CONTROLLED//PROPIN//BASIC

United States Department of Agriculture
Center for Veterinary Biologics

Testing Protocol

Identity Testing by Whole Genome Sequencing

Date:	**May 12, 2023**
Reference Number:	CVB-PRO-5108.02
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

Mention of trademark or proprietary product does not constitute a guarantee or warranty of the product by USDA and does not imply its approval to the exclusion of other products that may be suitable.

Table of Contents

1. Introduction

This document describes how Whole Genome Sequencing (WGS) is utilized at the Center for Veterinary Biologics (CVB) in identity testing to fulfill the intent of title 9 code of federal regulations (9 CFR), section 113.3(c). WGS may be used as a supplemental test for bacterial master seed identity. Identity testing of a Master Seed (MS) is not limited to the processes outlined here, but WGS is generally included in bacterial MS testing plans.

WGS data generated during identity testing of MS is also used to conduct antimicrobial resistance (AMR) genotype detection in support of the CVB Policy Regarding Presence of Antimicrobial Resistance Genes in Regulated Veterinary Biological Products.
WGS data may also be used to support the testing required for recombinant organisms (CVB-SOP-0154), such as to obtain sequence for genes inserted into expression vectors.

The CVB may utilize WGS results in specific situations to support sterility issues, but the primary use of WGS is to determine MS identity and for AMR genotype detection rather than extraneous agent testing. Extraneous agent testing for bacterial MS is described in 9 CFR 113.27, 113.28, and 113.30 and is outside the scope of this protocol.

2. Materials

2.1 DNA extraction

2.1.1 See CVB-SOP-5105

2.2 Sequencing Pipeline Specifications

2.2.1 See CVB-PRO-5107

2.3 Data analysis

2.3.1 General computing resources including internet access sufficient to conduct background research and utilize web-based analyses tools.

3. Preparation for the Test

3.1 Personnel qualification/training

Technical personnel must have a working knowledge of the use of general laboratory chemicals and equipment. They must have specific training and experience in DNA extraction and the reports generated from the National Veterinary Service Laboratories (NVSL) bioinformatics pipeline.

4. Test Procedure Overview

4.1 DNA Extraction

4.1.1 Extraction Method

MS samples are evaluated by the CVB lab to determine an appropriate method(s) of DNA extraction. The standard processes are described in detail in CVB-SOP-5105 and briefly below. Note, new and/or alternative methods are used as necessary to isolate genomic DNA of sufficient quantity and adequate quality for sequencing.

- **a.** For most bacterial submissions, the CVB performs DNA extractions directly (without subculture) on firm submitted MS samples using a Maxwell® RSC 48 System and the Maxwell® RSC Whole Blood DNA Kit.
- **b.** A preincubation with lysozyme is generally used prior to Maxwell® extraction for Gram-positive bacteria but not Gram-negative bacteria.
- **c.** Manual/specialized extraction methods and/or specific kits are often used for eukaryotic MS samples or samples that fail to extract well using the Maxwell system.
- **d.** Quality and quantity of extracted DNA is evaluated using spectrophotometric absorbance analysis (260nm/280nm ratio) or Qubit fluorometric quantification or equivalent methods.
- **e.** The ultimate determination of adequate quantity and quality of extracted DNA for submission for WGS is determined by comparison to current National Veterinary Service Laboratories (NVSL) submission guidelines and if necessary, consultation with NVSL technical experts.

4.2 Whole Genome Sequencing

4.2.1 NVSL Pipeline

MS DNA samples are submitted to NVSL for whole genome sequencing (WGS).

- **a.** The current standard sequencing option chosen by CVB is Nextera XT DNA Library Preparation with sequencing on an Illumina MiSeq System with 2 x 250bp Read Generation.
- **b.** NVSL also has Nanopore MinION sequencing available if necessary.
- **c.** Alternative equivalent or improved sequencing methods and/or providers may be substituted at any time.

4.2.2 WGS Report

A whole genome sequencing report is generated by NVSL summarizing an initial automated analysis of the data generated from the sequenced MS sample. This report provides quality information for raw and assembled sequence. Additionally, several automated bioinformatic tools have been incorporated into the NVSL WGS analysis pipeline and results from these tools are output into the WGS report. As methods and technology evolve, new, alternative, and additional programs/analyses/tools/versions may become part of the NVSL pipeline and WGS report overtime. Ultimately, it’s the responsibility of CVB subject matter experts to interpret sequencing pipeline results as a component of the overall MS identity evaluation. Listed below are current programs/analyses/tools that are used to analyze sequence data generated from MS samples.

a. SPAdes (the current standard assembler used by the NVSL pipeline, other assemblers may be utilized as necessary) (http://cab.spbu.ru/software/spades/)
b. MLST (NVSL pipeline) (https://github.com/tseemann/mlst)
c. MLST schemes and allelic profiles (MLST search database) (https://pubmlst.org/), MLST Typing (https://pubmlst.org/organisms/), rMLST (https://pubmlst.org/rmlst/)
d. Kraken2 (https://ccb.jhu.edu/software/kraken2/index.shtml?t=downloads)
Output visualized through KronaTools (https://github.com/marbl/Krona/wiki/KronaTools)
e. SeqSero2 (https://github.com/denglab/SeqSero2)
f. ABRicate (https://github.com/tseemann/abricate)
g. AMRFinder (https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/)

4.2.3 Data Storage

The NVSL WGS reports are generated and saved to the LUSTRE/Bioinfo Drive by the NVSL. The CVB saves a copy of the data and report to the subfolder of the appropriate fiscal year within \\aapiaamfiles\cvbdatabases\CVB PEL Shared\Sequencing Data before performing any analysis or edits.

5. Interpretation of the Test Results

Interpretation of WGS results takes place in two parts. First, test validity is determined using the sequence statistics outputs from the NVSL WGS report. If the CVB determines the test is "valid", additional evaluation will be conducted to analyze identity related results and determine the disposition of the valid test. CVB subject matter experts (SME) review all validity and disposition results and maintain the ability to modify acceptance criteria based on their best judgement through professional knowledge/experience, literature review, and/or consultation with other SMEs.

5.1 Validity

a. The results will generally be valid when the NVSL quality score for Sequence Statistics on the WGS report are in the green (high quality) color range. Scores in the yellow, orange, and red ranges may require additional actions and analyses to determine validity (see table below). An example score from the quality scale chromatogram is illustrated here (the grey oval on indicates chromatogram position):

b. Specific items considered when reviewing MS Sequence Statistics:

Criteria	Valid	Marginal	Action if Marginal	Invalid
Mean Score Reads*	≥ 30	27-29	1.Evaluate sequencer. Consider results valid if other quality criteria are acceptable and the sequencer had poor read quality upon consultation on a particular run, particularly for the Q30 Passing (R1 – R2). 2.Consider conditions that may hamper sequence quality including dilution buffer and previous sample treatment. 3.Resubmit if Action 1 does not explain poor sequence or item 2 can be corrected.	≤ 26
Q30 Passing for R1	>80%	75-80%		< 75%
Q30 Passing (R1 – R2)	≤ 30%	40-30%		> 40%

*A Mean Read Score of 30 equates to 99.9% accuracy of the base call.

5.2 Disposition Determination

Disposition for a valid test on a bacterial MS sample is primarily based on a combined analysis of MS sequence data using several of the following bioinformatics tools.

A. Bacterial Multilocus Sequence Typing (MLST), compares sequence variations across multiple genes to characterize a specific species based on allelic profile. CVB conducts its own MLST analysis independent of the NVSL pipeline result reported on the WGS report.
    a. Considered acceptable when a match is found to genus and species (based on typing if currently annotated in a bacterial database at (https://pubmlst.org/organisms/). Output should be consistent with firm identity statement.
    b. If further identification is required, e.g. strain or biovar, this tool may also be utilized for some organisms (Note: for *Salmonella*, see Appendix I).

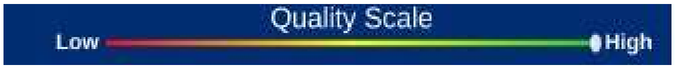
B. Ribosomal Multilocus Sequence Typing (rMLST), compares sequence variations across multiple ribosomal genes to identify/verify a specific species based on their unique allelic profile.

a. Considered acceptable if a match to genus and species (in currently annotated database at https://pubmlst.org/rmlst/). Output must be consistent with firm identity statement.
b. This tool may be used to support strain identification.

C. Kraken 2 Taxonomic Sequence Classification System, uses multiple short DNA sequences from genomes to identify closely related taxonomic groups (see https://ccb.jhu.edu/software/kraken2/index.shtml for further information).
   a. Kraken results are considered in context with MLST and rMLST results and should be consistent with the firm identity statement.
   b. Acceptability of Kraken results is determined according to the following table:

If **both** Bacterial MLST and rMLST determined genus and species match firm identification	If **only** rMLST provides genus and species match to firm identification
• > 85% genus match to predicted (excluding "No hits", i.e. Bacteria root) • < 3% Match to a single species other than predicted • < 5% Match to non-predicted species (as a group)	• > 85% genus match to predicted (excluding "No hits", i.e. Bacterial root) • **>30% species match to predicted** • < 3% Match to a single species other than predicted • < 5% Match to non-predicted species (as a group)

D. Assembly Statistics, evaluation of the sequence assembly based on quality scores and known information.
   a. The quality of sequence assemblies for data generated from the NVSL pipeline (specifically from the Illumina MiSeq System) is expected to vary based on organismal characteristics such as genome size, repetitive sequences, and high GC content inherent to specific MS samples.
   b. The NVSL pipeline has been tailored for bacterial analysis, and when necessary, the CVB will conduct additional or altered testing to obtain interpretable results for specific MS samples.
   c. In general, an assembly is considered acceptable if the quality scale on the WGS Report is in the green (see example below) but flexibility can be granted based on consultation between subject matter experts in NVSL and CVB.



d. If the other bioinformatic tool results (in A, B, or C above) are inconsistent with firm statements, assembly statistic values and/or lower assembly quality may help indicate whether a heterologous sample was submitted.

e. For a MS sample that is expected to be similar to other known, reasonably well characterized bacteria, the number of scaffolds in the assembly should generally be less than 400, the total length of the assembly should be approximately 10% of the expected genome size, and at least 95% of the assembly should be on contigs of 1000 bp or more in length.

f. For MS samples where the bioinformatic tool results in A, B, or C above are inconsistent with firm identity statements and sequence assembly statistics/quality are not as expected, further evaluation of the sample will be conducted.

g. Additional MS sample evaluation may include:

    i. Evaluation based on the organism and any special notes listed in Appendix I.

    ii. Consultation with experts with consideration of how well characterized the organism is in current databases and the percent genus match.

    iii. Requesting a contig BLAST against the RefSeq database to help determine if other strains or species are present.

    iv. Visually mapping contigs to a reference genome to identify problem sequence areas and unmapped reads.

    v. Long-read sequencing, such as Nanopore MinION sequencing of the MS sample.

    vi. Evaluating the sample preparation and extraction procedure for the specific MS sample submission for errors or deviations.

h. If the original MS sample submitted to the NVSL pipeline is suspected to be heterologous, the CVB may consider submitting the MS seed sample to the NVSL sequencing pipeline a second time after conducting a new DNA extraction for the MS sample.

E. When the MS sample has acceptable bacterial MLST results, ribosomal MLST results, and Kraken taxonomic sequence classification results combined with acceptable accessibly statistics and the CVB determined identity is in agreement with the firm identity statement, the test will be reported out in LSRTIS as **SATISFACTORY**.

F. When the MS sample has one or more item (bacterial MLST, ribosomal MLST, Kraken taxonomic sequence classification, or assembly statistics) that is not acceptable, a subject matter expert(s) (SME) will review available information and may also require additional testing/data analysis prior to a disposition determination. SMEs may consider the following:

a. Known analysis limitations
   i. Bacterial MLST profiles are not definitive for identification because genes can be shared between closely related bacterial groups.
   ii. *Enterobacterales* are poorly classified by Kraken due to shared genomic elements.

b. Contig Blast
   i. Considered acceptable if all matches >300 bp and >10X coverage depth are consistent with predicted Master Seed identity.
   ii. Possibly helpful when further identification is required/useful, e.g. strain or biovar determination.

c. Organism specific analysis
   i. Criteria referenced in peer-reviewed scientific literature may be utilized for an organism specific analysis.
   ii. Methods suggested per discussion with other experts.

d. Nanopore MinION sequencing of the MS sample
   i. The combination of short and long-read sequencing for the MS sample could be used to enhance genome completeness and accuracy if needed.

G. If the preponderance of evidence indicates that the identity of the MS sample is other than the firm submitted identity for the Master Seed or that the MS sample submitted to CVB is heterologous, the sample will be reported as **UNSATISFACTORY** in LSRTIS.

6. Report of Test Results

Test results are reported per standard operating procedures. A specific summary of percent identity from MLST schemes and Kraken results, allelic profiles, and results from additional analyses, such as contig BLASTs, should be included in the LSRTIS comments rather than in the results field unless the MS requires specification to a biovar or strain level that specifically is supported by an analysis. A summary of the testing report may be uploaded to the associated “Testing by Lab” Mail Log associated with the Master Seed (see CVB-WI-5279).

7. References

CVB-SOP-0154*, Recombinant Master Seed Testing*
CVB-SOP-5105, *Promega Maxwell RSC SOP*
CVB-WI-5279, *LSRTIS Master Seed Record - Sequencing and AMR*
CVB-PRO-5107, *Antimicrobial Resistance Genotype Detection*

Appendix I
Situational Specific Considerations

1. Obligate intracellular organisms

A low Kraken 2 Taxonomic Sequence Classification and poor assembly statistics are expected for such organisms. When MS DNA is isolated in the presence of a host (e.g. during infection of a host cell line) if possible, subtraction of host genomic DNA reads from the sequencing results should be conducted prior to data analysis of the MS sequence. As long as other contaminating organisms are not present, a VMO/Microbiologist/Biologist will use their best judgement, based on professional knowledge/experience, literature review, and/or consultation with other subject matter experts, to determine acceptance criteria for these MS.

2. Salmonella

The pubMLST database for Salmonella has known flaws at the serotype level. Other serotyping bioinformatics pipelines and manual serotyping should take precedence in final disposition.

3. Brucella

A full SNP analysis should be attempted on the MS sample for confirmation of species identity. This analysis is only available for select species/biovars. Note, the CVB does not conduct matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry analysis for *Brucella*.

4. Poorly annotated organisms

Fewer than 300 RefSeq entries suggest a poor sequence database. A VMO/Microbiologist/Biologist will use their best judgement, based on professional knowledge/experience, literature review, and/or consultation with other subject matter experts, to determine acceptance criteria for these MS.

5. Eukaryotic organisms

The NVSL sequencing pipeline is not optimized for eukaryotic organisms such a fungi or fungal-like organisms that have historically been submitted as MS candidates. Whole genome sequencing can still provide valuable information that can be used as a component of the identity disposition determination for these MS candidates. A VMO/Microbiologist/Biologist will use their best judgement, based on professional knowledge/experience, literature review, and/or consultation with other subject matter experts, to determine acceptance criteria for these MS.