CONTROLLED//PROPIN//BASIC

United States Department of Agriculture
Center for Veterinary Biologics

Testing Protocol

Antimicrobial Resistance Genotype Detection

Date:	**April 3, 2023**
Reference Number:	CVB-PRO-5107.01
Supersedes:	New
Contact:	Center for Veterinary Biologics, 515-337-6100

United States Department of Agriculture
Animal and Plant Health Inspection Service
P. O. Box 844
Ames, IA 50010

INTERNAL USE ONLY

Table of Contents

1. Introduction

This Testing Protocol (PRO) specifies the procedures for evaluating Master Seeds (MS) in biological products for antimicrobial resistance (AMR) genes in support of the Center for Veterinary Biologics (CVB) Policy Regarding Presence of Antimicrobial Resistance Genes in Regulated Veterinary Biological Products. Since risk is associated with the potential for horizontal transfer of resistance, assessment is based on output from the genotype detection pipeline used by the National Veterinary Services Laboratories (NVSL) Computational Biology and Informatics Section rather than phenotypic analysis. WGS data is generated during identity testing of MS, or as necessary, to conduct AMR genotype detection. Output from the NVSL pipeline is analyzed and interpreted by the CVB.

2. Materials

2.1 Equipment/Software necessary to run the AMR pipeline

Equivalents may be substituted if they achieve the same or better quality.

2.1.1 Hardware:

- Computer with MacOS, Linux, or Windows with the Ubuntu app installed

2.1.2 Bioinformatic tools/software:

- SPAdes (http://cab.spbu.ru/software/spades/) or another suitable assembler
- ABRicate (https://github.com/tseemann/abricate)
- Kraken2 (https://ccb.jhu.edu/software/kraken2/index.shtml?t=manual)
- MLST (https://github.com/tseemann/mlst)
- AMRFinder (https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/)
- SeqSero2 (https://github.com/denglab/SeqSero2)
- Python 3.6, or Later

2.1.3 Software necessary to analyze AMR pipeline output:

- PDF viewer, Microsoft Excel or other spreadsheet software, web browser for viewing html documents

3. Preparation for the Test

3.1 Personnel Qualifications/Training

Only individuals authorized and trained may run the AMR pipeline or analyze the output. Only bioinformaticists with basic Unix and Git understanding, along with user permission and the appropriate computing environment may run the AMR pipeline. Only individuals familiar with whole genome sequencing output and AMR analysis may interpret the final output.

4. Performance of Test

4.1 Pipeline Input

Initial processing and sequence assembly for CVB samples is described in CVB-PRO-5108. Sequence with valid assemblies may also be analyzed from external sources at the discretion of the CVB.

4.2 Input Format

The AMR genotype detection pipeline requires either raw FASTQ data or assembled data in FASTA format.

4.3 Failures and Errors

4.3.1 Output Script Evaluation

The output script should be evaluated for completeness and failures. It is the bioinformaticist's responsibility to determine the cause of any failures. The bioinformaticist will determine if the outputs are reasonable, however; it is the validator, or the subject matter expert's responsibility to check the accuracy of the output. See 5.1 for a detailed output guideline.

4.3.2 Output Sample Statistics in (sample_name)_(Date/time)_stats.xlsx

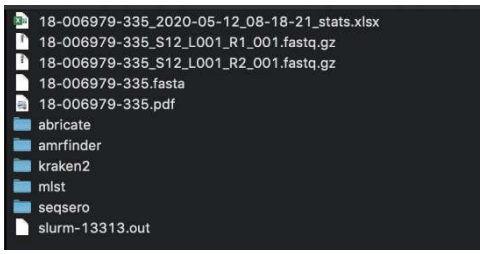
- The data in this file is summarized for quick and easy review of the output and indicates quality of the sequencing and assembly
- Common errors and issues that can be determined from the statistics
  - Insufficient FASTQ file size
  - Poor quality sequencing scores
  - Low average depth of coverage
  - Short read length
  - Fragmented assembly

5. Output Analysis and Reporting

5.1 Pipeline Output

- Original input FASTQ files
- Assembled FASTA file
- PDF sample report
- Excel sample statistics workbook
- Directories by application output

Example file structure:



5.2 WGS Report Output– software and versions used for each run are displayed on the report. See Appendix I for an example report.

5.1.1 Sequencing Statistics – This section is only applicable for samples ran from raw data ('fastq.gz' file extension) and may contain additional fields as appropriate.

- File names
- File sizes in gzipped format
- Mean read scores: Average Phred base score per file
- Q30 passing scores: Percent of reads with an average Phred score of 30 or greater per file
- Sequence depth: Calculated by the number of reads x read length of 240/assembled genome size
- Estimated genome size: Based on Multilocus Sequence Typing (MLST) information, if the species cannot be determined or the information is not available, 5 Mb will be used for bacterial samples
- Quality scale: See 5.3 Evaluating Quality Metrics for more information

5.1.2 Assembly Statistics – This section is applicable to both raw and assembled data and may contain fewer or additional fields as appropriate.

- Scaffolds: Total number of scaffolds/contigs in the assembly
- Total length: Length of all scaffolds/contigs summed
- Longest scaffold
- Scaffolds > 1K nt: Number of scaffolds/contigs that are larger than 1,000 bp
- Genome > 1K nt: Percent of Total Genome Length that the scaffolds/contigs greater than 1,000 bp comprise
- L50: In decreasing order by length, the number of scaffolds/contigs it takes to accumulate 50% of the Total Length
- N50: In decreasing order by length, the length of the contig where the sum of the lengths of the contigs is greater than or equal to 50% of the total length of all contigs
- Quality scale: See 5.3 Evaluating Quality Metrics for more information

5.1.3 Multilocus Sequence Typing (MLST) – Output from the software MLST, based on schemes defined on pubmlst.org

5.1.4 Serotyping for *Salmonella* – Output from SeqSero2. It is only displayed if the MLST indicates the sample is *Salmonella*

5.1.5 Antimicrobial Resistance Analysis– AMRFinder and ABRicate outputs are generated by the NVSL pipeline but only AMRFinder results are included in the WGS Report by default. ABRicate results are not applied in the CVB decision making process but may be utilized in developmental work.

5.3 Evaluating Quality Metrics

For general CVB guidelines on sequence quality see CVB-PRO-5108. There is no "lower limit" of sequence quality that precludes analysis, especially if the sequence was difficult to obtain. However, interpretation is done in the context of sequence quality. In addition to the quality metrics on each report, the statistics file for each sample contains redundant as well as additional statistics.

5.3.1 Statistics File: This file is contained within each isolate folder. The importance of each column's results may vary depending on the biology of the organism. Periodically, columns may be added and/or removed based on current scientific processes. Below are the most common fields used to evaluate sequencing and assembly.

- R1, R2 File size: Zipped file size of the reads. Typically for a 5MB genome, every 10MB should increase the depth of coverage by 5X
- R1, R2 Total reads: Total number of reads per file
- R1, R2 Mean length: Average read length after adapters are trimmed per file
- R1, R2 Mean quality: Average Quality score per file
- R1, R2 Passing Q30: Percent of reads with an average Quality score of 30 or greater per file
- Assembly contig count
- Contig count (< 300 bp)
- Contig count (301-1,000 bp)
- Contig count (>1 kb)
- Total length: Sum of the length of all contigs in the assembly
- Longest contig
- N50: In decreasing order by length, the length of the contig where the sum of the lengths of the contigs is greater than or equal to 50% of the total length of all contigs
- Mean coverage: Depth of coverage calculated by: (number of reads multiplied by length of reads)/total assembly length

5.3.2 Report: Information on understanding the visual Sequencing and Assembly Quality Scales.

- The Quality Scales are calculated using a Mahalanobis distance with a correlation matrix based on the quality metrics included in the report for both the sequencing statistics and assembly statistics. These parameters are tuned based on a historic sample dataset and centering around expected values. Note, it is simply a visual approximation of quality- no numeric score is reported from this calculation.
- The algorithm used to calculate the quality score is dependent on expected genome length from the MLST identification. If the species could not be determined or the estimated genome length cannot be determined, 5 megabase is used for bacterial samples. If this is not accurate, the quality scores will be negatively impacted for both sequencing and assembly.
- Other variables used to calculate the quality scores
  - Sequencing: Passing Q30, Total reads, Mean quality
  - Assembly: Longest contig, contigs >1,000kb, contig count, N50, L50, total length of contigs

- Example:

6. Report of Test Results

6.1 AMRFinder Result

Using the WGS report, resistance genes identified by AMRFinder from an acceptable Master Seed sequence are reported in the LSRTIS test "Result" field as follows:

- Genes detected with "AMR" in the "Element Subtype" column are reported with AMR: and the identity from the "Gene symbol" column of the WGS report.

- Genes detected with "POINT" in the "Element Subtype" column are reported with AMR, POINT: and the identity from the "Gene symbol" column on the WGS report.

- The disposition for the test is recorded as "Characterization" and verified.

After verifying the test, a Lab Microbiologist/VMO/Biologist or designee will notify the Reviewer and Risk Manager when an AMR gene(s) is detected in a Master Seed and will assist with any related questions.

Although they may appear in the NVSL WGS report, the CVB does not generally report in LSRTIS AMR genes identified beyond the "Core" AMRFinder database. The CVB may revise this position as databases evolve or at its discretion.

6.2 Role of CVB Reviewer

The CVB Reviewer for that Master Seed will determine additional action that may be needed based on the findings.

Appendix

USDA United States Department of Agriculture

Bacterial Whole Genome Sequencing Report

Month Day, Year

Sample ID: Example

Sequencing Technology

Nextera XT DNA Library Preparation
MiSeq 2 x 250 Read Generation

Sequence Statistics

Quality Scale: Low — High

Filename	Example.fastq.gz		Example.fastq.gz
File size	165.1 MB		190.1 MB
Mean Read Score	36.44		33.81
Q30 Passing	94.8%		80.2%
Sequence Depth	87.3X	Calculated by number of reads x 240/Genome Length	
Genome Length	5,000,000bp	Based on MLST identification	

Assembly Statistics

Quality Scale: Low — High

Scaffolds	Total length	Longest scaffold	Scaffolds >1K nt	Genome >1K nt	N50	L50
146	4,563,759	329,847	91	99.5	119,252	12

De novo assembly performed using SPAdes v3.13.0 .

Multi Locus Sequence Typing (MLST)

Organism ID: **Escherichia coli** | Schema-Sequence type: **ecoli-93**

MLST Detail: adk(6) fumC(11) gyrB(4) icd(10) mdh(7) purA(8) recA(6)

Data obtained using 2.19.0. Software website: https://github.com/tseemann/mlst

Information on MLST schemes and allelic profiles can be found at pubMLST.org

Serotyping for Salmonella Isolates

Predicted serotype(s)	**- -:-:-**
Predicted antigenic profile	-:-:-
Predicted subspecies	-

Note: The input genome cannot be identified as Salmonella. Check the input for taxonomic ID, contamination, or sequencing quality.

Data obtained using SeqSero. Software website: https://github.com/denglab/SeqSero2

Appendix

Antimicrobial Resistance Analysis

Results were obtained using AMRFinder. AMRFinder uses BLASTX to search a hierarchy of gene families with predetermined cutoffs.

AMRFinder

Contig id	Start	Stop	Gene symbol	Sequence name	Scope	Element Subtype	% Coverage of reference sequence	% Identity to reference sequence	Class	Subclass
NODE_14_length_105259_cov_53.422983	58238	60781	cyaA_S352T	Escherichia fosmidomycin resistant CyaA	core	POINT	100.0	99.29	FOSMIDOMYCIN	FOSMIDOMYCIN
NODE_23_length_68836_cov_32.953281	33138	34493	glpT_E448K	Escherichia fosfomycin resistant GlpT	core	POINT	100.0	99.78	FOSFOMYCIN	FOSFOMYCIN
NODE_3_length_312869_cov_48.266210	118299	119387	pmrB_E121K	Escherichia colistin resistant PmrB	core	POINT	100.0	99.72	COLISTIN	COLISTIN
NODE_58_length_6267_cov_1126.467915	2634	3491	blaTEM-1	broad-spectrum class A beta-lactamase TEM-1	core	AMR	100.0	100.0	BETA-LACTAM	BETA-LACTAM
NODE_58_length_6267_cov_1126.467915	5513	6133	tet(C)	tetracycline efflux MFS transporter Tet(C)	core	AMR	52.27	100.0	TETRACYCLINE	TETRACYCLINE

United States Department of Agriculture

Isolate ID: Example

Appendix

USDA United States Department of Agriculture

Isolate ID: Example

AMRFinder *Revision 3.10.16*

https://github.com/ncbi/amr/wiki

Definitions were taken from the AMRFinder documentation.

Target Identifier- This is from the FASTA defline for the DNA sequence

Contig id- Contig name

Start- 1-based coordinate of first nucleotide coding from protein in DNA sequence on contig

Stop- 1-based coordinate of last nucleotide coding for protein in DNA sequence on contig

Gene symbol- Gene or gene-family symbol for protein hit

Protein name- Full-text name for the protein

Method- Type of hit found by AMRFinder one of five options

ALLELE- 100% sequence match over 100% of length to a protein named at the allele level in the AMRFinder database

EXACT- 100% sequence match over 100% of length to a protein in the database that is not a named allele

BLAST- BLAST alignment is >90% of length and >90% identity to a protein in a the AMRFinder database

PARTIAL- BLAST alignment is >50% of length, but <90% of length and >90% identity

HMM- HMM was hit above the cutoff, but there was not a BLAST hit that met standards for BLAST or PARTIAL

Target length- The length of the query protein. The length of the BLAST hit for translated-DNA searches

Reference protein length- The length of the AMR Protein in the database (NA if HMM-only hit)

Scope- The AMRFinderPlus database is split into 'core' AMR proteins that are expected to have an effect on resistance and 'plus' proteins of interest added with less stringent inclusion criteria. These may or may not be expected to have an effect on phenotype.

Element subtype- Further elaboration of functional category into (ANTIGEN, BIOCIDE, HEAT, METAL, PORIN) if more specific category is available, otherwise the element is repeated

% Coverage of reference protein- % covered by blast hit (NA if HMM-only hit

% Identity to reference protein- % amino-acid identity to reference protein (NA if HMM-only hit)

Appendix

United States Department of Agriculture

Isolate ID: Example

Alignment length- Length of BLAST alignment in amino acids (NA if HMM-only hit)

Accession of closest protein- RefSeq accession for protein hit by BLAST (NA if HMM-only hit)

Name of closest protein- Full name assigned to the AMRFinder database protein (NA if HMM-only hit)

HMM id- Accession for the HMM

HMM description- The family name associated with the HMM