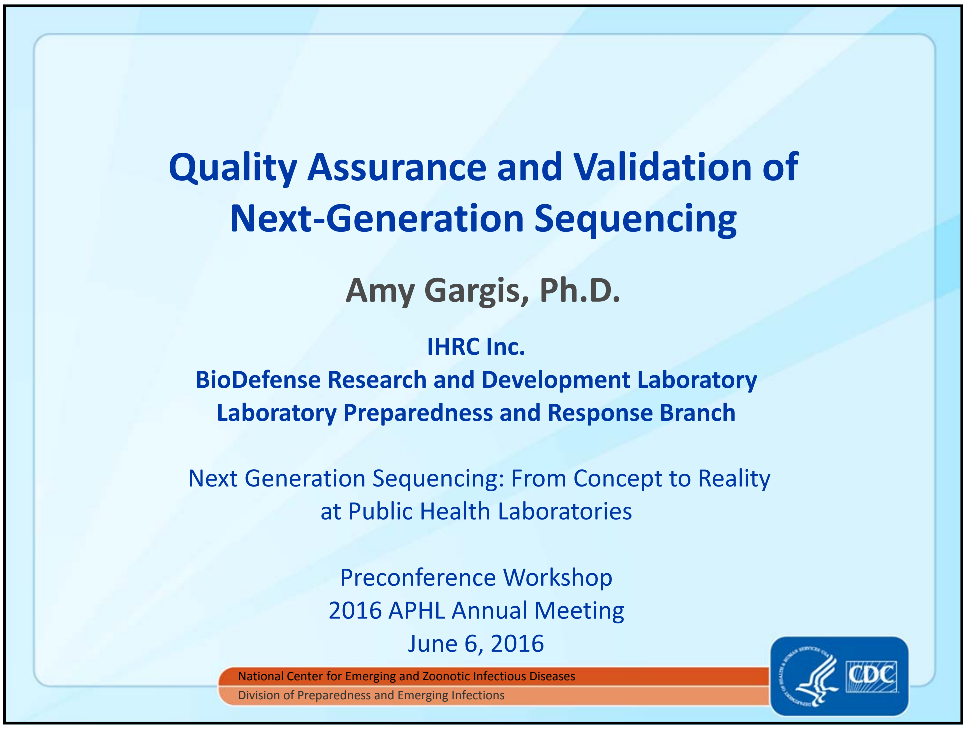


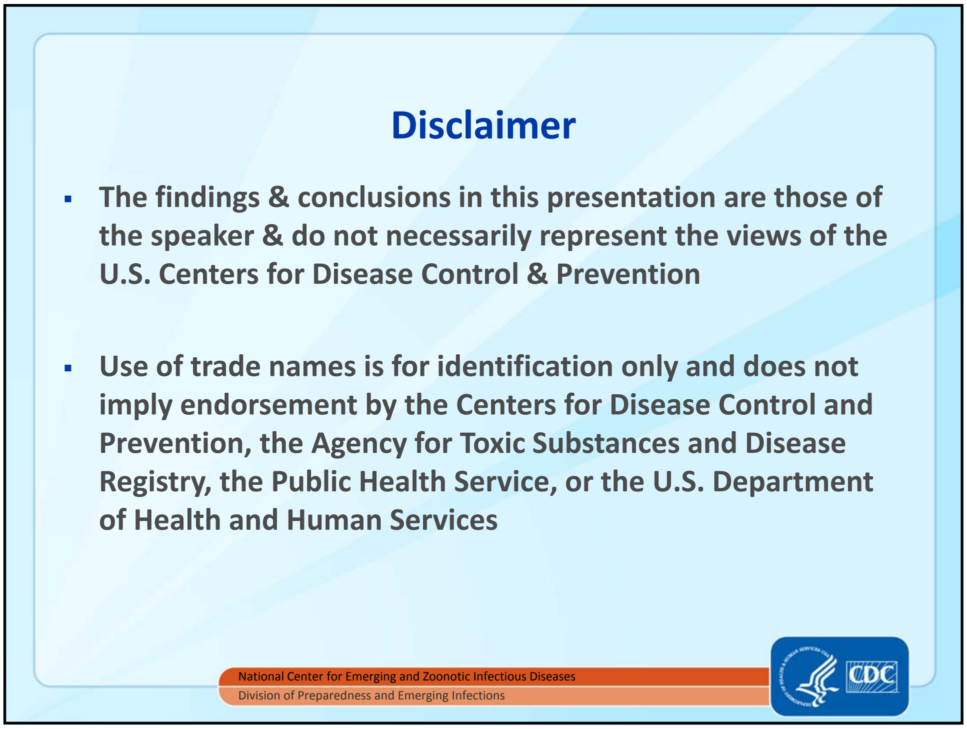
Quality Assurance and Validation of Next-Generation Sequencing

Amy Gargis, Ph.D.

IHRC Inc.
BioDefense Research and Development Laboratory
Laboratory Preparedness and Response Branch

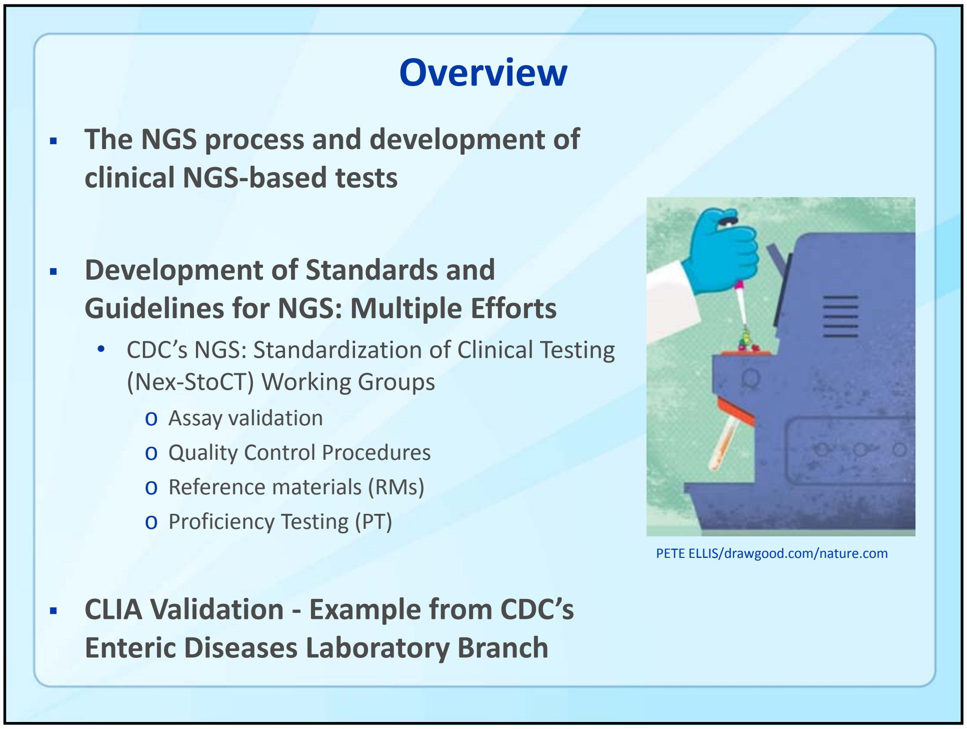
Next Generation Sequencing: From Concept to Reality
at Public Health Laboratories

Preconference Workshop
2016 APHL Annual Meeting
June 6, 2016



Disclaimer

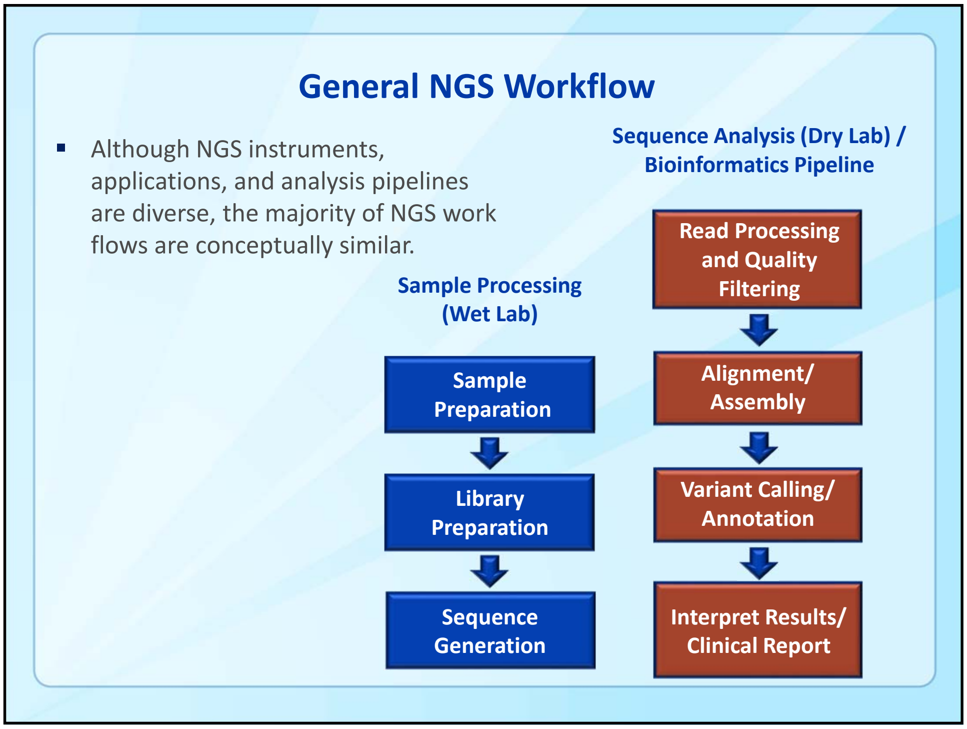
- The findings & conclusions in this presentation are those of the speaker & do not necessarily represent the views of the U.S. Centers for Disease Control & Prevention
- Use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention, the Agency for Toxic Substances and Disease Registry, the Public Health Service, or the U.S. Department of Health and Human Services



Overview

- The NGS process and development of clinical NGS-based tests
- Development of Standards and Guidelines for NGS: Multiple Efforts
  - CDC's NGS: Standardization of Clinical Testing (Nex-StoCT) Working Groups
    - Assay validation
    - Quality Control Procedures
    - Reference materials (RMs)
    - Proficiency Testing (PT)
- CLIA Validation - Example from CDC's Enteric Diseases Laboratory Branch

PETE ELLIS/drawgood.com/nature.com



General NGS Workflow

- Although NGS instruments, applications, and analysis pipelines are diverse, the majority of NGS work flows are conceptually similar.

Sample Processing (Wet Lab)
Sample Preparation
Library Preparation
Sequence Generation
Sequence Analysis (Dry Lab) / Bioinformatics Pipeline
Read Processing and Quality Filtering
Alignment/ Assembly
Variant Calling/ Annotation
Interpret Results/ Clinical Report

Development of Clinical NGS-based tests

- **The majority of clinical NGS tests are considered Laboratory Developed Tests (LDTs)**
  - *in vitro* diagnostic tests that are developed, manufactured by, and used within a single laboratory
- **Three instruments are FDA-cleared, but they can only be used for specific assays that were cleared by FDA*, otherwise the assay is an LDT**

FDA cleared For Cystic Fibrosis (2 Applications)*

Instrument / Universal Kit

Instrument

Instrument

- **In the US, LDTs are subject to the CLIA regulations, which require laboratories to establish analytical performance specifications of the assay (validation)**

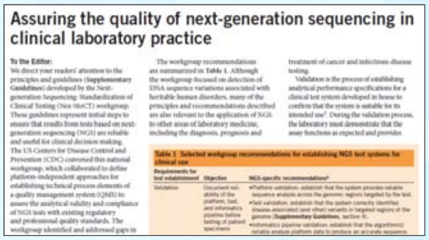
Center for Surveillance, Epidemiology and Laboratory Services
Division of Laboratory Systems (DLS)

- **Carries out CDC's responsibilities for the national CLIA* program; CMS and FDA also have CLIA responsibilities**
- **2010: DLS identified the need for standards and guidelines for clinical laboratory beginning to implement NGS**
- **CDC established two national workgroups:**
  - Next Generation Sequencing- Standardization of Clinical Testing (Nex-StoCT) Workgroups I and II
  - Develop a set of consensus principles and guidelines useful as a framework for implementing NGS into clinical settings

*Clinical Laboratory Improvement Amendments (CLIA), http://www.cdc.gov/ophss/csels/dls/clia.html

Next Generation Sequencing- Standardization of Clinical Testing (Nex-StoCT) Workgroups I and II

Recommendations Published



Assuring the quality of next-generation sequencing in clinical laboratory practice

To the Editor:
We direct your readers' attention to the principles and guidelines (**Supplementary Guidelines**) developed by the Next-generation Sequencing: Standardization of Clinical Testing (Nex-StoCT) workgroup. These guidelines represent initial steps to ensure that results from tests based on next-generation sequencing (NGS) are reliable and useful for clinical decision making. The US Centers for Disease Control and Prevention (CDC) convened this national workgroup, which collaborated to define platform-independent approaches for establishing technical process elements of a quality management system (QMS) to assure the analytical validity and compliance of NGS tests with existing regulatory and professional quality standards. The workgroup identified and addressed gaps in

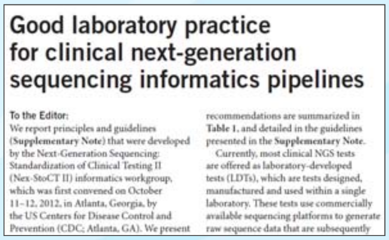
The workgroup recommendations are summarized in **Table 1**. Although the workgroup focused on detection of DNA sequence variations associated with heritable human disorders, many of the principles and recommendations described are also relevant to the application of NGS to other areas of laboratory medicine, including the diagnosis, prognosis and treatment of cancer and infectious-disease testing.

Validation is the process of establishing analytical performance specifications for a clinical test system developed in house to confirm that the system is suitable for its intended use[1]. During the validation process, the laboratory must demonstrate that the assay functions as expected and provides

Table 1 Selected workgroup recommendations for establishing NGS test systems for clinical use

Requirements for test establishment	Objective	NGS-specific recommendations[a]
Validation	Document reliability of the platform, test, and informatics pipeline before testing of patient specimens	•Platform validation: establish that the system provides reliable sequence analysis across the genomic regions targeted by the test. •Test validation: establish that the system correctly identifies disease-associated (and other) variants in targeted regions of the genome (**Supplementary Guidelines**, section 4). •Informatics pipeline validation: establish that the algorithm(s) reliably analyze platform data to produce an accurate sequence

Nat. Biotechnol. 2012; 30:1033-1036
+ Supplemental Guidelines



Good laboratory practice for clinical next-generation sequencing informatics pipelines

To the Editor:
We report principles and guidelines (**Supplementary Note**) that were developed by the Next-Generation Sequencing: Standardization of Clinical Testing II (Nex-StoCT II) informatics workgroup, which was first convened on October 11–12, 2012, in Atlanta, Georgia, by the US Centers for Disease Control and Prevention (CDC; Atlanta, GA). We present recommendations are summarized in **Table 1**, and detailed in the guidelines presented in the **Supplementary Note**.

Currently, most clinical NGS tests are offered as laboratory-developed tests (LDTs), which are tests designed, manufactured and used within a single laboratory. These tests use commercially available sequencing platforms to generate raw sequence data that are subsequently

Nat. Biotechnol. 2015;33: 689-693
+ Supplemental Guidelines

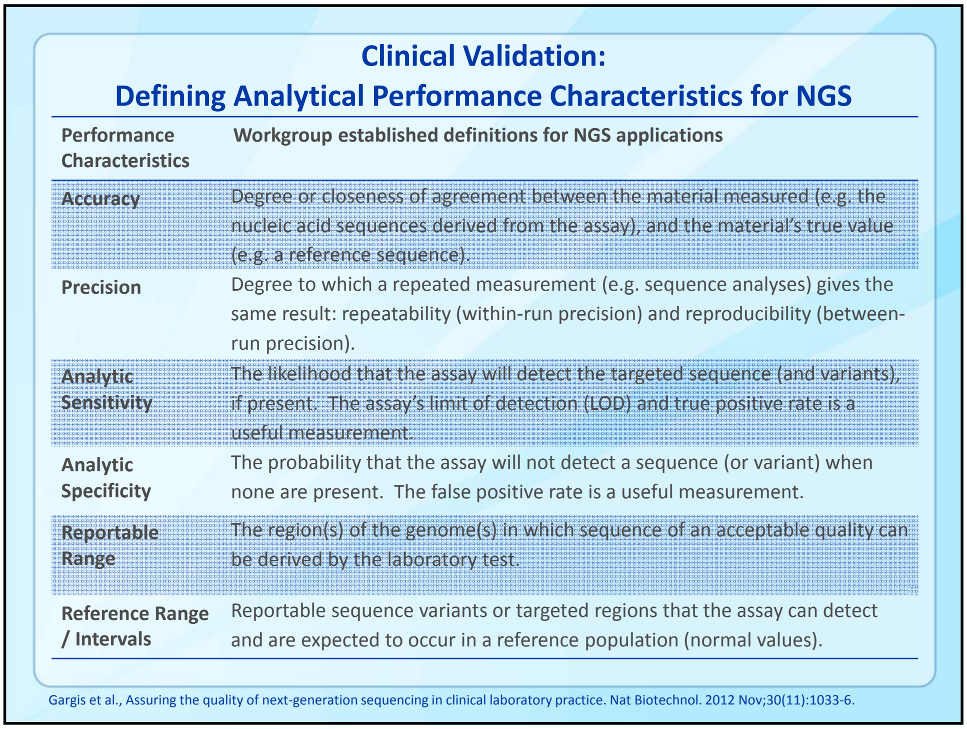
- **Focus of Nex-StoCT Guidelines:**
  - Test system validation, Quality control (QC), Reference materials (RMs), Proficiency testing(PT)/alternate assessment(AA), design and optimization of bioinformatics pipelines
  - Although these guidelines were developed for human genetic testing applications, many of the recommendations are applicable to clinical microbiology and public health NGS applications

Assay Validation

- **Regulatory requirements for clinical testing of human specimens are defined in the CLIA regulations***
  - Requires establishment of performance specifications to ensure the analytical validity of test results prior to patient testing

The definitions of performance characteristics described in CLIA do not readily translate to NGS due to the complexity and scale of the technology and data analyses

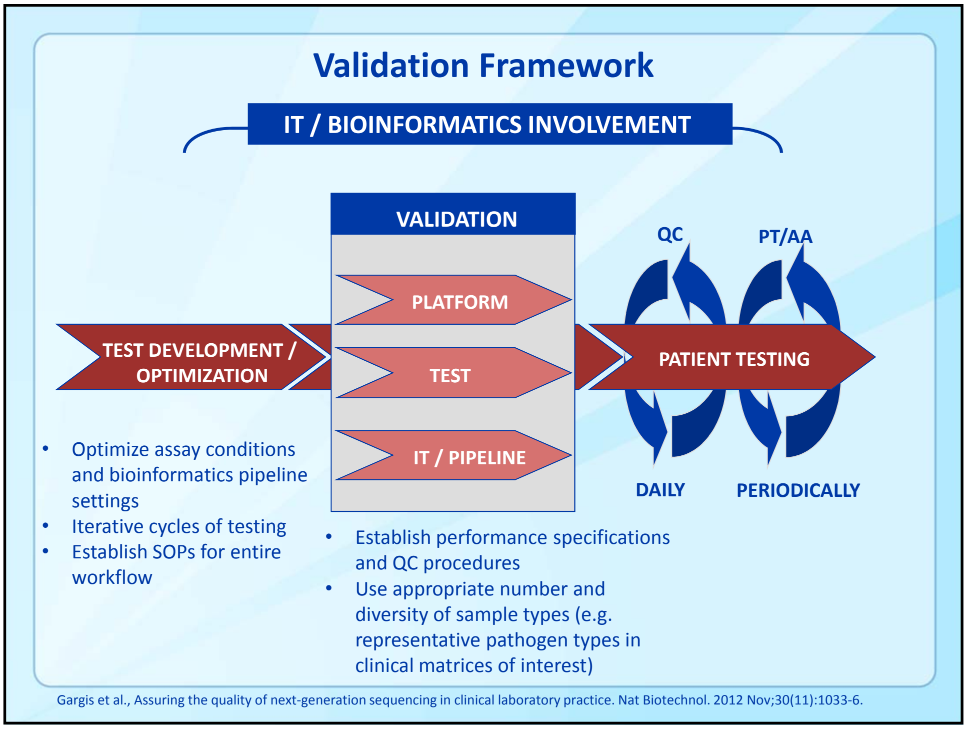
*Code of Federal Regulations. The Clinical Laboratory Improvement Amendments (CLIA). 42 CFR Part 493. (1256)



Clinical Validation: Defining Analytical Performance Characteristics for NGS

Performance Characteristics	Workgroup established definitions for NGS applications
Accuracy	Degree or closeness of agreement between the material measured (e.g. the nucleic acid sequences derived from the assay), and the material's true value (e.g. a reference sequence).
Precision	Degree to which a repeated measurement (e.g. sequence analyses) gives the same result: repeatability (within-run precision) and reproducibility (between-run precision).
Analytic Sensitivity	The likelihood that the assay will detect the targeted sequence (and variants), if present. The assay's limit of detection (LOD) and true positive rate is a useful measurement.
Analytic Specificity	The probability that the assay will not detect a sequence (or variant) when none are present. The false positive rate is a useful measurement.
Reportable Range	The region(s) of the genome(s) in which sequence of an acceptable quality can be derived by the laboratory test.
Reference Range / Intervals	Reportable sequence variants or targeted regions that the assay can detect and are expected to occur in a reference population (normal values).

Gargis et al., Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012 Nov;30(11):1033-6.



Validation Framework

IT / BIOINFORMATICS INVOLVEMENT

TEST DEVELOPMENT / OPTIMIZATION

VALIDATION

PLATFORM

TEST

IT / PIPELINE

PATIENT TESTING

QC

PT/AA

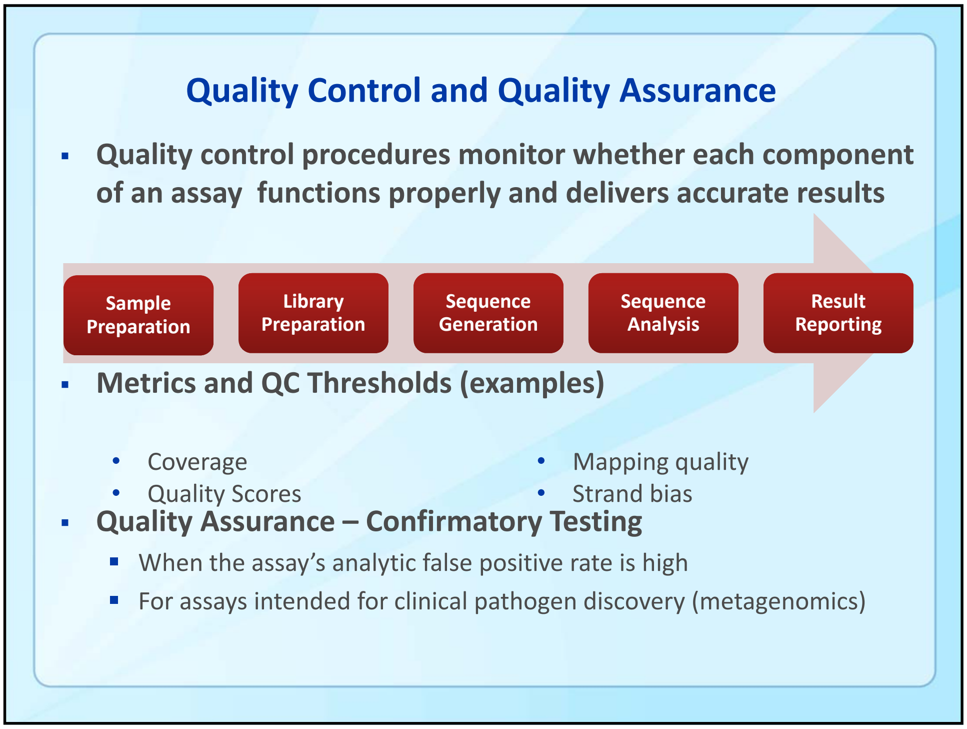
DAILY

PERIODICALLY

- Optimize assay conditions and bioinformatics pipeline settings
- Iterative cycles of testing
- Establish SOPs for entire workflow

- Establish performance specifications and QC procedures
- Use appropriate number and diversity of sample types (e.g. representative pathogen types in clinical matrices of interest)

Gargis et al., Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012 Nov;30(11):1033-6.

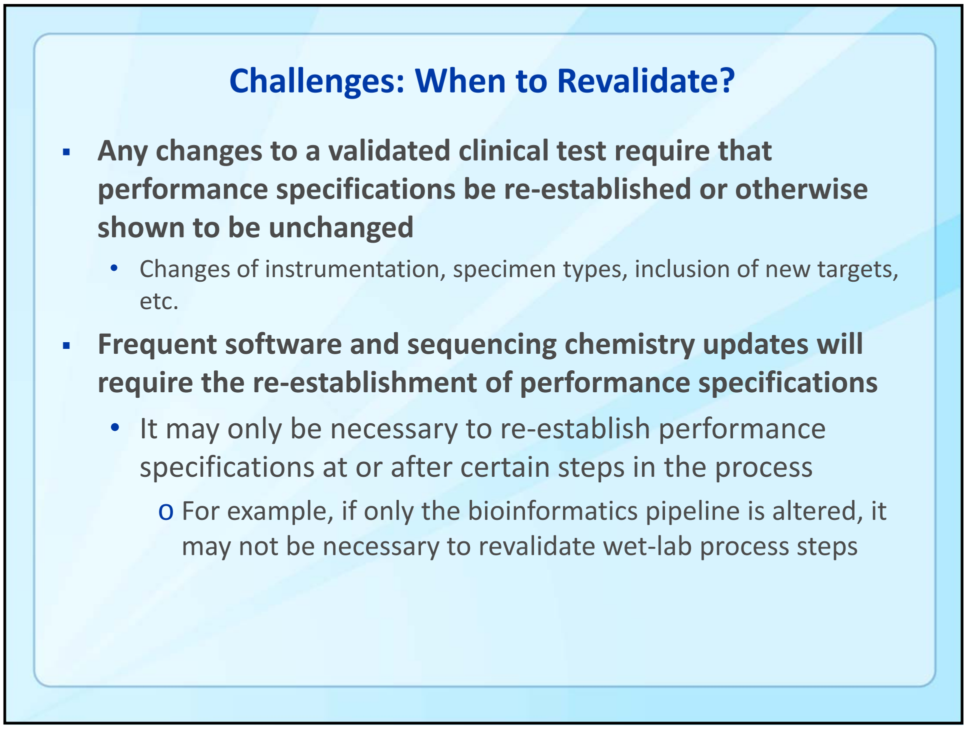


Quality Control and Quality Assurance

- Quality control procedures monitor whether each component of an assay functions properly and delivers accurate results

Sample Preparation
Library Preparation
Sequence Generation
Sequence Analysis
Result Reporting

- Metrics and QC Thresholds (examples)
  - Coverage
  - Quality Scores
  - Mapping quality
  - Strand bias
- Quality Assurance – Confirmatory Testing
  - When the assay's analytic false positive rate is high
  - For assays intended for clinical pathogen discovery (metagenomics)



Challenges: When to Revalidate?

- **Any changes to a validated clinical test require that performance specifications be re-established or otherwise shown to be unchanged**
  - Changes of instrumentation, specimen types, inclusion of new targets, etc.
- **Frequent software and sequencing chemistry updates will require the re-establishment of performance specifications**
  - It may only be necessary to re-establish performance specifications at or after certain steps in the process
    - For example, if only the bioinformatics pipeline is altered, it may not be necessary to revalidate wet-lab process steps

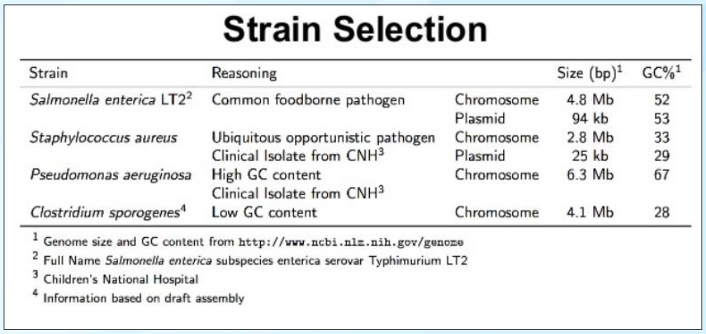
Reference Materials (RMs) useful for NGS

Used during test development, validation, for QC and PT to establish and monitor test quality

Type of Material	Considerations
Genomic DNA (from patient sample/ clinical isolate)	• Similar to patient's sample • Can be used as a reference for all phases of the testing process
Synthetic DNA	• Does not resemble patient sample • Can represent a broad range of sequences and variants
Electronic reference data files (e.g. curated benchmark datasets)	• Reference only for data analysis steps (not chemistry) • Data files may not be interoperable among different platforms

National Institute of Standards and Technology (NIST)

- ❑ Characterization of Bacterial Genomic RMs
- ❑ NIST plans to release the 4 strains as a single RM (anticipated release date, Fall 2016)



Strain Selection

Strain	Reasoning		Size (bp)[1]	GC%[1]
Salmonella enterica LT2[2]	Common foodborne pathogen	Chromosome	4.8 Mb	52
		Plasmid	94 kb	53
Staphylococcus aureus	Ubiquitous opportunistic pathogen	Chromosome	2.8 Mb	33
	Clinical Isolate from CNH[3]	Plasmid	25 kb	29
Pseudomonas aeruginosa	High GC content	Chromosome	6.3 Mb	67
	Clinical Isolate from CNH[3]			
Clostridium sporogenes[4]	Low GC content	Chromosome	4.1 Mb	28

[1] Genome size and GC content from http://www.ncbi.nlm.nih.gov/genome

[2] Full Name *Salmonella enterica* subspecies enterica serovar Typhimurium LT2

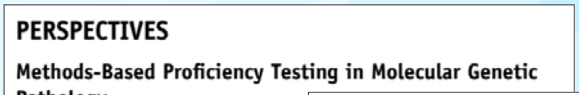
[3] Children's National Hospital

[4] Information based on draft assembly

N. D. Olson et al., Characterization of Bacterial Genomic Reference Materials. ASM 2015, Abstract: 1978

Proficiency Testing and Alternate Assessment

- **Clinical laboratories are required to demonstrate the independent assessment of test performance through PT/AA**
  - Use of methods-based evaluations of inter-laboratory performance rather than an analyte-specific PT
- **PT Programs**
  - College of American Pathologists – Methods- based PT
  - Global Microbial Identifier (GMI) Proficiency Test



PERSPECTIVES

Methods-Based Proficiency Testing in Molecular Genetic Pathology

Iris Schrijver,[*†] Nazneen Aziz,[‡] Lawrence J. Jennings,[§¶] Carolyn Su

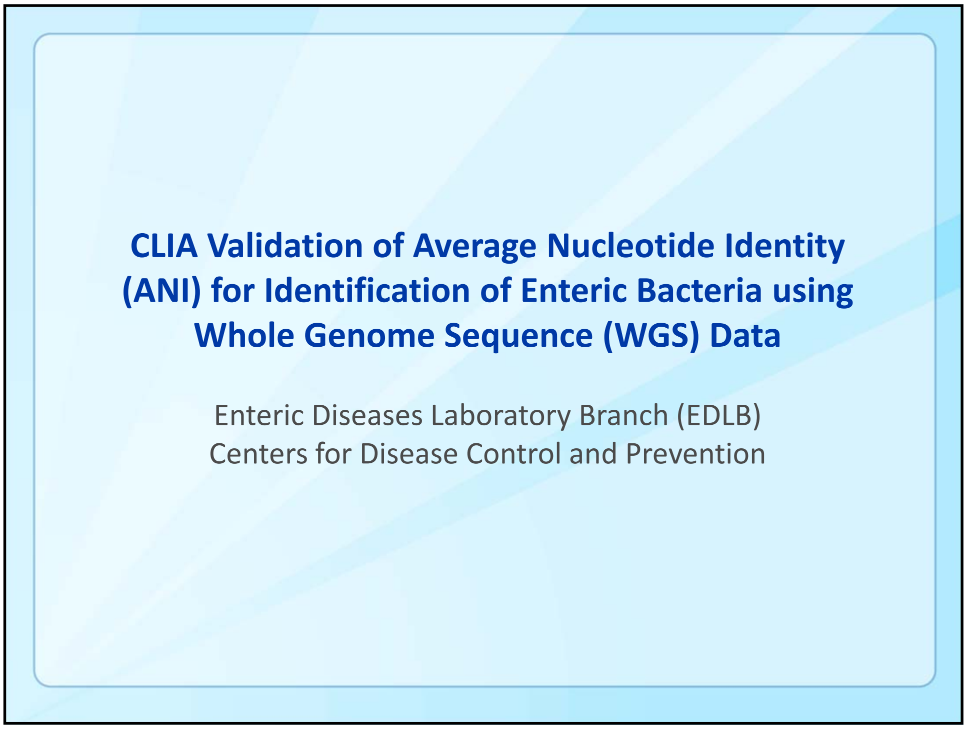
J Mol Diagn. 2014; 16:283-287

PROTOCOL for GMI Proficiency Test, 2015

http://www.globalmicrobialidentifier.org/

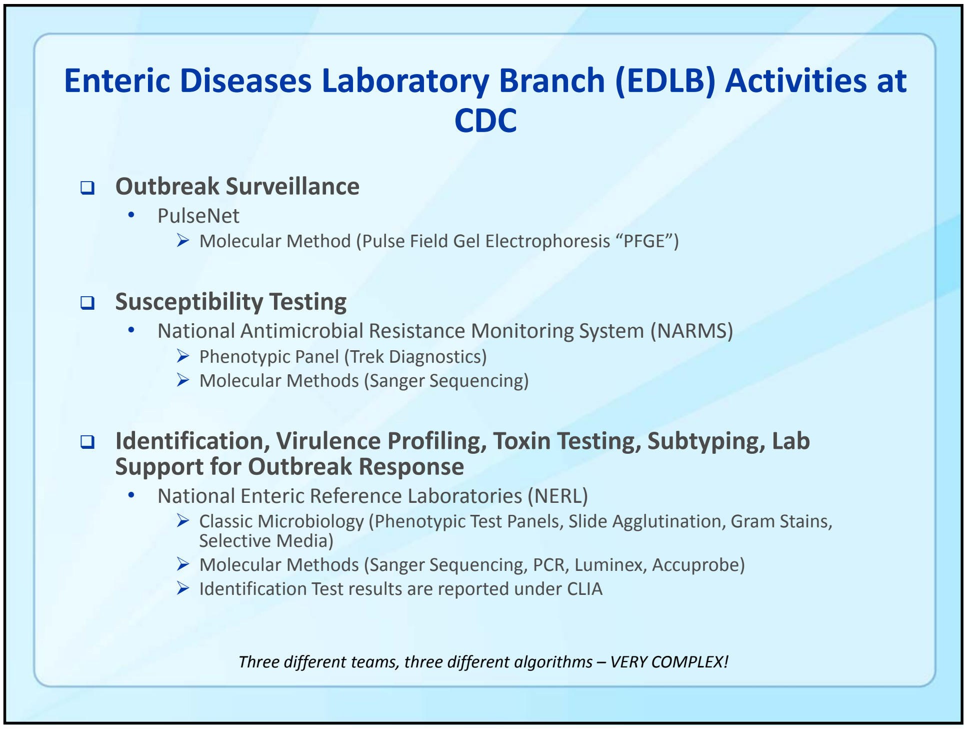
Guidance for Clinical NGS: Multiple Efforts

- **College of American Pathologists**
  - NGS Inspection Checklist (2012)
- **Clinical and Laboratory Standards Institute**
  - MM09 - Nucleic Acid Sequencing Methods in Diagnostic Laboratory Medicine (2014)
  - Includes Infectious Disease NGS applications
- **American Academy of Microbiology/ASM**
  - Colloquium on Applications of Clinical Microbial Next-Generation Sequencing (April, 2015)
  - Published a report with recommendations: includes considerations for QC and assay validation procedures and reference database needs (February 2016)
- **Food and Drug Administration**
  - Draft Guidance: Infectious Disease Next Generation Sequencing Based Diagnostic Devices: Microbial Identification and Detection of Antimicrobial Resistance and Virulence Markers (May 2016)



CLIA Validation of Average Nucleotide Identity (ANI) for Identification of Enteric Bacteria using Whole Genome Sequence (WGS) Data

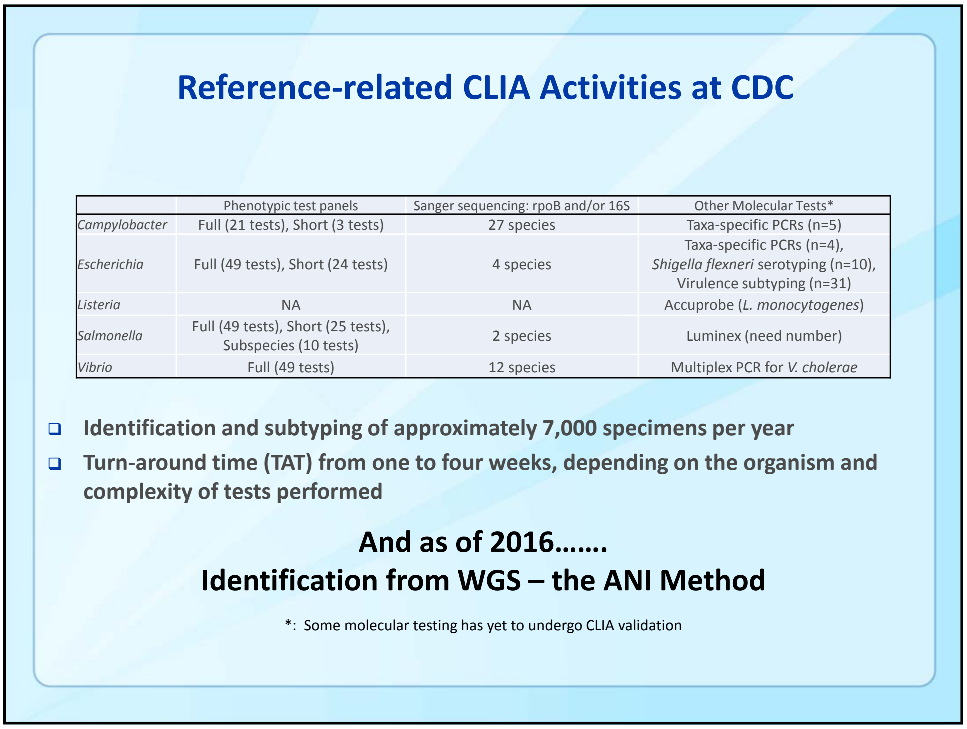
Enteric Diseases Laboratory Branch (EDLB)
Centers for Disease Control and Prevention



Enteric Diseases Laboratory Branch (EDLB) Activities at CDC

- **Outbreak Surveillance**
  - PulseNet
    - Molecular Method (Pulse Field Gel Electrophoresis “PFGE”)
- **Susceptibility Testing**
  - National Antimicrobial Resistance Monitoring System (NARMS)
    - Phenotypic Panel (Trek Diagnostics)
    - Molecular Methods (Sanger Sequencing)
- **Identification, Virulence Profiling, Toxin Testing, Subtyping, Lab Support for Outbreak Response**
  - National Enteric Reference Laboratories (NERL)
    - Classic Microbiology (Phenotypic Test Panels, Slide Agglutination, Gram Stains, Selective Media)
    - Molecular Methods (Sanger Sequencing, PCR, Luminex, Accuprobe)
    - Identification Test results are reported under CLIA

Three different teams, three different algorithms – VERY COMPLEX!



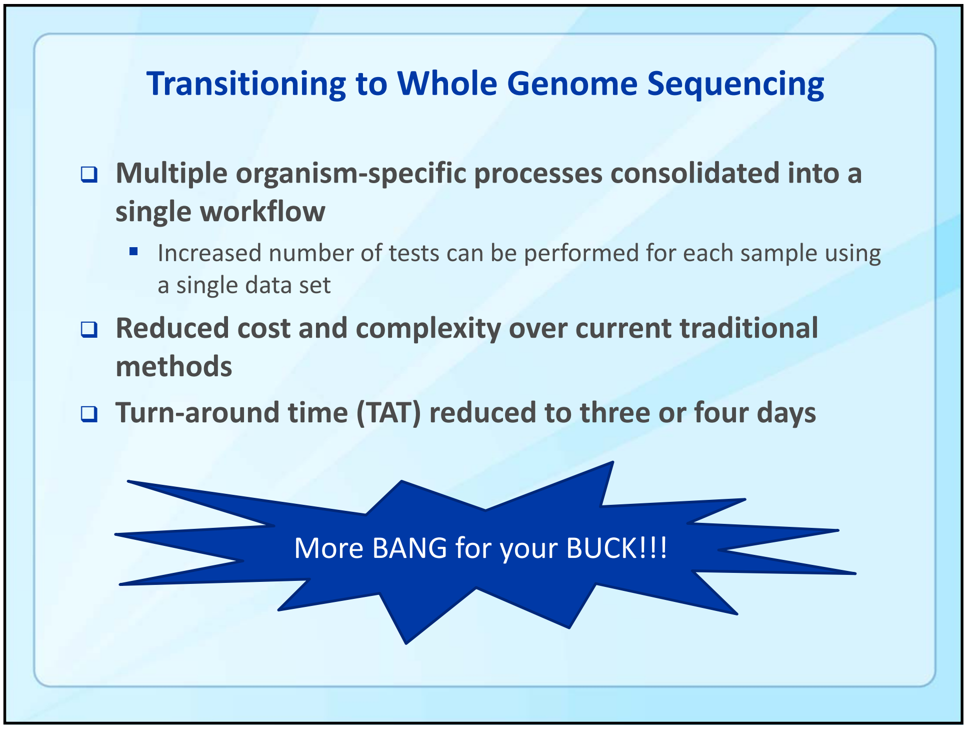
Reference-related CLIA Activities at CDC

	Phenotypic test panels	Sanger sequencing: rpoB and/or 16S	Other Molecular Tests*
Campylobacter	Full (21 tests), Short (3 tests)	27 species	Taxa-specific PCRs (n=5)
Escherichia	Full (49 tests), Short (24 tests)	4 species	Taxa-specific PCRs (n=4), *Shigella flexneri* serotyping (n=10), Virulence subtyping (n=31)
Listeria	NA	NA	Accuprobe (*L. monocytogenes*)
Salmonella	Full (49 tests), Short (25 tests), Subspecies (10 tests)	2 species	Luminex (need number)
Vibrio	Full (49 tests)	12 species	Multiplex PCR for *V. cholerae*

- ❑ **Identification and subtyping of approximately 7,000 specimens per year**
- ❑ **Turn-around time (TAT) from one to four weeks, depending on the organism and complexity of tests performed**

And as of 2016…….
Identification from WGS – the ANI Method

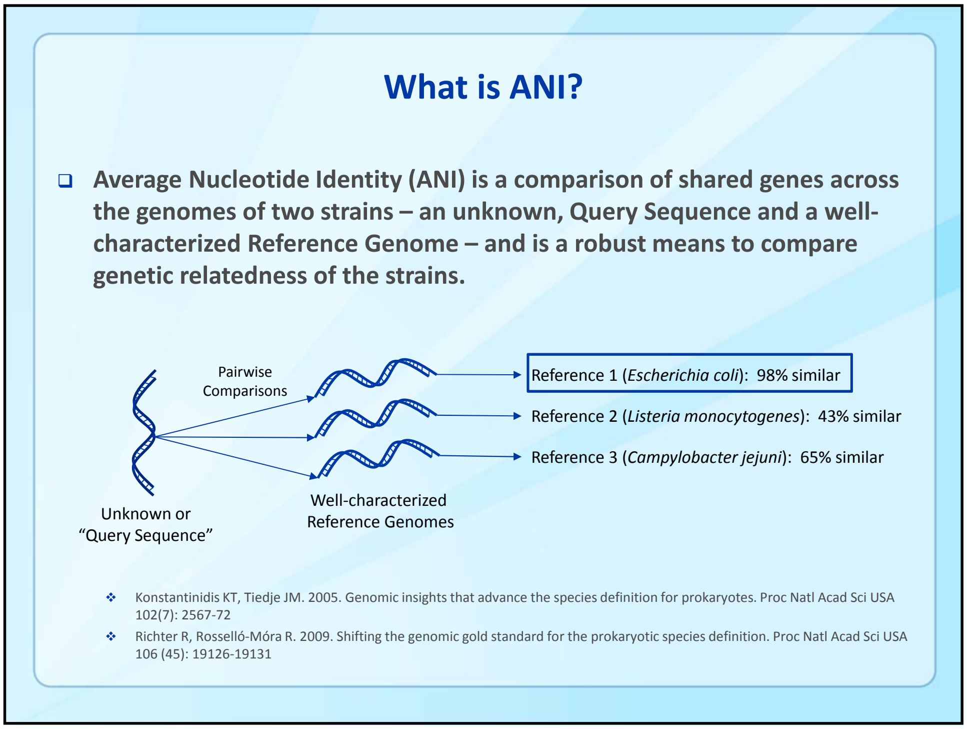
*: Some molecular testing has yet to undergo CLIA validation



Transitioning to Whole Genome Sequencing

- ❑ **Multiple organism-specific processes consolidated into a single workflow**
  - Increased number of tests can be performed for each sample using a single data set
- ❑ **Reduced cost and complexity over current traditional methods**
- ❑ **Turn-around time (TAT) reduced to three or four days**

More BANG for your BUCK!!!



What is ANI?

- Average Nucleotide Identity (ANI) is a comparison of shared genes across the genomes of two strains – an unknown, Query Sequence and a well-characterized Reference Genome – and is a robust means to compare genetic relatedness of the strains.

Pairwise Comparisons

Reference 1 (*Escherichia coli*): 98% similar

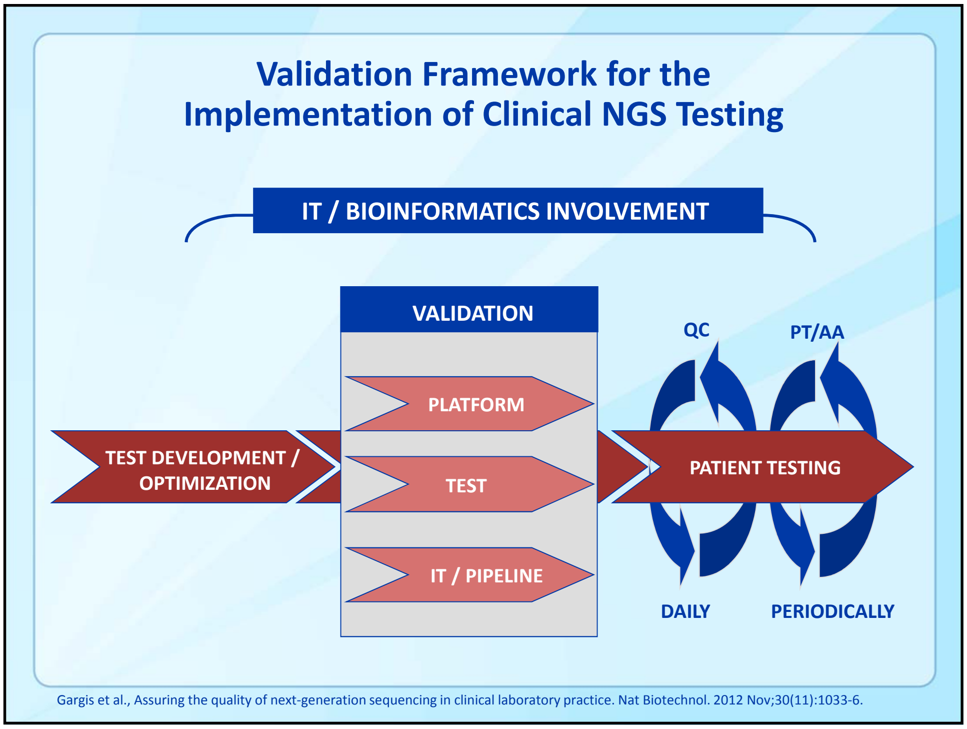
Reference 2 (*Listeria monocytogenes*): 43% similar

Reference 3 (*Campylobacter jejuni*): 65% similar

Unknown or "Query Sequence"

Well-characterized Reference Genomes

- Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA 102(7): 2567-72
- Richter R, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA 106 (45): 19126-19131



Validation Framework for the Implementation of Clinical NGS Testing

IT / BIOINFORMATICS INVOLVEMENT

TEST DEVELOPMENT / OPTIMIZATION

VALIDATION

PLATFORM

TEST

IT / PIPELINE

PATIENT TESTING

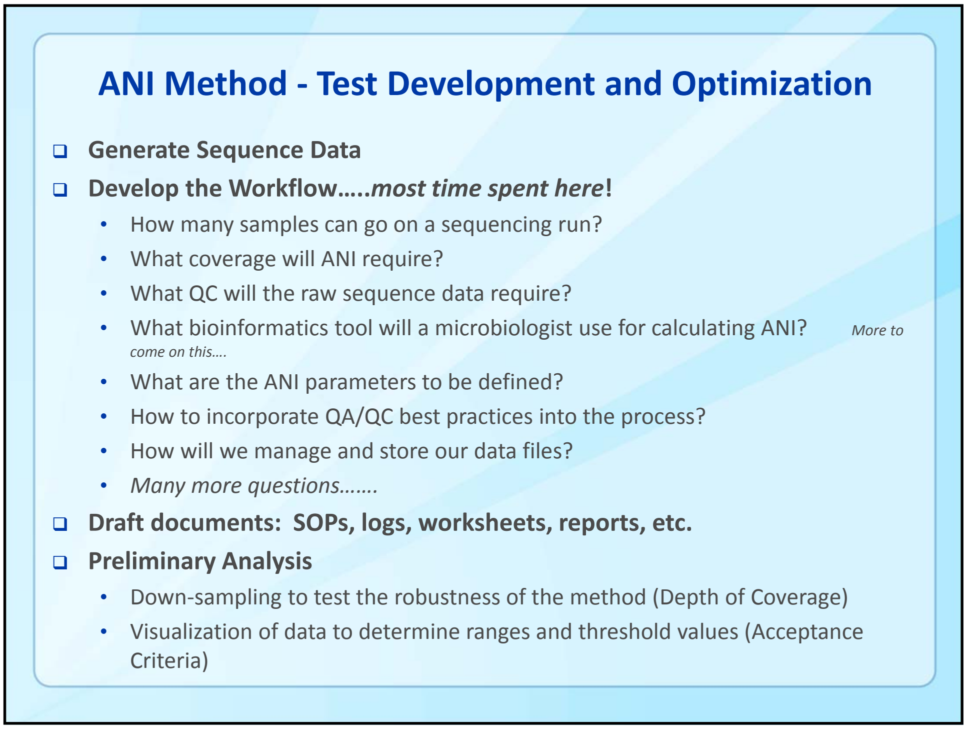
QC

PT/AA

DAILY

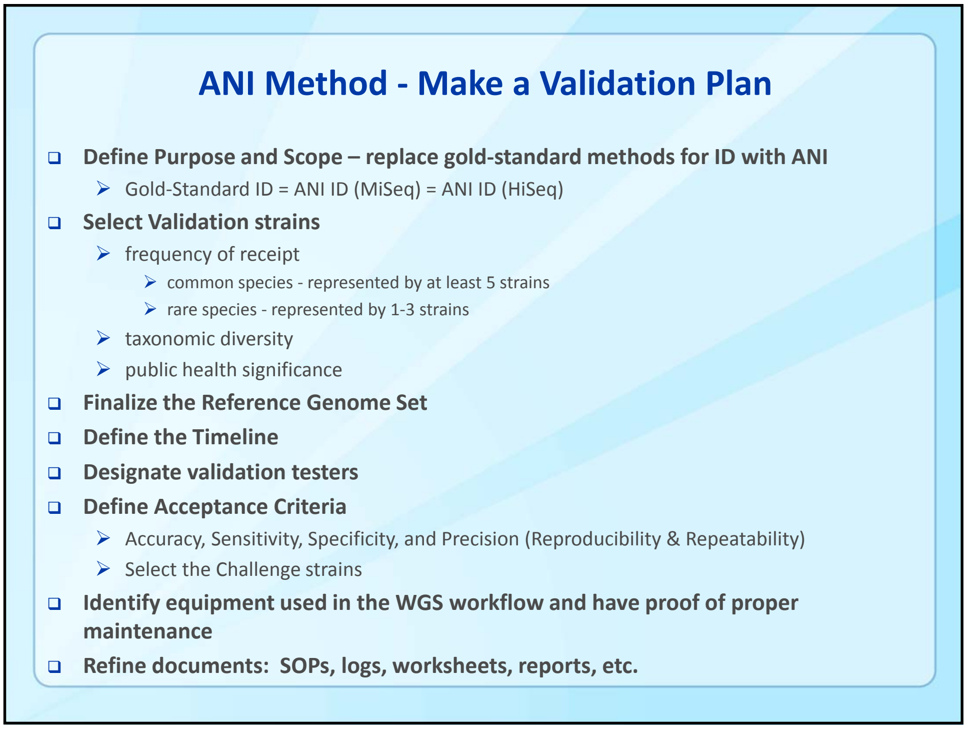
PERIODICALLY

Gargis et al., Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012 Nov;30(11):1033-6.



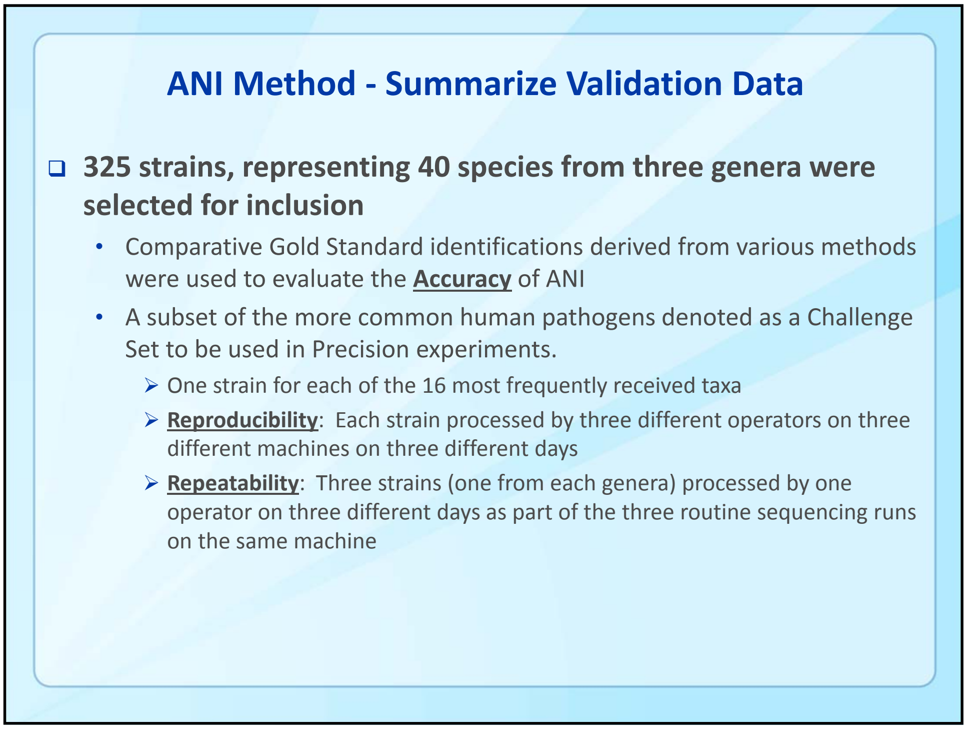
ANI Method - Test Development and Optimization

- ❑ **Generate Sequence Data**
- ❑ **Develop the Workflow…..*most time spent here*!**
  - How many samples can go on a sequencing run?
  - What coverage will ANI require?
  - What QC will the raw sequence data require?
  - What bioinformatics tool will a microbiologist use for calculating ANI? *More to come on this….*
  - What are the ANI parameters to be defined?
  - How to incorporate QA/QC best practices into the process?
  - How will we manage and store our data files?
  - *Many more questions…….*
- ❑ **Draft documents: SOPs, logs, worksheets, reports, etc.**
- ❑ **Preliminary Analysis**
  - Down-sampling to test the robustness of the method (Depth of Coverage)
  - Visualization of data to determine ranges and threshold values (Acceptance Criteria)



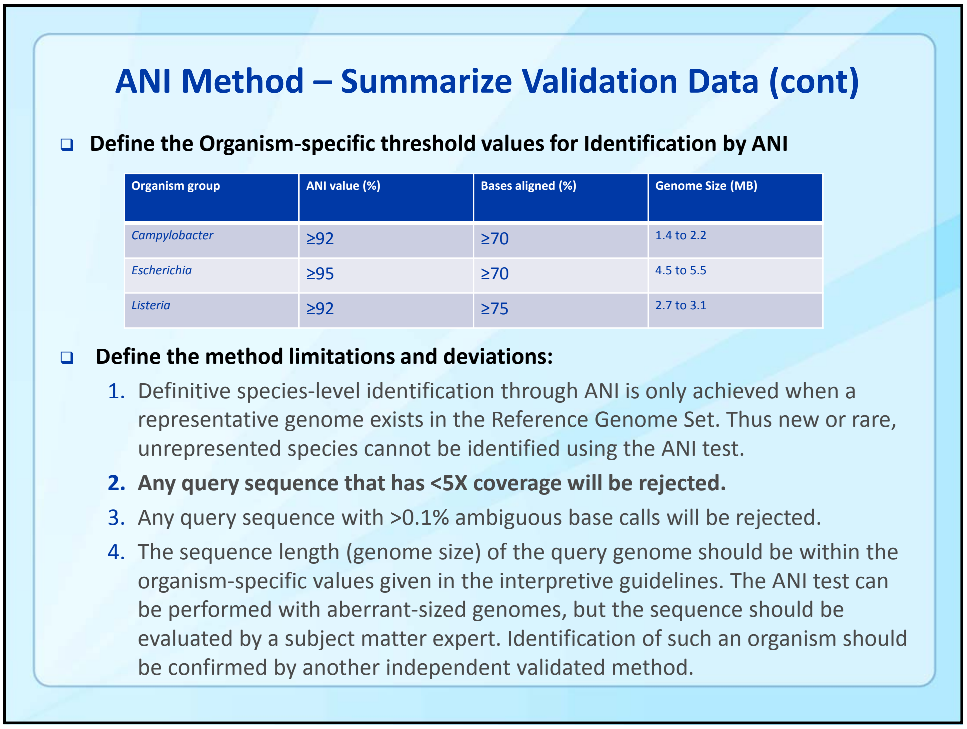
ANI Method - Make a Validation Plan

- ❑ **Define Purpose and Scope – replace gold-standard methods for ID with ANI**
  - ➢ Gold-Standard ID = ANI ID (MiSeq) = ANI ID (HiSeq)
- ❑ **Select Validation strains**
  - ➢ frequency of receipt
    - ➢ common species - represented by at least 5 strains
    - ➢ rare species - represented by 1-3 strains
  - ➢ taxonomic diversity
  - ➢ public health significance
- ❑ **Finalize the Reference Genome Set**
- ❑ **Define the Timeline**
- ❑ **Designate validation testers**
- ❑ **Define Acceptance Criteria**
  - ➢ Accuracy, Sensitivity, Specificity, and Precision (Reproducibility & Repeatability)
  - ➢ Select the Challenge strains
- ❑ **Identify equipment used in the WGS workflow and have proof of proper maintenance**
- ❑ **Refine documents: SOPs, logs, worksheets, reports, etc.**



ANI Method - Summarize Validation Data

- **325 strains, representing 40 species from three genera were selected for inclusion**
  - Comparative Gold Standard identifications derived from various methods were used to evaluate the **Accuracy** of ANI
  - A subset of the more common human pathogens denoted as a Challenge Set to be used in Precision experiments.
    - One strain for each of the 16 most frequently received taxa
    - **Reproducibility**: Each strain processed by three different operators on three different machines on three different days
    - **Repeatability**: Three strains (one from each genera) processed by one operator on three different days as part of the three routine sequencing runs on the same machine

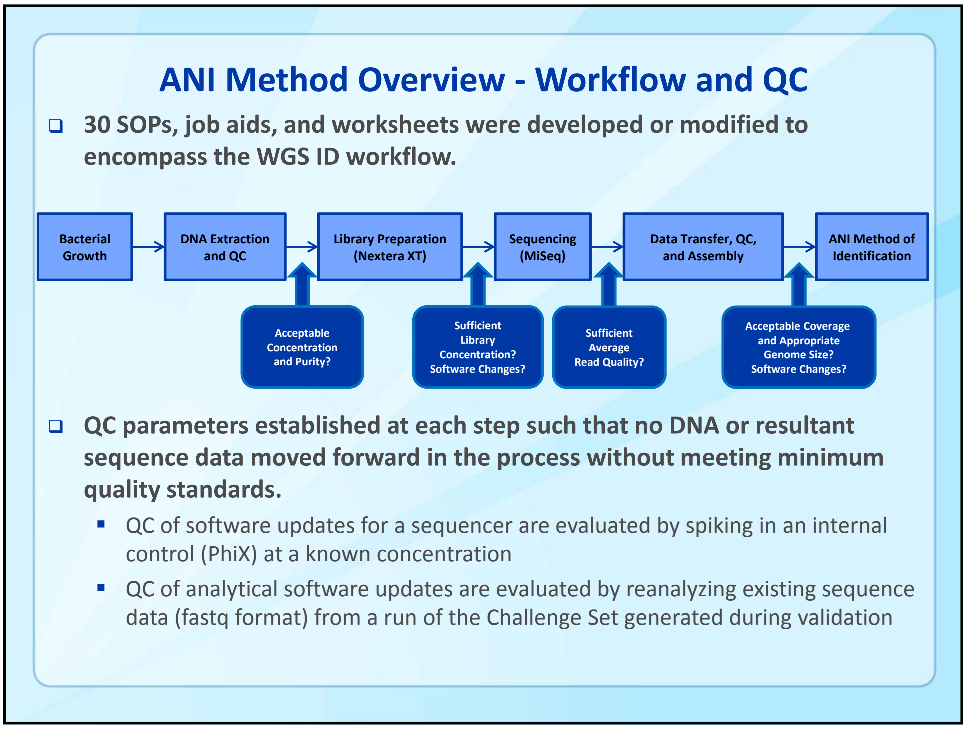


ANI Method – Summarize Validation Data (cont)

- **Define the Organism-specific threshold values for Identification by ANI**

Organism group	ANI value (%)	Bases aligned (%)	Genome Size (MB)
Campylobacter	≥92	≥70	1.4 to 2.2
Escherichia	≥95	≥70	4.5 to 5.5
Listeria	≥92	≥75	2.7 to 3.1

- **Define the method limitations and deviations:**
  1. Definitive species-level identification through ANI is only achieved when a representative genome exists in the Reference Genome Set. Thus new or rare, unrepresented species cannot be identified using the ANI test.
  2. **Any query sequence that has <5X coverage will be rejected.**
  3. Any query sequence with >0.1% ambiguous base calls will be rejected.
  4. The sequence length (genome size) of the query genome should be within the organism-specific values given in the interpretive guidelines. The ANI test can be performed with aberrant-sized genomes, but the sequence should be evaluated by a subject matter expert. Identification of such an organism should be confirmed by another independent validated method.



ANI Method Overview - Workflow and QC

- 30 SOPs, job aids, and worksheets were developed or modified to encompass the WGS ID workflow.

Bacterial Growth → DNA Extraction and QC → Library Preparation (Nextera XT) → Sequencing (MiSeq) → Data Transfer, QC, and Assembly → ANI Method of Identification

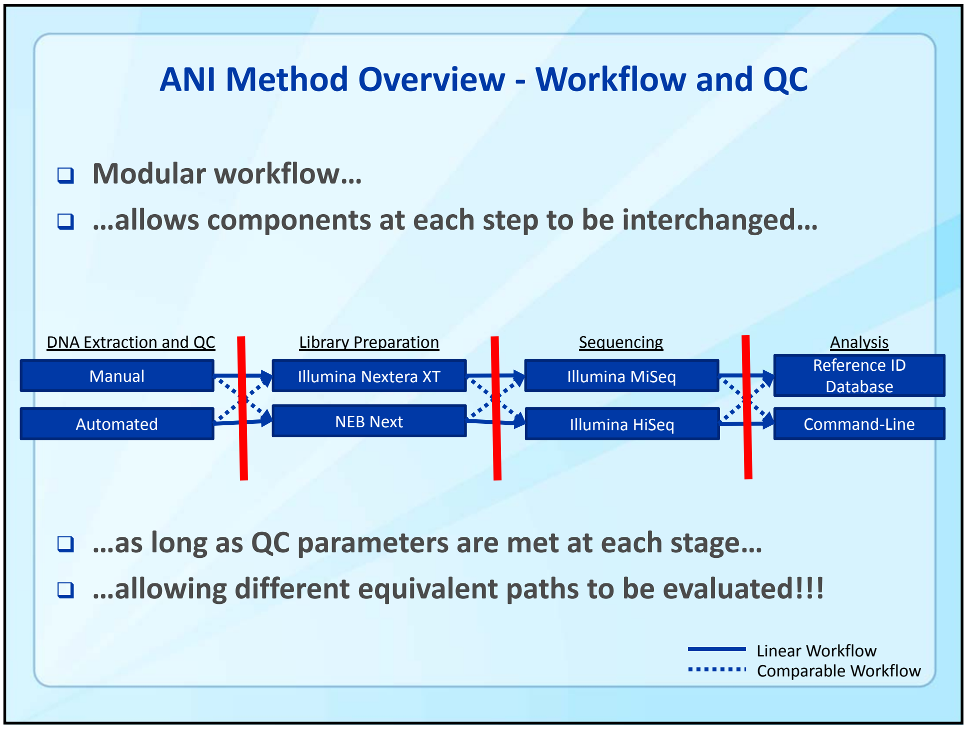
Acceptable Concentration and Purity?

Sufficient Library Concentration? Software Changes?

Sufficient Average Read Quality?

Acceptable Coverage and Appropriate Genome Size? Software Changes?

- QC parameters established at each step such that no DNA or resultant sequence data moved forward in the process without meeting minimum quality standards.
  - QC of software updates for a sequencer are evaluated by spiking in an internal control (PhiX) at a known concentration
  - QC of analytical software updates are evaluated by reanalyzing existing sequence data (fastq format) from a run of the Challenge Set generated during validation

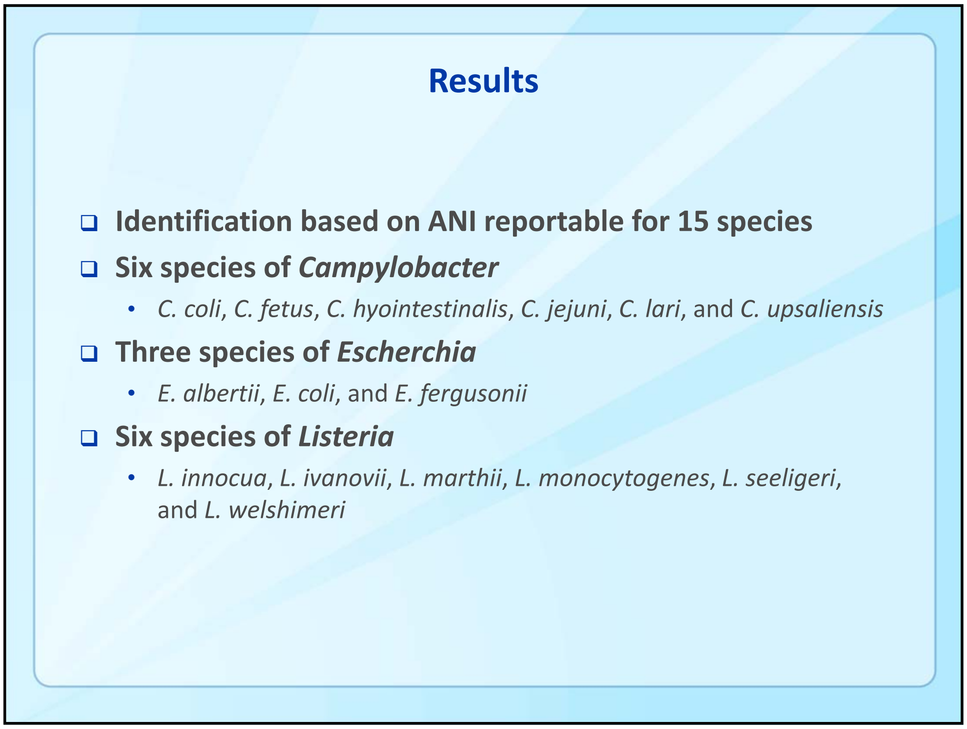


ANI Method Overview - Workflow and QC

- Modular workflow…
- …allows components at each step to be interchanged…

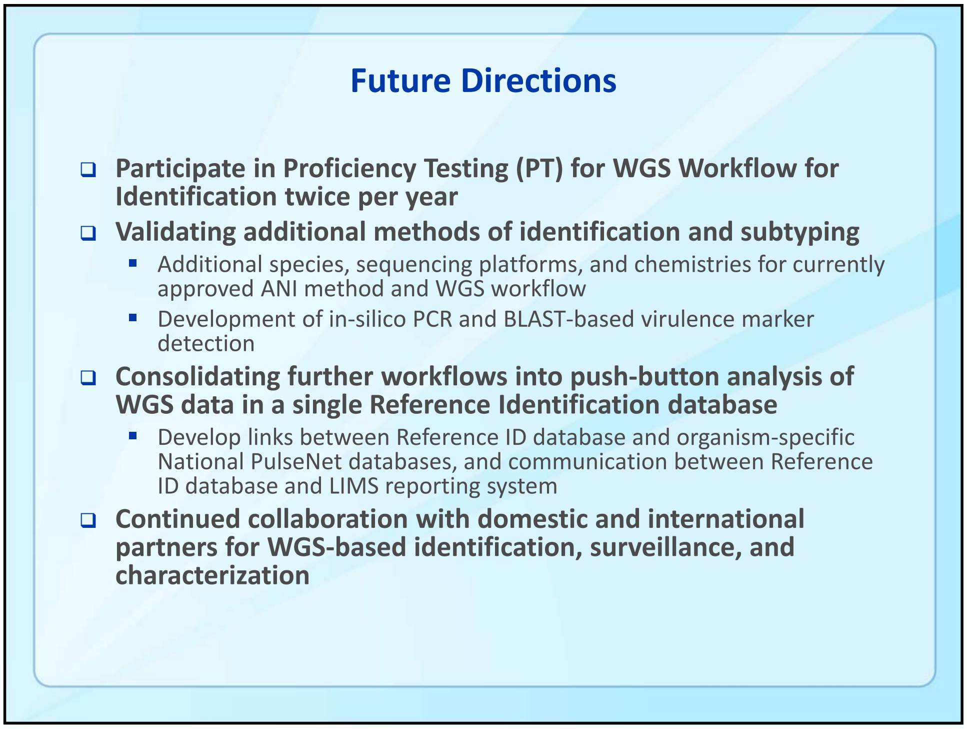
DNA Extraction and QC	Library Preparation	Sequencing	Analysis
Manual	Illumina Nextera XT	Illumina MiSeq	Reference ID Database
Automated	NEB Next	Illumina HiSeq	Command-Line

- …as long as QC parameters are met at each stage…
- …allowing different equivalent paths to be evaluated!!!



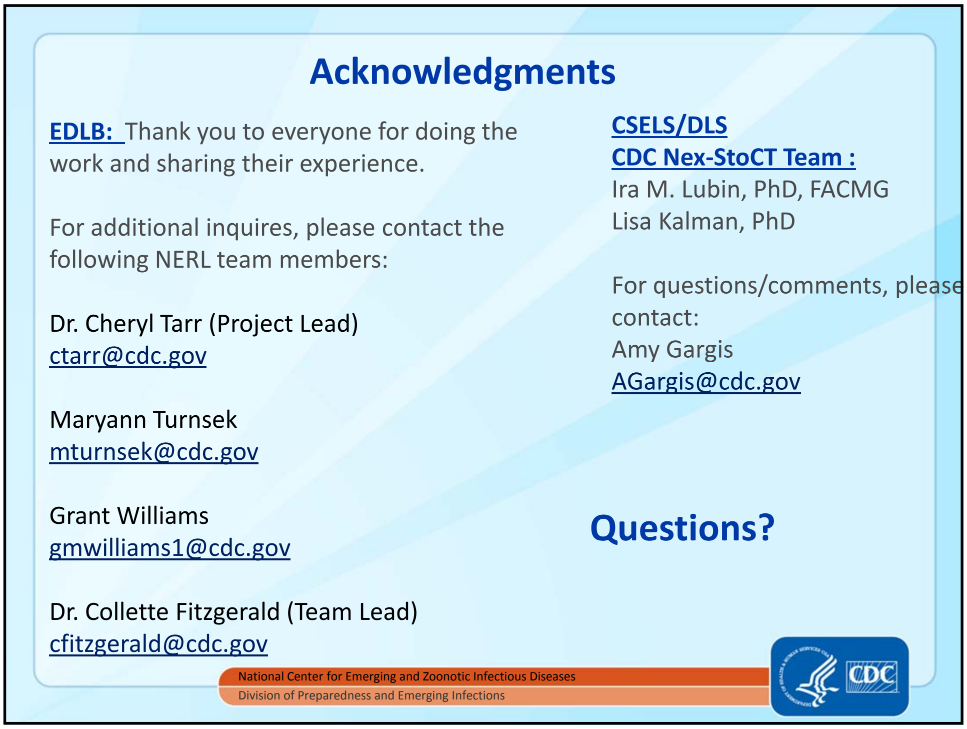
Results

- ❑ Identification based on ANI reportable for 15 species
- ❑ Six species of *Campylobacter*
  - *C. coli*, *C. fetus*, *C. hyointestinalis*, *C. jejuni*, *C. lari*, and *C. upsaliensis*
- ❑ Three species of *Escherchia*
  - *E. albertii*, *E. coli*, and *E. fergusonii*
- ❑ Six species of *Listeria*
  - *L. innocua*, *L. ivanovii*, *L. marthii*, *L. monocytogenes*, *L. seeligeri*, and *L. welshimeri*



Future Directions

- ❑ Participate in Proficiency Testing (PT) for WGS Workflow for Identification twice per year
- ❑ Validating additional methods of identification and subtyping
  - ▪ Additional species, sequencing platforms, and chemistries for currently approved ANI method and WGS workflow
  - ▪ Development of in-silico PCR and BLAST-based virulence marker detection
- ❑ Consolidating further workflows into push-button analysis of WGS data in a single Reference Identification database
  - ▪ Develop links between Reference ID database and organism-specific National PulseNet databases, and communication between Reference ID database and LIMS reporting system
- ❑ Continued collaboration with domestic and international partners for WGS-based identification, surveillance, and characterization



Acknowledgments

EDLB: Thank you to everyone for doing the work and sharing their experience.

For additional inquires, please contact the following NERL team members:

Dr. Cheryl Tarr (Project Lead)
ctarr@cdc.gov

Maryann Turnsek
mturnsek@cdc.gov

Grant Williams
gmwilliams1@cdc.gov

Dr. Collette Fitzgerald (Team Lead)
cfitzgerald@cdc.gov

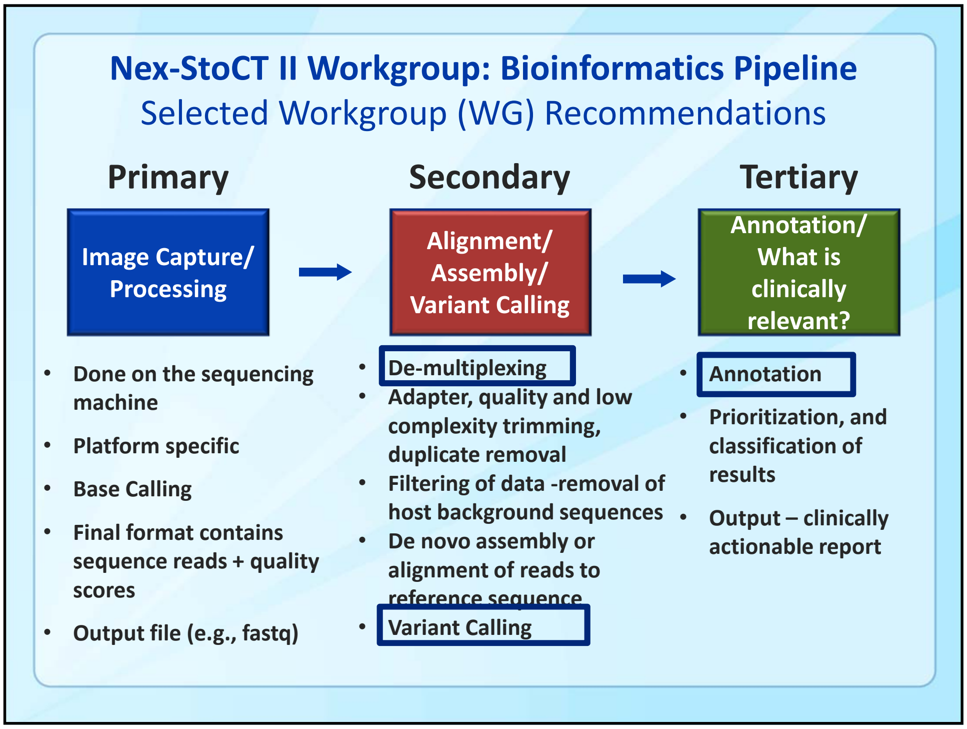
CSELS/DLS
CDC Nex-StoCT Team :
Ira M. Lubin, PhD, FACMG
Lisa Kalman, PhD

For questions/comments, please contact:
Amy Gargis
AGargis@cdc.gov

Questions?

National Center for Emerging and Zoonotic Infectious Diseases
Division of Preparedness and Emerging Infections

Extra Slides



Nex-StoCT II Workgroup: Bioinformatics Pipeline
Selected Workgroup (WG) Recommendations

Primary
Secondary
Tertiary
Image Capture/ Processing
Alignment/ Assembly/ Variant Calling
Annotation/ What is clinically relevant?

Primary

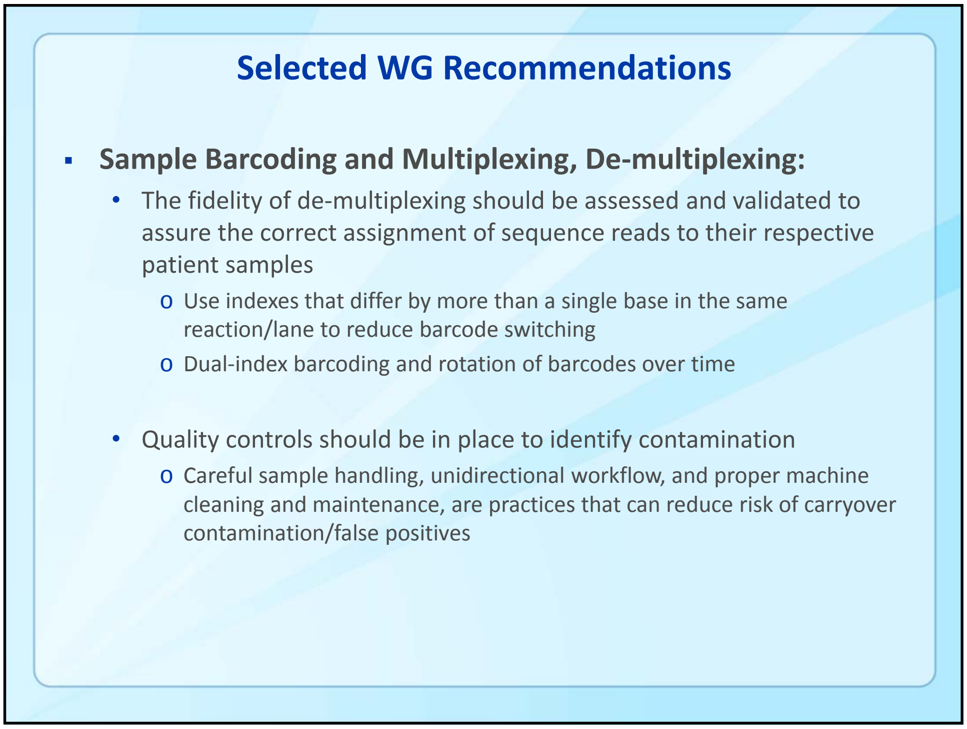
- Done on the sequencing machine
- Platform specific
- Base Calling
- Final format contains sequence reads + quality scores
- Output file (e.g., fastq)

Secondary

- De-multiplexing
- Adapter, quality and low complexity trimming, duplicate removal
- Filtering of data -removal of host background sequences
- De novo assembly or alignment of reads to reference sequence
- Variant Calling

Tertiary

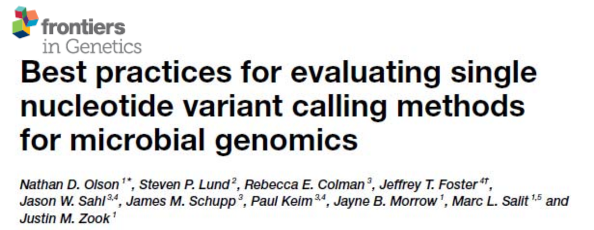
- Annotation
- Prioritization, and classification of results
- Output – clinically actionable report



Selected WG Recommendations

- **Sample Barcoding and Multiplexing, De-multiplexing:**
  - The fidelity of de-multiplexing should be assessed and validated to assure the correct assignment of sequence reads to their respective patient samples
    - Use indexes that differ by more than a single base in the same reaction/lane to reduce barcode switching
    - Dual-index barcoding and rotation of barcodes over time
  - Quality controls should be in place to identify contamination
    - Careful sample handling, unidirectional workflow, and proper machine cleaning and maintenance, are practices that can reduce risk of carryover contamination/false positives

Need for Standards in Variant Calling



frontiers in Genetics

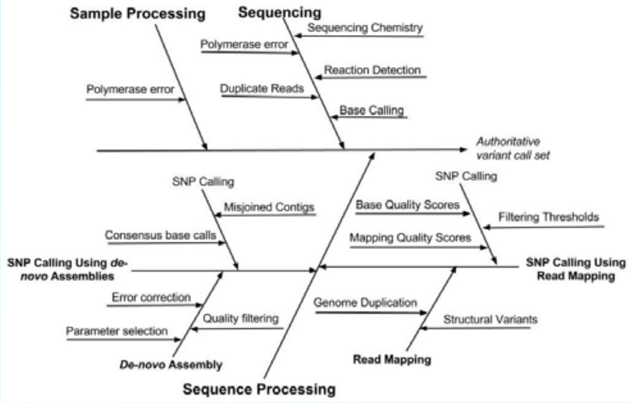
Best practices for evaluating single nucleotide variant calling methods for microbial genomics

Nathan D. Olson [1], Steven P. Lund [2], Rebecca E. Colman [3], Jeffrey T. Foster [4†], Jason W. Sahl [3,4], James M. Schupp [3], Paul Keim [3,4], Jayne B. Morrow [1], Marc L. Salit [1,5] and Justin M. Zook [1]*

235

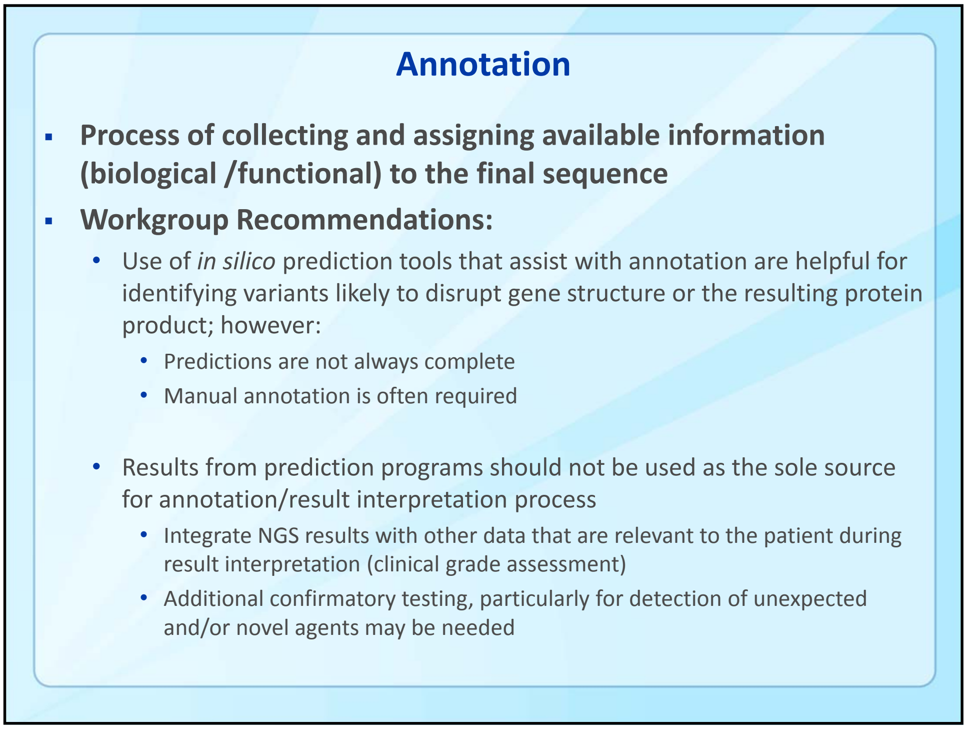
- **Essential for comparative genomics as it yields insights into nucleotide-level organismal differences**
- **A multistep process with a variety of potential error sources that may lead to incorrect variant calls**
- **Standardized methods for performance evaluation and reporting are needed**

Sources of Error for Sequencing and Variant Calling



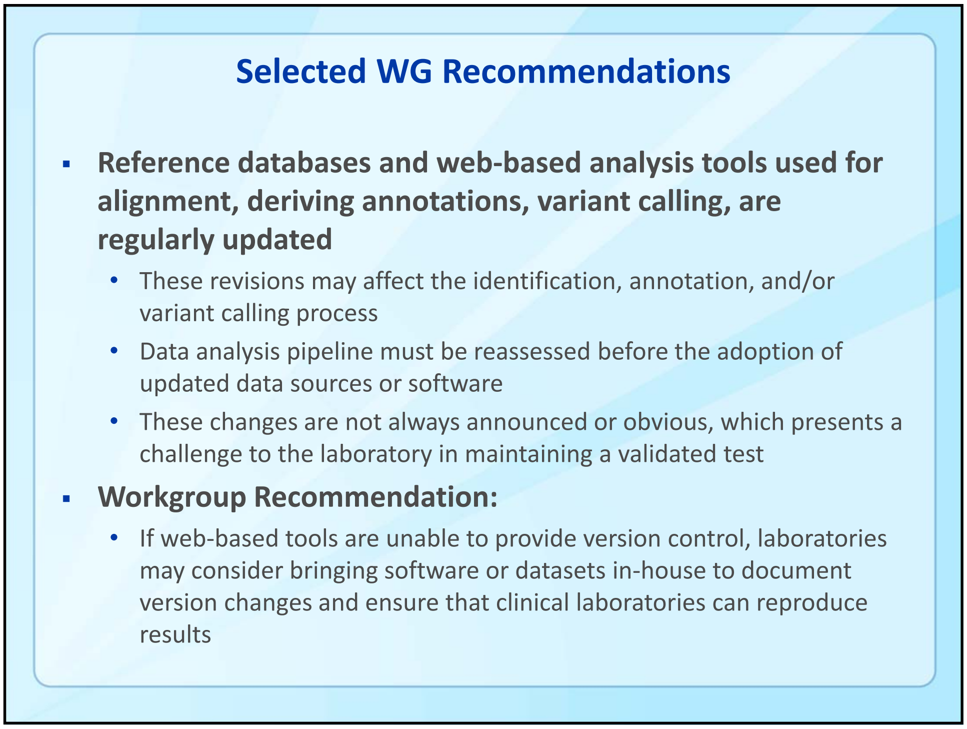
Front Genet. 2015; 6:235

- **Recommendations to optimize the quality of data used to generate variant calls:** minimize amplification during sequencing library preparation, perform paired-end sequencing, remove duplicate reads, realign around indels, and recalibrate base quality scores



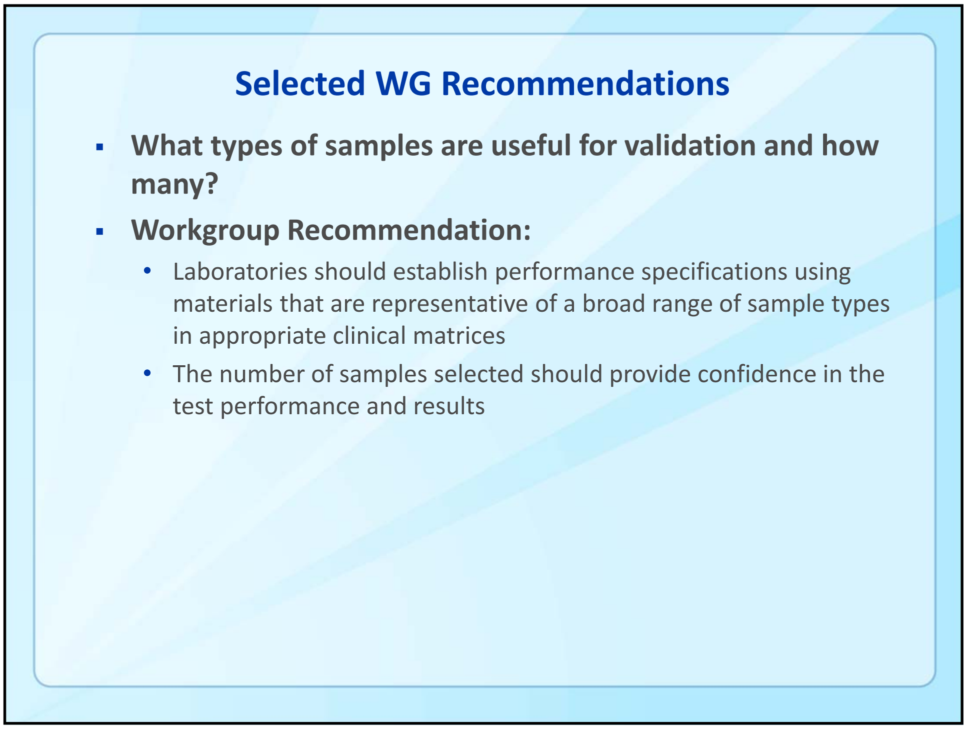
Annotation

- **Process of collecting and assigning available information (biological /functional) to the final sequence**
- **Workgroup Recommendations:**
  - Use of *in silico* prediction tools that assist with annotation are helpful for identifying variants likely to disrupt gene structure or the resulting protein product; however:
    - Predictions are not always complete
    - Manual annotation is often required
  - Results from prediction programs should not be used as the sole source for annotation/result interpretation process
    - Integrate NGS results with other data that are relevant to the patient during result interpretation (clinical grade assessment)
    - Additional confirmatory testing, particularly for detection of unexpected and/or novel agents may be needed



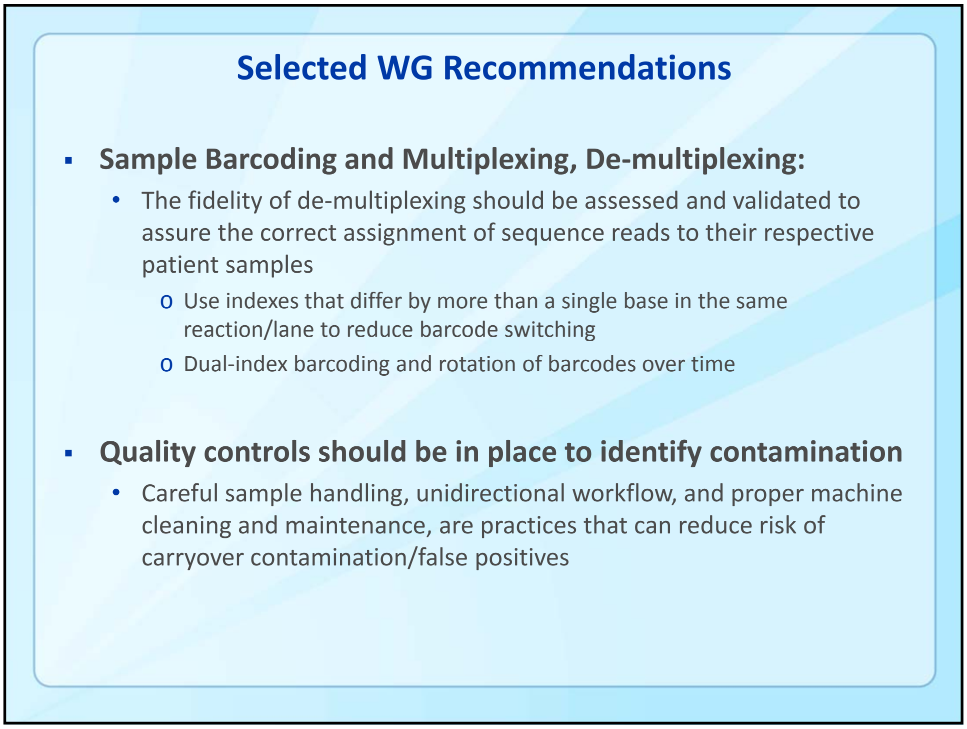
Selected WG Recommendations

- **Reference databases and web-based analysis tools used for alignment, deriving annotations, variant calling, are regularly updated**
  - These revisions may affect the identification, annotation, and/or variant calling process
  - Data analysis pipeline must be reassessed before the adoption of updated data sources or software
  - These changes are not always announced or obvious, which presents a challenge to the laboratory in maintaining a validated test
- **Workgroup Recommendation:**
  - If web-based tools are unable to provide version control, laboratories may consider bringing software or datasets in-house to document version changes and ensure that clinical laboratories can reproduce results



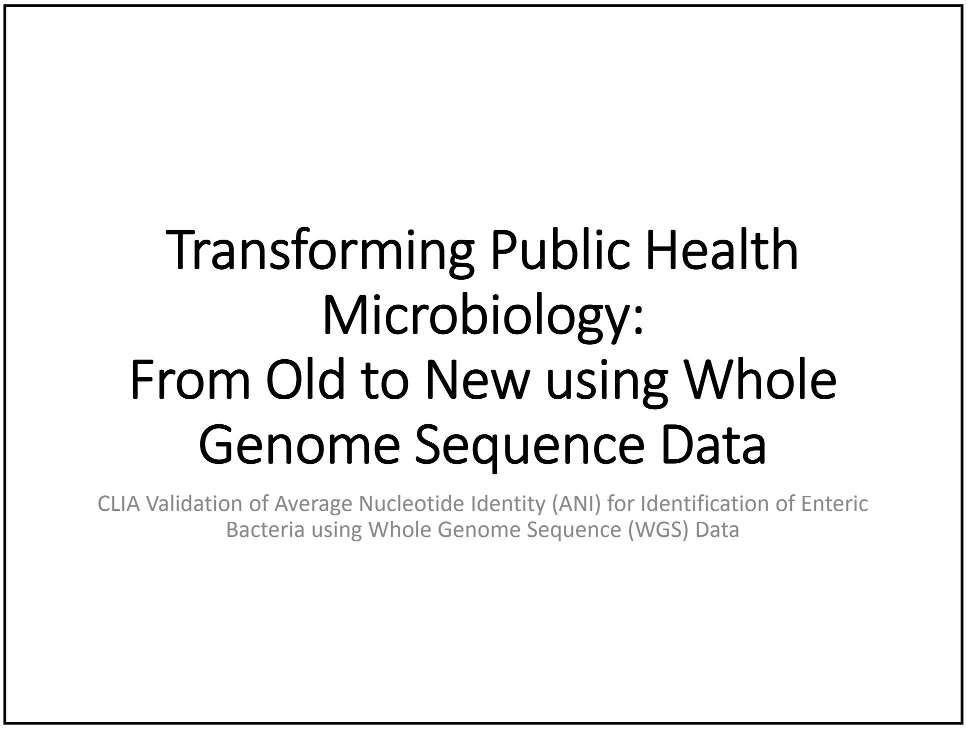
Selected WG Recommendations

- **What types of samples are useful for validation and how many?**
- **Workgroup Recommendation:**
  - Laboratories should establish performance specifications using materials that are representative of a broad range of sample types in appropriate clinical matrices
  - The number of samples selected should provide confidence in the test performance and results



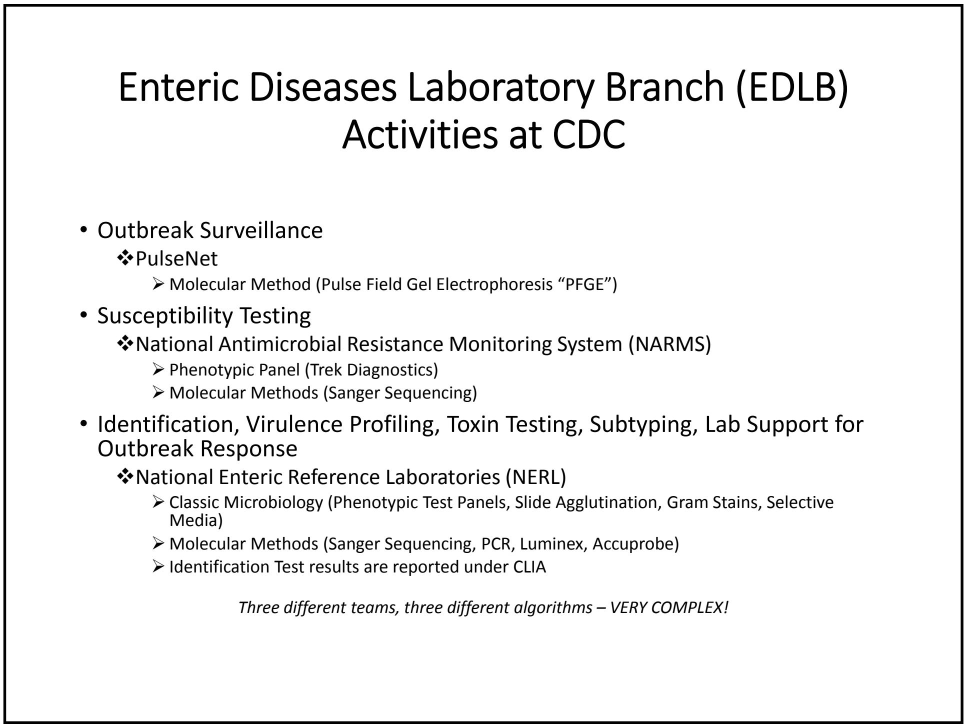
Selected WG Recommendations

- **Sample Barcoding and Multiplexing, De-multiplexing:**
  - The fidelity of de-multiplexing should be assessed and validated to assure the correct assignment of sequence reads to their respective patient samples
    - Use indexes that differ by more than a single base in the same reaction/lane to reduce barcode switching
    - Dual-index barcoding and rotation of barcodes over time
- **Quality controls should be in place to identify contamination**
  - Careful sample handling, unidirectional workflow, and proper machine cleaning and maintenance, are practices that can reduce risk of carryover contamination/false positives



Transforming Public Health Microbiology: From Old to New using Whole Genome Sequence Data

CLIA Validation of Average Nucleotide Identity (ANI) for Identification of Enteric Bacteria using Whole Genome Sequence (WGS) Data



Enteric Diseases Laboratory Branch (EDLB) Activities at CDC

- Outbreak Surveillance
  - ❖PulseNet
    - ➢ Molecular Method (Pulse Field Gel Electrophoresis "PFGE")
- Susceptibility Testing
  - ❖National Antimicrobial Resistance Monitoring System (NARMS)
    - ➢ Phenotypic Panel (Trek Diagnostics)
    - ➢ Molecular Methods (Sanger Sequencing)
- Identification, Virulence Profiling, Toxin Testing, Subtyping, Lab Support for Outbreak Response
  - ❖National Enteric Reference Laboratories (NERL)
    - ➢ Classic Microbiology (Phenotypic Test Panels, Slide Agglutination, Gram Stains, Selective Media)
    - ➢ Molecular Methods (Sanger Sequencing, PCR, Luminex, Accuprobe)
    - ➢ Identification Test results are reported under CLIA

Three different teams, three different algorithms – VERY COMPLEX!

Reference-related CLIA Activities at CDC

	Phenotypic test panels	Sanger sequencing: rpoB and/or 16S	Other Molecular Tests*
Campylobacter	Full (21 tests), Short (3 tests)	27 species	Taxa-specific PCRs (n=5)
Escherichia	Full (49 tests), Short (24 tests)	4 species	Taxa-specific PCRs (n=4), *Shigella flexneri* serotyping (n=10), Virulence subtyping (n=31)
Listeria	NA	NA	Accuprobe (*L. monocytogenes*)
Salmonella	Full (49 tests), Short (25 tests), Subspecies (10 tests)	2 species	Luminex (need number)
Vibrio	Full (49 tests)	12 species	Multiplex PCR for *V. cholerae*

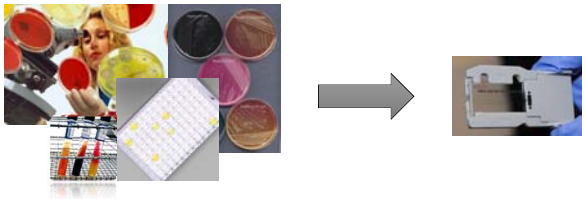
- Identification and subtyping of approximately 7,000 specimens per year
- Turn-around time (TAT) from one to four weeks, depending on the organism and complexity of tests performed

And as of 2016…….
Identification from WGS – the ANI Method

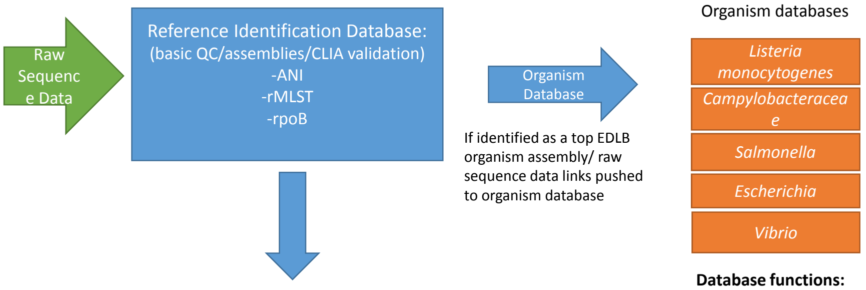
*: Some molecular testing has yet to undergo CLIA validation

WGS Vision

- Consolidating multiple laboratory workflows into one:
  - Identification – serotyping – virulence profiling – antimicrobial resistance characterization – plasmid characterization- subtyping
    - **Replacing - NOT supplementing current methods**
      - ✓ More Precise- Informative- Cost-efficient



WGS Vision – *work in progress*
EDLB's Consolidated Process



Raw Sequence Data

Reference Identification Database:
(basic QC/assemblies/CLIA validation)
-ANI
-rMLST
-rpoB

Organism Database

If identified as a top EDLB organism assembly/ raw sequence data links pushed to organism database

Organism databases

Listeria monocytogenes

Campylobacteraceae

Salmonella

Escherichia

Vibrio

If no database for organism then reference lab performs further characterization as needed:

- *Cronobacter*
- *Yersinia entercolitica*
- Etc.

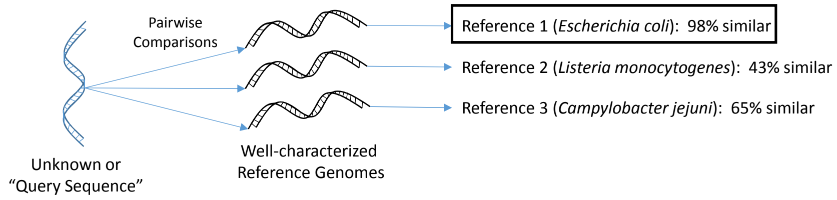
Database functions:

- Extended QC based on % core identified and # of alleles identified
- wgMLST for surveillance
- Organism specific 7-gene MLST
- Serotyping
- Identification of AMR genes
- Virulence gene id

CLIA Validation of Average Nucleotide Identity (ANI) for Identification of Enteric Bacteria using Whole Genome Sequence (WGS) Data

What is ANI?

- Average Nucleotide Identity (ANI) is a comparison of shared genes across the genomes of two strains – an unknown, Query Sequence and a well-characterized Reference Genome – and is a robust means to compare genetic relatedness of the strains.



Pairwise Comparisons

Reference 1 (*Escherichia coli*): 98% similar

Reference 2 (*Listeria monocytogenes*): 43% similar

Reference 3 (*Campylobacter jejuni*): 65% similar

Unknown or "Query Sequence"

Well-characterized Reference Genomes

❖ Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA 102(7): 2567-72
❖ Richter R, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA 106 (45): 19126-19131

ANI Method - Test Development and Optimization

- Generate Sequence Data
- Develop the Workflow…..*most time spent here*!
  - ❖How many samples can go on a sequencing run?
  - ❖What coverage will ANI require?
  - ❖What QC will the raw sequence data require?
  - ❖What bioinformatics tool will a microbiologist use for calculating ANI? *More to come on this….*
  - ❖What are the ANI parameters to be defined?
  - ❖How to incorporate QAQC best practices into the process?
  - ❖How will we manage and store our data files?
  - ❖*Many more questions…….*
- Draft documents: SOPs, logs, worksheets, reports, etc.
- Preliminary Analysis
  - ❖Down-sampling to test the robustness of the method (Depth of Coverage)
  - ❖Visualization of data to determine ranges and threshold values (Acceptance Criteria)

ANI Method – Test Development and Optimization The Enteric Reference ID Database (in BioNumerics)

“*The Microbiologist’s Tool*”

- Initial repository for all raw data and associated metadata
- “Click-button analysis”
  - ❖ Push-button assembly and analysis of QC metrics
  - ❖ ANI performed with the click of a button
    - ➢ Results imported into the database
  - ❖ Other alternative identification methods used when ANI returns no result
    - ➢ (i.e. no Reference Genome exists for the species of the Query Sequence)
- Linked to national PulseNet organism-specific databases
  - ❖ Push-button transcription into PulseNet databases for samples under the PulseNet umbrella
- Planned link to LIMS reporting system
  - ❖ Automated import and export between systems

ANI Method - Make a Validation Plan

- Define Purpose and Scope – replace gold-standard methods for ID with ANI
  - ➢ **Gold-Standard ID = ANI ID (MiSeq) = ANI ID (HiSeq)**
- Select Validation strains
  - ➢ frequency of receipt
  - ➢ taxonomic diversity
  - ➢ public health significance
- Finalize the Reference Genome Set
- Define the Timeline
- Designate validation testers
- Define Acceptance Criteria
  - ➢ Accuracy, Sensitivity, Specificity, and Precision (Reproducibility & Repeatability)
  - ➢ Select the Challenge strains
- Identify equipment used in the WGS workflow and have proof of proper maintenance
- Refine documents: SOPs, logs, worksheets, reports, etc.

ANI Method - Summarize Validation Data

- 325 strains, representing 40 species from three genera were selected for inclusion
  - ❖Comparative Gold Standard identifications derived from various methods were used to evaluate the Accuracy of ANI
  - ❖A subset of the more common human pathogens denoted as a Challenge Set to be used in Precision experiments.
    - ➢ One strain for each of the 16 most frequently received taxa
    - ➢ Reproducibility: Each strain processed by three different operators on three different machines on three different days
    - ➢ Repeatability: Three strains (one from each genera) processed by one operator on three different days as part of the three routine sequencing runs on the same machine

ANI Method – Summarize Validation Data (cont)

- **Define the Organism-specific threshold values for Identification by ANI**

Organism group	ANI value (%)	Bases aligned (%)	Genome Size (MB)
Campylobacter	≥92	≥70	1.4 to 2.2
Escherichia	≥95	≥70	4.5 to 5.5
Listeria	≥92	≥75	2.7 to 3.1

- **Define the method limitations and deviations**

1 – Definitive species-level identification through ANI is only achieved when a representative genome exists in the Reference Genome Set. Thus new or rare, unrepresented species cannot be identified using the ANI test.

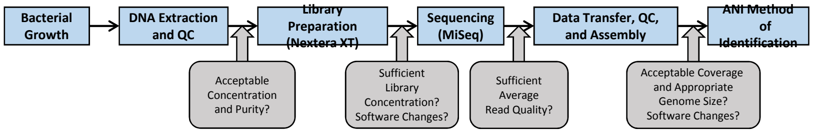
2 – Any query sequence that has <5X coverage will be rejected.

3 – Any query sequence with >0.1% ambiguous base calls will be rejected.

4 – The sequence length (genome size) of the query genome should be within the organism-specific values given in the interpretive guidelines. The ANI test can be performed with aberrant-sized genomes, but the sequence should be evaluated by a subject matter expert. Identification of such an organism should be confirmed by another independent validated method.

ANI Method Overview - Workflow and QC

- 30 SOPs, job aids, and worksheets were developed or modified to encompass the WGS ID workflow.



Bacterial Growth

DNA Extraction and QC

Library Preparation (Nextera XT)

Sequencing (MiSeq)

Data Transfer, QC, and Assembly

ANI Method of Identification

Acceptable Concentration and Purity?

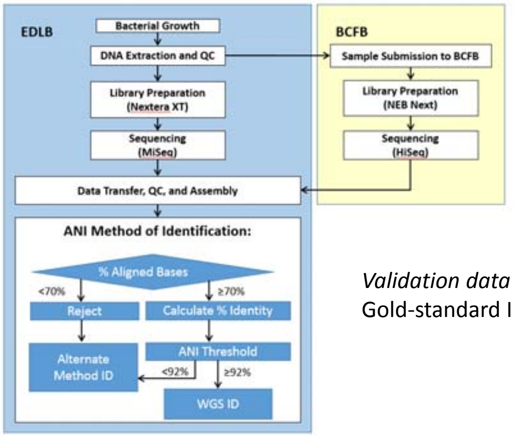
Sufficient Library Concentration? Software Changes?

Sufficient Average Read Quality?

Acceptable Coverage and Appropriate Genome Size? Software Changes?

- QC parameters established at each step such that no DNA or resultant sequence data moved forward in the process without meeting minimum quality standards.
  - QC of software updates for a sequencer are evaluated by spiking in an internal control (PhiX) at a known concentration
  - QC of analytical software updates are evaluated by reanalyzing existing sequence data (fastq format) from a run of the Challenge Set generated during validation

ANI Method – Process Flowchart



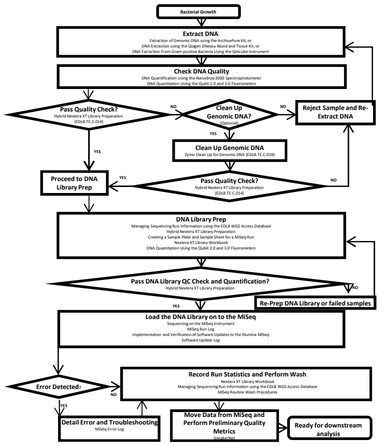
Regardless of the “wet lab” workflow…

- The QC for DNA remain the same (purity & concentration)
- The QC for the raw sequence data remain the same

Validation data shows….

Gold-standard ID = MiSeq ANI ID = HiSeq ANI ID

ANI Method – MiSeq Workflow



Bacterial Growth

Extract DNA
Extraction of Genomic DNA using the ArchivePure Kit, or
DNA Extraction using the Qiagen DNeasy Blood and Tissue Kit, or
DNA Extraction From Gram-positive Bacteria Using the QIAcube Instrument

Check DNA Quality
DNA Quantification Using the Nanodrop 2000 Spectrophotometer
DNA Quantitation Using the Qubit 2.0 and 3.0 Fluorometers

Pass Quality Check?
Hybrid Nextera XT Library Preparation
(EDLB.TE.C.014)

NO

Clean Up Genomic DNA?
(Optional)

NO

Reject Sample and Re-Extract DNA

YES

Clean Up Genomic DNA
Zymo Clean Up for Genomic DNA (EDLB.TE.C.018)

YES

Proceed to DNA Library Prep

YES

Pass Quality Check?
Hybrid Nextera XT Library Preparation
(EDLB.TE.C.014)

NO

DNA Library Prep
Managing Sequencing Run Information using the EDLB WGS Access Database
Hybrid Nextera XT Library Preparation
Creating a Sample Plate and Sample Sheet for a MiSeq Run
Nextera XT Library Workbook
DNA Quantitation Using the Qubit 2.0 and 3.0 Fluorometers

Pass DNA Library QC Check and Quantification?
Hybrid Nextera XT Library Preparation

NO

Re-Prep DNA Library or failed samples

YES

Load the DNA Library on to the MiSeq
Sequencing on the MiSeq Instrument
MiSeq Run Log
Implementation and Verification of Software Updates to the Illumina MiSeq
Software Update Log

Error Detected?

NO

Record Run Statistics and Perform Wash
Nextera XT Library Workbook
Managing Sequencing Run Information using the EDLB WGS Access Database
MiSeq Routine Wash Procedures

YES

Detail Error and Troubleshooting
MiSeq Error Log

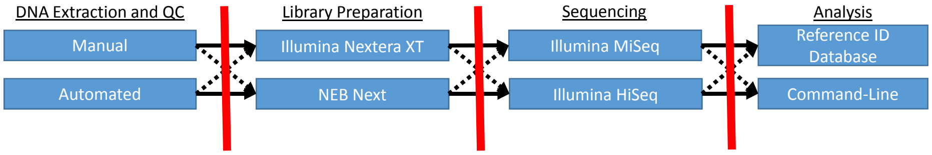
Move Data from MiSeq and Perform Preliminary Quality Metrics
SneakerNet

Ready for downstream analysis

- All procedures QMS approved at branch level
- MiSeq workflow can be used for different downstream methods of analysis

ANI Method Overview - Workflow and QC

- Modular workflow…
- …allows components at each step to be interchanged…



DNA Extraction and QC
Manual
Automated

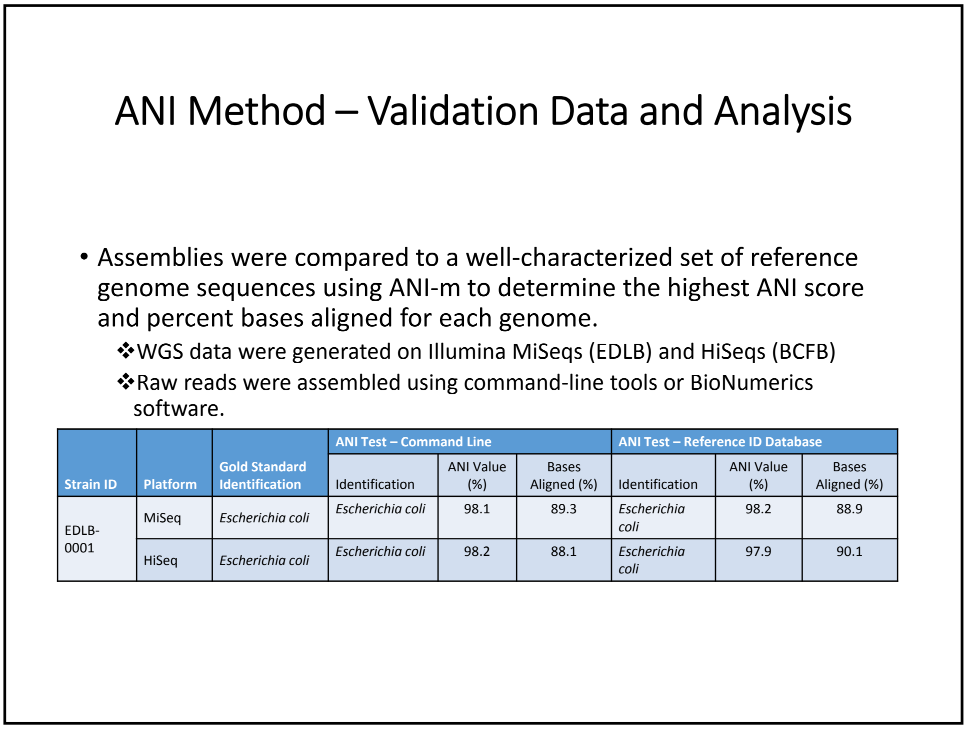
Library Preparation
Illumina Nextera XT
NEB Next

Sequencing
Illumina MiSeq
Illumina HiSeq

Analysis
Reference ID Database
Command-Line

- …as long as QC parameters are met at each stage…
- …allowing different equivalent paths to be evaluated!!!

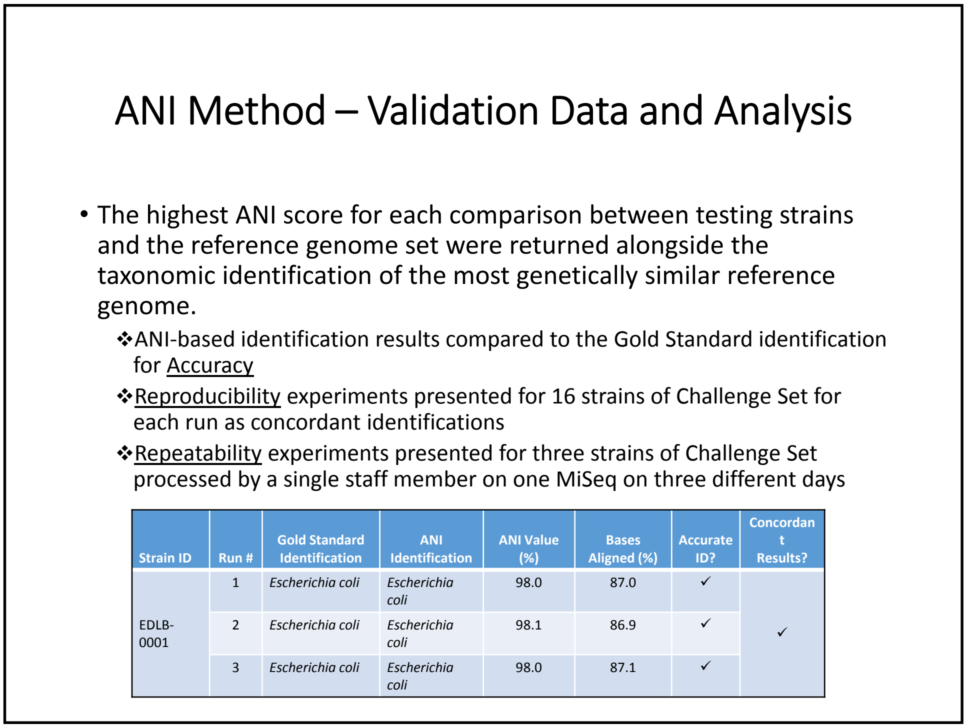
——— Linear Workflow
▪▪▪▪▪▪▪ Comparable Workflow



ANI Method – Validation Data and Analysis

- Assemblies were compared to a well-characterized set of reference genome sequences using ANI-m to determine the highest ANI score and percent bases aligned for each genome.
  - ❖WGS data were generated on Illumina MiSeqs (EDLB) and HiSeqs (BCFB)
  - ❖Raw reads were assembled using command-line tools or BioNumerics software.

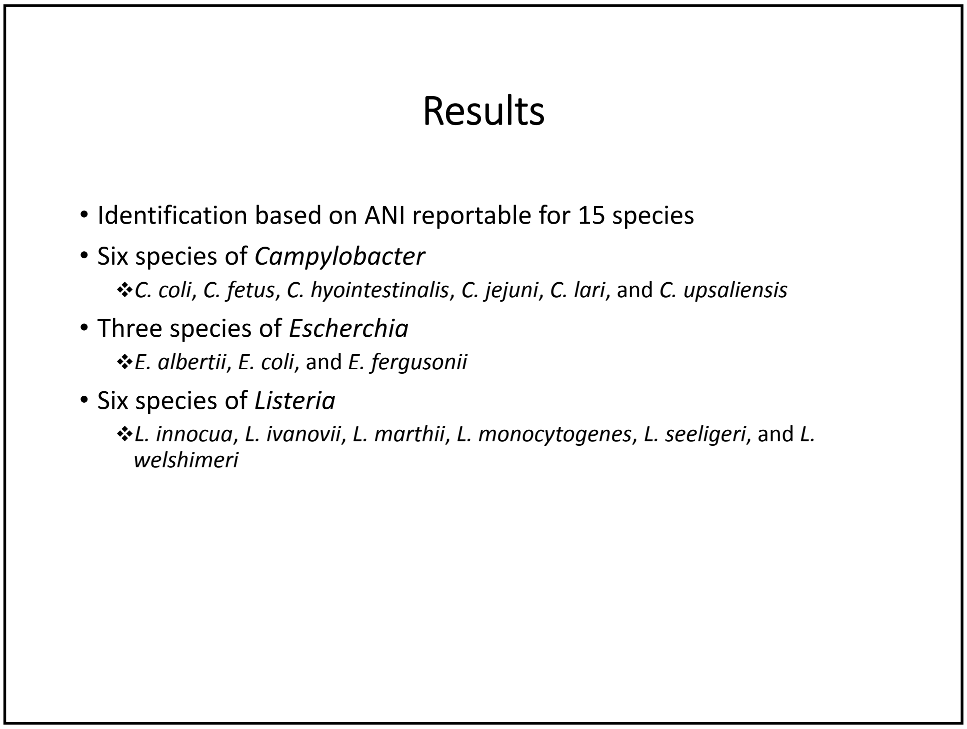
Strain ID	Platform	Gold Standard Identification	ANI Test – Command Line			ANI Test – Reference ID Database		
			Identification	ANI Value (%)	Bases Aligned (%)	Identification	ANI Value (%)	Bases Aligned (%)
EDLB-0001	MiSeq	*Escherichia coli*	*Escherichia coli*	98.1	89.3	*Escherichia coli*	98.2	88.9
	HiSeq	*Escherichia coli*	*Escherichia coli*	98.2	88.1	*Escherichia coli*	97.9	90.1



ANI Method – Validation Data and Analysis

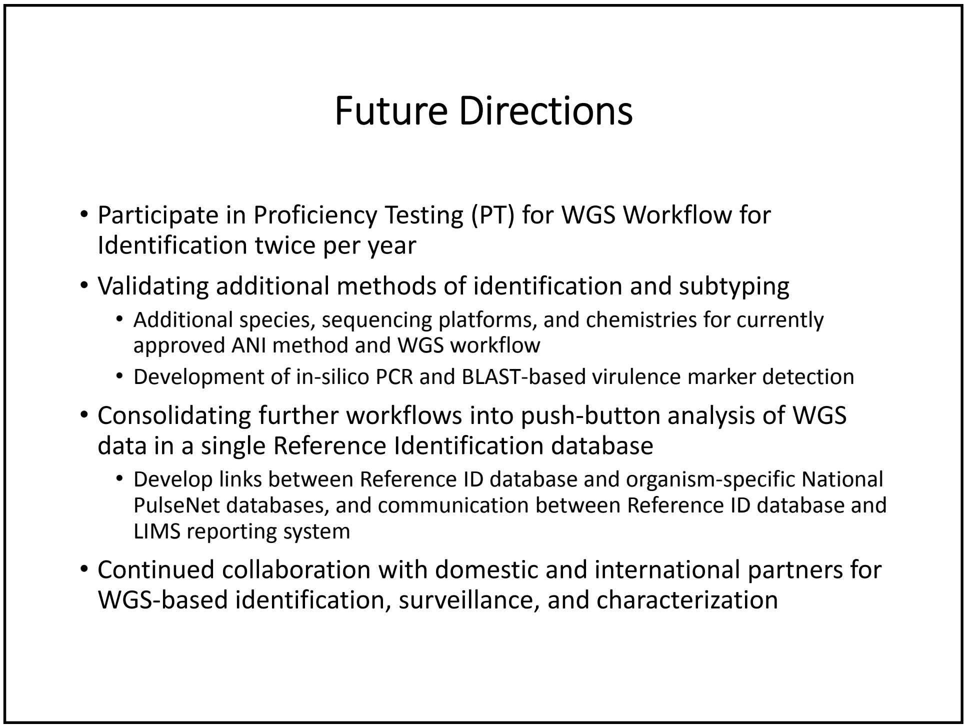
- The highest ANI score for each comparison between testing strains and the reference genome set were returned alongside the taxonomic identification of the most genetically similar reference genome.
  - ❖ANI-based identification results compared to the Gold Standard identification for Accuracy
  - ❖Reproducibility experiments presented for 16 strains of Challenge Set for each run as concordant identifications
  - ❖Repeatability experiments presented for three strains of Challenge Set processed by a single staff member on one MiSeq on three different days

Strain ID	Run #	Gold Standard Identification	ANI Identification	ANI Value (%)	Bases Aligned (%)	Accurate ID?	Concordant Results?
EDLB-0001	1	*Escherichia coli*	*Escherichia coli*	98.0	87.0	✓	✓
	2	*Escherichia coli*	*Escherichia coli*	98.1	86.9	✓	
	3	*Escherichia coli*	*Escherichia coli*	98.0	87.1	✓	



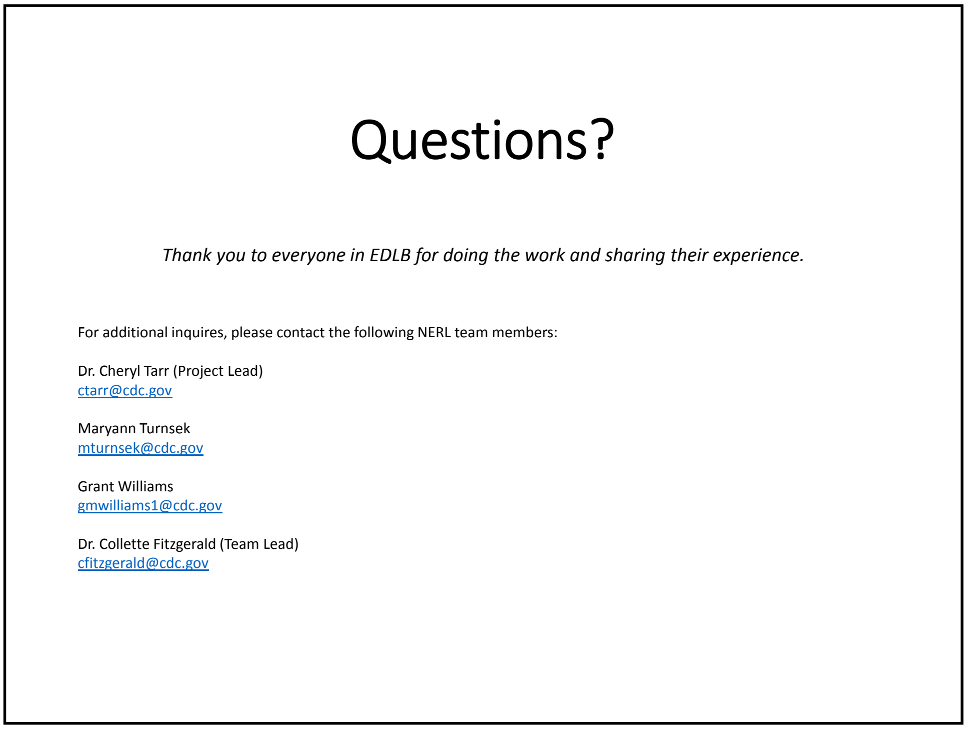
Results

- Identification based on ANI reportable for 15 species
- Six species of *Campylobacter*
  - ❖*C. coli*, *C. fetus*, *C. hyointestinalis*, *C. jejuni*, *C. lari*, and *C. upsaliensis*
- Three species of *Escherchia*
  - ❖*E. albertii*, *E. coli*, and *E. fergusonii*
- Six species of *Listeria*
  - ❖*L. innocua*, *L. ivanovii*, *L. marthii*, *L. monocytogenes*, *L. seeligeri*, and *L. welshimeri*



Future Directions

- Participate in Proficiency Testing (PT) for WGS Workflow for Identification twice per year
- Validating additional methods of identification and subtyping
  - Additional species, sequencing platforms, and chemistries for currently approved ANI method and WGS workflow
  - Development of in-silico PCR and BLAST-based virulence marker detection
- Consolidating further workflows into push-button analysis of WGS data in a single Reference Identification database
  - Develop links between Reference ID database and organism-specific National PulseNet databases, and communication between Reference ID database and LIMS reporting system
- Continued collaboration with domestic and international partners for WGS-based identification, surveillance, and characterization



Questions?

Thank you to everyone in EDLB for doing the work and sharing their experience.

For additional inquires, please contact the following NERL team members:

Dr. Cheryl Tarr (Project Lead)
ctarr@cdc.gov

Maryann Turnsek
mturnsek@cdc.gov

Grant Williams
gmwilliams1@cdc.gov

Dr. Collette Fitzgerald (Team Lead)
cfitzgerald@cdc.gov