EVALUATION AND OPTIMIZATION OF BIOINFORMATIC TOOLS FOR THE DETECTION OF HUMAN FOODBORNE PATHOGENS IN COMPLEX METAGENOMIC DATASETS

By

GRETTA MARIE SHARP

Bachelor of Science in Bioenvironmental Sciences
Texas A&M University
College Station, Texas
2012

Master of Science in Biology
The University of Texas at Tyler
Tyler, Texas
2016

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
In partial fulfillment of
The requirements for
The Degree of
DOCTOR OF PHILOSOPHY
July, 2019

EVALUATION AND OPTIMIZATION OF BIOINFORMATIC TOOLS FOR THE DETECTION OF HUMAN FOODBORNE PATHOGENS INCOMPLEX METAGENOMIC DATASETS

Dissertation Approved:

Li Maria Ma

Dissertation Adviser

Stephen Marek

Andres Espindola Camacho

Charles Chen

.

Name: GRETTA MARIE SHARP

Date of Degree: JULY 2019

Title of Study: EVALUATION AND OPTIMIZATION OF BIOINFORMATIC TOOLS FOR THE DETECTION OF HUMAN FOODBORNE PATHOGENS IN COMPLEX METAGENOMIC DATASETS

Major Field: PLANT PATHOLOGY

Abstract:

Foodborne human pathogens pose a significant risk to human health as each year one in six Americans becomes sick from one of over 31 known human foodborne pathogens. Due to the differences in their growth requirements, current detection assays can only detect one to a few of these pathogens per single assay. Metagenomics, an emerging field, allows for an entire community of organisms to be analyzed from DNA or RNA sequence data generated from a single sample, and therefore has the potential to detect any and all foodborne pathogens present in a single complex matrix. However, currently available bioinformatic pipelines for metagenomic sequence analysis require extensive time and high computer power inputs, often with unreliable results. The objectives of this study are 1) to evaluate community profiling bioinformatic pipelines, mapping pipelines and a novel pipeline created at Oklahoma State University, E-probe Diagnostic Nucleic-acid Analysis (EDNA), for the detection of *S. enterica* (as a model foodborne pathogen) in metagenomic data, 2) to optimize EDNA pipeline for sensitive detection of the *S. enterica* in metagenomic data, and 3) to simultaneously detect multiple foodborne pathogens from a single metagenomic sample. EDNA was able to detect *S. enterica* in metagenomic data in approximately five minutes compared to the other pipelines, which took between 2-500 hours. The optimized parameters for the EDNA pipeline were limited to using cleaned Illumina data with a read depth of one. The minimum BLAST E-value was set to 10^{-3} for curation. For detection the minimum percent identity was set to 95% and the minimum query coverage to 90% with an E-probe length of 80 nt. These new parameters significantly improved the sensitivity of the assay 100-fold, from 10^3 *S. enterica* cells detected by the original EDNA pipeline to just 10 cells. In the simultaneous detection of multiple foodborne pathogens, EDNA detected three additional pathogens *Listeria monocytogenes*, *Campylobacter jejuni* and Shiga toxin producing *Escherichia coli* at ten contamination levels in less than ten minutes and provided new detection insights into read abundance as it corresponds to pathogen cell numbers.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

CHAPTER III

CHAPTER IV

CHAPTER V

LIST OF FIGURES

Figure Page

CHAPTER III

CHAPTER IV

CHAPTER V

APPENDICES

CHAPTER I

INTRODUCTION

The contamination of food products by pathogenic bacteria by either accidental or nefarious means is a significant health concern worldwide. *Salmonella enterica* (Se), Shiga toxin producing *Escherichia coli* (STEC), *Listeria monocytogenes* (Lm) and *Campylobacter jejuni* (Cj) are associated with more hospitalizations and deaths than all other known bacterial pathogens (Scallan et al., 2011). Contamination of food products by these pathogens can occur at any point throughout food production chain including dispersal and preparation processes, which necessitates effective and efficient detection methods (Aruscavage et al., 2006; Abadias et al., 2006). Consumer consumption of minimally processed foods, especially fresh produce, has been documented in recent years (Barth et al., 2010). Unfortunately, this trend coincides with an increased incidence of foodborne illness associated with fresh produce (Sivapalasingam et al., 2004; Painter et al., 2013; CDC, 2013).

For surveillance of bacterial foodborne pathogens like *Salmonella,* the US Food and Drug Administration (FDA) uses standardized detection procedures outlined in the Bacteriological Analytical Manual (BAM). Fresh produce is considered a Category II

food since the product is subjected to a process that is lethal to *Salmonella* between the time of sampling and consumption. The standard procedures begin with the isolation of *Salmonella* from the sample matrix (Bacteriological Analytical Manual, 8th Edition, Revision A, 1998. Chapter 4.). Because the isolation procedure involves selecting for a specific pathogen, only one pathogen can be detected at a time and to detect any other pathogens in the sample, the entire process would need to be repeated. Using the biochemical protocol, the quickest turnaround time for *Salmonella* identification is 120 hrs. (5 days). Using the Real Time PCR method, the fastest turnaround time is 96 hrs. (4 days). None of these estimated times include serotyping, which would add between 2-3 days for shipping and processing strain typing laboratories. These time estimates do not reflect the time required to process a high volume of samples which would increase the overall time requirement.

The United States Department of Agriculture (USDA) is the other governmental regulator of food products, and their standardized procedures for *Salmonella* sampling and identification are included in the Microbiology Laboratory Guidebook (MLG). Unlike the FDA, the USDA only tests for *Salmonella* in meat products, dairy, and eggs. Despite this difference, the overall laboratory procedures between the two administrative bodies are incredibly similar, and the differences that exist are in product sampling strategies. Both the BAM and the MLG suggest using Real Time-PCR to decrease the time requirement.

Metagenomics emerged as an application of genomic analyses and was first used in the field of ecology, where it is necessary to sequence DNA from a whole community of organisms to gain insight about community structure and function. Before metagenomics, it was not possible to observe all of the members or potential gene interactions *in situ* in an environmental community, since many of the organisms in

environmental samples are not culturable or known. Metagenomic sequencing allows the direct genetic analysis of a complex environmental sample (Karlsson et al., 2013). Using this method for detection streamlines the identification process by removing the need for culturing (Nakamura, 2009; Nakamura, 2011). While most metagenomic studies have primarily focused on profiling microbial communities in a sample, metagenomics has the potential to detect all microbes, including pathogens, in a given sample (Stobbe et al., 2013; Yang 2011). A metagenomic approach has already been used to detect previously unknown pathogens in a variety of hosts, including mammals, insects, and plants using community profiling (Adams et al., 2009, Cox-Foster et al., 2007, Palacios et al., 2008, Roossinck et al., 2010). However, community profiling is time and computationally intensive and can lack the specificity needed to differentiate between closely related pathogenic and non-pathogenic organisms. When it is not necessary to know the composition of an entire community in metagenomic data and when the pathogen sequences or signatures are available, it is possible to target the pathogen sequences for detection, reducing the computational resource requirements of community profiling. This approach of utilizing these unique sequences is known as targeted detection from complex metagenomic samples. Because of the success of these methods and the ever-lowering price of next generation sequencing (NGS) technologies, detection of foodborne pathogens through metagenomic sequencing has now become a possibility (Nakamura, 2011). However, the pipelines necessary to analyze this type of metagenomic data have not been fully established.

The most common pipelines used to deal with sequence data are heuristic, like those found in the Basic Local Alignment Search Tool (NCBI, 2017). This tool from the National Center for Biotechnology Information (NCBI) uses short three-word k-mers of the query

sequences to identify similar sequences in the NCBI database. Even though this process is faster than searches requiring exact matches, the size of the database that must be searched compared to the query data can make this type of analysis cumbersome and is not ideal for metagenomic data. New pipelines like Kraken are being developed to assign taxa to metagenomic read data (Wood, 2014). However, because of the time required for creating a community profile, it is limited as a high throughput diagnostic technique because of its slow speed (Pop and Salzberg, 2008, Magi et al., 2010). Ideally, a diagnostic tool would be able to target unique regions of a pathogen, which would reduce the time necessary to reach a diagnostic decision.

E-probe Diagnostic Nucleic-acid Analysis (EDNA) is a tool developed at Oklahoma State University in conjunction with the USDA to bridge the gap between profiling-based methods and diagnostically realistic time requirements. This method builds on the Tool for Oligonucleotide Fingerprint Identification (TOFI) method of probe creation and simplifies it while making it compatible with metagenomic data by using the probes as search queries in BLAST. Similar to TOFI, this pipeline is entirely *in silico*, which reduces the cost. EDNA was initially utilized to detect plant pathogens. EDNA only requires genomes of the targets and can be used with incomplete genomes, although this reduces the specificity (Stobbe et al, 2012). This pipeline is also ideal for detection of human foodborne pathogens like *Salmonella enterica* because it presents a rapid detection that can be done with unassembled metagenomic sequence data.

The ability to combine metagenomic sequencing with a rapid bioinformatic detection tool presents an opportunity to improve the access and usability of both fields. This combination streamlines the detection process of complex metagenomic sequence data into a five-minute analysis of all possible pathogens in a single assay. Additionally,

the optimization of this tool for very low titer human foodborne pathogen detection confirms that this tool can be used in both the plant and human fields and could significantly improve upon the methods currently used by the FDA and USDA.

LITERATURE CITED

Abadias, M., Cañamas, T.P., Asensio, A., Angueram, M., & Viñas, I. (2006). Microbial quality of commercial 'Golden Delicious' apples throughout production and shelf-life in Lleida (Catalonia, Spain). International Journal of Food Microbiology. 108:404-409.

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., et al., 2009. Next generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol. Plant Pathol. 10, 537–545

Aruscavage, D., Lee, K., Miller, S., & Lejeune, J.T. (2006). Interactions affecting the proliferation and control of human pathogens on edible plants. Journal of Food Science. 71:R89R99.

Bacteriological Analytical Manual, 8th Edition, Revision A, 1998. Chapter 4

Barth, M., Hankinson, T., Zhuang, H., & Breidt, F. (2010). Microbiological Spoilage of Fruits and Vegetables. In W. H. Sperber & M. P. Doyle (Eds.), Compendium of the Microbiological Spoilage of Foods and Beverages (pp. 135-183): Springer New York.

Centers for Disease Control and Prevention. (2013). Multistate Outbreak of Shiga toxin producing Escherichia coli O157:H7 Infections Linked to Ready-to-Eat Salads (Final Update).

Cox-Foster, D.L., Conlan, S., Holmes, E.C., Palacios, G., Evans, J.D., et al., 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318, 283–287.

Karlsson, O. E., Hansen, T., Knutsson, R., Löfström, C., Granberg, F., & Berg, M. (2013). Metagenomic detection methods in biopreparedness outbreak scenarios. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, *11*(S1), S146-S157.

Nakamura, S., Yang, C. S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., ... & Mizutani, T. (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PloS one*, *4*(1), e4219.

Nakamura, S., Nakaya, T., & Iida, T. (2011). Metagenomic analysis of bacterial infections by means of high-throughput DNA sequencing. *Experimental Biology and Medicine*, *236*(8), 968-971..

Painter, J. A., Hoekstra, R. M., Ayers, T., Tauxe, R. V., Braden, C. R., Angulo, F. J., & Griffin, P. M. (2013). Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998-2008. Emerging Infectious Disease, 19(3): 407-415

Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., et al., 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. 358, 991–998.

Roossinck, M. J., Saha, P., Wiley, G. B., Quan, J., White, J. D., Lai, H., ChavarrÍA, F., Shen, G. & Roe, B. A. (2010). Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19, 81-88.

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., & Griffin, P. M. (2011). Foodborne illness acquired in the United States--major pathogens. Emerging Infectious Disease, 17(1): 7-15.

Stobbe, A. H., Daniels, J., Espindola, A. S., Verma, R., Melcher, U., Ochoa-Corona, F., ... & Schneider, W. (2013). E-probe Diagnostic Nucleic acid Analysis (EDNA): a theoretical approach for handling of next generation sequencing data for diagnostics. *Journal of microbiological methods*, *94*(3), 356-366.

Wood DE, Salzberg SL: Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 2014, 15:R46.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., ... & Wang, J. (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *Journal of clinical microbiology*, *49*(10), 3463-3469.

CHAPTER II

LITERATURE REVIEW

Global Food Infrastructure

The globalization of the food market has increased public awareness of food safety and food security (Yiannas et al., 2009). The constant expansion of the food supply chain has been enabled by improvements in technology, including food storage and handling on a global scale. However, because of the increase in supply chains and through-puts, the risk of pathogen contamination has also increased. Consumer's perceptions about food are also changing. More than ever, people are questioning where their food is coming from and how it has been grown. People are questioning conventional systems that have traditionally focused on high production and yield. These practices were epitomized by the Green Revolution that focused on food security and eliminating food shortages by cultivating high yielding varieties of food staples like wheat, rice, and corn. The Green Revolution combined these new varieties with new chemical fertilizers and irrigation as a "package of practices" for food stability worldwide. These practices greatly improved food security worldwide and are viewed as one of the most significant contributions to agriculture in recent history. However, as new

.

global issues take a seat in consumer consciousness, concerns about sustainability and environmental repercussions have prompted new smaller scale markets that are seeking to produce locally grown and “organic farm to table" products. These new additions mean that new markets and transport chains are evolving at all levels of the industry. The United Nations Food and Agriculture Organization (FAO) defines food security as a national responsibility that "exists when all people, at all times, have physical and economic access to sufficient, safe and nutritious food to meet their dietary needs and food preferences for an active and healthy life (FAO, 1996)." This definition is built on the four pillars of availability, access, utilization, and stability (FAO, 2006). This model promotes both security and safety. As far as food safety, most strategies have focused on reducing foodborne illness through combinations of good agricultural practices (GAPs), critical control point (HACCP) plans and incorporation of new technology for detection of pathogens. However, foodborne illness remains a critical issue worldwide (Kirk et al., 2007).

Foodborne Illness

The WHO defines a foodborne illness as a disease caused by an infectious or toxic agent that enters the body through food consumption (Kirk et al., 2007). Foodborne illness can be caused by parasites, bacteria, viruses, toxins, prions, and toxic chemicals (Solomon et al., 2006). Both biological and chemical agents can cause human foodborne illness; however, most are caused by human foodborne pathogens (CDC, 2013). Approximately 9.4 million illnesses, 55,961 hospitalizations, and 1,351 deaths occur each year in the United States, attributed to 31 major foodborne pathogens (Scallan et al., 2011). It is estimated that one in six Americans experience foodborne illness every year, with the global rate greatly exceeding that estimate (Havelaar et al., 2013). Identification of sources

of food contamination is vital in the implementation of effective control strategies (Kirk et al., 2007). These control strategies are divided into two groups. The first group of strategies focuses on preventing contamination, while the other focuses on decontamination of contaminated food sources (Thorns, 2000). Good agricultural practices (GAPs) have been effective in reducing the amount of contamination; however, due to the high throughput and processing of current agricultural products decontamination is still extremely critical (Goodburn and Wallace, 2013).

Identification of contamination and decontamination is particularly important for foods that are consumed raw or with minimal kill steps, like cooking. These foods include fresh fruits and vegetables, which, thanks to the global marketplace, are available all year round. Not surprisingly, this has led to an increase (30%) in the consumption of these foods over the past three decades (Barth et al., 2010). Fresh fruits and vegetables have also been increasingly linked to human foodborne illness outbreaks (Sivapalasingam et al., 2004; Painter et al., 2013). Outbreak investigations are complicated because many pathogens are present in the environment, and food is only one of many routes of infection. Additionally, these investigations suffer from extensive under-diagnosis and reporting (Painter et al., 2013; Scallan et al., 2011) Because of the risk to human health, outbreak investigations are the foundation of foodborne illness source attribution (Cole et al., 2014). A comprehensive review of US foodborne illness outbreaks (Painter et al., 2013) provides critical data about the foods and pathogens most commonly associated with foodborne illness. From 1998 to 2008 using 4,589 foodborne disease outbreaks attributed to known sources reported to the Centers for Disease Control and Prevention (CDC), illnesses were attributed to seventeen different food categories composed of both simple and complex foods, made up of plant and animal products. One of the most notable findings was that produce accounted for

almost half (46%) of the outbreaks (Painter et al., 2013). Among the six plant food categories, vegetables contributed to more illnesses (34%) than fruits and nuts twelve (12%), with leafy vegetables accounting for the most illnesses (22%). It was found that the yearly percentage of outbreaks associated with leafy green vegetables has increased from 6% to 11% (CDC, 2013). Some attribute this increase to improved pathogen detection methods and not new sources of contamination (Brooks et al., 2005; Johnson et al., 2006; Bettelheim, 2007; CDC, 2012, Brandl and Sundin, 2013). Pathogen contaminated produce is a top contributor to outbreak-associated illnesses, hospitalizations, and deaths in the United States (CDC, 2013; Cole et al., 2014).

Microbial Contamination of Fresh Produce

Human foodborne pathogen contamination can occur at any point along the chain of production, preparation, packaging, and distribution (Aruscavage et al., 2006). Regardless of the route of contamination, the transmission pathway is generally through the oral-fecal route, meaning that produce is contaminated with pathogens in the waste from humans or animals and then infects humans. The ability of enteric bacterial pathogens to survive on or within plants differs depending on the pathogens and the produce (Barak and Schroeder, 2012; Aruscavage et al., 2006). It has been observed that although produce surfaces are not ideal for enteric bacteria, they can both survive and increase in number (Barak and Schroeder, 2012). The point of contamination with fecal matter can occur through soil, water, fertilizer, animal activity, harvesting activity, processing, and human sanitation failures (Matthews, 2009; Gil et al., 2013). As part of HACCP plans, contamination events are classified at pre-harvest or post-harvest (Gil et al., 2013). Pre-harvest contamination is combated by GAPs to reduce the likelihood of pathogen introduction into the system (Gil et al., 2013). In the US, most outbreak investigations have

concluded that the majority of contamination occurs in post-harvest production. Attributed to practices like improper storage, processing, failure to decontaminate equipment, and cross contamination (Gil et al., 2013). The points of post-harvest sources of contamination are more easily controlled than pre-harvest sources due to the lack of environmental conditions (Gil et al., 2013).

Food Terrorism

In addition to accidental contamination of the food supply with human foodborne pathogens, intentional contamination, or the threat of intentional contamination is also a concern. Food terrorism is housed under the umbrella of biological terrorism, and biological security efforts focus on securing food cultivation, processing, and transportation. Biological terrorism has become a significant area of concern since the terrorist attacks on the American World Trade Center and the Amerithrax investigation of 2001 (DOJ, 2013). However, the threat of food terrorism is not a new idea and has been used as a military and political strategy for hundreds of years (CFSAN, 2003; Lepick et al., 1945). Unlike select agents, human foodborne pathogens are easier to obtain because they are often part of the natural microbial community associated with animal rearing and only become an issue when consumed by humans. In 1984, the pathogen *Salmonella enterica Typhimurium* was cultivated and used by the Rajneesh Cult to try to influence a local election in Oregon. This attack resulted in 751 illnesses and 45 hospitalizations in the area and remained one of the most significant known outbreaks of foodborne illness in the United States (CDC, 2013). In the United States, the Food and Drug Administration has issued an official rule for Mitigation Strategies to Protect Food Against Intentional Adulteration. This rule includes a Food Defense Vulnerability Assessment, where the focus is on potential avenues of intentional contamination in the handling, processing, and

transportation of food products. The need for new rapid and sensitive detection strategies is of great interest to these initiatives.

Major Sources of Contamination

The major pathogens associated with foodborne illness in the US include parasites, bacteria, and viruses. Bacterial pathogens contribute the most to the rate of human foodborne illness, which is due to a combination of their abundance, virulence, and environmental persistence. The major bacterial pathogens monitored by CDC surveillance systems include *Salmonella enterica, Shiga toxin-producing Escherichia coli (STEC), Listeria monocytogenes, Campylobacter jejuni, Yersinia pestis, Shigella* spp.*, Vibrio* spp.*,* and *Clostridium* spp. While all of these pathogens can be associated with meat and animal products, *Salmonella, E. coli (STEC),* and *Listeria* are commonly associated with plant-based products (Painter et al., 2013).

Salmonella enterica

Salmonella enterica is a Gram-negative, rod-shaped, non-spore forming, facultative anaerobic species of bacteria in the family Enterobacteriaceae. These environmentally persistent and ubiquitous bacteria have a nomenclature system composed of six subspecies and over 2500 different serovars (Erickson et al., 2012). The vast majority of human infections are caused by *Salmonella enterica* subspecies *enterica*. This group contains over 1500 different serovars capable of causing human illness (Pop et al., 2004). The serovars are categorized as typhoidal or non-typhoidal based on their different modes of pathogenesis in humans. Both types can occur on multiple food matrixes causing gastroenteritis and are the number one bacterial agent resulting in hospitalization (Scallan et al., 2011) (Painter et al., 2013; Cole et al., 2014). *Salmonella enterica* subsp. *enterica*

serovar Typhi is the typhoidal serovars that can cause the characteristic typhoid fever. The more common non-typhoidal infections are less severe but can result in severe complications (Eo'Donnell et al., 2014).

Shiga Toxin Producing *Escherichia coli (STEC)*

Escherichia coli (STEC) are also Gram-negative, rod-shaped, non-spore forming, facultatively anaerobic bacteria in the family Enterobacteriaceae. The well-known serotype *E. coli* O157:H7 is most commonly associated with foodborne illness, but additional virulent strains continue to be isolated and identified as the causal agents in multinational outbreaks (Luna-Gierke et al., 2014) (Betteheim, 2007; Luna-Gierke et al., 2014). It is unclear whether these new strains are a product of new isolation and detection capabilities or new emerging strains (Brooks et al., 2005; Johnson et al., 2006). STEC infections are of great concern due to the possible complication of hemolytic uremic syndrome (HUS), which affects the kidneys and is life threatening (Karmali, 1989). This pathogen is most often thought of as a contaminate in ground beef and meat products; however, it was also implicated in the human foodborne illness outbreaks in spinach (CDC 2016), fenugreek sprouts (CDC 2011), clover sprouts (CDC 2012) and precut salad (2013). This trend toward fresh produce is concerning, and research into the survival mechanisms on these products is ongoing (Leff and Fierer, 2013).

Listeria monocytogenes

Listeria monocytogenes is a species of Gram-positive, rod-shaped, non-spore forming, facultatively anaerobic bacteria in the family Listeriaceae. The human mortality of this pathogen is between 20-30% of cases in the US, making it the deadliest human foodborne pathogen (CDC 2012; Ramaswamy et al., 2007). Of the six species only, *L.*

monocytogenes has been identified as a causal agent of disease in humans. Of the thirteen serotypes, only three are associated with foodborne illness (1/2a, 1/2b, and 4b)(Ward et al., 2004; Giusti et al., 2010). *Listeria* is relatively rare, but because of its high virulence and severe complications like pneumonia, meningitis, septicemia and spontaneous abortion, it is treated as a pathogen of concern and monitored by the CDC (CDC, 2018) (Ramaswamy et al., 2007). This pathogen is most often associated with preserved products like cheese and deli meat; however, it has also been found on fresh produce (Bae et al., 2013; Kovacevic et al., 2013; Painter et al., 2014).

Campylobacter jejuni

Campylobacter jejuni is a common food contaminant estimated as the causal agent in 1.3 million cases of illness from food in the United States yearly (CDC, 2019). It is motile, Gram-negative, non-spore forming spiral shapes that thrive in microaerophilic environments. There are 34 recognized species of *Campylobacter* with *jejuni* and *coli* most often implicated in human disease. The two most cited subspecies of *Campylobacter jejuni* are *jejuni* and *doylei*. These bacteria are often associated with poultry contamination (Hirano et al., 1983).

Available Detection Technology

Culture-based methods are one of the oldest methods used to identify microorganisms (Priyanka, 2017). The method is widely available, requires limited expertise, and is cost effective (Priyanka, 2017). This method is limited by the slow turn-around rate due to the time required for the culture to grow in media, which for most foodborne pathogens can be from 18-24 hrs. (Priyanka, 2017). This time is not very conducive to the fast-paced food production and shipping industry in the USA that

continually seeks to reduce the time required to get food from the farm to the table. Another drawback is the potential lack of specificity of differential media, as well as the inability of some bacterial pathogens to be cultured at all (Fletcher et al., 2006).

Immunoassays are a popular method of identification that can significantly reduce the turn-around time of the identification (Kalapothakis, 2001; Priyanka, 2017). The most common immunoassay is called Enzyme-Linked Immunosorbent Assay (ELISA) (Kalapothakis, 2001). This method uses antibodies that bind to conjugates and produce a color change, indicating a positive result in less than 12 hrs. (Kalapothakis, 2001). A positive result is then normally confirmed using PCR (Kalapothakis, 2001; Priyanka, 2017). One of the drawbacks to this method is the potential of cross reactivity of the antibodies, which can cause difficulty differentiating between species in some assays (Kalapothakis, 2001; Priyanka, 2017).

Polymerase Chain Reaction (PCR) has long been considered the gold standard when it comes to diagnostics (Avaniss-Aghajani, 1994; Priyanka, 2017). This process is rapid and sensitive with detection limits as low as femtograms (10^{-15}g) (Priyanka, 2017). However, this method cannot distinguish between live and dead cells, which is very important in the food industry, since dead bacterial pathogens cannot cause disease (Avaniss-Aghajani, 1994; Priyanka, 2017). This method can also generate high false positive rates depending on the specificity of the primers used (Avaniss-Aghajani, 1994; Priyanka, 2017). In addition to specific limitations, the detection methods above are limited by their ability to only detect a single pathogen or group of pathogens at a time (Avaniss-Aghajani, 1994; Priyanka, 2017).

Pulse Field Gel Electrophoresis (PFGE) is the current standard used by the CDC to produce a DNA profile of bacterial isolates. The PFGE data is stored on the FoodNet system, which allows a new outbreak isolated to be quickly compared to the known outbreaks isolated. However, recently, there have been issues with different isolates generating the same PFGE profile (Jones et al., 2007). These issues are because not all of the isolated sequence is considered using this method, and similar sequences could generate the same profile.

Government Microbial Protocols

For surveillance of bacterial foodborne pathogens like *Salmonella,* the US Food and Drug Administration uses standardized detection procedures outlined in the Bacteriological Analytical Manual (BAM) (FDA, 2013). These procedures outline sampling practices that differ depending on the type of food product. Fresh produce is considered a Category II food because the product is subjected to a process that is lethal to *Salmonella* (also known as a kill step) between the time of sampling and consumption. The standard procedures begin with the isolation of *Salmonella* from the sample matrix (Bacteriological Analytical Manual, 8th Edition, Revision A, 1998. Chapter 4.). Because the isolation procedure involves selecting for a specific pathogen, only one pathogen can be selected for at a time and to detect any other pathogens in the sample; the entire process would need to be repeated. To increase the *Salmonella* titers, the sample is selectively enriched. Enrichment involves the use of two enrichment broths tetrathionate broth (TTB) and selenite cystine broth (SCB) incubated for 24 hours at 35°C. The enrichment is then plated on bismuth sulfite agar (BSA), xylose lysine desoxycholate agar (XLD-A) and Hektoen enteric agar (HEA). The plates are incubated for 24 hours at 35°C. After

enrichment, colony morphology is identified by eye by comparing the sample plates to controls. Suspected *Salmonella* colonies are selected, and the plates are re-incubated up to 48 hrs. The selected colonies are streaked onto triple sugar iron agar (TSI-A), and lysine iron agar (LIA) and the biochemical and serological reaction is examined by eye. To positively identify *Salmonella*, the mixed TSI cultures are streaked onto MacConkey agar, HE agar or XLD agar and incubated for 24 hours at 35°C. The pure cultures are then subjected to a urease test, which involves transferring the cultures into urea broth for 24 hours at 35°C and identifying the color change by eye. A serological polyvalent flagellar (H) test uses the urease negative growth TSI plates and observes agglutination after 24 hours of incubation at 35°C. To reduce the time required for biochemical testing, it is possible to use RT-PCR to identify presumptive positives. PCR would be completed after the isolation of pure cultures. Finally, the cultures can be submitted for serotyping, but they must be submitted as individual isolates from each somatic group. The samples are then sent to either the Arkansas Regional Laboratory or the Denver District Laboratory. These laboratories serotype all the samples, and this can bottleneck the serotyping turnaround time. Using the biochemical protocol, the quickest turnaround time for *Salmonella* identification would be 120 hrs. (5 days). Using the RT-PCR method, the fastest turnaround time would be 96 hrs., (4 days). Neither of these estimated times includes serotyping, which would add between 2-3 days for shipping and processing. These times also do not reflect the time required to process a high volume of samples, which is highly dependent on the total number of samples that would need to be processed.

The USDA is the other governmental regulator of food products, and their standardized procedures for *Salmonella* sampling and identification are included in the Microbiology Laboratory Guidebook (MLG) (USDA, 2007). Unlike the FDA, the USDA

only tests for *Salmonella* in meat products, dairy, and eggs. Despite this difference, the overall laboratory procedures between the two administrative bodies are extremely similar, and the differences that exist are in product sampling strategies. Both the BAM and the MLG suggest using RT-PCR to decrease the time requirement (Yoshitomi et al., 2015).

Sanger Sequencing

The ability to obtain and study the genetic sequence of an organism has made an enormous impact on the way scientific research is conducted and has given many insights into relationships between organisms (Bartels et al., 2014). The Sanger method of sequencing by nucleic acid chain reaction was pioneered by Fredrick Sanger in 1977 (Sanger et al., 1977; Metzker et al., 2010). This sequencing breakthrough provided the technological advancement that was needed to sequence the human genome (IHGSC, 2004; Lyon et al., 2013). The Sanger method was improved upon by using capillary electrophoresis to increase the speed of the sequences processed (Trainor, 1990). The most notable new developments in sequencing have Next Generation Sequencing (NGS), Massively Parallel Sequencing (MPS), or High Throughput Sequencing (HTS) which greatly increased the speed and allowed hundreds of thousands of reads to be produced in a single procedure. These new technologies advanced the field of science by increasing understanding of taxonomy, gene expression, and traditional genetics while enabling new fields of study like metagenomics (Mardis, 2008; 2013; Yandell et al., 2001).

Roche 454 pyrosequencing

In 2005, the Roche 454 pyrosequencer was developed. This technology allows for the creation of thousands of sequencing reads to be produced through emulsion polymerase chain reaction (emPCR) and pyrosequencing (Nakano et al., 2003; Elahi & Ronaghi, 2004).

It uses large-scale parallel pyrosequencing with a capability to sequence approximately 400-600 megabases (MB) of DNA in a ten-turn run (Gibbons et al., 2007). The library preparation is done by shredding the DNA into 300-800bp and blunting each end. Adaptors are then ligated to the fragment ends. The adapter containing the 5'-biotin tag is used for immobilizing the DNA library to the streptavidin-coated beads. Nick repair occurs and releases the non-biotinylated strand, which is used as the single-stranded template DNA (sstDNA), and emPCR amplification occurs, and the templates remain encapsulated in water-in-oil mixture beads. The sstDNA beads are added to the DNA Bead Incubation Mix and layered with Enzyme Beads on the PicoTiterPlate device, and the beads are placed into the well through centrifugations where the sequencing reaction occurs. Nucleotides are then washed over the plate and are added to the templates in parallel. In wells where the addition of a nucleotide occurs, the light reaction is quantified by a CCD camera. The signal strength is proportional to the number of single nucleotides incorporated. However, the lack of ability to detect more than eight consecutive single nucleotide stretches (homopolymer) is a drawback of this type of sequencing. Roche 454 sequencing was removed from the market in 2016 when it was found to be noncompetitive, but it can be used to compare the effect of using long versus short reads in metagenomic community studies. A direct comparison of the Illumina and Roche 454 sequencing was completed to identify how the two different platforms treated the data (Luo et al., 2012; Roossinck et al., 2010). The metagenomic sample tested was a complex freshwater planktonic community. The study summarized that despite differences in read length and sequencing protocols that both platforms overlapped in approximately 90% percent of the taxon assembled. It has been hypothesized that Roche 454 could be better for metagenomic community studies

since longer read lengths could provide a more complete picture of the community with less assembly (Xie et al., 2012).

Ion Torrent

The Ion Torrent improves upon the nucleotide addition method (Merriman et al., 2012). The sample preparation and amplification are similar to that of the Roche 454 platform, but instead of generating photons with each base addition, each microwell is a hypersensitive ion sensor, and as the base is added to the DNA strand, a hydrogen ion is released and detected. This method requires fewer reagents, thus reducing the cost of the method.

Illumina

Another NGS platform is Illumina (Rodrique et al., 2010). The nucleic acid preparation is similar to that of the Roche 454 platform, shredding of DNA, followed by adapter ligation. The Illumina method then uses massively parallel sequencing by leveraging clonal array formation and reversible terminator technology. Using the “bridge” technology, four fluorescently labeled nucleotides flow across the flow cell and when attaching to the nucleotide chain, release fluorescence that is base specific which is picked up by the device. Illumina is known for producing "short reads" that are from 50-150bps. Using very short reads without assembly may contribute to a high false positive rate in detection application since short reads are more likely to map to multiple areas in many genomes. "Long read" sequencing is less popular due to the higher cost.

Pacific Biosciences

Unlike the other methods, Pacific Biosciences does not require an amplification step. This method is often referred to as third generation sequencing technology (Eid et al, 2009). It uses single-molecule real-time (SMRT) sequencing on the original molecule (Eid et al., 2009). Small wells called zero-mode waveguides (ZMWs) house the DNA (Fichot et al., 2013). At the bottom of each well is a single polymerase enzyme that accepts fluorescently labeled nucleotides. The surface is washed with a mixture of uniquely fluorophore-labeled dNTPs, and as the bases are incorporated into the sequence, the fluorophore is detected at the bottom of the well.

Oxford Nanopore

This technology is also considered a third-generation sequencing method and offers direct DNA/RNA sequencing in real time and yields ultra-long reals up to 2 Mb. The small size of the Oxford Nanopore MinION makes portable sequencing a possibility. The MinION technology identifies bases by measuring changes in electrical conductivity as a single strand passes through the biological pore (Lu et al., 2016). Because of the long reads produced, there is less need for complex *de novo* assembly which necessary when assembling short reads like those from the Illumina platform (Lu et al., 2016).

Sequencing errors

Each sequencing platform utilizes and combines different technologies, and therefore, each has different strengths and are prone to different errors. When comparing NGS sequencing platforms for detection, it is essential to understand these errors. Miscalling bases and sequencing bias can lead to false negatives when a sample is, in fact, positive. Another issue is unequal amplification, also called preferential amplification. An example of this is GC bias, which means that GC rich nucleic acids are favored during

amplification and will be in greater abundance after sequencing compared to the original sequence.

Homo-oligomers

Homo-oligomers are long sequences (>8) of identical nucleotides. Sequencing methods that rely on the amplitude of a single, like the Roche 454 and Ion Torrent platforms, have a difficult time accurately preserving the number of homo-oligomers in a single run (Huse et al., 2007). To combat this problem, improvements have been made by coating the wells with metal to increase the amplitude possible in a single run (Huse et al., 2007; Voelkerding et al., 2009). Third generation sequencing technologies like Pacific Biosciences and Oxford Nanopore are able to deal with homo-oligomers much better than previous generations because they produce long reads and there is less to assemble and therefore less assembly bias (Lu et al., 2016).

Base Calling

Base miscalling is a common sequencing error that also occurs in nature. This error occurs when the wrong nucleotide is incorporated because of either the wrong nucleotide being incorporated into the synthetic strand or because of misinterpretation of the signal. Because in most platforms the signal of a single miscalled base is diluted by the overall clonal DNA cluster, it is not an issue, however, in SMRT sequencing, it is a problem and the PacBio error rate is 10-15% (Eid et al., 2009). The Oxford Nanopore technologies also have a higher error rate (5-15%) compared to second generation sequencing (Lu et al., 2016). Illumina is known for substitution base calling errors causing the sequence to fall out of phase. Machine learning filters the background noise to read the base more accurately (Mardis, 2013).

Sequencing Bias

Bias can also be introduced through ligation and amplification. In the Illumina platform, it has been observed that GC bias can occur in the adapter ligation steps, which can lead to low coverage of AT rich regions. Using an alternative ligase can mitigate this bias (Quail et al., 2008). For detection, GC bias could be used to favor targets that are GC rich to increase sensitivity.

Read Length

Read length can vary enormously from platform to platform. Illumina is known for producing the shortest average reads 50-150 bps, Ion Torrent produces read of 200 bps, Roche 454 produces 400bps. Third generation sequencing can produce the longest read length with PacBio averaging 10-15 kb and Oxford Nanopore averaging 900 kb (Eid et al., 2009; Lu et al., 2016).

Assembly

For most pipelines, assembly is a necessary step in sequencing analysis. This is especially true with short read data like Illumina. During assembly, sequencing reads are assembled into contiguous sequences (contigs) and scaffolds. Many assembly programs are available, and some are preferred for specific sequence data (Chaisson & Pevzner, 2008; Gnerre et al., 2011; Myers et al., 2000). Two main strategies exist for assembly referenced based assembly and de novo. Reference based assembly maps the reads to known genomes, while de novo bases assembly based on prediction algorithms. De novo assembly is computationally and time intensive (Pop et al., 2004). This time requirement reduces the speed of NGS pipelines, but it is necessary for most identification tools.

Newbler Assembler (de novo)

The Newbler assembler is available from Roche Life Sciences (Chaisson & Pevzner, 2008). It is specifically designed to work with Roche 454 reads and has a default of a sixteen seed minimum match before it extends to find the optimal match. Large contigs are identified, and the overlapping reads are compiled into a single contig. Assembly is useful for many bioinformatic tasks, but it takes time and could limit the quantification capacity of a diagnostic technique by removing read depth and obscuring copy number.

SOAP Assembler (de novo)

Short Oligonucleotide Analysis Package (SOAP) is a software package that can be used for assembly, alignment, and analysis of next generation sequence data. It is optimized for alignment of short reads and is favored by people working with Illumina datasets. It has been used to assemble large genomes like human and animal genomes. However, like Newbler, it requires extra time after assembly.

Mapping Assembly

Mapping to a sequence is another method of assembly. A mapping alignment can be done with either reads or contigs. When assembling reads to a genome, a genome must already be chosen for the alignment. This necessary foreknowledge is a limitation for metagenomic studies because it is a mixed sample and choosing genomes biases the results. Another limitation for use with metagenomic studies is the false assemblies that could occur because of shared genes. This means that even if a genome is not represented in a metagenomic sample, some reads could assemble to the genome because they are shared by many different organisms, which is an issue for pathogen detection (Iqbal et al., 2012).

Metagenomics

Metagenomics is the study of the genomic makeup of environmental samples and can be used to assess sample biodiversity (Breitbart et al., 2002; Daniel, 2005; Gill et al., 2006), gene expression (Frias-Lopez et al., 2008; Uchiyama et al., 2004), and gene interaction within an environment (Harrison, 1981; Jones et al., 2010). Metagenomic sequencing allows the direct genetic analysis of a complex environmental sample (Karlsson, 2013; Tucker et al., 2009). Using metagenomics streamlines the identification process by eliminating the need for culturing or isolation (Nakamura, 2009; Nakamura, 2011). These breakthroughs in the field of microbial ecology can also contribute to other microbial fields, such as microbial identification (Nakamura, 2009; Schloss et al., 2005). This method has been primarily used to profile whole microbial communities in environmental samples associated with soil, water, and humans/animals. The strength of this type of work revolves around the ability to "reconstruct" an entire community from a single sample. Utilizing metagenomics has played a crucial role in discovering uncultivable organisms and viruses in complex environmental samples (Nakamura, 2009). This has been key in uncovering viruses, as well as hard to culture pathogens. This method is not limited to presence or absence detection. By translating DNA reads into RNA or proteins, a more complete picture of community function and the genes involved can emerge. To get a quantitative view of community function, differential transcriptomics can be used to understand how inputs into soil, water or the human microbiome can influence the microbial community (Luo et al., 2017). This method also has almost limitless application for pathogen detection, since any genome present can be reconstructed from the sequence data. Using this method, it is possible to detect any pathogen present in the sample, but the

limitations arise from the limited information on the performance about the computational pipelines that can be used to process metagenomic data.

Community Profiling Methods

The two main strategies for sequence mapping are informative and non-informative. Informative searches involve identifying biologically informative genes, also called open reading frames (ORFs) or coding domains (CDs) in the sequence data (Das et al., 2018; Tyson et al., 2004). Using the ORFs is a popular strategy because it classifies sequences based on relevance and reduces redundant searches (Pookhao et al., 2015). This is extremely relevant in sequence data involving eukaryotic organisms where non-coding regions and regions containing identical strings of nucleotides are prevalent (Liu et al., 2013). Another benefit of this strategy is that it can reduce the amount of false positive mapping because the searches are limited to only well characterized gene regions. The main limitation of this strategy is that it relies on the identification and characterization of ORFs (Kolde et al., 2015). This is an issue with metagenomic research because many of the organisms in the mixed sample have not been well studied (Nagarajan et al., 2014). This means that many of the ORFs will not be able to be identified, and the ones that are may not be indefinable at an informative taxonomic level. This will likely improve as ORF databases increase.

Non-informative searches look for sequence similarity without regard to gene coding regions or open reading frames (Chattaway et al., 2017: Zhang et al., 2000). This method can often achieve a higher degree of taxonomic resolution because it is not dependent on the characterization of ORFs. However, it is more likely to result in a higher rate of false identification, depending on how the search algorithm identifies matches (Pop

et al., 2018). This type of search can often take longer than informative searches based on predicted ORFs because of the relative sizes of the databases. Regardless of which strategy is used, it is essential to understand how these different strategies compare; meaning that are different pipelines converging and resulting in similar taxonomic profiles at different levels of clarity or are different methods resulting in significantly different species abundance at all taxonomic levels?

The most commonly used bioinformatic pipelines for analyzing metagenomic data are heuristic like those found in the widely used Basic Local Alignment Search Tool (BLAST)(Altschul, 2009). This tool from NCBI uses short three-word k-mers of the query sequences to identify similar sequences in the NCBI database. Even though this process is much faster than Bayesian and strict alignments based on perfect matches, the relative size of the databases makes this type of analysis computationally cumbersome. The BLAST tool has an online platform that is used extensively for local sequence searches, but for large datasets, a high-performance computer is still needed, and it can take many days (Santamaria, 2012). Programs like the Diamond pipeline attempt to improve the speed of BLAST by formatting the NCBI protein database with a proprietary algorithm (Buchfink, 2015). Diamond was developed as a high throughput program for DNA protein coding sequences and protein sequence alignments, 20,000 times faster than traditional BLAST while retaining high sensitivity (Buchfink, 2015). Other programs, like Kraken2, assign taxonomic labels to DNA sequences using k-mer based binning. Kraken2 requires the use of the Bracken program for a re-estimation of read abundance (Wood, 2014). These pipelines can all result in a taxonomic profile, which can be used to estimate the approximate percentage of each taxon in the profile. The Kraken2 and Bracken programs require the construction of multiple scripts for running the analysis, as well as, extensive

computer resources and RAM. Additionally, alignment programs like Bowtie2 can be used as community profiling or by creating a custom database and used as a mapping assembly to genomes or sequences of interest.

Targeted Detection Methods

The Tool for Oligonucleotide Fingerprint Identification (TOFI) was created to generate a microarray *in silico* (Geyer et al., 2008; Stobbe, 2013; Stobbe, 2014; Satya et al., 2008). TOFI is an integrated, scalable, high-performance-computing tool that incorporates genome comparison and probe design software. It was designed as a high throughput method to simultaneously process multiple bacterial or viral genomes and identify fingerprints that are unique to each genome. It can also be used to find fingerprints that are common between genomes (Geyer et al., 2008). The TOFI pipeline includes three main steps. The first step is a comparison of pathogen sequence with those of near neighbors for unique fingerprinting, the second step is thermodynamic optimization, and the final step is a check for uniqueness with BLAST. The strength of this method is that it reduces that amount of data that needs to be queried by only searching for the fingerprinted regions. This method also suggests that by using the *in silico* fingerprinting method, hundreds of related genomes could be run in a single assay (Geyer et al., 2008). However, for detection, it is not necessary to do all of the work in gene expression that is proposed by this pipeline, and this pipeline is limited in its application with metagenomic data due to its reliance on thermodynamics, which is not a concern in metagenomics.

E-probe Diagnostic Nucleic-acid Analysis (EDNA) is a tool developed at Oklahoma State University in conjunction with the United State Department of Agriculture (USDA) to bridge the gap between profiling-based methods and diagnostically realistic

time requirements. This method builds on the TOFI method of probe creation and simplifies it while making it compatible with metagenomic data by using the probes as search queries in BLAST. EDNA is an *in silico* tool that allows for the creation of electronic probes (E-probes) based on a known pathogen sequence (Stobbe, 2013; Stobbe, 2014). The E-probes are created by selecting a target pathogen genome and comparing it to a closely related genome that acts as the inclusivity/exclusivity determinate. The E-probe length is then chosen, which is dependent on the type of target organism and the length of the genome; however, previous studies have found that E-probes lengths of 60-80nt seem to work well for most organisms (Stobbe et al., 2014; Stajich et al., 2002). This produces the raw E-probes that are then cleaned by aligning the raw E-probes on the NCBI database and removing off-target hits. The resulting E-probes can be stored and used to detect targets in any FASTA datasets. The E-probes can identify pathogens in sequence data, including large metagenomic data (Stobbe, 2014). While EDNA does not provide a taxon profile or a relative species abundance, it does have the potential to rapidly detect a pathogen in a metagenomic dataset by probe matches. It also has the benefit of being used for target detection in unassembled, non-quality checked sequence data (Stobbe et al., 2014). This method has been tested on viruses (RNA and DNA), bacteria, fungi, and oomycetes. Most of the targets used for detection have been plant pathogens; however, this technique has the potential to detect any target, including human pathogens from sequence data. This method provides an opportunity to detect human foodborne pathogens on non-host (fresh food substrates) which would be extremely beneficial to food safety.

Metagenomic Dataset Construction

Metagenomic mock datasets are simulations of real environmental data (Richter et al., 2008). These datasets are key in uncovering the limitations of currently available

metagenomic data analysis tools because they offer a way to test the output results against the inputs of an experiment (Richter, 2008). This has been a major problem in the evaluation of tools for metagenomic analysis because due to the nature of environmental samples, and the inputs are variable and exact quantities are unknown (Korem et al., 2015). Mock datasets allow for the creation of true positive and negative controls, something that is not possible in strict experiments using only metagenomic data from environmental samples. Without the use of true positive and negative samples, the experimental design is flawed, and conclusions derived from the study can be brought into question (Stobbe et al., 2012). This is not to say that mock datasets are a complete substitute for real environmental data sets, only that they are a resource that can be utilized for the testing of metagenomic analysis tools to better understand the outputs from studies with metagenomic data.

There are two main types of metagenomic mock datasets. The first type called an *in vitro* mock community dataset, is constructed by placing organisms in a simulated community before extracting the DNA or genetic material and sequencing the community (Fouhy, 2016; Fausser, 2011). This type of mock community is defined as a mixture of microbial cells, viruses or nucleic acids that were created *in vitro* to provide a simulation of the composition of a microbial sample (Castelino, 2014). This is considered a synthetic or laboratory mock community because it is not a community derived from a real environmental sample. Since the completion of the Human Genome Project and the Human Microbiome Project, this type of dataset has been used extensively to simulate the microbial community structure found in real environmental samples. Examples of these datasets are The Human Microbiome Project's BEI: HM-280, HM-281, HM-278D and HM -279D, these databases are available through BEI for researchers working on infectious diseases of humans (NIH HMMC web). Another popular mock community is the

Mock Bacteria ARchaea Community (MBARC-26) created for researchers working with archaea communities. However, this type of dataset is only an estimation of the community structure found in environmental metagenomic datasets and cannot completely replicate the relationships between community members (Wu, 2016). It should also be noted that since the community structure is calculated before sequencing, the actual amount of members is somewhat variable, due to extraction and sequencing errors (Miller, 2017).

The second type of mock metagenomic dataset is derived from *in silico* modeling that has been used to analyze programs in computer science (Richter et al., 2008). Many fields are now using these statistical and computer based *in silico* models to evaluate and optimize products and tools before implanting them in further studies. These are known as *in silico* mock metagenomic datasets. This type of dataset uses sequencing data and genomes from databases like NCBI. The quality of the sequencing and genome completeness is analyzed prior to the incorporation of each genome into the datasets. This allows stricter calculations of detection limits and specificity compared to other methods where levels could be confounded by pre-analysis errors. MetaSim is one of the most successfully used open access metagenomic data simulators available (Richter, 2008). MetaSim allows for common errors based on sequencing platform to be incorporated into the datasets to more realistically simulate a metagenomic data (Richter, 2008). This software works by generating collections of synthetic reads from specifically chosen genomes. The genome's representation, as well as, the number of reads from each genome can be designated during the taxon profile phase. The program then generates mate pairs based on platform models. More tools that enable experiments to mock metagenomic communities *in silico* are coming to the marketplace like InSilicoSeq (Gourle et al., 2018). This tool generates Illumina reads for simulating metagenomic samples. In addition to

providing more control on the mock community genome inputs, the cost of constructing an *in silico* mock metagenomic data set is minimal compared to other experiments that require extraction and sequencing. This is one reason why many fields, including food chemistry, have started regularly using *in silico* modeling for optimization studies (Lambert, 2012). This method also provides research at facilities that are not equipped to handle live human pathogens with the ability to conduct preliminary experiments containing sequence data from human pathogens without containment or health risks. The metagenomic analysis tools can then be evaluated by comparing the input data to the output data (Blagden, 2016). Like all modeling-based experiments, the tools used will then need to be validated using real metagenomic data from environmental and laboratory samples, because nothing can replace the use of real environmental data.

Both *in vitro* and *in silico* mock metagenomic data types are extremely useful in understanding how metagenomic analysis tools process and profile data. These tools are essential because completing metagenomic studies without an understanding of the biases and detection limits of the tools, can result in errors. If erroneous conclusions are made about metagenomic datasets due to the use of unvalidated tools, the understanding of metagenomic community structure can be obscured.

LITERATURE CITED

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., et al., 2009. Next generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol. Plant Pathol. 10, 537–545

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

Aruscavage, D., K. Lee, S. Miller, and J.T. Lejeune. (2006). Interactions affecting the proliferation and control of human pathogens on edible plants. Journal of Food Science. 71:R89R99.

Avaniss-Aghajani, E., Jones, K., Chapman, D., & Brunk, C. (1994). A molecular technique for identification of bacteria using small subunit ribosomal RNA sequences. *BioTechniques*, *17*(1), 144-6.

Bae, D., K.S. Seo, T. Zhang, and C. Wang. (2013). Characterization of a potential Listeria monocytogenes virulence factor associated with attachment to fresh produce. Applied and Environmental Microbiology. 79(22):6855-6861.

Barak, J.D., and B.K. Schroeder. (2012). Interrelationships of Food Safety and Plant Pathology: The Life Cycle of Human Pathogens on Plants. Annual Review of Phytopathology. 50(1):241266.

Bartels, M., Petersen, A., Warning, P., Nielsen, J., Larner-Svensson, H., Johansen, H., . . . Westh, H., (2014). Comparing whole-genome sequencing with Sanger sequencing for spa typing of methicillin-resistant Staphylococcus aureus. *Journal of Clinical Microbiology,52*(12), 4305-8.

Blagden, T., Schneider, W., Melcher, U., Daniels, J., & Fletcher, J. (2016). Adaptation and Validation of E-Probe Diagnostic Nucleic Acid Analysis for Detection of Escherichia coli O157:H7 in Metagenomic Data from Complex Food Matrices. *Journal of Food Protection,79*(4), 574-581.

Bettelheim, K.A., (2007). The Non-O157 Shiga-Toxigenic (Verocytotoxigenic) Escherichia coli; Under-Rated Pathogens. Critical Reviews in Microbiology/ 33(1):67-87.

.

Brandl, M., and G.W. Sundin. (2013). Focus on food safety: Human pathogens on plants. Phytopathology. 103:304-305.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P. & Rohwer, F (2003). Metagenomic Analyses of an Uncultured Viral Community from Human Feces. J Bacteriol 185, 6220-6223.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F.

Cebula, T. A., Payne, W. L. & Feng, P. (1995). Simultaneous identification of strains of Escherichia coli serotype O157: H7 and their Shiga-like toxin type by mismatch amplification mutation assay-multiplex PCR. J Clin Microbiol 33, 248-250.

Brooks, J.T., E.G., Sowers, J.G. Wells, K.D. Greene, P.M. Griffin, R.M. Hoekstra, and N.A. Strockbine. (2005). Non-O157 Shiga Toxin-Producing Escherichia coli Infections in the United States, 1983–2002. Journal of Infectious Diseases. 192(8):1422-1429.

Castelino, M., Eyre, S., Moat, J., Fox, G., Martin, P., Ho, P., . . . Barton, A., (2017). Optimisation of methods for bacterial skin microbiome investigation: Primer selection and comparison of the 454 versus MiSeq platform. *BMC Microbiology, 17*(1), 23.

doi:10.1371/journal.pone.0003373 http://www.plosone.org/article/info:doi/10.1371/journal.pone.0003373

Centers for Disease Control and Prevention. (2006). Update on multistate outbreak of E. coli O157:H7 infections from fresh spinach, October 6, 2006. Available at:http://www.cdc.gov/ecoli/2006/september/updates/100606.htm

Centers for Disease Control and Prevention. (2011). Food safety facts. Available at:http://www.cdc.gov/foodsafety/facts.html#what

Centers for Disease Control and Prevention. (2011b). Investigation Update: Outbreak of Shiga toxin-producing E. coli O104 (STEC O104:H4) infections associated with travel to Germany. Available at: http://www.cdc.gov/ecoli/2011/ecoliO104/

Centers for Disease Control and Prevention. (2012). Multistate outbreak of listeriosis linked to whole cantaloupes from Jensen Farms, Colorado. Available at: http://www.cdc.gov/listeria/outbreaks/cantaloupes-jensen-farms/082712/index.html

Centers for Disease Control and Prevention (CDC). (2012). National Shiga toxin-producing Escherichia coli (STEC) Surveillance Overview. Atlanta, Georgia: US Department of Health and Human Services.

Centers for Disease Control and Prevention. (2012b). Multistate outbreak of Shiga toxin producing Escherichia coli O26 infections linked to raw clover sprouts at Jimmy John's Restaurants (Final Update). Available at:http://www.cdc.gov/ecoli/2012/O26-02-12/

Centers for Disease Control and Prevention. (2013). Multistate outbreak of Shiga toxin-producing Escherichia coli O157:H7 infections linked to ready-to-eat salads (final update). Available at:http://www.cdc.gov/ecoli/2013/O157H7-11-13/

Center for Food Safety Applied Nutrition . Office of Regulations Policy. (2003). *Risk assessment for food terrorism and other food safety concerns*. College Park, Md.?]: CFSAN/Office of Regulations and Policy.

Chaisson, M. J. & Pevzner, P. A. (2008). Short read fragment assembly of bacterial genomes.

Genome Res 18, 324-330.

Cole, D., P.M. Griffin, K.E. Fullerton, T. Ayers, K. Smith, L.A. Ingram, and R.M. Hoekstra. (2014). Attributing sporadic and outbreak-associated infections to sources: blending epidemiological data. Epidemiology & Infection. 142(02):295-302.

Cox-Foster, D.L., Conlan, S., Holmes, E.C., Palacios, G., Evans, J.D., et al., 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318, 283–287.

Daniel, R., (2005). The metagenomics of soil. Nat Rev Micro 3, 470-478.

De Giusti, M., C. Aurigemma, L. Marinelli, L., Tu, D. De Medici, and S. Di Pasquale. (2010). The evaluation of the microbial safety of fresh ready-to-eat vegetables produced by different technologies in Italy. Journal of Applied Microbiology. 109: 996-1006.

Elahi, E. & Ronaghi, M. (2004). Pyrosequencing. In Bacterial Artificial Chromosomes, pp. 211219: Springer.

Eo'Donnell H, & Stephen Emcsorley. (2014). Salmonella as a model for non-cognate Th1 cell stimulation. *Frontiers in Immunology, 5*, Frontiers in Immunology, 01 December 2014, Vol.5.

Erickson, M.C., (2012). Internalization of fresh produce by foodborne pathogens. Annual Review of Food Science and Technology. 3:283-319.

http://www.fda.gov/Food/FoodIngredientsPackaging/IrradiatedfoodPackaging/ucm074734.htm

Fausser, Lee, Villari, Zeng, Zhang, Serikov, Gabriel. (2011). Numerical benchmarks TRIPOLI − MCNP with use of MCAM on FNG ITER bulk shield and FNG HCLL TBM mock-up experiments. *Fusion Engineering and Design, 86*(9), 2135-2138.

FDA, "Bacteriological Analytical Manual (BAM), U.S. Food and Drug Administration," 2013.
http://www.fda.gov/Food/FoodScienceResearch/LaboratoryMetods/BacteriologicalAnalyticalManualBAM/
ucm2006 949.htm

Fey, Axel, Eichler, Stefan, Flavier, Sebastien, Christen, Richard, Hofle, Manfred G., & Guzman, Carlos A. (2004). Establishment of a Real-Time PCR-Based Approach for Accurate Quantification of Bacterial RNA Targets in Water, Using Salmonella as a Model Organism. *Applied and Environmental Microbiology, 70*(6), 3618-3623.

Fichot, E., & Norman, R. (2013). Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome, 1*(1), 10.

Fletcher, J., Bender, C., Budowle, B., Cobb, W., Gold, S., Ishimaru, C., Luster, D., Melcher, U., Murch, R. & Scherm, H. (2006). Plant pathogen forensics: capabilities, needs, and recommendations. Microbiol Mol Biol R 70, 450-471.

Fouhy, F., Clooney, A., Stanton, C., Claesson, M., & Cotter, P. (2016). 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice, and sequencing platform. *BMC Microbiology, 16*(1), 123.

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W. & DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters. Proc. Natl. Acad. Sci. U. S. A. 105, 3805-3810.

Geyer Jeanne, Wasieloski Leonard, Padilla Susana, Bode Elizabeth, Kumar Kamal, Zavaljevski Nela, . . . Reifman Jaques. (2008). In silico microarray probe design for diagnosis of multiple pathogens. *BMC Genomics, 9*(1), 496.

Gibbons, A. (2007). Hoffmann-La Roche's PCR Push. *Science, 253*(5020), 627.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M. & Nelson, K. E. (2006). Metagenomic Analysis of the Human Distal Gut Microbiome. Science 312, 1355-1359.

Gourlé H, Karlsson-Lindsjö O, Hayer J, and Bongcam+Rudloff E, Simulating Illumina data with InSilicoSeq. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty630

Havelaar, A.H., A. Cawthorne, F. Angulo, D. Bellinger, T. Corrigan, A. Cravioto, and T. Kuchenmüller. (2013). WHO Initiative to Estimate the Global Burden of Foodborne Diseases. The Lancet. 381, S59.

Hirano, S.S., and C.D. Upper. (1983). Ecology and epidemiology of foliar bacterial plant pathogens. Annu. Rev. Phytopathol. 21:243–70.

Huse, S., Huber, J., Morrison, H., Sogin, M. & Welch, D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. Genome Bio 8, R143.

International Human Genome Sequencing Consortium (IHGSC) (2004). "Finishing the euchromatic sequence of the human genome." *Nature*. **431**, (7011): 931–945. Bibcode:2004Natur.431..931H. doi:10.1038/nature03001. PMID 15496913.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature Genetics 44, 226-232.

Jones, T.F., E. Scallan, and F.J. Angulo. (2007). FoodNet: Overview of a decade of achievement. Foodborne Pathogens and Disease. 4(1): 60-66.

Jones, W., 2010. High-throughput sequencing and metagenomics. Estuar. Coasts 33, 944–952.

Kalapothakis, E., Jardim, S., Magalhaes, A. C., Mendes, T. M., De Marco, L., Afonso, L. C. C., & Chávez-Olórtegui, C. (2001). Screening of expression libraries using ELISA: identification of immunogenic proteins from Tityus bahiensis andTityus serrulatus venom. *Toxicon*, *39*(5), 679-685.

Karlsson, O. E., Hansen, T., Knutsson, R., Löfström, C., Granberg, F., & Berg, M. (2013). Metagenomic detection methods in biopreparedness outbreak scenarios. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, *11*(S1), S146-S157.

Karmali, M.A., (1989). Infection by Verocytotoxin-producing Escherichia coli. Clin. Microbiol. Rev. 2:15-38.

Kimura, H., Morita, M., Yabuta, Y., Kuzushima, K., Kato, K., Kojima, S., Matsuyama, T.Kristensen, D. M., Mushegian, A. R., Dolja, V. V. & Koonin, E. V. (2010). New dimensions of the virus world discovered through metagenomics. Trends Microbiol 18, 11-19.

Kirk MD, Pires SM, Black RE, et al. World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis [published correction appears in PLoS Med. 2015 Dec;12(12):e1001940]. *PLoS Med*. 2015;12(12):e1001921. Published 2015 Dec 3. doi:10.1371/journal.pmed.1001921

Kovačević, M., J. Burazin, H. Pavlović, M. Kopjar, and V. Piližota. (2013). Prevalence and level of Listeria monocytogenes and other Listeria sp. in ready-to-eat minimally processed and refrigerated vegetables. World Journal of Microbiology and Biotechnology. 29(4):707-712.

Korem, Zeevi, Suez, Weinberger, Avnit-Sagi, Pompan-Lotan, Segal. (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science (New York, N.Y.), 349*(6252), 1101-1106.

Lambert, J., Yennawar, N., Gu, Y., & Elias, R. (2012). Inhibition of secreted phospholipase A2 by proanthocyanidins: A comparative enzymological and in silico modeling study. *Journal of Agricultural and Food Chemistry, 60*(30), 7417-20.

Leff, J.W., and N. Fierer. (2013). Bacterial Communities Associated with the Surfaces of Fresh Fruits and Vegetables. PLoS ONE. 8(3): e59310.

Lepick, O., (1945). French activities related to biological warfare, 1919-45. Biological and toxin weapons: research, development, and use from the Middle Ages to, 70-90.

Lu, Giordano, & Ning. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics, 14*(5), 265-279.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., Konstantinidis, K., Rodriguez-Valera, F., & Rodriguez-valera, Francisco. (2012). Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample (Illumina vs. Roche 454 Metagenomic Sequencing). *PLoS ONE, 7*(2), E30087.

Luna-Gierke, R.E., P.M. Griffin, L.H. Gould, K. Herman, C.A. Bopp, N. Strockbine, and R.K. Mody. (2014). Outbreaks of non-O157 Shiga toxin-producing Escherichia coli infection: USA. Epidemiology & Infection, FirstView, 1-11.

Lyons, E., Scheible, M., Sturk-Andreaggi, K., Irwin, J., & Just, R. (2013). A high-throughput Sanger strategy for human mitochondrial genome sequencing. *BMC Genomics, 14*(1), 881.

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., et al., 2010. Bioinformatics for next generation sequencing data. Genes 1, 294–307.

Matthews, K.R., (2009). Leafy vegetables. In Sapers, G.M., E. Solomon, and K.R. Matthews (Eds.). The produce contamination problem: Causes and solutions. Elsevier, Waltham, MA.

Mardis, E. R., (2008). Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics 9, 387-402.

Mardis, E. R., (2013). Next-Generation Sequencing Platforms. Annual Review of Analytical Chemistry 6.

Metzker, M.L., 2010. Sequencing technologies — the next generation. Nat. Rev. Genet. 11, 31–46.

Merriman, B., R&D Team, I., & Rothberg, J. (2012). Progress in Ion Torrent semiconductor chip based sequencing. *ELECTROPHORESIS, 33*(23), 3397-3417.

Miller, J. R., S. Koren, and G. Sutton. "Assembly algorithms for next-generation sequencing data." Genomics 95.6 (2010): 315-27. PubMed. Web. 25 May 2017

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D. & Bushman, F. D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21, 1616-1625.

MUMmer 2.1, NUCmer, and PROmer are described in "Fast Algorithms for Large-scale Genome Alignment and Comparision." A.L. Delcher, A. Phillippy, J. Carlton, and S.L. Salzberg, *Nucleic Acids Research* (2002), Vol. 30, No. 11 2478-2483.

Munir, S., Singh, S., Kaur, K. & Kapur, V. (2004). Suppression subtractive hybridization coupled with microarray analysis to examine differential expression of genes in virus infected cells. Biol Proced Online 6, 94-104.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H. & Remington, K. A. (2000). A wholegenome assembly of Drosophila. Science 287, 2196-2204.

Nakamura, S., Yang, C. S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., ... & Mizutani, T. (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PloS one*, *4*(1), e4219.

Nakamura, S., Nakaya, T., & Iida, T. (2011). Metagenomic analysis of bacterial infections by means of high-throughput DNA sequencing. *Experimental Biology and Medicine*, *236*(8), 968-971.

Nakano, M., Komatsu, J., Matsuura, S.-i., Takashima, K., Katsura, S. & Mizuno, A. (2003).

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/

Nilsson, R. H., Ryberg, M., Abarenkov, K., Sjökvist, E. & Kristiansson, E. (2009). The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. FEMS Microbiology Letters 296, 97-101.

Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., et al., 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. 358, 991–998.

Pop, M., Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. Trends Genet. 24, 142–149.

Postnikova, E., Baldwin, C., Whitehouse, C.A., Sechler, A., Schaad, N.W., et al., 2008. Identification of bacterial plant pathogens using multilocus polymerase chain reaction/ electrospray ionization-mass spectrometry. Phytopathology 98, 1156–1164.

Priyanka, B., Rajashekhar K. Patil, and Sulatha Dwarakanath. "A Review on Detection Methods Used for Foodborne Pathogens." *The Indian Journal of Medical Research* 144.3 (2016): 327–338. *PMC*. Web. 22 June 2017.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H.

& Turner, D. J., (2008). A large genome center's improvements to the Illumina sequencing system. Nat Methods 5, 1005-1010.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences, and Illumina MiSeq sequencers. BMC Genomics 13, 341.

Ramaswamy, V., V.M. Cresence, J.S. Rejitha, M.U. Lekshmi, K.S. Dharsana, S.P. Prasad, and H.M. Vijila. (2007). Listeria – review of epidemiology and pathogenesis. J Microbiol Immunol Infect. 40(1):4–13.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH, *MetaSim: a sequencing simulator for genomics and metagenomics.*, PLoS One, Oct. 8, 2008 [Abstract, cited in PMC]

Rodrigue, S., Materna, A., Timberlake, S., Blackburn, M., Malmstrom, R., Alm, E., . . . Gilbert, J., (2010). Unlocking Short Read Sequencing for Metagenomics (DNA Sequencing Pipeline). *PLoS ONE, 5*(7), E11840

Ronaghi, M., 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. 11, 3–11.

Roossinck, M. J., Saha, P., Wiley, G. B., Quan, J., White, J. D., Lai, H., ChavarrÍA, F., Shen, G. & Roe, B. A. (2010). Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19, 81-88.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463–5467. doi:10.1073/pnas.74.12.5463

Satya, Zavaljevski, Kumar, Bode, Padilla, Wasieloski, Biotechnology HPC Software Applications Inst Fort Detrick MD. (2008). In silico Microarray Probe Design for Diagnosis of Multiple Pathogens.

Schaad, N.W., Frederick, R.D., Shaw, J., Schneider, W.L., Hickson, R., et al., 2003. Advances in molecular-based diagnostics in meeting crop biosecurity and phytosanitary issues. Annu. Rev. Phytopathol. 41, 305–324.

Schloss, P. & Handelsman, J. (2005a). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Bio 6, 229

Sivapalasingam, S., C.R. Friedman, L. Cohen, and R.V. Tauxe. (2004). Fresh produce: A growing cause of outbreaks of foodborne illness in the united states, 1973 through 1997. Journal of Food Protection. 67(10):2342-2353.

Solomon, E.B., M. Brandl, and R.E. Mandrell. (2006). Behavior of human pathogens on produce. In: Matthews, K.R., (Ed.) Emerging Issues in Food Safety: Microbiology of Fresh Produce. Washington, D.C., ASM Press. p. 55-83.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., … Birney, E., (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome research*, *12*(10), 1611-8.

Stobbe, A. H., Daniels, J., Espindola, A. S., Verma, R., Melcher, U., Ochoa-Corona, F., ... & Schneider, W. (2013). E-probe Diagnostic Nucleic acid Analysis (EDNA): a theoretical approach for handling of next generation sequencing data for diagnostics. *Journal of microbiological methods*, *94*(3), 356-366.

Stobbe, A. H., Schneider, W. L., Hoyt, P. R., & Melcher, U. (2014). Screening metagenomic data for viruses using the e-probe diagnostic nucleic acid assay. *Phytopathology*, *104*(10), 1125-1129.

Singer, Andreopoulos, Bowers, Lee, Deshpande, Chiniquy, Woyke. (2016). Next generation sequencing data of a defined microbial mock community.

Thorns, C.J., (2000). Bacterial food-borne zoonoses. Rev Sci Tech. 19(1):226-239.

Trainor, G. L., (1990). DNA sequencing, automation, and the human genome. Anal Chem 62, 418-426.

Tucker, T., Marra, M., Friedman, J.M., 2009. Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Genet. 85, 142–154.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., et al., 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37–43.

USDA, FSIS. 2007. “FSIS Procedure.” *Laboratory Guidebook MLG 8A.03*.

Vaneechoutte, M., & Van Eldere, J. (1997). The possibilities and limitations of nucleic acid amplification technology in diagnostic microbiology. *Journal of medical microbiology*, *46*(3), 188-194.

Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. Clin Chem 55, 641-658.

Ward, T.J. L. Gorski, M.K. Borucki, R.E. Mandrell, J. Hutchins, and K. Pupedis. (2004). Intraspecific Phylogeny and Lineage Group Identification Based on the prfA Virulence Gene Cluster of Listeria monocytogenes. Journal of Bacteriology. 186(15):4994–5002.

Wood DE, Salzberg SL: Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 2014, 15:R46.

Wu, Y., Simmons, B., & Singer, S. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics, 32*(4), 605-607.

Xie, W., Meng, Q., Wu, Q., Wang, S., Yang, X., Li, N., . . . Vontas, John. (2012). Pyrosequencing the Bemisia tabaci Transcriptome Reveals a Highly Diverse Bacterial Community and a Robust System for Insecticide Resistance (Transcriptome Profiling of B. tabaci). *PLoS ONE, 7*(4), E35181.

Yandell, M., Evans, C. A. & Holt, R. A. (2001). The sequence of the human genome. Science Signaling 291, 1304.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., ... & Wang, J. (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *Journal of clinical microbiology*, *49*(10), 3463-3469.

Yiannas, F., (2009). Food safety culture: Creating a behavior-based safety management system. Springer, New York, NY.

Yoshitomi, K., Jinneman, K., Orlandi, P., Weagant, S., Zapata, R., & Fedio, W. (2015). Evaluation of rapid screening techniques for detection of Salmonella spp. from produce samples after pre-enrichment according to FDA BAM and a short secondary enrichment. *Letters in Applied Microbiology,61*(1), 7-12.

Zhang Z., Schwartz S., Wagner L., & Miller W. (2000), "A greedy algorithm for aligning DNA sequences" J Comput Biol 2000; 7(1-2):203-14. PubMed

CHAPTER III

EVALUATION OF BIOINFORMATIC PIPELINES FOR DETECTION OF *SALMONELLA ENTERICA* IN METAGENOMIC DATA

Abstract

Aim: Compared to the current pathogen detection methods, a metagenomics-based approach offers the potential to detect any and all known and unknown pathogens present in a complex sample in one assay. However, there are challenges that need to be addressed before pathogen detection from complex metagenomic data becomes practical. The aim of this study was to evaluate the influence of sequencing platforms, assembly and bioinformatic pipelines on the detection of foodborne pathogen *Salmonella enterica* in metagenomic data generated from fresh tomato surface wash.

Materials and Results: DNA was extract from the surface wash of commercial tomatoes with two *S. enterica* contamination levels (S1 and S2) and one control group (T1). Four community profiling bioinformatic pipelines (BLAST, DIAMOND, Kraken2, and Bowtie2) and one targeted pipeline, E- probe Diagnostic Nucleic-acid Analysis (EDNA) were used to analyze Illumina and 454 metagenomic cleaned clean reads and contigs for

detection of *Salmonella enterica* from the surface washes of tomatoes inoculated with two bacterial suspensions (10^3 or 10^6 cells per tomato). Detection limit and overall pipeline performance were compared. It was found that only Kraken and EDNA provided the speed necessary for rapid detection and only EDNA provided the sensitivity necessary for detection.

Conclusions: Among the bioinformatics pipelines evaluated, EDNA offers a faster and a more straight forward detection of human foodborne pathogens in metagenomic data.

Significance and Impact of the study: Utilizing metagenomics allows for an entire community of organisms to be analyzed from sequence data. Identifying bioinformatic tools for bacterial human pathogen detection from sequence data should enhance the safety of food products by expediting forensic trace-back investigations and determining the causal agents and sources of human diseases. The evaluation of speed and detection among databases and bioinformatic pipelines provides a further understanding of the benefits and limitations of currently available methods of pathogen detection in metagenomic sequence data.

Introduction

Foodborne human pathogens pose a significant risk to human health and welfare and are of particular concern to children, the elderly and those with compromised immune systems (Lund, 2011: Liu et al., 2018). Currently, 31 foodborne pathogens have been identified as the causal agents of diseases in humans (CDC, 2016). Fresh produce, an essential part of a healthy diet and often eaten raw, is particularly at risk for foodborne pathogen contamination due to the lack of pathogen killing steps such as cooking before consumption (Jung, 2014). In many developed countries, including the United States, improvements in

sanitation and farming practices have mitigated the levels of foodborne pathogen contamination on food. However, the CDC estimated that in 2016, foodborne pathogens resulted in 9.4 million illnesses, 55,961 hospitalizations, and 1,351 deaths in the United States (CDC, 2016). Bacterial pathogens make up the majority of the pathogens known to cause foodborne illnesses, and *Salmonella enterica* is listed as the top foodborne pathogen contributing to hospitalization (35%) and death (28%) in the United States (CDC, 2016). Because of the risk posed to human health, accurate and rapid detection of foodborne pathogens in complex food matrix is critical for routine quality control as well as foodborne outbreak investigations.

In the United States, the standard detection protocols for foodborne pathogens in various foods are developed and validated by two regulatory agencies: the U.S. Food and Drug Administration (FDA) and the U. S. Department of Agriculture (USDA). These assays are published as FDA's Bacteriological Analytical Manual (FDA, 2013) and USDA's Microbiology Laboratory Guidebook (USDA, 2007). The MLG assays are designed for detection of foodborne pathogens in meat, poultry, and certain egg products while FDA's assays for detection in all the remaining food matrices. Despite this difference, the overall laboratory procedures for the detection of each pathogen between the two regulatory agencies are very similar. In general, the standard assays consist of pre-enrichment, selective enrichment, plating on differential media, and biochemical, serological, or molecular tests for confirmation. Each assay could take 4-6 days to complete, and only one pathogen could be detected in each assay.

For rapid detection, Polymerase Chain Reaction (PCR) has long been considered the gold standard (Avaniss-Aghajani, 1994; Priyanka, 2017). This process is rapid (how long to complete?) and sensitive with detection limits as low as femtograms (10^{-15}g)

(Priyanka, 2017). For foodborne pathogens, the detection limit of multiplex PCR has been published as 10^3 CFU/ml (Yu et al., 2016). This method can also generate high false positive rates depending on the specificity of the primers used (Avaniss-Aghajani, 1994; Priyanka, 2017). In addition to their individual limitations, the detection methods above are limited by their ability to only detect a single pathogen or a small group of pathogens in one assay (Avaniss-Aghajani, 1994; Priyanka, 2017).

Metagenomics is the study of the genomic makeup of environmental samples and can be used to assess sample biodiversity (Breitbart et al., 2003; Gill et al., 2006; Hirano, 1983), gene expression (Frias-Lopez et al., 2008; Uchiyama et al., 2004), and gene interaction within an environment (Schwartz and Beaver 2011; Singh et al., 2013). Metagenomic sequencing allows the direct genetic analysis of a complex environmental sample (Karlsson, 2013). This streamlines the microbial identification process by eliminating the need for culturing or isolation (Nakamura, 2009; Nakamura, 2011). These breakthroughs in the field of microbial ecology can also contribute to other microbial fields, such as microbial identification (Nakamura, 2009). This method has been primarily used to profile whole microbial communities in environmental samples associated with soil, water, and humans/animals. The strength of this type of work revolves around the ability to "reconstruct" an entire community from a single sample. Metagenomics has played a vital role in discovering uncultivable organisms and viruses in complex environmental samples (Nakamura, 2009). This has been key in uncovering viruses as well as hard to culture pathogens. By aligning the assembled contigs to viral genomes, Yang et al (2011) were able to detect viral pathogens from clinical samples. This method is not limited to presence or absence detection. By translating DNA reads into RNA or proteins, a more complete picture of community function and the genes involved can emerge. In

order to get a quantitative view of community function, differential transcriptomics can be used to understand how inputs into soil, water or the human microbiome can influence the microbial community (Luo et al., 2017). This method also has almost limitless application for pathogen detection, because any genome present can be reconstructed from the sequence data (Wylezich et al., 2018). Using this method, it is possible to detect any pathogen present in the sample, but the limitations arise from the limited information on the performance about the computational pipelines that can be used to process metagenomic data.

Bioinformatic pipelines are formed by stringing together bioinformatic tools and programs (Golob et al., 2017). Each tool takes data in and performs a function on the data. The functions include trimming and refining the data, assembling overlapping reads into long contigs, mapping reads to sequences in databases, and assigning taxonomy to the mapped reads. These programs are necessary for bioinformatic work, however, because each tool is based on a specific algorithm the data that is processed by a specific program is marked with the inherent biases of the algorithm. To meet the demand of data processing, new bioinformatic tools are emerging constantly, but the inherent biases of the algorithms in these tools are not always obvious. One of the most informative tests to identify algorithm bias is by running the same data through multiple tools and pipelines and assessing the differences in output (Golob et al., 2017). Bias is of particular concern in pathogen detection because the degree of taxonomic resolution needed is extremely high. If a pipeline is unable to resolve taxonomy below the genus level it is not going to be useful in pathogen detection because many pathogenic and nonpathogenic species share the same genera.

Computational pipelines begin with sequence data. The most popular sequencing platform is Illumina due to the amount of data output compared to the cost (Mitra et al., 2010; Lawrence Berkeley National Laboratory 2010). The Illumina method uses massively parallel sequencing by leveraging clonal array formation and reversible terminator technology. Using the “bridge” technology, four fluorescently labeled nucleotides flow across the flow cell. When they attach to the nucleotide chain, they release fluorescence that is base specific and picked up by the device. Illumina is known for producing "short reads" that are from 50-150bps. Using very short reads without assembly may contribute to a high false positive rate in detection applications due to the fact that short reads are more likely to map to multiple areas in many genomes or be assigned to a species in greater abundance in the database depending on the algorithm used. "Long read" sequencing is less popular due to the higher cost. Roche 454 was one of the first commercial platforms for next generation sequencing. It used large scale parallel pyrosequencing with a capability to sequence approximately 400-600 megabases (Mb) of DNA in a ten turn run (Klein et al., 2011). The library preparation is done by shredding the DNA into 300-800bp and blunting each end. Adaptors are then ligated to the fragment ends. The adapter containing the 5’-biotin tag for immobilizing the DNA library to the streptavidin-coated beads. Nick repair occurs and releases the non-biotinylated strand, which is used as the single-stranded template DNA (sstDNA) and emPCR amplification occurs and the templates remain encapsulated in water-in-oil mixture beads. The sstDNA beads are added to the DNA Bead Incubation Mix and layered with Enzyme Beads on the PicoTiterPlate device and the beads are placed into the well through centrifugations were the sequencing reaction occurs. Nucleotides are then washed over the plate and are added to the templates in parallel. In wells where a nucleotide addition occurs, the light reaction is quantified by

a CCD camera. The signal strength is proportional to the number of single nucleotides incorporated. However, the lack of ability to detect more than eight consecutive single nucleotide stretches (homopolymer) in a drawback of this type of sequencing which Illumina does not share. Roche 454 sequencing was removed from the market in 2016 when it was found to be noncompetitive, but it can be used to compare the effect of using long versus short reads in metagenomic community studies. A direct comparison of the Illumina and Roche 454 sequencing was completed in order to identify how the two different platforms treated the data (Luo et al, 2012). The metagenomic sample tested was a complex freshwater planktonic community. The study summarized that despite differences in read length and sequencing protocols that both platforms overlapped in approximately 90% percent of the taxon assembled. It has been hypothesized that Roche 454 could be better for metagenomic community studies due to the fact that longer read lengths could provide a more complete picture of the community with less assembly (Mitra et al., 2010). However, the greater amounts of reads produced by an Illumina run could provide more targets for detection and a lower sequencing cost.

After sequencing, many pipelines choose to assemble reads into contigs which reduces the overall amount of data by consolidating many reads into a single contig and allows for longer sequences alignments. Creating contigs can also result in fewer false positive hits in BLAST because longer sequences with lower E-values are statistically more relevant. There is some debate about how this affects the reduction of reads and the relative abundance calculations but there is not a consensus (Knight et al., 2018). However, contigs creation has the benefit of resulting in more significant database searches due to the fact that longer queries do not result in as many false positive classifications as shorter reads (Jones et al., 2013; Carr et al., 2014). The contigs can then be used for informative or non-

informative database searches. Creating contigs takes time and the program used to create the contigs may impact the taxonomic assignment. It is important to observe the potential differences between pipelines that use mapped contigs verse clean read data.

The two main strategies for sequence mapping are informative and non-informative. Informative searches involve identifying biologically informative genes also called open reading frames (ORFs) or coding domains (CDs) in the sequence data (Das et al., 2018). Using the ORFs is a popular strategy because it classifies sequences based on relevance and reduces redundant searches (Pookhao et al., 2015). This is extremely relevant in sequence data involving eukaryotic organisms where non-coding regions and regions containing identical strings of nucleotides are prevalent (Liu et al., 2013). Another benefit of this strategy is that it can reduce the amount of false positive mapping because the searches are limited to only well characterized gene regions (Dos Santos et al., 2017). The main limitation of this strategy is that it relies on the identification and characterization of ORFs (Kolde and Vilo, 2015). This is an issue with metagenomic research because many of the organisms in the mixed sample have either not been well studied or are unknown (Nagarajan et al., 2014). This means that many of the ORFs will not be able to be identified and the ones that are may not be indefinable at an informative taxonomic level. This will likely improve as ORF databases increase This method has been used for functional analysis of microbial communities but not detection (Dos Santos et al., 2017). Because of the current limited taxonomic clarity, this method is not suitable for detection application from complex metagenomic data.

Non-informative searches look for sequence similarity without regard to gene coding regions or open reading frames (Chattaway et al., 2017). This method can often achieve a higher degree of taxonomic resolution because it is not dependent on the

characterization of ORFs. However, it is more likely to result in a higher rate of false identification depending on how the search algorithm identifies matches (Popic et al, 2018). This type of search can often take longer than informative searches based on predicted ORFs because of the relative sizes of the databases. Regardless of which strategy is used, it is important to understand how these different strategies compare. Meaning, are different pipelines converging and resulting in similar taxonomic profiles at different levels of clarity, or are different methods resulting in significantly different species abundance at all taxonomic levels?

The most commonly used bioinformatic pipelines for analyzing metagenomic data are heuristic like those found in the widely used Basic Local Alignment Search Tool (BLAST) (Altschul, 2009; Martins et al., 2015). This tool from NCBI uses short three-word k-mers of the query sequences to identify similar sequences in the NCBI database. Even though this process is much faster than Bayesian and strict alignments based on perfect matches, the large size of the databases makes this type of analysis computationally cumbersome. The BLAST tool has an online platform that is used extensively for local sequence searches but for large datasets a high-performance computer is still needed, and it can take many days (Santamaria, 2012). Programs like the Diamond pipeline attempts to improve the speed of BLAST by formatting the NCBI protein database with a proprietary algorithm (Buchfink, 2015). Diamond was developed as a high throughput program for DNA protein coding sequences and protein sequence alignments, 20,000 times faster than traditional BLAST while retaining high sensitivity (Buchfink, 2015). Other programs like Kraken2 assign taxonomic labels to DNA sequences using k-mer based binning. Kraken2 requires the use of the Bracken program for a re-estimation of read abundance (Wood, 2014). These pipelines can all result in a taxonomic profile, which can be used to estimate

the approximate percentage of each taxon in the profile. The Kraken2 and Bracken programs require the construction of multiple scripts for running the analysis, as well as, extensive computer resources and RAM.

Alternatively, E-probe Diagnostic Nucleic-acid Analysis (EDNA) is an *in silico* tool that allows for the creation of electronic probes (E-probes) based on a known pathogen sequence (Stobbe, 2013; Stobbe, 2014). The E-probes are created by selecting a target pathogen genome and comparing it to a closely related genome that acts as the inclusivity/exclusivity determinate (Figure 1). The E-probe length is then chosen, which is dependent on the type of target organism and the length of the genome. However, previous studies have shown that E-probes lengths of 60-80nt performed well for most microorganisms (Stobbe et al, 2014). This produces the raw E-probes that are then cleaned by aligning the raw E-probes on the NCBI database and removing off-target hits. The resulting E-probes can be stored and used to detect targets in any FASTA datasets. The E-probes are able to identify pathogens in sequence data including large metagenomic data (Stobbe, 2014). While EDNA does not provide a taxon profile or an approximate species abundance, it does have the potential to rapidly detect a pathogen in a metagenomic dataset by probe matches. It also has the benefit of being used for target detection in unassembled, non-quality checked sequence data (Stobbe et al., 2014). This method has been tested on viruses (RNA and DNA), bacteria, fungi, and oomycetes (Espindola et al, 2015). Most of the targets used for detection have been plant pathogens. However, this technique has the potential to detect any target including human pathogens from sequence data. This method provides an opportunity to detect human foodborne pathogen on non-host (fresh food substrates) which would be extremely beneficial to food safety.

Many variables in metagenomic approaches can influence the accuracy and speed in the detection of foodborne pathogens in a complex food sample. The aim of this study was to evaluate the influence of sequencing platforms, sequence input (clean reads vs. contigs), reference databases and bioinformatic pipelines on the detection of *Salmonella* in metagenomics data generated from commercial tomato surface wash (Figure 1).

Materials and Methods

DNA Extraction

Fresh Roma tomatoes were purchased from local commercial retailer located in Stillwater, OK. The tomatoes were spiked with *Salmonella* at 10^6 cell/tomato (S1), 10^3 cells/tomato (S2), and un-spiked control (T) inside the biosafety cabinets and left for air drying. For each treatment 27 tomatoes (9 tomatoes in three replicates) were taken; briefly, three tomatoes were placed in the stomacher bag containing 100 ml of UPB broth. To wash the bacteria/ native microflora from the surface of the tomatoes, 3 tomatoes were placed in the stomacher bags, shaken manually for 1 min, rubbing each tomato for 2 min, again shaking for 1 min. These tomatoes were removed, and another 3 tomatoes were placed in the same bag, washed in the similar way, removed, and another 3 tomatoes were washed in the same wash fluid. A total of 9 tomatoes were washed in same 100 ml of UPB broth. A total of 300 ml of wash fluid was collected for each treatment. Total DNA was extracted using the traditional method of DNA extraction, briefly- A total of 300 ml of the wash fluid from each of the treatment above was divided into 150 ml each in centrifuge bottles and centrifuged at 10,000 rpm for 50 mins. The pellet in each was removed by dispensing in 1ml of lysis solution [25 mM Tris, 10 mM EDTA and lysozyme (20 mg/ml)] and incubated

at 37 °C for 1 h. Sixty microliter of 10% SDS was added, and the mix was incubated at 56 °C for 30 min, followed by 2.5 µl of RNase A (20 m/ml) incubating at 37 °C for 30 mins, further 10 µl of Proteinase K (20 mg/ml; Promega) treatment was given at 56 °C for 30 mins. To the above lysate equal volume of phenol: chloroform: isoamyl alcohol (25:24:1, Sigma Aldrich, St. Louis, MO, USA) was added, mixed and centrifuged at 12,000 rpm for 15 mins. The above layer was removed carefully and extracted with equal volume of chloroform: isoamyl alcohol (24:1) and centrifuged again at 12,000 rpm for 15 mins. The supernatant was carefully separated, and the DNA was precipitated by adding 1/10 volume of sodium acetate (pH=5.2) and 2 volume of 100% chilled ethanol. The mix was precipitated by overnight incubation at -20°C. The pellet was finally collected by centrifugation at 12,000 rpm for 15 mins, washed twice with ice cold 70% ethanol, air dried in biosafety cabinet and finally, the pellet was dissolved in 50 µl of TE buffer. For the high-quality DNA for the Illumina run, the DNA was further purified and concentrated using the Zymo Research DNA clean and Concentrator kit (Zymo Research, Irvine, CA USA). For each treatment, DNA from 3 rounds of extractions were pooled together to produce the desired concentration for the Illumina run.

Illumina And Roche 454 Sequencing and Assembly

High-throughput sequencing was performed by BGI (Shenzhen, China) using the Illumina Hiseq 2000. After sequencing, the raw data was parsed to remove reads containing 'N' and adapters. The number and rate of cleaned reads were calculated. Contigs were created using the SOAP de novo program which utilizes the Bruiji graph tool that specializes in assembling NGS very short reads (Li et al., 2008).

Barcoded Roche 454 sequencing was completed using the Roche 454 GS Junior (OSU). Trimming of the raw reads and creation of contigs was completed using the Newbler (2.5pl) program for de novo sequence assembly (Miller, 2016). A copy of cleaned read data was used for the downstream cleaned read analyses. For contig creation, the program began assembly by finding overlapping reads by calculating the number of reads for alignment and building trees based on seeds of 16-mer lengths with each seed beings 12 bases upstream from the preceding seed. This was done to increase the speed of the calculation. If identical seeds are detected, the program extends the overlap between the reads until a minimum overlap of 40bps and a minimum alignment percentage of 90% is reached. The overlapping reads are compiled into a consequence sequence and the quality of each possible alignment of reads is calculated based on the consensus estimate which is based on the alignment reads called ‘nodes' and the reads between the nodes called ‘edges'. The optimal estimate is chosen based on the overall length of the contig and number of nodes versus edges in the contig. Scaffolds are calculated as a series of contigs and gaps between those contigs. Newbler provided reports that contained the identity and number of the scaffolds as well as contigs that were greater than 500bps in length. The reads and contigs from both platforms (Illumina and Roche 454) were used for downstream analysis.

BLAST-nt Pipeline

The data used was the cleaned reads and contigs from the three sample groups (S1, S2, and T1) from Illumina and 454 platforms. The non-redundant nucleotide (nt) database of NCBI was downloaded (Feb 2017). The BLAST+ program was run on the command line using the Oklahoma State University (Okstate) High Performance Computer (HPC) designated Cowboy (Altschul et al.,1990). The blastn application was used with the traditional parameters from the program that require an exact match of 11 nucleotides and the

BLOSUM62 scoring matrix. The BLAST nt data was parsed with a PERL script to obtain hits with e-values of 10^-9 or more stringent. Then MEGAN6 was used to assign the mapped reads and contigs to species level taxa. Species abundance of *S. enterica* was calculated as reads assigned to a specific taxon over all assigned reads. The percent abundance of *S. enterica* was compared to other pipelines. The speed of the pipeline was computed based on approximate run time without interruption.

Diamond-nr Pipeline

The Illumina and Roche 454 cleaned reads and contigs (S1, S2, and T1) were translated into protein coding sequences using the BLASTx program on the Okstate high performance computer "Cowboy". The NCBI non-redundant protein database (nr) was downloaded in FASTA format (Feb 2017). The Diamond program was loaded onto the Okstate HPC Cowboy (Feb 2017). Diamond was used to convert the nr database into a binary diamond database file. This database was used to align the data using the Standard Genetic code for the translation of the query. The standard alignment scoring matrix BLOSUM62 was used on the least sensitive mode. Alignments were parsed with a PERL script for hits with E-values of 10^{-9} or less. MEGAN6 was used for taxon assignment and a taxon abundance table was constructed. Species abundance of *S. enterica* was calculated as reads assigned to a specific taxon over all assigned reads. The percent abundance of *S. enterica* was calculated. The speed of the pipeline was computed based on approximate run time without interruption.

Kraken2 Pipeline

The data used was the cleaned reads and contigs from the three sample groups (S1, S2, and T1) from Illumina and 454 platforms. Kraken2 was imported into the Okstate HPC

Cowboy used to build the kraken2 Standard Database containing NCBI taxonomic information and complete RefSeq genomes for bacterial, archaeal and viral domains as well as the human genome and UniVec Core (Nov. 2018). The default parameters were used which included kmer (35) and length (31). The program uses a simple spaced seed algorithm ($s < l/4$) in order to increase alignment accuracy. The default for s=6 meaning that >1 position will be masked. This increases speed while maintaining a high level of accuracy for most alignments. The BLAST+ suit application *dustmasker* was used to mask low complexity sequences. The output from the kraken2 files was analyzed using the Bracken program (Wood et al., 2014). Bracken stands for Bayesian Re-estimation of AbundanCe with KrakEN. It is recommended by the creator of Kraken2 as a statistical method of computing the abundance of species from metagenomic DNA samples. It does this by estimating the number of reads originating for each species present in a sample. The percent abundance of *S. enterica* was calculated. The speed of the pipeline was computed based on approximate run time without interruption.

Bowtie2 Alignment of *Salmonella enterica* and Construction of Custom Database

In order to test how nonstandard parameters and custom databases affect the speed and detection of a target, an alignment to *S. enterica* was completed using a custom database containing only a *S. enterica* genome (NCBI Accession #AE006468.2) another custom database was created that contained only *S. bongori* (NCBI Accession #NZ_NAPQ01000027.1). The data used was the cleaned reads and contigs from the three sample groups (S1, S2, and T1) from Illumina and 454 platforms. The Bowtie2 program on the Okstate HPC Cowboy was used to create two custom databases containing an *S. enterica* genome (NCBI Accession #AE006468.2) and an *S. bongori* (NCBI Accession #NZ_NAPQ01000027.1). The Bowtie2 program was then used to locally align the

metagenomic reads and contigs to the *S. enterica* genome and *S. bongori* genomes. Alignment files were then visualized using the IGV and the CG View Server software package (Robinson et al., 2011). The percentage abundance of *S. enterica* and *S. bongori* were calculated as aligned reads over total reads and the overall speed of the alignment was evaluated as approximate run time without interruption.

E-probe Diagnostic Nucleic-acid Analysis (EDNA) Pipeline

The data used was the cleaned reads and contigs from the three sample groups (S1, S2, and T1) from Illumina and 454 platforms. The first step in EDNA is to create E-probes for *Salmonella enterica*, the reference sequence (NCBI Accession #AE006468.2) from NCBI was downloaded and compared to a reference sequence from *Salmonella borgori* (NCBI Accession #NZ_NAPQ01000027.1). This was done to establish the inclusivity/exclusivity determinate. The inclusivity/exclusivity determinate sequence chosen allows for the specificity of the probe set to be established. After comparison to the neighbor sequence using the MUMmer global sequence aligner, the sequence regions unique to *Salmonella enterica* were extracted using EDNA scripts and used to generate the probe set. BLAST on NCBI was used to identify if any of the probes hit on non-*Salmonella enterica* sequences in the NCBI database. The raw e-probes were mapped to the NCBI nt and genome databases If any sequences did hit on non-target organisms with an E-value of 10^{-9} or percent ID of >97%, they were removed from the probe set. *Salmonella enterica* probes of 60 and 80 nucleotides were created using the EDNA program. Six hundred and twenty-three probes were created with 60 nucleotides and 178 probes were created with 80 nucleotides. The probes were run against (Blastn) the Illumina and 454 metagenomic sequence clean reads and contigs. The matches were retrieved having percent identity >97% and e-value 10^-9 or more stringent. Hits occurred when a probe aligned to read in

the metagenomic data and was considered a positive match if the percent identity was at or above 97% with an E-value of 10^-9 or greater with a read depth of two or greater. The E-probe matches for each individual probe were counted and the number of times each individual E-probe resulted in a high-quality match was calculated as depth. *S. enterica* was compared to other pipelines. The speed of the pipeline was computed based on approximate run time without interruption. The 60nt and 80nt E-probes were mapped to the *S. enterica* genome using Integrated Genomics Veiwer and the CG View Server in order to visualize the matches in the genome and compare them with ORFs. The E-probes were also mapped to the *S. bongori* genome to observe any possible off target matches.

Results

Overview of Metagenomics Data Sets

From the 454 platform, the three samples were estimated to have an average of 22 Mb per sample with a Sample Standard Deviation (SSD) of 3.3. The Illumina paired-end sequencing resulted in over 23,000,000 clean reads per sample with an average read length of 100bp (Table 1). Using the SOAP aligner, an average of 35,354 contigs were created per sample with an average length of 46,949,175bps. The 454 single end sequencing resulted in over 109,000 cleaned reads per sample with an average read length of 402bps. Using the Newbler aligner an average of 7,616 contigs were created per sample with an average length of 722bps.

BLAST-nt Pipeline Profiles and Detection

All three samples ran for 500 hours regardless of platform or assembly (Table 2). Both the Illumina and 454 clean reads had similar patterns of *S. enterica* percent abundance. S1 had a percentage abundance of 30-32, S2 had a percentage abundance of 0.6-0.4 and T1 had range of percent abundance of 0.5-0.2. This pattern was also seen in the Illumina and 454 contigs data. Another trend seen across all sample sets was a decrease in the percent abundance of *S. enterica* from S1 to S2 and T1. The detection between S2 and T1 was difficult to distinguish even though S2 was spiked with S. enterica and T1 was not. In the community profile*, Pseudomonas* sp. *Raoultella* sp. (formerly *Klebsiella*) and *Clavibacter michiganensis* dominated the highest percentage abundances regardless of platform or assembly (Appendix). There was not a consensus on the percent abundance of these species, but they tended to be in high abundance in all BLAST pipelines. *Salmonella enterica* was also found in all samples regardless of platform or assembly. The percent abundance of this pathogen was the top hit in all S1 samples with however percent abundance in S1 and the lowest percent abundance in T1.

Diamond Pipeline Profiles and Detection

It took 24 hours to complete diamond pipeline profiling for each sample (Table 3). Similar to the results from the BLAST pipeline, the trend was for S1 to have the highest percentage abundance with S2 and T1 having lower and similar percent abundances. Similarly, to the BLAST community profiles*, Pseudomonas* sp. *Raoultella* sp (formerly *Klebsiella*) and *Clavibacter michiganensis* dominated the highest percentage abundances regardless of platform or assembly (Appendix). There was not a consensus on the percent abundance of these species, but they tended to be in high abundance in all pipelines. *Salmonella enterica* was also found in all samples regardless of platform or assembly. The percent abundance of this pathogen was the top hit in all S1 samples with lower percent

abundance in S2 and the lowest percent abundance in T1. The full taxon profiles for S1, S2, and the Control samples (T1) for both platforms are summarized in the supplementary information.

Kraken2/Bracken Profiles and Detection

All three samples ran for 2 hours (Table 4). Kraken2 is the only profiling pipeline developed for metagenomic profiling and it closely mirrored the results of the BLAST and DIAMOND pipeline and *S. enterica* percent abundances. It had the greatest total number of taxa assigned for the Illumina clean reads and contigs, while the 454 reads and contigs were similar to the other profiling pipelines. Similarly, to the BLAST and DIAMOND community profiles*, Pseudomonas* sp. *Raoultella* sp (formerly *Klebsiella*) and *Clavibacter michiganensis* dominated the highest percentage abundances regardless of platform or assembly (Appendix). There was not a consensus on the percent abundance of these species, but they tended to be in high abundance in all pipelines. The notable difference in the Kraken2 pipeline was that *S. enterica* was not in the top highest abundance in S1, although it was still close to the top and followed the sample percent abundance trend seen in the other pipelines.

Bowtie2 Alignment

The cleaned reads from samples S1, S2 and T1 were mapped to an *S. enterica* complete genome using Bowtie2. The alignment took 2 hours on the high performance computer. Alignment files were then visualized using the IGV software package. The percentage abundance of *S. enterica* was calculated as aligned reads over total reads and the overall speed of the alignment was evaluated as approximate run time without interruption (Table 5). Following the same trend seen in the previous profiling pipelines,

S1 had the highest percentage of *S. enterica* followed by S2 and T1. The alignment of the samples to *S. bongori* yielded nearly identical percentage abundances at each contamination level (Table 5).

EDNA Pipeline

The speed of detection using EDNA was 5 min regardless of platform or assembly. EDNA detected *Salmonella enterica* in the S1 clean read samples at both E-probe lengths (60nt and 80nt) regardless of whether the cleaned reads were sequenced using the Illumina or 454 platform (Table 6). The S1 contigs from the both the Illumina and 454 platforms also resulted in detection of *S. enterica*. In the S2 unassembled read samples, EDNA detected *Salmonella enterica* in the unassembled reads from the Illumina platform at both E-probe lengths but not from the 454 platform at either E-probe length (Table 6). Neither the contigs from the 454 nor from the Illumina platform resulted in *S. enterica* detection at either E-probe length in S2. No detection of *S. enterica* was found from the unassembled or contig data for the control sample (T1) set on either platform. The E-probes were mapped to the *S. enterica* genome and the open reading frames with Prokka annotation were correlated to the E-probes (Figures 2 & 3). No E-probes were able to be mapped to the *S. bongori* genome which was expected because the E-probes have been curated to only map to regions in *S. enterica* that are not found in *S. bongori.*

Comparison of Pipeline Profiles for *S. enterica* Detection

There is a consistent trend seen in all profiling pipelines. S1, the sample with the highest level of contamination, was consistently high for *S. enterica* across all pipelines, assemblies, and platforms. The percent of *S. enterica* in samples S2 and T1 were also similar in most of the pipelines.

Pipeline Speed Comparison

The difference in speed between the three profiling pipelines (BLAST-nt, Diamond and Kraken2/Bracken) was drastic for all samples (S1, S2 and T1). The BLAST-nt pipeline took 30,000 minutes for each sample (S1, S2, T1). The Diamond pipeline took 1440-120 minutes for each sample and the Kraken2/Bracken and Bowtie pipelines took 120 minutes. The EDNA pipeline took less than 5 min to detect *S. enterica* with 60nt and 80nt length E-probe sets. The speed was calculated after each pipeline was constructed from sample input to final output files.

Clean Read Verses Contigs *S. enterica* Detection

Contigs consistently resulted in less read/contig assignments and lower percentage of *S. enterica* in the S1 sample. The percentage abundances of *S. enterica* for samples S2 and T1 on both platforms were proportional with the clean reads. The EDNA pipeline showed that compared to clean reads the contigs resulted in negative detection at *S. enterica* lower titers (<1000 cells of *S. enterica*).

Discussion

S. enterica Detection Across Pipelines

This study focused on bioinformatic tools for detection of human foodborne pathogens in metagenomic sequence data. Using informative genes is popular in ecological surveys using whole genome metagenomic data because this type of analysis can be used to gain a greater understanding of the possible gene function in a metagenomic community, but it does not have enough taxonomic resolution to be used for detection at this time.

However, in this study there was not a notable difference in percent abundance of phyla between the Illumina and Roche 454 platforms. This supports the early work by Luo et al. who concluded that there was agreement between the Illumina and Roche 454 percent abundance profiles in freshwater metagenomic samples (Luo et al., 2012). This puts the burden of profile bias on the downstream analysis pipelines. Using non-informative methods means that a greater number kmers are available for analysis and therefore the greater number of regions likely results in the greater taxonomic clarity seen in the profiles from the non-informative datasets. Profiling methods like those used in the BLAST-nt, DIAMOND and Kraken2 pipelines, provide a more in-depth understanding of the microbial community and are able to estimate the abundance of all species with regard to the number of reads allocated to each taxon. The BLAST-nt pipeline tools are the oldest and most often utilized for sequence alignments. This tool was not originally intended for use with metagenomic data, it was used to find regions of local similarity between two sequences or small groups of sequences (Altschul, 1990). For this reason, the algorithms are not optimized for metagenomic data or unassembled reads. It is important to note that the profile of percent abundance of *S. enterica* was similar to the DIAMOND and Kraken2/Bracken and the standard deviation also was similar with respect to the type of data (Tables 2, 3, & 4). This indicates that there is a consensus between profiles and the database bias, while still present, is affecting the output of each pipeline in a similar way. In the community profiles, while there was not a consensus in species abundance, several species (*Pseudomonas* sp. *Raoultella* sp (formerly *Klebsiella*) and *Clavibacter michiganensis)* that were expected to be in the agricultural environment were found in high abundance across all profiling pipelines regardless of platform or assembly (supplemental). Compared to the Diamond pipeline the BLAST-nt pipeline resulted in more taxa

assignment for each sample even though the same scoring matrix was used. This is likely due to the fact that the Diamond uses translated protein coding sequences and not all kmers code for proteins. This limits the amount of data the Diamond pipeline sorted through compared to the BLAST-nt pipeline. Of the profiling pipelines, Kraken2 had the highest number of taxa assigned to the lowest percentage of *S. enterica*. This is likely due to the fact that since Kraken2 is designed for metagenomic datasets it assigns reads that match to more than one organism to the lowest common ancestor, meaning that read or contigs assigned to *S. enterica* by the other profiling pipelines are likely not represented at the species level due to overlapping between multiple taxa. The Bowtie2 alignment of the metagenomic clean reads and contigs to *S. enterica* provided not only a way to confirm the presence of *S. enterica* reads in the samples, but also a potential means for detection. This was done creating a custom dataset that contained only the target *S. enterica* genome. The creation of a limited custom database limits the targets that the data is aligned to. However, this neuters one of the strengths of the profiling method in that by reducing the number of targets, the ability to accurately predict the species abundance is limited. Additionally, this pipeline is also likely to mis-assign taxa as was seen when the reads were aligned to *S. bongori* and nearly identical percent abundances were observed (Table 5). In the case of detecting known pathogens, when a profile of the metagenomic community is not needed, methods like EDNA detection are preferable. Instead of detecting a target based on the percentage of reads over total assigned reads, EDNA gives a positive or negative result based on the number of E-probe matches set by the threshold parameters. EDNA only retrieves hits that match unique regions of the target sequence which alleviates the issue of false positives due to common gene regions. This is particularly important in detection of human foodborne pathogens because the fact that many human foodborne pathogens are

in high numbers in databases like NCBI increase the likelihood of a gene that is common among many bacterial species being taxonomically assigned to a human foodborne pathogen during profiling pipeline classification. EDNA also has the benefit of working optimally with unassembled Illumina data. Both contigs and clean reads from both platforms were run on the EDNA pipeline and it was found that EDNA was able to detect *S. enterica* at lower titers in the clean read data compared to the contig data. This is likely due to the fact that EDNA threshold was originally created for working with clean reads and the depth needed for a positive match was two or greater. Since the creation of contigs condense overlapping reads into a single sequence the depth needed for detection would need to be lowered and optimized for use with contigs.

Pipeline Speed

The BLAST-nt algorithm is too slow for detection application from complex metagenomic data sets. The Diamond pipeline was also not specifically built for metagenomic taxa assignment. It was designed to improve the speed of the original BLAST tool due to the way the algorithm partitions the data. The Diamond pipeline generated the profiles for all samples and platforms in less than 24 hours (Table 3). This speed is a great improvement on the BLAST tool, however, it is comparable to PCR and culture-based methods that are already well validated. The speed of the tool would need to be less than eight hours in order to improve upon other validated detection methods. The Kraken2/Bracken pipeline was created in order to assign taxa to metagenomic reads. Since the program was developed for this type of work, both the clean reads and contigs from both platforms were run in all three samples (Table 4). It was thought that perhaps using contigs would increase the speed of the pipeline, however, no increase in speed was found between the clean reads and contig *S. enterica* detection. Its overall percent abundance of

S. enterica in the clean read and contig data is similar, which was of concern because it was thought that the construction of contigs could obscure the percent abundance of some species due to the way the reads are combined into a single copy. Since the use of contigs did not increase the speed of the pipeline, it is probably unnecessary due to the time required to assemble the contigs after sequencing. The EDNA pipeline was the fastest of all the pipelines, requiring only 5 min to complete the detection of *S. enterica*. This difference in speed is likely due in part to the sizes of the databases used. For the BLAST-nt the entire non-redundant NCBI nt database was used in order to reduce potential database bias by only selecting for human foodborne pathogens or bacterial genomes. Similarity, the whole non-redundant NCBI protein database was used for the Diamond pipeline. The Kraken2 pipeline used a slightly smaller dataset and this was deemed appropriate because the Kraken2 pipeline is less susceptible to database bias because of the way that the algorithm assigns exact kmer matches to the lowest common ancestor. EDNA had the smallest database of all because it probed the metagenomic sample directly. EDNA is not subject to database bias because it uses unique target genome region for detection.

Limitations and Future Work

Like many metagenomic studies, there was not a "true" negative control (Miller, 2016). This was due to the fact that, while the Control sample (T1) was not spiked with *S. enterica*, there was no way to guarantee it was free of *S. enterica* reads. However, since culture based, and PCR methods did not detect the presence of *S. enterica* in the control sample this means that contamination was likely introduced after sampling for PCR and Culture based methods or due to the detection of nonliving cells. Assuming contamination or sample mislabeling occurred, it would have had to happen prior to sequencing, since samples were sent to separate facilitates for each platform and both platforms have similar

levels of contamination. It is also possible that a mistake was made when the data files were transferred from the sequencing facilities to the external storage hard-drive. The only thing that is certain is that no evidence of *S. enterica* was found prior to sequencing, but alignments show that the presence of *S. enterica* post sequencing. Because this was verified and accounted for, it does not affect the outcome of pipeline testing.

The Kraken2/Bracken and EDNA pipelines were the fastest at pathogen identification of the pipelines tested. However, they are difficult to directly compare due to the fact that EDNA currently lacks a quantitative capacity. Kraken2 also required substantial RAM to run part of the pipeline and was not as user-friendly as the other tree pipelines. EDNA also lacks a verified statistical test that would allow it to be compared to PCR for accuracy and precision. The evaluation of speed and detection among databases and bioinformatic pipelines provides a further understanding of the benefits and limitations of currently available methods of pathogen detection in metagenomic sequence data. Understanding potential pitfalls in bioinformatic analyses and the ability to optimize detection speed has important applications in the field of microbial forensics and biosecurity.

Based on this study, the effect of platform bias was minimal, but Illumina is better supported. The BLAST-nt and Diamond pipelines took longer to reach identification compared to Kraken2/Bracken and EDNA. Both Kraken2/Bracken are decent tools for detection from metagenomic sequence data. However, EDNA shows the most potential because of its ability to specify unique regions of the target genome. It will need to be optimized and the lack of quantitative capacity and statistical verification currently limit the ability of EDNA to be compared to the current PCR standards will need to be addressed by future studies.

This study had laid the groundwork for understanding the potential of current and emerging bioinformatic tools for human pathogen detection in complex metagenomic data. From this work, it is clear that tools developed for metagenomic analysis like Kraken2 and EDNA have the most potential for detection studies. With the strengths of Kraken2 being the generation of a metagenomic community profile and the weaknesses being the speed and lack of specificity, EDNA is a clear front runner for metagenomic detection studies, however, it will still need to be optimized to address the issues of low titer pathogen detection. These optimizations include increasing the sensitivity through assessing the E-value and percent ID thresholds, exploration of different E-probe lengths and lowing of the read depth.

LITERATURE CITED

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

Avaniss-Aghajani, E., Jones, K., Chapman, D., & Brunk, C. (1994). A molecular technique for identification of bacteria using small subunit ribosomal RNA sequences. *BioTechniques*, *17*(1), 144-6.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P. & Rohwer, F.

(2003). Metagenomic Analyses of an Uncultured Viral Community from Human Feces. J Bacteriol 185, 6220-6223.

Buchfink B, Xie C, Huson DH, "Fast and sensitive protein alignment using DIAMOND", *Nature Methods* 12, 59-60 (2015). doi:10.1038/nmeth.3176

Carr, Rogan & Elhanan Borenstein. (2014). Comparative analysis of functional metagenomic annotation and the mappability of short reads. *PLoS ONE, 9*(8), E105776.

Centers for Disease Control and Prevention. (2016). List of selected multistate foodborne outbreak investigations. Available at: http://www.cdc.gov/foodsafety/outbreaks/multistate-outbreaks/multistate-list.html. Accessed 6.10.17.

Chattaway, M., Schaefer, U., Tewolde, R., Dallman, T., & Jenkins, C. (2017). Identification of Escherichia coli and Shigella Species from Whole-Genome Sequences. *Journal of Clinical Microbiology, 55*(2), 616-623.

Das, Samarendra, Rai, Anil, Mishra, D.C., & Rai, Shesh N. (2018). Statistical approach for selection of biologically informative genes. *Gene,655*, 71-83.

Dos Santos, D., Istvan, P., Quirino, B., & Kruger, R. (2017). Functional Metagenomics as a Tool for Identification of New Antibiotic Resistance Genes from Natural Environments. *Microbial Ecology, 73*(2), 479-491.

Espindola, A., Schneider, W., Hoyt, P. R., Marek, S. M., & Garzon, C. (2015). A new approach for detecting fungal and oomycete plant pathogens in next generation sequencing metagenome data utilising electronic probes. *International journal of data mining and bioinformatics*, *12*(2), 115-128.

FDA, "Bacteriological Analytical Manual (BAM), U.S. Food and Drug Administration," 2013.http://www.fda.gov/Food/FoodScienceResearch/LaboratoryMetods/BacteriologicalAnalyticalManualBAM/ucm2006 949.htm

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W. & DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters. Proc. Natl. Acad. Sci. U. S. A. 105, 3805-3810.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M. & Nelson, K. E. (2006). Metagenomic Analysis of the Human Distal Gut Microbiome. Science 312, 1355-1359.

Golob, J., Margolis, E., Hoffman, N., & Fredricks, D. (2017). Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. *BMC Bioinformatics, 18*(1), 1-12.

Hirano, S.S., and C.D. Upper. (1983). Ecology and epidemiology of foliar bacterial plant pathogens. Annu. Rev. Phytopathol. 21:243–70.

Jones, M., & Blaxter, M. (2013). AfterParty: Turning raw transcriptomes into permanent resources. *BMC Bioinformatics, 14*, 301.

Jung, Y., Jang, H., & Matthews, K. R. (2014). Effect of the food production chain from farm practices to vegetable processing on outbreak incidence. *Microbial Biotechnology*, *7*(6), 517–527. http://doi.org/10.1111/1751-7915.12178

Karlsson, O. E., Hansen, T., Knutsson, R., Löfström, C., Granberg, F., & Berg, M. (2013). Metagenomic detection methods in biopreparedness outbreak scenarios. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, *11*(S1), S146-S157.

Klein, H., Bartenhagen, C., Kohlmann, A., Grossmann, V., Ruckert, C., Haferlach, T., & Dugas, M. (2011). R453Plus1Toolbox: An R/Bioconductor package for analyzing Roche 454 Sequencing data. *Bioinformatics, 27*(8), 1162-1163.

Knight, R., Vrbanac, A., Taylor, B., Aksenov, A., Callewaert, C., Debelius, J., . . . Dorrestein, P. (2018). Best practices for analysing microbiomes. *Nature Reviews. Microbiology, 16*(7), 410-422.

Kolde, Ravio & Jaak Vilo. (2015). GOsummaries: An R Package for Visual Functional Annotation of Experimental Data [version 1; referees: 2 approved]. *F1000Research, 4*, 574.

Lawrence Berkeley National Laboratory, & United States. Department of Energy. Office of Scientific Technical Information. (2010). *Analysis of Illumina Microbial Assemblies*. Berkeley, Calif. : Oak Ridge, Tenn.: Lawrence Berkeley National Laboratory ; distributed by the Office of Scientific and Technical Information, U.S. Dept. of Energy.

Liu, Geng, Kou, Xin, & Yang. (2018). Engineering nanomaterials-based biosensors for food safety detection. *Biosensors and Bioelectronics, 106*, 122-128.

Liu, Y., Guo, J., Hu, G., & Zhu, H. (2013). Gene prediction in metagenomic fragments based on the SVM algorithm. *BMC Bioinformatics, 14*(Suppl 5), S12.

Lund, B. M., & O'Brien, S. J. (2011). The Occurrence and Prevention of Foodborne Disease in Vulnerable People. *Foodborne Pathogens and Disease*, *8*(9), 961–973. http://doi.org/10.1089/fpd.2011.0860

Luo, D., Ziebell, S., & An, L. (2017). An informative approach on differential abundance analysis for time-course metagenomic sequencing data. *Bioinformatics, 33*(9), 1286-1292.

Martins de Sá, Contreras, & Cordeau. (2015). Exact and heuristic algorithms for the design of hub networks with multiple lines. *European Journal of Operational Research, 246*(1), 186-198.

Miller, R., Uyaguari-Diaz, M., McCabe, M., Montoya, V., Gardy, J., Parker, S., Patrick, D. (2016). Metagenomic Investigation of Plasma in Individuals with ME/CFS Highlights the Importance of Technical Controls to Elucidate Contamination and Batch Effects. *PLoS One, 11*(11), E0165691.

Mitra, S., Schubach, M., & Huson, D. (2010). Short clones or long clones? A simulation study on the use of paired reads in metagenomics. *BMC Bioinformatics, 11*(Suppl 1), S12.

Nagarajan, K., & Loh, K. (2014). Molecular biology-based methods for quantification of bacteria in mixed culture: Perspectives and limitations. *Applied Microbiology and Biotechnology, 98*(16), 6907-6919.

Nakamura, S., Yang, C. S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., ... & Mizutani, T. (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PloS one*, *4*(1), e4219.

Nakamura, S., Nakaya, T., & Iida, T. (2011). Metagenomic analysis of bacterial infections by means of high-throughput DNA sequencing. *Experimental Biology and Medicine*, *236*(8), 968-971.

Pookhao, N., Sohn, M., Li, Q., Jenkins, I., Du, R., Jiang, H., & An, L. (2015). A two-stage statistical procedure for feature selection and comparison in functional analysis of metagenomes. *Bioinformatics, 31*(2), 158-165.

Popic, V., Kuleshov, V., Snyder, M., & Batzoglou, S. (2018). Fast Metagenomic Binning via Hashing and Bayesian Clustering. *Journal of Computational Biology, 25*(7), 677-688.

Priyanka, B., Rajashekhar K. Patil, and Sulatha Dwarakanath. "A Review on Detection Methods Used for Foodborne Pathogens." *The Indian Journal of Medical Research* 144.3 (2016): 327–338. *PMC*. Web. 22 June 2017.

Santamaria, Monica, Bruno Fosso, Arianna Consiglio, Giorgio De Caro, Giorgio Grillo, Flavio Licciulli, Sabino Liuni, Marinella Marzano, Daniel Alonso-Alemany, Gabriel Valiente, Graziano Pesole; Reference databases for taxonomic assignment in metagenomics, *Briefings in Bioinformatics*, Volume 13, Issue 6, 1 November 2012, Pages 682–695, https://doi.org/10.1093/bib/bbs036

Schwartz, & Beaver. (2011). Evidence of a gene × environment interaction between perceived prejudice and MAOA genotype in the prediction of criminal arrests. *Journal of Criminal Justice, 39*(5), 378-384.

Singh, Chandra, Guhathakurta, Sinha, Chatterjee, Ahmed, . . . Rajamma. (2013). Genetic association and gene–gene interaction analyses suggest likely involvement of ITGB3 and TPH2 with autism spectrum disorder (ASD) in the Indian population. *Progress in Neuropsychopharmacology & Biological Psychiatry, 45*, 131-143.

Stobbe, A. H., Daniels, J., Espindola, A. S., Verma, R., Melcher, U., Ochoa-Corona, F., ... & Schneider, W. (2013). E-probe Diagnostic Nucleic acid Analysis (EDNA): a theoretical approach for handling of next generation sequencing data for diagnostics. *Journal of microbiological methods*, *94*(3), 356-366.

Stobbe, A. H., Schneider, W. L., Hoyt, P. R., & Melcher, U. (2014). Screening metagenomic data for viruses using the e-probe diagnostic nucleic acid assay. *Phytopathology*, *104*(10), 1125-1129.

Li, Ruiqiang, Li, Yingrui, Kristiansen, Karsten, & Wang, Jun. (2008). SOAP: Short oligonucleotide alignment program. *Bioinformatics,24*(5), 713-714.

Uchiyama, Naito, Yagi, Mizushima, Higashimura, Hirai, . . . Uchiyama, Kazuhiko. (2015). Peptidomic Analysis via One-Step Direct Transfer Technology for Colorectal Cancer Biomarker Discovery. *Journal of Proteomics & Bioinformatics,* (S5), 1.

USDA, FSIS. 2007. "FSIS Procedure." *Laboratory Guidebook MLG 8A.03*.

Wood DE, Salzberg SL: Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 2014, 15:R46.

Wylezich, C., Beer, M., Höper, D., & Papa, A. (2018). A Versatile Sample Processing Workflow for Metagenomic Pathogen Detection. *Scientific Reports, 8*(1).

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., ... & Wang, J. (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *Journal of clinical microbiology*, *49*(10), 3463-3469.

Yu, Zhai, Bie, Lu, Zhang, Tao, Zhao. (2016). Survey of five food-borne pathogens in commercial cold food dishes and their detection by multiplex PCR. *Food Control, 59*, 862-869.

TABLES

Table 1) Summary of sequencing and clean data output, number of cleans reads per sample, average read length per sample, number of contigs per sample.

Sequncing	Samples	# Raw Reads	# Clean Reads	% Clean Read Rate	Avg Read Length	# Contigs	% Aligned Reads	Average Length
Illumina Hiseq Paired-end	S1	24000000	23790000	99.94	100	25805	50	101,310
	S2	24000000	23920000	99.7	100	44903	43	54587
	T1	24000000	23910000	99.62	100	40734	42	53557
Roche 454 JR	S1	217102	135631	62	441	10942	34	730
	S2	231078	109647	47	415	6734	27	722
	T1	231469	120211	52	398	5173	19	713

Table 2) Summary of the BLAST pipeline took. The pipeline took 500hrs (30,000 min).

Sample	Illumina						Roche 454					
	Clean Reads			Contigs			Clean Reads			Contigs		
	Total Read #	Species Read #	Reads Assigned *S. enterica*	Total Contig #	Species Contig #	Reads Assigned *S. enterica*	Total Read #	Species Read #	Reads Assigned *S. enterica*	Total Contig #	Species Contig #	Reads Assigned *S. enterica*
S1	23,719,434	1825134	584042(32)	25805	25657	5387(21)	135631	29357	8807(30)	10942	10617	2654(25)
S2	23,927,370	2071615	12429(0.6)	44903	44572	267(0.6)	109647	10015	40(0.4)	6734	6470	19(0.3)
T1	23,877,420	641135	1282(0.2)	40734	40053	80(0.2)	120211	13671	68(0.5)	5173	4977	11(0.22)

Table 3) Summary of the DIAMOND pipeline. The pipeline took 24 hrs (1,440 min).

Sample	Illumina						Roche 454					
	Clean Reads			Contigs			Clean Reads			Contigs		
	Total Read #	Species Read #	sp. Reads Assigned *S. enterica*	Total Contig #	Species Contig #	sp. Contigs Assigned *S. enterica*	Total Read #	Species Read #	sp. Reads Assigned *S. enterica*	Total Contig #	Species Contig #	sp. Contigs Assigned *S. enterica*
S1	23,719,434	1056324	602105(57)	25805	25657	10519(41)	135631	21155	11846(56)	10942	10617	5733(54)
S2	23,927,370	1371428	21942(1.6)	44903	44572	713(1.6)	109647	9893	188(1.9)	6734	6470	110(1.7)
T1	23,877,420	602124	5419(0.9)	40734	40053	320(0.8)	120211	12261	196(1.6)	5173	4977	80(1.6)

Table 4) Summary of the Kraken2 pipeline. The pipeline took 240 min.

Sample	Illumina						Roche 454					
	Clean Reads			Contigs			Clean Reads			Contigs		
	Total Read #	Species Read #	Reads Assigned *S. enterica*	Total Contig #	Species Contig #	Reads Assigned *S. enterica*	Total Read #	Species Read #	Reads Assigned *S. enterica*	Total Contig #	Species Contig #	Reads Assigned *S. enterica*
S1	23,719,434	5561946	1,001,150(18)	25805	25657	5382(20)	135631	31357	4076(13)	10942	10617	1380(13)
S2	23,927,370	7002386	28,009(0.4)	44903	44572	178(0.4)	109647	11013	33(0.3)	6734	6470	19(0.3)
T1	23,877,420	861387	861(0.1)	40734	40053	40(0.1)	120211	13381	40(0.3)	5173	4977	14(0.3)

Table 5) Summary of the Bowtie2 pipeline. The pipeline took 240 min.

Sample	Genome	Illumina				Roche 454			
		Clean Reads		Contigs		Clean Reads		Contigs	
		Total Read #	*# S. enterica* reads	Total Contig #	*# S. enterica contigs*	Total Read #	*# S. enterica* reads	Total Contig #	*# S. enterica contigs*
S1	*S. enterica*	23,719,434	1,028,973(4%)	25805	5380(20%)	135631	4060(30%)	10942	1380(13%)
S2	*S. enterica*	23,927,370	17,624(0.07%)	44903	174(0.39%)	109647	33(0.03%)	6734	19(0.3%)
T1	*S. enterica*	23,877,420	5167(0.02%)	40734	40(0.1%)	120211	40(0.03%)	5173	14(0.3%)
S1	*S. bongori*	23,719,434	1,028,733(4%)	25805	5380(20%)	135631	4060(30%)	10942	1380(13%)
S2	*S. bongori*	23,927,370	17,610(0.07%)	44903	174(0.39%)	109647	33(0.03%)	6734	19(0.3%)
T1	*S. bongori*	23,877,420	5163(0.02%)	40734	40(0.1%)	120211	40(0.03%)	5173	14(0.3%)

Table 6) Summary of the EDNA pipeline detection of *S. enterica*. The pipeline took 5 min.

Sample	Illumina						454					
	Read #	Hit #		Contig #	Hit #		Read #	Hit #		Contig #	Hit #	
		60nt	80nt		60nt	80nt		60nt	80nt		60nt	80nt
S1	23,719,434	623	159	25805	212	32	135631	146	40	10942	30	5
S2	23,927,370	27	4	44903	0	0	109647	0	0	6734	0	0
T1	23,877,420	0	0	40734	0	0	120211	0		5173	0	0

FIGURES

Figure 1) Overview of pipeline workflow. Each pipeline's speed is an estimation of the workflow between the gray areas without interruption.

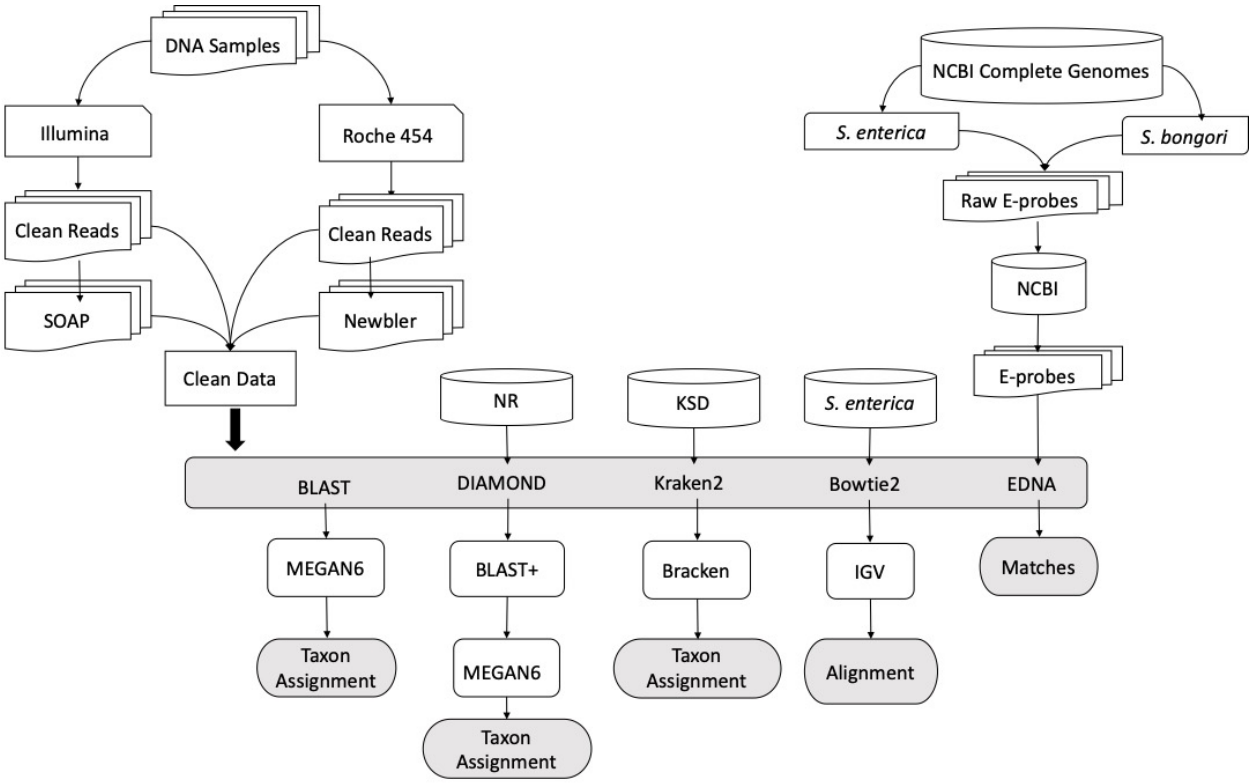


Figure 2) 60nt and 80nt E-probes mapped to an *S. enterica* genome using the CGView Server.

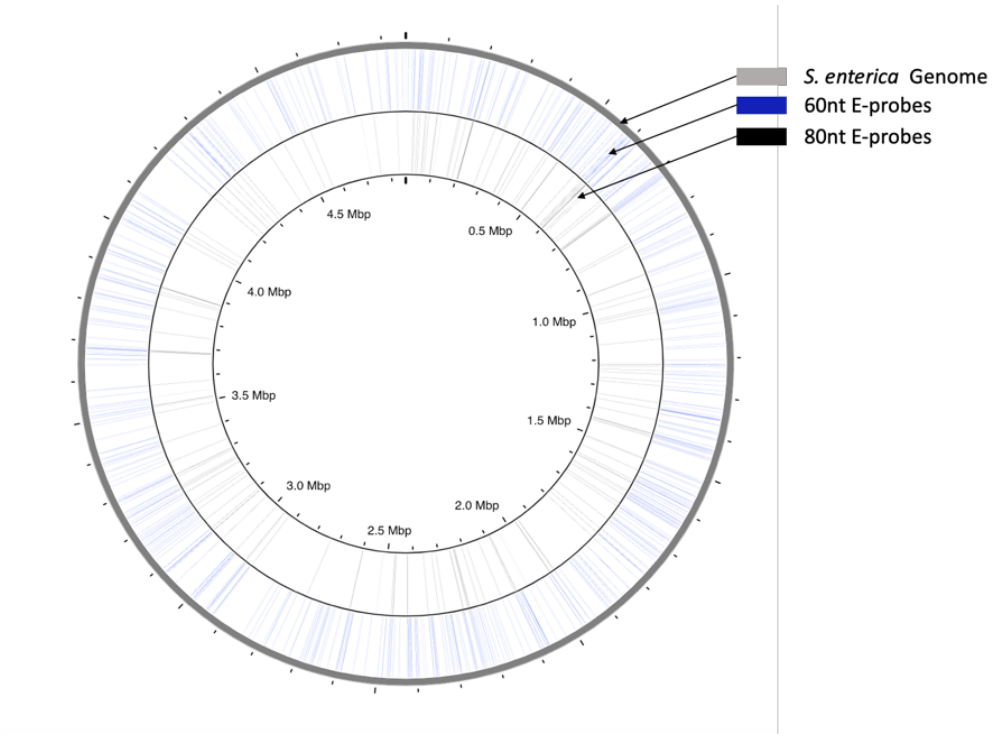
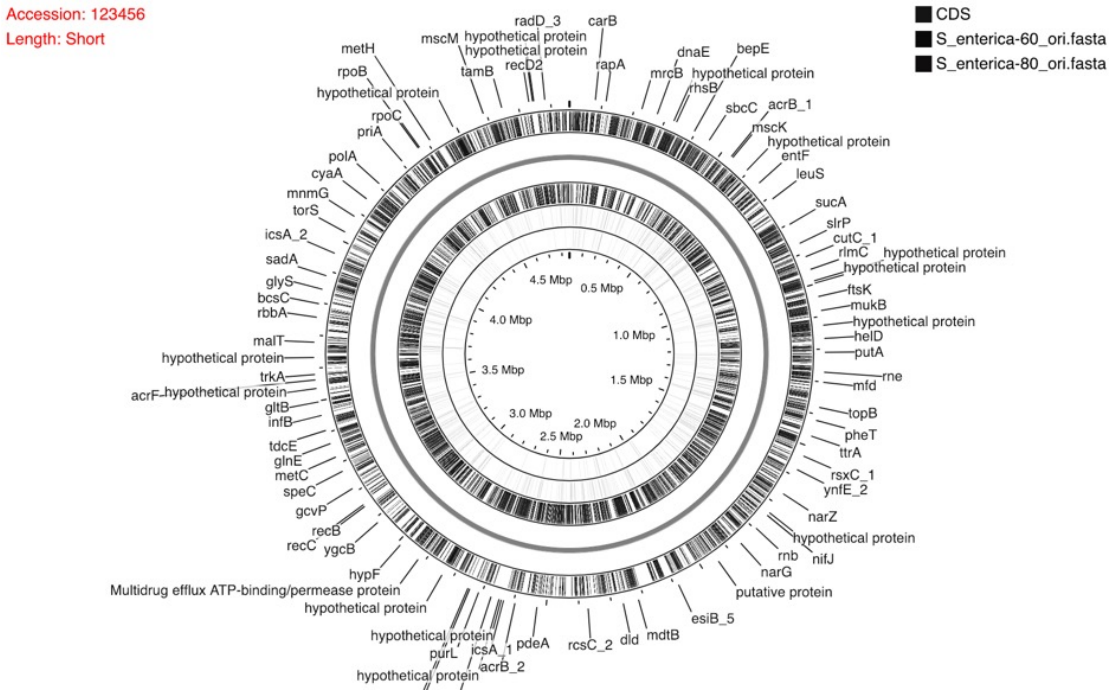


Figure 3) *S. enterica* genome with mapped 60nt and 80nt E-probes. Shown with Prokka annotation (CDS). Created on the CGView Server.



CHAPTER IV

OPTIMIZATION OF EDNA FOR HUMAN FOODBORNE PATHOGEN DETECTION IN COMPLEX METAGENOMIC DATA

.**Abstract**

Aim: The objective of this study is to optimize E-probe Diagnostic Nucleic-acid Analysis bioinformatics pipeline for rapid and sensitive detection of human foodborne bacterial pathogens in metagenomics datasets.

Materials and Methods: *In silico* complex metagenomic datasets were constructed in Illumina sequencing format using MetaSim. The datasets consisted of tomato genome (host plant) and ten species of bacteria commonly present on fresh tomato surfaces to serve as background in addition to *Salmonella bongori* (Inclusivity/Exclusivity determinate). *Salmonella enterica* was spiked at five concentrations with ten replicates for each concentration in these datasets. The laboratory samples consisted of DNA extracted from the surface washes of commercial tomatoes spiked with *Salmonella enterica* at two concentrations and sequenced using the Illumina platform. E-probe sets were constructed to test the impact of the E-probe length (60nt, 80nt and 100nt) and E-

value ($1x10^{-3}$, $1x10^{-6}$, $1x10^{-9}$) parameters. During the EDNA detection pipeline, the query coverage (90, 95 and 100) and percent identity (90, 95 and 100) were used to establish the sensitivity and specificity thresholds of the E-probe sets. **Results:** It was found that using unassembled Illumina data was the optimal data input. E-probes of 80nt lengths and curated with an E-value of $1x10^{-3}$ where able to detect *Salmonella enterica* when it made up at least 0.0018% of the metagenomic dataset. The optimal parameters for detection were a query coverage of 90% and a percent identity of 95%.

Significance and Impact of the study: E-probe Diagnostic Nucleic-acid Analysis (EDNA) is a targeted detection *in silico* probe-based bioinformatic pipeline that has the potential to quickly detect any pathogen present in a single complex sample through metagenomic data mining.

Introduction

Metagenomics emerged first as a new approach for genomic analysis in the field of ecology, where it was necessary to sequence a whole community of organisms in order to gain insight about community structure and function. Since many of the organisms in environmental samples are not culturable or known, it was not possible to observe all of the members or potential gene interactions *in situ* in an environmental community by culture methods before the metagenomic breakthrough. Metagenomic sequencing allows direct genetic analysis of a complex environmental sample (Karlsson et al., 2013). Using this method for detection streamlines the identification process by removing the need for culturing (Nakamura, 2009; Nakamura, 2011). While current metagenomic studies have primarily focused on profiling microbial communities in a sample, this approach has the

potential to detect any and all microbes in a given sample including pathogens (Stobbe et al., 2013; Yang 2011). A metagenomic approach has been used to detect previously unknown organisms and viruses in a variety of hosts, including mammals, insects, and plants through community profiling (Adams et al., 2009, Cox-Foster et al., 2007, Palacios et al., 2008; Adams et al., 2009, Roossinck et al., 2010). However, community profiling of metagenomics sequences is time demanding and computationally intensive and may lack the specificity needed to differentiate between closely related pathogenic and non-pathogenic organisms. For example, from previous work it was found that using BLAST under standard parameters took >500 hours on a high-performance computer to complete (My Paper1). Also, all of the community profiling pipelines analyzed detected target in the negative control samples which had been confirmed as negative for the target through plate streaking and PCR. This is likely due to the community profiling algorithms not having enough specificity when designating taxon. For detection of a particular target microorganism in a sample through metagenomics approach, it is not necessary to know the composition of an entire community, instead, the focus could be only on the sequences or signatures of the target microorganism during bioinformatics analysis which would reduce the intensive computational component of community profiling. This is a targeted detection from complex metagenomic samples and has been used to identify viral pathogens from clinical samples in order to rapidly and accurately identify human pathogenic viruses for outbreak investigations (Yang, 2011). Using metagenomic approach the samples were sequenced on the Illumina platform, and the Bowtie program was used to align the reads to the NCBI non-redundant nucleotide database and MEGAN was used to assign the alignments to the lowest common ancestor and the reference sequences were used to assemble the viral genomes. The limitation of this method for pathogen detection

was the amount of sample available from each individual which could have biased some of the samples since individual were only sampled once. The assembly of the pathogen is also unnecessary for detection and increases the computational load of the pipeline. However, because of the success of these methods and the ever-lowering price of next generation sequencing technologies (NGS), detection of foodborne pathogens through metagenomic sequencing has now become a possibility (Nakamura, 2011). However, the pipelines necessary to analyze this type of metagenomic data have not been fully elucidated.

In our previous study the potential of community profiling pipelines, custom databases and targeted region detection were examined for their use in bacterial human food pathogen detection. For community profiling it was found that the most commonly used pipelines like the Basic Local Alignment Search Tool are too computationally intensive to be used with complex metagenomic samples because of the time requirement >500 hours (NCBI 2017, Chapter III). Even pipelines such as Kraken that have been developed specifically for rapid assigning taxa to metagenomic read data have issues with specificity (Wood, 2014). Kraken consistently assigned reads as *Salmonella enterica* when this was not backed up by culture method or PCR as did the Bowtie pipeline using custom databases. The lack of specificity is a critical issue in pathogen detection because pathogenic and non-pathogenic species are often separated at the species level or below. From these studies it was found that only the targeted region detection pipeline E-probe Diagnostic Nucleic-acid Analysis (EDNA) met both the time and specificity criteria to move on to an optimization phase for human food borne pathogen detection.

E-probe Diagnostic Nucleic-acid Analysis (EDNA) is a tool developed at Oklahoma State University in conjunction with the United State Department of Agriculture

(USDA) to bridge the gap between profiling-based methods and diagnostically realistic time requirements. This method is inspired by the Tool for Oligonucleotide Fingerprint Identification (TOFI) method of simulated *in silico* microarray (Satya et al., 2008). Like TOFI, EDNA uses *in silico* probes creation but does not require the thermodynamic optimization and makes the probes compatible with metagenomic data by using the probes as search queries in BLAST. EDNA was originally utilized to detect plant pathogens, but it is also ideal for detection of human foodborne pathogens in unassembled metagenomic sequence data. This method works by creating electronic probes (E-probes) based on a known pathogen sequence (Stobbe, 2013; Stobbe, 2014). Target specific E-probes are created by choosing the genome of target organisms and aligning it to a closely related genome which will act as the inclusivity/exclusivity determinate using MUMmer for pairwise comparison (Delcher et al., 2002). The MUMmer program is used to find and identify the maximal matches in the global alignment of the two genomes and eliminate regions of similarity. The output is lengths of the target genome that do not overlap with the inclusivity/exclusivity determinate genome. The regions of the target genome are then shredded into E-probes using BioPerl scripts (Staijch et al., 2002). Based on the chosen length, longer regions are partitioned, and shorter regions are discarded. Previous studies have suggested that the size of the target genome and the similarity of the inclusivity/exclusivity determinate influence the number and size of unique regions and is likely to be greater in larger genomes (Stobbe et al., 2013). This has presented a problem in eukaryotic genomes where the larger number of genomes generated inhibit the speed of the downstream BLAST (Altschul et al., 1990). To get around this issue, previous studies suggested choosing an E-probe length of 60-80nt, although it is possible to make E-probes between 20-120nt (Stobbe et al, 2013). The E-probes are then curated by mapping them to

the nucleotide database of NCBI using BLAST. This is a critical step because even though only unique regions of the target were chosen compared to the closely related genome, it is likely that other genomes in the database share common or similar regions that can result in false positives in metagenomic data which is known as background noise. It is ideal to curate the E-probe set with as much of the potential background genomes as possible in order to reduce the background noise which is particular concern in metagenomic samples. The E-value is used to evaluate the E-probes and the lower the E-value is set the more E-probes will be removed. In BLAST the E-value is an estimate of how many times a nucleotide alignment score is expected to occur by random chance. After the BLAST, the curated E-probe library is ready to use for detection applications. The E-probe sets can be used simultaneously on the same sample potentially allowing for simultaneous detection of multiple pathogens (Geyer et al., 2008).

The matches are parsed based on the combined score of the Query Coverage (QC) and Percent Identity (%ID). The QC is a threshold parameter based on the number of nucleotides that have to match in order for the alignment of the E-probe read to be reported. The %ID is simultaneously measured and is similar to the match/mismatch parameter of BLAST. The %ID establishes a baseline percentage of nucleotides that have to be identical given a particular alignment length (QC). Both QC and %ID are calculated as percentages and reported as the "Score". The final step is the diagnostic call. In almost all diagnostic pipelines, it is assumed that there will be some level of false positives/false negatives and EDNA is no different. Several different strategies for determining the statistical relevance of the diagnostic call have been used. The first way is by running Decoy E-probes alongside the E-probes in every sample. A Decoy E-probe is created for each E-probe by using the reverse sequence of the E-probe. The Decoy E-probes are then subjected to the same E-

value, QC and %ID thresholds that the E-probes are. Then a Student's t Test is completed for the Score between each E-probe and Decoy E-probe couple. Since it is assumed that the Decoy E-probe sequence is not diagnostically relevant then the Score of positive E-probes should be significantly higher than the Score of the Decoy E-probe. This method was suggested because many of the metagenomic studies are not able to ensure that a metagenomic sample is a true negative control. When true negative and positive controls are available, it is possible to set the threshold for positive diagnostic calls relative to the difference in Score between the hits in the negative control versus the positive control. Not only is this method more diagnostically relevant since it is not assuming the absence of the Decoy E-probe sequences in the data, it is also computationally simpler. The only issue is obtaining or creating a relevant true negative control for each metagenomic sample set. The simplest and most cost-effective way to create a true negative is through the creation of *in silico* mock metagenomic datasets. In previous studies, the datasets have been created using the MetaSim program (Richter et al., 2008; Stobbe et al., 2012).

Metagenomic mock datasets are simulations of real environmental data (Richter et al., 2008). These datasets are key in uncovering the limitations of currently available metagenomic data analysis tools because they offer a way to test the output results against the inputs of an experiment (Richter, 2008). This has been a major problem in the evaluation of tools for metagenomic analysis, because due to the nature of environmental samples, the inputs are variable and exact quantities are unknown (Korem et al., 2015). Mock datasets allow for the creation of true positive and negative controls, something that is not possible in strict experiments using only metagenomic data from environmental samples. Without the use of true positive and negative samples, the experimental design is flawed, and conclusions derived from the study can be brought into question (Stobbe et al.,

2012). This is not to say that mock datasets are a complete substitute for real environmental data sets, only that they are a resource that can be utilized for the testing of metagenomic analysis tools in order to better understand the outputs from studies with metagenomic data.

There are two main types of metagenomic mock datasets. The first type called an *in vitro* mock community dataset, is constructed by placing organisms in a simulated community before extracting the DNA or genetic material and sequencing the community (Fouhy, 2016; Fausser, 2011). This type of mock community is defined as a mixture of microbial cells, viruses or nucleic acids that were created *in vitro* to provide a simulation of the composition of a microbial sample (Castelino, 2014). This is considered a synthetic or laboratory mock community because it is not a community derived from a real environmental sample. However, this type of dataset is only an estimation of the community structure found in environmental metagenomic datasets and cannot completely replicate the relationships between community members (Wu, 2016). It should also be noted that since the community structure is calculated prior to sequencing, the actual amount of members is somewhat variable, due to extraction and sequencing errors (Miller, 2017).

The second type of mock metagenomic dataset is derived from *in silico* modeling that has been used to analyze programs in computer science (Richter et al., 2008). Many fields are now using these statistical and computer based *in silico* models to evaluate and optimize products and tools before implanting them in further studies. These are known as *in silico* mock metagenomomic datasets. This type of dataset uses sequencing data and genomes from databases like NCBI. The quality of the sequencing and genome completeness is analyzed prior to incorporation of each genome into the datasets. This allows stricter calculations of detection limits and specificity compared to other methods

where levels could be confounded by pre-analysis errors. MetaSim was one of the most successfully used open access metagenomic data simulators available (Richter, 2008). MetaSim allows for common errors based on sequencing platform to be incorporated into the datasets in order to more realistically simulate a metagenomic data (NIH web). This software works by generating collections of synthetic reads from specifically chosen genomes. The genome's representation, as well as, the number of reads from each genome can be designated during the taxon profile phase. The program then generates mate pairs based on platform models. In addition to providing more control on the mock community genome inputs, the cost of constructing an *in silico* mock metagenomic data set is minimal compared to other experiments that require extraction and sequencing. This is one reason why many fields including food chemistry have started regularly using *in silico* modeling for optimization studies (Lambert, 2012). This method also provides research at facilities that are not equipped to handle live human pathogens with the ability to conduct preliminary experiments containing sequence data from human pathogens without containment or health risks. The metagenomic analysis tools can then be evaluated by comparing the input data to the output data (Blagden, 2016). Like all modeling-based experiments, the tools used will then need to be validated using real metagenomic data from environmental and laboratory samples, because nothing can replace the use of real environmental data.

Both *in vitro* and *in silico* mock metagenomic data types are extremely useful in understanding how metagenomic analysis tools process and profile data. These tools are extremely important because completing metagenomic studies without an understanding of the biases and detection limits of the tools, can result in errors. If erroneous conclusions

are made about metagenomic dataset due to the use of unvalidated tools, the understand of metagenomic community structure can be obscured.

In this study, EDNA was optimized for the detection of the model human foodborne pathogen *S. enterica*. The sensitivity and specificity thresholds and parameter optimization were tested first using *in silico* unassembled Illumina metagenomic samples constructed form NCBI genomes and then validated laboratory metagenomic samples spiked with the target pathogen.

Material and Methods

General Experiment Design

The experimental design included three complex metagenomic datasets from laboratory samples and six *in silico* complex metagenomic datasets with ten replications at each concentration (Figure 1). The laboratory samples were prepared and sequenced into Illumina unassembled metagenomic databases. The *in silico* mock databases from NCBI genomes including background host (*Solanum lycopersicum*), I/E genome, the top ten bacterial species from previously sampled and profiled communities and target pathogen genome into the MetaSim program where five mock dilutions of target and negative control databases were constructed. An E-probe set was constructed, and detection parameters identified for testing.

MetaSim Database Construction

The *in silico* metagenomic mock Illumina datasets were constructed to simulate massively parallel Illumina sequencing using the MetaSim program (Satya et al., 2008). In

order to simulate the complex background found in real metagenomic community samples, the genomes from the top ten bacterial species identified across the previous metagenomic community studies were extracted from the NCBI genome database (NCBI Accession #CP001191.1, NC_002947.4, NC_007005.1, NC_010407.1, NC_014121.1, NC_016830.1, NC_016845.1, NZ_CP007557.1, NZ_CP016889.1, NZ_LN907827) along with chromosome one of *Solanum lycopersicum* (NCBI Accession #CM001064.3) to further mimic the real metagenomic profiles. The inclusivity/exclusivity determinate genome (I/E) that was used to construct the E-probes was also included in the mock datasets to determine the specificity of the E-probe hits (NCBI Accession #CP006692.1). Six mock datasets were constructed including a negative control (Table 1). Each dataset was made to simulate a dilution of target pathogen in the complex metagenomic background. The dilutions were chosen by calculating the ratios between background community and the target in previously profiled metagenomic communities. Based on previous metagenomic reads of laboratory samples, each dataset contains 24,000,000 reads of 100 bps and the dilutions range from 0.00018-0.18 percent of target pathogen in the community which was the equivalent of 1-1,000 cells of *S. enterica*. Ten replicate databases were constructed for each dilution and negative control.

Laboratory Metagenomic Datasets

Fresh Roma tomatoes were purchased from local commercial retailer located in Stillwater, OK. The tomatoes were spiked with *Salmonella* at 10^6 cell/tomato (S1), 10^3 cells/tomato (S2), and un-spiked control (T) inside the biosafety cabinets and left for air drying. For each treatment 27 tomatoes (9 tomatoes in three replicates) were taken; briefly, three tomatoes were placed in the stomacher bag containing 100 ml of UPB broth. To wash the bacteria/ native microflora from the surface of the tomatoes, 3 tomatoes were placed in the

stomacher bags, shaken manually for 1 min, rubbing each tomato for 2 min, again shaking for 1 min. These tomatoes were removed, and another 3 tomatoes were placed in the same bag, washed in the similar way, removed, and another 3 tomatoes were washed in the same wash fluid. A total of 9 tomatoes were washed in same 100 ml of UPB broth. A total of 300 ml of wash fluid was collected for each treatment. Total DNA was extracted using the traditional method of DNA extraction, briefly- A total of 300 ml of the wash fluid from each of the treatment above was divided into 150 ml each in centrifuge bottles and centrifuged at 10,000 rpm for 50 mins. The pellet in each was removed by dispensing in 1ml of lysis solution [25 mM Tris, 10 mM EDTA and lysozyme (20 mg/ml)] and incubated at 37 °C for 1 h. Sixty microliter of 10% SDS was added, and the mix was incubated at 56 °C for 30 min, followed by 2.5 µl of RNase A (20 m/ml) incubating at 37 °C for 30 mins, further 10 µl of Proteinase K (20 mg/ml; Promega) treatment was given at 56 °C for 30 mins. To the above lysate equal volume of phenol: chloroform: isoamyl alcohol (25:24:1, Sigma Aldrich, St. Louis, MO, USA) was added, mixed and centrifuged at 12,000 rpm for 15 mins. The above layer was removed carefully and extracted with equal volume of chloroform: isoamyl alcohol (24:1) and centrifuged again at 12,000 rpm for 15 mins. The supernatant was carefully separated, and the DNA was precipitated by adding 1/10 volume of sodium acetate (pH=5.2) and 2 volume of 100% chilled ethanol. The mix was precipitated by overnight incubation at -20°C. The pellet was finally collected by centrifugation at 12,000 rpm for 15 mins, washed twice with ice cold 70% ethanol, air dried in biosafety cabinet and finally, the pellet was dissolved in 50 µl of TE buffer. For the high-quality DNA for the Illumina run, the DNA was further purified and concentrated using the Zymo Research DNA clean and Concentrator kit (Zymo Research, Irvine, CA USA).

For each treatment, DNA from 3 rounds of extractions were pooled together to get the desired concentration for the Illumina run.

E-probe Construction

A complete *S. enterica* genome (NCBI Accession #NC_003198.1) was downloaded from NCBI and the inclusivity/exclusivity determinate genome (NCBI Accession #CP006692.1) was also downloaded from NCBI. The MUMmer program was used to find the optimal global alignment between the genomes and the genome regions unique to *S. enterica* were binned. During the MUMmer alignment the maximum number of gaps was equal to zero and the minimal length of alignment was 15nt. The BioPerl program was used to divide the binned sequences into seven different E-probe sets with set lengths (60nt, 80nt, 100nt).

The E-probes were then curated by mapping them to the NCBI non-redundant nucleotide database and custom databases using BLAST. To test the effect of the E-value and background noise, sequences were retrieved after being curated at three different E-values (10^{-3}, 10^{-6}, 10^{-9}) and examined as separate E-probe sets. The process of E-probe creation was repeated from the MUMmer alignment to database curation, 100 times per E-probe set. This was done to test the hypothesis that under identical circumstances that the same E-probes sure be created.

Detection

The E-probe sets were aligned to the metagenomic data sets (mock and laboratory samples) and query coverage (QC) and percent identity (%ID) were measured at each intersection of 90%, 95% and 100%, with 27 total points of comparison and three E-probe

lengths (60nt, 80nt, 100nt). A hit was defined as any instance where a read had a counterpart E-probe and the count of hits of a particular E-probe is referred to as the hit depth.

The false positive threshold was established by comparing the hit alignments of the true negative control and observing the threshold parameters where no false positives were present. Using the True Negative/True Positive control method, the number of hits and hit depth for each E-probe in the True negative mock metagenomic data set and the true positive metagenomic datasets were calculated. Detection and hit number were compared in both the *in silico* and laboratory metagenomic datasets.

Results and Discussion

It was found that when using unassembled Illumina data, that E-probes with 80nt were able to detect the target when it made up at least 0.0018% of the metagenome using a query coverage of 90% and a percent identity of 95%. The next lowest detection (0.018%) was achieved using the same parameters but with E-probes of 100nt lengths. The 60nt E-probes were able to detect the target at (0.019%) using parameters of 90% query coverage and 100% percent identity.

In previous studies the E-value parameter was explored using viral, fungal and bacterial plant pathogens and it was found that there was not a significant difference using the E-values of 10^{-3} 10^{-6} and 10^{-9} (Stobbe et al., 2012; 2013). However, this had not been tested using human foodborne pathogens like *S. enterica* and it was hypothesized that more significant hits will be retrieved from the BLAST due to the high abundance of human pathogens in the non-redundant NCBI database compared to plant pathogens. The E-value had also not been evaluated on samples containing very low titers (less than 0.5%) (Stobbe

et al., 2012; 2013; Espindola et al, 2018). In this study the E-value parameter was tested at 10^{-3}, 10^{-6} and 10^{-9} and the number of false positives in the mock metagenomic negative control and laboratory metagenomic negative control was calculated (Table 2). Because of the type of curation, the threshold of 10^{-3} resulted in the removal of the greatest number of E-probes from the set and the lowest number of false positives followed by the 10^{-6} and 10^{-9}. This is due to the fact that more E-probe alignment occur that fall within the 10^{-3} level and are therefore removed from the set. Because an E-value of 10^{-3} resulted in the fewest number of false positives the resulting parameters were tested under this value.

Previous studies used assembled and unassembled data from both Illumina and Roche 454 sequencing platforms in the EDNA detection pipeline (Chapter III, Stobbe et al., 2012). It was thought that since Roche 454 had longer average read lengths, that longer E-probes could result in more significant matches at lower pathogen titers as could using assembled contigs. However, from previous work using EDNA in *S. enterica* detection, it was found that the lower amount of sequencing data produced by Roche 454 and contig creation compared to unassembled Illumina data contributed to less detection of *S. enterica* at lower pathogen titer. Like in all sequence mapping, longer sequences result in more significant matches because the longer the match the less likely it is to happen by chance (E-value) and the less likely it is to be a region common among many organisms (Zhang et al., 2000). The length of the nucleotides tested 60nt, 80nt and 100nt resulted in detection thresholds based on query coverage and percent identity (Table 3). The 60nt E-probes where not able to detect pathogen with enough sensitivity and specificity until the target made up 0.18% of the dataset in both the mock and laboratory samples. This is because using shorter E-probes results in a greater number of false positive alignments using the same detection parameters. The 100nt E-probes were able to detect the target when it made

up 0.018% of the sample or greater. The 80nt E-probes had the greatest sensitivity and specificity tested and they were able to detect the target at only 0.0018% of the dataset. The laboratory metagenomic samples exhibited the same pattern found in the *in silico* mock metagenomic datasets (Table 4). All E-probe sets showed an increase in number of hits with increasing number of pathogen reads. No detection was achieved with a percent identity lower than 95% which is likely due to the inclusion of the E/I genome in the datasets which has high similarity >90% to the target. The greatest number of hits was achieved when the E-probes were curated at the 10^{-3} E-value and run with detection parameters of percent identity of 95% a query coverage of 90% and 80nt length E-probes.

Based on previous work the difference in speed seems to be correlated with the number of alignment and therefore the number of E-probes in a set. Because of the way that the EDNA pipeline is constructed the shorter the length of the individual E-probes, the greater number of E-probes that will be produced. If an EDNA parameter increases the length of time required to reach a diagnostic call to greater than 2 hrs., then it would be possible to employ a profiling-based method and some of the strength of this diagnostic method would be lost (Miller et al., 2010). From previous work it was found that E-probe sets with fewer than one thousand E-probes were able to run all alignments in less than five minutes. This combined with the fact that more significant alignments occur with increasing length supports the findings that 80nt are optimal for this study. This was the second longest length tested, however because the 100nt length E-probes were the same length as the reads it was more challenging to get an alignment.

A metagenomics-based approach has many advantages for human foodborne pathogen diagnostics. Next generation sequencing (NGS) has made it possible to generate billions of sequences from a single nucleic acid sample that can be used to represent an

entire metagenomic community (Jones et al., 2010: Tyson et al. 2004) This allows for any pathogen present in a sample to be detected from a single assay. Metagenomic studies have been used in order to identify the causal agent of an unknown disease, but it is not a regularly used method (Adams et al., 2009, Cox-Foster et al., 2007, Palacios et al., 2008).

One of the biggest hindrances in using metagenomics in detection is the current cost per run. Metagenomic samples are often large and it is almost impossible to estimate coverage because the amount and identity of sequences are not known. The typical metagenomic diagnosis approach is nucleic acid extraction, sequencing, assembly and a BLAST of the assembled contigs. Based on current trends, it is likely that sequencing technologies will continue to drop in cost per run, due to advances in technology and greater access (Parameswaran et al., 2007).

However, as sequencing decreases in cost, increases in speed and increases in number of reads generated, the issues of downstream data handling becomes a bigger issue. These same advances in NGS will have an additional exponential growth effect on the databases (GenBank) that are used for the BLAST searching of sequence data, suggesting that the current metagenomic approach to pathogen diagnostics will eventually become too computationally intensive for everyday use.

Rapid detection pipelines like The Tool for Oligonucleotide Fingerprint Identification (TOFI) was created to generate a microarray *in silico* and provided a starting point for the EDNA pipeline (Geyer et al., 2008; Stobbe, 2013; Stobbe, 2014; Satys et al. 2008). TOFI is an integrated, scalable, high-performance-computing tool that incorporates genome comparison and probe design software. It was designed as a high throughput method to simultaneously process multiple bacterial and/or viral genomes and identify

fingerprints that are unique to each genome. It can also be used to find fingerprints that are common between genomes (Geyer et al., 2008). The TOFI pipeline includes three main steps. The first step is a comparison of pathogen sequence with those of near neighbors for unique fingerprinting, the second step is thermodynamic optimization and the final step is a check for uniqueness with BLAST. The strength of this method is that it reduces that amount of data that needs to be queried by only searching for the fingerprinted regions. This method also suggests that by using the *in silico* fingerprinting method, hundreds of related genomes could be run in a single assay (Geyer et al., 2008). However, for detection it is not necessary to do all of the work in gene expression that is proposed by this pipeline and this pipeline is limited in its application with metagenomic data due to its reliance on thermodynamics which is not a concern in metagenomics.

The EDNA system provides a simplified bioinformatic approach for managing the complexity and exponential growth of metagenomic sequencing. EDNA uses the sample as the searchable database and identifies unique regions of the target using E-probes for detection without the need for assembly. This streamlines the detection pipeline by removing the quality checking and assembly steps used by most data analysis pipelines. This technique has been demonstrated in plant pathogen studies where viral, fungal and bacterial plant pathogen E-probes were able to successfully detect multiple targets from a single metagenomic sample (Stobbe et al., 2012). It has also been effective in targeting the plant secondary metabolite aflatoxin from toxin-producing *Aspergillus flavus* (Espindola et al., 2018). Based on previous work, using EDNA for detection of human foodborne pathogens, it was established that the EDNA method has great potential for detection in human foodborne pathogens, but it was not optimized for this application.

In order to establish optimized parameters for human foodborne pathogen detection using the EDNA system, the pipeline was deconstructed, and each parameter was tested for its contribution to detection. Metagenomic data from the Illumina platform was chosen due to the lower cost per run compared to other methods, as well as, the accessibility of the technology. In addition to the laboratory metagenomic data sets, (S1, S2 and T1) *in silico* mock datasets were constructed that represented five simulated levels of *S. enterica* in Illumina metagenomic sample and a negative control with ten replicates for each concentration the target (M0-NC, M1, M2, M3, M4 and M5). These samples represented very low titers of target (< 0.5%). In previous studies, it was found that the standard parameters for detection correctly called positive sample positive except for those at very low titers (<0.5%). Because *S. enterica* found on fresh food substrates like tomato are likely in very low abundance, it was decided that the detection limit needed to be lower for the optimized parameters. It was suggested that in order to lower the detection threshold below < 0.5%, three parameters could be adjusted. These were E-probe number, length and parsing the E-value (Stobbe et al., 2012).

Earlier studies concluded that the number of hits (any instance where an individual e-probe finds a counterpart or counterparts in the database) and hit depth (cumulative total of e-probe/counterpart finds) were correlated to the number of e-probes available for a pathogen, to the pathogen abundance, to the E-value threshold used when parsing the data, and inversely correlated to the length of the E-probes. Because of this, it was hypothesized that a greater number of E-probes could increase the number of matches. In order to increase the number of E-probes, the overall length of the E-probes needed to decrease. However, using variable length E-probes and E-probes < 60nt significantly reduced the speed and were removed from the optimization study. It was also observed that longer E-

probes had a reduced number of false positives in the negative control. When all E-probe lengths were "normalized" by calculating the percent of *S. enterica* detection in each sample by dividing the number of matches by the total number of E-probes in a set, it was found that there was no significant difference in the level of detection between sets. It was found that by artificially reducing the number of E-probes to the 60nt, 80nt and 100nt length sets, that detection was greatly reduced. Additionally, at very low titers, E-probes with 60nt were found to overestimate the hit depth and resulted in an overestimation of *S. enterica* abundance in each sample. This is likely due to more than one E-probe matching reads since the read lengths were 100bps long. E-probes that were longer than 100nt did not provide any increase in detection compared to E-probes that were 80nt and 100nt. It was therefore decided that E-probes of 80nt were optimal because that length provided the greatest number of E-probes without reducing the speed of the pipeline or overestimating the abundance of the target. The curating the E-probes with and E-value of 10^{-3} reduced background noise and false positives compared to curating with E-values of 10^{-6} or 10^{-9}.

In the original parameters for EDNA, a read depth of two or greater was required for detection. This was based on the error rate for Illumina sequencing. However, since in human food borne pathogen detection a false negative can result in serious human health ramifications, the detection limit was lowered to a depth of one. In addition to the human help implications, it was rationalized that with the other optimizations to EDNA, like the percent identity of 97% or higher, that the match/mismatch of a nucleotide was unlikely to result in a greater number of false positives.

The number of false positives in a sample was of great concern because previous work had listed that as an issue in detection at very low pathogen titer. This work was able to achieve a stable threshold for the false positive rate that is based on the biological

similarity between the target genome and the I/E genome and not a limitation from the EDNA pipeline. It was originally thought that the more complex background of the laboratory metagenomic samples could result in higher false positive rates due to increased background signal. However, the same amount of false positive was found in both samples which could indicate that the mock metagenomic samples are an adequate representation of the complexity found in the laboratory samples.

The optimization of EDNA for human foodborne pathogens successfully detected *S. enterica* in all positive samples and resulted in a biologically base false positive alignment rate due to the I/E genome in the negative sample. Preliminary data suggests that the optimized parameters for *S. enterica* detection can be transferred to other human foodborne pathogens of concern. It should also be noted that biological group of *S. enterica* used for this study theoretically includes all subspecies of *S. enterica* and excludes all species, subspecies and strains of *S. bongori*. This was not expressly tested by this study and should be confirmed by future work. Additionally, since *S. enterica* is a diverse group, work to analyze the proportion of the *S. enterica* pan genome verses accessory genome represented by the E-probes could shed more light on the potential detection capability of this E-probe set.

The diagnostic positive/negative call is arguably the most important parameter in a diagnostic test. For molecular techniques, like PCR, the presence or absence of a product is easily determined. However, using quantitative measurements like those in fluorescence or absorbance in ELISA, the determination involves statistical analysis. The Decoy method is meant to be similar to molecular quantitative methods. For ELISA, a common approach is to make a diagnostic call by comparing the fluorescence value of a well to those of a set

of negative control wells and define the threshold cutoff as a certain number of standard deviations over background. The original EDNA design proposed converting these methods for use with NGS. Decoy E-probe sets were developed for *S. enterica*, and these Decoy E-probe sets were used to determine the chances that a random sequence would find a counterpart in a eukaryotic host background by chance. The problem with this approach was that the Decoy E-probes were more likely to match in complex metagenomic data than in a simple eukaryotic host. The Decoy method versus the true negative/positive method yielded similar results. However, there is concern that since the Decoy method measured an assumed negative that it could contribute to a higher false Positive/False negative rate depending on the specific Decoy E-probes created. It is also computationally more extensive without supplying a better test compared to the true negative/positive method. The true negative method simply removes the false positives created by background noise by comparing the true negative to the true positive. The only confounding factor is the creation of a true negative/positive experimentally or simulating an adequate background for a mock true negative. The ability to combine metagenomic sequencing with a rapid bioinformatic detection tool presents an opportunity to improve the access and usability of both fields. This streamlines the detection process of complex metagenomic sequence data into a five-minute analysis of all possible pathogens in a single assay. Additionally, the optimization of this tool for very low titer human foodborne pathogen detection confirms that this tool can be used in both the plant and human fields and could greatly improve upon the methods currently used by the FDA and USDA.

LITERATURE CITED

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., et al., 2009. Nextgeneration sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol. Plant Pathol. 10, 537–545

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

Blagden, T., Schneider, W., Melcher, U., Daniels, J., & Fletcher, J. (2016). Adaptation and Validation of E-Probe Diagnostic Nucleic Acid Analysis for Detection of Escherichia coli O157:H7 in Metagenomic Data from Complex Food Matrices. *Journal of Food Protection,79*(4), 574-581.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., et al., 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. 99, 14250–14255.

Castelino, M., Eyre, S., Moat, J., Fox, G., Martin, P., Ho, P., . . . Barton, A. (2017). Optimisation of methods for bacterial skin microbiome investigation: Primer selection and comparison of the 454 versus MiSeq platform. *BMC Microbiology, 17*(1), 23.doi:10.1371/journal.pone.0003373 http://www.plosone.org/article/info:doi/10.1371/journal.pone.0003373

Cox-Foster, D.L., Conlan, S., Holmes, E.C., Palacios, G., Evans, J.D., et al., 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318, 283–287.

Daquigan, N., Grim, C., White, J., Hanes, D., & Jarvis, K. (2016). Early Recovery of from Food Using a 6-Hour Non-selective Pre-enrichment and Reformulation of Tetrathionate Broth. *Frontiers in Microbiology, 7*, 2103.

MUMmer 2.1, NUCmer, and PROmer are described in "Fast Algorithms for Large-scale Genome Alignment and Comparision." A.L. Delcher, A. Phillippy, J. Carlton, and S.L. Salzberg, *Nucleic Acids Research* (2002), Vol. 30, No. 11 2478-2483.

Espindola, Andres S., William Schneider, Kitty F Cardwell, Yisel Carrillo, Peter R Hoyt, Stephen M Marek, Carla D Garzon. (2018). Inferring the presence of aflatoxin-producing Aspergillus flavus strains using RNA sequencing and electronic probes as a transcriptomic screening tool. *PLoS ONE, 13*(10), E0198575.

Fausser, Lee, Villari, Zeng, Zhang, Serikov, . . . Gabriel. (2011). Numerical benchmarks TRIPOLI − MCNP with use of MCAM on FNG ITER bulk shield and FNG HCLL TBM mock-up experiments. *Fusion Engineering and Design, 86*(9), 2135-2138.

Fey, Axel, Eichler, Stefan, Flavier, Sebastien, Christen, Richard, Hofle, Manfred G., & Guzman, Carlos A. (2004). Establishment of a Real-Time PCR-Based Approach for Accurate Quantification of Bacterial RNA Targets in Water, Using Salmonella as a Model Organism. *Applied and Environmental Microbiology, 70*(6), 3618-3623.

Fouhy, F., Clooney, A., Stanton, C., Claesson, M., & Cotter, P. (2016). 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology, 16*(1), 123.

Geyer Jeanne, Wasieloski Leonard, Padilla Susana, Bode Elizabeth, Kumar Kamal, Zavaljevski Nela, . . . Reifman Jaques. (2008). In silico microarray probe design for diagnosis of multiple pathogens. *BMC Genomics, 9*(1), 496.

Gourlé H, Karlsson-Lindsjö O, Hayer J and Bongcam+Rudloff E, Simulating Illumina data with InSilicoSeq. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty630

Hope Eo'Donnell, & Stephen Emcsorley. (2014). Salmonella as a model for non-cognate Th1 cell stimulation. *Frontiers in Immunology, 5*, Frontiers in Immunology, 01 December 2014, Vol.5.

Jones, W., 2010. High-throughput sequencing and metagenomics. Estuar. Coasts 33, 944–952.

Karlsson, O. E., Hansen, T., Knutsson, R., Löfström, C., Granberg, F., & Berg, M. (2013). Metagenomic detection methods in biopreparedness outbreak scenarios. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, *11*(S1), S146-S157.

Korem, Zeevi, Suez, Weinberger, Avnit-Sagi, Pompan-Lotan, . . . Segal. (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science (New York, N.Y.), 349*(6252), 1101-1106.

Lambert, J., Yennawar, N., Gu, Y., & Elias, R. (2012). Inhibition of secreted phospholipase A2 by proanthocyanidins: A comparative enzymological and in silico modeling study. *Journal of Agricultural and Food Chemistry, 60*(30), 7417-20.

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., et al., 2010. Bioinformatics for next generation sequencing data. Genes 1, 294–307.

Metzker, M.L., 2010. Sequencing technologies — the next generation. Nat. Rev. Genet. 11, 31–46.

Miller, J. R., S. Koren, and G. Sutton. "Assembly algorithms for next-generation sequencing data." Genomics 95.6 (2010): 315-27. PubMed. Web. 25 May 2017

Nakamura, S., Yang, C. S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., ... & Mizutani, T. (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PloS one*, *4*(1), e4219.

Nakamura, S., Nakaya, T., & Iida, T. (2011). Metagenomic analysis of bacterial infections by means of high-throughput DNA sequencing. *Experimental Biology and Medicine*, *236*(8), 968-971.

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/

Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., et al., 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. 358, 991–998.

Pop, M., Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. Trends Genet. 24, 142–149.

Postnikova, E., Baldwin, C., Whitehouse, C.A., Sechler, A., Schaad, N.W., et al., 2008. Identification of bacterial plant pathogens using multilocus polymerase chain reaction/ electrospray ionization-mass spectrometry. Phytopathology 98, 1156–1164.

Reis-Filho, J., 2009. Next-generation sequencing. Breast Cancer Res. 11, 1–7.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH, *MetaSim: a sequencing simulator for genomics and metagenomics.*, PLoS One, Oct. 8, 2008 [Abstract, cited in PMC]

Ronaghi, M., 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. 11, 3–11.

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., et al., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol. Ecol. 19, 81–88.

Satya, Zavaljevski, Kumar, Bode, Padilla, Wasieloski, Biotechnology Hpc Software Applications Inst Fort Detrick MD. (2008). In silico Microarray Probe Design for Diagnosis of Multiple Pathogens.

Schaad, N.W., Frederick, R.D., Shaw, J., Schneider, W.L., Hickson, R., et al., 2003. Advances in molecular-based diagnostics in metting crop biosecurity and phytosanitary issues. Annu. Rev. Phytopathol. 41, 305–324.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome research*, *12*(10), 1611-8.

Stobbe, A. H., Daniels, J., Espindola, A. S., Verma, R., Melcher, U., Ochoa-Corona, F., ... & Schneider, W. (2013). E-probe Diagnostic Nucleic acid Analysis (EDNA): a theoretical approach for handling of next generation sequencing data for diagnostics. *Journal of microbiological methods*, *94*(3), 356-366.

Stobbe, A. H., Schneider, W. L., Hoyt, P. R., & Melcher, U. (2014). Screening metagenomic data for viruses using the e-probe diagnostic nucleic acid assay. *Phytopathology*, *104*(10), 1125-1129.

Singer, Andreopoulos, Bowers, Lee, Deshpande, Chiniquy, Woyke. (2016). Next generation sequencing data of a defined microbial mock community.

Tucker, T., Marra, M., Friedman, J.M., 2009. Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Genet. 85, 142–154.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., et al., 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37–43.

Wood DE, Salzberg SL: Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 2014, 15:R46.

Wu, Y., Simmons, B., & Singer, S. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics, 32*(4), 605-607.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., ... & Wang, J. (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *Journal of clinical microbiology*, *49*(10), 3463-3469.

Yoshitomi, K., Jinneman, K., Orlandi, P., Weagant, S., Zapata, R., & Fedio, W. (2015). Evaluation of rapid screening techniques for detection of Salmonella spp. from produce

samples after pre-enrichment according to FDA BAM and a short secondary enrichment. *Letters in Applied Microbiology,61*(1), 7-12.

Zhang Z., Schwartz S., Wagner L., & Miller W. (2000), "A greedy algorithm for aligning DNA sequences" J Comput Biol 2000; 7(1-2):203-14. PubMed

TABLES

Table 1) *In silico* mock Illumina metagenomic datasets created with MetaSim. The taxonomic profile contains the target pathogen *Salmonella enterica*, I/E *Salmonella bongori* and the top ten bacterial species from previously profiled tomato fruit surface communities and the background host *Solanum lycopersicum*. Each dataset was replicated ten times which is not shown. The only change in the dataset is the concentration of S. *enterica.*

Profile		M0 NC = 0 Cells S.enterica		M1 = 1 Cell S. enterica		M2 = 10 Cells S. enterica		M3 = 100 Cells S. enterica		M4 = 1000 Cells S. enterica		M5 = 10000 Cells S. enterica	
Name	NCBI Accession #	# of reads	% of sample	# of reads	% of sample	# of reads	% of sample	# of reads	% of sample	# of reads	% of sample	# of reads	% of sample
Salmonella enterica	NC_003198.1	0	0	45	0.00018	437	0.0018	4497	0.018	45234	0.18	444921	1.8
Salmonella bongori	CP006692.1	20859	0.09	20858	0.09	20764	0.09	20484	0.09	20839	0.9	20495	0.09
Solanum lycopersicum	CM001064.3	92529	0.4	92529	0.4	92853	0.4	92607	0.4	92672	0.4	90965	0.4
Rhizobium leguminosarm	CP001191.1	2139380	9	2139380	9	2139823	9	2139561	9	2136954	9	2101825	9
Pseudomonas putida	NC_002947.4	2915164	12	2915161	12	2917600	12	2915156	12	2912300	12	2861370	12
Pseudomonas syringae	NC_007005.1	2876125	12	2876123	12	2876695	12	2876329	12	2871473	12	2822729	12
Clavibacter michiganensis	NC_010407.1	1538512	6	1538512	6	1536223	6	1537871	6	1536155	6	1510081	6
Enterobacter cloacae	NC_014121.1	2509662	10	2509660	10	2507210	10	2510390	10	2503602	10	2457752	10
Pseudomonas fluorescens	NC_016830.1	3231296	13	3231296	13	3229201	13	3228287	13	3225847	13	3169883	13
Klebsiella pneumoniae	NC_016845.1	2517282	10	2517282	10	2517152	10	2517149	10	2508876	10	2472335	10
Citrobacter freundii	NZ_CP007557.1	2406478	10	2406470	10	2406325	10	2402908	10	2401666	10	2358873	10
Pantoea agglomerans	NZ_CP016889.1	1971560	8	1971560	8	1973325	8	1973117	8	1966693	8	1938332	8
Erwinia sp.	NZ_LN907827	1781153	7	1781153	7	1782392	7	1781644	7	1777689	7	1750439	7

Table 2) False positives rates in the *in silico* mock negative control at $1x10^{-3}$ $1x10^{-6}$ and $1x10^{-9}$. The use of the least stringent E-value $1x10^{-3}$ resulted in the removal of the greatest number of E-probes from the sets during curation and was correlated to the lowest false positive rate during detection followed by $1x10^{-6}$ and $1x10^{-9}$.

M0a-NC #rdsSE=0	Ev= 10^-3	Length				
		60nt	80nt	100nt		
QC	90	P	P	P	90	%ID
	95	P	P	P	90	
	100	P	P	P	90	
	90	P	N	N	95	
	95	P	N	N	95	
	100	P	N	N	95	
	90	N	N	N	100	
	95	N	N	N	100	
	100	N	N	N	100	
M0a-NC #rdsSE=0	Ev= 10^-6	Length				
		60nt	80nt	100nt		
QC	90	P	P	P	90	%ID
	95	P	P	P	90	
	100	P	P	P	90	
	90	P	P	P	95	
	95	P	N	N	95	
	100	P	N	N	95	
	90	N	N	N	100	
	95	N	N	N	100	
	100	N	N	N	100	
M0a-NC #rdsSE=0	Ev= 10^-9	Length				
		60nt	80nt	100nt		
QC	90	P	P	P	90	%ID
	95	P	P	P	90	
	100	P	P	P	90	
	90	P	P	P	95	
	95	P	P	P	95	
	100	P	N	N	95	
	90	N	N	N	100	
	95	N	N	N	100	
	100	N	N	N	100	

Table 3) The *in silico* mock metagenomic datasets show twenty-seven detection intersections from testing E-probe length (60nt, 80nt and 100nt) against QC (90%, 95% and 100%) and %ID (90%, 95% and 100%). Nine additional replicates not shown. The negative control shows that the false positive threshold for the 60nt is at 90% QC and 100% ID which was only able to result in detected target when the target made up 0.18% of the dataset or greater. The 80nt E-probes had the most optimal threshold with a QC of 90% and a %ID of 95 with the lowest level of detection being when the target made up 0.0018% of the databases. The 100nt E-probes were able to achieve detection at 90% QC and %ID of 95% when the target was at least 0.018% of the databases.

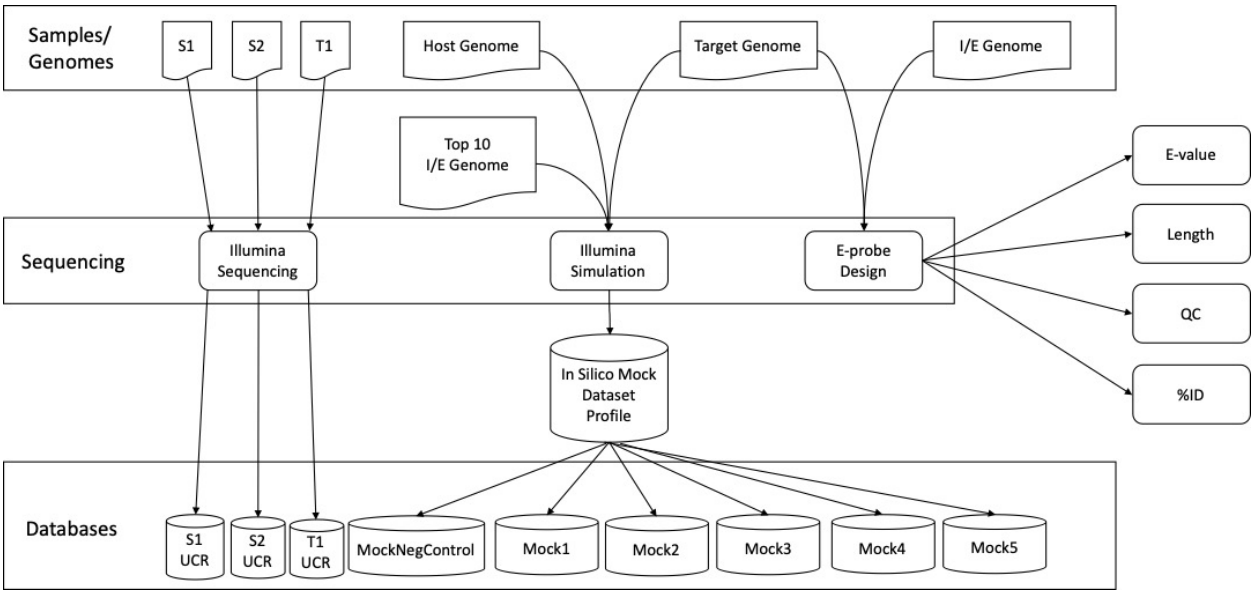
M0a-NC		Length				
#rdsSE=0		60nt	80nt	100nt		
QC	90	P	P	P	90	%ID
	95	P	P	P	90	
	100	P	P	P	90	
	90	P	N	N	95	
	95	P	N	N	95	
	100	P	N	N	95	
	90	N	N	N	100	
	95	N	N	N	100	
	100	N	N	N	100	
M1	1 cell	Length				
#rdsSE=48	%SE=0.00019	60nt	80nt	100nt		
QC	90	FP	FP	FP	90	%ID
	95	FP	FP	FP	90	
	100	FP	FP	FP	90	
	90	FP	N	N	95	
	95	FP	N	N	95	
	100	FP	N	N	95	
	90	N	N	N	100	
	95	N	N	N	100	
	100	N	N	N	100	
M2	10 cells	Length				
#rdsSE=437	%SE=0.0018	60nt	80nt	100nt		
QC	90	FP	FP	FP	90	%ID
	95	FP	FP	FP	90	
	100	FP	FP	FP	90	
	90	FP	5H/3HD	N	95	
	95	FP	4H/1HD	N	95	
	100	FP	4H/1HD	N	95	
	90	N	N	N	100	
	95	N	N	N	100	
	100	N	N	N	100	
M3	100 cells	Length				
#rdsSE=4497	%SE=0.019	60nt	80nt	100nt		
QC	90	FP	FP	FP	90	%ID
	95	FP	FP	FP	90	
	100	FP	FP	FP	90	
	90	FP	27H/1HD	5H/1HD	95	
	95	FP	24H/1HD	N	95	
	100	FP	19H/1HD	N	95	
	90	N	N	N	100	
	95	N	N	N	100	
	100	N	N	N	100	
M4	1,000 cells	Length				
#rdsSE=45234	%SE=0.19	60nt	80nt	100nt		
QC	90	FP	FP	FP	90	%ID
	95	FP	FP	FP	90	
	100	FP	FP	FP	90	
	90	FP	271	29	95	
	95	FP	241	15	95	
	100	FP	171	10	95	
	90	27	10	N	100	
	95	24	5	N	100	
	100	N	N	N	100	
M5	10000 cells	Length				
#rdsSE=444921	%SE=1.9	60nt	80nt	100nt		
QC	90	FP	FP	FP	90	%ID
	95	FP	FP	FP	90	
	100	FP	FP	FP	90	
	90	FP	2669	231	95	
	95	FP	2561	191	95	
	100	FP	2100	15	95	
	90	31	15	3	100	
	95	29	10	1	100	
	100	10	5	N	100	

Table 4) The laboratory metagenomic datasets showing twenty-seven detection intersections from testing E-probe length (60nt, 80nt and 100nt) against QC (90%, 95% and 100%) and %ID (90%, 95% and 100%). These tests exhibit the same pattern in the above Table 3. In this table, S2 is approximately equivalate to Table 3 M4 where the database contains approximately 0.18% of target from 1,000 cells of *S. enterica*. In this Table S1 contains over the amount of Table 3 M5 which is also reflected in the number of hits, however no change in detection threshold occurred.

T1		Length				
		60	80	100		
QC	90	FP	FP	FP	90	%ID
	95	FP	FP	FP	90	
	100	FP	FP	FP	90	
	90	FP	N	N	95	
	95	FP	N	N	95	
	100	FP	N	N	95	
	90	N	N	N	100	
	95	N	N	N	100	
	100	N	N	N	100	
S2	1000 cells SE	Length				
		60	80	100		
QC	90	FP	FP	FP	90	%ID
	95	FP	FP	FP	90	
	100	FP	FP	FP	90	
	90	FP	260	21	95	
	95	FP	253	15	95	
	100	FP	173	10	95	
	90	29	10	N	100	
	95	23	5	N	100	
	100	N	N	N	100	
S1	1000000 cells SE	length				
		60	80	100		
QC	90	FP	FP	FP	90	%ID
	95	FP	FP	FP	90	
	100	FP	FP	FP	90	
	90	FP	3769	1521	95	
	95	FP	2567	1393	95	
	100	FP	2303	57	95	
	90	570	551	15	100	
	95	331	59	10	100	
	100	10	5	N	100	

FIGURES

Figure 1) Overview of the experimental design and pipeline construction. (Far left) laboratory sample preparation and sequencing into Illumina unassembled metagenomic databases. (Middle) Construction of the *in silico* mock databases from NCBI genomes including background host (*Solanum lycopersicum*), I/E genome, the top ten bacterial species from previously sampled and profiled communities and target pathogen genome into the MetaSim program where five mock dilutions of target and negative control databases were constructed (Ten replicates for each concentration and negative control not shown). (Far right) E-probe construction and detection parameters identified for testing.



CHAPTER V

SIMULTANEOUS DETECTION OF HUMAN FOODBORNE PATHOGENS USING EDNA

Abstract

Aim: The objective of this study is to evaluate the range of the optimized detection capacity of the EDNA system for bacterial human foodborne pathogens by comparing the detection capability of the optimized parameters for the model bacterial pathogen *Salmonella enterica* to three additional human foodborne pathogens of concern *Campylobacter jejuni*, *Escherichia coli* O157:H7(STEC), *Listeria monocytogenes*.

Materials and Methods: Unassembled metagenomic DNA sequence data from the Illumina platform was simulated using the MetaSim program. *In silico* complex metagenomic samples were constructed at five concentrations of the four pathogens with ten replicates for each concentration using the MetaSim Illumina algorithm in order to mimic the complex metagenomic communities found from the community profiling of metagenomic laboratory samples. Using the optimized model parameters for S. *enterica* as a reference for sensitivity and specificity, the detection compacity of the three new pathogen E-probes sets will be compared for hit number, sensitivity and specificity.

.

Results: No difference in detection was observed when the number of read for each target made up at least 0.0018% of the dataset. However, because of the difference in genome length among the four pathogens, the number of the reads were not equivalent to the estimated cell number.

Significance and Impact of the study: E-probe Diagnostic Nucleic-acid Analysis (EDNA) is a probe-based bioinformatic pipeline that has the potential to rapidly and simultaneously detect any and all pathogens present in a single complex sample through metagenomic data mining.

Introduction

Foodborne human pathogens pose a significant risk to human health and welfare. According to the data gathered by the Center of Disease Control (CDC), there are currently 31 pathogens that have been identified on food as the causal agents of disease in humans (CDC, 2016). Improvements in sanitation and farming practices have mitigated the levels of pathogen contamination on food, however, the CDC estimates that in 2016 foodborne pathogens have resulted in 9.4 million illnesses, 55,961 hospitalizations, and 1,351 deaths in the United States (CDC, 2016). Bacterial pathogens make up a majority of the pathogens known to cause foodborne illnesses and *Salmonella enterica* is listed as the top foodborne pathogen contributing to hospitalization (35%) and death in at risk groups (28%) in the United States (CDC, 2016).

Salmonella is a popular organism for pathogen modeling studies (Preeti et al., 2012). It is representative of Gram-negative facultative anaerobic bacteria (Eo'Donnell and

Emcsorley, 2014). The Salmonella genus has high sequence similarity (96-99%) and is divided into two species *Salmonella enterica* and *Salmonella bongori*, with eight subspecies. Of the two species, *S. bongori* is considered significantly smaller having only one subspecies known as subspecies V. The other seven subspecies (I, II, IIIb, IV, VI and VII) belong to *S. enterica*. Subspecies I is specific to warm blooded animals, while the other six subspecies are found in cold blooded animals. There are over 2,500 serovars that have been identified with *Salmonella enterica* servovars Tyhimurium and Typhi specifically of concern in humans (Fey at al, 2004). It is estimated by the CDC that *Salmonella* species are the causal agents of 1.2 million illnesses, 23,000 hospitalizations and 450 deaths annually in the united states (CDC, 2019). Because of the risk to human health and subsequent surveillance by government agencies, Salmonella is widely available for research. It also has a high rate of growth making it an ideal model pathogen for laboratory study (Fey et al., 2004).

Escherichia coli (STEC) are also Gram negative, rod-shaped, non-spore forming, facultative anaerobic bacteria in the family Enterobacteriaceae. The well-known serotype *E. coli* O157:H7 is most commonly associated with foodborne illness, but additional virulent strains continue to be isolated and identified as the causal agents in multinational outbreaks (Luna-Gierke et al., 2014; Johnson et al., 2006) (Sodha et al., 2014; Luna-Gierke et al., 2014). It is unclear whether these new strains are a product of new isolation and detection capabilities or new emerging strains (Brooks et al., 2005; Johnson et al., 2006). The CDC estimates that *E.coli* STEC is the causal agent of 95,400 illness and yearly in the united states and STEC infections are of great concern due to the possible complication of hemolytic uremic syndrome (HUS) which affects the kidneys and is life threatening (Karmali, 1989). This pathogen is most often thought of as a contaminate in ground beef

and meat products; however, it was also implicated in the human foodborne illness outbreaks in spinach (CDC 2016), fenugreek sprouts (CDC 2011), clover sprouts (CDC 2012) and precut salad (2013). This trend toward fresh produce is concerning and research into the survival mechanisms on these products is ongoing (Leff and Fierer, 2013).

Listeria monocytogenes is a species of Gram positive, rod-shaped, non-spore forming, facultative anaerobic bacteria in the family Listeriaceae. The human mortality of this pathogen is between 20-30% of 1,600 cases annually in the US making it the mostly deadly human foodborne pathogen (CDC 2012; Ramaswamy et al., 2007). Of the six species only, *L. monocytogenes* has been identified as a causal agent of disease in humans. Of the thirteen serotypes, only three are associated with foodborne illness (1/2a, 1/2b and 4b)(Ward et al., 2004; Painter et al., 2013). *Listeria* is relatively rare but because of its high virulence and serious complications like pneumonia, meningitis, septicemia and spontaneous abortion it is treated as a pathogen of concern and monitored by the CDC (CDC, 2018) (Scallan et al., 2011; Ramaswamy et al., 2007). This pathogen is most often associated with preserved products like cheese and deli meet, however it has also been found on fresh produce (Bae et al., 2013; Kovacevic et al., 2013; Painter et al., 2014).

Campylobacter jejuni is a common food contaminate estimated as the causal agent in 1.3 million cases of illness from food in the United States yearly (CDC, 2019). It is a motile Gram-negative non-spore forming spiral shaped that thrives in microaerophilic environments. There are 34 recognized species of *Campylobacter* with *jejuni* and *coli* most often implicated in human disease. The two most cited subspecies of *Campylobacter jejuni* are *jejuni* and *doylei*. These bacteria are often associated with poultry contamination (Parker et al., 2007).

EDNA Optimization

Builds on the fundamentals of the Tool for Oligonucleotide Fingerprint Identification (TOFI) and streamlines it for use with metagenomic data. The TOFI tool was created to simulate a microarray *in silico* (Geyer et al., 2008; Stobbe, 2013; Stobbe, 2014; Satya et al., 2008). The strength of this method is that it reduces that amount of data that needs to be queried by only searching for unique fingerprinted regions. This method also suggests that by using the *in silico* fingerprinting method, hundreds of related genomes could be run in a single assay (Geyer et al., 2008). However, for detection it is not necessary to do all of the work in gene expression that is proposed by the TOFI pipeline and the pipeline is limited in its application with metagenomic data. E-probe Diagnostic Nucleic-acid Analysis (EDNA) is a tool developed at Oklahoma State University in conjunction with the United State Department of Agriculture (USDA) to bridge the gap between profiling-based methods and diagnostically realistic time requirements. Similar to TOFI, this pipeline is completely *in silico* which reduces the cost. EDNA was originally utilized to study plant pathogens due to the fact that many of the organisms and viruses in plant pathogen systems are not well characterized and the amount of unculturable and unknown pathogens are likely higher than in human and animal systems. EDNA requires genomes of the targets and can be used with incomplete genomes, although this reduces the specificity. This pipeline is also ideal for detection of human foodborne pathogens like *Salmonella enterica* because it presents a rapid detection that can be done with unassembled metagenomic sequence data which greatly reduces computational time after sequencing and has great potential for in field use. The EDNA system was optimized for *S. enterica* detection and the optimal pipeline is as follows. By following the standard workflow of choosing a representative target sequence of *S. enterica* and an inclusivity

exclusivity determinate genome *S. bongori*. The two sequences were aligned using the MUMmer program for pairwise comparison (Delcher et al., 2002). The MUMmer program is used to find and identify the maximal matches in the global alignment of the two genomes and eliminate regions of similarity. The output is lengths of the target genome that do not overlap with the inclusivity/exclusivity determinate genome. The regions of the target genome are then shredded into 80nt length E-probes using BioPerl (Staijch et al., 2002). The E-probes are then curated by mapping them to the nucleotide database of NCBI using BLAST as well as the genomes of the organisms in the negative control samples if they were not complete or present in the NCBI database. The E-probes were removed from the dataset if they aligned to non-target reads at an E-value of $1x10^{-3}$. This is a critical step because even though only unique regions of the target were chosen compared to the closely related genome, it is likely that other genomes in the database share common or similar regions that can result in high scoring false positives in metagenomic data. This is a major issue for community profiling pipelines, but one that EDNA has been able to alleviate though optimized curation. After the BLAST, the curated E-probe library is ready to use for detection applications.

EDNA is designed to detect targets in unassembled metagenomic sequence data. The E-probes are mapped to the sequence data using BLAST and the hits are filtered based on the combined score of the Query Coverage (QC) and Percent Identity (%ID). The QC is a threshold parameter based on the number of nucleotides that have to match in order for the alignment of the E-probe read to be reported. The %ID is simultaneously measured and is similar to the match/mismatch parameter of BLAST. The %ID establishes a baseline percentage of nucleotides that have to be identical given a particular alignment length (QC). Both QC and %ID are calculated as percentages and reported as the “Score”. The

optimal threshold settings for QC was 90 while the %ID needed to be at least 95% for *S. enterica*. The final step is the diagnostic call, meaning does the sample contain the pathogen or is it negative for the pathogen? In almost all diagnostic pipelines, it is assumed that there will be some level of false positives/false negatives and EDNA is no different. When true negative and positive controls are available, it is possible to set the threshold for positive diagnostic calls relative to the difference in Score between the hits in the negative control versus the positive control. Using these optimized parameters EDNA was able to detect the *S. enterica* target when it was as low as 0.0018% of a complex metagenomic dataset of 24,000,000 reads of 100 bps in length. Theoretically, EDNA can be used for simultaneous detection of multiple pathogens (Geyer et al., 2008; Stobbe et al., 2012). But this has not been tested using bacterial foodborne pathogen or complex metagenomic data.

In previous studies *in silico* complex metagenomic datasets have been used to assess the detection limit, sensitivity and specificity of the EDNA optimization parameters. The *in silico* findings where then compared to the detection from laboratory metagenomic datasets and they were found to follow the same patterns without divergence which could indicate that the simulated complex background in the *in silico* datasets was an adequate representation of the laboratory samples. Metagenomic mock datasets are simulations of real environmental data (Richter et al., 2008). These datasets are key in uncovering the limitations of currently available metagenomic data analysis tools because they offer a way to test the output results against the inputs of an experiment (Richter, 2008). This has been a major problem in the evaluation of tools for metagenomic analysis, because due to the nature of environmental samples, the inputs are variable and exact quantities are unknown (Korem et al., 2015). Mock datasets allow for the creation of true positive and negative controls, something that is not possible in strict experiments using only metagenomic data

from environmental samples. Without the use of true positive and negative samples, the experimental design is flawed, and conclusions derived from the study can be brought into question (Stobbe et al., 2012). This is not to say that mock datasets are a complete substitute for real environmental data sets, only that they are a resource that can be utilized for the testing of metagenomic analysis tools in order to better understand the outputs from studies with metagenomic data.

There are two main types of metagenomic mock datasets. The first type called an *in vitro* mock community dataset, is constructed by placing organisms in a simulated community before extracting the DNA or genetic material and sequencing the community (Fouhy, 2016; Fausser, 2011). This type of mock community is defined as a mixture of microbial cells, viruses or nucleic acids that were created *in vitro* to provide a simulation of the composition of a microbial sample (Castelino, 2014). This is considered a synthetic or mock community because it is not a community derived from a real environmental sample. Since the completion of the Human Genome Project and the Human Microbiome Project, this type of dataset has been used extensively to simulate the microbial community structure found in real environmental samples. Examples of these datasets are The Human Microbiome Project's BEI: HM-280, HM-281, HM-278D and HM -279D, these databases are available through BEI for researchers working on infectious diseases of humans (NIH HMMC web). Another well-known mock community is the Mock Bacteria ARchaea Community (MBARC-26) created for researchers working with archaea communities. However, this type of dataset is only an estimation of the community structure found in environmental metagenomic datasets and cannot completely replicate the relationships between community members (Wu, 2016). It should also be noted that since the community structure is calculated prior to sequencing, the actual

number of members is somewhat variable, due to extraction and sequencing errors (Miller, 2017).

The second type of mock metagenomic dataset is derived from *in silico* modeling that has been used to analyze programs in the computer science field (Richter et al., 2008). Many fields are now using these statistical and computer based *in silico* models to evaluate and optimize products and tools before implanting them in further studies. These are known as *in silico* mock metagenomomic datasets. This type of dataset uses sequencing data and genomes from databases like NCBI. The quality of the sequencing and genome completeness is analyzed prior to incorporation of each genome into the datasets. This allows stricter calculations of detection limits and specificity compared to other methods where levels could be confounded by pre-analysis errors. MetaSim was one of the most successfully used open access metagenomic data simulators available (Richter, 2008). MetaSim allows for common errors based on sequencing platform to be incorporated into the datasets in order to more realistically simulate a metagenomic data (NIH web). This software works by generating collections of synthetic reads from specifically chosen genomes. The genomes representation, as well as, the number of reads from each genome can be designated. The program then generates mate pairs based on platform models. More tools that enable experiments to mock metagenomic communities *in silico* are coming to the marketplace like InSilicoSeq (Gourle et al., 2018). This tool generates Illumina reads for simulating metagenomic samples. In addition to providing more control on the mock community genome inputs, the cost of constructing an *in silico* mock metagenomic data set is minimal compared to other experiments that require extraction and sequencing. This is one reason why many fields including food chemistry have started regularly using *in silico* modeling for optimization studies (Lambert, 2012). This method also provides

research at facilities that are not equipped to handle live human pathogens with the ability to conduct preliminary experiments containing sequence data from human pathogens without containment or health risks. The metagenomic analysis tools can then be evaluated by comparing the input data to the output data (Blagden, 2016). Like all modeling-based experiments, the tools used will then need to be validated using real metagenomic data from environmental samples, because nothing can replace the use of real environmental data.

Both *in vitro* and *in silico* mock metagenomic data types are extremely useful in understanding how metagenomic analysis tools process and profile data. These tools are extremely important because completing metagenomic studies without an understanding of the biases and detection limits of the tools, can result in errors. If erroneous conclusions are made about metagenomic dataset due to the use of unvalidated tools, the understand of metagenomic community structure can be obscured. Both in silico mock databases and laboratory databases were used in the optimization of EDNA for *S. enterica*.

In this study, the optimized parameters of EDNA for the detection of the model human foodborne pathogen *S. enterica* will be used to construct E-probes for three additional human foodborne pathogens of concern (*E. coli* STEC, *Listeria monocytogenes* and *Campylobacter jejuni*) (Figure 1). The detection limit and possible areas of model bias will be examined by comparing the detection of *S. enterica* verses its possible reads to the four other pathogens and their equivalent possible reads. To do this *in silico* mock datasets will be constructed for side by side testing.

Material and Methods

The *in silico* metagenomic mock Illumina datasets were constructed to simulate massively parallel Illumina sequencing using the MetaSim program (Satya et al., 2008). Based on previous work, datasets were constructed to simulate the complex background found in real metagenomic community samples, the genomes from the top ten bacterial species identified across the previous metagenomic community studies were extracted from the NCBI genome database (NCBI Accession #CP001191.1, NC_002947.4, NC_007005.1, NC_010407.1, NC_014121.1, NC_016830.1, NC_016845.1, NZ_CP007557.1, NZ_CP016889.1, NZ_LN907827) along with chromosome one of *Solanum lycopersicum* (NCBI Accession #CM001064.3) to further mimic the real metagenomic profiles. The inclusivity/exclusivity determinate genomes (I/E) for each of the four pathogens that was used to construct the E-probes was also included in the mock datasets to determine the specificity of the E-probe hits (NCBI Accession #CP006692.1, CP001665.1, NC_003212.1, NZ_CP019977.1). Four mock datasets were constructed including a negative control. Each dataset was made to simulate a dilution of target pathogen in the complex metagenomic background. The dilutions were chosen by calculating the ratios between background community and the target in previously profiled metagenomic communities. This takes into account the factors that influence detection like incomplete extractions and limitations in sequencing depth. Based on previous metagenomic community profiles, it was decided that each dataset should contain 24,000,000 reads of 100 bps and the dilutions ranged from the equivalent of 1-1,000 cells of each pathogen. Ten replicate databases were constructed for each dilution and negative control.

E-probe Construction

Complete genomes of *S. enterica, E. coli* (STEC), *L. monocytogenes* and *Campylobacter jejuni* (NCBI Accessions #CM001064.3, #NC_002695.2, #NC_003210.1, #NC_009495.1) were downloaded from NCBI and the inclusivity/exclusivity determinate genomes (NCBI Accession #CP006692.1, #CP001665.1, #NZ_NYPG01000001-16.1, CP006905.1) were also downloaded from NCBI. The MUMmer program was used to find the optimal global alignment between the genomes and the corresponding inclusivity/exclusivity determinate genome, the regions unique to each target genome were binned. During the MUMmer alignment the maximum number of gaps was equal to zero and the minimal length of alignment was 15nt.

The BioPerl program was used to divide the binned sequences into lengths of 80 nucleotides. The E-probe sets were then curated by mapping them to the NCBI non-redundant nucleotide database using BLAST. Additionally, the E-probes were mapped to a database of complete genomes in the negative control which consisted of the background for all of the metagenomic datasets used and the I/E determinate genomes. The process of E-probe creation was repeated from the MUMmer alignment to database curation, 100 times per E-probe set. This was done to test the hypothesis that under identical circumstances that the same E-probes sure be created. BLAST alignments of the four E-probe sets were completed against each corresponding target genome and visualized using the CG View program.

Results

E-probes

All but one of the targets and I/E pairs were at the species level. By examining the alignments of E-probes to the target genome *S. enterica, L. monocytogenes* and *Campylobacter jejuni* a similar pattern of E-probes can be see spread relatively evenly across the genome (Figure 2). This is in contrast to the *E. coli* (STEC) alignment which shows greater clustering of E-probes. This is likely due to the fact that *E. coli* (STEC) was the only set created using a pathogenic target subspecies and a non-pathogenic target of the same subspecies. The pairs were chosen for biological inclusivity/exclusivity reasons. Meaning that it represented either clinically or biologically relevant groups for detection. All of the target genomes had different lengths, however they corresponded to their I/E genomes using the same ratios (Table 1). Most importantly, the number of E-probes in each pathogen set were not significantly different (Table 1).

Detection

By observing the number of reads and calculating the percentage that each target made up in the dataset it is possible to see that EDNA was able to achieve detection when the number of reads of each pathogen made up at least 0.0012% of the dataset (Table 2). However, in food microbiology it is necessary to correlate the number of reads to the number of cells. This means that species with smaller genomes will have fewer reads at the same cell dilution and it will be necessary to have a greater number of cells to achieve detection (Table 3).

Discussion

A metagenomics-based approach has many advantages for human foodborne pathogen diagnostics. Next generation sequencing (NGS) has made it possible to generate billions of sequences from a single nucleic acid sample that can be used to represent an entire metagenomic community (Jones et al., 2010; Tyson et al., 2004). This allows for any pathogen present in a sample to be detected from a single assay. Metagenomic studies have been used in order to identify the causal agent of an unknown disease, but it is not a regularly used method (Adams et al., 2009; Cox-Foster et al., 2007; Palacios et al., 2008).

One of the biggest hindrances in using metagenomics in detection is the current cost per run. Metagenomic samples are often large and it is almost impossible to estimate coverage because the amount and identity of sequences are not known. The typical metagenomic diagnosis approach is nucleic acid extraction, sequencing, assembly and a BLAST of the assembled contigs. Based on current trends, it is likely that sequencing technologies will continue to drop in cost per run, due to advances in technology and greater access (Parameswaran et al., 2007).

However, as sequencing decreases in cost, increases in speed and increases in number of reads generated, the issues of downstream data handing become a bigger issue. These same advances in NGS will have an additional exponential growth effect on the databases (GenBank) that are used for the BLAST searching of sequence data, suggesting that the current metagenomic approach to pathogen diagnostics will eventually become too computationally intensive for everyday use.

The EDNA system provides a simplified bioinformatic approach for managing the complexity and exponential growth of metagenomic sequencing. EDNA uses the sample

as the searchable database and identifies unique regions of the target using E-probes for detection without the need for assembly. This streamlines the detection pipeline by removing the quality checking and assembly steps used by most data analysis pipelines. This technique has been demonstrated in plant pathogen studies where viral, fungal and bacterial plant pathogen E-probes were able to successfully detect multiple targets from a single metagenomic sample (Stobbe et al., 2012). It has also been effective in targeting the plant secondary metabolite aflatoxin from toxin-producing *Aspergillus flavus* (Espindola et al., 2018). Based on previous work, optimizing EDNA for detection of human foodborne pathogens, EDNA can be used as a tool for simultaneous bacterial pathogen detection from complex metagenomic data.

The thresholds for sensitivity and specificity set by the EDNA optimization parameters using *S. enterica* as a model were able to detect the target when it made up at least 0.0018% of the sample. The new E-probe sets showed the same ability to detect target as low as 0.0012% of the sample. However, differences in genome size among the target sets affects the number of reads in the set and the percentage of target at a specific cell number. In the *E coli* (STEC) E-probe set another subspecies was used as the I/E determinate which phylogenetically make to the sequences more similar compared to the other E-probe sets. This did not seem to affect detection or E-probe number, however in the alignment of the E-probes to the target sequence it was observed that the E-probes clustered more closely together.

The ability to combine metagenomic sequencing with a rapid bioinformatic detection tool presents an opportunity to improve the access and usability of both fields. This streamlines the detection process of complex metagenomic sequence data into a five-minute analysis of all possible pathogens in a single assay. Additionally, the optimization

of this tool for very low titer human foodborne pathogen detection confirms that this tool can be used in both the plant and human fields and could greatly improve upon the methods currently used by the FDA and USDA

LITERATURE CITED

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., et al., 2009. Nextgeneration sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol. Plant Pathol. 10, 537–545

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

Blagden, T., Schneider, W., Melcher, U., Daniels, J., & Fletcher, J. (2016). Adaptation and Validation of E-Probe Diagnostic Nucleic Acid Analysis for Detection of Escherichia coli O157:H7 in Metagenomic Data from Complex Food Matrices. *Journal of Food Protection,79*(4), 574-581.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., et al., 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. 99, 14250–14255.

Castelino, M., Eyre, S., Moat, J., Fox, G., Martin, P., Ho, P., . . . Barton, A. (2017). Optimisation of methods for bacterial skin microbiome investigation: Primer selection and comparison of the 454 versus MiSeq platform. *BMC Microbiology, 17*(1), 23.doi:10.1371/journal.pone.0003373 http://www.plosone.org/article/info:doi/10.1371/journal.pone.0003373

Cox-Foster, D.L., Conlan, S., Holmes, E.C., Palacios, G., Evans, J.D., et al., 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318, 283–287.

Daquigan, N., Grim, C., White, J., Hanes, D., & Jarvis, K. (2016). Early Recovery of from Food Using a 6-Hour Non-selective Pre-enrichment and Reformulation of Tetrathionate Broth. *Frontiers in Microbiology, 7*, 2103.

MUMmer 2.1, NUCmer, and PROmer are described in "Fast Algorithms for Large-scale Genome Alignment and Comparision." A.L. Delcher, A. Phillippy, J. Carlton, and S.L. Salzberg, *Nucleic Acids Research* (2002), Vol. 30, No. 11 2478-2483.

Espindola, Andres S., William Schneider, Kitty F Cardwell, Yisel Carrillo, Peter R Hoyt, Stephen M Marek, Carla D Garzon. (2018). Inferring the presence of aflatoxin-producing Aspergillus flavus strains using RNA sequencing and electronic probes as a transcriptomic screening tool. *PLoS ONE, 13*(10), E0198575.

Fausser, Lee, Villari, Zeng, Zhang, Serikov, . . . Gabriel. (2011). Numerical benchmarks TRIPOLI − MCNP with use of MCAM on FNG ITER bulk shield and FNG HCLL TBM mock-up experiments. *Fusion Engineering and Design, 86*(9), 2135-2138.

Fey, Axel, Eichler, Stefan, Flavier, Sebastien, Christen, Richard, Hofle, Manfred G., & Guzman, Carlos A. (2004). Establishment of a Real-Time PCR-Based Approach for Accurate Quantification of Bacterial RNA Targets in Water, Using Salmonella as a Model Organism. *Applied and Environmental Microbiology, 70*(6), 3618-3623.

Fouhy, F., Clooney, A., Stanton, C., Claesson, M., & Cotter, P. (2016). 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology, 16*(1), 123.

Geyer Jeanne, Wasieloski Leonard, Padilla Susana, Bode Elizabeth, Kumar Kamal, Zavaljevski Nela, . . . Reifman Jaques. (2008). In silico microarray probe design for diagnosis of multiple pathogens. *BMC Genomics, 9*(1), 496.

Gourlé H, Karlsson-Lindsjö O, Hayer J and Bongcam+Rudloff E, Simulating Illumina data with InSilicoSeq. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty630

Hope Eo'Donnell, & Stephen Emcsorley. (2014). Salmonella as a model for non-cognate Th1 cell stimulation. *Frontiers in Immunology, 5*, Frontiers in Immunology, 01 December 2014, Vol.5.

Jones, W., 2010. High-throughput sequencing and metagenomics. Estuar. Coasts 33, 944–952.

Karlsson, O. E., Hansen, T., Knutsson, R., Löfström, C., Granberg, F., & Berg, M. (2013). Metagenomic detection methods in biopreparedness outbreak scenarios. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, *11*(S1), S146-S157.

Korem, Zeevi, Suez, Weinberger, Avnit-Sagi, Pompan-Lotan, . . . Segal. (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science (New York, N.Y.), 349*(6252), 1101-1106.

Lambert, J., Yennawar, N., Gu, Y., & Elias, R. (2012). Inhibition of secreted phospholipase A2 by proanthocyanidins: A comparative enzymological and in silico modeling study. *Journal of Agricultural and Food Chemistry, 60*(30), 7417-20.

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., et al., 2010. Bioinformatics for next generation sequencing data. Genes 1, 294–307.

Metzker, M.L., 2010. Sequencing technologies — the next generation. Nat. Rev. Genet. 11, 31–46.

Miller, J. R., S. Koren, and G. Sutton. "Assembly algorithms for next-generation sequencing data." Genomics 95.6 (2010): 315-27. PubMed. Web. 25 May 2017

Nakamura, S., Yang, C. S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., ... & Mizutani, T. (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PloS one*, *4*(1), e4219.

Nakamura, S., Nakaya, T., & Iida, T. (2011). Metagenomic analysis of bacterial infections by means of high-throughput DNA sequencing. *Experimental Biology and Medicine*, *236*(8), 968-971.

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/

Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., et al., 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. 358, 991–998.

Pop, M., Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. Trends Genet. 24, 142–149.

Postnikova, E., Baldwin, C., Whitehouse, C.A., Sechler, A., Schaad, N.W., et al., 2008. Identification of bacterial plant pathogens using multilocus polymerase chain reaction/ electrospray ionization-mass spectrometry. Phytopathology 98, 1156–1164.

Reis-Filho, J., 2009. Next-generation sequencing. Breast Cancer Res. 11, 1–7.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH, *MetaSim: a sequencing simulator for genomics and metagenomics.*, PLoS One, Oct. 8, 2008 [Abstract, cited in PMC]

Ronaghi, M., 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. 11, 3–11.

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., et al., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol. Ecol. 19, 81–88.

Satya, Zavaljevski, Kumar, Bode, Padilla, Wasieloski, Biotechnology Hpc Software Applications Inst Fort Detrick MD. (2008). In silico Microarray Probe Design for Diagnosis of Multiple Pathogens.

Schaad, N.W., Frederick, R.D., Shaw, J., Schneider, W.L., Hickson, R., et al., 2003. Advances in molecular-based diagnostics in metting crop biosecurity and phytosanitary issues. Annu. Rev. Phytopathol. 41, 305–324.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome research*, *12*(10), 1611-8.

Stobbe, A. H., Daniels, J., Espindola, A. S., Verma, R., Melcher, U., Ochoa-Corona, F., ... & Schneider, W. (2013). E-probe Diagnostic Nucleic acid Analysis (EDNA): a theoretical approach for handling of next generation sequencing data for diagnostics. *Journal of microbiological methods*, *94*(3), 356-366.

Stobbe, A. H., Schneider, W. L., Hoyt, P. R., & Melcher, U. (2014). Screening metagenomic data for viruses using the e-probe diagnostic nucleic acid assay. *Phytopathology*, *104*(10), 1125-1129.

Singer, Andreopoulos, Bowers, Lee, Deshpande, Chiniquy, Woyke. (2016). Next generation sequencing data of a defined microbial mock community.

Tucker, T., Marra, M., Friedman, J.M., 2009. Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Genet. 85, 142–154.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., et al., 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37–43.

Wood DE, Salzberg SL: Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 2014, 15:R46.

Wu, Y., Simmons, B., & Singer, S. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics, 32*(4), 605-607.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., ... & Wang, J. (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *Journal of clinical microbiology*, *49*(10), 3463-3469.

Yoshitomi, K., Jinneman, K., Orlandi, P., Weagant, S., Zapata, R., & Fedio, W. (2015). Evaluation of rapid screening techniques for detection of Salmonella spp. from produce

samples after pre-enrichment according to FDA BAM and a short secondary enrichment. *Letters in Applied Microbiology,61*(1), 7-12.

Zhang Z., Schwartz S., Wagner L., & Miller W. (2000), "A greedy algorithm for aligning DNA sequences" J Comput Biol 2000; 7(1-2):203-14. PubMed

TABLES

Table 1) The number of Hits and Hit depth of each E-probe set in each concentration of pathogen in the *in silico* complex metagenomic datasets. Nine replicates for each concentration not shown.

MM0-NC 0 cells		Length 80nt					
		SE	EC	LM	CJ		
QC	90	FP	FP	FP	FP	90	%ID
	95	FP	FP	FP	FP	90	
	100	FP	FP	FP	FP	90	
	90	N	N	N	N	95	
	95	N	N	N	N	95	
	100	N	N	N	N	95	
	90	N	N	N	N	100	
	95	N	N	N	N	100	
	100	N	N	N	N	100	
MM1a 1 cell		Length 80nt					
		SE	EC	LM	CJ		
QC	90	FP	FP	FP	FP	90	%ID
	95	FP	FP	FP	FP	90	
	100	FP	FP	FP	FP	90	
	90	N	N	N	N	95	
	95	N	N	N	N	95	
	100	N	N	N	N	95	
	90	N	N	N	N	100	
	95	N	N	N	N	100	
	100	N	N	N	N	100	
MM2a 10 cells		Length 80nt					
		SE	EC	LM	CB		
QC	90	FP	FP	FP	FP	90	%ID
	95	FP	FP	FP	FP	90	
	100	FP	FP	FP	FP	90	
	90	5H/3HD	9H/1HD	3H/1HD	N	95	
	95	4H/1HD	7H/1HD	1H/1HD	N	95	
	100	4H/1HD	7H/1HD	1H/1HD	N	95	
	90	N	N	N	N	100	
	95	N	N	N	N	100	
	100	N	N	N	N	100	
MM3a 100 cells		Length 80nt					
		SE	EC	LM	CB		
QC	90	FP	FP	FP	FP	90	%ID
	95	FP	FP	FP	FP	90	
	100	FP	FP	FP	FP	90	
	90	27H/1HD	49H/1HD	17H/1HD	9H/1HD	95	
	95	24H/1HD	42H/1HD	5H/1HD	5H/1HD	95	
	100	19H/1HD	33H/1HD	5H/1HD	5H/1HD	95	
	90	N	N	N	N	100	
	95	N	N	N	N	100	
	100	N	N	N	N	100	
MM4a 1000 cells		Length 80nt					
		SE	EC	LM	CJ		
QC	90	FP	FP	FP	FP	90	%ID
	95	FP	FP	FP	FP	90	
	100	FP	FP	FP	FP	90	
	90	271H/1HD	491H/1HD	181H/1HD	175H/1HD	95	
	95	241H/1HD	431H/1HD	173H/1HD	171H/1HD	95	
	100	171H/1HD	335H/1HD	95H/1HD	89H/1HD	95	
	90	10H/1HD	15H/1HD	10H/1HD	10H/1HD	100	
	95	5H/1HD	5H/1HD	5H/1HD	5H/1HD	100	
	100	N	N	N	N	100	

Table 2) The read number and cell number in each of the *in silico* complex metagenomic dataset correlated to the number of hits and total percentage of the datasets.

Profile	MM-NC			MM1 = 1 Cell			MM2 = 10 Cells			MM3 = 100 Cells			MM4 = 1000 Cells		
Org.	Read #	Sample %	Hits	Read #	Sample %	Hits	Read #	Sample %	Hits	Read #	Sample %	Hits		Sample %	Hits
SE	0	0	0	45	0.00018	0	407	0.0017	5	4473	0.018	27	44872	0.18	271
I/E SB	20439			20581			20692			20611			20772		
EC	0	0	0	58	0.00024	0	514	0.002	9	5219	0.022	49	51760	0.21	491
I/E EC	21588			21398			21664			21525			21596		
LM	0	0	0	21	0.000087	0	288	0.0012	3	2777	0.011	17	27134	0.11	181
I/E LI	14138			14295			14143			14247			13916		
CJ	0	0	0	15	0.000062	0	137	0.00057	0	1511	0.0062	9	15206	0.06	175
I/E CC	9717			9621			9686			9578			9516		

Table 3) The genome sizes of each target pathogen and I/E genome and resulting E-probe number.

Org.	**Genome Size**
S. enterica	4.86 Mb
I/E *S. bongori*	4.4 Mb
E. coli (STEC)	5.5 Mb
I/E *E. coli*	5.2 Mb
L. mono	2.9 Mb
I/E *L. innocua*	2.9 Mb
C. jejuni	1.6 Mb
I/E *C. coli*	2 Mb
Org.	**E-probe #**
SE	405
EC	411
LM	397
CJ	371

FIGURES

Figure 1) Overview of the creation of E-probes for *S. enterica, E. coli* (STEC), *L. monocytogenes* and *Campylobacter jejuni* and detection in complex metagenomic datasets using the EDNA pipeline. Target genomes for the four pathogens were extracted from NCBI as well as their I/E determinate genomes. Alignments were completed using the MUMmer program and overlapping regions of the genomes were removed. BioPerl was used to shred the remaining sequences of target into 80nt length segments. For curation, the sequences were used as BLAST queries and run against the NCBI nucleotide database, as well as, the fasta files for the negative control and non-target alignments with an E-value of $1x10^{-3}$ or more stringent were removed. The curated E-probe sets were then saved in fasta format. For detection, the E-probe sets were run simultaneously as BLAST queries against the *in silico* mock metagenomic datasets. The alignments were scored using a QC of 90% and a %ID of 95%. The target containing dataset hits were compared to the hits in the negative control to identify false positives and detection limits.

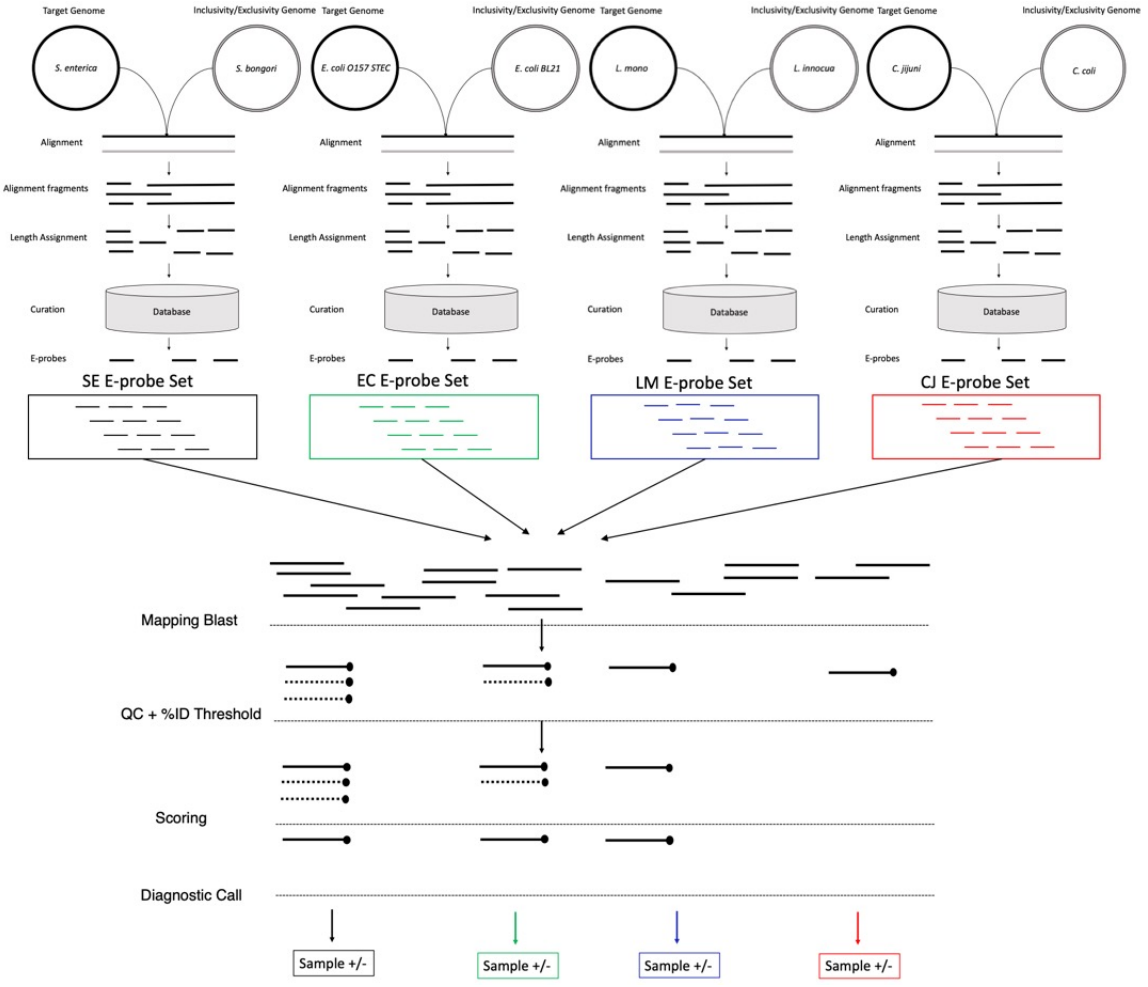
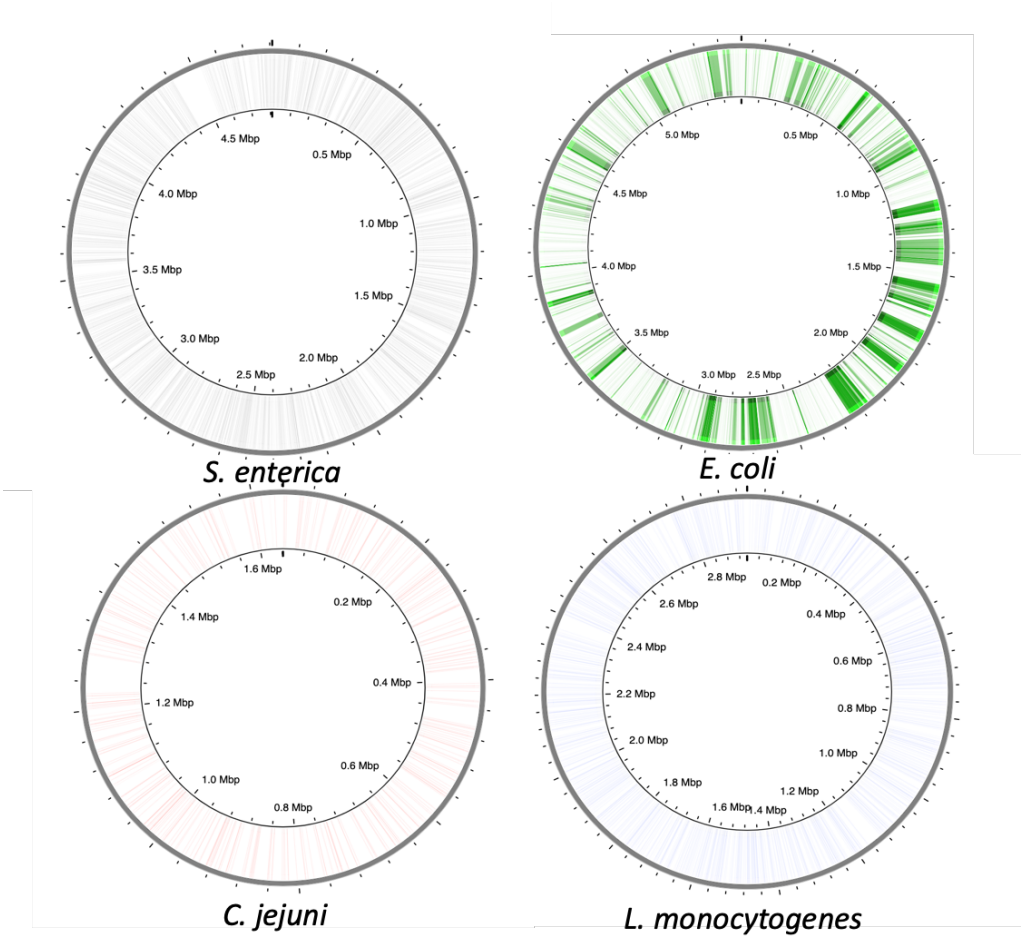


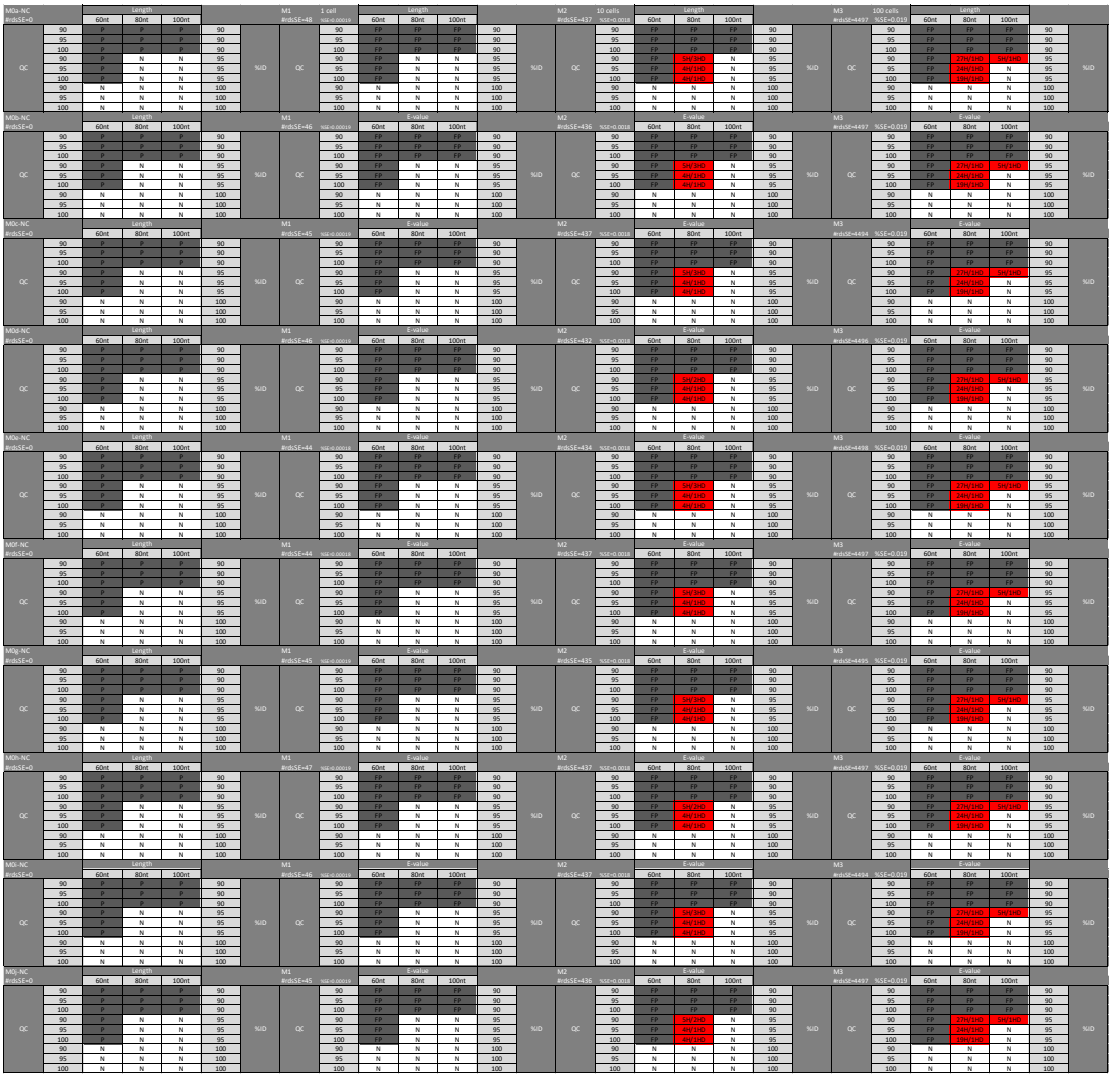
Figure 2) Alignment of each E-probe set to their corresponding target genome. The S. *enterica, L. monocytogenes* and *C. jejuni* E-probe sets show a similar pattern of dispersal across the respective target genome. *E. coli* (STEC) E-probes show a greater degree of clustering compared to the other sets. Alignments done using BLAST on the CG View Server.



APPENDICES

TABLES

Table 1) Ten replications of the *in silico* mock metagenomic datasets show twenty-seven detection intersections from testing E-probe length (60nt, 80nt and 100nt) against QC (90%, 95% and 100%) and %ID (90%, 95% and 100%). Nine additional replicates not shown. The negative control shows that the false positive threshold for the 60nt is at 90% QC and 100% ID which was only able to result in detected target when the target made up 0.18% of the dataset or greater. The 80nt E-probes had the most optimal threshold with a QC of 90% and a %ID of 95 with the lowest level of detection being when the target made up 0.0018% of the databases. The 100nt E-probes were able to achieve detection at 90% QC and %ID of 95% when the target was at least 0.018% of the databases.



FIGURES

Figure 1) T1 Illumina taxon assignments with 10,000 alignments or greater graphed as a function of percent identity

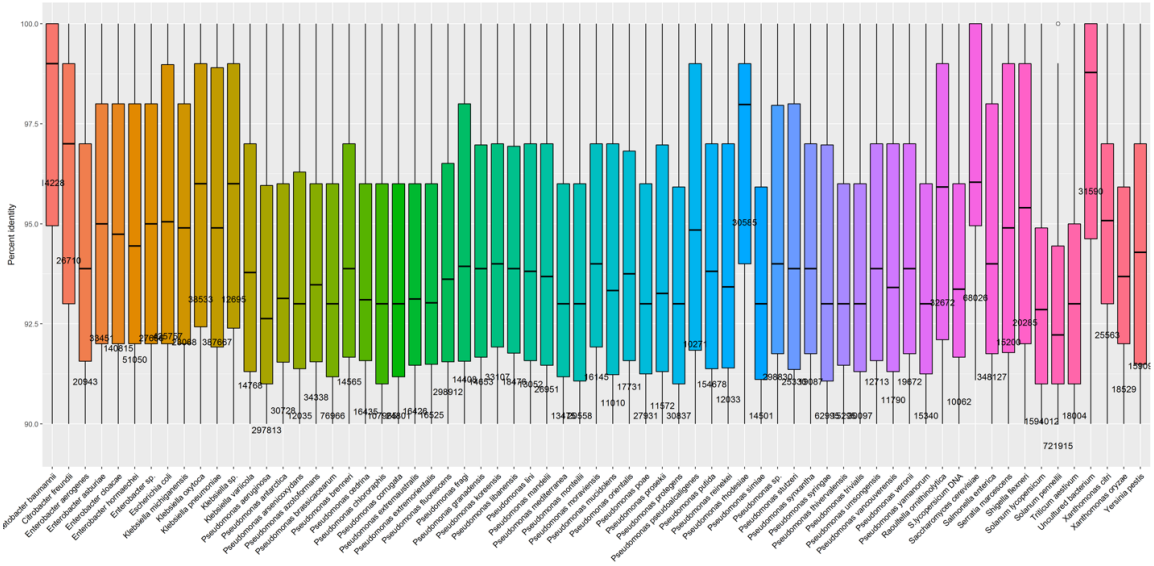


Figure 2) S2 Illumina taxon assignments with 10,000 alignments or greater graphed as a function of percent identity

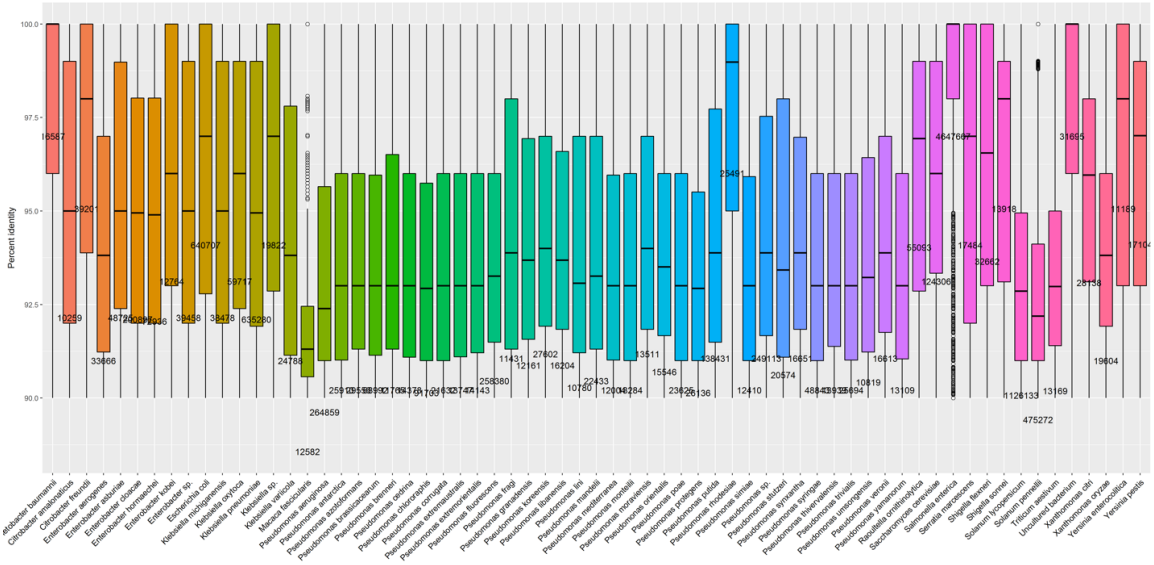


Figure 3) S1 Illumina taxon assignments with 10,000 alignments or greater graphed as a function of percent identity.

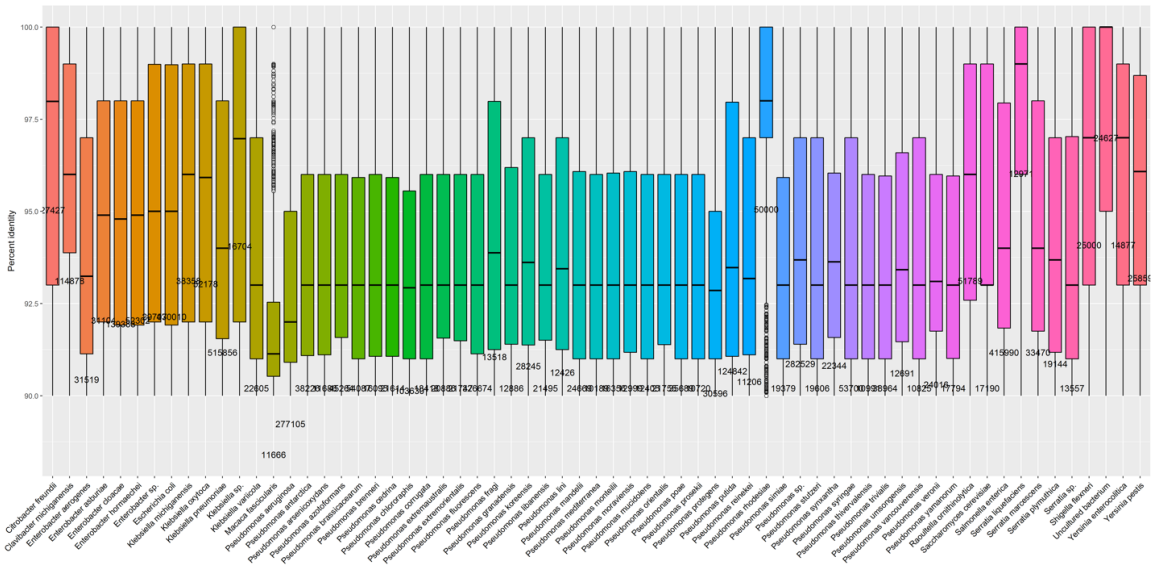


Figure 4) T1 454 taxon assignments with 10,000 alignments or greater graphed as a function of percent identity.

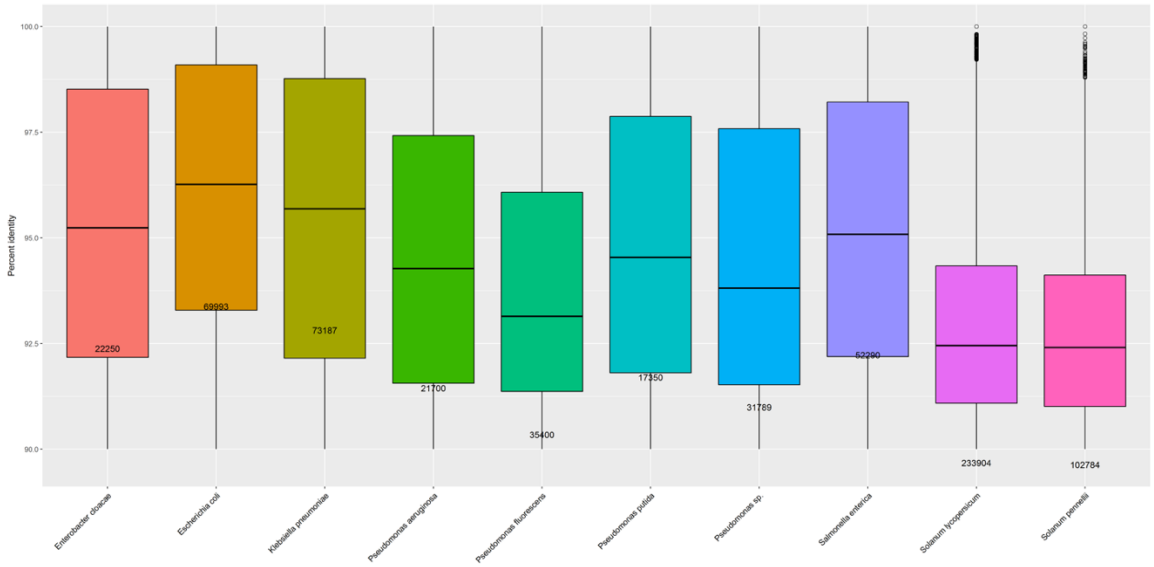


Figure 5) S2 454 taxon assignments with 10,000 alignments or greater graphed as a function of percent identity.

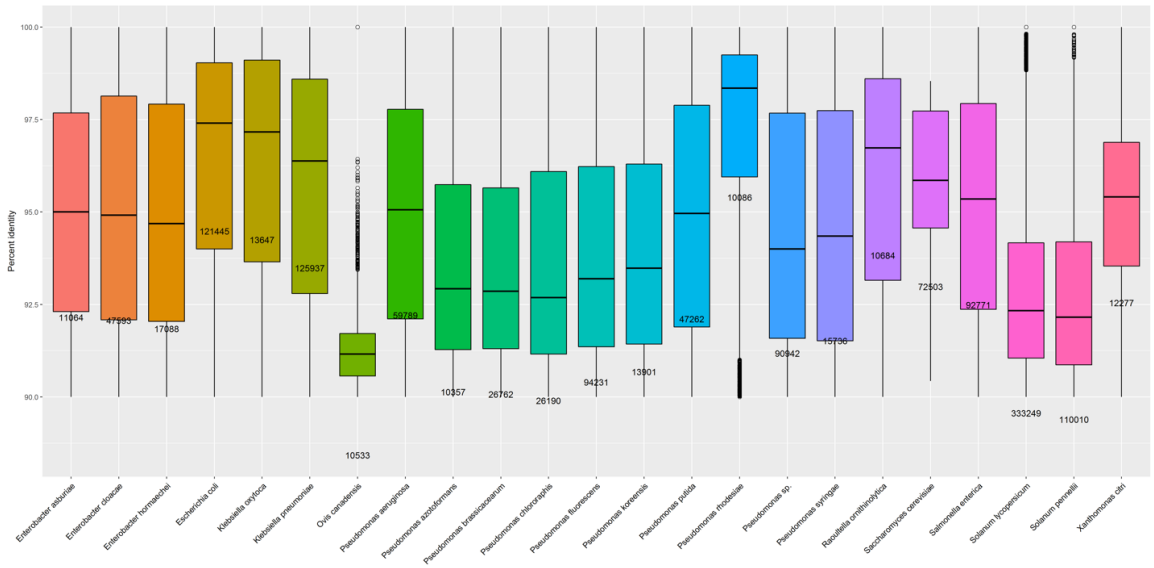
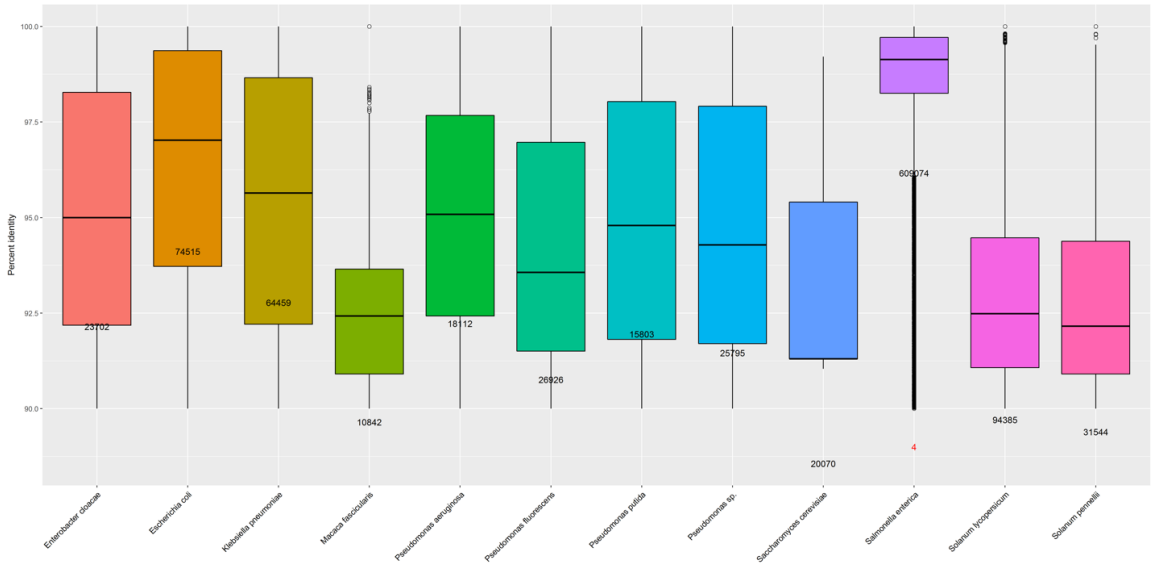


Figure 6) S1 454 taxon assignments with 10,000 alignments or greater graphed as a function of percent identity.



VITA

Gretta Marie Sharp

Candidate for the Degree of

Doctor of Philosophy

Dissertation: EVALUATION AND OPTIMIZATION OF BIOINFORMATIC TOOLS FOR THE DETECTION OF HUMAN FOODBORNE PATHOGENS IN COMPLEX METAGENOMIC DATASETS

Major Field: Plant Pathology

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Plant Pathology at Oklahoma State University, Stillwater, Oklahoma in July, 2019.

Completed the requirements for the Master of Science in Biology at the University of Texas at Tyler, Tyler, Texas in 2016.

Completed the requirements for the Bachelor of Science in Bioenvironmental Sciences and Agronomy at Texas A&M University, College Station Texas in 2012.

Experience:

Graduate Research Assistant. National Institute for Microbial Forensics and Food and Agricultural Biosecurity (NIMFFAB), Henry Bellmon Research Center, Oklahoma State University, Stillwater, Oklahoma, August 2016 to July 2019.

Visiting Scientist at the Federal Bureau of Investigation Laboratory Division, Stafford Virginia, 2019