Determining Limit of Detection of High Throughput Sequencing Diagnostics by Including Internal Controls in E-probes

Tyler Dang[1], Huizi Wang[2], Andres Espindola[3], Josh Habiger[2], Sohrab Bodaghi[1], Georgios Vidalakis[1], and Kitty Cardwell[3]

[1]Department of Microbiology and Plant Pathology, University of California, Riverside, CA
[2]Department of Statistics, Oklahoma State University, Stillwater, OK
[3]Institute of Biosecurity and Microbial Forensics, Oklahoma State University, Stillwater, OK

Introduction

The cost for high-throughput sequencing (HTS) has decreased significantly and has made it possible for the application of this technology for routine plant diagnostics. However, there are constraints with the use of HTS as a diagnostic tool which include the need for dedicated personnel with a bioinformatics background for data analysis and the lack of a standardized analysis pipeline that makes evaluating and validating results generated at different HTS laboratories difficult. E-probe Diagnostic Nucleic Acid Analysis (EDNA) is an in-silico bioinformatic tool that utilizes short curated electronic probes (e-probes) designed from pathogen specific sequences which allow users to detect and identify single or multiple pathogens of interest in raw HTS datasets. This platform streamlines the bioinformatic data analysis into a GUI interface as a plant diagnostic tool used by diagnosticians. In this study, we describe the process for the development, validation, and use of e-probes for detection and identification of a wide range of taxonomically unique citrus pathogens that include *Citrus exocortis viroid*, *Citrus tristeza virus*, and '*Candidatus* Liberibacter asiaticus'. We demonstrate the process for evaluating the analytical and diagnostic sensitivity and specificity metrics of the in-silico EDNA assays. In addition, we show the importance of including background noise (internal controls) to generate variance in non-infected samples for a valid statistical test using quadratic discriminant analysis. The fully validated EDNA assays, from this study, can be readily integrated into existing citrus testing programs that utilize HTS.

Aim

- Generate HTS data from known pathogen-infected and non-infected citrus samples from greenhouse and field sources.
- Develop and validate a probability algorithm to generate a Limit of detection (LOD) for each e-probe.
- Determine if the newly developed non-host internal control (IC) e-probes improved the variance in the non-infected samples.

Method

- HTS was performed on infected and non-infected citrus stem tissues using the Illumina© platforms.
- Internal control (IC) e-probes were designed from housekeeping genes of *Pistacia*, *Malus*, and *Prunus*, then combined with e-probes specific to the pathogen of interest.
- HTS data were analyzed using the MiFi® platform and the hits, scores, and diagnostic results were obtained.
- Scores and hits were generated for each e-probe sequence, which were used to calculate the total score for each pathogen.
- IC e-probes were curated by ranking and retaining the e-probes with the lowest scores.
- The probability of a positive or negative hit was calculated using the scores obtained from infected and non-infected samples.
- The calculations were repeated for individual pathogens.
- **Definitions:**
  - ✓ **Hits:** reads that bind with e-probe sequence(s) based on a minimum e-value.
  - ✓ **Scores:** calculated for each hit based on percent identity and query coverage.

Results

$$LOD = x = \frac{\left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right) - \sqrt{\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 \sigma_2^2} - \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) \times 2log\frac{\sigma_2}{\sigma_1}}}{\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)}$$

Equation illustrating the Quadratic Discriminant Analysis used to estimate the limit of detection (LOD). The LOD is the value of the total score.

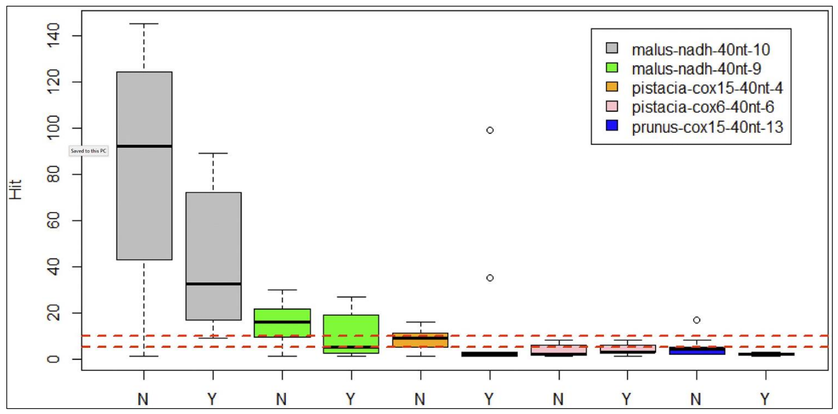

Fig. 1. Boxplot of the five internal control (IC) e-probes shows the distribution of hits for the internal controls across the non-infected (N) and infected samples (Y). The IC e-probes with the lowest hits were the best candidates.

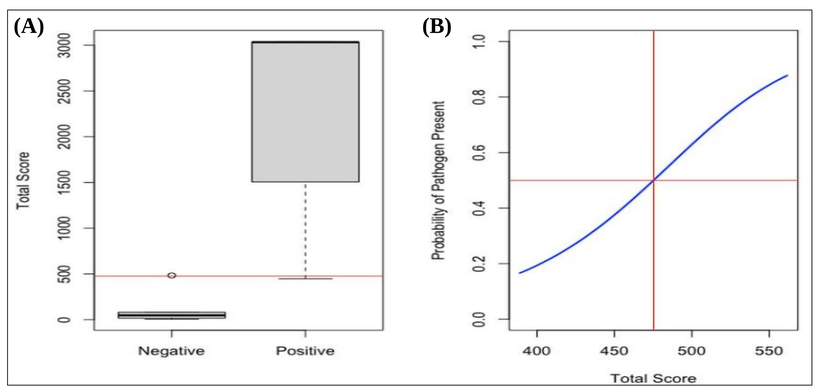

Fig. 2. (A) Boxplot of Citrus exocortis viroid (CEVd) e-probes at 20 nucleotides (nt): Distribution of the total scores between healthy and CEVd positive controls using 20 nt CEVd and internal control (IC) e-probes. LOD score indicated in red. **(B) Probably curve of CEVd e-probes at 20 nt:** The LOD for CEVd 20 nt e-probes were calculated. A score of 475.5 would indicate 50/50 chance of "positive" diagnostic result.

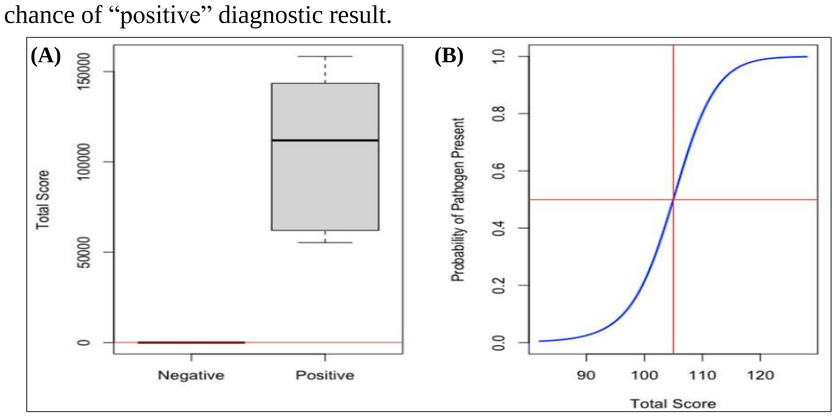

Fig. 3. (A) Boxplot of Citrus tristeza virus (CTV) e-probes: Distribution of the total scores between non-infected and CTV-positive controls using 20 nt CTV and internal control (IC) e-probes. LOD score indicated in red. **(B) Probably curve of CTV e-probes with 20 nt e-probes:** The LOD for CTV 20 nt e-probes were calculated. A score of 116.7 would indicate 50/50 chance of "positive" diagnostic result.

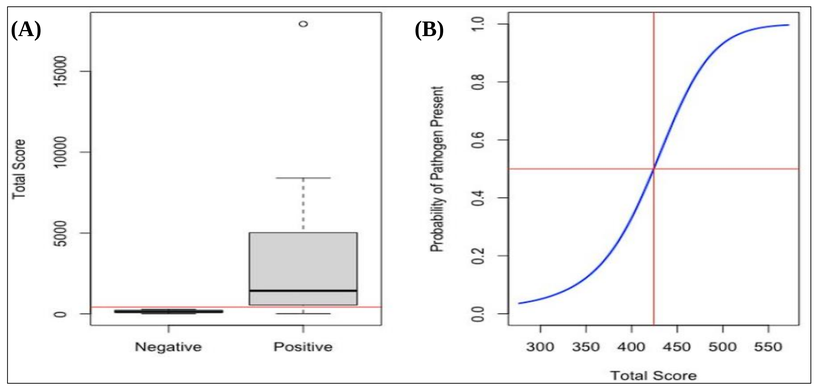

Fig. 4. (A) Probably curve of '*Candidatus* Liberibacter asiaticus' (CLas) e-probes: Distribution of the total scores between non-infected and CLas-positive controls using 40 nt CLas and internal control (IC) e-probes. LOD score indicated in red. **(B) Probably curve of CLas e-probes with 40 nt e-probes:** The LOD for CLas 40 nt e-probes were calculated. A score of 424.2 would indicate 50/50 chance of "positive" diagnostic result.

Conclusions and Future Direction

- The implementation of additional non-host IC e-probes to all e-probe sets for pathogen detection is required in order to obtain the proper variance for LOD calculations with non-infected samples.
- As a proof of concept, the LODs were calculated for citrus-specific pathogens such as a viroid, virus, and prokaryote. These values indicate when the chance of a "positive" diagnostic result is at 50%.
- E-probes to detect additional citrus pathogens are currently under development.
- The EDNA online platform can be accessed at: https://bioinfo.okstate.edu/login/index.php and can utilize datasets from various sequencing platforms such as Oxford Nanopore, Illumina, etc.

References

- Dang et al. 2022. Developmental and statistical validation e-probe diagnostic nucleic acid analysis assay for the detection of citrus pathogens from raw high throughput sequencing data. PhytoFrontiers.
- Espindola, A. S., and Cardwell, K. F. 2021. Microbe Finder (MiFi®): Implementation of an interactive pathogen detection tool in metagenomic sequence data. Plants.10(2): 250.
- Dang, T., Espindola, A., Vidalakis, G., and Cardwell, K. 2022. An *in-silico* detection of a Citrus Viroid from raw high-throughput sequencing data. In *Viroids: Methods and Protocols*, eds. Ayala L. N. Rao, Irene Lavagi-Craddock, and Georgios Vidalakis. New York, NY: Springer US, p. 275–283.

Funding Sources

- Citrus Research Board: Grant# **5300-205**.
- APHIS Plant Protection and Quarantine

UNIVERSITY OF CALIFORNIA
Agriculture and Natural Resources
ucanr.edu