CLIA Compliant Validation of WGS in a Public Health Laboratory

Sara Blosser, PhD, D(ABMM)
Director of Clinical Microbiology
317-921-5894
sblosser@isdh.in.gov

Indiana State Department of Health

Outline

- Introduction
- ISDH validation:
  - Study strategy
  - Accuracy
  - Precision

Introduction

- NGS was introduced to Indiana public health in 2013 when PulseNet began using NGS to detect *Listeria* outbreaks.
- Now many CDC programs incorporate NGS, usually with their own analysis pipelines (eg., GHOST, *Legionella,* AMR).
- How do we bring in NGS on our own? How do we perform CLIA validation for this data?

Challenges to Adopting NGS

- **Bioinformatics:**
  - Infrastructure—We do not have a network or the computing power to perform full range of bioinformatics analyses.
  - Bioinformaticians—We do not have the financial ability to hire a full-time bioinformatician.
- **Validation cost** is relatively high.
- **Proficiency testing** for clinical and public health laboratories is still in development.

What Guidelines are Available?

- **NGS Standardization of Clinical Testing (Nex-StoCT) workgroup (CDC):**
  - Developed principles and guidelines for validation, QC, PT, and reference materials.
  - Workgroup focused on heritable human disorders, but may offer insight to NGS implementation in clinical and public health laboratories.

- **CAP Molecular Pathology Checklist**:
  - Includes wet and dry bench portions

- **CLSI MM09 (Nucleic Acid Sequencing Methods in Diagnostic Laboratory Medicine)**

Gargis AS, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice.

The ISDH CLIA-Compliant Validation

- ISDH goal: Validate the MiSeq and CLC Bio Genomic Workbench using the Nextera FLEX kit for outbreak analysis

- The ISDH validation was modeled after a paper written by the California Department of Public Health (Kozyreva *et al.)*.

Kozyreva, V. K., Truong, C. L., Greninger, A. L., Crandall, J., Mukhopadhyay, R., & Chaturvedi, V. (2017). Validation and Implementation of Clinical Laboratory Improvements Act-Compliant Whole-Genome Sequencing in the Public Health Microbiology Laboratory. *Journal of clinical microbiology*, *55*(8), 2502-2520.

How Did Our Validation Differ from California?

California

- 31 isolates
- In-house pipeline
- Additional analysis available

Indiana

- 10 isolates
- Commercial software
- Limited analysis

ISDH Validation Study Strategy

- Isolates were chosen based on frequency of receipt and likelihood of need.

- 10 total isolates, including:
  - 5 *Enterobacteriaceae* (CP-CREs)
  - 1 VPD (*H. influenzae)*
  - 1 Gram positive (*Staph aureus)*
  - 1 Actinomycete (*Nocardia* sp.)
  - 1 *Mycobacterium* (non-MTBC)
  - 1 'Enteric' (*Salmonella*)

ISDH Validation Study Strategy

- Ran technical triplicates on the same cartridge for each isolate (all 10 isolates):
  - Single extraction per isolate
  - Independent library preps
  - Single sequencing run

- Operator variance (5 isolates):
  - Used the same extract as the technical triplicates
  - Library prep performed by second analyst
  - Single sequencing run

What Did We Measure?

- **Accuracy:**
  - Platform
  - Application
  - Bioinformatics pipeline
- **Repeatability:**
  - Platform
  - Application
- **Reproducibility:**
  - Platform
  - Application

ACCURACY

Platform Accuracy: Definition

- "The accuracy of base calls against a reference sequence."
- The accuracy of the platform is established by determining the proximity of agreement between base calling made by the MiSeq sequencer (measured value) and NCBI reference sequence (gold standard).

Platform Accuracy: DIY

- Technical triplicates for each isolate were sequenced.
- Reads were mapped to the reference sequence and SNPs identified.
- SNPs which occur in all replicates are considered true - not a sequencing error.
- Calculation: % agreement with reference

- Due to computing limitations, raw sequence reads were analyzed by Kelsey Florek at Wisconsin State Laboratory of Hygiene.

Platform Accuracy: How did we do?

- 9 isolates had 100% agreement with the reference for all 3 technical replicates

- 1 isolate had 100% agreement for 1 technical replicate, and 99.99996429% and 99.99996429% agreement for the other 2

- **Conclusion:**

  Average % Agreement to Reference: 99.99%

Application Accuracy: Definition

- Agreement of the assay results for validation samples with the assay results for the gold standard.

Application Accuracy: DIY

- Assessed four applications:
  - 16S identification
  - Resistance typing
  - *in silico* MLST
  - hqSNP genotyping

- Only one replicate was used per strain for 16S and resistance typing.

Application Accuracy: How Did We Do?

16S Identification

Isolate Number	Gold Standard Identification	16S Identification
ARLN 0347	*Klebsiella pneumoniae*	*Klebsiella pneumoniae*
ARLN 0076	*Klebsiella pneumoniae*	*Klebsiella pneumoniae*
ARLN 0349	*E. coli*	*E. coli*
ATCC BAA 2146	*Klebsiella pneumoniae*	*Klebsiella pneumoniae*
ARLN 0056	*Acinetobacter baumannii*	*Acinetobacter baumannii*
ATCC 10211	*Haemophilus influenzae*	*Haemophilus influenzae*
ATCC 29213	*Staphylococcus aureus*	*Staphylococcus aureus*
18Myc0103	*Nocardia farcinica*	*Nocardia farcinica*
ATCC 25291	*Mycobacterium avium*	*Mycobacterium avium*
C17090866	*Salmonella enterica*	*Salmonella enterica*

Conclusion: All 10 sequenced isolates used for analysis with 100% agreement.

Resistance Typing

ARLN/ATCC Number	ARLN/ATCC Resistance Mechanisms	Detected Resistance Mechanisms
ARLN 0347	KPC-3	KPC-3
ARLN 0076	VIM-1	VIM-1
ARLN 0349	mcr-1, CTXM-14, CTXM-55, TEM-1B	mcr-1, CTXM-14, CTXM-55, TEM-1B
ATCC BAA 2146	NDM-1	NDM-1
ARLN 0056	OXA-23, OXA-66	OXA-23, OXA-66

Conclusion: Five ARLN/ATCC isolates (9 markers) used for analysis with 100% agreement for all markers.

in silico MLST

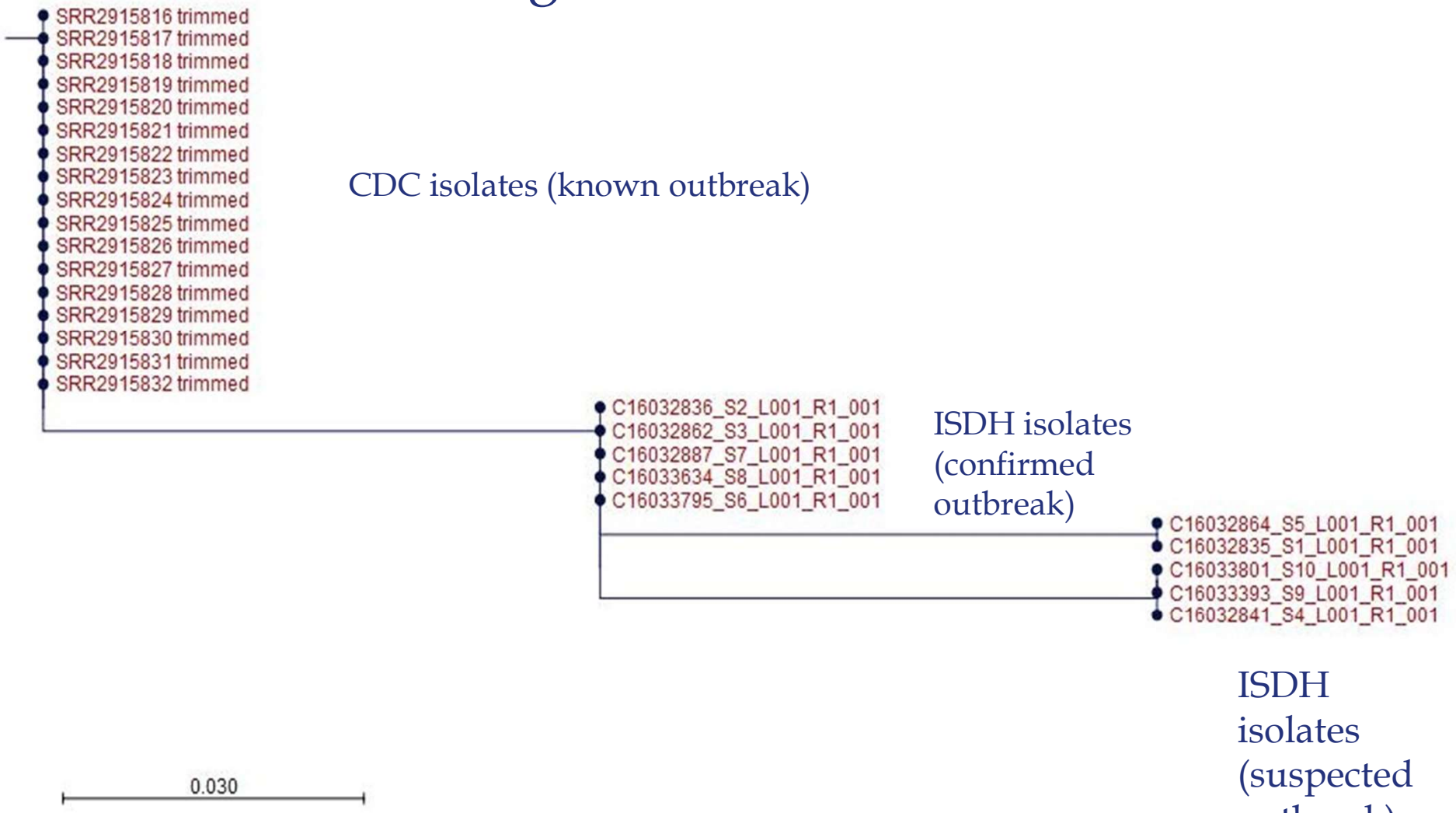
- ISDH does not conduct MLST testing, and therefore did not know the MLST schemes of the validation isolates.
- To determine the accuracy of the *in silico* MLST tool, sequences used in the Kozyreva *et al.* paper were pulled from NCBI.
- MLST tool was considered in agreement if results obtained from CLC Bio matched published scheme.
- **Conclusion:** 15 sequences were analyzed with 100% agreement.

hqSNP Genotyping

- Genotyping analysis performed in CLC Bio to determine if software was able to correctly cluster isolates of known relatedness.
- Ten (10) previously sequenced KPC positive ISDH isolates were analyzed alongside 17 KPC positive sequences provided by CDC as part of known outbreak.

Conclusion:

100% agreement between known clusters



CDC isolates (known outbreak)

ISDH isolates (confirmed outbreak)

ISDH isolates (suspected outbreak)

Pipeline Accuracy: Definition

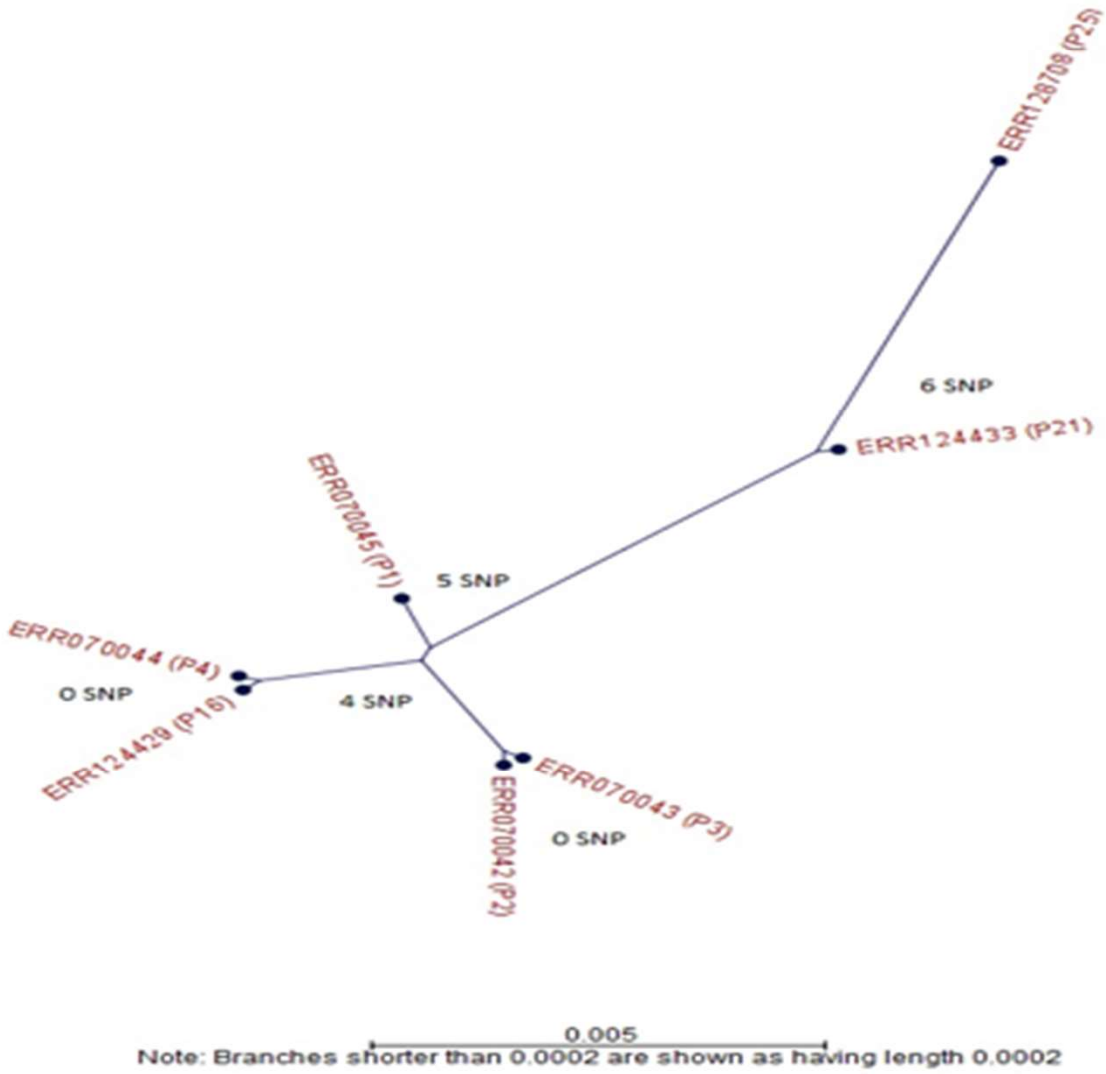
- Clustering suggested by previous investigators matches clustering achieved by ISDH pipeline.
- Inclusion of epidemiologically unrelated sequences should not cluster with the previous studies' findings.

Pipeline Accuracy: DIY

- Sequences were pulled from the literature and run though the ISDH pipeline.
  - Raw WGS reads of isolates from well-characterized MRSA outbreak (Harris, *et al.)*
- Demonstrate that we are able to reproduce published SNP analysis and phylogenetic trees.

S.R. Harris, *et. al.* 2013. Whole-genome sequencing for analysis of and outbreak of methicillin-resistant *Staphylococcus aureus:* a descriptive study.

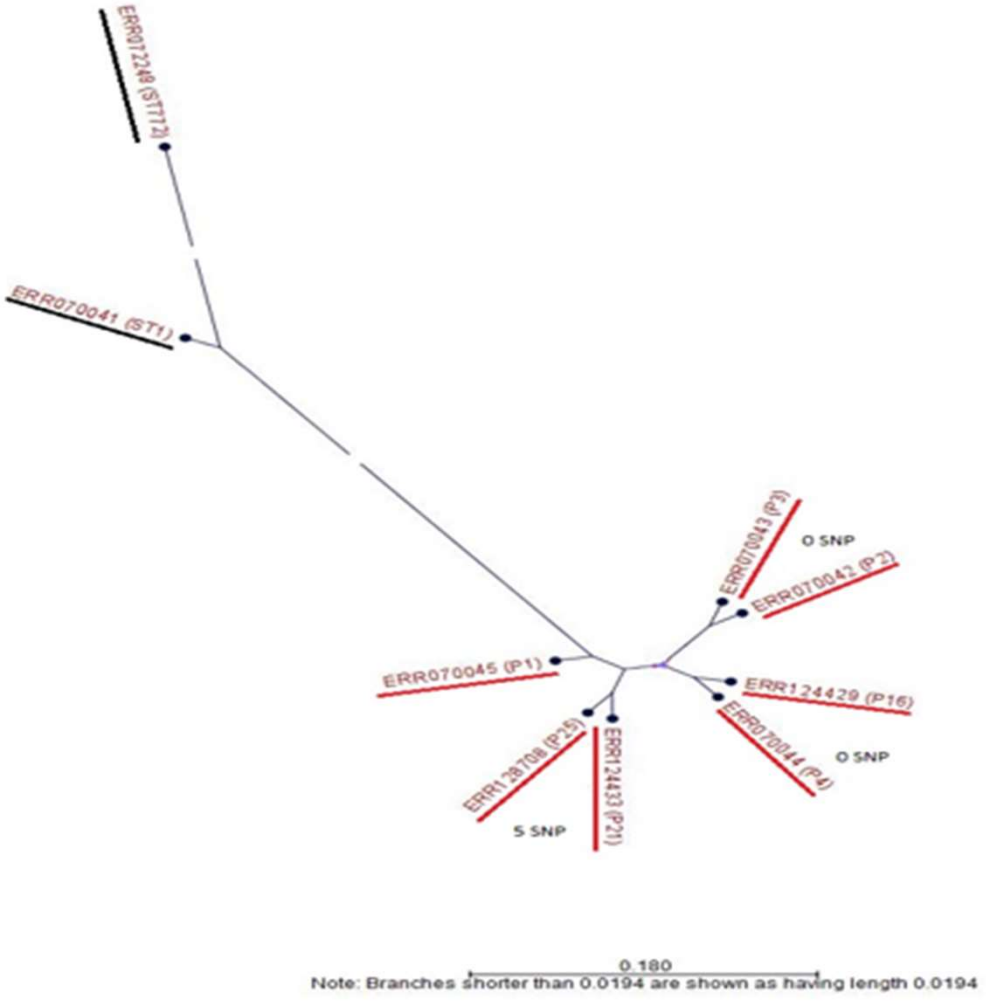
Phylogenetic Tree Created in CLC Bio



Clustering is reproducible with 100% agreement, although topology differs.

CLC Bio Tree with Unrelated Isolates

Epidemiologically unrelated isolates were added to analysis. These unrelated isolates did not cluster with outbreak isolates.



Pipeline Accuracy: How Did We Do?

	Paper	Validation	Match?
# of clusters	1 cluster, 2 unrelated isolates	1 cluster, 2 unrelated isolates	yes
# of isolates/ cluster	7 isolates	7 isolates	yes
# of SNPs within cluster	1-11	4-11	yes

Analytical Sensitivity and Specificity: Definition & DIY

- **True positive**—# of related sequences that cluster (CDC sequences)
- **True negative**—# of unrelated sequences that do not cluster with the related sequences (ISDH sequences)
- **False negative**—# of related sequences (CDC sequences) that cluster with unrelated sequences (ISDH sequences)
- **False positive**—# of unrelated sequences (ISDH sequences) that cluster with related sequences (CDC sequences)

Analytical Sensitivity and Specificity: How Did We Do?

- # TP: 17
- # TN: 10
- # FP: 0
- # FN: 0

- Overall sensitivity and specificity = 100%

PRECISION

Platform Repeatability: Definition

- "Within-run precision"
- Sequenced the same isolate multiple times under the same conditions (*i.e.* same cartridge)
- Performed SNP analysis on all within-run replicates for each validation isolate

Platform Repeatability: DIY

Isolate	Length of Covered Genome (bp)	Total # of SNP Difference for Within-run Replicates	Repeatability
ARLN 0056	3,645,732	0	100%
ARLN 0076	4,789,070	0	100%
ARLN 0349	4,506,303	2	99.99995562%
ARLN 0347	5,022,339	0	100%
C17030866	4,731,694	0	100%
BAA 2146	5,127,545	0	100%
ATCC 10211	1,796,677	0	100%
ATCC 29213	2,692,122	0	100%
ATCC 25291	4,658,151	2	99.9999571%
18Myc0103	5,363,168	5	99.99990677%

Platform Repeatability: How Did We Do?

- California study calculated repeatability per isolate/replicate – requiring 0 errors per millions of bp.
- By this calculation, ISDH repeatability was only 90%.
  - ARLN 0349 and ATCC 25291 each had 2 SNPs in 1 of 3 technical replicates
  - 18MYC103 had 5 SNPs in 1 of 3 technical replicates
  - So by the California calculations, only 27 of 30 runs were repeatable.

- ISDH calculated this as a rate per base pair: average repeatability per bp was 99.99998194%

- **Conclusion:** Platform repeatability per base pair is >99.9%

Platform Reproducibility: Definition

- "Between-run precision"
- Sequenced the same isolate under different conditions (*i.e.*, different cartridges, different library prep, different analyst)
- Performed SNP analysis on all replicates for 5 validation isolates

Platform Reproducibility: DIY

Isolate	Length of Covered Genome (bp)	Total # of SNP Difference for Between-run Replicates	Reproducibility
ARLN 0056	3,645,732	0	100%
ARLN 0076	4,789,070	0	100%
ARLN 0349	4,506,303	3	99.99993343%
ARLN 0347	5,022,339	0	100%
BAA 2146	5,127,545	0	100%

Platform Reproducibility: How Did We Do?

- Similar to Repeatability, California calculated this per isolate/replicate.
- By this method, ISDH had 95% reproducibility:
  - ARLN 0349 had a 3 SNP difference between runs

- ISDH calculated this as a rate per base pair: average repeatability per bp 99.99998668%

- **Conclusion:** Platform reproducibility per base pair is >99.9%

Application Repeatability and Reproducibility: Definition

- Difference between replicates of the same run (repeatability) and between runs (reproducibility) for the analysis component
  - Resistance typing
  - 16S identification
  - MLST
- All analysis tools showed 100% repeatability and reproducibility.

Repeatability: DIY Resistance Typing

ARLN/ATCC Number	ARLN/ATCC Resistance Mechanisms	Detected Resistance Mechanisms
ARLN 0347-1	KPC-3	KPC-3
ARLN 0347-2	KPC-3	KPC-3
ARLN 0347-3	KPC-3	KPC-3
ARLN 0076-1	VIM-1	VIM-1
ARLN 0076-2	VIM-1	VIM-1
ARLN 0076-3	VIM-1	VIM-1
ARLN 0349-1	mcr-1, CTXM-14, CTXM-55, TEM-1B	mcr-1, CTXM-14, CTXM-55, TEM-1B
ARLN 0349-2	mcr-1, CTXM-14, CTXM-55, TEM-1B	mcr-1, CTXM-14, CTXM-55, TEM-1B
ARLN 0349-3	mcr-1, CTXM-14, CTXM-55, TEM-1B	mcr-1, CTXM-14, CTXM-55, TEM-1B
ATCC BAA 2146-1	NDM-1	NDM-1
ATCC BAA 2146-2	NDM-1	NDM-1
ATCC BAA 2146-3	NDM-1	NDM-1
ARLN 0056-1	OXA-23, OXA-66	OXA-23, OXA-66
ARLN 0056-2	OXA-23, OXA-66	OXA-23, OXA-66
ARLN 0056-3	OXA-23, OXA-66	OXA-23, OXA-66

Reproducibility: DIY Resistance Typing

ARLN/ ATCC Number	Operator 1 Mechanisms Detected	Operator 2 Mechanisms Detected
ARLN 0347	KPC-3	KPC-3
ARLN 0076	VIM-1	VIM-1
ARLN 0349	mcr-1, CTXM-14, CTXM-55, TEM-1B	mcr-1, CTXM-14, CTXM-55, TEM-1B
ATCC BAA 2146	NDM-1	NDM-1
ARLN 0056	OXA-23, OXA-66	OXA-23, OXA-66

Repeatability: DIY 16S Identification

Organism	# of Attempts	# of Successful IDs
Klebsiella pneumoniae	9	9
Escherichia coli	3	3
Acinetobacter baumannii	3	3
Haemophilus influenzae	3	3
Staphylococcus aureus	3	3
Nocardia farcinica	3	3
Mycobacterium avium	3	3
Salmonella enterica	3	3

Reproducibility: DIY16S Identification

ARLN/ATCC Number	Operator 1 16S Identification	Operator 2 16S Identification
ARLN 0347	*Klebsiella pneumoniae*	*Klebsiella pneumoniae*
ARLN 0076	*Klebsiella pneumoniae*	*Klebsiella pneumoniae*
ARLN 0349	*E. coli*	*E. coli*
ATCC BAA 2146	*Klebsiella pneumoniae*	*Klebsiella pneumoniae*
ARLN 0056	*Acinetobacter baumannii*	*Acinetobacter baumannii*

Repeatability: DIY MLST

Organism	Sequence Type	# of Attempts	# of Successful IDs
Klebsiella pneumoniae	ST 11	6	6
	ST 86	3	3
Escherichia coli	ST 354	3	3
Acinetobacter baumannii	ST 423	3	3
Haemophilus influenzae	ST 6	3	3
Staphylococcus aureus	ST 4618	3	3
Salmonella enterica	ST 19	3	3

Reproducibility: DIY MLST

ARLN/ATCC Number	Operator 1 MLST	Operator 2 MLST
ARLN 0347 (*K. pneumoniae*)	ST 11	ST 11
ARLN 0076 (*K. pneumoniae*)	ST 86	ST 86
ARLN 0349 (*E. coli*)	ST 354	ST 354
ATCC BAA 2146 (*K. pneumoniae*)	ST 11	ST 11
ARLN 0056 (*A. baumanii*)	ST 423	ST 423

Summary of Assay Repeatability & Reproducibility

Type	# Attempted	# Met
Repeatability		
• Resistance	15	15
• 16S	30	30
• MLST	24	24
Reproducibility		
• Resistance	10	10
• 16S	10	10
• MLST	10	10

Questions?

Acknowledgements

- ISDH Project Lead: Cassie Campion
- Bioinformatics assistance: Kelsey Florek, Wisconsin State Laboratory of Hygiene
- KPC outbreak sequences: Tom de Man, CDC

Contact Info:
317-921-5894
sblosser@isdh.in.gov