THE USE OF NEXT GENERATION SEQUENCING TO DETECT PLANT PATHOGENIC PROKARYOTES

By

JON DANIELS

Bachelor of Science in Biology: Medical/Molecular

Rogers State University

Claremore, Oklahoma

2009

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 2013

THE USE OF NEXT GENERATION SEQUENCING TO

DETECT PLANT PATHOGENIC PROKARYOTES

Thesis Approved:

Francisco Ochoa Corona, Ph.D.

Thesis Adviser

Jacqueline Fletcher, Ph.D.

Stephen Marek, Ph.D.

ACKNOWLEDGEMENT

I would like to extend my gratitude to my major professor and mentor, Francisco Ochoa Corona. His passion for family, health, creative thought, and ability to always find ways to collaborate with any science discipline were an inspiration for me. I would also like to thank each of my committee members. Jacqueline Fletcher and Stephen Marek for providing valuable guidance in writing this thesis along with helping me understand the importance of networking and professionalism. I would like to thank William Schneider (Bill) for hosting my family and me for two summers and for teaching me about opportunities to work in government and industry. Bill is a true friend and someone I was fortunate to get to know. I would like to thank Diana Sherman and Trenna Blagden for teaching me laboratory techniques and helping with problems as they arose. Diana's and Trenna's guidance was a major factor in the success of my experiments. I would also like to thank Ian Moncrief and Sharon Andreason for all of their help and friendship.

While my committee and mentor made sure I stayed on course academically, my family provided me with the love and support I needed. To my wife, Maggie Daniels, I know it was hard but I'm grateful to have someone as wonderful as you to take care of everything while I was away. Your ability to teach our children how to succeed and surpass their classmates is a special gift. Your unconditional love for me will always be cherished. To my children Michaela, Lauren, and Bella, you are the greatest gifts I've ever been given and I'm proud of each of you and all that you are accomplishing.

This research was funded by the USDA-CSREES Plant Biosecurity Program, grant number 2010-85605-20542.

Name: JON DANIELS

Date of Degree: DECEMBER, 2013

Title of Study: THE USE OF NEXT GENERATION SEQUENCING TO DETECT PLANT PATHOGENIC PROKARYOTES

Major Field: ENTOMOLOGY AND PLANT PATHOLOGY

Abstract: Increasing importation of commodities from countries abroad increases the risk of introduction of exotic plant pathogens. Although individual pathogen assays are available, current screening methods have limited ability to detect multiple plant pathogens concurrently. The advent of next generation sequencing (NGS) technology allows for the creation of a single assay to detect simultaneously, any and all microbes in a sample, including pathogens that have been genetically modified. In this project, bioinformatic pipelines, streamlined PC programs, were developed to generate mock sample databases used to simulate 454 runs, query "electronic probe" (e-probe) design and BLAST searches. Pathogen specific queries, ranging in lengths from 20 nt to 140 nt, were created for detection of the bacterial select agents, *Xanthomonas oryzae* pv. *oryzae* and *Ralstonia solanacearum* race 3 biovar 2, as well as for *Candidatus* Liberibacter asiaticus and *Xylella fastidiosa* 9a5c (not select agents). The query sets were used to BLAST mock sample databases with one host, grapevine (*Vitis* vinifera), for all pathogen sequences at various ratios. All four bacterial pathogens were readily detectable *in silico*, suggesting that NGS technology has advantages beyond those of existing pathogen detection assays. To test *in silico* results pathogen specific e-probes, ranging in lengths from 15 nt to 60 nt, were created for detection of *Ralstonia solanacearum* race 3 biovar 2, and *Pseudomonas syringae* pv. *tomato* DC3000. The e-probe sets were used to query NGS sequencing data of diseased hosts, potato inoculated with Rs r3b2, and tomato inoculated with DC3000. Both bacterial pathogens were readily detectable; suggesting NGS data can be used, when combined with e-probes, as a prokaryotic plant pathogen detection assay. This research merges bioinformatics and plant pathology for addressing national security needs of a quick detection tool for any pathogen in a single assay for the agriculture industry.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

CHAPTER II

CHAPTER III

CHAPTER IV

LIST OF FIGURES

Figure Page

CHAPTER III

CHAPTER IV

Figure Page

CHAPTER I

INTRODUCTION

Agricultural biosecurity is a concern on many levels: local, regional, state, national and international. The U.S. agricultural sector, which includes both plants and animals, in managed and un-managed ecosystems including crops, forestry, range lands and aquatic systems, is vulnerable in the U.S. This vulnerability is due in part to the lack of security and surveillance systems and the enormous amount of land this industry uses (Harl 2002). Varying definitions of biosecurity have been published but blending those of the Food and Agriculture Organization of the United Nations, U.S. Office of Science and Technology Policy and the New Zealand Ministry for Primary Industries, presents biosecurity as an integrated series of strategies that combines policy and regulations to assess risk factors of food safety, animal and plant health, and environmental impact in an effort to prevent transmission of harmful biological agents to persons or the environment (FAO 2003; Guy 2013; OSTP 2013).

To achieve a goal of maintaining and sustaining plant and animal health, biosecurity agencies must engage subject matter experts in order to make decisions based on prioritization of microbial agents that threaten human, plant, animal and environmental health. In addition, agricultural biosecurity agencies must be in communication with growers and the public to maintain trust and use the agency's resources to collect data to enable appropriate responses and response times in the event a disease outbreak occurs.

Within the United States, the responsibility of protecting agricultural interest is divided among the United States Department of Agriculture (USDA), Department of Homeland Security, Department of Defense, Federal Burial of Investigation and others, in a collaborative effort to provide logistical support needed in protecting American agricultural interests. To assist in developing of portions of biosecurity protocols, the federal government collaborates with land-grant universities, agribusinesses, Cooperative Extension personnel, and other relevant organizations (Parker 2003).

Plant pathogens pose a unique biosecurity threat for many reasons. A majority of plant pathogenic microorganisms do not sicken humans directly, but can be harmful indirectly by damaging food crops and ornamentals. Unlike humans, plants cannot be vaccinated against diseases. If one plant, in a group of susceptible plants, becomes infected with a pathogen, the surrounding plants cannot simply move. In the current era of dense monoculturing and low genetic diversity, a pathogen can easily spread throughout a susceptible crop. For a would-be perpetrator, information is readily available online discussing the propagation and dissemination of plant pathogenic microorganisms (Champoiseau and Momol 2008; Sullivan et al. 2011). Besides intentional introduction of plant pathogens, non-intentional dispersion of plant pathogens occurs.

Factors contributing to unintentional pathogen introduction include wind, rain, flooding and hurricanes (Aylor 2003). In addition to weather, insects play a role in dispersal of phytopathogens within a field (Brault et al. 2010; Backus et al. 2012). Most agricultural goods including plant and animal products are transported from state to state and country to country, increasing the likelihood of exotic pathogen introduction to the U.S. As mentioned previously, the high levels of genetic uniformity and high plant densities characteristic of modern cropping systems pose added risks of a pathogen(s) severely damaging or destroying an entire crop. These risks are compounded by the varying degrees of pathogen virulence, making it critical, with certain pathogens, to quickly identify pathovar, biovar, or race. Thus, diagnosticians rely upon plant pathogen detection and identification tools that are specific, as well as being rapid and inexpensive (Brault et al. 2010; Meyer 2003).

Many immuno- and nucleic acid based assays are available for detection of plant pathogens (Schaad et al. 2003). For rapid, inexpensive pathogen detection, immunoassays such as enzyme-linked immunosorbance (ELISA) and immune-strip tests can be used but these often lack sensitivity required for a biosecurity application. In contrast, nucleic acid based tests such as the polymerase chain reaction (PCR) and multilocus PCR offer the high degree of sensitivity required for biosecurity applications, but are limited in the total number of pathogens they will detect (Postnikova et al. 2008). Ideally, a biosecurity assay could quickly detect any and all classes (prokaryote, eukaryotes, and viruses) of pathogens, including unknowns, in a given sample, at a degree of sensitivity comparable to that of nucleic acid based detections. Such approaches have been applied to the detection of known and unknown plant viruses in mammals, insects and plants (Adams et al. 2009, Roossinck et al. 2009, Cox-Foster et al. 2007; Palacios et al. 2008), leading Stobbe et al. (2013) to hypothesize that metagenomics combined with NGS has the potential to be used as a plant pathogen detection tool.

Current NGS technologies produce enormous amounts of data, which, depending on the methods used to gather DNA, can contain the genomic profile of all organisms in a given sample in their natural environment (Chen and Pachter 2005). Together; the advances in metagenomics and NGS will assist biosecurity agencies in lessening the risks of disease outbreak from exotic and native plant pathogens by providing a powerful screening tool. Combining these two technologies will facilitate development of a plant pathogen detection system that will meet current needs and future needs. To achieve this, a partnership between academia and the government has been made.

The National Institute for Microbial Forensics & Food and Agricultural Biosecurity (NIMFFAB) at Oklahoma State University partners with U.S. and international agricultural biosecurity entities to address current and future biological threats to crops and food safety. In one initiative, researchers at NIMFFAB and the Foreign Disease and Weeds Research Laboratory of the United States Department of Agriculture-Agricultural Research Service (USDA-ARS), are collaborating to develop novel technologies to monitor, detect and identify plant or foodborne pathogens in complex samples. As a whole, the project addresses bacterial, viral, and fungal pathogens, but the research presented in this thesis focuses on the prokaryotic plant pathogens.

The objectives are as follows:

1. To create bioinformatic pipelines, streamlined computer programs, for mock sample database generation used in simulating 454 sequencer runs, query using specifically designed "electronic probes," and BLAST searches.
    a. *Vitis vinfera* (wine grape) was used as a host for mock database development
    b. *Xanthomonas oryzae* pv. *Oryzae* PXO99A, *Ralstonia solanacearum* GMI1000 as a substitute for *R. solanacearum* r3b2, *Candidatus* Liberibacter asiaticus psy62, and *Xylella fastidiosa* 9a5c (8.1b) were used as targeted bacteria

2. To demonstrate the ability to use metagenomics methodology combined with NGS and electronic probes to identify targeted bacterial plant pathogens from raw sequence data.
    a. Inoculations of potato plants with *R. solanacearum* r3b2 will be done by USDA-ARS at their onsite BSL-3 facility
    b. Inoculations of tomato plants with *P. syringae* pv. *tomato* str. DC3000 (DC3000) were done by NIMFFAB. The addition of DC3000 was due to ready availability to the bacterium and host.

LITERATURE CITED

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

Aylor D. 2003. Spread of plant diseases on a continental scale: role of aerial dispersal of pathogens. Ecology 84:1989-97.

Backus E, Andrews K, Shugart H, Carl G, Labavitch J, Alhaddad H. 2012. Salivary enzymes are injected into xylem by the glassy-winged sharpshooter, a vector of *Xylella fastidiosa*. J Insect Physiol. 58:949-59.

Brault V, Uzest M, Monsion B, Jacquot E, Blanc S. 2010. Aphids as transport devices for plant viruses. Comptes Rendus Biologies 333:524-38.

Champoiseau P, Momol T. 2008. Bacterial Wilt of Tomato. In: Florida UO. The United States Department of Agriculture - National Research Initiative Program.

Chen K, Pachter L. 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol 1:10.1371.

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

Food and Agriculture Organization of the United Nations (FAO). 2003. Biosecurity in food and agriculture. Report on the 17th session of the committee on agriculture. COAG/2003/9.

Guy H, Ministry for Primary Industries (MPI), New Zealand. 2013. Ministerial statement of responsibility: Statement of intent 2013-2018. C.5 SOI (2013) SOI.

Harl, Neil E. 2002. U.S. Agriculture, food production is threatened by bioterrorism attacks. Ag Lender. http://www.econ.iastate.edu/~harl/USAgThreatened.pdf.

Meyer J. 2003. Insect vectors of plant pathogens. In.: Department of Entomology North Carolina Stae University. http://www.cals.ncsu.edu/course/ent425/text18/plantvectors.html.

Office of Science and Technology Policy (OSTP), White House. 2013. Biosecurity. www.whitehouse.gov/administration/eop/ostp/nstc/biosecurity.

Parker, H. S. 2003. Agricultural Bioterrorism: A Federal Strategy to Meet the Threat. McNair Paper 65. Washington, D.C., Institute for National Strategic Studies, National Defense University.

Palacios G, Druce J, Du L, Tran T, Birch C. 2008. A new Arenavirus in a cluster of fatal transplant-associated diseases. N.E. J. Med. 358:991-998.

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

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

Schaad N, Abrams J, Madden L, Frederick R, Luster D, Damsteegt V, Vidaver A. 2006. An assessment model for rating high-threat crop pathogens. Phytopathology 96:616-621.

Sullivan M, Daniells E, Southwick C, 2011. CPHST Pest Datasheet for *Xanthomonas oryzae* pv. *oryzae*. USDA-APHIS-PPQ-CPHST.

Waage J, and Mumford J. 2008. Agricultural biosecurity. Philosophical Transactions of the Royal Society B: Biological Sciences 363:863-876.

CHAPTER II

LITERATURE REVIEW

History: Vulnerability of the United States to terrorism and biocrimes

The September 11, 2001, attacks on the World Trade Center and the Pentagon and subsequent anthrax mail attacks and, more recently, the Boston Marathon bombing on April 15, 2013, demonstrated the vulnerability of the United States to acts of terrorism (Flynn 2002; Comfort and Kapucu 2006; Speckhard 2013). The September 11th events and the Boston Marathon bombing were the result of a few radical individuals; these, together with the anthrax incident, show that mass economic and civilian casualties can result from the actions of only one or a few individuals. In fact, the latter crime was attributed to an American scientist who had passed stringent governmental clearances. Such events indicate that individuals or non-state groups can bypass tactical methods of traditional warfare and use unpredictable and increasingly psychologically devastating approaches that undermine governments, creating a sense of insecurity for citizens of the targeted nation or region (Bradley et al. 2004; Blendon et al. 2002; Miller et al. 2013).

Security implementations

As a result of the September 11th World Trade Center/Pentagon and anthrax incidents, the U.S. government implemented new security programs to identify weaknesses in America's critical infrastructures and to make changes necessary to reduce significantly the chance that a future attack would be successful (Shawn 2004). The U.S. Department of Homeland Security and other biosecurity agencies around the globe are tasked with the responsibility of identifying threats and weakness within their national infrastructures and insuring the continuing growth and longevity of their respective economies, while protecting their citizens. However, even with the U.S. government's implementation of new and more stringent security procedures, the U.S. agriculture sector continues to be vulnerable to both direct and indirect threats.

The agricultural industry, which includes animals, food crops, forestry, range lands and water resources, provides opportunity for addressing biosecurity concerns due to the lack of security and surveillance systems and to the enormous amount of land that this industry uses (Harl 2002). Enhanced monitoring and screening or surveillance of plant and animal samples is critical in maintaining a robust biosecurity program (Bunn et al. 2011; Fisher et al. 2012; Guy 2013).

Bioterrorism

Bioterrorism is the threat or intentional release of biological agents with the goal of generating fear, intimidation, or harm to a population or specific group for religious, political, and/or economic purposes (Budowle 2005). The ultimate objective is to undermine a government or to achieve personal objectives by releasing microorganisms, toxins, or other deadly bio-organisms (ADHS 2012; Budowle 2005). In the United States, biological agents used in bioterrorism acts against humans are separated into three main categories, according to the Centers for Disease Control and Prevention (CDC 2010): A, high priority, B, moderate priority, and C, low priority. Category A agents are infrequently observed in the United States and are considered a "national security risk". They are transmitted from one person to another, have an elevated death rate and create social unrest. Category B agents can be disseminated fairly easy and have a slightly less ability to cause illness and death than category A. Lastly, Category C agents are the third highest priority due to their ease of dissemination and propagation. Additionally, category C agents can cause illness and death but not comparable to those in categories A and B (CDC 2010). While bioterrorism is typically defined as a direct attack on a government and/or its citizens, this definition fails to consider other forms of bioterrorism that have equal potential to cause civil unrest or the continual evolution of national security needs.

Agroterrorism

Agroterrorism is the intentional introduction of a plant or animal pathogen for the purpose of undermining government stability, generating fear, or causing economic losses and social instability (Monke 2007). Within the United States Department of Agriculture (USDA), the Animal and Plant Health Inspection Services (APHIS) is given the responsibility for implementing the Agriculture Bioterrorism Protection Act of 2002, which provides guidelines for determining agriculture select agents and toxins (APHIS 2008). Agriculture and veterinary select agents and toxins are those that are determined to have a potential to pose a severe threat to plant health or plant products, or animal health or animal products (APHIS 2008). Considerations for classifying an agent or toxin as an agriculture select agent includes; effects from exposure of agents or toxins to marketability and production of plant or animal products; the pathogenicity of the agent or toxin and the methods it is transferred to animals or plants; the ability to treat and prevent illnesses caused by the agent or toxin; and any additional criteria the USDA Secretary deems important for protection of animal or plant health, or animal or plant products (APHIS 2008). The agricultural select agents list includes *Bacillus anthracis*, *Ralstonia solanacearum*, *Enterovirus* 71, *Hendra virus* and many others (Federal Register 2012). Egypt, Iran, North Korea, Syria, and the United States, along with many other countries are thought to have or had a history of biological weapons programs dedicated to the development of agents for the offensive purpose of agroterrorism and according to the Biological Weapons Convention, under the regulation of the United Nations, biological warfare programs are prohibited (MIIS 2009; UN 2012).

An agricultural attack generally has several key objectives: decreasing food output for both human and animal consumption, significant national and/or global economic losses relating to the agriculture industry and forestry lands, possible export/import trade embargoes, and undermining governments by instilling a lack of confidence in the safety of the food supply (CIDRAP 2010). As seen in the 2008 Middle East food riots, the current world economic instability, along with high food prices, has the potential to escalate tension among nations and destabilize weakened governments (McMichael 2009). This issue is compounded by the presence of endemic plant pathogens in crops, which farmers must address in order to maintain a profitable economic threshold. A purposeful introduction of new plant pathogens poses a significant risk due to there being a lack of natural suppression factors and possible resulting in an uncontrolled disease outbreak (Schwartz et al. 2006).

The use of a plant pathogen as a weapon is usually health-risk free for the perpetrators because, unlike human and zoonotic pathogens, most plant pathogens are harmless to humans. Furthermore, there are numerous plant pathogens, which cause various diseases, giving a perpetrator the opportunity to sequester, propagate and disperse or engineer agriculturally devastating bacterial, fungal or viral strains while leaving little to no evidence (RAND 1999). Having very little evidence makes attribution of a perpetrator extremely difficult.

History of bioterrorism

The historical use of bioterrorism tactics dates back to ancient times; however, more recent events have occurred in the U.S. (Abbott 1990; Breeze 2004; Johnson 2013). In a 2003 report by the U.S. Government Accountability Office, Bhagwan Shree Rajneesh was the first person to commit a biocrime on U.S. soil and an example of the danger of an enemy within. The cult leader settled with many followers in Wasco County, Oregon in the 1980's (Abbott 1990). After disputes arose among local officials, in an effort to sway political outcomes, cult members introduced *Salmonella* to several local restaurant salad bars in hopes of affecting an upcoming election. This crime resulted in 750 persons becoming ill (Dyckman 2003). More recently, in 2003, an employee of a Michigan supermarket purposefully introduced Black Leaf 40, an insecticide for sucking insects on plants, and for lice and mites on chickens, leading to illness in approximately 100 individuals (CDC 2003). Fortunately, in both cases the contamination was contained to a small region. Together, the examples illustrate the vulnerability of America's agriculture industry even in areas protected by security measures more stringent than those in place for field crops, and how the actions of a few individuals can cause physical and psychological anguish to hundreds and possibly thousands of people.

In the event that individuals or groups, foreign and/or domestic, purposely introduced either enteric-human or plant pathogens into U.S. food plants , the consequences have the potential to cause harm, as noted by the Gilmore Commission in a 1999 report to the President and Congress which stated, "…concerted biological attack against an agriculture target offers terrorists a virtually risk-free form of assault, which has a high probability of success and which also has the prospect of obtaining political objectives, such as undermining confidence in the ability of a government or giving terrorists an improved bargaining position" (RAND 1999). Whether considering unintentional food contamination, natural disease outbreaks, or bioterrorist acts it is critical to have methodologies in place that are thoroughly validated in pathogen detection and identification. By having certified protocols for detection of relevant pathogens, response and recovery time is greatly reduced.

To lessen the risk posed by bacterial select agents and non-select agents the Foreign Disease and Weeds Research Laboratory of the United States Department of Agriculture-Agricultural Research Service (USDA-ARS) and the National Institute for Microbial Forensics &

Food and Agricultural Biosecurity (NIMFFAB) at Oklahoma State University, are collaborating to develop novel technologies to monitor, detect and identify bacterial plant pathogens in complex samples. Bacterial pathogens of interest include *Xylella fastidiosa* 9a5c, *Xanthomonas oryzae* pv. *oryzae*, *Ralstonia solanacearum* race 3 biovar 2 and *Candidatus* Liberibacter asiaticus (Table 2).

The bacterial pathogens were chosen based on the availability of the genome or expressed sequence tags (ESTs), which are short (500 -800 nt) sub-sequences of cDNA; the economic importance of the pathogens, select agent status; and the availability to be propagates at the USDA-ARS (Fort Detrick, Maryland) containment facility.

Bacterial Plant Pathogens

Xylella fastidiosa 9a5c

The disease citrus variegated chlorosis (CVC) affects a variety of citrus species (Pooler 1995; Brlansky et al. 2008; Redak et al. 2004). The causal organism, *Xylella fastidiosa*, was classified by the United States Department of Agriculture, Animal and Plant Health Inspection Service (APHIS) as a select agent; however, in 2012 it was removed from the list. The decision to remove *X. fastidiosa* was based on the potential of the bacterium to cause mass causalities or devastating effects on the economy, critical infrastructure, or public health (Federal Register 2012).Additionally, evaluations of the bacterium assessing morbidity and mortality, low infectious dose, availability of countermeasures, and risk of deliberate misuse including historical documentation of weaponization were performed by experts who study the bacterium (Federal Register 2012). The principle strategies in place for controlling this pathogen are introduction prevention and development of cost effective early detection and identification systems (Ancona et al. 2010; Brlansky et al. 2008). *X. fastidiosa* 9a5c is a fastidious, Gram-negative, xylem-limited bacterium phenotypically identical to other strains of *X. fastidiosa* (Hartung et al. 1994). The host range includes plum, almond, coffee, oak, citrus, peach, oleander, and grapevine. Transmission in the U.S. occurs by various xylem feeding insects including, most notably, the glassy-winged sharpshooters (*Homalodisca vitripennis*) and blue-green sharpshooters (*Graphocephala atropunctata*) (Chatterjee et al. 2008).

There are three primary steps involved in vector-to-plant transmission of *X. fastidiosa* (Janes and Obradovic 2010). After the xylem feeding sharpshooter ingests the bacterium from an infected plant, *X. fastidiosa* attaches to the lining of the vector's foregut. Finally, the vector feeds on a new host plant, inoculating it with *X. fastidiosa* and completing the transmission cycle.

With many vector borne bacterial plant pathogens, the first and final steps of ingestion and transmission to the host are active processes separated by a latent period during which the bacteria multiply; however, *X. fastidiosa* does not need a latent period and is termed a foregut-borne pathogen (Nault 1997; Purcell and Finlay 1979). Transmission also occurs months after initial bacterial acquisition (Hill and Purcell 1994). The latent period is due to *X. fastidiosa* colonizing and forming biofilms, made up of extracellular polymeric substances (EPS), within the vector foregut, hours after initial acquisition (Lorite et al. 2013; Marques et al. 2002). Attachment within the sharpshooter or plant is mediated by two forms of pili; the short Type I and the longer Type IV. Type I pili are necessary in bacterial attachment and biofilm formation, which enhance bacterial survival. Type IV pili, which are clustered at just one pole of the cell, facilitate upstream translocation within the plant xylem vessels via twitching motility (Chatterjee et al. 2008; Meng et al. 2005).To release bacteria contained in a biofilm inside the vector foregut, salivary enzymes EGase and other cell-wall degrading enzymes loosen the matrix allowing bacteria to pass into the plant as the insect feeds (Backus et al. 2012). Only a few bacteria are needed for transmission to occur (Hill and Purcell 1995). This represents the highly infectious nature of *X. fastidiosa* when it is established in a vector and its ability to damage and/or destroy large amounts of crop as the vector moves from one leaf/plant to another.

Xanthomonas oryzae* pv. *oryzae

Bacterial leaf blight (BLB) or bacterial blight (BB) and bacterial leaf streak (BLS) are major diseases of rice (*Oryzae sativa*) around the globe. The causal organisms are, *Xanthomonas oryzae* pv. *oryzae* and *Xanthomonas oryzae* pv. *Oryzicola.* The casual organism, *X.oryzae* pv. *oryzae,* is classified by APHIS as a select agent. As with all plant pathogenic microbial select agents, the principle management strategy is prevention of introduction into the U.S. through all borders and ports. *X. oryzae* is a yellow, slime-producing, Gram-negative rod that translocates throughout the plant vascular tissue after infection. Two closely related pathovars, *oryzae* and *oryzicola,* are similar in many aspects, but pv. *oryzae* (Xoo) causes BB by colonizing plant vascular tissues while pv. *oryzicola* infects parenchyma cells, causing BLS (Nino-Liu et al. 2006). The typical mode of entry is through stomata on the leaves and wounds on the stems and roots (Ou 1985). Secondary inoculum consists of bacteria that ooze from the hydathodes, where they congregated, and are exuded onto the leaf surface (Mew et al. 1993). *X. oryzae* is transmitted primarily by rain, wind, and flooding. Natural movement of the pathogen is limited to short distances; however human influence such as the movement of infected seeds has led to distant outbreaks of disease (Hsieh et al. 1974). Current distribution of the pathovars is illustrated in Table 1.

While *X. oryzae* is found around the globe where rice is grown, in some countries its distribution is limited to a particular region. For example, in Australia Xoo is found only in Northern Territories and Queensland, while in Asia *X. oryzae* pv. *oryzicola* is limited to tropical areas. Because it causes major diseases of rice, a staple food around the world, for which there is worldwide demand, especially among Asian countries. *X. oryzae* poses significant risk to crop security across the globe.

Ralstonia solanacearum

Previously known as *Pseudomonas solanacearum*, *Ralstonia solanacearum* is a motile, soil borne, Gram negative, rod-shaped bacterium with polar flagella (Denny 2006). The bacteria enter the plant through wounds below ground caused by nematodes and cracks in lateral root emergence, and quickly move to the aerial parts of the plant through the vascular system (Mansfield et al. 2012). *R. solanacearum* is disseminated primarily in the soil through contaminated water sources, infected planting material, contaminated equipment and personnel (Janse 1996).

R. *solanacearum*, as a species, has a very broad plant host range and causes wilting diseases in over 450 plant species in tropical, subtropical and warm temperate regions (Genin and Boucher 2004; Hayward 1991). The bacteria are divided into subcategories based on their host range (five races) and their biochemical utilization patterns (up to five biovars) (Table 3). Because of its broad host range, *R. solanacearum* has a high potential of invading uninfected regions through trade and interstate or local commerce (Champoiseau et al. 2010; CABI), and has been labeled the most destructive plant pathogen with damages reaching over $1 billion in global losses each year (Mansfield et al. 2012; Elphinstone 2005). The fact that most *R. solanacearum* strains do not travel long distances keeps infection areas mostly localized. Long distance movement requires unnatural intervention (man-made transportation).

Of the five *R. solanacearum* races, race 3 biovar 2 (r3b2), which is classified by the USDA-APHIS as a select agent, is unique in its ability to tolerate cooler temperatures and higher altitudes and was first officially identified in the Netherlands in 1992 (Jansen 1996; Messiha et al. 2009). Even before it was classified a select agent; this bacterium was considered a risk for use as a bioterrorism agent (Lambert 2002). *R. solanacearum* r3b2, which causes brown rot of potato and tomato, was documented as entering the United States in 1999, 2000, 2003 and 2004 through importation of infected geranium cuttings from Africa, Central America, Kenya and Guatemala (Champoiseau et al. 2010). Successful *R. solanacearum* r3b2 eradication procedures were performed in each case, and as of 2013 no widespread outbreaks of the pathogen within the U.S. borders has been reported.

The virulence of *R. solanacearum* is based upon three primary factors: extracellular polysaccharides, a Type III secretion system and Type IV pili. Other important factors include cell wall degrading enzymes, oxidative stress genes, and quorum sensing (Schell 2000; Flores-Cruz and Allen 2011). The most important of these is exopolysaccharides, which clog and colonize the plant vascular system and are used by the bacteria as a barrier against host defenses (Saile et al. 1997; Milling et al. 2011). Exopolysaccharide mutants were found to be non-pathogenic *in vivo* and *in planta*, while a non-mutants colonized plant tissues, leading to wilting (Araud-Razou et al. 1998). The Type III secretion system (T3SS), which moves bacterial effector proteins into the host cell, is required for both disease and the hypersensitive response in susceptible plants, (Cornels and Gijsegem 2000). When the *hrp* genes encoding the T3SS are silenced, the bacteria become non-pathogenic (Buttner and Bonas 2002). The *R. solanacearum* type IV pili generate twitching motility during initial invasion and colonization (Tans-Kersten et al. 2001and Liu et al. 2001). Non-motile mutants (lacking type IV pili) failed to cause measurable disease; however, when the same non-motile mutants were injected directly into tomato plant tissues disease presented similarly to the wild type (Tans-Kersten et al. 2001).

***Candidatus* Liberibacter asiaticus**

The disease commonly referred to as citrus greening or citrus huanglongbing (HLB) is attributed to three species of fastidious, phloem-limited Gram-negative alpha-proteobacteria having a worldwide distribution. *Candidatus* Liberibacter americanus, *Candidatus* Liberibacter africanus, and *Candidatus* Liberibacter asiaticus together are responsible for damaging citrus crops around the globe, leading to the destruction of millions of citrus trees (Bove 2006). Of the three species, only *Ca*. L. asiaticus was identified in the U.S., where it was found by both PCR and next generation sequencing in both symptomatic and non-symptomatic citrus leaf tissues (Sagaram et al. 2009; Tyler et al. 2009). Disease symptoms, which are often non-uniform on the tree, include blotchy mottling of leaves with varying shades of yellow and green, and small and disfigured fruit. The tree canopy can range from full foliage to none, depending on disease severity. Damage to growers from HLB is caused by poor fruit yields, short tree life-span and unmarketable, small, disfigured fruits. The citrus crop in Florida alone is valued at $9 billion dollars annually. The primary means by which *Ca*. L. asiaticus is spread from one citrus plant to another in the U.S. is by the vector *Diaphorina citri*, the Asian citrus psyllid (ACP). The other known vector of HLB, not found in the U.S., is *Trioza erytreae,* the African psyllid.

Because 60-100% of ACP acquire the HLB bacteria during nymphal stages and up to 40% as adults, all life stages are a concern to growers and pose a threat to the citrus industry (Pelz-Stelinski et al. 2010). Once ACP acquires *Ca*. L. asiaticus it is maintained for up to12 weeks, which is very close to the insects' 90 day lifespan (Hung et al. 2004). There are conflicting reports about whether *Ca*. L. asiaticus propagates within the ACP. Inoue et al. (2009) exposed ACP fifth instar nymphs to *Ca*. L. asiaticus for 24 hours; qPCR at days 10, 15, and 20 revealed a 25, 360 and 130 fold increase, respectively, in *Ca*. L. asiaticus titers compared to day 1. But, Pelz-Stenlinski et al. (2010) found that *Ca*. L. asiaticus titers in adult ACPs decreased over time as the insects fed on healthy plants. There is agreement, however, that nymphal ACPs are the principle means of spreading the pathogen, suggesting that early application of integrated pest management (IPM) strategies can be instrumental in maintaining profitability for U.S. citrus growers.

Pseudomonas syringae pv. *tomato* str. DC3000 (DC3000)

Pseudomonas syringae are rod shaped and Gram-negative plant pathogenic bacteria with polar flagella. The current 50 pathovars are divided into races based on their degree of host specificity (Gardan et al. 1999). Diseases caused by *Pseudomonas syringae* include bacterial speck of tomato, brown spot of bean, blight of soybean and canker of kiwi. Of the various pathovars and strains, *P. syringae* pv. *tomato* str. DC3000 (DC3000) is the most commonly used to study virulence mechanisms in both model systems (*Arabidopsis thaliana*) and commercially relevant crops such as tomato.

Dissemination of DC3000 occurs by animals, people, insects, agricultural tools, soil particles and contaminated water (Bashan 1986). DC3000 will persist from one season to the next in crop debris and within weeds such as nightshade and groundcherry (Davis et al. 2008). Historically, DC3000 has played a unique and vital role in understanding basic virulence mechanisms of plant pathogenic prokaryotes.

During the 1980’s little was known about the pathogenicity genes of phytopathogenic bacteria. In Cuppels (1986) reported that *Pseudomonas syringae* pv. *tomato* strain DC3000 could be transformed with a rifampicin-resistance gene (Cupples 1986). Because of Cupples work, DC3000, a genetically modified bacterium, is now one of several model bacteria used universally for molecular interaction studies and also makes a good surrogate for an agroterrorism agent. For organisms that could be used as agroterrorism agents Schaad et al. (2006) list a rating criteria. From the criteria, DC3000 meets the following; produces toxins, able to be manipulated, targets multiple host, easy to propagate and disseminate, lack of chemical control and has a high degree of virulence; which makes DC3000 an acceptable surrogate (Schaad et al. 2006). Additionally, DC3000 is not listed by APHIS as a select agent and does not require special permitting as needed for select agent organisms.

Plant Pathogen Detection Systems

Plant pathogens are detected by a wide range of assays (Schaad et al. 2003). Immunoassays, such as enzyme-linked immunosorbance assay (ELISA) and immune-strip tests, and nucleic acid based assays, such as real time polymerase chain reaction (rtPCR) or DNA microarray hybridization, are popular methods for plant pathogen detection. The former offer quick, inexpensive, means of detection but lack the sensitivity required for biosecurity and forensics applications, while the latter, such as rtPCR or end-point PCR (PCR), offer a degree of sensitivity required for biosecurity applications but are limited in the total number of pathogens they can detect (Postnikova et al. 2008). Both immunoassays and nucleic acid based methods require pre-characterization of a targeted pathogens proteins or genomic sequences for detection, which makes it very difficult to detect uncharacterized plant pathogens. Being able to detect multiple pathogens at the same time in a quick and cost effective manner is another limitation of current pathogen detection systems. DNA microarrays, SSRs, and MSLTs are all capable of detecting multiple pathogens, but require previous characterization and are still limited in the total number of pathogens they will screen. Additionally, all of these methods consume the original starting sample, leaving limited opportunity to reuse the material to search for additional pathogens. Having a single method for use on any plant material, for detection of any and all plant pathogens simultaneously, will greatly reduce the time it takes biosecurity agencies or local diagnostic labs to identify a pathogen and limit the spread of a disease. Recent methods/technologies of metagenomics and next generation sequencing show promise as plant pathogen detection tools.

Metagenomics

The new field of metagenomics emerged in the late 1990’s as an exciting example of how new technology leads to innovation (Handelsman et al. 1998). Chen and Pachter (2005) define metagenomics as the use of current genomic techniques to study communities of microorganisms in their natural environment, bypassing the need for isolation and cultivation of individual species. By this definition, metagenomic analysis differs from traditional detection approaches in that an entire microbial community is characterized simultaneously, offering an opportunity to discover unknown organisms and fastidious bacteria, obligate fungi and viruses that cannot easily be detected or isolated *in vivo*. Metagenomic sampling provides a true representation of a microbial environmental community at a particular moment in time, much like taking a photograph, except metagenomics captures a genomic snap-shot.

Metagenomic studies begin by extraction of total nucleic acid from an environmental sample and sequencing it by next generation sequencing (NGS) technology, from which sequences are used to build a sample-sequence database (SSD) or genomic library. NGS is a term used to describe various platforms (Ion Torrent, 454 pyrosequencing, Illumina, and SOLiD) that produce millions of nucleotide sequences concurrently in an ultra-high-throughput process. In the metagenomic approach, characterization of this entire nucleic acid library provides insight on ecology, evolution, and function, enzymatic proteins and antibiotic characteristics (Anonymous 2007). More importantly, metagenomics coupled with NGS will lead to the design of innovative tools for detection and identification of all classes of plant pathogenic microorganisms in a single assay.

Traditionally, a single organism would be isolated and propagated from an environmental sample and then sequenced, resulting in a nearly complete genome that may allow identification of the species. Genes can be annotated with a certain degree of confidence, and, if needed, Koch’s postulates can be performed to establish a causal relationship between a microbe and a disease. However, there are disadvantages to the traditional sequencing approach. Not all organisms can be isolated or grown in culture. There are costs in media prep for propagation and equipment to obtain optimal microbial growing conditions. Because not all microbes are able to be propagated in a lab setting, we do not obtain a true representation of everything going on in a particular sample. Additionally, there is a cost in time to prepare the isolate for sequencing, which could lead to a pathogen outbreak and cause economic losses for a grower.

Metagenomics, as an alternative to traditional sequencing, offers the advantage that microorganisms do not need to be isolated and propagated. Because the genomes of all organisms are included in the sequencing, we are able to capture unknowns and fastidious organisms. The cost is moderated by the lack of a need to propagate the sample; nucleic acid extractions are made directly from the sample. Additionally, we capture biochemical pathways not yet known.

There are disadvantages to using metagenomics. In traditional sequencing we are able to thoroughly characterize an organism; however, with metagenomics we only capture short fragments. Having small fragments makes it difficult to thoroughly characterizing a single microbe with a high degree of confidence. Because a metagenome is made up multiple organisms, there will likely be genomic information not yet known; therefore assembling an entire genome for a singular organism is not possible. Even with the disadvantages, the ability to capture genomic information for all organisms in a sample provide opportunities for detection of plant pathogens both known and unknown. Aside from a specific use in pathogen detection, metagenomics plays a critical role industrial sustainability.

There are three primary industries using metagenomics for long-term sustainability and discovery of novel compounds. The medical biotechnology industry, plant or agriculture biotechnology industry and all other industries not covered by medical or agricultural biotechnologies. Together, the industries are estimated to spend $3.74 billion by 2015 on enzymes (protease and carbohydrates) used in detergents, food applications, agriculture, textile, pharmaceuticals and many others (Anonymous. 2013; Lorenz and Eck 2005).

Current public and political awareness of climate changed and globalization of economies has led to the demand of lessening the environmental impact and improved sustainability for all industries across the globe. To meet this demand, metagenomics is being used to explore environmental communities of microorganism in hopes of discovering novel biocatalyst (Lorenz et al. 2002). Considering the 2015 estimate of industry spending billions on enzymes, the continual demand of clean and sustainable resource, and need for novel plant pathogenic detection tools, metagenomics methodology will continue to have a critical role in the future.

Next generation sequencing

NGS is a relatively recent technology that allows for the generation of huge amounts of sequence data from a given sample (Ronaghi 2001). NGS is one of three different sequencing technologies, commonly referred to as first, second and third generation sequencing used in today's research, each with their own set of advantages and disadvantages. First generation sequencing technology or Sanger sequencing works by fragmenting DNA and inserting it into plasmids, which are then cloned to produce enough starting material for the sequencer (Sanger et al. 1977; Sanger and Coulson 1975). For sequencing, a dye-termination method is used, which allows quick sequencing of one reaction by having different light wavelengths for each ddNTP. Advantages of first generation sequencing include; large genome fragments with less total volume of data that allows for assembly of a genome at lower costs as compared to other sequencing technologies. Disadvantages to Sanger sequencing include; speed limitations, cloning bias or the inability to clone certain genes, and issues with incorporating repeat regions, both of which result in an incomplete genome (Sorek et al. 2007; Wooley et al. 2010).

Second generation sequencing is commonly used today and referred to as next generation sequencing (NGS) and is far more productive than traditional Sanger sequencing (Pop and

Salzberg 2008; Magi et al. 2010; Metzker 2010). There are multiple NGS platform technologies that differ in read length (20 nt to approximately 1000 nt) and number (100,000 to 1 million), which combine to generate a range of overall sequence data (Tucker et al. 2009). The particular second generation sequencing technology described here is the 454 pyrosequencing by Roche Applied Sciences. Margulies et al. (2005) discuss the workflow of 454. Briefly, the first step is to randomly fragmenting genomic DNA and attaching adapters witch then bind to beads for emulsion PCR. Upon emulsion completion there are millions of unique DNA copies, which are denatured to make single stranded DNA. These fragments are placed into wells where they are mixed with enzymes for pyrophosphate sequencing. The final step includes a series of nucleotide flushes followed by washes, which work in unison as a massive and parallel sequencing reaction. Thousands of individual DNA fragment are being sequenced during this last phase. The combination of NGS technology and metagenomics offers many advantages over first generation sequencing.

When combining NGS sequencing technology with environmental samples, also known as metagenomics (see above), a highly processive form of shotgun sequencing in which any and all nucleic acids in a sample are potential candidates for sequencing templates is observed (Jones 2010; Tyson et al. 2004). This methodology has been applied to several types of environmental samples including seawater, bilge water, marines, intestinal tracts of various animals and contaminated water sources (Tyson et al. 2004; Daniel 2005; Breitbart et al. 2003; Gill et al. 2006; Tringe and Rubin 2005). In theory, NGS combined with metagenomic methodology could be applied to disease diagnostics as a means to search for unknown pathogens. Similar, combined, NGS and metagenomics approaches have been applied to the detection of known and unknown plant viruses in mammals, insects and plants (Adams et al. 2009, Roossinck et al. 2009, Cox-Foster et al., 2007; Palacios et al., 2008).

The amount of data (400MB – 28GB) produced from an NGS run is computationally demanding and requires computer clusters to assemble (Metzker 2010; Reis-Filho 2009). The cost of building and maintaining computer cluster systems can be too high for many research labs. A metagenomic approach used for pathogen detection will contain a majority of sequence from the host, which results in pathogen sequence making up a small percentage of the total reads (Roossinck et al. 2009; Adams et al. 2009). For a diagnostician, the host sequences that make up most of a plant metagenome sample are essentially irrelevant. What is important for diagnostician is the ability to capture a genomic overview of everything in the sample, which can then be screened for the presence of pathogen genomic fragments. The positive identification of pathogen genomic material obtained from a NGS run would result in confirmation of a pathogen being present.

LITERATURE CITED

Abbott C. 1990. Utopia and bureaucracy: The fall of Rajneeshpuram, Oregon. Pacif Hist Rev 59:77-103.

Adams I, Glover R, Monger W, Mumford R, Jackeviciene E, 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Pl Pathol 10:537-545.

Ancona V, Appel D, Figueiredo P. 2010. *Xylella fastidiosa*: A model for analyzing agricultural biosecurity. Biosecurity and Bioterrorism: Biodefense strategy, practice, and science. Mary Ann Leibert, Inc. http://online.liebertpub.com/doi/pdf/10.1089/bsp.2009.0021

Animal and Plant Health Inspection Service (APHIS), United States Department of Agriculture. 2008. Rules and regulation. Federal Register. 7 CFR 331; 9 CFR 121. APHIS-2007-0033.

Animal Health and Plant Inspection Service (APHIS), United States Department of Agriculture. 2012. Agriculture select agent. www.aphis.usda.gov/programs/ag_selectagent/ag_bioterr_toxinlist.shtml

Anonymous. 2007. The New Science of Metagenomics: Revealing the secrets of our microbial planet. The National Academies Press.

Anonymous. 2013. Industry enzymes: A global strategic business report. Global Industry Analysts, Inc. http://www.strategyr.com/Industrial_Enzymes_Market_Report.asp

Araud-Razou I, Vasse J, Montrozier H, Etchebar C, Trigalet A. 1998. Detection and visualization of the major acidic exopolysaccharide of *Ralstonia solanacearum* and its role in tomato root infection and vascular colonization. Eur J Plant Pathol 104:795-809.

Arizona Department of Health Services (ADHS). 2012. Bioterrorism, November 23, 2012. Available from http://www.azdhs.gov/phs/emergency-preparedness/bioterrorism/

Backus E, Andrews K, Shugart H, Carl G, Labavitch J, Alhaddad H. 2012. Salivary enzymes are injected into xylem by the glassy-winged sharpshooter, a vector of *Xylella fastidiosa*. J. Insect Physiol 58:949-59.

Bashan, Y. 1986. Field dispersal of *Pseudomonas syringae* pv. *tomato*, *Xanthomonas campestris* pv. *vesicatoria*, and *Alternaria macrospora* by animals, people, birds, insects, mites, agricultural tools, aircraft, soil particles, and water sources. Can J Bot 64:276-281.

Blendon R, Benson J, DesRoches C, Pollard W, Parvanta C, Herrmann M. 2002. The impact of anthrax attacks on the American public. Med Gen Med 4:1-4.

Bove J. 2006. Hauanglongbing: A destructive, newley-emerging, century-old disease of citrus. J Plant Pathol 88:7-37.

Brlansky R, Davis M, Mizell R, Mossler M, Chang C, Fletcher J, Huber D, Petersend T, Vidalakis G, Wright G, Xiong Z, Caravetta J, Crocker R, Dixon W, Halber S, Meadows M, Schubert T, Kosta K, Polek M, Chand-goyal T, Rosenblatt D, Baker H, Dell D, Boratynski T, Brown L, Bulluck R, Divan C, Gomes P, Hernadez J, Levitt J, Levy L, Li W, Manson P, Varona E, Chen J, Damsteegt V, Hall D, Hartung J, Krueger R, Lee R, Smith K. 2008. Citrus variegated chlorosis caused by *Xylella fastidiosa* (CVC strain). Recovery plan.
http://www.ars.usda.gov/SP2UserFiles/Place/00000000/opmp/Citrus%20CVC%20110408.pdf

Breeze R. 2004. Biosecurity and bioterrorism: Biodefense strategies, practice, and science. Thomas Reuters. 11:1-14.

Breitbart M, Hewson I, Felts B, Mahaffy J, Nulton J. 2003. Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185:6220-6223.

Budowle B, Randall M, Ranajit C. 2005. Microbial Forensics: The Next Forensics Challenge. Int J Legal Med 119:317-330.

Bunn D, Beltran-Alcrudo D, Cardona C. 2011. Integrating surveillance and biosecurity activities to achieve efficiencies in national avian influenza programs. Prev Vet Med 98:292-4.

Büttner D, Bonas U. 2002. Getting across--bacterial type III effector proteins on their way to the plant cell. EMBO J 21:5313-5322.

CABI, Centre for Agricultural Biosciences International. Data Sheets on Quarantine Pests *Ralstonia solanacearum*: EPPO quarantine pest.

CABI, EPPO. 1997. Data sheets on quarantine pests; *Xanthomonas oryzae.* Prepared by CABI and EPPO for the European Union under Contract 90/399003.

Centers for Disease Control and Prevention (CDC). 2003. Nicotine poisoning after ingestion of contaminated ground beef - Michigan, 2003. MMWR Weekly 52:413-416.

———. 2005. Outbreaks of *Salmonella* infections associated with eating roma tomatoes - United States and Canada, 2004. MMWR Weekly 54:325-328.

———. Bioterrorism agents/diseases. Emergency preparedness and response.. 2010. Available from http://www.bt.cdc.gov/agent/agentlist-category.asp.

Champoiseau P, Jones J, Momol T, Pingsheng J, Allen C, Norman D, Harmon C, Miller S, Schubert T, Bell D, Floyd J, Kaplan D, Bulluck R, Smith K, Cardwell K. National Plant Disease Recovery System (NPDRS). 2010. Recovery plan for *Ralstonia solanacearum* race 3 biovar 2 causing brown rot of potato, bacterial wilt of tomato, and southern wilt of geranium.
Chatterjee S, Almeida RPP, Lindow S. 2008. Living in two worlds: The plant and insect lifestyles of *Xylella fastidiosa*. Phytopathology 46:243-71.

Center for Infectious Disease Research and Policy (CIDRAP). 2010. Overview of Agricultural Biosecurity. Regents of the University of Minnesota.

Chen K, Pachter L. 2005 Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol 1:10.1371

Comfort L, Kapucu N. 2006. Inter-organizational coordination in extreme events: The World Trade Center attacks, September 11, 2001. Nat Hazards 39:309-327.

Cornelis G, Van Gijsegem F. 2000. Assembly and function of type III secretory systems. Annu Rev Microbiol 54:735-74.

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

Cupp O. Shawn, Walker II D, Hillison J. 2004. Agroterrorism in the U.S.: Key security challenge for the 21st century. Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science 2:97-105.

Cuppels A. 1986. Generation and characterization of Tn5 insertion mutations in *Pseudomonas syringae* pv. *Tomato*. Appl Environ Bicrobiol 51:323-327.

Daniel, Rolf. 2005. The metagenomics of soil. Nat Rev Micro 3:470-478.

Daughtrey M. 2007. Southern bacterial wilt, caused by *Ralstonia solanacearum*. Cornell University. Department of Plant Pathology. Long Island Horticultural Research and Extension Services.

Davis, R, Davis U, Miyao G, Subbarao K, Stapleton JJ. 2009. IPM pest management guidelines: tomato. University of California Agriculture and Natural Resources. Publication 3470. http://ucanr.org/sites/ipm/pdf/pmg/pmgtomato.pdf

Denny, Tim. 2006. Plant-Associated Bacteria. Plant pathogenic *Ralstonia* species. Edited by S. Gnanamanickam: Springer Netherlands, 573-574.

Denny T, Hayward A. 2001. Laboratory Guide for the Identification of Plant Pathogenic Bacteria.3rd ed. APS Press, St. Paul p. 151-174.

Dyckman, Lawrence. 2003. Bioterrorism: A threat to agriculture and the food supply: Government Accountability Office (GAO).

Elphinstone, J. 2005. The current bacterial wilt situation: a global overview. In: Bacterial Wilt Disease and the *Ralstonia solanacearum* Species Complex. Edited by C. Allen, P. Prior and A. Hayward. St Paul, MN: APS Press. 9–28.

Federal Register. 2012. Agricultural Bioterrorism Protection Act of 2002; Biennial review and republication of the select agent and toxin list; Amendments to the select agent and toxin regulations; Final rule Department of Agriculture, Animal and Plant Health Inspection Service. 194. Docket APHIS-2009-0070.

Fisher M, Henk D, Briggs C, Brownstein J, Madoff L, Gurr S. 2012. Emerging fungal threats to animal, plant and ecosystem health. In. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited, 484:186-94.

Flores-Cruz Z, Allen C. 2011. Necessity of OxyR for the hydrogen peroxide stress response and full virulence in *Ralstonia solanacearum*. Appl Environ Microbiol 77:6426-32.

Flynn S. 2002. America the vulnerable. Council on Foreign Relations. Foreign Aff 81.60-74.

Gardan L, Shafik H, Belouin S, Broch R, Grimont F, Grimont PaD. 1999. DNA relatedness among the pathovars of *Pseudomonas syringae* and description of *Pseudomonas tremae* sp. nov. and *Pseudomonas cannabina* sp. nov. Int J Syst Evol Microbiol 49:469-78.

Genin S, Boucher C. 2004. Lessons learned from the genome analysis of *Ralstonia solanacearum*. Annu Rev Phytopathol 42:107-134.

Gill S, Pop M, Deboy R, Eckburg P, Turnbaugh P, Samuel B, Gordon J, Relman D, Fraser-Liggett C, Nelson K. 2006. Metagenomic analysis of the human distal gut microbiome. Science 312:1355-9.

Gizjen M. 2008. Diane Cuppels and the history of *Pseudomonas syringae* pv. *tomato* DC3000. IS-MPMI Report. 1:4–5. http://www.ismpminet.org/newsletter/pdf/0801.pdf.

Guy H, Ministry for Primary Industries (MPI), New Zealand. 2013. Ministerial statement of responsibility: Statement of intent 2013-2018. C.5 SOI (2013) SOI.

Handelsman J, Rondon M, Brady S, Clardy J, Goodman R. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 5:R245-R9.

Harl, Neil. 2002. U.S. Agriculture, food production is threatened by bioterrorism attacks. Ag Lender, p.10-11. http://www.econ.iastate.edu/~harl/USAgThreatened.pdf.

Hartung J, Beretta J, Brlansky R, Spisso J, Lee R. 1994. Citrus variegated chlorosis bacterium axenic culture, pathogenicity, and serological relationships with other strains of *Xylella fastidiosa*. Phytopathology 84:591-7.

Hayward A. 1991. Biology and epidemiology of bacterial wilt caused by *Pseudomonas solanacearum*. Ann Rev Phytopathol 29:65-87.

Hill B, Purcell A. 1995. Multiplication and Movement of *Xylella fastidiosa* within grapevine and four other plants. Ecology and Epidemiology 85:1368-1372.

Hill B, Purcell A. 1994. Acquisition and retention of *Xylella fastidiosa* by an efficient vector, *Graphocephala atropunctata*. Phytopathology 85:209-12.

Hung T, Hung S, Chen C, Hsu M, Su H. 2004. Detection by PCR of *Candidatus* Liberibacter asiaticus, the bacterium causing citrus huanglongbing in vector psyllids: application to the study of vector–pathogen relationships. Plant Pathol 53:96-102.

Inoue H, Ohnishi J, Ito T, Tomimura K, Miyata S, Iwanami T, Ashihara W. 2009. Enhanced proliferation and efficient transmission of *Candidatus* Liberibacter asiaticus by adult *Diaphorina citri* after acquisition feeding in the nymphal stage. Ann Appl Biol 155:29-36.

Janse J. 1996. Potato brown rot in western Europe – history, present occurrence and some remarks on possible origin, epidemiology and control strategies. EPPO Bulletin 26:679-95.

Janes J, Obradovic. 2010. *Xylella fastidiosa*: its biology, diagnosis, control, and risks. J Plant Pathology 92:35-48.

Johnson T. 2013. A history of biological warfare from 300 B.C.E. to the present. https://www.aarc.org/resources/biological/history.asp.

Jones W. 2010. High-Throughput Sequencing and Metagenomics. Estuaries and Coasts 33, 944-52.

Lambert C. 2002. Agricultural Bioterrorism Protection Act of 2002: possession, use, and transfer of biological; agents and toxins; interim and final rule (7 CFR Part 331). Federal Registry 67, 76908–76938.

Lemay A, Redlin S, Fowler G, Dirani M. 2003. Pest data sheet: *Ralstonia solanacearum* race 3 biovar 2. USDA-APHIS-PPQ. Center for Plant health Science and Technology Plant Epidemiology and Risk Analysis Laboratory.

Liu H, Kang Y, Genin S, Schell M, Denny T. 2001. Twitching motility of *Ralstonia solanacearum* requires a type IV pilus system. Microbiology 147:3215-29.

Lorenz P, Liebeton K, Niehaus F, Eck J. 2002. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. Curr Opin Biotech13:572-7.

Lorite G, De Souza A, Neubauer D, Mizaikoff B, Kranz C, Cotta M. 2013. On the role of extracellular polymeric substances during early stages of *Xylella fastidiosa* biofilm formation. Colloids and Surfaces B: Biointerfaces 102:519-25.

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

Mansfield, J., Genin S, Magori S, Citovsky V, Sriariyanum M, Ronald P, Dow MAX, Verdier V, Beer S, Machado M, Toth IAN, Salmond G, Foster G. 2012. Top 10 plant pathogenic bacteria in molecular plant pathology. Mol Plant Pathol 13:614-629.

Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L, Berka J, Braverman M, Chen Y, Chen Z, Dewell S, Du L, Fierro J, Gomes X, Godwin B, He W, Helgesen S, Ho C, Irzyk G, Jando S, Alenquer M, Jarvie T, Jirage K, Kim J, Knight J, Lanza J, Leamon J, Lefkowitz S, Lei M, Li J, Lohman K, Lu H, Makhijani V, McDade K, McKenna M, Myers E, Nickerson E, Nobile J, Plant R, Puc B, Ronan M, Roth G, Sarkis G, Simons J, Simpson J, Srinivasan M, Tartaro K, Tomasz A, Vogt K, Volkmer G, Wang S, Wang Y, Weiner M, Yu P, Begley R, Rothberg J. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-80.

Marques L, Ceri H, Manfio G, Reid D, Olson M. 2002. Characterization of biofilm formation by *Xylella fastidiosa in vitro*. Plant Dis 86:633-8.

McMichael P. 2009. A food regime analysis of the ‘world food crisis’. Agric Hum Values 26:281-295.

Metzker M. 2010. Sequencing technologies - the next generation. Nat Rev Genet 11:31-46.

Meng Y, Yaxin L, Cheryl G, Guixia H, James T, Thomas B, and HC Hoch. 2005. Upstream migration of *Xylella fastidiosa* via pilus-driven twitching motility. Bacteriology 187:5560-5567.

Messiha N, Van Bruggen A, Franz E, Janse J, Schoeman-Weerdesteijn M, Termorshuizen A, Van Diepeningen A. 2009. Effects of soil type, management type and soil amendments on the survival of the potato brown rot bacterium *Ralstonia solanacearum*. Appl Soil Ecol 43:206-215.

Mew T, Alavarez A, Leach J, Swings J. 1993. Focus on bacterial blight of rice. Plant Dis 77:5-12.

Miller M, Wolf E, Hein C, Prince L, Reardon A. 2013. Psychological effects of the marathon bombing on Boston-area veterans with posttraumatic stress disorder. J Traum Stress. 1573-6598.

Milling A, Babujee L, Allen C. 2011. *Ralstonia solanacearum* extracellular polysaccharide is a specific elicitor of defense responses in wilt-resistant tomato plants. PLoS ONE 6:15853.

Monke J. 2007. Congressional Research Service Report for Congress, Agroterrorism: Threat and preparedness. http://www.fas.org/sgp/crs/terror/RL32521.pdf.

Monterey Institute of International Studies (MIIS). 2009. Agriculture Biowarefare: State programs to develop offensive capabilities. James Martin Center for Nonproliferation Studies. http://cns.miis.edu/cbw/agprogs.htm.

Nault R. 1997. Arthropod transmission of plant viruses: a new synthesis. Annals of the Entomological Society of America, 90:521–541.

Niño-Liu D, Ronald P, Bogdanove A. 2006. *Xanthomonas oryzae* pathovars: model pathogens of a model crop. Mol Plant Pathol 7:303-24.

Ou S. 1985. Rice Disease. Kew Surrey: Commonwealth Agriculture Bureau. p.76-77.

Palacios G, Druce J, Du L, Tran T, Birch C. 2008. A new A*renavirus* in a cluster of fatal transplant-associated diseases. N.E. J. Med 358:991-998.

Pelz-Stelinski K, Brlansky R, Ebert T, Rogers M. 2010. Transmission parameters for *Candidatus* Liberibacter *asiaticus* by Asian citrus psyllid (*Hemiptera*: *Psyllidae*). J Econ Entomol 103:1531-41.

Pooler M, Hartung J, 1995. Specific PCR detection and identification of *Xylella fastidiosa* strains causing citrus variegated chlorosis. Curr Microbiol 31:377-81.

Pop M, Salzberg S. 2008. Bioinformatics challenges of new sequencing technology. Trends Genet 24:142-9.

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

Purcell A, Finlay A, 1979. Evidence for noncirculative transmission of Pierce's disease bacterium by sharpshooter leafhoppers. The American Phytopathology Society 69:393-5.
Redak R, Purcell A, Lopes J, Blua M, Mizell R, Andersen P. 2004. The biology of xylem fluid-feeding insect vectors of *Xylella fastidiosa* and their relation to disease epidemiology. Annu Rev Entomol 49: 243-70.

Research and Development Corporation (RAND). 1999. First Annual Report to the president and the congress of the advisory panel to assess domestic response capabilities for terrorism involving weapons of mass destruction. In Assessing the Threat: (AKA: "Gilmore Commission"). http://biotech.law.lsu.edu/blaw/general/terror.pdf.

Reis-Filho J (2009) Next-generation sequencing. Breast Cancer Res. 11:1-7.

Richter, D, Felix O, Alexander A, Ramona S, and Daniel H. 2008. MetaSim—A sequencing simulator for genomics and metagenomics. PLoS ONE 3:3373.

Ronaghi M. 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. 11:3-11.

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

Saile E, Mcgarvey J, Schell M, Denny T. 1997. Role of extracellular polysaccharide and endoglucanase in root invasion and colonization of tomato plants by *Ralstonia solanacearum*. Phytopathology 87:1264-71.

Sagaram U, Deangelis K, Trivedi P, Andersen G, Lu S-E, Wang N. 2009. Bacterial diversity analysis of Huanglongbing pathogen-infected citrus, using phyloChip arrays and 16S rRNA gene clone library sequencing. Appl Environ Microb 75:1566-74.

Sanger F, Coulson A. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441-8.

Sanger F, Nicklen S, Coulson A. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74:5463-7.

Schell M. 2000. Control of virulence and pathogenicity genes of *Ralstonia solanacearum* by an elaborate sensory network. Ann Rev Phytopathol 38:263-292.

Schaad N, Abrams J, Madden L, Frederick R, Luster D, Damsteegt V, Vidaver A. 2006. An assessment model for rating high-threat crop pathogens. Phytopathology 96:616-621.

Schwartz M, Hoeksema J, Gehring C, Johnson N, Klironomos J, Abbott L, Pringle A. 2006. The promise and the potential consequences of the global transport of mycorrhizal fungal inoculum. Ecol Lett 9:501-515.

Sorek R, Zhu Y, Creevey C, Francino M, Bork P, Rubin E. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318:1449-52.

Speckhard A. 2013. The Boston Marathon Bombers: the lethal cocktail that turned troubled youth to terrorism. Perspectives on Terrorism. 3:64-78.

Stein B, Elliott M, Jaycox L, Collins R, Berry S, Klein D, Schuster M. 2004. A National longitudinal study of the psychological consequences of the September 11, 2001 terrorist attacks: Reactions, impairment, and help-seeking. Psychiatry: Interpersonal and Biological Processes. Nat Hazards 67:105-117.

Tans-Kersten J, Huang H, Allen C. 2001. *Ralstonia solanacearum* needs motility for invasive virulence on tomato. J Bacteriol 183:3597-605.

Tringe S, Rubin E. 2005. Metagenomics: DNA sequencing of environmental samples. Nat. Rev. Gen 6:805-814.

Triplett L, Hamilton J, Buell C, Tisserat N, Verdier V, Leach J. 2011. Genomic analysis of *Xanthomonas oryzae* isolates from rice grown in the United States reveals substantial divergence from known *X. oryzae* pathovars. Appl Environ Microbiol 12:3930-3937.

Tucker T, Marra M, Friedman J. 2009. Massively parallel sequencing: The next big thing in genetic medicine. Am J Hum Genet 85:142-54.

Tyler H, Roesch L, Gowda S, Dawson W, Triplett E. 2009. Confirmation of the sequence of *Candidatus* Liberibacter *asiaticus* and assessment of microbial diversity in Huanglongbing-infected citrus phloem using a metagenomic approach. Mol Plant Microbe In 22:1624-34.

Tyson G, Chapman J, Hugenholtz P, Allen E, Ram R, Richardson P, Solovyev V, Rubin E, Rokhsar D, Banfield J. 2004 Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43.

United Nations (UN). Disarmament. Biological and Toxic Weapons Convention. 2012. Seventh review conference of the States parties to the convention on the prohibition of the development, production and stockpiling of bacteriological (biological) and toxin weapons and on their destruction. BWC/CONF.VII/7.

Vijaya Satya R, Zavaljevski N, Kumar K, Reifman J. 2008. A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. BMC Bioinformatics 9:185.
Wallis C, Stone A, Sherman D, Damsteegt V, Gildow F, Schneider W. 2007. Adaptation of plum pox virus to a herbaceous host (*Pisum sativum*) following serial passages. J Gen Virol 88:2839-45.

Wooley J, Godzik A, Friedberg I. 2010. A primer on metagenomics. PLoS Comput Biol 6:1371.

Zhao W, Zhu S, Liao X, Chen H, Tan T. 2007. Detection of *Xanthomonas oryzae* pv. *oryzae* in seeds using a specific TaqMan probe. Mol Biotechnol 35:119-27.

TABLES

Table 1. *Xanthomonas oryzae* pathovars *oryzae* (Xoo) and *oryzicola*: major diseases, host tissues colonized and global distribution.

Bacteria	**Disease**	**Area of plant infected**	**Distribution**
Xanthomonas oryzae* pv. *oryzae	Bacterial leaf blight (BLB) Bacterial blight (BB) Kresek	**Vascular tissue -** characterized by marginal leaf lesions	Russia, Ukraine, Asia, Africa, Mexico, U.S. (Xoo-like bacterium) Central America, Caribbean, South America, Australia
Xanthomonas oryzae* pv. *oryzicola	Bacterial leaf streak (BLS)	**Parenchyma cell**s - characterized by leaf streaking	Asia, Africa, Australia
The information in the table above was gathered from: CABI, EPPO. 1997. Data sheets on quarantine pests; *Xanthomonas oryzae.* and Triplett L, et al. 2011. Appl Environ Microbiol 12:3930-3937.			

Table 2. Prokaryotic plant pathogens discussed in this study.

Pathogen	**NCBI accession #**	**Notes**
Xylella fastidiosa 9a5c	NC_002488.3 NC_002489 NC_002490	Causal agent of citrus variegated chlorosis (CVC)
Xanthomonas oryzae pv. *oryzae* PXO99A	NC_010717.1	Causal agent of leaf blight
Ralstonia solanacearum race 3 biovar 2 (UW551)	*GCA_000167955.1	Causal agent of brown rot
Candidatus Liberibacter asiaticus	NC_012985.3	Causal agent of citrus greening 'Huanglongbing'
Pseudomonas syringae pv. *tomato* DC3000	NC_004578.1 NC_004633.1 NC_004632.1	Causal agent of bacterial speck
*Genome not fully assembled		

Table 3. *Ralstonia solanacearum* races, biovars, hosts and geographical distribution.

***Ralstonia solanacearum* species complex**			
Race	**Biovar**	**Host**	**Geographical Distribution**
1	1,3,4	Wide	Asia, Australia, Americas
2	1	Banana	Caribbean, Brazil, Philippines
3	**2**	Potato, Tomato, Geranium & other species	Global except for the U.S. & Canada
4	3,4	Ginger	Asia
5	5	Mulberry	China
The information in the table aove was athered from: Daughtrey M. 2007. Southern bacterial wilt, caused by *Ralstonia solanacearum*. Denny T, Hayward A. 2001. Laboratory Guide for the Identification of Plant Pathogenic Bacteria.3rd ed. APS Press Lemay A, et al. 2003. Pest data sheet: *Ralstonia solanacearum* race 3 biovar 2.			

CHAPTER III

E-PROBE DIAGNOSTIC NUCLEIC ACID ANALYSIS (EDNA): A THEORETICAL APPROACH FOR HANDLING OF NEXT GENERATION SEQUENCING DATA FOR DIAGNOSTICS

PUBLISHED WORK

This chapter is a published peer-reviewed manuscript with modifications to fit the thesis format of Oklahoma State University's Graduate College requirements. The manuscript is reproduced in its entirety with the permission of the Journal of Microbiological Methods. My contributions to the manuscript were the work on prokaryotic plant pathogens and significant portions of background information included in the introduction, pathogen detection assays, and metagenomics. Co-authors Anthony Stobbe and Andres Espindola performed the portions of the work on viruses and eukaryotic plant pathogens. To all co-authors, they have my thanks and acknowledgments for their contributions.

Stobbe A, Daniels J, Espindola A, Ruchi V, Melcher U, Ochoa-Corona F, Garzon C, Fletcher J, Schneider W. 2013. E-probe diagnostic nucleic acid analysis (EDNA): A theoretical approach for handling of next generation sequencing data for diagnostics. J Microbiol Meth 94:356-366.

CHAPTER III

E-PROBE DIAGNOSTIC NUCLEIC ACID ANALYSIS (EDNA): A THEORETICAL APPROACH FOR HANDLING OF NEXT GENERATION SEQUENCING DATA FOR DIAGNOSTICS

ABSTRACT

Plant biosecurity requires rapid identification of pathogenic organisms. While there are many pathogen-specific diagnostic assays, the ability to test for large numbers of pathogens simultaneously is lacking. Next generation sequencing (NGS) allows one to detect all organisms within a given sample, but has computational limitations during assembly and similarity searching of sequence data which extend the time needed to make a diagnostic decision. To minimize the amount of bioinformatic processing time needed, unique pathogen-specific sequences (termed e-probes) were designed to be used in searches of unassembled, non-quality checked, sequence data. E-probes have been designed and tested for several select phytopathogens, including an RNA virus, a DNA virus, bacteria, fungi, and an oomycete, illustrating the ability to detect several diverse plant pathogens. E-probes of 80 or more nucleotides in length provided satisfactory levels of precision (75%). The number of e-probes designed for each pathogen varied with the genome size of the pathogen. To give confidence to diagnostic calls, a statistical method of determining the presence of a given pathogen was developed, in which target e-probe signals (detection signal) are compared to signals generate by

a decoy set of e-probes (background signal). The E-probe Diagnostic Nucleic acid Assay (EDNA) process provides the framework for a new sequence-based detection system that eliminates the need for assembly of NGS data.

INTRODUCTION

Agricultural biosecurity is a priority for ensuring uninterrupted international and interstate trade, which in turn ensures an abundant food supply. With increased movement of commodities across state and national borders, the risk of introduction of exotic plant pathogens has risen significantly over the past few decades (Gamliel et al. 2008). To compound this risk, the lag time from pathogen introduction to appearance of disease symptoms provides opportunity for diseases to spread, limiting abilities for containment and eradication (Gamliel et al. 2008). Particularly for plant pathogens, for which vaccines are impossible and post infection therapies are limited and expensive, early detection and correct diagnoses are critical. Currently, plant pathogens are detected primarily by immunoassays, such as enzyme-linked immunosorbance assay (ELISA) and immune-strip tests, and nucleic acid based assays, such as real time PCR or microarray hybridization (Schaad et al. 2003). Immunoassays are relatively simple and quick, but may lack both the level of sensitivity required for agrosecurity applications and the ability to detect multiple pathogen species in a single assay (Schaad et al. 2003; Postnikova et al. 2008). Nucleic acid based techniques for detection and identification of plant pathogens, such as end-point polymerase chain reaction (PCR) and quantitative real-time PCR (qPCR) are more sensitive and selective than immunoassays, but they too may be limited in the number of pathogenic organisms that can be detected simultaneously (Postnikova et al. 2008). Both immunoassays and nucleic acid-based tests require previous characterization of the pathogen on either the protein or sequence level, and therefore lack the ability to detect uncharacterized plant pathogens. Although individual pathogen nucleic acid and immunoassays are readily available, current screening methods have limited ability to detect multiple plant pathogens concurrently in an efficient and cost effective manner. DNA microarrays, PCR-electrospray ionization/MS, multilocus sequencing typing, and simple sequence repeat assays all have the capacity to search for multiple pathogens and/or multiple diagnostic targets, but require existing pathogen characterization, which relies upon continuous development and maintenance of reference databases (Schaad et al. 2003; Postnikova et al. 2008).

Next generation sequencing (NGS) is a relatively recent technology that allows for the generation of very large amounts of sequence data from a given sample (Ronaghi 2001). Because various NGS platform technologies differ in read length (20 bp to approximately 1000 bp) and in the total number of reads (100,000 to 1 million), the amount of overall sequence data produced varies widely (Tucker et al., 2009). The productivity of NGS technology far exceeds that of traditional Sanger sequencing (Pop and Salzberg 2008; Magi et al. 2010; Metzker 2010). NGS of environmental samples has enabled the field of metagenomics, in which any and all nucleic acids in a sample are potential candidates for sequencing templates. Thus, NGS generates a sequencing profile that represents any and all organisms present within the sample (Jones 2010; Tyson et al. 2004). Metagenomics has been applied to several types of environmental samples including, seawater, ship bilge water, intestinal tracts of various animals and contaminated environments such as acid mine drainage systems (Tyson et al. 2004; Daniel 2005; Breitbart et al. 2003; Gill et al. 2006; Tringe and Rubin 2005). A metagenomic approach also could be applied to disease diagnostics, providing the benefit that NGS could detect any and all microbes in a given sample. A metagenomic approach has already been used to detect previously unknown pathogens in a variety of organisms, including mammals, insects, and plants (Adams et al. 2009; Cox-Foster et al. 2007; Palacios et al. 2008). In addition, NGS can be used to discover unknown pathogens and microbes, and has already been applied to the detection of both known and unknown plant viruses (Adams et al. 2009; Roossinck et al. 2010).

The advantage of NGS over other sequencing technologies is the volume (400MB – 28GB) of data generated (Metzker 2010; Reis-Filho 2009). From a different perspective, the volumes of data generated by NGS could be a detriment to a diagnostician, as bioinformatic processing becomes a limiting factor in high throughput applications (Pop and Salzberg 2008; Magi et al. 2010). For example, consider 200 liters of seawater containing over 5000 different viruses (Breitbart et al. 2002). If a metagenomics approach is used for plant pathogen detection within this sample, plant pathogen-specific sequences will likely make up only a small percentage of the total reads (Adams et al. 2009; Roossinck et al. 2010). In contrast, plants infected with viruses may have a much higher percentage of the total nucleic acid comprised of pathogen sequences (Kreuze et al. 2009). The host sequences that would make up the majority of an infected plant metagenome sample are essentially unimportant for diagnosis.

The novel assay developed in this research (Figure 1), and reported herein, termed E-probe Detection of Nucleic acid Analysis (EDNA), is a bioinformatic pipeline that minimizes and ignores irrelevant sequence data thereby focusing on specific pathogen-associated sequences. Mock sample databases (MSDs), simulating 454-pyrosequencing runs from plant pathogen infected plants, were generated. Rather than assessing the presence or absence of pathogens by BLAST of all sequences against a curated database, such as the nucleotide sequence databases of GenBank, the NGS metagenomic data was assessed using pathogen unique sequences termed target e-probes, incorporating local BLAST searches of designed e-probes against databases of raw sequence reads on local computer systems. This modified bioinformatics approach resulted in the rapid detection of pathogen-associated sequences without extensive analysis of the metagenome.

MATERIALS AND METHODS

Pathogens and their sequences

The plant pathogens studied here belong to three general groups, viral, prokaryotic, and eukaryotic pathogens. The chosen systems represent a wide variety of plant pathogens and have global economic importance (Table 1). Two viruses were used: Plum pox virus, a single stranded RNA virus, and Bean golden mosaic virus, which is a bipartite DNA virus. Prokaryotic pathogens included *Xylella fastidiosa* 9a5c, the causal bacterium of citrus variegated chlorosis, *Xanthomonas oryzae* pv. *oryzae*, which causes bacterial blight in rice, and *Ralstonia solanacearum* race 3 biovar 2, a select agent that causes wilting of a variety of crops including potatoes and tomatoes, *Candidatus* Liberibacter asiaticus, a bacterium responsible for citrus greening, and *Spiroplasma citri*, which causes citrus stubborn disease. Eukaryotic pathogens included: *Puccinia graminis* a rust fungus, causing the stem rust of wheat and affecting a very broad host range including 365 cereals and grasses in 54 genera (Hodson et al., 2005); *Phytophthora ramorum*, a stramenopile with a wide host range of 23 species in 12 plant families (Rizzo, 2003; Tyler et al. 2006); and *Phakopsora pachyrhizi*, which causes soybean rust, a widespread pathogen that now can be found in Africa, Asia, Australia, South America and Hawaii (Miles et al. 2003). For each pathogen, a near neighbor was chosen based on a close phylogenetic relationship, and the availability of complete genome sequence (Table 1). Grapevine, *Vitis vinifera* (GenBank: PRJNA33471), was chosen as the host background due to the availability of its genome sequence, and its genome size, which is within the range of those of full plant genomes. While grapevine is not a natural host for many of the chosen pathogens, it serves well as an example of background sequences in which the target pathogen sequences exist.

Experimental Flow

The principle behind EDNA is to minimize the bioinformatic processing by eliminating post-sequencing assembly, quality checks, and extensive BLAST searching of individual sequence reads. Rather than a traditional metagenome-based analysis of sequencing data, a simple sample database composed of raw unassembled sequence reads is generated. E-probes are then used to query the sequence database to assess the presence or absence of the target pathogen, in effect simulating a microarray or traditional hybridization assay *in silico*.

E-Probe Design

Pathogen-specific sequence queries were designed using a modified version of the Tool for Oligonucleotide Fingerprint Identification (TOFI) (Vijaya Satya et al. 2008). The basic TOFI pipeline includes three basic steps: comparison of pathogen sequences with those of near neighbors, thermodynamics optimization, and a BLAST search check for uniqueness. The EDNA query design process is similar, with the following changes. For *in silico* querying, the e-probe thermodynamics optimization step is omitted because the thermodynamic properties of the unique sequences are irrelevant. Parameters of interest to a BLAST search and/or important to a successful NGS run were added in its place. In the BLAST parameter step, the query sequence length was restricted to standardize e-values from the BLAST search and candidate e-probes containing a homo-oligomer (five or more of the same nucleotide in tandem) were removed because of the inherent miscalling of homo-oligomers in many NGS platforms. To test the optimal length of e-probes the BLAST check step was omitted, and the preliminary e-probes were used in the optimization of e-probe length. After optimization of e-probe length, a BLAST check and manual editing were reintroduced to assure specificity (Table 1). Any e-probes that hit a species different than the target with an E-value of 1x10-10 or below were removed from the final e-probe set.

Near neighbor comparisons were conducted as published (Vijaya Satya et al. 2008) with a maximum number of gaps equal to zero, a minimum probe length equal to 20 nt, and a maximum probe length equal to 4000 nt. The near neighbor selection was performed based on two criteria: complete genome availability in NCBI Genbank and close relationship to the target pathogen. The BLAST parameter step has two possible variables, the length of the designed query and the number of nucleotides that would be considered a homo-oligomer. A range of query lengths were designed, at intervals of 20 (20, 40, 60, 80, 100, 120, and 140) nucleotides, while the number of nucleotides considered to be a homo-oligomer was held constant at five.

Mock database construction

To test the designed queries, a data set consisting of both known host and pathogen genome segments was generated. Simulation of massively parallel sequencing was performed using MetaSim software (Vijaya Satya et al. 2008). The simulation includes planned mistakes in base calling, as well as a range of read lengths, both of which are common for 454, or Illumina sequencing. The resulting databases contained 10,000 simulated reads, each approximately 400 ± 30 nucleotides, or 62 nucleotides, respectively. Abundance values (representing the given amount of nucleic acid within a sample) for host genomic sequences were set at a default of 100, while host mitochondrial and chloroplast sequences were given an abundance value of 1000, meaning that for every genomic sequence there will be 10 mitochondrial and chloroplast sequences. This value was chosen arbitrarily. Pathogen abundance values were varied to generate a number of reads corresponding to the percent of the database that is made up of pathogen sequences (i.e. 25% pathogen sequences is equivalent to 2500 pathogen reads in a 10,000 read database). The databases were placed into categories based on the pathogen sequence percentage: those with 15-25% pathogen sequences were considered high, with 5-15% medium, with 0.5-5% low, and with less than 0.5% very low. These percentages were chosen arbitrarily. Each category contained three databases, which were considered as replicates within the category.

Querying Mock Databases

MSDs were queried using BLASTn with an e-value set at 50. Pathogen-specific e-probe sets were used as queries, and the MSDs served as reference databases. A match was defined as an instance where an individual e-probe was found in an MSD, such that the total number of matches must be equal to or less than the total number of e-probes. A hit was defined as any instance where a MSD read had a counterpart e-probe. A single match could be made up of multiple hits. Once the query search was conducted, the data was parsed according to different e-values thresholds to find an e-value threshold with minimal false positives, with steps at $1x10^{-3}$, $1x10^{-6}$, and $1x10^{-9}$.

The decision to designate a sample as positive or negative for a pathogen is crucial for any diagnostic assay. The criterion used to determine a positive sample in this assay was the presence of pathogen-specific sequences. It was likely that many of these sequences would be similar to sequences that belong to either the plant host, or to a different microbe that resides in the sample. Each e-probe set is designed to be unique to a specific pathogen. The signals of these sets were compared to the signals of decoy sets, which represent background signal. To generate a decoy set of e-probes, the designed target set of e-probes was reversed in sequence. Each set was then used as queries in a BLASTn search against the MSD. Each probe in both sets was given a score based on the e-value and the percent coverage of the top n hit(s), where n equals [50, 10, 5, 1].

The two sets of scores were then compared using a T-test. Three tiers of diagnostic calls were used in the statistical test, positive (p-value <= 0.05), suspect (p-value <=0.1) and negative (p value > 0.1). No significant difference between the two sets indicated no evidence for the presence of pathogen sequences, and the sample was designated negative for the pathogen.

RESULTS

General

Plant pathogenic query production was analyzed in relation to genome size for two viruses, five bacteria, two fungi and one stramenopile. The targeted viral (Plum pox virus and Bean golden mosaic virus), fungal (*Puccinia graminis* and *Phakopsora pachyrhizi*) and stramenopile (*Phytophthora ramorum*) plant pathogens were compared to near neighbors of the same species. For the bacteria, the *Ca*. Liberibacter asiaticus near neighbor was from the same species, while those of the other 3 bacteria were from a closely related species (*X. oryzae* paired with X. *fastidiosa* and vice versa). Fungal pathogens Puccinia graminis and *Phakopsora pachyrhizi* had the same near neighbor, *Puccinia triticina*. In addition, *P. pachyrhizi* was found to be broadly similar in biological attributes to *P. triticina* (Pivonia and Yang, 2006). In the case of *Phytophthora ramorum*, *P. infestans* was used as near neighbor (Table 1). The lack of a spiroplasma related to *S. citri* resulted in the selection of a near neighbor that was related at the order level (Table 1). The genome sizes of the pathogens used ranged from 5.23 knt to 88 Mnt, and the number of queries ranged from 4 to 21,790. As the genome size of the plant pathogen increased so did the total number of queries for the targeted pathogen. The total length of the combined e-probes was proportional to the total number of e-probes, and to the genome size. The percentage of genome covered ranged from 1.74 to 6.57 without any correlation with genome size or total query number (Table 1).

The number of hits at a threshold of 1x10-3, 1x10-6, or 1x10-9 received for each pathogen was determined (Figures 2-4). The number of positive hits rose with the size of the pathogen genome. As expected, the number of hits increased also with increasing pathogen proportions. At lower proportions, there was an increase in the standard deviation of the number of hits. A general similarity of the number of hits can be seen for each pathogen type, with prokaryotic pathogens having the greatest variability across pathogens.

The number of matches was compared to pathogen abundance in the MSDs. A match was defined as a single query found within a MSD, such that one match could represent multiple hits. As the pathogen abundance increased, the number of matches increased, as expected. The number of hits was nearly always greater than the number of matches, demonstrating that single queries frequently generated multiple hits in a MSD (Figures 5-7). The number of prokaryotic pathogen e-probe matches was related to the number of e-probes available for the pathogen, in other words, the more e-probes designed for a given pathogen, the more matches were attained in a BLAST search. For example, a *Ca*. L. asiaticus e-probe set of 80 nt length consists of 502 e-probes, and when queried with a low pathogen ratio MSD, received 169 matches. *X. oryzae* contained 2597 e-probes with 345 matches. In contrast, the number of matches for *P. ramorum* (1645) was less than the number of matches for *P. graminis* (1998), despite the greater number of queries for the former. For the viral pathogens a match was found for every query available in high, normal and low pathogen abundance MSDs, and the number of matches in very low abundance MSDs was approximately half of the number of available queries (2 matches/ 4 e-probes in the case of BGMV) (Figures 5-7, Table 1).

Optimization of e-probe length

To determine the optimum e-probe length, precision was calculated for each of the e-probe sets (Table 2), in which each hit is either a true positive (a pairing of e-probe and pathogen sequence), or false positive (a pairing of e-probe and non-pathogen sequence). We calculated the precision as the number of pathogenic hits (True positive) divided by the total number of hits (hits to pathogen or hits to host). For each of the pathogens, e-probe lengths below 80 nt were substandard (precision less than 75%) as queries of very low pathogen ratio (<0.5%) MSDs. Viral e-probe sets had high precision, most likely due to the minimal similarity between viral and eukaryotic sequences. For prokaryotic and eukaryotic pathogens, at abundances greater than 0.5%, the specificity was greater than 80.4% at any e-probe length. With the very low abundance MSDs, the precision varied between 14.1 and 100%.

The effect of varying e-probe lengths from 20 – 140 nt on the matches generated by searches on the MSDs was determined. As expected, for each pathogen, match numbers decreased as the length of the e-probes increased, because the number of longer e-probes designed was much lower than that for shorter e-probes. In general, each pathogen type (virus, bacterial, and eukaryotic) had a similar number of matches for each member within a group (Figures 5-7). One exception was *X. oryzae*, which showed no such downward trend (Figure 6). Almost all pathogens were detected using every query length. The other exception was *R. solanacearum* in very low pathogen abundance MSDs, in which an average of a single match was found for the majority of query lengths (40, 80, 100, 120, and 140 nt). *P. ramorum* and *P. graminis* showed the smallest number of matches of all the pathogens when very low pathogen proportion MSDs were queried with 140 nt e-probes. This low number of matches could be due to the random selection of sequences when constructing MSDs because fungal and stramenopile genomes are larger than viral and bacterial genomes, allowing the presence of portions of the genome in the MSDs that have a low density of e-probe sequences. This phenomenon is most likely to occur for low pathogen proportions and large pathogen genomes.

E-value threshold

All four categories of mock databases (high, medium, low, and very low) were queried using the 80 nt e-probes for all of the target pathogens. Pathogens reads were detected via e-probe based BLAST search routinely with a threshold e-value of 1x10-3. Using 80 nt queries, all of the pathogens also were detected in very low abundance databases, in some but not all replicates (Figures 2-4, Supplemental Table 1).

Some e-probes generated false positive matches, i.e. instances when the e-probe sequence found a host counterpart in the MSD. The number of false positive matches was directly related to the e-values used in the BLASTn searches of the MSDs, with higher e-values generating more false positives. Overall, the eukaryotic pathogen simulations with a threshold e-value of 1x10-3 generated the highest number of false positive matches and hits (Supplemental Table 1). Bacterial pathogen simulations also generated false positives; however these were fewer (5 or fewer per database). No false positives at a threshold e-value of 1x10-3 were observed in viral MSDs. The e-value was adjusted during the parsing step by using three different threshold e-values of 1x10-3, 1x10-6, and 1x10-9. When the pathogens were analyzed using lower e-values, the number of false positives per database decreased from an average of 1 for prokaryotic e-probe sets, and 8 for eukaryotic e-probe sets to 0 for both.

Using the threshold values of either 1x10-6 or 1x10-9 also decreased the total number of matches and hits; particularly for fungal pathogens, i.e. for *P. graminis*, the number of matches decreased from 1998 matches (e-value of 1x10-3) to 1530 matches (1x10-9). Among prokaryotic pathogens, the greatest decrease in total matches and hits was observed with *X. oryzae*, which decreased from 2597 to 1832 at e-values of 1x10-3 to 1x10-9, respectively. This difference of 765 fewer e-probes did not lessen the effectiveness of pathogen detection. Instead it decreased the number of false positives due to the greater stringency placed on the bioinformatics system. For viruses, the total number of matches was not affected by increased stringency (lower e-values); however the total number of hits was reduced with lower e-value BLASTn (Supplementary Table 2). Mock sample databases also were generated using read lengths of 62 nt and with the error model found for a typical Illumina run (Richter et al., 2008). EDNA analysis showed similar results to the 454 simulations (data not shown).

BLAST check comparison

False positives were reduced in number by removing e-probes that have similarity to known sequences in NCBI. Each 80 nt e-probe set was used as queries in a search against the NCBI GenBank nt database. E-probes with hits at an e-value of 1x10-10 or lower were removed from the probe set. This decreased the number of probes per set by up to 50% (Table 1). Comparing the performance of BLAST-checked e-probe sets showed a slight reduction in the number of false positive hits, with a larger reduction in the number of matches and total hits (Supplemental Table 1).

Determination of Positive and Negatives

Using the above method, we were able to correctly call samples positive for all positive samples except for those at a very low abundance (<0.5% pathogen reads) (Table 3). At this abundance there were mixed results, at times calling the sample positive while other times calling it negative. *R. solanacearum* was not detected in very low abundance MSDs. Pathogen negative MSDs (MSDs without pathogens) were all negative or suspect for viruses, *S. citri*, and *R. solanacearum*. False positives were most common in eukaryotic pathogens. When the number of top hits (n in equation 1) was lowered in the scoring step, the pathogen negative MSDs were correctly identified in some, but not all, replicates (Table 3).

Discussion

There are multiple advantages to using a metagenomics-based approach to pathogen diagnostics. Advances in NGS have made it possible to generate billions of bases of sequence for any given sample, creating metagenomes that represent a complete profile of all organisms in a given nucleic acid sample, including host, endophytes and pathogens (Jones, 2010; Tyson et al., 2004). This presents the very real probability that any and all microbes in any given sample could be identified. Metagenomics approaches have been used in multiple instances to suggest the cause of unknown diseases (Adams et al. 2009; Cox-Foster et al. 2007; Palacios et al. 2008), but two factors would seem to preclude the use of metagenomic sequencing as an everyday diagnostic tool.

The first detriment to adopting metagenomics-based diagnostics is the current per run cost. The typical approach to a metagenome diagnosis is nucleic acid extraction, sequencing, sequence assembly, and BLAST analysis of the assembled contigs. An examination of recent history suggests that sequencing technologies will likely become less expensive, due to the technologies becoming faster, more accessible and the sequencing more processive over time, outpacing Moore's Law. This prediction suggests that NGS costs may not be a long term restraint, particularly when combined with barcoding (Parameswaran et al. 2007). However, the very same advances that drive down per sample costs of sequencing create additional data handling problems. As NGS becomes less expensive, faster and the length of reads increases, the number of bases sequenced in a single run will increase exponentially. These same advances in NGS will have an additional exponential growth effect on the databases (i.e. GenBank and its subsidiaries) 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 objective of this work was to find a simplified bioinformatic approach for dealing with the exponential growth and complexity of NGS metagenome data, which could be handled on a standard personal computer without extensive computational delays. To do this, we developed a protocol (EDNA) in which the input NGS data would be treated as the searchable database, and this sequence database would be queried by diagnostic signature sequences (e-probes) without the need for assembly or quality checks. This approach allows the user to limit and control both the size of the searchable database and the size of the searching query set.

The EDNA approach was tested using a series of MSDs representing potential metagenomes with pathogen sequences in a plant background. Representatives of multiple taxonomic groups of plant pathogens were used, including an RNA virus, a DNA virus, a spiroplasma, prokaryotes, a stramenopile, and a fungus. Diagnostic e-probe sequences were selected at a range of lengths, and used to query MSDs with differing levels of pathogen abundance (from 0.5% pathogen reads to 25% pathogen reads). EDNA was successful at detecting all pathogens at low, medium and high levels (everything above 0.5% pathogen reads in the MSD). The number of matches (any instance where an individual e-probe finds a counterpart or counterparts in the database) and hits (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. Below the low pathogen threshold, the EDNA results were mixed, suggesting that EDNA has a threshold of detection in its current format. However it should be noted that the limit of detection could be improved to suit user needs by adjusting the number of e-probes, the length of the e-probes and/or the parsing E-value.

Not surprisingly, EDNA generated some false positive hits and matches. The number of false positives appeared to remain relatively the same regardless of the pathogen abundance (Supplemental Table 1), and were problematic only in the very low abundance MSDs. Viruses were completely free of false positives at all concentrations of pathogen reads, which might be expected considering the lack of related sequences in the host setting. Prokaryotes have chloroplast and mitochondrial counterparts in the host MSD, and there were occasional false positive hits and matches using prokaryotic e-probes. Overall, eukaryotic pathogen e-probes were the most problematic, as might be expected when confronted with a eukaryotic host background. Very low pathogen abundance simulations were not distinguished from pathogen-free MSDs, and generated the highest number of false positive matches and hits. However, EDNA is flexible enough to generate higher precision, by raising the E-value threshold required for calling a positive hit. Both *P. graminis* and *P. ramorum* showed fewer (zero or one) false positive hits when the E-value was lowered to $1x10^{-9}$, and the prokaryotic pathogen e-probes were completely specific when the parsing E-value was lowered to $1x10^{-6}$. Larger, more complex genomes and the conservation of genes and sequences between pathogen and host (eukaryotic pathogens) require lower E-value cutoff levels. It should also be noted that some of the near neighbors were less related to the target organisms, a limitation driven by the lack of available sequenced genomes. Improved near neighbors, which should become available as more pathogen genomes are sequenced, will also improve precision.

A second approach for improving specificity involved improving the screening of potential e-probes. Clearly, as genome size increases the number of e-probes generated increases in proportion. Removal of a number of e-probes from the larger pathogen genome screens would likely not affect the overall limit of detection. The e-probes from all pathogens were searched against GenBank, as is done in primer selection, to eliminate a number of false positive generating e-probes. This strategy may be of limited use for plant pathogens, however, as the majority of environmental microbes in a typical plant metagenome have no GenBank counterpart (Pivonia and Yang 2006). The addition of a healthy control BLAST, searching healthy control asymptomatic host environmental sample sequence databases for the presence of potential false positive queries might eliminate some e-probes that would react to host or endophyte sequences not available in GenBank. Regardless, much like limit of detection, EDNA precision could be adjusted up or down as needed in the e-probe design (by adjusting e-probe length or near neighbor selection) or during database searching (adjusting E-value threshold). As an added advantage, adjusting E-value threshold and choosing “general” e-probes could allow for searching for related organisms that are not the specific target organism.

A key to any diagnostic method is determining the level of positive “signal” necessary to confirm that a pathogen is present in a given sample. For molecular techniques such as PCR, the presence or absence of a product is easily distinguished. However when the positive/negative decision is based on a quantitative measurement, such as fluorescence or absorbance in ELISA, the determination involves some level of statistical analysis. The number of matches and hits returned from a sequence database query within the proposed EDNA concept is not entirely dissimilar to these quantitative approaches, in which it is critical to distinguish between a true signal (e.g. matches that represent pathogen sequences) and a false “signal” (e.g. matches where query sequence is identical or nearly identical to non-pathogen sequence). In ELISA, a common approach is to make a diagnostic decision by comparing the fluorescence value of a sample well to those of a set of negative control wells, with a cutoff defined as a certain number of standard deviations over background. To utilize a similar approach for NGS, a basal level of false positives (erroneous query matches) was determined. Decoy probe sets were developed for every pathogen, and these decoy e-probe sets were used to determine the chances that a relatively random sequence would find a counterpart in a eukaryotic host background by chance. The decoy comparison method was particularly successful with virus pathogens, and less successful with eukaryotic pathogens. This finding indicates that statistical approaches could be developed to assess the accuracy of positive/negative determinations in NGS-based diagnostics. As in other diagnostic assays, the balance between specificity and limit of detection is a necessity in this bioinformatics approach to diagnostics.

The theoretical ability of next generation sequencing coupled with bioinformatics to detect highly consequential plant pathogens (EDNA), at varying abundances, and in a complex host sample was validated. The advantage of the EDNA system is that it can be adjusted or designed to address a range of applications and/or the scientific needs in a variety of fields including bioinformatics, epidemiology, detection and diagnostics of human, animal, and plant pathogens, monitoring and surveillance, quarantine, and microbial forensics. EDNA alleviates the computational work load routinely associated with classic metagenomic assembly and BLAST-based approaches; allowing plant pathologists to use personal computers for running bioinformatic pipelines without investing in large and expensive cluster systems of bioinformatic infrastructure. The EDNA approach could be usable for all types of pathogens in all types of hosts, and could work with any NGS platform. The flexibility given by the possibility to periodically modify or build custom tailored databases of e-probe sets plus the lower computational requirements favor the implementation of endless applications limited only by the imagination of the scientific community.

LITERATURE CITED

Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N. 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Pl Pathol 10:537-545.

Breitbart M, Hewson I, Felts B, Mahaffy J, Nulton J, Salamon P, Rohwer. 2003. Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185:6220-6223.

Breitbart M, Salamon P, Andresen B, Mahaffy J, Segall A, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. Proc Nat Acad Sci USA 99:14250-14255.

Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran N, Quan P, Briese T, Hornig M, Geiser D, Martinson V, vanEnglesdrop D, Kalsstein A, Drysdale A, Hui J, Zhai J, Cui L, Hutchison S, Simons J, Egholm M, Pettis J, Lipkin W. 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.

Gamliel A, Gullino ML, Stack JP. 2008. Crop biosecurity: local, national, regional and global perspectives. In: Gullino ML, Fletcher J, Gamliel A, Stack JP, editors. Crop Biosecurity: Springer Netherlands. pp. 37-61.

Gill S, Pop M, DeBoy R, Eckburg P, Turnbaugh P, Buck S, Gordon J, Relman D, Fraser-Liggett C, Nelson K. 2006. Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.

Hodson DP, Singh, R.P., Dixon, J.M. 2005. An initial assesment of the potential impact of stem rust (race Ug99) on wheat producing regions of Africa and Asia using GIS. 7th International Wheat Conference. Mar del Plata, Argentina. pp. 142.

Jones W .2010. High-throughput sequencing and metagenomics. Estuaries and Coasts 33:944-952.

Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R. 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1-7.

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

Metzker ML 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. 11:31-46.

Miles MR, Frederick, R.D., and Hartman, G.L. 2003. Soybean rust: Is the U.S. soybean crop at risk? APS Feature Story: American Phytopathological Society.

Palacios G, Druce J, Du L, Tran T, Birch C, Briese C, Quan P, Hui J, Marshall J, Simons J, Eqholm M, Paddock C, Shieh W, Coldsmith C, Zaki S, Catton M, Lipkin W. 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. N.E. J. Med 358:991-998.

Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M, Fire A. 2007. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nuc Ac Res 35:130.

Pivonia S, Yang XB 2006. Relating epidemic progress from a general disease model to seasonal appearance time of rusts in the United States: Implications for soybean rust. Phytopathology 96: 400-407.

Pop M, Salzberg SL 2008. Bioinformatics challenges of new sequencing technology. Trends Genet 24:142-149.

Postnikova E, Baldwin C, Whitehouse CA, Sechler A, Schaad NW, 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 (2008) MetaSim-A sequencing simulator for genomics and metagenomics. PLoS ONE 3:3373.

Rizzo DM, Garbelotto. 2003. Sudden Oak death: endanfering California and Oregon. Front Ecol Environ 1:197-204.

Ronaghi M 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res 11:3-11.

Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, et al. 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19:81-88.

Schaad NW, Frederick RD, Shaw J, Schneider WL, Hickson R, Petrillo M, Luster D. 2003. Advances in molecular-based diagnostics in metting crop biosecurity and phytosanitary. Issues. Ann Rev Phytopath 41:305-324.

Tringe SG, Rubin EM 2005. Metagenomics: DNA sequencing of environmental samples. Nat Rev Gen 6:805-814

Tucker T, Marra M, Friedman JM 2009. Massively Parallel Sequencing: The next big thing in genetic medicine. Am J Hum Gen 85142-154.

Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, Chapman J, Damasceno CM, Dorrance AE, Dou D, Dickerman AW, Dubchak IL, Garbelotto M, Gijzen M, Gordon SG, Govers F, Grunwald NJ, Huang W, Ivors KL, Jones RW, Kamoun S, Krampis K, Lamour KH, Lee MK, McDonald WH, Medina M, Meijer HJ, Nordberg EK, Maclean DJ, Ospina-Giraldo MD, Morris PF, Phuntumart V, Putnam NH, Rash S, Rose JK, Sakihama Y, Salamov AA, Savidor A, Scheuring CF, Smith BM, Sobral BW, Terry A, Torto-Alalibo T, Win J, Xu Z, Zhang H, Grigoriev IV, Rokhsar S, Boore L. 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313:1261-1266.

Tyson GW, Chapman J, Hugenholtz P, Allen E, Ram RJ, Richardson P, Solovyev W, Rubin E, Rokhsar D, Banfield J. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-4.

Vijaya R, Zavaljevski N, Kumar K, Reifman J. 2008. A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. BMC Bioinformatics 9:185.

TABLES

Table 1. Comparison of the amount of genome coverage of e-probes across tested pathogens.

Name	**NCBI accession #**	**Near Neighbor**	**NCBI accession #**	**Original Sequence Size (kb)**	**# 80nt e-probes**	**Total kbps**	**Genome % coverage**
Bean golden mosaic virus	NC_004042 NC_004043	Abutilon mosaic virus	NC_001928 NC_001929	5.23	4	0.32	6.12%
Plum pox virus	NC_001445	Pepper mottle virus	NC_001517	9.74	8	0.64	6.57%
Spiroplasma citri	115252846, 110005886 110005766, 110005758 11000748, 110005735 110005716, 110005696 110005687, 110005683 110005675, 110005664 110005652, 110005641 110005622, 110005605 110005592, 110005560 110005522, 110005436 110005327, 110005285 110005260, 110005199 110005145, 110005138 110005098, 110005060 110005027, 110004948 110004868, 110004796 110004744, 110004631 110004607, 110004455 110004127, 110004055 110003907M	*Mycobacterium bovis*	NC_008769	1525.76	423	33.84	2.22%
Ca. L. asiaticus	NC_012985	*Agrobacterium tumefaciens*	AE007869	1226.70	114	9.12	0.74%

Name	NCBI accession #	Near Neighbor	NCBI accession #	Original Sequence Size (kb)	# 80nt e-probes	Total kbps	Genome % coverage
Xanthomonas oryzae	CP000967	*Xylella fastidiosa*	NC_002488 NC_002489 NC_002490	2679.31	1459	116.72	4.36%
Xylella fastidiosa	NC_002488 NC_002489 NC_002490	*Xanthomonas oryzae*	CP000967	5240.08	2597	207.76	3.96%
Ralstonia solanacearum	NC_003295 NC_003296	*Ralstonia pickettii*	NC_010682 NC_010678 NC_010683	3716.41	1418	113.44	3.05%
Puccinia graminis	AAWC01000001 AAWC01004563	*Puccinia triticina*	ADAS01000001 ADAS01038776	66652.40	20573	1645.84	2.47%
Phytophora ramorum	AAQX01000001 AAQX01007589	*Phytophora infestants*	AATU01000001 AATU01018288	88644.63	21790	1743.2	1.97%

Continuation of Table 1 from page 56.

Table 2. Table showing the precision (in percentage) at varying probe lengths and different pathogenic concentrations.

Name	**E-probe length**	**15-25%**	**5-15%**	**.05-5%**	**< 0.5%**
Bean golden mosaic virus	20	100	100	100	100
	40	100	100	100	100
	60	100	99.97	100	100
	80	100	100	100	100
	100	100	100	100	100
	120	100	100	100	100
	140	100	100	100	100
Plum pox virus	20	100	100	100	100
	40	100	100	100	100
	60	100	100	100	100
	80	100	100	100	100
	100	100	100	100	100
	120	100	100	100	100
	140	100	100	100	100
Spiroplasma citri	20	97.66	94.32	80.38	33.36
	40	98.89	98.14	91.37	51.1
	60	98.94	98.75	93.91	54.44
	80	99.56	99.38	96.2	78.59
	100	99.73	99.03	93.37	72.44
	120	99.78	99.28	97.4	68.33
	140	99.53	98.84	99.02	63.89
Ca. L. asiaticus	20	98.97	98.31	92.42	55.58
	40	99.48	99.27	96.35	54.79
	60	99.26	98.72	96.42	62.05
	80	99.74	99.84	98.06	81.24
	100	99.63	99.05	96.44	63.49
	120	99.49	99.33	97.17	57.08
	140	99.33	99.12	96.47	40.12
Xanthomonas oryzae	20	99.96	100	99.58	84.2
	40	100	99.78	99.58	87.91
	60	99.95	99.81	99.51	84.21
	80	99.93	99.95	99.87	93.72
	100	99.98	99.89	99.87	93.91
	120	99.9	99.89	99.86	94.57
	140	99.98	99.95	99.87	100

Name	E-probe length	15-25%	5-15%	.05-5%	< 0.5%
Xylella fastidiosa	20	99.96	99.83	99.39	98.1
	40	99.97	99.87	100	97.09
	60	99.93	99.52	99.72	96.41
	80	99.91	99.71	99.68	94.98
	100	99.86	99.67	99.63	94.42
	120	99.89	99.61	99.56	93.07
	140	99.87	99.53	99.52	93.07
Ralstonia solanacearum	20	100	98.89	99.52	97.94
	40	99.91	99.83	99.42	95.38
	60	99.9	99.87	98.78	93.1
	80	100	100	99.42	92.86
	100	100	100	99.02	90.91
	120	100	100	98.57	75
	140	100	100	98	75
Phytophthora ramorum	20	99.45	98.95	96.41	24.78
	40	99.75	99.57	97.66	30.58
	60	99.66	99.37	95.68	14.14
	80	99.76	99.68	98.52	48.94
	100	98.04	100	100	100
	120	99.75	99.26	98.11	45.45
	140	99.43	99.22	95.77	28.57
Puccinia graminis	20	98.28	96.52	87.8	30.54
	40	99.36	98.65	94.12	44.22
	60	99.17	97.87	92.69	35.86
	80	99.69	99.35	97.77	56.9
	100	99.71	99.2	98.5	60.78
	120	99.75	99.28	98.07	66.67
	140	99.91	99.45	98.21	57.14

Continuation of Table 2 from page 58.

Table3. P-values of EDNA diagnostic call.

		15-25%			**5-15%**			**0.5-5%**			**<0.5%**			**0%**		
BGMV	Top 50	0.031	0.031	0.000	0.026	0.022	0.000	0.000	0.000	0.001	0.007	0.004	0.384	0.077	0.765	0.243
	Top 10	0.000	0.034	0.000	0.000	0.042	0.003	0.001	0.006	0.001	0.008	0.005	0.582	0.151	0.327	0.611
	Top 5	0.012	0.012	0.000	0.000	0.000	0.000	0.007	0.005	0.018	0.008	0.045	0.654	0.432	0.396	0.590
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.006	0.004	0.788	0.769	0.978	0.936
PPV	Top 50	0.000	0.000	0.000	0.001	0.001	0.001	0.000	0.009	0.035	0.374	0.018	0.052	0.334	0.310	0.096
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.026	0.397	0.019	0.057	0.562	0.629	0.153
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.390	0.020	0.057	0.681	0.953	0.489
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.376	0.020	0.007	0.904	0.384	0.947
S. citri	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.164	0.202	0.001	0.970	0.431	0.277
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.102	0.001	0.673	0.786	0.170
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.052	0.109	0.001	0.910	0.277	0.383
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.083	0.098	0.001	0.904	0.384	0.947
Ca. L. asiaticus	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.007	0.001	0.027	0.009	0.027
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.017	0.006	0.198	0.003	0.009
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.017	0.023	0.021	0.308	0.003	0.039
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.035	0.030	0.042	0.631	0.005	0.029
R. solanacearum	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.605	0.648	0.011	0.061	0.174	0.056
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.586	0.057	0.025	0.256	0.656	0.208
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.081	0.012	0.223	0.105	0.448	0.231
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.073	0.008	0.067	0.218	0.953	0.392

		15-25%			5-15%			0.5-5%			<0.5%			0%		
X. oryzea	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.060	0.811	0.002	0.000	0.000	0.000
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.824	0.173	0.650	0.000	0.001	0.002
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.004	0.074	0.521	0.157	0.398
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.001	0.033	0.016	0.016	0.089
X. fastidiosa	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.745	0.306	0.025	0.316	0.222	0.271
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.018	0.003	0.000	0.006
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.007	0.004	0.000	0.027
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.026	0.031	0.001	0.514
P. graminis	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.001	0.000	0.000	0.000
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.333	0.428	0.894	0.413	0.009	0.020
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P. ramorum	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.083	0.508	0.000	0.000	0.000
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.479	0.049	0.000	0.014	0.000
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.350	0.004	0.000	0.338	0.007	0.019
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.257

Continuation of Table 3 from page 60.

FIGURES

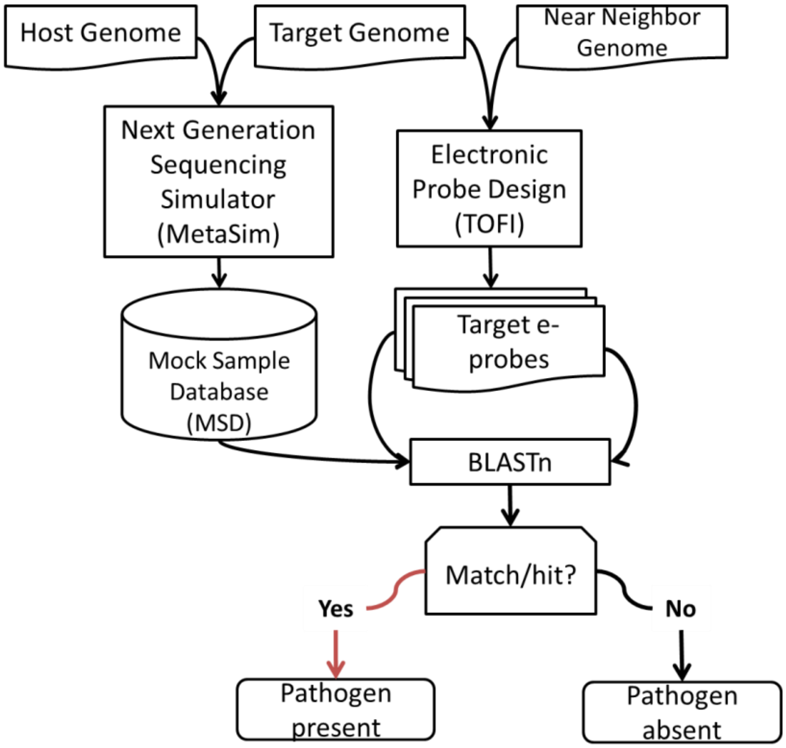

Figure 1. Experimental flow of E-probe Diagnostic Nucleic acid Assay pipeline (EDNA).

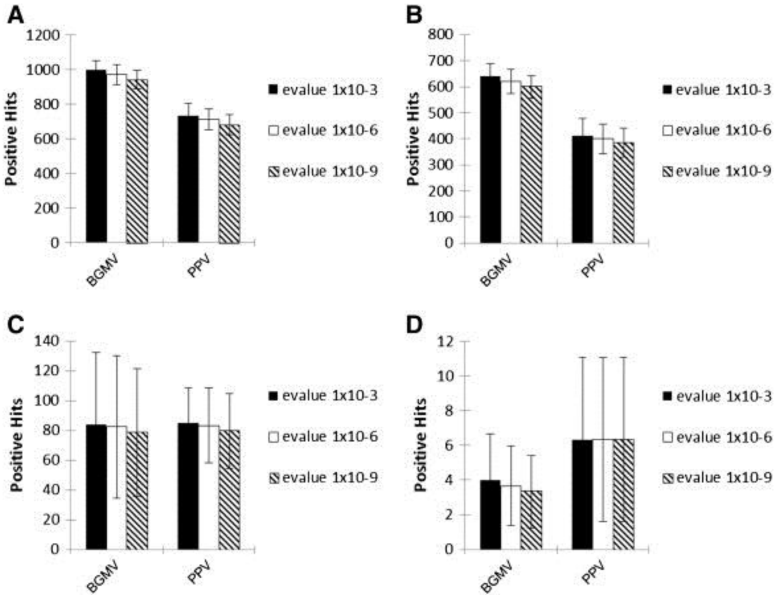

Figure 2. The total number of hits from a BLAST search of 80 base target virus e-probe sets against MSDs containing grapevine and target pathogen sequences at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) < 0.5% pathogen read abundances.

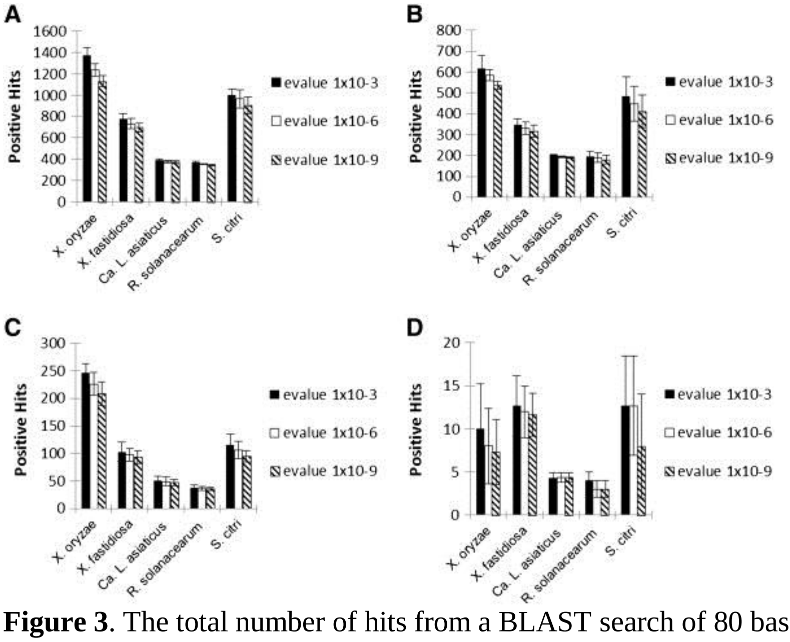

Figure 3. The total number of hits from a BLAST search of 80 base target prokaryotic pathogen e-probe sets against MSDs containing grapevine and target pathogen sequences at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) < 0.5% pathogen read abundances.

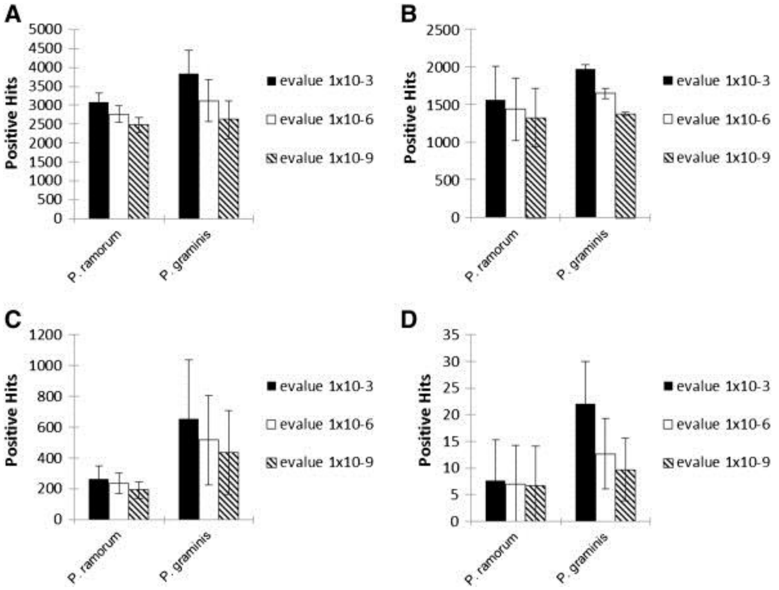

Figure 4. The total number of hits from a BLAST search of 80 base eukaryotic pathogens e-probe sets against MSDs containing grapevine and target pathogen sequences at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) < 0.5% pathogen read abundances.

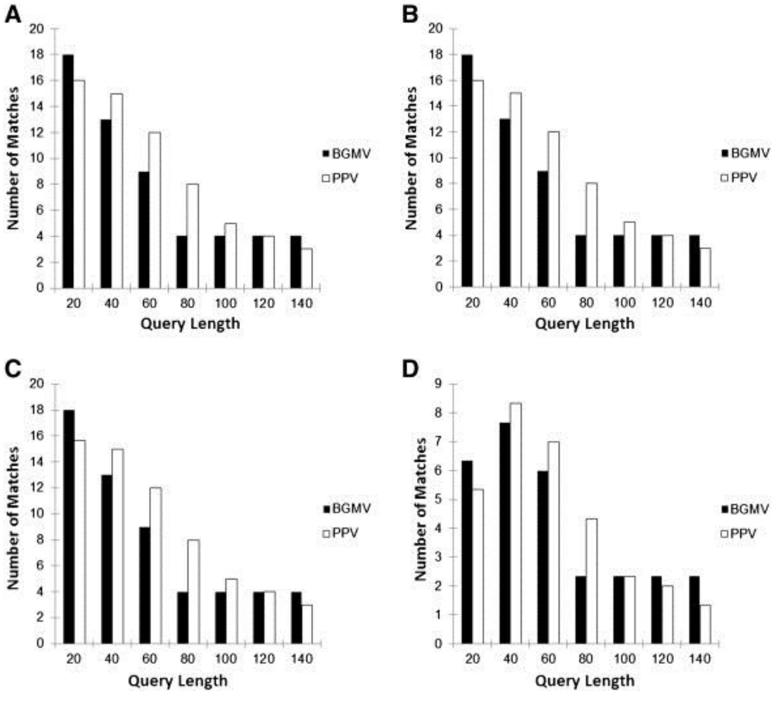

Figure 5. Number of matches (positive e-probes) for each given length of e-probes, for target viruses at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) < 0.5% pathogen read abundances.

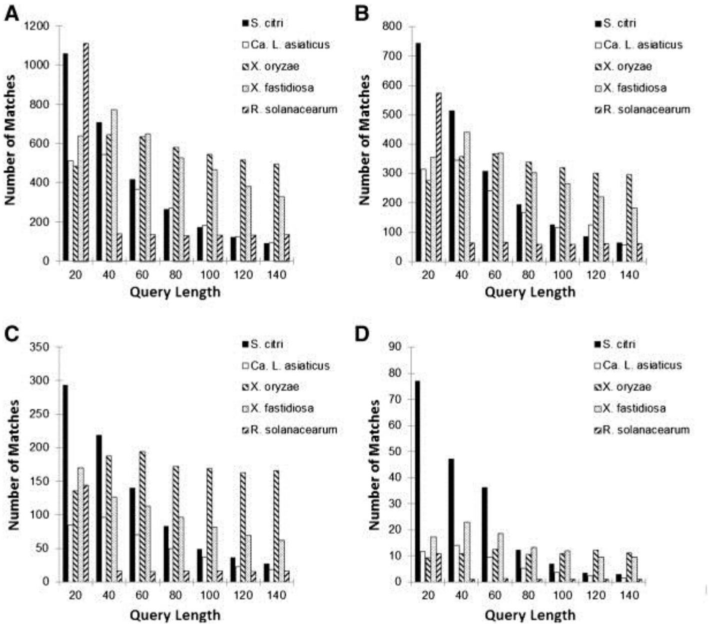

Figure 6. Number of matches (positive e-probes) for each given length of e-probes, for target prokaryotic pathogens at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) < 0.5% pathogen read abundances.

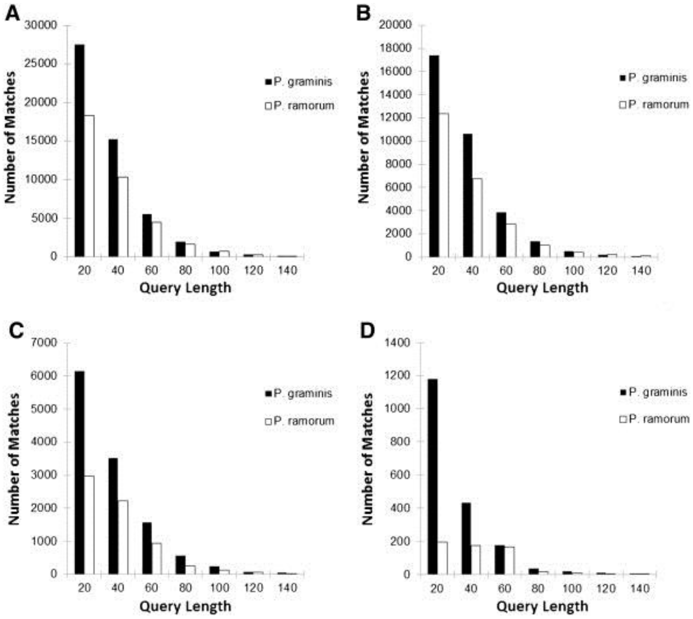

Figure 7. Number of matches (positive e-probes) for each given length of e-probes, for target eukaryotic pathogens at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) < 0.5% pathogen read abundances.

CHAPTER IV

A NOVEL TOOL FOR DETECTION OF PROKAYOTIC PLANT PATHOGENS USING NEXT GENERATION SEQUENCING AND EDNA

ABSTRACT

Biosecurity agencies around the globe require plant pathogen detection to prevent or lessen the risk of pathogen introductions. The detection and identification systems need to be readily adjusted as to meet current and future biological threats. Although individual pathogen assays abound, and some multiplex assays have been developed, current screening methods are limited in the total number of pathogens they detect concurrently. Combining bioinformatics with next generation sequencing (NGS) allows for the creation of a single assay to detect, simultaneously, any and all microbes in a sample, including pathogens that have been genetically modified. The adaptation of bioinformatic pipelines for query “electronic probe” generation and screening using BLASTn for plant tissues infected with *Ralstonia solanacearum* race 3 biovar 2 (Rs r3b2) and *Pseudomonas syringae* pv. *tomato* DC3000 (DC3000), this research facilitates the detection of these two bacteria in a complex host sample. Pathogen specific queries, ranging in lengths from 15 nt to 80 nt, were created for detection of Rs r3b2 and DC3000. The e-probe sets were used to query NGS data of diseased hosts, potato inoculated with Rs r3b2, and tomato inoculated with DC3000. The NGS data against which the e-probes were tested contained sequences from multiple bacteria, fungi, plant genomes, mitochondrial genomes, and a chloroplast genome, typical of a metagenomic sample from an infected plant. Both bacterial pathogens were readily detectable; suggesting NGS data can be used for the screening of targeted prokaryotes by using e-probes. This research merges bioinformatics and plant pathology for addressing both agricultural and national security detection and diagnostic needs.

INTRODUCTION

Today's intensive movement of agricultural commodities from state to state or from one country to another increases the risk of exotic plant pathogen introductions. Biosecurity agencies need tools to screen all high-threat plant pathogens in a single assay. Current plant pathogen screening tools include immunoassays, which are rapid and allow for high volumes of processing, but may lack the sensitivity standards required in cases where high consequence pathogens are being screened for presence or absence. Nucleic acid based tests, such as quantitative polymerase chain reaction (qPCR), are highly sensitive and selective when compared to immunoassays, but are limited in the total number of pathogens screened per reaction (Kim et al. 2008; Zhang et al. 2007). Additionally, if there is very little diseased plant material available, the diagnostics could be limited in the number of test that can be performed due to degradation of the original sample that occurs when processing it for immuno- or nucleic acid assays, which could result in non-diagnosis or false negatives. An alternative would be using a single protocol for both processing of plant material for next generation sequencing (NGS) and screening the resulting data with a bioinformatics tool as discussed in Stobbe et al. (2013). To test this approach two prokaryotic plant pathogens were used: *Ralstonia solanacearum* race 3 biovar 2 (Rs r3b2), which has been designated by the USDA's Animal and Plant Health Inspection Service (APHIS) as a select agent, and *Pseudomonas syringae* pv. *tomato* str. DC3000 (DC3000), a model organism to study virulence mechanisms in both *Arabidopsis thaliana* and tomato (Xin and He 2013). Rs r3b2 strain is the causal pathogen of bacterial wilt or brown rot of potato and tomato, and DC3000 is the causal pathogen of bacterial speck of tomato (Champoiseau and Momol 2008; Zhao et al. 2003). In this research, *Solanum tuberosum* (potato) inoculated with Rs r3b2 and *Solanum lycopersicum* (tomato) inoculated with DC3000 will be used as host and target pathogen, respectively.

To detect a variety of plant pathogens of various taxa, including prokaryotes, eukaryotes and viruses in a single assay next generation sequencing (NGS) data obtained from metagenomic sampling of diseased plant tissues was used. The enormous amount of sequence data generated during an NGS run is a limiting factor in current research using NGS because of the high computational demand required for assembly and annotation. High computational demands overwhelm most computers not set up on cluster systems, in which algorithms divide the workload among numerous processors, thereby reducing the total time and computational load required to assemble the NGS data (Karypis et al. 1999). While working in the cluster environment reduces processing time and computational load, there are costs of setting up and maintaining the cluster. Also, with most academic based cluster systems there are wait times to submit jobs for processing as well as a risk of error due to script coding mistakes, both of which can hamper data processing and result in flawed output files. A plausible alternative to circumvent such problems is to remove the assembly and annotation steps, which require high performance computing. Ideally, research scientists or diagnostic labs would use their own dedicated computer, far less costly than a cluster system, to screen an NGS run of suspect plant tissue. A recent report by Stobbe et al. (2013) described such approach for detection of numerous classes of plant pathogens.

Rather than time-consuming assembly and annotation, an unassembled raw NGS data file can be queried electronically with generated e-probes for particular pathogens of interest (Stobbe et al. 2013). This approach requires far less computational processing used in assembling and annotating of NGS data, and can be performed on a personal computer. E-probe Diagnostic

Nucleic acid Analysis (EDNA) methodology works by first generating electronic probes (e-probes) for a particular prokaryotic, eukaryotic and/or viral pathogen. To generate e-probes the entire genome of a target pathogen, and that of a near neighbor genome, are downloaded and processed through bioinformatic pipelines, in which the user sets e-probe length. The resulting output file, containing e-probes, contains several hundred to thousands unique target pathogen digital sequences depending on the genome size and similarity of the target pathogen and near neighbor (Stobbe et al. 2013). The file containing e-probes is used to query a raw NGS data file by BLASTn. The resulting BLASTn file is parsed at e-values of 1e-3, 1e-6, or 1e-9, and the resulting data file is analyzed for total matches, hits, and number of e-probes used in the query. Depending on the presence or absence of matches, a diagnostics of positive (pathogen present) or negative (pathogen absent) is made. Additionally, EDNA can run on a laptop computer, providing greater mobility than a desktop or cluster computer for data analysis.

This research describes the adaptation and biological validation of EDNA-NGS assay for the detection of the prokaryotic plant pathogens Rs r3b2 and DC3000 in a plant metagenomic sample.

Materials and Methods

General procedures. In an effort to optimize extraction of total nucleic acids from purified bacterial cultures and symptomatic plants, multiple extractions were used including a phenol-chloroform procedure and multiple commercial kits obtained from Qiagen. Total nucleic acids from Rs r3b2 infected potato tubers and pure bacterial cultures extracted by phenol-chloroform separation (Wallis et al. 2007) were provided by William Schneider, Foreign Disease and Weeds Science, United States Department of Agriculture-Agricultural Research Service (USDA-ARS), Ft. Detrick, MD. Total nucleic acids from DC3000 infected tomato plant leaves and pure bacterial cultures were extracted by commercial kits including; DNeasy and RNeasy Mini Plant Kits, Blood and Tissue Kit, and RNeasy Mini Kit (Qiagen, Valencia, CA) (Table 1). All nucleic acid samples were processed using whole genome amplification (WGA) and whole transcriptome amplification (WTA) (Sigma-Aldrich, St. Louis, MO) or a combined WGA/WTA protocol provided by Diana Sherman, Foreign Disease-Weeds Science, United States Department of Agriculture-Agricultural Research Service (USDA-ARS), Fort Detrick, MD and then sent to Biochemistry and Molecular Biology Recombinant DNA and Protein Core Facility (Oklahoma State University, Stillwater, OK) or Foreign Disease and Weeds Laboratory United States Department of Agriculture-Agricultural Research Service (USDA-ARS) (Fort Detrick, MD) for sequencing on a 454 pyrosequencer (Roche GS Junior, 454 Life Sciences, Branford, CT). For an overview of procedures used in processing pure bacterial cultures and plant samples, refer to Figure1. All post sequencing data were processed according to Stobbe et al. (2013).

Generation of e-probes. E-probes were generated for detection of the plant pathogens Rs r3b2 and DC3000 using a laptop computer with an Intel Core i5 processor and 8 GB RAM. Because genome sequence of Rs r3b2 was not available at the time of this research, the related strain Rs GMI1000 race 1 (Salanoubat et al. 2002), was used in its place. The DC3000's (Buell et al. 2003) genome was available. For both pathogens the entire genomes consisted of chromosomes and plasmids. Near neighbor selection was based on phylogenetic relationships and the availability of complete genome data. *R. pickettii* 12D and *P. aeruginosa* PAO1were chosen as near neighbors for comparison with Rs GMI1000 and DC3000, respectively (Table 2). The first bioinformatic script used to generate e-probes aligns the target and near neighbor, then identifies unique sequences to the target pathogen at a predetermined length. The resulting file goes through an additional filtering step to increase specificity. The filtering script works by querying all e-probes, generated in the first step, against the entire NCBI nucleotide database, using BLAST. The resulting file contains each e-probe with an attached label of the particular organism(s) it matched. A third script searchers the labels to identify the e-probes the user wants to keep; all non-target labeled e-probes are removed, leaving the user with e-probes only matching the target pathogen.

Tomato plant growth

Glamour tomato seeds (Victory Seeds, Molalla, OR) were potted using Miracle-Gro Potting Mix (Scotts, Marysville, OH) and placed in a Conviron E8 growth chamber (Conviron, Manitoba, Canada). Plants were grown at 23°C with humidity set at 80% and light intensity set to 40µMOL. Plants were allowed to grow three to four weeks post emergence to obtain mature leaves which were inoculated with DC3000 using sterile swabs.

Bacterial cultures

DC3000 growth and inoculation. A single DC3000 colony was transferred from a King's B agar streak plate (Schaad 1980) to 10 ml King's B broth and incubated at 28°C with constant shaking at 120 rpm using an orbital shaker (Thermo Scientific, Forma 420, Houston, TX) for 48 hr. Subcultures were made by transferring 1ml broth culture to 10ml King's B broth after 48 hours as needed for experimentation. Subcultures had a 4.6 x 10^8 CFU/ml average (Schaad 1980). For swab inoculations, broth cultures were centrifuged at 5000 x g for 5 minutes using a Fisher Scientific, Marathon 6K. Supernatant was removed leaving a bacterial pellet. Sterile cotton swabs were used to gather the pellet and inoculate seven to eight week old tomato plants by rubbing and wounding the underside of tomato leaves. Inoculated tomato plants were covered using 1 gal clear plastic storage bags for 72 hrs. Symptomatic tomato leaves were gathered two to three weeks post inoculation. Tissues of infected leaves were harvested by using razor blades soaked in 70% EtOH, flamed briefly and allowed to cool. Leaves were placed on weighing paper, cut into smaller pieces, and weighed to obtain ≤ 100mg to be used in extraction procedures.

Total nucleic acid extraction

Rs r3b2 nucleic acid extraction. Total nucleic acids were extracted from Rs r3b2-inoculated potato tubers and from pure cultures of Rs r3b2 using phenol-chloroform as described by Wallis et al. (2007) and were provided by William Schneider, Foreign Disease and Weeds Science, United States Department of Agriculture-Agricultural Research Service (USDA-ARS), Ft. Detrick, MD.

DC3000 DNA extraction. Extraction of DC3000 DNA was done using a Qiagen Blood and Tissue Kit following the manufacture's Gram-negative bacteria protocol, from fresh (48hr at 28°C) King's B broth cultures except that 50µl of buffer AE was used instead of the recommended 200µl during the final elution step to increase final concentration.

DC3000 RNA extraction. Extraction of DC3000 RNA was done using a Qiagen RNeasy Mini Kit following the manufacture's protocol, from fresh King's B broth DC3000 cultures (48hr at 28°C). To increase final concentration, the first 30µl eluate from the initial elution was reapplied to the filter to elute a second time per Qiagen RNeasy Mini Kit protocol.

DNA extraction from DC3000 infected tomato plant tissues. To extract DNA from symptomatic tomato plants, the Qiagen Blood and Tissue Kit was used as described above (DC3000 DNA extraction) because of availability and successful extraction from DC3000 cultures.

RNA extraction from DC3000 infected tomato plant tissues. To obtain RNA from infected tomato leaf tissues, the Qiagen RNeasy Plant Mini Kit was used according to the manufacture's protocol including taking the eluate from the initial elution and reapplying it to the filter to elute a second time to increase final concentration.

Removal of plant leaf rRNA

A Ribo-Zero Magnetic Kit (Plant Leaf) (Illumina, Epicentre, Madison, WI) was used to remove plant ribosomal material from extracted total nucleic acids from the tubers of symptomatic potato plants previously inoculated with Rs r3b2 by using plant specific primers according to the manufacture’s protocol (Sooknanan et al. 2011).

Amplification of extracted total nucleic acids

Modified whole genome amplification and whole transcriptome amplification (WGA/WTA). The GenomePlex Whole Genome Amplification Kit (WGA) and TransPlex Whole Transcriptome Amplification Kit (WTA) (Sigma-Aldrich, St. Louis, MO) were performed on all Rs r3b2 infected potato leaf samples, following a combined protocol provided by Diana Sherman, Foreign Disease-Weeds Science, United States Department of Agriculture-Agricultural Research Service (USDA-ARS), Fort Detrick, MD. A concentration of 300ng/µl of total nucleic acids from Rs r3b2 infected potato tuber sample was added to 1.6µl 10x fragmentation buffer and enough nuclease-free water to bring the volume to 16.5µl. The mixture was incubated 4 min at 95ºC and then placed on ice. A volume of 5µl of synthesis buffer and 2.5µl stabilization buffer were added and the tube was incubated at 95ºC for 2 min and then placed on ice. A volume of 1µl of library enzyme was added and the tube was placed in a Biometra T-Professional thermocycler (Goettingen, Germany) at 24ºC for 15 min, 42ºC for 2 hr and 95ºC for 5 min. A WTA master mix (300µl water, 37.5µl WTA amplification master mix, 7.5µl dNTP mix, 5 µl Titanium Taq) was added to 25µl of the 375µl extract prepared above. The sample was divided into 75µl aliquots (5 PCR tubes) and placed in a Biometra thermocycler at 95ºC, 3 min, and then cycled 20 times at 94ºC for 20 sec and 65ºC for 5 min.

DNA only WGA amplification. For DNA extracted using Qiagen kits, a GenomePlex Whole Genome Amplification Kit (WGA) (Sigma-Aldrich, St. Louis, MO) was used following the manufacture's protocol except that the DNA concentration was increased from one to 20ng/µl to produce greater concentrations.

RNA only WTA amplification. For RNA extracted using Qiagen kits, a TransPlex Whole Transcriptome Amplification Kit (WTA) (Sigma-Aldrich, St. Louis, MO) was used following the manufacture's protocol.

Bead sizing

Removing small fragments of genomic material in a sample to be sequenced by the 454 Junior pyrosequencer reduces the instrument's bias towards tiny fragments and increases the proportion of usable reads. In place of the nebulization step used in the Roche 454 Junior pyrosequencer protocol a bead sizing protocol provided by Diana Sherman, Foreign Disease-Weeds Science, United States Department of Agriculture-Agricultural Research Service (USDA-ARS), Fort Detrick, MD, was used. Rs r3b2 samples, including both the Ribo-Zero treated and non-treated samples were subjected to bead sizing following WGA/WTA. A volume of 140µl Agencourt AMPure XP (Beckman Coulter, Brea, CA) magnetic beads were combined with 3µg amplified nucleic acid and water to bring the volume of the reaction to 240µl. The samples were mixed by vortexing for 5 min, spun briefly (1-2 sec) on a Fisher Scientific mini centrifuge at 2000 x g, and placed in a magnetic rack for approximately 5 min to allow binding of nucleic acid fragments of 200bp to 10kb to attach to the beads. The supernatant was removed, 100µl 70% EtOH was added and mixed by vortexing, and, after a brief spin (1-2 sec) at 2000 x g the tube was placed into a magnetic rack (DynaMag-2, Invitrogen, Oslo, Norway) (this step starting with removing the supernatant and adding EtOH was repeated one time).With the tube still in the magnetic rack, the supernatant was removed and the beads allowed to dry for 5-10 min to allow the EtOH to evaporate. The adhering nucleic acid fragments were eluted from the beads by adding 20µl of TE buffer, vortexing, and brief spin (1-2 sec) at 2000 x g, and replaced onto the magnetic rack. Supernatant, containing the nucleic acids, was collected in a sterile 1.5ml tube.

Sequencing

Five separate NGS runs were performed on a Roche 454 Junior pyrosequencer. Three separate sequencing runs with Rs r3b2 as the target pathogen, and two separate sequencing runs with DC3000 as the target pathogen were completed. Material for sequencing was processed according to the manufacturer's protocol (454 Life Sciences, Roche, Bradford, CT) except for the omission of the nebulizer step, which removes significant amounts of DNA and could therefore remove critical target pathogen sequences present in comparatively low titers compared to host DNA, personal communications with Diana Sherman, USDA-ARS. A bead sizing step was performed on Rs r3b2 nucleic acid containing material as a substitute to nebulizing prior to sequencing. DC3000 containing samples were neither nebulized nor the bead sized.

Roche barcoding

Two distinct barcodes were added to one 454 Junior pyrosequencing run containing one tube of DC3000 culture total nucleic acids and one tube of healthy tomato plant total nucleic acids. Barcodes RL11 (ACTATACGAGT) and RL12 (ACTCGCGTCGT) were attached to the total nucleic acid samples from DC3000 and healthy tomato, respectively.

RESULTS

Generation of e-probes

The DC3000 genome size was 6.54Mb with a GC% of 58.3 and that of Rs GMI1000 was 5.81Mb with a GC% of 67. Genome sizes of the near neighbor bacteria were similar in size, with

P. aeruginosa PAO1 at 6.26Mb and a GC% of 66.6 and *R. pickettii* 12D at 5.69Mb and a GC% of 63.3. E-probes with lengths of 15, 20, 25, 40, and 60 nt were generated for DC3000 by comparing the target pathogen to the neighbor using only the initial script discussed in Stobbe et al. (2013). Similarly, e-probes of 20, 25, 40, and 60nt were generated for Rs GMI1000. The first 20 e-probes (15, 20, 25, 40, and 60 nt) generated for DC3000 were queried against NCBI's nucleotide database using BLASTn. Querying the first 20 15 nt e-probes generated for DC3000 produced no matches to the target pathogen. The 15 nt e-probes matched a variety of organisms including the Gram positive bacterium *Bifidobacterium* spp. and *Capra hircus* (goat) and the Gram negative bacterium *Pseudomonas aeruginosa* and *Macaca fascicularis* (crab-eating monkey). Of the 20-nt e-probes, only 5 of the first 20 yielded matches to the target pathogen. The remaining DC3000 e-probes matched multiple non-target organisms as diverse as *Pseudomonas*, *Azotobacter vinelandii* (a soil borne N_2 fixer) and *Chrysemys picta* (painted turtle). E-probes of 25 nt yielded more matches with the target pathogen; in fact, only one non-target match (*Chondrus crispus*, Irish moss) was observed. One 25-nt e-probe (ACCTAGATGTCTCTTAGTCGCGTCT) yielded a score, e-value and coverage with matches to two non-targets. For all remaining 18 25-nt DC3000 e-probes the top match was with the target pathogen. For the 40-nt DC3000 e-probes, only two of the first twenty matched the non-target species *Pseudomonas syringae* pv. *maculicola*, while the remaining 40 nt e-probes matched the target pathogen. Of the 80-nt DC3000 e-probes, only one matched a non-target organism, *Pseudomonas syringae* PT14; the remaining 80-nt e-probes matched the target pathogen. Comparing DC3000 e-probes of all lengths, the percentage of probes matching the target pathogen increased as probe length increased above 20 nt. With the lack in specificity of the e-probes obtained by only using the initial script, additional filters to remove non-target e-probes were used. The final e-probes, after the additional filtering, only contained target specific e-probes for DC3000 and Rs GMI1000.

Extractions

***R. solanacearum* r3b2**. All Rs r3b2 total nucleic acid material was provided by Aaron Sechler, Foreign Disease-Weeds Science, United States Department of Agriculture-Agricultural Research Service (USDA-ARS), Fort Detrick, MD. Concentrations of two nucleic acid extractions from Rs r3b2 inoculated potato tubers were 169.6ng/µl and 260.3ng/µl (Table 1). The 260.3ng/µl sample was processed with the RiboZero kit to remove plant ribosomal nucleic acid, yielding a nucleic acid concentration of 17.2ng/µl. The sample with a nucleic acid concentration of 169.6ng/µl was not processed with the Ribo-Zero kit.

Pseudomonas syringae DC3000. DNA extracted and purified from DC3000 cultures at 4.6 x 10^8 CFU/ml was present at 688.4ng/µl, while RNA concentration was 155.1ng/µl. Nucleic acid extracted from symptomatic tomato leaves using the Qiagen Blood and Tissue Kit was 784.1ng/µl; and RNA obtained using the RNeasy Mini Plant Kit was 719.4ng/µl (Table 1).

Non-inoculated tomato plants. Nucleic acids obtained using the DNeasy Plant Mini Kit from healthy tomato leaves, were 294.3ng/µl and 219.7ng/µl, while that obtained using the Qiagen RNeasy Mini Plant Kit was 819ng/µl and 1365 ng/µl (Table 1).

Amplification of total nucleic acid

WGA/WTA combined protocol. Post extraction, a sample of Rs r3b2 inoculated potato tubers had a nucleic acid concentration of 169.6ng/µl; after WGA/WTA amplification the concentration was 365.4ng/µl. A similar sample from which cytoplasmic rRNA was removed using the Ribo-Zero kit had an initial concentration of 17.2ng/µl; after WGA/WTA amplification the concentration was 154ng/µl .

DNA only WGA amplification. All DNAs subjected to the WGA protocol after extraction were adjusted to 20ng/µl. DNA from DC3000 cultures, or from symptomatic or non-symptomatic tomato leaves, and extracted using the Blood and Tissue Kit, had post-WGA concentrations of 134.4ng/µl, 133.4 ng/µl, and 138.2ng/µl, respectively.

RNA only WTA amplification. All RNAs amplified using the WTA protocol were adjusted to 300ng/µl. RNA from DC3000 cultures, extracted using the RNeasy Mini Kit had a final concentration of 551ng/µl post WTA. RNA from symptomatic and non-symptomatic tomato plants, obtained using the RNeasy Mini Plant Kit, had final concentrations of 575.5 ng/µl and 547.2ng/µl, respectively.

Bead sizing

For the Rs r3b2 sample not processed with the Ribo-Zero kit, the initial concentration was 365.4ng/µl post WGA/WTA with a final concentration of 28.8ng/µl post bead sizing. For the sample processed with the Ribo-Zero kit, the initial concentration was 154ng/µl post WGA/WTA and a final concentration of 19.1ng/µl post bead sizing.

454 pyrosequencing

Five NGS runs were performed on a 454 Junior pyrosequencer on samples including; potato tuber nucleic acids mixed with Rs r3b3 nucleic acids at a 4 to 1 ratio, treated and non-treated Rs r3b2 infected potato samples with Ribo-Zero, barcoded DC3000 culture and tomato plant total nucleic acids, and DC3000 infected tomato plant. With most of the data close to or exceeding Roche's recommended values, all five NGS runs were considered successful (Table 3). For total raw wells, Roche recommends ≤300,000 and all five sequencing runs met this criteria. The recommended read length is >300 bp and all five NGS runs exceeded this value with the potato tuber and Rs r3b3 4 to 1 mix having the shortest average read length of 322.1 bp and the DC3000 infected tomato plant sample having the longest average read length of 433.9 bp. The number of passed filter reads and passed filtered bases for potato tuber and Rs r3b3 nucleic acids mixed, Rs r4b3 infected potato samples treated with Ribo-Zero, and barcoded DC3000 and tomato plant all were below recommended values; however, all produced a significant amount of data (Table 3).

E-probe queries of 454 pyrosequencing

All of the 454 pyrosequencing data output files, termed sample sequence databases (SSDs), were formatted then queried using e-probes and BLASTn. The total numbers of matches are shown in Figures 3 - 7. A match was defined as an instance in which an individual e-probe aligned with a sequence in a SSD, such that the total number of matches was equal to or less than the total number of e-probes. After the query search was conducted the data was parsed according to four different e-value thresholds at 1e-1, 1e-3, 1e-6, and 1e-9.

DISCUSSION

When using current molecular approaches for pathogen detection a diagnostician's decision about whether a sample is positive or negative for a particular pathogen is dependent on a reporter label, in the case of immunoassays, or a fluorescent probe (rtPCR) or small DNA agarose band fragment, in the case of a nucleic acid based approach. In both of these cases, pre-characterization of pathogen protein or nucleic acid sequences is required, and a limited number of different pathogens can be detected in a single assay (Postnikova 2008). Other considerations, such as primer thermodynamics, buffer and $MgCl_2$ concentrations, melting temperatures, non-specific binding, and non-antigen binding should be optimized to avoid possible false positives and false negatives, which could cause costly delays and even erroneous conclusions. The use of 454 pyrosequencing merged with bioinformatics avoids some of the concerns associated with traditional diagnostic approaches. However, NGS, combined with a bioinformatics approach, cannot completely replace other techniques used in diagnostic labs; rather, it is a new tool with the capability of screening a sample for all classes of pathogens in a single assay.

Advantages to using NGS and bioinformatics for pathogen detection include the generation of a complete metagenomic profile of the sample, which includes genetic information on all organisms (pathogens, endophytes, and organelles, both known and unknown) in the infected plants as well as the host itself (Jones 2010; Tyson et al. 2004).

In this study, the 454 Junior pyrosequencer sequencing runs produced millions of sequence bases for each sample, creating a snapshot of all biological material present in the sample at that given moment. The output file from each NGS run was saved as a digital file, accessible indefinitely. Data stored in this way can be manipulated for microbial detection and identification in the future. As we discover new pathogens, or re-discover existing ones, we can explore previous NGS sequences to re-assess possible roles of pathogens in disease outbreaks of the past. Other novel applications are sure to emerge as the costs of NGS technology continue to decline and potential uses continue to arise.

Because EDNA analyses all nucleic acid sequences in a sample it avoids some of the common pitfalls of traditional immunological and molecular diagnostic technologies. Features such as tertiary folding, nucleotide bond strength, percent GC, and polymerase activity are irrelevant.

The objective of this work was to test a simplified bioinformatics approach for dealing with the complexity of NGS metagenomic data as described in EDNA by querying raw NGS data with e-probes, for the purpose of detecting and identifying prokaryotic plant pathogens. To achieve this goal, a SSD was formatted to be searchable, much like using NCBI. The formatted SSD was queried by diagnostic signature sequences (e-probes) without the need for assembly or quality checks. All bioinformatic steps can be performed on a personal laptop computer.

The selection of e-probes appropriate for a given target pathogen is critical for querying the NGS run. E-probes of 15, 20, 25, 40 and 60 nt were generated for DC3000. As a quality check a few e-probes of each size were used in BLASTn searches on NCBI. During e-probe generation the entire genome, including the chromosome(s) and all plasmids, are used for both the target pathogen and near neighbor. In contrast, traditional molecular primer/probe design considers only a small section of the genome, such as 16S rRNA, 18S rRNA or the ITS regions. Because entire genomes are used to design e-probes, it is expected that a few sequences could generate matches with universal or common genes in non-target organisms. However, we anticipate that a majority will match only our targeted pathogen and with the use of additional filters in the e-probe design suboptimal e-probes are removed.

When considering DC3000 e-probes of lengths between 15 to 60 nt we expected and observed, among the first 20 e-probes at each length, that as probe length increases the likelihood of matching with a non-target decreases. The first 20 DC3000 e-probes of 15 nt failed to match with the target, which is not surprising considering that the probability of 15 nt matching a random sequence in the NCBI database is greater than 60 nt sequences matching randomly to sequence in the NCBI database. Even in molecular nucleic acid based approaches, primers of 15 nt or less are undesirable due to the likelihood that they will bind to non-target sequences. In this study the 20 nt DC3000 e-probes were no more suitable than the 15 nt e-probes. Not until the length was increased to 25 nt and longer did we observe consistent matches to the target pathogen. Since lengths equal to or greater than 25 nt are suitable for molecular routine detection, since the first twenty15 nt e-probes for DC3000 yielded poor results, Rs r3b2 were designed at lengths of only 20, 25, 40, and 60 nt.

Several total nucleic acid extraction were compared. The basic phenol-chloroform procedure (Wallis et al. 2007) used on samples containing Rs r3b2 is relatively inexpensive compared to kit extractions. The kits used for samples containing DC3000 cost $155 to $333, and since every sample required two kits, one each for DNA and RNA, the total cost reaches $500 to $600. Extraction kits generally yield greater nucleic acid yields than phenol-chloroform separations, but higher nucleic acid concentrations may not result in a better sequencing read. Considering all of the factors, the use of commercial extraction kits did not significantly improve the performance of EDNA for detecting prokaryotes in our experiments.

When plants are imported, arrive at a diagnostic laboratory or are purchased for home use they may be carrying pathogens even if no disease symptoms are visible (Lemay 2003). Pathogen titers vary from plant to plant and even within different tissues of a single plant. Because the plant host genome will make up a majority of the nucleic acid obtained from the initial extractions it is important to limit the downstream bias as the sample is prepared for 454 sequencing. To address this issue, the WGA kit, which chemically fragments all nucleic acids to smaller fragments, reduces large host DNA fragments to be closer in size to the bacterial and viral genomes. If a pathogen is present in low titer it may be possible to take advantage of cellular communication among the pathogens for amplification. WTA was used to enhance this ‘transcriptome noise.’ Together, the WGA/WTA treatments reduce host bias and increase pathogen transcriptome activity on the molecular level.

Before sequencing, post WGA/WTA Rs r3b2 samples were subjected to bead sizing, in which smaller fragments (≥ 200 bp, smaller than the 300-500 bp size considered optimal for Roche 454 sequencing (Margulies et al. 2005)) observed in gel smears of the WTA/WGA samples were removed. Sequencing of these modified Rs r3b2 samples was done without a nebulizing step since fragment size reduction was done by bead sizing. DC3000 samples were processed without bead sizing. The 454 protocol, without the nebulizing step, was exactly the same for samples of both Rs and DC3000. There was not enough difference between the two to warrant the additional cost and procedure of performing bead sizing (Table 3). The addition of processing steps, such as removal of host RNA by Ribo-Zero and removal of sub-optimal DNA fragments by bead sizing, do not improve sequencing results and may even reduce sequencing efficiency.

Bioinformatic analysis of the sequencing data began with formatting the raw 454 output (sample sequence databases or SSD) using a formatting script that allows the file to be queried by BLAST. Next, a BLASTn query using target e-probes generated a file that was parsed at various e-values by searching the BLASTn output file for every match. Every match was assigned an e-value. The parser script was programmed so that when a match was detected at or below the set e-value threshold of 1e-1, 1e-3, 1e-6, or 1e-9 it accepted the matching e-probe, identifying a portion of the target pathogen's genome.

The 454 pyrosequencing SSD of barcoded samples of pure DC3000 and healthy tomato tissues generated 15,582 DC3000 reads and 47,057 reads for the healthy tomato. Out of the 18,788 20 nt e-probes, only 1 match was obtained after parsing at an e-value of 1e-1. No matches were observed when the parsing criterion was set to a more stringent e-value of 1e-3. This result is not surprising considering that BLAST searches with the 20 nt e-probes identified only a few matches to the target. The larger e-probes of 25, 40, and 60 nt were more effective, with 59, 232, and176 matches, respectively, when parsed at 1e-1. In order to assess the e-probes that matched the barcoded SSD, all 25 nt e-probes that matched the target and parsed at 1e-1 were checked by BLASTn on the NCBI database. Only one e-probe (AAAGTCAAAGTCAAAGTCAAAGTCA) out of the 59 e-probes matched a non-target sequence. The remaining 58 e-probes all matched the target DC3000 when queried on the NCBI webpage. However, this e-value provides little stringency and would potentially accept non-specific matches. Decreasing the e-value to a more stringent 1e-3 also decreased the total number of matches to 2, 3, 2, and 0 for e-probes at lengths of 25, 40, 60 and 80, respectively (Figure 3). Because the total number of matches decreased with increasing stringency, there is more confidence in calling this sample positive for DC3000. Similarly, Stobbe et al (2013) reported that optimal parsing occurred at 1e-6 or even 1e-9 due to false positives found at higher e-values.

Developing diagnostics *in silico,* as described in Stobe et al. 2013, is quick and relatively inexpensive; however, assays developed in this manner must always be validated experimentally. As discussed above, the barcoded SSD revealed that pathogen sequences were present in the sequenced sample. In a real application a diagnostician might not know what pathogens, if any, are present. Experiments done where DC3000 was used to inoculate a tomato plant allowed assessment of the potential to detect prokaryotic pathogens in a complex metagenomic sample. For example, when the same e-probes queried in the barcoded SSD run were used the symptomatic tomato leaf tissue was found positive for DC3000. When the 20 nt e-probes were tested they did not generated matches, but the 25, 40 and 60 nt e-probes generated high numbers of matches. For example, 46, 825 and 225 for 25, 40, and 60 nt e-probes, respectively when parsed at an e-value of 1e-1. Interestingly, the 40 and 60 nt e-probes generated a higher number of matches when parsed at 1e-3 and 1e-9, whereas the barcoded SSD run did not. This result could be due in part to the absence of barcoding. Table 3 shows that the total passed filtered bases or nt bases that were considered of good quality by the sequencing software for the non-barcoded infected tomato plant SSD were nearly three times that of the barcoded SSD run. By not barcoding there is greater sequencing coverage within the sample.

Rs r3b2, a select agent, is a major concern to the potato industry in the U.S. The ability to detect plant pathogenic select agents is critical to the U.S. biosecurity efforts. Being able to detect this pathogen using e-probes of 25, 40, 60 and 80 nt with parsing at 1e-1, 1e-3, and 1e-6 was demonstrated (Figures 5-7). These sequencing runs include healthy potato and Rs r3b2 total nucleic acids mixed at a 4 to 1 ratio, and Rs r3b2 infected potato tuber that was treated with Ribo-Zero and Rs r3b2 infected potato tuber that was untreated. GMI1000 e-probes of all sizes generated high numbers of matches, at all e-probe sizes when parsed at 1e-1. Using the Spiked SSD at a 4 to 1 ratio and the same e-probes but parsing at the more stringent e-value of 1e-3, the total numbers of matches were reduced and the confidence of calling a match as a true identifier of the target pathogen increased. Both the 40 and 25 nt Rs GMI1000 e-probes generated the greatest number of matches when parsing at more stringent e-values, suggesting the shorter lengths might be more appropriate than the 60 and 80 nt e-probes.

A combined WGA/WTA protocol was performed on the two 454 sequencing runs of Rs r3b2 infected potato tubers including one SSD not processed through Ribo-Zero and one SSD processed through Ribo-Zero. AMPure bead sizing was done on both SSDs to remove small fragments. The Ribo-Zero kit was used on one SSD to remove host rRNA. The totals of raw sequencing data (Table 3) shows that the RiboZero-treated SSD generated only half the number of passed filter bases and passed filter reads compared to the sample without such treatment, suggesting that RiboZero processing reduces the chances of detecting the target pathogen sequences; however, when the same e-probes were used to query both SSDs there was little difference in detection at a parsed e-value of 1e-1 (Figures 6 and 7). Contrarily, when more stringent e-values of 1e-3 and 1e-6 were used there is a noticeable increase in matches with the sample not treated with Ribo-Zero kit. Even though the Ribo-Zero kit contains plant specific primers that remove most of the plant ribosomal RNA (Epicentre 2013), there is a potential of reducing pathogen ribosomal material, which could lessen the detection ability of the system.

An additional assessment of both the Rs GMI1000 and DC3000 e-probes was needed to analyze the test specificity. The 454 pyrosecuencing SSD of DC3000 infected tomato plant was queried using Rs GMI1000 e-probes at lengths of 25, 40, 60, and 80 nt. All queries were parsed at e-values of 1e-1, 1e-3, 1e-6, and 1e-9. The observed results indicate an elevated number of matches when parsing at 10^{-1}; however, there were also elevated numbers of matches of 38, 31, and 5 for the 25, 40, and 60 nt e-probes when parsed at 1e-3, and 4 and 11 matches with 25 and 40 nt e-probes when parsed at 1e-6. Typically, very few, if any, matches are expected. This suggests the possibility of a *Ralstonia solanacearum* species being present in the original extracted material. To test the specificity of the DC3000 e-probes, the 454 pyrosequencing run of

Rs r3b2 infected potato tuber not processed with Ribo-Zero was used. For e-probes at a length of 25, 40, 60, and 80 nt and parsed at 1e-1, there was an observed high number of matches of 78, 279, 225, and 50, respectively. Increasing the parser stringency to lower e-values generated only 3 matches with the 40 nt e-probe when parsed at 1e-3.

CONCLUSION

The ability of the NGS to be used as a diagnostic tool for detection of prokaryotic plant pathogens was demonstrated. Extraction of sample nucleic acid by traditional and inexpensive phenol-chloroform separation was as effective and more cost-efficient than the use of commercial kits. Reduction of host background using a RiboZero kit was costly ($90/reaction, as of 12/02/2013) and provided no observable benefit in detecting the pathogen Rs r3b2. Similarly, bead sizing with AMpure beads provided little to no benefit. NGS, combined with EDNA, is a valuable tool for rapid screening of multiple pathogens at little cost. However, the current bioinformatics and manipulation of computer scripts to develop e-probes and query NGS data requires training to operate. Further development of simple and user-friendly programs that automate the design of e-probes and querying of NGS data will be necessary for this technology to become more usable. Additional work to validate the detection threshold of the EDNA system and to address specificity concerns on the e-probes will be required along with adaptation of EDNA for strain typing and detection of genetically modified organisms.

LITERATURE CITED

Buell C, Joardar V, Lindeberg M, Selengut J, Paulsen T, Gwinn M, Dodson R, Deboy R, Durkin A, Kolonay J, Madupu R, Daugherty S, Brinkac L, Beanan M, Haft D, Nelson W, Davidsen T, Zafar N, Zhou L, Liu J, Yuan Q, Khouri H, Fedorova N, Tran B, Russell D, Berry K, Utterback T, Van Aken S, Feldblyum T, D'Ascenzo M, Deng W, Ramos A, Alfano J, Cartinhour S, Chatterjee A, Delaney T, Lazarowitz S, Martin G, Schneider D, Tang X, Bender C, White O, Fraser C, Collmer A. 2003. The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. Proc Natl Acad Sci 100:10181-10186.

Champoiseau P, Momol T. 2008. Bacterial Wilt of Tomato. In: Florida UO, ed.:The United States Department of Agriculture - National Research Initiative Program.

Cronn R, Knaus B, Liston A, Maughan P, parks M, Syring J, Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Botany 99:291-31

Jones W. 2010. High-throughput sequencing and metagenomics. Estuaries and Coasts 33:944-952.

Karypis G, Eui-Hong H, Kumar V, 1999. Chameleon: hierarchical clustering using dynamic modeling. Computer 32:68-75.

Kim J, Lee G, Kim J, Kwon J, Kwon S. 2008. The development of rapid real-time PCR detection system for *Vibrio parahaemolyticus* in raw oyster. Lett Appl Microbiol 46:649–654.

Lemay A, Redlin S, Fowler G, Dirani M. 2003. Pest data sheet: Ralstonia solanacearum race 3 biovar 2. USDA-APHIS-PPQ. Center for Plant health Science and Technology Plant Epidemiology and Risk Analysis Laboratory.

Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L, Berka J, Braverman M, Chen Y, Chen Z, Dewell S, Du L, Fierro J, Gomes X, Godwin B, He W, Helgesen S, Ho C, Irzyk G, Jando S, Alenquer M, Jarvie T, Jirage K, Kim J, Knight J, Lanza J, Leamon J, Lefkowitz S, Lei M, Li J, Lohman K, Lu H, Makhijani V, McDade K, McKenna M, Myers E, Nickerson E, Nobile J, Plant R, Puc B, Ronan M, Roth G, Sarkis G, Simons J, Simpson J, Srinivasan M, Tartaro K, Tomasz A, Vogt K, Volkmer G, Wang S, Wang Y, Weiner M, Yu P, Begley R, Rothberg J. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-80.
Postnikova E, Baldwin C, Whitehouse CA, Sechler A, Schaad N. 2008. Identification of bacterial plant pathogens using multilocus polymerase chain reaction/electrospray ionization-mass spectrometry. Phytopathology 98:1156-1164.

Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, Arlat M, Billault A, Brottier P, Camus J, Cattolico L, Chandler M, Choisne N, Claudel-Renard C, Cunnac S, DemangeN, Gaspin C, Lavie M, Moisan A, Robert C, Saurin W, Schiex T, Siguier P, Thebault P, Whalen M, Wincker P, Levy M, Weissenbach J, Boucher A. 2002. Genome sequence of the plant pathogen *Ralstonia solanacearum*. Nature 415:497-502.

Schaad N. 1980. Laboratory Guide for the Identification of Plant Pathogenic Bacteria. The American Phytopathological Society, St. Paul, MN. 72p. 3.

Sooknanan R, Agnes R, Hitchen J, Khanna A. 2011 Improved technology for ribosomal RNA removal and directional RNA-Seq library preparation. Epicentre Biotechnologies. http://www.epibio.com/applications/rna-sequencing/rrna-removal/ribo-zero-rrna-removal-kits-(plant)?details.

Stobbe, A, Daniels J, Espindola A, Ruchi V, Melcher U, Ochoa-Corona F, Garzon C, Fletcher J, 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 94:356-366.

Tyson G, Chapman J, Hugenholtz P, Allen E, Ram R, Richardson P, Solovyev V, Rubin E, Rokhsar D, Banfield J. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43.

Wallis C, Stone A, Sherman D, Damsteegt V, Gildow F, Schneider W. 2007. Adaptation of plum pox virus to a herbaceous host (*Pisum sativum*) following serial passages. J Gen Virol 88:2839-2845.

Xin X, He S. 2013. *Pseudomonas syringae* pv. *tomato* DC3000: A model pathogen for probing disease susceptibility and hormone signaling in plants. Ann Rev Phytopathol 51:473-498.

Zhang G, Lu Z, Wan F, Lovei G. 2007. Real-time PCR quantification of *Bemisia tabaci* (Homoptera: Aleyrodidae) B-biotype remains in predator guts. Mol Ecol Notes 7:947–954.

Zhao Y, Thilmony R, Bender C, Schaller A, He S, Howe G. 2003. Virulence systems of *Pseudomonas syringae* pv. *tomato* promote bacterial speck disease in tomato by targeting the jasmonate signaling pathway. Plant J 36:485-99.

TABLES

Table 1. Commercial kits and non-commercial methods used for nucleic acid extraction with nucleic acid concentrations and cost per reaction from samples containing bacteria, infected plant tissue and healthy plant tissue.

Sample ID	Extraction protocol	Nucleic acid concentration	260/280	260/230	DNA/RNA	Cost per reaction
Potato tuber	[a]Phenol/chloroform	30	1.79	.94	DNA/RNA	n/a
R. solanacearum r3b2 infected potato tuber	[a]Phenol/chloroform	169.6 260.3	1.62 1.61	1.26 1.24	DNA/RNA	n/a
P. syringae DC3000	Qiagen Blood & Tissue Kit	688.4 1730.3	2.07 2.17	2.3 2.35	DNA	[b]$3.10
P. syringae DC3000	Qiagen RNeasy Mini Kit	155.1 159.7	1.92 1.96	1.29 1.46	RNA	[b]$5.52
DC3000 infected tomato plant	Qiagen Blood & Tissue Kit	498.4 784.1	1.31 1.24	0.58 0.58	DNA	[b]$3.10
DC3000 infected tomato plant	Qiagen RNeasy Mini Plant Kit	719.4 792	1.89 1.93	0.68 0.74	RNA	[b]$6.66
Tomato plant	Qiagen DNeasy Plant Mini Kit	294.3 237.2	1.40 1.42	0.91 0.93	DNA	[b]$4.16
Tomato plant	Qiagen RNeasy Plant Mini Kit	1365.0 819.0	1.65 2.01	0.95 1.5	RNA	[b]$6.66

[a]Wallis C, et al. 2007. J Gen Virol 88:2839-2845.

[b]Pricing as of10/14/2013

Table 2. Target pathogens and near neighbors, with accession number, used for generation of e-probes. Accession numbers are from GenBank and accessed through the National Center for Biotechnology Information (NCBI).

Target pathogen	**Accession #**	**Near neighbor**	**Accession #**
R. solanacearum GMI1000	NC_003295.1 NC_003296.1	*Ralstonia pickettii* 12D	NC_012856.1 NC_012857.1 NC_012855.1 NC_012849.1 NC_012851.1
P. syringae pv. *tomato* DC3000	NC_004578.1 NC_004633.1 NC_004632.1	*Pseudomonas* *aeruginosa* PAO1	NC_002516.2
Accession numbers will link to chromosomes and plasmids when entered on NCBI webpage.			

Table 3. Results of five separate 454 Junior pyrosequencing runs. One sequencing run contained a 4:1 mixture of potato tuber to *Ralstonia solanacearum* r3b2 (Rs r3b2) total nucleic acids, respectively. One sequencing run was with a potato tuber infected with Rs r3b2, while another run was with a potato infected with Rs r3b2 processed through a Ribo-Zero kit that removes host RNA. A barcoded sequencing run was performed using tomato plant and *Pseudomonas syringae* DC3000 total nucleic acids. The final sequencing run was of a tomato plant infected with DC3000.

	4:1 Potato tuber : Rs r3b2	Potato infected with Rs r3b2 **No RiboZero**	Potato infected with Rs r3b2 **RiboZero**	Barcoded Tomato plant & DC3000	Tomato plant infected with DC3000	Roche recommended values
Total raw wells	229,810	232,938	228,689	226,692	235,492	**≤ 300,000**
Average read length	322.1	353.9	324.5	391.9	433.9	**> 300 bp**
Number of passed filter reads	64,927	111,693	51,938	64,719	160,254	**≥ 88,000**
Total passed filter bases	20,911,623	39,531,357	16,851,367	26,109,600	69,537,825	**> 27 million**

FIGURES

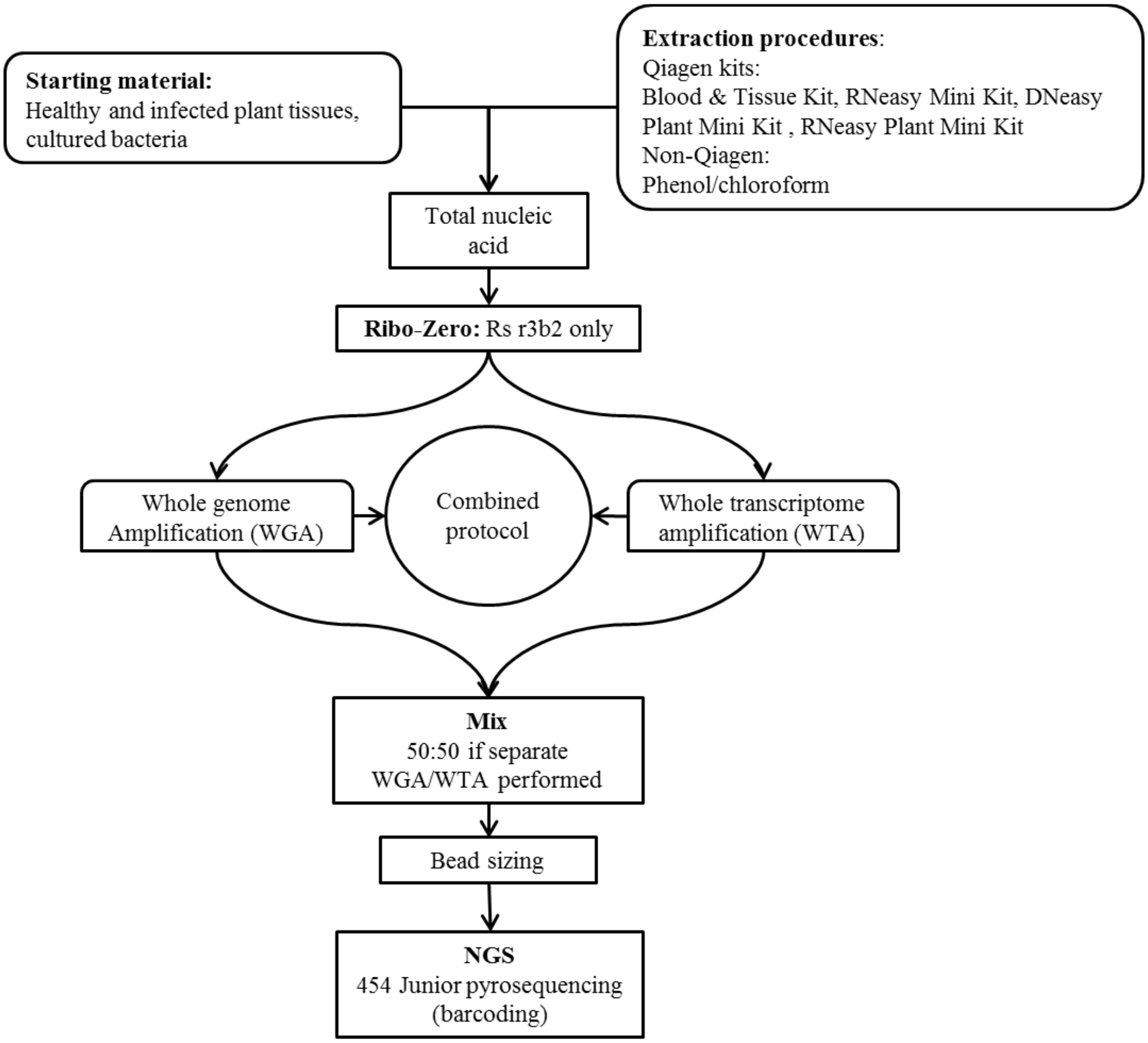
Starting material:
Healthy and infected plant tissues, cultured bacteria
Extraction procedures:
Qiagen kits:
Blood & Tissue Kit, RNeasy Mini Kit, DNeasy Plant Mini Kit , RNeasy Plant Mini Kit
Non-Qiagen:
Phenol/chloroform
Total nucleic acid
Ribo-Zero: Rs r3b2 only
Whole genome Amplification (WGA)
Combined protocol
Whole transcriptome amplification (WTA)
Mix
50:50 if separate WGA/WTA performed
Bead sizing
NGS
454 Junior pyrosequencing (barcoding)

Figure 1. Experimental workflow used in processing *Ralstonia solanacearum* race 3 biovar 2 and *Pseudomonas syringae* pv. *tomato* DC3000 cultures, healthy potato and tomato plants, potato plants infected with Rs r3b2 and tomato plants infected with DC3000. Total nucleic acids were obtained and processed through WGA/WTA amplification and sized with AMPure XP beads prior to sequencing.

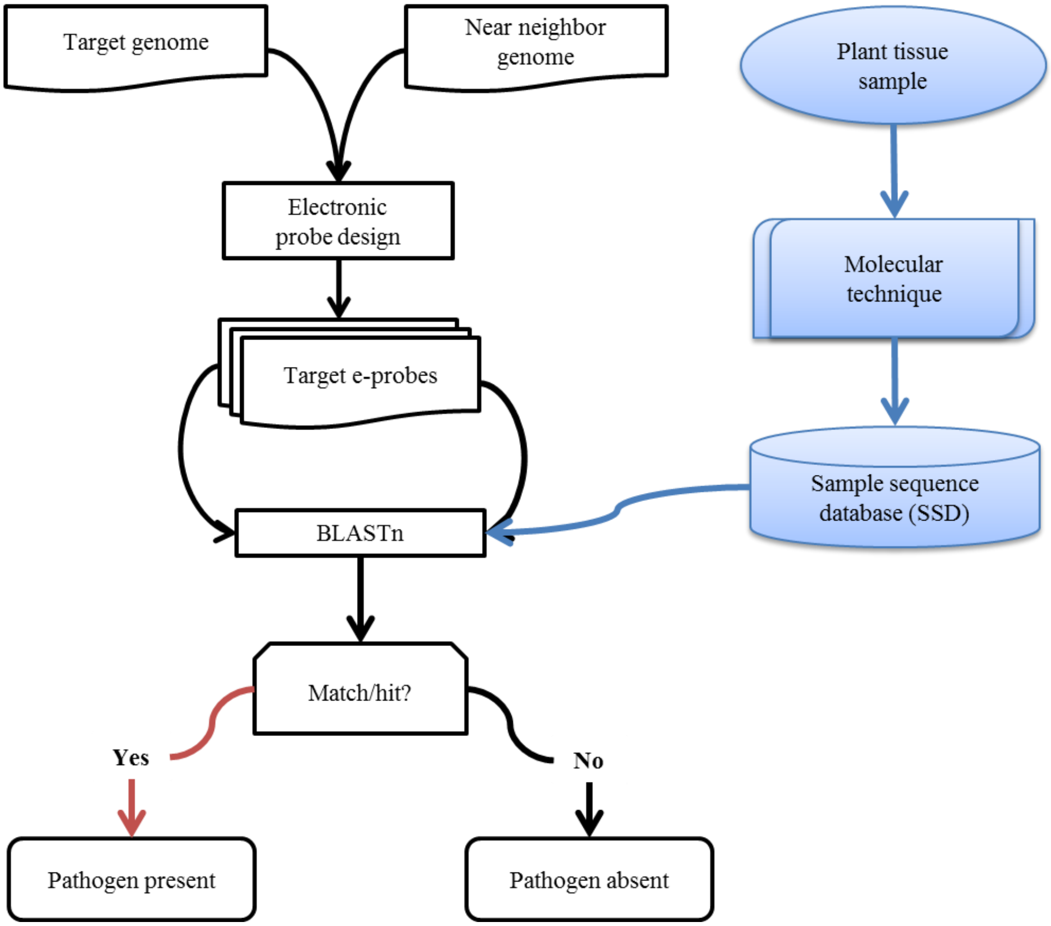

Figure 2. Use of e-probe Diagnostic Nucleic acid Analysis (EDNA) to design electronic probes and query a next generation sequencing database. Plant sample tissue is obtained from symptomatic and non-symptomatic plants.

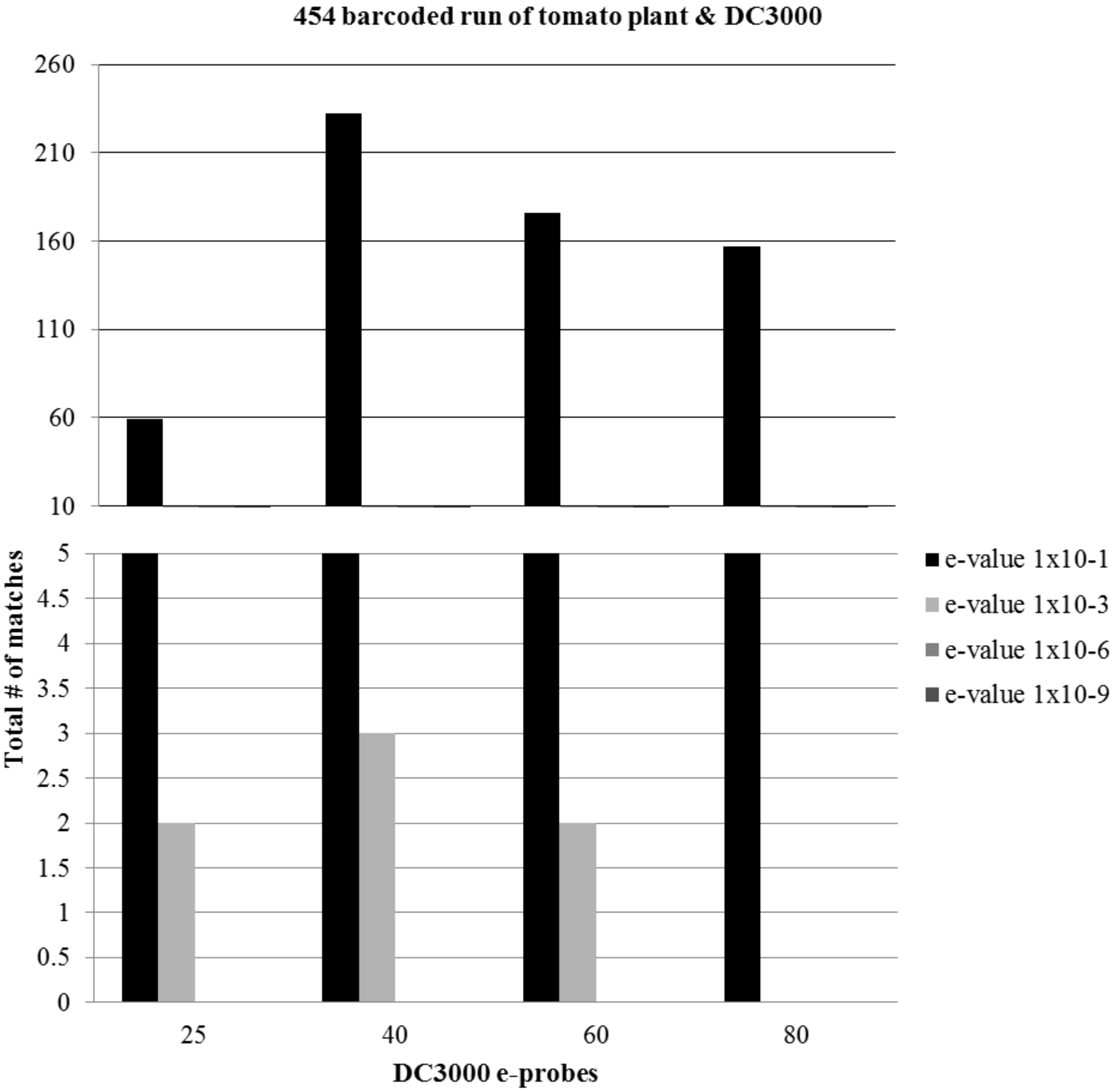

Figure 3. Results of an EDNA search showing total matches using *Pseudomonas syringae* pv. *tomato* DC3000 e-probes of a barcoded 454 pyrosequences run of healthy tomato and *Pseudomonas syringae* pv. *tomato* DC3000 culture.

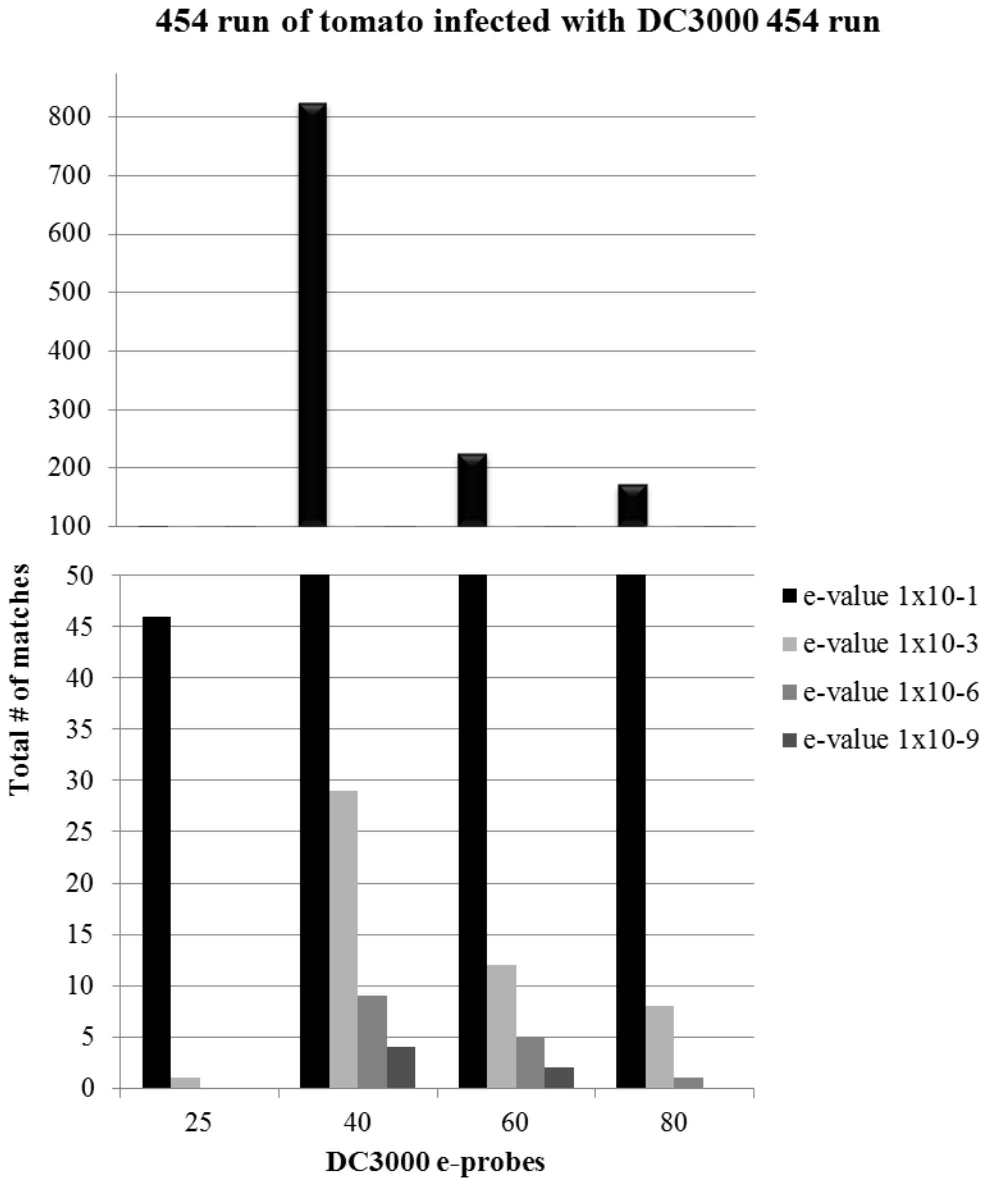

Figure 4. Results of an EDNA search showing total matches using *Pseudomonas syringae* pv. *tomato* DC3000 e-probes, of a 454 pyrosequence run of symptomatic tomato plant infected with *Pseudomonas syringae* pv. *tomato* DC3000.

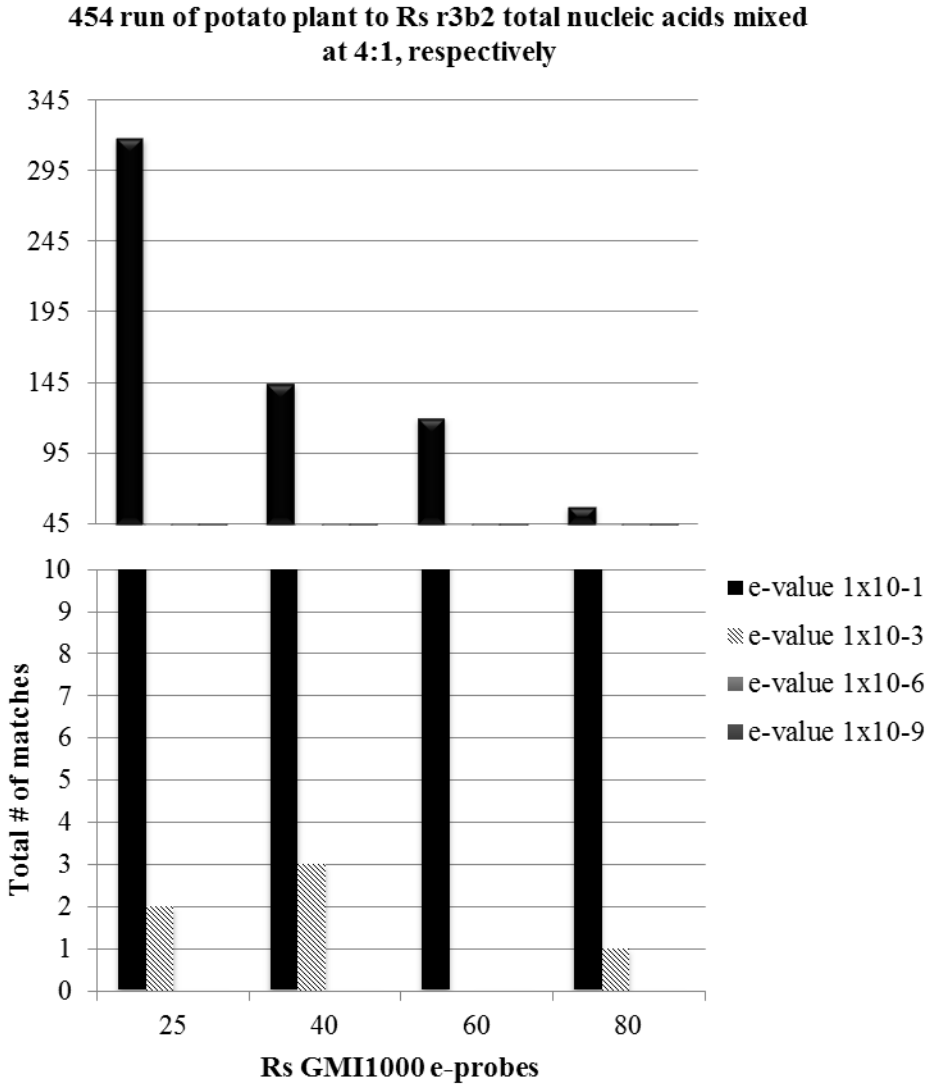

Figure 5. Results of an EDNA search showing total matches using *Ralstonia solanacearum* GMI1000 e-probes, of a 454 pyrosequence run of potato leaf and *Ralstonia solanacearum* race 3 biovar 2 total nucleic acids mixed at a 4 to 1 ratio, respectively.

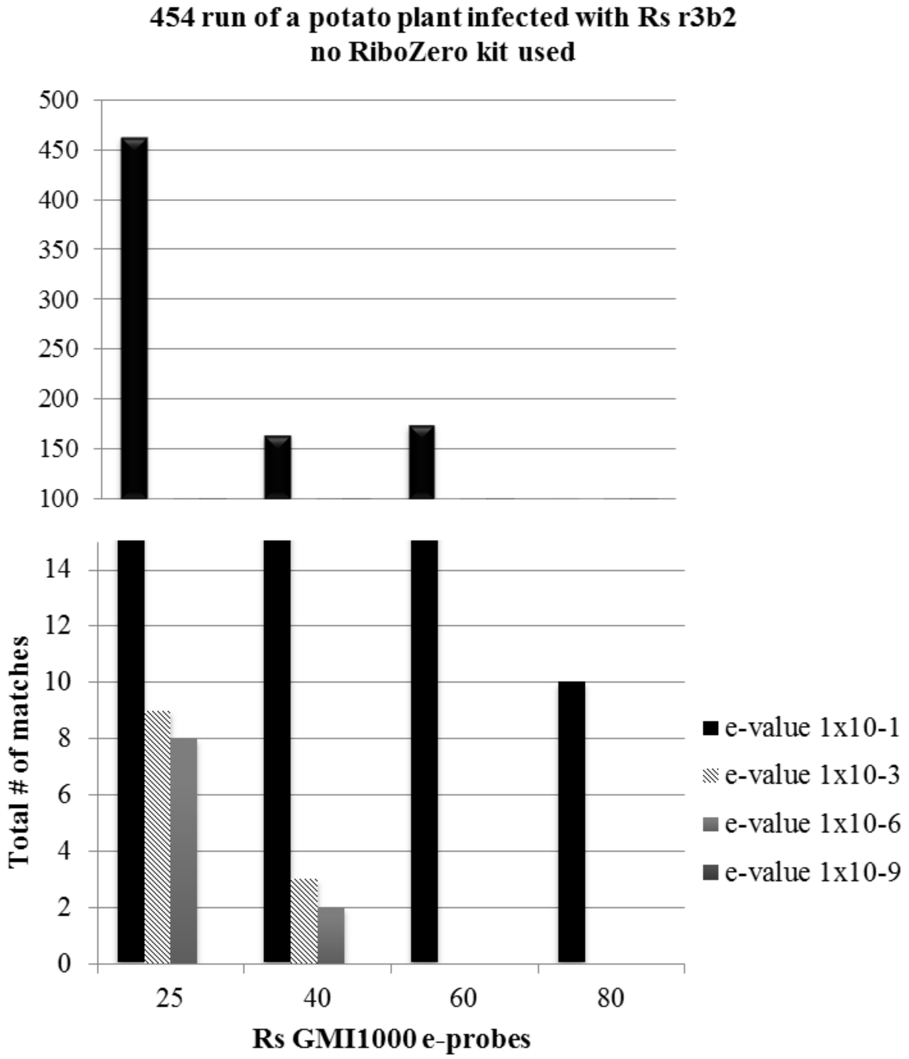

Figure 6. Results of an EDNA search showing total matches using *Ralstonia solanacearum* GMI1000 e-probes, of a 454 pyrosequence run of a symptomatic potato plant infected with *Ralstonia solanacearum* race 3 biovar 2.

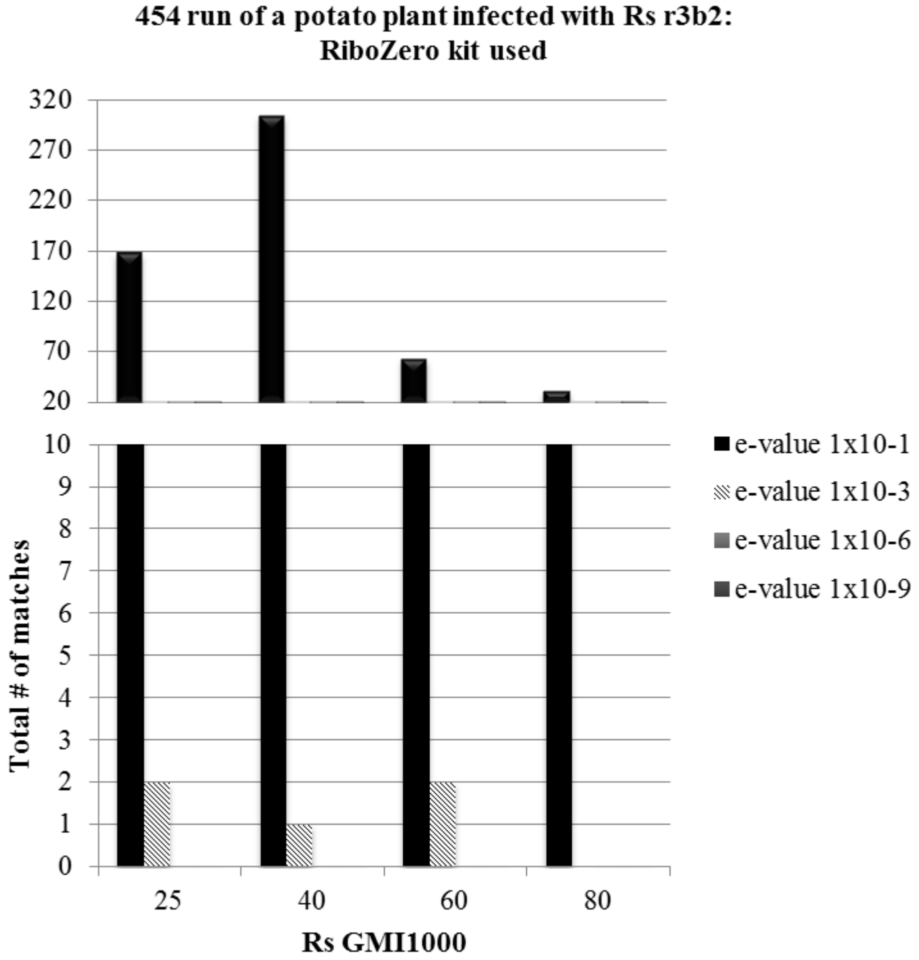

Figure 7. Results of an EDNA search showing total matches using *R. solanacearum* GMI1000 e-probes, of a 454 sequencing run of a symptomatic potato plant inoculated with *R. solanacearum* r3b2 and processed through a Ribo-Zero kit to remove plant rRNA.

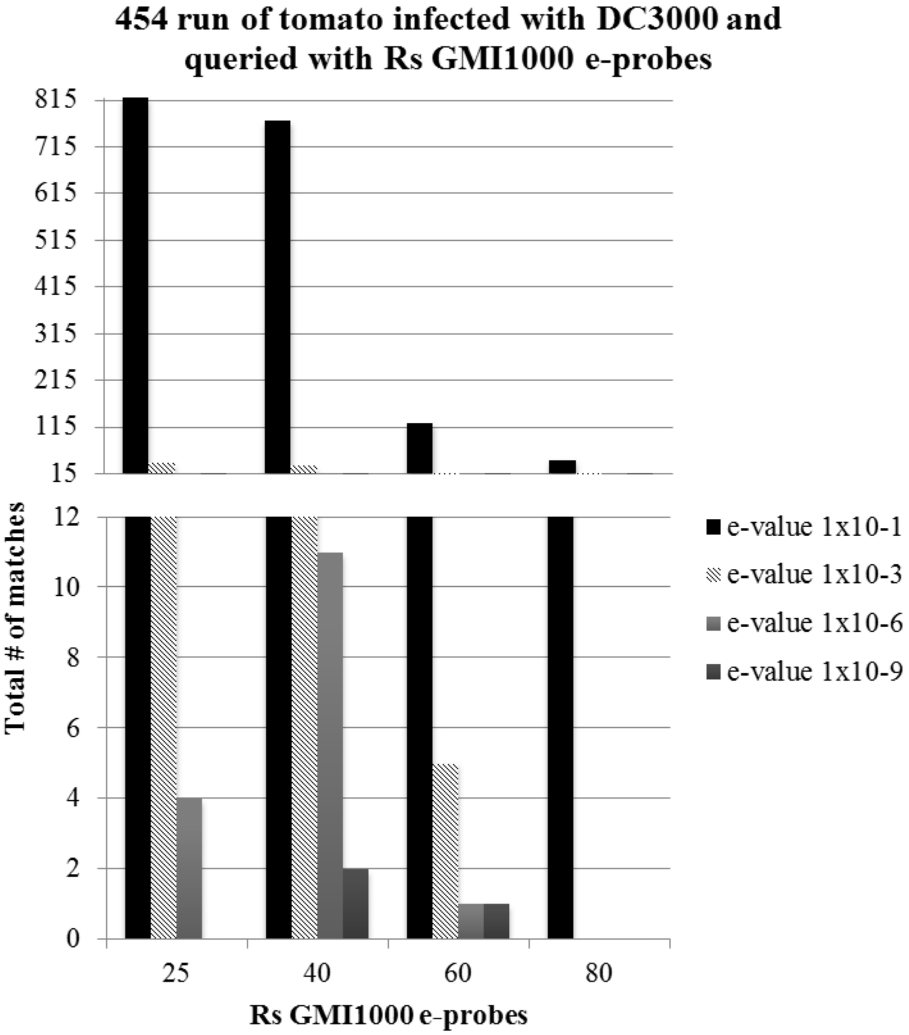

Figure 8. Results of an EDNA search showing total matches using *R. solanacearum* GMI1000 e-probes, of a 454 pyrosequence run of a symptomatic tomato plant infected with *Pseudomonas syringae* pv. *tomato* DC3000.

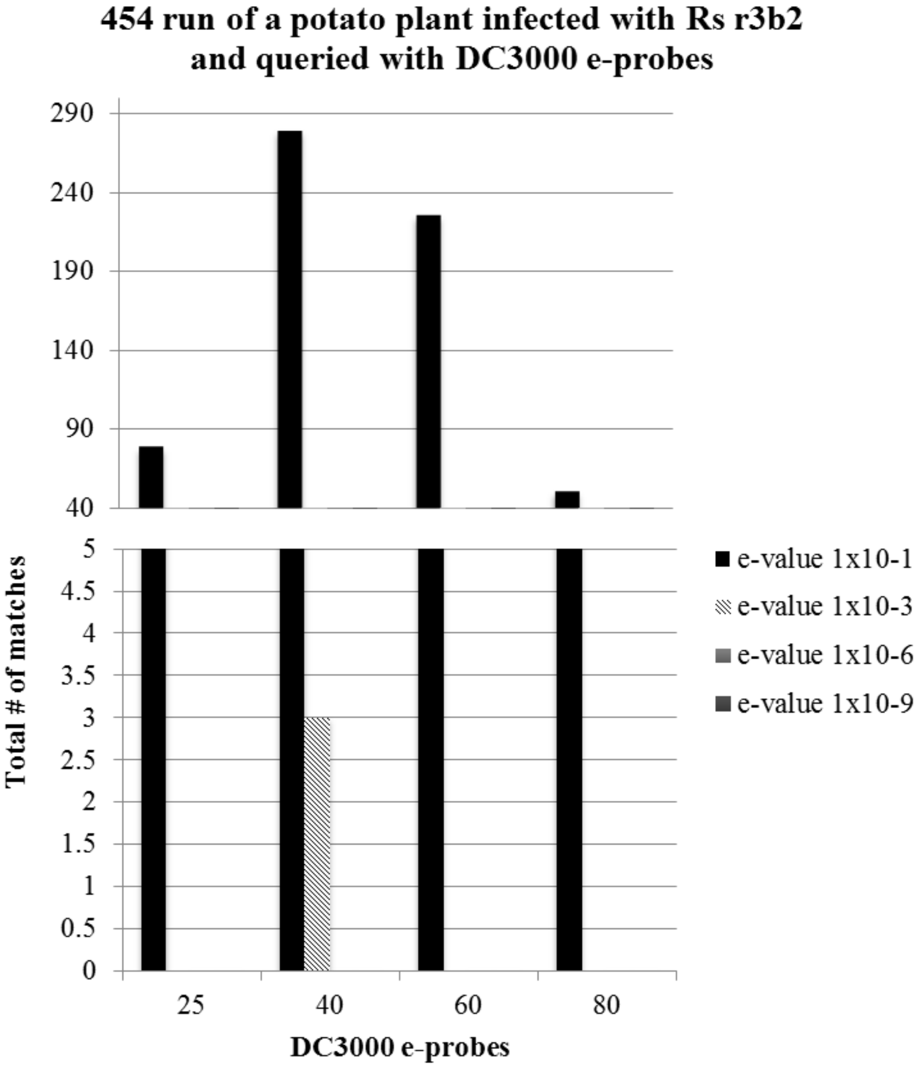

Figure 9. Results of an EDNA search showing total matches using *Pseudomonas syringae* pv. *tomato* DC3000 e-probes, of a 454 pyrosequence run of a symptomatic potato plant infected with *Ralstonia solanacearum* race 3 biovar 2.

VITA

Jon Michael Daniels

Candidate for the Degree of

Master of Science

Thesis: THE USE OF NEXT GENERATION SEQUENCING TO DETECT ALL CLASSES OF PLANT PATHOGENIC MICROORGANISMS

Major Field: Entomology and Plant Pathology

Biographical:

Education:

Completed the requirements for the Master of Science in Entomology and Plant Pathology at Oklahoma State University, Stillwater, Oklahoma in December, 2013.

Completed the requirements for the Bachelor of Science in Biology at Rogers State University, Claremore, Oklahoma in 2009.

Experience:

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

Undergraduate Independent Researcher. Oral Biofilm Destruction via Bacteriophage. Department of Biology, Roger State University, Claremore, Oklahoma.

Professional Memberships:

American Phytopathological Society