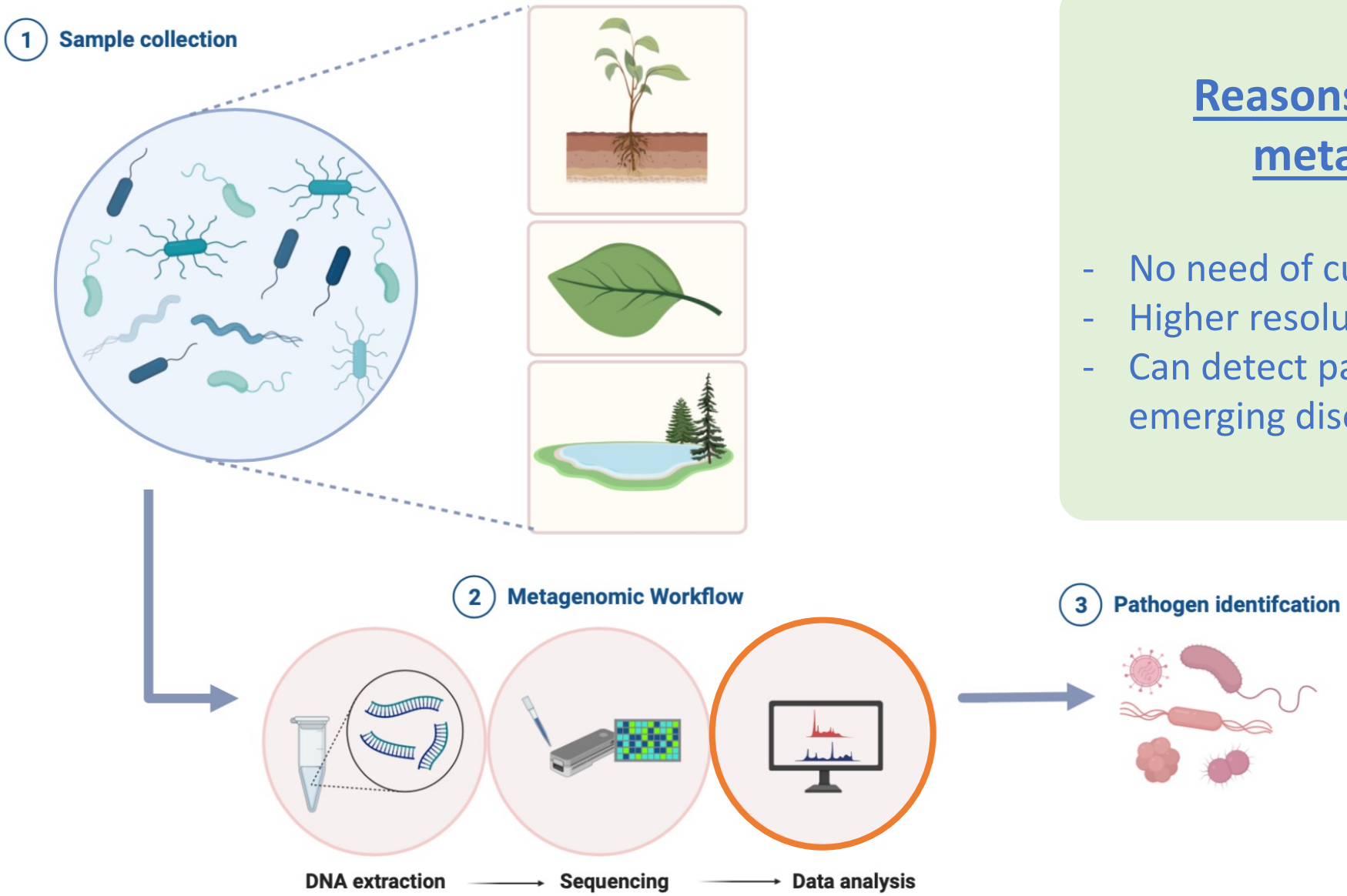
# Metagenomics for Pathogen Detection

Parul Sharma

ICPPB2024

Workshop: (Meta)Genomics for pathogen identification

# Metagenomics for pathogen detection

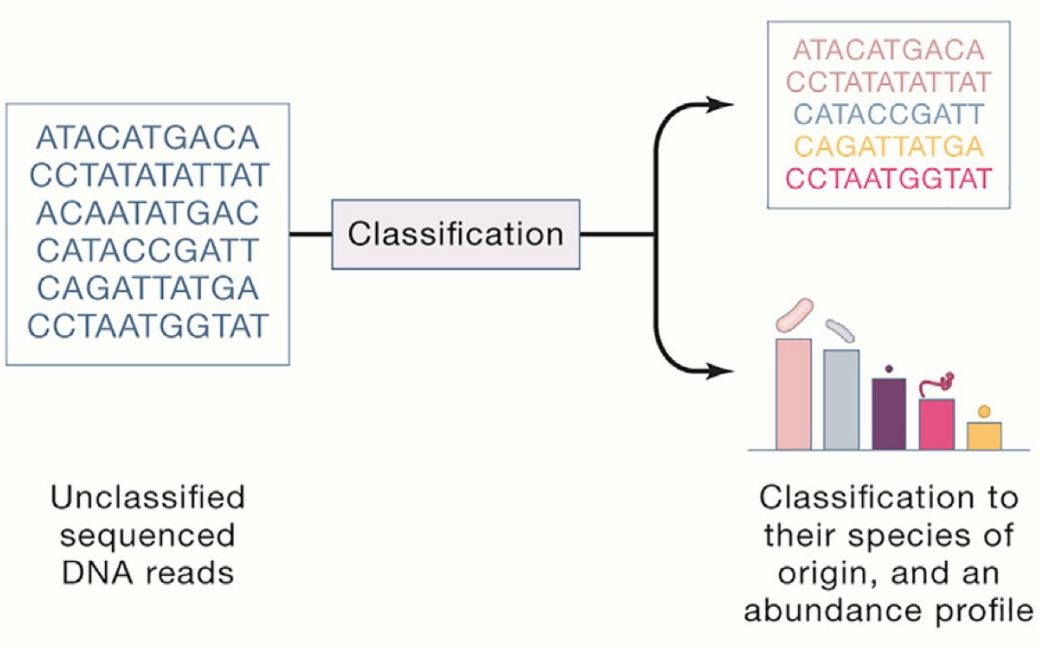


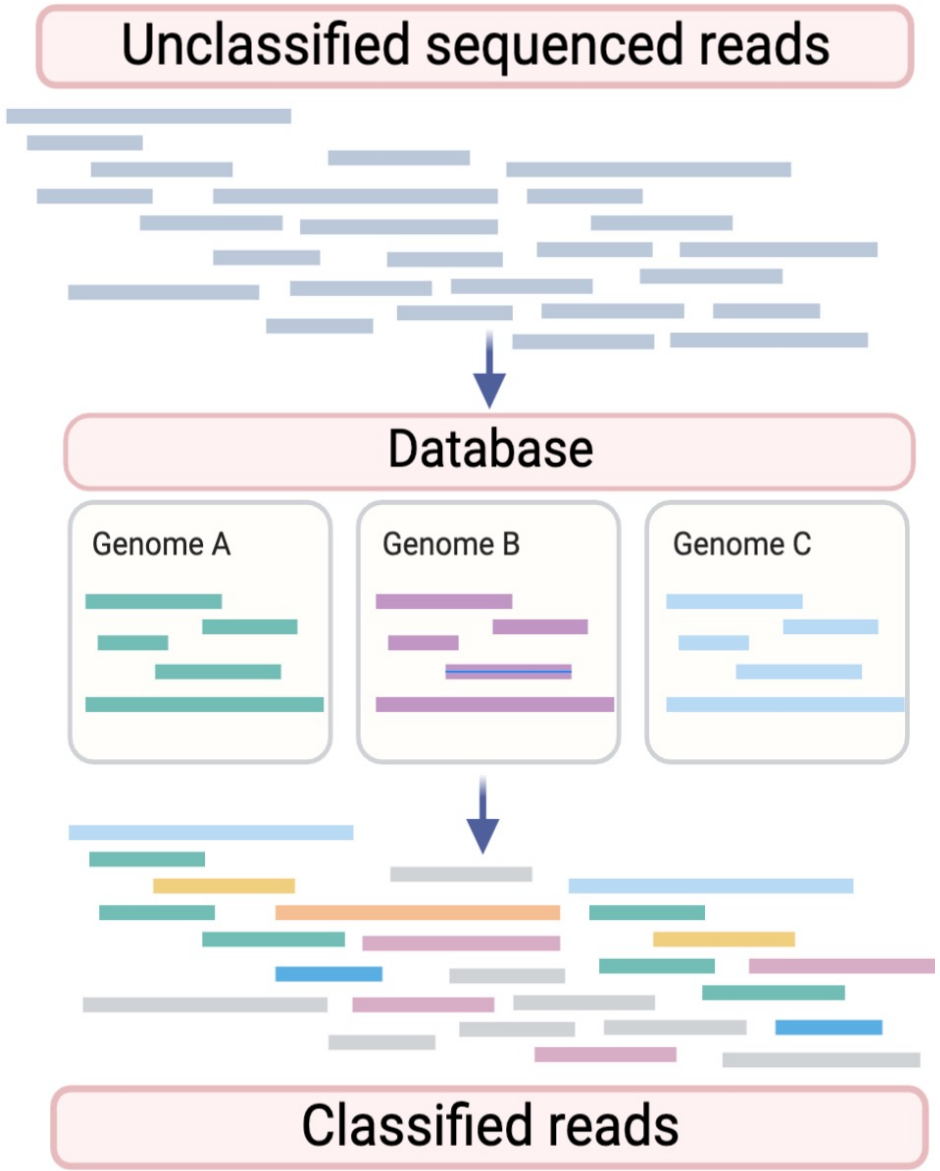

## Reasons for choosing metagenomics:

- No need of culturing the pathogen
- Higher resolution
- Can detect pathogens from newly emerging diseases

Created using BioRender

# Metagenomic Classification




## Issue:

**"Genomes share similarity!"**

- Between species below 95% similarity
- Within species above 95% similarity

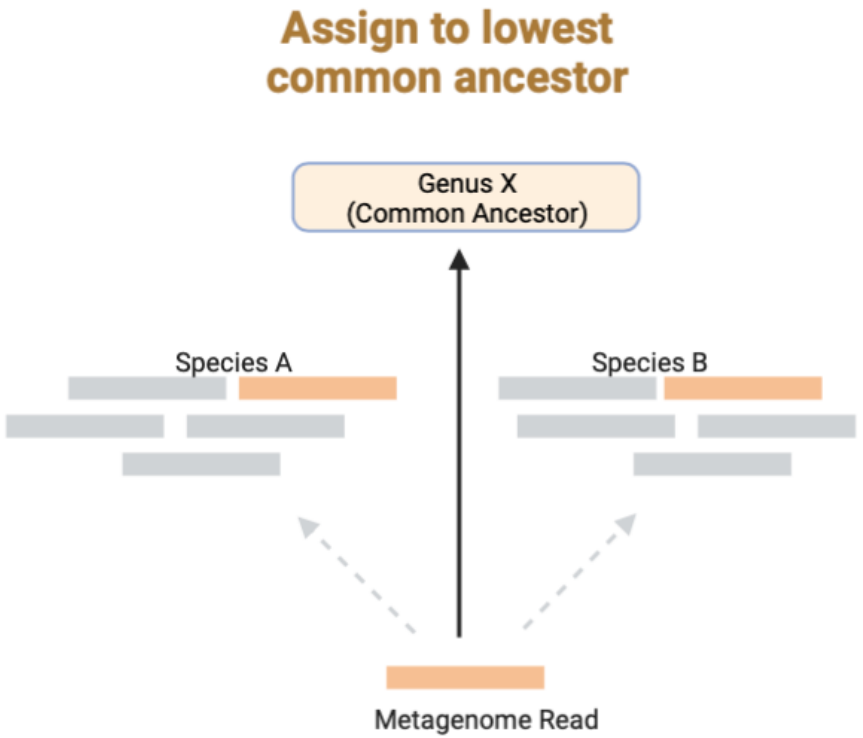
**How to determine the correct assignment in cases of conflict?**

Created using BioRender

# Using taxonomy!

Taxonomy is a classification system that allows for clear communication about organisms


To improve accuracy of read assignment




- Taxonomy is a hierarchical
- To improve the accuracy of read assignment, Lowest Common Ancestor (LCA) is used.

Created using BioRender

# Challenges with taxonomy…

- Taxonomic assignments are inconsistent between different Databases (NCBI v/s GTDB)
- Taxonomic assignments are frequently updated by scientists, but the changes are not always reflected in the public databases
- Most classification tools use NCBI taxonomy which is inaccurate
- Neither NCBI nor GTDB assign Taxonomy IDs at the sub-species level

**Thus, pathogen detection at the subspecies/strain level is difficult to attain using current classification tools.**

# LIN system for genome-similarity based taxonomy

Life Identification Numbers or LINs are used to classify bacteria based on a distance-based approach using sequence similarity between genomes. Each position of a LIN corresponds to a similarity threshold as below:

## LIN Assignment

| | | 70 | 80 | 90 | 95 | 96 | 97 | 98 | 99 | 99.9 |
|---|---|---|---|---|---|---|---|---|---|---|
| | Genomes | A | B | C | D | E | F | G | H | I |
| | G1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 97.5% similar to G1 | G2 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 |
| 96.1% similar to G1 | G3 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 |
| 98.2% similar to G2 | G4 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **1** | 0 |

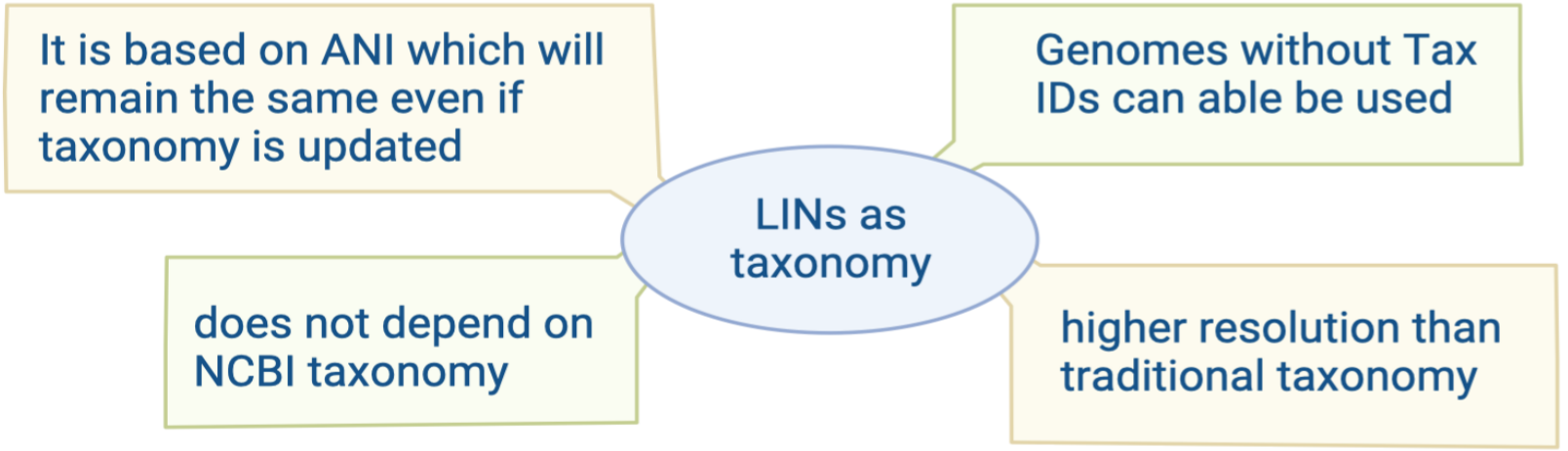
| | 70 | 75 | 80 | 85 | 90 | 95 | 96 | 97 | 98 | 98.5 | 99 | 99.25 | 99.5 | 99.75 | 99.9 | 99.925 | 99.95 | 99.975 | 99.99 | 99.999 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Genomes | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T |
| G1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G6 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

LINgroup at 98.5% ANI
*Ralstonia sonalacearum*
Sequevar 1

**LINgroup**

# Comparison of ranks in taxonomy vs LINS

## Advantages of using LIN taxonomy




The current default implementation of LINs describe genomes on 20 LIN positions from 70-99.999% similarity. Using these 20 positions as taxonomic ranks will give us 14 more taxonomic ranks for sub-species comparison.

14 more taxonomic ranks at the sub-species level

LIN-Ranks

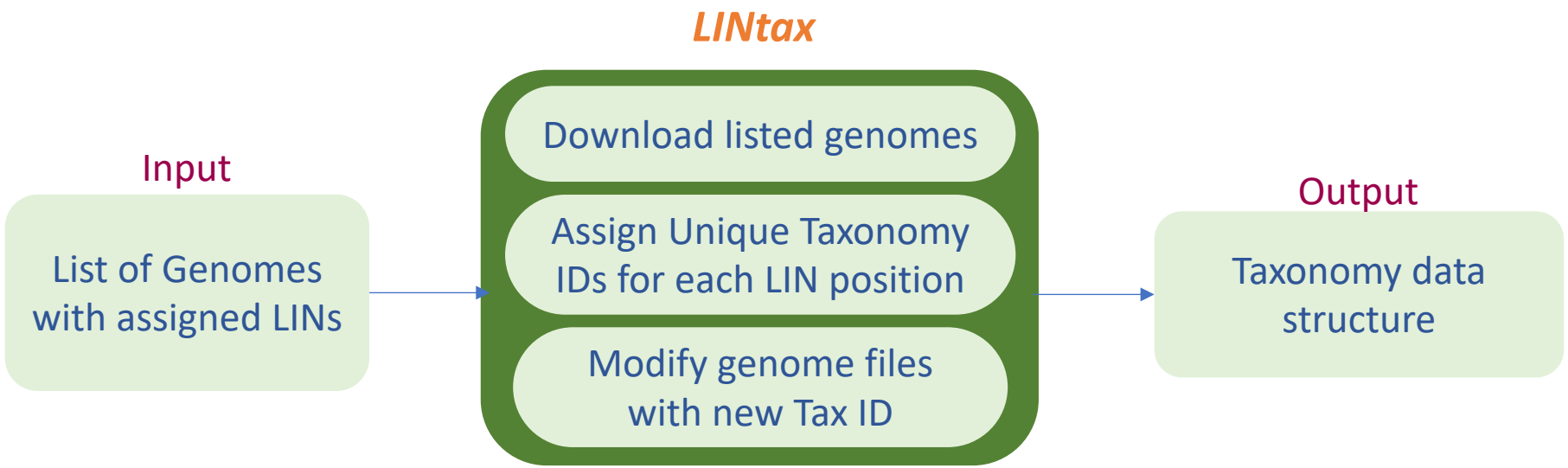
| ANI | 70 | 75 | 80 | 85 | 90 | 95 | 96 | 97 | 98 | 98.5 | 99 | 99.25 | 99.5 | 99.75 | 99.9 | 99.925 | 99.95 | 99.975 | 99.99 | 99.999 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Position | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |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

Genus — Species

NCBI-Ranks

Created using BioRender

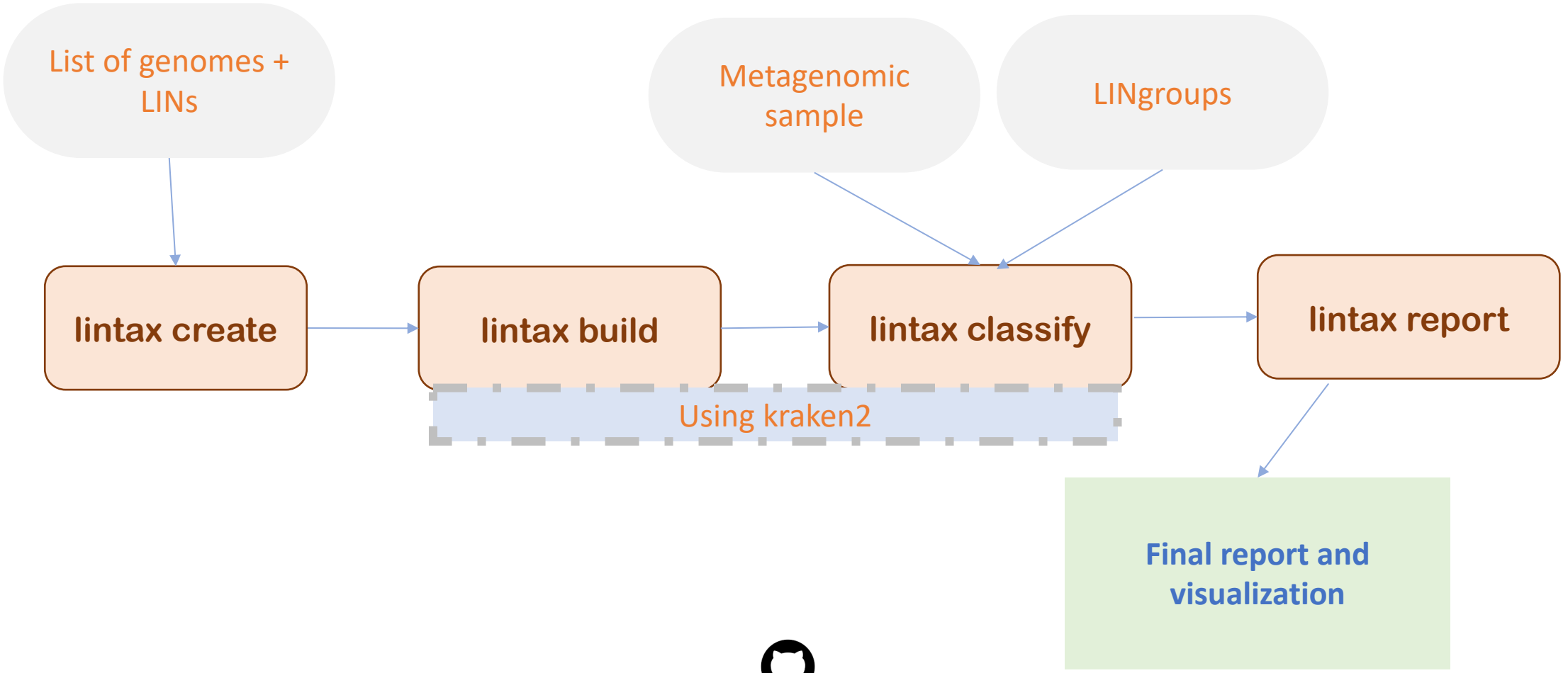
# Develop a tool for metagenomic classification using custom taxonomy databases based on genome-similarity thresholds




GitHub

parul-sharma/LINtax

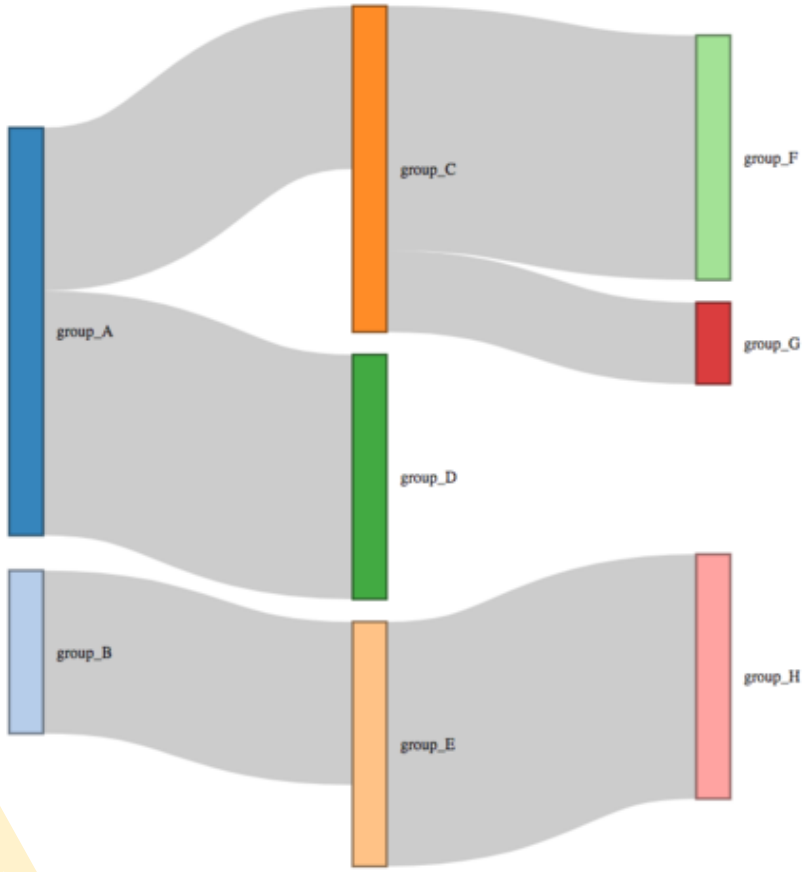
# Pipeline for using LINtax for strain-level pathogen detection




GitHub
parul-sharma/LINtax

# Summary report and visualization with LIN

| LINgroup_Name | Assigned reads | % assigned reads | Unique Assigned reads | % unique assigned reads | Total reads length |
|---|---|---|---|---|---|
| Group_A | | | | | |
| group_D | | | | | |
| group_C | | | | | |
| group_F | | | | | |
| group_G | | | | | |
| Group_B | | | | | |
| group_E | | | | | |
| group_H | | | | | |



# Ralstonia case study: Can we identify the pathogen?

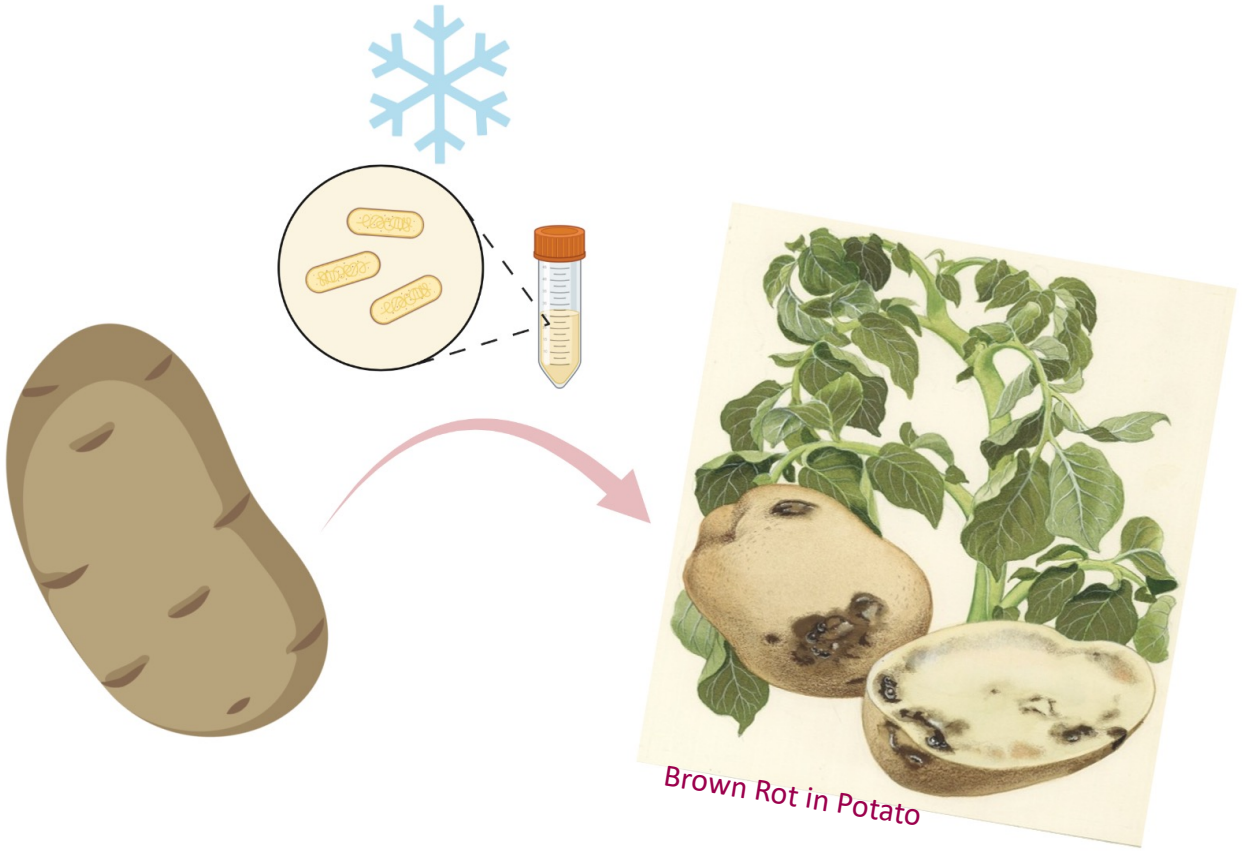
- Strain of *Ralstonia solanacearum*
- Formerly called **Race 3 biovar 2** (R3bv2)
- Can cause **100% crop loss**
- Capable of causing **disease at cooler temperatures** (20°C).
- Other strains capable to attack plants only at tropical temperatures (28°C).
- Not present in the US; threat to US agriculture
- Highly regulated under Select Agent Program

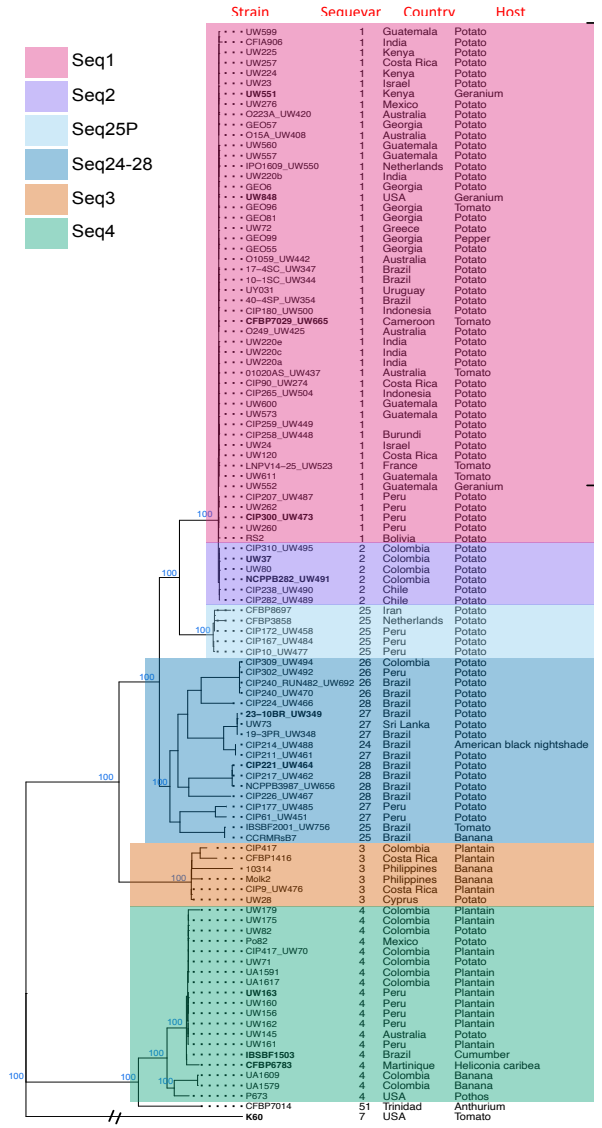
FEDERAL SELECT AGENT PROGRAM

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.redbubble.com%2Fpeople%2Fbiosecuritycoll%2Fworks%2F26922428-bacterial-wilt-of-potato%3Fp%3Dposter&psig=AOvVaw2SlQCimmsHHLCCKJZ_EOc_&ust=1583387605199000&source=images&cd=vfe&ved

Meng, F., Babujee, L., Jacobs, J. M., & Allen, C. (2015). Comparative transcriptome analysis reveals cool virulence factors of Ralstonia solanacearum race 3 biovar 2. *PLoS One*, *10*(10).

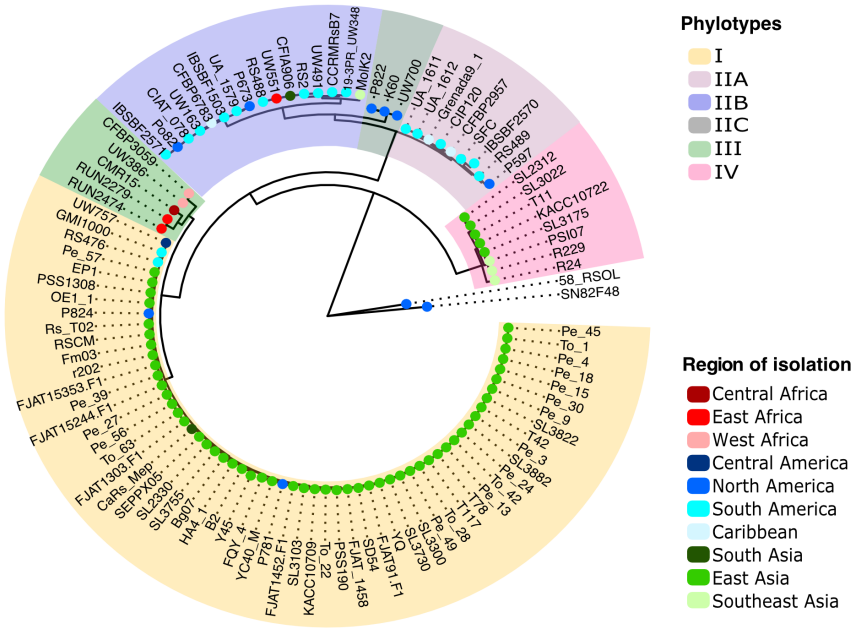
# Ralstonia case study: The objective

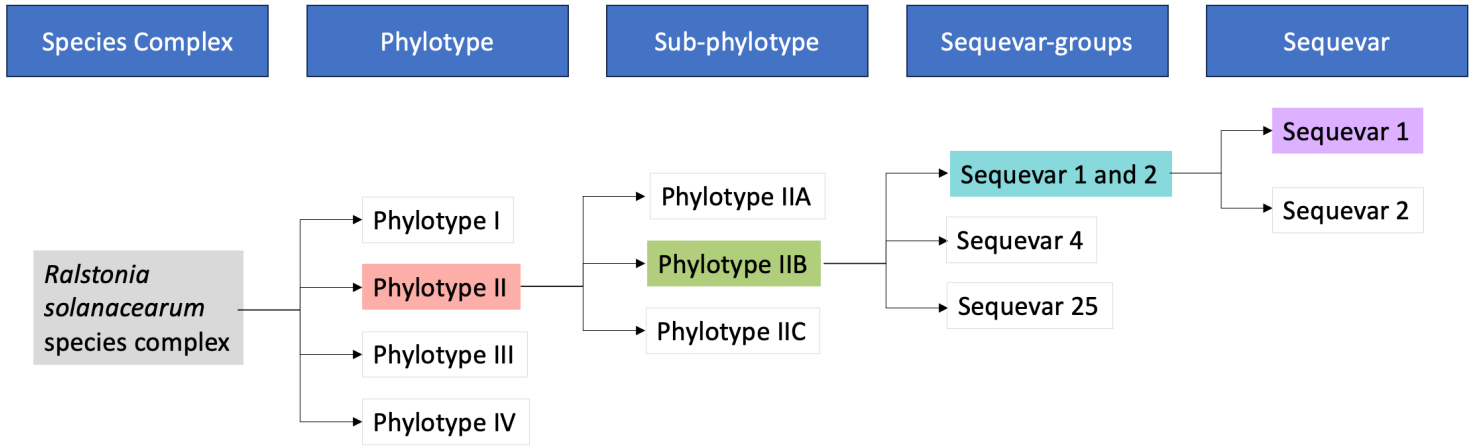


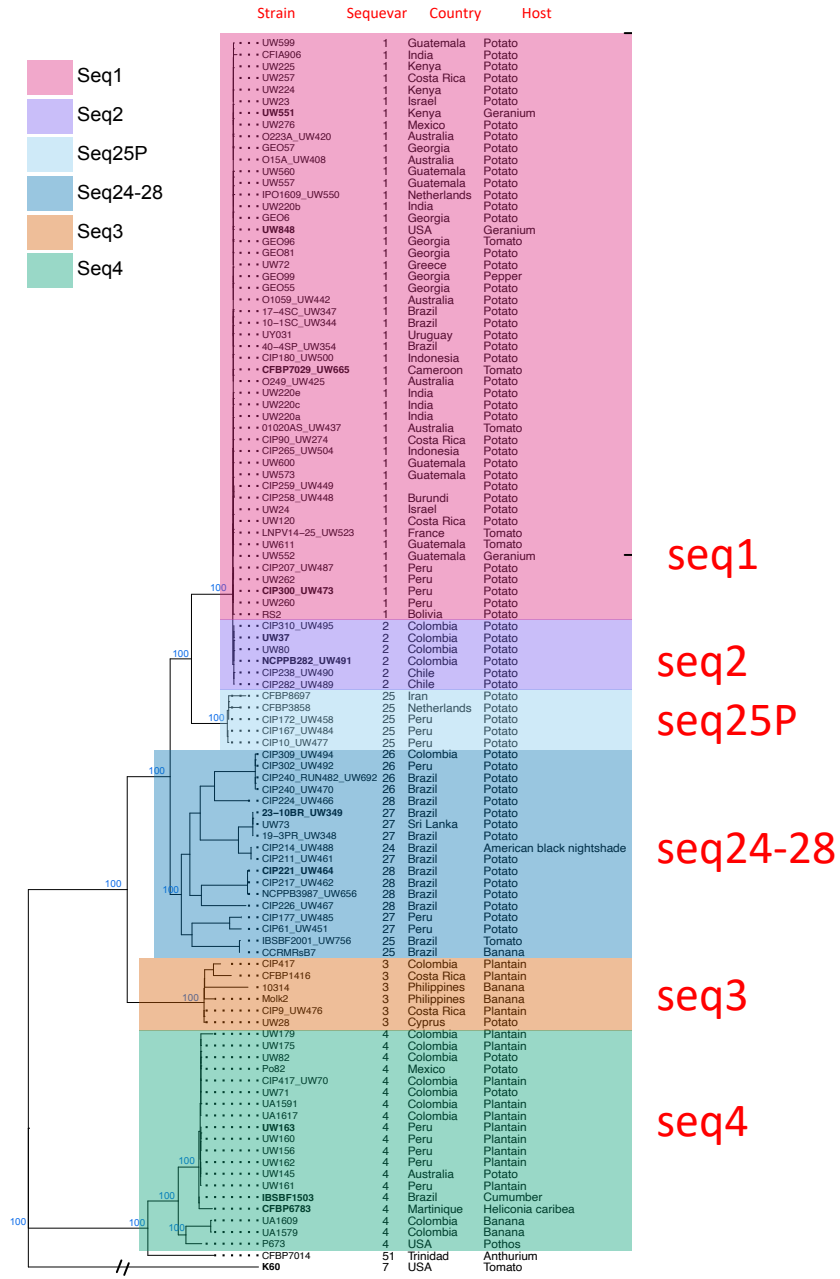




***Identify seq1 strains!***

# Ralstonia case study: Selecting reference genomes



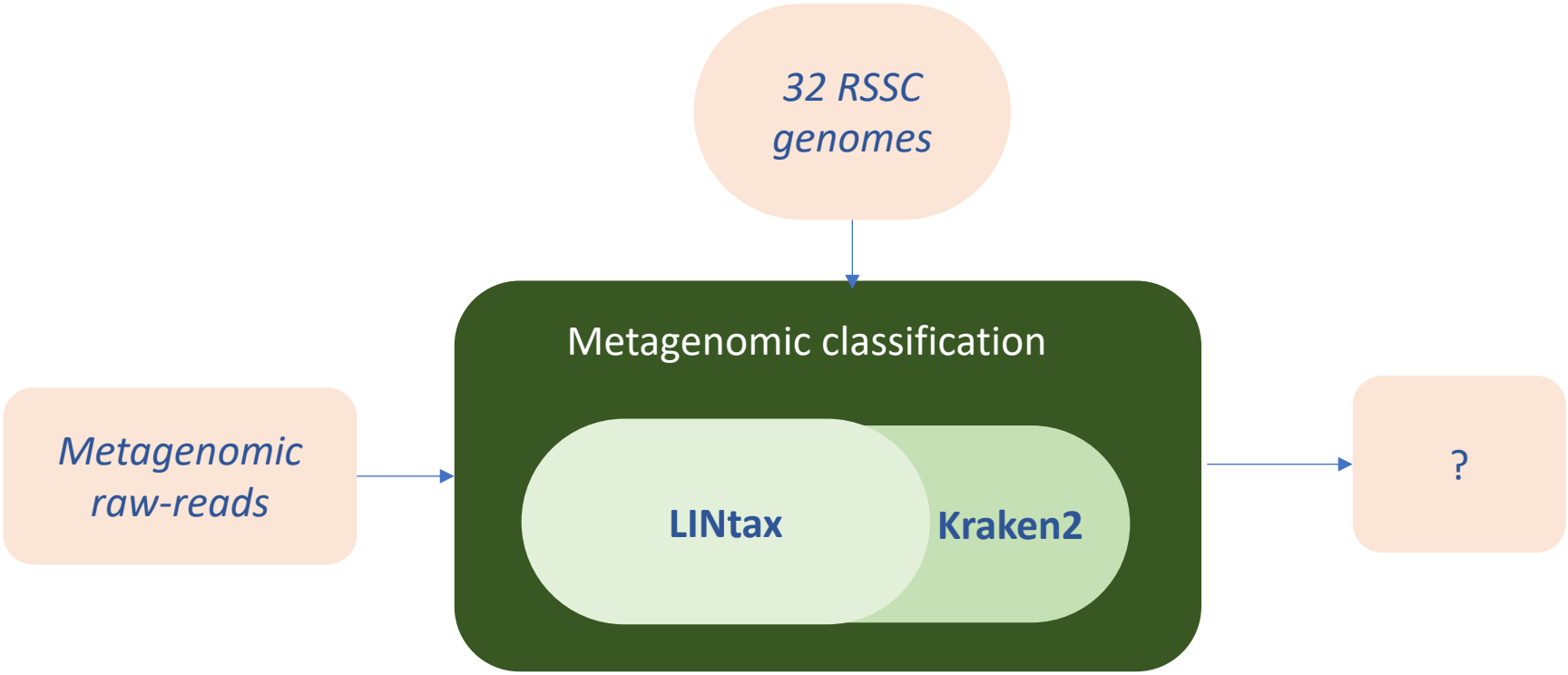



# *Ralstonia* case study: Example

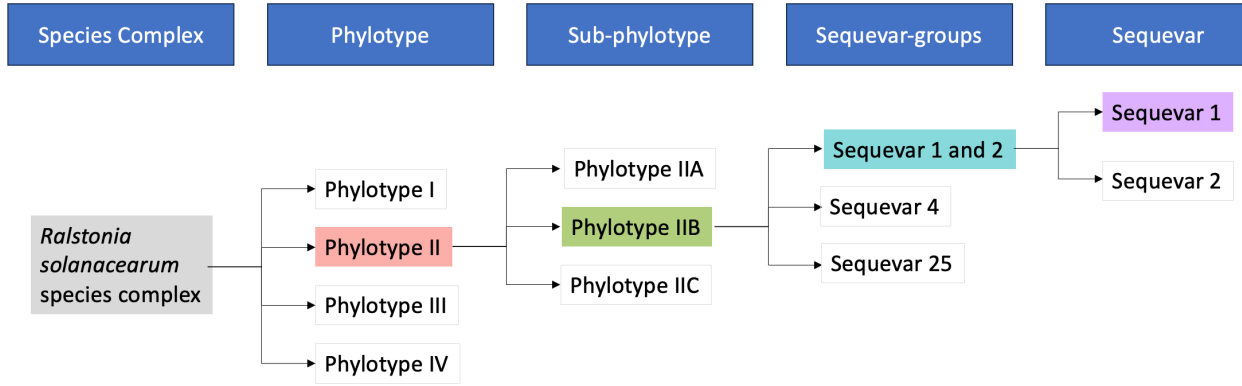
**Input Query :**

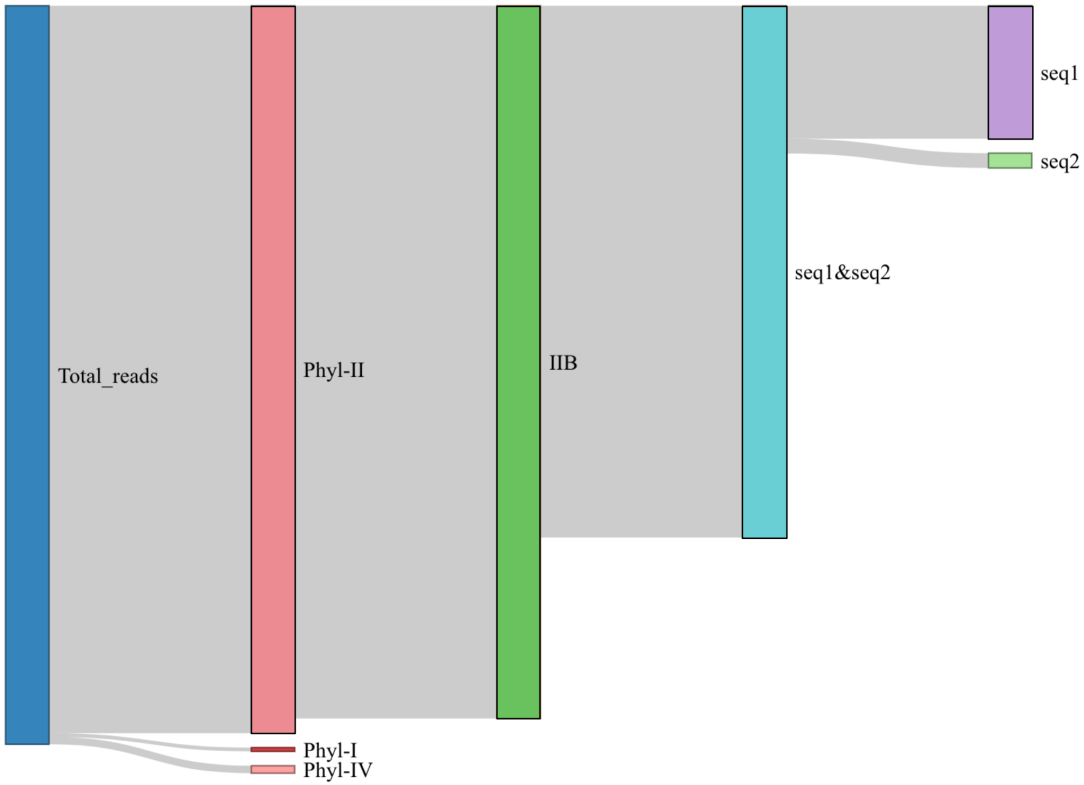
Metagenome sequencing samples of plant DNA spiked with seq1 strain DNA



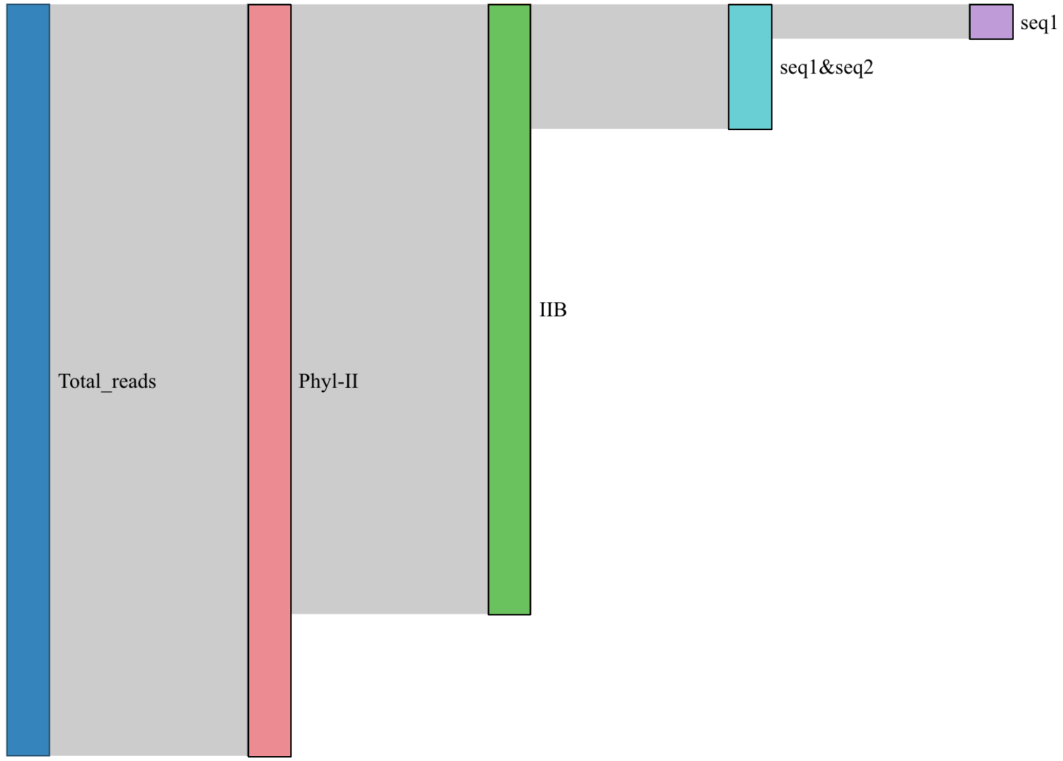
# Example:

*Plant DNA + 75pg/µL of seq1 strain*




Confidence threshold = 0




Confidence threshold = 10%

Sample provided by Eric Newberry at USDA-APHIS

# How does this compare with other tools?

## Example:

*Plant DNA + 75pg/microL of seq1 strain*

**Kraken2 with NCBI taxonomy**

| *Taxonomic Level* | *Taxon* | *# reads* |
|---|---|---|
| genus | *Ralstonia* | 181 |
| species | *Ralstonia solanacearum* | 181 |
| strain | *Ralstonia solanacearum* CMR15 | 3 |

**Kraken2 with LINtax**

| *Taxonomic Level* | *Taxon* | *# reads* |
|---|---|---|
| Species complex | *Ralstonia solanacearum* species complex | 554 |
| Species [Phylotype] | *Ralstonia solanacearum* [Phylotype II] | 197 |
| Sub-phylotype | Phylotype IIB | 193 |
| Sequevar-groups | Sequevar 1 and 2 | 144 |
| sequevar | Sequevar 1 | 36 |

## Key Findings

- We can achieve **precise** metagenomic classification from raw reads
- **Strain-level** pathogen identification achieved
- Using the **confidence-threshold** parameter reduces the false positive assignments

# Let’s do a Demo!

# What about the ‘infected field sample’?

**Sample: ‘barcode16’**

*From an infected tomato plant from Buckingham, Virginia*

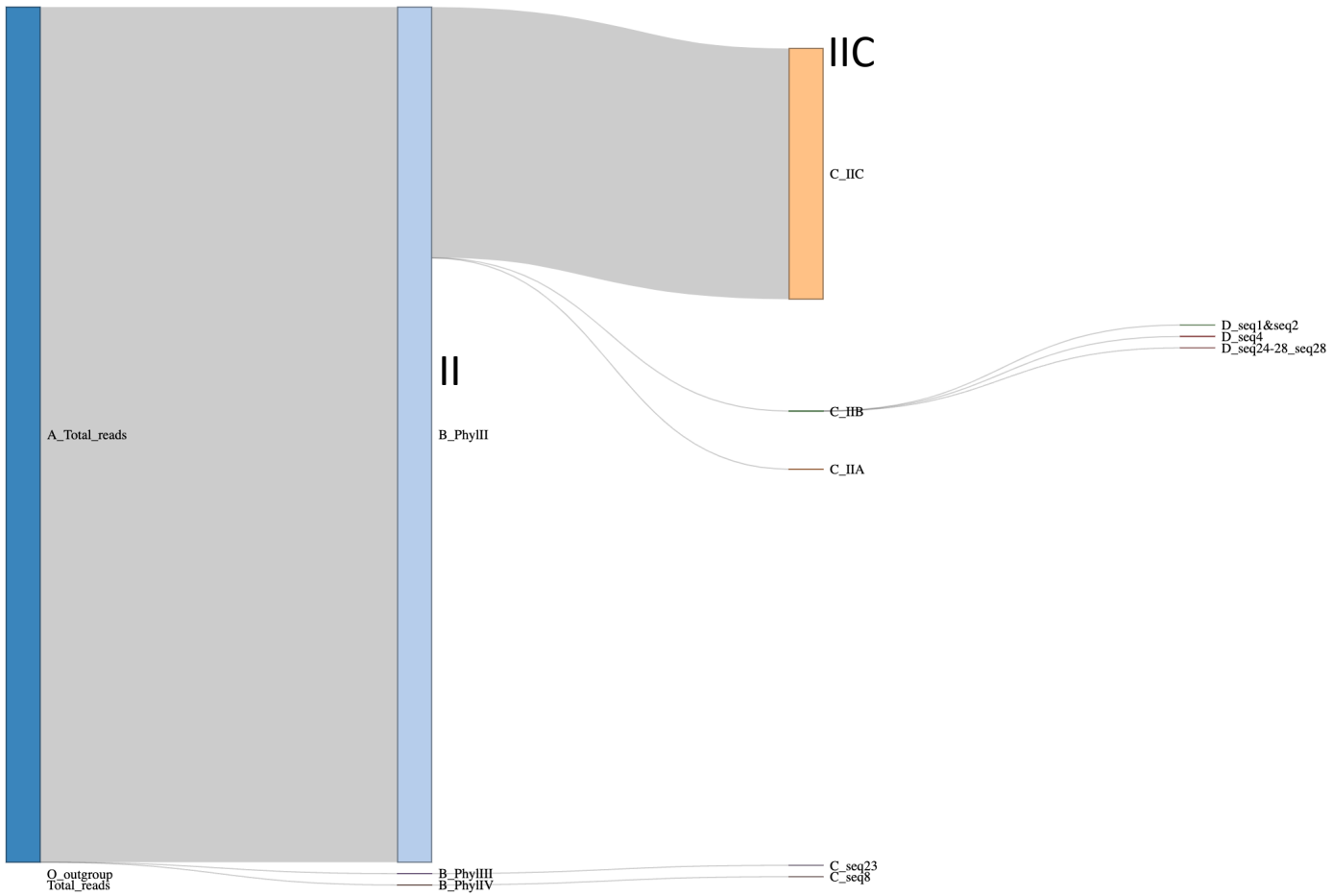
**Results from lintax classify ------**

**Lintax classify --db Kraken-db --lins**

barcode16_LINreport.csv > data

```
LINgroup_Name,LINgroup_prefix,Assigned_reads,Percentage_assigned_reads,Unique_Assigned_reads,Percentage_unique_assigned_reads,Total
A_Total_reads;B_PhylII,"14,1,0,0,0,3,0",223279,86.4825817846602,157779,61.112488283277436,329653945
A_Total_reads;B_PhylII;C_IIA,"14,1,0,0,0,3,0,1",5,0.0019366483588841806,0,0.0,0
A_Total_reads;B_PhylII;C_IIC,"14,1,0,0,0,3,0,2",65445,25.348790369435044,0,0.0,0
A_Total_reads;B_PhylII;C_IIB,"14,1,0,0,0,3,0,0",50,0.01936648358884181,6,0.0023239780306610167,1714
A_Total_reads;B_PhylII;C_IIB;D_seq1&seq2,"14,1,0,0,0,3,0,0,0,0,1,0,0,0,0",1,0.0003873296717768361,1,0.0003873296717768361,260
A_Total_reads;B_PhylII;C_IIB;D_seq4,"14,1,0,0,0,3,0,0,1,0,0,0",33,0.012781879168635593,0,0.0,0
A_Total_reads;B_PhylII;C_IIB;D_seq24-28_seq28,"14,1,0,0,0,3,0,0,0,0,5,0",10,0.0038732967177683613,0,0.0,0
A_Total_reads;B_PhylIII,"14,1,0,0,0,0,1",2,0.0007746593435536722,1,0.0003873296717768361,279
A_Total_reads;B_PhylIII;C_seq23,"14,1,0,0,0,0,1,1,0,0,0,0",1,0.0003873296717768361,0,0.0,0
A_Total_reads;B_PhylIV,"14,1,0,0,0,2,0",1,0.0003873296717768361,0,0.0,0
A_Total_reads;B_PhylIV;C_seq8,"14,1,0,0,0,2,0,0,0,1,0,0",1,0.0003873296717768361,0,0.0,0
O_outgroup,"14,1,0,1,2",11,0.004260626389545198,0,0.0,0
Total_reads,,258178,,,,
```

# Visualizing the results help




*Most of the reads classified at the lineage:*
*Phylotype II --> IIC*

*Hence, this is not a select agent pathogen!*

# Conclusion

A pipeline for pathogen identification using metagenomic sequencing data

1: **Metagenome classification and taxonomy**

2: LINtax to create custom taxonomy

3: **Identified Select Agent pathogen in our Case Study**

Overall, this work serves as a useful resource for diagnostic surveillance of plant pathogens

# ACKNOWLEDGEMENTS


**The LIN team**:

- Boris Vinatzer
- Lenwood Heath
- Reza Mazloom
- Long Tian

**The Ralstonia team:**

- Caitilyn Allen (UW-Madison)
- Tiffany Lowe-Power (UC-Davis)
- Eric Newberry (USDA APHIS)



Any questions?


Advanced Research Computing (ARC) at Virginia Tech for providing computational resources