Screening Metagenomic Data for Viruses Using the E-Probe Diagnostic Nucleic Acid Assay

A. H. Stobbe, W. L. Schneider, P. R. Hoyt, and U. Melcher

First, third, and fourth authors: Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater 74078; and second author: USDA-ARS, Foreign Disease-Weed Science Research Unit, Fort Detrick, MD 21702.
Current address of A. H. Stobbe: Department of Plant Pathology, Pennsylvania State University, State College 16802.
Accepted for publication 24 March 2014.

ABSTRACT

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

Next generation sequencing (NGS) is not used commonly in diagnostics, in part due to the large amount of time and computational power needed to identify the taxonomic origin of each sequence in a NGS data set. By using the unassembled NGS data sets as the target for searches, pathogen-specific sequences, termed e-probes, could be used as queries to enable detection of specific viruses or organisms in plant sample metagenomes. This method, designated e-probe diagnostic nucleic acid assay, first tested with mock sequence databases, was tested with NGS data sets generated from plants infected with a DNA (*Bean golden yellow mosaic virus,* BGYMV) or an RNA (*Plum pox virus,* PPV) virus. In addition, the ability to detect and differentiate among strains of a single virus species, PPV, was examined by using probe sets that were specific to strains. The use of probe sets for multiple viruses determined that one sample was dually infected with BGYMV and *Bean golden mosaic virus*.

The global trading of plant material has increased the introduction of foreign plant diseases in the last few decades (24), leading to a need for enhanced surveillance and detection of pathogens in imported plants. Currently pathogens are detected within imported plant material by visual, nucleic acid-based, and protein-based methods. Protein-based assays such as enzyme-linked immunosorbent assays, western blots, and immuno-strip tests are not multiplexed easily to test for several pathogens simultaneously, but proteomic methods such as "mud-pit" (19) are used to test for multiple pathogens. Of the two major nucleic acid-based assays, polymerase chain reaction (PCR) and microarrays, only microarrays are readily multiplexed (3,11,14,29).

Metagenomics can provide sequence information on the entire organismal and viral makeup of a sample. Metagenomics has led to the identification of previously unknown species of microorganisms, as well as offered insights into their ecological distribution. Metagenomics is facilitated by next generation sequencing (NGS), a powerful technology that allows the acquisition of hundreds of thousands of short sequence reads from the large number of organisms within a given sample (8,27). This immense sequencing capability can be a boon to diagnosticians who are interested in the detection and identification of specific pathogens. NGS has been used to identify pathogens in various systems (1,13,20) and has the advantage of being able to detect and identify many different pathogens within a sample. Two significant drawbacks have kept this technology from being used to compare short sequence reads to known sequences for diagnostic purposes: the length of computational analysis time, and the amount of computational power needed.

Corresponding author: U. Melcher;
E-mail address: u-melcher-4@alumni.uchicago.edu

http://dx.doi.org/10.1094/PHYTO-11-13-0310-R

The speed at which sequence data is generated and placed in curated sequence databanks has been increasing, and will continue to do so (6,12). In response, steps have been taken to increase the efficiency of search algorithms that use sequence reads as queries of the ever-enlarging databanks (15). The e-probe diagnostic nucleic acid assay (EDNA) pipeline (23) overcomes the sequence databank size problem by switching roles of databank and sequence reads. It uses short pathogen-specific sequences as queries against a databank of raw sequence data from the sample. These short sequence queries, termed e-probes, allow the user to choose only the pathogens of interest and thus reduce the computational time needed to detect and identify a pathogen. Previously, the EDNA pipeline concept was tested and validated with simulated data sets generated from plant genomes and the genomes of viruses, bacteria, fungi, and oomycetes.

In the study described herein, the EDNA pipeline was applied to metagenomic data sets obtained from virus-infected plant material. The ability to differentiate among closely related strains of viral pathogens was also tested using *Plum pox virus* (PPV) strains 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) (21). 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 with near neighbor sequences using the Nucmer script of the Mummer software package (5). Sequences having similarity to the near neighbors were removed, leaving only unique target sequences, which are used as queries against the NCBI nonredundant nucleotide database to ensure specificity to the target organism. Any candidate probe which received a hit with an e-value of 1×10^{-9} or lower but was not

identifiably from the target organism was removed. The same modified pipeline was used in the initial testing of EDNA (23), except that e-probe lengths were not limited to a specific size, and were instead allowed to vary between 30 nt and infinity. For controls a decoy set of e-probes were generated 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 with the near neighbor *Abutilon mosaic virus* (NC_001928.2, NC_001929.2). PPV (NC_001445.1) was compared to its near neighbor *Pepper mottle virus* (NC_001517.1). Five PPV strains, C (including the newly identified CR strain [7]), D, EA, M, and W, were used in the design of the e-probe sets (16–18). Each strain was considered as a target pathogen, with all other strains considered as near neighbors, for a total of five e-probe sets (Table 1). The strain specific e-probes were designed by using the methods described above.

In silico testing of the specificity of the PPV strain e-probes was carried out as previously described using mock sample databases (23).

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 from BGYMV-infected bean were generously provided by Judith Brown. Total nucleic acids of leaf discs of *Prunus persica* infected with PPV were obtained as described (26). Four samples of PPV-infected tissue were used, two of which were infected with different passages of the Penn-3 isolate of the PPV D strain, one with the an isolate of the M strain and another with the El Amar isolate of the EA strain. The presence of the viruses was confirmed with quantitative reverse transcription-PCR as described (22). Each total nucleic acid sample was amplified using a Whole Transcriptome Amplification Kit (Sigma-Aldrich, St. Louis, MO), as per the manufacturer's instructions, followed by size-selection using AMPure Beads (New England BioLabs Inc., Ipswich, MA) to recover DNA greater than 100 bp. The resulting cDNA library was sequenced using the Roche 454 Jr. platform, excluding nebulization.

To determine the percentage of reads originating from the pathogen, the raw sequencing results were queried with the pathogen's genome and enumerated. 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 nonredundant database, and parsing of the query results. For the "traditional" approach, the trimming and filtering were performed with the iPlant discovery environment (9) using the FASTX Trimmer and FASTX Quality Filter. The assembly of the contigs was performed using the Roche de novo Assembler. Querying the NCBI nonredundant database was performed with the mpiBLAST+ software (4) on a high performance computing cluster at Oklahoma State University. The MEGAN software package was used in the identification of organisms that contributed to the metagenome (10). Additional analysis was performed by mapping sequencing reads to the reference genomes of pathogens present by using Reference Mapper (Roche, Basel, Switzerland) according to the manufacturer's instructions.

TABLE 1. *Plum pox virus* strain isolates used

Isolate name	Strain	Accession
SwC	C	Y09851.2
RU-17sc	CR	KC020124
RU-18sc	CR	KC020125
RU-30sc	CR	KC020126
PENN-1	D	AF401295.1
PENN-2	D	AF401296.1
Cdn 4	D	AY953263.1
PENN-4	D	DQ465243.1
NAT	D	NC_001445.1
El Amar	EA	AM157175.1
El Amar	EA	DQ431465.1
PS	M	AJ243957.1
SK 68	M	M92280.1
BOR-3	Rec	AY028309.2
AbTk	T	EU734794.1
W3174	W	AY912055.1

The second analysis method used the EDNA pipeline. The FASTA file was extracted from the .SFF output file of the Roche 454 Jr. sequencing, 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 h represents a hit, n is the total number of top hits used, $Eval$ is the e-value of the hit, and the $\%cov.$ is the percent coverage of the e-probe used in the hit.

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

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 its target score was more than 10 standard deviations above the decoy 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 one compute node (12 cores) of the OSU "Cowboy" 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 (9).

RESULTS

E-probe sets. PPV e-probe sets were used to search mock sample databases of each of the five strains for which they were designed and for a mixture of Rec and T strains (recombinants of strains D and M). Each set was able to correctly identify the strain for which it was designed. Surprisingly, the searches of the Rec-T combined mock sample database with each strain-specific e-probe set gave a positive diagnostic call only with the M e-probe set. For the BGMV and BGYMV probe sets, several probes gave false positives in the in silico testing. Consequently, they were removed from the sets.

Sequences. The sequencing files are summarized in Table 2. Samples from plants infected with PPV strains were barcoded and sequenced on two 454 Jr. plates (PPV-MT0, PPV-M paired on one plate, PPV-EA, PPV-MT4 paired on the other), while the BGYMV sample was sequenced on a single plate. These sequence data sets consist of between 9,250 and 45,295 reads, with a range of average read lengths (296 to 412 nt). The percentage of the reads that matched to known pathogens in a BLAST search ranged from 0.35 to 6.80%. The average percentage of pathogen reads was much lower in the PPV samples than in the BGYMV sample, with the exception of PPV-EA (Table 2).

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 percentage of the metagenome to which the pathogen contributed (Fig. 1). The analysis of the BGYMV 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 levels from 0.35 to 6.80%. A strain identification was made 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. With the EDNA pipeline, each of the probe sets successfully detected the presence of virus pathogens in each processed raw sequence data set, based on results of the two tests mentioned above (Table 3). Examination of confidence levels for the *t* test revealed sensitivity to the number of top hits considered with uncertainty decreasing as the number of hits considered increased for BGMV and BYGMV probes, while uncertainty increased when the PPV-EA data set was probed with PPV probes. On using the top 50 hits, the identification could not be made with confidence. In addition to the EDNA approach, the BGMV and BGYMV genomes were used as 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, while the complete sequences of BGYMV DNA segment A and both segments of BGMV were mapped, only 86.1% of the BGYMV DNA B segment was mapped. A larger number of high quality variants were found for the BGYMV reference (81 variants for DNA A, 178 for DNA B) when compared with BGMV (3 and 0, respectively).

Strain-specific e-probes. Strain-specific e-probes were used as queries of the sample data sets shown in Table 4, using the same methods described above. The strain-specific e-probes were more cross-reactive to other strains than the genus level e-probes were to other genera, but still are able to differentiate between the strains. The samples from D, EA, and M strain infected plants tested positive with probes for these strains in both statistical tests. The W probe sets did not recognize sequence data sets created from D-, EA-, or M-infected plants. The C probe set contained a sufficient number of probes with enough similarity to EA and M sequences to allow $P < 0.05$ discrimination of the e-probe set from the decoy set. The same was true of the M e-probe set

TABLE 2. Summary of sequence data generated

454 run name	Host	Known pathogen	Number of reads	Total bp	Average read length	Pathogen reads
BGYMV	Bean	BGYMV	45,295	13,423,738	296	5.05%
PPV-EA[a]	*Prunus*	PPV-EA	36,374	13,491,357	371	6.80%
PPV-M[b]	*Prunus*	PPV-M	9,250	3,808,884	412	0.35%
PPV-P30[c]	*Prunus*	PPV-D	42,418	16,100,234	380	1.34%
PPV-P34[c]	*Prunus*	PPV-D	30,121	12,244,317	406	0.53%

[a] GenBank DQ431465.
[b] Greek isolate described in Varveri et al. (25).
[c] Penn-3 isolate (D strain, GenBank DQ465242). "P30" and "P34" are passage number designations.

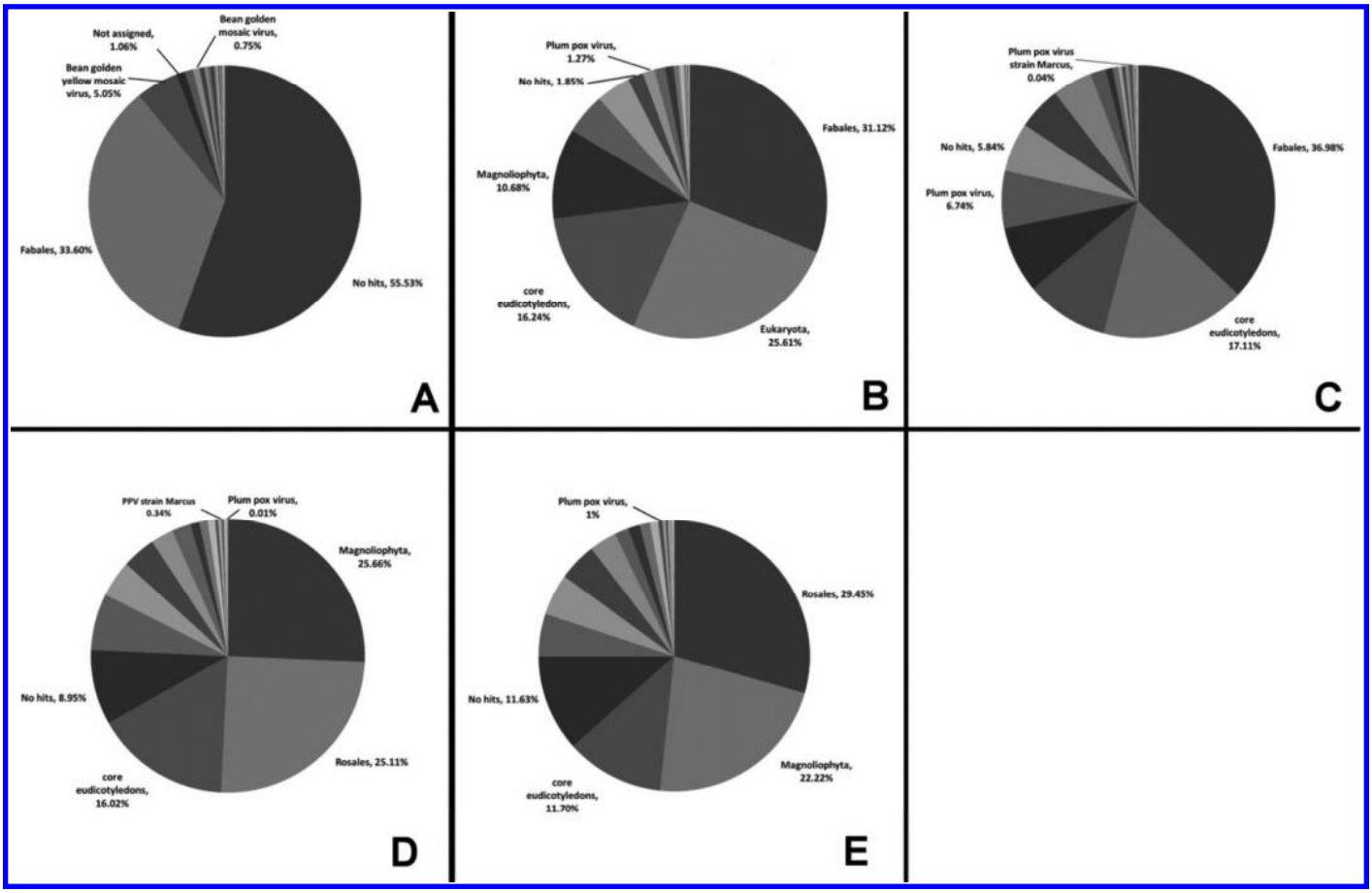

Fig. 1. MEGAN identification of reads for the **A,** *Bean golden yellow mosaic virus*, **B,** *Plum pox virus* (PPV)-MT0, **C,** PPV-MT4, **D,** PPV-EA, and **E,** PPV-M samples. Groups which represent less than 10% of the reads are unlabeled.

with one of the D strain infected plant samples. However, in both misidentification cases, no individual probes were positive. Due to the prevalence of recombinant strains, the genomic locations of the strain-specific e-probes is important to consider. Mapping the positions of the e-probes onto the PPV genome (Fig. 2) shows that each set of the strain-specific e-probes spans across the entirety of its genome.

DISCUSSION

Previous validation of the EDNA approach to testing metagenomic nucleic acid samples for the presence of viral sequence used, as BLAST search targets, simulated data sets assembled with varying levels of viral sequences relative to plant sequences (23). In the present study, validation was extended to samples from virus-infected plants. When the metagenomic data sets from these plants were analyzed by BLASTn search of available sequences followed by MEGAN analysis of the taxonomic distribution of reads, it was determined that the data sets contained from 0.35 to 6.8% virus-derived reads (Table 2). These values are within the range of values (less than 0.5 to 25%) used in constructing the simulated databases previously used for validation.

The MEGAN analysis also suggested that one of the samples originated from a plant infected with two viruses, BGMV and BGYMV. Other virus-infected plant samples analyzed by metagenomic methods also provide evidence of multiple infection. For example, the metagenome of one grapevine sample yielded evidence of the presence of seven distinct RNA genomes (virus and viroid) (2). Further, in a sequence-based survey of 1,305 noncultivated plants of the Tallgrass Prairie Preserve (28), 37% of plants testing positive for viral sequences were deduced to harbor more than one virus (U. Melcher, *unpublished data*). Increasingly, multiple infection of plants by viruses needs to be considered in disease diagnosis.

The previously constructed PPV e-probe set (23) was constructed by using the sequence of the D strain isolate NAT (Table 1) and, as near neighbor, *Pepino mosaic virus*. The finding that

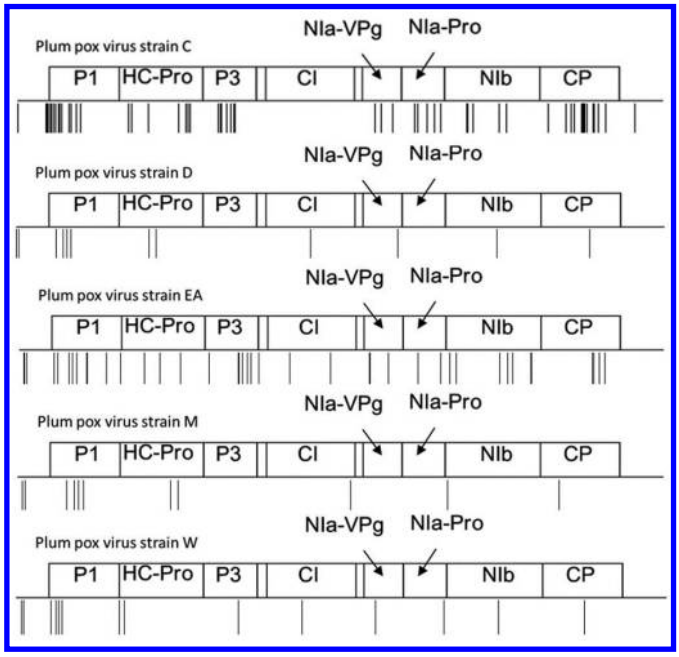

Fig. 2. Positions of the e-probes on the *Plum pox virus* genome.

TABLE 3. Probing of sequence data sets from *Plum pox virus* (PPV)- and *Bean golden yellow mosaic virus* (BGYMV)-infected plants with species-specific probes by E-probe detection nucleic acid assay

		BGYMV[b]		PPV-P30		PPV-P34		PPV-EA		PPV-M	
Probe set	Top hits[a]	*P* value	Positive probes	*P* value	Positive probes	*P* value	Positive probes	*P* value	Positive probes	*P* value	Positive probes
BGMV[c]	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**

[a] Number of top hits used in calculation of scores using the *t* test method and in calculating the mean and standard deviation of decoy probe scores for the standard deviation test for positive probes.
[b] Bold entries indicate a positive diagnostic call, while nonbold entries indicate a negative diagnostic call.
[c] *Bean golden mosaic virus.*

TABLE 4. Probing of sequence data sets from *Plum pox virus* (PPV)-infected plants with strain-specific probes by E-probe detection nucleic acid assay

		PPV-P30[b]		PPV-P34		PPV-EA		PPV-M	
E-probe set	Top hits[a]	*P* value	Positive probes	*P* value	Positive probes	*P* value	Positive probes	*P* value	Positive probes
C set	10	0.31	0/93	0.86	0/93	**<0.05**	0/93	**<0.05**	0/93
D set	10	**<0.05**	**22/22**	**<0.05**	**19/22**	0.11	0/22	0.18	0/22
EA set	10	0.44	0/49	0.44	0/49	**<0.05**	**31/49**	0.39	0/49
M set	10	**<0.05**	0/11	0.45	0/11	0.05	0/11	**<0.05**	**5/11**
W set	10	0.36	0/14	0.15	0/14	0.88	0/14	0.18	0/14

[a] Number of top hits used in calculation of scores using the *t* test method and in calculating the mean and standard deviation of decoy probe scores for the standard deviation test for positive probes.
[b] Bold values indicate a positive diagnostic call, while plain values indicates a negative diagnostic call.

this e-probe set recognized reads from read data sets of samples infected with isolates of a selection of PPV strains (D, EA, and M; Table 3) demonstrates that the design strategy yielded an e-probe set that recognized PPV infection regardless of isolate or strain. The strategy to develop strain-specific probe sets, by using the other PPV strain sequences as near neighbors was effective, generating a significant number of probes (11 to 93). The use of these e-probe sets in BLAST searches was also effective, clearly recognizing samples infected with the strain to which they were designed and only rarely giving a false positive result using the *t* test method, while finding no individual e-probe matches with material infected with other strains.

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. By reversing the roles of known sequences and metagenomic data sets relative to traditional metagenomic approaches, EDNA has great potential diagnostic use. Reducing the size of the known sequence data set (selected e-probes versus the entire GenBank contents) makes scanning plant-derived sequence data sets for selected known pathogens a rapid and effective process for diagnostic and screening use.

ACKNOWLEDGMENTS

This study was funded by the USDA-CSREES Plant Biosecurity Program (grant number 2010-85605-20542) and additionally supported through instrumentation funded by the National Science Foundation (grant number OCI–1126330) and the Oklahoma Agricultural Experiment Station. H. Hwang and the Bioinformatics Core Facility are acknowledged for providing exceptional consistency and skill while preparing libraries and performing pyrosequencing.

LITERATURE CITED

1. Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M., and Boonham, N. 2009. Next-generation sequencing and metagenomic analysis: A universal diagnostic tool in plant virology. Mol. Plant Pathol. 10:537-545.
2. Al Rwahnih, M., Daubert, S., Golino, D., and Rowhani, A. 2009. Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology 387:395-401.
3. Call, D. R., Borucki, M. K., and Loge, F. J. 2003. Detection of bacterial pathogens in environmental samples using DNA microarrays. J. Microbiol. Methods 53:235-243.
4. Darling, A., Carey, L., and Feng, W.-C. 2003. The design, implementation, and evaluation of mpiBLAST. Proc. ClusterWorld 2003.
5. Delcher, A. L., Salzberg, S. L., and Phillippy, A. M. 2003. Using MUMmer to identify similar regions in large sequence sets. Curr. Prot. Bioinform. 00:10.3:10.3.1-10.3.18.
6. Fritz, M. H.-Y., Leinonen, R., Cochrane, G., and Birney, E. 2011. Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res. 21:734-740.
7. García, J. A., Glasa, M., Cambra, M., and Candresse, T. 2014. *Plum pox virus* and sharka: A model potyvirus and a major disease. Mol. Plant Pathol. 15:226-241.
8. Gilbert, J. A., and Dupont, C. L. 2011. Microbial metagenomics: Beyond the genome. Annu. Rev. Marine Sci. 3:347-371.
9. Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., and Lenards, A. 2011. The iPlant collaborative: Cyberinfrastructure for plant biology. Frontiers in Plant Science 2. Frontiers Editorial Office, Lausanne, Switzerland.
10. Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. 2007. MEGAN analysis of metagenomic data. Genome Res. 17:377-386.
11. Iqbal, S. S., Mayo, M. W., Bruno, J. G., Bronk, B. V., Batt, C. A., and Chambers, J. P. 2000. A review of molecular recognition technologies for detection of biological threat agents. Biosens. Bioelectron. 15:549-578.
12. Kodama, Y., Shumway, M., and Leinonen, R. 2012. The sequence read archive: Explosive growth of sequencing data. Nucleic Acids Res. 40:D54-D56.
13. Koonin, E. V., and Dolja, V. V. 2012. Expanding networks of RNA virus evolution. BMC Biol. 10:54.
14. Lazcka, O., Campo, F., and Munoz, F. X. 2007. Pathogen detection: A perspective of traditional methods and biosensors. Biosens. Bioelectron. 22:1205-1217.
15. Li, H., and Homer, N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 11:473-483.
16. Maiss, E., Timpe, U., Brisske, A., Jelkmann, W., Casper, R., Himmler, G., Mattanovich, D., and Katinger, H. 1989. The complete nucleotide sequence of plum pox virus RNA. J. Gen. Virol. 70:513-524.
17. Matic, S., Elmaghraby, I., Law, V., Varga, A., Reed, C., Myrta, A., and 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.
18. Nemchinov, L., and Hadidi, A. 1996. Characterization of the sour cherry strain of plum pox virus. Phytopathology 86:575-580.
19. Padliya, N. D., and Cooper, B. 2006. Mass spectrometry-based proteomics for the detection of plant pathogens. Proteomics 6:4069-4075.
20. Roossinck, M. J. 2012. Plant virus metagenomics: Biodiversity and ecology. Annu. Rev. Genet. 46:359-369.
21. Satya, R. V., Zavaljevski, N., Kumar, K., and Reifman, J. 2008. A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. BMC Bioinform. 9:185.
22. Schneider, W. L., Sherman, D. J., Stone, A. L., Damsteegt, V. D., and Frederick, R. D. 2004. Specific detection and quantification of *Plum pox virus* by real-time fluorescent reverse transcription-PCR. J. Virol. Methods 120:97-105.
23. Stobbe, A. H., Daniels, J., Espindola, A. S., Verma, R., Melcher, U., Ochoa-Corona, F., Garzon, C., Fletcher, J., and 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 94:356-366.
24. Tatem, A. J., Hay, S. I., and Rogers, D. J. 2006. Global traffic and disease vector dispersal. Proc. Natl. Acad. Sci. USA 103:6242-6247.
25. Varveri, C., Zintzaras, E., Dimou, D., Papapanagiotou, A., and Di Terlizzi, B. 2004. Monitoring and spatiotemporal analysis of PPV-M spread in two apricot orchards in Southern Greece. Ann. Benaki Phytopathol. Inst. 20:1-9.
26. Wallis, C. M., Stone, A. L., Sherman, D. J., Damsteegt, V. D., Gildow, F. E., and Schneider, W. L. 2007. Adaptation of plum pox virus to a herbaceous host (*Pisum sativum*) following serial passages. J. Gen. Virol. 88:2839-2845.
27. Willner, D., and Hugenholtz, P. 2013. From deep sequencing to viral tagging: Recent advances in viral metagenomics. BioEssays 35:436-442.
28. Wren, J. D., Roossinck, M. J., Nelson, R. S., Scheets, K., Palmer, M. W., and Melcher, U. 2006. Plant virus biodiversity and ecology. PLoS Biol. 4:e80.
29. Ye, Y., Mar, E.-C., Tong, S., Sammons, S., Fang, S., Anderson, L. J., and Wang, D. 2010. Application of proteomics methods for pathogen discovery. J. Virol. Methods 163:87-95.