Using genomics to identify plant-associated microbes

Michael J Stulberg, PhD

Molecular Biologist

The findings and conclusions in this presentation are those of the author and should not be construed to represent any official USDA or U.S. Government determination or policy

United States Department of Agriculture

USDA Organizational Chart

Secretary

Deputy Secretary

- Assistant Secretary for Civil Rights
- Assistant Secretary for Congressional Relations
- Assistant Secretary for Administration
  - Office of the Chief Information Officer
- Office of the Inspector General
- Office of the General Counsel
- Office of the Chief Financial Officer
- Office of the Chief Economist
- Office of Budget and Program Analysis
- Office of Communications
- Office of Partnerships and Public Engagement
- Office of Tribal Relations
- Office of Hearings and Appeals

Under Secretary for Food Safety

- Food Safety and Inspection Service

Under Secretary for Marketing and Regulatory Programs

- Agricultural Marketing Service
- Animal and Plant Health Inspection Service

Under Secretary for Natural Resources and Environment

- Forest Service

Under Secretary for Food, Nutrition, and Consumer Services

- Food and Nutrition Service

Under Secretary for Farm Production and Conservation

- Farm Service Agency
- Risk Management Agency
- Natural Resources Conservation Service

Under Secretary for Trade and Foreign Agricultural Affairs

- Foreign Agricultural Service
- Codex Alimentarius Commission

Under Secretary for Research, Education, and Economics

- National Institute of Food and Agriculture
- Economic Research Service
- National Agricultural Statistics Service
- Agricultural Research Service

Under Secretary for Rural Development

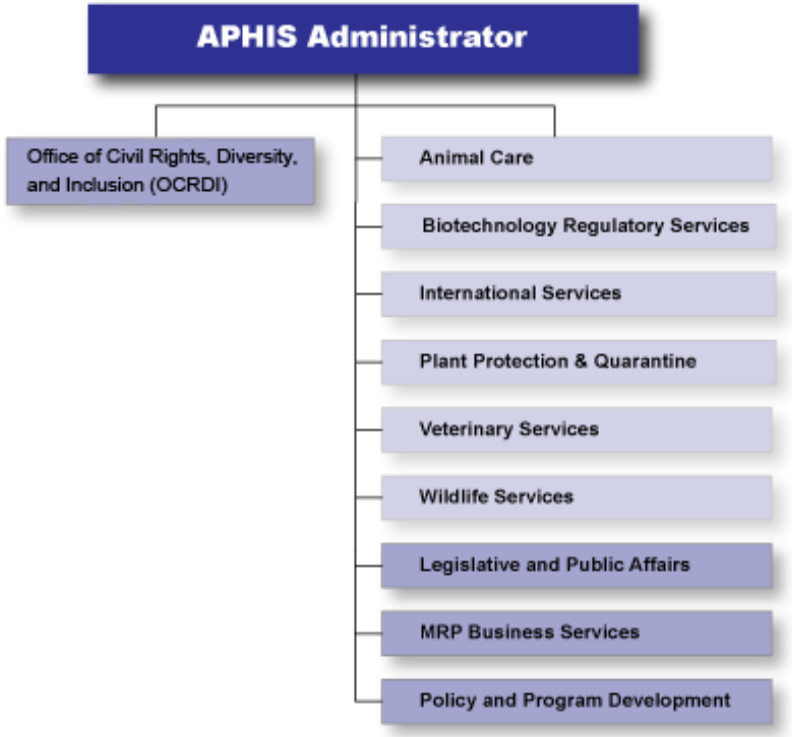
- Rural Housing Service
- Rural Utilities Service
- Rural Business Cooperative Service

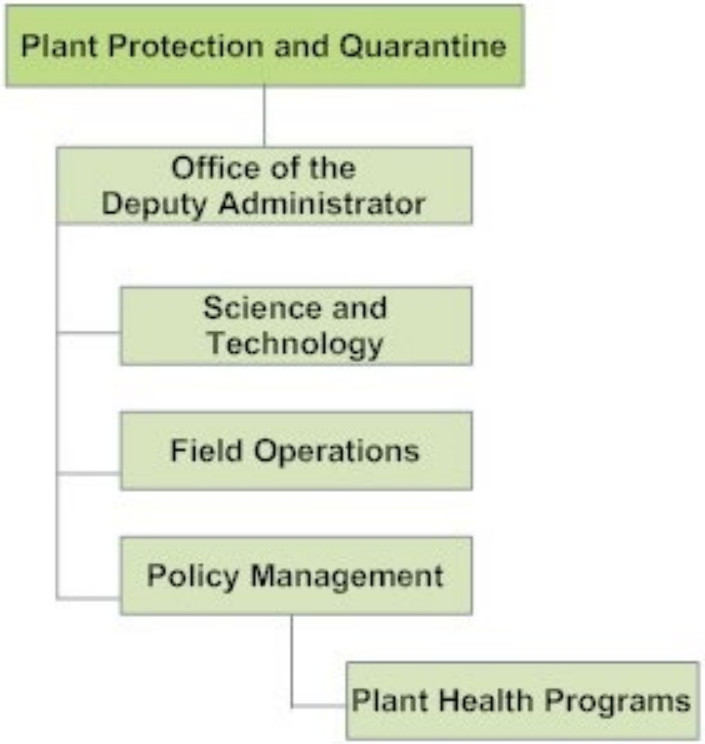
Updated 09/16/2019

This organization chart displays the names of USDA offices, agencies, and mission areas. Each office, agency, and mission area is placed within a cell connected by lines to show the structure and hierarchy (Under Secretary, Deputy Secretary, or Secretary) for which they fall under. An HTML version that lists USDA Agencies and Offices and USDA MissionAreas is also available on usda.gov.

The Secretary's Memorandum 1076-031 was signed August 12, 2019 effectuating a change to Rural Development.

APHIS org chart with PPQ highlighted





Overview

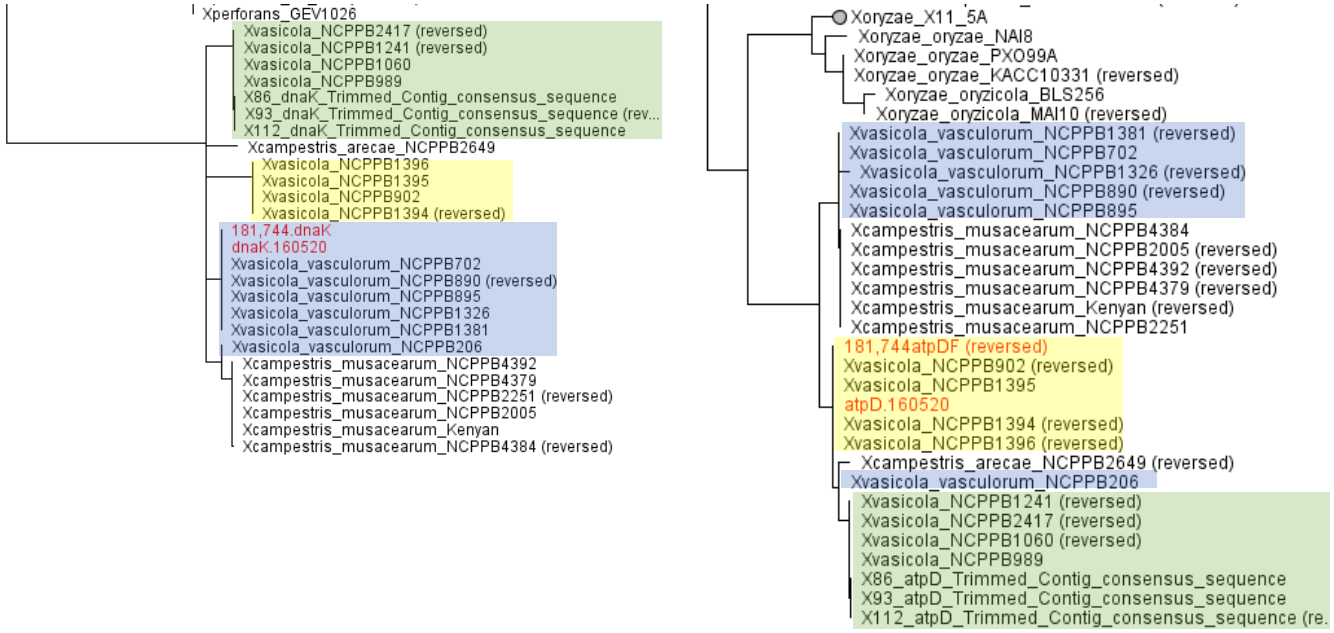
- Massive parallel sequencing to improve confirmatory diagnostics
- Challenges of using genome sequencing for microbe identification
- Solutions through validation initiatives

Bacterial leaf streak of corn in 2016

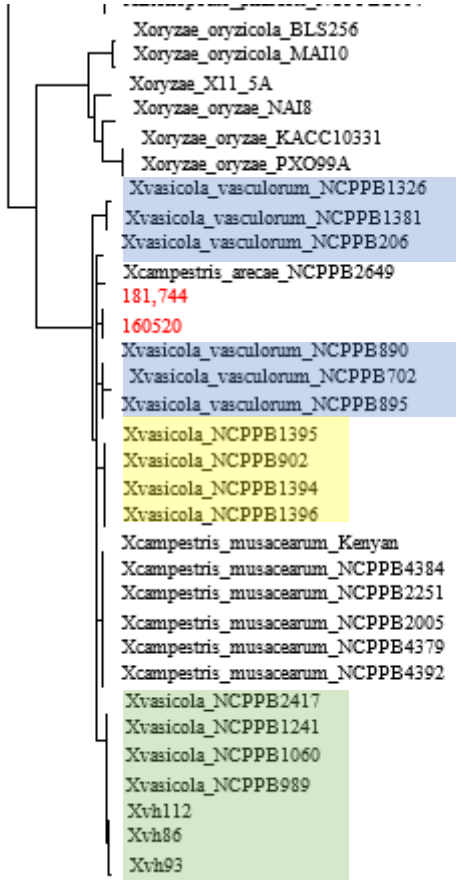
http://cropprotectionnetwork.org/

House-keeping genes provide unclear answer

Single gene phylogenies

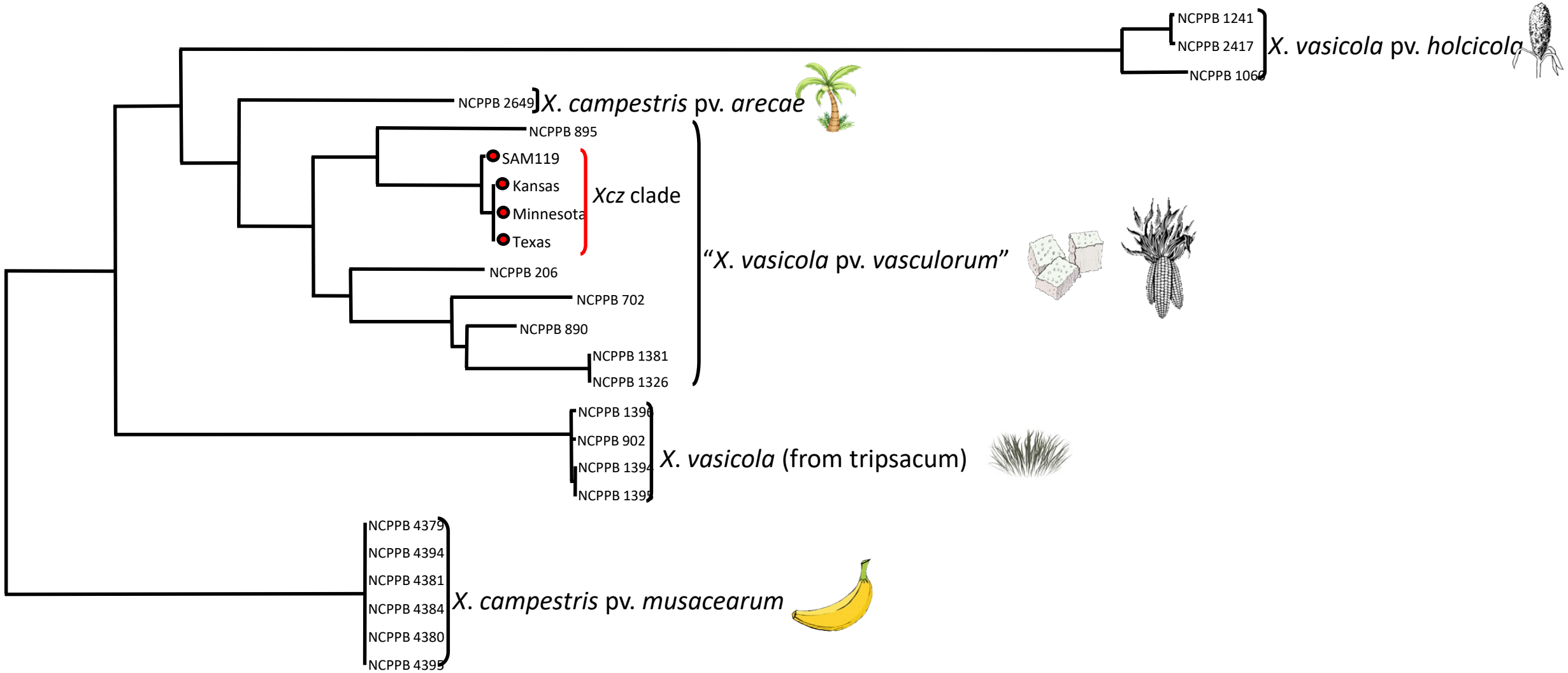


MLST



Higher taxonomic resolution using genomic analysis

Whole genome phylogenetics



There are questions beyond the genome sequence

- Should we call this *Xvv*?
  - The U.S. strain Xvv702 is 99% identical to other Xvv strains
  - Xvv702 is 98% identical to *X. campestris* pv. *musacearum* and *X. vasicola* pv. *holcicola*
  - *Species cutoff is considered to be 95%*
- Do all Xv have the same host range?
- Seed transmissible?

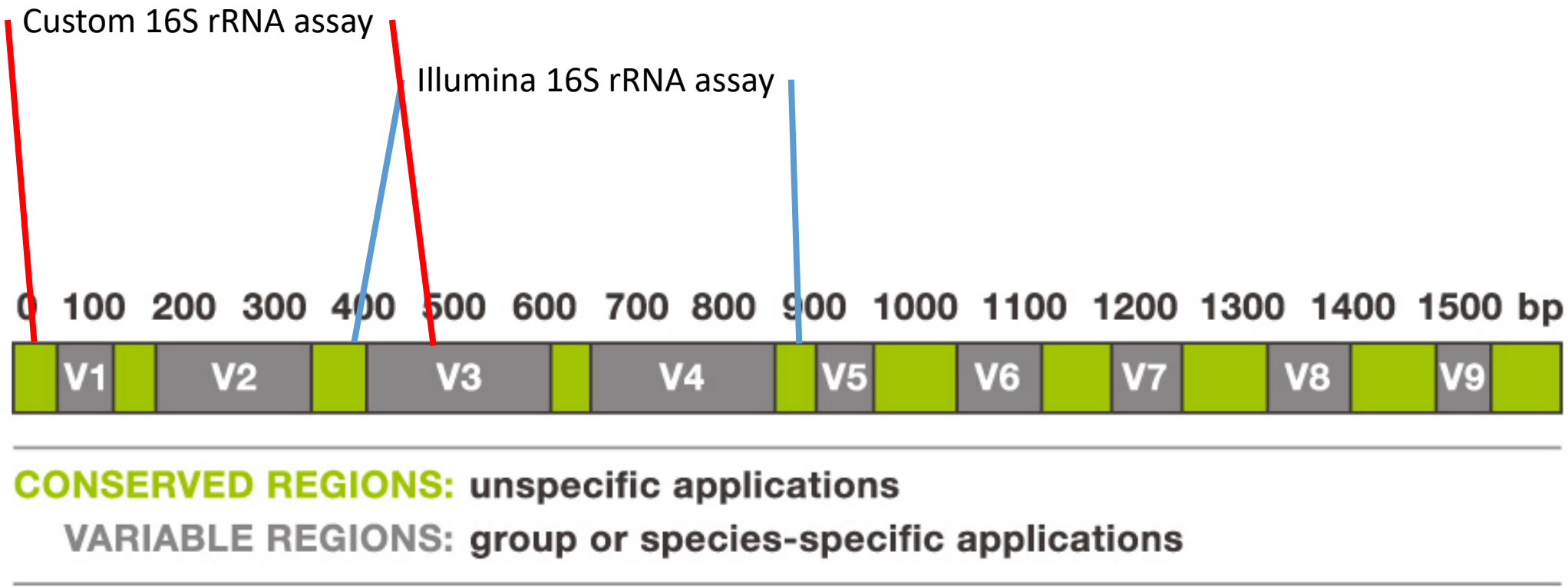
MPS can detect citrus greening

- Citrus greening caused by *Candidatus* Liberibacter asiaticus
- Major threat to the multi-billion dollar U.S. citrus industry
- The bacterium has low-titer and unevenly distributed in the tree
- Pathogen is unculturable
- New, early-detection technologies need confirmatory tools

Photo by T. Gottwald 2007

Division of Plant Industry, Florida Department of Agriculture

Partial 16S rRNA sequencing for identification

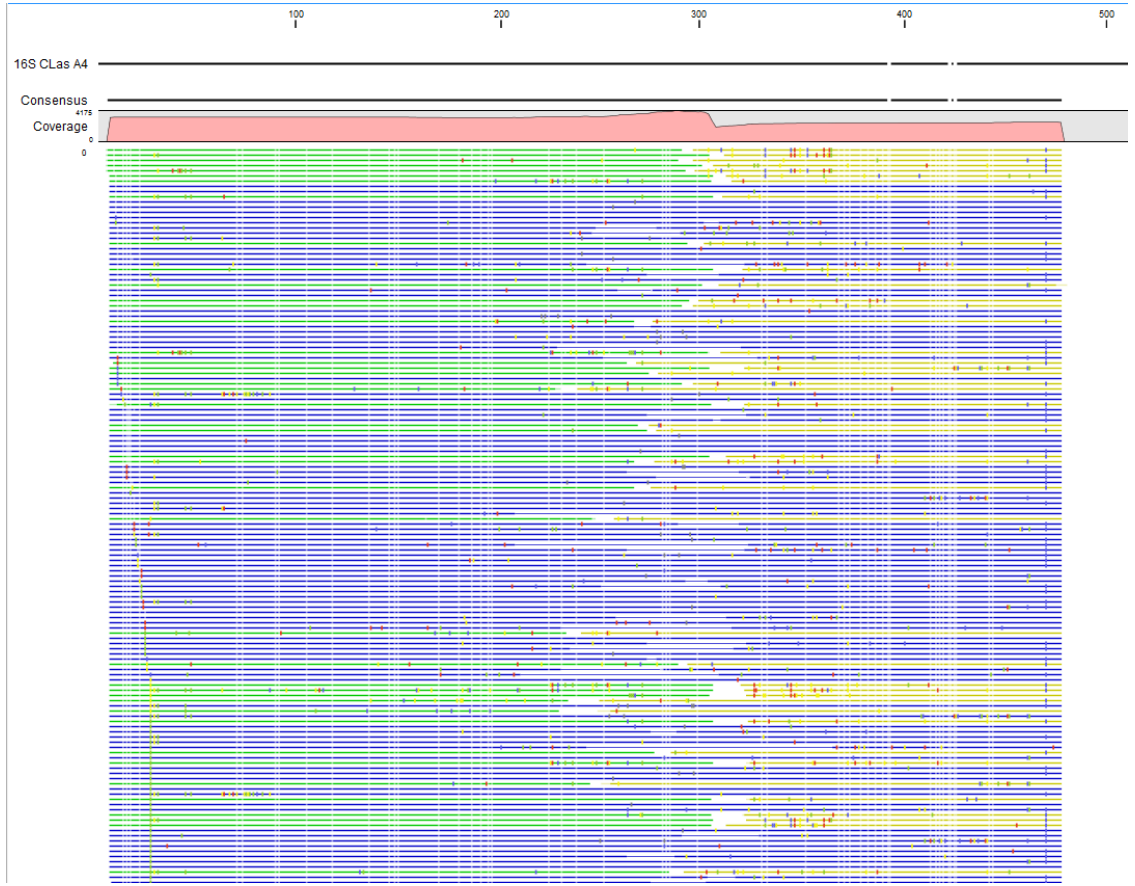


Successful map-to-reference

Rows: 4

Filter

Name	Consensus length	Total read count	Average coverage	Reference sequence	Reference length
16S CLas A4 mapping	470	5916	955.22	16S CLas A4	1503



Overview

- Massive parallel sequencing to improve confirmatory diagnostics
- Challenges of using genome sequencing for microbe identification
- Solutions through validation initiatives

Challenges of using genome sequencing

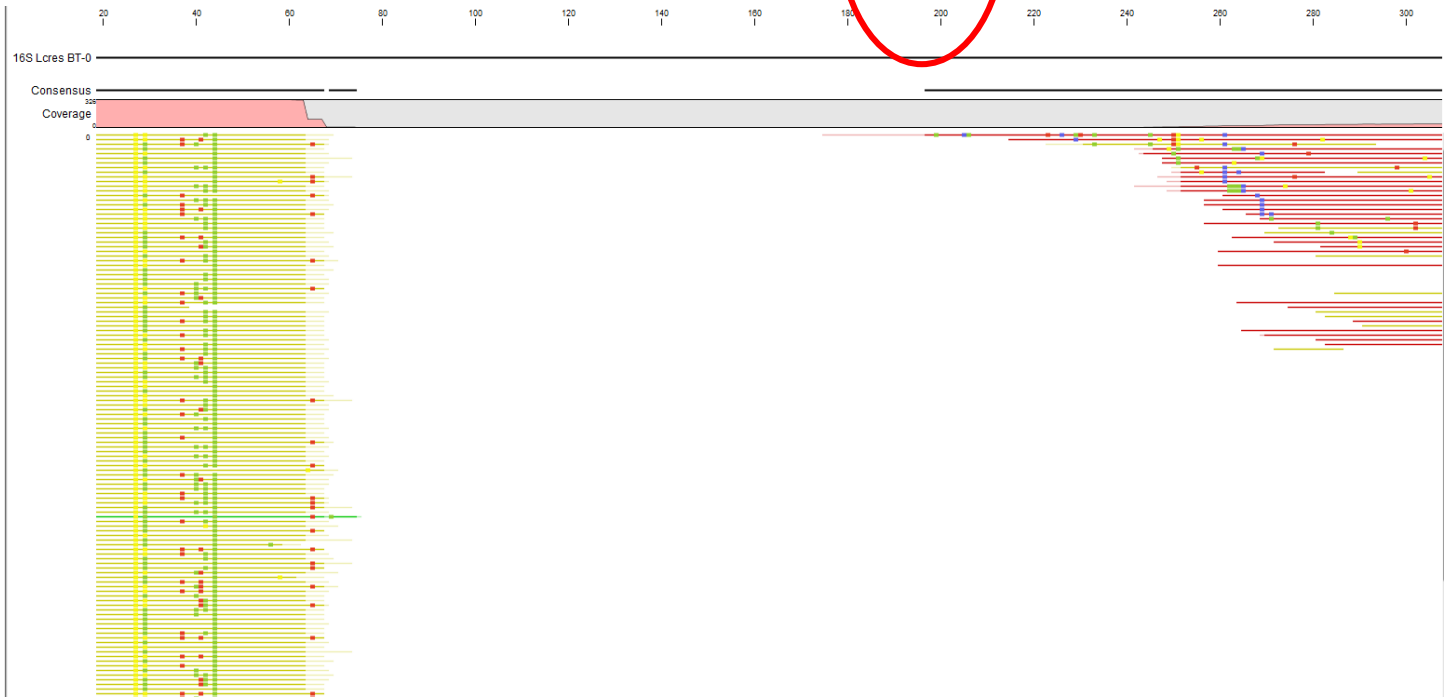
- Genome identification works well when:
  - Good organism database exists
  - Pathogen present in high titer
- Risk needs to be assessed for different/new organisms
- How do you know when you have done enough sequencing in a metagenomics sample to be sure something is not there?

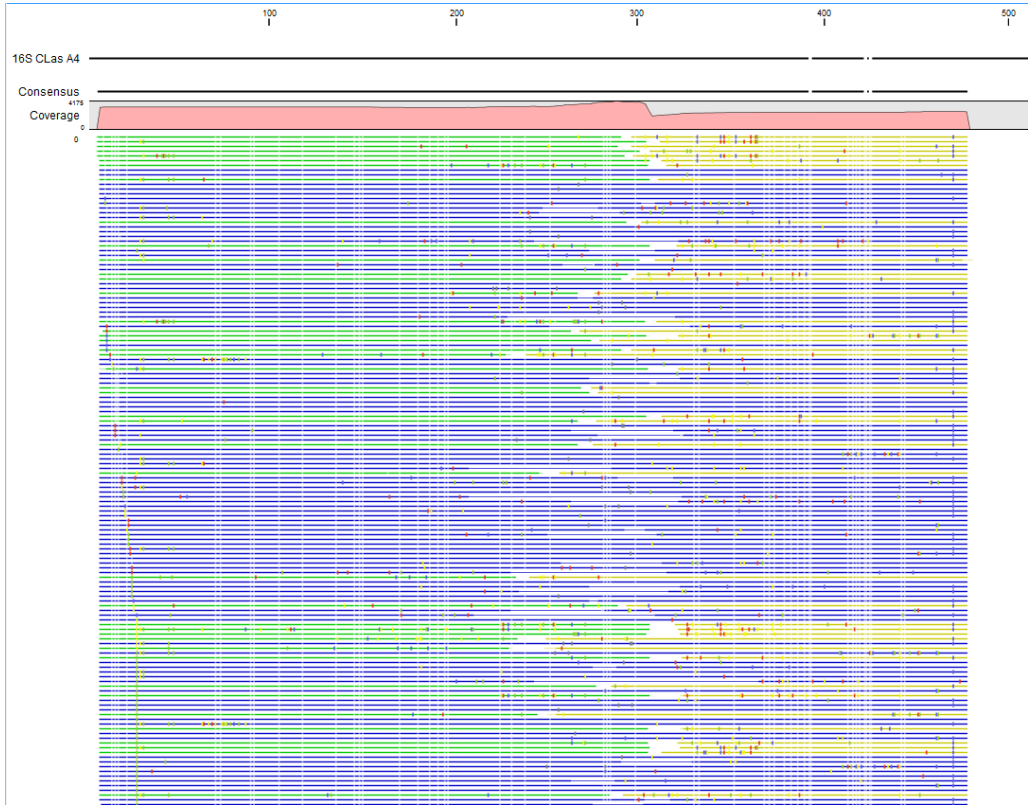
Unsuccessful map-to-reference

Rows: 4

Filter

Name	Consensus length	Total read count	Average coverage	Reference sequence	Reference length
16S CLas A4 mapping	470	5916	955.22	16S CLas A4	1503
16S CLaf PTSAPSY mapping	470	314	31.35	16S CLaf PTSAPSY	1501
16S CLam Sao Paulo mapping	264	462	16.85	16S CLam Sao Paulo	1484
16S Lcres BT-0 mapping	342	457	20.59	16S Lcres BT-0	1483





Tricky when something is new

Name	Consensus length	Total read count	Average coverage	Reference sequence	Reference length
16S CLas A4 mapping	0	0	0.00	16S CLas A4	1503
16S CLaf PTSAPSY mapping	19	1	0.01	16S CLaf PTSAPSY	1501
16S CLam Sao Paulo mapping	281	16	1.84	16S CLam Sao Paulo	1484
16S Lcres BT-0 mapping	368	275	46.63	16S Lcres BT-0	1483

Microbiome – HLB case example

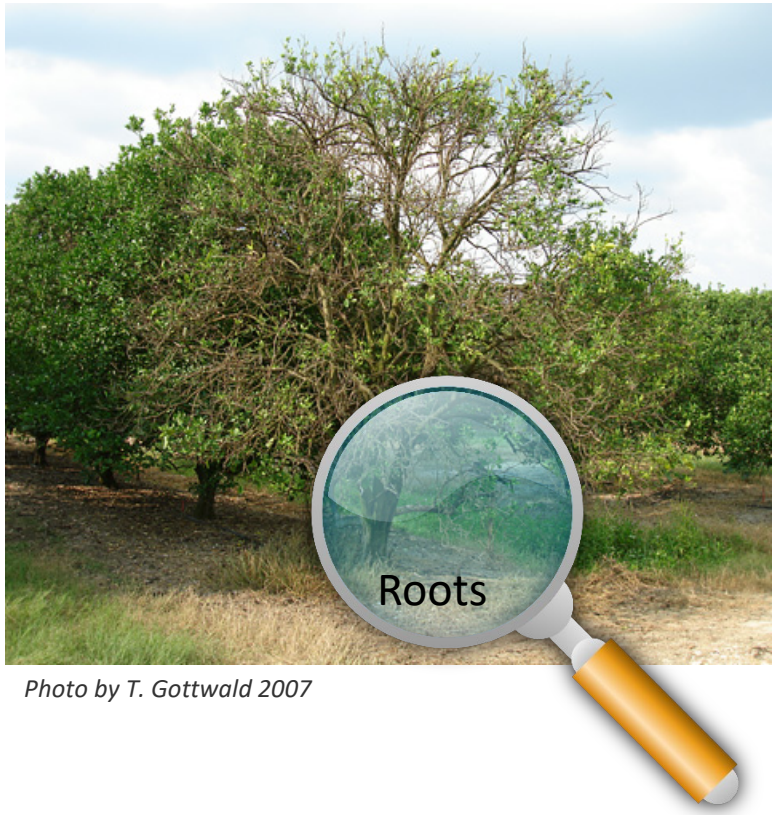
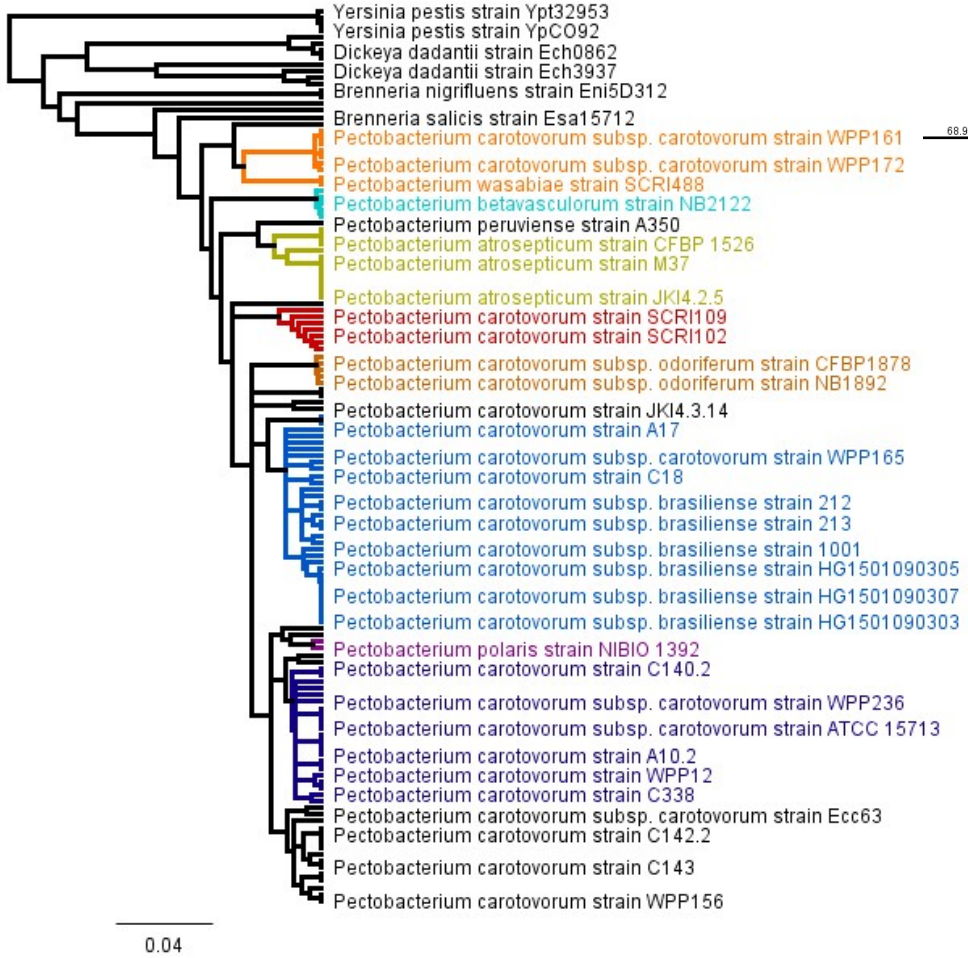
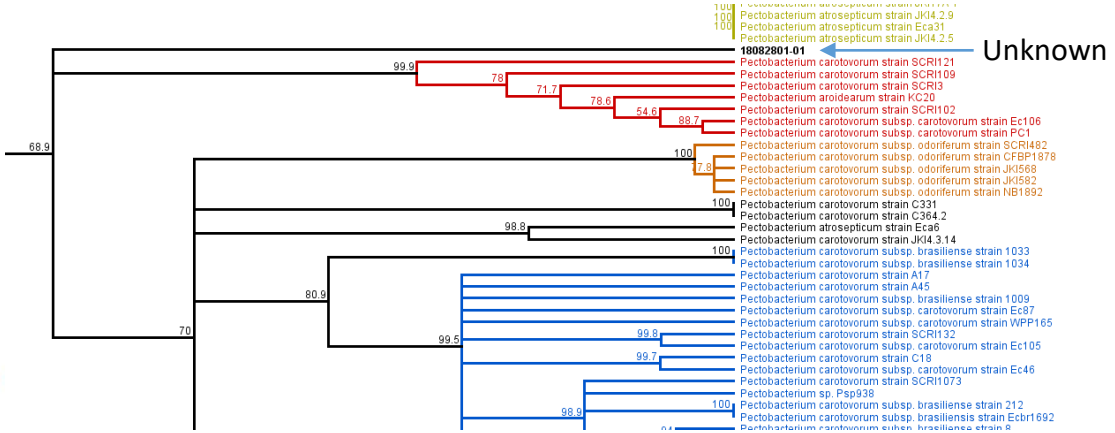


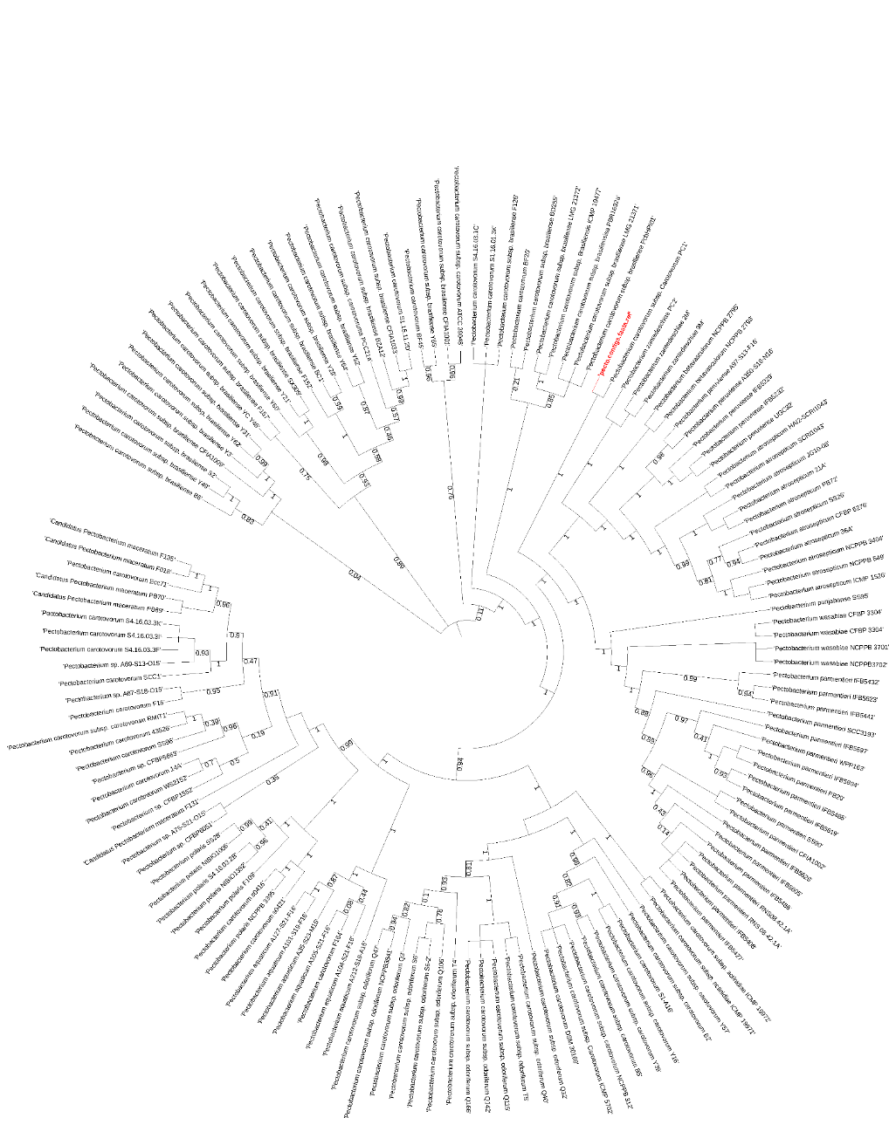
Photo by T. Gottwald 2007

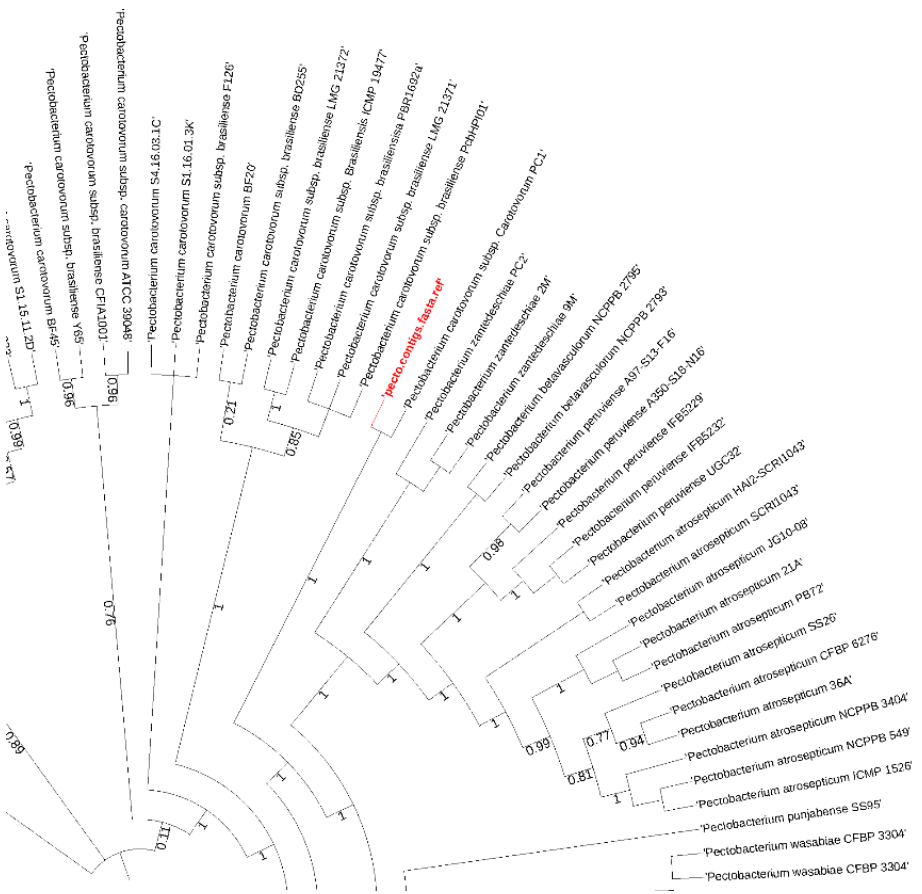
No standard pathogen analysis





Incomplete databases

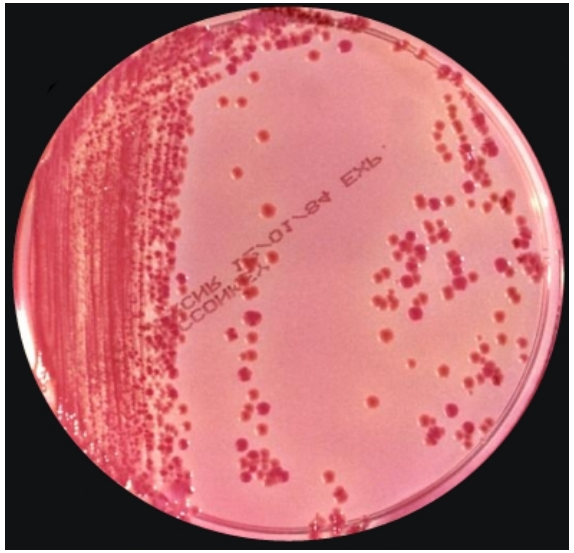




ANI to nearest neighbor is 91.5%

Pathogen biased databases

Jeffrey W. Lotz, Florida Department of Agriculture and Consumer Services, Bugwood.org



https://ppdictionary.com/bacteria/gnbac/pneumoniae.htm

Three bacteria isolated from mexican fruit fly

Name	Reference genome	Accession	Total reads	Mapped reads	Ratio of mapped reads over total
AG01-Klebsiella-spp-Silvestre-KPS	Klebsiella pneumoniae	CP014123.1	10,238,524	8,443,522	82.61%
AG01-Klebsiella-spp-Silvestre-KPS	Serratia marcescens UMH8, 7, 5	CP018927.1	10,238,524	2,510,532	25.09%
AG02-Klebsiella-oxytoca-planta-KOP	Klebsiella michiganensis M1	CP008841.1	7,136,680	6,155,236	86.47%
AG02-Klebsiella-oxytoca-planta-KOP	K. oxytoca KONIH5	CP026275.1	7,136,680	6,075,025	85.36%
AG03-Klebsiella-spp-Planta-KPP	Klebsiella pneumoniae	CP014123.1	7,786,294	5,538,373	71.74%
AG03-Klebsiella-spp-Planta-KPP	Enterobacter hormaechei	CP010384.1	7,786,294	2,695,441	35.56%

Low pathogen titer challenges

Sample	Reads	Reads mapped	Percentage of reads
Lib-1	5,746,156	5,322	0.07%
Lib-2	7,832,186	21,103	0.25%

Lib-1

Name	Consensus length	Total read count	Average coverage	Reference sequence	Reference length
dnaG mapping	28	51	0.48	dnaG	1947
gyrB mapping	15	1	6.29E-3	gyrB	2384
mutS mapping	17	3	0.02	mutS	2677
nusG mapping	0	0	0.00	nusG	524
rplA mapping	474	4	0.94	rplA	696
rpoB mapping	16	1	3.87E-3	rpoB	4139
tufB mapping	0	0	0.00	tufB	1178

Lib-2

Name	Consensus length	Total read count	Average coverage	Reference sequence	Reference length
dnaG mapping	39	37	0.34	dnaG	1947
gyrB mapping	0	0	0.00	gyrB	2384
mutS mapping	0	0	0.00	mutS	2677
nusG mapping	0	0	0.00	nusG	524
rplA mapping	0	0	0.00	rplA	696
rpoB mapping	15	1	3.62E-3	rpoB	4139
tufB mapping	41	3	0.05	tufB	1178

Challenges knowing when something is *not* there

Figure 1. The ZymoBIOMICS Spike-in Control I is added to samples and undergoes NGS-based microbial analysis.

Overview

- Massive parallel sequencing to improve confirmatory diagnostics
- Challenges of using genome sequencing for microbe identification
- Solutions through validation initiatives

Lack of reproducibility

Two-thirds of researchers fail to reproduce another scientist's experiments

- Study tried replicating five experiments: two were repeated, two were inconclusive, and one failed.
- A Nature survey reported over 70% of researchers have tried and failed to reproduce another scientist's experiments
- Scientific culture promotes impact over substance, flashy findings over dull, confirmatory work that most of science is about.
- There are important differences between replication and reproducibility

Feb 2017 - https://www.bbc.com/news/science-environment-39054778

USDA United States Department of Agriculture

Using statistics to boost assay performance

Verification and Validation of XX assay to detect XXX organism

Participants:

PI	
Team member	

Fitness for intended purpose

Highlight purpose.

Purpose	Tier validation	Report type
Assay conversion	N/A	Verification – Repeatability table with variable comparison
Specialist use	Tier 1	Verification – precision (no reproducibility), accuracy
Specialists, in-house diagnostician use	Tier 2	Validation – precision, accuracy, reproducibility within lab
Specialists, in-house diagnosticians, outside laboratories use	Tier 3	Validation – precision, accuracy, reproducibility ring test

Explanation of purpose (Introduction)

This should provide context about the nature of this assay and to what degree the report is demonstrating verification or validation. It should discuss whether this is a modification of an existing assay (such as switching a reagent or assay machine), application of an existing technology (hydrolysis probe based rtPCR), or implementation of an entirely new assay.

Can include minimum background explaining the need for the assay or modification.

Please refer to SOP-B-09.01 to review overall procedures. Particularly, section 5.5 and 5.6 cover what sections of this document are necessary for qualitative and real-time assays.

Materials

The material section should primarily be tables listing the information about the samples used, broken into separate tables based on purpose. For example a table listing pure cultures used, a separate table for diagnostic samples used, and yet another table with the analytes used for LOD testing. Please include simple references in the table below to find the material in this section.

Methods

The methods section should be chiefly concerned with the methods used to validate the assay, with a brief outline of the methods of the assay itself.

Precision

The degree of dispersion (such as variance, standard deviation, or coefficient of variation) within a series of measurements of the same sample tested under specified conditions.

Calculations

	Formula	Excel formula
Arithmetic Mean (μ)	$SUM(x_1, x_2 \ldots x_n)/n$	=AVERAGE()
Standard Deviation (σ)	$SQRT[SUM((x_1-\mu)^2+(x_2-\mu)^2+\ldots(x_n-\mu)^2)/n)]$	=STDEV()
Coefficient of Variation (CV)	$[\sigma/\mu]$	=STDEV()/AVERAGE()
95% Confidence interval (CI)	$\bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$	=confidence(0.05, STDEV, n)

Add additional tables for additional assay targets in Precision category.

Repeatability

Level of agreement between replicates of the same sample in the same exact conditions by the same operator, equipment, and reagents.

Table : Assay #1 targeting X.

Analyte	Quantity	Mean Reported Value	St Dev	N	%Positive
Analyte #1	*high*				
Analyte #1	*medium*				
Analyte #1	*low*				

Analyte = identifier. Try having 3-5 different analytes (samples, gBlock, cultures, etc).

Quantity = how much of assayed material is present (concentration, mass, high/low, gold-standard Ct value, etc.). Try having 3-5 different quantities that span the range (per analyte).

Reported value = the result the assay directly reports (e.g. Ct, optical density, absorbance). Average of assay replicates (try to replicate 3-5 times).

Technical replicates = repeats of same reaction mix run at the same time.

Assay replicates = repeat of the same biological material using different reaction mixtures. Can be an average of technical replicates.

Standard Deviation (St Dev) = the extent of deviation for a group as a whole.

N = number of times the assay was performed on the sample.

% Positive = the percentage of analytes which reported a positive result.

Linearity

Ability within a given range to obtain test results which are directly proportional to the concentration (amount) of analyte in the sample.

Extra lines are for multiple analytes and/or repeats, and also for different backgrounds (*e.g.* water, plant DNA).

Use appropriate table for single or multiplex assay and delete unused table.

Table for singleplex:

Assay (background)	R^2	Slope	Amplification

Table for multiplex:

Assay (background)	R^2	
	Single	M

Insert linearity plots

Intermediate precisi

Level of agreement be
the same lab. For exar
gent lots in any combi
other (for example, ea

Coefficient of variatio
values where the % p
dence interval (CI).

Levels of Variation: *e.g. Three testers;*	
Analyte	Quantity
Analyte #1	*high*
Analyte #1	*medium*
Analyte #1	*low*
Average % CV	

Reproducibility

Ability of a test metho
the same method in di

Testing Laboratory (Lab #1) Levels of Variation:						
Analyte	Quantity	Expected value	Mean Reported Value	Coefficient of Variation	N	%Positive
Average % CV						

Accuracy

Sensitivity (Analytical) = Limit of Detection (LOD)

Estimated lowest amount of an analyte in a specified matrix that can be reliably distinguished from absence of the analyte (a blank). Positive results are expressed as a percent and the probability of detection.

LOD = Lowest analyte quantity detected.

Suggested cutoff = Instrument value of LOD (± 2 Standard Deviations)

Matrix: *(e.g. water)*

Table: Assay limit of detection highlighted.

Analyte	Quantity	Assay #1 results

Specificity (Analytical)

Analytical specificity refers to the ability of an assay to identify a specific organism or analyte, rather than any other.

Table: Pure cultures tested and the result from the assay (or assays) with standard deviation or error.

Analyte	Assay #1 result

Selectivity

The capability to discriminate between the organism of interest and other organisms and components of the sample, such as host tissue.

Table :

Matrix	Analyte	Quantity	Assay result

Sensitivity (Diagnostic)

Proportion of known infected reference samples that test positive in the assay; infected plants that test negative are considered to have false-negative results.

Sensitivity = # true positives / (# true positives + false negatives) x 100

Calculate based on results using defined cutoffs.

Specificity (Diagnostic)

Proportion of known uninfected reference plants that test negative in the assay; uninfected reference plants that test positive are considered to have false-positive results.

Specificity = # of true negatives / (# true negatives + false positives) x 100

Calculate based on results using defined cutoffs.

Conclusion

Explain any anomalies in the validation data, justify assay modification

Appendix

Make note of location of raw data.

When submitting for review, please make raw data accessible for reviewers.

https://www.apsnet.org/edcenter/disimpactmngmnt/Pages/AssayValidationGlossary.aspx

End of document

Precision measurements

PRECISION: The degree of dispersion (such as variance, standard deviation or coefficient of variation) within a series of measurements.

- REPEATABILITY
- LINEARITY
- INTERMEDIATE PRECISION
- REPRODUCIBILITY

Accuracy measurements

ACCURACY: Assessment of nearness of a test value to the expected value. The expected value may be obtained from a known reference standard (plant pest, pathogen, or biomolecule associated with either), reagent of known activity, or well-documented titer.

- SENSITIVITY (ANALYTICAL) – LIMIT OF DETECTION
- SPECIFICITY (ANALYTICAL)
- SENSITIVITY (DIAGNOSTIC)
- SPECIFICITY (DIAGNOSTIC)
- SELECTIVITY

Precision measurement example

INTERMEDIATE PRECISION: Level of agreement between replicates of the same sample in similar conditions by the same lab (ICH Q2, 2005). For example, the sample is tested by analyst A and analyst B, or tested on instrument ABC and DEF, or tested using reagent lots UVW and XYZ on different days, in any combination (VIM, 2007).

Mixed Model analysis

Effect Summary

Source	LogWorth	PValue
Log CFU/reaction	168.588	0.00000
Platform	5.018	0.00001
Technician	0.978	0.10530

Summary and Needs

- PPQ S&T using genomics for identification and diagnostic improvement
- Remainder of PPQ does not have bioinformaticians at the ready
- Need information to perform risk assessment
  - Identification, pathogenicity, transmissibility, etc.

Acknowledgment

APHIS PPQ S&T Beltsville Laboratory

Discussion questions

- How do you measure microorganism risk?
- How do you prove a negative?
- How do you ensure protection of phytobiome diversity?