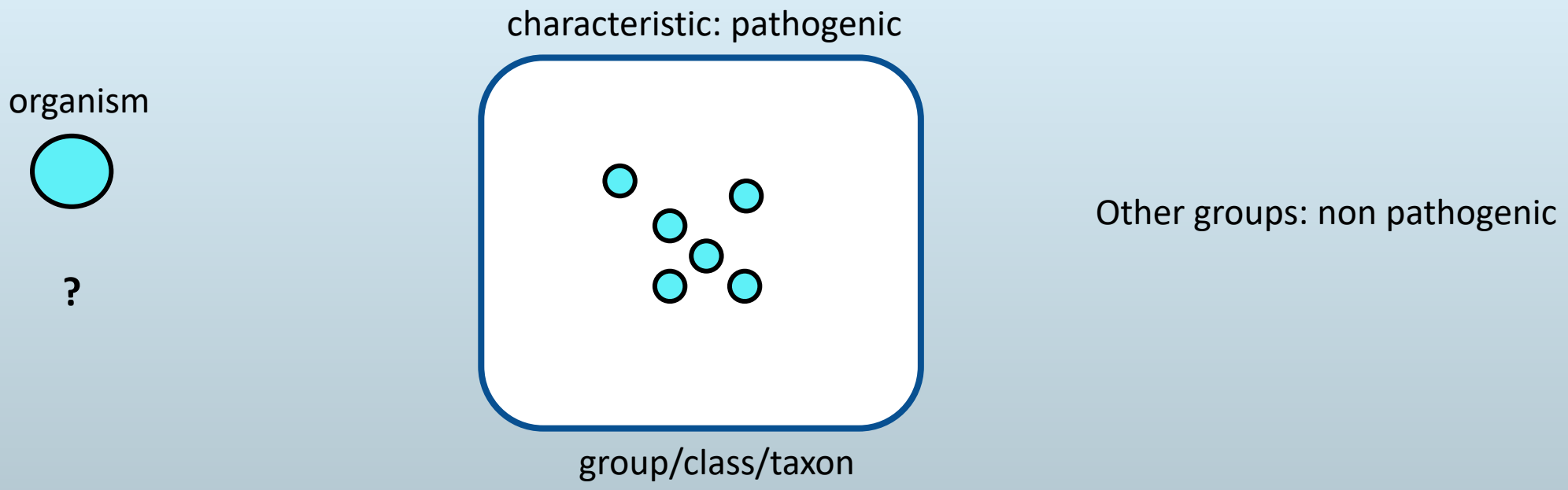
Why do we need taxonomy: **Identification**

- the science of assigning an organism to a group/class/taxon that is circumscribed in such a way that being <u>identified</u> as a member <u>predicts</u> the organism's <u>characteristics</u> that distinguish it from all organisms that are outside of that group.



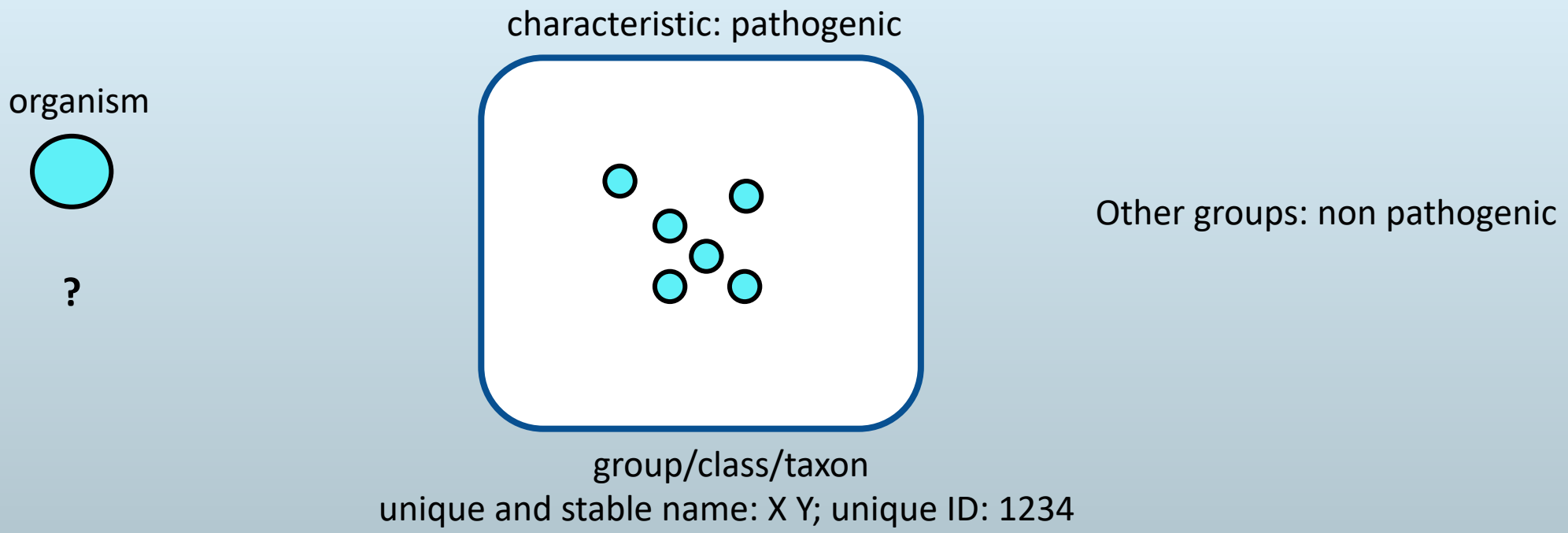
The **Why** definition of taxonomy: **Identification**

- For a group to be useful, it should have characteristics that are relevant to science and/or society so that research can be performed or fast and effective action can be taken after identification occurred.



The **Why** definition of taxonomy: **Identification**

- For effective science and action to occur, we need <u>coordination and collaboration</u>, for which we need <u>clear communication.</u> Therefore, we need such groups to have <u>unique names and/or identifiers</u>.



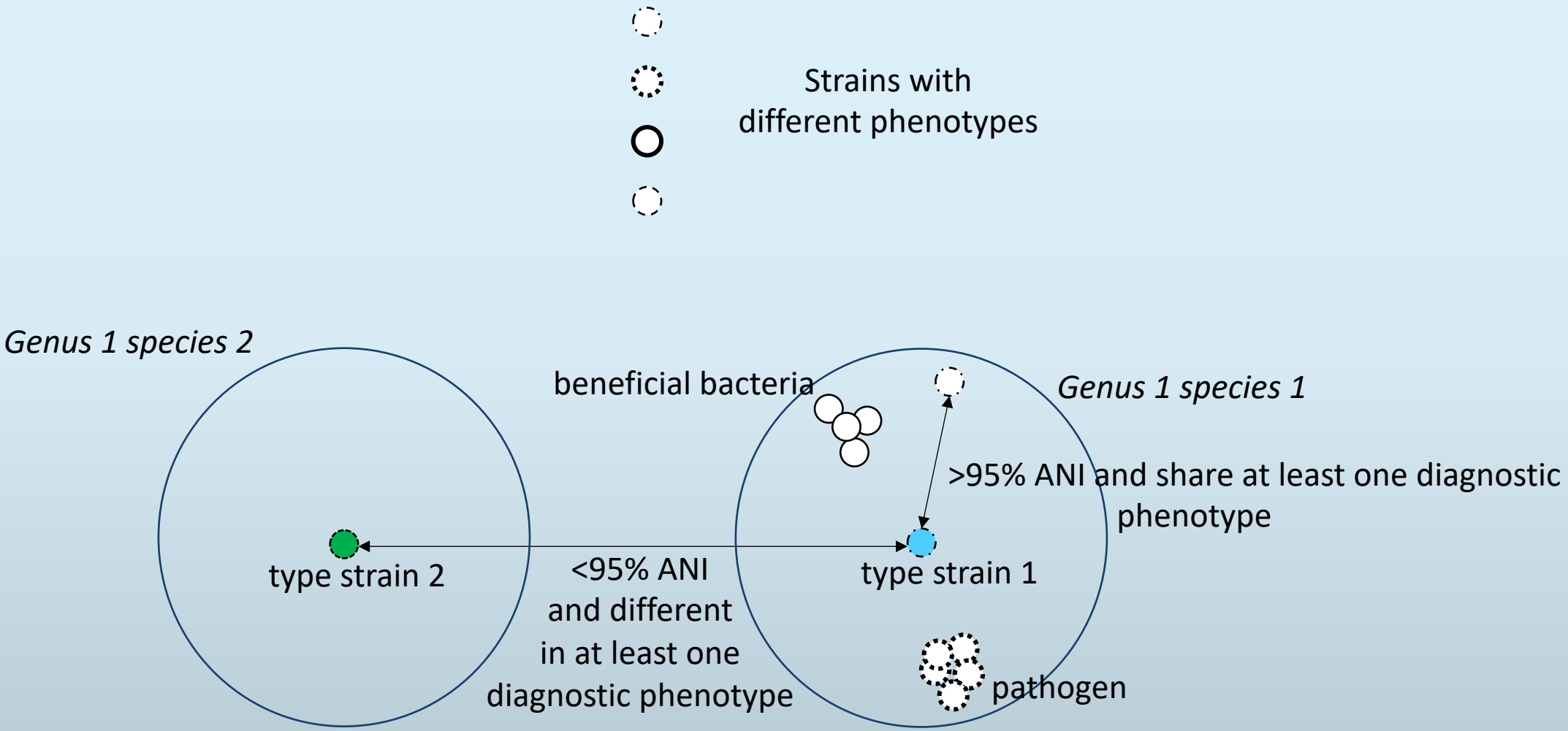
Taxonomy needs to be fast: Novel Coronavirus nCoV2019

- The first few weeks after the sequence of SARS-CoV-2 had been obtained, nobody knew what to call it.
- The first reports about this new virus referred to it as “Novel Coronavirus” or “nCoV2019”
- We had to wait for a taxonomic committee to decide what to call it
- Only after that committee met, we got a name for the pathogen and the disease
- Because it took weeks, none of the early communication included the terms “SARS-CoV-2” and “COVID-19”
- You cannot use these search terms to google and find early reports!
- We need faster taxonomy! We need immediate classification and names!

How to evaluate the quality of taxonomy?

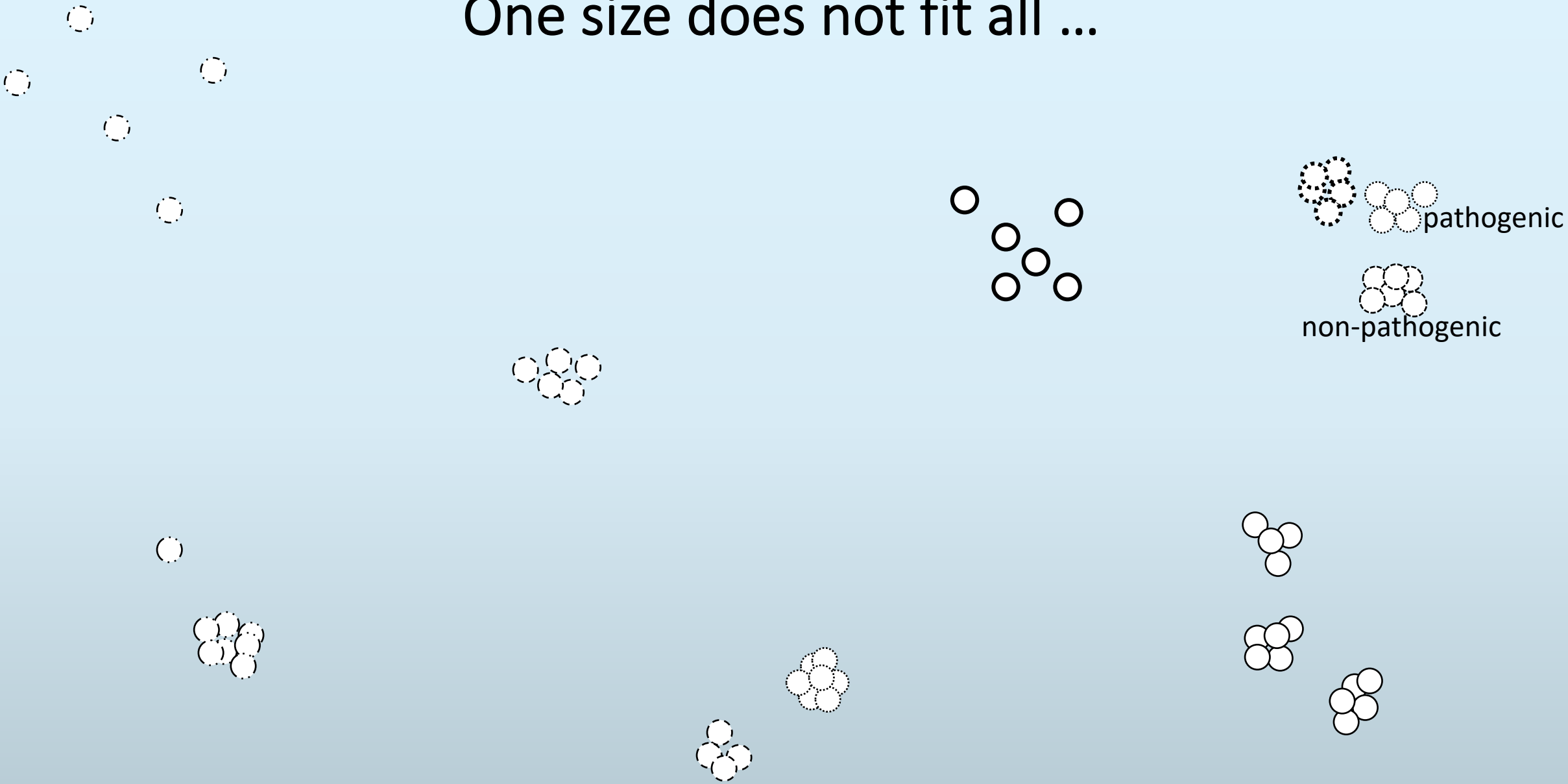
- Does it allow me to identify an organism in a way to predict relevant characteristics that I need to know to perform scientific research, reap benefits, protect from danger, establish effective regulations that maintain human and environmental health, … and so on.
- Current rank-based taxonomy using the species as basic unit definitely plays an important part in all that but it is not enough.

Today's taxonomy – the operational species concept

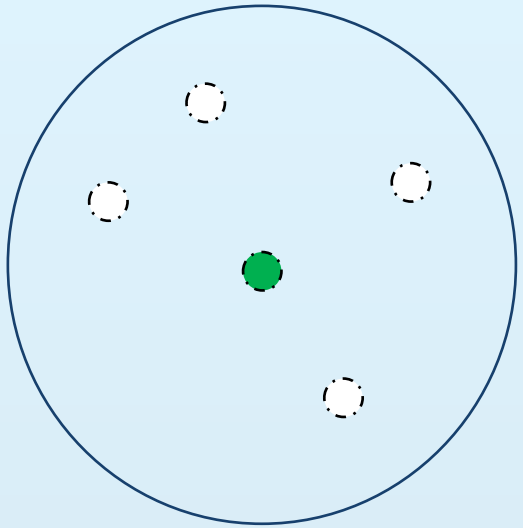


ANI: Average Nucleotide Identity

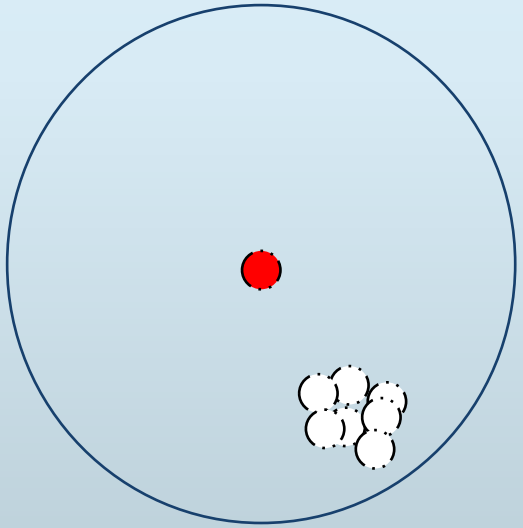
One size does not fit all …



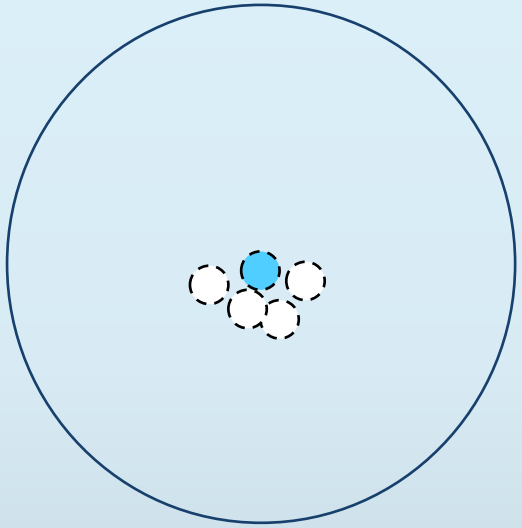
One size does not fit all …



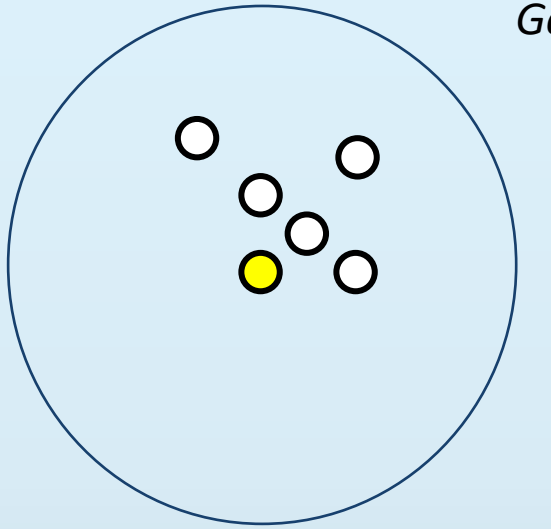
<u>Genus 2</u> species 1



Genus 1 species 2

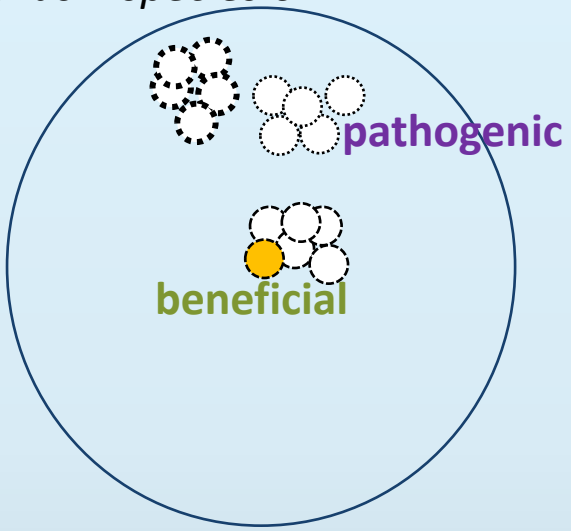


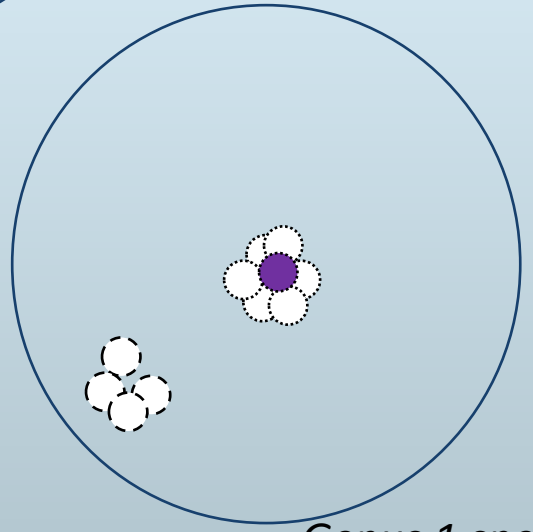
Genus 1 species 1



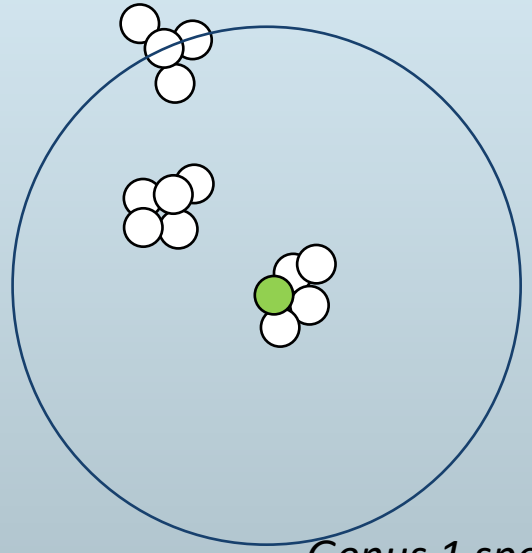
Genus 1 species 5

Genus 1 species 6





Genus 1 species 4



Genus 1 species 7

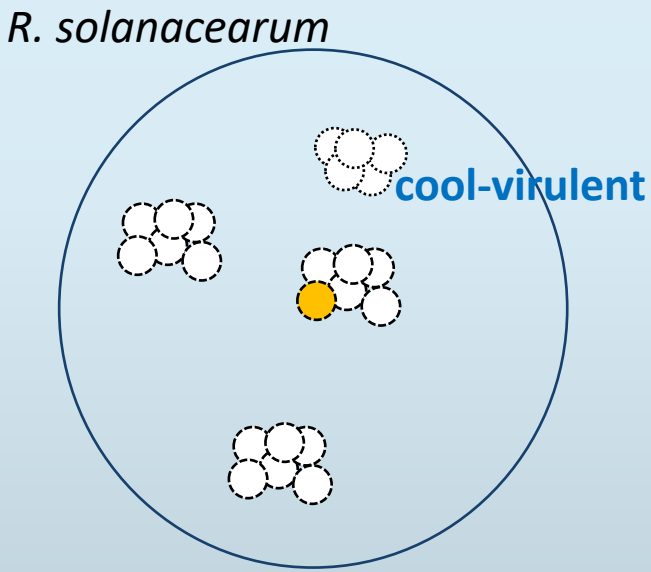
Species and regulatory agencies

- 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* [7]
64. *Rathayibacter toxicus*
65. *Sclerophthora rayssiae* [7]
66. *Synchytrium endobioticum*
67. *Xanthomonas oryzae*

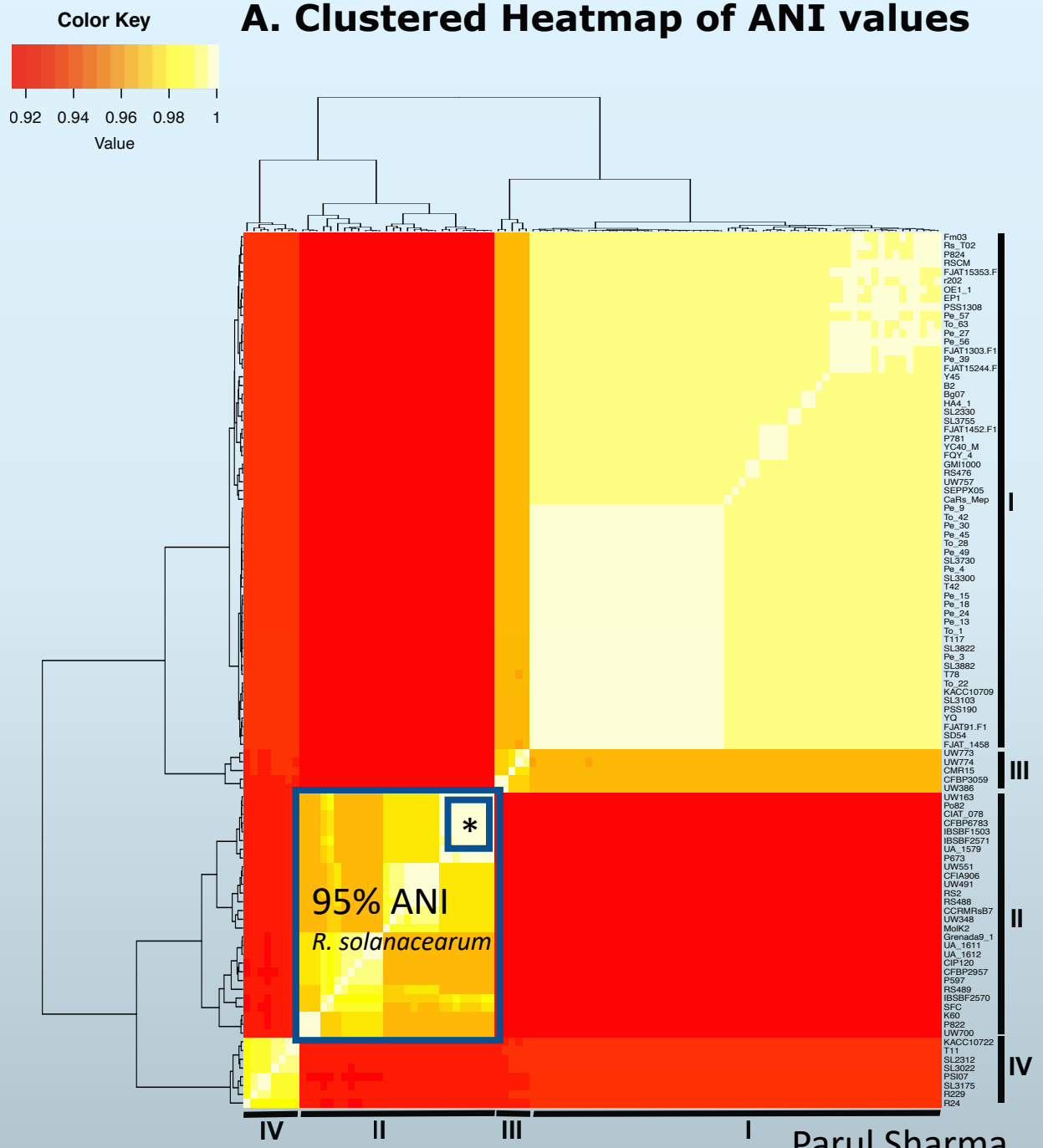
https://www.selectagents.gov/SelectAgentsandToxinsList.html



So what do we need to do?

- **Discovery:** there are cool-virulent *R. solanacearum* strains
- **Description/phenotyping:** which strains of *R. solanacearum* are cool-virulent and which are not?
- **Classification:** precisely define/circumscribe the group of *R. solanacearum* strains that are cool-virulent and the ones that are not.
- **Naming:** give this group a unique name/ID to communicate about it.
- **Identification:** provide tools to quickly and easily identify a bacterial strain as a member of the cool-virulent group of *R. solanacearum*
- **Spoiler alert:** it will not be a species with a 95% ANI threshold and will not have a latin *genus species* name!

A. Clustered Heatmap of ANI values

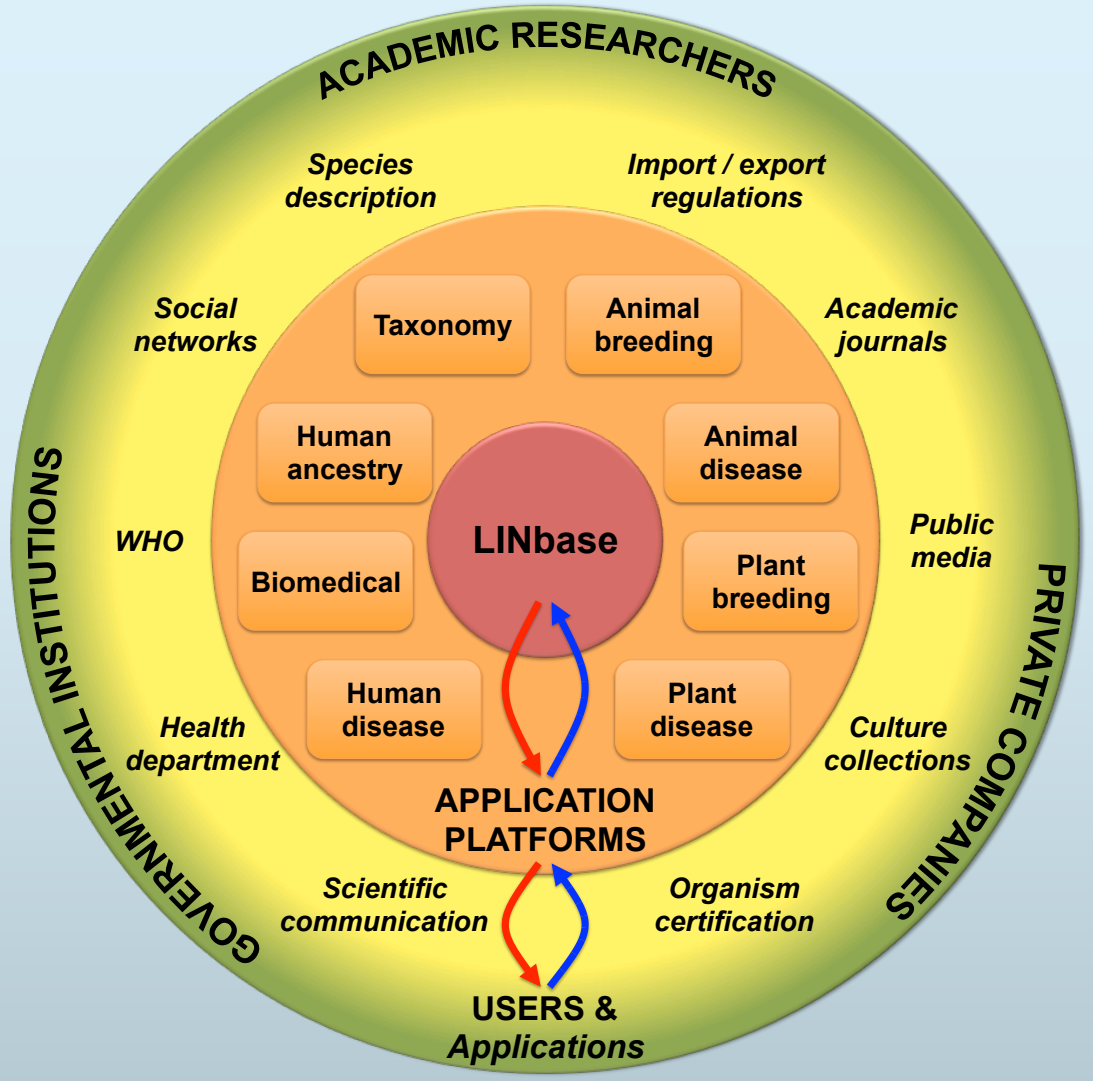


Parul Sharma

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 <u>named groups</u> with <u>distinct phenotypes</u> 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!**

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 (<u>expanding hierarchical taxonomy from the species almost to the individual</u>).

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

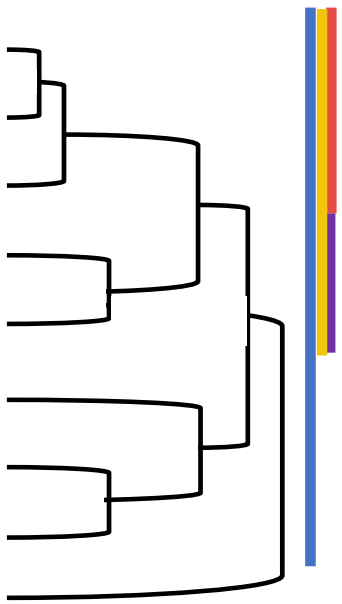
and how can LINs be used
to precisely identify unknown organisms as members of groups (that have phenotypes we care about)?

LINgroups:

<u>any</u> 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



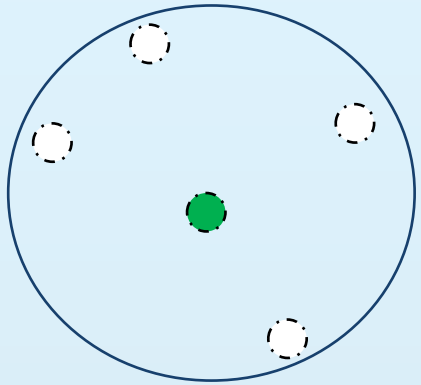
LINgroup: $0_A0_B0_C0_D0_E0_F$
Ralstonia solanacearum
Phylogroup IIB; cool virulent lineage

LINgroup: $0_A0_B0_C0_D0_E1_F$
Ralstonia solanacearum
Phylogroup IIA

LINgroup: $0_A0_B0_C0_D0_E$
Ralstonia solanacearum species

LINgroup: $0_A0_B0_C$
Ralstonia solanacearum species complex

LINgroups fit all sizes

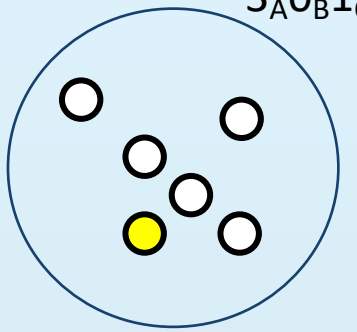


Genus 2 species 1

$1_A0_B1_C0_D0_E0_F$

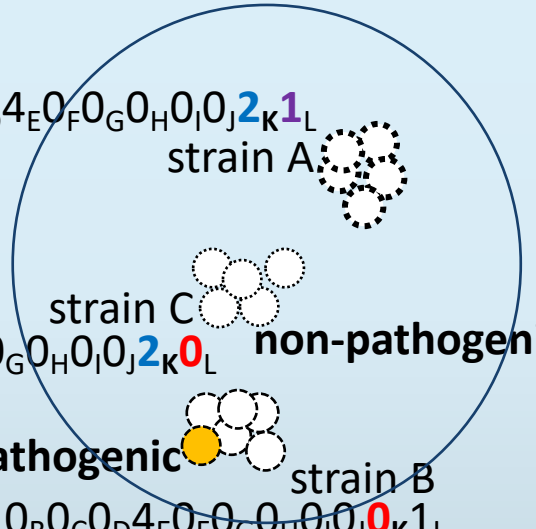
Genus 1 species 5

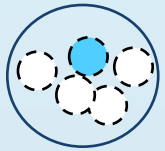
$3_A0_B1_C0_D0_E0_F8_G$



Genus 1 species 6

$3_A0_B0_C0_D4_E0_F0_G0_H0_I$

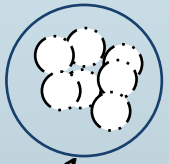




Genus 1 species 1

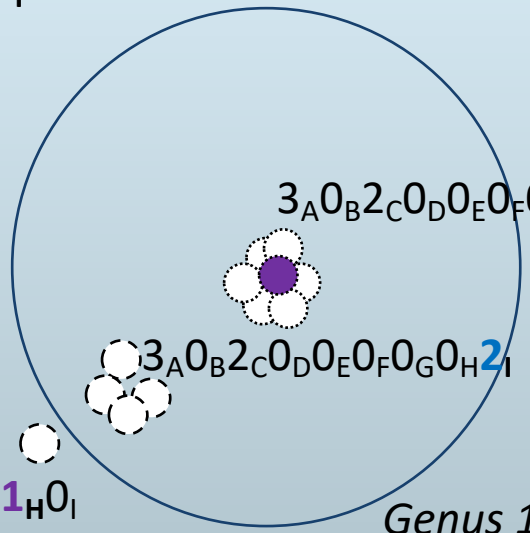
$3_A0_B1_C0_D0_E0_F0_G0_H0_I$

$3_A0_B1_C0_D0_E0_F4_G0_H0_I3_J$



Genus 1 species 2

$3_A0_B1_C0_D0_E0_F4_G0_H0_I$



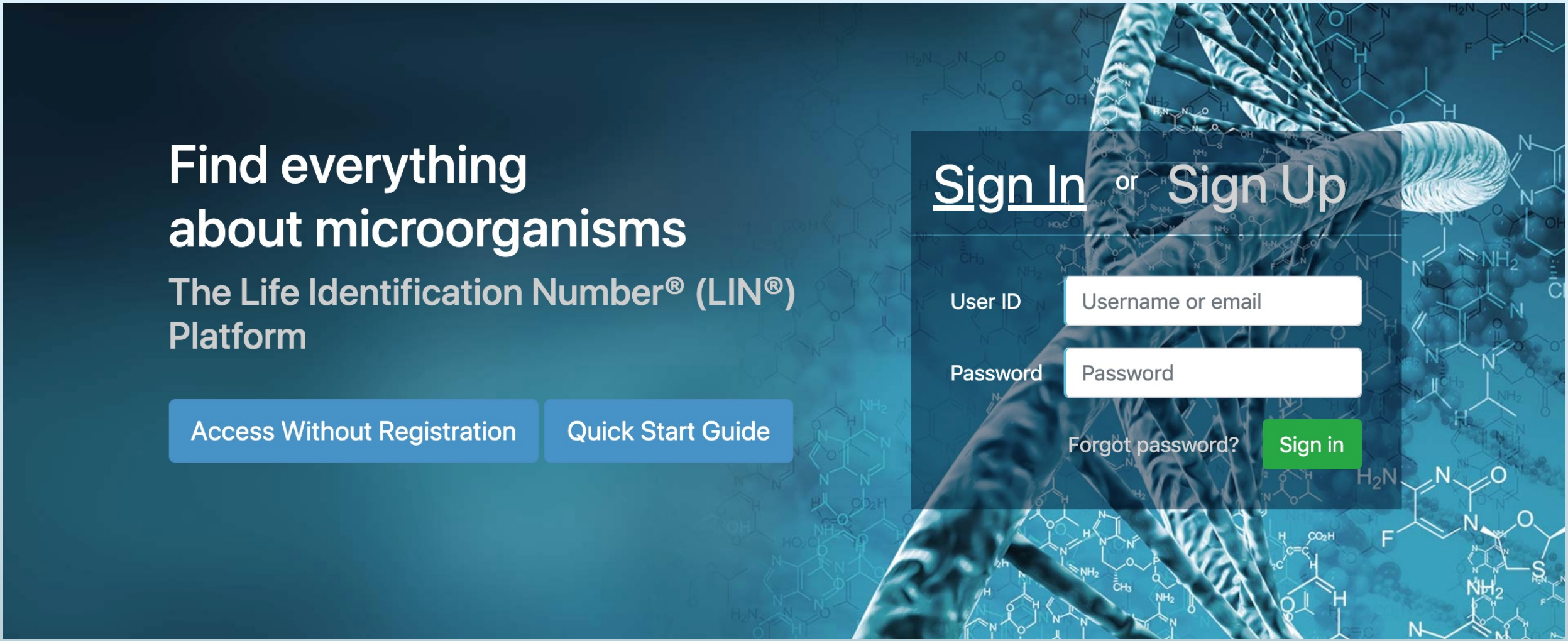
Genus 1 species 4

$3_A0_B2_C0_D0_E0_F0_G$

LINs and LINgroups have been implemented in:

LINbase

linbase.org


Find everything about microorganisms
The Life Identification Number® (LIN®) Platform
Access Without Registration
Quick Start Guide
Sign In
or
Sign Up
User ID
Username or email
Password
Password
Forgot password?
Sign in

LIN®base

Home | Upload | Search | Identify | About us | Sign out

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

Edit

LIN Database

Upload | Search | Identify

Submissions

No genome submitted.

Recent activities

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

<< < 1 2 3 4 5 > >>

Describing a LINgroup

0	3	0	0	0	0	0	0	0	0	1	0	0	0	1	0	Ralstonia	solanacearum	biovar 1 phylotype IIB sequevar 3	UW28	N/A
0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Ralstonia	solanacearum	phylotype II sequevar 1	NCPPB 909	N/A
0	3	0	0	0	0	1	0	0	0	0	0	0	0	1	0	Ralstonia	solanacearum	phylotype II sequevar 1	UY031	N/A
0	3	0	0	0	0	1	0	0	0	0	0	0	0	1	1	Ralstonia	solanacearum	biovar 2	UW23	
0	3	0	0	0	0	1	0	0	0	0	0	0	0	10	0	Ralstonia	solanacearum	phylotype II sequevar 1	UW551	
0	3	0	0	0	0	1	0	0	0	0	0	0	0	10	1	Ralstonia	solanacearum	biovar 2 phylotype IIB	UW425	
0	3	0	0	0	0	1	0	0	0	0	0	0	0	10	2	Ralstonia	solanacearum	biovar 2 phylotype IIB	UW408	

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

LINgroup

The LIN is the <u>immediately</u> assigned "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	3	0	0	0	0	1	0	0	0	0					

Type	Non-taxonomic group
Name	Pandemic lineage of cool-virulent strains **this name is not that important** **because the LIN is the stable identifier**
Description	Preliminary, unofficial circumscription of the pandemic lineage of cool-virulent strains; not an official circumscription. Do not use to make decisions about
URL	

Only as good as the phenotyping!

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.

Identify

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

There is no genome-based taxonomy without phenotyping

- There is no genome-based taxonomy without extensive phenotyping
- I need to know exactly which group of strains is cool-virulent and which is not.
- Only after I have the reliable phenotypic data, can I make the precise circumscription.
- For biosecurity this means, “someone” needs to fund the research to do the tedious phenotyping for pathogens and beneficials so that we can go from **reliable description** to **precise classification** to **precise identification** to **effective regulation** leading to **effective action**!

LINbase is becoming the genomeRxiv

- Currently, LINbase contains ~25,000 genome sequences, 20,000 species circumscriptions, a few dozen within-species pathogen groups
- It is functional and open to use but somewhat glitchy and slow
- LINbase will become genomeRxiv:
  - All high quality, public genome sequences
  - Simultaneous identification based on validly published species (NCBI taxID), GTDB species clusters, and specialty within-species groups
  - Fast identification and sharing of genome signatures (no need to share unpublished genome sequences)
  - Alerts to users when similar signatures are uploaded by other users: collaborations; outbreak discovery
  - Primers for selected groups for genome-**in**dependent identification

"For every complex problem there is an answer that is clear, simple and **wrong**" by H.L. Mencken

linbase.org @vinatzerLab vinatzer@vt.edu