Leveraging Third Generation Sequencing And Novel Sequence Analysis Algorithms For Rapid And Efficient Amplicon-Based Detection Of Foodborne Pathogens

by

ALEXANDRA N. FUTRAL

(Under the Direction of Hendrik den Bakker)

ABSTRACT

Current databases for taxonomic classification of bacteria are very large, computationally expensive and allow little to no customization. This study served to decrease computational time and memory required to taxonomically identify foodborne pathogens through incorporating the novel ColorID algorithm with the Nanopore MinION™ and to determine the limit of detection of *Salmonella* Enteritidis and *Listeria monocytogenes*. Various fecal samples were prepared into 16S libraries and sequenced via the MinION™. The sequencing speed, computational power and efficiency were determined with ColorID using an "all-bacteria" database compared with a database consisting of relevant foodborne pathogens. Analyses were compared to bioinformatic pipelines within QIIME2 for Illumina data. The MinION™/ColorID method using a "pathogen-specific" database was more computationally efficient than an "all-bacteria" database or QIIME2. The limit of detection of the MinION™/ColorID method was 1.7

log and 4.1 log CFU/ml for *Salmonella* Enteritidis and *Listeria monocytogenes*, respectively. These findings could greatly reduce computational time and resources needed to detect pathogens, which could be used for many applications related to food safety.

INDEX WORDS: *Listeria, Salmonella*, MinION™, MiSeq, ColorID, QIIME2, sklearn, Vsearch, 16S, Goose, k-mer, database

Leveraging Third Generation Sequencing And Novel Sequence Analysis Algorithms For Rapid And Efficient Amplicon-Based Detection Of Foodborne Pathogens

by

ALEXANDRA N. FUTRAL

B.S. Biology, University of Georgia, 2014

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2020

Leveraging Third Generation Sequencing And Novel Sequence Analysis Algorithms For Rapid And Efficient Amplicon-Based Detection Of Foodborne Pathogens

by

ALEXANDRA N. FUTRAL

Major Professor:	Hendrik den Bakker
Committee:	Michael Rothrock, Jr.
	Abhinav Mishra
	Manpreet Singh

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
May 2020

DEDICATION

This project is dedicated to my family, friends and colleagues that have helped me through every moment. Mainly, I dedicate this project to my late father, Donald Futral, who passed in May of 2019. He supported me through the trials of grad school, listened to presentation practices with enthusiasm, and his spirit inspired me to finish what I started. This project represents a continuation of the Futral contributions from my late grandfather, J Gordon Futral, who essentially paved the way (literally) for the agricultural engineering department. When I look at the road named after him, "Gordon Futral Court," it makes me proud to be a part of the system that he helped create. I thank my mother, Lynn Futral, for her constant praise of my work, though quite embellished I might say. Lastly, I would like to thank my fiancé, Jack Muka, for his continuous support and advice.

ACKNOWLEDGEMENTS

First, I would like to thank my co-advisors, Dr. Henk den Bakker and Dr. Michael Rothrock Jr. for their constant support, wisdom and advice for this project. I thank Dr. den Bakker for providing me valuable lessons and allowing me to learn without feeding me answers and trusting in me to complete this project. I thank Dr. Rothrock for being a support system from a distance and providing different perspectives for this project. I would also like to thank Dr. Manpreet Singh and Dr. Abhinav Mishra for serving on my committee and being available whenever needed.

I would like to thank Amy Mann for guiding me through my lab work, ordering essential materials, preparing DNA samples and teaching me essential practices. Thank you to David Mann in Dr. Xiangyu Deng's lab and Meghan den Bakker in Dr. Francisco Diez's lab for allowing me to use essential equipment and reagents that were of utmost importance in my project.

I would like to thank everyone at the University of Georgia Griffin Campus for providing a wonderful atmosphere to work and collaborate. Thank you to Wayne Harvester for setting up essential Zoom appointments and checking in on me for IT purposes. I thank the administrative staff for providing necessary coffee to get me through my mornings. Lastly, I would like to thank my food science friends/colleagues at that provided constant support and made things more manageable throughout classes as well as research.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Methods to identify the source of a foodborne outbreak are necessary to prevent foodborne infections, which could lead to hospitalizations or death (Grutzke et al., 2019). Additionally, the detection of pathogens in various matrices could prevent outbreaks from happening initially; however, it is usually very difficult and expensive to use detection tools on a routine basis to prevent outbreaks. When outbreaks do occur, it can take days to weeks to receive confirmation of a pathogen in a sample to initiate a traceback (Bibby, Ma & Stachler, 2017).

The purpose of this study is to find a method of taxonomic classification and detection that is inexpensive, easy to use and computationally efficient in terms of speed and power to identify pathogens by combining the Nanopore MinION™ with the novel ColorID algorithm. Although there have been numerous studies on the MinION™ in combination with other taxonomic classifiers (Benítez-Páez et al., 2016; Bibby, Ma & Stachler, 2017; Kai et al., 2019; Nygaard et al., 2020), there have not been many that balance both speed and computational power of taxonomic classification. Typically, a classifier/algorithm prioritizes speed over computational efficiency or vice-versa (Ainsworth, Sternberg, Raczy, & Butcher, 2017). The combination of these qualities could not only provide rapid methods to find the source of an outbreak, but also provide routine cost-efficient monitoring in processing environments.

The current study was performed by comparing the taxonomic classification accuracy of the MinION™/ColorID sequencing method with that of QIIME2 for Illumina MiSeq data (the current gold standard). The computational resources (in terms of speed and memory) were then compared between the two methods. Furthermore, the computational resources of ColorID analysis were tested with a reference database using "all-bacteria" versus one consisting of relevant foodborne pathogens. The limit of detection of this method was tested using *Listeria monocytogenes* and *Salmonella* Enteritidis, two pathogens absent in the initial analysis of the microbiome of goose feces. These pathogens were spiked in controlled quantities (~6, 4 and 2 log CFU/ml) into goose fecal samples. Subsequent DNA extraction, MinION™ sequencing and ColorID analysis was then conducted to determine the limit of detection. We hypothesized that the MinION™/ColorID sequencing method would be faster and more computationally efficient than the MiSeq/QIIME2 sequencing method, and that the "all-bacteria" database would be more computationally efficient than a database consisting of relevant foodborne pathogens. The reduced computational resources could eliminate the need for a computer and allow operations on a smartphone or home laptop, which could be used outside of the laboratory. If verified, these expectations, combined with the expected detection of certain foodborne pathogens, could result in the creation of an expedient and convenient system to use within the food industry and beyond.

CHAPTER 2

LITERATURE REVIEW

Identification of Foodborne Pathogens

Foodborne pathogens, including bacteria, viruses, fungi and some parasites, have become a significant health problem recognized by the public and governmental agencies. Certain foodborne bacteria, such *as Salmonella enterica, Listeria monocytogenes, Campylobacter jejuni, Escherichia coli O157:H7, Staphylococcus aureus, Vibrio* spp. and *Bacillus cereus* are leading causes of foodborne illness (Zhao, Wang & Oh, 2014). The Center for Disease Prevention and Control (CDC) estimate that as many as 128,000 hospitalization and 3000 deaths occur annually from contaminated food and drinking water (Vidic et al., 2019). Identification and characterization of bacteria is extremely important in the realm of food safety for many reasons including health implications, tracing foodborne illnesses back to a source to food processes and understanding the fermentations of foods (Rhoads, Wolcott, Sun, & Dowd, 2012).

Conventional methods for foodborne pathogen detection, while specific in terms of minimizing false positives, are expensive and time consuming, as they typically involve preliminary identification with selective media, which can take up to 3 days as well as the biochemical identification which can take a week or longer. As expected, these methods are also labor intensive, as they involve preparation of media, inoculation and colony counting, all of which increase the potential to introduce human error (Law, Ab Mutalib, Chan, & Lee, 2014). They also require specific conditions, such as optimal

composition of the media, specific incubation temperatures, and defined atmospheric conditions (Vidic et al. 2019). As stated, these methods depend on the microorganism to grow in the media, a condition which has presented a problem for detecting viable, but non-culturable bacteria (Law et al. 2014). These conventional methods thus have high specificity at the cost of decreased sensitivity.

Molecular methods, in contrast, have been shown to be both sensitive and specific, with the ability to detect relevant bacteria that have not been well described (Rhoads, Wolcott, Sun, & Dowd, 2012).

Multiplex PCR is a nucleic acid-based method of detection that is faster than traditional PCR and can amplify multiple gene targets. It can detect up to five or six pathogens simultaneously (Law et al. 2014, Zhao et al. 2013). However, it generally requires a pre-enrichment step to increase the number of cells and prevent detecting DNA from dead bacteria (Vidic et al., 2019).

The incorporation of DNA sequencing methods allows improvements to traditional identification techniques, eliminating the need to isolate colonies, allowing uncultivable microorganisms to be studied and decreasing turn-around time (Rhoads, Wolcott, Sun, & Dowd, 2012). Conventional techniques, including gram staining and culture-based methods, can identify only identify around 0.1% of the bacterial communities that exist and require isolation of pure cultures. As a result, they do not give us insight into an important aspect of food matrices, known as the microbiome (Rhoads, Wolcott, Sun, & Dowd, 2012).

DNA Sequencing

DNA sequencing began with Sanger sequencing, named after Frederick Sanger, who, in 1975, determined a rapid method for determining the base pairs of DNA constituting a sequence. This was done through priming synthesis with DNA polymerase and chain-terminating dideoxynucleotides. This method was widely used for approximately four decades. In 2003, the Human Genome Project came with refinements in Sanger's method of sequencing (Pareek, Smoczynski, & Tretyn, 2011). These refinements ultimately led to the development of next generation sequencing (NGS) technologies. These NGS technologies have revolutionized many fields in terms of clinical diagnoses, outbreak investigations, forensics, antimicrobial resistance and more. The versatility of these instruments has allowed numerous modifications to the technologies and software to interpret these results (Jagadeesan et al., 2019). Next generation sequencing technologies have enabled more extensive characterization of bacterial genomes as well as taxonomic characterization of various microbiomes (Cao, Fanning, Proos, Jordan, & Srikumar, 2017).

In terms of food safety, NGS technologies are typically used in two ways: Whole Genome Sequencing (WGS) and Shotgun Metagenomics or Amplicon-based sequencing. WGS involves the complete sequencing of the entire genome of a single species from an isolate. This is useful for the surveillance of foodborne pathogens and outbreak detection/root cause analysis of contamination events because of the increase in molecular information gathered across the entirety of the genome allowing determination of genetic relatedness between strains. This approach has steadily replaced traditional

forms of microbial identification, which rely on subtyping sequence changes across a small portion of the genome (Jagadeesan et al., 2019).

Shotgun metagenomics, in contrast to WGS, allows random sequencing of the genome of the entire microbiome without bacterial culture. Amplicon sequencing or metabarcoding, is another approach that does not require bacterial culture, but involves amplification and sequencing of specific target genes by PCR amplification (Jagadeesan et al., 2019). These targets of the PCR amplification, also known as amplicons, are commonly sections of the 16S RNA gene in bacteria and the 18S RNA gene in fungi. The 16S rDNA gene is typically used in studies of microbial diversity because the structure consists of alternating conserved and variable regions. The conserved regions evolve slowly, which allows universal primers to amplify these genes across different taxa, while the variable regions evolve more rapidly, allowing for the detection of taxonomic differences (Jovel et al., 2016).

The Illumina MiSeq

The Illumina MiSeq and Illumina HiSeq are currently the most widely used platforms encompassed by NGS (Allali et al., 2017; Jagadeesan et al., 2019; Pareek, Smoczynski, & Tretyn, 2011), although other sequencing platforms producing similar data exist, such as Ion Torrent and SOLiD, owned by Life Technologies™. The MiSeq works by sequencing clusters of DNA in parallel to provide a strong signal strength. The technology is a ‘sequencing by synthesis’ method, in which new bases are integrated into copies of the original DNA template and are measured and recorded (Jovel et al., 2014), which allows high throughput of DNA; however, since the MiSeq sequences DNA in

parallel, it produces short read lengths around 100-300 base pairs long. These reads can still be assembled into incomplete genomes, which can be used for many applications.

The MiSeq produces reads with an accuracy of over 99%; however, because it produces relatively short read lengths, the short read length may result in a decrease in sequence information, which affects correct classification of bacteria species (Kai et al., 2019). In addition, the MiSeq as well as other NGS platforms require a high start-up cost and can take days to weeks from extraction to analysis of DNA (Bibby, Ma, & Stachler, 2017).

Single-molecule sequencing, in contrast to parallel sequencing, has the capability to produce long reads of up to 100 kb in length. These technologies are very useful for determining complex genomic regions involving genomic rearrangements and repetitive regions. Oxford Nanopore is a company that creates technologies which perform single-molecule sequencing (Jagadeesan et al., 2019). The ability to sequence a single molecule at a time enables new functions such as real-time data analysis and has distinguished these technologies from traditional NGS technologies. Therefore, they are collectively referred to as third generation sequencing (Benítez-Páez, Portune, & Sanz, 2016).

Oxford Nanopore MinION™

The Nanopore MinION™ developed by Oxford Nanopore in 2014 is a third generation DNA sequencer that has several advantages over traditional platforms, in that it is the size of a USB thumb drive and operates from a computer by a USB port, allowing much greater portability (Benítez-Páez, Portune, & Sanz, 2016). It also has a much smaller start-up cost at $1000 for the MinION™ itself. It produces much longer reads (up to 100kb) than second generation sequencers such as the Miseq (up to 300

bases) and can target the entire coding region of the 16S rDNA gene, which provides detection with more accuracy and sensitivity than the MiSeq (Kai et al., 2019).

The MinION™ works through a sensor measuring changes in ionic current while DNA passes through one of the hundreds of nanopores. This change allows identification of the different DNA bases by the different charges they generate (Magi, Semeraro, Mingrino, Giusti, & D'Aurizio, 2018). In addition, the MinION™ provides real-time analysis of the results, since it outputs reads as it continues to sequence (Laver, 2014). The accuracy and DNA throughput of the MinION™ have increased through several updates to the flow cell chemistry. The updates in software and chemistry are continuous and the recent R9.4.1 flow cell chemistry has an accuracy of 85-90% (Maestri et al., 2019).

Although the error rate has improved as a result of newer versions of the technology, it is still lower than the 99% performed by other short read sequencers. This low read accuracy can be a problem for analysis of samples with single nucleotide variations (SNVs). Sequencing errors associated with Nanopore sequencing are generally attributed to the lack of signal change in homopolymers and the inconsistent speed through which DNA is threaded through the pore. In addition, the errors can be compounded during signal interpretation. However, consensus sequences can be obtained by sequencing the same samples through genome assembly, leading to accuracies of more than 99%. (Rang et al., 2018).

As stated previously, the 16S rDNA gene is used due to its' conserved nature and because all bacteria possess it. Therefore, one marker can be used to identify the majority of taxa found in a microbiome. The applicability of the 16S marker for

identification and differentiation of microbial organisms is preferred to shotgun metagenomics in terms of cost efficiency and microbiome characterization. Although shotgun metagenomics provides detection to the level of species and strain and can be used to study many aspects of the microbiome beyond taxonomic composition, 16S amplicon sequencing greatly decreases the cost of microbiome sequencing and can utilize specific software programs that are not available for shotgun metagenomic approaches (Laver, 2014).

Metagenomic Analyses for the MiSeq platform

Once sequenced, the data (sequences or 'sequence reads') need to be compared against a reference database to determine the taxonomy that each sequence corresponds to. The 16S region, although conserved, differs enough to distinguish among different taxonomic affiliations (Bokulich et al., 2018). There are numerous software programs to analyze 16S data, including QIIME and MOTHUR, in addition to databases, such as Greengenes and RDP to train the sequences to taxonomic information. One significant caveat with the 16S rDNA analysis is, since 16S has conserved regions that vary little between species, the taxonomic resolution can be poor at the species level and will not assign a species level assignment up to a third of the time (Laver, 2014). Another caveat with the 16S rDNA analysis is that many bacterial species harbor multiple copies of the 16S gene in the same genome, which in some cases have diverged from each other by up to 5%. This is significant because if it has diverged enough from a single reference sequence, an accurate taxonomic assignment may not be possible (Laver, 2014).

QIIME2, also known as Quantitative Insights into Microbial Ecology version 2, processes raw 16S reads sequenced through the MiSeq. QIIME2 uses plugins, such as

DADA2 (Callahan et al., 2016) and deblur (Amir et al., 2017), for quality control purposes by clustering/filtering reads into amplicon sequence variants (ASVs) to reduce sequence errors and dereplicate sequences. The filtered data is converted to taxonomic information through machine-learning and alignment-based classifiers through the plugin feature-classifier (Bokulich et al., 2018). Many different classifiers can be trained on sequence information, each with their own strengths and weaknesses for classification of 16S amplicons (Bokulich et al. 2018). QIIME2 and MOTHUR are standard software packages for analyzing 16S amplicon sequence data for the MiSeq. However, these software packages are not compatible with newer long read technologies, such as the MinION™.

Metagenomic Analyses for the Nanopore MinION™

Since the early days of metagenomics, several taxonomic classifiers have been developed to analyze the data sequenced. Kraken (Wood & Salzberg, 2014) was one of the first computational programs designed to quickly identify the reads in a metagenomic sample. Kraken is an exact alignment classifier which works by matching k-mers (smaller sub-sequences within a sequence read) to the lowest common ancestor in the database, and using a taxonomy based algorithm to infer the classification of an individual sequence read based on the information obtained from the individual kmers it contains (Wood & Salzberg, 2014). The length of the k-mers can be adjusted with trade-offs. Longer k-mers may have a decreased sensitivity and fail to match from sequencing errors or differences between same species due to mutations; however, they come with greater specificity. Shorter k-mers may have higher sensitivity but result in decreased specificity from matches to multiple genomes (Breitwieser, Lu, & Salzberg, 2019).

Although quick, Kraken requires 93 GB of RAM (Random-access memory) for just over 4000 genomes, which is more than the amount of memory on modern desktop computers. Improvements in RAM usage and speed have been made for this algorithm (Wood, Lu, & Langmead, 2019), but are still not sufficient for analysis on modern laptops or desktop computers.

As new classifiers have developed, increased computational efficiency has become a new focus in addition to rapid classification. The number of prokaryotic and viral genomes that have been sequenced has risen to 447,833 by October 2017 and is expected to double every two years (Bradley, Den Bakker, Rocha, Mcvean, & Iqbal, 2019). This rapid increase in the number of genomes available presents computational problems in terms of minimizing data processing and memory requirements for fast analysis of sequencing from sample to classification. Analysis for metagenomic classification initially used BLAST to compare reads with sequences deposited on a database called GenBank. However, with the increase in the number of genomes in the database, the computational power required became impractical (Breitwieser, Lu, & Salzberg, 2019).

One program that addresses this problem is known as Centrifuge. Centrifuge is based on the Ferragina-Manzini (FM) index and is modified to compress indices by removing similar genomes among the same species through a concatenation method. Utilizing the FM index also removes the sensitivity/specificity trade-off problem faced by Kraken and other similar programs by allowing the search of k-mers at any length (Kim, Song, Breitwieser, & Salzberg, 2016)

ColorID is a novel sequence algorithm that employs an exact search method rather than alignment to classify raw sequences (den Bakker, 2018) This utilizes an internal data structure known as the Bitsliced Genomic Signature Index (BIGSI), which is similar to the Sequence Bloom Tree (SBT), a search engine data structure which indexes for raw sequence data in the form of k-mers. Unlike SBT, BIGSI does not increase storage requirements with an increase in unique k-mers, but linearly increases with the number of datasets. This is important for bacteria, which horizontally transfer DNA, resulting in variety among the same species. Therefore, using the BIGSI data structure with bacteria is efficient because of the low k-mer sharing. This results in a significant decrease in storage space (Bradley, Den Bakker, Rocha, Mcvean, & Iqbal, 2019).

The BIGSI stores the k-mers in a bloom filter, which determines whether the k-mers are present in the dataset. Multiple k-mers must be present in the sequence, which reduces the false negative rate to near zero. Meanwhile, the false positive rate is user-determined through certain parameters related to the bloom filter (Bradley, Den Bakker, Rocha, Mcvean, & Iqbal, 2019).

ColorID incorporates the BIGSI data structure and is written in the Rust programming language to make it user friendly. ColorID can be used to search the sequence data in the form of a compressed fastq.gz file against a reference database to determine the percentage of k-mers shared or to classify the reads using a simple majority-rule algorithm. This reference database can be custom made to have any number/combination of species. If a database, also known as an index, is smaller, it will be more computationally efficient due to the properties of the BIGSI (den-Bakker, 2018). Though ColorID contains many characteristics that distinguish it from other popular data

analysis software platforms, it has yet to be studied with actual sequence data straight from the MinION™ sequencing platform.

NGS and third generation sequencing technologies have revolutionized our understanding of the gut microbiome. 16S rDNA sequencing has provided information about the taxonomic composition of a food microbiome, whereas whole metagenomic strategies have provided species-level to strain-level characterization (Cao, Fanning, Proos, Jordan, & Srikumar, 2017).

Gut Microbiome of Geese

The microbiome is a community of microorganisms that inhabit an environment. Microbiomes are very dynamic, varying between organisms and environments, and can be influenced by factors such as diet and climate. Thus, it is necessary to monitor population fluctuations of microorganisms within a microbiome to improve our understanding of complex food matrices (The Kavli Foundation, 2020).

The gut microbiome of any animal is essential not only for health in terms of food digestion, metabolism regulation, immune protection (Wu et al., 2018), but also in terms of food safety and public health in animals involved in food production (Rothrock & Locatelli, 2019). The microbiome refers to all the microorganisms, including their genes and metabolites belonging to an organism or a niche. Poultry gut microbiota are of unique importance due to characteristics associated with unique diets and physiological features, such as a short gastrointestinal tract and high energy demands for flight (Wang et al., 2018). Poultry gut microbiota differ immensely between species. However, analyses have shown that four phyla dominate*: Firmicutes, Proteobacteria*,

Actinobacteria, and *Bacteroidetes* (Wang et al., 2018; Drovetski et al., 2018; Rothrock & Locatelli, 2019).

The preharvest environment contributes significantly to the ecology of the poultry gut microbiome. Within poultry production facilities, environmental factors such as hygiene within hatcheries (Stanley et al., 2013), housing type (Ludvigsen et al., 2016), litter quality (Torok et al., 2009; Dumas et al, 2011) as well as external factors, such as diet and feed additives affect the gut microbiome (Pan & Yu, 2014; Walugembe et al., 2015; Videnska et al., 2013; Costa et al., 2017). Among wild animals, the gut microbiome could be influenced by pollution as well as other human activities (Wu et al., 2018). Seasonal changes also affect poultry microbiome composition, with studies showing that fewer bacteria genera are found in the winter than in the spring or summer (Oakley et al., 2018)

The gut microbiome of wild geese as well as other types of flying birds are different than land dwelling birds, such as chickens. This is due to excess weight inhibiting flying; however, quick digestion and excretion keeps weight at a minimum. Wild geese feed on cellulose rich diets and the throughput time from digestion to excretion varies from a few minutes to over an hour (Mattocks, 1971). However, the cecum retains digests for a much longer time than other gut regions (Mattocks, 1971). Therefore, microbiota from the cecum are more abundant, though less diverse, due to opportunities for stability of richness throughout many generations of bacteria (Wang et al., 2018). Microbiomes differ not only between species of poultry and within species of poultry, but also regionally among areas of the gut.

CHAPTER 3

MATERIALS AND METHODS

Relative Abundance and Computational Efficiency Analysis

Sample selection

Eight chicken, cow and goose fecal DNA samples were selected from a previous project which had been 16S amplicon and shotgun metagenomic sequenced on the Illumina MiSeq platform. These samples were sequenced with the MinION™ on the basis of containing an adequate DNA concentration, which was necessary to perform downstream processes.

16S library preparation

The samples described were previously extracted using the Zymo Quick-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research, Irvine, CA, USA) and quantified using qubit fluorometer 3.0 (Life Technologies by Invitrogen™, Carlsbad, CA, USA). To perform a single 16S rRNA gene sequencing run for multiple samples (i.e., a multiplex run), all samples were diluted to 10ng in 10µl of Nuclease Free water (IBI scientific, Dubuque, Iowa, USA) as the input DNA. PCR amplification of 16SrRNA genes was conducted using the 16S Barcoding Kit (SQK-RAB204; Oxford Nanopore Technologies, Oxford, UK) with LongAmp Taq 2X Master Mix (New England BioLabs® Inc., Ipswich, MA, USA) using a BioRad SmartSpec™3000 thermocycler. The thermocycler parameters were as follows: 95°C for 1 min; 25 cycles of 95°C for 20 sec, 55°C for 30 sec, 65°C for 2 min, and a final extension step at 65°C for 5 min. PCR products were

purified using AMPureXP beads (Beckman Coulter, Brea, CA, USA) on a DynaMag™ magnetic stand (Invitrogen™, Carlsbad, CA, USA), with two wash steps using 70% Ethanol in nuclease-free water and eluted using 10mM Tris-HCl pH 8.0 (Quality Biological™, Gaithersburg, MD, USA) with 50mM NaCl (Millipore Corporation, Burlington, MA, USA). DNA samples were quantified using Qubit 3.0 fluorometer and Qubit dsDNA BR assay kit (Invitrogen™ by ThermoFisher Scientific, Waltham, MA, USA). Samples were pooled to a total of 75fmol and brought to 10µl volume with 10mM Tris-HCl pH 8.0 with 50 mM NaCl. 1µl of Rapid Adaptor (RAP) was added to the barcoded DNA and the library was placed on ice until ready to sequence.

16S sequencing

A FLO-MIN106 flow cell with R9.4 sequencing chemistry (Oxford Nanopore Technologies, Oxford, UK) was opened, placed in the portable MinION™ (Oxford Nanopore) and checked for available number of pores via the MinKNOW software (ONT Version 19.06.7) on a Dell intel Core i7 laptop with 15.5GB of RAM and a 1TB solid state hard drive. The library, consisting of Sequencing Buffer (SQB), Loading Beads (LB), Nuclease-free water and DNA library was added to a 1.5ml DNA LoBind tube for a total volume of 75µl. The flow cell was primed with 800 µl priming mix, consisting of 30µl Flush Tether (FLT) and one tube of Flush Buffer (FLB) (EXP-FLP001) followed by an additional 200µl priming mix. The prepared library was mixed and added dropwise onto the SpotON sample port. The sequencing run was started with the experiment titled "16S_Barcoding" and run for 22 hours 51 min and 41s. The initial bias voltage was set to -180mV on the MinKNOW 'run options' tab, the default voltage recommended by Oxford Nanopore for the start of sequencing. After sequencing, the flow cell was washed

using Flow Cell Wash Kit (EXP-WSH002) to prevent carryover into further experiments on the same flow cell.

ColorID analysis

Following sequencing and basecalling of the 8 samples via the MinION™, fastq files were demultiplexed using Guppy v.3.4.1 (Oxford Nanopore Technologies, 2019). Taxonomic identification of raw sequence data was done using ColorID, specifically using the subcommand read_id (den-Bakker, 2018). Taxonomic identification of raw sequence data from the same eight previously sequenced MiSeq samples was also completed using ColorID for the purposes of comparing the relative abundances received between both MiSeq and MinION™ sequencers.

The speed and computational efficiency with which ColorID provides taxonomic assignment was determined via the /usr/bin/time -v command. The speed/computational efficiency was compared with an index consisting of 15,192 bacterial and some archaeal full 16S rRNA gene sequences (Refseq+RDP database; Alishum, 2019), as well one consisting of only relevant foodborne pathogens. This database consisted of 107 isolates from the following genera and species: *Listeria spp., Salmonella spp., Escherichia coli, Klebsiella pneumoniae, Citrobacter spp., Yersinia spp., Campylobacter spp., Serratia spp., Staphylococcus spp.,* and *Clostridium spp*. The time was recorded for the speed and the RAM usage in MB was recorded in each case and compared to that of QIIME2.

QIIME2 analysis

Bioinformatic analysis of the previously sequenced raw 16S MiSeq reads was performed with QIIME 2 2019.7 (Bolyen et al. 2019). The DADA2 plugin was used to infer Amplicon Sequence Variants (ASVs), which dereplicate sequences and filter out

additional sequencing errors (Callahan et al. 2016) (via q2-dada2). Taxonomic classification of sequences was completed using two classifiers within QIIME2, sklearn and Vsearch. These two classifiers were chosen due to drastically different time (**Table 4**) required to taxonomically classify sequences as determined using the /usr/bin/time -v command on these classifiers. Sklearn (Pedregosa et al., 2011) is a pre-trained Naïve Bayes Python tool that was used by QIIME2 as the default method in the 'Moving Pictures Tutorial' (*Moving Pictures Tutorial*, 2019) to taxonomically classify sequences. The classifier was trained on the Greengenes 13_8 97% database (McDonald et al., 2012), which was selected by downloading the Greengenes 16SrRNA database under the 'Data Resources' tab in QIIME2. Following the 'Moving Pictures Tutorial,' the 'taxa-bar-plots.qsv' file was viewed, which showed the relative abundances of all bacteria in a bar plot. 'Level 6' was selected, corresponding to genus level. A .csv file was downloaded and the genera with the highest relative abundances were selected for comparison with ColorID. The respective counts were divided by the total number of counts to calculate the relative abundance.

QIIME2 taxonomic classification speed was also measured using Vsearch v. 2020.2.0 (Rognes et al., 2016), which is a consensus-based alignment plugin used by QIIME2 available under the 'feature-classifier' plugin (Bokulich et al., 2018). Reference sequences were also clustered at 97% similarity; however, taxonomic labels were required to assess a consensus taxonomy. Therefore, the Greengenes 97% taxonomy was downloaded as well and used as input. The default parameters were used, except for the 'p-maxaccepts VALUE', which was set at 'all' to keep all hits greater than the default

similarity, which was set at a proportion of 0.8 (80% identity to query sequence). The time to classify sequences was also measured using the /usr/bin/time -v command.

Limit of Detection Experiment

Strain selection for limit of detection experiments

To determine the limit of detection of the MinION™ for relevant pathogens, two strains were selected that were not found in any of the samples during the analysis. These strains were *Listeria monocytogenes*-2011L-2626 and *Salmonella* Enteritidis CC52. 10ml of Brain Heart Infusion (BHI, Acumedia ®, Lansing, Michigan) for *Listeria* and Tryptic Soy Broth (TSB, Acumedia ®) for *Salmonella* were transferred via serological pipette into two 50ml Centrifuge Tubes (VWR ®, Radnor, PA). The tubes were then inoculated with the strains and placed in a 37°C incubator overnight.

Optical Density and growth experiment

An Optical Density (OD) versus time experiment was conducted to grow strains to mid-log phase. Following overnight incubation, *Salmonella* and *Listeria* cultures were taken out of the 37°C walk-in incubator and sub-cultured (100µl) into 10ml of TSB and BHI in two 50ml centrifuge tubes. Optical Density at 600nm was measured with a Bio-Rad SmarSpec™3000 spectrophotometer at time 0 for each strain. Subsequent measurements were recorded every 30 minutes until 90 minutes, where measurements were recorded every hour until cultures reached stationary phase. A graph of OD_{600} versus Time (hours) was drawn on logarithmic paper and a central point was selected manually as the optimal growth time for each strain.

Serial dilution (*Salmonella* and *Listeria* pure culture)

Following overnight incubation, *Salmonella* and *Listeria* cultures were sub-cultured (100µl) into 10ml of TSB and BHI in two 50ml centrifuge tubes. Cultures were grown to mid-logarithmic phase based on prior collected data and the OD_{600} was checked to ensure the cultures were in mid-logarithmic phase. *Salmonella* and *Listeria* cultures were serially diluted by 7 logs in duplicate and plated onto sterile 100 x 15mm petri dishes (VWR™, Radnor, PA, USA) with Tryptic Soy Agar (TSA, Neogen ®, Lansing, MI, USA) for *Salmonella* and BHI with Plate Count Agar (Difco™, Sparks, MD, USA) for *Listeria* using aseptic techniques. The plates were placed in the 37°C for overnight incubation. The resulting colonies were counted and averaged to determine the estimated concentration (CFU/ml) of the cultures.

DNA extraction (*Salmonella* and *Listeria* pure culture)

Immediately following serial dilution and plating, *Salmonella* (0.5ml, 8 log) and *Listeria* (0.5ml, 8 log) were subjected to DNA extraction, according to the gram negative and gram positive protocols, respectively, for the Qiagen DNeasy ® Blood & Tissue Kit (Qiagen, Hilden, Germany). The resulting extracted DNA was quantified using Qubit fluorometer 3.0 (Invitrogen).

Fecal sampling

Four fecal samples were obtained from different non-migratory Canada geese (*Branta canadensis*) located on campus at the University of Georgia in Griffin, GA. The samples were labelled and placed in a -20°C freezer until further processing. Due to a large amount of grass in one of the samples, three samples total were used.

Spiking experiment

Each of the 3 fecal samples were subjected to DNA extraction using a Zymo Quick-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research, Irvine, CA, USA). The standardized protocol, given by the instruction manual for the kit (Zymo Research; Catalog No. D6010) was adjusted as follows: different amounts of each sample were weighed out using an OHAUS Adventurer™ Pro Analytical Balance and the samples were homogenized (Vortex Genie 2™, model G-560, Fisher Scientific) on 'high' for 20 minutes. The amount of fecal material was adjusted to obtain a DNA yield of at least 1ng/µl for each sample, as determined via qubit fluorometer 3.0, which corresponds to the concentration needed for library preparation via the 16S barcoding kit.

For trial 1, 0.35g of one goose fecal sample (G-FS-1) was weighed out in 3 ZR BashingBead Lysis tubes (Zymo) for *Listeria* spikes, enough for one replicate consisting of 3 spikes (6,4,2 log CFU/ml). 0.30g of another goose fecal sample (G-FS-2) was weighed out into 6 additional lysis tubes for *Salmonella*, which was enough for two replicates of 3 spikes (6,4,2 log CFU/ml). Samples were placed in -20°C freezer until DNA extraction.

Following overnight incubation at 37°C and corresponding growth to mid-log phase, *Listeria* and *Salmonella* samples were serially dilution by 6 log and plated onto BHI plates with Plate Count Agar and TSA plates, respectively in duplicate using aseptic techniques. The plates were placed in the 37°C incubator for overnight incubation. The resulting colonies were counted and averaged to determine the estimated concentration (CFU/ml) of the cultures.

Immediately following serial dilution and plating, BashingBead Buffer was added to the prepared fecal samples in ZR BashingBead Lysis tubes, in accordance with the Zymo protocol. *Listeria* and *Salmonella* spikes were added from different dilutions, as shown in **Figure 1** to the different lysis tubes for intended dilutions of 6, 4, and 2 log CFU/ml (**Figure 2**). The spike in amounts were added based on the plate counts received from pure culture and corresponded to 9.62µl and 10.56 µl for *Listeria* and *Salmonella*, respectively. The plate counts described above were used to determine the actual concentrations of the three dilutions. Following spikes, the DNA was extracted using the Zymo Quick-DNA Fecal/Soil Microbe Miniprep Kit optimized protocol described previously. The eluted purified DNA was quantified using Qubit fluorometer 3.0 and was recorded.

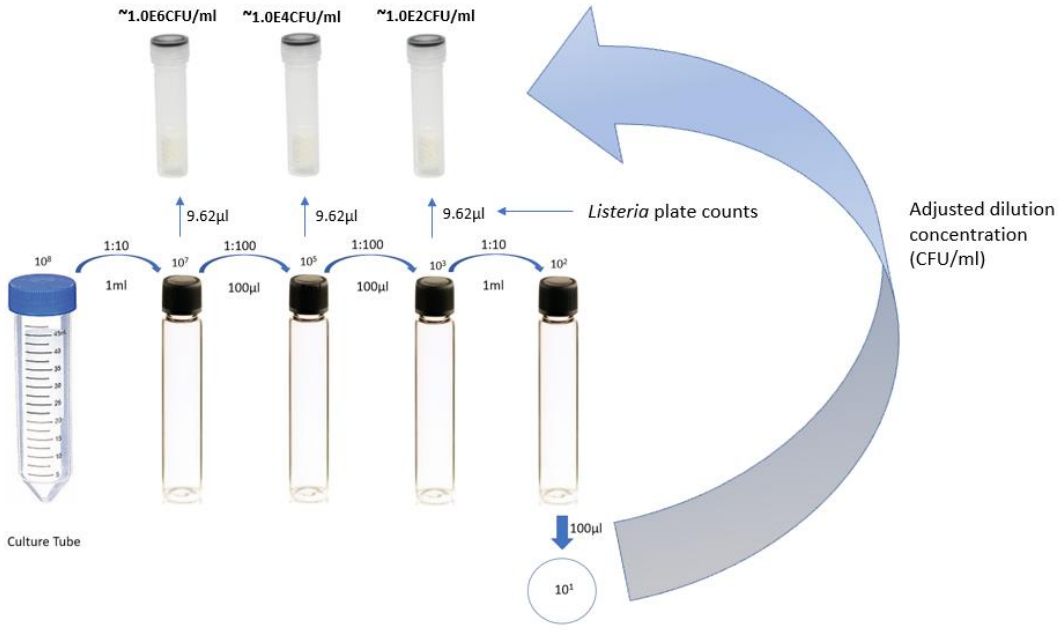

Figure 1.
6-log serial dilution scheme and three *Listeria* spikes (9.62µl; determined from previous plate counts) yielding intended dilutions of 6,4,2 log CFU/ml. Actual plate counts adjusted the intended dilutions**.**

$$\frac{cfu}{ml} of\ 1st\ dilution\ X\ Spike\ In\ amount\ (ml) \approx 1.0E6\ \frac{cfu}{ml}$$

$$\frac{cfu}{ml} of\ 2nd\ dilution\ X\ Spike\ In\ amount\ (ml) \approx 1.0E4\ \frac{cfu}{ml}$$

$$\frac{cfu}{ml} of\ 3rd\ dilution\ X\ Spike\ In\ amount\ (ml) \approx 1.0E2\ \frac{cfu}{ml}$$

Figure 2.
Equations used to determine spike-in amounts from the first three pure culture dilutions to give intended dilutions of 6, 4 and 2 log CFU/ml.

For trial 2, the serial dilution/plating and spike-in process was repeated with two replicates for *Salmonella* and two replicates for *Listeria* (six individual samples each) for a total of twelve samples. 0.17g of G-FS-2 was used for each of the twelve samples. Plate counts and DNA concentrations were recorded.

Native barcoding using 16S primers

The Native Barcoding Kit was used along with 16S primers to increase throughput of DNA from the 16S rRNA gene. As performed during library preparation using the 16S Barcoding Kit, the DNA from the spikes was diluted to 10ng in 10µl of Nuclease Free water. Individual samples were barcoded and prepped for PCR according to the 16S barcoding protocol, but with 2µl of forward primer (27F) and 2µl of reverse primer (1492R) (Frank et al., 2008) in a thin walled 0.2ml strip tube. Amplified DNA was cleaned up and eluted.

One µg of eluted DNA was diluted to 49µl with Nuclease-free water and DNA repair and end-prep steps were performed with NEBNext FFPE DNA Repair Mix (2µl), NEBNext FFPE DNA Repair Buffer (3.5µl) and ultra II End Prep Buffer (3.5µl)(New England BioLabs® Inc.) being added to each diluted sample. The reaction was incubated

in the thermocycler at 20°C for 30 min with the heated lid off. Three microliters of ultra II End Prep-Enzyme Mix were then added. The samples were then incubated for 5 min at 20°C and 5 min at 65°C in the thermocycler to complete the DNA repair and end-prep step. DNA barcoding and ligation and pooling of libraries as well as adaptor ligation and clean-up was done according to the Native Barcoding genomic DNA Kit (SQK-LSK109; Oxford Nanopore Technologies, Oxford, UK), with unique barcodes provided in the Native Barcoding Expansion Kit (EXP-NBD104) and Blunt/TA Ligase MasterMix (New England BioLabs). The Adaptor ligation and clean-up steps were performed using the Short Fragment Buffer (SFB) to retain DNA fragments around 1,500 base pairs, corresponding to the length of the 16S DNA fragment. DNA was eluted in elution buffer (EB) and quantified.

Native barcode/16S sequencing

To sequence the first trial of *Listeria/Salmonella* spiked DNA, a flow cell (ID: FAK51328) was primed and the prepared library was sequenced. The sequencing run continued for ~3.5 hours until around 1 million reads were generated. The initial bias voltage was set to -180mV. After sequencing, the flow cell was washed and stored at 4°C. The second trial of *Listeria/Salmonella* spiked DNA was sequenced with flow cell (ID: FAK53605). The experiment ran for ~4 hr with an initial bias voltage set to -180mV. This flow cell was also washed and stored at 4°C.

ColorID analysis

Taxonomic identification of raw sequence data sequenced from the MinION™ was done using ColorID, specifically using the subcommand read_id. Sequences from accepted reads were blasted against the NCBI database (Altschul et al., 1990) for

different minimum phred quality scores (Johnson et al., 2008). Based on the accuracy of the taxonomic assignment generated via BLASTn, a relative percentage of accepted hits of k-mers supporting the classification divided by the number of k-mers used as input for the classification was determined to minimize the number of false positives using Spyder v. 4.0.1 (Python Software Foundation). Based on this analysis, a relative abundance of *Salmonella* and *Listeria* was determined for each trial/dilution. The overall limit of detection for *Salmonella* and *Listeria* was determined from these results.

CHAPTER 4

RESULTS

Relative Frequencies

The raw data from the eight previously processed and sequenced MiSeq samples were analyzed by ColorID and QIIME2. Three out of eight samples are shown, one each from Chicken, Cow and Goose samples, since samples from the same animal produced microbiomes with similar composition. The relative frequencies of the most abundant genera (genera with a relative abundance of more than 2% within the three selected samples) were compared between the ColorID and QIIME2 results. The results of this comparison is shown in **Figure 3**. The total number of reads attributed to each of the genera were divided by the total number of reads classified for each sample to give the relative frequency for that genus.

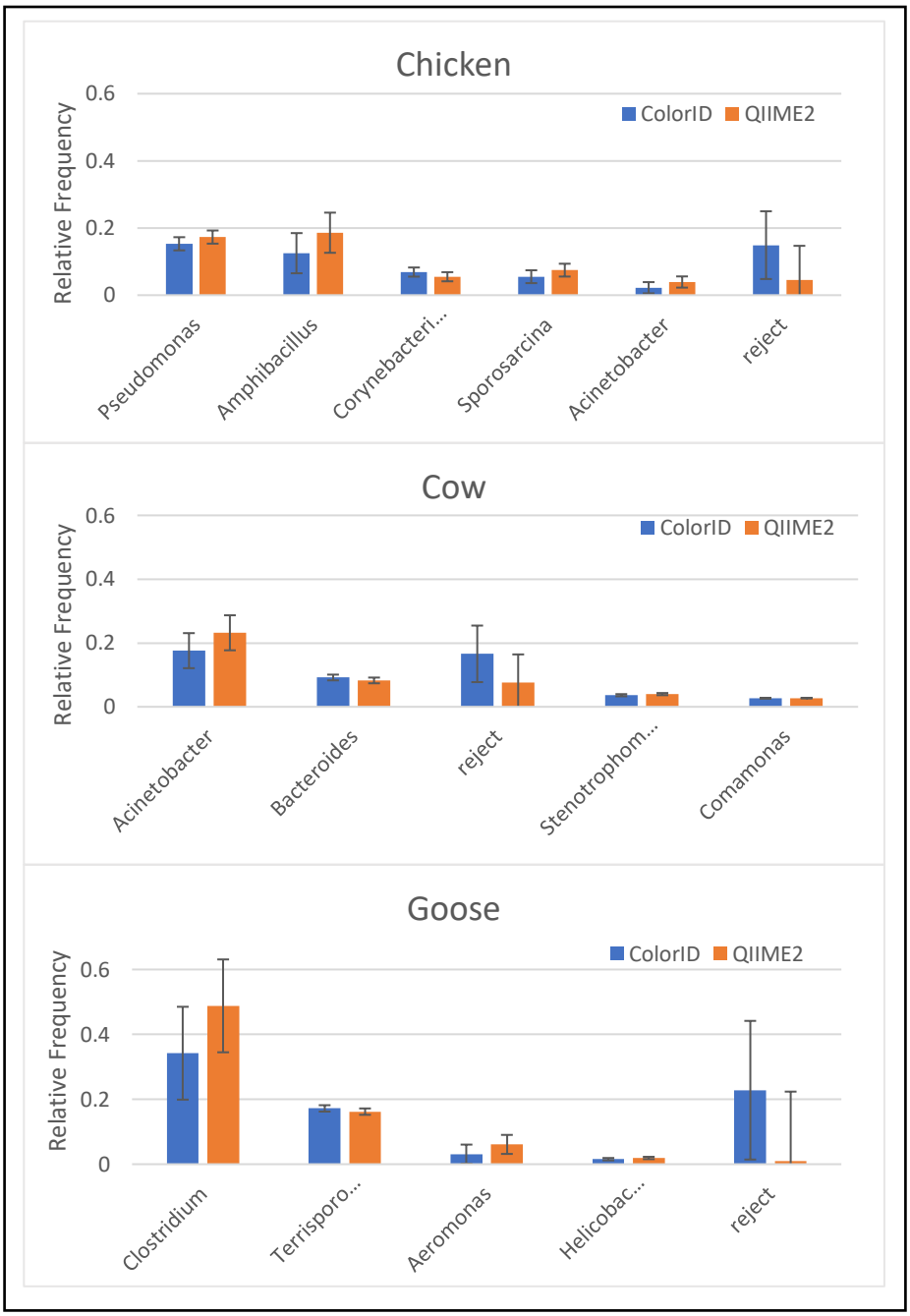

Figure 3.
Bar graph showing comparison of relative frequencies of the most abundant genera/rejected reads (taxa which made up at least 2 % of the total reads as determined by read counts) from Chicken, Cow and Goose samples (one each shown) analyzed by ColorID vs. QIIME2. Error bars indicate the 95% confidence interval (n=2) among replicates

Regression analyses were performed via JMP Pro 15 for each sample as shown in **Figure 4**. The line indicates the expected relative abundance of each genera within the QIIME2 pipeline given the observed relative abundance of genera in the ColorID results. The points lying outside the 95% confidence interval (shown in pink) correspond to the relative abundance of reads that were rejected via ColorID and QIIME2 analyses. Reads were rejected if no taxonomic classification was given.

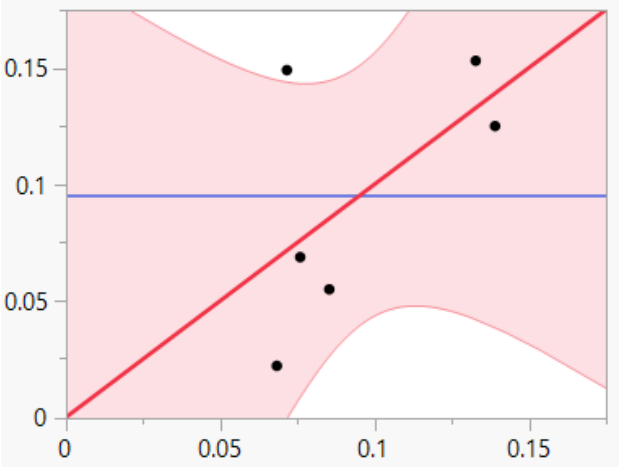

Figure 4.
Scatterplot graph showing the 95% confidence interval (pink area) for the regression fitting of genera from sample 'Chicken' that were analyzed by ColorID (y-axis) and QIIME2 (x-axis), R-squared = 0.34, p = 0.22.

These analyses, represented in **Table 1**, showed that, when the rejected reads were included in the analysis, there was no correlation between genera observed by ColorID and QIIME2-based analysis (P = 0.222, 0.127, 0.141 for chicken, cow, and goose, respectively). However, when the rejected reads were removed from analysis, there was a correlation between genera observed by ColorID and QIIME2-based analysis (p = 0.019, 0.016, 0.022 for chicken, cow, and goose feces, respectively). Since there was a correlation in relative abundance among ColorID and QIIME2 analyzed data despite the number of rejected reads from ColorID and QIIME2 analysis being

significantly different, it would be appropriate to compare the bioinformatic pipelines in terms of computational efficiency to determine whether ColorID is more efficient than QIIME2.

Table 1.
Results (P-value and R^2) from regression analyses of relative abundances of genera from 'Chicken, Cow and Goose' samples analyzed by ColorID versus QIIME2 with and without 'rejected' reads

Sample	Rejected Reads?	R^2	P-value
Chicken	Y	0.34	0.2216
Chicken	N	0.88	0.019
Cow	Y	0.59	0.127
Cow	N	0.97	0.0161
Goose	Y	0.57	0.1413
Goose	N	0.96	0.0216

Y – Yes, rejected reads were included in analyses; N – No, rejected reads were not included in analyses

Similarly, the relative frequencies of the most abundant genera of the same three samples that were sequenced by MiSeq and MinION™ and analyzed by ColorID were compared as shown in **Figure 5**. Detailed results of one of the regression analyses is shown in **Figure 6**.

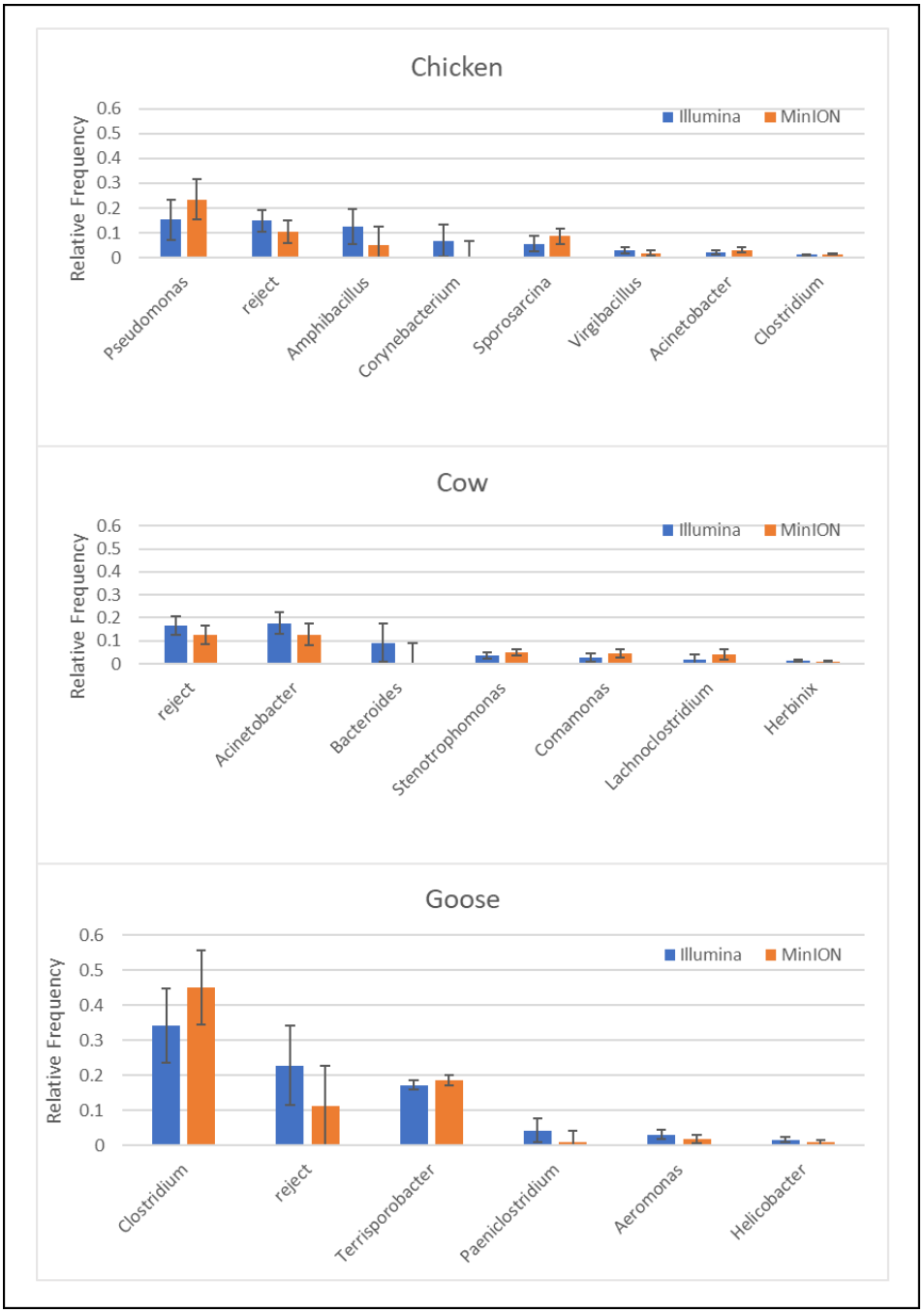

Figure 5. Bar graph showing comparison of relative frequencies of the most abundant categories in the read classification (genera or the category 'rejected reads' with a relative frequency of more than 1% as determined by read counts) from Chicken, Cow and Goose samples (one each shown) analyzed by MiSeq vs. MinION™. Error bars indicate the 95% confidence interval (n=2) among replicates.

Regression analyses showed that, for most samples, there was a correlation between genera (determined by relative abundance) observed by MiSeq and MinION™

based analysis at $p \leq 0.05$ with or without including the reads that were rejected in the analysis. Since there was a correlation between the sequencers based on relative abundance, it would be appropriate to compare the sequencers in terms of sequencing and computational efficiency to determine whether the MinION™ is more efficient than the MiSeq.

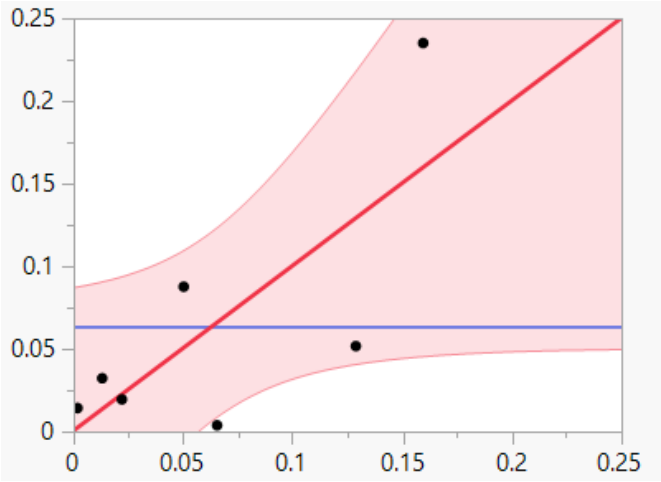

Figure 6.
Scatterplot graph showing the 95% confidence interval (pink area) for the regression fitting of genera from sample 'Chicken' that were sequenced by MiSeq (x-axis) and MinION™ (y-axis), R-squared = 0.55, p = 0.05.

Table 2.
Results (P-value and R^2) from regression analyses of relative abundances of genera from 'Chicken, Cow and Goose' samples sequenced by MiSeq versus MinION™ with and without 'rejected' reads

Sample	Rejected Reads	R^2	P-value
Chicken	Y	0.53	0.0411
Chicken	N	0.55	0.0547
Cow	Y	0.66	0.0266
Cow	N	0.51	0.1087
Goose	Y	0.85	0.0091
Goose	N	0.992	0.0003

Y – Yes, rejected reads were included in analyses; N – No, rejected reads were not included in analyses

Sequencing Efficiency

Four 16S MinION™ sequencing runs were selected to demonstrate the sequencing efficiency of the Nanopore MinION™ as shown in **Table 3**. The average number of reads generated per second was 71.3. These results were compared to the general output of the Miseq, which is greater than 20 million reads generated within 65 hours according to the 16S metagenomics Sequencing Library preparation protocol of Illumina (Illumina, 2013). This yields 85.5 reads generated per second. When accounting for the shorter lengths of MiSeq reads (averaging 300 base pairs), this would equate to around 17.1 full length 16S reads generated per second, which is statistically different ($p<0.001$) from 71.3 as determined using a Z-test for the mean with $\alpha = 0.05$. This

demonstrates that the MinION™ was more efficient in terms of reads generated per second.

Table 3.

Summary of reads generated per second of run time from four 16S MinION™ sequencing runs

	Run time	Reads generated	Reads generated per second
16S_Barcoding	22h 51m	5.72M	69.5
Goose_Native_1	3h 29m	929.4K	74.1
Goose_Native_2	4h 11m	1.05M	69.7
Native_16S	3h 43m	958.8K	71.7

Computational Speed and Efficiency

QIIME2 versus ColorID

To determine whether ColorID was faster and more computationally efficient than the sklearn and Vsearch classifiers within QIIME2, the time to classification of the reads and RAM usage were determined. For QIIME2, sklearn, the default classifer used in the 'Moving-pictures tutorial,' classified reads in 54 seconds with a RAM usage of 1,265 MB. The time of classification using the plugin, Vsearch, available in the 'feature-classifier' plugin was 6 hours 12 minutes with a RAM usage of 1,850 MB. ColorID classified the sequences in 4 minutes 22 seconds on average with a RAM usage of 281 MB per sample. Using sklearn, the time for classification was significantly less as determined by a Z-test for mean with $\alpha = 0.05$ ($p<0.001$) than ColorID; however, the RAM usage was significantly less with ColorID than sklearn within QIIME2 (Z-test: $p<0.001$). Using Vsearch, both the time for classification and the RAM usage were

significantly greater than ColorID (Z-test: $p<0.001$). These findings are summarized in **Table 4**.

Table 4.
Comparison of time and memory (RAM) of running QIIME2 with two different classifiers and ColorID

	QIIME2		ColorID Average
	Vsearch	sklearn	
Time/Run (sec)	22,036	54	262.2
RAM (MB)	1,853	1,266	281.4
Threads	5	5	8

ColorID 'all bacteria' database versus 'pathogen-specific' database

To determine whether ColorID using an "pathogen-specific" database consisting of 107 isolates was faster and more computationally efficient than ColorID using an "all-bacteria" database consisting of 15192 isolates, the average time, reads per second and RAM usage were compared between sequencing platforms (MiSeq and MinION™) as shown in **Table 5**. The average time, RAM usage and number of classified reads per second from ColorID analysis of MinION™ sequenced data using an 'all-bacteria' database were all significantly different as determined by a Z-test for mean, with $\alpha = 0.05$ ($p<0.001$) from a 'pathogen-specific' database. This indicates that ColorID using a 'pathogen-specific' database is much faster and more computationally efficient than an 'all-bacteria' database.

To determine whether the same trend was true for MiSeq sequenced data, a Z-test analysis was conducted as before. The average time, RAM usage and number of

classified reads per second from ColorID analysis of MiSeq sequenced data using an 'all-bacteria' database were all significantly different at $p<0.001$ from a 'pathogen-specific' database. This shows that the 'pathogen-specific' database is faster than the 'all-bacteria' database using both MinION™ and MiSeq data.

Although the average classified reads per second appears higher for MiSeq than MinION™ sequences, typically only the V3-V4 region of the 16S rDNA gene is sequenced, resulting in approximately 300 base pair fragments. Thus, compared to the full length (~1500bp) gene sequenced by the MinION™, the number of "full length" reads classified per second by the ColorID/MiSeq method would be approximately $1/5^{th}$ of the reported values shown in **Table 5**.

Table 5.
Comparison of Average time, memory and reads classified per second of running ColorID using an 'all-bacteria' database versus a 'pathogen-specific' database for data sequenced by MinION™ and Illumina

	MinION™			MiSeq		
	'all-bacteria' database average	'pathogen-specific' database average	P-Value*	'all-bacteria' database average	'pathogen-specific' database average	P-Value*
Time (sec)	43.8	1.3	<0.001	262.2	11.6	<0.001
RAM (MB)	288.1	69.2	<0.001	281.4	62.2	<0.001
Reads classified per second	557	17,319	<0.001	1,309	30,309	<0.001
Threads	8	8	N/A	8	8	N/A

*Test for mean between 'all-bacteria' and 'pathogen-specific' databases from both sequencing platforms determined using Z-test ($\alpha=0.05$)

Limit of Detection

Serial dilution (*Salmonella* and *Listeria* pure culture)

Following incubation at 37°C to mid-log phase, *Salmonella* and *Listeria* yielded an OD_{600} of 0.677 and 0.477, respectively. Following serial dilution/plating and DNA extraction, the average culture concentration for *Salmonella* was 8.85 log CFU/ml and 8.89 log CFU/ml for *Listeria*.

Spiking experiment

For trial 1, fecal samples from different geese were used (G-FS-1 for *Listeria* and G-FS-2 for *Salmonella*). Trial 1 incorporated *Salmonella* spikes into samples consisting of 0.3g G-FS-2 for intended dilutions of 6,4 & 2 log CFU/ml. Two replicates were performed for each dilution for a total of 6 samples. *Listeria* was spiked into samples consisting of 0.35g G-FS-1 for the same intended dilutions. One replicate was performed (3 samples) due to limited amounts of G-FS-1. Final OD_{600} of *Salmonella* culture was 0.683 and serial dilution/plating resulted in average culture concentration of 5.01E8 log CFU/ml for replicate 1 and 5.62E8 CFU/ml (8.75log) for replicate 2. Final OD_{600} for *Listeria* culture was 0.423 with an average culture concentration of 9.0E8 CFU/ml (6.14log). Final adjusted spike-in concentrations are shown in **Table 8.** *Listeria* spikes are not shown due to significant *Salmonella* cross-contamination.

Trial 2 incorporated *Salmonella* and *Listeria* spikes into samples consisting of 0.17g of G-FS-2. Final OD_{600} of *Salmonella* culture was 0.832, higher than desired optical density. However, average culture concentration was 5.25E8 CFU/ml (8.72 log) for replicate 1 and 4.17E8 CFU/ml (8.64 log) for replicate 2, which was similar to trial 1. Final OD_{600} of *Listeria* culture was 0.462 and serial dilution/plating resulted in average

culture concentrations of 8.91E8 CFU/ml (8.95 log) and 1.44E9 CFU/ml (9.16 log) for replicates 1 and 2, respectively. Spike-in concentrations for trial 2 are shown in **Table 9**.

Relative frequency of *Listeria* and *Salmonella*

Analysis of reads from *Salmonella/Listeria* spikes showed that the relative frequency of *Listeria/Salmonella* differed depending on the quality (phred) score used as a parameter for classification. To determine what was the optimum phred score to use and remove false positives, a quality filtering step was necessary. To obtain sets of samples that were known to be *Listeria monocytogenes* and *Salmonella* Enteritidis, we obtained MinION™ reads from the *Salmonella/Listeria* pure cultures extracted previously and analyzed the results with ColorID. Spyder v. 4.0.1 (Python) analysis of reads showed that the relative frequency (proportion of k-mers supporting classification / k-mers used as input; a/b in **Table 6**) producing correct classifications at $p \leq 0.05$ depended on the phred score used. Generally, the higher the phred score used as a parameter for ColorID, the higher the k-mer-hit proportion is that produced a correct classification at $p \leq 0.05$. **Table 7** shows that, for a quality score of 5, 95% of the time a k-mer hit proportion of 0.32 or higher will correctly identify *Listeria monocytogenes*. The same was true for *Salmonella* Enteritidis. Based on this information, all reads below this proportion were filtered out to reduce the number of false positives. Although for *Listeria/Salmonella*, a phred score of 8 produced the highest proportion of k-mer hits that correctly identify the microorganism, the sensitivity decreased at this score. Therefore, a k-mer ratio of 0.32 with a phred score of 5 was chosen for both *Listeria/Salmonella* for all subsequent analyses.

Table 6.
Example ColorID output for *Listeria monocytogenes* illustrating how to determine relative frequency

Taxonomic classification	K-mers supporting classification	K-mers used as input for classification	Accept/Reject	Number of Accept/Reject
Listeria monocytogenes_B	119_a	244_b	Accept	1
Listeria monocytogenes	377_a	582_b	Accept	1

Note. a/b indicates relative frequency

Table 7.
Proportion of k-mer hits that produce correct classification at $p \leq 0.05$ for different phred scores

Listeria monocytogenes_B			*Salmonella enterica_C*		
Phred score	P-Value	Proportion	Phred score	P-Value	Proportion
Q0	0.062192935	0.24	Q0	0.064105567	0.18
Q1	0.062192935	0.24	Q1	0.064105567	0.18
Q2	0.059922194	0.24	Q2	0.057799443	0.18
Q3	0.058330134	0.26	Q3	0.051827385	0.26
Q4	0.055704999	0.26	Q4	0.052587797	0.3
Q5	0.053687935	0.32	Q5	0.061122936	0.32
Q6	0.056538208	0.3	Q6	0.05813163	0.34
Q7	0.053900519	0.36	Q7	0.06065063	0.38
Q8	0.061392106	0.38	Q8	0.059080768	0.38

Note: Highlighted phred score shows proportion used. *Salmonella enterica*_C is synonymous with *Salmonella* Enteritidis

The relative frequency for each sample, determined by the filtered reads of *Listeria/Salmonella* divided by the total number of reads classified, is shown in **Tables 8 & 9**. **Table 8** for trial 1 shows that *Salmonella* Enteritidis could be detected at 3.72 log CFU/ml, but there is question as to whether it could be detected at 1.72 log CFU/ml, since only 1 read was detected in over 100,000 reads for 1.72 log CFU/ml in replicate 1, whereas no reads were detected for 1.77 log CFU/ml for replicate 2.

Table 8.
Trial 1-*Salmonella* spikes into 0.3g goose fecal samples and corresponding relative frequencies of positive reads/total reads detected by ColorID

	Spike-in concentrations (log CFU/ml)	Spike-in concentrations (CFU/ml)	Relative frequency*
Salmonella 1 (8.7 log CFU/ml)	5.72	5.25E+05	0.001858
	3.72	5.25E+03	0.00002057
	1.72	5.25E+01	0.00000985
Salmonella 2 (8.75 log CFU/ml)	5.77	5.89E+05	0.001318
	3.77	5.89E+03	0.00005156
	1.77	5.89E+01	0

* Relative frequency represents filtered reads / total reads

Table 9 for trial 2 shows that *Salmonella* Enteritidis could be detected as low as 1.7-1.8 log CFU/ml. *Listeria monocytogenes* could be detected as low as 4.14 log CFU/ml, but could not be detected below 3.94 log CFU/ml. Based on the results for *Salmonella* between Trials 1 and 2, it appears that the amount of fecal material effects the limit of detection, as it could be detected between 1.7-1.8 log CFU/ml in 0.17g goose feces, but could not between 1.7-1.8 log CFU/ml in 0.3g of the same sample of goose feces. However, more studies would be needed to verify this.

For each dilution shown in **Table 9**, *Salmonella* relative frequency is higher than *Listeria* at a similar dilution (+ 0.14 log CFU/ml). For example, the relative frequency at 5.8-6.0 log CFU/ml for *Salmonella* is ~0.02, whereas *Listeria* is ~0.0008. This shows that detection of *Listeria* is not as efficient as *Salmonella* using the same processes.

Table 9.
Trial 2-*Salmonella* Enteritidis and *Listeria monocytogenes* spikes of 0.17g goose fecal samples and corresponding relative frequencies of positive reads / total reads detected by ColorID

	Spike-in concentrations (log CFU/ml)	Spike-in concentrations (CFU/ml)	Relative frequency*
Salmonella 1 (8.72 log CFU/ml)	5.8	6.31E+05	0.01897
	3.8	6.31E+03	0.0001432
	1.8	6.31E+01	0.00005189
Salmonella 2 (8.72 log CFU/ml)	5.7	5.01E+05	0.02074
	3.7	5.01E+03	0.0002064
	1.7	5.01E+01	0.00008857
Listeria 1 (8.95 log CFU/ml)	5.94	8.71E+05	0.000853
	3.94	8.71E+03	0
	1.94	8.71E+01	0
Listeria 2 (9.16 log CFU/ml)	6.14	1.38E+06	0.0006589
	4.14	1.38E+04	0.0000292
	2.14	1.38E+02	0

* Relative frequency represents filtered reads / total reads

CHAPTER 5

DISCUSSION

Relative Frequency and Computational Resources

Relative frequencies of genera from samples analyzed by ColorID vs. QIIME2

Excluding rejected reads, **Figure 3** shows that the relative frequencies of genera among samples analyzed by ColorID and QIIME2 were similar, and this was confirmed by the regression analysis shown in **Figure 4** and **Table 1.** Although the number of rejected reads for ColorID was high compared to QIIME2, the majority of this difference can be attributed to inherent differences in programs. Instead of not providing a taxonomic assignment to a sequence, QIIME2 generally assigns it to its' least common ancestor (Bokulich et al., 2018). Also, since the 16S rRNA gene can have multiple copies in a bacterium and these copies can evolve independently of one another, it is likely that some copies varied enough to inhibit taxonomic classification (Grützke et al., 2019).

The observation that ColorID produces more rejected reads than QIIME2 without affecting the overall composition of the genera shows that ColorID is more efficient in terms of taxonomic classification accuracy. It produces similar relative frequencies of reads from genera as QIIME2 but does not spend the time classifying reads that may produce false positives.

Relative frequencies of genera from samples sequenced by Nanopore MinION™ vs. Miseq

Figure 5 shows a similarity in the ColorID-generated relative frequencies of genera among samples sequenced by the MinION™ versus MiSeq. This was confirmed by the regression analysis shown in **Figure 6** and **Table 2.** According to these analyses, the relative frequencies of reads from genera for most samples were not different ($P \leq 0.05$) between both sequencers, with or without rejected reads being included in the analysis. This makes sense due to ColorID being used to analyze reads generated from both sequencers.

Despite the similarities, there were some differences among select genera from certain samples such as *Sporosarcina* and *Amphibacillus* from 'Chicken' and *Bacteroides* from 'Cow' samples. Nevertheless, there appears to be no difference in the relative abundances of reads from genera sequenced by MinION™ and MiSeq. Since both methods produce similar results, the MinION™ could be compared to MiSeq in terms of sequencing efficiency. Similarly, since both QIIME2 and ColorID produce similar results, they could be compared to each other in terms of computational efficiency.

Computational resources

As shown in **Table 4**, ColorID analysis of Goose, Chicken and Cow fecal samples using an 'all-bacteria' database was faster and used less RAM than Vsearch within QIIME2 ($P<0.001$) Since Vsearch is a consensus-based classifier that requires consensus alignment between the 'query' sequence (reads in the form of ASVs) and the reference database as well as to corresponding taxonomy (Rognes et al., 2016), it was expected that

ColorID would be more efficient. Sklearn was faster than ColorID using an 'all-bacteria' database ($P<0.001$); however, it used much more RAM than ColorID ($P<0.001$).

Analysis of the same samples sequenced by the MinION™ using an "all-bacteria" database consisting of 15192 isolates yielded an average classification time of 43.8 seconds with an average RAM usage of 288.1 Mb. Although the amount of RAM required for ColorID analysis increases with an increasing number of accessions in the database, ColorID can use custom user-defined databases. Therefore, for a particular environment such as poultry, it may be advantageous to customize the database to classify only pathogens associated with poultry. As shown in **Table 5**, analysis of the samples sequenced by the MinION™ using a "pathogen-specific" database consisting of 107 isolates yielded an average classification time of 1.3 seconds with an average RAM usage of 69.2Mb. The average number of reads classified per second was over 17k. This showed that ColorID with a customizable, smaller database was extremely efficient.

Spiking Experiments/Limit of Detection

Limit of detection for *Salmonella* Enteritidis and *Listeria monocytogenes*

Multiplex PCR, abbreviated MqPCR, is a method of detection is a detection method that was found to be able to detect up to 5-6 bacteria/pathogens simultaneously (Law et al. 2014, Zhao et al. 2013). The limit of detection for *Salmonella* Enteritidis and *Listeria monocytogenes* using this method was found in studies to be around 10^3 CFU/ml in artificially contaminated pork (Silva et al. 2011, Guan et al. 2013). Similar values were found for *Salmonella* and *Listeria* detected by qPCR without enrichment (Zhao et al. 2013). However, in a study by Hu et al. the detection limits for *Salmonella enterica* and

Listeria monocytogenes in stool ranged from 1.3×10^3-1.6×10^4 CFU/g using multiplex qPCR (Hu et al. 2014, Law et al. 2015).

In this study, the limit of detection for the Nanopore MinION™/ColorID sequencing method was 4.96×10^1 CFU/ml (1.7 log CFU/ml) for *Salmonella* Enteritidis spikes and 1.39×10^4 CFU/ml (4.14 log CFU/ml) for *Listeria monocytogenes* in spiked Goose stool samples (**Tables 8 and 9**). The limit of detection of *Salmonella* was much lower using the MinION™/ColorID method than that of multiplex qPCR or qPCR without enrichment (Hu et al. 2014, Law et al. 2015). *Listeria*, however, was slightly higher. This could possibly be due to a failure of the DNA extraction protocol to yield good quality DNA for *Listeria*, a gram-positive organism (Vidic et al. 2019). This is supported by a study showing that gram-positive organisms, containing thick cell walls, are more resistant to lysis for DNA extraction (Kai et al., 2018). This would result in less 16S rRNA gene amplification and possibly a biased relative abundance. Use of a lysis method that is compatible with gram positive organisms such as *Listeria* as well as optimized for the matrix used (feces) may be beneficial in providing successful universal DNA extraction (Vidic et al. 2019).

Typically, salmonellosis, or gastritis resulting from *Salmonella* Enteritidis infection, requires an infectious dose of around 10,000 cells to cause illness; however, in some cases, numbers as low as 100 cells can cause illness (Blackburn, 2009). This current study showed that the MinION™/ColorID method could detect less than 2 log CFU/ml, which, in combination with a reduction/elimination plan could prevent illnesses that otherwise would not have been detected by some other detection methods, such as MqPCR or qPCR without enrichment.

Trial 2 yielded a limit of detection between 1.7-1.8 log CFU/ml for *Salmonella* Enteritidis with spikes into 0.17g of goose fecal samples. The number of filtered reads was enough to exclude the possibility of these reads arising from misclassification (i.e., false positives) or from contamination of other samples. The inability to detect *Salmonella* in 0.3g fecal samples at the same concentration as those detected in 0.17g fecal samples suggests that it may be harder to detect pathogens in low abundance as the total microbial load increases. This observation would need more data/replications to corroborate; however, it may be important information to consider for future studies.

Although detection of pathogenic 16S rDNA yields comparable results to that of qPCR and MqPCR (Multiplex qPCR) (Hu et al. 2014, Law et al. 2015), it should not be used primarily for detection. The 16S rRNA gene can and often does have more than one copy per bacterium and this number can differ between species. Although there are methods available to normalize this bias when calculating relative abundance, the 16S rRNA genes within a bacterium can evolve independently of one another, resulting in enough genetic diversity to prevent the ability to distinguish among species. Therefore, genus level identification is reached, but this is often not sufficient as *Listeria monocytogenes* causes illness, whereas other *Listeria* species may not. Metagenomic detection methods, such as shotgun sequencing, may fare better for both species and genus level resolution, as it gets all genetic information within a sample (Grützke et al., 2019). However, an advantage of the MinION™/ColorID sequencing method is that it is highly adaptable for metagenomic shotgun approaches as well, since the specific reference databases can be adjusted to reflect gDNA instead of 16S DNA. This ability to customize the database according not only to the probable composition of the

microbiome studied but also to the detection approach shows the versatility of the MinION™/ColorID sequencing method.

There are several limitations in analyses of 16S rDNA that the ColorID/MinION™ sequencing method addresses. Since the 16S rRNA gene does not contain a single hypervariable region that discriminates between all species, approaches that single out 1-2 regions, such as the V3-V4 region commonly used by the Illumina Miseq sequencing method, are not as sufficient for genus/species level resolution as the MinION™ (Grutzke et al. 2019). The MinION™, since it can sequence reads of up to 10-30Gb in length (Oxford Nanopore Technologies, 2019), can sequence the whole 16S fragment, allowing for higher discrimination.

Another limitation of bioinformatic analyses with larger databases is that they are generally not representative of certain populations, as they contain easy to isolate, abundant organisms. Certain communities may be underrepresented and therefore have no taxonomically similar neighbors, possibly prohibiting accurate taxonomic labels (Nygaard et al. 2020; Shah et al. 2019). Again, the ability to create custom databases with ColorID is extremely important and addresses these problems as different databases can be created depending on the expected composition of the population of interest.

A limitation of classifiers, such as naïve Bayes and RDP classifiers that require high-precision to detect pathogens, is that, although they will minimize false-positives, the number of false negatives may be higher. This may result in a large number of unclassified sequences or classified sequences, but to a least common ancestor. In contrast, when the number of false negatives is minimized (high recall), which is important in environments with a large quantity of unidentified species, the precision

suffers slightly (Bokulich et al. 2018). ColorID uses a data structure called BIGSI that eliminates any false negatives, allowing the false positive rate to be controlled by two inherent user-defined parameters contained in the structure (Bradley et al. 2017; den Bakker 2018). Also, as shown in **Figure 4 and Table 1**, ColorID rejects reads, while producing similar relative frequency of reads from genera sequenced by MinION™ and MiSeq, showing that it excludes reads that may produce false positives, making it more efficient.

Although there may be limitations with the MinION™/ColorID sequencing method when it comes to detecting pathogens from Illumina MiSeq data and detecting pathogens to species level via 16S analysis, the main observation from this experiment is that this method has the potential to significantly reduce computational time and resources to detect foodborne pathogens. The reduction in resources in terms of memory usage can allow on-site analysis of food production environments via laptop or iPhone without the use of internet. The reduction in library preparation time, sequencing and analysis could also greatly decrease the time to produce actionable data which could be used for various applications within the food safety sector.

Future Research and Modifications

Research on molecular sequencing and bioinformatic analysis methods are complex and all steps in the process must be carefully controlled. Future modifications to this experiment may be necessary in certain areas, such as relative abundance and limit of detection to gain adequate conclusions.

Since QIIME2 classifies many reads to a least common ancestor, a more accurate comparison method at the order or family taxonomic level could be used to compare

relative abundances. This would allow greater confirmation that ColorID and QIIME2 produce similar results, strengthening the case for ColorID.

To determine whether the DNA extraction process did affect *Listeria* detection, a future study could be done in which one set of fecal samples are spiked by pure culture and the other by the "equivalent" amount of DNA extracted from pure culture. This may eliminate the bias of *Listeria* cells surviving the extraction process via the Zymo kit and inhibiting correct interpretation of the limit of detection. The process of determining what the equivalent concentrations of DNA to pure culture were for both *Salmonella* Enteritidis and *Listeria monocytogenes* as well as time/resources became limiting factors in starting/completion of this portion of the study.

In addition, different goose fecal samples with various levels of microbial diversity could be used to assess how microbial diversity affects detection. Different goose samples at various amounts per sample may also contribute to this future study. To conclude the modification of limit of detection, it would also be beneficial to have more spike-in dilutions to provide more accurate limit of detection measurements.

Despite the need for more robust relative abundance and limit of detection experiments, this study shows that the MinION™/ColorID sequencing method can greatly reduce computational time and resources to detect foodborne pathogens. Oxford nanopore has recently come out with the MinIT™, a companion device control accessory for the Nanopore MinION™ that comes preconfigured with software needed for sequencing and analysis. This device can be operated by smartphone or laptop in remote settings, without the need for a laboratory. This technology combined with the ColorID algorithm could greatly reduce computational resources even more. Many outbreaks

originate from farms or processing facilities, so stakeholders could use this product for routine environmental monitoring, traceback investigations, among other applications, without sending samples to a lab.

CHAPTER 6

CONCLUSION

16S raw Illumina data from chicken, goose and cow fecal samples revealed a correlation between genera analyzed by ColorID versus QIIME2, with the exception of the rejected reads. A correlation between genera was also observed when 16S raw data from Illumina Miseq versus Nanopore MinION™ from the same fecal samples were analyzed by ColorID, with the exception of certain genera. There are some limitations and potential issues when comparing relative abundance data between bioinformatic platforms, such as the portion of 16S variable region chosen affecting classification. Still, this data is extremely useful for comparing platforms.

The MinION™ sequencing efficiency demonstrated in this study was greater than literature reports of MiSeq efficiency. Moreover, the ColorID algorithm in combination with the MinION™ sequencer produced a computationally efficient method of taxonomic classification/detection. This was apparent when comparing the computational speed and memory between the MinION™/ColorID method with QIIME2 analysis of MiSeq data. In addition, computational efficiency increased ~32x with ColorID analysis using a database consisting of relevant pathogens compared with all bacteria.

The limit of detection using the MinION™/ColorID sequencing method was 1.7-1.8 log CFU/ml for *Salmonella* Enteritidis and ~4 log CFU/ml for *Listeria monocytogenes*. These results clearly indicate the effectiveness of this method in detecting *Salmonella* over *Listeria*. Procedures to standardize detection between distantly

related organisms may prove very useful to make the MinION™/ColorID method a more viable option of detection comparable to qPCR and multiplex qPCR (MqPCR). The ability to detect microorganisms among a complex microbiome with very little computational resources may allow this method to surpass others in terms of versatility, which could benefit a wide variety of stakeholders.

.

REFERENCES

Ainsworth, D., Sternberg, M. J. E., Raczy, C., & Butcher, S. A. (2017). K-SLAM: Accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic Acids Research, 45*(4), 1649-1656. doi:10.1093/nar/gkw1248

Alishum, A. (2019). DADA2 formatted 16S rRNA gene sequences for both bacteria & archaea (Version 1) [RefSeq+RDP]. Zenodo. http://doi.org/10.5281/zenodo.2541239

Allali, I., Arnold, J. W., Roach, J., Cadenas, M. B., Butz, N., Hassan, H. M., . . . Azcarate-Peril, M. A. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiology, 17*(1), 194. doi:10.1186/s12866-017-1101-8

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J.. (1990). Basic local alignment search tool. *J. Mol. Biol*. 215:403–410. [PubMed] [Google Scholar]

Benítez-Páez, A., Portune, K. J., & Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™™ portable nanopore sequencer. *GigaScience, 5*(1), 4. doi:10.1186/s13Al742-016-0111-z

Benítez-Páez, A., & Sanz, Y. (2017). Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™™ portable nanopore sequencer. *GigaScience, 6*(7), 1-12. doi:10.1101/117143

Bibby, K., Ma, X., & Stachler, E. (2017). Evaluation of Oxford Nanopore MinION™™ Sequencing for 16S rRNA Microbiome Characterization. doi: http://dx.doi.org/10.1101/099960

Blackburn, C. de W. (2009). *Foodborne pathogens: hazards, risk analysis and control*. Retrieved from https://www.sciencedirect.com/topics/food-science/salmonella

Bokulich, N. A., Kaehler, B., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., . . . Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. doi:10.1186/s40168-018-0470-z

Bolyen E, Rideout JR, Dillon MR, et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology 37: 852–857*. https://doi.org/10.1038/s41587-019-0209-9

Bradley, P., Den Bakker, H. C., Rocha, E. P. C., Mcvean, G., & Iqbal, Z. (2019). Ultra-fast search of all deposited bacterial and viral genomic data. *Nature Biotechnology, 37*(2), 152-159. doi:10.1038/s41587-018-0010-1

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics, 20*(4), 1125-1136. doi:10.1093/bib/bbx120

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Cao, Y., Fanning, S., Proos, S., Jordan, K., & Srikumar, S. (2017). A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. *Frontiers in Microbiology, 8*, 1829. doi:10.3389/fmicb.2017.01829

Costa, M. C., Bessegatto, J. A., Alfieri, A. A., Weese, J. S., João Filho, A., and Oba, A. (2017). Different antibiotic growth promoters induce specific changes in the cecal microbiota membership of broiler chicken. PLoS ONE 12:e0171642. doi: 10.1371/journal.pone.0171642

Den Bakker, Henk. (2018). ColorID: An experiment with writing code in Rust and the BIGSI data-structure. GitHub repository. https://github.com/hcdenbakker/colorid

Dumas, M. D., Polson, S. W., Ritter, D., Ravel, J., Gelb J. Jr., Morgan, R., et al. (2011). Impacts of poultry house environment on poultry litter bacterial community composition. PLoS ONE 6:e24785. doi: 10.1371/journal.pone.0024785

Frank, J. A., Reich, C. I., Sharma, S., Weisbaum, J. S., Wilson, B. A., & Olsen, G. J. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Applied and environmental microbiology*, *74*(8), 2461–2470. https://doi.org/10.1128/AEM.02272-07

Grützke, J., Malorny, B., Hammerl, J. A., Busch, A., Tausch, S. H., Tomaso, H., & Deneke, C. (2019). Fishing in the soup - pathogen detection in food safety using metabarcoding and metagenomic sequencing. *Frontiers in Microbiology, 10*, 1805. doi:10.3389/fmicb.2019.01805

Guan,Z.P.,Jiang,Y.,Gao,F.,Zhang,L.,Zhou,G.H.,and Guan,Z.J.(2013). Rapidand simultaneous analysis of five foodborne pathogenic bacteria using multi-plex PCR.Eur.Food Res.Technol.237, 627–637. doi: 10.1007/s00217-013-2039-1

Hu,Q.,Lyu,D.,Shi,X.,Jiang,Y.,Lin,Y.,Li,Y.,et al.(2014).A modified molecu-lar beacons-based multiplex real-time PCR assay for simultaneous detection ofeight foodborne pathogens in a single reaction and its application. FoodbornePathog.Dis.11, 207–214. doi:10.1089/fpd .2013.1607

Illumina. (2013). 16S metagenomic sequencing library preparation protocol: preparing 16S ribosomal RNA gene amplicons for the Illumina MiSeq system. Part no. 15044223 Rev B. Illumina, San Diego,

CA: https://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf. [Google Scholar]

Jagadeesan, B., Gerner-Smidt, P., Allard, M. W., Leuillet, S., Winkler, A., Xiao, Y., . . . Grant, K. (2019). The use of next generation sequencing for improving food safety: Translation into practice. *Food Microbiology, 79*(6), 96-115. doi:10.1016/j.fm.2018.11.005

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research, 36*(Web Server issue), W5–W9. https://doi.org/10.1093/nar/gkn201

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., . . . Wong, G. K. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology, 7*, 459. doi:10.3389/fmicb.2016.00459

Kai, S., Matsuo, Y., Nakagawa, S., Kryukov, K., Matsukawa, S., Tanaka, H., . . . Hirota, K. (2019). Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION™™ nanopore sequencer. *FEBS Open Bio, 9*(3), 548-557. doi:10.1002/2211-5463.12590

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research, 26*(12), 1721-1729. doi:10.1101/gr.210641.116

Laver, T. W. (2014). *Evaluating metagenomic quantifications from next-generation sequencing data* Available from Dissertations & Theses Europe Full Text: Literature & Language. Retrieved from https://search.proquest.com/docview/1788101775

Law, J. W., Ab Mutalib, N., Chan, K., & Lee, L. (2014). Rapid methods for the detection of foodborne bacterial pathogens: Principles, applications, advantages and limitations. *Frontiers in Microbiology, 5*, 770. doi:10.3389/fmicb.2014.00770

Louca,S.,Doebeli,M.,and Parfrey,L.W.(2018).Correcting for 16S rRNAgene copy numbers in microbiome surveys remains an unsolved problem.Microbiome 6:41. doi: 10.1186/s40168-018-0420-9

López-García, A., Pineda-Quiroga, C., Atxaerandio, R., Pérez, A., Hernández, I., García-Rodríguez, A., & González-Recio, O. (2018). Comparison of mothur and QIIME for the analysis of rumen microbiota composition based on 16S rRNA amplicon sequences. *Frontiers in Microbiology, 9*, 3010. doi:10.3389/fmicb.2018.03010

Ludvigsen, J., Svihus, B., and Rudi, K. (2016). Rearing room affects the non-dominant chicken cecum microbiota, while diet affects the dominant microbiota. Front. Vet. Sci. 3:16. doi: 10.3389/fvets.2016.00016

Maestri, Cosentino, Paterno, Freitag, Garces, Marcolungo, . . . Delledonne. (2019). A rapid and accurate MinION-based workflow for tracking species biodiversity in the field. *Genes, 10*(6), 468. doi:10.3390/genes10060468

Magi, A., Semeraro, R., Mingrino, A., Giusti, B., & D'Aurizio, R. (2018). Nanopore sequencing data analysis: State of the art, applications and challenges. *Briefings in Bioinformatics, 19*(6), 1256. doi:10.1093/bib/bbx062

Mattocks, J. (1971). Goose feeding and cellulose digestion. *Wildfowl, 22*(22), 107-113. Retrieved rom https://wildfowl.wwt.org.uk/index.php/wildfowl/article/view/427/427

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, *6*(3), 610–618. https://doi.org/10.1038/ismej.2011.139

"Moving Pictures" tutorial – QIIME 2 2018.8.0 documentation. [Cited 29 Mar 2020]. Available: https://docs.qiime2.org/2018.8/tutorials/moving-pictures/

Nearing,J.T.,Douglas,G.M.,Comeau,A.M.,and Langille,M.G.I.(2018).Denoising the Denoisers:an independent evaluation of microbiome sequenceerror-correction approaches.PeerJ 6:E5364. doi:10.7717/peerj .5364

Nygaard, A. B., Tunsjø, H. S., Meisal, R., & Charnock, C. (2020). A preliminary study on the potential of nanopore MinION™ and illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. *Scientific Reports, 10*(1), 3209. doi:10.1038/s41598-020-59771-0

Oakley, B. B., Vasconcelos, E. J. R., Diniz, Pedro P V P, Calloway, K. N., Richardson, E., Meinersmann, R. J., . . . Berrang, M. E. (2018). The cecal microbiome of commercial broiler chickens varies significantly by season. *Poultry Science, 97*(10), 3635-3644. doi:10.3382/ps/pey214

Overview of QIIME2 Plugin Workflows – QIIME 2 v.2020.2. Retrieved March 29, 2020, from https://docs.qiime2.org/2020.2/tutorials/overview/#let-s-get-oriented-flowcharts

Oxford Nanopore Technologies. (2019). MinION™. Retrieved from March 30, 2020, from https://nanoporetech.com/products/MinION™

Pan, D., and Yu, Z. (2014). Intestinal microbiome of poultry and its interaction with host and diet. Gut Microbes 5, 108–119. doi: 10.4161/gmic.26945

Pareek, C., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics, 52*(4), 413-435. doi:10.1007/s13353-011-0057-x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research,* Retrieved from https://hal.inria.fr/hal-00650905

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology, 19*(1), 90. doi:10.1186/s13059-018-1462-9,

Rhoads, D. D., Wolcott, R. D., Sun, Y., & Dowd, S. E. (2012). Comparison of culture and molecular identification of bacteria in chronic wounds. *International Journal of Molecular Sciences, 13*(3), 2535-2550. doi:10.3390/ijms13032535

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ,* 4:e2584. https://doi.org/10.7717/peerj.2584

Rossmanith, P., & Wagner, M. (2011). The challenge to quantify listeria monocytogenes– a model leading to new aspects in molecular biological food pathogen detection. *Journal of Applied Microbiology, 110*(3), 605-617. doi:10.1111/j.1365-2672.2010.04915.x

Rothrock, M. J., & Locatelli, A. (2019). Importance of farm environment to shape poultry-related microbiomes throughout the farm-to-fork continuum of pasture-raised broiler flocks. *Frontiers in Sustainable Food Systems, 3* doi:10.3389/fsufs.2019.00048

Shah, N., Meisel, J. S., & Pop, M. (2019). Embracing ambiguity in the taxonomic classification of microbiome sequencing data. *Frontiers in Genetics, 10*, 1022. doi:10.3389/fgene.2019.01022

Silva,D.S.P.,Canato,T.,Magnani,M.,Alves,J.,Hirooka,E.Y.,and de Oliveira,T.C.R.M.(2011).Multiplex PCR for the simultaneous detection of Salmonellaspp.and Salmonella Enteritidis in food.Int.J.Food Sci.Tech.46, 1502–1507. doi:10.1111/j .1365-2621.2011.02646.x

Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P. J., . . . Spang, R. (2016). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome, 4*(1), 28. doi:10.1186/s40168-016-0175-0

Stanley, D., Geier, M. S., Hughes, R. J., Denman, S. E., and Moore, R. J. (2013). Highly variable microbiota development in the chicken gastrointestinal tract. PLoS ONE 8:e84290. doi: 10.1371/journal.pone.0084290

The Kavli Foundation. (2020). About the microbiome. Retrieved from https://www.kavlifoundation.org/about-microbiome

Torok, V. A., Hughes, R. J., Ophel-Keller, K., Ali, M., and Macalpine, R. (2009). Influence of different litter materials on cecal microbiota colonization in broiler chickens. Poult. Sci. 88, 2474–2481. doi: 10.3382/ps.2008-00381

Valenzuela-González, F., Martínez-Porchas, M., Villalpando-Canchola, E., & Vargas-Albores, F. (2016). Studying long 16S rDNA sequences with ultrafast-metagenomic sequence classification using exact alignments (kraken). *Journal of Microbiological Methods, 122,* 38-42. doi:10.1016/j.mimet.2016.01.011

Vidic, J., Vizzini, P., Manzano, M., Kavanaugh, D., Ramarao, N., Zivkovic, M., . . . Gadjanski, I. (2019). Point-of-need DNA testing for detection of foodborne pathogenic bacteria. *Sensors (Basel, Switzerland), 19*(5), 1100. doi:10.3390/s19051100

Videnska, P., Faldynova, M., Juricova, H., Babak, V., Sisak, F., Havlickova, H., et al. (2013). Chicken faecal microbiota and disturbances induced by single or repeated therapy with tetracycline and streptomycin. BMC Vet. Res. 9:30. doi: 10.1186/1746-6148-9-30

Walugembe, M., Hsieh, J., Koszewski, N., Lamont, S., Persia, M., and Rothschild, M. (2015). Effects of dietary fiber on cecal short-chain fatty acid and cecal microbiota of broiler and laying-hen chicks. Poult. Sci. 94, 2351–2359. doi: 10.3382/ps/pev242

Wang, W., Liu, Y., Yang, Y., Wang, A., Sharshov, K., Li, Y., . . . Li, L. (2018). Comparative analyses of the gut microbiota among three different wild geese species in the genus anser. *Journal of Basic Microbiology*, *58*(6), 543-553. doi:10.1002/jobm.201800060

Wood, Derrick E, Jennifer Lu, and Ben Langmead. 2019. "Improved Metagenomic Analysis with Kraken 2." *Genome Biology* 20 (1): 257. doi:10.1186/s13059-019-1891-0.

Wu, Y., Yang, Y., Cao, L., Yin, H., Xu, M., Wang, Z., . . . Deng, Y. (2018). Habitat environments impacted the gut microbiome of long-distance migratory swan geese but central species conserved. *Scientific Reports*, *8*(1), 13314-11. doi:10.1038/s41598-018-31731-9

Zhao, X., Lin, C.-W., Wang, J., & Oh, D. H. (2014). Advances in Rapid Detection Methods for Foodborne Pathogens. *Journal of Microbiology and Biotechnology*, *24*(3), 297–312. https://doi-org.proxy-remote.galib.uga.edu/10.4014/jmb.1310.10013