MASSIVELY PARALLEL SEQUENCING (MPS) AS A DIAGNOSTIC AND FORENSIC ANALYSIS TOOL FOR IMPORTANT FUNGI AND CHROMISTA PLANT PATHOGENS

By

ANDRES ESPINDOLA

Bachelor of Science/Science in Biotechnology

Escuela Politécnica del Ejército

Quito, Ecuador

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
July, 2013

MASSIVELY PARALLEL SEQUENCING (MPS) AS A DIAGNOSTIC AND FORENSIC ANALYSIS TOOL FOR IMPORTANT FUNGI AND CHROMISTA PLANT PATHOGENS

Thesis Approved:

Dr. Carla Garzón

Thesis Adviser

Dr. Stephen Marek

Dr. Peter Hoyt

ACKNOWLEDGEMENTS

I would like to extend my deepest thanks to Dr. Carla Garzon, my major professor, for being patient and for her encouraging advisement. I'm very grateful for the opportunity she gave me to study in this prestigious program at Oklahoma State University. I would also like to express my gratefulness to Dr. Stephen Marek and Dr. Peter Hoyt for being part of my advisory committee, for their invaluable input, for their advice and for the knowledge that you imparted in me while working on my research. I would also like to thank Dr. William Schneider for guiding me through important steps on my research and for giving me the opportunity to participate during two summers in internships in the USDA-ARS Disease-Weed Science Research Unit.

I wish to acknowledge Dr. Jacqueline Fletcher, for giving me the opportunity of being part of the National Institute of Microbial Forensics and Food & Agricultural Biosecurity (NIMFFAB) during my research, and for allowing me to participate actively throughout the development of this project. Thanks to all the members of this project, particularly: Dr. Francisco Ochoa-Corona, Dr. Ulrich Melcher, Tony and Jon. Finally, I'd like to thank everyone who was part of Dr. Carla Garzon's lab team, principally Patricia Garrido and Gabriela Orquera for being always ready to help me when needed.

Name: ANDRES SEBASTIAN ESPINDOLA

Date of Degree: JULY, 2013

Title of Study: MASSIVELY PARALLEL SEQUENCING (MPS) AS A DIAGNOSTIC AND FORENSIC ANALYSIS TOOL FOR IMPORTANT FUNGI AND CHROMISTA PLANT PATHOGENS

Major Field: PLANT PATHOLOGY

Abstract: Different mechanisms are involved in the movement of plant pathogens such as rainwater, wind, vectors, etc. Eukaryotic plant pathogens, and principally fungal and chromista are mostly dispersed by air-, water- or soilborne spores. Early stage infections caused by fungal/oomycete spores may not be detected until signs or symptoms are developed visibly. Although specific and sensitive molecular and serological diagnostic methods have been developed to increase the capability of plant pathogen detection, their scope is limited to a single or a few organisms. In addition, as a part of molecular-based detection methods, sequencing has also being widely utilized as a diagnostic tool, principally Sanger sequencing. Its specificity and sensitivity depends of the targeted locus uniqueness and copy number, and its detection scope is also narrowed to one sample per reaction. The detection sensitivity and specificity of PCR based detection methods is improved when using more than one locus. Massively parallel sequencing (MPS) randomly sequences numerous various DNA fragments from one single DNA sample in parallel. As a result, multiple loci are targeted, and among them, unique signatures of pathogens are highly likely to be identified from the sequencing output. In this study, MPS is proposed as a diagnostic and forensic analysis tool for important fungal and chromista plant pathogens. The pathogens used in this study are *Puccinia gramins* f. sp. *tritici*, *Phakopsora pachyrhizi* (Fungi)*, Phytophthora ramorum,* and *Pythium ultimum* (Chromista). The first objective was to simulate 454 sequencing runs to standardize optimal e-probe length and statistical analyses for viability of the method. The second objective was to test the diagnostic method using databases created from real infected plant samples. The diagnostic method was successful in the detection of the four pathogens in Mock Sample Sequencing Databases (MSSDs) and in real sequencing data output even at low pathogen abundances.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Movement of plant pathogens from areas where they are endemic to areas where they have never been before represents enormous risks (10). Various mechanisms are involved in the global dissemination of plant diseases. Currently, one of the most important and risky is international trading of plant propagation materials. However, the implementation of regulations of imports and exports is a difficult task due to various factors, such as the limited number of trained regulatory personnel in ports of entry, large volume of the agricultural products that are shipped worldwide, and the uncontrolled movement of plant materials by tourists and immigrants. These factors, and other logistical problems, limit the number of plants that can be tested. Screening a small percentage of the total imports may not be enough to detect the presence of certain plant pathogens. Eukaryotic plant pathogens (fungi and chromista, or stramenopiles) are the most ubiquitous and and damaging plant diseases. In addition, these pathogens can be difficult to diagnose cleanly, as they share homology with numerous endophytic organisms as well as their eukaryotic hosts.

The United States is an important producer of plants that are exported to different countries. The raw monthly income depending on agricultural trade of the U.S. is

$10,122,718,358 as of August 2011 (2). Soybeans, corn and wheat are the 3 most exported commodities in the U.S. (3), and the impact of introducing plant diseases potentially affecting these crops could be devastating for the U.S. economy and even national security.

The U.S. exported 33,438,497 metric tons of wheat between 2010 and 2011. Similar exports can be expected for 2011-2012 unless production is severely affected by the spread of wheat pathogens, broad infection and disease development (4), and unfavorable weather. Wheat is affected by multiple diseases, among which, rust diseases are particularly devastating. Wheat stem rust is an important disease of wheat, caused by the pathogen *Puccinia graminis* f. sp. *tritici* (12) (Fungi, Basidiomycota). *P. graminis* is an obligate biotroph, non-culturable, therefore its study is limited to laboratories where the pathogen can be cultured on live tissue and manipulated. Etiology analyses of the pathogen have shown that *P. graminis* f. sp. *tritici* affects principally wheat stems. Severe infections interrupt nutrient flow to the developing heads, which results in shriveled grains. Affected stems are prone to lodging, leading to yield losses (12).

Soybean is as another important food and oil crop, with 39,993,124 metric tons exported in 2010-2011 (4). An important soybean disease is soybean rust, a disease caused by the fungal pathogen *Phakopsora pachyrhizi* (Fungi, Basidiomycota). This pathogen has been detected in Asia, Africa (15), South America, and recently in North America (5), causing devastating losses in most soybean producing countries. Soybean rust is the focus of a significant surveillance and monitoring program in the United States.

Another pathogen of great importance, because of its economic and environmental impact, is *Phytophthora ramorum* (Chromista, Oomycota). Commonly known as sudden

oak death (SOD), this pathogen affects Rhododendron, Azaleas, Oaks and Tanoaks, among many other crops (17). This pathogen has been detected in Europe and in the U.S. (Oregon, Washington, and California). Recent analysis have found that 2 mating types (A1 & A2) are present both in Europe and the United States, potentially allowing sexual reproduction, which would lead to the advent of new and more aggressive strains of *P. ramorum*. However, sexual reproduction between these lineages has not been demonstrated yet (8).

Another oomycete of interest is *Pythium ultimum,* a soilborne plant pathogen that causes root-rots, damping-off of seedlings, and post-harvest rots. Although this pathogen is not a regulated organism, it is ubiquitous in distribution, has a broad host range, and can produce severe losses in susceptible crops, particularly when valuable seed is lost. *P. ultimum* has been used as an oomycete model system in numerous studies and is the only *Pythium* species with a sequenced genome available as of July, 2013 (13).

Because of the threat to agriculture and food supply that plant diseases represent, prevention and limitation of outbreaks are highly desirable. In addition, because plants lack immune systems and because of the manner of agricultural production, preventative or post-infection control measures are rare and expensive. As a result, the early detection of plant pathogen outbreaks is the preferred approach for addressing emerging diseases. Various methods are used to detect plant pathogens and avoid their introduction into and dispersal within U.S. territories, starting with screening of plant materials that enter into the U.S. In spite of tight regulations and routine screenings at ports of entry, occasionally pathogens escape surveillance and quarantines (16). Plant materials infected with regulated

organisms are destroyed, causing losses to American businesses. Therefore, a critical step in decision making is the early, sensitive and accurate detection of plant pathogens.

Several methods have been developed for detection of regulated plant pathogens. Immunological assays and PCR-based methods have been developed to high profile plant pathogens, such as *P. ramorum* (11). Immunological assays are less accurate than PCR based methods, but are cheaper for high throughput screening, and thus are usually used for preliminary screening of samples in the field and in diagnostic clinics. However, immunological assays require the development of antibodies to surface or excreted proteins that are specific to the pathogen, and these can be difficult to identify and produce. Currently, real-time PCR assays allow specific and accurate detection of *P. ramorum* at levels as low as 12 fg of pathogen DNA (7). Diagnoses of *P. pachyrhizi, P. ultimum* and *P. graminis* are also possible with real-time PCR, yielding high resolution and reproducible results (1; 9).

Unfortunately, PCR based methods have several limitations. Most of the currently available PCR assays target one or few pathogens at a time (simplex and multiplex real-time PCR) by amplifying sequences of a limited number of genes. PCR based detection can fail if DNA sequences targeted by the primers change due to random mutation or under suboptimal reaction conditions. Hence, when a stringent detection of regulated pathogens is necessary, a single locus may not be enough. Furthermore, molecular diagnostics can produce unreliable results with certain pathogens if the methods have not been properly validated by testing on closely related species, if the pathogen titer is below the detection threshold of the method, or when plant or DNA samples are of poor quality. In spite of the margin of error of these methods, the large flow of plants and plant products coming from

different parts of the world does not permit delays in decisions regarding product entrance or quarantine.

In order to improve the odds of detection of plant pathogens, new methods must be developed that allow rapid and accurate screening of plant materials for multiple pathogens simultaneously. A promising new approach to plant pathogen detection uses analysis of metagenomic information generated by MPS, also called next generation sequencing. Next Generation Sequencing (NGS) allows the production of large amounts of data from a single sequencing run. Used for detection, this methodology can target several loci and screen data sets for multiple pathogens at a time. NGS has resulted in metagenomics, a new approach to the study of ecology and ecosystems. Metagenomics is simply the sequencing of the genetic content of an entire sample, including all the organisms contained therein. For a plant, the metagenome would consist of sequences from the host plant plus any and all microbes associated with the plant. Metagenomics approaches can be used for the detection of pathogens (Stobbe et al *in press*). The disadvantage to this type of approach is the cumbersome level of data and analysis. Typical metagenomic approaches assemble contigs, discard singleton sequences, and blast the entire assembly results against Genbank. This project aims to develop an improved NGS based protocol for simultaneous detection of multiple fungal and oomycete pathogens from infected plant metagenomes.

The pathogen genome was analyzed to design unique queries (e-probes) that will permit the identification of the pathogen when mixed in a sample with different organisms in the output data from a 454 sequencing using a BLASTn-search. Statistical analysis was carried out on the results to assess the accuracy and sensitivity of detection. Similar studies have been developed with human distal gut microbiome, showing promising detection

results (6). The advantages of this approach are a significant reduction in the amount of data handling, such that the entire procedure can be handled in seconds on a typical laptop.

The advent of NGS has permitted the enhancement of methods for the study of plant diseases. Additionally, with the prices of sequencing progressively diminishing, the analysis of human and plant metagenomes is becoming more affordable (14). Analyzed samples can harbor unknown plant pathogens, which will also be present in the sequencing output data. Further analysis of these sequences could reveal undiscovered organisms of potential economic relevance. Also, NGS based methods would permit the analysis of several samples within few hours, therefore, facilitating sample screening at ports of entry, allowing to test samples for various pathogens at the same time, saving time and resources. NGS based methods would also have unlimited multiplexing capacity, and could theoretically be applied to any class of pathogen, from viruses to bacteria to eukaryotes. While the cost of a single NGS analysis is currently too high for regular detection of individual pathogens, it is already economical for instances where a single sample needs to be tested for a wide range of pathogens, such as the certification of imported breeding stock at quarantine facilities.

LITERATURE CITED

1. Barnes, C. W., and L. J. Szabo. 2007. Detection and Identification of Four Common Rust Pathogens of Cereals and Grasses Using Real-Time Polymerase Chain Reaction. Phytopathology 97(6):717-727 doi:10.1094/phyto-97-6-0717.
2. ERS. Total value of U.S. agricultural trade and trade balance, monthly. E. R. Service, ed., United States.
3. ———. 2012. Top 25 export commodities, with level of processing, by calendar year, current dollars. Pages 1 C. B. U.S. Department of Commerce, ed., United States.
4. Foreing, A., Service. 2011. Exports by marketing year - Wheat. Export Sales Query Online, United States.

5. Frederick, R. D., C. L. Snyder, G. L. Peterson, and M. R. Bonde. 2002. Polymerase Chain Reaction Assays for the Detection and Discrimination of the Soybean Rust Pathogens Phakopsora pachyrhizi and P. meibomiae. Phytopathology 92(2):217-227 doi:10.1094/phyto.2002.92.2.217.
6. Gill, S. R., M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. 2006. Metagenomic Analysis of the Human Distal Gut Microbiome. Science 312(5778):1355-1359 doi:10.1126/science.1124234.
7. Hayden, K. J., D. Rizzo, J. Tse, and M. Garbelotto. 2004. Detection and Quantification of Phytophthora ramorum from California Forests Using a Real-Time Polymerase Chain Reaction Assay. Phytopathology 94(10):1075-1083 doi:10.1094/phyto.2004.94.10.1075.
8. Ivors, K. L., K. J. Hayden, P. J. M. Bonants, D. M. Rizzo, and M. Garbelotto. 2004. AFLP and phylogenetic analyses of North American and European populations of Phytophthora ramorum. Mycological research 108:378-392 doi:10.1017/s0953756204009827.
9. Kageyama, K., A. Ohyama, and M. Hyakumachi. 1997. Detection of Pythium ultimum using polymerase chain reaction with species-specific primers. Plant Disease 81(10):1155-1160 doi:10.1094/pdis.1997.81.10.1155.
10. Klinkowski, M. 1970. Catastrophic Plant Diseases. Annual Review of Phytopathology 8(1):37-60 doi:doi:10.1146/annurev.py.08.090170.000345.
11. Kox, L. F. F., I. R. van Brouwershaven, B. T. L. H. van de Vossenberg, H. E. van den Beld, P. J. M. Bonants, and J. de Gruyter. 2007. Diagnostic values and utility of immunological, morphological, and molecular methods for in planta detection of Phytophthora ramorum. Phytopathology 97(9):1119-1129 doi:10.1094/phyto-97-9-119.
12. Leonard, K. J., and L. J. Szabo. 2005. Stem rust of small grains and grasses caused by Puccinia graminis. Molecular Plant Pathology 6(2):99-111 doi:10.1111/j.1364-3703.2005.00273.x.
13. Levesque, C. A., H. Brouwer, L. Cano, J. Hamilton, C. Holt, E. Huitema, S. Raffaele, G. Robideau, M. Thines, J. Win, M. Zerillo, G. Beakes, J. Boore, D. Busam, B. Dumas, S. Ferriera, S. Fuerstenberg, C. Gachon, E. Gaulin, F. Govers, L. Grenville-Briggs, N. Horner, J. Hostetler, R. Jiang, J. Johnson, T. Krajaejun, H. Lin, H. Meijer, B. Moore, and P. Morris. 2010. Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. Genome Biology 11(7):R73.
14. Service, R. F. 2006. The Race for the $1000 Genome. Science 311(5767):1544-1546 doi:10.1126/science.311.5767.1544.
15. Twizeyimana, M., P. S. Ojiambo, K. Sonder, T. Ikotun, G. L. Hartman, and R. Bandyopadhyay. 2009. Pathogenic Variation of Phakopsora pachyrhizi Infecting Soybean in Nigeria. Phytopathology 99(4):353-361 doi:10.1094/phyto-99-4-0353.
16. VanDersal, #160, and J. M. 2007. Managing plant diseases offshore. Vol. 36. CSIRO Publishing, Collingwood, AUSTRALIE.
17. Werres, S., R. Marwitz, W.A. Man in 't Veld, A.W. De Cock, P.J.M. Bonants, M. De Weerdt, K. Themann, E. Ilieva, and R.P. Baayen. . 2001. Phytophthora ramorum

sp. nov: a new pathogen on Rhododendron and Viburnum. . Mycological Research 105(10):1155-1165.

Chapter 2

LITERATURE REVIEW

Overview

This project aims to detect plant pathogens in a cost-effective manner using Next Generation Sequencing and Bioinformatics tools. The techniques that were developed during this research used 4 important eukaryotic (fungal and chromista) plant pathogens as model systems, opening the doors to similar investigations in other organisms. Fungal and chromista plant pathogens are important threats to agriculture and food supply worldwide due to their potentially devastating effects, and their ease for dissemination through infected plant materials (i.e. cuttings, seed-borne inoculums, etc.), and inocula dissemination (i.e. air-, water-, and soilborne, insect vectored), and infection through natural openings (i.e. stomata, lenticels, hydatodes) and wounds caused by insects and other animals, or by human manipulation (Agrios 2005). Fungal plant pathogens, including fungi and oomycetes, cause more plant diseases than any other group, including about 8,000 pathogenic species (Ellis, Boehm et al. 2008). Fungal plant pathogens may cause diseases necrotic diseases such as blights, die backs, vascular wilts,

soft rots, as well as mildews, and rusts, among others (2), targeting underground as well as aerial plant parts. The pathogens that will be studied during this research are the oomycetes *Pythium ultimum* and *Phytophthora ramorum* (Chromista), and the rust fungi *Phakopsora pachyrhizi* and *Puccinia graminis* (Fungi).

Fungal and Chromista Plant pathogens review

Puccinia graminis f. sp. tritici

Relevance, occurrence and distribution

Wheat stem rust is a disease affecting wheat, rye, barley and oat. The causal agent is *Puccinia graminis* f. sp. *tritici* Erikss. & Henning, an obligate biotroph. The first epidemics of Stem rust happened in United States and Canada in 1916, and a few years later in Australia (69).

Stem rust is a warm weather disease and the pathogen does not adapt well to cooler environmental conditions. Australia stem rust epidemics occurred intermittently and mainly in warmer areas through the mid-20th century (60). In India, stem rust affects wheat crops in areas with warm growing season (38). In severe infections, the nutrient flow is interrupted to developing heads by haustoria mother cells growing towards the sap flow (**Error! Not a valid bookmark self-reference.**). As a result of nutrient flow interruption, grain become shriveled (2).

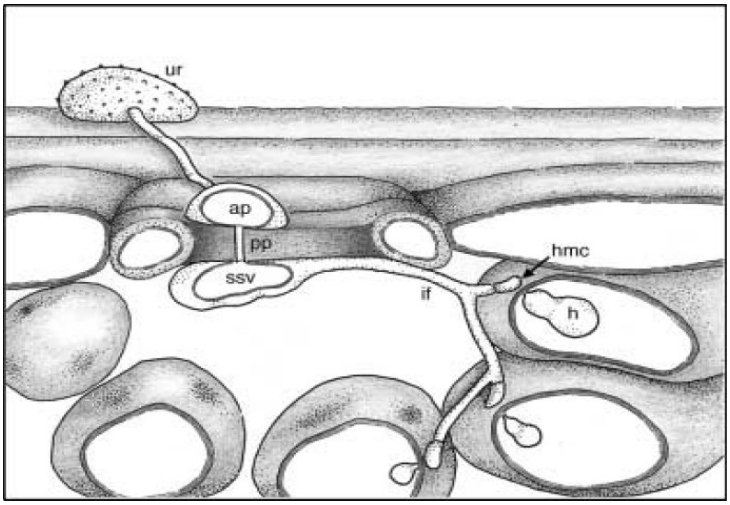

Figure 2-1:Infection structures formed during the early stages of uredinial development by P. graminis: ur = urediniospore, ap = appressorium, pp = penetration peg, which passes between the guard cells and through the stoma, s = stoma, ssv = substomatal vesicle, ih = infection hypha, hcm = haustorial mother cell, h = haustorium(41).

Biology and Life Cycle

Puccinia graminis is heteroecious rust, with a life cycle that includes 5 spore stages: teliospores, aeciospores, basidiospores, urediniospores and spermatia (Figure 2-2). Almost all research on infection processes of *P. graminis* is focused on the uredinial stage because it has the highest economic impact (2; 41). Both, a sexual and an asexual stage are observed during the rust life cycle. The sexual stage happens on the aecial host (*Berberis* spp.) while the asexual stage occurs on the gramineous host. In temperate environments, *P. graminis* produces thick-walled, two-celled teliospores that serve as resting spores. Each teliospore cell contains two haploid nuclei when first formed, but karyogamy occurs early in teliospore maturation. Teliospore stalks remain intact on the wheat stem and the spores are

not detached from the telial pustule. Instead, they remain dormant in the infected stem until they germinate in synchrony with the new leaf growth in the alternate host *Berberis* spp. (41). Meiosis in the teliospore happens before dormancy. During the spring, one or both cells of the teliospore produce promycelium or basidium (Figure 2-2). When meiosis is complete, the resulting four haploid nuclei are separated from each other in the promycelium by three transverse septa. Mitosis happens at the tip of each basidiospore, and it results in two identical haploid nuclei per mature basidiospore.

Mature basidiospores are carried by the wind to infect alternative hosts (*Barberis* spp.), of which the most studied is the common barberry (*B. vulgaris*). The alternative host infection results in the production of flask shaped pycnia, most of the time on the upper surface of the barberry leaf. Pycnial nectar is produced from thin-walled pycniospores. The nectar is attractive to insects that, along with rain splashing, serve to disseminate pycniospores among pycnia. Pycniospores, are the male gametes and each consist mainly of a single haploid nucleus. Flexuous hyphae, which extend out of the top of the flask-shaped pycnia, serve as female gametes. Two mating types, commonly designated + and – , have been identified (62).

Studying this organism requires the infection of plant hosts to produce biomass for downstream applications (62). Nonetheless, in 1993 a research reported a method to culture Urediniospores in liquid media using heat shock and pH regulation (32).

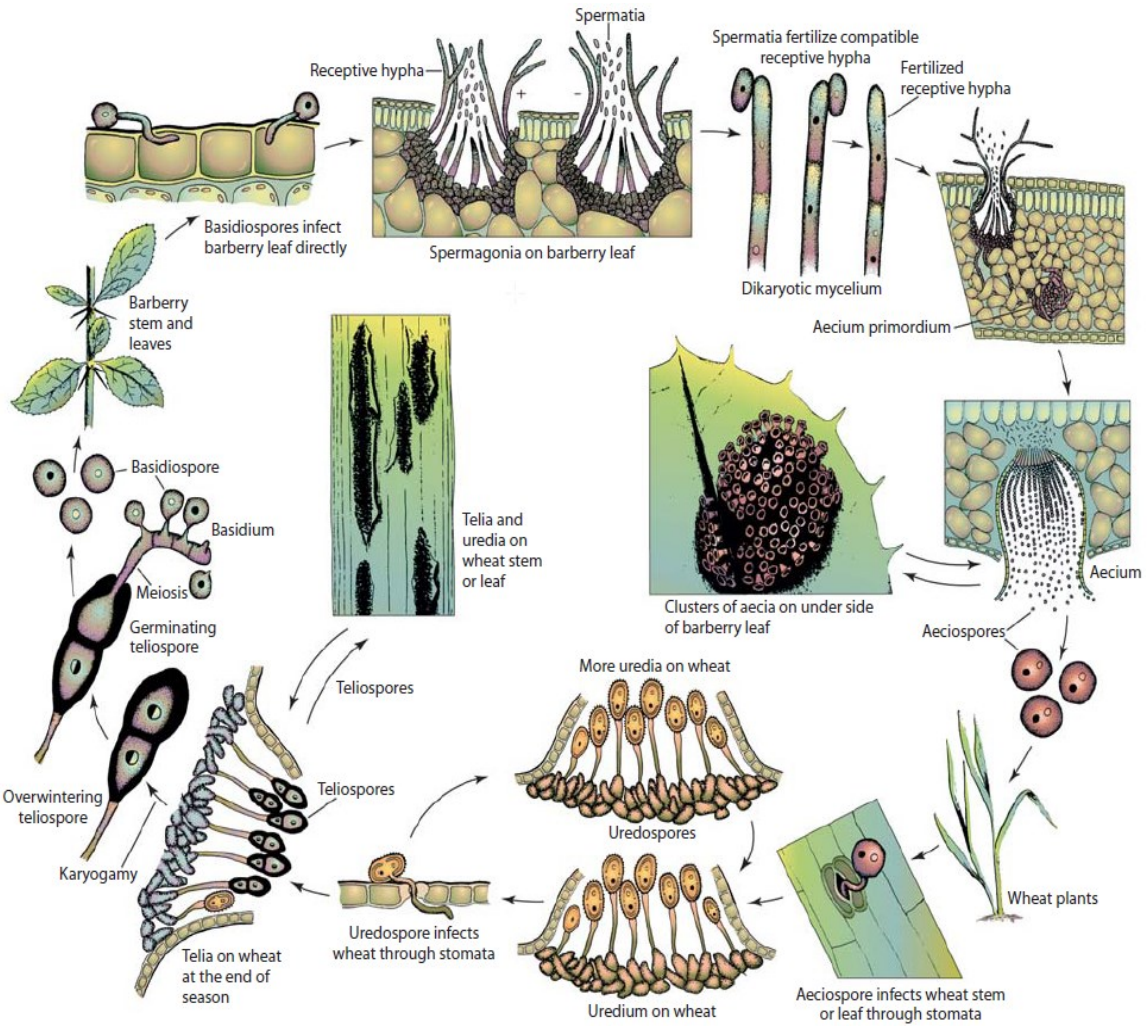

Figure 2-2: Disease cycle of stem rust of wheat caused by Puccinia graminis f. sp tritici (2).

Host Range

The host range of this pathogen is broad, including 365 species of cereals and grasses in 54 genera (41), which can be attributed to the ability of their 5 spore stages to infect particular host plants, depending of the different genes that are active during these stages. During the asexual stage, a gramineous host (Wheat, *Triticum aestivum*) is affected and during the sexual stage the alternate dicot host is affected (Barberry, *Berberis* spp.).

Genome sequence

The whole genome sequence for *Puccinia graminis f. sp. tritici* has been sequenced by the Broad Institute of Harvard and MIT. The genome was produced from two plasmid libraries (4kb and 10kb inserts) and a fosmid library (58). The assembled genome has been public since 2007 (19). *Puccinia triticina* genome was produced from 454 libraries (fragment and 3kb inserts) as well as a fosmid library sequenced using Sanger sequencing. The assembled genome has been public since 2009 (34). Both genomes were used in this project. *P. graminis* and *P. triticina* are closely related species that attack the same host (88).

Diagnostics

To successfully manage *P. graminis* in wheat an early and accurate diagnosis of the disease is necessary. Wheat stem rust is mainly detected using visual keys; however rust fungal species are often difficult to identify in early stages of disease development (7). Currently exists a real-time PCR assay that includes a general set of primers for *Puccinia* spp. and specific fluorescent probes for identification of each *Puccinia* spp. (7) from urediniospores. Unfortunatelly, due to the emergence of new and more aggressive strains of *P. graminis*, like the UG99 strain, pathogen identification to species level is not enough to implement effective quarantine protocols to prevent their introduction to the U.S. Hence, currently there is a need for accurate and sensitive methods for fungal strain identification.

Phakopsora pachyrhizi

Relevance, occurrence and distribution

Phakopsora pachyrhizi (Syd. & P. Syd.) is an obligate fungal plant pathogen and the causal agent of the disease Soybean Rust. It is one of the most damaging diseases of soybean (*Glycine max*), causing yield losses of approximately 10 to 50% (2). Taxonomically it is placed in the phylum basidiomycota and belongs to the order Uredinales. The disease has been reported in Japan, Australia, Hawaii, parts of Central Africa, Central and South America, and the Caribbean islands (2). It has been found in the continental United States, in Louisiana, North Carolina, Alabama, Kentucky, Illinois, Texas, and Florida, since 2004 (67; 77). The prevalence of the pathogen in the U.S. territory is attributed to its capability of overwintering on Kudzu (*Pueraria lobata*), which is a plant native to southern Japan and south east China; however it is an invasive species also found in the Southern United States, along the Gulf of Mexico (54). After overwintering on Kudzu in the South, spores can be spread to Northern soybean growing regions (77).

Biology and Life Cycle

P. pachyrhizi is a microcyclic rust fungi that does not have a known aecial host. Urediniospores are the common and principal inoculum for the disease. Air-borne urediniospores can be disseminated over hundreds of miles in few days (47). The environmental conditions determine the moment when the spores initiate the infection cycle. The optimum range of temperatures for urediniospore germination is 15 ºC to 28 ºC. High humidity is also a key factor for germination (46). A single germ tube is produced upon germination, which grows across the soybean leaf until an appressorium is formed. Appressoria are formed over anticlinal walls or over the center of epidermal cells, but rarely

over stomata, in contrast to the pattern of many other rusts. Penetration is trough wounds rather than through natural openings in the leaf tissue. From the appressorium cone, penetration hyphae arise and pass through the cuticle to emerge in the intercellular space where septa are formed to produce primary hyphae. A primary hypha grows between the mesophyll cells, where it forms the haustorium. The haustorium will absorb nutrients from the host. After this first step in the infection cycle, additional hyphae emerge and spread through the apoplast. Necrosis is visible on the leaves as yellow mosaic discolorations after five days of infection. Discolorations have been observed in both, resistant and susceptible cultivars (75). Urediniospores develop in Uredinia that form from hyphae aggregates. Urediniospore production can be observed only after 3 weeks. Additional uredinia can arise on the margins of the initial infection, extending spore production for up to 8 weeks. Therefore, from the initial infection, there is a first generation of spores after 3 weeks, and a second generation that can appear after 8 weeks of infection. However, evidence suggests that sporulation can be extended for 15 weeks after first germination (40). The urediniospores produced in the uredia are transported by wind to other soybean plants. Healthy leaves are infected with windborne urediniospores. Plants classified as resistant will develop dark, reddish-brown lesions with few or no spores, while susceptible plants will produce uredia and high number of spores (30).

Host Range

P. pachyrhizi has a broad host range that comprises at least 31 species in 17 genera and 42 species in 19 genera of leguminous plants (8; 27; 53). There are two important and very similar species of *Phakopsora, P. pachyrhizi* and *P. meibomiae;* both species causing soybean rust. *P. meibomiae* is a minor and less aggressive pathogen on soybean (10).

However, until recently taxonomists had not done distinction between these two species. Both were treated as a single species until morphological differences and genetic analysis confirmed their differences (24; 53). Both species are able to sporulate when artificial inoculation is performed. Individual inoculations on different potential hosts showed that *P. pachyrhizi* has a broader host range (18 species in 12 genera) than *P. meibomiae* (53).

Genome sequence

P. pachyrhizi genome sequencing is underway, and there are traces (reads) of the whole genome sequencing project that are publicly available (57). Transcriptome studies have reported 2.4 million DNA sequences representing portions of potential *P. pachyrhizi* genes, but this information has been difficult to corroborate since there is not a fully assembled genome available yet (76). However, the mitochondrial genomes of *P. pachyrhizi* and *P. meibomiae* have been sequenced and are publicly available (72).

Diagnostics

Because of its high sensitivity, specificity and efficiency, PCR has been continuously used for detecting *P. pachyrhizi* and *P. meibomiae* in soybean plants (81). However, PCR cannot be employed as a viability test. Viability tests are important to verify if the pathogen is still able to infect another susceptible host. Fungal viability tests are often performed when the pathogen has been submitted to detrimental environmental conditions. For example, fungicide testing involves an assessment of the presence of the fungal pathogen and its viability after a treatment. Particularly for *P. pachyrhizi,* spore viability is tested on petri dishes with water agar (WA) by evaluating the spore's ability to germinate. Normally, viability will be evident after 5 days, when signs of hyphal growth can be observed (81).

Immunoassays have been recently developed for *P. pachyrhizi (39)*. Of which only one method allows viability determination, based on indirect immunofluorescence using fluorescein isothiocyanate-labeled secondary antibodies. Although these methods are sensitive and specific they target one pathogen at a time. Thereafter, if a host is infected by multiple pathogens, only that with most obvious symptoms and signs will be tested for and detected.

Pythium ultimum

Relevance, occurrence and distribution

Oomycetes are eukaryotic diploid organisms in the Kingdom Chromista. Most of the known Oomycete species cause plant diseases, including soilborne diseases, such as damping-off or root rots, as well as airborne diseases, like blights, leaf spots and die backs. Like most *Pythium* species, *P. ultimum* is an important soil inhabitant plant pathogen that causes root rots in a broad range of hosts (42). Being a facultative saprophyte and opportunistic pathogen, it may cause damage to crops when environmental conditions are favorable for infection and colonization, particularly when plant health is compromised by environmental stresses.

Biology and Life Cycle

Pythium ultimum is diploid during its vegetative cycle. The vegetative cells reproduce asexually and the sac-like sporangium can be used as a short term resting structure in soil (71). Hypha are coenocytic and multinucleate, often homokaryotic. The sexual cycle occurs mostly by selfing. Although outcrossing is infrequent, heterokaryosis can be achieved through mutations during asexual reproduction (23). There are no detailed

studies of mitosis performed for *Pythium* spp. The sexual stage is usually short and prone to partitioning errors (23). Sexual reproduction occurs when the haploid sexual structures, a spherical oogonium and a club-shaped anteridium, fuse to produce a diploid zygote that matures to form an oospore, which serves as a long term resting structure (Figure 2-3). Sporangia have been found to survive up to 11 months in soils (32). While oospores can survive in soil, under dry conditions, for up to 12 years (3).

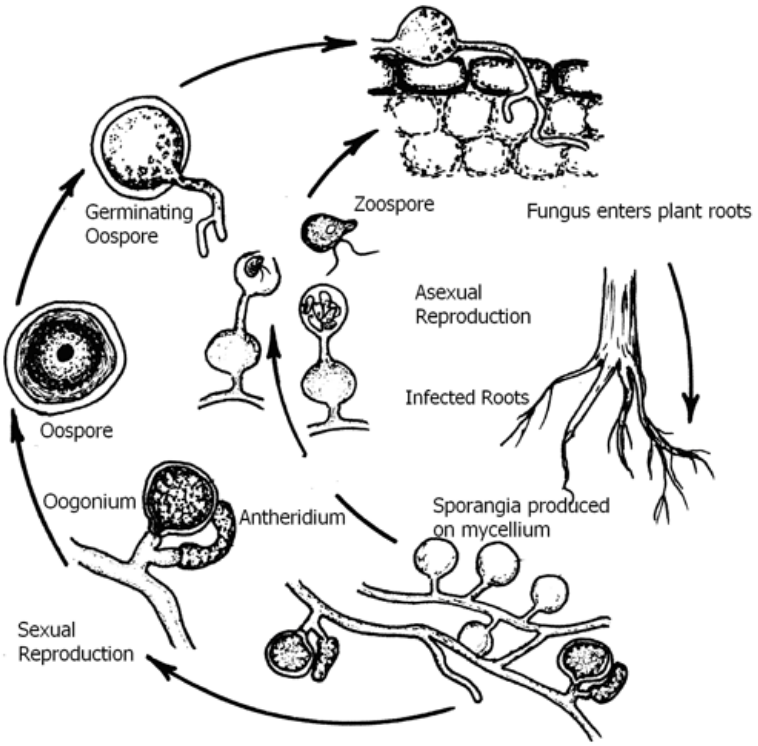

Figure 2-3: Pythium spp. disease cycle (1)

Host Range

Pythium spp. lack host specificity beyond the preference of some species for monocotyledon or dicotyledon hosts. Although, gene-for-gene interactions and a cultivar-race differential responses have been found for some *Phytophthora* and downy mildew species that have narrow host range; *Pythium* spp. have not shown reliable evidence of gene-for-gene interactions or cultivar-race differences. *Pythium* spp. are mostly

necrotrophs and saprophytes. *Pythium ultimum* has a broad host range with 333 reported hosts (http://nt.ars-grin.gov/fungaldatabases/).

Genome sequence

The soilborne plant pathogen *P. ultimum* has been selected for this study because *P. ultimum*'s relevance in agriculture worldwide, it has been used as a model system for genetic studies, and its genome has been sequenced and is publicly available (43). The genome of *P. ultimum* is about 42.8 Mb and approximately 15,300 genes have been annotated using transcripts, protein homology and ESTs databases (43). The genomes of *P. ultimum* and *Phytophthora infestants* have extensive sequence similarity. (42). Transcriptome analysis showed that although abiotic stress and the presence of a host induce the up-regulation of genes, 86% of the genes are expressed constitutively in the absence of these factors. Certain effector genes (i.e. RXLR and Crinkler effectors) similar those of other oomycetes were found in the *P. ultimum* genome; however, proteome analysis didn't confirm their expression.(42).

Diagnostics

Currently, *P. ultimum* infections are routinely diagnosed based on plant symptoms (i.e. yellowing, wilting), and signs (i.e. mycelia and oospores), and species identification is done based on morphology and real-time PCR (16). Another preferred method for *Pythium* species identification is sequencing of the ITS region using ITS1-ITS4 primers (White et al. 1990) followed by BLASTn search on NCBI database. Biochemical methods for detecting *P. ultimum* involve monoclonal antibodies (MAb E5) that react to a cell wall glycoprotein of 46 kDa (5).

Phytophthora ramorum

Relevance, occurrence and distribution

Phytophthora ramorum, the causal agent of Sudden Oak Death and Ramorum blight, is a devastating pathogen in Oregon, California, Washington state and British Columbia, causing economic losses to nurseries and killing oak trees in forests (28; 35; 36). The most visible symptoms in trees are cankers in the trunk and an exudation of a brown or black or red color (17). Symptoms in *Rhododendron* plants include brownish to black discoloration on leaves and the dieback of shoot tips (17).

Biology and Life Cycle

Phytophthora spp. have a disease cycle that usually involves infection of the host tissue, production of spores (sporangia with zoospores), release and movement of infective zoospores and finally reinfection of host tissue. Chlamydospores (mitospores) and oospores (meiospores) can be formed to function as resting structures and reservoir of inoculum for the pathogen. The movement of *Phytophthora* spp. zoospores is through rain splash, watercourses, soil dust, but rarely through wind without rain (61).

Spore movement takes place during the mid- to late rainy season in California. *P. ramorum* cannot be transmitted by sporadic summer rains or soil and litter during the hot dry summer months (17). Rainwater can transport the pathogen 5 to 10 m from the inoculum source, also hikers can transport the pathogen (17). Rates of transmission are affected by each step in the disease cycle. Also, transmission is very likely to vary depending on the host species that *P. ramorum* is attacking (17).

Host Range

P. ramorum has a wide and increasing host range. More than 117 taxa have been identified as hosts. The USDA's fungal database has 98 pathogen-hosts combinations registered for *P. ramorum* of 26 genera in 17 plant families. Because of their economic impact, the most important hosts are Oaks (*Quercus* spp.), Tanoaks (*Notholithocarpus densiflorus*), Rhododendron (*Rhododendron* spp.), Camellias (*Camellia* spp.), Roses (*Rosa gymnocarpa*) and Magnolias (*Magnolia spp.*) (4).

Genome sequence

Phytophthora ramorum draft genome was originally sequenced during 2006 (78). The size of the assembled genome is 54.5 Mb (7.588 contigs); however, the estimated genome size is 65 Mbp. Genome analyses contributed to the discovery of proteins that are associated with plant infection like hydrolases, ABC transporters, protein toxins, and proteinase inhibitors. In addition, a superfamily of 700 proteins with similarity to well-known Oomycete avirulence genes was found.

Diagnostics

P. ramorum symptoms and signs vary depending of the host. Although the symptoms and signs are important to detect the pathogen at a glance, more accurate techniques have been developed. Such techniques include both molecular and protein based diagnostic methods.

For the specific detection of *P. ramorum,* the method has to be sensitive and reliable. As a part of a widespread environmental screening throughout California, a qPCR assay was developed and used for the detection of *P. ramorum.* The method was able to

detect less that 12 fg of pathogen DNA in a given sample (31). Subsequently, other diagnostic methods have been developed and implemented. Specificity and sensitivity tests of end-point PCR, TaqMan real-time PCR and immunoassays have shown that the diagnosis of *P. ramorum* is more reliable when sample collection is conducted during wet and warm weather because these conditions are favorable for the pathogen (80).

In addition, a different real-time qPCR study allowed the quantification of the pathogen in samples. An evaluation of the method showed that this technique had a detection limit of 50 fg, which corresponds to very low concentrations of the pathogen, and it could be used as a rapid screening method for detection of *P. ramorum* in plant tissues (13). Other methods based on detection of mitochondrial DNA have been developed and used successfully, allowing to identify this pathogen at the species level, discriminating it from other pathogens like *Phytophthora pseudosyringae,* a pathogen that causes similar symptoms as *P. ramorum* (74). The later method has a limit of detection of 1 fg of genomic DNA. Strain discrimination is important since variation in aggressiveness and mating type need to be monitored for prevention of epidemics in natural and agricultural environments, particularly in areas where conditions are favorable for the development of the disease (2).

Diagnosis of Plant Pathogens

Timely detection, identification and quantification of pathogens are crucial requirement for the preparation of effective disease management. The conventional methods of isolating pathogens from infected plants and identifying them based on taxonomical criteria are time-consuming and labor-intensive, and therefore become expensive, even though no sophisticated equipment is required. Designing specific primers for plant pathogen detection has been possible mostly due to the study of molecular

genetics that has provided information on the nucleotide sequences. In some instances pathogenicity related genes have been identified (51).

Fungal pathogens and FLOs have been identified during the last two decades using molecular techniques. The most common detection method is nested PCR (85). One of the most promising PCR techniques used for identifying fungi is focused in amplifying the internal transcribed spacer (ITS) region. ITS contains two introns, variable non-coding regions, that are nested within the rDNA repeat between the highly conserved small subunit, 5.8S, and large subunit rRNA genes (25; 26). Several features make ITS a convenient target region for molecular identification of fungi: (i) the size of the product is between 600 and 800 bp and can be easily amplified with "universal primers", (ii) the presence of multiple copies of rDNA in fungal organisms makes the ITS region easy to amplify from dilute and highly degraded samples and (iii) several studies have confirmed that ITS is highly variable among fungal and Oomycete species (14; 15; 25; 26).

The current disadvantage of using ITS region for identification of fungi and Oomycete organisms is that the ITS region is also present in plants, protists and animals (85), and it can be amplified with some universal primers, which could lead to false positive identification when a sample contains DNA from other organisms. The use of pure cultures isolated from diseased plants is strongly suggested when trying to identify plant pathogens using the ITS region.

Equally important are immunological and biochemical methods that require relatively less time, labor and personnel training. Clonal antibodies that target specific pathogen proteins have been widely utilized by plant diagnosticians (6). Protein profiles in

electrophoretic gels, fatty acid and nutritional profiling have shown to be useful for cheap identification of pathogens (51). Serological techniques like immunostrips and ELISA (Enzyme-linked immunosorbent assay), have demonstrated to be practical, rapid, reliable and sensitive for the detection of plant pathogens. A serological technique used previous to ELISA was radioimmunoassay, using radioactivity (87). Unfortunately, serological techniques may not discriminate among closely related species, and cannot distinguish among strains of fungal species.

The final objective of timely plant pathogen detection is to provide opportune information for crop disease management. Disease intensity and potential crop losses are nowadays assessed with specific detection techniques like qPCR and ELISA. Propagative plant material has been successfully screened because of the effectiveness of some detection methods. The selection and application of appropriate diagnosis methods is important to prevent diseases to reach devastating proportions, and to prevent the use of excessive chemicals when they may not be necessary (51).

Genome Sequencing and disease diagnosis

DNA sequencing was first successfully achieved in the bacteriophage ϕX174 using the “plus and minus” method which was a relatively rapid and simple method for sequencing (66). This method involved sequencing separately the "plus" and the "minus" strands and them comparing their sequences. Although the method was considered more rapid and simple than other available techniques it was not completely accurate and further confirmatory data was necessary. Other techniques, like specific chemical degradation of the DNA and ribo-substitution, have been developed, but the most successful is DNA sequencing with chain-terminating inhibitors (66) also called Sanger sequencing, denoting

the first author's name. The principle of Sanger sequencing relies on the inhibitory activity of 2',3'-dideoxythymidine triphosphate (ddTTP) on DNA polymerase I, as well as the other three ribonucleosides variants: dideoxycytosine triphosphate (ddCTP), dideoxyadenine triphosphate (ddATP) and dideoxyguanine triphosphate ddGTP, which have the same inhibitory activity. The chain termination is dependent on these dideoxiribonucleoside triphosphates being incorporated into the growing oligonucleotide chain. The polymerase I is inhibited when a dideoxi-nucleotide is added to compliment the template DNA. As a result, the oligonucleotide extension is terminated and oligonucleotides with different lengths are created. The length of the oligonucleotides is driven by the DNA template sequence. The oligonucleotides are visualized in an acrylamide gel using electrophoresis. Initially the Sanger sequencing technique used to be performed in 4 different pools of ddNTPs (ddATP, ddCTP, ddTTP, ddGTP) run in polyacrylamide gels in adjacent wells. However, now it is performed in one single pool using ddNTPs labeled with different fluorescent dyes. Afterwards, capillary electrophoresis is performed in one single column. At the bottom of the gel a laser excites the fluorescent dyes in the fragments as they pass and detectors collect the emission intensities at four different wavelengths (70). The computer analysis converts the gel image to an inferred base sequence for each template. The analysis consists of four different steps: (i) lane tracking [gel boundaries are identified], (ii) lane profiling [each of the four signals is summed across the lane width to create a profile], (iii) trace processing [signal processing methods are used to smooth the signal estimates, reduce noise and correct for dye effects on fragment mobility] and (iv) base-calling [the processed trace is translated into a sequence of bases] (22; 70). Chromatograms consisting of four curves of

different colors are the result of Sanger sequencing, each curve representing the signal for one of the four bases and drawn left to right in the direction of increasing time to detection. An ideal chromatogram would consist of evenly spaced non-overlapping peaks. The subjectivity of the user while analyzing chromatograms might alter sequences and create a bias. Phred scores (scores assigned to each nucleotide base call in automated sequencer traces) have been utilized to avoid the subjectivity of users analyzing chromatograms (22).

Among the most important contributions of Sanger sequencing to science is the sequence of the human genome (79), as well as targeted sequencing (shotgun sequencing) using molecular cloning as well as amplicon sequencing (48; 65). The discovery of dispersed tandem-repetitive ‘minisatellites’ and 'microsatellites' in genomes, also known as Simple Sequence Repeats (SSRs), which are highly polymorphic in most organisms, including humans (37), has allowed their use for DNA fingerprinting or DNA typing. This approach produces information that can be used for population genetic studies, as well a in other applications, like in forensics and paternity testing.

However, some disadvantages of Sanger sequencing have also slowed down the life sciences pathway. The most important disadvantage has been the time that takes to sequence a whole genome of a single organism. The Sanger sequencer can process only few sequencing reads per run and the reads normally cannot be longer than a PCR product. Even though the method is very sensitive, the efficiency is not in accordance with current needs.

Sanger sequencing dominated the DNA sequencing field since 1977, almost 30 years. Despite many improvements made to Sanger sequencing, the limitations of this

technology triggered the search of new and improved technologies for sequencing large numbers of genomes. Recent efforts have focused mostly on developing new techniques for sequencing whole genomes. Today, Sanger sequencing is frequently called “first generation sequencing” and newer methods are referred to as "next-generation sequencing" (NGS) (49).

Various NGS strategies were developed the first decade of the 2000s. Most of these strategies prepare the template DNA (sizing the fragmented genome for maximum efficiency) as a first step to follow with sequencing and imaging. A third step implies a bioinformatics analysis that comprises alignment and assembly. The major advantage of NGS is the ability to produce an enormous volume of data at relatively low per base cost (49). This advantage permits analyses that were not possible with Sanger sequencing. For example, traditional transcriptome studies required a previous knowledge of the genes to be analyzed and the utilization of microarrays. Now they have been replaced with sequencing methods that permit the sequencing of the total RNA in one single run, allowing the identification and quantification of undiscovered transcripts, alternative splicing and sequence changes in known genes (82; 86). Comparative genomics and evolutionary studies have been possible because now full genomes of related organisms can be sequenced at low costs.

Commercially available platforms for NGS have similar principles, but different approaches and outcomes. The first platform available was Roche 454 pyrosequencing (45); a technique that was able to avoid cloning requirements using a method called emulsion PCR (emPCR) (73). On the other hand, the Illumina/Solexa technology avoids cloning by attaching single-stranded DNA fragments to a solid surface known as a single-

molecule array or flow cell (9). Ultimately, three major new sequencing platforms were released in 2011: Ion Torrent's Personal Genome Machine (PGM), Pacific Biosciences' RS (PacBio) and the Illumina MiSeq (59). Ion torrent detects protons that are released while the nucleotides are incorporated during sequencing by synthesis. The DNA fragments are ligated to 3-micron diameter beads called Ion Sphere Particles, forming the emPCR (64). PacBio uses a technique called single molecule real-time (SMRT) sequencing (20). Here, DNA polymerase molecules are bound to a DNA template, and then they are attached to the bottom of 50 nm-wide wells termed zero-mode waveguides (ZMWs). The sequencing signal is captured from the bottom of the well and is produced by γ-phosphate fluorescently labeled nucleotides that are incorporated into the DNA chain. In all the techniques previously mentioned, while the DNA fragments are sequenced, output data (reads) are generated (one read per DNA fragment). Read length is one of the main features that differentiate among platforms. User preferences need to be in accordance with the sequencing output in order to perform the required downstream analysis. Longer reads are preferred for biological applications like *de novo* genome assembly and 16S metagenomic studies. On the contrary, small reads are preferred for variant discovery by whole-genome re-sequencing or whole exome capture and gene discovery in metagenomics (50).

454 Pyrosequencing

A limitation of Sanger sequencing is the requirement of *in vivo* amplification of DNA fragments. The amplification step is usually achieved with cloning; although currently, samples can be sequenced directly from PCR amplicons. Both, cloning and PCR amplification are labor intensive when a high number of samples need to be sequenced

(50). In addition, the resulting sequence is limited to the amplified or cloned regions adjacent to the primers used for sequencing.

454 Pyrosequencing avoids cloning or single-fragment amplification. Instead, it uses a highly efficient *in vitro* emPCR. During this process, each DNA fragment is ligated to adaptors (forward and reverse) and later attached to streptavidin beads. When streptavidin beads are mixed with the 454 beads DNA fragments are captured into separate emulsion droplets. The droplets act as individual amplification reactors producing approximately 10^7 clonal copies of a unique DNA template per bead (45). All the beads are subsequently transferred into a picotiter plate containing hundreds of thousands of wells where pyrosequencing reactions are carried out in parallel, tremendously increasing the sequencing throughput. Pyrosequencing (52; 63) is a sequencing by synthesis technique that measures the release of inorganic pyrophosphate (PPi) using chemiluminescense. The beads bearing DNA fragments are immobilized into the picotiter wells (one bead per well) and dNTPs pools are added to the picotiter plate one at a time. PPi is released when nucleotides are incorporated to the DNA fragments bound to the beads. The reaction mix contains an enzyme that produces chemiluminescense every time PPi is released. Finally the sequence of the DNA fragment is determined from a "pyrogram" that is produced during the sequencing process. The "pyrogram" contains a series of pictures of the picotiter plate while chemiluminescense was irradiated from each well. The pictures are compiled into one single data file of .sff file extension. SFF files can be used for further analyses of the sequencing data. However, error correction and elimination of uncertain sequences is imperative before performing any further bioinformatics analysis.

Sequencing Economics

Sequencing prices have been continuously dropping since the first human genome sequence was published (79). The cost per raw Megabase of DNA sequence was close to $10,000 in 2001 (Figure 2-4). However, the development of new sequencing techniques has allowed the cost per raw Megabase to drop to as low as $0.1, which indicates a reduction of 99.999% in sequencing costs (119).

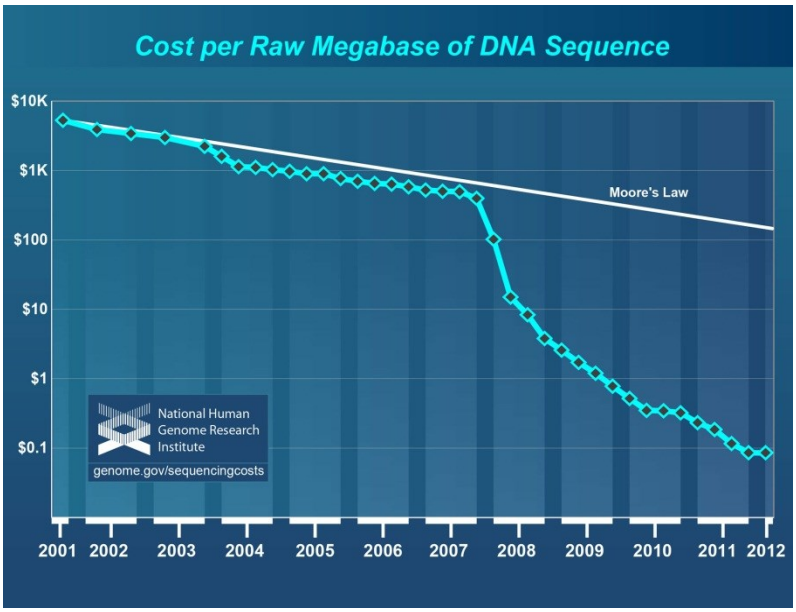

Figure 2-4: Sequencing costs reduction since year 2001 to 2012 (84).

Besides the low cost for determining whole genomes, the advantage of utilizing NGS is that the time to obtain high volumes of data is significantly reduced to a few hours. The sequencing capacity of a Roche 454jr sequencer single run is 25 million bases with a Phred quality score higher than that of Sanger sequencing by capillary electrophoresis (44).

NGS for Fungal and Chromista plant pathogen diagnosis

It was previously stated that it is necessary rapid and accurate diagnosis of plant pathogens to perform a proper and timely control and management of plant diseases. Molecular, biochemical and serological tests are highly accurate and inexpensive, although they have limitations when detecting multiple pathogens simultaneously and accurately.

Currently there is not any diagnostic technique for the detection of fungal and chromista plant pathogens using NGS and bioinformatics. Although there are techniques publicly available to identify organisms in NGS sequencing output (29; 33), they are not designated specifically for the organisms that are used in this study. The aim of this project is to develop a newly diagnostic tool based on NGS data to detect multiple fungal and chromista plant pathogens accurately, sensitively and specifically. Sequencing costs are decreasing as new sequencing techniques appear available, which will soon make this new diagnostic tool more cost effective than currently available protocols (84).

The new protocol was designed to use unique genomic signatures, referred to as e-probes, for accurate and specific detection of pathogens using bioinformatics tools, without requiring genome assembly. The genome size of fungal and chromista plant pathogens allow the design of multiple unique DNA signatures from currently available genome sequences and ESTs data. The detection method takes advantage of newly identified e-probes to detect the presence of pathogen sequences in plant metagenomic databases generated by NGS. The diagnostic tool benefits from an improved sequencing efficiency (454 pyrosequencing). Roche 454 pyrosequencing produces read lengths of approximately 400 bp which perfectly fits the minimal assembly needs of this project (50).

Genome assembly is a difficult task when the data has been produced by NGS. There are several sequencing assemblers that use either *de novo* or referenced genome assembly (11; 12; 18; 55; 56; 83). Computer memory or thread number has limited some assemblers when using organisms with genomes larger than bacterial genomes (68). The use of e-probes to evaluate the presence of a certain organism in NGS data minimizes the resources needed in NGS computational analyses because only certain portions of the whole genome are used and genome assembly is not necessary.

LITERATURE CITED

1. Agriculture, B. C. M. o. 2013. Pythium Diseases of Greenhouse Vegetable Crops.
2. Agrios, G. 2005. Plant Pathology. Edited by Fifth. Elsevier.
3. Allen, T. W., A. Martinez, and L. L. Burpee. Pythium blight of turfgrass. APSnet.
4. APHIS List of Regulated Hosts and Plants Proven or Associated with Phytophthora ramorum. 2012. USDA APHIS.
5. Avila-Rodriguez, F. J. 1994. Serological characterization of Pythium ultimum. 9507805, The University of Nebraska - Lincoln, United States -- Nebraska.
6. Avila, F., Schoedel, Barbara. 2009. ELISA and ImmunoStrip® for detection of Phytophthora ramorum, P. kernoviae, and other Phytophthora species. Pages 95-96. U.S. Department of Agriculture, Forest Service, Pacific Southwest Research Station. , Albany, CA.
7. Barnes, C. W., and L. J. Szabo. 2007. Detection and Identification of Four Common Rust Pathogens of Cereals and Grasses Using Real-Time Polymerase Chain Reaction. Phytopathology 97(6):717-727 doi:10.1094/phyto-97-6-0717.
8. Barnes, C. W., L. J. Szabo, and V. C. Bowersox. 2009. Identifying and Quantifying Phakopsora pachyrhizi Spores in Rain. Phytopathology 99(4):328-338 doi:10.1094/phyto-99-4-0328.
9. Bennett, S. 2004. Solexa ltd. Pharmacogenomics 5(4):433-438.
10. Bonde, M. R., S. E. Nester, C. N. Austin, C. L. Stone, R. D. Frederick, G. L. Hartman, and M. R. Miles. 2006. Evaluation of virulence of Phakopsora pachyrhizi and P. meibomiae isolates. Plant Disease 90(6):708-716.
11. Butler, J., I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe. 2008. ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome research 18(5):810-820.
12. Chaisson, M. J., and P. A. Pevzner. 2008. Short read fragment assembly of bacterial genomes. Genome research 18(2):324-330.
13. Chandelier, A., K. Ivors, M. Garbelotto, J. Zini, F. Laurent, and M. Cavelier. 2006. Validation of a real-time PCR method for the detection of Phytophthora ramorum1. EPPO Bulletin 36(2):409-414 doi:10.1111/j.1365-2338.2006.01020.x.

14. Chen, W., J. W. Hoy, and R. W. Schneider. 1992. Species-specific polymorphisms in transcribed ribosomal DNA of five< i> Pythium</i> species. Experimental Mycology 16(1):22-34.
15. Cooke, D. E. L., A. Drenth, J. M. Duncan, G. Wagels, and C. M. Brasier. 2000. A molecular phylogeny of Phytophthora and related oomycetes. Fungal Genetics and Biology 30(1):17-32 doi:10.1006/fgbi.2000.1202.
16. Cullen, D. W., I. K. Toth, N. Boonham, K. Walsh, I. Barker, and A. K. Lees. 2007. Development and Validation of Conventional and Quantitative Polymerase Chain Reaction Assays for the Detection of Storage Rot Potato Pathogens, Phytophthora erythroseptica, Pythium ultimum and Phoma foveata. Journal of Phytopathology 155(5):309-315 doi:10.1111/j.1439-0434.2007.01233.x.
17. Davidson, J. M., A. C. Wickland, H. A. Patterson, K. R. Falk, and D. M. Rizzo. 2005. Transmission of Phytophthora ramorum in Mixed-Evergreen Forest in California. Phytopathology 95(5):587-596 doi:10.1094/phyto-95-0587.
18. Dohm, J. C., C. Lottaz, T. Borodina, and H. Himmelbauer. 2007. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome research 17(11):1697-1706.
19. Duplessis, S., C. A. Cuomo, Y.-C. Lin, A. Aerts, E. Tisserant, C. Veneault-Fourrey, D. L. Joly, S. Hacquard, J. Amselem, and B. L. Cantarel. 2011. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proceedings of the National Academy of Sciences 108(22):9166-9171.
20. Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, and B. Bettman. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133-138.
21. Ellis, S. D., M. J. Boehm, and T. k. Mitchel. 2008. Fungal and Fungal-like Diseases of Plants. Seventh, ed. Ohio State University.
22. Ewing, B., L. D. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. Genome research 8(3):175-185.
23. Francis, D. M., and D. A. S. Clair. 1997. Population Genetics of Pythium ultimum. Phytopathology 87(4):454-461 doi:10.1094/phyto.1997.87.4.454.
24. Frederick, R. D., C. L. Snyder, G. L. Peterson, and M. R. Bonde. 2002. Polymerase chain reaction assays for the detection and discrimination of the soybean rust pathogens Phakopsora pachyrhizi and P-meibomiae. Phytopathology 92(2):217-227 doi:10.1094/phyto.2002.92.2.217.
25. Gardes, M., and T. Bruns. 1993. ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. Molecular ecology 2(2):113-118.
26. Gardes, M., T. J. White, J. A. Fortin, T. D. Bruns, and J. W. Taylor. 1991. Identification of indigenous and introduced symbiotic fungi in ectomycorrhizae by amplification of nuclear and mitochondrial ribosomal DNA. Canadian Journal of Botany 69(1):180-190.
27. Goellner, K., M. Loehrer, C. Langenbach, U. W. E. Conrath, E. Koch, and U. Schaffrath. 2010. Phakopsora pachyrhizi, the causal agent of Asian soybean rust. Molecular Plant Pathology 11(2):169-177 doi:10.1111/j.1364-3703.2009.00589.x.

28. Grunwald, N. J., E. M. Goss, M. M. Larsen, C. M. Press, V. T. McDonald, C. L. Blomquist, and S. L. Thomas. 2008. First report of the European lineage of Phytophthora ramorum on viburnum and Osmanthus spp. in a California Nursery. Plant Disease 92(2):314-314 doi:10.1094/pdis-92-2-0314b.
29. Haft, D. H., and A. Tovchigrechko. 2012. High-speed microbial community profiling. Nature Methods 9(8):793-794.
30. Hartwig, E., and K. Bromfield. 1983. Relationships among three genes conferring specific resistance to rust in soybeans. Crop Science 23(2):237-239.
31. Hayden, K. J., D. Rizzo, J. Tse, and M. Garbelotto. 2004. Detection and Quantification of Phytophthora ramorum from California Forests Using a Real-Time Polymerase Chain Reaction Assay. Phytopathology 94(10):1075-1083 doi:10.1094/phyto.2004.94.10.1075.
32. Hendrix, F. F., and W. A. Campbell. 1973. Pythiums as Plant Pathogens. Annual Review of Phytopathology 11(1):77-98 doi:doi:10.1146/annurev.py.11.090173.000453.
33. Huson, D. H., S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome research 21(9):1552-1560.
34. Institute, B. 2009. Puccinia triticina 1-1 BBBD Race 1. B. Institute, ed. NCBI.
35. Ivors, K., M. Garbelotto, I. D. E. Vries, C. Ruyter-Spira, B. T. Hekkert, N. Rosenzweig, and P. Bonants. 2006. Microsatellite markers identify three lineages of Phytophthora ramorum in US nurseries, yet single lineages in US forest and European nursery populations. Molecular Ecology 15(6):1493-1505 doi:10.1111/j.1365-294X.2006.02864.x.
36. Ivors, K. L., K. J. Hayden, P. J. M. Bonants, D. M. Rizzo, and M. Garbelotto. 2004. AFLP and phylogenetic analyses of North American and European populations of Phytophthora ramorum. Mycological research 108:378-392 doi:10.1017/s0953756204009827.
37. Jeffreys, A. J., V. Wilson, and S. L. Thein. 1985. Hypervariable'minisatellite' regions in human DNA. Nature 314(6006):67-73.
38. Joshi, L. M., and L. T. Palmer. 1973. Epidemiology Of Stem, Leaf And Stripe Rusts Of Wheat In Northern India. Plant Disease Reporter 57(1):8-12.
39. Jurick, W., C. L. Harmon, J. Marois, D. Wright, and K. Lamour. 2007. A comparative analysis of diagnostic protocols for detection of the Asian soybean rust pathogen, Phakopsora pachyrhizi. Plant Health Progress.
40. Koch, E., and H. H. Hoppe. 1988. Development Of Infection Structures By The Direct-Penetrating Soybean Rust Fungus (Phakopsora-Pachyrhizi Syd) On Artificial Membranes. Journal of Phytopathology-Phytopathologische Zeitschrift 122(3):232-244 doi:10.1111/j.1439-0434.1988.tb01012.x.
41. Leonard, K. J., and L. J. Szabo. 2005. Stem rust of small grains and grasses caused by Puccinia graminis. Molecular Plant Pathology 6(2):99-111 doi:10.1111/j.1364-3703.2005.00273.x.
42. Levesque, C. A., H. Brouwer, L. Cano, J. Hamilton, C. Holt, E. Huitema, S. Raffaele, G. Robideau, M. Thines, J. Win, M. Zerillo, G. Beakes, J. Boore, D. Busam, B. Dumas, S. Ferriera, S. Fuerstenberg, C. Gachon, E. Gaulin, F. Govers, L. Grenville-Briggs, N. Horner, J. Hostetler, R. Jiang, J. Johnson, T. Krajaejun, H.

Lin, H. Meijer, B. Moore, and P. Morris. 2010. Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. Genome Biology 11(7):R73.

43. Levesque, C. A., H. Brouwer, L. Cano, J. P. Hamilton, C. Holt, E. Huitema, S. Raffaele, G. P. Robideau, M. Thines, J. Win, M. M. Zerillo, G. W. Beakes, J. L. Boore, D. Busam, B. Dumas, S. Ferriera, S. I. Fuerstenberg, C. M. M. Gachon, E. Gaulin, F. Govers, L. Grenville-Briggs, N. Horner, J. Hostetler, R. H. Y. Jiang, J. Johnson, T. Krajaejun, H. Lin, H. J. G. Meijer, B. Moore, P. Morris, V. Phuntmart, D. Puiu, J. Shetty, J. E. Stajich, S. Tripathy, S. Wawra, P. van West, B. R. Whitty, P. M. Coutinho, B. Henrissat, F. Martin, P. D. Thomas, B. M. Tyler, R. P. De Vries, S. Kamoun, M. Yandell, N. Tisserat, and C. R. Buell. 2010. Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. Genome Biology 11(7) doi:10.1186/gb-2010-11-7-r73.

44. Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376-380 doi:http://www.nature.com/nature/journal/v437/n7057/suppinfo/nature03959_S1.html.

45. Margulies, M., M. Egholm, W. Altman, S. Attiya, J. Bader, L. Bemben, J. Berka, M. Braverman, Y. Chen, Z. Chen, S. Dewell, L. Du, J. Fierro, X. Gomes, B. Godwin, W. He, S. Helgesen, C. Ho, G. Irzyk, S. Jando, M. Alenquer, T. Jarvie, K. Jirage, J. Kim, J. Knight, J. Lanza, J. Leamon, S. Lefkowitz, M. Lei, J. Li, K. Lohman, H. Lu, V. Makhijani, K. McDade, M. McKenna, E. Myers, E. Nickerson, J. Nobile, R. Plant, B. Puc, M. Ronan, G. Roth, G. Sarkis, J. Simons, J. Simpson, M. Srinivasan, K. Tartaro, A. Tomasz, K. Vogt, G. Volkmer, S. Wang, Y. Wang, M. Weiner, P. Yu, R. Begley, and J. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376 - 380.

46. Melching, J. S. 1983. Effect Of Dew Temperature, Duration, And Frequency On Lesion Development On Wayne Soybean Inoculated With Phakopsora-Pachyrhizi, The Cause Of Soybean Rust. Phytopathology 73(6):967-967.

47. Melching, J. S., K. R. Bromfield, and C. H. Kingsolver. 1979. Infection, Colonization, And Uredospore Production On Wayne Soybean By 4 Cultures Of Phakopsora-Pachyrhizi, The Cause Of Soybean Rust. Phytopathology 69(12):1262-1265 doi:10.1094/Phyto-69-1262.

48. Messing, J., R. Crea, and P. H. Seeburg. 1981. A system for shotgun DNA sequencing. Nucleic acids research 9(2):309-321.

49. Metzker, M. L. 2010. Sequencing technologies [mdash] the next generation. Nat Rev Genet 11(1):31-46.
50. Morozova, O., and M. A. Marra. 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92(5):255-264.
51. Narayanasamy, P. 2001. Plant pathogen detection and disease diagnosis. Vol. 83. CRC.
52. Nyrén, P., B. Pettersson, and M. Uhlen. 1993. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. Analytical biochemistry 208(1):171-175.
53. Ono, Y., P. Buriticá, and J. F. Hennen. 1992. Delimitation of Phakopsora, Physopella and Cerotelium and their species on Leguminosae. Mycological research 96(10):825-850 doi:10.1016/s0953-7562(09)81029-0.
54. Park, S., Z. Y. Chen, A. K. Chanda, R. W. Schneider, and C. A. Hollier. 2008. Viability of Phakopsora pachyrhizi urediniospores under simulated southern Louisiana winter temperature conditions. Plant Disease 92(10):1456-1462 doi:10.1094/pdis-92-10-1456.
55. Pevzner, P. A., and H. Tang. 2001. Fragment assembly with double-barreled data. Bioinformatics 17(suppl 1):S225-S233.
56. Pevzner, P. A., H. Tang, and M. S. Waterman. 2001. An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences 98(17):9748-9753.
57. Posada-Buitrago, M., J. Boore, and R. Frederick. 2009. Soybean Rust Genome Sequencing Project. www.plantmanagementnetwork.com.
58. Project, P. G. S., and B. I. o. H. a. MIT. 2012. Puccinia graminis whole genome.
59. Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. Bmc Genomics 13 doi:10.1186/1471-2164-13-341.
60. Rees, R. G. 1972. Uredospore Movement And Observations On Epidemiology Of Wheat Rusts In North-Eastern Australia. Australian Journal of Agricultural Research 23(2):215-& doi:10.1071/ar9720215.
61. Ristaino, J. B., and M. L. Gumpertz. 2000. New frontiers in the study of dispersal and spatial analysis of epidemics caused by species in the genus Phytophthora. Annual Review of Phytopathology 38(1):541-576.
62. Roelfs, A. P. 1988. Genetic-Control Of Phenotypes In Wheat-Stem Rust. Annual Review of Phytopathology 26:351-367 doi:10.1146/annurev.py.26.090188.002031.
63. Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén. 1996. Real-time DNA sequencing using detection of pyrophosphate release. Analytical biochemistry 242(1):84-89.
64. Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, and M. Edwards. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348-352.
65. Sambrook, J., and D. W. Russell. 2001. Molecular cloning: a laboratory manual. Vol. 1. CSHL press.

66. Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74(12):5463-5467.
67. Schneider, R. W., C. A. Hollier, H. K. Whitam, M. E. Palm, J. M. McKemy, J. R. Hernandez, L. Levy, and R. DeVries-Paterson. 2005. First report of soybean rust caused by Phakopsora pachyrhizi in the continental United States. Plant Disease 89(7):774-774 doi:10.1094/pd-89-0774a.
68. Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and İ. Birol. 2009. ABySS: a parallel assembler for short read sequence data. Genome research 19(6):1117-1123.
69. Singh, R. P., D. P. Hodson, J. Huerta-Espino, Y. Jin, P. Njau, R. Wanyera, S. A. Herrera-Foessel, and R. W. Ward. 2008. Will stem rust destroy the world's wheat crop? Pages 271-309. in: Advances in Agronomy, Vol 98 D. L. Sparks, ed. Elsevier Academic Press Inc, San Diego.
70. Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, and L. E. Hood. 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321(6071):674-679.
71. Stanghellini, M., and J. Hancock. 1971. The sporangium of Pythium ultimum as a survival structure in soil. Phytopathology 61(15):7-164.
72. Stone, C. L., M. L. P. Buitrago, J. L. Boore, and R. D. Frederick. 2010. Analysis of the complete mitochondrial genome sequences of the soybean rust pathogens Phakopsora pachyrhizi and P. meibomiae. Mycologia 102(4):887-897 doi:10.3852/09-198.
73. Tawfik, D. S., and A. D. Griffiths. 1998. Man-made cell-like compartments for molecular evolution. Nature biotechnology 16(7):652-656.
74. Tooley, P. W., F. N. Martin, M. M. Carras, and R. D. Frederick. 2006. Real-Time Fluorescent Polymerase Chain Reaction Detection of Phytophthora ramorum and Phytophthora pseudosyringae Using Mitochondrial Gene Regions. Phytopathology 96(4):336-345 doi:10.1094/phyto-96-0336.
75. Tremblay, A. 2011. Soybean Rust: Five Years of Research.
76. Tremblay, A., P. Hosseini, S. Li, N. W. Alkharouf, and B. F. Matthews. 2012. Identification of genes expressed by Phakopsora pachyrhizi, the pathogen causing soybean rust, at a late stage of infection of susceptible soybean leaves. Plant Pathology 61(4):773-786 doi:10.1111/j.1365-3059.2011.02550.x.
77. Twizeyimana, M., and G. L. Hartman. 2012. Pathogenic Variation of Phakopsora pachyrhizi Isolates on Soybean in the United States from 2006 to 2009. Plant Disease 96(1):75-81 doi:10.1094/pdis-05-11-0379.
78. Tyler, B. M., S. Tripathy, X. Zhang, P. Dehal, R. H. Y. Jiang, A. Aerts, F. D. Arredondo, L. Baxter, D. Bensasson, and J. L. Beynon. 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313(5791):1261-1266.
79. Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J.

Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. 2001. The Sequence of the Human Genome. Science 291(5507):1304-1351 doi:10.1126/science.1058040.

80. Vettraino, A. M., S. Sukno, A. Vannini, and M. Garbelotto. 2010. Diagnostic sensitivity and specificity of different methods used by two laboratories for the detection of Phytophthora ramorum on multiple natural hosts. Plant Pathology 59(2):289-300 doi:10.1111/j.1365-3059.2009.02209.x.

81. Villavicencio, A., G. B. Fanaro, M. M. Araujo, S. Aquino, P. V. Silva, and J. Mancini-Filho. 2007. Detection of Phakopsora pachyrhizi by polymerase chain reaction (PCR) and use of germination test and DNA comet assay after e-bearn processing in soybean. Radiation Physics and Chemistry 76(11-12):1878-1881 doi:10.1016/j.radphyschem.2007.03.021.

82. Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10(1):57-63.
83. Warren, R. L., G. G. Sutton, S. J. M. Jones, and R. A. Holt. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23(4):500-501.
84. Wetterstrand, K. 2011. DNA Sequencing COsts: Data from NHGRI Large-Scale Genome Sequencing Program.
85. White, T. J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR protocols, a guide to methods and applications.
86. Wold, B., and R. M. Myers. 2008. Sequence census methods for functional genomics. Nature methods 5(1):19-21.
87. Yalow, R. S., and S. A. Berson. 1960. Immunoassay Of Endogenous Plasma Insulin In Man. The Journal of Clinical Investigation 39(7):1157-1175.
88. Zambino, P. J., and L. J. Szabo. 1993. Phylogenetic-Relationships Of Selected Cereal And Grass Rusts Based On rDNA Sequence-Analysis. Mycologia 85(3):401-414 doi:10.2307/3760702.

Chapter 3

A NEW APPROACH FOR DETECTING FUNGAL AND STRAMENOPILE PLANT PATHOGENS IN NEXT GENERATION SEQUENCING METAGENOME DATA UTILIZING ELECTRONIC PROBES

INTRODUCTION

Plant pathogen and pest dispersal to new areas has significant economic, ecological and evolutionary consequences that have the potential of being irreversible in the structure and functions of specific ecosystems, principally agricultural ecosystems (9; 11; 12). The development of rapid and accurate diagnostic methods for plant pathogens is crucial for the implementation of trading regulations. New method development is crucial for high impact pathogens, such as *Puccinia graminis* f. sp *tritici*, *Phakopsora pachyrhizi,* and *Phytophthora ramorum,* while evaluation of new methods on model systems, such as *Pythium ultimum,* can demonstrate the application of the newly developed methods to a broader range of pathogens, particularly ubiquitous soil inhabitants that can be often found in disease complexes.

Phytophthora ramorum is the causal agent of Sudden Oak Death disease (SOD) and Ramorum blight, causing severe symptoms in susceptible oaks as well as foliar symptoms on a wide range of herbaceous

and woody host (6; 7; 13; 18; 19; 39) principally in coastal forests in California (30) and southern Oregon (14; 26), but also in commercial greenhouses and nurseries. The accurate diagnosis of *P. ramorum* might be affected by the presence of other *Phytophthora* spp. that cause similar symptoms and have similar morphology (16) (24).

Puccinia graminis f. sp. *tritici* is the causal agent of wheat stem rust, a disease affecting wheat, rye, barley and oat. Almost all research on infection processes of *P. graminis* is focused on the uredinial stage because it has the highest economic impact (1; 20). The successful management of *P. graminis* f. sp. *tritici* in wheat requires an early and accurate diagnosis of the disease. Various factors are used for detecting the pathogen, principally visual keys are utilized. However, rust fungal species are often difficult to identify in early stages of disease development using morphology (4) and these crops can host various rust species.

Phakopsora pachyrhizi is an obligate fungal plant pathogen and the causal agent of Soybean rust. It is one of the most damaging diseases in soybean (*Glycine max*) causing yield losses of approximately 10 to 50% (1). *Phakopsora pachyrhizi* can be detected using PCR (38) and immunoassays (25) (5) from plant tissues presenting symptoms or directly from urediniospores, the only spore type produced by this microcyclic rust.

Pythium ultimum is an important and ubiquitous soil inhabitant plant pathogen that causes damping off and root rots in a broad range of hosts (21). Being a facultative saprophyte and opportunistic pathogen, it may cause damage to crops when environmental conditions are favorable for infection and colonization, particularly at early stages of seed and seedling development, and when plant health is compromised by environmental

stresses. Currently, *P. ultimum* infections are routinely diagnosed based on plant symptoms (i.e. yellowing, wilting), and signs (i.e. mycelia and oospores), and species identification is done based on morphology of sexual and asexual reproductive structures, and PCR based assays (10). Another preferred method for *Pythium* species identification is sequencing of the ITS region using ITS1-ITS4 primers (White et al. 1990) followed by BLASTn searches on the National Center for Biotechnology Information (NCBI) database. Biochemical methods for detecting *P. ultimum* involve monoclonal antibodies (MAb E5) that react to a cell wall glycoprotein of 46 kDa (3).

Diagnostic tools for the presence of Oomycete and fungal plant pathogens exist but they are limited in their scope and versatility. The recent development of new sequencing technologies have allowed the development of new approaches to detect plant pathogens (Stobbe et al. in press). The purpose of this research is to validate a newly developed plant pathogen diagnostic tool for detection and accurate identification of fungi and oomycetes. The new tool is called E-probe Diagnostic Nucleic acid Analysis (EDNA). The proposed approach uses unique-DNA pathogen signature sequences (e-probes), 20 to 140bp long, to detect plant pathogens from host plant metagenomic data using bioinformatics methods (Stobbe et al. in press). Roche 454 pyrosequencing was used to generate metagenomic databases, but the method should be compatible with other NGS platforms. Initially, 454 mock sample sequencing databases (MSSD) were generated, and used to assess the performance of the bioinformatic tool *in silico*. The evaluation of the technique implies the use of unique-DNA pathogen signatures termed electronic-probes (e-probes). The objective of this study was to optimize the EDNA bioinformatic tool for the analysis of metagenomic data from plants infected by fungi and/or oomycetes.

MATERIALS AND METHODS

Initially, 454 mock sample sequencing databases (MSSD) were generated, and used to assess the performance of the bioinformatic tool *in silico*. The evaluation of the technique utilizes unique-DNA pathogen sequence signatures termed electronic-probes (e-probes). These e-probes were generated using a modified version of a high throughput pipeline that was originally created to design microarray-based pathogen diagnostic assays called Tools for Oligonucleotide Fingerprint Identification (TOFI) (37). Once the e-probes were generated, they were searched for in the MSSD using BLAST+ (8).

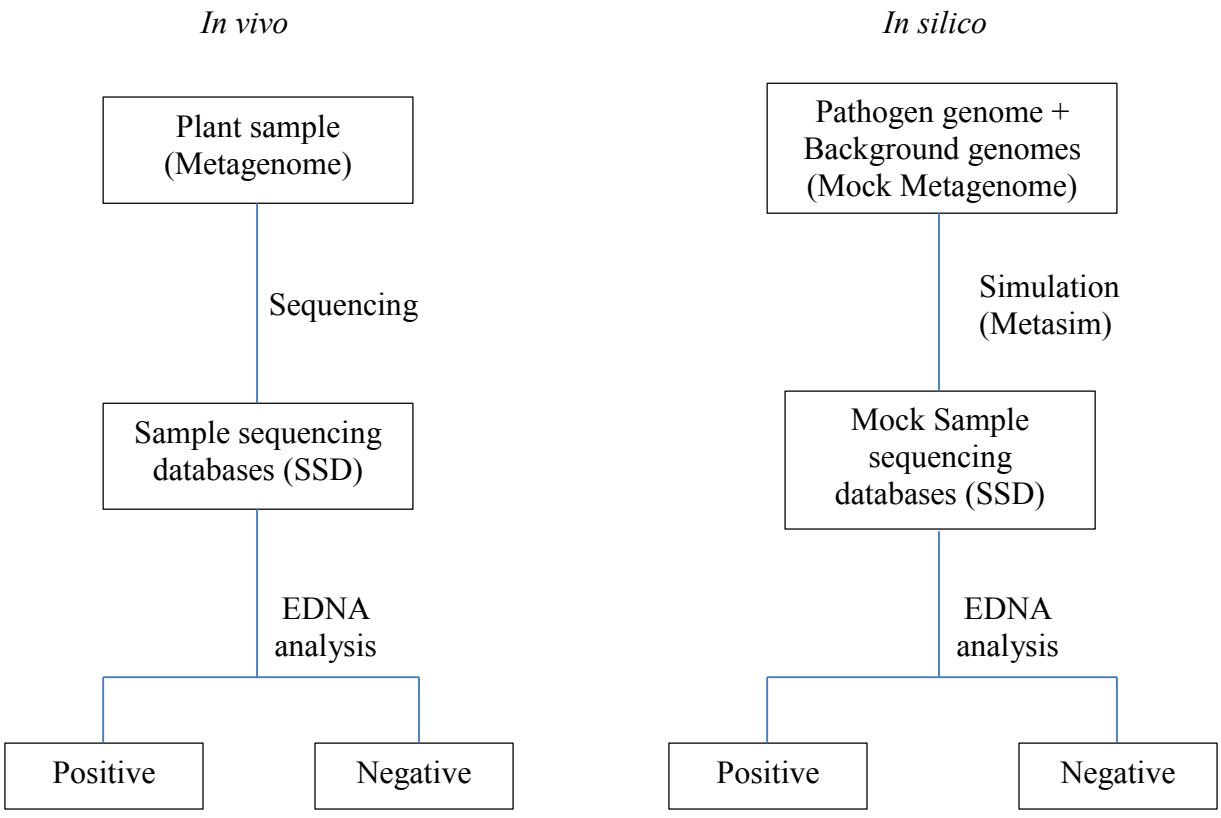

Figure *3-1:* EDNA concept in vivo and in silico for the diagnosis of eukaryotic plant pathogens

In silico assessment of EDNA in eukaryotic plant pathogens

E-probe Diagnostic Nucleic acid Analysis (EDNA) was developed to be used as a diagnostic tool, however, prior *in silico* assessment is necessary. EDNA was developed by a bioinformatics team from the National Institute for Microbial Forensics and Food & Agricultural Biosecurity (NIMFFAB) (Stobbe et al. in press). The *in silico* approach tested the accuracy, sensitivity and specificity of EDNA. This assessment included MSSDs constructed with different pathogen-read ratios (Table 3-1). Different pathogen/host ratios were used to evaluate the sensitivity of the technique using BLAST+ and e-probes.

E-Probe design

E-probe design at the species level required two genomes (Table 3-1), the target genome and a near neighbor genome. The target genome acted as a template to generate e-probes and the near neighbor helped to eliminate redundant genome regions in the target genome. Both, the elimination of redundant genome regions and the e-probe development were performed by a tool implemented to design pathogen diagnostic fingerprints termed Tools for Oligonucleotide Fingerprint identification (TOFI) (37). The identification tool was originally built in Perl language for use with both high-performance computing (HPC) and/or in personal computers. It included various versions: TOFI alpha was the first version of the program and it is a personal computer version (34); TOFI beta included several optimizations and significantly reduced the overall execution time of the pipeline (37); and a final version that included the parallel implementation for HPC (28). EDNA modified TOFI's pipeline by eliminating the UNAFold stage that involved microarray probe selection based on melting temperatures, and two-state folding or hybridization calculations. Instead, e-probes with varying lengths (40, 60, 80, 100, 120, 140, 160, 180

and 200 nt) were developed when comparing the target genome (pathogen genome) against the near neighbor genome (Table 3-1). TOFI takes advantage of SNP finding to select non-redundant areas of the target genome to develop the unique e-probes. E-probe databases are subjected to uniqueness assessment by BLASTn (2), which involves pairwise alignments of every e-probe with the nucleotide database of NCBI.

Table 3-1: Target Genome information used for the e-probe design of the four different pathogens

Organism	GenBank ID	Near Neighbor	Length	Source
***Phytophthora ramorum* strain Pr102**	AAQX00000000.1	*Phytophthora infestants*	66,652,401 bp	Full genome scaffold assembly
Phakopsora pachyrhizi	83921866-392996738	*Melampsora larici populina*	n/a	ESTs
***Pythium ultimum* strain DAOM BR144**	ADOS00000000.1	*Phytophthora infestants*	44,913,463 bp	Full genome scaffold assembly
***Puccinia graminis f. sp. tritici* CRL 75-36-700-3**	AAWC00000000.1	*Puccinia triticina*	81,600,488 bp	Full genome scaffold assembly

Certain 454 pyrosequencing errors are created when DNA contains homopolymeric regions with a length of 5-6 nt (31). To make these errors irrelevant while using EDNA, sequence regions with homopolymers were eliminated from the target genome sequence before it was processed by the modified TOFI. Therefore, e-probes lacked homopolymers.

All the e-probe databases were subjected to curation. This consisted of the elimination of e-probes that could cause false positive results, making the database more specific. Curation eliminated e-probes that were redundant in genomic data of other organisms obtained from public databases. Two perl scripts called

falsepositive_eliminator.pl and parser_falsepositive_eliminator.pl were used in this task. E-probe databases were pairwise aligned with sequences available through the nucleotide database (nt) from NCBI, and any e-probe that aligned with an e-value score lower than $1x10^{-3}$ was eliminated from the e-probe database. Higher e-values were not utilized to avoid the elimination of excessive number of pathogen-specific e-probes. While parsing the hits and matches from the previous alignment, pathogen specific sequences for the four fungal and stramenopile pathogens were excluded from this search.

The output of TOFI was a set of unique e-probes that was used later on during the EDNA assessment. E-probes with varying lengths (40-200 nt) were evaluated with simulated sequencing databases. The use of varying e-probe lengths was justified by the presence of varying read lengths and randomization of 454 pyrosequencing library fragments during DNA fragmentation. The optimum e-probe size was identified for each species using a newly developed criteria for match scoring when using EDNA. The scoring criteria as well as the diagnostic criteria are both fully described in the discussion section.

Mock Sample Sequencing Databases

Because of the relatively high cost of NGS, programs have been created to attempt the reduction of errors during experiment design to avoid failed sequencing runs. In this study an *in silico* assessment of EDNA's performance was conducted prior the use of real sequencing data. For that purpose, 454 sequencing data was simulated using MetaSim (29) by combining genomic information of a host and the individual pathogens at varying proportions. The sequencing databases obtained from these simulations were called Mock Sample Sequencing Databases (MSSDs).

The variable parameter in MSSDs was pathogen read abundance (Table 3-2). Pathogen read abundance included four different abundances (high, medium, low and very low). The background read abundance will be dependent on the pathogen read abundance and the total number of reads, which were limited to 10,000 for the assessment *in silico* (Table 3-2).

The genomes of *P. ultimum* (22), *P. ramorum* (35) and *P. graminis* (27) were used for making MSSDs, while ESTs libraries were used for *P. pachyrhizi* databases. The background host genome was grapevine (*Vitis vinifera*) (36). This plant host was selected because the public grapevine genome has 12x of coverage and most of its genome has been annotated, which facilitates further bioinformatics and statistical analysis. One hundred replicates were done of every MSSD.

Table 3-2: Molecular parameters for the construction of MSSDs

	454 reads		
Read abundances	Pathogen reads	Host reads	Total Reads
High	15%-25%	85%-75%	10,000
Medium	5%-15%	95%-85%	10,000
Low	0.5%-5%	99.5%-95%	10,000
Very Low	0.01%-0.5%	99.99%-99.5%	10,000
Negative Control	0%	100%	10,000

EDNA diagnosis

EDNA is a bioinformatics tool designed to avoid genome assembly and whole genome alignments to detect a plant pathogen in a stream of DNA sequences (Stobbe et al. in press). The present study optimized EDNA for the detection of eukaryotic plant pathogens, particularly fungi and oomycetes. MSSDs were subjected to analyses with EDNA in the Cowboy supercomputer at Oklahoma State University. Pairwise sequence alignment was performed between e-probes and MSSDs using BLASTn. The presence of the pathogen was detected when alignments (hits) were found between the specific pathogen e-probes and the database. Hits with e-values equal or lower than $1x10^{-9}$ and percent identity 95% or higher were considered high score hits (HSH) and were counted towards a positive match. Because of the stringency of these requirements, a positive match was called a High Quality Match (HQM). An e-probe was considered high-quality if it had multiple HSHs with the database. The HQMs must have an acceptable depth in order to confirm the presence of the pathogen in the metagenome. A depth of 4x or higher could be considered reliable based on the sequencing error rate that 454 pyrosequencing yields, since it has been observed an insertion and deletion error rate of approximately 3.3%, and substitution errors with a rate of 0.5% (23). Even with a higher error rate, it has been observed that consensus accuracies of 99.99% are achieved with a depth of coverage of four or more (23). However, depending of the pathogen biology, genome size and titer, depths lower than 4x may be considered for defining HQMs.

Negative controls did not contain any pathogen sequence in their database. Consequently, negative controls were expected to show zero matches. In addition to negative control MSSDs, non-specific e-probes were evaluated with positive MSSDs. Non-

specific e-probes were termed decoy e-probes and shuffled e-probes. Decoy e-probes were generated by inversing pathogen specific e-probes, which potentially would convert them into non-specific e-probes, and by shuffling the sequences, to avoid possible DNA inversions, which can be found in genetically variable populations of some fungi (15). Shuffled e-probes were generated from pathogen specific e-probes with a Perl script that randomly shuffled the nucleotide positions. Non-specific probes pairwise aligned with MSSDs were expected to have zero matches. It is stated in Stobbe et al. *in press* a statistical analysis where decoy e-probes are used as a negative control. Although this statistical analysis has shown to reduce significantly false positive results with preliminary data of fungal and Oomycete plant pathogens, improvements in the negative control concept needed to be adapted in order to be a reliable negative control method. Therefore, here, decoy e-probes are mostly used to look for the presence of inversions of specific chromosome areas in these four plant pathogens.

False positive calls were those EDNA calls that showed a positive result in a MSSD that lacked pathogen reads. True positive calls were those MSSDs that were called positive when the database contained the pathogen reads. Also, true negative calls were those MSSDs that were called negative when the pathogen reads were not present in the MSSDs. Finally, the False Negative calls were the MSSDs that although are known to contain the pathogen, they show absence of pathogen reads in the MSSD.

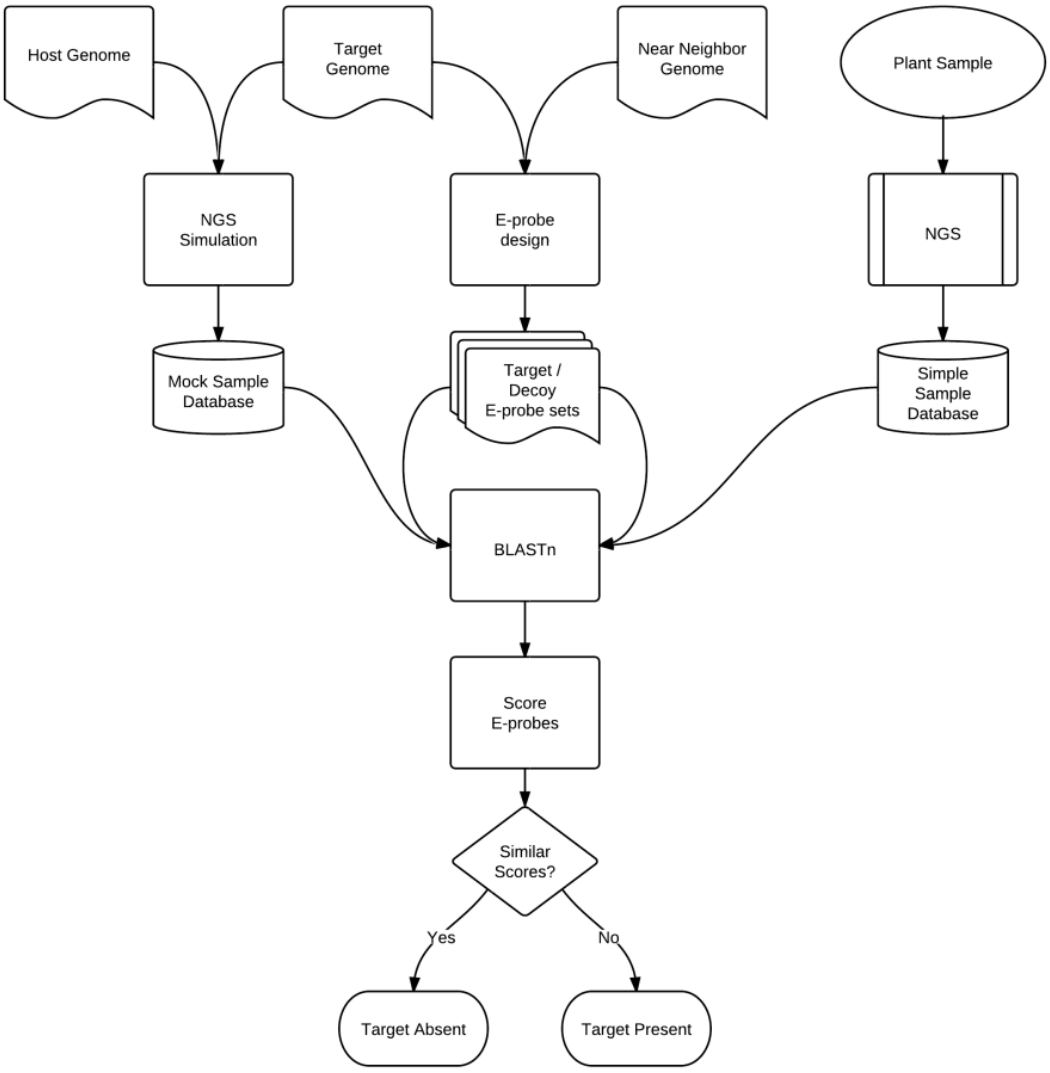

Figure 3-2. EDNA deployment for the detection of plant pathogens with Next Generation Sequencing (33)

Sensitivity and Specificity analysis

Sensitivity and specificity tests were conducted to compare e-probe lengths to select the optimal length and the limit of detection for each pathogen. These values were determined based on EDNA's effectiveness to detect the pathogen in MSSDs at different pathogen read abundances using probes of different lengths.

These tests allowed assessing the reliability to the proposed detection/identification model. The specificity analysis formula was $Sp = \frac{TN}{(FP+TN)}$, where TN is the number of true negative calls and FP is the number of false positive calls. The sensitivity analysis formula was $Sn = \frac{TP}{(TP+FN)}$, where TP is the number of true positive calls (aligning with pathogen reads) and FN (aligning with non-pathogen reads in control MSSDs) is the number of false negative calls.

For specificity and sensitivity analyses, the variable e-probe length and pathogen read abundances were used as reference. Therefore, separate analyses were conducted for the 9 different e-probe lengths (40, 60, 80, 100, 120, 140, 160, 180, and 200 nt) as well as for the different pathogen read abundances (High, Medium, Low and Very Low) in the MSSDs (Table 3-2).

RESULTS AND DISCUSSION

E-probe design

E-probe length was a limiting factor for the number of e-probes designed. As the e-probe length increases, the number of e-probes that the modified TOFI was able to design decreased. Although initially e-probes lengths of up to 200 nt were considered due to the large reads that 454 pyrosequencing provides, e-probe length range was decreased because sensitivity started to decrease at larger e-probe lengths (Figure 3-4). The number of e-probes varied among different e-probe lengths (Figure 3-2). Various parameters are measured to select the best e-probe length for pathogen detection. The parameters used in this study included sensitivity and specificity, and data processing time. Each parameter was measured and compared. E-probes 40 nucleotide long were produced in the scale of

hundreds of thousands for the four fungal and oomycete pathogens studied. The use of high numbers of e-probes was good for sensitivity, due to the subsequently higher number of unique signatures available to detect each pathogen; however, such a large data set made the computing process time consuming. Overall, e-probes 60 nt long provided optimal sensitivity, specificity, and data processing time (Figure 3-4 and Figure 3-5). E-probe database curation decreased the number of e-probes in low percentage. The final e-probes were considered unique and were expected to detect the pathogen in a metagenome 454 sequencing database.

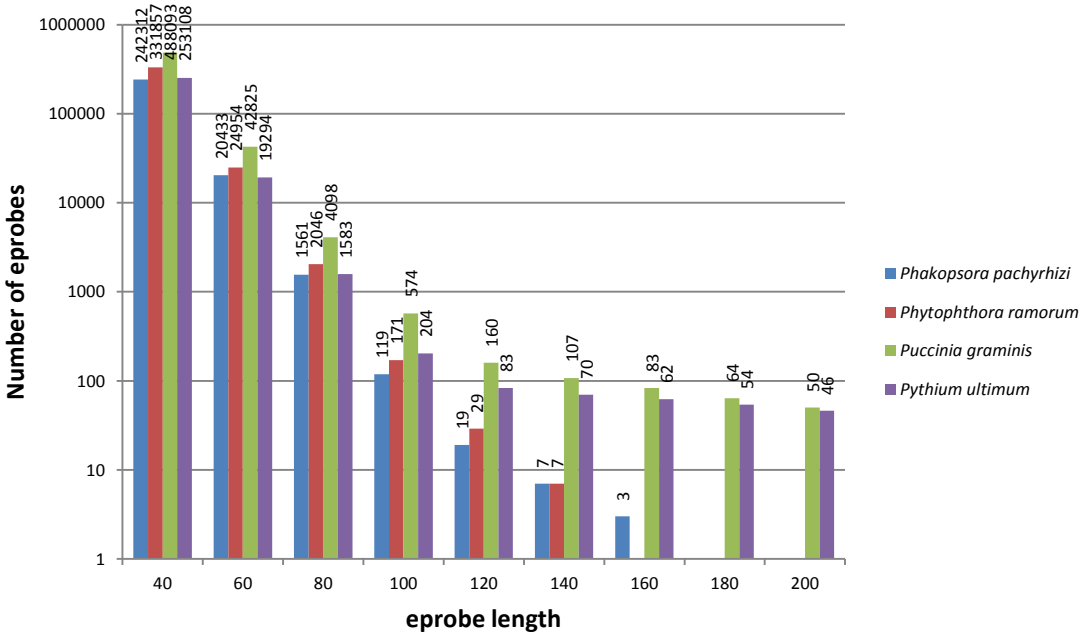

Figure 3-3. Variation of the number of e-probes designed among pathogens and e-probe length

Mock Sample Sequencing Database design

One hundred Mock Sample Sequencing databases were constructed for each of the four plant pathogens and for each pathogen read abundance (H, M, L and VL pathogen

read abundances). The high amount of replicates for MSSDs was justified by the need of a statistical validity for the EDNA trial with eukaryotic plant pathogens. MSSDs had 10,000 reads each, and pathogen read abundances were variable. The 454 simulation settings produced 200 cycles with an average of 509 base pairs per read. The average substitution rate in MSSDs was zero while the average insertion rate was 2.29% and the average deletion rate was 0.62%. Error values in MSSDs were in accordance with sequencing errors reported for 454 pyrosequencing (23).

Diagnostics with EDNA

Approximately 2,000 EDNA analyses were performed to provide a statistically valid sample size of 454 pyrosequencing metagenome databases. All the pathogens were detected at high, medium, and low read abundances. Very low read abundances produced ambiguous results due to false positives while performing cross analyses with the other three pathogens. In order to call a MSSD either positive or negative for a specific pathogen, a HQM limit of detection needed to be determined (Table 3-3). The detection limit (lowest HQM number to call a MSSDs positive for the presence of a pathogen) was obtained by pairwise alignment of all the pathogen e-probes against all the MSSDs. Any sample containing HQM equal or lower than HQM false positive limit were considered negative (Table 3-3). The HQM False positive limit has been calculated based on 2,000 MSSDs subjected to EDNA analysis. Although, the numbers of replicates are high, the HQM false positive limit could vary depending on the total number of MSSD utilized for the analyses, as well as for previous quality alignment consideration while eliminating non-pathogen specific e-probes. However, the validity that gives that high number of replicates might suggest considering this value a constant.

Table 3-3. False positive High quality matches in four eukaryotic plant pathogens using EDNA: Pha = *P. pachyrhizi;* Ram=*P. ramorum*

Organism	HQM False Pos. Limit	Organism w/ ambigueties
Phytophthora ramorum strain Pr102 (Ram)	25	Pha
Phakopsora pachyrhizi (Pha)	100	Ram
Pythium ultimum DAOM BR144 (Ult)	5	Ram
Puccinia graminis f. sp. tritici CRL 75-36-700-3 (Puc)	1	Ram

The HQM false positive limit (twilight zone) is a variable value that was adjusted depending on sensitivity and specificity yields. Therefore, for each pathogen, different twilight zones were calculated. A reason for this is because the e-probes and the EDNA approach could contain false positive HQMs that are considered noise, likewise in real time PCR the user has to learn to distinguish noise fluorescence from DNA amplification fluorescence. For qPCR there is software to automatize that task. Specifically for EDNA it doesn't have to be automatized unless various users need to design e-probes and validate their results. A datasheet with the calculations needed to determine this value are provided in Appendix 1.

An equation that includes HQM and HQM false positive limit (FPHQM) (Table 3-3) permits a more user friendly diagnostic call (C).

$$C = \frac{HQM}{FPHQM}$$

In the equation, if C is higher than 1, the MSSD is considered to be positive, conversely, if C is equal or lower than 1, the pathogen is considered to be absent in the SSD.

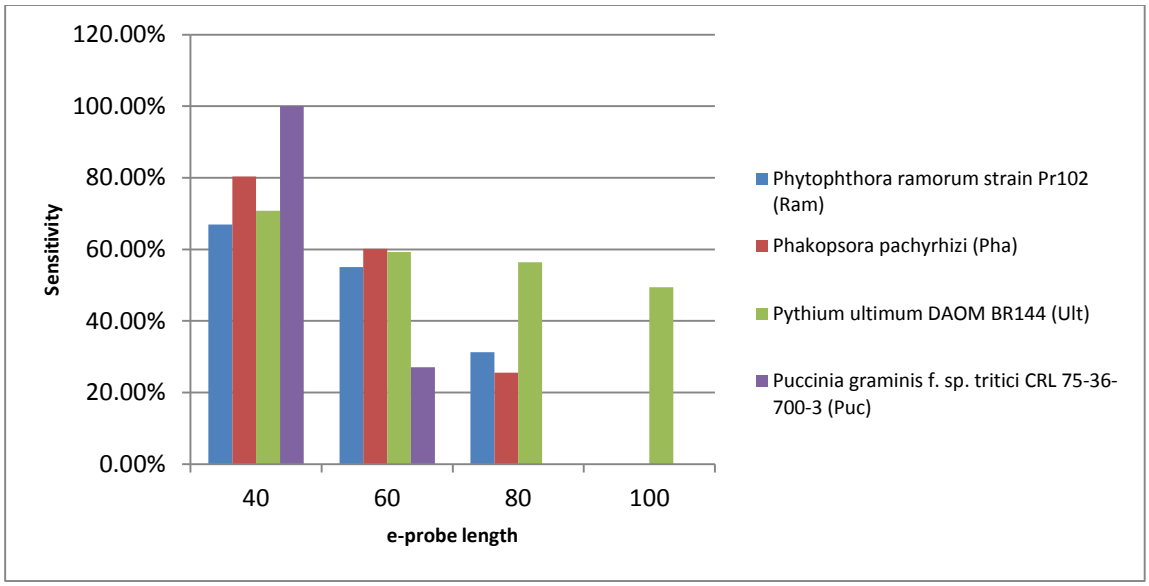

Figure 3-4. Relationship among e-probe length and sensitivity while using EDNA in four eukaryotic plant pathogens and four different pathogen abundances combined

The sensitivity of EDNA decreased while e-probe length increased (Figure 3-4). This phenomenon may be attributed to the number of e-probes contained in each database. Since the number of e-probes decreases when the e-probe length increases, the feasibility to detect the pathogen decreases tremendously. Therefore, high sensitivity values are restricted mostly to e-probes with lengths of either 40 nt or 60 nt. On the other hand, specificity of the diagnostic tool varied between 71.29% and 100% (Figure 3-5). The specificity of the test did not decrease prominently since the e-probes were meant to be very specific for each of the four plant pathogens. However, the best e-probe lengths having acceptable specificity were between 40 and 100 nt e-probe lengths. In order to select a diagnostic tool, both specificity and sensitivity must be considered. In this case, e-probes 60 nt long had the highest combined values of sensitivity and specificity for the four pathogens.

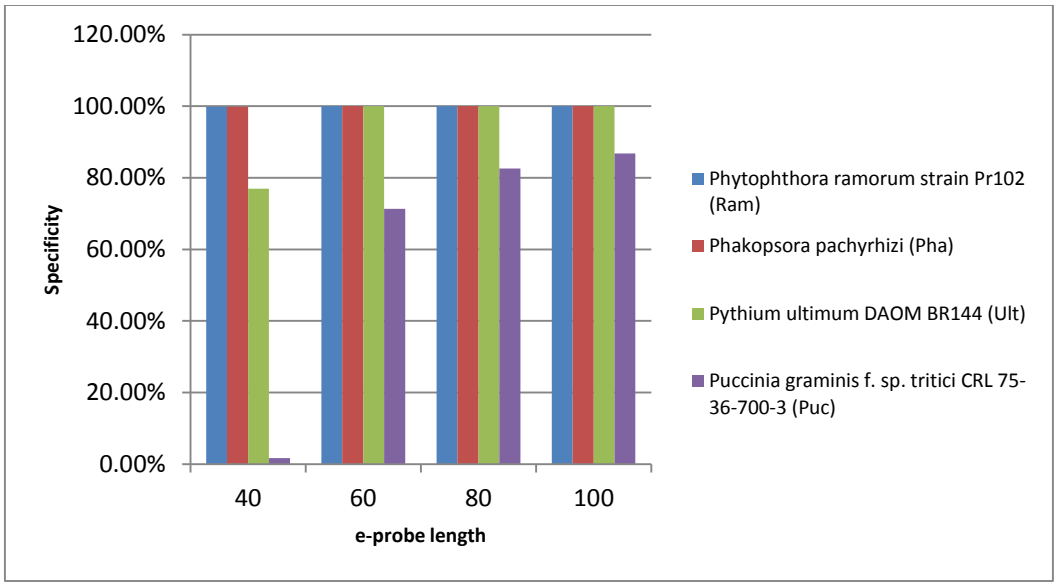

Figure 3-5. Relationship among e-probe length and specificity while using EDNA as a diagnostic tool in four eukaryotic plant pathogens with four different pathogen abundances

EDNA was a reliable system to detect plant pathogens in a stream of DNA sequences like 454 pyrosequencing output databases. It detected eukaryotic plant pathogens with high sensitivity and specificity when utilizing e-probe lengths between 40 and 60 nt at high, medium, and low pathogen read abundance. At very low pathogen read abundance detection was unreliable. However, specificity was maintained at 100% even at very low pathogen abundance. Conclusively, the randomness of NGS when sequencing large metagenomes plays an important role in sensitivity of EDNA. The likelihood of pathogen specific reads to be found in a metagenome decreases as the metagenome is larger and the pathogen titer is lower. On the other hand, specificity is not database dependent, applying EDNA as a diagnostic tool maintained a high specificity due to the highly specific e-probes designed. Various bioinformatics filters allowed keeping only pathogen specific e-probes in our databases. All these factors influenced the pathogen detection using

EDNA. The sensitivity of EDNA reduced as the length of e-probes increased and as the abundance of pathogen reads reduced in MSSDs.

MSSDs that contained 10,000 total reads were used in this study although 454 pyrosequencing is capable of sequencing approximately 150,000 reads in one single run. The objective of using lower number of total reads was to demonstrate that the pathogens could be detected if approximately 15 barcoded samples were analyzed in a single 454 pyrosequencing run. Eventually, NGS will become cheaper and there will be no need of barcoding samples. Although, EDNA could be compared with bioinformatics tools that were developed principally to identify organisms in NGS output databases like Metaphlan and MEGAN (17) (32), our tool offers the assurance of the pathogen presence in the database. While other tools only provide the number of reads belonging to the target organisms (17), EDNA uses specific signatures of the pathogen and can realistically decide whether the pathogen is present or not in the original sample. There are not studies where the detection of Fungal or Oomycete plant pathogens was performed using NGS output databases, this fact makes EDNA the pioneer in the utilization of NGS data to detect eukaryotic plant pathogens.

LITERATURE CITED

1. Agrios, G. 2005. Plant Pathology. Edited by Fifth. Elsevier.

2. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology 215(3):403-410 doi:10.1016/s0022-2836(05)80360-2.

3. Avila-Rodriguez, F. J. 1994. Serological characterization of Pythium ultimum. 9507805, The University of Nebraska - Lincoln, United States -- Nebraska.

4. Barnes, C. W., and L. J. Szabo. 2007. Detection and Identification of Four Common Rust Pathogens of Cereals and Grasses Using Real-Time Polymerase Chain Reaction. Phytopathology 97(6):717-727 doi:10.1094/phyto-97-6-0717.

5. Baysal-Gurel, F., M. L. L. Ivey, A. Dorrance, D. Luster, R. Frederick, J. Czarnecki, M. Boehm, and S. A. Miller. 2008. An immunofluorescence assay to detect urediniospores of Phakopsora pachyrhizi. Plant Disease 92(10):1387-1393 doi:10.1094/pdis-92-10-1387.

6. Beales, P. A., A. Schlenzig, and A. J. Inman. 2004. First report of ramorum bud and leaf blight (Phytophthora ramorum) on Syringa vulgaris in the UK. Plant Pathology 53(4):525-525 doi:10.1111/j.1365-3059.2004.01033.x.

7. Beales, P. A., T. Brokenshire, A. V. Barnes, V. C. Barton, and K. J. D. Hughes. 2004. First report of ramorum leaf blight and dieback (Phytophthora ramorum) on Camellia spp. in the UK. Plant Pathology 53(4):524-524 doi:10.1111/j.1365-3059.2004.01028.x.

8. Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. BMC bioinformatics 10(1):421.

9. Castello, J. D., D. J. Leopold, and P. J. Smallidge. 1995. PATHOGENS, PATTERNS, AND PROCESSES IN FOREST ECOSYSTEMS. Bioscience 45(1):16-24 doi:10.2307/1312531.

10. Cullen, D. W., I. K. Toth, N. Boonham, K. Walsh, I. Barker, and A. K. Lees. 2007. Development and Validation of Conventional and Quantitative Polymerase Chain Reaction Assays for the Detection of Storage Rot Potato Pathogens, Phytophthora erythroseptica, Pythium ultimum and Phoma foveata. Journal of Phytopathology 155(5):309-315 doi:10.1111/j.1439-0434.2007.01233.x.

11. Enserink, M. 1999. Predicting invasions: Biological invaders sweep in. Science 285(5435):1834-1836 doi:10.1126/science.285.5435.1834.

12. Everett, R. A. 2000. Patterns and pathways of biological invasions. Trends in Ecology & Evolution 15(5):177-178 doi:10.1016/s0169-5347(00)01835-8.

13. Giltrap, P. M., A. J. Inman, V. C. Barton, A. V. Barnes, C. R. Lane, K. J. D. Hughes, J. Tomlinson, M. L. Dean, and K. Izzard. 2004. First report of ramorum dieback (Phytophthora ramorum) on Hamamelis virginiana in the UK. Plant Pathology 53(4):526-526 doi:10.1111/j.1365-3059.2004.01034.x.

14. Goheen, E., E. Hansen, A. Kanaskie, M. McWilliams, N. Osterbauer, and W. Sutton. 2002. Sudden oak death caused by Phytophthora ramorum in Oregon. Plant Disease 86(4):441-441.

15. Hane, J. K., T. Rouxel, B. J. Howlett, G. H. J. Kema, S. B. Goodwin, and R. P. Oliver. 2011. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. Genome biology 12(5):R45.

16. Hansen, E. M., P. W. Reeser, W. Sutton, and L. M. Winton. 2003. First report of A1 mating type of Phytophthora ramorum in North America. Plant Disease 87(10):1267-1267 doi:10.1094/pdis.2003.87.10.1267a.

17. Huson, D. H., S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome research 21(9):1552-1560.

18. Lane, C. R., P. A. Beales, K. J. D. Hughes, J. A. Tomlinson, A. J. Inman, and K. Warwick. 2004. First report of ramorum dieback (Phytophthora ramorum) on container-grown English yew (Taxus baccata) in England. Plant Pathology 53(4):522-522 doi:10.1111/j.1365-3059.2004.01022.x.

19. Lane, C. R., P. A. Beales, K. J. D. Hughes, R. L. Griffin, D. Munro, C. M. Brasier, and J. F. Webber. 2003. First outbreak of Phytophthora ramorum in England, on Viburnum tinus. Plant Pathology 52(3):414-414 doi:10.1046/j.1365-3059.2003.00835.x.

20. Leonard, K. J., and L. J. Szabo. 2005. Stem rust of small grains and grasses caused by Puccinia graminis. Molecular Plant Pathology 6(2):99-111 doi:10.1111/j.1364-3703.2005.00273.x.

21. Levesque, C. A., H. Brouwer, L. Cano, J. Hamilton, C. Holt, E. Huitema, S. Raffaele, G. Robideau, M. Thines, J. Win, M. Zerillo, G. Beakes, J. Boore, D. Busam, B. Dumas, S. Ferriera, S. Fuerstenberg, C. Gachon, E. Gaulin, F. Govers, L. Grenville-Briggs, N. Horner, J. Hostetler, R. Jiang, J. Johnson, T. Krajaejun, H. Lin, H. Meijer, B. Moore, and P. Morris. 2010. Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. Genome Biology 11(7):R73.

22. Levesque, C. A., H. Brouwer, L. Cano, J. P. Hamilton, C. Holt, E. Huitema, S. Raffaele, G. P. Robideau, M. Thines, J. Win, M. M. Zerillo, G. W. Beakes, J. L. Boore, D. Busam, B. Dumas, S. Ferriera, S. I. Fuerstenberg, C. M. M. Gachon, E. Gaulin, F. Govers, L. Grenville-Briggs, N. Horner, J. Hostetler, R. H. Y. Jiang, J. Johnson, T. Krajaejun, H. Lin, H. J. G. Meijer, B. Moore, P. Morris, V. Phuntmart, D. Puiu, J. Shetty, J. E. Stajich, S. Tripathy, S. Wawra, P. van West, B. R. Whitty, P. M. Coutinho, B. Henrissat, F. Martin, P. D. Thomas, B. M. Tyler, R. P. De Vries, S. Kamoun, M. Yandell, N. Tisserat, and C. R. Buell. 2010. Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. Genome Biology 11(7) doi:10.1186/gb-2010-11-7-r73.

23. Margulies, M., M. Egholm, W. Altman, S. Attiya, J. Bader, L. Bemben, J. Berka, M. Braverman, Y. Chen, Z. Chen, S. Dewell, L. Du, J. Fierro, X. Gomes, B. Godwin, W. He, S. Helgesen, C. Ho, G. Irzyk, S. Jando, M. Alenquer, T. Jarvie, K. Jirage, J. Kim, J. Knight, J. Lanza, J. Leamon, S. Lefkowitz, M. Lei, J. Li, K. Lohman, H. Lu, V. Makhijani, K. McDade, M. McKenna, E. Myers, E. Nickerson, J. Nobile, R. Plant, B. Puc, M. Ronan, G. Roth, G. Sarkis, J. Simons, J. Simpson, M. Srinivasan, K. Tartaro, A. Tomasz, K. Vogt, G. Volkmer, S. Wang, Y. Wang, M. Weiner, P. Yu, R. Begley, and J. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376 - 380.

24. Martin, F. N., and P. W. Tooley. 2003. Phylogenetic relationships of Phytophthora ramorum, P-nemorosa, and P-pseudosyringae, three species recovered from areas in California with sudden oak death. Mycological research 107:1379-1391 doi:10.1017/s0953756203008785.

25. Miranda, B. S., E. M. Linares, S. Thalhammer, and L. T. Kubota. 2013. Development of a disposable and highly sensitive paper-based immunosensor for early diagnosis of Asian soybean rust. Biosensors and Bioelectronics.

26. Parke, J. L., R. G. Linderman, N. K. Osterbauer, and J. A. Griesbach. 2004. Detection of Phytophthora ramorum blight in Oregon nurseries and completion of Koch's postulates on Pieris, Rhododendron, Viburnum and Camellia. Plant Disease 88(1):87-87 doi:10.1094/pdis.2004.88.1.87a.

27. Project, P. G. S., and B. I. o. H. a. MIT. 2012. Puccinia graminis whole genome.

28. Ravi Vijaya, S. 2009. Parallel Implementation of a Bioinformatics Pipeline for the Design of Pathogen Diagnostic Assays. Pages 213-218 K. Kamal, Z. Nela, and R. Jaques, eds.

29. Richter, D. C., F. Ott, A. F. Auch, R. Schmid, and D. H. Huson. 2008. MetaSim—A Sequencing Simulator for Genomics and Metagenomics. PLoS ONE 3(10):e3373 doi:10.1371/journal.pone.0003373.

30. Rizzo, D. M., M. Garbelotto, J. M. Davidson, G. W. Slaughter, and S. T. Koike. 2002. Phytophthora ramorum as the cause of extensive mortality of Quercus spp. and Lithocarpus densiflorus in California. Plant Disease 86(3):205-214 doi:10.1094/pdis.2002.86.3.205.

31. Ronaghi, M. 2001. Pyrosequencing Sheds Light on DNA Sequencing. Genome Research 11(1):3-11 doi:10.1101/gr.150601.

32. Segata, N., L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Meth 9(8):811-814 doi:http://www.nature.com/nmeth/journal/v9/n8/abs/nmeth.2066.html#supplementary-information.

33. Stobbe, A., J. Daniels, A. Espindola, R. Verma, U. Melcher, F. Ochoa-Corona, C. Garzon, J. Fletcher, and W. Schneider. 2013. E-probe Diagnostic Nucleic acid Analysis

(EDNA): A theoretical approach for handling of next generation sequencing data for diagnostics. Microbiological Methods.

34. Tembe, W., N. Zavaljevski, E. Bode, C. Chase, J. Geyer, L. Wasieloski, G. Benson, and J. Reifman. 2007. Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays. Bioinformatics 23(1):5-13.

35. Tyler, B. M., S. Tripathy, X. Zhang, P. Dehal, R. H. Y. Jiang, A. Aerts, F. D. Arredondo, L. Baxter, D. Bensasson, and J. L. Beynon. 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313(5791):1261-1266.

36. Velasco, R., A. Zharkikh, M. Troggio, D. A. Cartwright, A. Cestaro, D. Pruss, M. Pindo, L. M. FitzGerald, S. Vezzulli, J. Reid, G. Malacarne, D. Iliev, G. Coppola, B. Wardell, D. Micheletti, T. Macalma, M. Facci, J. T. Mitchell, M. Perazzolli, G. Eldredge, P. Gatto, R. Oyzerski, M. Moretto, N. Gutin, M. Stefanini, Y. Chen, C. Segala, C. Davenport, L. Demattè, A. Mraz, J. Battilana, K. Stormo, F. Costa, Q. Tao, A. Si-Ammour, T. Harkins, A. Lackey, C. Perbost, B. Taillon, A. Stella, V. Solovyev, J. A. Fawcett, L. Sterck, K. Vandepoele, S. M. Grando, S. Toppo, C. Moser, J. Lanchbury, R. Bogden, M. Skolnick, V. Sgaramella, S. K. Bhatnagar, P. Fontana, A. Gutin, Y. Van de Peer, F. Salamini, and R. Viola. 2007. A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLoS ONE 2(12):e1326 doi:10.1371/journal.pone.0001326.

37. Vijaya Satya, R., N. Zavaljevski, K. Kumar, and J. Reifman. 2008. A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. BMC Bioinformatics 9(1):185.

38. Villavicencio, A., G. B. Fanaro, M. M. Araujo, S. Aquino, P. V. Silva, and J. Mancini-Filho. 2007. Detection of Phakopsora pachyrhizi by polymerase chain reaction (PCR) and use of germination test and DNA comet assay after e-bearn processing in soybean. Radiation Physics and Chemistry 76(11-12):1878-1881 doi:10.1016/j.radphyschem.2007.03.021.

39. Werres, S., and D. De Merlier. 2003. First detection of Phytophthora ramorum mating type A2 in Europe. Plant Disease 87(10):1266-1266 doi:10.1094/pdis.2003.87.10.1266c.

Chapter 4

E-PROBE DIAGNOSTICS FOR NUCLEIC ACID ANALYSES (EDNA) VALIDATION IN SAMPLE SEQUENCING DATABASES FROM 454 PYROSEQUENCING OF EUKARYOTIC PLANT PATHOGENS

INTRODUCTION

Molecular biology diagnostic tools, like immunoassays and nucleic acid-based analyses have permitted a high degree of sensitivity and specificity in diagnostics (7; 11; 17). Nevertheless, the emergence of new strains of pathogens, principally among those reproducing sexually, may decrease the efficacy of currently available diagnostic tests. The possible inaccuracy of molecular-based diagnostics emerges in part from the single target dependence of these methods, meaning that they target either a specific protein motif or a specific gene or locus in the target genome. Furthermore, the presence of SNPs on the target nucleic acid sequences may affect the annealing of diagnostic primers during PCR reactions.

Fungi and chromista plant pathogens are among the most economically important because of their impact on agriculture and natural ecosystems.

These plant pathogens cause enormous crop losses worldwide (1). The focus of this study was to validate the new bioinformatic tool EDNA (Stobbe et al., in press) for the detection of fungal and stramenopile plant pathogens from 454 sequencing output data. The pathogens targeted were the rust pathogens *Puccinia graminis* f sp. *tritici* (Wheat Stem Rust) and *Phakopsora pachyrhizi* (Soybean Rust), and the chromista *Phytophthora ramorum* (Sudden Oak death and Ramorum blight) and *Pythium ultimum* (Pythium damping-off, root rot).

Diagnostic techniques for the four pathogens have been limited to either molecular or serological analyses (19) (3); (14) (2); *(11)*; (12). Molecular methods rely mostly on unique areas of the genome that permit the specific amplification of the target genome, allowing the visualization of either a band in an agarose gel or fluorescence when using qPCR. All of these techniques are effective and specific for the detection of their target pathogens. However, all of these assays face the same limitations that all PCR and immunological assays face, namely requirement for a priori characterization and reagents (antibodies or primer sequences) and limited multiplexing capacity. Next generation sequencing (NGS) has spawned the study of metagenomics, the sequencing of all nucleic acids from all organisms in a given environmental sample. Metagenomics has been applied to diagnostics (10) and plant pathogen detection (16), but never to the detection of eukaryotic plant pathogens. In addition, traditional metagenomic approaches to handling data were cumbersome, and not tailored to the needs of diagnosticians.

E-probe for diagnostics nucleic acid Analysis (EDNA) has been previously validated in Mock Sample Sequencing Databases (chapter III) showing promising results

for the detection of fungal and chromista plant pathogens (8). EDNA is a bioinformatics-based diagnostic tool that utilizes unique signatures of the pathogen genome to detect the eukaryotic plant pathogens in NGS metagenomic databases. Previous assessment of EDNA *in silico* demonstrated that the bioinformatic tool was highly specific and reliable at variable pathogen concentrations (pathogen read abundance above 0.5%) in NGS output data bases, also called Sample Sequencing Databases (SSDs) (16).

NGS has become a widely used tool, enhancing sequencing results and speeding data gathering. In addition, the costs of sequencing are decreasing due to service provider competition. However, the increased efficiency of NGS creates data handling and management issues that limit the speed and effectiveness of the process. Due to the enormous amount of data generated by NGS, it is fundamental to perform specialized bioinformatic analyses to efficiently retrieve the targeted information. This study focuses on the detection of fungal and chromista plant pathogens from real NGS output databases.

Genetic features of each pathogen, like genome size and/or ploidy are crucial when using EDNA. Unlike other NGS databases analysis programs like MEGAN4 (10) and Metaphlan (15), EDNA uses unique signatures of the pathogen genome (e-probes) reducing the time consuming pairwise alignment of the total SSD with the nr database on NCBI. Genome size is important for e-probe design, it was previously shown that the large genome size of eukaryotic organisms allows the design of larger e-probes and higher number of e-probes (chapter III). Another advantage of EDNA is the elimination of the time consuming and computing intense assembling process.

EDNA was previously tested with simulated SSDs containing sequencing reads of fungi, viruses and bacteria, successfully identifying important plant pathogens (Stobbe et al. in press). Although, viral and bacterial pathogens can be detected using EDNA, eukaryotic plant pathogens have a higher likelihood of being detected due to their larger genomes, which increase the proportion of pathogen reads being analyzed. However, the difficulties with eukaryotic plant pathogen detection via NGS and EDNA lie in the degree of relatedness between the pathogen and the eukaryotic host as well as the high likelihood of uncharacterized eukaryotic endophytes commonly found associated with plants. EDNA had not been validated using actual sequencing runs from fungi and chromista infected plants, a critical step described in this manuscript.

MATERIALS AND METHODS

Sample preparation

Pythium ultimum (P17) was provided by the Soilborne plant pathogens laboratory in Oklahoma State University. The isolate was originally retrieved I n Guilford Gardens in Chambersburg, PA. PARP agar cultures 72h old were used to inoculate potato slices. The inoculated potato slices were kept in humid chambers for 5 days until DNA extraction. Whole DNA was extracted from infected potato using the DNease Plant mini kit (Qiagen, Austin, TX), without previous tuber surface disinfection or mycelium isolation. *Phytophthora ramorum* infected rhododendron leafs showing dark spots and mycelium were use to obtain whole nucleic acids. The DNA and the RNA of this sample were extracted separately using the QDNAeasy Plant Mini kit and the RNeasy Plant Mini kit (Qiagen). After the nucleic acid extractions, both DNA and RNA were mixed together to

take advantage of the highly expressed RNA sequences from *P. ramorum* during an active infection. Additionally, DNA of *Phakopsora pachyrhizi* infected soybean was provided by Dr. Kerry Pedley (USDA Foreign Disease-Weed Science in Ft. Detrick, MD); and DNA of *Puccinia graminis* f. sp. *tritici* infected wheat was provided by Dr. Les Szabo (USDA Cereal Disease Laboratory).

Fungi and Chromista infected plant samples contained nucleic acids from all the microbiota present in the sample at the moment of the nucleic acid extraction. In order to amplify non-representative genome sequences, the whole nucleic acid was amplified using random hexamers. Whole Genome Amplification (WGA) was performed with *P. ultimum* and *P. pachyrhizi* DNA using the Sigma Aldrich WGA Kit 1 for *P. pachyrhizi* and the Genomi phi WGA from General Electric Healthcare. For *P. ultimum*. *P. graminis* and *P. ramorum* samples were amplified using the WGA kit (Genomi Phi, Buckingh amshire, UK) and the Whole Transcriptome Amplification (WTA) Transplex (Sigma Aldrich, St. Louis, MO).

NGS was performed in a 454 Titanium Genome Sequencer, at the Oklahoma State University Nucleic Acid and Protein Core Facility. Four different full plate runs were completed; each run contained a different plant pathogen. All the runs were configured to perform 200 cycles. The raw reads were trimmed using the program that the sequencer is equipped with. The E-probe Diagnosis for Nucleic Acid (EDNA) script was run in the output FASTA file containing only high quality reads (chapter III).

EDNA diagnosis in Sample Sequencing Databases (SSDs)

Electronic Probe Diagnosis for Nucleic Acid (EDNA) is a tool that permits the identification of specific organisms in metagenomic sequencing data (16). The EDNA pipeline was optimized for utilization with fungal and stramenopile plant pathogens (chapter III). The e-probe databases selected for eukaryotic pathogens included lengths from 40 nt to 60 nt which were selected based on previous sensitivity and specificity analyses with MSSDs containing a total number of reads of 10,000. The eukaryotic e-probes were designed by aligning the pathogen genome against the nearest neighbor organism genome available using the UNIX script Tools for oligonucleotide fingerprinting (TOFI). TOFI was originally created for the design of probes suitable for microarray analysis (18). However, the pipeline was modified for e-probe design. The original TOFI script was downloaded from http://www.bhsai.org/downloads/tofi_beta.tar.gz . Its principle is to make pairwise comparisons of the target genome with a non-target genome (near neighbor) eliminating common regions among these genomes and providing unique fingerprints for the target genome (pathogen genome). The pairwise comparisons were performed by the program Mummer (13). E-probes were Blastn-searched against the NCBI's nucleotide database, and e-probes showing similarity any another organism in the database were eliminated (16).

EDNA aligned the 454 sequencing raw reads with curated and validated e-probes 40 and 60 nt long. The number of high quality matches (HQM) were recovered and verified with the HQM false positive limit, previously calculated (8). If the number of HQMs was higher than the HQM false positive limit, the sample was considered positive for the

pathogen analyzed. Conversely, if the HQM number was lower than the HQM false positive limit, the sample was considered negative.

Detection parameters

Each e-probe detected in the sequencing raw reads were scored based on selected BLASTn search parameters (chapter III). The parameters used were both e-value and percent identity. If an e-probe had hit in a read, the hit may be kept based on its e-value and max identity. If the maximum identity was higher than 95% and the e-value lower than $1x10^{-9}$, the hit was counted for the analyses. Good sequencing outputs had a coverage of 4x or higher. Therefore, 4 hits were enough to tell that the e-probe was truly present and therefore a pathogen unique sequence was found in the database.

A consequence of the genome breakage process that occurs during 454 sequencing and posterior DNA amplification of the fragments is that many reads will be repetitive in the output database (SSDs), therefore one single e-probe may have more than one hit while others may have none. The hits used in this analysis were high scoring hits (HSH), which gave further validity to the pairwise alignments performed by the EDNA protocol.

As described previously, a SSDs diagnostic call is dependent on the false positive HQMs limit, a constant calculated using MSSDs statistics (chapter III). The selection of these values is important when using EDNA and, they are calculated right after e-probe design. False positive HQMs might be a result of sharing genes with the genomes of other unrelated organisms, and are unique for each pathogen but they might be variable within a species (8) (6) (5) (4). The sharing of genes between organisms is a problem that has to be faced by the diagnosticians. When organisms share genes, molecular-based diagnostic techniques have to be tested carefully to avoid false positives. Particularly in EDNA, since

it uses whole genome sequences and the chances of e-probes hitting shared genes are much higher than in PCR based or inmmunoassays. The reduction of false positive HQMs was performed by pairwise alignment of e-probe databases with the nucleotide database on NCBI. Since NCBI's records are limited, there is a margin of error in the analysis. As a consequence there may be shared e-probes that will not be eliminated and a tolerance false positive HQM number must to be used (Espindola, et al in press). The calculation of this FPHQM limit is taken from the highest number of FPHQM that are found for each set of e-probes.

Decoy e-probes were designed utilizing two bio-perl scripts developed for this research. The script decoymaker.pl changed the sense of the e-probe sequence, instead of being 3'—5' it was 5'—3' (Stobbe et al. in press). This change in all e-probes was performed to create a negative control environment. These e-probes were not expected to hit any positive sample, however in eukaryotic organisms, transposable elements and DNA rearrangements may result in some inverted decoy e-probes having hits in the SSDs. Therefore, another bio-perl script was used, called shuffledeprobe.pl, which produced another set of negative e-probes by shuffling the sequences of the original positive e-probes (Espindola et al. in press).

RESULTS AND DISCUSSION

Calling a metagenomic sample positive or negative for certain organisms can be a difficult task. Currently available metagenome based diagnostic methods align the whole sequencing database to the nucleotide database on NCBI. The output often results in a great number of high scoring alignments with different organisms. These alignments can tell the user that the organism is present, however, misidentification may happen due to lack of

specificity for certain sequences uploaded to NCBI. A high number of sequences that are on GenBank and NCBI can be shared by multiple organisms. If such is the case, a positive high scoring hit can be found, but it may not be the pathogen of interest. The advantage of EDNA with eukaryotic pathogens is that it utilizes unique signatures designed from the reference pathogen genome or EST databases (16) and eliminates all the redundant and useless sequences, while identifying the pathogen in less time than currently used methods, since it doesn't require contig assembly (Figure 4.1).

Four SSDs were retrieved after 454 pyrosequencing infected plant samples. The .sff files were quality trimmed using mothur, and subsequently the EDNA diagnostic tool was used for specific pathogen detection (Figure 4-1). Eight different e-probe databases were used, 2 per each SSD (40 nt and 60 nt). The diagnostic analysis effectively detected the pathogens presence in the four samples respectively (Table 4-1). However, not all e-probe lengths were able to detect all the pathogens.

When using real sequencing data, 60 nt length e-probe databases are not able to detect the pathogen for *P. pachyrhizi* (Table 4-1). This false negative result might be caused by the genetic data used for e-probe development of this pathogen. For this organism the e-probes were generated from ESTs and whole genome was not used since one is still not available on public databases. The disadvantage of using ESTs for e-probe design is that this genetic information usually is collected from some specific stages of the pathogen development and, does not contain data from other stages.

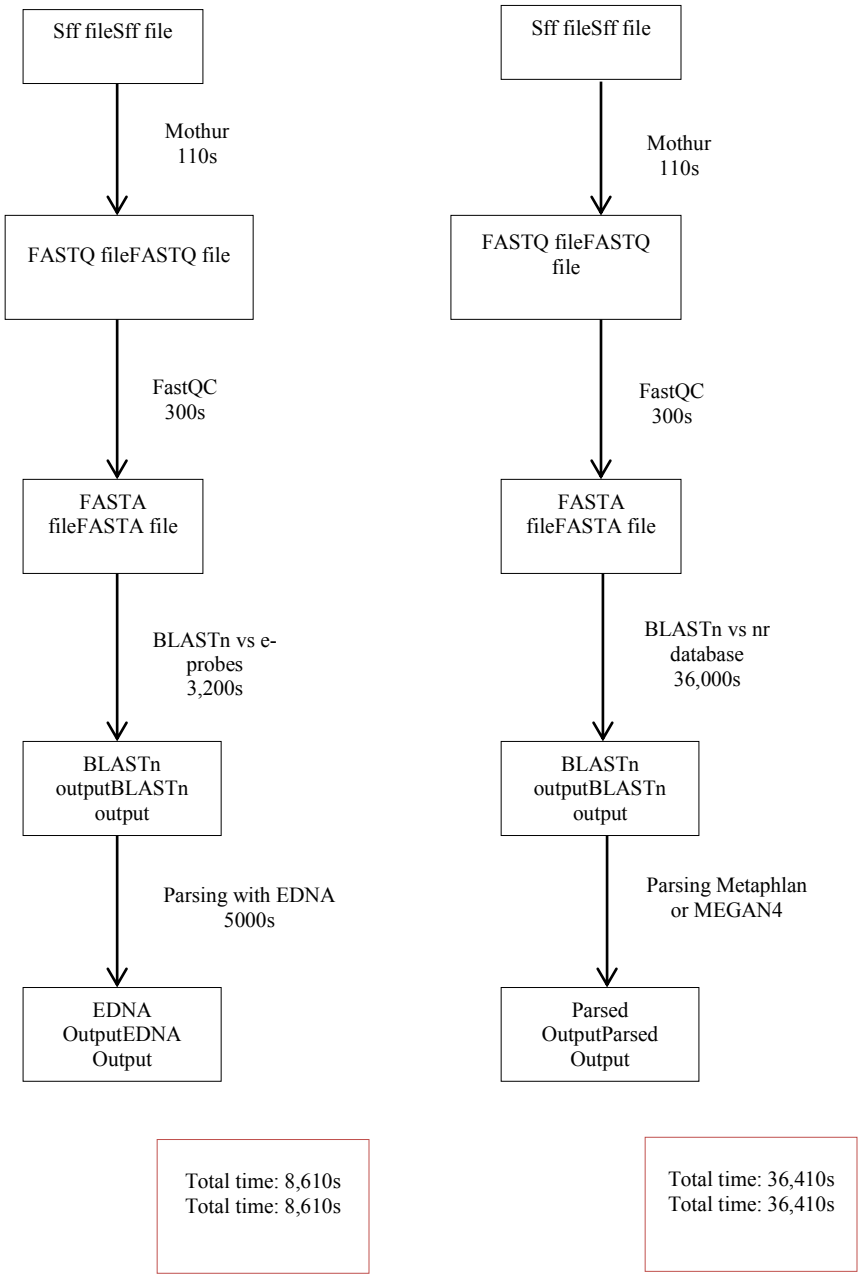

Figure 4-1. EDNA vs. other metagenome analyses programs time consume comparison

This bias in data collection might cause lack of important parts of the pathogen genome while e-probe designs. This reduces consistently the likelihood of having e-probes

representing the total genome of the pathogen. Therefore, the sensitivity of EDNA can be reduced. This issue was not evident during *in silico* validation since the simulated data were generated to resemble 454 runs and not EST data. This increases the likelihood of finding the pathogen in the database.

Due to eukaryotic large genomes, high detection rates with EDNA were expected from SSDs, this is directly related to high titers of the pathogen genome. In previous bioinformatic analyses, EDNA was able to detect reliably eukaryotic plant pathogens in MSSDs with pathogen reads abundances higher than 0.5% (8). In this study, biotroph fungal organisms like *P. graminis* and *P. pachyrhizi* had fewer number of HQM than expected, based on the simulated data. Similar observations were made with the hemibiotroph chromista *P. ramorum.* These biotroph/hemibiotroph pathogens were not found in high ratios (plant/pathogen). Because of their biology, the titer of these pathogens on infected plants can remain low. Nonetheless, the symptoms in the host may be very prominent and may cause severe yield losses and host devastation in natural ecosystems.

Conversely, the saprophytic oomycete plant pathogen *P. ultimum* was found in high titer in infected soybean. It is possible that the noticeable the contrast in terms of number of pathogen reads among biotroph and saprophytic was due to their feeding habits.
This suggests that the pathogen titer on the saprophytic pathogen was high, while the pathogen read abundance in biotrophic and hemibiotrophic pathogens were low to very low. Although a high number of HQM was obtained with their respective e-probes, decoy or shuffled e-probes show zero HQM using the same SSDs.

Decoy and shuffled e-probes were developed to use them as a supplemental negative control for the analyses. Both e-probe types were used in the analysis to select the best negative and positive bioinformatics control (Figure 4-2 and Figure 4-3) and to determine the optimum e-value for detection of eukaryotic pathogens from SSDs. The e-value of $1x10^{-3}$ was not specific enough for this diagnosis since it showed some hits in the negative control. However, lower e-values ($1x10^{-6}$ and $1x10^{-9}$) showed zero hits with both shuffled and decoy e-probes.

Eukaryotic plant pathogens also contain various genes that are commonly shared among eukaryotic organisms; these genes can be either physiological or only genetic keys (9). Such genes may turn the identification process difficult when utilizing a metagenome which is mainly composed of eukaryotic sequences. To reduce biases caused by this phenomenon, e-probe databases can be updated on a regular basis to avoid false positive calls and allow the user to have an updated e-probe database. In addition, the scoring system used by EDNA when it was first proposed as a diagnostic tool relied only on comparisons with decoy e-probes. In this study, the implementation of shuffled e-probes as well as using the near neighbor as negative controls provided higher specificity of EDNA.

Table 4-1. EDNA diagnosis for two Fungal and two Chromista plant pathogens infecting specific hosts. High Quality Matches and False positive HQM limit are presented.

Sample	e-probe length	HQM	FPHQM	Call (C)	Diagnostic
Wheat+*Puccinia graminis*	40	144	1	144	POSITIVE
Wheat+*Puccinia graminis*	60	23	1	23	POSITIVE
Soybean+*Phakopsora pachyrhizi*	40	127	100	1.27	POSITIVE
Soybean+*Phakopsora pachyrhizi*	60	21	100	0.21	NEGATIVE
Rhododendron+*Phytophthora ramorum*	40	50248	25	2009.92	POSITIVE

Rhododendron+*Phytophthora ramorum*	60	4568	25	182.72	POSITIVE
Soybean+*Pythium ultimum*	40	88248	5	1676.6	POSITIVE
Soybean+*Pythium ultimum*	60	8383	5	1676.6	POSITIVE

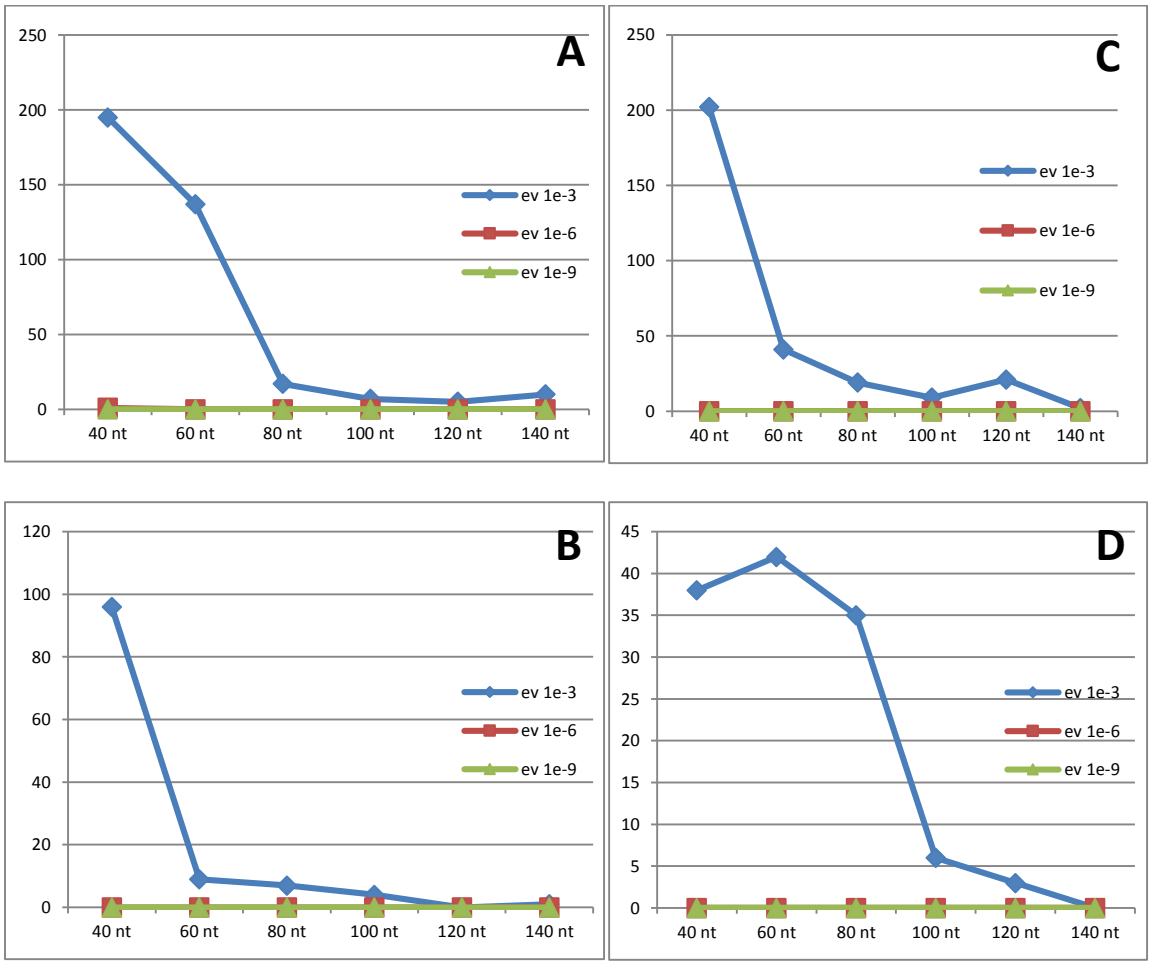

Figure 4-2: Total number of hits with shuffled e-probes using EDNA: A). P. graminis B) P. pachyrhizi, C) P. ramorum, D) P. utimum

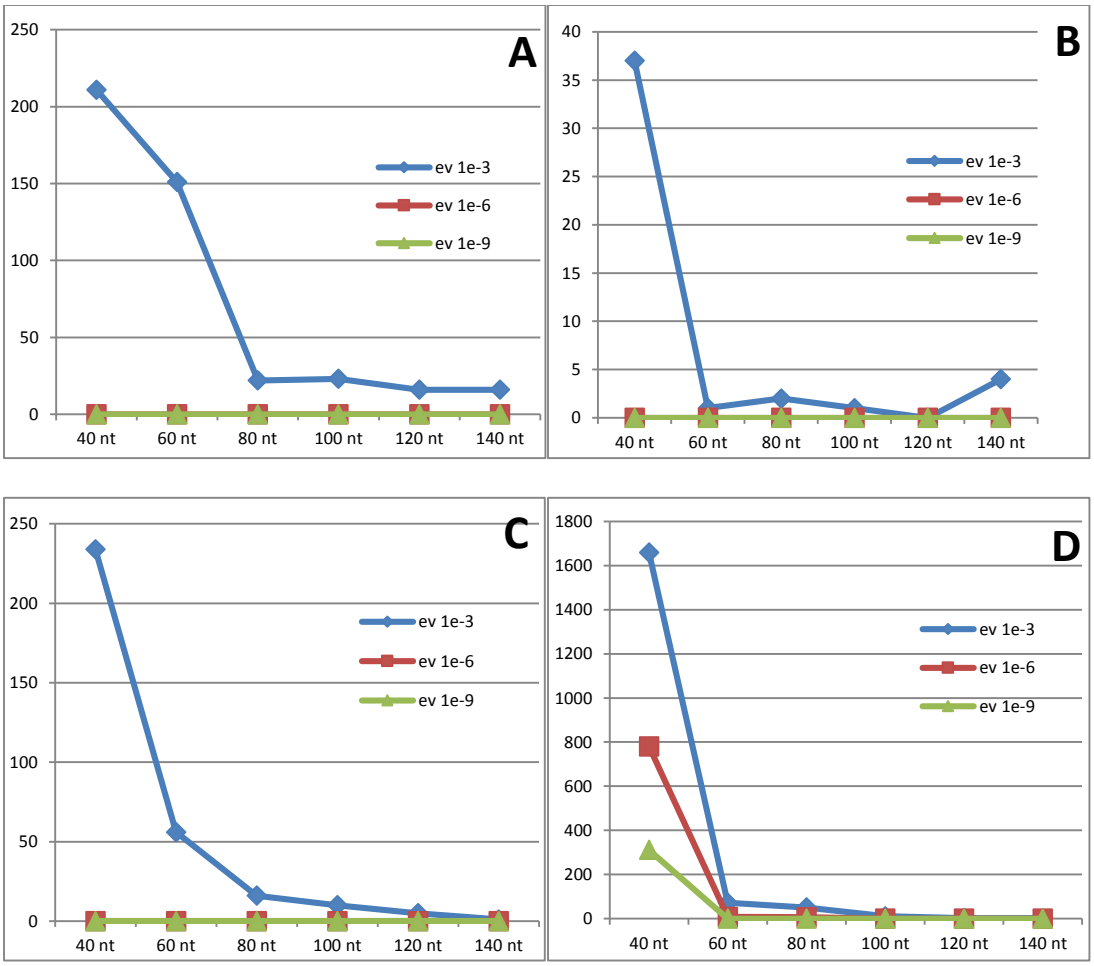

Figure 4-3 Total number of hits with Decoy e-probes using EDNA: A). P. graminis B) P. pachyrhizi, C) P. ramorum, D) P. utimum

Using e-value as a defining parameter for EDNA detection permits the utilization of different sizes of SSDs (10,000; 20,000, 150,000 reads). The e-value takes into account various factors for the alignment, and one of these is the database size to calculate probabilistic data and to produce a score. Highly important was the utilization of different e-values. Many manuals suggest using BLASTn with an e-value of 0.001 which for this study was extremely high. E-values higher than $1x10^{-6}$ produced false positives, which might be critical when trying to detect pathogens in asymptomatic plant tissue.

LITERATURE CITED

1. Agrios, G. 2005. Plant Pathology. Edited by Fifth. Elsevier.
2. Barnes, C. W., and L. J. Szabo. 2007. Detection and Identification of Four Common Rust Pathogens of Cereals and Grasses Using Real-Time Polymerase Chain Reaction. Phytopathology 97(6):717-727 doi:10.1094/phyto-97-6-0717.
3. Barnes, C. W., L. J. Szabo, J. L. Johnson, V. C. Bowersox, and K. S. Harlin. 2006. Detection of Phakopsora pachyrhizi spores in rain using a real-time PCR assay. Phytopathology 96(6):S9-S9.
4. Berbee, M. L., and J. W. Taylor. 1993. Dating the evolutionary radiations of the true fungi. Canadian Journal of Botany 71(8):1114-1127.
5. Bowman, B. H., J. W. Taylor, A. G. Brownlee, J. Lee, S.-D. Lu, and T. White. 1992. Molecular evolution of the fungi: relationship of the Basidiomycetes, Ascomycetes, and Chytridiomycetes. Molecular Biology and Evolution 9(2):285-296.
6. Bruns, T. D., R. Vilgalys, S. M. Barns, D. Gonzalez, D. S. Hibbett, D. J. Lane, L. Simon, S. Stickel, T. M. Szaro, and W. G. Weisburg. 1992. Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. Molecular Phylogenetics and Evolution 1(3):231-241.
7. Cullen, D. W., I. K. Toth, N. Boonham, K. Walsh, I. Barker, and A. K. Lees. 2007. Development and Validation of Conventional and Quantitative Polymerase Chain Reaction Assays for the Detection of Storage Rot Potato Pathogens, Phytophthora erythroseptica, Pythium ultimum and Phoma foveata. Journal of Phytopathology 155(5):309-315 doi:10.1111/j.1439-0434.2007.01233.x.
8. Espindola, A., C. Garzon, and W. Schneider. 2013. A new approach for detecting Fungal and Stramenopile plant pathogens in Next Generation Sequencing Metagenome data utilizing Electronic Probes. Microbiological Methods.
9. Galagan, J., S. Calvo, K. Borkovich, E. Selker, N. Read, D. Jaffe, W. FitzHugh, L. Ma, S. Smirnov, S. Purcell, B. Rehman, T. Elkins, R. Engels, S. Wang, C. Nielsen, J. Butler, M. Endrizzi, D. Qui, P. Ianakiev, D. Bell-Pedersen, M. Nelson, M. Werner-Washburne, C. Selitrennikoff, J. Kinsey, E. Braun, A. Zelter, U. Schulte, G. Kothe, G. Jedd, W. Mewes, C. Staben, E. Marcotte, D. Greenberg, A. Roy, K. Foley, J. Naylor, N. Stange-Thomann, R. Barrett, S. Gnerre, M. Kamal, M. Kamvysselis, E. Mauceli, C. Bielke, S. Rudd, D. Frishman, S. Krystofova, C. Rasmussen, R. Metzenberg, D. Perkins, S. Kroken, C. Cogoni, G. Macino, D. Catcheside, W. Li, R. Pratt, S. Osmani, C. DeSouza, L. Glass, M. Orbach, J. Berglund, R. Voelker, O. Yarden, M. Plamann, S. Seiler, J. Dunlap, A. Radford, R. Aramayo, D. Natvig, L. Alex, G. Mannhaupt, D. Ebbole, M. Freitag, I. Paulsen, M. Sachs, E. Lander, C. Nusbaum, and B. Birren. 2003. The genome sequence of the filamentous fungus Neurospora crassa. Nature 422(6934):859 - 868.
10. Huson, D. H., S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome research 21(9):1552-1560.
11. Kageyama, K., A. Ohyama, and M. Hyakumachi. 1997. Detection of Pythium ultimum using polymerase chain reaction with species-specific primers. Plant Disease 81(10):1155-1160 doi:10.1094/pdis.1997.81.10.1155.

12. Kong, P., C. X. Hong, P. W. Tooley, K. Ivors, M. Garbelotto, and P. A. Richardson. 2004. Rapid identification of Phytophthora ramorum using PCR-SSCP analysis of ribosomal DNA ITS-1. Letters in Applied Microbiology 38(5):433-439 doi:10.1111/j.1472-765X.2004.01510.x.
13. Kurtz, S., A. Phillippy, A. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. Salzberg. 2004. Versatile and open software for comparing large genomes. Genome Biology 5(2):R12.
14. Nazari, K., M. Mafi, A. Yahyaoui, R. Singh, and R. Park. 2009. Detection of wheat stem rust (Puccinia graminis f. sp. tritici) race TTKSK (Ug99) in Iran. Plant Disease 93(3):317-317.
15. Segata, N., L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Meth 9(8):811-814 doi:http://www.nature.com/nmeth/journal/v9/n8/abs/nmeth.2066.html#supplementary-information.
16. Stobbe, A., J. Daniels, A. Espindola, R. Verma, U. Melcher, F. Ochoa-Corona, C. Garzon, J. Fletcher, and W. Schneider. 2013. E-probe Diagnostic Nucleic acid Analysis (EDNA): A theoretical approach for handling of next generation sequencing data for diagnostics. Microbiological Methods.
17. Vettraino, A. M., S. Sukno, A. Vannini, and M. Garbelotto. 2010. Diagnostic sensitivity and specificity of different methods used by two laboratories for the detection of Phytophthora ramorum on multiple natural hosts. Plant Pathology 59(2):289-300 doi:10.1111/j.1365-3059.2009.02209.x.
18. Vijaya Satya, R., N. Zavaljevski, K. Kumar, and J. Reifman. 2008. A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. BMC Bioinformatics 9(1):185.
19. Villavicencio, A., G. B. Fanaro, M. M. Araujo, S. Aquino, P. V. Silva, and J. Mancini-Filho. 2007. Detection of Phakopsora pachyrhizi by polymerase chain reaction (PCR) and use of germination test and DNA comet assay after e-bearn processing in soybean. Radiation Physics and Chemistry 76(11-12):1878-1881 doi:10.1016/j.radphyschem.2007.03.021.

APPENDICES

Apendix 1: Sensitivity and Specificity values calculated for each e-probe length for different fungi and chromista plant pathogens

	40	40	60	60	80	80
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Phytophthora ramorum strain Pr102 (Ram)	66.96%	100.00%	55.07%	100.00%	31.30%	100.00%
Phakopsora pachyrhizi (Pha)	80.35%	99.81%	60.12%	100.00%	25.51%	100.00%
Pythium ultimum DAOM BR144 (Ult)	70.81%	76.98%	59.25%	100.00%	56.36%	100.00%
Puccinia graminis f. sp. tritici CRL 75-36-700-3 (Puc)	100.00%	1.65%	27.06%	71.29%	0.00%	82.58%

Continued appendix 1

100	100	120	120	140	140
Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
0.00%	100.00%	0.00%	100.00%	0.00%	100.00%
0.00%	100.00%	0.00%	100.00%	0.00%	100.00%
49.42%	100.00%	0.00%	100.00%	0.00%	100.00%
0.00%	86.78%	0.00%	94.77%	0.00%	99.90%

Apendix 2: E-probe design parameters and output values

Pathogen	e-probe length	# eprobes	min-match
Pythium ultimum	40	473	13
Phytophthora ramorum	40	769	13
Phakopsora pachyrhizi	40	1699	13
Puccinia graminis	40	3085	13
Phakopsora pachyrhizi	40	242312	14
Phakopsora pachyrhizi	60	20433	14
Puccinia graminis	60	101	13
Pythium ultimum	60	34	13
Phytophthora ramorum	60	0	13
Phakopsora pachyrhizi	60	45	13
Phytophthora ramorum	40	331857	14
Phytophthora ramorum	60	24954	14
Puccinia graminis	40	488093	14
Puccinia graminis	60	42825	14
Pythium ultimum	40	253108	14
Pythium ultimum	60	19294	14

VITA

Andres Sebastian Espindola
Candidate for the Degree of
Master of Science/Arts

Thesis: MASSIVELY PARALLEL SEQUENCING (MPS) AS A DIAGNOSTIC AND FORENSIC ANALYSIS TOOL FOR IMPORTANT FUNGI AND CHROMISTA PLANT PATHOGENS

Major Field: Entomology and Plant Pathology

Biographical:

Andrés Espíndola was born in Ambato, Ecuador and studied his Bachelor's Degree in Biotechnology in the Escuela Politécnica del Ejército (ESPE). Soon after he graduated, he joined ESPE as a Teaching Assistant in the subjects of Microbiology and Phytochemistry. After one year of working as Teaching Assistant, a research assistant position was offered by Oklahoma State University to work under Dr. Carla Garzon in the Soilborne Plant Pathogens Lab and NIMFFAB. While working as RA, the position also permitted to pursuit a Master's degree in the Department of Entomology and Plant Pathology.

Education:

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

Completed the requirements for the Bachelor of Science in Biotechnology Engineering at the Escuela Politécnica del Ejército in June, 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, January 2011 to uly 2013.
Undergraduate Research Assistant. Centro de Investigaciones Cientificas, Escuela Politécnica del Ejército. Quito, Ecuador, June 2008 to June 2009.

Professional Memberships:

American Phytopathological Society
Oomycete Molecular Genetic Network