

THE ISOLATION, CHARACTERIZATION, AND IDENTIFICATION OF A NOVEL  
SPECIES OF BACTERIUM IN THE ENTEROBACTERIACEAE FAMILY FROM KEPHART  
PRONG, GREAT SMOKY MOUNTAINS NATIONAL PARK

A thesis presented to the faculty of the Graduate School of Western Carolina University in  
partial fulfillment of the requirements for the degree of Master of Science in Biology

By

Lisa Marie Dye

Director: Dr. Seán O'Connell  
Associate Professor and Department Head  
Department of Biology

Committee Members: Dr. Sabine Rundle, Biology  
Dr. Malcolm Powell, Biology

April 2017

## ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Seán O’Connell, for allowing me the honor of working in his lab. He took me in when no one else would, and he didn’t mind that I didn’t have any wet lab skills. This work would not have been possible without him. I would also like to thank my committee members, Dr. Malcolm Powell and Dr. Sabine Rundle for their guidance and assistance. I would like to thank Dr. Kathy Mathews for her patience and time in helping me with the phylogenetic analysis. I would like to thank Kacie Fraser, Tori Carlson, and Rob McKinnon for their help in the lab. Also, I would like to thank the Biology Department of Western Carolina University for the use of equipment and funding, along with the Graduate School of Western Carolina University for funding. Finally, I would like to give special thanks to Dr. Barry G. Hall, author of *Phylogenetic Trees Made Easy*. He was invaluable in teaching me, long distance, how to use MEGA7.

## TABLE OF CONTENTS

LIST OF TABLES .....	iv
LIST OF FIGURES .....	v
ABSTRACT .....	vi
INTRODUCTION .....	1
Importance of Microorganisms .....	1
Taxonomic Classification of Microorganisms .....	2
Methods of Classification .....	5
All Taxa Biodiversity Inventory .....	6
Purpose .....	8
MATERIALS AND METHODS .....	9
Sampling at Kephart Prong in GSMNP .....	9
Phenotypic Characterization .....	9
DNA Extraction, PCR, and Sequencing .....	10
Genomic and Phylogenetic Characterization .....	11
RESULTS .....	15
Phenotypic Results .....	15
Genotypic Results .....	18
Phylogenetic Results .....	25
DISCUSSION .....	31
Bacterial Taxonomy – A Polyphasic Approach .....	31
Phenotypic Analysis .....	33
Genotypic Analysis .....	34
Phylogenetic Analysis .....	40
CONCLUSIONS AND POSSIBLE FUTURE WORK .....	45
REFERENCES .....	50
APPENDIX .....	58

## LIST OF TABLES

Table 1. Environmental parameters and phenotypic characteristics of isolate LD2.....	16
Table 2. Differential characteristics of LD2 and closely related genera.....	17
Table 3. DNA sequence matches for a bacterial isolate .....	19
Table 4. MLST results using the WGS of LD2 and the scheme for <i>Yersinia</i> spp .....	20
Table 5. MLST results using the WGS of LD2 and the scheme for <i>Y. ruckeri</i> .....	21
Table 6. List of significant genes and their putative function.....	22

## LIST OF FIGURES

Figure 1. Primary and secondary structures of 16S rRNA from <i>Escherichia coli</i> .....	4
Figure 2. Evolutionary relationships and the phylogenetic tree of life .....	5
Figure 3. Map of Great Smoky Mountains National Park. ....	7
Figure 4. Species level results of One Codex analysis .....	23
Figure 5. Genus level results of One Codex analysis .....	24
Figure 6. Phylogram for “Enterobacteriales” showing distances and the placement of LD2.....	26
Figure 7. Most likely tree for “Enterobacteriales” .....	27
Figure 8. Most likely tree for Enterobacteriaceae .....	28
Figure 9. Branch lengths and bootstrap values for Enterobacteriaceae .....	29
Figure 10. Majority rule bootstrap consensus tree for Enterobacteriaceae.....	30

## ABSTRACT

### THE ISOLATION, CHARACTERIZATION, AND IDENTIFICATION OF A NOVEL SPECIES OF BACTERIUM IN THE ENTEROBACTERIACEAE FAMILY FROM KEPHART PRONG, GREAT SMOKY MOUNTAINS NATIONAL PARK

Lisa Marie Dye, M.S. in Biology

Western Carolina University (April 2017)

Director: Dr. Seán O’Connell

The purpose of this study was to examine a single bacterial species isolated from Great Smoky Mountains National Park (GSMNP), characterize its growth requirements, and identify it down to the species level. A polyphasic approach that examined phenotypic, genotypic, and phylogenetic characteristics was used. Phenotypic analysis revealed that the isolate is Gram-negative, rod-shaped, non-motile, oxidase negative, catalase positive, and grows in the presence and absence of oxygen. Growth was observed at temperatures ranging from 4°C to 37°C, with optimum growth at 30°C based on visual observation of colony mass. The pH range for growth was pH7-9, with optimum growth at pH9 based on visual observation of colony mass. The isolate can tolerate up to 1% NaCl in the nutrient media. Genotypic analysis utilizing 16S rDNA sequences and whole genome sequencing (WGS) identified the isolate as a member of the order “Enterobacteriales” and the family Enterobacteriaceae. Phylogenetic analysis supported the isolate’s position in both taxa, but did not cluster the isolate with any specific genera. On the basis of phenotypic, genotypic, and phylogenetic properties, the isolate LD2 represents a novel species of a new genus.

## INTRODUCTION

### **Importance of Microorganisms**

Microorganisms have enormous ecological, medical, and practical importance. They are beneficial to the environment, the food industry, biofuel production, bioremediation, industrial microbiology, biotechnology, and human welfare (Madigan et al. 2015). Agriculture benefits from the cycling of nutrients by microorganisms. Nitrogen fixing bacteria convert atmospheric nitrogen into ammonia that the plants can use as a nitrogen source (Berersen and Turner 1968). Other bacteria are instrumental in the sulfur cycle (Pfennig and Widdel 1982). Microorganisms also inhabit the rumen of animals such as cows (Bryant 1959). They convert the cellulose from grass into fatty acids that can be used by the animal. There are other microorganisms that inhabit the human gastrointestinal tract which assist in digestion and vitamin synthesis. The very oxygen we breathe is, in part, the result of microbial activity. In the absence of microorganisms, higher life forms could not be sustained (Madigan et al. 2015).

There are also negative effects of microorganisms that are very important. The primary harmful effects of microbes upon our existence and civilization is that they are an important cause of disease in animals and crop plants, and they are agents of spoilage and decomposition of our foods, textiles and dwellings. Certain types of bacteria can cause human diseases, such as, typhoid fever, syphilis, cholera, and tuberculosis (Madigan et al. 2015). Given the vast array of microbial influences, discovery and identification of new bacterial species could lead to future benefits from unique products that might help humans or our planet. Discovery of new pathogens can help us prepare to ward off future infection and infestations.

## **Taxonomic Classification of Microorganisms**

Carl Linnaeus, also known as Carolus Linnaeus, is often referred to as the Father of Taxonomy. Part of Linnaeus' innovation was the grouping of species into higher taxa that were based on shared morphological similarities. In Linnaeus' original system, species were grouped into genera, genera were grouped into orders, orders into classes, and classes into kingdoms. Kingdom was the highest level of classification and there were only two: plants and animals (Linnaeus 1756). This two-kingdom system persisted even after the discovery of the diverse microbial world. Taxonomists simply placed bacteria in the plant kingdom after discovering they possessed a rigid cell wall (Reece et al. 2011).

In 1866, Ernst Haeckel formally challenged the plant/animal division of the living world. He recognized that singled-celled life forms called protists did not fit either category. Haeckel depicted the tree of life as having three main branches: Plantae, Protista, and Animalia (Haeckel 1866). Taxonomic schemes with more than three kingdoms started to appear around 1957 with the work of one of the most influential ecologists of his time, Robert Whittaker. He started with a three-kingdom system, but over the course of ten years of critical reflection, he refined his system based on cell biology and the distinction between prokaryotes and eukaryotes. He eventually arrived at a five-kingdom system that became a standard feature of biology textbooks. His five kingdoms were: Monera (prokaryotes), Protista (a diverse kingdom consisting mostly of unicellular organisms), Plantae, Fungi, and Animalia. This system set the prokaryotes apart from the eukaryotes by placing them in their own kingdom. Biologists and educators found this system attractive because it seemed to capture the fundamental properties of living organisms (Hagen 2012).

In 1977, Carl Woese postulated a revolutionary new taxonomic scheme based on phylogenetic relationships rather than visible morphological similarities. He used the small-subunit rRNA gene (16S rRNA of bacteria and 18S rRNA of eukaryotes) as a universal marker for phylogenetic reconstruction (Albers et al. 2013, Fox et al. 1977). The 16S rRNA gene is a section of prokaryotic DNA found in all bacteria and archaea. This gene codes for an rRNA which makes up part of the ribosome. The ribosome is composed of two subunits, the large subunit (LSU) and the small subunit (SSU). These two subunits sandwich the mRNA as it feeds through the ribosome for translation. Woese realized that rRNA genes make excellent candidates for phylogenetic analysis because they are (1) found in all known life forms, (2) functionally constant, and (3) highly conserved (Madigan et al. 2015). Their highly conserved nature is due to their important function of translating mRNA into proteins. However, there are portions of the genes that are more conserved than others. This is due to the structure of the ribosome itself. The RNA strand creates bonds with itself in some places (conserved regions) while other portions are looped and unbounded (hypervariable regions). The hypervariable regions have been more tolerant of mutations over time, and have therefore accumulated more changes within the nucleotide sequence (Figure 1).

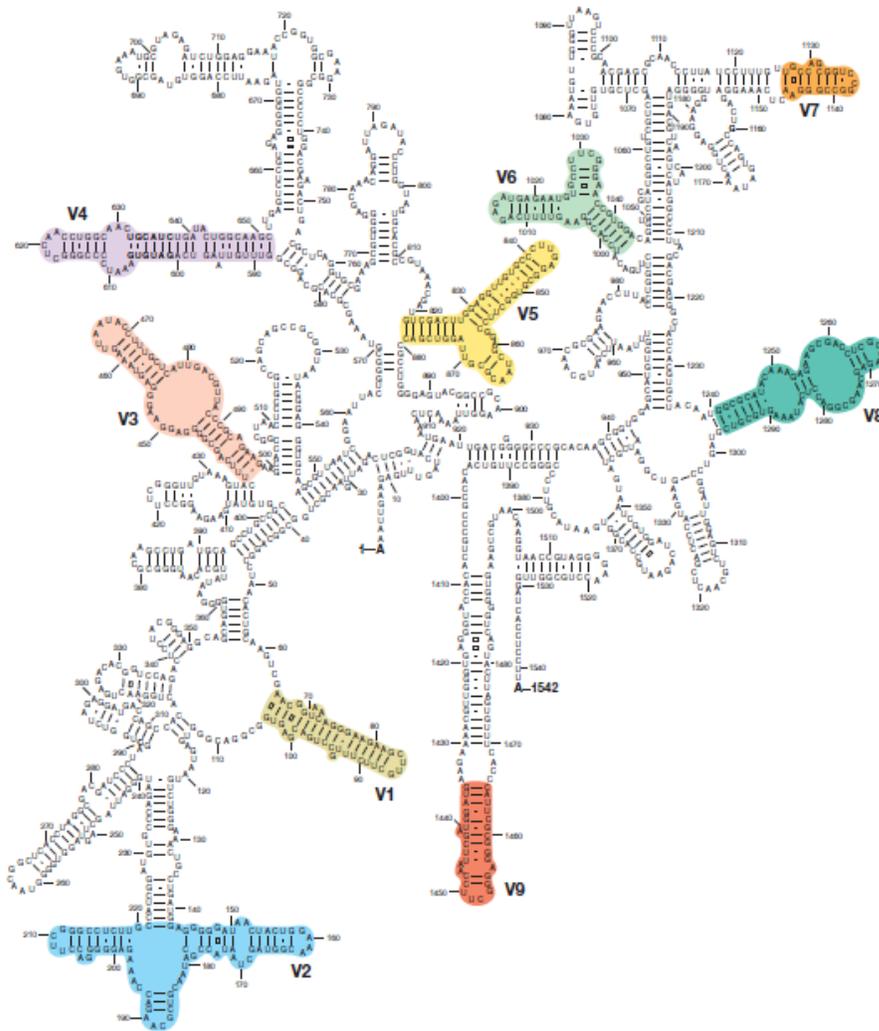


Figure 1. Primary and secondary structures of 16S rRNA from *Escherichia coli*. The molecule is composed of conserved and variable regions. The positions of the variable regions are indicated in color (Madigan et al. 2015).

Woese discovered that within the kingdom Monera existed two distinct groups of organisms that were no more related to one another than they were to eukaryotes. To remedy this situation, he proposed that a new formal system of taxonomy be established in which, above the level of kingdom, there exists a new taxon called a domain. His three domains were Bacteria, Archaea, and Eucarya, now known as Eukarya (Figure 2; Woese et al. 1990).

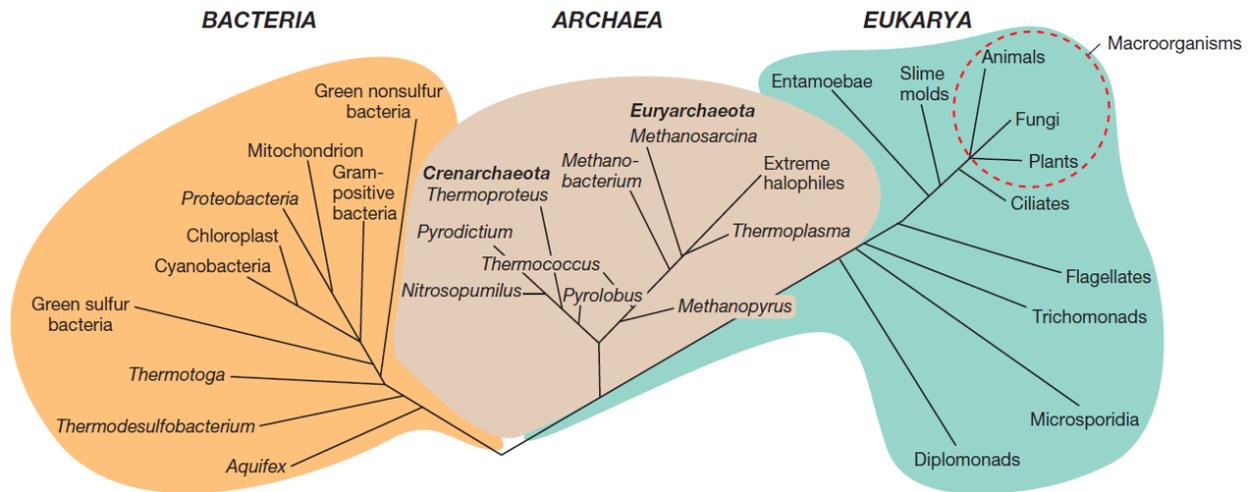


Figure 2. Evolutionary relationships and the phylogenetic tree of life defined by rRNA gene sequencing showing the three domains of life: Bacteria, Archaea, and Eukarya. Only a few representative groups are shown in each domain (Madigan et al. 2015).

### Methods of Classification

There are basically two methods of determining bacterial diversity in soil and water samples. The first, and oldest, is culture-based techniques used to characterize phenotypic aspects of the isolate. This method measures various morphological, metabolic, physiological, and environmental parameters that can lead to a better understanding of how the microorganism functions (O'Connell et al. 2007). Culturing a microorganism remains the only way to fully characterize its properties and predict its impact on an environment (Madigan et al. 2015). One of the shortcomings of culture-based techniques is that there is a large discrepancy between the number of bacterial colonies that form on solid media and the total number of bacterial cells actually present in the sample (Joseph et al. 2003). It has been estimated that only 0.1-1% of soil bacteria are accessible by conventional culture-dependent techniques, which leaves most of the phylogenetic diversity unstudied (Zhang and Xu 2008).

The second way to estimate bacterial diversity is through culture-independent genetic analyses of microbial communities. Carl Woese and others pioneered the analysis of various bacteria using DNA sequencing, specifically the 16S rDNA genes (Fox et al. 1977). The invention of PCR and automated DNA sequencing has led to the accumulation of a large amount of sequence data on the rDNA genes of many organisms (Woo et al. 2008). A drawback to this method is that it does not usually reveal the metabolic, physiological, and biochemical activities of the organism that help define its role in the environment (O'Connell et al. 2007).

### **All Taxa Biodiversity Inventory**

Great Smoky Mountains National Park (GSMNP) is a reserve that is located on the mountainous border between North Carolina and Tennessee. It is roughly 2200 km<sup>2</sup> in size and is considered one of the most biologically diverse areas in the temperate zone. It is designated as an International Biosphere Reserve and a World Heritage Site, and it is the largest federally protected area in the eastern United States (Nichols and Langdon 2007).

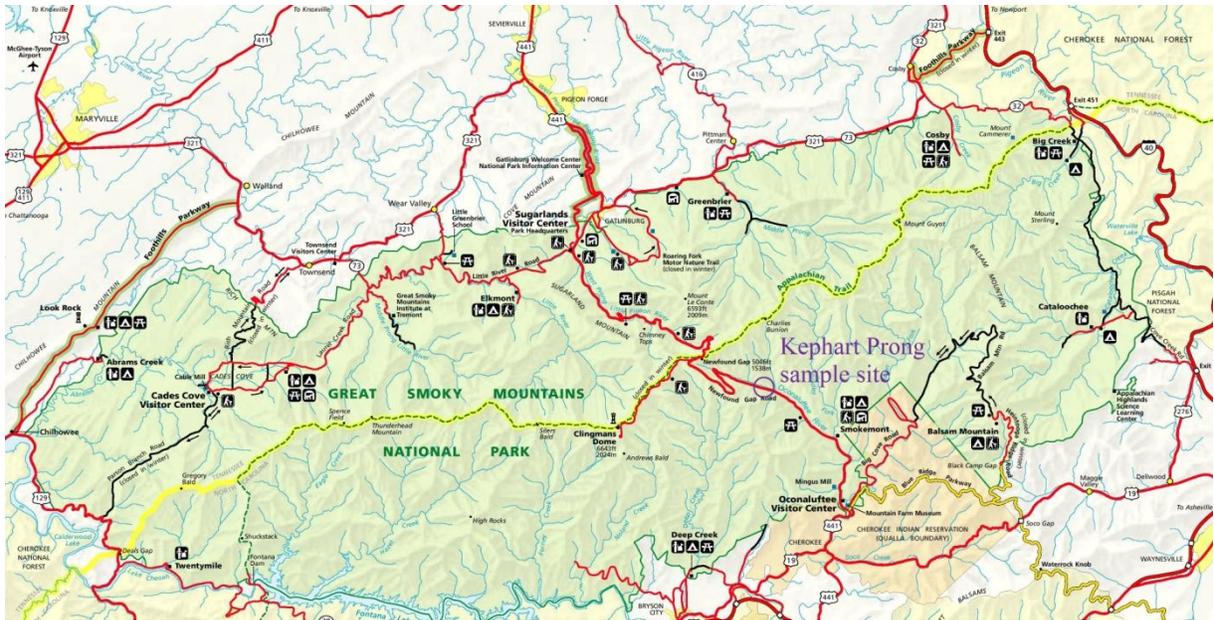


Figure 3. Map of Great Smoky Mountains National Park (GSMNP) showing the location of the sampling site at Kephart Prong trail (National Park Service 2017).

The mission of the National Park Service (2015) is to preserve the natural resources of the park. How can we preserve our resources if we do not even know what kinds of resources we have, where they are found, their rarity, or have an understanding of their ecological role? (White and Langdon 2006). Taking an inventory would appear to be a precursory requirement to resource management. In 1997, a group of natural resource professionals decided to correct this problem by implementing an All Taxa Biodiversity Inventory (ATBI) in GSMNP (White and Langdon 2006). The concept of an ATBI is to determine all species present within a defined area within a certain time frame. The goal of this effort is to collect and disseminate useful information of all species including 1) documenting as many species as possible, 2) mapping the general distribution for each taxon in the give area, 3) determining the relative abundance for each taxon, and 4) compiling a synopsis of each taxon’s natural history and ecological role (Nichols & Langdon 2007). The ATBI currently being conducted in GSMNP includes

microorganisms. It has been estimated that there may be as many as 10 million species in 10 grams of soil (Gans et al. 2005). Unfortunately, only a few thousand species of prokaryotes in the world have been identified and classified (Janssen 2006). In 2002, O'Connell et al. (2007) began the task of inventorying bacteria in the soil and streams of GSMNP. They documented 69 genera from eleven different phyla. The waters were dominated by species from Bacteroidetes, while soils had more cultured representatives from the phylum Firmicutes (O'Connell et al. 2007). Inventorying microbial species may seem like an insurmountable task given the sheer numbers previously estimated. However, these smallest of life forms are not inconsequential; they constitute the bulk of Earth's biomass and are key reservoirs of essential nutrients for life (Madigan et al. 2015).

### **Purpose**

The purpose of this study was to examine a single bacterial species isolated from GSMNP, characterize its growth requirements, and identify it down to the species level using 16S rDNA sequence analysis. Based on early Ribosomal Database Project (RDP) results, I hypothesized that I have discovered a new species of bacterium. Study of the biochemical capabilities and genomic characteristics of this species could help define its role in the environment and lead to better management of the soils and waters in this reserve.

## MATERIALS AND METHODS

### **Sampling at Kephart Prong in GSMNP**

Water samples were aseptically taken from a stream near the Kephart Prong trail in GSMNP (GPS N35.35.222 W83.21.431) in September of 2015. A sterile centrifuge tube was submerged in the stream and filled to 80% volume. The sample was then placed in a cooler on ice and returned to the laboratory where it was kept at 4°C until culture work began (O'Connell et al. 2007).

The sample was then serially diluted in 0.85% sterile saline using 10 fold dilutions from  $10^{-1}$  through  $10^{-6}$ . 100µL of each dilution was spread-plated on R2A media plates. The plates were then incubated in the dark at room temperature. After one week, all plates were assessed for growth and a single colony was selected and isolated. The isolated colony was streaked onto R2A agar plates. This process was repeated 4 to 6 times until a pure culture representing a single species was obtained (O'Connell et al. 2007). The isolate is referred to as LD2 in this thesis.

### **Phenotypic Characterization**

Routine cultivation was conducted at 25°C with R2A medium unless indicated otherwise. The Gram reaction was determined using a Gram staining kit according to the manufacturer's instructions. The isolate was negative stained and observed under a light microscope to determine cell size and morphology. It was re-streaked onto nutrient agar plates with 1%, 5%, and 10% w/v NaCl to test for salinity tolerance, and pH3, pH5, and pH9 plates to test for pH tolerance. Metabolic tests were performed on the isolate including an anaerobic test, a catalase test, and an oxidase test. Temperature tests were performed at 4°C, 25°C, 37 °C, and 50 °C to determine the growth temperature range for the isolate. A motility test was performed using

semisolid agar and tetrazolium dye, as well as the hanging drop method. Flagella staining was performed using a kit from Presque Isle Cultures (Erie, PA) as well as the Ryu method (Heimbrook 1989).

The isolate was tested using an EnteroPluri-*Test*® (Liofilchem, Italy) which is a multiple test system designed to identify bacteria based on their ability to: ferment glucose, adonitol, lactose, arabinose, dulcitol, and sorbitol; decarboxylate lysine and ornithine; reduce sulfur; produce hydrogen sulfide, indole, and acetoin; deaminate phenylalanine; hydrolyze urea; and utilize citrate. The tubes were inoculated then incubated for 48 hours at 25°C. All environmental and biochemical tests were performed in triplicate.

### **DNA Extraction, PCR, and Sequencing**

DNA was extracted from the isolate using the Ultra Clean Microbial DNA Isolation Kit (Mo Bio Inc., Solana Beach, CA) modified for use with a bead beater (2,500 RPM for 60 seconds). Partial gene fragments for 16S rRNA were targeted using PCR for sequencing and identification purposes. The total reaction consisted of 50µL with the following ingredients: 2.5X MasterMix (Promega Corporation, Madison, WI) diluted with water to 1X concentration, 0.25µM each primer (bacterial-specific primers 341F and 907R, and 1µL DNA template.)

A touchdown PCR approach was employed, consisting of 5 minutes of initial denaturation at 94°C, then 30 cycles of denaturation at 94°C for 1 minute, annealing for 1 minute, and extension for 3 minutes at 72°C. The annealing temperature in the first two rounds was at 65°C, followed by one round each at 1°C lower than the previous round, and finally eighteen rounds at 55°C. A final extension for 7 minutes at 72°C was employed, and then amplicons were stored at 4°C until they could be sequenced (S. O'Connell, personal communication).

The Big Dye Terminator Version 3.1 Cycle Sequencing Kit and 3130 Automated Sequencer (Applied Biosystems, Inc., Foster City, CA) were used to sequence a partial section of DNA that codes for the 16S rRNA of the isolate LD2. At a later time, the entire 16S rRNA gene was sequenced by the company Genewiz (South Plainfield, NJ).

The entire genome was sequenced by MR DNA (Molecular Research LP, Shallowater, TX). The genomic library was prepared using Nextera DNA Sample preparation kit (Illumina) following the manufacturer's user guide. The initial concentration of DNA was evaluated at 144.40 ng/ $\mu$ L using the Qubit® dsDNA HS Assay Kit (Life Technologies). Fifty ng DNA was used to prepare the library. The samples underwent the simultaneous fragmentation and addition of adapter sequences. These adapters are utilized during a limited-cycle (5 cycles) PCR in which unique indices were added to the sample. Following the library preparation, the final concentration of the libraries (8.92 ng/ $\mu$ L for LD2) were measured using the Qubit® dsDNA HS Assay Kit (Life Technologies), and the average library size (636 bp for LD2) was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies). The libraries were then pooled in equimolar ratios of 2nM, and 10pM of the library pool was clustered using the cBot (Illumina) and sequenced paired end for 500 cycles using the HiSeq 2500 system (Illumina).

### **Genomic and Phylogenetic Characterization**

The 16S sequence from the isolate was uploaded to the Ribosomal Database Project (RDP) myRDP website (Cole et al. 2014). The RDP Classifier (Wang et al. 2007) was used to classify the sequence from the domain level down to the genus level. Sequence matches were selected using the RDP Seqmatch program (Cole et al. 2014) with the following parameters: Strain, Type; Source, Isolates; Size,  $\geq 1200$ ; Quality, Good. The Basic Local Alignment Search Tool (BLAST; Altschul et al. 1997) was also used to find similar sequences for phylogenetic

analysis. The search set parameters were set to: Database, 16S ribosomal RNA sequences (Bacteria and Archaea); Limit to, Sequences from type material; Optimize for, Highly similar sequences (megablast). The Bergey's Manual of Systematic Bacteriology was referenced to compare proposed species matches using metabolic, physiological, and biochemical attributes (Brenner et al. 2005).

An integrated database called EzBioCloud (<http://www.ezbiocloud.net>) holds the taxonomic hierarchy of bacteria and archaea that are represented by quality controlled 16S rRNA gene and genome sequences. The database has integrated search tools that were used to determine the G+C content of the entire genome and to find the 16S rRNA sequence with the highest pairwise nucleotide sequence similarity value (Tindall et al. 2010, Yoon et al. 2017).

Multilocus sequence typing (MLST) was performed using the whole genome sequence and the online server and MLST 1.8 software from the Center for Genomic Epidemiology (CGE) at <https://cge.cbs.dtu.dk/services/MLST/>. MLST configuration was run using both *Yersinia* spp. and *Yersinia ruckeri* schemes and utilizing assembled genome/contigs (Larsen et al. 2012). The results are listed in Tables 4 and 5. One Codex is an online data platform for microbial genomics, supporting taxonomic and functional analysis of whole genome sequences as well as 16S rDNA sequences (Minot et al. 2015). The whole genome sequence for the isolate LD2 was uploaded and analyzed on this platform.

Two software programs were used to perform phylogenetic analysis of the 16S rRNA sequence: Phylogenetic Analysis Using Parsimony (PAUP\*; Swofford 2002) and Molecular Evolutionary Genetics Analysis (MEGA) version 7 (Kumar et al. 2016). However, PAUP\* has not been updated in several years and does not have a graphic interface or accompanying documentation. In addition, PAUP\* took many days of computing time to run a maximum

likelihood analysis. Therefore, MEGA7 was used for the final phylogenetic analysis. Two datasets were compiled using BLASTn: one to determine the placement of the isolate LD2 within the order Enterobacteriales, and one to determine the isolate's placement within the genus *Yersinia*. *Haemophilus influenzae* was used as the outgroup for both datasets. The datasets were aligned using Multiple Sequence Comparison by Log-Expectation (MUSCLE; Edgar 2004) from within MEGA7. Gap penalties were -400 for opening a gap in the alignment and 0 for extending a gap. Maximum iterations was set at 8, and the clustering method was UPGMB. The alignment was visually inspected using the Alignment Explorer to check for obviously misaligned sites. Duplicate sequences were eliminated by computing the pairwise distances using the following parameters: Substitution type – nucleotide; Model/Method – number of differences. The average identity was checked to estimate the reliability of the alignment by computing the overall mean distance among pairs of sequences using parameters: Scope – overall mean; Substitution Type – nucleotides; Model/Method – p distance. P-distance is the proportion ( $p$ ) of nucleotide sites at which the two sequences to be compared are different. It is obtained by dividing the number of nucleotide differences by the total number of sites compared. It does not make any correction for multiple substitutions at the same site or differences in evolutionary rates among sites (Nei and Kumar 2000). Before using the Maximum Likelihood method to create a tree, a log-likelihood test was performed by running the “Find Best DNA Models” within MEGA in order to determine the best evolutionary model, as well as the best rate among sites. Analysis preferences were set to: Analysis – Model Selection (ML); Tree to use – Automatic; Substitution Type – Nucleotide; Gap/Missing Data Treatment – Partial deletion; Site Coverage Cutoff (%) – 95. The resulting best model with the highest log likelihood was the

Hasegawa-Kishino-Yano model (Hasegawa et al. 1985) with rate among sites set to Gamma distributed with Invariant sites (HKY + G + I) model.

The evolutionary history was inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model (Hasegawa et al. 1985). The tree with the highest log likelihood was selected. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites. The rate variation model allowed for some sites to be evolutionarily invariable. Phylogeny was tested using the bootstrap method. The bootstrap consensus tree inferred from 1000 bootstrap replications is taken to represent the evolutionary history of the taxa analyzed (Felsenstein 1985).

## RESULTS

### Phenotypic Results

The appearance of the chosen colony was white to light pink in color and had a shiny, opaque consistency. The colony was circular with an entire margin and a raised elevation. The growth was also viscous and sticky. It is a Gram-negative, rod-shaped organism measuring 0.5-1.0µm in diameter x 2.0-3.0µm in length. It is non-motile at 25°C and 30°C. Flagella staining showed negative results for any type of flagella. The isolate LD2 grows between 4°C and 37°C with an optimal temperature of 30°C as assessed by visual observation of colony mass. It can tolerate up to 1% NaCl in the nutrient media. The pH range is from 7 to 9, with optimal growth at pH9 assessed by visual observation of colony mass. The isolate is a facultative anaerobe: it grew in both 0% and 20% oxygen. The environmental parameters are summarized in Table 1.

The isolate LD2 is catalase positive and oxidase negative. The Enteropluri-Test© was positive for glucose fermentation and Voges-Proskauer (acetoin production). The results of the Enteropluri-Test© were interpreted using the Enteropluri-Test© Codebook to identify the bacterium. After deciphering the code, the results identified the isolate LD2 as one of three possible bacteria: *Shigella* spp., *Pantoea agglomerans*, or *Yersinia pseudotuberculosis*. The biochemical characteristics are summarized in Table 2.

Table 1. Environmental parameters and phenotypic characteristics of isolate LD2 and closely related genera. Unless stated otherwise, information obtained from Brenner et al. 2005. Information for *Shigella flexneri* from Zaika 2005. FA = facultative anaerobe, ND = No data.

<b>Characteristic</b>	Isolate LD2	<i>Yersinia ruckeri</i>	<i>Yersinia kristensenii</i>	<i>Yersinia pseudotuberculosis</i>	<i>Budvicia diplopodorum</i>	<i>Budvicia aquatica</i>	<i>Shigella flexneri</i>	<i>Pantoea agglomerans</i>
<b>Gram staining</b>	-	-	-	-	-	-	-	-
<b>Neg. staining</b>								
Shape	Rods	Rods	Rods	Rods	Rods	Rods	Rods	Rods
Size width	0.5-1.0 x	0.5 x	0.5 x	0.5 x	0.8 x	0.8 x	0.7 x	0.5–1.3x
length	2.0-3.0 µm	2.0 µm	2.0 µm	2 µm	3.0 µm	3.0 µm	2 µm	1.0–3.0 µm
<b>Motility test</b>	-	+ 25-28°	+ 25-28°	+ 25-28°	-	+ 22 °C	-	+
<b>Temperature</b>								
4°C	+	+	+	+	+	+	ND	ND
25°C	+	+	+	+	+	+	+	+
30°C	+	+	+	+	+	+	+	+
37°C	+	+	+	+	-	+	++	+
50°C	-	-	-	-	-	-	ND	-
<b>Salt</b>								
0%	+	+	+	+	+	+	+	ND
1%	+	+	+	+	+	+	+	ND
5%	-	+	+	-	-	-	+	ND
10%	-	-	-	-	-	-	ND	ND
<b>pH</b>								
pH3	-	-	-	-	ND	ND	+	ND
pH5	-	+	+	+	ND	ND	+	ND
pH7	+	+	+	+	ND	ND	ND	ND
pH9	+	+	+	+	ND	ND	ND	ND
<b>O<sub>2</sub> utilization</b>	FA	FA	FA	FA	FA	FA	FA	FA

16

Table 2. Differential characteristics of LD2 and closely related genera. (+ = 90% or more of the strains are positive; - = 10% or less of the strains are positive; d = 11–89% of the strains are positive; ND = no data). *Y. pseudotuberculosis*, *S. flexneri*, and *P. agglomerans* were identified as the closest matches using the EnteroPluri-Test©. The remaining strains were selected using RDP and BLASTn.

	<b>Isolate LD2</b>	<i>Yersinia ruckeri</i>	<i>Yersinia kristensenii</i>	<i>Yersinia pseudotuberculosis</i>	<i>Budvicia diplopodorum</i>	<i>Budvicia aquatica</i>	<i>Shigella flexneri</i>	<i>Pantoea agglomerans</i>
<b>EnteroPluri-Test©</b>								
<b>Glucose</b>	+	+	+	+	+	+	+	+
<b>Gas</b>	-	-	d	-	-	+	d	-
<b>Lysine</b>	-	d	-	-	-	-	-	-
<b>Gas</b>	-	ND	ND	ND	ND	ND	-	ND
<b>Ornithine</b>	-	+	+	-	-	-	-	-
<b>Gas</b>	-	ND	ND	ND	ND	ND	-	ND
<b>Indole</b>	-	-	d	-	-	-	d	-
<b>Adonitol</b>	-	-	-	-	ND	-	-	-
<b>Lactose</b>	-	-	-	-	-	d	-	d
<b>Arabinose</b>	-	-	d	+	-	+	d	+
<b>Sorbitol</b>	-	d	+	-	-	-	d	-
<b>Voges-Proskauer</b>	+	-	-	-	-	-	-	+
<b>Dulcitol</b>	-	-	-	-	-	-	-	-
<b>Phenylalanine</b>	-	-	-	-	ND	-	-	d
<b>Urease (urea)</b>	d	-	+	+	-	+	-	-
<b>Citrate</b>	-	-	-	-	+	+	-	d
<b>Other Tests</b>								
<b>Spirit Blue</b>	-	ND	ND	ND	ND	ND	ND	ND
<b>DNase</b>	-	-	-	-	-	-	ND	ND
<b>Gel. hydrolysis</b>	-	d	-	-	-	-	-	d
<b>Mannitol</b>	-	ND	+	+	+	+	+	ND
<b>MacConkey</b>	-	ND	ND	+	-	+	ND	ND
<b>KCN</b>	-	d	-	-	ND	-	-	ND
<b>Catalase</b>	+	+	+	+	+	+	+	+
<b>Oxidase</b>	-	-	-	-	-	-	-	-

## Genotypic Results

The isolate LD2 was classified from the domain to the genus level based on the small subunit of the 16S rDNA sequence using the RDP Classifier (Wang et al. 2007). The results along with bootstrap confidence values are as follows: Domain Bacteria (100%); phylum Proteobacteria (100%); class Gammaproteobacteria (100%); order “Enterobacteriales” (100%); family Enterobacteriaceae (100%); genus *Yersinia* (52%).

The RDP SeqMatch program was used to find the closest sequences to the isolate based on the fraction of shared seven-base sequence fragments (words) between the isolate and reference sequences (S\_ab score). SeqMatch is more accurate than BLAST at identifying database sequences that are closely related to query rRNA sequences (Cole et al. 2014). *Yersinia ruckeri* and *Yersinia kristensenii* were the top closest matches in RDP SeqMatch, BLASTn, and EzBioCloud searches. *Y. ruckeri* also had the highest Max Score of 2396 in the BLASTn search. This score is calculated from the sum of the match rewards and the mismatch, gap open and extend penalties independently for each sequence. The top closest genomic matches for the 16S rRNA sequence of the isolate LD2 are listed in Table 3.

Table 3. DNA sequence matches for a bacterial isolate obtained from water samples at Kephart Prong, GSMNP, North Carolina, using the RDP, BLAST, and EzBioCloud software programs. The RDP seqmatch score (S<sub>ab</sub>) is the number of unique 7-base oligomers shared between the isolate sequence and a given RDP sequence divided by the lowest number of unique oligos in either of the two sequences. The BLAST % identity is the extent to which two sequences have the same nucleotide at the same positions in an alignment, expressed as a percentage. EzBioCloud pairwise similarity % is calculated according to Myers and Miller (1988).

Sequence Matches		
RDP (S <sub>ab</sub> score)	BLAST (% Identity) <i>Max Score</i>	EzBioCloud (Pairwise Similarity%)
1. <i>Y. ruckeri</i> ATCC 29473 (89.2%)	<i>Y. ruckeri</i> ATCC 29473 (98%) 2396	<i>Y. ruckeri</i> ATCC 29473 (97.76%)
2. <i>Y. kristensenii</i> ATCC 33638 (88.7%)	<i>Y. kristensenii</i> ATCC 33638 (98%) 2379	<i>Y. kristensenii</i> ATCC 33638 (97.76%)
3. <i>Y. pestis</i> NCTC 5923(87.9%)	<i>Y. pseudotuberculosis</i> CCUG 5855 (98%) 2368	<i>Y. bercovieri</i> ATCC 43970 (97.33%)
4. <i>Y. bercovieri</i> ATCC 43970 (87.9%)	<i>Y. pestis</i> NCTC 5923(97%) 2368	<i>Y. pseudotuberculosis</i> NBRC 105692 (97.25%)
5. <i>Y. similis</i> Y228 (87.8%)	<i>Y. watersii</i> 12-219N1 (97%) 2359	<i>Y. wautersii</i> 12-219N1 (97.25%)
6. <i>Y. massiliensis</i> CCUG 53443 (87.8%)	<i>Y. bercovieri</i> CNY 7506 (97%) 2357	<i>Y. pekkanenii</i> CIP110230 (97.18%)
7. <i>Y. pseudotuber*</i> ATCC 29833 (87.5%)	<i>Y. similis</i> Y228 (97%) 2351	<i>Y. similis</i> Y228 (97.18%)
8. <i>Y. aldovae</i> ATCC 35236 (87.5%)	<i>Y. aleksiciae</i> DSM 14987 (97%) 2346	<i>Rhanella woolbedingensis</i> FRB 227 (97.17%)
9. <i>Y. pekkanenii</i> AYV7 (87.5%)	<i>Y. massiliensis</i> 50640 (97%) 2346	<i>Y. aldovae</i> ATCC 35236 (97.11%)
10. <i>Y. wautersii</i> 12-219N1 (87.5%)	<i>Y. frederiksenii</i> CNY 6175 (97%) 2346	<i>Y. mollaretii</i> ATCC 43969 (97.11%)
11. <i>Y. enterocolitica</i> ATCC 9610 (87.2%)	<i>Y. pekkanenii</i> AYV7 (97%) 2340	<i>Y. massiliensis</i> CCUG 53443 (97.11%)
12. <i>Y. frederiksenii</i> ATCC 33641 (87.2%)	<i>Y. mollaretii</i> CNY 7263 (97%) 2340	<i>Y. frederiksenii</i> ATCC 33641 (97.11%)
13. <i>Y. intermedia</i> ATCC 29909 (87.2%)	<i>Y. intermedia</i> CNY 3953 (97%) 2335	<i>Y. aleksiciae</i> DSM 14987 (97.11%)
14. <i>Y. mollaretii</i> ATCC 43969 (87.2%)	<i>Y. aldovae</i> ATCC 35236 (97%) 2335	<i>Rahnella bruchi</i> FRB 226 (97.1%)
15. <i>Y. rohdei</i> ATCC 43380 (87.1%)	<i>Ewingella americana</i> CIP 81.94 (97%) 2330	<i>Rahnella variigena</i> CIP 105588 (97.05%)

EzBioCloud determined the GC content of the entire genome of the isolate to be 47.40%. In addition, the Identify tool returned the Top-hit taxon of *Yersinia kristensenii* ATCC 33638 and *Yersinia ruckeri* ATCC 29473 each with a 16S rRNA sequence highest pairwise nucleotide sequence similarity value of 97.76%.

MLST allele sequences and sequence type (ST) profile tables are stored in online databases hosted at five different sites around the world. The University of Oxford collects data from all databases and makes it easily accessible at [pubmlst.org](http://pubmlst.org). The Center for Genomic Epidemiology's (CGE) web-based method for MLST automatically updates allele sequences and schemes weekly. In total, 66 bacterial MLST schemes are currently available (Larsen et al. 2012). The *Yersinia* MLST scheme used to generate Table 4 used fragments of the following seven house-keeping genes:

*aarF* - putative ubiquinone biosynthesis protein UbiB

*dfp* - bifunctional phosphopantothencycysteine decarboxylase/phosphopantothenate synthase

*galR* - DNA-binding transcriptional regulator

*glnS* - glutaminyl-tRNA synthetase

*hemA* - glutamyl-tRNA reductase

*rfaE* - bifunctional heptose 7-phosphate kinase/heptose 1-phosphate adenylyltransferase

*speA* - arginine decarboxylase

Table 4. MLST results using the WGS of LD2 and the scheme for *Yersinia* spp. Locus: MLST locus against which the input sequence has been aligned. % Identity: Percentage of nucleotides that are identical between the best matching MLST allele in the database and the corresponding sequence in the genome. HSP Length: Length of the alignment between the best matching MLST allele in the database and the corresponding sequence in the genome, also called the high-scoring segment pair (HSP). Allele Length: Length of the best matching MLST allele in the database. Gaps: Number of gaps in the HSP. Allele: Name of the best matching MLST allele.

Locus	% Identity	HSP Length	Allele Length	Gaps	Allele
<i>aarF</i>	78.29	479	500	0	<i>aarf_33</i>
<i>dfp</i>	81.10	328	455	0	<i>dfp_9</i>

<i>galR</i>	86.11	36	500	0	<i>galr_27</i>
<i>glnS</i>	83.89	180	442	0	<i>glns_80</i>
<i>hemaA</i>	79.67	482	490	0	<i>hema_55</i>
<i>rfaE</i>	77.78	414	429	0	<i>rfae_23</i>
<i>speA</i>	78.32	452	452	3	<i>spea_47</i>

The *Yersinia ruckeri* MLST scheme used to generate Table 5 used fragments of the following six house-keeping genes:

*dnaJ* – a molecular chaperone

*glnA* - glutamine synthetase

*gyrB* - DNA gyrase B subunit

HSP60 - encoding a 60-kDa heat shock protein

*recA* - DNA repair and recombination

*thrA* - aspartokinase

Table 5. Multilocus sequence results using the whole genome sequence of LD2 and the scheme for *Y. ruckeri*. Locus: MLST locus against which the input sequence has been aligned. % Identity: Percentage of nucleotides that are identical between the best matching MLST allele in the database and the corresponding sequence in the genome. HSP Length: Length of the alignment between the best matching MLST allele in the database and the corresponding sequence in the genome, also called the high-scoring segment pair (HSP). Allele Length: Length of the best matching MLST allele in the database. Gaps: Number of gaps in the HSP. Allele: Name of the best matching MLST allele.

Locus	% Identity	HSP Length	Allele Length	Gaps	Allele
<i>dnaJ</i>	78.72	625	632	12	<i>dnaj_1</i>
<i>glnA</i>	80.68	409	416	0	<i>glna_8</i>
<i>gyrB</i>	81.18	356	454	0	<i>gyrb_2</i>
<i>hsp60</i>	85.46	447	509	0	<i>hsp60_4</i>
<i>recA</i>	81.05	459	472	0	<i>reca_1</i>
<i>thrA</i>	94.44	18	303	0	<i>thra_1</i>

The results of the WGS of the isolate LD2, including the entire sequence and an annotated Excel file, are listed in the appendix. The annotated file was visually scanned for

putative genes that may have an impact on the environment or be involved in pathogenicity.

Table 6 is a synopsis of these genes.

Table 6. List of significant genes and their putative function. This list was compiled from the annotated whole genome sequence generated by MR DNA (Molecular Research LP). The Excel file is available as a link in the appendix.

# of Genes	Putative Function	Significance
18	Hemin transport/Siderophore	Formation of soluble Fe <sup>3+</sup> in environment, acquisition of iron from host organisms
3	Hemolysin	Lysis of red blood cell membrane
44	Type III, IV, VI secretion systems	Role in the pathogenesis
3	Virulence factor	Role in the pathogenesis
3	Invasins	Damage host cells
11	Multidrug transport system	Antibiotic resistance
6	Macrolide (multidrug resistance)	Antibiotic resistance
1	Plasmid-related proteins	Antibiotic resistance
6	Bundle-forming pilus	Possible gene transfer
10	Prophage integrase	Viral genes
26	Phage specific genes	Viral genes
76	Mobile element protein	Transposons
9	Insertion sequence protein	Transposons
12	Co, Cu, Zn, Cd, As resistance	Toxic metal resistance
72	Flagella-specific genes	Motility

One Codex is an online data platform for microbial genomics, supporting taxonomic and functional analysis of whole genome sequences as well as 16S rDNA sequences (Minot et al. 2015). The whole genome sequence for the isolate LD2 was uploaded and analyzed on this platform. A total of 28 reads out of 42, or 66.67%, were classified using the One Codex database. Twenty-two reads were classified at the species level (Figure 4). Sixteen of those reads were classified as *Serratia* sp. DD3, 2 reads as *Serratia fonticola*, and 1 read each for *Serratia marcescens*, *Alteromonas* sp. SN2, *Erwinia iniecta*, and bacterial symbiont BFo 1 of *Frankliniella occidentalis*.

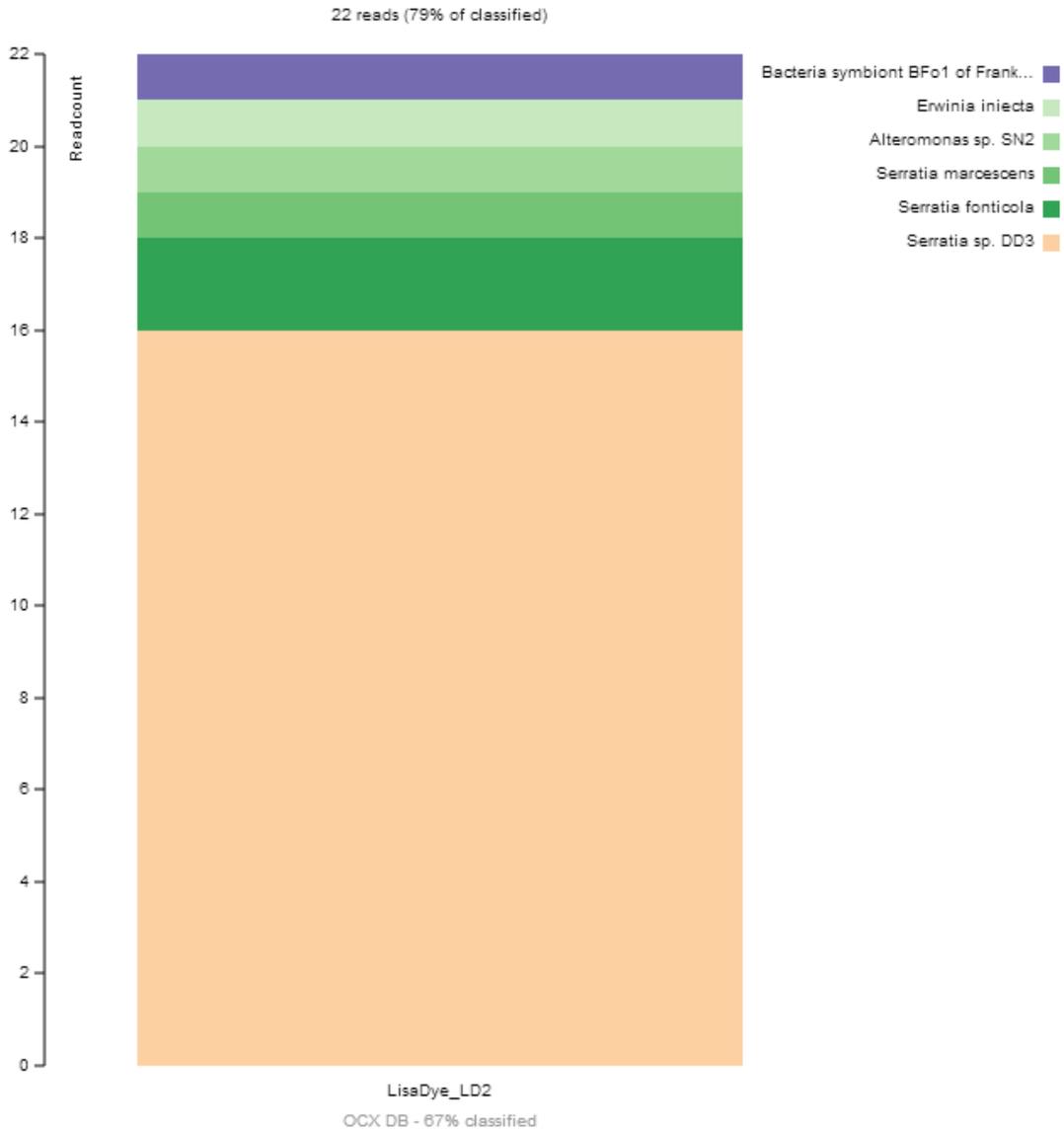


Figure 4. Species level results of One Codex analysis. A total of 28 of 42 reads for the isolate LD2 were classified using the One Codex database. Twenty-two reads were classified at the species level. The majority of these reads (16) were classified as belonging to the species *Serratia* sp. DD3.

Figure 5 shows the results of One Codex analysis at the genus level. A total of 28 reads were classified at this level, with 21 reads (75%) classified as the genus *Serratia*. Two reads were classified as genus *Klebsiella*, and one read each for *Alteromonas*, *Pseudomonas*, *Erwinia*,

and *Salmonella*. One read was classified at a species level, but the taxonomy doesn't have a genus for that species (S. Minot, personal communication).

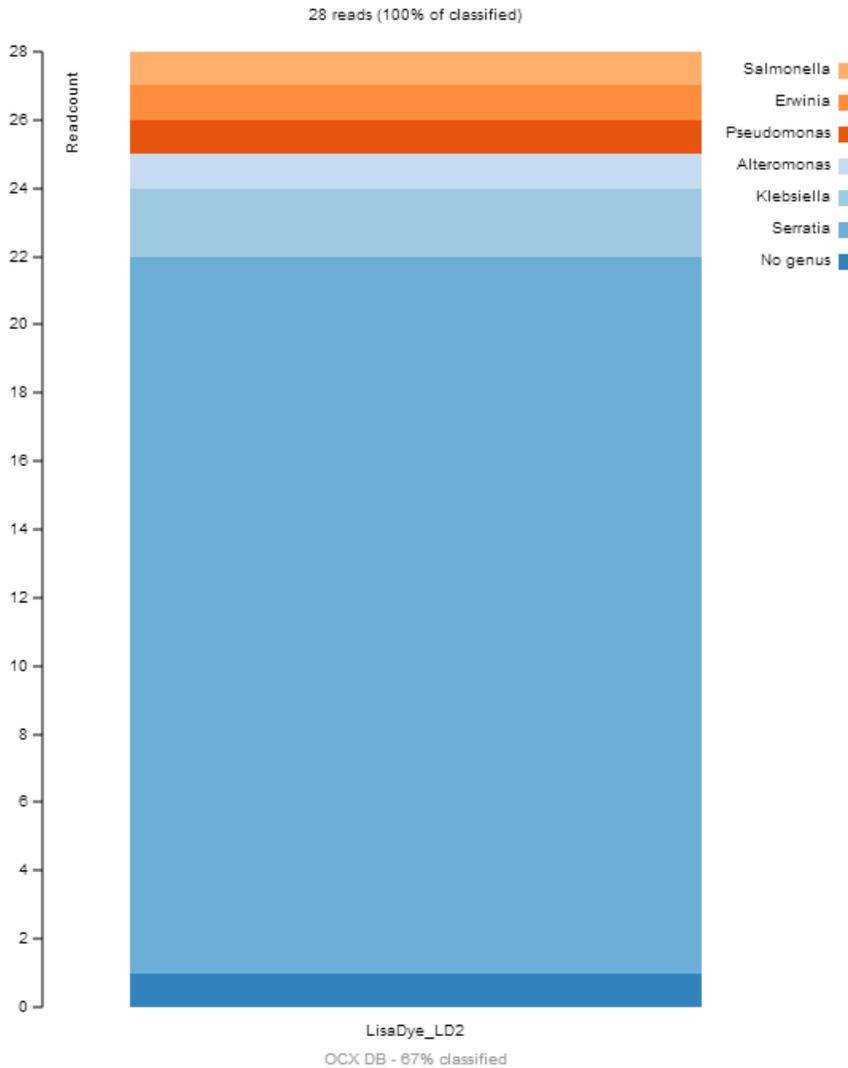


Figure 5. Genus level results of One Codex analysis. A total of 28 reads were classified at the genus level. The majority of the reads (21) were classified as belonging to the genus *Serratia*.

The 16S rDNA sequence was also uploaded to One Codex. The results classified that sequence as belonging to *Salmonella enterica*.

## Phylogenetic Results

Figure 6 is the most likely tree using the maximum likelihood method. It shows the placement of the isolate LD2 within the order “Enterobacteriales”. The isolate LD2 is clustered together with *Yersinia ruckeri* with a 94% bootstrap confidence value. The tree is drawn to scale to show the distance between the outgroup *Haemophilus influenzae*. All of the strains used in the data set, except for *Haemophilus influenzae*, are from the order “Enterobacteriales”. *H. influenzae* belongs to the order Pasteurellales. The bootstrap values are very low on some of the branches leading to LD2. Figure 7 is the same as Figure 6 except it is shown in cladogram format to clarify the branching order and bootstrap values.

Figures 8, 9, and 10 use a different dataset in which all the sequences are from the family Enterobacteriaceae. Now that it has been demonstrated that the isolate belongs in the order “Enterobacteriales”, these trees help clarify the isolates position within the family Enterobacteriaceae by using sequences that are more closely related to the isolate. Figure 8 is the most likely tree using the maximum likelihood method. It is drawn to scale and shows that the isolate LD2 is clustered among closely related species within the family Enterobacteriaceae. Figure 9 shows that the isolate LD2 is no longer grouped with *Yersinia ruckeri*, but is in a sister clade along with *Budvicia*. Because the bootstrap values are low in this figure, an additional tree (Figure 10) is shown. Branches with less than a 50% bootstrap value are collapsed, demonstrating that the isolate LD2 is now a polytomy, which is a section of a phylogeny in which the relationships cannot be fully resolved to dichotomies, thus presenting an unlikely picture of many branches appearing simultaneous at the same point in evolutionary time.

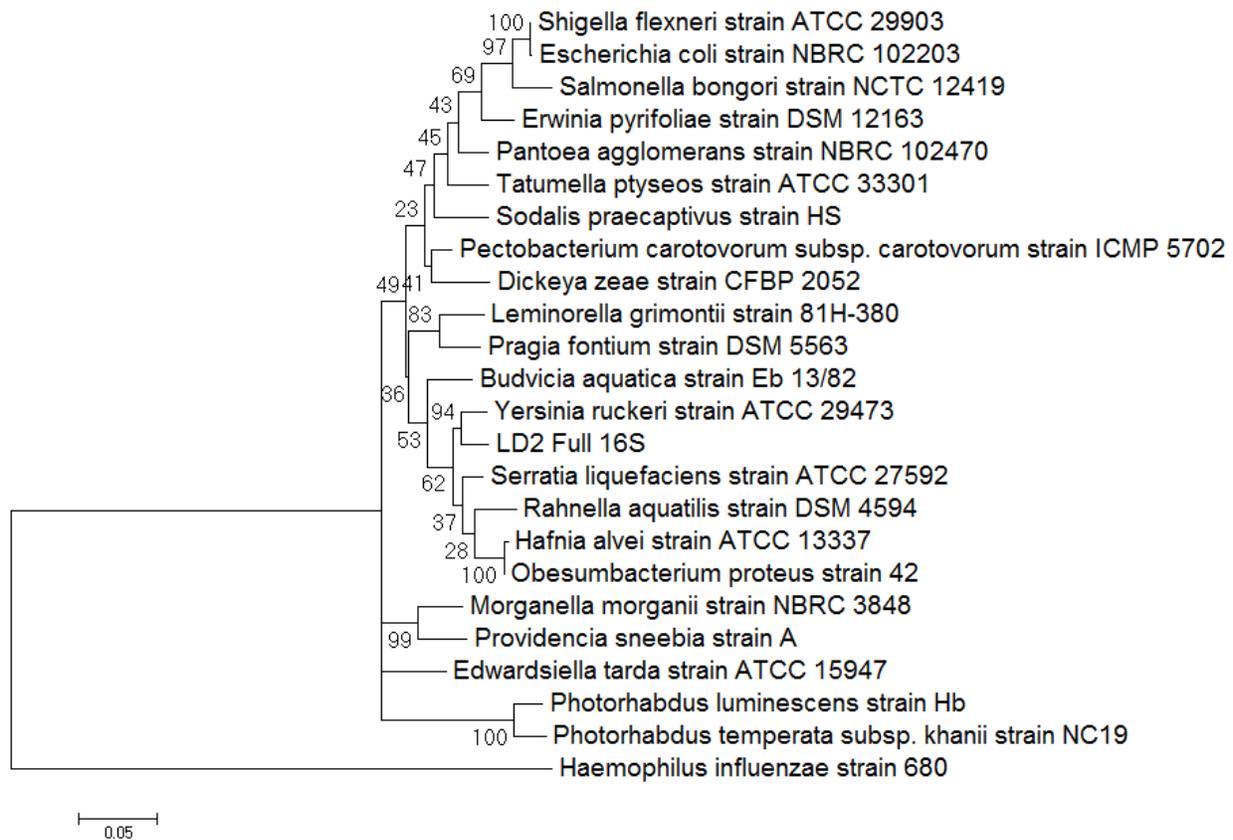


Figure 6. Phylogram for “Enterobacteriales” showing distances and the placement of LD2 in the order. The evolutionary history was inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model (Hasegawa et al. 1985). The tree with the highest log likelihood (-5820.8899) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.3727)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 39.9926% sites). The analysis involved 24 nucleotide sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 1349 positions in the final dataset. The tree is rooted on *Haemophilus influenzae*. Evolutionary analyses were conducted in MEGA7 (Kumar et al. 2016).

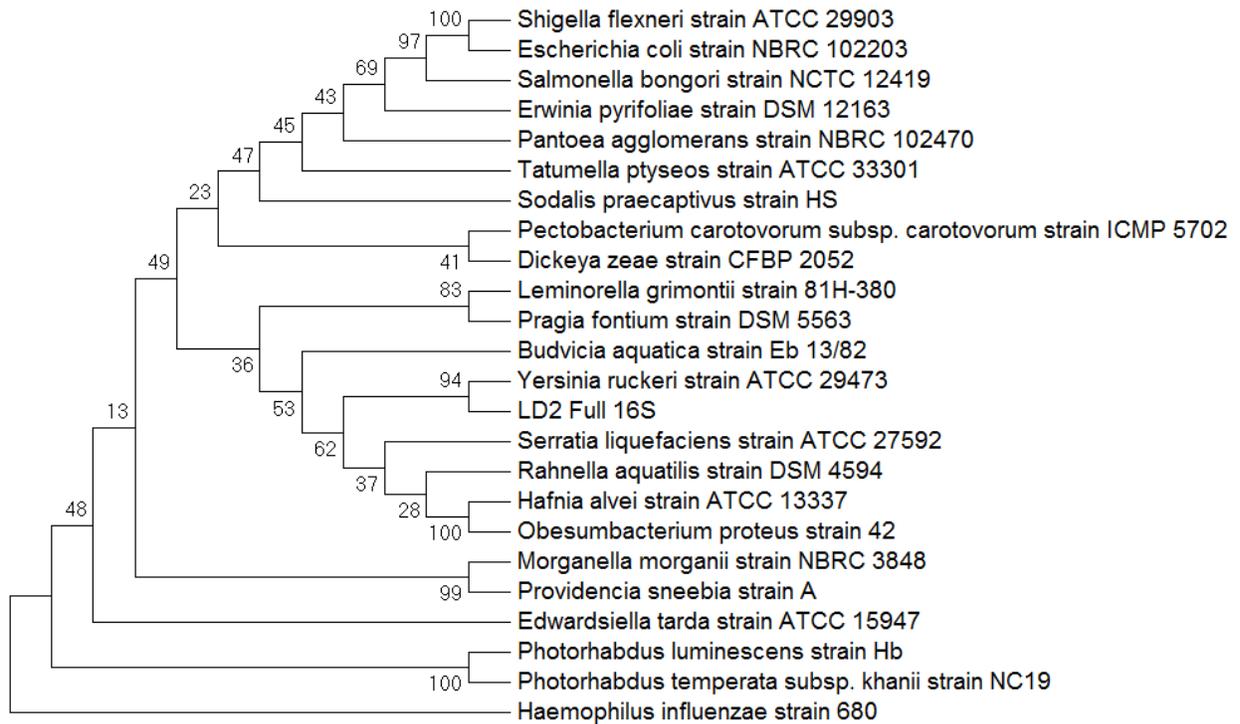


Figure 7. Most likely tree for “Enterobacteriales” showing the position of the isolate LD2 within the order. This is the same tree as Figure 6 shown in topology only format so the branching order and bootstrap values can be seen more clearly. The tree is not drawn to scale. (Kumar et al. 2016).

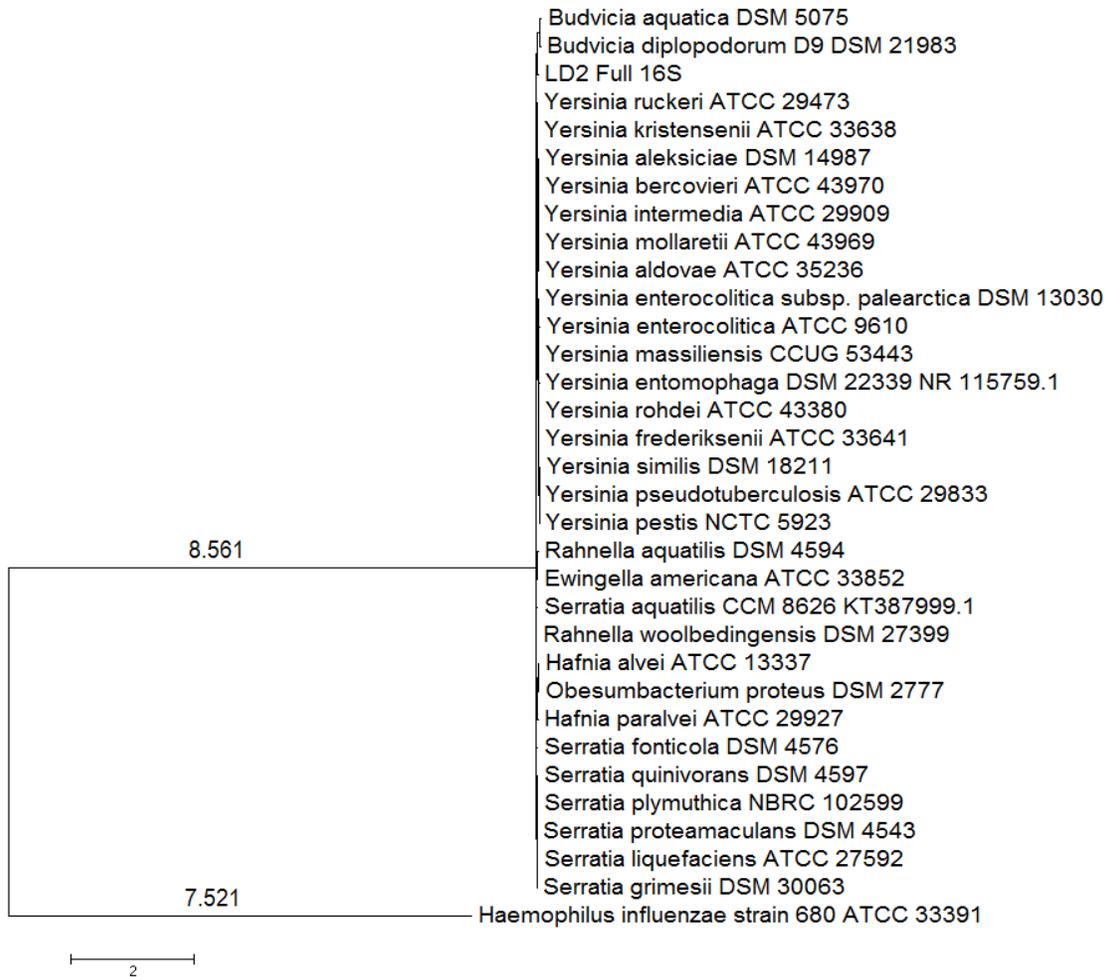


Figure 8. Most likely tree for Enterobacteriaceae showing the position of the isolate LD2 within the family. The evolutionary history was inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model (Hasegawa et al. 1985). The tree with the highest log likelihood (-4680.4809) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.2335)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 41.7694% sites). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 33 nucleotide sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 1379 positions in the final dataset. The tree is rooted on *Haemophilus influenzae*. Evolutionary analyses were conducted in MEGA7 (Kumar et al. 2016).

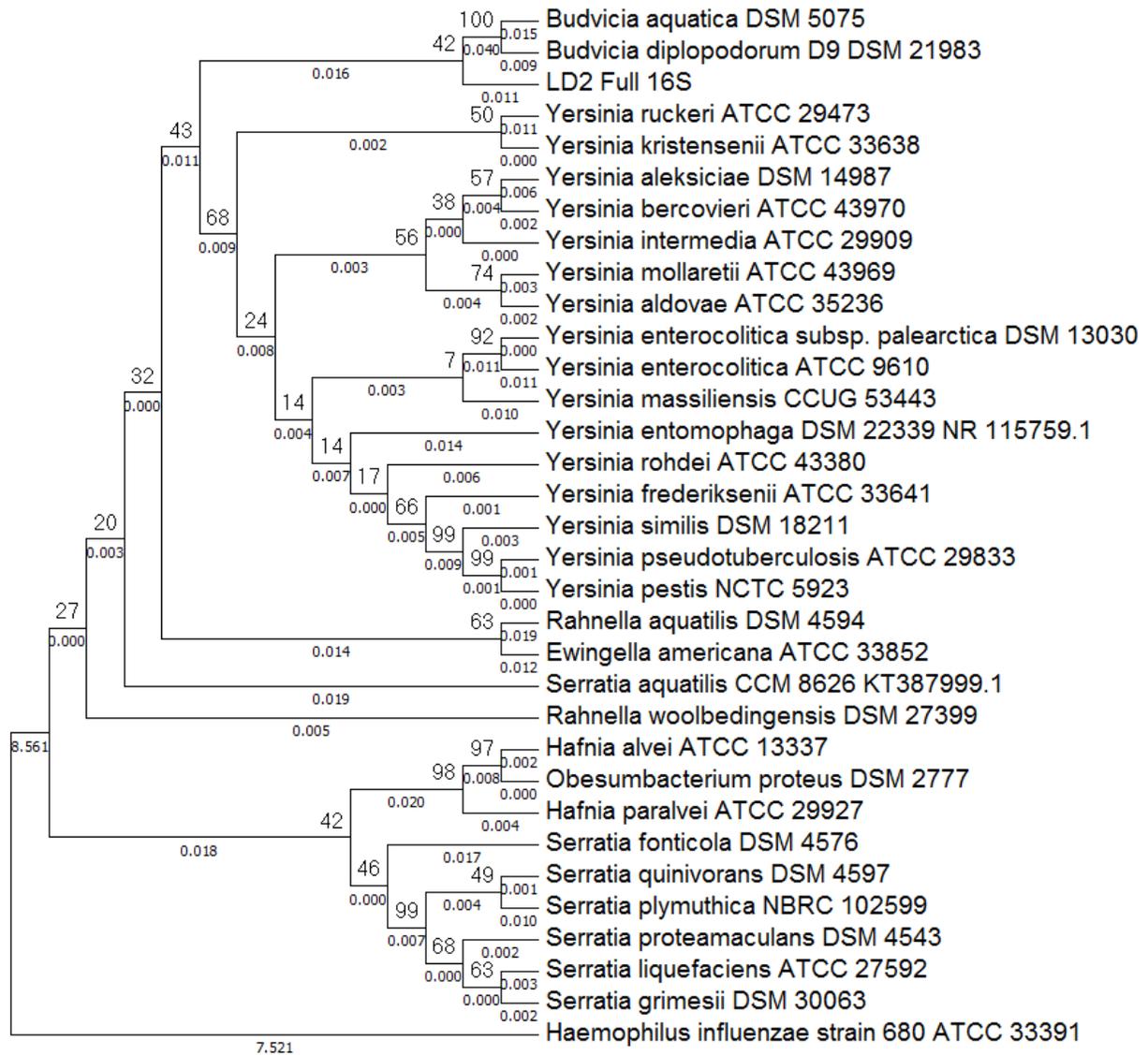


Figure 9. Branch lengths and bootstrap values for Enterobacteriaceae. Most likely tree showing the position of the isolate LD2 within the family Enterobacteriaceae with branch lengths and bootstrap values shown. The tree is not drawn to scale. Parameters are the same as Figure 8. (Kumar et al. 2016).

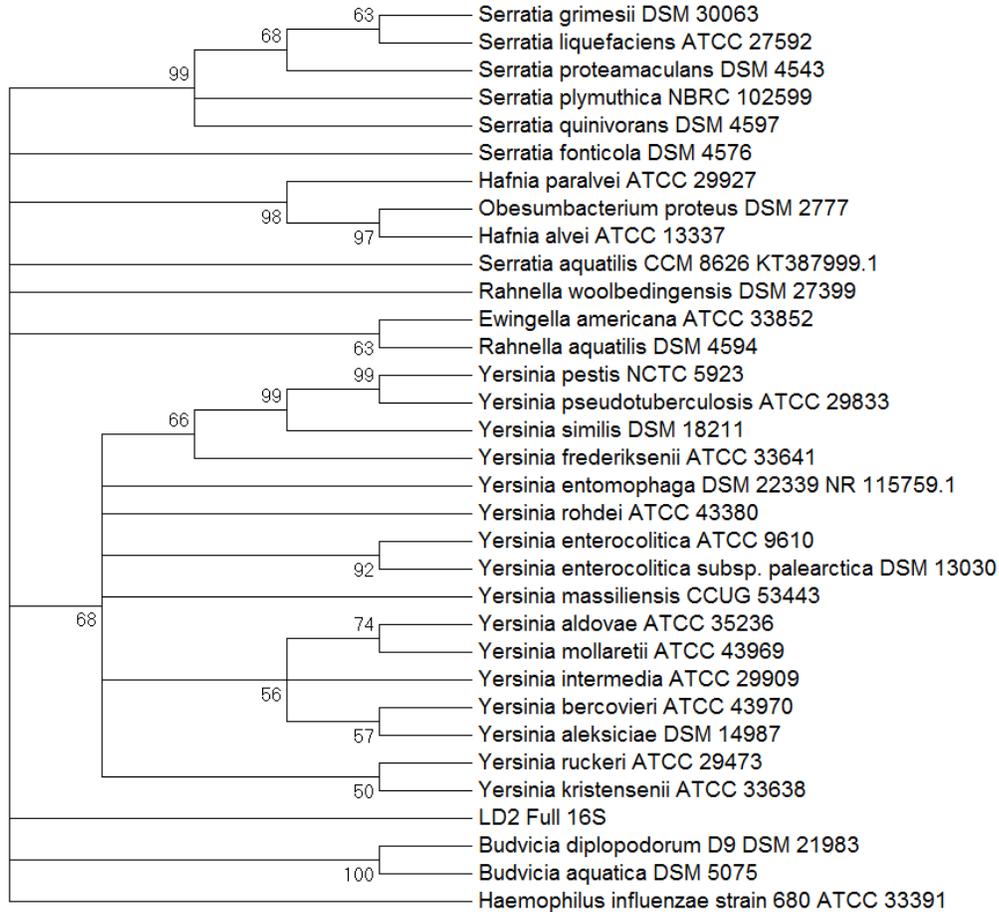


Figure 10. Majority rule bootstrap consensus tree for Enterobacteriaceae showing the uncertainty of the branching pattern within the family. The evolutionary history was inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model (Hasegawa et al. 1985). The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed (Felsenstein 1985). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.2335)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 41.7694% sites). The analysis involved 33 nucleotide sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 1379 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al. 2016).

## DISCUSSION

### **Bacterial Taxonomy – A Polyphasic Approach**

There is currently no widely accepted concept of species for prokaryotes (Gevers et al. 2005). Prokaryotes are asexual, so the classic definition of a species as a group of organisms that can interbreed and produce fertile offspring cannot be applied (Emerson et al. 2008). If the term species is going to be used to classify organisms into a taxonomic rank, microbiologists must agree to some guidelines in order to provide stability, reproducibility, and coherency in taxonomy. Microorganisms were traditionally classified on the basis of morphological, physiological, and biochemical methods. However, the advent of modern molecular and genetic techniques created a blurred image of microbial taxonomy which needed further clarification (Prakash 2007).

The International Committee on Systematics of Prokaryotes (ICSP), formerly the International Committee on Systematic Bacteriology (ICSB), is the body that oversees the nomenclature of prokaryotes, determines the rules by which prokaryotes are named and whose Judicial Commission issues opinions concerning taxonomic matters and revisions to the Bacteriological Code. ICSP is a committee under the Bacteriology and Applied Microbiology Division of the International Union of Microbiological Societies (IUMS). Their publications include the International Journal of Systematic and Evolutionary Microbiology (IJSEM), the International Code of Nomenclature of Bacteria (Bacteriological Code) and the Approved Lists of Bacterial Names (ICSP 2017). In 1987, an ad hoc committee of the ICSB was convened for a workshop on reconciliation of approaches to bacterial systematics. Their objective was to arrive at a common ground of understanding in the process of bacterial systematics. There was general

agreement that the complete deoxyribonucleic acid (DNA) sequence would be the reference standard to determine phylogeny and that phylogeny should determine taxonomy. During this workshop, a practical approach was developed to define a species by a polyphasic approach; one that deployed a number of methods for the complete characterization of microbes. The committee recommended that the phylogenetic definition of a bacterial species generally would include strains with approximately 70% or greater DNA-DNA Hybridization (DDH) relatedness and with 5°C or less  $\Delta T_m$ , and that phenotypic characteristics should agree with this definition (Wayne et al. 1987).

An early comparative study between DDH and 16S rRNA gene sequence similarity revealed that 97% 16S rRNA gene sequence similarity corresponded to 70% DDH (Stackebrandt & Goebel, 1994). This demarcation value of 16S rRNA gene sequence similarity has been widely used in bacterial classification as an alternative to the laborious DDH, and it is now generally accepted that DDH is only required when 16S rRNA gene sequence similarity between two strains is over 97% (Tindall et al., 2010), and even higher thresholds of 98.7–99.0% have been recommended (Stackebrandt and Ebers, 2006).

Given the modern advances in technology, the methods now employed for bacterial systematics include: phenotypic characteristics such as biochemical assays, as well as physiological and morphological tests; genotypic characteristics such as the complete 16S rRNA gene sequencing, DDH studies with related organisms, analyses of molecular markers, multilocus sequence typing (MLST) and whole genome sequence analysis; and phylogenetic analysis (Prakash 2007).

## Phenotypic Analysis

Phenotypic analysis includes aspects such as cell shape, size, physiological and biochemical tests, as well as methods of chemical analysis of the cell (Tindall et al. 2010). Tables 1 and 2 were used to compare phenotypic characteristics between several similar species of bacteria. *Yersinia* and *Budvicia* species were selected based on the similarity of the 16S rRNA gene. The isolate was also matched to known species of bacteria in the Enterobacteriaceae family by using the Codebook and results for the EnteroPluri-Test© (Liofilchem, 2013). Using this method, the isolate LD2 was identified as either *Pantoea agglomerans* (formerly known as *Enterobacter agglomerans*), *Shigella flexneri*, or *Yersinia pseudotuberculosis*. *P. agglomerans* is biochemically the most similar to the isolate based on the characteristics in Tables 1 and 2. However, data for growth in the presence of NaCl and various pH levels was not available. The differences are that *P. agglomerans* is positive for arabinose fermentation, and 11-89% of strains of *P. agglomerans* are positive for lactose fermentation, citrate utilization, and phenylalanine deamination. In addition, *P. agglomerans* is motile at 36°C. The isolate LD2 was non-motile at 25°C and 30°C, and did not grow well at 36°C. In addition to biochemical similarities, *P. agglomerans* shares a morphological characteristic with LD2; many strains produce mucoid colonies that stick to the agar (Brenner et al. 2005).

Strains of the *P. agglomerans* species complex are found on the above ground surfaces of plants and within healthy plant tissues and seeds. Nitrogen fixing strains have been found in the rhizosphere of wheat and sorghum. In fact, these bacteria are typical of the innermost part of the rhizosphere of wheat. Strains of *P. agglomerans* have been isolated from water, paper mill process water, soil and decaying wood. They are frequently isolated from damaged plant tissues and lesions, although they are rarely considered pathogenic (Brenner et al. 2005).

Based on phenotypic characteristics and environmental parameters, it would appear that the isolate is most closely related to *P. agglomerans*. However, it must be taken into consideration that phenotypic characteristics of strains are usually highly dependent on growth conditions, (temperature, growth phase, growth medium). Lang et al. (2013) observed that incubation time and test conditions were critical criteria when evaluating biochemical reactions involving the API 20 E strip. In addition, phenotypes observed in the laboratory environment may not accurately represent phenotypes present in the natural environment. Therefore care should be taken in using phenotypic characteristics in systematic analyses (Madigan et al. 2015).

### **Genotypic Analysis**

The Ribosomal Database Project (RDP) grew out of Carl Woese's vision of how rRNA comparison methods could transform evolutionary phylogenetic analysis in the biological sciences. The RDP provides data, tools, and services related to rRNA sequences to the research community. Since its inception, the project has grown from a few hundred to several million rRNA gene sequences (Cole and Tiedje 2014). Today, rRNA-based analysis remains as a central method used in microbiology to explore microbial diversity as well as a day-to-day method for bacterial identification (Wang et al. 2007).

The RDP Classifier is a tool that rapidly and accurately assigns sequences into taxa with a confidence estimate value for each assignment called a bootstrap value. It uses a naïve Bayesian classification method that is capable of classifying near-full-length sequences as well as 400-base segments to the genus level with an overall accuracy above 88.7%. (Cole et al. 2014). Using the 16S rRNA sequence of the isolate, the Classifier identified the isolate down to the family level as Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriales, Enterobacteriaceae with a 100% confidence bootstrap values. Although it also identified the

isolate as belonging to the genus *Yersinia*, the bootstrap confidence value was only 52%. It is important to understand that the bootstrap method is an estimate of reproducibility, not accuracy. Bootstrap values of less than 70% are not taken very seriously (Hall 2011).

RDP's SeqMatch tool finds the closest RDP 16S rRNA sequences to a query based on the fraction of shared seven-base sequence fragments (words) between the query and reference sequences (S\_ab score). SeqMatch works well on partial and full-length sequences and, according to Cole et al. (2014), is more accurate than BLAST at identifying database sequences that are closely related to query rRNA sequences. Table 3 indicates the top two highest matches in SeqMatch were *Yersinia ruckeri* and *Yersinia kristensenii* with S\_ab scores of 89.2% and 88.7% respectively. The BLAST search also returned *Y. ruckeri* and *Y. kristensenii* as the top two sequences with the highest percent identities of 98% each. The BLAST percent identity is the extent to which two sequences have the same nucleotide at the same positions in an alignment, expressed as a percentage. Tindall et al. (2010) gave recommendations for sequence analysis of the 16S rRNA gene. They caution that pairwise nucleotide sequence similarities should be scrutinized according to the method of calculation. They state that pairwise similarity values obtained from local alignment programs, such as BLAST and FASTA, should not be used. These programs are primarily useful for database searches. They recommend several programs for similarity calculations including EzBioCloud (Tindall et al. 2010). Coincidentally, EzBioCloud also returned *Y. ruckeri* and *Y. kristensenii* as the two sequences with the highest pairwise nucleotide sequence similarities of 97.76% each.

*Yersinia ruckeri* is the causative agent of enteric redmouth disease (ERM), one of the most important infectious diseases in rainbow trout (*Oncorhynchus mykiss*) aquaculture (Huang et al. 2013). Since the first report of *Y. ruckeri* infection in rainbow trout in the USA, the

pathogen has been isolated from multiple other fish species worldwide (Kumar et al. 2015). ERM has also affected North Carolina trout farmers who have historically reported losses of up to 30% of their trout to disease. *Y. ruckeri* was reported as the primary cause of the trout loss (AREERA 2004). Additionally, *Y. ruckeri* has been isolated from animals other than fish, including muskrat, kestrel, sea gulls, turtles, and humans. These numerous reports demonstrate that *Y. ruckeri* has a wide host range and geographical distribution (Kumar et al. 2015).

The current definition of a bacterial species is that a value of 70% or less DDH and a pairwise similarity in the 16S rRNA sequence of less than 97% between two organisms is taken as evidence that the two are distinct species. The results from RDP, BLAST, and EzBioCloud all returned pairwise nucleotide sequence similarities greater than 97% (percent identities for RDP were not shown, but were higher than 97%). This would seem to suggest that the isolate may not be a distinct species from either *Y. ruckeri* or *Y. kristensenii*. However, even though bioinformatic comparison of 16S rDNA may provide an objective and reliable way of identifying a given strain, it has a critical limitation in its use at the species level. Even almost identical 16S rDNA may not guarantee that two strains belong to the same species (Yoon et al. 2017). Several groups of organisms have been identified which share nearly identical 16S rRNA sequences but in which DNA hybridization is significantly lower than 70%, thus indicating that they represent individual species (Stackebrandt and Goebble 1994). Where 16S rRNA gene sequence similarity values are more than 97%, other methods such as DNA–DNA hybridization or analysis of gene sequences with a greater resolution must be used (Tindall et al. 2010).

Sequencing a single gene such as the 16S rRNA gene, which has been used for the molecular analysis of many bacteria, is not optimal. Interspecies recombinations in 16S rRNA genes, initially thought to be very rare, have recently been inferred to occur in at least some

bacterial species, which underscores the importance of not focusing on a single gene during studies to determine the phylogeny/taxonomy of a bacterial species (Kotetishvili et al. 2005). Stackebrandt et al. (2002) encouraged investigators to propose new species based upon other genomic methods or techniques provided that they can demonstrate that, within the taxa studied, there is a sufficient degree of similarity between the technique used and DDH. They further suggest that a method of great promise is the evaluation of protein-coding gene sequence analysis for its ability to genomically delineate a species and differentiate it from neighboring species previously detected by other methods such as 16S rDNA analysis. They agreed that an informative level of phylogenetic data would be obtained from the determination of a minimum of five genes under stabilizing selection for encoded metabolic functions (housekeeping genes; Stackebrandt et al. 2002).

Multilocus sequence typing (MLST) is a method for characterizing isolates of bacterial species using the sequences of internal fragments of several (usually seven) house-keeping genes. Approximately 450-500 bp internal fragments of each gene are used. For each house-keeping gene, the different sequences present within a bacterial species are assigned as distinct alleles. Different allele sequences at each locus are assigned numbers and, for each isolate, the alleles at each of the seven loci define the allelic profile or sequence type (ST). Each isolate of a species is therefore unambiguously characterized by a series of seven integers which correspond to the alleles at the seven house-keeping loci. Most bacterial species have enough variation within house-keeping genes to create several alleles per locus, allowing billions of distinct allelic profiles to be distinguished using only seven house-keeping loci (PubMLST website).

In this study, the Center for Genomic Epidemiology MLST results using the whole genome sequence and the scheme for the genus *Yersinia* returned percent identity matches for all

seven genes below 87%. The scheme specific for *Yersinia ruckeri* returned a 94.44% identity for the *thrA* gene; the remaining identities were all below 86%. MLST was also run against the *Yersinia pseudotuberculosis* scheme (results not shown). The percent identities were all below 81%. There were no schemes available for the genera *Budvicia*, *Shigella*, or *Pantoea*. The program gave a warning stating that one or more loci do not match perfectly to any previously registered MLST allele. They recommend verifying the results by traditional methods for MLST, meaning the use of primers, PCR, and sequencing followed by analysis. Despite the fact that MLST and multilocus sequence analysis (MLSA) have become accepted and widely used methods in prokaryotic taxonomy, no common generally accepted recommendations have been devised to date for either the whole area of microbial taxonomy or for taxa-specific applications of individual schemes (Glaeser and Kämpfer 2015). However, given the fact that the standard for comparison of the individual 16S rRNA gene is 97% or better, the low percent identities returned by the CGE MLST seem to indicate that the isolate is not a species of *Yersinia*. In addition, a study done by Kotetishvili et al. (2005) stated that MLST was better suited for determining genetic relatedness among *Yersinia* species than was 16S rRNA analysis.

The G+C content of a bacterial chromosome is an important index for the identification and classification of bacteria. The minimum amount of genomic information required for the description of a novel bacterial species must include its phylogenetic classification, DNA–DNA relatedness, and the mol% G+C content of DNA (Fournier et al. 2006; Stackebrandt et al., 2002). Before the advent of whole genome sequencing, G+C content was determined using a number of biochemical assays such as buoyant density centrifugation, thermal denaturation methods, and high-performance liquid chromatography (Mesbah et al. 1989). However, determination of the DNA G+C content of prokaryotic genomes using traditional methods is time-consuming and

results may vary from laboratory to laboratory, depending on the technique used (Fournier et al. 2006). The recent advent of DNA sequencing technologies facilitates the use of genome sequencing data that provide means for more informative and precise classification and identification of bacteria. An integrated database developed by EzBioCloud (<http://www.ezbiocloud.net>) takes advantage of the accumulating genome sequencing data and offers many integrated search tools on their website (Yoon et al. 2017). These tools were used to determine the GC content of the entire genome of the isolate, and returned a value of 47.70%.

The G+C content of bacterial chromosomal DNA ranges from 25 to 80 mol% (Xu et al. 2000). Of the genera that were similar to the isolate in Table 1, the genus *Yersinia* has a mol% G+C of 46-50%, *Budvicia* is 46 +/- 1%, *Shigella* is 49-53%, and *Pantoea* is 49.7 – 60.6% (Brenner et al. 2005).) EzBioCloud returned a 47.40% GC content, which appears to be consistent with *Yersinia* and *Budvicia*.

One Codex is a web-based computational platform for identifying microbes. The One Codex metagenomic classification system is powered by a database containing roughly 40,000 whole genomes. One Codex classifies unknown nucleotide sequences according to the set of signature sequences within it that are unique to a specific taxonomic group. Each read is first broken into the complete set of overlapping sequences of length 31bp that comprise it (k-mers). These k-mers are compared against an exhaustive database that contains known k-mers that are unique to a specific taxonomic grouping. Each read or contig is then assigned to the microbial clade it most closely resembles, and the complete sample is summarized as a collection of these signature sequences found in the uploaded query sequence (Minot et al. 2015). The results of the entire genome analysis using One Codex (Figures 4 and 5) suggest that the isolate LD2 is a member of the genus *Serratia* and possibly a close match to *Serratia* sp. DD3. However, the

G+C content of the isolate does not match the G+C content of the genus *Serratia*, which is 52-60% (Brenner et al. 2005).

### **Phylogenetic Analysis**

There are several steps involved in making a phylogenetic tree based on molecular sequence data:

1. Identify and acquire the sequences that are to be included in the tree.
2. Align the sequences.
3. Estimate the tree by one of several methods.
4. Draw the tree (Hall 2011)

In this study, two datasets of sequences were assembled based on work done by Adeolu et al. (2016). They completed comprehensive comparative genomic analyses of the members of the order "Enterobacteriales" which included phylogenetic reconstructions based on 1548 core proteins, 53 ribosomal proteins, 4 multilocus sequence analysis proteins, as well as examining the overall genome similarity amongst the members of this order. The results of these analyses all supported the existence of 7 distinct monophyletic groups of genera within the order "Enterobacteriales". In addition, they performed analyses of protein sequences called conserved signature insertions/deletions (CSIs) which independently supported their monophyletic groups. CSIs are insertions or deletions (indels) that are uniquely present in a related group of organisms. On the basis of their analyses, they made a proposal for the order Enterobacterales ord. nov. which consists of seven families: Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morganellaceae fam. nov., and Budviciaceae fam. nov. (Adeolu et al. 2016).

The first dataset used for phylogenetics in this study was assembled by selecting three genera within each of the proposed new families of the Enterobacterales ord. nov. (spelling is correct for the proposed new order). Then, at least one species was selected for each genera, and each species was the same type species used in the Adeolu et al. (2016) study. The second dataset was compiled after the phylogenetic results from the first dataset indicated that the isolate clustered within the *Yersinia* clade. Sequences were selected from the top match lists generated from RDP and BLAST as shown in Table 3.

Each dataset was aligned in MEGA7. MEGA7 offers two methods for aligning nucleotide sequences: ClustalW (Thompson et al. 1994) and MUSCLE (Multiple Sequence Comparison by Log-Expectation; Edgar 2004). Although ClustalW is more widely used, MUSCLE is slightly more accurate (Nuin et al. 2006). MUSCLE is also 2-5 times faster using typical-size data sets, and over 80,000 times faster for a set of 5000 sequences of average length 350 (Edgar 2004, Hall 2011).

In order to reconstruct evolutionary trees, some assumptions must be made about the nucleotide substitution process. Models state those assumptions and determine the way in which a program calculates branch lengths. Branch lengths are intended to indicate the amount of genetic change between an ancestor and its descendants (Hall 2011). In this study, the Maximum Likelihood method was used to generate the phylogenetic trees, and the Hasegawa-Kishino-Yano model was specified, which distinguishes between transitional substitution rates among purines and transversional substitutions rates among pyrimidines (Hall 2011). This model was chosen using MEGA's built-in "Find Best DNA/Protein Models (ML)" feature. It finds the best model by using the nucleotides in the data set and determining the highest log likelihood of each model. For each of these models, MEGA provides the estimated values of the shape parameter of the

Gamma distribution, the proportion of invariant sites, and the substitution rates between bases or residues, as applicable (Tamura et al. 2011).

*Haemophilus influenzae* was chosen as the outgroup to root all of the phylogenetic trees. This was based on a study done by Williams et al. (2010). They determined the phylogeny of the bacterial class Gammaproteobacteria using a set of 356 protein families for the entire class. Their results showed that the order Pasteurellales was a sister group to the order “Enterobacteriales”. The taxonomy of *Haemophilus* is: Domain, Bacteria; Phylum, Proteobacteria; Class, Gammaproteobacteria; Order, Pasteurellales; Family, Pasteurellaceae; Genus, *Haemophilus*. Adeolu et al. (2016) also used members of the family Pasteurellaceae as outgroups in determining the phylogeny of the order “Enterobacteriales”.

Figure 6 shows the most likely tree generated using the maximum likelihood method and the data set for the order “Enterobacteriales”. This phylogram shows the distance between the outgroup *Haemophilus influenzae* and the rest of the data set. *Haemophilus influenzae* is in the order Pasteurellales, and the isolate LD2 is grouped in the order “Enterobacteriales”. The tree is drawn to scale with branch lengths measured in the number of substitutions per site; that is, the number of substitutions as calculated by the user-selected model of nucleotide substitutions divided by the length of the sequence (Tamura et al. 2011). It indicates that the isolate LD2 clusters with *Yersinia ruckeri* with a high bootstrap value of 94%. Figure 7 is the same tree as Figure 6 shown in topology only format so that bootstrap values and branching order can be seen more clearly. A maximum parsimony tree was also run (not shown) that confirmed the grouping of the isolate with *Y. ruckeri* with a lower bootstrap value of 62%.

Figure 8 is the most likely tree run from the data set containing several closely related genera in the Enterobacteriaceae family. It is a phylogram which shows the distance between

different genera and species. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This figure emphasizes two things: 1) that there is a great distance between the outgroup *Haemophilus influenzae* and all the rest of the sequences in the Enterobacteriaceae family, and 2) there is very little distance between the sequences of the closely related genera of the family Enterobacteriaceae, and LD2 is very close distance-wise to those sequences. This would suggest that the placement of LD2 within the family of Enterobacteriaceae is supported by this phylogenetic tree.

Figure 9 is the same tree as in Figure 8 except that it is shown in topology only format so that bootstrap values and branching order can be seen more clearly. This tree contradicts the most likely tree from the dataset for the order “Enterobacteriales” (Figures 6 and 7). It does not group the isolate with *Yersinia ruckeri*, but instead, shows it as a clade of its own between the *Budvicia* clade and a clade that contains both *Y. ruckeri* and *Y. kristensenii*. *Y. ruckeri* is the genetically most distant species within the *Yersinia* genus (Kotetishvili et al. 2005). This might explain why LD2 clustered with *Y. ruckeri* in the data set for the order “Enterobacteriales” since *Y. ruckeri* is the only representative of the genus *Yersinia* in that set. When the isolate is grouped with other closer members of the family Enterobacteriaceae, it does not cluster with *Y. ruckeri*, but in fact, appears to be a clade of its own.

Figure 10 shows the majority rule bootstrap consensus tree from the data set for the family Enterobacteriaceae. It was run under the same program and parameters as Figures 8 and 9. All branches with bootstrap values less than 50% are collapsed. This implies that the branching order of the isolate is unresolved with regard to the other members of the Enterobacteriaceae family. Adeolu et al. (2016) noted that phylogenetic trees produced based on the 16S rRNA gene sequence exhibited a limited ability to resolve the clades that were identified

in genome, ribosomal protein, and MLSA based phylogenetic trees. Additionally, the branching of the genera and species within the order “Enterobacteriales” in 16S rRNA gene based phylogenies shows considerable stochasticity depending on the algorithms used and the organisms analyzed. Overall, the results obtained in their study substantiated previous suggestions that the 16S rRNA gene possesses limited utility in accurate phylogenetic reconstruction of inter-genus level relationships within the order “Enterobacteriales” (Adeolu et al. 2016).

## CONCLUSIONS AND POSSIBLE FUTURE WORK

During this study, a polyphasic approach was deployed in order to characterize and identify a novel species of bacterium isolated from Kephart Prong, Great Smoky Mountains National Park. A number of methods were used for the complete characterization of the isolate including phenotypic, genotypic, and phylogenetic analysis. Phenotypic analysis revealed that the isolate LD2 is Gram-negative, rod-shaped, oxidase negative, catalase positive, and grows in the presence and absence of oxygen. These are the typical characteristics that define members of the family Enterobacteriaceae (Brenner et al. 2005). Comparison of biochemical and environmental characteristics indicated that the isolate was most similar to *Pantoea agglomerans*; however, much data was not available for this species. Phenotypic evidence did not support the inclusion or exclusion of the isolate as a member of *Pantoea*, *Yersinia*, *Shigella*, or *Budvicia*.

Genotypic analysis using the RDP classifier identified the isolate down to the family Enterobacteriaceae with a 100% bootstrap value. The bootstrap value for assignment to the genus *Yersinia* was only 52%; not a significant value. RDP Seqmatch, BLAST, and EzBioCloud all indicated that the isolate had the highest pairwise nucleotide sequence similarity with *Y. ruckeri* and *Y. kristensenii*. The G+C content of the isolate appears to be consistent with genera *Yersinia*, and *Budvicia*. MLST analysis using schemes for both *Yersinia* ssp. and *Y. ruckeri* returned low percent identities for all genes in the schemes except the *thrA* gene at 94%. This is still below the threshold used for the 16S rRNA, although no standards have been recommended for MLST. These results would seem to indicate that the isolate is not a member of the genus *Yersinia*. In

addition, the results of One Codex whole genome analysis would suggest that the isolate LD2 may be a member of the genus *Serratia*.

Phylogenetic analysis appears to support identification of LD2 in the order “Enterobacteriales” and the family Enterobacteriaceae. Phylogenetic analysis was unresolved at placing the isolate within a genus; however, analysis of the entire genome via One Codex seems to indicate the isolate may be a member of the genus *Serratia*. Therefore, on the basis of phenotypic, genotypic, and phylogenetic properties, isolate LD2 likely represents a novel species and likely a novel genus.

The significance of this project can be underscored by the possible impact this isolate may have on the environment or on other organisms as indicated by putative genes in Table 6. Eighteen genes are associated with hemin transport and/or siderophores. Siderophore is a word taken from the Greek language that literally means “iron carrier”. Iron is essential for almost all life for processes such as respiration and DNA synthesis. Despite being one of the most abundant elements in the Earth’s crust, the bioavailability of iron in many environments is limited by the very low solubility of the ferric ion. Siderophores are small compounds with a high-affinity for ferric iron. They are secreted by bacteria into the extracellular environment to form soluble  $Fe^{3+}$  complexes that can be taken up by active transport mechanisms (Neilands 2015). Siderophores are also important for some pathogenic bacteria in their acquisition of iron. In some hosts, iron is tightly bound to proteins such as hemoglobins. Bacteria can release siderophores to scavenge ferric iron from iron proteins (Hider and Kong 2010). This would suggest that the isolate LD2 lives within a host organism as a parasite.

Other significant genes listed in the annotated genome of LD2 were three hemolysin genes. These genes function in the lysis of red blood cell membranes. There were also three

invasions genes. Most invasins are proteins that act locally to damage host cells and/or have the immediate effect of facilitating the growth and spread of a pathogen (Todar 2012). This evidence, taken together with many genes for type III, IV, VI secretion systems as well as virulence factor and multidrug resistance genes suggest this isolate is a possible pathogen.

This project started out as part of the ATBI effort to identify all taxa in GSMNP. These efforts add to a growing database of species locations, habitats, genetic diversity, population density, symbiotic relationships, and organismal interactions. It is a way to discover new species that interact with their habitat, to identify new threats in time to act, and to understand how to manage and protect a complex and valuable ecosystem like the Smoky Mountains. This study has indeed led to the discovery of a new species of bacteria that could be beneficial, or conversely pathogenic, in the environment and to other organisms including humans. Species in the family Enterobacteriaceae are found worldwide. Many species are of considerable economic importance due to pathogenicity to commercial crops and stock including poultry farms and fisheries (Brenner et al., 2005). Benefits of investigating this new isolate could be the discovery of economically useful biochemicals and determining what role this organism plays in the environment. In addition, this study advances the science of documenting microbial diversity within western North Carolina.

Future work that may be added to this study could include biochemical tests to determine if the isolate is a member of the genus *Serratia*. These tests would include the ability to ferment fructose, d-galactose, maltose, d-mannitol, d-mannose, ribose, l-fucose, and trehalose and utilize them as sole carbon sources. L-sorbose is not fermented or utilized as sole carbon source by members of *Serratia*. All species but *S. fonticola* fail to ferment or utilize dulcitol (LD2 cannot)

and tagatose (Brenner, et al., 2005). The isolate could also be grown on media incorporated with various antibiotics to determine its resistance.

The genome-based methods used by Adeolu et al. (2016) seemed to produce reliable results in determining the phylogeny and taxonomy of the “Enterobacteriales”. Of particular interest is their use of conserved signature insertions/deletions, which are specifically shared by the members of the clades they identified and independently support their monophyly and distinctness. Perhaps this is a feasible project that could be carried out to help identify isolates from GSMNP. In addition, *in silico* primers could be used to extract MLST gene sequences from the whole genome sequence of the isolate using different schemes. Then these sequences could be used to produce phylogenetic trees to compare to the trees already produced.

This study adds to the growing amount of information amassing as part of the ATBI in Great Smoky Mountains National Park. This information may enlighten park officials in areas of soil and water management. This new species of bacterium was taken from a water sample that feeds into the Oconaluftee River. Genetic analysis indicated this isolate may be closely related to *Yersinia ruckeri*, which is a species of bacterium that causes enteric redmouth disease in rainbow trout. The annotated sequence file for the isolate indicated the presence of many genes that are associated with pathogenicity and parasitism. Western North Carolina has a fishing tourism industry that relies upon a healthy population of rainbow trout. Further investigation of this isolate could help protect this vital industry.



## REFERENCES

- Adeolu M, Alnajar S, Naushad S, Grupta R. 2016. Genome-based phylogeny and taxonomy of the 'Enterobacteriales': proposal for Enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morganellaceae fam. nov., and Budviciaceae fam. nov. *International Journal of Systematic and Evolutionary Microbiology*. 66:5575–5599.
- Albers SV, Forterre P, Prangishvili D, Schleper, C. 2013. The legacy of Carl Woese and Wolfram Zillig: from phylogeny to landmark discoveries. *Nature Reviews Microbiology*. 11(10):713–719.
- Altschul SF, Madden TL, Schaffer, AA, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*. 25:3389–3402.
- AREERA Annual Report of Accomplishments and Results. 2004 Mar [accessed 2017 Mar 6].  
<https://portal.nifa.usda.gov/web/areera/Reports/2003/Nc/Combined.CES.NC.pdf>
- Berersen FJ, Turner, GL. 1968. Comparative studies of nitrogen fixation by soybean root nodules, bacteroid suspensions and cell-free extracts. *Microbiology*. 53:205–220.
- Brenner DJ, Krieg NR, Staley JT, Garrity GM, Boone DR, De Vos P, Goodfellow M. (Eds.). 2005. *Bergey's manual of systematic bacteriology: Volume two the Proteobacteria Part B the Gammaproteobacteria (Second ed.)*.
- Bryant MP. 1959. Bacterial species of the rumen. *Microbiology and Molecular Biology Reviews*. 23:125–148.
- Campbell NA, Reece JB. 2005. *Biology*. San Francisco: Pearson, Benjamin Cummings.

- Cole JR, Tiedje JM. 2014. History and impact of RDP. *RNA Biology*. 11(3):239–243.
- Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, and Tiedje JM. 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*. 42(Database issue):633–642.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 32(5):1792–1797.
- Emerson D, Agulto L, Liu H, Liu L. 2008. Identifying and characterizing bacteria in an era of genomics and proteomics. *BioScience*. 58(10):925.
- Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*. 39:783-791.
- Fournier P-E, Suhre K, Fournous G, Raoult D. 2006. Estimation of prokaryote genomic DNA G+C content by sequencing universally conserved genes. *International Journal of Systematic and Evolutionary Microbiology*. 56(5):1025–1029.
- Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR. 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences USA*. 74:4537–4541.
- Gans J, Wolinsky M, Dunbar J. 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*. 309:1387–1390.
- Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Peer YVD, Vandamme P, Thompson FL, et al. 2005. Opinion: Re-evaluating prokaryotic species. *Nature Reviews Microbiology*. 3(9):733–739.
- Glaeser SP, Kämpfer P. 2015. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology*. 38(4):237–245.

- Haeckel E. 1866. *Generelle Morphologie der Organismen*.
- Hagen JF. 2012 Five kingdoms, more or less: Robert Whittaker and the broad classification of organisms. *BioScience*. 62(1):67-74.
- Hall BG. 2011. *Phylogenetic trees made easy: a how-to manual*. Sunderland, MA: Sinauer Associates.
- Hasegawa M, Kishino H, and Yano T. 1985. Dating the human-ape split by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*. 22:160-174.
- Heimbrook ME, Wang WL, Campbell G. 1989. Staining Bacterial Flagella Easily. *Journal of Clinical Microbiology*. 27(11):2612–2615.
- Hider RC, Kong X. 2010. Chemistry and biology of siderophores. *Natural Products Reports*. 27: 637–657
- Huang Y, Runge M, Michael G, Schwarz S, Jung A, Steinhagen D. 2013. Biochemical and molecular heterogeneity among isolates of *Yersinia ruckeri* from rainbow trout (*Oncorhynchus mykiss*, Walbaum) in north west Germany. *BMC Veterinary Research*. 9(1):215.
- International Committee on Systematics of Prokaryotes. 2017. [Internet, cited 2017, February 26]. Available from <http://www.the-icsp.org>
- Janssen PH. 2006. Identifying the dominant soil bacterial taxa in libraries of 16S rDNA and 16S rRNA genes. *Applied and Environmental Microbiology*. 72:1719–1728.
- Joseph SJ, Hugenholtz P, Sangwan P, Osborne CA, Janssen PH. 2003. Laboratory cultivation of widespread and previously uncultured soil bacteria. *Applied and Environmental Microbiology*. 69(12):7210–7215.

- Kim M, Oh H-S, Park S-C, Chun J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*. 64(Pt 2):346–351.
- Kotetishvili M, Kreger A, Wauters G, Morris JG, Sulakvelidze A, Stine OC. 2005. Multilocus sequence typing for studying genetic relationships among *Yersinia* species. *Journal of Clinical Microbiology*. 43(6):2674–2684.
- Kumar G, Menanteau-Ledouble S, Saleh M, El-Matbouli M. 2015. *Yersinia ruckeri*, the causative agent of enteric redmouth disease in fish. *Veterinary Research*. 46(1):1-10.
- Kumar S, Stecher G, Tamura K. 2016. Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*. 33(7):1870–1874.
- Lang E, Schumann P, Knapp BA, Kumar R, Sproer C, Insam H. 2013. *Budvicia diplopodorum* sp. nov. and emended description of the genus *Budvicia*. *International Journal of Systematic and Evolutionary Microbiology*. 63(Pt 1):260–267.
- Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM and Lund O. 2012. Multilocus sequence typing of total genome sequenced bacteria. *Journal of Clinical Microbiology* 50(4): 1355-1361.
- Linnaeus C. *Systema Naturae*. Ninth edition. Theodor Haak, Leiden.
- Liofilchem. 2013. *EnteroPluri-Test*©: Codebook [Internet]. Available from [https://www.bd.com/ds/technicalCenter/inserts/L010569\(02\).pdf](https://www.bd.com/ds/technicalCenter/inserts/L010569(02).pdf)
- Madigan MT, Martinko JM, Bender KS, Buckley DH, Stahl DA. 2015. *Brock biology of microorganisms*. New York: Pearson.

- Mesbah M, Premachandran U, Whitman WB. 1989. Precise measurement of the G C content of deoxyribonucleic acid by high-performance liquid chromatography. *International Journal of Systematic Bacteriology*. 39(2):159–167.
- Minot SS, Krumm N, Greenfield NB. 2015. One Codex: A sensitive and accurate data platform for genomic microbial identification [Internet]. *BioRxiv*; [cited 2017 Apr 6]. Available from <http://biorxiv.org/content/biorxiv/early/2015/09/28/027607.full.pdf>
- Myers EW, Miller W. 1988. Optimal alignments in linear space. *Bioinformatics*. 4(1):11–17.
- National Park Service. 2015. What we do [Internet]. [cited 2017 Mar 10]. Available from <http://www.nps.gov/aboutus/index.htm>
- National Park Service. 2017. Maps [Internet]. [cited 2017 Mar 10]. Available from <https://www.nps.gov/grsm/planyourvisit/maps.htm>
- Nei M, Kumar S. 2000. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.
- Neilands JB. 1995. Siderophores: Structure and function of microbial iron transport compounds. *Journal of Biological Chemistry*. 270(45):26723–26726.
- Nichols BJ, Langdon KR. 2007. The Smokies All Taxa Biodiversity Inventory: History and progress. *Southeastern Naturalist*. 1:27–34.
- Nuin PA, Wang Z, Tillier ER. 2006. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*. 7:471.
- O'Connell SP, York EA, Collins MB, Rosbach DT, Black KR, Haney WB. 2007. An initial inventory of bacteria found within the soils and waters of Great Smoky Mountains National Park: A Search for Species in Our Own Backyard. *Southeastern Naturalist*. Special Issue 1:57-72

- Pfennig N, Widdel F. 1982. The bacteria of the sulphur cycle. *The Philosophical Transactions of the Royal Society B*. 298:433–441.
- Prakash O, Verma M, Sharma P, Kumar M, Kumari K, Singh A, Kumari H, Jit S, Gupta SK, Khanna M, et al. 2007. Polyphasic approach of bacterial classification: An overview of recent advances. *Indian Journal of Microbiology*. 47(2):98–108.
- PubMLST. Multilocus Sequence Typing [Internet]. [cited 2017 Mar 6]. Available at <https://pubmlst.org/general.shtml>
- Stackebrandt E, Ebers J. 2006. Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today*. 33:152–155.
- Stackebrandt E, Garrity GM, Trüper HG, Whitman WB, Grimont PAD, Nesme X, Frederiksen W, Vauterin L, Kämpfer P, Rosselló-Mora R, et al. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology*. 52(3):1043–1047.
- Stackebrandt E, Goebel BM. 1994. Taxonomic Note: A Place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology*. 44(4):846–849.
- Swofford DL. 2002. PAUP\*. *Phylogenetic Analysis Using Parsimony (\*and Other Methods)*. Version 4. Sunderland, Massachusetts: Sinauer Associates.
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*. 28(10):2731–2739.

- Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. 22(22):4673–4680.
- Tindall BJ, Busse H-J, Ludwig W, Rosselló-Móra R, Kämpfer P. 2010. Notes on the characterization of prokaryote strains for taxonomic purposes. *International Journal of Systematic and Evolutionary Microbiology*. 60(1):249–266.
- Todar K. 2012. Colonization and invasion by bacterial pathogens [Internet]. Madison (WD); [cited 2017 Mar 25]. Available from [http://textbookofbacteriology.net/colonization\\_3.html](http://textbookofbacteriology.net/colonization_3.html)
- Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*. 73(16):5261–5267.
- Wayne LG, Moore WEC, Stackebrandt E, Kandler O, Colwell RR, Krichevsky MI, Truper HG, Murray RGE, Grimont PAD, Brenner DJ, et al. 1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic and Evolutionary Microbiology*. 37(4):463–464.
- White P, Langdon K. 2006. The ATBI in the Smokies: An Overview. *The George Wright Forum*. 23(3):18–25.
- Williams KP, Gillespie JJ, Sobral BWS, Nordberg EK, Snyder EE, Shallom JM, Dickerman AW. 2010. Phylogeny of Gammaproteobacteria. *Journal of Bacteriology*. 192(9):2305–2314.

- Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*. 87(12):4576–4579.
- Woo PCY, Lau SKP, Teng JLL, Tse H, Yeun KY. 2008. Then and now: Use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *European Society of Clinical Microbiology and Infectious Diseases*. 14:908–934.
- Xu HX, Kawamura Y, Li N, Zhao L, Li TM, Li ZY, Shu S, Ezaki T. 2000. A rapid method for determining the G C content of bacterial chromosomes by monitoring fluorescence intensity during DNA denaturation in a capillary tube. *International Journal Of Systematic And Evolutionary Microbiology*. 50(4):1463–1469.
- Yoon SH, HaSM, Kwon S, Lim J, Kim Y, Seo H, Chun J. 2017. Introducing EzBioCloud: A taxonomically united database of 16S rRNA and whole genome assemblies. *International Journal of Systematic and Evolutionary Microbiology*. doi:10.1099/ijsem.0.001755.
- Zaika LL, Phillips JG. 2005. Model for the combined effects of temperature, pH and sodium chloride concentration on survival of *Shigella flexneri* strain 5348 under aerobic conditions. *International Journal of Food Microbiology*. 101(2):179–187.
- Zhang L, Xu Z. 2008. Assessing bacterial diversity in soil. *Journal of Soil and Sediments*. 8:379–388.

## APPENDIX

An Excel file of the annotated genome of the isolate LD2 can be viewed by click on the link below:

<https://drive.google.com/file/d/0Byeoacs59RSTeEpjX0VJTFJJZ2c/view?usp=sharing>