VIRUS DETECTION IN A METAGENOMIC SEQUENCE DATASET: METHODS AND APPLICATIONS

By

ANTHONY STOBBE

Bachelor of Science in Health Science
Southwestern Oklahoma State University
Weatherford, Oklahoma
2006

Master of Science in Biochemistry
Oklahoma State University
Stillwater, Oklahoma
2012

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2013

VIRUS DETECTION IN A METAGENOMIC SEQUENCE DATASET: METHODS AND APPLICATIONS

Dissertation Approved:

Ulrich Melcher

Dissertation Adviser

Patricia Canaan

Peter Hoyt

Jacqueline Fletcher

William Schneider

ACKNOWLEDGEMENTS

I would like to extend my deepest thanks to my major professor and mentor, Ulrich Melcher. His passion for science and discovery was one reason I joined his laboratory, and remained a source of motivation for me over these past four years. I would also like to thank each of my committee members. Bill Schneider taught me the ins and outs of the scientific community. Thank you for support and confidence boosts when they were needed. Jacqueline Fletcher showed me the importance of professionalism and networking. It is obvious that she cares for the future of all of the students that she interacts with. I would like to thank Patricia Canaan, who lent me my first book on writing scripts, setting me down the path of bioinformatics, as well as Peter Hoyt whose continued efforts to make Oklahoma State a center for bioinformatics in Oklahoma has led to great experiences for the future bioinformaticians.

While my committee made sure I stayed on course academically, my friends and family provided great support when times were tough. They are always ready to celebrate successes, commiserate failures, and provide much needed relaxation. I cannot imagine a better group of colleagues and friends than the ones that I have made during my time at Oklahoma State University. I must give a special thank you to my long suffering wife, who has listened to me when I needed an ear, talked to me when I needed to listen, and has loved me when it was difficult to do so. I love you dearly.

Name: ANTHONY STOBBE

Date of Degree: JULY, 2013

Title of Study: VIRUS DETECTION IN A METAGENOMIC SEQUENCE DATASET: METHODS AND APPLICATIONS

Major Field: BIOCHEMISTRY AND MOLECULAR BIOLOGY

Abstract:

Global trade of plant material has increased the introduction of foreign plant viruses in nations across the world. This has led to the need for increased abilities to detect plant viruses. Surveys assessing the viral biodiversity of environments often use Next Generation Sequencing (NGS) as a tool for virus discovery. NGS has yet to be used as a diagnostic tool due to the computational and time requirements of analyzing NGS data. The purpose of this work is first to show the importance of virus discovery, and second to describe the development and validation of a bioinformatic pipeline designed to detect and identify both DNA and RNA viruses by using pathogen-specific probes to detect virus sequence signals in a metagenomic sample dataset. Finally, Chapter VI shows how the bioinformatic pipeline can be used for the detection of unknown viruses by using general probes.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

CHAPTER III

CHAPTER IV

CHAPTER V

CHAPTER VI

LIST OF FIGURES

Figure Page

CHAPTER III

CHAPTER V

PREFACE

This thesis contains one published peer-reviewed manuscript (Chapter III) and one manuscript which have been accepted for publication (Chapter IV), both of which are reproduced here with the journal's permission, with edits on the request of my committee. I have performed the majority of the experimental work presented here, as well as written and revised the manuscripts mentioned above, although Jon Daniels and Andres Espindola performed a significant portion of the work presented in Chapter IV. For that they have my thanks and acknowledgments.

Stobbe, A. H., Melcher, U., Palmer, M. W., Roossinck, M. J. & Shen, G. (2012). Co-divergence and host-switching in the evolution of tobamoviruses. *J Gen Virol* **93**, 408-418.

Stobbe, A. H., Daniels, J., Espindola, A. S., Verma, R., Melcher, U., Ochoa-Corona, F., Garzon, C., Fletcher, J. & Schneider, W. (2013). E-probe Diagnostic Nucleic acid Analysis (EDNA): A theoretical approach for handling of next generation sequencing data for diagnostics. *J. Microbiol. Methods* (in press).

CHAPTER I

INTRODUCTION

Metagenomics is the study of genomic content of a community of organisms. It has been enabled by Next Generation Sequencing (NGS), in which hundreds of thousands of short sequences can be obtained from a given sample containing a mixture of organisms. Since obtaining a pure culture of a microbe is no longer necessary, sequence information has been found for several previously uncharacterized organisms, including viruses (Breitbart *et al.*, 2002; Edwards & Rohwer, 2005; Kristensen *et al.*, 2010; Minot *et al.*, 2011; Rodriguez-Brito *et al.*, 2010; Rosario *et al.*, 2009; Suttle, 2005; Wren *et al.*, 2006). Thus, metagenomics is a useful tool to characterize an environment and the organisms found within that environment.

NGS studies have illuminated the microbial ecology of specific environments, including ocean water, bilge water, grassland preserves, lake water, soil, human gut flora, and many others (Adams *et al.*, 2009; Breitbart *et al.*, 2003; Breitbart *et al.*, 2002; Cox-Foster *et al.*, 2007; Daniel, 2005; Rosario *et al.*, 2009; Wren *et al.*, 2006).

These genetic “fishing expeditions” have addressed many of the basic questions about the ecology of the environment studied, such as what organisms are present, at what abundance, and where they are located. Many of these studies revealed the presence of many previously uncharacterized viruses (Minot *et al.*, 2011; Rodriguez-Brito *et al.*, 2010; Suttle, 2005; Wren *et al.*, 2006).

These surveys also bring to light questions and ideas about the role of viruses in their environments. Some suggest that marine viruses simply serve as genetic reservoirs for the community (Kristensen *et al.*, 2010), others remind us that not all viruses are antagonistic, but can act as mutualists (Márquez *et al.*, 2007). A large number of viruses can be found in asymptomatic plants (Muthukumar *et al.*, 2009; Ooi *et al.*, 1997; Robertson, 2005). Does this mean that these viruses are simply associated with their hosts with little to no interaction between them, or are they a source for emergent diseases?

Viruses have been known to jump from one host to another, often causing different, more severe symptoms. The most famous case of a viral host jump is the Chimp-Human jump of *Human immunodeficiency virus* (Gao *et al.*, 1999). This phenomenon is not limited to animal viruses, as several plant viruses have been known to infect a wide variety of hosts, causing different symptoms (Cropley, 1968; Holmes, 1946). This fact, taken with the movement of viruses, which has been expedited due to the mass global trade of plant material and food stuffs, has created a scenario in which viruses are exposed to a plethora of hosts, any one of which they may infect causing vast economic damage. In the United States alone, 65% of loss of crops is due to exotic or foreign pathogens (Pimentel, 1993). While the United States exports more food than it imports (USBC, 2012) and is not in eminent danger of food shortages, developing nations are threatened by this problem (Godfray *et al.*, 2010). The majority of the world’s population growth is centered in these developing nations (Bongaarts, 2009). The increased detection of viruses and curbing of their movement could lead to a decrease in both crop loss and hunger in the world.

In 2005 the Plant Virus Ecology and Biodiversity project began to explore the Tallgrass Prairie Preserve in northeastern Oklahoma for its virus biodiversity. In this survey, hundreds of distinct virus strains were discovered, the majority of which had not been described previously (Muthukumar *et al.*, 2009). One of these strains shared high sequence similarity with the *Tobamaovirus* family of single-stranded RNA (Stobbe *et al.*, 2012), which includes the tobamovirus *Passionfruit mosaic virus.* Tobamoviruses are relatively ubiquitous single stranded positive stranded RNA viruses widely studied in academic and agriculture settings, and the tobamovirus found in the Tallgrass Prairie Preserve is but one of many tobamoviruses that have been discovered and characterized over the past decade (Adkins *et al.*, 2003; Adkins *et al.*, 2007; Min *et al.*, 2006; Song *et al.*, 2006; Song & Ryu, 2011).

Codivergence is a process of evolution in which two associated organisms speciate concurrently. Codivergence differs from coevolution in that selective pressures that the organisms exert on each other do not contribute to the speciation. Several virus families, such as the mastreviruses (Wu *et al.*, 2008), may have codiverged alongside their host plants. The tobamovirus family members may have codiverged with their respective host plants based on similarities between the virus and host phylogenetic trees (Gibbs, 1980; 1999; Gibbs *et al.*, 2010; Lartey *et al.*, 1996). The tobamoviruses have been separated into 3 subgroups based on their phylogenetic tree (Lartey *et al.*, 1996). Members of each subgroup share susceptible host orders. For example, members of subgroup 1 infect astrids, while those that are in subgroup 2 infect rosids. This congruency of trees suggests codivergence, and therefore would place the origin of tobamoviruses at the same time as the astrid/rosid split, 100 million years ago. The substitution rate of the tobamoviruses, assuming codivergence, has been calculated to be on the order of 10^{-7} to 10^{-9} .

By sequencing isolates of recent serial passages, as well as isolates from decades old preserved samples, a substitution rate of 10^{-4} was calculated (Pagán *et al.*, 2010). This substitution

rate agrees with the rate of substitution of other RNA viruses, and would not allow the possibility of codivergence, but it is important to note that this study relied upon the GenBank collection of tobamovirus sequences, which is biased towards viruses of agricultural significance. The possibility exists that agricultural crops, with their long history of human selection, represents a different environment than what might be expected in a natural host/virus relationship. This discrepancy in substitution rates merits further study to determine the evolutionary history of this family of viruses. Since the last examination of codivergence in the tobamovirus family, several new species have been discovered and characterized. In addition, new phylogenetic techniques, such as those based on Bayesian models, have been deployed.

Presented in Chapter III of this thesis is a re-evaluation of the codivergence theory within the tobamoviruses. Tobamovirus genomes were collected from GenBank and aligned, and an analysis using the BEAST software package was performed. Assuming the split between supergroups 1 and 2 coincided with the astrid/rosid split, a substitution rate of 10^{-8} substitutions/site/year was calculated. Statistical testing using the BaTS software package shows that the virus' hosts are clustered together on the phylogenetic tree. These two pieces of information, the substitution rate and the grouping of hosts within clades, gives credence to the codivergence hypothesis (Stobbe *et al.*, 2012).

This re-evaluation would not have been possible without the discovery of several tobamovirus species over the past decade. With the increase in use of NGS and surveys in plant biodiversity, several new virus species have been and will continue to be found. A major bottleneck in these surveys is the bioinformatic processing of the sequencing information. With the decrease in the cost of conventional and next generation sequencing, there is an increase in sequencing information. This information is stored in curated databases such as GenBank, which are, in turn, used in the analysis of sequencing data. The large amount of time and computational power needed to search these curated databases is a difficulty in virus discovery and makes this

use of NGS as a diagnostic tool unreasonable. Further, this is a problem that will continue to exacerbate itself, as the exponential increases in sequence data per NGS run continue in step with the exponential growth of the reference database, GenBank.

Screening for and diagnosis of plant disease are extremely important for agricultural industries. As there are very few treatment options for plant disease, the best management method includes early detection and destruction of the infected plant material. Current diagnostic tools, such as qPCR or ELISA, are difficult to multiplex (Schaad *et al.*, 2003). While efforts to create a microarray diagnostic approach have been made, the use of NGS would provide the diagnostic community the ability to detect all microbial organisms within a given plant sample, including viruses, bacteria and eukaryotic pathogens.

Chapters IV and V describe the theory and validation of the E-probe Diagnostic Nucleic acid Assay (EDNA). The step requiring the most time in identifying sequence reads, due to the sheer size of these databases, is querying curated databases. For the EDNA pipeline short pathogen specific sequences (termed e-probes) are designed. These e-probes are analogous to PCR or microarray probes; with the major difference being that e-probes are not physical molecules, but are instead a simple string of characters. Continuing the analogy with microarray, the e-probes will match pathogen sequences within the sequencing data, much like microarray probes will anneal to their target sequences. In addition to the species level e-probes, e-probes were also designed to detect and differentiate *Plum pox virus* (PPV) strains. To validate the pipeline, total nucleic acid samples of plants infected with *Bean golden mosaic virus* (BGMV) or, separately, 5 isolates of PPV were obtained and sequenced using the Roche 454 Jr. platform. These viruses were chosen for two reasons: first, to test the pipelines ability to detect both RNA and DNA viruses, and second, because they are both viruses of economic interest. The EDNA pipeline performed as well as the “traditional method” in terms of detection, and surpassed the ”traditional method” in terms of speed (2100 times faster).

One drawback to the EDNA pipeline is that one must know the genome sequence of the target pathogens before assaying for the pathogens. This makes searching for new uncharacterized species and strains difficult. Chapter V of this work describes our attempts to use a more generalized probe set, designed originally for a microarray platform, to screen NGS datasets of potentially infected plant material.

In conclusion, this work describes both the need for, and merit of, new virus discovery. To accommodate this need, a bioinformatic pipeline (EDNA) was created to assist in parsing the large amount of sequencing data that is generated during biodiversity surveys. EDNA has the ability to identify target pathogens on a family, species, and strain level. This feature has applications not only in assisting researchers in virus discovery, but also in detection of known select agents, giving diagnosticians a new powerful tool for screening incoming plant material.

REFERENCES

Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M. & Boonham, N. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Mol. Plant. Pathol.* **10**, 537-545.

Adkins, S., Kamenova, I., Achor, D. & Lewandowski, D. J. (2003). Biological and Molecular Characterization of a Novel Tobamovirus with a Unique Host Range. *Plant Dis* **87**, 1190-1196.

Adkins, S., Kamenova, I., Rosskopf, E. N. & Lewandowski, D. J. (2007). Identification and Characterization of a Novel Tobamovirus from Tropical Soda Apple in Florida. *Plant Dis* **91**, 287-293.

Bongaarts, J. (2009). Human population growth and the demographic transition. *Philos. Trans. R. Soc. Lond., B* **364**, 2985-2990.

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

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. & Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14250-14255.

Cox-Foster, D. L., Conlan, S., Holmes, E. C., Palacios, G., Evans, J. D., Moran, N. A., Quan, P.-L., Briese, T., Hornig, M., Geiser, D. M., Martinson, V., vanEngelsdorp, D., Kalkstein, A. L., Drysdale, A., Hui, J., Zhai, J., Cui, L., Hutchison, S. K., Simons, J. F., Egholm, M., Pettis, J. S. & Lipkin, W. I. (2007). A Metagenomic Survey of Microbes in Honey Bee Colony Collapse Disorder. *Science* **318**, 283-287.

Cropley, R. (1968). The identification of plum pox (Sharka) virus in England. *Plant Pathol* **17**, 66-70.

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

Edwards, R. A. & Rohwer, F. (2005). Viral metagenomics. *Nat Rev Microbiol* **3**, 504-510.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M. & Shaw, G. M. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397**, 436-441.

Gibbs, A. (1980). How ancient are the Tobamoviruses. *Intervirology* **14**, 101-108.

Gibbs, A. (1999). Evolution and origins of tobamoviruses. *Philos Trans R Soc Lond Ser B-Biol Sci* **354**, 593-602.

Gibbs, A. J., Fargette, D., Garcia-Arenal, F. & Gibbs, M. J. (2010). Time - the emerging dimension of plant virus studies. *J Gen Virol* **91**, 13-22.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M. & Toulmin, C. (2010). Food security: the challenge of feeding 9 billion people. *Science* **327**, 812-818.

Holmes, F. O. (1946). A comparison of the experimental host ranges of Tobacco-etch and Tobacco-mosaic viruses. *Phytopathology* **36**, 643-659.

Kristensen, D. M., Mushegian, A. R., Dolja, V. V. & Koonin, E. V. (2010). New dimensions of the virus world discovered through metagenomics. *Trends microbiol* **18**, 11-19.

Lartey, R. T., Voss, T. C. & Melcher, U. (1996). Tobamovirus evolution: Gene overlaps, recombination, and taxonomic implications. *Mol Biol Evol* **13**, 1327-1338.

Márquez, L. M., Redman, R. S., Rodriguez, R. J. & Roossinck, M. J. (2007). A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. *Science* **315**, 513-515.

Min, B., Chung, B., Kim, M., Ha, J., Lee, B. & Ryu, K. (2006). Cactus mild mottle virus is a new cactus-infecting tobamovirus. *Arch Virol* **151**, 13-21.

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

Muthukumar, V., Melcher, U., Pierce, M., Wiley, G. B., Roe, B. A., Palmer, M. W., Thapa, V., Ali, A. & Ding, T. (2009). Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. *Virus Res* **141**, 169-173.

Ooi, K., Ohshita, S., Ishii, I. & Yahara, T. (1997). Molecular phylogeny of geminivirus infecting wild plants in Japan. *J Plant Res* **110**, 247-257.

Pagán, I., Firth, C. & Holmes, E. (2010). Phylogenetic Analysis Reveals Rapid Evolutionary Dynamics in the Plant RNA Virus Genus Tobamovirus. *J Mol Evol* **71**, 298-307.

Pimentel, D. (1993). *Habitat factors in new pest invasions*: New York: John Wiley & Sons.

Robertson, N. (2005). A newly described plant disease complex involving two distinct viruses in a native Alaskan lily, Streptopus amplexifolius. *Botany* **83**, 1257-1267.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., Desnues, C., Dinsdale, E. & Edwards, R. (2010). Viral and microbial community dynamics in four aquatic environments. *The ISME journal* **4**, 739-751.

Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009). Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol* **11**, 2806-2820.

Schaad, N. W., Frederick, R. D., Shaw, J., Schneider, W. L., Hickson, R., Petrillo, M. D. & Luster, D. G. (2003). Advances in Molecular-based Diagnostics in Metting Crop Biosecurity and Phytosanitary Issues. *Annu Rev Phytopathol* **41**, 305-324.

Song, Y., Min, B., Hong, J., Rhie, M., Kim, M. & Ryu, K. (2006). Molecular evidence supporting the confirmation of Maracuja mosaic virus as a species of the genus Tobamovirus and production of an infectious cDNA transcript. *Arch Virol* **151**, 2337-2348.

Song, Y. S. & Ryu, K. H. (2011). The complete genome sequence and genome structure of passion fruit mosaic virus. *Arch Virol* **156**, 1093-1095.

Stobbe, A. H., Melcher, U., Palmer, M. W., Roossinck, M. J. & Shen, G. (2012). Co-divergence and host-switching in the evolution of tobamoviruses. *J Gen Virol* **93**, 408-418.

Suttle, C. A. (2005). Viruses in the sea. *Nature* **437**, 356-361.

USBC (2012). Statistical Abstract of the United States 2012. Edited by USBC. Washington D.C.: U.S. Government Printing Office.

Wren, J. D., Roossinck, M. J., Nelson, R. S., Scheets, K., Palmer, M. W. & Melcher, U. (2006). Plant virus biodiversity and ecology. *PLoS biology* **4**, e80.

Wu, B. L., Melcher, U., Guo, X. Y., Wang, X. F., Fan, L. J. & Zhou, G. H. (2008). Assessment of codivergence of Mastreviruses with their plant hosts. *BMC Evol Biol* **8**.

CHAPTER II

REVIEW OF THE LITERATURE

Sanger sequencing

Since Fredrick Sanger published his method of sequencing by nucleic acid chain termination (Sanger *et al.*, 1977), the ability to obtain the genetic sequence of an organism has had many profound effects in all areas of science (Schuster, 2008). Terminating the chain reaction of a DNA polymerase with di-deoxynucleotides created a simple mechanism for stopping and labeling DNA molecules at every position in a given fragment. This, in combination with improved electrophoresis techniques, generated the ability to completely sequence the genetic material of an organism. Sanger sequencing was the driving force behind the complete sequencing of the human genome (Collins *et al.*, 2004; Venter *et al.*, 2001). There have been several improvements on the traditional Sanger sequencing, including the use of capillary electrophoresis, which increased the number of sequences processed (Trainor, 1990).

Over time the need for increased sequencing led to the development of new sequencing technologies. These technologies, termed either Massively Parallel Sequencing (MPS), High Throughput Sequencing (HTS) or Next Generation Sequencing (NGS), gave researchers the ability to sequence hundreds of thousands of sequencing reads (short sequence results ranging from 35 bp to over 1 kbp) in a single procedure. These new technologies opened many new and exciting fields of study, such as metagenomics, as well as giving new techniques and methods for traditional genetics and gene expression studies (Mardis, 2008). Several platforms have been made available since 2005, four of which are explained below.

Roche 454 pyrosequencing

The increased need for sequencing technologies led to the release of what is now known as Roche 454 pyrosequencer in 2005. This technology has utilized two important features which allow for the sequencing of several thousand sequence reads; emulsion polymerase chain reaction (emPCR) (Nakano *et al.*, 2003) and pyrosequencing (Elahi & Ronaghi, 2004). To obtain copious amounts of unique strands of DNA to be sequenced, the nucleic acid is first sheered to a size of 300 to 500 bp and then ligated to an adapter made up of a primer region and an adapter region. The primer region is paired to an oligonucleotide, which is covalently bond to a polystyrene bead. The ligated DNA, PCR reagents, sequencing beads, and oil are then emulsified, creating several thousand microreaction vesicles. The ligated DNA is added in such a concentration that approximately one piece of nucleic acid is present in each microreaction vesicle. PCR is then performed as usual, resulting in each bead containing numerous copies of a unique, covalently linked piece of nucleic acid.

The sequencing beads are then washed and packed into a picotiter plate that consists of several thousand small wells, each packed with a single sequencing bead, beads linked with a luciferase enzyme, a sulfurylase enzyme, and reagents for DNA synthesis. Once in the sequencing instrument, each of the four types of dNTP is washed sequentially and repeatedly over the plate. If the dNTP is incorporated into the sequence a pyrophosphate is released and is

consequently used by the sulfurylase to generate ATP, which is used by the luciferase, generating photons. The plate is washed and a different dNTP is added. A photo detector above the plate reads the photons and a base call is made for that sequencing bead. This process is continued for a number of cycles, resulting in the sequence of the nucleic acid of each sequencing bead.

Ion Torrent

An improvement to the nucleotide addition detection was seen with the Ion Torrent (Rothberg *et al.*, 2011). The sample preparation and amplification are similar to that of the Roche 454 platform. The Ion Torrent diverges from the Roche 454 by instead of generating photons with each base addition, each microwell is a hypersensitive ion sensor, and as the base is added to the DNA strand, a hydrogen ion is released and detected. This method of detection reduces the overall cost of the sequencing procedure by reducing the amount of reagents needed to generate a sequencing signal.

Illumina Solexa

Another NGS platform is Illumina Solexa (Bentley *et al.*, 2008). The nucleic acid preparation is similar to that of the Roche 454 platform, sheering of the DNA, followed by ligation of an adapter. However, the DNA molecules are amplified on a surface linked to oligonucleotides, instead of on sequencing beads. Amplification of these DNA molecules forms DNA clusters on the surface with enough density for the sequencing signal to be read by a fluorometer. The sequencing signal is generated by washing the plate with all four nucleotides, each of which is labeled with a unique fluorophore at the 3' end of the nucleotide, effectively halting synthesis. A fluorescent reading is taken and the base for each DNA cluster is called, after which the fluorescent group is chemically removed. This process is repeated, generating sequence reads of up to 30 to 250 bp (Quail *et al.*, 2012). The sequencing process is then repeated for the opposite strand, giving two sequence reads per DNA cluster. These two reads are paired together as they are from the same molecule. These paired end reads are useful in downstream assembly and for SNP typing by offering positional information from the two reads.

Pacific Biosciences

Most platforms require an amplification step, in which unique DNA molecules are amplified in close proximity to one another in order for the sequencing signal to be strong enough for detection. Pacific Biosciences uses single molecule real time (SMRT) sequencing, which does not use amplified DNA but rather the original DNA molecule (Eid *et al.*, 2009). The DNA is placed on a surface covered in small wells called zero-mode waveguides (ZMWs) (Levene *et al.*, 2003), that are designed to detect only light at the bottom of the well. At the bottom of each zero-wave guide is a single polymerase enzyme that has been engineered to accept fluorescently labeled nucleotides while retaining the properties of a wild-type polymerase (Korlach *et al.*, 2008). The surface is washed with a mixture of uniquely fluorophore labeled dNTPs, and as the bases are incorporated into the sequence the fluorophore is detected.

Sequencing Errors

Because each platform uses different techniques, each is prone to different forms of error. In using NGS as a diagnostic tool, a sequencing error made either by miscalling bases, or bias in sequencing can lead to a false negative diagnostic call. Sequencing bias is preferential amplification of specific nucleic acid molecules; for example, a GC bias indicates that GC rich nucleic acids are amplified preferentially and then sequenced. Below, several different forms of error are discussed.

Homo-oligomers

In the Roche 454 and Ion Torrent platforms, the detection of homo-oligomers is difficult (Huse *et al.*, 2007) because of the incorporation of multiple bases in a single nucleotide cycle, which increases the amount of sequencing signal (either light or ion) proportionally up to 3-4 times before diminishing. Given this method of homo-oligomer base calling, mistakes can be made in the number of bases added. Improvements have been made in the chemistry by coating

the sequencing wells with metal to promote more signal (Huse *et al.*, 2007; Voelkerding *et al.*, 2009).

Base Calling

A miscall of bases can be attributed to two different sources; the incorporation of the wrong nucleic acid base into the synthesized strand, and confusion or difficulty in reading the sequencing signal. The former is an issue only for SMRT sequencing, as other sequencing platforms use clonal DNA clusters, diluting any signal of a single misincorperation event. The major source of error for SMRT sequencers is the incorporation time is too short to be detected reliably (Eid *et al.*, 2009). In the Illumina platform, the major causes of base miscalling are substitution errors that cause the sequence to fall out of phase. Efforts have been made to filter the noise using machine learning (Erlich & Mitra, 2008; Mardis, 2013). The ability to detect and correctly attribute the sequencing signal to the incorporation of a base is entirely dependent on the instrument. This is only an issue for those sequencing reads which are in low abundance, as the major method of determining sequence quality is the consensus sequence made up of many individual sequencing reads.

Sequencing Bias

The ligation and amplification steps of many of these platforms permit bias to be introduced. Illumina platforms have been shown to have a GC bias in the adapter ligation step leading to low or no coverage of AT rich regions of a genome (Aird *et al.*, 2011). This problem can cause difficulties in genome sequencing and metagenomic as well as RNA expression experiments. The use of an alternate ligase has diminished this bias (Quail *et al.*, 2008). Generally, sequencing bias is something to be avoided, but in cases where target pathogens are GC rich, using the sequencing bias to increase sensitivity is a possibility.

Read length

Due to the chemistries involved in each of the sequencing platforms, the average read lengths differ greatly from platform to platform. These average read lengths range from 150 bp

reads for the Illumina platform to 1.5 kbp for the Pacific Biosciences platform, with the Ion torrent and Roche 454 platforms at 200 bp and 400 bp respectively. While not technically a sequencing error, longer reads would give a higher signal for detecting pathogens, which would be beneficial for NGS to be used in a diagnostic method.

Assembly

Before meaningful analysis can be performed on NGS data, the sequence reads are typically assembled into larger contiguous sequences (contigs). Since the late nineties, dozens of assembly programs have been developed (Batzoglou *et al.*, 2002; Chaisson & Pevzner, 2008; Gnerre *et al.*, 2011; Myers *et al.*, 2000; Simpson *et al.*, 2009; Zerbino & Birney, 2008). While de novo assembly assists (and is necessary in the vast majority of cases) with the identification of organisms, it is both computationally and time intensive though efforts are being made to reduce the time needed to assemble sequencing data (Pop *et al.*, 2004). The time requirement for assembly limits the use of NGS as a diagnostic tool as the speed of diagnostic assays is almost as important as the accuracy or sensitivity. The assembly of viral genomes is not complicated when compared to those of prokaryotic and eukaryotic organisms, to the higher genome complexity and repetitive sequences. The ability to find a signal within the raw sequence reads which can lead to the diagnosis of disease inducing pathogens is the subject of a major portion of this thesis.

Newbler assembler (de novo)

There are several freely available assembly programs, Newbler being one available from Roche Life Sciences (Chaisson & Pevzner, 2008). Newbler matches reads by using a default of 16 base seeds, and extending to find a maximal match. Once large contigs are found, a contig map is generated, linking contigs that share reads or paired end reads. These maps may link several contigs together (i.e. contig A and B both map to the 5' end of contig C) and require the map to be collapsed to its simplest form. While assembly of a draft genome is useful in identifying new and novel microbes, it is a bottleneck within a bioinformatic pipeline, slowing the analysis and thus delaying any diagnostic call to be made.

Mapping assembly

Another form of assembly is to map the sequencing reads or contigs onto a previously assembled genome. The difficulty here is that one must know which genome (or genomes) is present before any mapping can occur. It would be beneficial to have this type of assembly done after a diagnostic call is made for two reasons. First, confirmation of the diagnostic call, and second, to continue with any downstream analysis needed, such as SNP typing or phylogenetic analysis for forensic purposes (Iqbal *et al.*, 2012).

Metagenomics

With new sequencing technologies, questions can now be answered that were impossible (or impractical) to answer previously. Metagenomics, the study of the genomic makeup of select ecosystems, has been used as a tool for determining the biodiversity (Breitbart *et al.*, 2002; Daniel, 2005; Gill *et al.*, 2006; Harrison, 1981; Wren *et al.*, 2006), gene expression (Frias-Lopez *et al.*, 2008; Uchiyama *et al.*, 2004), and gene interaction within a given environment (Ezenwa *et al.*, 2006; Harrison, 1981; Schwartz *et al.*, 2012; Singh *et al.*, 2004; Webster *et al.*, 2007). Looking at not only a single organism, but all of the organisms within a given sample gives insights into how these microbes interact with their neighbors (Singh *et al.*, 2004). Gene interactions among and between organisms are important factors in ecosystems. In many cases, these microbes cannot be cultured, making these species difficult to identify, let alone characterize (Schloss & Handelsman, 2005a). For determining biodiversity of prokaryotic microbes, the 16S RNA may be targeted (Ward *et al.*, 1990), while internal transcribed spacer (ITS) regions may be used for eukaryotic microbes, such as fungi (Gardes & Bruns, 1993; Gräser *et al.*, 1999; Nilsson *et al.*, 2009). This practice has become common and several resources are

available to assist researchers with microbial surveys (Ashelford *et al.*, 2002; DeSantis *et al.*, 2006; Schloss & Handelsman, 2005b; Schloss *et al.*, 2009).

Viruses are genetically diverse compared to other cellular organisms in nature (Baltimore, 1971; Koonin, 1991; Koonin *et al.*, 2006) and no single gene can be targeted for such surveys. Instead one must either purify virus-like particles (Melcher *et al.*, 2008), or be selective in the nucleic acid extraction (such as purifying double-stranded RNA (Roossinck *et al.*, 2010), targeting small RNA resulting from plant infecting viruses, or rolling circle amplification (RCA) (Dean *et al.*, 2001)). Efforts to catalog virus biodiversity have taken multiple approaches, often relying on more than one extraction procedure (Breitbart *et al.*, 2002; Wren *et al.*, 2006). Each method has its own biases; for example, purifying virus-like particles is arguably the best method to obtain all types of viruses within a given sample, but even this method fails to capture viral genomes that have either failed to or have not assembled into a virion. Small RNA sequencing can miss viruses that are particularly adept at suppression of silencing, and not all viruses produce viable levels of dsRNA.

Virus discovery

Since the discovery of “filterable agents” in 1892 (Beijerinck, 1898), viruses have been an intriguing class of microbes which has been found infecting nearly every form of life on this planet. In every type of environment where life can be found, there also seems to be one or more viruses in the ecosystem. Our knowledge of what viruses exist is skewed to those that affect us medically, socially, or economically, but as one begins to look beyond those agents that cause disease, one can find that there are a plethora of unknown viruses in the world (Suttle, 2005). Metagenomic surveys have cataloged the virus biodiversity of several environments, including marine samples from multiple oceans, human gut flora, bee hives, bilge water, and plant samples from various regions (Angly *et al.*, 2006; Breitbart *et al.*, 2003; Cox-Foster *et al.*, 2007; Daniel,

2005; Delwart, 2007; Gill *et al.*, 2006; Kristensen *et al.*, 2010; Minot *et al.*, 2011; Wren *et al.*, 2006). In each of these surveys, novel viruses are found. It is very likely that these viromes are an integral part of the ecosystems ofwhich they are members, but before questions of how these populations interact with their environment it is necessary to know what viruses are present.

It is difficult to use presence of symptoms as the major method of cataloging the virus biodiversity of a given area. Methods used for detecting and identifying viruses can be generalized into two categories: nucleic-acid based and protein based. Protein based assays, such as enzyme-linked immunosorbent assay (ELISA), typically use antibodies specific for a viral protein (Hamblin *et al.*, 1986). Because of this, it is difficult to multiplex protein based assays, which in turn leads to an increase in the amount of reagents as well as time needed to make the identification. Family specific antibodies have been made in the case of viruses (Hammond, 1991). Attempts to use proteomics to detect and identify microbes have been made (Cooper *et al.*, 2003), but the technology was used primarily to study the physiology of viruses (Alfonso *et al.*, 2004; Baas *et al.*, 2006; Zheng *et al.*, 2008).

Nucleic acid based assays

Differential centrifugation

One method of purifying viral nucleic acid from host or other nucleic acid is using either a cesium gradient to isolate virus particles (Yamamoto *et al.*, 1970), or ultra-centrifugation to obtain a pellet containing virus particles (Scott, 1963), after which the nucleic acid in the virus fraction is amplified (Melcher *et al.*, 2008). This method does not discriminate or select for any particular type of virus, though viral nucleic acid that has not been packaged into a virion will be missed, as will viroid nucleic acid.

Polymerase Chain Reactions

As the name implies, nucleic acid based tests, such as microarrays or PCR, rely on the presence of pathogen (virus) specific nucleic acid sequences to determine the presence of the organism (Thomson & Dietzgen, 1995). Specific PCR assays are difficult to multiplex (Henegariu *et al.*, 1997) and are used mostly for strain differentiation (Cebula *et al.*, 1995; Mahony *et al.*, 2007; Oliveira & de Lencastre, 2002; Paton & Paton, 1998). In diagnostic settings quantitative real time PCR (q-PCR) assays can be used to test quickly for the presence of virus nucleic acid (Kimura *et al.*, 1999; Kuypers *et al.*, 2006; Schaad & Frederick, 2002; Spackman *et al.*, 2002). q-PCR reactions use either fluorescent dyes or probes to track the amplification of nucleic acids cycle by cycle, and as the fluoresce of a sample rises above a set threshold (such as 10 standard deviations above the background fluoresce) the sample is called positive (Schaad & Frederick, 2002). q-PCRs have been shown to detect nucleic acid at picogram/µl concentrations (Arif & Ochoa-Corona, 2012). q-PCRs also have the ability to quickly detect RNA viruses by including a reverse transpiration step in the cycle program (Spackman *et al.*, 2002). In terms of diagnostics, PCR assays rely on the sequence similarity of the primers and probes to determine presence of a pathogen. In most cases, the numbers of sequences tested for are three or less.

Whole Genome/Transcriptome Amplification (WGA/WTA)

In many cases, a sample's extracted nucleic acid is at a concentration too low to be used in NGS. By using random primers for amplifying either DNA or transcribing RNA to cDNA, one is able to generate large amounts of nucleic acid (Cheung & Nelson, 1996). Adding another round of amplification adds more bias, as random primers could bind preferentially (Mardis, 2013). In addition to sequencing bias, additional rounds of amplification add even more errors due to error prone polymerase. The sequencing errors due to amplification can be mitigated by the use of high fidelity polymerase (Lahr & Katz, 2009). By itself, WGA/WTA does not differentiate virus from host nucleic acid, it is merely a method of increasing the amount of the total nucleic acid extracted from a sample.

Double stranded RNA preparation

A large pecentage of plant viruses are either double stranded RNA (dsRNA) or single stranded (both plus and minus strands in their replication) (Roossinck *et al.*, 2010). Three decades ago, this fact was capitalized upon by selecting for and purifying the dsRNA from a plant sample (Dodds *et al.*, 1984). The process is essentially a phenol/chloroform extraction with a cellulose column, but centrifuge at low speeds for a short time (i.e. 200 rpm for 30 seconds). This procedure has been modified to accept tissues, from different plant species, that contain substances able to disrupt the extraction process, such as phenolics. The purified dsRNA can then be used as a template in a reverse transcription reaction to obtain cDNA for sequencing.

Rolling Circle Amplification

Much as dsRNA extraction selects for RNA viruses, rolling circle amplification (RCA) selects for circular DNA viruses, such as the geminiviruses. This method of amplification utilizes the PhiX29 polymerase, which displaces the 5' end of DNA (Dean *et al.*, 2001). The displaced strand then becomes the template for another polymerase, leading to the rapid amplification of any circular DNA in the extract. If the virus of interest is known, an additional step of digestion with a restriction enzyme which only cuts once in the virus genome can be used. The resulting fragment should be the exact length of the virus genome, and can be gel purified.

Suppressive Subtractive Hybridization

An additional method for purifying and enriching virus nucleic acid is to remove non-virus nucleic acid by suppressive subtractive hybridization (SSH) (Diatchenko *et al.*, 1996). While SSH is used typically to discover differentially expressed genes (He *et al.*, 2005; Munir *et al.*, 2004; Patzwahl *et al.*, 2001), the method has used for comparison of two metagenomic

samples (Galbraith *et al.*, 2004) and could be used for enrichment of pathogenic nucleic acid. This method of enrichment uses two separate pools of nucleic acids, a tester and a driver. The testers are separated into two pools and each are ligated with a separate adapter primers. The two pools are hybridized with an excess of the driver pool, resulting in six possible DNA molecules: 1) tester homohybrid (primer A with primer A or primer B with primer B, 2) tester-driver hybrid, 3) driver hybrid, 4) single stranded driver, 5) single stranded tester and 6) tester-tester heterohybribs. An amplification step using both primer A and primer B will result in exponential amplification of the tester-tester hybrids, but only linear or no amplification for the other combinations. DNA extraction from healthy host tissue will yield the host DNA, mitochondrial (and plastid in the case of plants) DNA, and any DNA from microbes associated with that host. While subtractive hybridization does remove much of the host and other non-viral nucleic acids, it can also remove virus nucleic acids due to cross reactivity. The ligation and amplification steps included in this method compounds both sequencing bias and error, though as mentioned before, use of high fidelity polymerase may help offset those errors. A similar process can be used to remove rRNA from a DNA sample, reducing the host sequence background even further (Sooknanan *et al.*, 2010).

Microarrays

Microarrays have been used for the diagnosis and discrimination of several virus families (Baxi *et al.*, 2006; Boriskin *et al.*, 2004; Hadidi *et al.*, 2004; Wang *et al.*, 2002). The ability to use the multiplexing ability to assay several viruses simultaneously has been beneficial in the identification of several animal and plant virus genera. The probes used in this microarray assays are designed to either be specific or general, specific for differentiation of closely related species or strains (Wang *et al.*, 2002). General probes are designed to be degenerate and able to detect viruses of a single genus or family.

Food Safety and Biosecurity

The U.S. import and export trade of agriculture products contributed over 21 billion dollars to the GDP (82 billion in exports and 61 billion in imports) (USBC, 2012), reflecting the mass globalization of trade, which has led to the artificial spread of several tons of plant material (USBC, 2012). It comes as no surprise that as the movement and trade of plant material increases, so too does the introduction of several exotic plant pathogens increase (Pimentel *et al.*, 2005). These exotic plant pathogens contribute to almost two-thirds (65%) of loss of crop due to plant pathogens (Pimentel, 1993). Imported agriculture products need to be screened to reduce the introduction of these pathogens. In 2012, 62 million metric tons of agriculture products were imported into the U.S. (USBC, 2012), of this only 1% was screened visually, and a percentage of visually screened material is further tested for pathogens (Barrionuevo, 2007). Currently, the primary methods of screening include ELISA and qPCR, as well as quarantine of plant germplasm. Using metagenomic detection methods can reduce the number of assays required to be performed in order to test for all select agents. Additionally, a metagenomic approach to detection will give sequencing information of pathogens of interest which are not select agents.

In addition to the natural introduction of plant pathogens, there is a possibility of the intentional introduction of plant pathogens with malicious intent. The concept of agro-terrorism has been dismissed by some (Young *et al.*, 2008), but bioweapons programs which have focused on plant pathogens were active in at least five countries, the United States (Whitby, 2002), France (Lepick, 1945), Iraq (Whitby & Rogers, 1997), the former Soviet Union (Alibek, 1999), and Japan (Rogers *et al.*, 1999). In addition Islamic militants have shown interest in the weaponization of wheat rust (Fletcher *et al.*, 2006). The goal of terrorism is to disrupt the way of

life of the target. A threat to the food supply of a nation has been shown to cause a lowered confidence in the government, damage to a country's economy and in some cases even riots in the populous (Stack *et al.*, 2010). Even if the critics of agroterrorism are correct, the need to quickly and easily detect select agents in imported and exported food stuffs would limit the movement of harmful pathogens, while simultaneously acting as a deterrent for any form of malicious intentional introduction of pathogens.

Pathogens of Interest

For the work presented herein, three virus groups, the tobamoviruses, begomoviruses, and potyviruses, were chosen for a variety of reasons. The discovery of Passionfruit mosaic tobamovirus (PaMV) in the Tallgrass Prairie Preserve led to a reexamination of the evolutionary history of this virus genus. Bean golden mosaic virus (BGMV) and Plum pox virus (PPV) are pathogens of significant economic importance. For this reason these pathogens were chosen as targets in the development and validation of a bioinformatic pipeline to detect and diagnosis pathogens in a metagenomic sample. Each virus is described in depth below.

Tobamoviruses

Tobamoviruses, a member of the Virgaviridae family, is a postive sense single-stranded RNA virus. The genome consists of four open reading frames (ORFs) of which ORF 2 contains a read through stop codon leading to the expression of a RNA dependent RNA polymerase. ORFs 1, 3 and 4 encode for a methyltransferase/helicase, movement and coat protein respectively. Of these, only ORF 1, 2 and 3 are necessary for movement between cells.

Tobamoviruses are transmitted via mechanical inoculation and infect a wide variety of hosts ranging from cacti to hibiscus. The major groups (sometimes referred to as subgroups) infect solanaecous plants, cucurbits, and brassicas (Lartey *et al.*, 1996). The grouping of viruses coinciding with their hosts suggests that they may have codiverged with their hosts. Codivergence

is similar to co-evolution, with the difference of genetic drift being the major determinant of speciation. Other viruses, such as the Mastreviruses (Wu *et al.*, 2008), and the Polyomaviruses (Perez-Losada *et al.*, 2006) may also have codiverged with their hosts.

Begomoviruses

Bean golden mosaic virus (BGMV) is a member of the family *Geminiviridae*. The genome of BGMV is split between two circular single-stranded DNA molecules (termed DNA A and DNA B). Like most begomoviruses, BGMV is bipartite, while there are examples of monopartite begomoviruses, which lack DNA B while DNA A contains a copy of a movement protein which allows the virus to be spread within a plant. Both DNA A and DNA B contain a constant region, which forms a stem loop structure and is thought to be the origin of replication for both genomes. The replicase gene is located on DNA A, and the resulting protein is responsible for rolling circle amplification of the genome, while the hosts DNA polymerase is responsible for DNA replication (Gutierrez, 1999). This life cycle means that BGMV sequences exist as both DNA and RNA, an important feature when selecting a nucleic acid extraction procedure for detection.

Begomoviruses are carried by their vector, a whitefly *Bemisia tabaci*. Interestingly *B. tabaci* is a cryptic species comprised of several different biotypes (Brown *et al.*, 1995) that differ mostly in protein expression and behavior, though minor genetic polymorphisms have been documented (Cervera *et al.*, 2000). The introduction of a biotype foreign to North America has almost completely displaced the native biotype A (Costa *et al.*, 1993). This evidence that while BGMV is currently only found in South America, the possibility of the introduction of this pathogen to the U.S. Efforts have been made to produce a resistant variety of *Phaseolus vulgaris* (Bonfim *et al.*, 2007).

Potyviruses

A major concern for growers in the state of Pennsylvania and New York is *Plum pox virus* (PPV), the causal agent of Sharka (Atanassov, 1932). PPV is a member of the potyviridae family, and as such has a single-stranded positive sense RNA genome that encodes a single poly-protein, which, after translation, self cleaves into ten separate proteins (Maiss *et al.*, 1989). The HC-Pro gene assists in the transmission of PPV by its aphid vector. The economic impact of PPV is vast, as the virus causes 80-100% fruit drop in areas where PPV is found, leading to 6.4 million dollars of loss in the U.S. alone (Cambra *et al.*, 2006). Understandably, PPV is a concern for stone fruit growers. No treatment is currently available; so the most effective method of prevention is early detection and eradication of the infected plants.

There are 7 different strains of PPV; PPV-C infects cherries (Nemchinov & Hadidi, 1996), PPV-D and PPV-M infect peach, plum and apricot, PPV-EA is currently limited to Egypt (Matic *et al.*, 2011), PPV-W originated in Canada (Stobbs *et al.*, 2005), and PPV-T is a recently described recombinant (Serçe *et al.*, 2009). The seventh strain, PPV-Rec (Recombination), is a commonly found recombinant of PPV-D and PPV-M, with the 5' end of the genome coming from PPV-D, and the remainder from PPV-M (Glasa *et al.*, 2005). These strains are genetically distinct, infect different hosts, and are transmitted with different efficiencies.

REFERENCES

Aird, D., Ross, M., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D., Nusbaum, C. & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Bio* **12**, R18.

Alfonso, P., Rivera, J., Hernáez, B., Alonso, C. & Escribano, J. M. (2004). Identification of cellular proteins modified in response to African swine fever virus infection by proteomics. *Proteomics* **4**, 2037-2046.

Alibek, K. (1999). The Soviet Union's Anti-Agricultural Biological Weapons. *Ann NY Acad Sci* **894**, 18-19.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S. & Liu, H. (2006). The marine viromes of four oceanic regions. *PLoS biology* **4**, e368.

Arif, M. & Ochoa-Corona, F. (2012). Comparative assessment of 5′ A/T-rich overhang sequences with optimal and sub-optimal primers to increase PCR yields and sensitivity. *Mol Biotechnol*, 1-10.

Ashelford, K. E., Weightman, A. J. & Fry, J. C. (2002). PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Res* **30**, 3481-3489.

Atanassov, D. (1932). Plum pox. A new virus disease. *Annals of the University of Sofía Faculty of Agriculture and Silviculture, 11: 49-69* **11**, 49-69.

Baas, T., Baskin, C., Diamond, D. L., Garcia-Sastre, A., Bielefeldt-Ohmann, H., Tumpey, T., Thomas, M., Carter, V., Teal, T. & Van Hoeven, N. (2006). Integrated molecular signature of disease: analysis of influenza virus-infected macaques through functional genomics and proteomics. *J Virol* **80**, 10813-10828.

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriol Rev* **35**, 235.

Barrionuevo, A. (2007). Food imports often escape scrutiny. *New York Times* **1**.

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P. & Lander, E. S. (2002). ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12**, 177-189.

Baxi, M. K., Baxi, S., Clavijo, A., Burton, K. M. & Deregt, D. (2006). Microarray-based detection and typing of foot-and-mouth disease virus. *The Veterinary Journal* **172**, 473-481.

Beijerinck, M. (1898). Concerning a Contagium Viwm Fluidum as Cause of the Spot Disease of Tobacco Leaves.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L. & Bignell, H. R. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59.

Bonfim, K., Faria, J. C., Nogueira, E. O., Mendes, É. A. & Aragão, F. J. (2007). RNAi-mediated resistance to Bean golden mosaic virus in genetically engineered common bean (Phaseolus vulgaris). *Mol Plant Microbe In* **20**, 717-726.

Boriskin, Y. S., Rice, P. S., Stabler, R. A., Hinds, J., Al-Ghusein, H., Vass, K. & Butcher, P. D. (2004). DNA microarrays for virus detection in cases of central nervous system infection. *J Clin Microbiol* **42**, 5811-5818.

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

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. & Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14250-14255.

Brown, J., Frohlich, D. & Rosell, R. (1995). The sweetpotato or silverleaf whiteflies: biotypes of Bemisia tabaci or a species complex? *Annu Rev Entomol* **40**, 511-534.

Cambra, M., Capote, N., Myrta, A. & Llácer, G. (2006). Plum pox virus and the estimated costs associated with sharka disease. *EPPO Bulletin* **36**, 202-204.

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

Cervera, M., Cabezas, J., Simon, B., Martinez-Zapater, J., Beitia, F. & Cenis, J. (2000). Genetic relationships among biotypes of Bemisia tabaci (Hemiptera: Aleyrodidae) based on AFLP analysis. *B Entomol Res* **90**, 391-396.

Chaisson, M. J. & Pevzner, P. A. (2008). Short read fragment assembly of bacterial genomes. *Genome Res* **18**, 324-330.

Cheung, V. G. & Nelson, S. F. (1996). Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 14676-14679.

Collins, F., Lander, E., Rogers, J., Waterston, R. & Conso, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945.

Cooper, B., Eckert, D., Andon, N. L., Yates III, J. R. & Haynes, P. A. (2003). Investigative proteomics: identification of an unknown plant virus from infected plants using mass spectrometry. *J Am Soc Mass Spectr* **14**, 736-741.

Costa, H., Brown, J., Sivasupramaniam, S. & Bird, J. (1993). Regional distribution, insecticide resistance, and reciprocal crosses between the A and B biotypes of Bemisia tabaci. *Insect Sci Appl* **14**, 255.

Cox-Foster, D. L., Conlan, S., Holmes, E. C., Palacios, G., Evans, J. D., Moran, N. A., Quan, P.-L., Briese, T., Hornig, M., Geiser, D. M., Martinson, V., vanEngelsdorp, D., Kalkstein, A. L., Drysdale, A., Hui, J., Zhai, J., Cui, L., Hutchison, S. K., Simons, J. F., Egholm, M., Pettis, J. S. & Lipkin, W. I. (2007). A Metagenomic Survey of Microbes in Honey Bee Colony Collapse Disorder. *Science* **318**, 283-287.

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

Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. (2001). Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**, 1095-1099.

Delwart, E. L. (2007). Viral metagenomics. *Rev Med Virol* **17**, 115-131.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microb* **72**, 5069-5072.

Diatchenko, L., Lau, Y., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N. & Sverdlov, E. D. (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 6025-6030.

Dodds, J. A., Morris, T. J. & Jordan, R. L. (1984). Plant viral double-stranded RNA. *Annu Rev Phytopathol* **22**, 151-168.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P. & Bettman, B. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138.

Elahi, E. & Ronaghi, M. (2004). Pyrosequencing. In *Bacterial Artificial Chromosomes*, pp. 211-219: Springer.

Erlich, Y. & Mitra, P. P. (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods* **5**, 679-682.

Ezenwa, V. O., Godsey, M. S., King, R. J. & Guptill, S. C. (2006). Avian diversity and West Nile virus: testing associations between biodiversity and infectious disease risk. *P Roy Soc B-Biol Sci* **273**, 109-117.

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

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

Galbraith, E. A., Antonopoulos, D. A. & White, B. A. (2004). Suppressive subtractive hybridization as a tool for identifying genetic diversity in an environmental metagenome: the rumen as a model. *Environ Microbiol* **6**, 928-937.

Gardes, M. & Bruns, T. D. (1993). ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. *Mol Ecol* **2**, 113-118.

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

Glasa, M., Paunovic, S., Jevremovic, D., Myrta, A., Pittnerová, S. & Candresse, T. (2005). Analysis of recombinant Plum pox virus (PPV) isolates from Serbia confirms genetic homogeneity and supports a regional origin for the PPV-Rec subgroup. *Arch Virol* **150**, 2051-2060.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P. & Sykes, S. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1513-1518.

Gräser, Y., El Fari, M., Vilgalys, R., Kuijpers, A., De Hoog, G., Presber, W. & Tietz, H.-J. (1999). Phylogeny and taxonomy of the family Arthrodermataceae (dermatophytes) using sequence analysis of the ribosomal ITS region. *Med Mycol* **37**, 105-114.

Gutierrez, C. (1999). Geminivirus DNA replication. *Cell Mol Life Sci* **56**, 313-329.

Hadidi, A., Czosnek, H. & Barba, M. (2004). DNA microarrays and their potential applications for the detection of plant viruses, viroids, and phytoplasmas. *J Plant Pathol* **86**, 97-104.

Hamblin, C., Barnett, I. & Hedger, R. (1986). A new enzyme-linked immunosorbent assay (ELISA) for the detection of antibodies against foot-and-mouth disease virus I. Development and Method of ELISA. *Journal of Immunological Methods* **93**, 115-121.

Hammond, J. (1992). Potyvirus serology, sequences and biology. In *Potyvirus Taxonomy*, pp. 123-138: Springer.

Harrison, B. (1981). Plant virus ecology: ingredients, interactions and environmental influences. *Annals of Applied Biology* **99**, 195-209.

He, N., Qin, Q. & Xu, X. (2005). Differential profile of genes expressed in hemocytes of White Spot Syndrome Virus-resistant shrimp (< i> Penaeus japonicus</i>) by combining suppression subtractive hybridization and differential hybridization. *Antiviral Research* **66**, 39-45.

Henegariu, O., Heerema, N., Dlouhy, S., Vance, G. & Vogt, P. (1997). Multiplex PCR: critical parameters and step-by-step protocol. *Biotechniques* **23**, 504-511.

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

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

Kimura, H., Morita, M., Yabuta, Y., Kuzushima, K., Kato, K., Kojima, S., Matsuyama, T. & Morishima, T. (1999). Quantitative analysis of Epstein-Barr virus load by using a real-time PCR assay. *J Clin Microbiol* **37**, 132-136.

Koonin, E. V. (1991). Genome replication/expression strategies of positive-strand RNA viruses: a simple version of a combinatorial classification and prediction of new strategies. *Virus genes* **5**, 273-281.

Koonin, E. V., Senkevich, T. G. & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol Direct* **1**, 29.

Korlach, J., Bibillo, A., Wegener, J., Peluso, P., Pham, T. T., Park, I., Clark, S., Otto, G. A. & Turner, S. W. (2008). Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. *Nucleosides, Nucleotides and Nucleic Acids* **27**, 1072-1082.

Kristensen, D. M., Mushegian, A. R., Dolja, V. V. & Koonin, E. V. (2010). New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* **18**, 11-19.

Kuypers, J., Wright, N., Ferrenberg, J., Huang, M.-L., Cent, A., Corey, L. & Morrow, R. (2006). Comparison of real-time PCR assays with fluorescent-antibody assays for diagnosis of respiratory virus infections in children. *J Clin Microbiol* **44**, 2382-2388.

Lahr, D. & Katz, L. A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* **47**, 857-866.

Lartey, R. T., Voss, T. C. & Melcher, U. (1996). Tobamovirus evolution: Gene overlaps, recombination, and taxonomic implications. *Mol Biol Evol* **13**, 1327-1338.

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

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G. & Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682-686.

Mahony, J., Chong, S., Merante, F., Yaghoubian, S., Sinha, T., Lisle, C. & Janeczko, R. (2007). Development of a respiratory virus panel test for detection of twenty human respiratory viruses by use of multiplex PCR and a fluid microbead-based assay. *J Clin Microbiol* **45**, 2965-2970.

Maiss, E., Timpe, U., Brisske, A., Jelkmann, W., Casper, R., Himmler, G., Mattanovich, D. & Katinger, H. (1989). The complete nucleotide sequence of plum pox virus RNA. *J Gen Virol* **70**, 513-524.

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

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

Matic, S., Elmaghraby, I., Law, V., Varga, A., Reed, C., Myrta, A. & James, D. (2011). Serological and molecular characterization of isolates of Plum pox virus strain El Amar to better understand its diversity, evolution, and unique geographical distribution. *J Plant Pathol* **93**, 303-310.

Melcher, U., Muthukumar, V., Wiley, G. B., Min, B. E., Palmer, M. W., Verchot-Lubicz, J., Ali, A., Nelson, R. S., Roe, B. A., Thapa, V. & Pierce, M. L. (2008). Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from Ambrosia psilostachya. *J Virol Methods* **152**, 49-55.

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

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

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

Nakano, M., Komatsu, J., Matsuura, S.-i., Takashima, K., Katsura, S. & Mizuno, A. (2003). Single-molecule PCR using water-in-oil emulsion. *J Biotechnol* **102**, 117-124.

Nemchinov, L. & Hadidi, A. (1996). Characterization of the sour cherry strain of plum pox virus. *Phytopathology* **86**, 575-580.

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

Oliveira, D. C. & de Lencastre, H. (2002). Multiplex PCR strategy for rapid identification of structural types and variants of the mec element in methicillin-resistant Staphylococcus aureus. *Antimicrob Agents Ch* **46**, 2155-2161.

Paton, A. W. & Paton, J. C. (1998). Detection and Characterization of Shiga ToxigenicEscherichia coli by Using Multiplex PCR Assays forstx 1, stx 2, eaeA, Enterohemorrhagic E. coli hlyA, rfb O111, andrfb O157. *J Clin Microbiol* **36**, 598-602.

Patzwahl, R., Meier, V., Ramadori, G. & Mihm, S. (2001). Enhanced expression of interferon-regulated genes in the liver of patients with chronic hepatitis C virus infection: detection by suppression-subtractive hybridization. *J Virol* **75**, 1332-1338.

Perez-Losada, M., Christensen, R. G., McClellan, D. A., Adams, B. J., Viscidi, R. P., Demma, J. C. & Crandall, K. A. (2006). Comparing Phylogenetic Codivergence between Polyomaviruses and Their Hosts. *J Virol* **80**, 5663-5669.

Pimentel, D. (1993). *Habitat factors in new pest invasions*: New York: John Wiley & Sons.

Pimentel, D., Zuniga, R. & Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecol Econ* **52**, 273-288.

Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. (2004). Comparative genome assembly. *Brief Bioinform* **5**, 237-248.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H. & Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005-1010.

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

Rogers, P., Whitby, S. & Dando, M. (1999). Biological warfare against crops. *Sci Am* **280**, 62-67.

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

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J. & Edwards, M. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463-5467.

Schaad, N. W. & Frederick, R. D. (2002). Real-time PCR and its application for rapid plant disease diagnostics. *Can J Plant Pathol***24**, 250-258.

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

Schloss, P. D. & Handelsman, J. (2005b). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microb* **71**, 1501-1506.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H. & Robinson, C. J. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microb* **75**, 7537-7541.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature* **200**, 16-18.

Schwartz, S., Friedberg, I., Ivanov, I. V., Davidson, L. A., Goldsby, J. S., Dahl, D. B., Herman, D., Wang, M., Donovan, S. M. & Chapkin, R. S. (2012). A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biol* **13**, r32.

Scott, H. (1963). Purification of cucumber mosaic virus. *Virology* **20**, 103-106.

Serçe, Ç. U., Candresse, T., Svanella-Dumas, L., Krizbai, L., Gazel, M. & Çağlayan, K. (2009). Further characterization of a new recombinant group of *Plum pox virus* isolates, PPV-T, found in orchards in the Ankara province of Turkey. *Virus Res* **142**, 121-126.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. & Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117-1123.

Singh, B. K., Millard, P., Whiteley, A. S. & Murrell, J. C. (2004). Unravelling rhizosphere–microbial interactions: opportunities and limitations. *Trends Microbiol* **12**, 386-393.

Sooknanan, R., Pease, J. & Doyle, K. (2010). Novel methods for rRNA removal and directional, ligation-free RNA-seq library preparation. *Nat Methods| Application Notes*.

Spackman, E., Senne, D. A., Myers, T., Bulaga, L. L., Garber, L. P., Perdue, M. L., Lohman, K., Daum, L. T. & Suarez, D. L. (2002). Development of a real-time reverse

transcriptase PCR assay for type A influenza virus and the avian H5 and H7 hemagglutinin subtypes. *J Clin Microbiol* **40**, 3256-3260.

Stack, J., Suffert, F. & Gullino, M. (2010). Bioterrorism: A Threat to Plant Biosecurity? In *The Role of Plant Pathology in Food Safety and Food Security*, pp. 115-132: Springer.

Stobbs, L., Van Driel, L., Whybourne, K., Carlson, C., Tulloch, M. & Van Lier, J. (2005). Distribution of Plum pox virus in residential sites, commercial nurseries, and native plant species in the Niagara region, Ontario, Canada. *Plant Dis* **89**, 822-827.

Suttle, C. A. (2005). Viruses in the sea. *Nature* **437**, 356-361.

Thomson, D. & Dietzgen, R. G. (1995). Detection of DNA and RNA plant viruses by PCR and RT-PCR using a rapid virus release protocol without tissue homogenization. *J Virol Methods* **54**, 85-95.

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

Uchiyama, T., Abe, T., Ikemura, T. & Watanabe, K. (2004). Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol* **23**, 88-93.

USBC (2012). Statistical Abstract of the United States 2012. Edited by USBC. Washington D.C.: U.S. Government Printing Office.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A. & Holt, R. A. (2001). The sequence of the human genome. *Science Signaling* **291**, 1304.

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

Wang, D., Coscoy, L., Zylberberg, M., Avila, P. C., Boushey, H. A., Ganem, D. & DeRisi, J. L. (2002). Microarray-based detection and genotyping of viral pathogens. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 15687-15692.

Ward, D. M., Weller, R. & Bateson, M. M. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community.

Webster, C. G., Coutts, B., Jones, R., Jones, M. & Wylie, S. (2007). Virus impact at the interface of an ancient ecosystem and a recent agroecosystem: studies on three legume-infecting potyviruses in the southwest Australian floristic region. *Plant Pathol* **56**, 729-742.

Whitby, S. & Rogers, P. (1997). Anti-crop biological warfare-implications of the Iraqi and US programs. *Defense Analysis* **13**, 303-317.

Whitby, S. M. (2002). *Biological warfare against crops*: Palgrave Publishers Ltd.

Wren, J. D., Roossinck, M. J., Nelson, R. S., Scheets, K., Palmer, M. W. & Melcher, U. (2006). Plant virus biodiversity and ecology. *PLoS biology* **4**, e80.

Wu, B. L., Melcher, U., Guo, X. Y., Wang, X. F., Fan, L. J. & Zhou, G. H. (2008). Assessment of codivergence of Mastreviruses with their plant hosts. *BMC Evol Biol* **8**.

Yamamoto, K. R., Alberts, B. M., Benzinger, R., Lawhorne, L. & Treiber, G. (1970). Rapid bacteriophage sedimentation in the presence of polyethylene glycol and its application to large-scale virus purification. *Virology* **40**, 734-744.

Young, J., Allen, C., Coutinho, T., Denny, T., Elphinstone, J., Fegan, M., Gillings, M., Gottwald, T., Graham, J. & Iacobellis, N. (2008). Plant-pathogenic bacteria as biological weapons-real threats? *Phytopathology* **98**, 1060-1065.

Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829.

Zheng, X., Hong, L., Shi, L., Guo, J., Sun, Z. & Zhou, J. (2008). Proteomics analysis of host cells infected with infectious bursal disease virus. *Mol Cell Proteomics* **7**, 612-625.

CHAPTER III

CO-DIVERGENCE AND HOST-SWITCHING IN THE EVOLUTION OF TOBAMOVIRUSES

Summary

The proposed phylogenetic structure of the genus Tobamovirus supports the idea that these viruses have codiverged with their hosts since radiation of the hosts from a common ancestor. The determinations of genome sequence for two strains of *Passionfruit mosaic virus* (PafMV), a tobamovirus from plants of the family Passifloraceae (order Malpighiales) from which only one other tobamovirus (*Maracuja mosaic virus*; MarMV) had been characterized, combined with the development of Bayesian analysis methods for phylogenetic inference, provided an opportunity to reassess the codivergence hypothesis. The sequence of one PafMV strain, PafMV-TGP, was discovered during a survey of plants of the Tallgrass Prairie Preserve for their virus content. Its nucleotides are only 73% identical to those of MarMV. A conserved, open reading frame not found in other tobamovirus genomes, and encoding a cysteine-rich protein, was found in MarMV and both PafMV strains. Phylogenetic tree construction, using an alignment of the nucleotide sequences of PafMV-TGP and other tobamoviruses, resulted in a major clade containing isolates exclusively from rosid plants. Asterid-derived viruses were

found exclusively in a second major clade that also contained an orchid-derived tobamovirus and tobamoviruses infecting plants of the order Brassicales. With a few exceptions, calibrating the virus tree with dates of host divergence at two points resulted in predictions of divergence times of family-specific tobamovirus clades that were consistent with the times of divergence of the host plant orders.

Introduction

Hallmark virus genes are those whose relatives can be recognized in a wide variety of virus genomes but are not found in any host organism genome (Koonin *et al.*, 2006). Their recognition suggests that viruses have existed since life began. Recent work finding insertions of virus sequences in multiple mammalian genomes (Belyi *et al.*, 2010) supports the antiquity of viruses, and suggests the presence of viruses at least 40-50 million years ago (mya). Consistent with an ancient origin of viruses, the species diversity of angiosperm viruses may have arisen largely during long association of viruses and hosts during the angiosperms' divergence and proliferation. Such diversification is designated “codivergence” to emphasize a lesser role of selection in the process. The angiosperm virus genus Tobamovirus of the family Virgaviridae (Adams *et al.*, 2009) has been proposed to have evolved by codivergence (Gibbs, 1980, 1999; Lartey *et al.*, 1996).

The type member of the genus Tobamovirus (Fauquet *et al.*, 2005), *Tobacco mosaic virus* (TMV), figured prominently in the recognition of the existence of filterable infectious agents now called viruses (Schothof *et al.*, 1999). The TMV genome consists of a positive-sense single-stranded RNA (ssRNA) molecule of 6.4 knt that is packaged in a non-enveloped rod-shaped coat approximately 300 nm in length. Particle preparations also contain shorter rods, corresponding to encapsidated subgenomic RNAs that contain the virus origin of assembly (Fukuda *et al.*, 1981). The TMV genome has four open reading frames (ORFs), encoding a 126 kDa replicase

component that has methyl-transferase (MT), helicase (Hel) and RNA silencing suppressor domains (Vogler *et al.*, 2007), a 183 kDa protein which is a read-through product of the 126 kDa protein ORF and has an additional RNA-dependent RNA polymerase (RdRp) domain, a movement protein (MP), and a coat protein (CP) (Fig. 1). The 126 kDa and/or the 183 kDa protein and the MP are necessary for TMV intercellular movement (Harries *et al.*, 2009).

Codivergence of tobamovirus lineages with those of their hosts is suggested by apparently slow rates of sequence change and the structure of phylogenetic trees based on both nucleotide and amino acid sequences. Several lines of evidence support a slow rate of accumulation of nucleotide substitutions within species of the tobamoviruses. The initial sequence determination of the TMV genome was performed in two laboratories (Dawson *et al.*, 1986; Goelet *et al.*, 1982) on samples that had been propagated separately in tobacco for 30 years and yet they yielded virtually identical sequences. Sequence comparisons of archival *Tobacco mild green mosaic virus* (TMGMV) specimens from wild *Nicotiana glauca* covering a span of close to 100 years showed little evidence of divergence at this time scale (Fraile *et al.*, 1997).

Phylogenetic evidence showing clustering according to host taxonomy also supports the codivergence hypothesis for tobamovirus species. Initial support was obtained from a comparison of CP amino acid sequences of known tobamoviruses (Gibbs, 1980) and was extended by a comparison of nucleotide sequences of each coding region (Gibbs, 1999; Lartey *et al.*, 1996). Where distinct virus species have been isolated from members of the same plant order (Cucurbitales, Malvales, Fabales, Solanales, Brassicales and Lamiales), almost invariably their sequences have clustered on the same branch of the viral phylogenetic trees (Gibbs, 1999; Lartey *et al.*, 1996; Min *et al.*, 2009). For the Cucurbitales, Malvales and Fabales, the branches do not contain viruses isolated from other plant orders. Viruses of Brassicales and Lamiales co-habit the same branch, designated subgroup III (Lartey *et al.*, 1996) to distinguish it from the branch associated with plants of the Solanales (subgroup I) and other tobamoviruses (subgroup II). A

virus isolated from orchids has one genome portion that branches with the viruses from Solanales members and another that is found on the Brassicales-Lamiales branch, suggesting it is a recombinant (Lartey *et al.*, 1996). Comparative sequence analysis also revealed that all lineages of tobamoviruses except that of the recombinant tobamovirus had impressively uniform rates of sequence evolution, consistent with genetic drift minimally influenced by selective events. Patterns of codivergence of other virus species with the species of their hosts have been identified (Perez-Losada *et al.*, 2006). Cases of codivergence generally suggest nucleotide substitution rates in the vicinity of 10^{-8} substitutions/site/year (Gibbs *et al.*, 2010).

The availability of nucleotide sequences from numerous dated isolates of single virus species combined with the development of Bayesian phylogenetic analysis has allowed estimation of the rates of nucleotide substitution in a diversity of virus species (Duffy & Holmes, 2008; Fargette *et al.*, 2008; Kang *et al.*, 2009; Moore & Donoghue, 2009; Perez-Losada *et al.*, 2006; Ramsden *et al.*, 2008). The calculated substitution rates are similar to mutation rates estimated experimentally by analysis of progeny from plants inoculated with cloned sequences. The rates are in the range of 10^{-3} to 10^{-5} substitutions/site/year (Duffy & Holmes, 2008; Fargette *et al.*, 2008). Extrapolation of such rates linearly over time to represent the evolution of species led to the conclusion that current virus taxa arose recently rather than having codiverged with their hosts. Recently, Bayesian analysis has been extended to selected individual species of the genus Tobamovirus and identified a similarly high substitution rate (Pagán *et al.*, 2010) for those species. A similarly high rate of evolution for the virus species in the genus, extrapolated from that analysis, would be inconsistent with the codivergence hypothesis.

Plants of the family Passifloraceae (order Malipighiales) host a variety of viruses, including the tobamoviruses *Maracuja mosaic virus* (MarMV) (Song *et al.*, 2006) and *Passionfruit mosaic virus* (PafMV) (Song & Ryu, 2011). An isolate of the latter (PafMV-FL, NC_015552) was thought originally to be a strain of MarMV, MarMV-FL. While this

manuscript was in preparation, its sequence was reported (Song & Ryu, 2011), revealing that the virus is a distinct species in the genus Tobamovirus and proposing the *Passionfruit mosaic virus* name. The Plant Virus Biodiversity and Ecology (PVBE) project focused on determining the distribution of plant viruses in close to 600 native or non-crop plant species in the Tallgrass Prairie Preserve of Osage Co., Oklahoma (TGP) through sequence analysis (Wren *et al.*, 2006). One sequence identified in the study apparently was from a virus strain of PafMV, which we refer to here as PafMV-TGP. Characterization of the PafMV-TGP genome revealed a novel, conserved C-rich ORF of unknown function also found in PafMV-FL and MarMV. The availability of these new tobamovirus genomes coupled with those of other newly sequenced tobamovirus genomes (Min *et al.*, 2006; Srinivasan *et al.*, 2005) allowed a new examination of the codivergence hypothesis for tobamoviruses.

Results

PafMV-TGP

In the PVBE project, viral sequences in plant extracts were enriched by either isolation of double-stranded RNA (dsRNA) or the preparation of virus-like particles (VLP) prior to nucleic acid extraction (producing VLP-VNA; see Methods). Both approaches to identifying viral sequences in plant extracts yielded unequivocal identification of two plants with evidence of the presence of a member of the genus *Tobamovirus*. One plant, 05TGP00580 (sampled on 24 July, 2005 from 36.848^{o} N, 96.420^{o} W), the only sample of the *Passiflora incarnata* species analyzed, produced a high percentage of tobamovirus sequence reads in both methods (VLP-VNA, 85.2% of 2,931 reads; dsRNA, 66.0% of 280 reads). The other, 07TGP00004 (sampled on 8 June, 2007 from 36.838 ^{o} N, 96.443 ^{o} W), was a sample of *Vernonia baldwinii* represented by 0.7% of 597 reads.

Sequences obtained (SuppFile1) were sufficiently abundant to allow assembly of 6696 nt into one large contig, missing only short sequences (less than 100 nt) at the 5'- and 3' ends, probably including the first 11 codons of the 5'-most ORF. The high percentages of nucleotide and predicted amino acid identity for PafMV-TGP compared to PafMV-FL (Table 1) clearly identified the two strains as belonging to the same species. Among known tobamovirus genomes, the next closest known relative of the sequence was that of MarMV, exhibiting 72.7 % nucleotide sequence identity. The PafMV-TGP genome (Fig. 1) contains similar ORFs in an order similar to that of other members of the tobamovirus genus, but with the following notable differences. A stretch of 374 nt separates the 185 kDa ORF termination codon from the MP initiation codon. The 185K ORF termination codon is followed, starting four nts 3' of it, by an ORF of 594 nt. It thus overlaps the MP ORF out of frame for 220 nt. This ORF is present also in the PafMV-FL and MarMV genomes although it was not described in the published reports (Song et. al., 2006; Song and Ryu, 2011) and is not annotated in GenBank. The ORF can encode a hypothetical cysteine-rich protein located between nucleotide number 4809 and 5403. In a BLASTp search of protein sequences, the amino acid sequence of the predicted product of the ORF lacked similarity to proteins of characterized functions. Conservation of this ORF in both PafMV strains and MarMV suggests that it is a functional ORF. The PafMV-TGP and MarMV MP regions overlap with the CP regions in a different frame for 118 nt, similar to the overlap previously noted for crucifer-associated tobamoviruses (Lartey *et al.*, 1996). The overlap regions involving each of the last three coding regions perhaps account for the higher percent nucleotide identity for these regions (Table 1). Of 6,098 PafMV-TGP positions covered by more than one contig or singleton read, only 29 (0.5%) exhibited evidence of polymorphism. Of those 29, only two were informatively polymorphic (Suppfile2: Table 1). The highest densities of polymorphic positions were in the ORF for the putative cysteine-rich polypeptide and the C-terminal part of the movement protein, 8.4 and 7.5 polymorphisms per kilonucleotide, respectively, compared to 3.1,

3.2, and 3.7 polymorphisms per kilonucleotide, respectively, for the MT-Hel domain, the RdRp domain and the CP ORF.

Phylogenetics

A Bayesian likelihood tree of the replicase (MT/Hel-RdRp) ORFs of the tobamoviruses (Fig. 2A) clearly defined clades containing the Lamiales-, Solanales-, Fabales-, Malvales- and Cucurbitales- associated lineages. This definition was obtained also by maximum likelihood analysis (data not shown). The asterid Lamiales-associated clade, corresponding to subgroup III, also included viruses associated with the asterid Ericales and the rosid Brassicales orders. Viruses infecting plants of these orders were interspersed in the topology of the subgroup III clade. The Solanales-associated clade, corresponding to subgroup I, included *Rehmannia mosaic virus* (ReMV), a close relative of TMV whose host is a member of the Lamiales. At a deeper level, subgroup I and subgroup III clades clustered together separately from subgroup II viruses. The subgroup I-III branch was subtended by one containing a single member clade consisting of *Cactus mild mottle virus* (CMMoV, host order Caryophyllales), while the other single member clade, that of *Frangipani mosaic virus* (FrMV, host order Gentianales of the asterid clade) appeared basal to the subgroup II clade. However, the confidence intervals were such that the two single member clades and all order-specific subgroup II branches may have originated at about the same time. Among these branches, PafMV-TGP appeared as sister to MarMV in a Malpighiales-associated branch. The MT/Hel ORF gave a tree identical in topology and similar in proportions to that generated for the replicase ORF, as expected, due to the latter being a continuation of the MT/Hel ORF (Fig. 1). Greater variation was seen in the MP and CP trees, most likely due to the shorter length of their ORFs and greater heterogeneity in their evolutionary rates as shown below.

Dating

If the substitution rate remains constant throughout the history of a gene, the gene can be said to be clock-like and can be used to extrapolate dates of divergence. The potential clock-like nature of the substitutions within the tobamovirus ORFs was addressed by examining the ucld.stdev parameter of BEAST (standard deviation of the relaxed substitution clock rates; Table 2). This parameter reflects the rate heterogeneity of the lineages. When ucld.stdev >1, the heterogeneity in rates is high; when ucld.stdev = 0, rates are perfectly clock-like. The ucld.stdev of the nucleotides in the four ORFs ranged from 0.253 to 0.477, with the RdRp ORF being the most clock-like.

If virus clades diverged from each other near the same time that their host orders did, we would expect a linear correlation between age of plant orders and virus clade divergence distances. The ages of seven plant orders (Cucurbitales, Solanales, Malvales, Malpighiales, Lamiales, Caryophylales and Fabales) were well correlated with the ages of the viral clades associated with them (Fig. 3A, $R^2 = 0.795$, slope = 0.670). The strong correlation motivated dating the Bayesian trees, using plant divergence estimates (Magallon & Castillo, 2009). We used a uniform distribution for divergences of plants of the orders Cucurbitales (120.22 – 120.32 mya) and Solanales (77.42 – 77.52 mya) from other taxa. Using these points for calibration did not change the topology significantly, but improved the correlation coefficient (Fig. 3B, $R^2 = 0.8921$, slope = 0.98). Omitting the Cucurbitales and Solanales from the calculation of correlation did not change the correlation coefficient significantly ($R^2 = 0.8994$, slope = 0.96). Estimates of ages of the nodes of the viral replicase derived from this correlation are given in Table 3. The data allowed calculation of the divergence resulting in rosid- and asterid/Caryophyllales-associated groups to have occurred approximately 109 - 130 mya.

Four virus-host order age relationships were outliers in the correlation plot (Fig. 3): orders Asparaginales, Brassicales, Lamiales and Gentianales. The position of the Asparaginales – *Odontoglossum ringspot virus* (ORSV) correlation can be attributed to the recombinant nature of this virus, mentioned above. Also as mentioned above, the viruses of the Brassicales and Ericales are in a clade with viruses derived from the Lamiales. FrMV was extrapolated to have separated 16 mya earlier than the plants of the Gentianales diverged from other asterids.

Where estimatable, crown ages (times since first evidence of divergence in a lineage) for virus branches were an average of 27 million years less than those for the host crown ages (Table 3), except for the order Malvales and their viruses, but were well within the 95% highest posterior density. Both Malpighiales-associated virus species were isolated from plants of the genus Passiflora and would therefore be expected to have diverged later than the radiation of the plant orders. The substitution rates of each branch diverging within the interspersed Lamiales/Brassicales-associated clade showed little heterogeneity in their calculated rates of evolution (Supplemental file 4).

BaTS Analysis of Association.

If members of the genus Tobamovirus did codiverge alongside their hosts (a form of allopatric speciation), an association of the branching patterns of the primary natural hosts and their viruses should be seen (Kitchen *et al.*, 2011). To test this, the tips of the posterior set of trees found using the above method were labeled with the host order from which the virus was derived. This labeled tree set was tested using Bayesian Tip-Significance (BaTS) software (Parker *et al.*, 2008), which tests the association of the primary host and the virus employing three independent statistical tests, association index, parsimony score and maximum monophyletic clade. Both the association index and parsimony score tests showed a strong association of states within the trees. The maximum monophyletic clade test showed strong association ($p < 0.05$) of the orders

Cucurbitales-, Malvales-, Fabales-, and Malpighiales-associated viruses (Table 4). Viruses of subgroup I (Solanales-associated) were less strongly associated ($p < 0.5$) likely due to the inclusion of ReMV, a Lamiales-associated virus. Analysis using the posterior set of trees (PSTs) attained from the CP and MP regions showed subgroup I to have a strong association ($p < 0.05$). Gentianales-, Caryophyales-, and Ericales-derived viruses only had one representative tip and, because of this, were not able to associate with other tips. The Lamiales-, and Brassicales-associated viruses showed no statistical association ($p = 1$).

Discussion

Codivergence

Both RNA and DNA viruses have been hypothesized to have codiverged with their hosts (Perez-Losada *et al.*, 2006; Wu *et al.*, 2008). The principal observations supporting codivergence hypotheses in general are the congruence of viral phylogenetic tree reconstructions with those of the host organisms with which the viruses are associated. As the number of characterized tobamoviruses has increased (Min *et al.*, 2006; Song & Ryu, 2011; Srinivasan *et al.*, 2005) the validity of the generalization that virus phylogenetic trees resemble those of their isolation plant hosts has been strengthened. The present analysis strongly supports a codivergence hypothesis for tobamoviruses. PafMV was confidently assigned to a clade of tobamoviruses associated with hosts in the Malpighiales (Song & Ryu, 2011; this work). In the survey of plants of the TPP, only a specimen of the genus *Passiflora* accumulated PafMV-TGP to high levels. Plants of an order are most often hosts to a single monophyletic clade of tobamoviruses found in that order.

The virus tree (Fig. 2) mirrored the plant species tree in most regards. Malvales-, Cucurbitales-, Fabales-, and Malpighiales-associated viruses each formed monophyletic clades. All four of their plant orders are rosids and no asterid-associated viruses were in the branch that included these clades, indicating that supra-order association also exists. Two single member

clades, CMMoV and FrMV, appeared to diverge from common ancestors with the rosid clade and a clade containing all other known asterid-associated viruses. The phylogenetic distances of the points of divergence of seven of nine order-specific virus clades were significantly correlated with the ages of the plant orders (Fig. 3B). The oldest branch point within each order-specific virus clade occurred within 10 million years after diversification of the order (Table 3). The splitting of virus lineages following splitting of the host lineage is consistent with codivergence. Furthermore, and most importantly, BaTS analysis showed (Table 4) that the virus tree topology and its association with host taxonomy is highly unlikely to have resulted from random processes. Combined, these observations suggest strongly that the apparent relationship of virus clades with the phylogeny of their natural hosts reflects an important evolutionary phenomenon.

Alternatives to Codivergence

Establishment of an evolutionary relationship underlying tree similarity requires four assumptions. It assumes that a virus species' host of initial isolation reflects the hosts in which it spent most of its time evolving. Consistent with the assumption, annotations of known natural hosts of tobamoviruses in the International Committee for the Taxonomy of Viruses (2006) database reveal no taxonomically widespread distribution of natural hosts. In the TGP plant community, PafMV-TGP was found only twice during this study. Of about 450 specimens of six frequently sampled plant species, only one was definitively positive for the virus. Of 550 plant species tested, only two yielded evidence of the virus. The *P. incarnata* specimen had a high concentration of the tobamovirus with no obvious symptoms of infection, suggesting that it was the source of the virus found in the other plant. *P. incarnata* cuttings may have been transported to Florida from Arkansas (Hill *et al.*, 1992), the state neighboring that of the PVBE study, a likely source of PafMV-FL. These observations suggest that PafMV established a long-term productive association with one host lineage while still occurring in plants of other lineages in ways that do not contribute to evolution of the virus. TGP viruses such as the *Asclepias asymptomatic*

tymovirus (Min *et al.*, 2011) and *Asclepias virus TGP-2* (Thapa *et al.*, 2012), a proposed member of the Secoviridae, were detected at substantial levels in many host species, although populations were highest in *Asclepias viridis*. Thus, assuming that the host of initial isolation is an indicator of the host plant lineage in which tobamoviruses evolved is reasonable, but may not be valid for other viruses.

A second assumption in asserting codivergence based on tree similarity is that codivergence could be accompanied by occasional successful infection and establishment in a plant lineage other than the lineage of origin. Such species jumps may have happened several times in tobamovirus evolution. Some putative jumps apparent in Fig. 2 are not statistically supportable. The Fabales-associated viruses branched with those associated with the Malpighiales, while the Fabales plant family is thought to branch with the Cucurbitales (Magallon & Castillo, 2009). However, both alternate interpretations (Fabales-associated viruses with Cucurbitales-associated viruses, or, in the plant tree, the Fabales with the Malpighiales) are possible since the node separations were inadequate for confident placement (within the 95% confidence limits of the deduced trees). Confidence levels also mitigate against attaching significance to the difference in virus and host branching patterns for FrMV and CMMoV. One incontrovertible exception to absolute tree congruence is ReMV, a virus of a plant species in the Lamiales whose closest relative is the Solanales-associated TMV. Another exception occurs in the branch of the tree that includes viruses associated with all three asterid orders: those infecting Ericales, Solanales and Lamiales. The branch has two subbranches. All Solanales-associated viruses included in the study occur on one subbranch (subgroup 1), also occupied by ReMV. The other subbranch (subgroup 3) contains viruses from multiple orders: the remaining Lamiales, the rosid Brassicales and ORSV (discussed later). Close relatives of these viruses have also been identified from the asterid orders Solanales (Petunia (Sabanadzovic *et al.*, 2008)) and Ericales (Actinidia and Impatiens) (Chavan *et al.*, 2009; Heinze *et al.*, 2006). Within this mixed order

subbranch there is no substructure according to plant host, suggesting that this lineage has acquired the ability to be successful in multiple hosts. Placement of this subbranch as sister to the primarily asterid Solanales-associated subbranch suggests that the lineage arose in the astrid Lamiales and subsequently gained the ability to survive and spread in the rosid Brassicales and other lineages. That tobamoviruses from Brassicales and Lamiales do not form separate clades, but instead are part of a single clade suggests that this new ability was attained after the divergence of subgroup III from subgroup I.

A third assumption in asserting codivergence based on tree similarity is that rates of nucleotide substitution are much slower in the process of evolution of viral species than they are during evolution within a species. In this study, as has been done in others (Rector *et al.*, 2007), calibrating the viral tree using estimated dates of plant order divergence (Magallon and Castillo, 2009) gave substitution rates of the order of 10^{-8} to 10^{-9} substitutions/site/year. This rate is compatible with the observation that sequences of tobamoviruses from century-old herbarium specimens are not appreciably different from modern sequences (Fraile *et al.*, 1997). Slow evolutionary substitution rates for species evolution have also been proposed for other viruses (Gibbs *et al.*, 2010; Wu *et al.*, 2008). Rates estimated by Bayesian analysis for the evolution of sequences within species using dated isolates are orders of magnitude higher than the interspecies inferences assuming codivergence (Harkins *et al.*, 2009; Pagán *et al.*, 2010). For tobamoviruses, Pagán *et al.* (2010) found rates of 10^{-4} to 10^{-5} substitutions/site/year by comparing tobamovirus samples of individual species from the asterid clade and CGMMV from the past 60 years. Using just CP sequences of selected species, they extrapolated a maximum predicted age of the last common ancestor of known tobamoviruses of 10^{5} years rather than the 10^{8} years inferred from the codivergence hypothesis. In support of the younger estimate, rates of the magnitude 10^{-4} to 10^{-5} substitutions/site/year have also been obtained upon inoculation of plants with cloned genomes followed by later harvest and sequencing (Ge *et al.*, 2006; Schneider and Roossinck, 2001). The

Pagán *et al.* (2010) study of genus evolution differed from the one presented here in that it focused on the smallest section of the tobamovirus genome (less than 10% of the total length), rather than the replicase and included only one species outside the Solanales-Lamiales associated clade as defined in the present study.

There are three reasons why it is risky to extrapolate results of Bayesian analysis of substitution during intraspecies evolution linearly to the events that gave rise to those species. First, substitution profiles for within species variation differ from those of between species variation. The types of substitutions dominating substitution profiles of tobamoviruses vary with the taxonomic level of the comparison (Melcher, 2010). G<->A and T<->C transitions predominate among recently diverged pairs of sequences but accumulate at rates similar to those of other substitutions with more diverged pairs. Second, a variety of observations suggest that the nucleotide populations of viruses in their natural hosts is low, implying stability over large evolutionary scales (Acosta-Leal *et al.*, 2011). Purifying selection may be particularly strong and bottlenecks may be comparatively wide. These views are consistent with the comparatively low density of polymorphisms among the sequences retrieved for PafMV-TGP. Third, the apparent discrepancy between interspecies and intraspecies substitution rates can be reconciled easily by recognizing that the best evolution models used in Bayesian estimations are usually those that invoke a category of sites that are invariant. However, the proportion of sites that are actually invariant will decrease with increasing phylogenetic distance. Substitution rates within the evolution of individual species focus on a limited number of changing sites. These play only a small role in phylogenetic inference at the interspecies level because the sites are close to being saturated with changes at the longer times. The short-term invariant sites do undergo substitution during interspecies evolution and it is those substitutions that are important in the phylogenetic reconstruction of species evolution. An additional manner by which studies of historical strains can lead to misleading rates is that strains of a virus may disappear for a time and then reappear in

a later year, such as has been reported for TGP carmovirus 3 (Scheets *et al.*, 2011) such that the differences between them do not reflect sequence evolution on the time scale postulated but on a longer one resulting from evolution of strains. It bears noting, correspondingly, that the wide difference in rates among sites means that Bayesian estimates of divergence times based on codivergence with hosts are exaggeratedly high for the most recent nodes.

A fourth assumption of the tobamovirus-host codivergence hypothesis is that an ancestral tobamovirus existed before the radiation of dicotyledonous plants. The recent analysis of the genome of *Chara australis virus* (CAV), a virus associated with a brown alga (Gibbs *et al.*, 2011) suggests this assumption is warranted. The CP gene of CAV is homologous to those of known tobamoviruses, but is about 1.7 fold more ancient than the radiation of tobamoviruses of embryophyta. Parsimony analysis of the CP ORF nucleotide sequence (data not shown), but not the amino acid sequences (Gibbs *et al.*, 2011), placed the CAV branch in the vicinity of the origin of tobamovirus diversification. This origin can be deduced from the tree (Fig.2B) from the region where the confidence limits of many of the order-specific branches overlap. The Fabales-, Malpighiales- Cucurbitales-, Caryophylales- and Gentianales associated virus clades and the combined Solanales-Lamiales- associated virus clade have similarly deep branches suggesting that the root of the dicotyledonous tobamovirus tree is in this region. However, it is important to recognize that there are many plant orders from which no tobamovirus has been detected to date. Since several orders harboring tobamoviruses have only recently been reported, this absence could be due to inadequate sampling. An alternative, suggested earlier for the order Brassicales (Lartey *et al.*, 1996), is that the original tobamovirus lineages in those orders now without tobamoviruses have died out.

Generation of codivergence

The above discussion shows that reasonable assumptions support the designation of the similarity between viral and natural host trees as codivergence. If codivergence did actually occur, one must ask: what is the cause of codivergence? Some would argue that it is the result of limited sampling, asserting that as additional tobamoviruses are found, they will be discovered to fill in the tree so that the association of viral and host trees disappears. Exactly the opposite has occurred. Early suggestions of codivergence only identified two rosid-associated viruses, one in the Fabales and the other in the Cucrbitales. Since then four additional Cucrbitales-associated viruses and one Fabales infecting virus have been described and new branches for Malvales-associated viruses and Malpighiales associated virus each have more than one clade member. While this manuscript was being prepared, a report of a tobamovirus from a plant of the order Caryophyllales appeared with phylogenetic placement as sister to CMMoV, also from a Caryophyllales plant (Kim *et al.*, 2011). Additional Solanales-derived and Lamiales-associated viruses have, for the most part, been placed on the appropriate branches.

A second argument against attributing tree similarity to codivergence is that the apparent coincidences of host and viral phylogenetic trees reflect adaptation of the virus to a specific host environment (Holmes, 2008). Viruses in similar host environments (same order, for example) would therefore naturally be more similar than viruses from diverse environments (different orders, for example). While it has been shown that adaptive sequence changes occur when a virus is experimentally transferred from one host to another, the number of such changes is usually small and thus of minor influence in shaping phylogenetic trees (Wallis et al, 2007). Tobamoviruses provide at least two examples of the failure of adaptation to provide major phylogenetic signals. The orchid associated tobamovirus, ORSV, has been shown to be a recombinant and some acceleration of evolution associated with recombination was detected (Lartey *et al.*, 1996). Nevertheless, adequate phylogenetic signal remains in the recombined

segments to place them with high confidence in known tobamovirus clades. Second, viruses of the Lamiales/Brassicales-associated virus clade infect both rosids and asterids. There are two possible explanations. One requires at least two distinct host "jumps" from hosts in the Lamiales to hosts in the Brassicales. However, such jumps, if they occurred, could not have been accompanied by any adaptive changes since sequences of tobamoviruses isolated from plants of the order Lamiales are virtually identical to those isolated from plants of the order Brassicales. The alternate explanation is that an earlier event allowed the subgroup to infect plants of multiple orders. The evolutionary rate of a virus is expected to increase after a host jump (Smith *et al.*, 2009) because adaptive substitutions are selected for after the jump. Consistent with this view, jumps in host species are known to increase the intraplant diversity of a virus population (Schneider & Roossinck, 2001). However, the evolutionary rates for each branch within subgroup 3 do not suggest any specific host crossing event, but instead suggest that the subgroup gained the ability to infect both rosids and asterids approximately between 61 and 21 mya. Because adaptation is unlikely to have played a major role in causing codivergence, we suggest the following model. At the time before the radiation of the dicotyledonous orders, there were large populations of their last common ancestor. These ancestors harbored large populations of diverse copies of a tobamovirus, the diversity being derived by gradual genetic drift from a common ancestor. The large population and the environmental conditions at the time, increased the probability of multiple near simultaneous mutations occurring in the same genome, allowing it to establish a new lineage on a new fitness plane.

Methods

PVBE Methods

Plant sampling and the isolation and extraction of VLP-VNA have been described previously (Melcher *et al.*, 2008), as has the extraction of dsRNA from these plant samples, its

conversion to dsDNA and its subsequent sequencing (Roossinck *et al.*, 2010). VLP-VNA sequences reported here were determined as described for dsRNA (Roossinck *et al.*, 2010). Assembled sequences of the putative PafMV-TGP inferred from the contig data were used to identify additional sequences of these viruses among unassembled sequences. These singleton sequences were used in building the consensus sequence. Because of the chance that an occasional read could have been assigned mistakenly to the wrong plant sample in a sequencing pool, a plant sample was called positive using the following criteria: the virus reads were >0.4% of the total reads of the sample, and there was confidence that the read tags were not misread (leading to attribution of the read to the wrong sample).

Phylogenetic Methods

Sequences for this study were retrieved from GenBank (See Suppfile2 Table 2). Clustal W was used to create a multiple sequence alignment using default values (Larkin *et al.*, 2007). Alignments were viewed using BioEdit (Hall, 1999). Initial studies to place PafMV-TGP into a subgroup used the PHYLIP package. Parsimonious and distance trees were made with *Tobacco rattle virus* (TRV) as an out-group. The evolutionary rates of the tobamoviruses were found using Bayesian methods with BEAST v 1.5.3 and a most likely tree was made (Drummond & Rambaut, 2007). Log files were viewed to ascertain convergence using Tracer v1.5. Trees were viewed using FigTree v1.3.0. XML files for BEAST are available in suppfile4. All priors for the undated tree were determined by BEAST. For the dated trees, calibration dates were chosen using the speciation times for the order Cucurbitales and Solanales hosts (120.22 – 120.32 mya and 77.42 – 77.52 mya respectively) (Magallon & Castillo, 2009) and used a uniform prior distribution. A GTR model of evolution with a gamma distribution of rates and invariant sites was found to be the most probable model via JModelTest (Posada, 2008) and was used in this study. All other priors were determined by BEAST. Four ORFs were analyzed with BEAST, the 120 kDa ORF which encodes for the MT/Hel protein, the 180 kDa replicase ORF which includes the MT/Hel

and RdRp, the MP ORF, and the CP ORF. Each of these was run to 10 million states with a burn-in of 1 million. PSTs were manually edited to include the host derived states (see Table 4 for states) and a burn-in of 1 million states. These PSTs were then analyzed using BaTS with 1,000 replicates.

Acknowledgements

The work was supported by funding from NSF-EPSCoR (EPS-0447262) , USDA- NIFA (2010-85605-20542) and the Oklahoma Agricultural Experiment Station whose Director has approved the manuscript for publication. The authors thank Plant Virus Biodiversity and Ecology colleagues for gathering of materials and assistance in nucleic acid preparation, Graham B. Wiley and Bruce A. Roe of the Advanced Center for Genome Technology for 454 sequencing, Andrew Doust, Adrian Gibbs and Akhtar Ali for reading drafts of the manuscript, and Mike Deom and Adrian Gibbs for providing sequences.

REFERENCES

Acosta-Leal, R., Duffy, S., Xiong, Z., Hammond, R. & Elena, S. F. (2011). Advances in plant virus evolution: translating evolutionary insights into better disease management. *Phytopathology* **101**, 1136-1148.

Adams, M. J., Antoniw, J. F. & Kreuze, J. (2009). Virgaviridae: a new family of rod-shaped plant viruses. *Arch Virol* **154**, 1967-1972.

Belyi, V. A., Levine, A. J. & Skalka, A. M. (2010). Sequences from Ancestral Single Stranded DNA Viruses In Vertebrate Genomes: The Parvoviridae And Circoviridae Are More Than 40-50 Million Years Old. *J Virol*, JVI.01789-01710.

Chavan, R. R., Pearson, M. N. & Cohen, D. (2009). Partial characterisation of a novel Tobamovirus infecting Actinidia chinensis and A. deliciosa (Actinidiaceae) from China. *Euro J Plant Pathol* **124**, 247-259.

Dawson, W. O., Beck, D. L., Knorr, D. A. & Grantham, G. L. (1986). cDNA cloning of the complete genome of tobacco mosaic virus and production of infectious transcripts. *P Natl Acad Sci USA* **83**, 1832-1836.

Drummond, A. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214.

Duffy, S. & Holmes, E. C. (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *J Virol* **82**, 957-965.

Fargette, D., Pinel, A., Rakotomalala, M., Sangu, E., Traore, O., Sereme, D., Sorho, F., Issaka, S., Hebrard, E., Sere, Y., Kanyeka, Z. & Konate, G. (2008). Rice yellow mottle virus, an RNA plant virus, evolves as rapidly as most RNA animal viruses. *J Virol* **82**, 3584-3589.

Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U. & Ball, L. A., (eds) (2005). *Virus Taxonomy*. London: Elsevier.

Fraile, A., Escriu, F., Aranda, M., Malpica, J., Gibbs, A. & Garcia-Arenal, F. (1997). A century of tobamovirus evolution in an Australian population of Nicotiana glauca. *J Virol* **71**, 8316-8320.

Fukuda, M., Meshi, T., Okada, Y., Otsuki, Y. & Takebe, I. (1981). Correlation between particle multiplicity and location on virion RNA of the assembly initiation site for viruses of the Tobacco mosaic virus group. *P Natl Acad Sci USA-BIO* **78**, 4231-4235.

Ge, L., Zhang, J., Zhou, X. & Li, H. (2007). Genetic structure and population variability of Tomato yellow leaf curl China virus. *J Virol* **81**, 5902-5907.

Gibbs, A. (1980). How ancient are the Tobamoviruses. *Intervirology* **14**, 101-108.

Gibbs, A. (1999). Evolution and origins of tobamoviruses. *Philos Trans R Soc Lond Ser B-Biol Sci* **354**, 593-602.

Gibbs, A. J., Fargette, D., Garcia-Arenal, F. & Gibbs, M. J. (2010). Time - the emerging dimension of plant virus studies. *J Gen Virol* **91**, 13-22.

Gibbs, A. J., Torronen, M., Mackenzie, A. M., Wood, J. T., Armstrong, J. S., Kondo, H., Tamada, T. & Keese, P. L. (2011). The enigmatic genome of Chara australis virus. *J Gen Virol* **92**, 2679-2690.

Goelet, P., Lomonossoff, G. P., Butler, P. J., Akam, M. E., Gait, M. J. & Karn, J. (1982). Nucleotide sequence of tobacco mosaic virus RNA. *P Natl Acad Sci USA* **79**, 5818-5822.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41**, 95-98.

Harkins, G. W., Delport, W., Duffy, S., Wood, N., Monjane, A. L., Owor, B. E., Donaldson, L., Saumtally, S., Triton, G., Briddon, R. W., Shepherd, D. N., Rybicki, E. P., Martin, D. P. & Varsani, A. (2009). Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virol J* **6**, (16 July 2009).

Harries, P. A., Park, J.-W., Sasaki, N., Ballard, K. D., Maule, A. J. & Nelson, R. S. (2009). Differing requirements for actin and myosin by plant viruses for sustained intercellular movement. *P Natl Acad Sci USA* **106**, 17594-17599.

Heinze, C., Lesemann, D. E., Ilmberger, N., Willingmann, P. & Adam, G. (2006). The phylogenetic structure of the cluster of tobamovirus species serologically related to ribgrass mosaic virus (RMV) and the sequence of streptocarpus flower break virus (SFBV). *Arch Virol* **151**, 763-774.

Holmes, E. C. (2008). Evolutionary History and Phylogeography of Human Viruses. *Annu Rev Microbiol* **62**, 307-328.

Kang, H. J., Bennett, S. N., Sumibcay, L., Arai, S., Hope, A. G., Mocz, G., Song, J. W., Cook, J. A. & Yanagihara, R. (2009). Evolutionary insights from a genetically divergent hantavirus harbored by the European common mole (Talpa europaea). *PLoS One* **4**, e6149.

Kim, N., Hong, J., Song, Y., Chung, B., Park, J. & Ryu, K. H. (2012). The complete genome sequence of a member of a new species of tobamovirus (rattail cactus necrosis-associated virus) isolated from Aporcactus flagelliformis. *Arch Virol* **157**, 185-187.

Kitchen, A., Shackelton, L. A. & Holmes, E. C. (2011). Family level phylogenies reveal modes of macroevolution in RNA viruses. *P Natl Acad Sci USA* **108**, 238-243.

Koonin, E. V., Senkevich, T. G. & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol Direct* **1**, 29.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

Lartey, R. T., Voss, T. C. & Melcher, U. (1996). Tobamovirus evolution: Gene overlaps, recombination, and taxonomic implications. *Mol Biol Evol* **13**, 1327-1338.

Magallon, S. & Castillo, A. (2009). Angiosperm Diversification Through Time. *Am J Bot* **96**, 349-365.

Melcher, U. (2010). Assessing constancy of substitution rates in viruses over evolutionary time. *BMC bioinformatics* **11**, S3.

Melcher, U., Muthukumar, V., Wiley, G. B., Min, B. E., Palmer, M. W., Verchot-Lubicz, J., Ali, A., Nelson, R. S., Roe, B. A., Thapa, V. & Pierce, M. L. (2008). Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from Ambrosia psilostachya. *J Virol Methods* **152**, 49-55.

Min, B.-E., Feldman, T. S., Ali, A., Wiley, G., Muthukumar, V., Roe, B. A., Roossinck, M., Melcher, U., Palmer, M. W. & Nelson, R. S. (2012). Molecular characterization, ecology, and epidemiology of a novel Tymovirus in Asclepias viridis from Oklahoma. *Phytopathology* **102**, 166-176.

Min, B., Song, Y. & Ryu, K. (2009). Complete sequence and genome structure of cactus mild mottle virus. *Arch Virol* **154**, 1371-1374.

Min, B. E., Chung, B. N., Kim, M. J., Ha, J. H., Lee, B. Y. & Ryu, K. H. (2006). Cactus mild mottle virus is a new cactus-infecting tobamovirus. *Arch Virol* **151**, 13-21.

Moore, B. R. & Donoghue, M. J. (2009). A Bayesian approach for evaluating the impact of historical events on rates of diversification. *Proc Natl Acad Sci U S A* **106**, 4307-4312.

Pagán, I., Firth, C. & Holmes, E. (2010). Phylogenetic Analysis Reveals Rapid Evolutionary Dynamics in the Plant RNA Virus Genus Tobamovirus. *J Mol Evol* **71**, 298-307.

Parker, J., Rambaut, A. & Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infect Genet Evol* **8**, 239-246.

Perez-Losada, M., Christensen, R. G., McClellan, D. A., Adams, B. J., Viscidi, R. P., Demma, J. C. & Crandall, K. A. (2006). Comparing Phylogenetic Codivergence between Polyomaviruses and Their Hosts. *J Virol* **80**, 5663-5669.

Posada, D. (2008). jModelTest: Phylogenetic Model Averaging. *Mol Biol Evol* **25**, 1253-1256.

Ramsden, C., Holmes, E. C. & Charleston, M. A. (2008). Hantavirus Evolution in Relation to its Rodent and Insectivore Hosts: No Evidence for Co-divergence. *Mol Biol Evol*, msn234.

Rector, A., Lemey, P., Tachezy, R., Mostmans, S., Ghim, S. J., Van Doorslaer, K., Roelke, M., Bush, M., Montali, R. J., Joslin, J., Burk, R. D., Jenson, A. B., Sundberg, J. P., Shapiro, B. & Van Ranst, M. (2007). Ancient papillomavirus-host co-speciation in Felidae. *Genome Bio* **8**.

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

Sabanadzovic, S., Abou Ghanem-Sabanadzovic, N., Henn, A. & Lawrence, A. (2008). Characterization of a petunia strain of turnip vein-clearing virus. *J Plant Pathol* **90**, 505-509.

Scheets, K., Blinkova, O., Melcher, U., Palmer, M. W., Wiley, G. B., Ding, T. & Roe, B. A. (2011). Detection of members of the *Tombusviridae* in the Tallgrass Prairie Preserve, Osage County, Oklahoma, USA. *Virus Res* **160**, 256-263.

Schneider, W. L. & Roossinck, M. J. (2001). Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions. *J Virol* **75**, 6566-6571.

Schothof, K.-B. G., Shaw, J. G. & Zaitlin, M. (1999). *Tobacco Mosaic Virus, One Hundred Years of Contributions to Virology*. St. Paul: APS Press.

Smith, G. J. D., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., Peiris, J. S. M., Guan, Y. & Rambaut, A. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122-1125.

Song, Y. S., Min, B. E., Hong, J. S., Rhie, M. J., Kim, M. J. & Ryu, K. H. (2006). Molecular evidence supporting the confirmation of Maracuja mosaic virus as a species of the genus Tobamovirus and production of an infectious cDNA transcript. *Arch Virol* **151**, 2337-2348.

Song, Y. S. & Ryu, K. H. (2011). The complete genome sequence and genome structure of passion fruit mosaic virus. *Arch Virol* **156**, 1093-1095.

Srinivasan, K. G., Narendrakumar, R. & Wong, S. M. (2002). Hibiscus virus S is a new subgroup II tobamovirus: evidence from its unique coat protein and movement protein sequences. *Arch Virol* **147**, 1585-1598.

Thapa, V., Melcher, U., Wiley, G. B., Doust, A., Palmer, M. W., Roewe, K., Roe, B. A., Shen, G., Roossinck, M. J. & Wang, Y. M. (2012). Detection of members of the *Secoviridae* in the Tallgrass Prairie Preserve, Osage County, Oklahoma, USA. *Virus Res* **167**, 34-42.

Vogler, H., Akbergenov, R., Shivaprasad, P. V., Dang, V., Fasler, M., Kwon, M. O., Zhanybekova, S., Hohn, T. & Heinlein, M. (2007). Modification of small RNAs associated with suppression of RNA silencing by tobamovirus replicase protein. *J Virol* **81**, 10379-10388.

Wren, J. D., Roossinck, M. J., Nelson, R. S., Scheets, K., Palmer, M. W. & Melcher, U. (2006). Plant virus biodiversity and ecology. *PLoS biology* **4**, e80.

Wu, B. L., Melcher, U., Guo, X. Y., Wang, X. F., Fan, L. J. & Zhou, G. H. (2008).

Assessment of codivergence of Mastreviruses with their plant hosts. *BMC Evol Biol* **8**..

Table 1. Comparison of nucleotide and predicted amino acid sequences of two strains of *Passionfruit mosaic virus*.

Region	Nucleotide Identity*	Amino Acid Identity*
Methyl transferase-helicase (125K) †	94.9 (3062/3267)	98.8 (1077/1089)
Replicase (184K)	94.1 (4511/4794)	98.9 (1583/1601)
Putative C-rich protein	96.5 (573/594)	96.0 (190/198)
Movement protein	96.4 (899/933)	99.7 (310/311)
Coat protein	97.8 (522/534)	97.8 (174/178)
Overall	94.9	n.a.

*Percentage identity (number identical/number of positions) comparing PafMV-TGP (JF807914) and PafMV-FL (NC_015552)

†First 89 nt residues of PafMV-TGP missing

Table 2: Analysis of clock-like behavior of tobamoviral ORFs.

ORF	ucld.stdev* (aa)	95% HPD	ucld.stdev (nt)	95% HPD
MT-Hel	0.331	0.235 -0.435	0.317	0.232 – 0.415
RdRp†	NA	NA	0.267	0.192 – 0.350
RdRp	0.271	0.188 – 0.362	0.253	0.191 – 0.326
CP	0.228	2.96E-3 – 0.407	0.376	0.208 – 0.539
MP	0.404	0.216 – 0.612	0.477	0.297 – 0.680

*This value is a representation of the heterogeneity of rates found in a tree. ORFs with values close to zero are considered to be clock-like, while those whose values are close to, or above one are considered not clock-like.

†Uncalibrated tree without priors.

Table 3: Divergence dates* for tobamovirus clades.

Clade	Stem Divergence Time (mya)	95% HPD♦	Crown Divergence Time (mya)	95%HPD
Asterid- / Rosid-associated split	NA	NA	118.007	109.018 – 129.513
Cucurbitales-associated †	102.271	102.223 – 102.318	67.584	57.050 – 77.094
Solanales-associated †	77.470	77.424 – 77.519	66.851	62.269 – 70.655
Malvales-associated	86.916	79.043 – 94.528	40.690	27.490 – 54.818
Lamiales/Brassicales-associated	77.470	77.424 – 77.519	61.653	54.494 – 68.918
Caryophyllales-associated	107.520	96.068 – 120.544	NA	NA
Malpighiales-associated	86.916	79.043 – 94.528	32.346	22.679 – 41.924
Fabales-associated	96.015	85.340 – 109.361	45.617	32.680 – 57.961
Gentianales-associated	96.015	84.340 – 109.361	NA	NA

* Deduced by BEAST from the replicase alignment

♦Highest posterior density

† Clade divergences used as calibration points

Table 4: Statistical evidence for host association within a tobamovirus clade*.

Statistic (state) †	observed mean	lower 95% CI	upper 95% CU	null mean	lower 95% CI	upper 95% CI	significance
AI	1.487	1.311	1.552	2.980	2.467	3.372	0.000
PS	13.000	13.000	13.000	20.426	18.968	22.000	0.000
MC (Ericales) ‡	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MC (Solanales)	2.011	2.000	2.000	1.547	1.000	3.000	0.437
MC (Caryophyllales) ‡	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MC (Fabales)	2.000	2.000	2.000	1.015	1.000	1.000	0.015
MC (Brassicales)	1.000	1.000	1.000	1.058	1.000	1.782	1.000
MC (Cucurbitales)	5.000	5.000	5.000	1.221	1.000	2.000	0.001
MC (Gentianales) ‡	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MC (Malvales)	2.000	2.000	2.000	1.018	1.000	1.000	0.014
MC (Malpighiales)	2.000	2.000	2.000	1.020	1.000	1.000	0.019
MC (Orchidales) ‡	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MC (Lamiales)	1.000	1.000	1.000	1.117	1.000	2.000	1.000

* Deduced with BaTS using the using the PSTs from the above BEAST analysis

†Association index (AI), parsimony score (PS), and Maximum monophyletic clade (MC)

‡ Only one state included

Figure captions

Figure 1. Genome structure of *Passionfruit mosaic virus*-TGP compared to that of *Tobacco mosaic virus*. Arrows indicate open reading frames. Sizes of 120 and 180 kDa polypeptides are approximate due to lack of the 5' end of the nucleotide sequence.

Figure 2. Bayesian-likelihood phylogenetic trees using the replicase gene of tobamoviruses without (A) and with (B) dating priors. The dated tree (B) is in millions of years (scale bar equals 20 million years). Posterior probabilities are shown on the nodes. Names of viruses and sources of their sequences are given in Supplemental File 3.

Figure 3. Patristic distances of tobamovirus clades as a function of the divergence times of their associated host plants. Patrisitic distances were derived from branch lengths from Figures 2A and 2B, respectively for panels A and B (without and with dating priors).

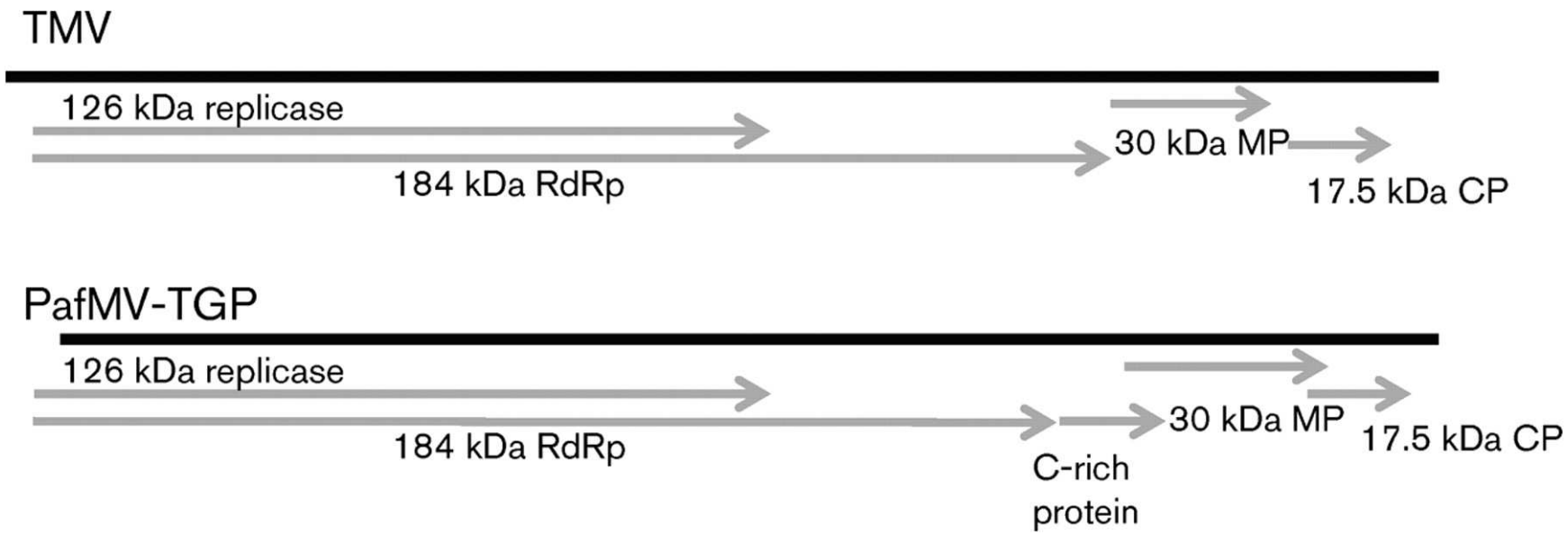
TMV
126 kDa replicase
184 kDa RdRp
30 kDa MP
17.5 kDa CP
PafMV-TGP
126 kDa replicase
184 kDa RdRp
C-rich
protein
30 kDa MP
17.5 kDa CP

Figure 1

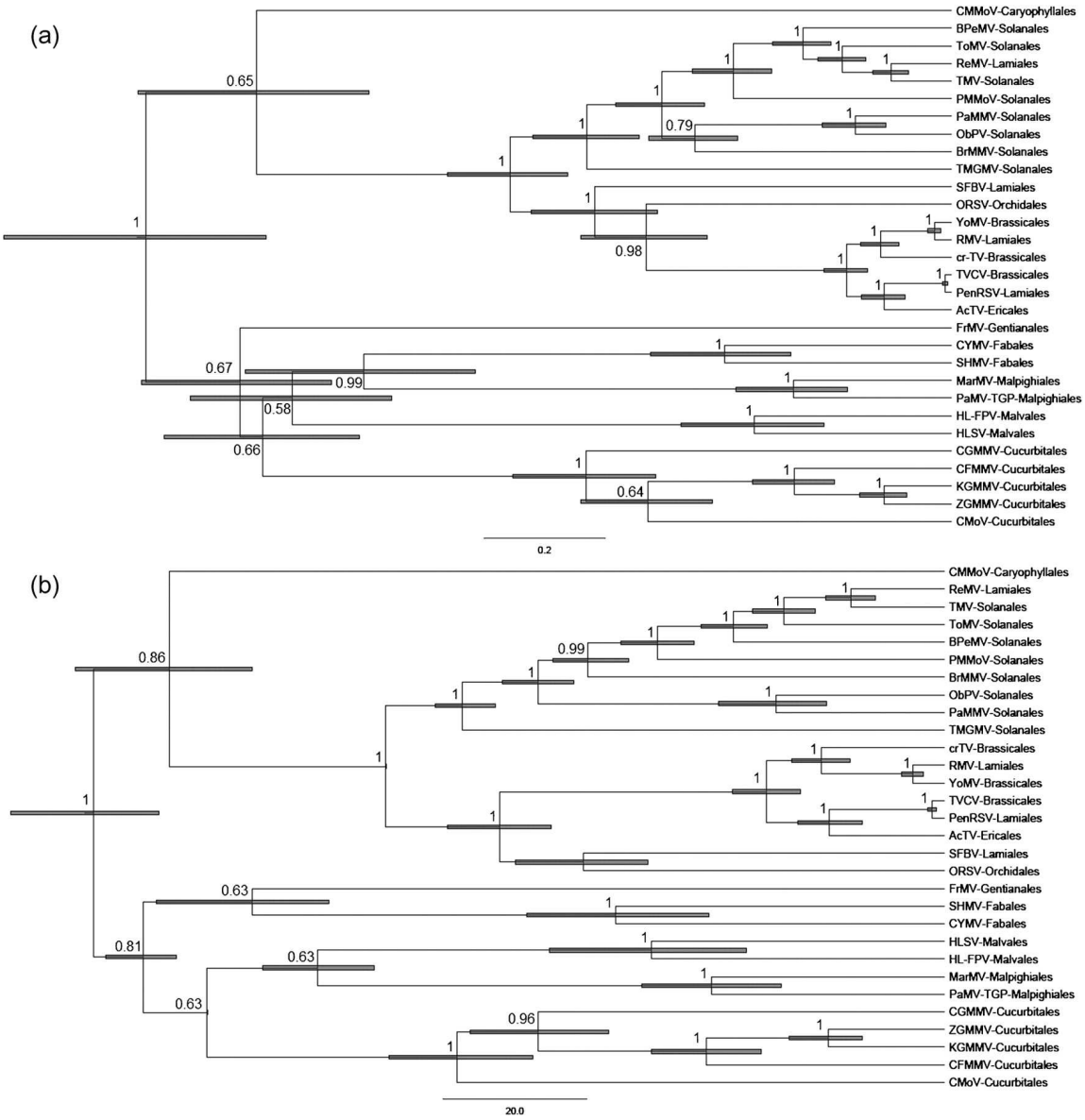
(a)
CMMoV-Caryophyllales
BPeMV-Solanales
ToMV-Solanales
ReMV-Lamiales
TMV-Solanales
PMMoV-Solanales
PaMMV-Solanales
ObPV-Solanales
BrMMV-Solanales
TMGMV-Solanales
SFBV-Lamiales
ORSV-Orchidales
YoMV-Brassicales
RMV-Lamiales
cr-TV-Brassicales
TVCV-Brassicales
PenRSV-Lamiales
AcTV-Ericales
FrMV-Gentianales
CYMV-Fabales
SHMV-Fabales
MarMV-Malpighiales
PaMV-TGP-Malpighiales
HL-FPV-Malvales
HLSV-Malvales
CGMMV-Cucurbitales
CFMMV-Cucurbitales
KGMMV-Cucurbitales
ZGMMV-Cucurbitales
CMoV-Cucurbitales
0.65
0.79
0.98
0.67
0.99
0.58
0.66
0.64
0.2
(b)
CMMoV-Caryophyllales
ReMV-Lamiales
TMV-Solanales
ToMV-Solanales
BPeMV-Solanales
PMMoV-Solanales
BrMMV-Solanales
ObPV-Solanales
PaMMV-Solanales
TMGMV-Solanales
crTV-Brassicales
RMV-Lamiales
YoMV-Brassicales
TVCV-Brassicales
PenRSV-Lamiales
AcTV-Ericales
SFBV-Lamiales
ORSV-Orchidales
FrMV-Gentianales
SHMV-Fabales
CYMV-Fabales
HLSV-Malvales
HL-FPV-Malvales
MarMV-Malpighiales
PaMV-TGP-Malpighiales
CGMMV-Cucurbitales
ZGMMV-Cucurbitales
KGMMV-Cucurbitales
CFMMV-Cucurbitales
CMoV-Cucurbitales
0.86
0.99
0.63
0.81
0.63
0.63
0.96
20.0

Figure 2

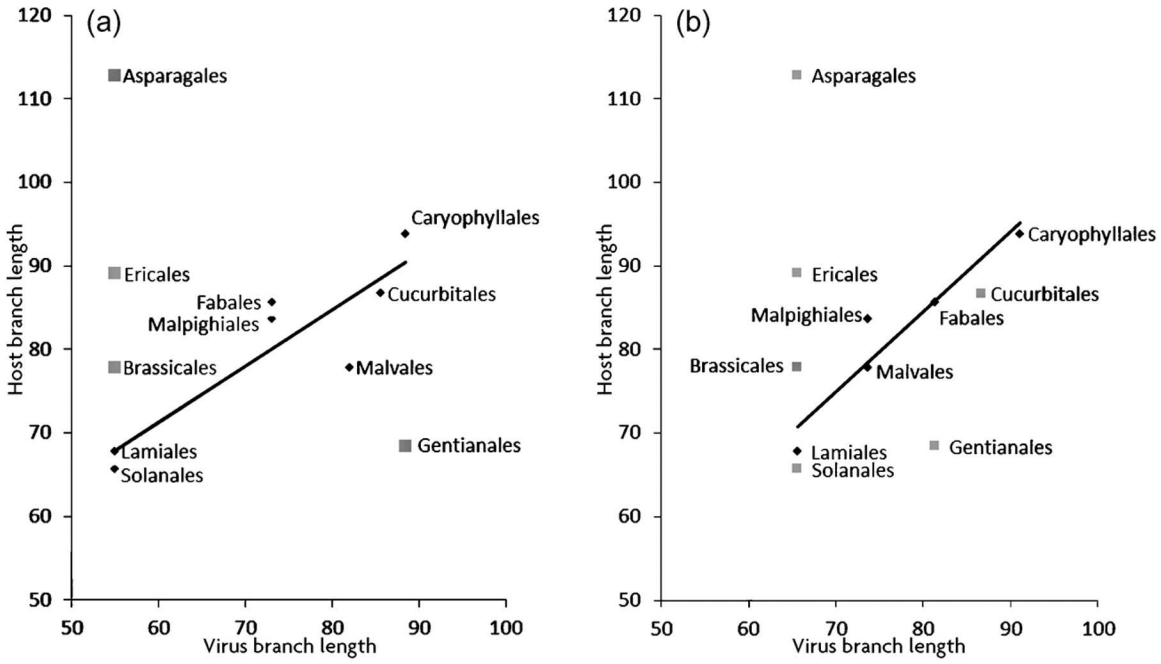
(a)
Asparagales
Caryophyllales
Ericales
Fabales
Malpighiales
Cucurbitales
Brassicales
Malvales
Gentianales
Lamiales
Solanales
Host branch length
Virus branch length
120
110
100
90
80
70
60
50
50
60
70
80
90
100
(b)
Asparagales
Caryophyllales
Ericales
Cucurbitales
Malpighiales
Fabales
Brassicales
Malvales
Lamiales
Gentianales
Solanales
Host branch length
Virus branch length

Figure 3

CHAPTER IV

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

Abstract

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

designed for each pathogen varied with the genome size of the pathogen. To give confidence to diagnostic calls, a statistical method of determining the presence of a given pathogen was developed, in which target e-probe signals (detection signal) are compared to signals generated by a decoy set of e-probes (background signal). The E-probe Diagnostic Nucleic acid Assay (EDNA) process provides the framework for a new sequence-based detection system which eliminates the need for assembly of NGS data.

Author Summary

Humans, agricultural animals and plants all face threats arising from disease outbreaks caused by extant and emerging pathogens. A single streamlined assay capable of detecting any and all microbes in a given sample would represent a powerful tool for diagnostic analysis. Due to advances in next generation sequencing and metagenomics, the use of nucleic acid sequence based diagnostics as a broad range diagnostic is becoming more feasible. With the ever growing amount of sequence data, it is quickly becoming unrealistic to perform wholesale searches of next generation sequence outputs against curated databases. The E-probe Diagnostic Nucleic acid Analysis offers a simple and fast approach to detecting sequences belonging to pathogens of interest within a metagenomic sequence background. We have shown the basic concept of designing pathogen specific sequences (which we have termed e-probes) to be used in searching raw next generation sequencing data to be faster than a search against known curated databases, as well as successful in identifying many types of pathogens of interest including viruses, bacteria, fungi and oomycetes.

Introduction

Agricultural biosecurity is a priority for ensuring uninterrupted international and interstate trade, which in turn ensures an abundant food supply. With increased movement of commodities across state and national borders, the risk of introduction of exotic plant pathogens

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

Next generation sequencing (NGS) is a relatively recent technology that allows for the generation of very large amounts of sequence data from a given sample (Ronaghi, 2001). Because

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

The advantage of NGS over other sequencing technologies is the volume (400MB – 28GB) of data generated (Metzker, 2009; Reis-Filho, 2009). From a different perspective, the volumes of data generated by NGS could be a detriment to a diagnostician, as bioinformatic processing becomes a limiting factor in high throughput applications (Magi *et al.*, 2010; Pop & Salzberg, 2008). For example, consider 200 liters of seawater containing over 5000 different viruses (Breitbart *et al.*, 2003). If a metagenomics approach is used for plant pathogen detection within this sample, pathogen-specific sequences will likely make up only a small percentage of the total reads (Adams *et al.*, 2009; Roossinck *et al.*, 2010). In contrast, plants infected with

viruses may have a much higher percentage of the total nucleic acid comprised of pathogen sequences (Kreuze *et al.*, 2009). The host sequences that would make up the majority of an infected plant metagenome sample are essentially unimportant for diagnosis.

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

Materials and Methods

Pathogens and Their Sequences

The plant pathogens studied here belong to three general groups, viral, prokaryotic, and eukaryotic pathogens. The chosen systems represent a wide variety of plant pathogens and have global economic importance (Table 1). Two viruses were used: *Plum pox virus*, a single stranded RNA virus, and *Bean golden mosaic virus*, which is a bipartite DNA virus. Prokaryotic pathogens included *Xylella fastidiosa* 9a5c, the causal bacterium of citrus variegated chlorosis, *Xanthomonas oryzae* pv. oryzae, which causes bacterial blight in rice, and *Ralstonia solanacearum* race 3 biovar 2, a select agent that causes wilting of a variety of crops including potatoes and tomatoes, *Candidatus Liberibacter asiaticus*, a bacterium responsible for citrus

greening, and *Spiroplasma citri*, which causes citrus stubborn disease. Eukaryotic pathogens included: *Puccinia graminis* a rust fungus, causing the stem rust of wheat and affecting a very broad host range including 365 cereals and grasses in 54 genera (Hodson, 2005); *Phytophthora ramorum*, a stramenopile with a wide host range of 23 species in 12 plant families (Rizzo 2003); and *Phakopsora pachyrhizi*, which causes soybean rust, a widespread pathogen that now can be found in Africa, Asia, Australia, South America and Hawaii (Miles, 2003). For each pathogen, a near neighbor was chosen based on a close phylogenetic relationship, and the availability of complete genome sequence (Table 1). Grapevine, *Vitis vinifera* (GenBank Accession: PRJNA33471), was chosen as the host background due to the availability of its genome sequence, and its genome size, which is within the range of those of full plant genomes. While grapevine is not a natural host for many of the chosen pathogens, it simply serves as an example of background sequences in which the target pathogen sequences exist.

Experimental Flow

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

E-Probe Design

Pathogen-specific sequence queries were designed using a modified version of the Tool for Oligonucleotide Fingerprint Identification (TOFI) (Vijaya Satya *et al.*, 2008). The basic TOFI pipeline includes three basic steps: comparison of pathogen sequences with those of near neighbors, thermodynamics optimization, and a BLAST search check for uniqueness. The EDNA

query design process is similar, with the following changes. For in silico querying, the e-probe thermodynamics optimization step is omitted because the thermodynamic properties of the unique sequences are irrelevant. Parameters of interest to a BLAST search and/or important to a successful NGS run were added in its place. In the BLAST parameter step, the query sequence length was restricted to standardize e-values from the BLAST search and the candidate e-probes containing a homo-oligomer (five or more of the same nucleotide in tandem) were removed because of the inherent miscalling of homo-oligomers in many NGS platforms. To test the optimal length of e-probes the BLAST check step was omitted, and the preliminary e-probes were used in the optimization of e-probe length. After optimization of e-probe length, a BLAST check and manual editing were reintroduced to assure specificity (Table 1). Any e-probes that hit a species different than the target with an E-value of $1x10^{-10}$ or below were removed from the final e-probe set.

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

Mock Database Construction

To test the designed queries, a data set consisting of both known host and pathogen genome segments was generated. Simulation of massively parallel sequencing was performed using MetaSim software (Richter *et al.*, 2008). The simulation includes planned mistakes in base

calling, as well as a range of read lengths, both of which are common for 454, or Illumina pyrosequencing. The resulting database contained 10,000 simulated reads, each approximately 400 ± 30 nucleotides, or 62 nucleotides, respectively. Abundance values (representing the given amount of nucleic acid within a sample) for host genomic sequences were set at a default of 100, while host mitochondrial and chloroplast sequences were given an abundance value of 1000, meaning that for every genomic sequence there will be 10 mitochondrial and chloroplast sequences. This value was chosen arbitrarily. Pathogen abundance values were varied to generate a number of reads corresponding to the percent of the database that is made up of pathogen sequences (i.e. 25% pathogen sequences is equivalent to 2500 pathogen reads in a 10,000 read database). The databases were placed into categories based on the pathogen sequence percentage: those with 15-25% pathogen sequences were considered high, with 5-15% medium, with 0.5-5% low, and with less than 0.5% very low. These percentages were chosen arbitrarily.

Querying Mock Databases

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

The decision to designate a sample as positive or negative for a pathogen is crucial for any diagnostic assay. The criterion used to determine a positive sample in this assay was the presence of pathogen-specific sequences. It was likely that many of these sequences would be

similar to sequences that belong to either the plant host, or to a different microbe that resides in the sample. Each e-probe set is designed to be unique to a specific pathogen. The signals of these sets were compared to the signals of decoy sets which represents background signals. To generate a decoy set of e-probes, the designed target set of e-probes was reversed in sequence. Each set was then used as queries in a BLASTn search against the MSD. Each probe in both sets was given a score based on the e-value and the percent coverage of the top n hit(s), where n equals [50, 10, 5, 1] (Equation 1, where n is the hit number, Eval is the e-value of the nth hit, and %cov is the percent of the e-probe contributing to the high scoring segment pairing.).

$$\sum_{h=1}^{n} -logEval[h] * (\%cov.[h])$$

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

Results

Plant pathogenic query production was analyzed in relation to genome size for two viruses, five bacteria, two fungi and one stramenopile. The targeted viral (*Plum pox virus* and *Bean golden mosaic virus*), fungal (*Puccinia graminis* and *Phakopsora pachyrhizi*) and stramenopile (*Phytophthora ramorum*) plant pathogens were compared to near neighbors of the same species. For the bacteria, the Ca. *Liberibacter asiaticus* near neighbor was from the same species, while those of the other 3 bacteria were from a closely related species (*X. oryzae* paired with *X. fastidiosa* and vice versa). Fungal pathogens *Puccinia graminis* and *Phakopsora pachyrhizi* had the same near neighbor, *Puccinia triticina*. In addition, *P. pachyrhizi* was found to be broadly similar in biological attributes to *P. triticina* (Pivonia and Yang 2006). In the case of

Phytophthora ramorum, *P. infestans* was used as near neighbor (Table 1). The lack of a spiroplasma related to *S. citri* resulted in the selection of a near neighbor that was related at the order level (Table 1). The genome sizes of the pathogens used ranged from 5.23 knt to 88 Mnt, and the number of queries ranged from 4 to 21,790. As the genome size of the plant pathogen increased so did the total number of queries for the targeted pathogen. The total length of the combined e-probes was proportional to the total number of e-probes, and to the genome size. The percentage of genome covered ranged from 1.74 to 6.57 without any correlation with genome size or total query number (Table 1).

The number of hits at a threshold of $1x10^{-3}$, $10x^{-6}$, or $10x^{-9}$ received for each pathogen was determined (Figure 2-4). The number of hits rose with the size of the pathogen genome. As expected, the number of hits increased with increasing pathogen proportions. At lower proportions, there was an increase in the standard deviation of the number of hits. A general similarity of the number of hits can be seen for each pathogen type, with prokaryotic pathogens having the greatest variability across pathogens.

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

former. For the viral pathogens a match was found for every query available in high, normal and low pathogen abundance MSDs, and the number of matches in very low abundance MSDs was approximately half of the number of available queries (2 matches/ 4 e-probes in the case of BGMV) (Figures 5-7, Table 1).

Optimization of E-probe Length

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

The effect of varying e-probe lengths from 20 – 140 nt on the matches generated by searches on the MSDs was determined. As expected, for each pathogen, match numbers decreased as the length of the e-probes increased, because the number of longer e-probes designed was much lower than that for shorter e-probes. In general, each pathogen type (virus, bacterial, and eukaryotic) had a similar number of matches for each member within a group (Figures 5-7). One exception was *X. oryzae*, which showed no such downward trend (Figure 6). Almost all pathogens were detected using every query length. The other exception was *R. solanacearum* in very low pathogen abundance MSDs, in which an average of a single match was found for the majority of query lengths (40, 80, 100, 120, and 140 nt). *P. ramorum* and *P.*

graminis showed the smallest number of matches of all the pathogens when very low pathogen proportion MSDs were queried with 140 nt e-probes. This low number of matches could be due to the random selection of sequences when constructing MSDs because fungal and stramenopile genomes are larger than viral and bacterial genomes, allowing the presence of portions of the genome in the MSDs that have a low density of e-probe sequences. This phenomenon is most likely to occur for low pathogen proportions and large pathogen genomes.

E-value Threshold

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

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

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

BLAST Check Comparison

False positives were reduced by removing e-probes that have similarity to known sequences in NCBI. Each 80 nt e-probe set was used as queries in a search against the NCBI GenBank nt database. E-probes with hits at an e-value of $1x10^{-10}$ or lower were removed from the probe set. This decreased the number of probes per set by up to 50% (Table 1). Comparing the performance of BLAST-checked e-probe sets with probes not checked with BLAST showed a slight reduction in the number of false positive hits, with a larger reduction in the number of matches and total hits (Supplemental Table 1).

Determination of Positive and Negatives

Using the above method, we were able to correctly call samples positive for all positive samples except for those at a very low abundance (<0.5% pathogen reads) (Table 3). At this abundance there were mixed results, at times calling the sample positive while other times calling

it negative. *R. solanacearum* was not detected at very low abundance MSDs. Pathogen negative MSDs (MSDs without pathogens) were all negative or suspect for viruses, *S. citri*, and *R. solanacearum*. False positives were most common in eukaryotic pathogens. When the number of top hits (n in equation 1) was lowered in the scoring step, the pathogen negative MSDs were correctly identified in some, but not all, replicates (Table 3).

Discussion

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

The first detriment to adopting metagenomics-based diagnostics is the current per run cost. The typical approach to a metagenome diagnosis is nucleic acid extraction, sequencing, sequence assembly, and BLAST analysis of the assembled contigs. An examination of recent history suggests that sequencing technologies will likely become less expensive, faster and more accessible, and processive over time, outpacing Moore's Law, suggesting that NGS costs may not be a long term restraint, particularly when combined with barcoding (Parameswaran *et al.*, 2007). However, the very same advances that drive down per sample costs of sequencing create additional data handling problems. As NGS becomes less expensive, faster and the length of reads increases the number of bases sequenced in a single run will increase exponentially. These

same advances in NGS will have an additional exponential growth effect on the databases (i.e. GenBank and its subsidiaries) that are for the BLAST searching of sequence data, suggesting that the current metagenomic approach to pathogen diagnostics will eventually become too computationally intensive for everyday use.

The objective of this work was to find a simplified bioinformatic approach for dealing with the exponential growth and complexity of NGS metagenome data, which could be handled on a standard personal computer without extensive computational delays. To do this, we developed a protocol (EDNA) in which the input NGS data would be treated as the searchable database, and this sequence database would be queried by diagnostic signature sequences (e-probes) without the need for assembly or quality checks. This approach allows the user to limit and control both the size of the searchable database and the size of the searching query set.

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

detection could be improved to suit user needs by adjusting the number of e-probes, the length of the e-probes and/or the parsing E-value.

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

A second approach for improving specificity involved improving the screening of potential e-probes. Clearly, as genome size increases the number of e-probes generated increases in proportion. Removal of a number of e-probes from the larger pathogen genome screens would likely not affect the overall limit of detection. The e-probes from all pathogens were searched against GenBank, as is done in primer selection, to eliminate a number of false positive generating e-probes. This strategy may be of limited use for plant pathogens, however, as the majority of environmental microbes in a typical plant metagenome have no GenBank counterpart.

The addition of a healthy control BLAST, searching healthy control asymptomatic host environmental sample sequence databases for the presence of potential false positive queries might eliminate some e-probes that would react to host or endophyte sequences not available in GenBank. Regardless, much like limit of detection, EDNA precision could be adjusted up or down as needed in the e-probe design (by adjusting e-probe length or near neighbor selection) or during database searching (adjusting E-value threshold).

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

diagnostic assays, the balance between specificity and limit of detection is a necessity in this bioinformatics approach to diagnostics.

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

Acknowledgments

This work was funded by the USDA-CSREES Plant Biosecurity Program, grant number 2010-85605-20542. The authors would like to thank Dr. Rakesh Kaundal for the critical review of this manuscript. The *Phytophthora ramorum* sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/.

REFERENCES

Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M. & Boonham, N. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Mol. Plant. Pathol.* **10**, 537-545.

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

Cox-Foster, D. L., Conlan, S., Holmes, E. C., Palacios, G., Evans, J. D., Moran, N. A., Quan, P.-L., Briese, T., Hornig, M., Geiser, D. M., Martinson, V., vanEngelsdorp, D., Kalkstein, A. L., Drysdale, A., Hui, J., Zhai, J., Cui, L., Hutchison, S. K., Simons, J. F., Egholm, M., Pettis, J. S. & Lipkin, W. I. (2007). A Metagenomic Survey of Microbes in Honey Bee Colony Collapse Disorder. *Science* **318**, 283-287.

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

Gamliel, A., Gullino, M. L. & Stack, J. P. (2008). Crop Biosecurity: Local, National, Regional and Global Perspectives. In *Crop Biosecurity*, pp. 37-61. Edited by M. L. Gullino, J. Fletcher, A. Gamliel & J. P. Stack: Springer Netherlands.

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

Jones, W. (2010). High-Throughput Sequencing and Metagenomics. *Estuaries and Coasts* **33**, 944-952.

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

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F. & Brandi, M. L. (2010). Bioinformatics for next generation sequencing data. *Genes* **1**, 294-307.

Metzker, M. L. (2009). Sequencing technologies—the next generation. *Nat Rev Genet* **11**, 31-46.

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

Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., Briese, T., Conlan, S., Quan, P.-L., Hui, J., Marshall, J., Simons, J. F., Egholm, M., Paddock, C. D., Shieh, W.-J., Goldsmith, C. S., Zaki, S. R., Catton, M. & Lipkin, W. I. (2008). A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases. *New Engl J Med* **358**, 991-998.

Parameswaran, P., Jalili, R., Tao, L., Shokralla, S., Gharizadeh, B., Ronaghi, M. & Fire, A. Z. (2007). A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* **35**, e130.

Pop, M. & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet* **24**, 142-149.

Postnikova, E., Baldwin, C., Whitehouse, C. A., Sechler, A., Schaad, N. W., Sampath, R., Harpin, V., Li, F., Melton, R., Blyn, L., Drader, J., Hofstadler, S. & Schneider, W. L. (2008). Identification of Bacterial Plant Pathogens Using Multilocus Polymerase

Chain Reaction/Electrospray Ionization-Mass Spectrometry. *Phytopathology* **98**, 1156-1164.

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

Richter, D. C., Ott, F., Auch, A. F., Schmid, R. & Huson, D. H. (2008). MetaSim—A Sequencing Simulator for Genomics and Metagenomics. *PLoS One* **3**, e3373.

Ronaghi, M. (2001). Pyrosequencing Sheds Light on DNA Sequencing. *Genome Res* **11**, 3-11.

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

Schaad, N. W., Frederick, R. D., Shaw, J., Schneider, W. L., Hickson, R., Petrillo, M. D. & Luster, D. G. (2003). Advances in Molecular-based Diagnostics in Metting Crop Biosecurity and Phytosanitary Issues. *Annu Rev Phytopathol* **41**, 305-324.

Tringe, S. G. & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples.

Tucker, T., Marra, M. & Friedman, J. M. (2009). Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. *The American Journal of Human Genetics* **85**, 142-154.

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

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

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

Name	**Source**	**Near Neighbor**	**Source**	**Original Sequence Size (kb)**	**# 80 bases e-probes preliminary *(BLAST check)***	**Total probe length (kb)**	**Genome % coverage**
Bean golden mosaic virus	NC_004042 NC_004043	Abutilon mosaic virus	NC_001928 NC_001929	5.23	4 *(2)*	0.32 *(0.16)*	6.12% *(3.06%)*
Plum pox virus	NC_001445	Pepper mottle virus	NC_001517	9.74	8 *(5)*	0.64 *(0.40)*	6.57% *(4.11)*
Spiroplasma citri	115252846 110005886 110005766 110005758 11000748 110005735 110005716 110005696 110005687 110005683 110005675 110005664 110005652 110005641 110005622 110005605 110005592 110005560 110005522 110005436 110005327 110005285	*Mycobacterium bovis*	NC_008769	1525.76	423 *(309)*	33.84 *(24.72)*	2.22% *(1.62%)*

	110005260 110005199 110005145 110005138 110005098 110005060 110005027 110004948 110004868 110004796 110004744 110004631 110004607 110004455 110004127 110004055 110003907M						
Ca. L. asiaticus	NC_012985	*Agrobacterium tumefaciens*	AE007869	1226.70	502 *(469)*	40.16 *(37.52)*	3.27% *(3.06%)*
Xanthamonas oryzae	CP000967	*Xylella fastidiosa*	NC_002488 NC_002489 NC_002490	2679.31	2597 *(1832)*	207.76 *(146.56)*	7.75% *(5.47%)*
Xylella fastidiosa	NC_002488 NC_002489 NC_002490	*Xanthomonas oryzae*	CP000967	5240.08	1459 *(1041)*	116.72 *(83.28)*	2.23% *(1.59%)*
Ralstonia solanacearum	NC_003295 NC_003296	*Ralstonia pickettii*	NC_010682 NC_010678 NC_010683	3716.41	1964 *(1418)*	157.12 *(113.44)*	(4.23%) *(3.05%)*
Puccinia graminis	AAWC01000001 AAWC01004563	*Puccinia triticina*	ADAS01000001 ADAS01038776	66652.40	21790 *(21635)*	1743.20 *(1730.80)*	2.66% *(2.65%)*
Phytophora ramorum	AAQX01000001 AAQX01007589	*Phytophora infestants*	AATU01000001 AATU01018288	88644.63	21286 *(18945)*	1702.88 *(1515.60)*	1.92% *(1.71%)*

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

Name	E-probe length	15-25%	5-15%	0.5-5%	<0.5%
BGMV	20	100	100	100	100
	40	100	100	100	100
	60	100	99.97	100	100
	80	100	100	100	100
	100	100	100	100	100
	120	100	100	100	100
	140	100	100	100	100
PPV	20	100	100	100	100
	40	100	100	100	100
	60	100	100	100	100
	80	100	100	100	100
	100	100	100	100	100
	120	100	100	100	100
	140	100	100	100	100
Spiro	20	97.66	94.32	80.38	33.36
	40	98.89	98.14	91.37	51.1
	60	98.94	98.75	93.91	54.44
	80	99.56	99.38	96.2	78.59
	100	99.73	99.03	93.37	72.44
	120	99.78	99.28	97.4	68.33
	140	99.53	98.84	99.02	63.89
Liberibacter	20	98.97	98.31	92.42	55.58
	40	99.48	99.27	96.35	54.79
	60	99.26	98.72	96.42	62.05
	80	99.74	99.84	98.06	81.24
	100	99.63	99.05	96.44	63.49
	120	99.49	99.33	97.17	57.08
	140	99.33	99.12	96.47	40.12
Xanthomonas	20	99.96	100	99.58	84.2
	40	100	99.78	99.58	87.91
	60	99.95	99.81	99.51	84.21
	80	99.93	99.95	99.87	93.72
	100	99.98	99.89	99.87	93.91
	120	99.9	99.89	99.86	94.57
	140	99.98	99.95	99.87	100
Xylella	20	99.96	99.83	99.39	98.1
	40	99.97	99.87	100	97.09
	60	99.93	99.52	99.72	96.41
	80	99.91	99.71	99.68	94.98
	100	99.86	99.67	99.63	94.42
	120	99.89	99.61	99.56	93.07
	140	99.87	99.53	99.52	93.07
Ralstonia	20	100	98.89	99.52	97.94
	40	99.91	99.83	99.42	95.38
	60	99.90	99.87	98.78	93.10
	80	100	100	99.42	92.86

	100	100	100	99.02	90.91
	120	100	100	98.57	75.00
	140	100	100	98.00	75.00
Phytophthora ramorum	20	99.45	98.95	96.41	24.78
	40	99.75	99.57	97.66	30.58
	60	99.66	99.37	95.68	14.14
	80	99.76	99.68	98.52	48.94
	100	98.04	100	100	100
	120	99.75	99.26	98.11	45.45
	140	99.43	99.22	95.77	28.57
Puccinia graminis	20	98.28	96.52	87.8	30.54
	40	99.36	98.65	94.12	44.22
	60	99.17	97.87	92.69	35.86
	80	99.69	99.35	97.77	56.9
	100	99.71	99.2	98.5	60.78
	120	99.75	99.28	98.07	66.67
	140	99.91	99.45	98.21	57.14

Table 3: P-values of EDNA diagnostic call

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

	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.586	0.057	0.025	0.256	0.656	0.208
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.081	0.012	0.223	0.105	0.448	0.231
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.073	0.008	0.067	0.218	0.953	0.392
X. oryzea	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.060	0.811	0.002	0.000	0.000	0.000
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.824	0.173	0.650	0.000	0.001	0.002
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.004	0.074	0.521	0.157	0.398
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.001	0.033	0.016	0.016	0.089
X. fastidiosa	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.745	0.306	0.025	0.316	0.222	0.271
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.018	0.003	0.000	0.006
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.007	0.004	0.000	0.027
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.026	0.031	0.001	0.514
P. graminis	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.001	0.000	0.000	0.000
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.333	0.428	0.894	0.413	0.009	0.020
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P. ramorum	Top 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.083	0.508	0.000	0.000	0.000
	Top 10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.479	0.049	0.000	0.014	0.000
	Top 5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.350	0.004	0.000	0.338	0.007	0.019
	Top 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.257

CHAPTER V

E-PROBE DIAGNOSTIC NUCLEIC ACID ASSAY (EDNA): A USEFUL TOOL FOR SCREENING METAGENOMIC DATA FOR VIRUSES OF INTEREST.

Abstract

Next Generation Sequencing (NGS) is not commonly used in diagnostics, possibly due to the large amount of time and computational power needed to identify each sequence in a NGS data set. By using pathogen specific sequences, termed e-probes, as queries in a search of unassembled sequence data; it is possible to identify a specific virus in a plant sample. This method, designated E-probe Diagnostic Nucleic acid Assay (EDNA) has been tested with both DNA (*Bean golden mosaic virus*), and RNA (*Plum pox virus*) virus infected plant material. In addition, the ability to detect and differentiate among strains of a single virus species is shown by using probe sets that are specific to the strain. Multiple viruses in plant samples can also be identified as long as probe sets for each virus are used. The EDNA pipeline is over 2400 times faster than a traditional metagenomic analysis, a feature desirable in a diagnostic setting.

Introduction

The global trading of plant material has increased the introduction of foreign plant diseases in the last few decades (Tatem *et al.*, 2006), leading to a need for enhanced surveillance and detection of pathogens in imported plants. Methods currently used to detect pathogens within imported plant material are visual, nucleic acid based, and protein based. Examples of protein based assays such as ELISA, western blots, and immuno-strip tests are not easily multiplexed easily to test for several pathogens simultaneously, but proteomic methods such as 'mud-pit,' have been used to test for multiple pathogens. Of the two major nucleic acid based assays, real time PCR and microarrays, only microarrays are readily multiplexed (Call *et al.*, 2003; Iqbal *et al.*, 2000; Lazcka *et al.*, 2007; Ye *et al.*, 2010).

With the advent of metagenomics, the ability to obtain sequence information on the entire organismal makeup of a sample has become commonplace. This advance has led to the identification of previously unknown species of microorganisms, as well as offered insights into their ecological distribution. Using metagenomics, multiplexing introduces a problem of an overwhelmingly large amount of data, commonly referred to as "big data", a term that refers to both the movement of data from one server to another and the analysis of large data sets.

Metagenomics is reliant on Next Generation Sequencing (NGS), a powerful technology that allows the acquisition of hundreds of thousands of short sequence reads from the majority, if not all, of the organisms within a given sample (Gilbert & Dupont, 2011; Willner & Hugenholtz, 2013). This immense sequencing capability can be a boon to diagnosticians who are interested in the detection and identification of specific pathogens. NGS, which has been used to identify pathogens in various systems (Adams *et al.*, 2009; Koonin & Dolja, 2012; Roossinck, 2012), has the advantage of being able to detect and identify many different pathogens within a sample; however, two significant draw backs that have kept this technology from being used for

diagnostic purposes are the length of time and amount of computational power needed to compare the short sequence reads to known sequences.

The speed at which sequence data is generated and placed in curated sequence databanks has been increasing, and will likely continue to do so (Hsi-Yang Fritz *et al.*, 2011; Kodama *et al.*, 2012). Steps have been taken to increase the efficiency of search algorithms (Li & Homer, 2010). One of these steps, the E-probe Detection Nucleic acid Assay (EDNA) pipeline (Stobbe *et al.*, 2013), uses short pathogen-specific sequences as queries against raw sequence data from a given sample. These short sequence queries, termed e-probes, allow the user to choose only the pathogens of interest and thus reduce the time needed to detect and identify a pathogen. In the work described herein, it is shown that the EDNA pipeline performs as well as a traditional metagenomic pipeline with respect to detection, and surpasses it in terms of computational speed. In addition, the ability to differentiate between closely related strains of viral pathogens has been shown using *Plum pox virus* (PPV) as an example.

Materials and Methods

E-probe design

For the detection of virus sequences in a metagenomic sample, pathogen-specific sequences were identified using a modified version of the microarray probe software Tool for Oligonucleotide Fingerprint Identification (TOFI) (Satya *et al.*, 2008). The thermodynamic determinants of TOFI were removed because e-probes are character strings, and will not be converted to oligonucleotides. The EDNA version of TOFI works in two steps. First, the target sequences are compared to near neighbor sequences using the Nucmer script in the Mummer software package (Delcher *et al.*, 2003). Sequences having similarity to the near neighbors were removed, leaving only unique target sequences, which are used as queries against the NCBI non-redundant nucleotide database to ensure specificity to the target organism. Any candidate probe which received a hit with an e-value of $1x10^{-9}$ or lower to any organism that does not share a

name with the target was removed. The same modified pipeline was used in the initial testing of EDNA (Stobbe *et al.*, 2013), with the following difference. The lengths of the e-probes were not limited to a specific size, but were instead allowed to vary between 30 nt and infinity. A decoy set of e-probes were generated by using the reverse sequence of each e-probe.

The target pathogens *Bean golden mosaic virus* (BGMV; NC_004042.1, NC_004043.1) and *Bean golden yellow mosaic virus* (BGYMV; NC_001438.1, NC_001439.1) were compared to the near neighbor *Abutilon mosaic virus* (NC_001928.2, NC_001929.2). PPV (NC_001445.1) was compared to it's near neighbor *Pepper mottle virus* (NC_001517.1). Five PPV strains, C, D, EA, M, and W, were used in the design of the e-probe sets (Maiss *et al.*, 1989; Matic *et al.*, 2011; Nemchinov & Hadidi, 1996). Each strain was considered as a target pathogen, with all other strains considered as near neighbors, for a total of 5 e-probe sets (Table 1). Otherwise, the design of strain specific e-probes was the same as described above.

In silico testing of the specificity of the strain specific e-probes was carried out as previously described (Stobbe *et al.*, 2013). In this test each set was able to correctly identify the strain for which it was designed. Surprisingly, when mock sample databases were generated using PPV isolates of the Rec strain (a recombinant of strains D and M) and queried with each strain-specific e-probe set, only the M e-probe set gave a positive diagnostic call. For the BGMV and BGYMV probe sets, any probes that gave false positives in the *in silico* testing were removed from the sets.

Whole transcriptome amplification and 454 Jr. sequencing

Total nucleic acid was extracted from plant tissue in order to detect both RNA and DNA viruses. Total nucleic acids extracted of BGMV-infected bean were generously provided by Dr. Judith Brown. Total nucleic acids of leaf discs of *Prunus persica* infected with PPV were obtained as described (Wallis *et al.*, 2007). Four samples of PPV-infected tissue were used, two

of which were infected with the D strain of PPV, one with the M strain and another with the EA strain. The presence of the viruses was confirmed with real time-PCR as described (Schneider *et al.*, 2002). Each total nucleic acid sample was amplified using a Sigma Whole Transcriptome Amplification Kit, as per the manufacturer's instructions, followed by size-selection using AMPure beads (New England BioLabs Inc.). The resulting cDNA library was sequenced using the Roche 454 Jr. platform, excluding nebulization.

The number of pathogen sequences within each sample was first enumerated by querying the raw sequencing results against the pathogen's genome, in order to determine the level of detection. The sequencing results were then analyzed using two methods. The first was a "traditional" bioinformatic approach to NGS data, which includes trimming and filtering the sequence reads to remove portions of poor quality, followed by de novo assembly of the sequence reads into contigs, and then query of the contigs against the NCBI non-redundant database, and parsing of the results of the query. For the "traditional" approach, the trimming and filtering was performed in the iPlant discovery environment (Goff *et al.*, 2011) using the FASTX Trimmer and FASTX Quality Filter. The assembly of the contigs was performed using the Roche *de novo* Assembler. Querying the non-redundant database was performed with the mpiBLAST software (Darling *et al.*, 2003) on the Pistolpete high performance computing cluster. The MEGAN software package was used in the identification of organisms that contributed to the metagenome (Huson *et al.*, 2007).

The second analysis method used the EDNA pipeline. The FASTA file was extracted from the output file of the Roche 454 Jr. sequencing (SFF file), and the sequencing primers from the 5' and 3' ends were trimmed. This FASTA file then served as a database and was queried using the previously designed e-probe sets, both target and decoy. The BLAST result was then parsed and scored using the following equation, in which n is the number of top hits to use, Eval is the e-value of the hit, and the %cov. is the percent coverage of the e-probe used in the hit.

$$\sum_{h=1}^{n} -logEval[h] * (\%cov.[h])$$

The target e-probe scores were compared to the decoy scores using two statistical tests. The first was a simple t-test. The second found the average and standard deviation of the decoy scores, and called a probe positive if the target score was more than 10 standard deviations above the average. This two-pronged strategy offers two ways to view the results. The t-test offers a view of the entire e-probe set, while the standard deviation offers a probe by probe view. A p-value of less than 0.05 was considered to be positive for pathogen presence, while a p-value of less than 0.1 and greater than 0.05 was considered suspect.

Each analysis was performed on a high performance computer cluster, which consists of 252 standard compute nodes, each with dual Intel Xeon E5-2620 "Sandy Bridge" hex core 2.0 GHz CPUs, with 32 GB of 1333 MHz RAM or using the iPlant Discovery Environment (Goff *et al.*, 2011). The run time of each was recorded (Table 2). The time taken to move data was not included.

Results

The sequencing files are summarized in Table 3. Samples from plants infected with PPV strains were barcoded and sequenced on two 454 Jr. plates (PPV-DT0, PPV-M paired on one plate, PPV-EA, PPV-DT4 paired on the other), while the BGMV sample was sequenced on a single plate. These sequence datasets consist of between 9,250 and 45,295 reads, with a range of average read lengths (296-412 nt). The percent of the reads that matched to known pathogen in a BLAST search ranged from 0.35% to 6.80%. The average percent of pathogen reads was much lower in the PPV samples than in the BGMV sample, with the exception of PPV-EA (Table 3).

Traditional metagenomic approach

Traditional metagenomic analysis was able to identify each of the pathogens whose sequences were known to be present in the data samples, as well as the percent of the metagenome to which the pathogen contributed (Figure 1). The analysis of the BGMV data shows that the third and sixth most prevalent organisms were BGYMV (5.05%), and BGMV (0.75%) respectively. Analysis of data samples of PPV strains showed each strain's presence at various levels (0.35-6.80%). A strain was called in the two cases (PPV strain M and strain D). Using the Investigate option of the MEGAN software, all of the PPV samples were identified at the strain level. Identification of the host species, however, was unsatisfactory with many of the reads being assigned to the wrong family, order, and phylum.

EDNA pipeline approach

After processing the raw sequence files with the EDNA pipeline, each of the probe sets successfully detected the presence of virus pathogens in each data set, based on results of the two tests mentioned above. When the PPV e-probes were used to query the PPV-EA sample with fifty top hits, the sample received a negative diagnostic call p-value (Table 4). In addition, the BGMV and BGYMV genomes were used reference sequences and the reads of the 454 sequence data were mapped to each genome. 0.8% of the reads mapped to BGMV, while 4.8% of the reads mapped to BGYMV. Interestingly, only 86.1% of the BGYMV DNA B segment was mapped. A large number of high quality variants were found for the BGYMV reference (81 variants for DNA A, 178 for DNA B), when compared to BGMV (3 and 0 respectively).

The EDNA analysis requires both fewer steps and less time (avg.14 seconds) when compared to the "traditional" metagenomics approach (average of 9.2 hours) (Table 2), making the EDNA analysis over 2400 times faster than the "traditional" approach. When EDNA was run

on a laptop computer, obtaining a diagnostic call was obtained as quickly as on a computing cluster.

Strain-specific e-probes

Strain-specific e-probes were used as queries to the sample datasets shown in Table 5, using the same method described above. The strain-specific e-probes are less specific than the genus level e-probes, but still are able to differentiate between the strains. The C, EA, and M strain e-probe sets gave non-specific association of e-probes, while the D set gave a suspect call on a positive sample. Due to the prevalence of both D and M strain recombinants found in Europe, it is beneficial to know the genomic locations of the strain specific e-probes. Mapping the positions of the e-probes onto the PPV genome shows that each set of the strain-specific e-probes span across the entirety of the genome.

Discussion

NGS offers a powerful tool for diagnostics. The ability to obtain sequence information from every organism within a sample gives an in depth look at what microorganisms may be associated with a disease. As with many other processes using large datasets, the computational power and time needed to analyze the datasets are extremely large, and are therefore limiting in diagnostics, for which a diagnostic call is required in a timely manner. The EDNA pipeline can give a diagnostic call over 2400 times faster than a metagenomic approach because it searches only for the presence of sequences specific to the pathogen of interest, and ignores other organisms. The EDNA pipeline's output is a simple list of pathogens tested, a p-value, the number of positive probes, and a diagnostic call, while the metagenomic approach requires a subjective diagnostic call by a user, based on a review of the results, which in turn can lead to variability between users and limit desirable automation of the diagnostic process.

For many diagnosticians access to supercomputers is limited or nonexistent, but the time to a diagnostic call by EDNA, performed using a laptop, is still 2400 times faster than a metagenomic approach on a supercomputer. The EDNA approach can be used with minimal computational resources. The fact that it also does not require a curated database removes the need to maintain an updated copy of the curated database for the analysis, although it is still needed for the e-probe design process. As sequencing becomes more common, the amount of data stored in these curated databases will grow, and the length of time to query these databases will likewise grow. By moving the diagnostic process to a local machine, the EDNA pipeline removes the need for users to have access to supercomputers, curated databases, or networks, a feature that facilitates it's applications in the developing world, as well as during times when access to supercomputers or curated databases has been frequently interrupted.

A multitude of sequencing platforms is available. Which, if any, of these will be used in the future is unknown, but one commonality among each of these platforms is their output of sequence reads in fastA format. The EDNA pipeline was designed for use with any sequencing platform once the data is in fastA format, which can be then formatted into a local BLAST database. Regardless of sequencing platform, the EDNA pipeline can be used to identify specific pathogens within a metagenomic dataset. The EDNA pipeline is flexible and will remain relevant as sequencing technology grows.

Our data shows that EDNA is clearly more effective as a diagnostic tool than traditional metagenomic approaches. However, the true comparison for EDNA as an everyday diagnostic tool isn't really other metagenomics approaches, but the more conventionally used plant virus diagnostic tools of PCR, reverse-transcription PCR and ELISA. The obvious advantage to PCR and ELISA based virus detection are the limited costs per assay and the lack of any need for bioinformatic comparisons. These widely used detection tools are very much limited to answering diagnostic questions for well characterized viruses that the investigator knows to look for.

Certainly these are preferred for wide scale screenings where the virus of interest is already identified. However, in cases where an unexpected infected plant sample reaches a diagnositic lab without any *a priori* knowledge of etiology, it could take hundreds of individual viral assays to determine a potential cause. A single EDNA analysis could provide a much quicker answer. While the price constraints of NGS are currently limiting for everyday survey use, EDNA based metagenome analysis may already be a viable option in areas where agricultural products of interest need to be tested for multiple pathogens, preferably simultaneously, like a import quarantine facility. In addition, it's logical to assume that the greater scientific community drive for cheaper and faster sequencing technologies will only serve to drive down the costs of NGS metagenome based viral discovery and diagnostics in the future.

Acknowledgements

This work was supported in part through instrumentation funded by the National Science Foundation through grant OCI–1126330.

REFERENCES

Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M. & Boonham, N. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Mol. Plant. Pathol.* **10**, 537-545.

Call, D. R., Borucki, M. K. & Loge, F. J. (2003). Detection of bacterial pathogens in environmental samples using DNA microarrays. *J. Microbiol. Methods* **53**, 235-243.

Darling, A., Carey, L. & Feng, W.-c. (2003). The design, implementation, and evaluation of mpiBLAST. *Proceedings of ClusterWorld* **2003**.

Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. (2003). Using MUMmer to identify similar regions in large sequence sets. *Current Protocols in Bioinformatics*, 10.13. 11-10.13. 18.

Gilbert, J. A. & Dupont, C. L. (2011). Microbial metagenomics: beyond the genome. *Annual review of marine science* **3**, 347-371.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., Muir, A., Merchant, N., Lowry, S., Mock, S., Helmke, M., Kubach, A., Narro, M., Hopkins, N., Micklos, D., Hilgert, U., Gonzales, M., Jordan, C., Skidmore, E., Dooley, R., Cazes, J., McLay, R., Lu, Z., Pasternak, S., Koesterke, L., Piel, W. H., Grene, R., Noutsos, C., Gendler, K., Feng, X., Tang,

C., Lent, M., Kim, S.-j., Kvilekval, K., Manjunath, B. S., Tannen, V., Stamatakis, A., Sanderson, M., Welch, S. M., Cranston, K., Soltis, P., Soltis, D., O'Meara, B., Ane, C., Brutnell, T., Kleibenstein, D. J., White, J. W., Leebens-Mack, J., Donoghue, M. J., Spalding, E. P., Vision, T. J., Myers, C. R., Lowenthal, D., Enquist, B. J., Boyle, B., Akoglu, A., Andrews, G., Ram, S., Ware, D., Stein, L. & Stanzione, D. (2011). The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Frontiers in Plant Science* **2**.

Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G. & Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* **21**, 734-740.

Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**, 377-386.

Iqbal, S. S., Mayo, M. W., Bruno, J. G., Bronk, B. V., Batt, C. A. & Chambers, J. P. (2000). A review of molecular recognition technologies for detection of biological threat agents. *Biosensors and Bioelectronics* **15**, 549-578.

Kodama, Y., Shumway, M. & Leinonen, R. (2012). The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res* **40**, D54-D56.

Koonin, E. V. & Dolja, V. V. (2012). Expanding networks of RNA virus evolution. *BMC biology* **10**, 54.

Lazcka, O., Campo, F. & Muñoz, F. X. (2007). Pathogen detection: A perspective of traditional methods and biosensors. *Biosensors and Bioelectronics* **22**, 1205-1217.

Li, H. & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* **11**, 473-483.

Maiss, E., Timpe, U., Brisske, A., Jelkmann, W., Casper, R., Himmler, G., Mattanovich, D. & Katinger, H. (1989). The complete nucleotide sequence of plum pox virus RNA. *J Gen Virol* **70**, 513-524.

Matic, S., Elmaghraby, I., Law, V., Varga, A., Reed, C., Myrta, A. & James, D. (2011). Serological and molecular characterization of isolates of Plum pox virus strain El Amar to better understand its diversity, evolution, and unique geographical distribution. *J Plant Pathol* **93**, 303-310.

Nemchinov, L. & Hadidi, A. (1996). Characterization of the sour cherry strain of plum pox virus. *Phytopathology* **86**, 575-580.

Roossinck, M. J. (2012). Plant Viral Metagenomics. *Annu Rev Genet* **46**.

Satya, R. V., Zavaljevski, N., Kumar, K. & Reifman, J. (2008). A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. *BMC bioinformatics* **9**, 185.

Schneider, W., Sherman, D., Stone, A., Damsteegt, V. & Frederick, R. (2002). Specific Detection and Quantification of Plum Pox Potyvirus by Real-Time Fluorescent Rt-Pcr. *American Phytopathological Society Abstracts.*

Stobbe, A. H., Daniels, J., Espindola, A., Verma, R., Melcher, U., Ochoa-Corona, F., Garzon, C., Fletcher, J. & Schneider, W. L. (2013). E-probe Diagnostic Nucleic acid Analysis (EDNA): A theoretical approach for handling of next generation sequencing data for diagnostics. . *J. Microbiol. Methods* **in submission**

Tatem, A. J., Hay, S. I. & Rogers, D. J. (2006). Global traffic and disease vector dispersal. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6242-6247.

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

Willner, D. & Hugenholtz, P. (2013). From deep sequencing to viral tagging: Recent advances in viral metagenomics. *BioEssays*.

Ye, Y., Mar, E.-C., Tong, S., Sammons, S., Fang, S., Anderson, L. J. & Wang, D. (2010). Application of proteomics methods for pathogen discovery. *J Virol Methods* **163**, 87-95.

Table 1: List of strain isolates used including accession number.

Isolate name	Strain	Accession
SC	C	X81083.1
SwC	C	Y09851.2
PENN1	D	AF401295.1
PENN2	D	AF401296.1
Cdn 4	D	AY953263.1
PENN4	D	DQ465243.1
Euro D	D	NC_001445.1
El Amar	EA	AM157175.1
El Amar 1	EA	DQ431465.1
PS	M	AJ243957.1
SK 68	M	M92280.1
BOR-3	Rec	AY028309.2
AbTk	Rec	EU734794.1
Canadian	W	AY912055.1

Table 2: Table of runtimes of each step of the analysis. This does not include time to move data from one location to another.

EDNA	Time (HH:MM:SS)	"Traditional"	Time (HH:MM:SS)
Extract fastA	00:00:00	Extract fastQ	00:00:56
EDNA Pipeline	00:00:14	FastQC	00:01:07
		Filter & Trim reads	00:00:58
		BLASTn - GenBank nt	09:18:04
Total	00:00:14		09:21:05

Table 3: Sequence data set summary.

454 run Name	Host	Known Pathogen	Number of reads	Total Bp	Average read length	% Pathogen Reads
BGMV	Bean	BGMV	45,295	13,423,738	296	5.02%
PPV- EA	Prunus	PPV-EA	36,374	13,491,357	371	6.80%
PPV-M	Prunus	PPV-M	9,250	3,808,884	412	0.35%
PPV-MT0	Prunus	PPV-D	42,418	16,100,234	380	1.34%
PPV-MT4	Prunus	PPV-D	30,121	12,244,317	406	0.53%

Table 4: EDNA p-values and number of positive probes. Bolded sections indicate a positive diagnostic call, while non-bolded indicates a negative diagnostic call.

Probe Set	Top Hits	BGMV		PPV-MT0		PPV-MT4		PPV-EA		PPV-M	
		Pval	#Pos	Pval	#Pos	Pval	#Pos	Pval	#Pos	Pval	#Pos
BGMV	1	**0.018**	**21/21**	1	0/21	1	0/21	1	0/21	1	0/21
	5	**0.018**	**21/21**	1	0/21	1	0/21	1	0/21	1	0/21
	10	**0.017**	**21/21**	1	0/21	1	0/21	1	0/21	1	0/21
	50	**0.016**	**21/21**	1	0/21	1	0/21	1	0/21	1	0/21
BGYMV	1	**0.033**	**17/27**	1	0/27	1	0/27	1	0/27	1	0/27
	5	**0.032**	**17/27**	1	0/27	1	0/27	1	0/27	1	0/27
	10	**0.032**	**17/27**	1	0/27	1	0/27	1	0/27	1	0/27
	50	**0.024**	**17/27**	1	0/27	1	0/27	1	0/27	1	0/27
PPV	1	0.551	0/64	**0.000**	**62/64**	**0.001**	**39/64**	**0.003**	**28/64**	**0.000**	**63/64**
	5	0.331	0/64	**0.000**	**62/64**	**0.000**	**37/64**	**0.007**	**29/64**	**0.000**	**63/64**
	10	0.993	0/64	**0.000**	**62/64**	**0.000**	**36/64**	**0.020**	**26/64**	**0.000**	**63/64**
	50	0.107	0/64	**0.003**	**62/64**	**0.000**	**31/64**	0.122	**22/64**	**0.000**	**63/64**

Table 5: Strain-specific e-probe results. Bolded sections indicate a positive diagnostic call, italicized scores indicate a suspect diagnostic call, while the absence of these modifiers indicate a negative diagnostic call.

E-probe set	Top hits	PPV-MT0		PPV-MT4		PPV-EA		PPV-M	
		Pval	#Pos	Pval	#Pos	Pval	#Pos	Pval	#Pos
C set	10	0.468	*1/88*	0.41	*1/88*	0.158	0/88	0.985	0/88
D set	10	*0.057*	**6/10**	**0.002**	**8/10**	0.189	0/10	0.352	0/10
EA set	10	0.787	0/203	0.026	*2/203*	**0.000**	**142/203**	0.224	0/203
M set	10	0.798	0/96	**0.009**	*3/96*	0.971	*1/96*	**0.000**	**41/96**
W set	10	0.213	0/119	0.498	0/119	0.139	0/119	0.114	0/119

Figure captions:

Figure 1: MEGAN identification of reads for the BGMV (A), PPV-MT0 (B), PPV-MT4 (C), PPV-EA (D), and PPV-M (E). Some groups which represent less than 10% of the reads are unlabeled.

Figure 2: Positions of the e-probes on the PPV genome.

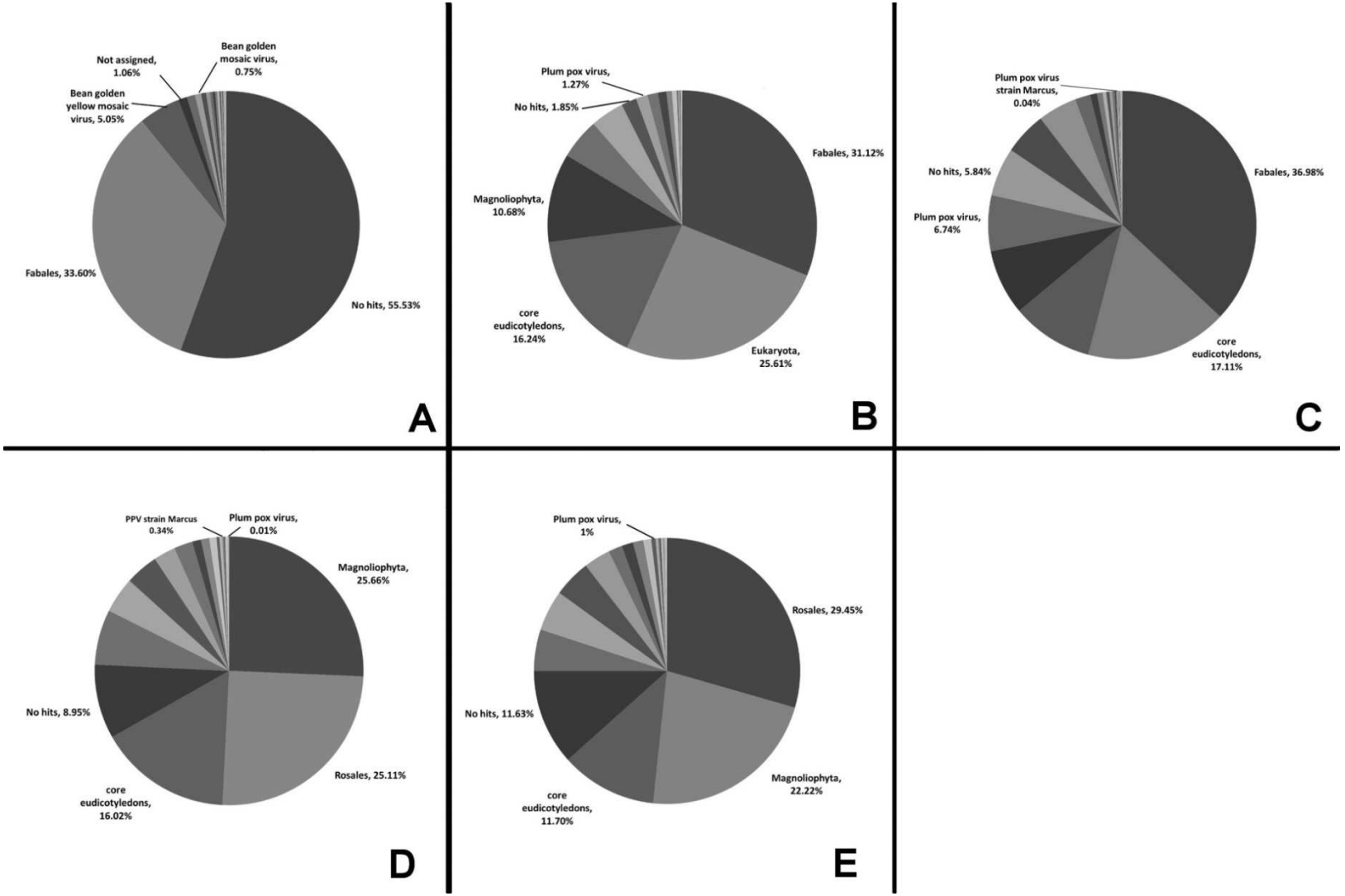
Not assigned, 1.06%
Bean golden mosaic virus, 0.75%
Bean golden yellow mosaic virus, 5.05%
Fabales, 33.60%
No hits, 55.53%
A
Plum pox virus, 1.27%
No hits, 1.85%
Fabales, 31.12%
Magnoliophyta, 10.68%
core eudicotyledons, 16.24%
Eukaryota, 25.61%
B
Plum pox virus strain Marcus, 0.04%
No hits, 5.84%
Fabales, 36.98%
Plum pox virus, 6.74%
core eudicotyledons, 17.11%
C
PPV strain Marcus 0.34%
Plum pox virus, 0.01%
Magnoliophyta, 25.66%
No hits, 8.95%
Rosales, 25.11%
core eudicotyledons, 16.02%
D
Plum pox virus, 1%
Rosales, 29.45%
No hits, 11.63%
Magnoliophyta, 22.22%
core eudicotyledons, 11.70%
E

Figure 4

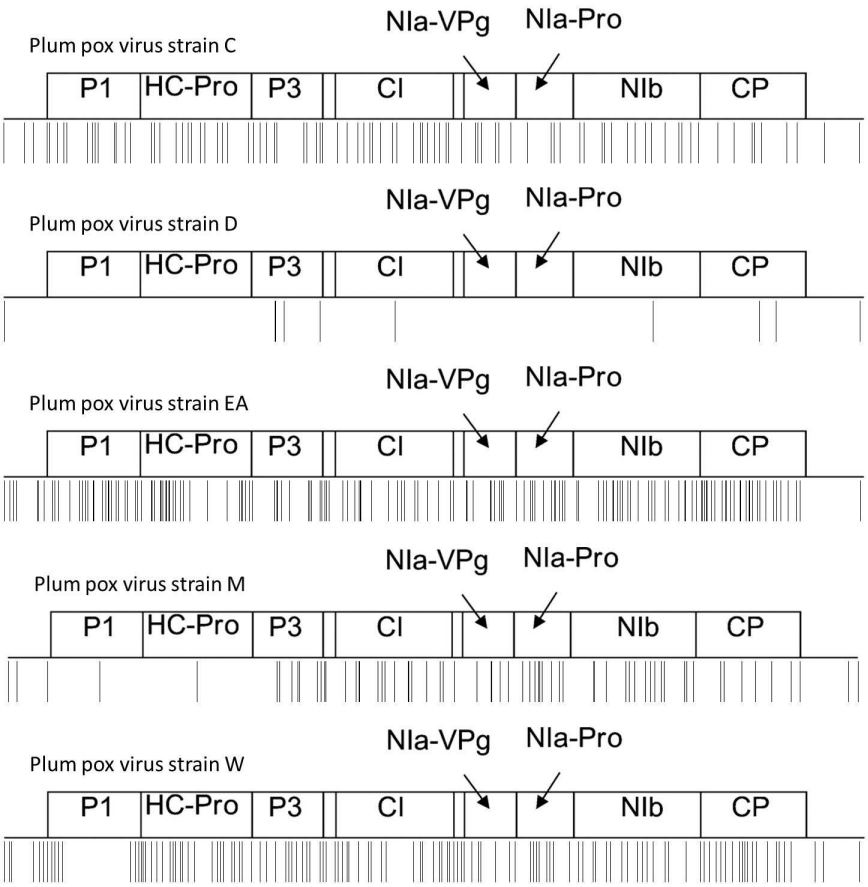
Plum pox virus strain C
NIa-VPg
NIa-Pro
P1
HC-Pro
P3
CI
NIb
CP
Plum pox virus strain D
NIa-VPg
NIa-Pro
P1
HC-Pro
P3
CI
NIb
CP
Plum pox virus strain EA
NIa-VPg
NIa-Pro
P1
HC-Pro
P3
CI
NIb
CP
Plum pox virus strain M
NIa-VPg
NIa-Pro
P1
HC-Pro
P3
CI
NIb
CP
Plum pox virus strain W
NIa-VPg
NIa-Pro
P1
HC-Pro
P3
CI
NIb
CP

Figure 5

CHAPTER VI

THE USE OF NON-SPECIFIC E-PROBES FOR GENERAL DETECTION OF VIRUSES.

Introduction

Previous chapters focused on the use of species- or strain-specific e-probes for detection and diagnosis, and the EDNA pipeline was introduced as a novel approach to detect and diagnose plant pathogens. There are two major limitations of the EDNA pipeline, however. The first is the e-probe design, for which there is a requirement for the sequences of both the pathogen genome and a phylogenetically near neighbor. The second is that the pipeline will detect only viruses for which e-probes have been designed and implemented. While there are many known plant viruses of economic importance, many may be missed if either the virus is unknown, or presence of the virus was not tested for.

The diagnostic community has developed many biological assays to test for a wide variety of viruses in both animal and plant systems (Chen *et al.*, 2011; Zhang *et al.*, 2010). To further explore the applications of Next Generation Sequencing (NGS) in a diagnostic setting, probes designed for use in a general virus microarray were converted to e-probes that can be used in the EDNA pipeline for detection on the family level, which is useful for detection of related

species of economically damaging viruses. This method would be extremely useful for government agencies that regulate and ensure the quality of plant material imported into the U.S., such as the USDA-Animal and Plant Health Inspection Service (APHIS).

Methods

Design of general e-probes

The sequences used as e-probes in this work originated from probes designed by Dr. Kael Fisher at the University of Utah School of Medicine, in collaboration with Drs. John Hammond of the USDA agriculture research service (ARS) and Claude Fauquet of the Danforth Center. The probes were to be used in a general virus detection microarray, designed to house over 9,300 unique probes, corresponding to over 1,300 virus and viroid species (Bagewadi *et al.*, 2010a; Bagewadi *et al.*, 2010b). The taxonomic information associated with these species are available in GenBank as of 2009. The microarray has been validated using *Wheat streak mosaic virus* (Brown, 2011). One aspect of this e-probe set which differs from previous ones, is the inclusion of a number of plant housekeeping genes to serve as positive controls. Since the of this probe set was originally designed for microarray, care was taken to ensure proper thermodynamics during hybridization, so each microarray probe (and thus each e-probe) is 60 bp long. No changes to the sequence of the probes were made in the conversion to e-probes.

Datasets

Three 454 sequence datasets were provided by USDA-APHIS in fastA file format, labeled RL-20, RL-24, and RL-26. Sample sequencing was performed using the Roche 454 Jr. platform in the USDA-ARS laboratory in Ft. Detrick, Maryland, and the only information available on the samples, their preparation or potential symptoms, was that each sample came from a different plant suspected of being infected with a virus. Each dataset was analyzed using the EDNA pipeline as described in chapters IV and V in this thesis with the general e-probe set

described above, with no change to the program. Once identification was made, the sequence of the type member of the virus family was used as a scaffold for mapping reads. Two methods were used to assemble the sequencing reads into genomes. In the first, the type member genome was used as a query in a BLAST search of the dataset, extracting the reads that resulted in a hit with an e-value of 10^{-3}, and assembling the reads into contigs using the CAP3 program. The second method was to use the Roche GS Reference Mapper software to map reads onto the type member genomes.

Results

EDNA detection

The EDNA pipeline revealed 10 and 12 positive plant control e-probes for RL-20 and RL-24, respectively, but no positive virus e-probes (Table 1). Eighteen positive plant control e-probes were found positive for RL-26, with 5 separate viral e-probes (3 "Tymovirales", and 2 "Alphaflexiviridae, Potexvirus"). The fact that the alphaflexiviridae is a family within the order Tymovirales, suggests that there is a single virus within the sample dataset. Further analysis was performed to obtain the full genome of virus(es) recognized by the e-probes. Type members for each e-probe classification were chosen: *Turnip yellow mosaic virus* (TYMV; NC_004063.1) for the "Tymovirales" probes, and *Potato virus X* (PVX; NC_011620.1) for the "Alphaflexiviridae, Potexvirus" probes, to be used as scaffolds in the mapping assembly of the genomes.

Mapping results

When queried with the TYMV genome, 11 sequence reads were recovered. These reads were assembled into a single contig of 582 bp. When this contig was used as a query against the nt Genbank database, several of the top hits were to PVX, with 95% identity. Querying with the PVX genome recovered 359 reads, assembled into 15 contigs, each matching PVX with over 90% identity. Mapping reads from the dataset to the type member genomes resulted in zero reads being

mapped to TYMV. Eight contigs consisting of 317 reads were mapped to PVX, with 92.3% of the PVX genome covered. This evidence strongly suggests that the EDNA detection of the detection of PVX is true, an interpretation corroborated by a commercially available potex virus antibody detection assay evidence of flexuous rods in the sample (Jorge Abad; personal correspondence).

Discussion

The EDNA pipeline was able to detect the presence of a member of the family *Alphaflexiviridae*, most likely a potexvirus. With further downstream analysis, 92.3% of the viral genome was recovered and assembled. This evidence, together with the biological evidence, leads to a definitive conclusion that the plant sample is infected with a potexvirus, most likely PVX. Almost the full genome was recovered and can be used in the design of additional e-probes to be tested with future samples. In addition, if the sample was suspected of being infected with a virus intentionally, SNP typing is available to assist in attribution.

As with many forms of detection, further validation of EDNA is needed, either by the recovery of addition reads to assemble a genome, or biological evidence such as EM images or PCR assays. PCR assays would especially helpful to ensure that the assembly of the virus genome is correct. Future work is needed, including optimizing the general e-probes, determining a limit of detection, and adapting the test for viroid genomes.

REFERENCES

Bagewadi, B., Fischer, K., Henderson, D., Jordan, R., Wang, D., Perry, K., Melcher, U., Hammond, J. & Fauquet, C. (2010a). Universal plant virus microarray development and validation. In *Phytopathology*, pp. S154-S154: AMER PHYTOPATHOLOGICAL SOC 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA.

Bagewadi, B., Henderson, D., Jordan, R., Perry, K., Melcher, U., Wang, D., Fischer, K., Hammond, J. & Fauquet, C. (2010b). DNA microarray based universal plant virus detection and identification. In *Phytopathology*, pp. S10-S10: AMER PHYTOPATHOLOGICAL SOC 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA.

Brown, T. P. (2011). The Development and Characterization of Molecular Tools for Microbial Forensics In *Biochemistry and Molecular Biology*, p. 199: Oklahoma State University.

Chen, E. C., Miller, S. A., DeRisi, J. L. & Chiu, C. Y. (2011). Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens. *Journal of visualized experiments: JoVE*.

Zhang, Y., Yin, J., Li, G., Li, M., Huang, X., Chen, H., Zhao, W. & Zhu, S. (2010). Oligonucleotide microarray with a minimal number of probes for the detection and identification of thirteen genera of plant viruses. *J Virol Methods* **167**, 53-60.

Table 1: List of positive probes for each sequencing data set.

RL-20		RL-24		RL-26	
Probe name	Description	Probe name	Description	Probe name	Description
Zm_18S_rRNA-2	Plant Control	Zm_18S_rRNA-2	Plant Control	9629255_nt4776.60	Tymovirales
Zm_18S_rRNA-2.11	Plant Control	Zm_18S_rRNA-2.11	Plant Control	169219512_nt4170.60	Tymovirales
Zm_18S_rRNA-4	Plant Control	Zm_18S_rRNA-3	Plant Control	189458579_nt4414.60	Tymovirales
Zm_18S_rRNA-4.11	Plant Control	Zm_18S_rRNA-3.11	Plant Control	37905677_nt4201.60	Alphaflexiviridae, Potexvirus
Zm_25S_rRNA-5	Plant Control	Zm_18S_rRNA-4	Plant Control	37905677_nt4226.60	Alphaflexiviridae, Potexvirus
Zm_25S_rRNA-5.11	Plant Control	Zm_18S_rRNA-4.11	Plant Control	Arabidopsis_rbcL	Plant Control
Zm_25S_rRNA-6	Plant Control	Zm_25S_rRNA-5	Plant Control	Arabidopsis_rbcL.11	Plant Control
Zm_25S_rRNA-6.11	Plant Control	Zm_25S_rRNA-5.11	Plant Control	Potato_eF1a	Plant Control
Zm_5.8S_rRNA-1	Plant Control	Zm_25S_rRNA-6	Plant Control	Potato_eF1a.11	Plant Control
Zm_5.8S_rRNA-1.11	Plant Control	Zm_25S_rRNA-6.11	Plant Control	Zm_18S_rRNA-2	Plant Control
		Zm_25S_rRNA-7	Plant Control	Zm_18S_rRNA-2.11	Plant Control
		Zm_25S_rRNA-7.11	Plant Control	Zm_18S_rRNA-3	Plant Control
				Zm_18S_rRNA-3.11	Plant Control
				Zm_18S_rRNA-4	Plant Control
				Zm_18S_rRNA-4.11	Plant Control
				Zm_25S_rRNA-5	Plant Control
				Zm_25S_rRNA-5.11	Plant Control
				Zm_25S_rRNA-6	Plant Control
				Zm_25S_rRNA-6.11	Plant Control
				Zm_25S_rRNA-7	Plant Control
				Zm_25S_rRNA-7.11	Plant Control
				Zm_5.8S_rRNA-1	Plant Control
				Zm_5.8S_rRNA-1.11	Plant Control

APPENDICES

The following supplemental material from Chapters III, IV, and V have been made available on the ProQuest/UMI Dissertation Publishing website.

CHAPTER III

Supplementary Figure S1. The Brassicales-associated clade of the Bayesian-likelihood tree of theRdRp ORF of tobamoviruses.

Supplementary Table S1. Positions of polymorphic residues in PafMV-TGP RNA

Supplementary Table S2. Virus names, abbreviations and GenBank accession numbers

CHAPTER IV

Supplementary Table S1. Comparison of EDNA results with e-probes before, and after the BLAST check

CHAPTER V

ChapterV_EDNAPipeline.rar The EDNA pipeline and probe design scripts

VITA

Anthony Harold Stobbe

Candidate for the Degree of

Doctor of Philosophy

Thesis: VIRUS DETECTION IN A METAGENOMIC SEQUENCE DATASET, METHODS AND APPLICATIONS

Major Field: Biochemistry and Molecular Biology

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Biochemistry and Molecular Biology at Oklahoma State University, Stillwater, Oklahoma in July, 2013.

Completed the requirements for the Master of Science in Biochemistry and Molecular Biology at Oklahoma State University, Stillwater, Oklahoma in 2012.

Completed the requirements for the Bachelor of Science in Health Sciences at Southwestern Oklahoma State University, Weatherford, Oklahoma in 2006.

Experience:

Graduate Research Assistant in Virology Laboratory of Dr. Ulrich Melcher, Biochemisty and Molecular Biology Department, Oklahoma State University 2009-2013
Visiting Scientist in Virology Laboratory of Dr. William Schneider, USDA-ARS, Ft. Detrick Maryland 2011-2012

Professional Memberships:

American Phytopathology Society (2010-2013)
American Society for Virology (2013)