A new approach for detecting fungal and oomycete plant pathogens in next generation sequencing metagenome data utilising electronic probes

Andres Espindola*

Entomology and Plant Pathology Department,
Oklahoma State University,
127 Noble Research Center Stillwater OK 74078, USA
Email: andres.espindola@okstate.edu
*Corresponding author

William Schneider

USDA-ARS FDWSRU,
1301 Ditto Ave. Fort Detrick MD 21702, USA
Email: William.Schneider@ARS.USDA.GOV

Peter R. Hoyt

Department of Biochemistry and Molecular Biology,
246 Noble Research Center Stillwater OK 74078, USA
Email: peter.r.hoyt@okstate.edu

Stephen M. Marek

Department of Entomology and Plant Pathology,
Oklahoma State University,
127 Noble Research Center Stillwater OK 7408, USA
Email: stephen.marek@okstate.edu

Carla Garzon

Department of Entomology and Plant Pathology,
127 Noble Research Center Stillwater OK 74078, USA
Email: carla.garzon@okstate.edu

Abstract: Early stage infections caused by fungal/oomycete spores may not be detected until signs or symptoms develop. Serological and molecular techniques are currently used for detecting these pathogens. Next-generation sequencing (NGS) has potential as a diagnostic tool, due to the capacity to target multiple unique signature loci of pathogens in an infected plant metagenome. NGS has significant potential for diagnosis of important eukaryotic plant pathogens. However, the assembly and analysis of huge amounts of sequence is laborious, time consuming, and not necessary for diagnostic purposes. Previous work demonstrated that a bioinformatic tool termed Electronic probe Diagnostic Nucleic acid Analysis (EDNA) had potential for greatly simplifying detecting fungal and oomycete plant pathogens in simulated metagenomes. The initial study demonstrated limitations for detection accuracy related to the analysis of matches between queries and metagenome reads. This study is a modification of EDNA demonstrating a better accuracy for detecting fungal and oomycete plant pathogens.

Keywords: sequencing; EDNA; electronic probe; e-probe; *Puccinia graminis*; *Pythium ultimum*; *Phakopsora pachyrhizi*; *Phytophthora ramorum*; 454 Roche; diagnostics; fungi; pucciniomycetes; *Chromalveolata*; oomycetes.

Reference to this paper should be made as follows: Espindola, A., Schneider, W., Hoyt, P.R., Marek, S.M. and Garzon, C. (2015) 'A new approach for detecting fungal and oomycete plant pathogens in next generation sequencing metagenome data utilising electronic probes', *Int. J. Data Mining and Bioinformatics*, Vol. 12, No. 2, pp.115–128.

Biographical notes: Andres Espindola is a PhD student with the Entomology and Plant Pathology Department at Oklahoma State University, Stillwater, OK. He is currently working at the soilborne plant diseases laboratory with Dr. Carla Garzon as his Principal Investigator. His work focuses on the diagnosis of fungal and oomycete plant diseases using bioinformatics tools in Next Generation Sequencing Data as well as metagenome data analysis using Next Generation Sequencing Data. He received his Master's degree from Oklahoma State University in 2013. He has done internships at the Foreign Disease/Weed Science Research Unit on Fort Detrick, MD working with Next Generation Sequencing samples and data analysis.

William Schneider is a research plant pathologist with the US Department of Agriculture Agricultural Research Service. He is stationed at the Foreign Disease/Weed Science Research Unit on Fort Detrick, MD. He received his PhD from Michigan State and did Post-doctoral research at the Noble Foundation in Ardmore, OK prior to beginning work with the USDA. His research focuses on plant pathogen diagnostics, plant virus evolution, and virus/vector relationships. Examples of related recent publications include Stobbe et al., 2013 (*Journal of Microbiological Methods* 94: 356–266) and Roy et al., 2013 (*Journal of Data Mining in Genomics and Proteomics* 4: 1–6).

Peter R. Hoyt is a Senior Research Scientist and Director of the Bioinformatics Certificate Graduate Program at Oklahoma State University in the Department of Biochemistry and Molecular Biology. He received his PhD in Human Genetics and Cell Biology at the University of Texas Medical Branch in Galveston, Texas. Recent publications include "Screening Metagenomic Data for Viruses Using the E-Probe Diagnostic Nucleic Acid Assay" Stobbe et al. in press, "Draft Genomes of Heterogeneous Vancomycin-Intermediate Staphylococcus aureus Strain MM66 and MM66 Derivatives with Altered Vancomycin Resistance Levels" Matyi et al., *Genome Announc.* 2014 Jul 10;2(4), and "Draft genomes of Staphylococcus aureus strains isolated from raw milk samples" Matyi et al., *J Dairy Sci.* 2013 Jun;96(6):3535–3542. His research includes stem-cell determination mechanisms, antibiotic resistance transmission and evolution, and microbiome metagenomics.

Stephen M. Marek, Associate Professor, PhD Plant Pathology (UC-Davis), MS Agronomy (Univ Missouri), studies the biology and functional genomics of plant pathogenic fungi and teaches mycology at Oklahoma State University. He is currently investigating switchgrass rust, *Phymatotrichopsis* root rot, Phoma leaf spot and *Fusarium* diseases. He is member of the American Phytopathological Society and Mycological Society of America.

Carla Garzon is an Associate Professor of Plant Pathology at the Oklahoma State University. She received her doctorate in Plant Pathology from the Pennsylvania State University (PSU). Additionally, she was a postdoctoral researcher at PSU and the Ohio State University, and a Visiting Assistant Professor at the College of Wooster. Her research program focuses on diverse aspects of the biology, population genetics, phylogenetics, epidemiology, and molecular detection of Soilborne fungal and Oomycete plant pathogens, and management of the diseases they cause.

1 Introduction

Plant pathogen and pest dispersal into alien ecosystems has significant economic, ecological and evolutionary consequences. These introductions have the potential of being irreversible in the structure and functions of specific ecosystems, principally agricultural ecosystems (Castello et al., 1995; Enserink, 1999; Everett, 2000). Plants lack immune systems, and post infection therapies are limited, expensive and frequently unfriendly to the environment. Hence, the development of rapid and accurate diagnostic methods for plant pathogens is crucial for the implementation of trading regulations. New method development is crucial for high impact pathogens, while evaluation of new methods on model systems can demonstrate the application of the newly developed methods to a broader range of pathogens. This is particularly true for fungal and oomycete plant pathogens, which are the most common and highest impact pathogens of crops (Strange, and Scott, 2005). In addition, diagnostics for fungal and oomycete plant pathogens are often complicated by the need to distinguish either fungal or oomycete pathogen from the eukaryotic host background, compounded by the rarity of pathogen genome sequence availability.

Diagnostic tools for the presence of oomycete and fungal plant pathogens exist but they are limited in their scope and versatility. The most frequently used plant pathogen technologies are either antibody based assays, such as ELISA, or nucleic acid detection assays, such as the many variations of PCR. These systems are useful in specific cases where the goal is detection of one or a few well characterised pathogens. However, these assays are completely dependent on the availability of sequence information or antibody reagents, and the multiplex capacity of these assays is limited. The recent development of new nucleic acid sequencing technologies has allowed the development of innovative approaches to detect plant pathogens. Next generation sequencing (NGS) is a relatively recent technology that generates large amounts of sequence data from a given sample (Ronaghi, 2001). The productivity of NGS technology far exceeds that of traditional Sanger sequencing (Magi et al., 2010), enabling the field of metagenomics, which generates a sequencing profile that represents any and all organisms present within the sample (Jones, 2010; Tyson et al., 2004). 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).

The major advantage in NGS metagenomic approaches to pathogen discovery and detection is the sheer volume of sequence information available for analysis. However, this strength also is as a limitation of these approaches, as data handling becomes cumbersome. The typical metagenomics approach involves NGS via a single or multiple platforms followed by sequence filtering, assembly and BLAST comparison to a database such as Genbank (Huson et al., 2011; Segata et al., 2012). This process is laborious and time consuming, requiring significant computing resources. Data management problems will only become worse as inevitably sequencing becomes more processive and reference sequence databases grow larger.

One approach to dealing with the limitations of metagenomic based diagnostics is called Electronic-probe Diagnostic Nucleic acid Analysis, or EDNA (Stobbe et al., 2013). EDNA limits post-sequencing manipulation of data, completely removing assembly. In addition, the NGS data is queried by a limited number of signature sequences or electronic probes (e-probes), rather than comparing the entire sequence data set against a comprehensive database. The original intent for developing EDNA was to create a single manageable tool by which NGS could be used for the detection of any plant pathogen: virus, prokaryote or eukaryote. EDNA was successful in achieving its goal when tested using simulated metagenomes (Stobbe et al., 2013). However, while the data demonstrated that a one-size-fits-all bioinformatic approach may work, such an approach may not be ideal or appropriate for dealing with such diverse classes of pathogens. In addition, the limited number of simulated databases analysed made it difficult to do accurate projections of efficiency for any particular class of pathogens. The objective of this study was to optimise the EDNA bioinformatic tool for the analysis of metagenomic data from plants infected by fungi and/or oomycetes.

2 Materials and methods

2.1 Generation of Mock Sample Sequencing Databases (MSSDs)

The development of E-probe Diagnostic Nucleic acid Analysis (EDNA) as a diagnostic tool required prior *in silico* assessment (Figure 1). Mock sample sequence databases (MSSDs) representing simulated metagenomes were generated using three high impact fungal and oomycete plant pathogens as previously described (Stobbe et al., 2013). *Puccinia graminis* f. sp. *tritici* (Fungi, Pucciniomycetes) is the causal agent of wheat stem rust, a disease severely affecting wheat, but can also affect rye, barley and oats. *Phakopsora pachyrhizi* (Fungi, Pucciniomycetes) is an obligate plant pathogen and the causal agent of Soybean rust. *Phytophthora ramorum* (Chromalveolata, Oomycetes) is the causal agent of Sudden Oak Death disease (SOD) and Ramorum blight, and *Pythium ultimum* (Chromalveolata, Oomycetes) is the causal agent of damping off and root rots in a broad range of hosts (Levesque et al., 2010). This assessment included MSSDs constructed with different pathogen-read ratios (Table 2), created by a metagenomics simulator called Metasim (Richter et al., 2008). Different pathogen/host ratios were used to evaluate the sensitivity of the technique using BLAST+ and e-probes.

The genomes of *P. ultimum* (Levesque et al., 2010)*,* *P. ramorum* (Tyler et al., 2006) and *P. graminis* (Puccinia Group Sequencing, 2012) were used for making MSSDs, while EST libraries were used for *P. pachyrhizi* databases*.* The background host genome was grapevine (*Vitis vinifera*) (Velasco et al., 2007). This plant host was selected because the public grapevine genome has 12x coverage and most of it has been annotated, which facilitates further bioinformatics and statistical analysis. One hundred replicates were done of every MSSD. The variable parameter in MSSDs was pathogen read abundance (Table 2). Pathogen read abundance included four different abundances (high, medium, low and very low). The background read abundance was dependent on the pathogen read abundance and the total number of reads, which were limited to 10,000 per MSSD for the assessment *in silico* (Table 2).

2.2 E-Probe design

E-probe design at the species level required two genomes (Table 2), the target genome and a near neighbour genome. The target genome acted as a template to generate e-probes and the near neighbour helped to eliminate redundant genome regions in the target genome. Both, the elimination of redundant genome regions and the e-probe development were performed by a tool implemented to design pathogen diagnostic fingerprints termed Tool for Oligonucleotide Fingerprint identification (TOFI) (Vijaya Satya et al., 2008). The identification tool was originally built in Perl language for use with both high-performance computing (HPC) and/or personal computers. It included various versions: TOFI alpha was the first version of the program and it is a personal computer version (Tembe et al., 2007); TOFI beta that included several optimisations and significantly reduced the overall execution time of the pipeline (Vijaya Satya et al., 2008); and a final version that included the parallel implementation for HPC (Ravi Vijaya, 2009).

EDNA modified TOFI's pipeline by eliminating the UNAFold stage that involved microarray probe selection based on melting temperatures, and two-state folding or hybridisation calculations. Instead, e-probes with varying lengths (40, 60, 80, 100, 120, 140, 160, 180 and 200 nt) were developed when comparing the target genome (pathogen genome) against the near neighbour genome (Table 1). TOFI takes advantage of SNP finding to select non-redundant areas of the target genome to develop the unique e-probes. In certain NGS platforms errors are created when DNA contains homopolymeric regions with a length greater than 5-6 nt (Ronaghi, 2001). To make these errors irrelevant while using EDNA, sequence regions with homopolymers were eliminated from the target genome sequence before it was processed by the modified TOFI. Therefore, e-probes lacked homopolymers.

All the e-probe databases were curated so that e-probes resulting in false positives or were redundant with public genomic data from other organisms were eliminated, making the database more specific. Two Perl scripts called falsepositive_eliminator.pl and parser_falsepositive_eliminator.pl were used in this task. E-probe databases were pairwise aligned with sequences available through the partially non-redundant nucleotide collection (nr/nt) database (nt) from the National Center for Biotechnology Information (NCBI), and any e-probe that aligned with an e-value score lower than $1x10^{-3}$ was eliminated from the e-probe database. Higher e-values were not utilised to avoid the elimination of excessive number of pathogen-specific e-probes. Parsing the hits and matches from the previous alignments allowed to discriminate among alignments that certainly belonged to the pathogen (using GI numbers) and alignments that showed ambiguous (pathogen and third organism) alignments.

Table 1 Target genome information used for the e-probe design of the four different pathogens

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

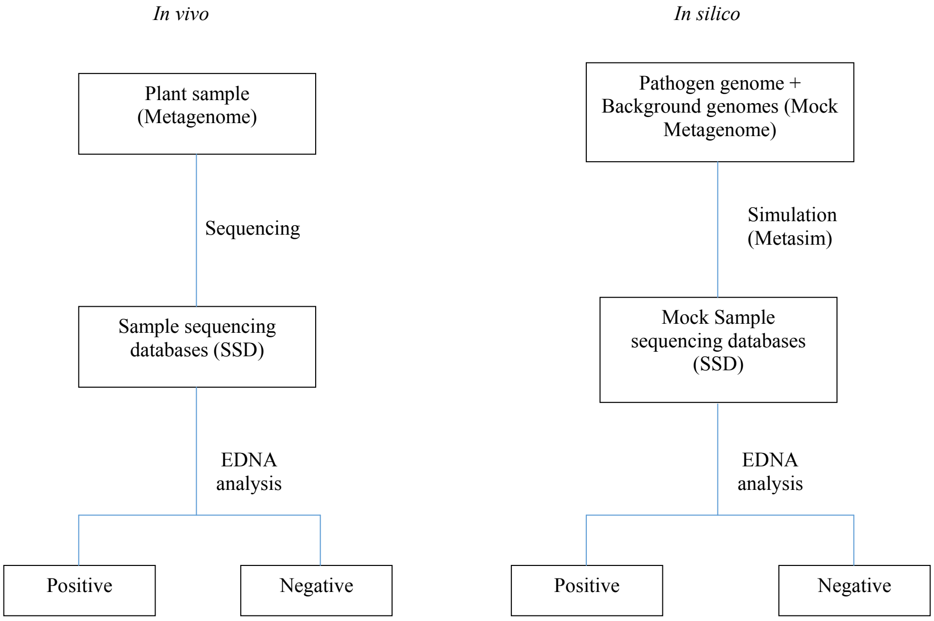
The output of TOFI was a set of unique pathogen-specific e-probes that were used for EDNA assessment. E-probes with varying lengths (40–200 nt) were re-evaluated against MSSDs. The use of varying e-probe lengths was justified by the presence of varying read lengths and randomisation of 454 pyrosequencing library fragments during DNA fragmentation. The optimal e-probe sizes were identified for each species using a newly developed criteria for scoring base pair matches when using EDNA.

2.3 EDNA scoring

MSSDs were subjected to analyses with EDNA using the “Cowboy” supercomputer at Oklahoma State University. The EDNA work flow is demonstrated in Figure 1. Pairwise sequence alignment was performed between e-probes and MSSDs using BLASTn (Camacho et al., 2009). The presence of the pathogen was detected when the pathogen-specific e-probes and sequences in the MSSDs aligned (hits). An individual e-probe may have multiple hits in a MSSD, so the term matches is used to describe the total number of e-probes that have one or more hits in the BLASTn analysis. Hits with e-values equal or less than $1\text{x}10^{-9}$ and percent identity 95% or higher were considered high score hits (HSH) and were counted towards a positive match. An e-probe was considered high-quality if it had multiple HSHs with MSSDs. Multiple high-quality HSHs resulted in a positive match called a High Quality Match (HQM). The HQMs must have a coverage depth of at least 4x in order to confirm the presence of the pathogen in the metagenome. This coverage depth was considered reliable since the error rates of 454 pyrosequencing yields have an insertion and deletion error rate of approximately 3.3%, and a substitution error rate of 0.5% (Margulies et al., 2005). Even with a higher error rate, it has been observed that consensus accuracies of 99.99% are achieved with a coverage depth of 4x or more (Margulies et al., 2005). However, depending of the pathogen biology, genome size and at high titers, coverage depths less than 4x may be sufficient for defining HQMs.

Negative control MSSDs did not contain any pathogen sequences and thus were not expected to show any matches. In addition to negative control MSSDs, non-specific e-probes were evaluated with positive MSSDs. Two types of non-specific e-probes were generated, “decoy” e-probes and “shuffled” e-probes. Decoy e-probes were generated by reversing the DNA sequence of pathogen specific e-probes, which should convert them into non-specific e-probes, but could match naturally occurring DNA inversions, found in genetically variable populations of some fungi (Hane et al., 2011). Shuffled e-probes were generated from pathogen specific e-probes with a Perl script that randomly shuffled the nucleotide positions.

Figure 1 EDNA concept in vivo and in silico for the diagnosis of eukaryotic plant pathogens


In vivo
In silico
Plant sample (Metagenome)
Sequencing
Sample sequencing databases (SSD)
EDNA analysis
Positive
Negative
Pathogen genome + Background genomes (Mock Metagenome)
Simulation (Metasim)
Mock Sample sequencing databases (SSD)
EDNA analysis
Positive
Negative

In Stobbe et al. (2013) positive and negative results were assessed using statistical analysis where decoy e-probes were used as a negative control. The principle behind this decoy test was to determine a BLASTn “background” level, based on the idea that any comparison of two large sequence sets will result in occasional alignments. This method was useful in providing a statistical level of confidence in a positive/negative call, but even in limited testing it was apparent that different parameters were needed depending on the class of pathogen. Therefore, in this study, decoy e-probes were mostly used to look for the presence of inversions of specific chromosome areas in the four fungal and oomycete plant pathogens.

False positive calls were positive EDNA calls in a MSSD that lacked pathogen reads. True positive calls were positive EDNA calls when a MSSD contained target pathogen reads. Also, true negative calls were negative EDNA calls when pathogen reads were not present in a MSSDs. Finally, false negative calls were negative EDNA calls in MSSDs that were known to contain pathogen reads.

2.4 Sensitivity and specificity analysis

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

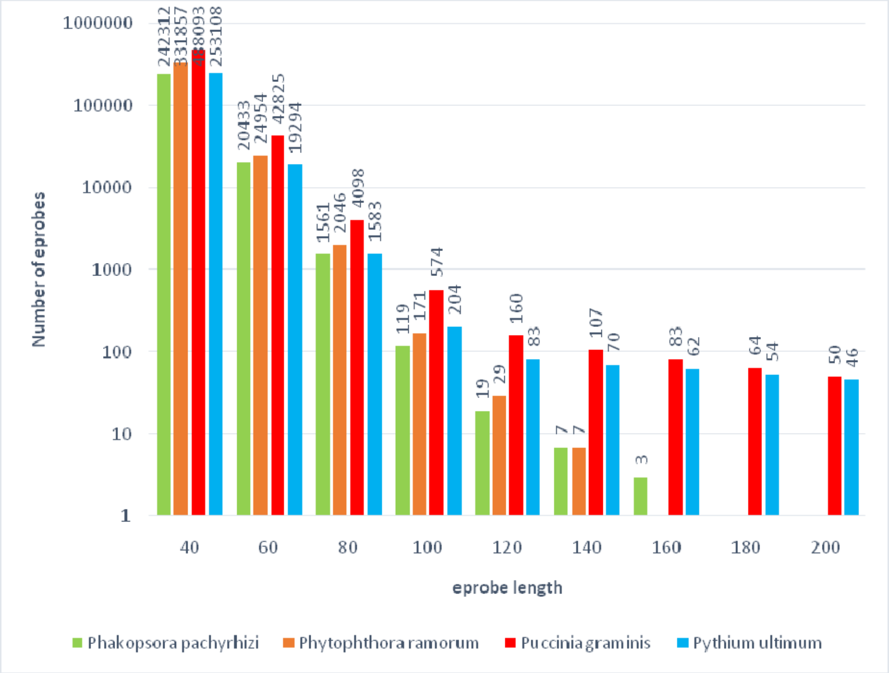
These tests assessed the reliability of the proposed detection/identification model. The specificity analysis formula was $S_p = \dfrac{TN}{(FP + TN)}$, where TN is the number of true negative calls and FP is the number of false positive calls. The sensitivity analysis formula was $S_n = \dfrac{TP}{(TP + FN)}$, where TP is the number of true positive calls and FN is the number of false negative calls.

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

Table 2 Molecular parameters for the construction of MSSDs

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

Figure 2 Variation of the number of e-probes designed among pathogens and e-probe length (see online version for colours)



3 Results and discussion

3.1 E-probe design

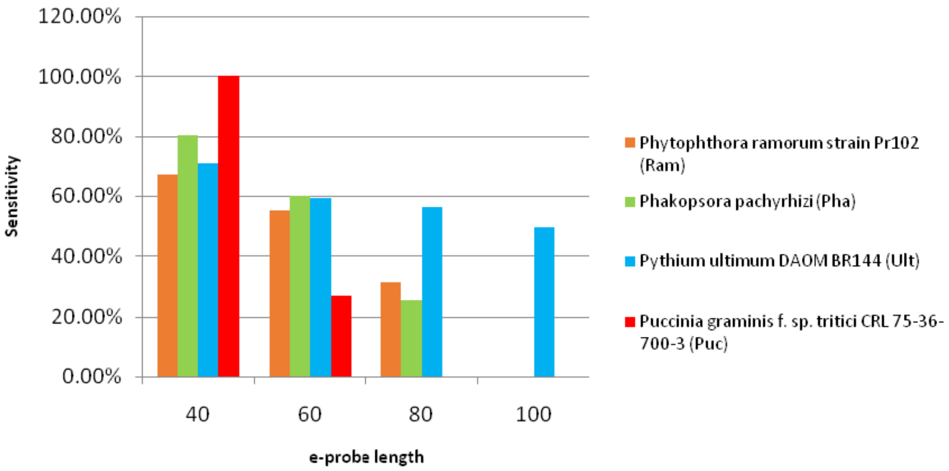
E-probe length was a limiting factor for the number of e-probes designed. As the e-probe length increases, the number of e-probes that the modified TOFI was able to design decreased (Figure 3). The number of e-probes at each of the lengths varied among different pathogens (Figure 3). Various parameters were measured to select the best e-probe length for pathogen detection, including sensitivity, specificity, and data processing time. E-probes 40 nucleotides in length were produced on the scale of hundreds of thousands for the fungal and oomycete pathogens studied. As e-probe length increased, the likelihood of finding pathogen specific sequences decreased, resulting in lower numbers of longer e-probes for all pathogens (Figure 3). The effect was more dramatic for *P. ramorum* and *P. pachyrhizi*, where no e-probes were generated at lengths greater than 160 nucleotides. The numbers of e-probes generated for these two pathogens was similar to the other two pathogens (*P. ultimum* and *P. graminis*) at shorter lengths (100 nucleotides and smaller), but notably fewer e-probes were generated for *P. pachyrhizi* and *P. ramorum* at lengths of 120 nucleotides and above. This would be expected for *P. pachyrhizi*, where the amount of available sequence used for e-probe selection was limited due to the use of an EST library instead of a complete genome. Expressed sequences are a small portion of a complete eukaryotic genome, and have the added disadvantage of potential functional conservation, especially between near neighbours. This is of particular interest in developing e-probe sets for eukaryotic plant pathogens, as the availability of complete genomes is very limited, and e-probe design will frequently rely on a limited number of expressed sequences. The reasons behind the limited number of long e-probes for *P. ramorum* is not entirely clear, although it is likely related to the closeness of near neighbours and/or genome size.

For all probe lengths, *P. graminis* generated the most e-probes. The genome sizes related to the number of e-probes showed that 40nt e-probes used 23.92% of the total genome for *Puccinia graminis,* similarly for *P. ramorum* only 19.86% of the genome was utilised and for *P. ultimum* 22.46% (data not shown). When the e-probe size increased, the portion of the genome used for the e-probe design decreased to a proportion between 2 and 3% (data not shown).

3.2 Mock sample sequencing database design

One hundred MSSDs were constructed for each of the four plant pathogens at each pathogen read abundance (H, M, L and VL pathogen read abundances; Table 2). Each MSSDs was comprised of 10,000 reads total, with the variable pathogen read abundances. Each MSSD was constructed to simulate typical 454 settings producing 200 cycles with an average of 509 base pairs per read. In terms of sequencing errors, the average substitution rate in MSSDs was zero while the average insertion rate was 2.29% and the average deletion rate was 0.62%. Error values in MSSDs were in accordance with sequencing errors reported for 454 pyrosequencing (Margulies et al., 2005).

Figure 3 Relationship among e-probe length and sensitivity while using EDNA in four eukaryotic plant pathogens and four different pathogen abundances combined (see online version for colours)



3.3 Diagnostics with EDNA

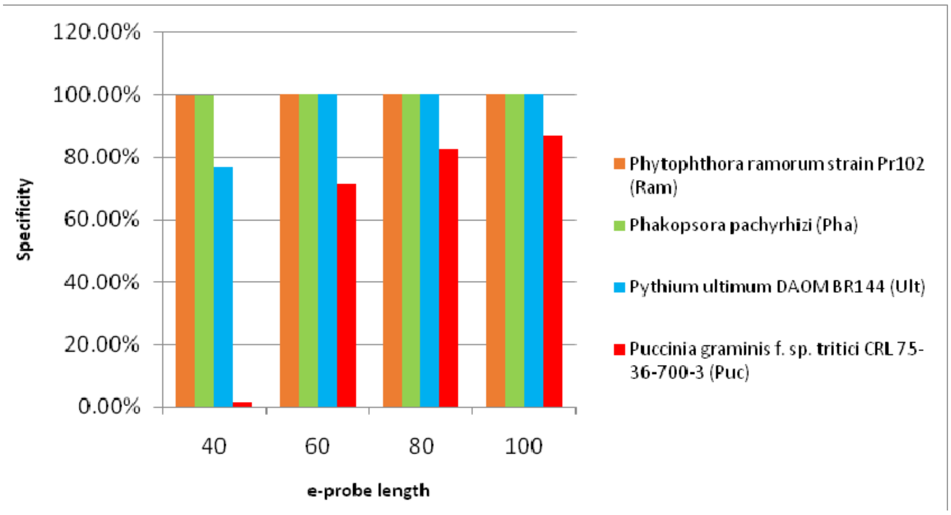
Approximately 2,000 EDNA analyses were performed to provide a statistically valid sample size of 454 pyrosequencing metagenome databases. All the pathogens were detected at high, medium, and low read abundances. Very low read abundances produced ambiguous results due to false positives while performing cross analyses with the other three pathogens. In order to call a MSSD either positive or negative for a specific pathogen, a HQM limit of detection needed to be determined (Table 3). The detection limit (lowest HQM number to call a MSSDs positive for the presence of a pathogen) was obtained by pairwise alignment of all the pathogen e-probes against all the MSSDs. Any sample containing HQM equal or lower than HQM false positive limit were considered negative (Table 3). The HQM False positive limit has been calculated based on 2,000 MSSDs subjected to EDNA analysis. Particularly for this case, negative controls were used. The negative controls were MSSDs that contained no pathogen sequences on them and had the same properties as the ones having the pathogen (400 negative controls MSSDs in total). Although, the numbers of replicates were high, the HQM false positive limit could vary depending on the total number of MSSD utilised for the analyses, as well as for previous quality alignment consideration while eliminating non-pathogen specific e-probes. However, the validity that gives that high number of replicates might suggest considering this value a constant.

Although initially e-probes lengths of up to 200 nt were considered due to the large reads that 454 pyrosequencing provides, e-probe length range was decreased because sensitivity started to decrease at larger e-probe lengths (Figure 4). The use of large numbers of e-probes was good for sensitivity, due to the subsequently larger number of unique signatures available to detect each pathogen. However, such a large data set made the computing process time consuming. Overall, e-probes 60 nt long provided optimal sensitivity, specificity, and data processing time (Figure 4). E-probe database curation decreased the number of e-probes in low percentage. The final e-probes were considered unique and were expected to detect the pathogen in a metagenome 454 sequencing database.

Table 1 False positive High quality matches in four eukaryotic plant pathogens using EDNA: *Pha = P. pachyrhizi*; *Ram=P. ramorum*

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

Figure 4 Relationship among e-probe length and specificity while using EDNA as a diagnostic tool in four eukaryotic plant pathogens with four different pathogen abundances (see online version for colours)



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

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

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

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

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

4 Conclusions

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

MSSDs that contained 10,000 total reads were used in this study although 454 pyrosequencing (Roche GS Junior) is capable of sequencing approximately 150,000 reads in one single run. The objective of using lower number of total reads was to demonstrate that the pathogens could be detected if approximately 15 barcoded samples were analysed in a single 454 pyrosequencing run. Eventually, NGS will become cheaper and there will be no need of barcoding samples.

Although, EDNA could be compared with bioinformatics tools that were developed principally to identify organisms in NGS output databases like Metaphlan and MEGAN (Huson et al., 2011; Segata et al., 2012), our tool offers the assurance of the pathogen presence in the database, while other tools only provide the number of reads belonging to the target organisms (Huson et al., 2011). EDNA uses specific signatures of the pathogen and can realistically decide whether the pathogen is present or not in the original sample. To our knowledge, there are no studies where the detection of specific fungal or oomycete plant pathogens was validated using NGS output databases, this fact makes EDNA the pioneer in the utilisation of NGS data for detection of eukaryotic plant pathogens.

References

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', *Molecular Plant Pathology*, Vol. 10, No. 4, pp.537–545.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) 'BLAST+: architecture and applications', *BMC bioinformatics*, Vol. 10, No. 1, p.421.

Castello, J.D., Leopold, D.J. and Smallidge, P.J. (1995) 'Pathogens, patterns, and processes in forest ecosystems', *Bioscience*, Vol. 45, No. 1, pp.16–24.

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., van Engelsdorp, D., Kalkstein, A.L., Drysdale, A., Hui, J., Zhai, J., Cui, L., Hutchison, S.K., Simons, J.F., Egholm, M., Pettis, J.S. and Lipkin, W.I. (2007) 'A metagenomic survey of microbes in honey bee colony collapse disorder', *Science*, Vol. 318, No. 5848, pp.283–287.

Enserink, M. (1999) 'Predicting invasions: biological invaders sweep in', *Science*, Vol. 285, No. 5435, pp.1834–1836.

Everett, R.A. (2000) 'Patterns and pathways of biological invasions', *Trends in Ecology & Evolution*, Vol. 15, No. 5, pp.177–178.

Hane, J.K., Rouxel, T., Howlett, B.J., Kema, G.H.J., Goodwin, S.B. and Oliver, R.P. (2011) 'A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi', *Genome Biology*, Vol. 12, No. 5, p.R45.

Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N. and Schuster, S.C. (2011) 'Integrative analysis of environmental sequences using MEGAN4', *Genome research*, Vol. 21, No. 9, pp.1552–1560.

Jones, W. (2010) 'High-throughput sequencing and metagenomics', *Estuaries and Coasts*, Vol. 33, No. 4, pp.944–952.

Levesque, C.A., Brouwer, H., Cano, L., Hamilton, J.P., Holt, C., Huitema, E., Raffaele, S., Robideau, G.P., Thines, M., Win, J., Zerillo, M.M., Beakes, G.W., Boore, J.L., Busam, D., Dumas, B., Ferriera, S., Fuerstenberg, S.I., Gachon, C.M.M., Gaulin, E., Govers, F., Grenville-Briggs, L., Horner, N., Hostetler, J., Jiang, R.H.Y., Johnson, J., Krajaejun, T., Lin, H., Meijer, H.J.G., Moore, B., Morris, P., Phuntmart, V., Puiu, D., Shetty, J., Stajich, J.E., Tripathy, S., Wawra, S., van West, P., Whitty, B.R., Coutinho, P.M., Henrissat, B., Martin, F., Thomas, P.D., Tyler, B.M., De Vries, R.P., Kamoun, S., Yandell, M., Tisserat, N. and Buell, C.R. (2010) 'Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire', *Genome biology*, Vol. 11, No. 7.

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F. and Brandi, M.L. (2010) 'Bioinformatics for next generation sequencing data', *Genes*, Vol. 1, No. 2, pp.294–307.

Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M., Chen, Y., Chen, Z., Dewell, S., Du, L., Fierro, J., Gomes, X., Godwin, B., He, W., Helgesen, S., Ho, C., Irzyk, G., Jando, S., Alenquer, M., Jarvie, T., Jirage, K., Kim, J., Knight, J., Lanza, J., Leamon, J., Lefkowitz, S., Lei, M., Li, J., Lohman, K., Lu, H., Makhijani, V., McDade, K., McKenna, M., Myers, E., Nickerson, E., Nobile, J., Plant, R., Puc, B., Ronan, M., Roth, G., Sarkis, G., Simons, J., Simpson, J., Srinivasan, M., Tartaro, K., Tomasz, A., Vogt, K., Volkmer, G., Wang, S., Wang, Y., Weiner, M., Yu, P., Begley, R. and Rothberg, J. (2005) 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature*, Vol. 437, No. 7057, pp.376–380.

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. and Lipkin, W.I. (2008) 'A new arenavirus in a cluster of fatal transplant-associated diseases', *New England Journal of Medicine*, Vol. 358, No. 10, pp.991–998.

Puccinia Sequencing Project, Harvard, B.I.o. and MIT 'Puccinia graminis whole genome' [online] http://www.broadinstitute.org/annotation/genome/puccinia_group/Credits.html.

Richter, D.C., Ott, F., Auch, A.F., Schmid, R. and Huson, D.H. (2008) 'MetaSim—A Sequencing Simulator for Genomics and Metagenomics', *PloS one*, Vol. 3, No. 10, p.e3373.

Ronaghi, M. (2001) 'Pyrosequencing Sheds Light on DNA Sequencing', *Genome research*, Vol. 11, No. 1, pp.3–11.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012) 'Metagenomic microbial community profiling using unique clade-specific marker genes', *Nat Meth*, Vol. 9, No. 8, pp.811–814.

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', *Journal of Microbiological Methods*, Vol. 94, No. 3, pp.356–366.

Strange, R.N. and Scott, P.R. (2005) 'Plant Disease: A Threat to Global Food Security', *Annual Review of Phytopathology*, Vol. 43, No. 1, pp.83–116.

Tembe, W., Zavaljevski, N., Bode, E., Chase, C., Geyer, J., Wasieloski, L., Benson, G. and Reifman, J. (2007) 'Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays', *Bioinformatics*, Vol. 23, No. 1, pp.5–13.

Tyler, B.M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R.H.Y., Aerts, A., Arredondo, F.D., Baxter, L., Bensasson, D. and Beynon, J.L. (2006) 'Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis', *Science*, Vol. 313, No. 5791, pp.1261–1266.

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. and Banfield, J.F. (2004) 'Community structure and metabolism through reconstruction of microbial genomes from the environment', *Nature*, Vol. 428, No. 6978, pp.37–43.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., FitzGerald, L.M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J.T., Perazzolli, M., Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Demattè, L., Mraz, A., Battilana, J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J.A., Sterck, L., Vandepoele, K., Grando, S.M., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S.K., Fontana, P., Gutin, A., Van de Peer, Y., Salamini, F. and Viola, R. (2007) 'A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety', *PloS one*, Vol. 2, No. 12, p.e1326.

Vijaya Satya, R., Zavaljevski, N., Kumar, K. and Reifman, J. (2008) 'A high-throughput pipeline for designing microarray-based pathogen diagnostic assays', *BMC Bioinformatics*, Vol. 9, No. 1, p.185.