Journal of Microbiological Methods 94 (2013) 356–366

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

E-probe Diagnostic Nucleic acid Analysis (EDNA): A theoretical approach for handling of next generation sequencing data for diagnostics

Anthony H. Stobbe [a], Jon Daniels [b], Andres S. Espindola [b], Ruchi Verma [a], Ulrich Melcher [a], Francisco Ochoa-Corona [b], Carla Garzon [b], Jacqueline Fletcher [b], William Schneider [c,*]

[a] Oklahoma State University, Department of Biochemistry and Molecular Biology, United States
[b] Oklahoma State University, Department of Entomology and Plant Pathology, United States
[c] USDA ARS, Foreign Disease-Weed Science Research Unit, United States

ARTICLE INFO

Article history:
Received 26 March 2013
Received in revised form 2 July 2013
Accepted 2 July 2013
Available online 16 July 2013

Keywords:
Bioinformatics
Next-generation sequencing
Pathogen detection

ABSTRACT

Plant biosecurity requires rapid identification of pathogenic organisms. While there are many pathogen-specific diagnostic assays, the ability to test for large numbers of pathogens simultaneously is lacking. Next generation sequencing (NGS) allows one to detect all organisms within a given sample, but has computational limitations during assembly and similarity searching of sequence data which extend the time needed to make a diagnostic decision. To minimize the amount of bioinformatic processing time needed, unique pathogen-specific sequences (termed e-probes) were designed to be used in searches of unassembled, non-quality checked, sequence data. E-probes have been designed and tested for several selected 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 Analysis (EDNA) process provides the framework for a new sequence-based detection system that eliminates the need for assembly of NGS data.

Published by Elsevier B.V.

1. Introduction

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

Next generation sequencing (NGS) is a relatively recent technology that allows for the generation of very large amounts of sequence data from a given sample (Ronaghi, 2001). Because various NGS platform technologies differ in read length (20 bp to approximately 1000 bp) and in the total number of reads (100,000 to 1 million), the amount of

* Corresponding author at: 1301 Ditto Avenue, Bldg 1301, Fort Detrick, MD 21702-5023, United States. Tel.: +1 301 619 7312.
E-mail address: william.schneider@ars.usda.gov (W. Schneider).

0167-7012/$ – see front matter. Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.mimet.2013.07.002

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

The advantage of NGS over other sequencing technologies is the volume (400 MB–28 GB) of data generated (Metzker, 2010; Reis-Filho, 2009). From a different perspective, the volumes of data generated by NGS could be a detriment to a diagnostician, as bioinformatic processing becomes a limiting factor in high throughput applications (Pop and Salzberg, 2008; Magi et al., 2010). For example, consider 200 l of seawater containing over 5000 different viruses (Breitbart et al., 2002). If a metagenomics approach is used for plant pathogen detection within this sample, plant pathogen-specific sequences will likely make up only a small percentage of the total reads (Adams et al., 2009; Roossinck et al., 2010). In contrast, plants infected with viruses may have a much higher percentage of the total nucleic acid composed 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 (Fig. 1), and reported herein, termed E-probe Detection of Nucleic acid Analysis (EDNA), is a bioinformatic pipeline that minimizes and ignores irrelevant sequence data thereby focusing on specific pathogen-associated sequences. Mock sample databases (MSDs), simulating 454-pyrosequencing runs from plant pathogen infected plants, were generated. Rather than assessing the presence or absence of pathogens by BLAST of all sequences against a curated database, such as the nucleotide sequence databases of GenBank, the NGS metagenomic data was assessed using pathogen unique sequences termed target e-probes, incorporating local BLAST searches of designed e-probes against databases of raw sequence reads on local computer systems. This modified bioinformatic approach

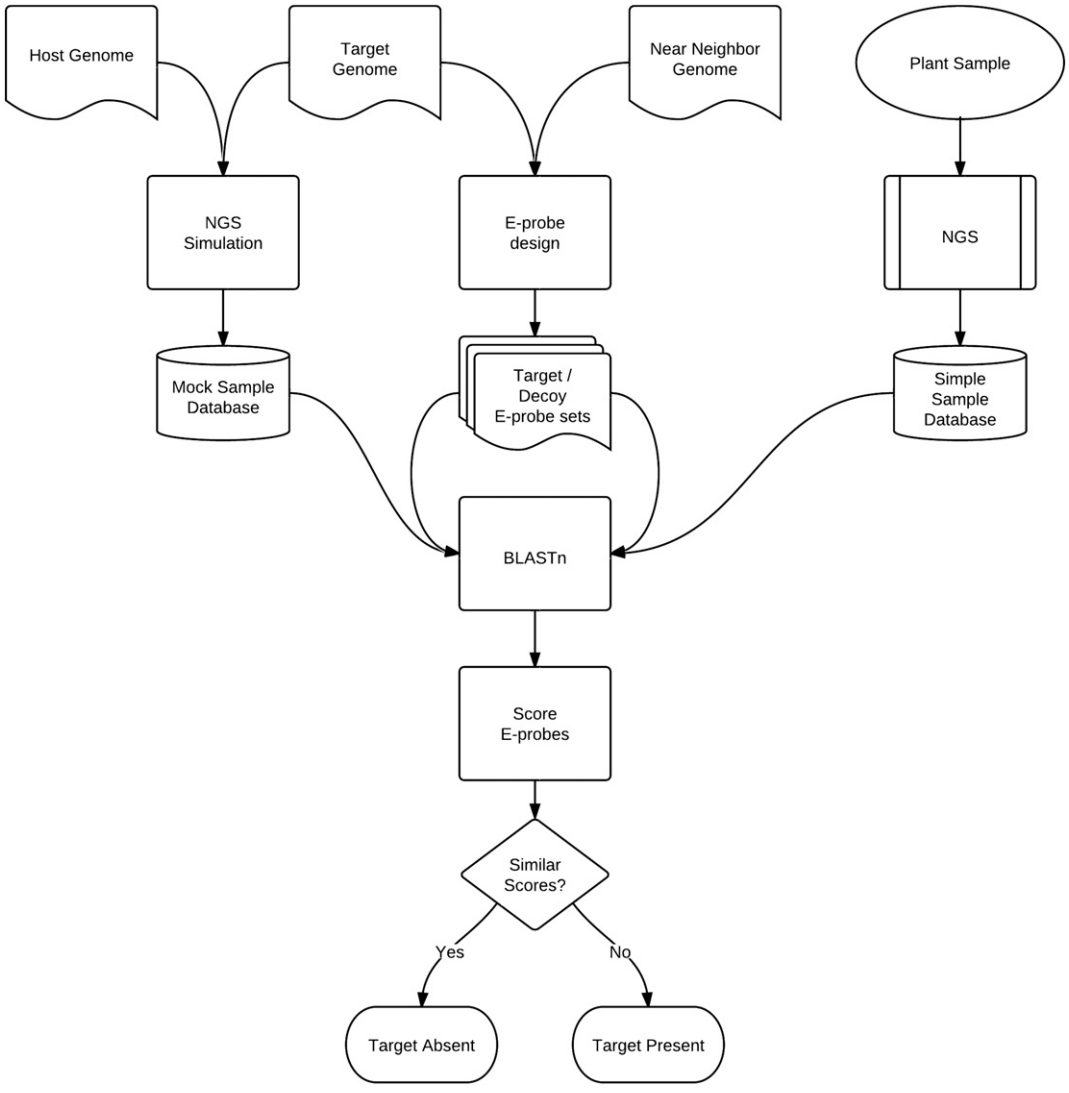

Fig. 1. Experimental flow of E-probe Diagnostic Nucleic acid Analysis pipeline.

resulted in the rapid detection of pathogen-associated sequences without extensive analysis of the metagenome.

2. Materials and methods

2.1. Pathogens and their sequences

The plant pathogens studied here belong to three general groups, viral, prokaryotic, and eukaryotic pathogens. The chosen systems represent a wide variety of plant pathogens and have global economic importance (Table 1). Two viruses were used: *Plum pox virus*, a single stranded RNA virus, and *Bean golden mosaic virus*, which is a bipartite DNA virus. Prokaryotic pathogens included *Xylella fastidiosa* 9a5c, the causal bacterium of citrus variegated chlorosis, *Xanthomonas oryzae* pv. oryzae, which causes bacterial blight in rice, and *Ralstonia solanacearum* race 3 biovar 2, a select agent that causes wilting of a variety of crops including potatoes and tomatoes, *Candidatus Liberibacter asiaticus*, a bacterium responsible for citrus greening, and *Spiroplasma citri*, which causes citrus stubborn disease. Eukaryotic pathogens included: *Puccinia graminis* a rust fungus, causing the stem rust of wheat and affecting a very broad host range including 365 cereals and grasses in 54 genera (Hodson et al., 2005); *Phytophthora ramorum*, a stramenopile with a wide host range of 23 species in 12 plant families (Rizzo and Garbelotto, 2003; Tyler et al., 2006); and *Phakopsora pachyrhizi*, which causes soybean

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 110005260 110005199 110005145 110005138 110005098 110005060 110005027 110004948 110004868 110004796 110004744 110004631 110004607 110004455 110004127 110004055 110003907M	*Mycobacterium bovis*	NC_008769	1525.76	423 (*309*)	33.84 (*24.72*)	2.22% (*1.62%*)
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 [29]	AAWC01000001 AAWC01004563	*Puccinia triticina*	ADAS01000001 ADAS01038776	66,652.40	21,790 (*21,635*)	1743.20 (*1730.80*)	2.66% (*2.65%*)
Phytophora ramorum	AAQX01000001 AAQX01007589	*Phytophora infestants*	AATU01000001 AATU01018288	88,644.63	21,286 (*18,945*)	1702.88 (*1515.60*)	1.92% (*1.71%*)

rust, a widespread pathogen that now can be found in Africa, Asia, Australia, South America and Hawaii (Miles et al., 2003). For each pathogen, a near neighbor was chosen based on a close phylogenetic relationship, and the availability of complete genome sequence (Table 1). Grapevine, *Vitis vinifera* (GenBank: PRJNA33471), was chosen as the host background due to the availability of its genome sequence, and its genome size, which is within the range of those of full plant genomes. While grapevine is not a natural host for many of the chosen pathogens, it serves well as an example of background sequences in which the target pathogen sequences exist.

2.2. 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.

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

2.4. Mock database construction

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

2.5. 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 an 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-value thresholds to find an e-value threshold with minimal false positives, with steps at 1×10^{-3}, 1×10^{-6}, and 1×10^{-9}.

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

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

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

3. Results

3.1. General

Plant pathogenic query production was analyzed in relation to genome size for two viruses, five bacteria, two fungi and one stramenopile. The targeted viral (*Plum pox virus* and *Bean golden mosaic virus*), fungal (*P. graminis* and *P. pachyrhizi*) and stramenopile (*P. ramorum*) plant pathogens were compared to near neighbors of the same species. For the bacteria, the *Ca. L. asiaticus* near neighbor was from the same

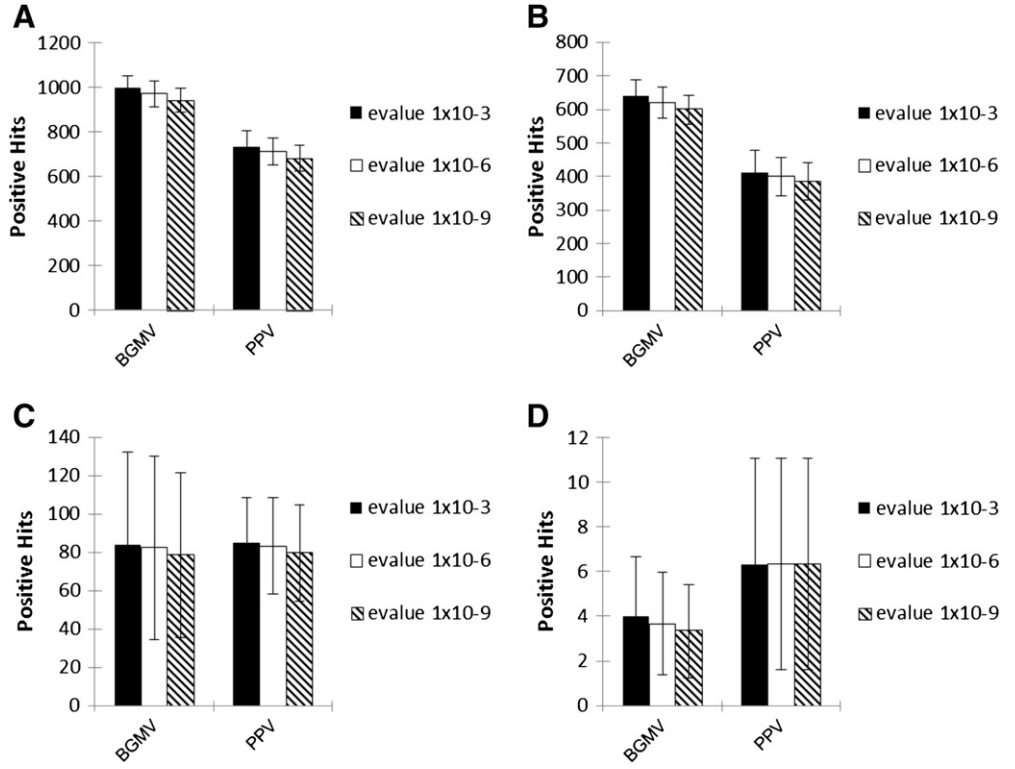

Fig. 2. The total number of hits from a BLAST search of 80 base target virus e-probe sets against MSDs containing grapevine and target pathogen sequences at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) <0.5% pathogen read abundances.

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 *P. graminis* and *P. 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 *P. ramorum*, *P. infestans* was used as near neighbor (Table 1). The

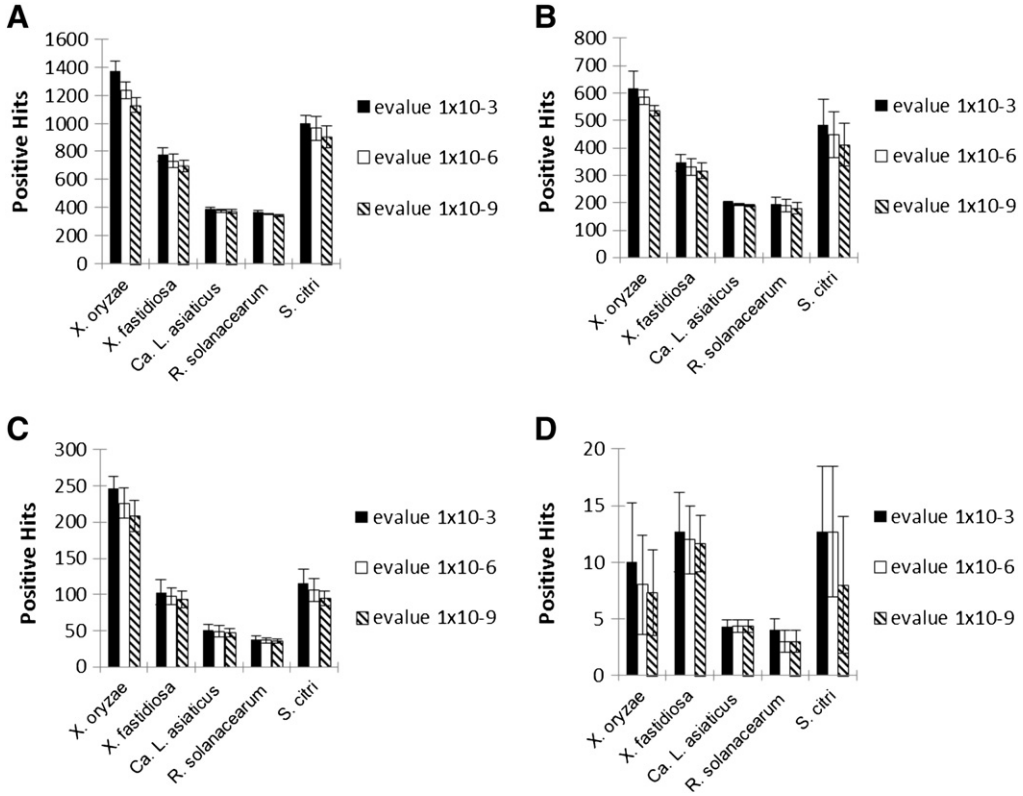

Fig. 3. The total number of hits from a BLAST search of 80 base target prokaryotic pathogen e-probe sets against MSDs containing grapevine and target pathogen sequences at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) <0.5% pathogen read abundances.

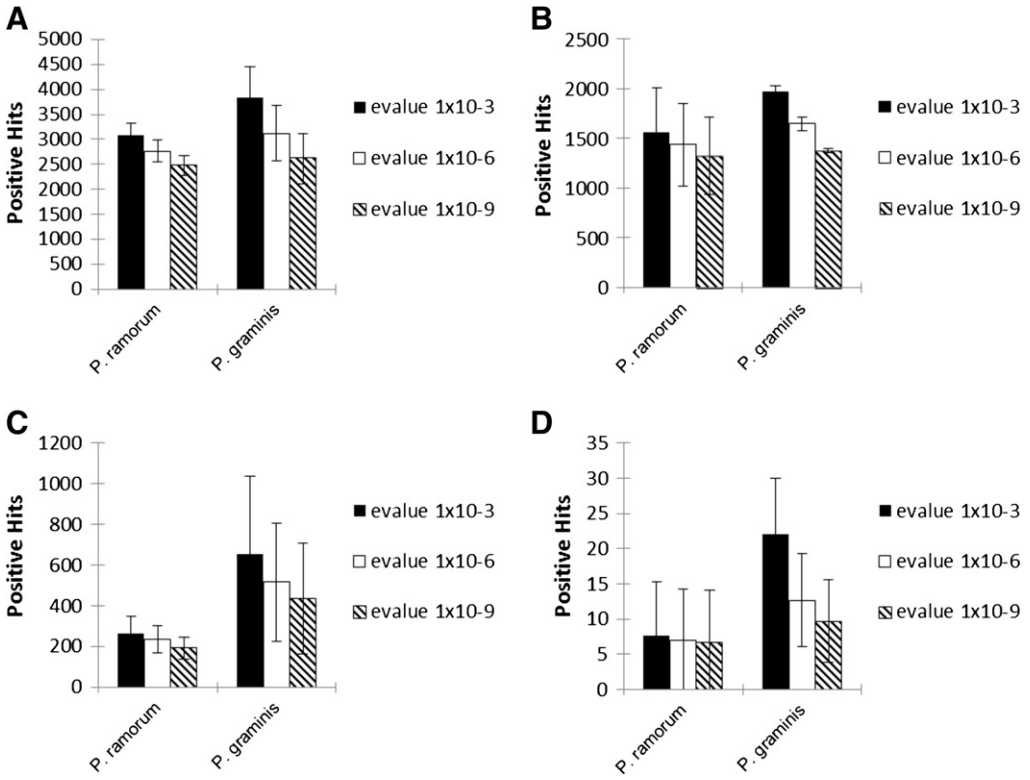

Fig. 4. The total number of hits from a BLAST search of 80 base eukaryotic pathogens e-probe sets against MSDs containing grapevine and target pathogen sequences at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) <0.5% pathogen read abundances.

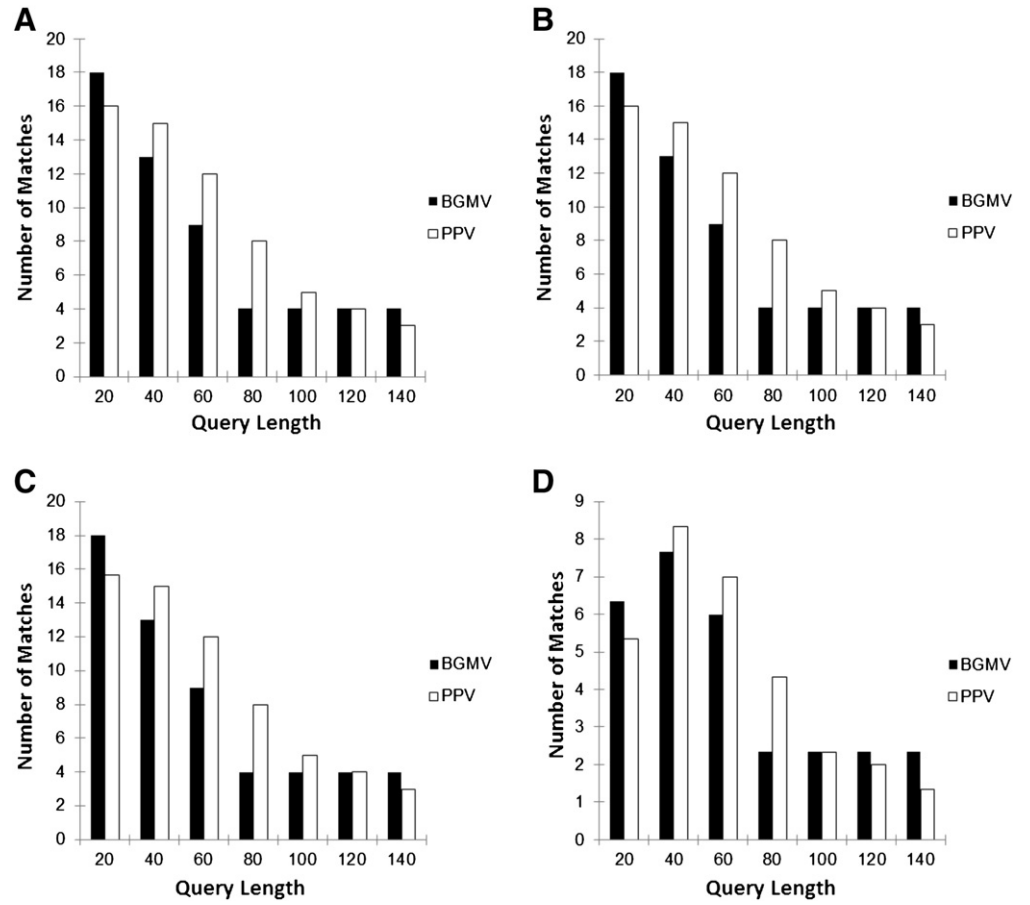

Fig. 5. Number of matches (positive e-probes) for each given length of e-probes, for target viruses at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) <0.5% pathogen read abundances.

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 1×10^{-3}, 1×10^{-6}, or 1×10^{-9} received for each pathogen was determined (Figs. 2–4). The number of positive hits rose with the size of the pathogen genome. As expected, the number of hits also 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 an 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 an MSD (Figs. 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) (Figs. 5–7, Table 1).

3.2. Optimization of e-probe length

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

The effect of varying e-probe lengths from 20 to 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,

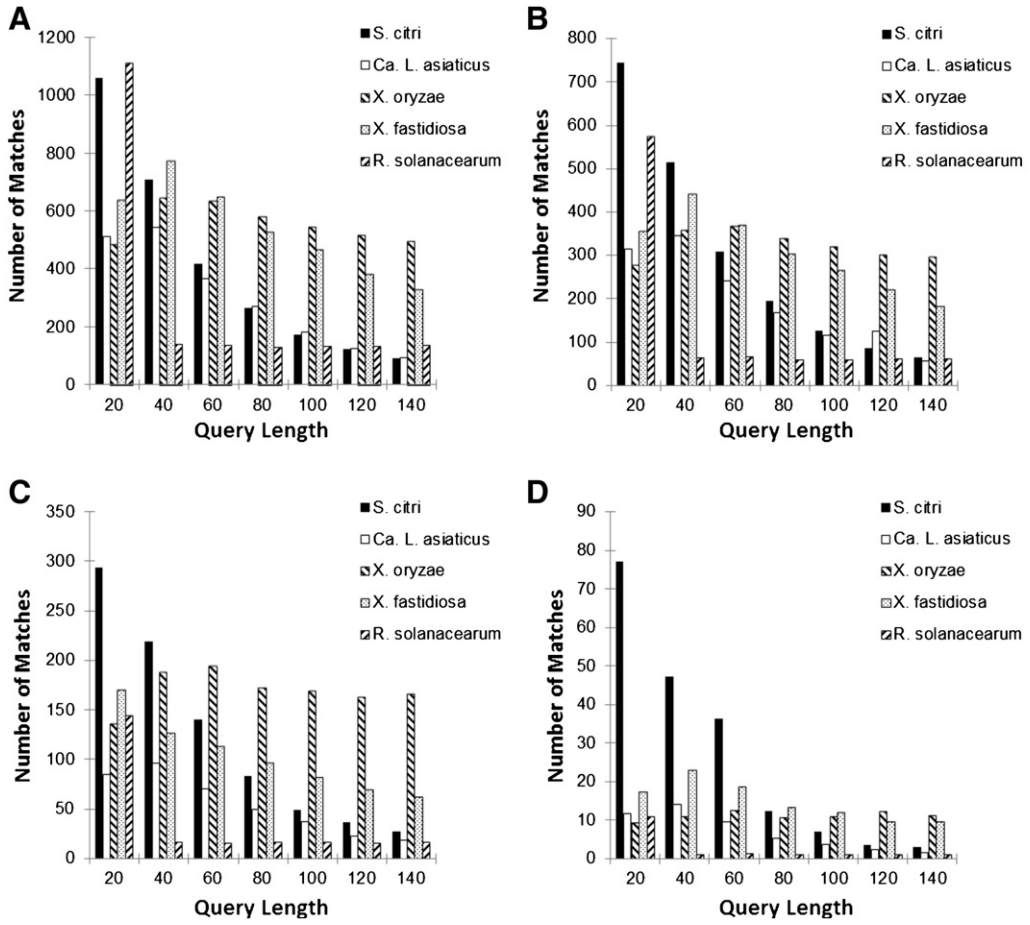

Fig. 6. Number of matches (positive e-probes) for each given length of e-probes, for target prokaryotic pathogens at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) <0.5% pathogen read abundances.

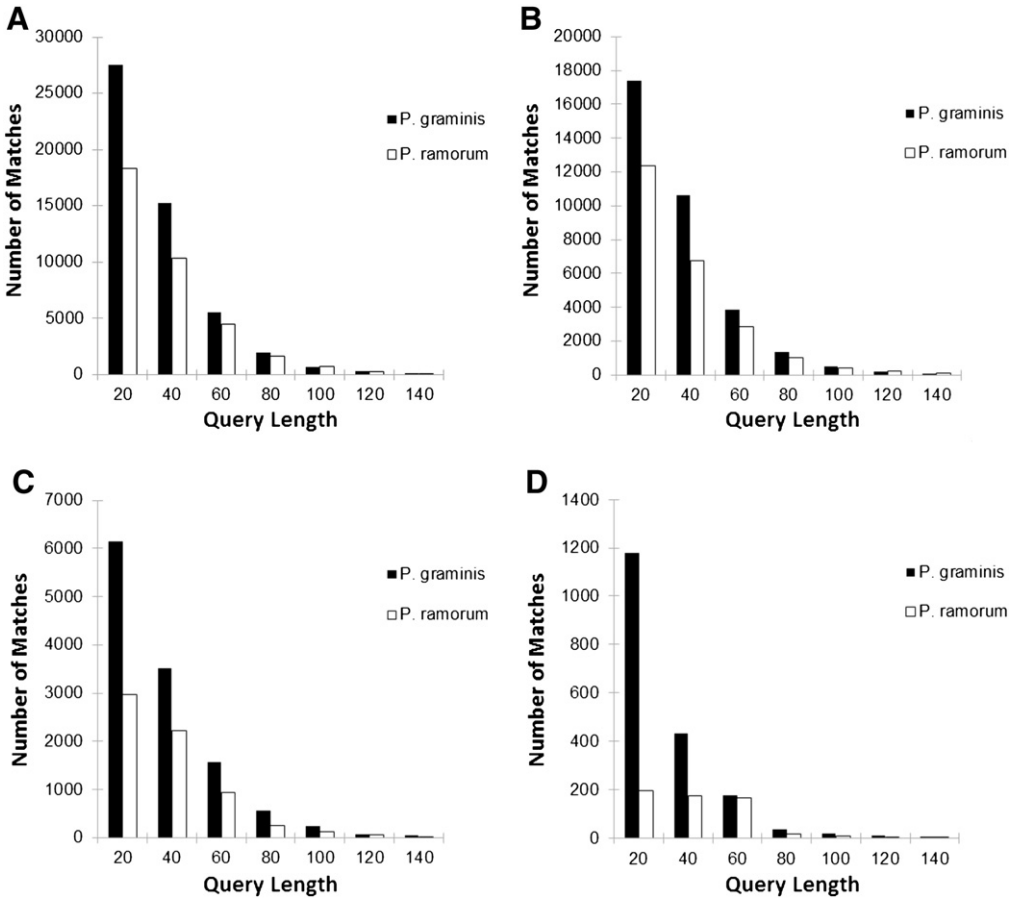

Fig. 7. Number of matches (positive e-probes) for each given length of e-probes, for target eukaryotic pathogens at (A) 15–25%, (B) 5–15%, (C) 0.5–5% and (D) <0.5% pathogen read abundances.

each pathogen type (virus, bacterial, and eukaryotic) had a similar number of matches for each member within a group (Figs. 5–7). One exception was *X. oryzae*, which showed no such downward trend (Fig. 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.

3.3. 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 1×10^{-3}. Using 80 nt queries, all of the pathogens were also detected in very low abundance databases, in some but not all replicates (Figs. 2–4, Supplementary Table 1).

Some e-probes generated false positive matches, i.e. instances when the e-probe sequence found a host counterpart in the MSD. The number of false positive matches was directly related to the e-values used in the BLASTn searches of the MSDs, with higher e-values generating more false positives. Overall, the eukaryotic pathogen simulations with a threshold e-value of 1×10^{-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 1×10^{-3} were observed in viral MSDs. The e-value was adjusted during the parsing step by using three different threshold e-values of 1×10^{-3}, 1×10^{-6}, and 1×10^{-9}. When the pathogens were analyzed using lower e-values, the number of false positives per database decreased from an average of 1 for prokaryotic e-probe sets, and 8 for eukaryotic e-probe sets to 0 for both.

Using the threshold values of either 1×10^{-6} or 1×10^{-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 1×10^{-3}) to 1530 matches (1×10^{-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 1×10^{-3} to 1×10^{-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 bioinformatic 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 62 nt and with the error model found for a typical Illumina run (Richter et al., 2008). EDNA analysis showed similar results to the 454 simulations (data not shown).

3.4. BLAST check comparison

False positives were reduced in number by removing e-probes that have similarity to known sequences in NCBI. Each 80 nt e-probe set

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

was used as queries in a search against the NCBI GenBank nt database. E-probes with hits at an e-value of 1×10^{-10} or lower were removed from the probe set. This decreased the number of probes per set by up to 50% (Table 1). Comparing the performance of BLAST-checked e-probe sets showed a slight reduction in the number of false positive hits, with a larger reduction in the number of matches and total hits (Supplemental Table 1).

3.5. Determination of positives and negatives

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

4. Discussion

There are multiple advantages to using a metagenomic-based approach to pathogen diagnostics. Advances in NGS have made it possible to generate billions of bases of sequence for any given sample, creating metagenomes that represent a complete profile of all organisms in a given nucleic acid sample, including host, endophytes and pathogens (Jones, 2010; Tyson et al., 2004). This presents the very real probability that any and all microbes in any given sample could be identified. Metagenomic 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 metagenomic-based diagnostics is the current per run cost. The typical approach to a metagenome diagnosis is nucleic acid extraction, sequencing, sequence assembly, and BLAST analysis of the assembled contigs. An examination of recent history suggests that sequencing technologies will likely become less expensive, due to the technologies becoming faster, more accessible and the sequencing more processive over time, outpacing Moore's Law. This prediction suggests that NGS costs may not be a long term restraint, particularly when combined with barcoding (Parameswaran et al., 2007). However, the very same advances that drive down per sample costs of sequencing create additional data handling problems. As NGS becomes less expensive, faster and the length of reads increases, the number of bases sequenced in a single run will increase exponentially. These same advances in NGS will have an additional exponential growth effect on the databases (i.e. GenBank and its subsidiaries) that are used for the BLAST searching of sequence data, suggesting that the current metagenomic approach to pathogen diagnostics will eventually become too computationally intensive for everyday use.

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

The EDNA approach was tested using a series of MSDs representing potential metagenomes with pathogen sequences in a plant background. Representatives of multiple taxonomic groups of plant pathogens were used, including an RNA virus, a DNA virus, a spiroplasma, prokaryotes, a stramenopile, and a fungus. Diagnostic e-probe sequences were selected at a range of lengths, and used to query MSDs with differing levels of pathogen abundance (from 0.5% pathogen reads to 25% pathogen reads). EDNA was successful at detecting all pathogens at low, medium and high levels (everything above 0.5% pathogen reads in the MSD). The number of matches (any instance where an

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

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 1×10^{-9}, and the prokaryotic pathogen e-probes were completely specific when the parsing e-value was lowered to 1×10^{-6}. Larger, more complex genomes and the conservation of genes and sequences between pathogen and host (eukaryotic pathogens) require lower e-value cutoff levels. It should also be noted that some of the near neighbors were less related to the target organisms, a limitation driven by the lack of available sequenced genomes. Improved near neighbors, which should become available as more pathogen genomes are sequenced, will also improve precision.

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

A key to any diagnostic method is determining the level of positive "signal" necessary to confirm that a pathogen is present in a given sample. For molecular techniques such as PCR, the presence or absence of a product is easily distinguished. However when the positive/negative decision is based on a quantitative measurement, such as fluorescence or absorbance in ELISA, the determination involves some level of

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

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mimet.2013.07.002.

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 a critical review of this manuscript.

References

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

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

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., et al., 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., et al., 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318, 283–287.

Daniel, R., 2005. The metagenomics of soil. Nat. Rev. Microbiol. 3, 470–478.

Gamliel, A., Gullino, M.L., Stack, J.P., 2008. Crop biosecurity: local, national, regional and global perspectives. In: Gullino, M.L., Fletcher, J., Gamliel, A., Stack, J.P. (Eds.), Crop Biosecurity. Springer Netherlands, pp. 37–61.

Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., et al., 2006. Metagenomic analysis of the human distal gut microbiome. Science 312, 1355–1359.

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

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

Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., et al., 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., et al., 2010. Bioinformatics for next generation sequencing data. Genes 1, 294–307.

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

Miles, M.R., Frederick, R.D., Hartman, G.L., 2003. Soybean Rust: Is the U.S. Soybean Crop at Risk? APS Feature Story.American Phytopathological Society.

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

Parameswaran, P., Jalili, R., Tao, L., Shokralla, S., Gharizadeh, B., et al., 2007. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acid Res. 35, e130.

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

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

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

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

Richter, 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.

Rizzo, D.M., Garbelotto, 2003. Sudden oak death: endangering California and Oregon. Front. Ecol. Environ. 1, 197–204.

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

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

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

Tringe, S.G., Rubin, E.M., 2005. Metagenomics: DNA sequencing of environmental samples. Nat. Rev. Genet. 6, 805–814.

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

Tyler, B.M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R.H., Aerts, A., Arredondo, F.D., Baxter, L., Bensasson, D., Beynon, J.L., Chapman, J., Damasceno, C.M., Dorrance, A.E., Dou, D., Dickerman, A.W., Dubchak, I.L., Garbelotto, M., Gijzen, M., Gordon, S.G., Govers, F., Grunwald, N.J., Huang, W., Ivors, K.L., Jones, R.W., Kamoun, S., Krampis, K., Lamour, K.H., Lee, M.K., McDonald, W.H., Medina, M., Meijer, H.J., Nordberg, E.K., Maclean, D.J., Ospina-Giraldo, M.D., Morris, P.F., Phuntumart, V., Putnam, N.H., Rash, S., Rose, J.K., Sakihama, Y., Salamov, A.A., Savidor, A., Scheuring, C.F., Smith, B.M., Sobral, B.W., Terry, A., Torto-Alalibo, T.A., Win, J., Xu, Z., Zhang, H., Grigoriev, I.V., Rokhsar, D.S., Boore, J.L., 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313, 1261–1266.

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

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