Application of Whole Genome Sequencing for the identification and characterization of a *Vibrio parahaemolyticus* outbreak strain

Kun Connie Liu*, Wen-Hsin Cindy Wu and Jinxin Hu

* Email: kun.liu@fda.hhs.gov; Phone: 1 425 487 5388

The Pacific Northwest Laboratory, Office of Regulatory Science, Office of Regulatory Affairs, the U. S. Food and Drug Administration (FDA)

Abstract

With advancements in modern technology and reduction in costs, whole genome sequencing (WGS) has become an available powerful tool for microbial identification and characterization. In ORS laboratories, WGS is routinely performed for field isolates of *Salmonella* species, *Listeria monocytogenes*, and Shiga toxin-producing *E. coli* (STEC). This study aims to extend WGS analysis for the identification and characterization of *Vibrio parahaemolyticus*, a foodborne bacterial pathogen with steadily increasing human infections. In this work, WGS and bioinformatics analysis were performed for an archived *Vibrio parahaemolyticus* outbreak strain R10-B2-71. After bacterial DNA extraction and library preparation, shotgun WGS was performed on an Illumina MiSeq instrument following recommended protocols. During data analysis, a draft genome was determined using a *de novo* assembler, and genome characteristics were analyzed with various bioinformatics tools. A total of 1,461,426 passing-filter reads were generated by the MiSeq with an estimated coverage of 554×. After primer trimming and decontamination, the genome of R10-B2-71 contains 5,270,223 bp with a GC content of 45.0%. There are 550 contigs longer than 200 bp with a N50 of 130,921 bp. A total of 4741 open reading frames (ORF) were annotated, including 105 virulence associated genes and 115 RNAs. SpeciesFinder 2.0 analysis based on WGS data correctly identified the species of this strain, consistent with previous lab analytical results using standard assays and the bioMérieux VITEK MS ID system. The genome assembly was submitted to GenBank and assigned an accession number GCA_001704915.1. These results have shown the potential of WGS and genomic analysis for identification and characterization of *Vibrio parahaemolyticus* as great tools to prepare for future outbreaks and improve public health.

Introduction

Vibrio parahaemolyticus is a leading cause of foodborne illnesses associated with the consumption of raw shellfish, responsible for approximately 35,000 infections each year in the United States. The average annual incidence of *Vibrio* infections increased steadily by 54% between 2006 and 2017. It is believed that the transition of *V. parahaemolyticus* caused illnesses from a regional to a national pattern is connected to the emergence of pathogenic isolates with epidemic potential. A study by Centers for Disease Control and Prevention (CDC) has identified that international trade of shellfish may be involved in the dispersal of *V. parahaemolyticus* populations into the United States nationwide.

The application of next-generation sequencing (WGS) is critical for the US Food and Drug Administration (FDA) to identify, characterize, and track foodborne bacterial pathogens. Previously *V. parahaemolyticus* analyses include traditional microbial assays, Multi-Locus Sequence Typing (MLST), and Pulsed Field Gel Electrophoresis (PFGE). WGS provides great resolution of bacterial genome, however, it has not been widely used to identify or characterize *Vibrio* species compared to the extent on studying *Salmonella*, *E. coli*, or *Listeria monocytogenes* at FDA regulatory labs. Here, the shotgun WGS, *de novo* assembly, and genomic characterization of a *V. parahaemolyticus* outbreak strain are reported. This strain was isolated from oyster in the Washington State during an outbreak in 1997. The study demonstrates the usefulness of WGS and downstream genomic analyses for identification and characterization of *V. parahaemolyticus* during investigations of foodborne outbreaks to safeguard public health.

Materials and Methods

Bacteria strain and media

An archived *Vibrio parahaemolyticus* strain R10-B2-71 was previously isolated from oyster in Washington State during an outbreak in 1997. The isolation and characterization methods were performed following FDA Bacteriological Analytical Manual (BAM) Chapter 9. Bacteria were cultivated with Trypticase soy agar (TSA) with 5% sheep blood for VITEK MS analysis and Brain heart infusion (BHI) broth for WGS as describe below following growth conditions specified in BAM Chapter 9.

Species confirmation with the VITEK MS system

Sample preparation for VITEK MS analysis was performed following instrument manual. Spectra of matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI TOF) were analyzed with the VITEK MS *in vitro* diagnostics database (IVD) v3.2.0 and automatically reported in the integrated MYLA tool (BioMérieux, Marcy-l'Étoile, France). The strain used for quality control (QC) was *E. coli* American Type Culture Collection (ATCC) 8739 following manufacturer's instructions. For unambiguous species identification, the threshold of confidence scores were set to equal or greater than 99 per manufacturer's recommendation.

Genomic DNA extraction, WGS, and bioinformatic analyses

Genomic DNA was extracted with a QIAcube (QIAGEN Inc., Valencia, CA) and quantified with a Qubit 3.0 fluorimeter and a dsDNA BR Assay Kit (ThermoFisher Scientific Inc., Waltham, MA). The library was prepared with a Nextera XT kit (Illumina, San Diego, CA) following manufacturer's instructions. The genome was sequenced on MiSeq with 12 samples per flow cell in 2 x 251 sequencing cycles. Sequencing quality metrics were obtained using QUAST without reference genome and Illumina Sequencing Analysis Viewer 1.8.37. After on-instrument primer trimming, *de novo* assembly was performed with CLC Genomics Workbench v9.0 (Qiagen Inc.). Contigs shorter than 200bp were removed and deposited to GenBank after passing the NCBI contamination screening pipeline. Annotation was performed with the Rapid Annotation using Subsystem Technology (RAST) server. Bioinformatic analyses were operated using various tools on the GalaxyTrakr and Center for Genomic Epidemiology (CGE) websites.

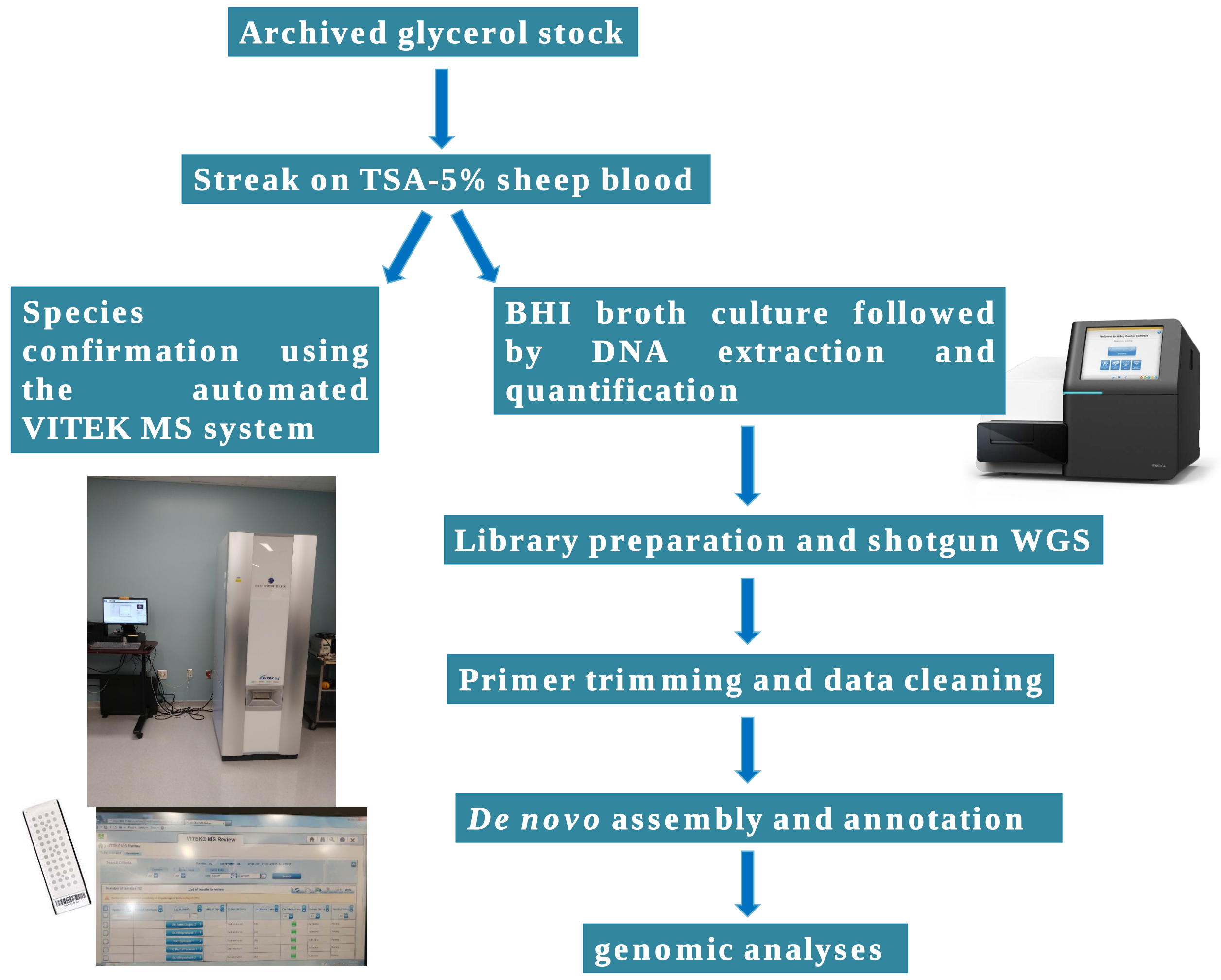

Archived glycerol stock
Streak on TSA-5% sheep blood
Species confirmation using the automated VITEK MS system
BHI broth culture followed by DNA extraction and quantification
Library preparation and shotgun WGS
Primer trimming and data cleaning
De novo assembly and annotation
genomic analyses

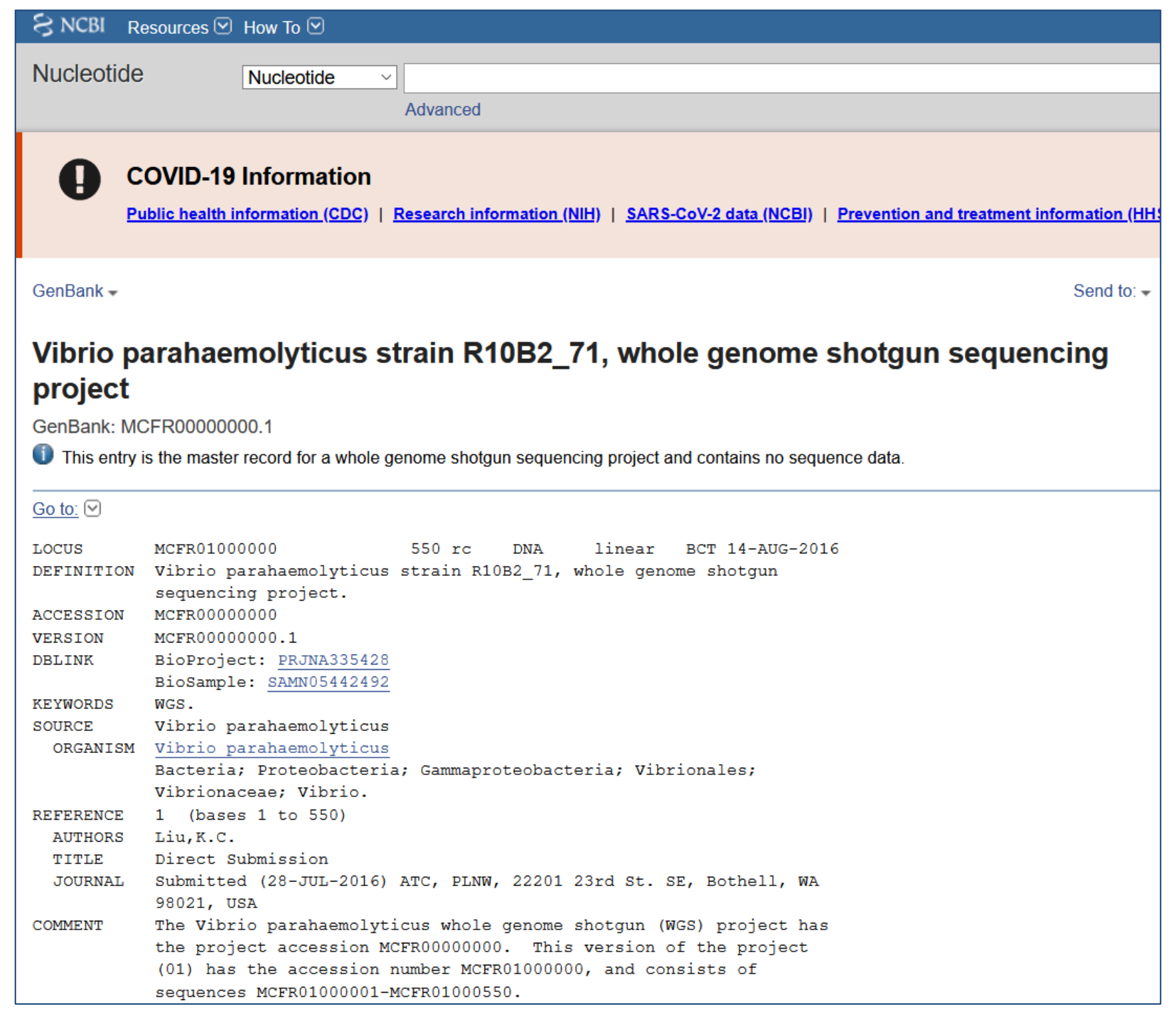
Figure 1. A brief workflow of the present study.

Results and Discussion

- The strain species was verified with VITEK MS (Table 1).
- Shotgun WGS was completed on a MiSeq (Table 2, lanes 1-3).
- A draft genome was *de novo* assembled and deposited to the National Center for Biotechnology Information (NCBI) GenBank after passing the contamination screening pipeline (Figure 2 and Table 2).
- Genome annotation identified virulence associated genes (Figure 3). WGS-based speciation was consistent with the lab result based on biochemical assays (Figure 4). There was no plasmid recognized (Table 2). The MLST sequence type was determined to be 1556 (Figure 5).

Table 1. Species confirmation with the VITEK MS microbial ID system.

Strain ID	VITEK MS biological replicate 1		VITEK MS biological replicate 2		VITEK MS species ID
	species	confidence score	species	confidence score	
R10-B2-71	*V. parahaemolyticus*	99.9	*V. parahaemolyticus*	99.9	*V. parahaemolyticus*
QC: ATCC 8739	*E. coli*	99.9	NA	NA	QC satisfactory



COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)

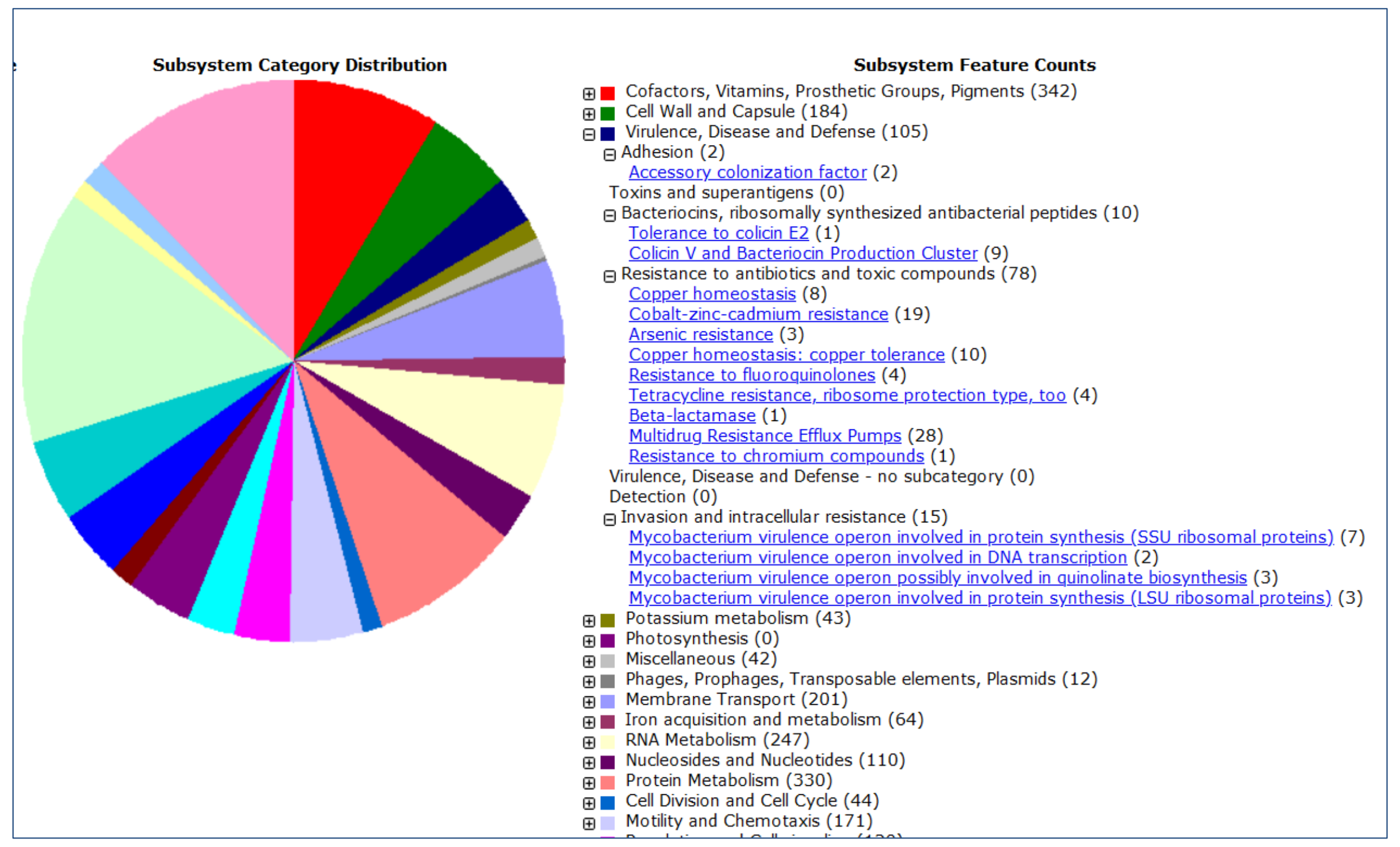
Vibrio parahaemolyticus strain R10B2_71, whole genome shotgun sequencing project

GenBank: MCFR00000000.1

This entry is the master record for a whole genome shotgun sequencing project and contains no sequence data.

```
LOCUS       MCFR01000000             550 rc    DNA     linear   BCT 14-AUG-2016
DEFINITION  Vibrio parahaemolyticus strain R10B2_71, whole genome shotgun
            sequencing project.
ACCESSION   MCFR00000000
VERSION     MCFR00000000.1
DBLINK      BioProject: PRJNA335428
            BioSample: SAMN05442492
KEYWORDS    WGS.
SOURCE      Vibrio parahaemolyticus
  ORGANISM  Vibrio parahaemolyticus
            Bacteria; Proteobacteria; Gammaproteobacteria; Vibrionales;
            Vibrionaceae; Vibrio.
REFERENCE   1  (bases 1 to 550)
  AUTHORS   Liu,K.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-JUL-2016) ATC, PLNW, 22201 23rd St. SE, Bothell, WA
            98021, USA
COMMENT     The Vibrio parahaemolyticus whole genome shotgun (WGS) project has
            the project accession MCFR00000000.  This version of the project
            (01) has the accession number MCFR01000000, and consists of
            sequences MCFR01000001-MCFR01000550.
```

Figure 2. The *de novo* assembly deposited to NCBI GenBank.

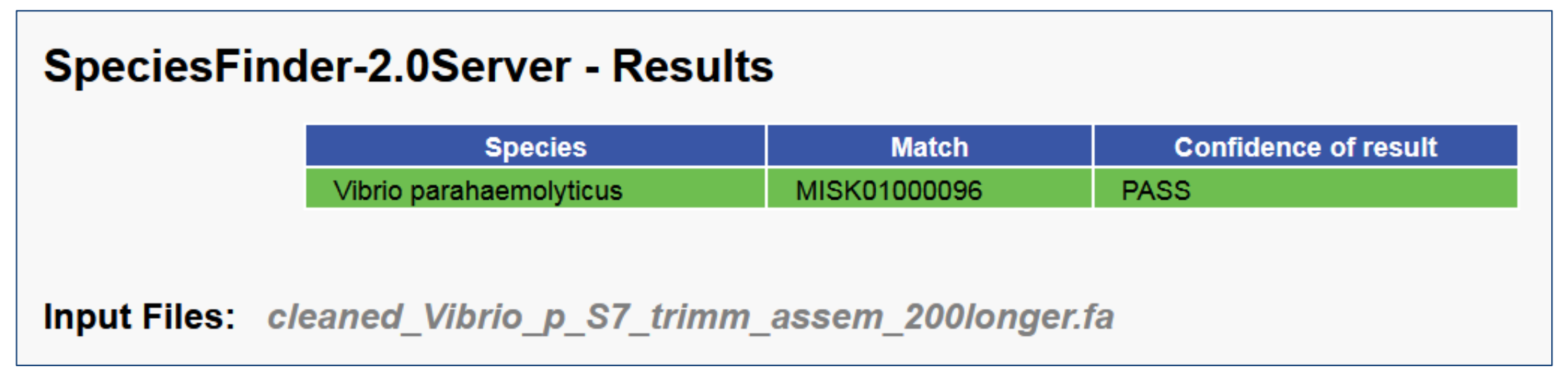


- Cofactors, Vitamins, Prosthetic Groups, Pigments (342)
- Cell Wall and Capsule (184)
- Virulence, Disease and Defense (105)
  - Adhesion (2)
    - Accessory colonization factor (2)
  - Toxins and superantigens (0)
  - Bacteriocins, ribosomally synthesized antibacterial peptides (10)
    - Tolerance to colicin E2 (1)
    - Colicin V and Bacteriocin Production Cluster (9)
  - Resistance to antibiotics and toxic compounds (78)
    - Copper homeostasis (8)
    - Cobalt-zinc-cadmium resistance (19)
    - Arsenic resistance (3)
    - Copper homeostasis: copper tolerance (10)
    - Resistance to fluoroquinolones (4)
    - Tetracycline resistance, ribosome protection type, too (4)
    - Beta-lactamase (1)
    - Multidrug Resistance Efflux Pumps (28)
    - Resistance to chromium compounds (1)
  - Virulence, Disease and Defense - no subcategory (0)
  - Detection (0)
  - Invasion and intracellular resistance (15)
    - Mycobacterium virulence operon involved in protein synthesis (SSU ribosomal proteins) (7)
    - Mycobacterium virulence operon involved in DNA transcription (2)
    - Mycobacterium virulence operon possibly involved in quinolinate biosynthesis (3)
    - Mycobacterium virulence operon involved in protein synthesis (LSU ribosomal proteins) (3)
- Potassium metabolism (43)
- Photosynthesis (0)
- Miscellaneous (42)
- Phages, Prophages, Transposable elements, Plasmids (12)
- Membrane Transport (201)
- Iron acquisition and metabolism (64)
- RNA Metabolism (247)
- Nucleosides and Nucleotides (110)
- Protein Metabolism (330)
- Cell Division and Cell Cycle (44)
- Motility and Chemotaxis (171)

Figure 3. Genome annotation using the RAST server. The virulence, disease and defense associated genes are expanded on the right column.

Table 2. WGS data summary and statistics of genomic analyses.

Samples per flow cell	12
Q score (all cycles in the run)	>=Q30: 7.3 G (85.7%)
Passing filter reads	1,461,426
Genome coverage	554x
GC%	45.0%
Genome Length (bp)	5,270,223
Contig number (>=200 bp)	550
Contig N50 (bp)	130,921
GenBank nucleotide sequence #	MCFR00000000.1
GenBank assembly #	GCA_001704915.1
Annotated coding sequences	4741
Annotated RNA	115
annotated virulence-associated genes	105
Plasmid identified	0

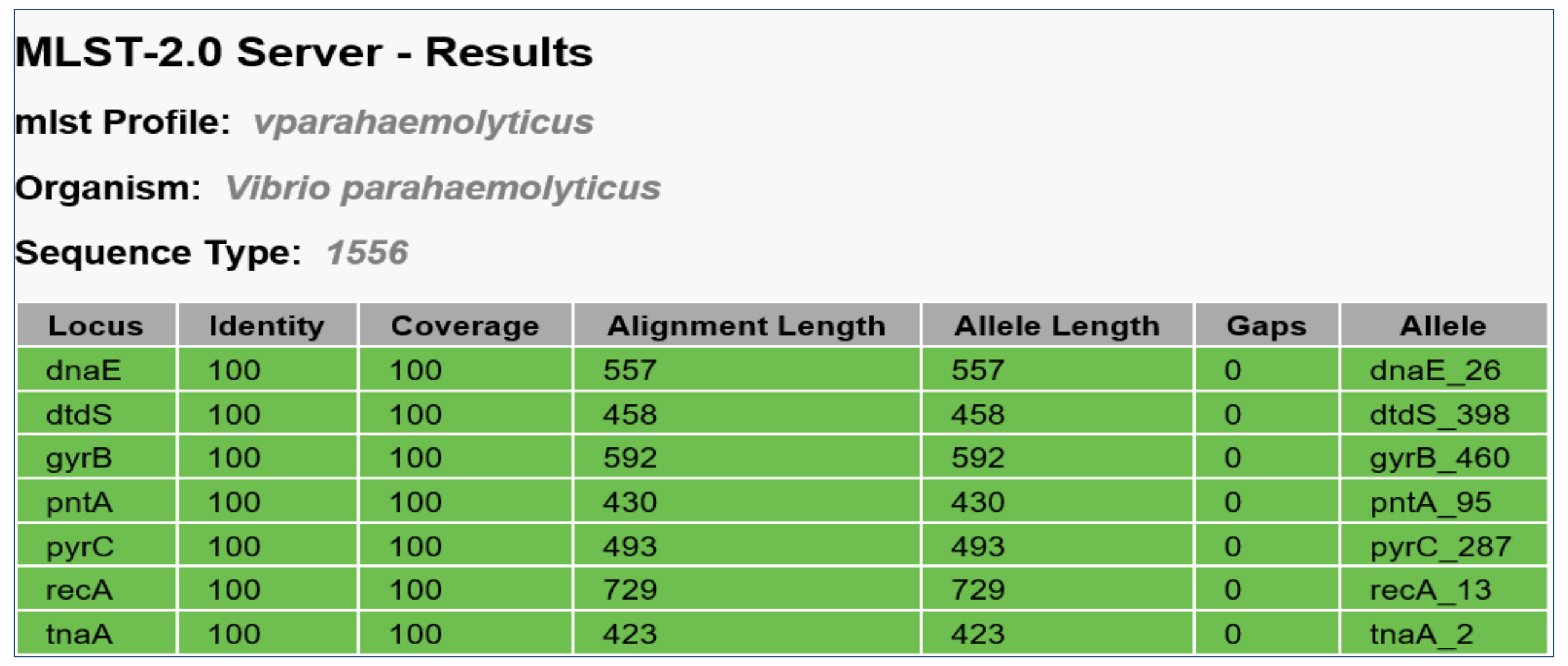


SpeciesFinder-2.0Server - Results

Species	Match	Confidence of result
Vibrio parahaemolyticus	MISK01000096	PASS

Input Files: *cleaned_Vibrio_p_S7_trimm_assem_200longer.fa*

Figure 4. Speciation using WGS data based SpeciesFinder 2.0.



MLST-2.0 Server - Results

mlst Profile: *vparahaemolyticus*

Organism: *Vibrio parahaemolyticus*

Sequence Type: *1556*

Locus	Identity	Coverage	Alignment Length	Allele Length	Gaps	Allele
dnaE	100	100	557	557	0	dnaE_26
dtdS	100	100	458	458	0	dtdS_398
gyrB	100	100	592	592	0	gyrB_460
pntA	100	100	430	430	0	pntA_95
pyrC	100	100	493	493	0	pyrC_287
recA	100	100	729	729	0	recA_13
tnaA	100	100	423	423	0	tnaA_2

Figure 5. Sequence typing using MLST 2.0 on the CGE server.

Conclusion

This study evaluated the applications of whole genome sequencing and analyses in the identification and characterization of a *Vibrio parahaemolyticus* outbreak strain R10-B2-71. The speciation result based on WGS data was consistent with two standard biochemical methods. The research outcomes show that WGS analyses can be applied to speciate and characterize *V. parahaemolyticus* field isolates, and will greatly facilitate regulatory science on strain characterization and outbreak investigation.

Acknowledgement

The authors thank Ms. Karen Jinneman for scientific discussion, Ms. Khamphet Nabe for assistance on sample information, and the PNL Microbiology Branch for technical support.

Disclaimer

Reference to any commercial materials or equipment does not, in any way, constitute approval, endorsement, or recommendation by the FDA. All views and opinions expressed in the presentation are those of the presenter and do not necessarily represent views or official position of the FDA.