DETECTION OF THE SELECT AGENT *CONIOTHYRIUM GLYCINES*, CAUSAL PATHOGEN OF RED LEAF BLOTCH OF SOYBEANS USING HIGH-THROUGHPUT SEQUENCING DATA

By

DANIEL ALEXIS CARRERA LOPEZ

Bachelor of Science in Biotechnology Engineering

Universidad de las Fuerzas Armadas ESPE

Sangolqui, Ecuador

2019

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2023

DETECTION OF THE SELECT AGENT

CONIOTHYRIUM GLYCINES, CAUSAL PATHOGEN OF

RED LEAF BLOTCH OF SOYBEANS USING HIGH-

THROUGHPUT SEQUENCING DATA

Thesis Approved:

Kitty F. Cardwell, Ph.D.

Thesis Adviser

Andres Espindola, Ph.D.

Stephen Marek, Ph.D.

Francisco Ochoa-Corona, Ph.D.

ACKNOWLEDGEMENTS

I want to thank my family, especially my parents and siblings, for their constant support, understanding, love, and bits of advice. Having me far from home was tough for them, but they always showed me their best smile. I am genuinely grateful to Nadia Suquillo for always being there for me and for all her love and patience; she never gave up and was a motivation for me. This achievement would not have been possible without them.

I want to thank my advisor, Dr. Kitty Cardwell, for her constant support, guidance, and encouragement during my MS program. I would also like to thank my committee members, Dr. Andres Espindola, Dr. Stephen Marek, and Dr. Francisco Ochoa-Corona, for their help and advice. I am very grateful to Dr. Fernanda Proano-Cuenca for always being there, for her friendship and mentorship, and for guiding and supervising this research.

To all my friends, especially Nicolas, Ishtar, Camila, and Gabriela, thank you for making this experience enjoyable and filled with good moments I will remember forever. I would also like to acknowledge the staff and faculty of the Department of Entomology and Plant Pathology and the Institute of Biosecurity and Microbial Forensics at Oklahoma State University.

Lastly, I would like to thank my grandfather, Juanito Elias. I am sure you were with me all this time, taking care of me and watching me from the sky.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: DANIEL ALEXIS CARRERA LOPEZ

Date of Degree: MAY 2023

Title of Study: DETECTION OF THE SELECT AGENT CONIOTHYRIUM GLYCINES, CAUSAL PATHOGEN OF RED LEAF BLOTCH OF SOYBEANS USING HIGH-THROUGHPUT SEQUENCING DATA.

Major Field: ENTOMOLOGY AND PLANT PATHOLOGY

Abstract: The select agent *Coniothyrium glycines*, the causal pathogen of the disease red leaf blotch of soybeans, has not been identified in the U.S. Although *C. glycines* is listed as a select agent by the Federal Government, little information about its biology, evolution, and genomics is available, which poses a challenge to developing diagnostic tools. This research aimed to expand the general molecular and genomic knowledge of *C. glycines* and apply the generated information for detecting the pathogen using high-throughput sequencing (HTS) data. During this research, fourteen *C. glycines* isolates obtained from Zambia and Zimbabwe were used. A multilocus phylogenetic analysis was performed using two non-coding and two coding genes, revealing that isolates from matching locations form monophyletic clades. The results also suggested the movement of the fungus across borders since some isolates from different countries had the same common ancestor. Based on the topology of the phylogenetic tree, five representative isolates were selected, and their whole genome was assembled using Oxford Nanopore Technologies and Illumina sequencing data. Finally, the generated assemblies and the MiFi® web application were used to develop and validate three e-probe sets for the detection and differentiation of *C. glycines* isolates. The limit of detection (LOD), sensitivity, and specificity was estimated using *in silico, in vitro,* and *in vivo* approaches. LOD was influenced by the number of e-probes, sequencing read length, and sequencing platform. Once the LOD was exceeded, the sensitivity and specificity were 100%, allowing a reliable detection and discrimination of *C. glycines* isolates. The obtained results contribute to the molecular and genomic knowledge of *C. glycines,* facilitating future research on this organism. Additionally, these findings provide valuable guidelines for using HTS-based e-probe detection and discrimination of *C. glycines* to improve current biosecurity measures.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

Introduction

Agricultural land covers about 37% of the earth’s land surface, and the sector has remained strong despite the current decline in the world economy. In 2021, more than $150 billion in agricultural commerce was conducted, which has increased the cost of several commodities globally (Gullino et al. 2017; Meyer 2021). Protecting agriculture from biotic factors like plant pathogens and pests is challenging due to the large amount of land devoted to agriculture, its economic importance, and the need to maintain food supply stability (Karunarathna et al. 2021; Waage and Mumford 2008).

Emerging plant pathogens, such as bacteria, viruses, fungi, oomycetes, and nematodes, constantly threaten agriculture in many countries (Fones et al. 2020; Gullino et al. 2017). They cause yield losses, affect consumer confidence, reduce the nutritional value, and impede international commerce (Fletcher et al. 2020). The incidence of these pathogens has increased in recent years, mainly due to globalization, climate change, and ecological modifications. Detection methods have been inadequate, which may have led to underestimating the problem (Avila-Quezada et al. 2018; Fisher et al. 2012; Karunarathna et al. 2021). Disease detection and prevention are essential to minimize their negative impact on productivity and ensure agricultural sustainability (Ristaino et al. 2021).

Soybean (*Glycine max* L.) is a legume considered a critical commodity globally, with more than 6% of total agricultural land used for production. It produces 71% of the entire plant-based protein meal and 29% of the oil (Bateman et al. 2020; Lin et al. 2022). In 2022, soybean production in the U.S. recorded a total of 4.3 billion bushels, with an average yield of 50 bushels per acre, representing an annual worth of over $60 billion. Soybean is the second most valuable crop in the U.S. after corn (Hartman et al. 2011; United States Department of Agriculture 2023). Plant diseases and pests can affect the yield of soybeans, with diseases alone causing up to 20% yield losses. More than 200 pathogens are known to affect soybeans, and at least 35 can significantly impact the economy (Faske et al. 2014; Lin et al. 2022).

Red leaf blotch (RLB) of soybeans, caused by the fungal pathogen *Coniothyrium glycines* (R.B. Stewart) Verkley & Gruyter, is a disease that can affect soybeans from seedling to maturity stage (de Gruyter et al. 2013; Hartman et al. 1987; Stewart 1957). The disease shows irregular, circular, small, and dark leaf spots, which enlarge and become necrotic, surrounded by a chlorotic halo. RLB can cause lesions in various soybean tissues, including foliage, petioles, pods, and stems. The most severe outcome is leaf abscission, which can generate up to 75% defoliation of some soybean varieties, reducing their photosynthetic abilities and yield (Datnoff 1987; Hartman and Murithi 2022; Levy et al. 1990; Murithi et al. 2022).

Coniothyrium glycines is a fungal plant pathogen undergoing several taxonomic revisions (Hartman et al. 2011). It was first reported in Ethiopia in 1953 (Stewart 1957) and has since spread to sub-Saharan African countries (EPPO 2023). It is not present in the U.S. and is listed as a select agent by the Federal Government due to its potential economic and food safety impact. Currently, no diagnostic tool is available, highlighting the need to develop molecular diagnostic tools to ensure biosecurity (Hartman et al. 2011; Murithi et al. 2022; Proano-Cuenca et al. 2023).

Traditional techniques for identifying plant diseases involve observing characteristic structures or culturing tentative causal organisms, which require expertise and are time-consuming (Gullino et al. 2017; Sankaran et al. 2010). These methods have limitations when different organisms share the same characteristics or cannot be cultured (Buja et al. 2021). Molecular diagnosis using nucleic acids, such as PCR (Polymerase Chain Reaction), LAMP (Loop-mediated Isothermal Amplification), RPA (Recombinase Polymerase Amplification), RT-PCR (Retro-transcribed PCR), and qPCR (Quantitative PCR), is currently the best choice due to their speed, reliability, scaling capabilities, and high specificity. However, they require bulky instruments, experienced personnel, and prior knowledge of the organism for primer or probe development (Fang and Ramasamy 2015).

DNA-based approaches like barcoding and phylogenetic analysis are also commonly used to differentiate similar or identical species (Raja et al. 2017). DNA barcoding compares an unknown sequence with a reference sequence database, while phylogenetic analysis clusters the unknown samples within an evolutionary context with other homologous sequences (Hibbett and Taylor 2013; Kapli, Yang, and Telford 2020; Naranjo-Ortiz and Gabaldón 2019). These approaches aid in understanding species' evolution and features but are limited by the quality of DNA sequences and prior knowledge of the organisms (Raja et al. 2017).

The analysis of nucleic acid sequences has improved the identification of plant pathogens, especially with next-generation sequencing (NGS). NGS is becoming increasingly advanced and cost-effective, generating large amounts of data that can be analyzed with bioinformatic pipelines. This makes it suitable for detecting new causal agents without prior knowledge of their identity (Ansorge 2009; Engelthaler and Litvintseva 2020; Gullino et al. 2017). NGS is valuable for pathogen discovery, gene discovery, *de novo* genome assemblies, and tracking plant pathogen movement (Ristaino et al. 2021). However, the limitations of NGS are related to the amount and purity of genetic material used, high computational resources, and dependency on nucleic acid and genomic databases (Loman et al. 2012). Detecting a pathogen through sequencing requires confirmation with cultures or Koch’s postulates. Furthermore, new legal frameworks are needed to approve the use of NGS in forensic investigations (Gilchrist et al. 2015; Ristaino et al. 2021).

NGS has revolutionized the detection of plant pathogens, but bioinformatic processing is a bottleneck for data analysis (Hu et al. 2021). E-probe Diagnostic Nucleic acid Analysis (EDNA) is a bioinformatic pipeline based on short DNA sequences known as e-probes, that eliminates the need for bioinformatic expertise and computational resources to detect one or multiple targets within raw sequencing data (Espindola et al. 2018; Stobbe et al. 2013). EDNA has been used and validated in several studies, mainly for detecting plant pathogens within metagenomic raw sequencing data (Dang et al. 2022; Espindola et al. 2015, 2022; Pena-Zuniga 2020; Proano-Cuenca, Espíndola, and Garzon 2022; Stobbe et al. 2014).

MiFi® is a web application that hosts the EDNA bioinformatic pipeline and offers two main components: MiProbe® and MiDetect®. Users can use MiProbe® to design specific e-probes by uploading target genomic information and selecting desired e-probe lengths. MiDetect® allows testing the designed e-probes against any sequencing data using BLAST and providing a positive or negative outcome based on a statistical test. MiFi® provides a database for storing unique e-probes for the target (Espindola and Cardwell 2021).

This research aimed to detect the select agent *Coniothyrium glycines*, the causal pathogen of red leaf blotch of soybeans using high-throughput sequencing data. To fulfill this objective, molecular and evolutionary knowledge was addressed. First, a multilocus phylogenetic analysis was performed on fourteen *C. glycines* isolates. Later, five representative isolates were selected for genome sequencing using Oxford Nanopore Technologies (ONT) and Illumina sequencing. Finally, using MiFi®, three e-probe sets were developed and validated to detect and differentiate *C. glycines* within metagenomic data. Three validation approaches were used (*in silico, in vitro,* and *in vivo)* to estimate the limit of detection (LOD), sensitivity, and specificity of each e-probe set.

This project provides essential molecular and genomic information on *C. glycines* that can help understand its biology, ecology, and evolution. This knowledge can be utilized to develop disease management strategies to reduce crop losses and improve food security. Furthermore, using e-probes and HTS data can aid in identifying and differentiating *C. glycines*, which can enhance current biosecurity measures.

References

Ansorge, Wilhelm J. 2009. “Next-Generation DNA Sequencing Techniques.” *New Biotechnology* 25(4):195–203. doi: 10.1016/j.nbt.2008.12.009.

Avila-Quezada, Graciela Dolores, Jesus Fidencio Esquivel, Hilda Victoria Silva-Rojas, Santos Gerardo Leyva-Mir, Clemente de Jesús Garcia-Avila, Andrés Quezada-Salinas, Lorena Noriega-Orozco, Patricia Rivas-Valencia, Damaris Ojeda-Barrios, and Alicia Melgoza-Castillo. 2018. “Emerging Plant Diseases under a Changing Climate Scenario: Threats to Our Global Food Supply.” *Emirates Journal of Food and Agriculture* 30(6):443–50. doi: 10.9755/ejfa.2018.v30.i6.1715.

Bateman, Nick R., Angus L. Catchot, Jeff Gore, Don R. Cook, Fred R. Musser, and J. Trent Irby. 2020. “Effects of Planting Date for Soybean Growth, Development, and Yield in the Southern USA.” *Agronomy* 10(4):596. doi: 10.3390/agronomy10040596.

Buja, Ilaria, Erika Sabella, Anna Grazia Monteduro, Maria Serena Chiriacò, Luigi De Bellis, Andrea Luvisi, and Giuseppe Maruccio. 2021. “Advances in Plant Disease Detection and Monitoring: From Traditional Assays to In-Field Diagnostics.” *Sensors* 21(6):2129. doi: 10.3390/s21062129.

Dang, Tyler, Huizi Wang, Andrés S. Espíndola, Joshua Habiger, Georgios Vidalakis, and Kitty Cardwell. 2022. “Development and Statistical Validation of E-Probe Diagnostic Nucleic Acid Analysis (EDNA) Detection Assays for the Detection of Citrus Pathogens from Raw High Throughput Sequencing Data.” *PhytoFrontiers^{TM}* 1–51. doi: 10.1094/PHYTOFR-05-22-0047-FI.

Datnoff, L. 1987. "Effect of Red Leaf Blotch on Soybean Yields in Zambia." *The American Phytopathological Society* 71(2):1–4.

Engelthaler, David M., and Anastasia P. Litvintseva. 2020. "Genomic Epidemiology and Forensics of Fungal Pathogens." Pp. 141–54 in *Microbial Forensics*. Elsevier.

EPPO. 2023. "EPPO Global Database." *Global Database*. Retrieved February 19, 2023 (https://gd.eppo.int/).

Espindola, Andres, and Kitty Cardwell. 2021. "Microbe Finder (MiFi®): Implementation of an Interactive Pathogen Detection Tool in Metagenomic Sequence Data." *Plants* 10(2):250. doi: 10.3390/plants10020250.

Espindola, Andres, Kitty Cardwell, Frank N. Martin, Peter R. Hoyt, Stephen M. Marek, William Schneider, and Carla D. Garzon. 2022. "A Step Towards Validation of High-Throughput Sequencing for the Identification of Plant Pathogenic Oomycetes." *Phytopathology®* 112(9):1859–66. doi: 10.1094/PHYTO-11-21-0454-R.

Espindola, Andres, William Schneider, Kitty Cardwell, Yisel Carrillo, Peter Hoyt, Stephen Marek, Hassan Melouk, and Carla Garzon. 2018. "Inferring the Presence of Aflatoxin-Producing Aspergillus Flavus Strains Using RNA Sequencing and Electronic Probes as a Transcriptomic Screening Tool" edited by R. A. Wilson. *PLOS ONE* 13(10):e0198575. doi: 10.1371/journal.pone.0198575.

Espindola, Andres, William Schneider, Peter R. Hoyt, Stephen M. Marek, and Carla Garzon. 2015. "A New Approach for Detecting Fungal and Oomycete Plant Pathogens in next Generation Sequencing Metagenome Data Utilising Electronic Probes." *International Journal Data Mining and Bioinformatics* 12(2):1–14.

Fang, Yi, and Ramaraja Ramasamy. 2015. "Current and Prospective Methods for Plant Disease Detection." *Biosensors* 5(3):537–61. doi: 10.3390/bios5030537.

Faske, Travis, Terry Kirkpatrick, Jing Zhou, and Ioannis Tzanetakis. 2014. “Chapter 11: Soybean Diseases.” Pp. 1–18 in *Arkansas Soybean Production Handbook*. Vol. 4. University of Arkansas.

Fisher, Matthew C., Daniel. A. Henk, Cheryl J. Briggs, John S. Brownstein, Lawrence C. Madoff, Sarah L. McCraw, and Sarah J. Gurr. 2012. “Emerging Fungal Threats to Animal, Plant and Ecosystem Health.” *Nature* 484(7393):186–94. doi: 10.1038/nature10947.

Fletcher, Jacqueline, Neel G. Barnaby, James Burans, Ulrich Melcher, Douglas G. Luster, Forrest W. Nutter, Harald Scherm, David G. Schmale, Carla S. Thomas, and Francisco M. Ochoa Corona. 2020. “Forensic Plant Pathology.” Pp. 49–70 in *Microbial Forensics*. Elsevier.

Fones, Helen N., Daniel P. Bebber, Thomas M. Chaloner, William T. Kay, Gero Steinberg, and Sarah J. Gurr. 2020. “Threats to Global Food Security from Emerging Fungal and Oomycete Crop Pathogens.” *Nature Food* 1(6):332–42. doi: 10.1038/s43016-020-0075-0.

Gilchrist, Carol A., Stephen D. Turner, Margaret F. Riley, William A. Petri, and Erik L. Hewlett. 2015. “Whole-Genome Sequencing in Outbreak Analysis.” *Clinical Microbiology Reviews* 28(3):541–63. doi: 10.1128/CMR.00075-13.

de Gruyter, J., J. H. C. Woudenberg, M. M. Aveskamp, G. J. M. Verkley, J. Z. Groenewald, and P. W. Crous. 2013. “Redisposition of Phoma-like Anamorphs in Pleosporales.” *Studies in Mycology* 75:1–36. doi: 10.3114/sim0004.

Gullino, Maria Lodovica, James P. Stack, Jacqueline Fletcher, and John D. Mumford. 2017. *Practical Tools for Plant and Food Biosecurity*. Vol. 8. 1st ed. edited by M. L. Gullino, J. P. Stack, J. Fletcher, and J. D. Mumford. Cham: Springer International Publishing.

Hartman, G., L. Datnoff, C. Levy, J. Sinclair, D. Cole, and F. Javaheri. 1987. “Red Leaf Blotch of Soybeans.” *Plant Disease* 113–18.

Hartman, G., and H. M. Murithi. 2022. “Coniothyrium Glycines (Red Leaf Blotch).” *CABI Compendium* CABI Compendium. doi: 10.1079/CABICOMPENDIUM.17687.

Hartman, Glen, James Haudenshield, Kent Smith, and Paul Tooley. 2011. *Recovery Plan for Red Leaf Blotch of Soybean Caused by Phoma Glycinicola*.

Hibbett, David S., and John W. Taylor. 2013. “Fungal Systematics: Is a New Age of Enlightenment at Hand?” *Nature Reviews Microbiology* 11(2):129–33. doi: 10.1038/nrmicro2963.

Hu, Taishan, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. 2021. “Next-Generation Sequencing Technologies: An Overview.” *Human Immunology* 82(11):801–11. doi: 10.1016/j.humimm.2021.02.012.

Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. 2020. “Phylogenetic Tree Building in the Genomic Age.” *Nature Reviews Genetics* 21(7):428–44.

Karunarathna, Samantha C., Sajeewa S. N. Maharachchikumbura, Hiran A. Ariyawansa, Belle Damodara Shenoy, and Rajesh Jeewon. 2021. “Editorial: Emerging Fungal Plant Pathogens.” *Frontiers in Cellular and Infection Microbiology* 11.

Levy, C., G. S. Mahuku ?, J. R. Tattersfield $ And, and Desiri~e L. Cole. 1990. *Method of Assessment of Red Leaf Blotch on Soybeans Used to Evaluate Cultivar Susceptibility and Chemical Control*.

Lin, Feng, Sushil Satish Chhapekar, Caio Canella Vieira, Marcos Paulo Da Silva, Alejandro Rojas, Dongho Lee, Nianxi Liu, Esteban Mariano Pardo, Yi Chen Lee, Zhimin Dong, Jose Baldin Pinheiro, Leonardo Daniel Ploper, John Rupe, Pengyin Chen, Dechun Wang, and Henry T. Nguyen. 2022. “Breeding for Disease Resistance in Soybean: A Global Perspective.” *Theoretical and Applied Genetics* 135(11):3773–3872.

Loman, Nicholas J., Raju V Misra, Timothy J. Dallman, Chrystala Constantinidou, Saheer E. Gharbia, John Wain, and Mark J. Pallen. 2012. “Performance Comparison of Benchtop

High-Throughput Sequencing Platforms." *Nature Biotechnology* 30(5):434–39. doi: 10.1038/nbt.2198.

Meyer, Seth. 2021. *The Outlook for U.S. Agriculture-2021. Building on Innovation: A Pathway to Resilience*.

Murithi, Harun M., Michelle Pawlowski, Tizazu Degu, Deresse Hunde, Molla Malede, Tonny Obua, Hapson Mushoriwa, Danny Coyne, Phinehas Tukamuhabwa, and Glen L. Hartman. 2022. "Evaluation of Soybean Entries in the Pan-African Trials for Response to Coniothyrium Glycines, the Cause of Red Leaf Blotch." *Plant Disease* 106(2):535–40. doi: 10.1094/PDIS-05-21-1017-RE.

Naranjo-Ortiz, Miguel A., and Toni Gabaldón. 2019. "Fungal Evolution: Diversity, Taxonomy and Phylogeny of the Fungi." *Biological Reviews* 94(6):2101–37. doi: 10.1111/brv.12550.

Pena-Zuniga, Lizbeth Daniela. 2020. "EDNA-HOST: Detection of Global Plant Viromes Using High Throughput Sequencing." PhD, Oklahoma State University, Stillwater.

Proano-Cuenca, Fernanda, Daniel Carrera-Lopez, Douglas Luster, Kurt Zeller, and Kitty Cardwell. 2023. "Genome Sequence Resources for Five Isolates of Coniothyrium Glycines, Causal Pathogen of Red Leaf Blotch of Soybeans." *PhytoFrontiers*^{TM} 1–15. doi: 10.1094/PHYTOFR-10-22-0113-A.

Proano-Cuenca, Fernanda, Andrés S. Espíndola, and Carla Garzon. 2022. "Detection of Phytophthora, Pythium, Globisporangium, Hyaloperonospora and Plasmopara Species in High-Throughput Sequencing Data by in Silico and in Vitro Analysis Using Microbe Finder (MiFi®)." *PhytoFrontiers*^{TM} 1–73. doi: 10.1094/PHYTOFR-04-22-0039-FI.

Raja, Huzefa A., Andrew N. Miller, Cedric J. Pearce, and Nicholas H. Oberlies. 2017. "Fungal Identification Using Molecular Tools: A Primer for the Natural Products Research Community." *Journal of Natural Products* 80(3):756–70. doi: 10.1021/acs.jnatprod.6b01085.

Ristaino, Jean B., Pamela K. Anderson, Daniel P. Bebber D, Kate A. Brauman E, Nik J. Cunniffe, Nina V Fedoroff, Cambria Finegold, Karen A. Garrett, Christopher A. Gilligan, Christopher M. Jones K, Michael D. Martin, Graham K. Macdonald, Patricia Neenan, Angela Records, David G. Schmale, Laura Tateosian, and Qingshan Wei. 2021. “The Persistent Threat of Emerging Plant Disease Pandemics to Global Food Security.” *PNAS* 118(23):1–9. doi: 10.1073/pnas.2022239118/-/DCSupplemental.

Sankaran, Sindhuja, Ashish Mishra, Reza Ehsani, and Cristina Davis. 2010. “A Review of Advanced Techniques for Detecting Plant Diseases.” *Computers and Electronics in Agriculture* 72(1):1–13. doi: 10.1016/j.compag.2010.02.007.

Stewart, Robert B. 1957. “An Undescribed Species of Pyrenochaeta on Soybean.” *Mycologia* 49(1):115–17. doi: 10.1080/00275514.1957.12024619.

Stobbe, A. H., W. L. Schneider, P. R. Hoyt, and U. Melcher. 2014. “Screening Metagenomic Data for Viruses Using the E-Probe Diagnostic Nucleic Acid Assay.” *Phytopathology* 104(10):1125–29. doi: 10.1094/PHYTO-11-13-0310-R.

Stobbe, Anthony H., Jon Daniels, Andres S. Espindola, Ruchi Verma, Ulrich Melcher, Francisco Ochoa-Corona, Carla Garzon, Jacqueline Fletcher, and William Schneider. 2013. “E-Probe Diagnostic Nucleic Acid Analysis (EDNA): A Theoretical Approach for Handling of next Generation Sequencing Data for Diagnostics.” *Journal of Microbiological Methods* 94(3):356–66. doi: 10.1016/j.mimet.2013.07.002.

United States Department of Agriculture. 2023. *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2022 Summary*.

Waage, J. K., and J. D. Mumford. 2008. “Agricultural Biosecurity.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1492):863–76. doi: 10.1098/rstb.2007.2188.

CHAPTER II

Literature review

1. Agricultural Biosecurity

1.1. Definition and importance

Globalization has strengthened the economy of countries based on the revenue of their international commerce. At the same time, it has increased the likelihood of spreading pathogens and pests worldwide and introducing them into agricultural production areas (Anderson et al. 2004; Karunarathna et al. 2021). Land dedicated to agriculture is estimated to cover around 37% of the earth's surface, and its industry has been resilient during the current global economic shrinkage. In 2021, agricultural trade accounted for more than $150 billion, driving the price of multiple commodities worldwide (Gullino et al. 2017; Meyer 2021). Due to the amount of land dedicated to agriculture, its economic importance, and food supply stability, one of the challenges is to protect it from biotic factors, such as plant pathogens and pests, which account for up to 30% and 20% of yield losses during post- and pre-harvest activities, respectively (Karunarathna et al. 2021; Waage and Mumford 2008).

Emerging plant pathogens, including bacteria, viruses, fungi, oomycetes, and nematodes, represent a constant threat to productivity in multiple countries where agriculture is essential not just for their economic welfare but to their national food safety stability (Fones et al. 2020; Gullino et al. 2017). The effect of plant diseases is not only reflected in yield losses; they also break consumer confidence, reduce nutritional food value, and impede international commerce (Fletcher et al. 2020). Unfortunately, the incidence of emerging plant pathogens has increased over recent years, mainly due to globalization, climate change, ecological modifications, vector spreading, mutations, and excessive use of different chemical controls. This phenomenon might have been overlooked or underestimated due to inadequate detection methods (Avila-Quezada et al. 2018; Fisher et al. 2012; Karunarathna et al. 2021).

The implementation of laws and regulations is needed to reduce the spread of plant diseases and to have in place control and management strategies (Fletcher et al. 2020). In 2007, FAO (Food and Agriculture Organization of the United Nations) compiled policies and regulations to protect agriculture, food, and the environment from biological risks, strengthening agricultural biosecurity. The term biosecurity has had multiple interpretations. In general, it represents an integrated approach to analyzing and managing crucial dangers to human health, animal welfare, plant health, and the well-being of the environment (Waage and Mumford 2008). Similarly, plant or agricultural biosecurity is defined as the combination of measures that aim to protect national boundaries from introducing external or invasive pests and diseases and to stop the internal spreading of those biological threats (Gullino et al. 2017).

The terrorist attacks on the U.S. in 2001, especially those involving letters contaminated with anthrax spores, increased the awareness of the use of pathogenic organisms as potential biological weapons (Waage and Mumford 2008). Agricultural systems represent a target for intentional biological attacks due to the economic and food safety distress that could be triggered. Introducing invasive species or foreign diseases could devastate the national economy and reduce food availability (Fletcher et al. 2020).

1.2. Biosecurity measures in the United States

1.2.1.Surveillance systems

Daily a huge number and volume of plants and their derivates considered a threat due to the possibility of harboring and spreading pests and plant pathogens, are moved through ports and borders (Alonso, Parnell, and Van den Bosch 2016). Hence, regulatory and control agencies perform inspections, surveillance, and monitoring of those goods and keep track of the diseases and pests previously reported within the country (Madden and Wheelis 2003). The aim of performing those activities is to be prepared and respond promptly to tentative introductions or outbreaks in the U.S. (Fletcher et al. 2020).

Surveillance is an organized data collection system that aims to support detection goals and use the gathered data efficiently to optimize a hazard response (Cook et al. 2022). Surveillance and detection depend on the plant system, target pathogen, and location. Currently, the U.S. applies a surveillance strategy where all potentially dangerous events (either deliberate or accidental) are monitored and are performed based on "at-risk" areas prioritizing pathogens within the Select Agent Program List (Fletcher et al. 2020; Gilchrist et al. 2015).

Prevention is preferred to avoid dispersing new pathogen incursions, regardless of being accidentally, naturally, or deliberately triggered. For this strategy to be effective, rapid, and accurate diagnosis/detection is needed (Gullino et al. 2017). Early detection plays a vital role in surveillance activities and implementing control and management strategies, which, if not implemented, can diminish food security, and dramatically affect the economy (Karunarathna et al. 2021).

1.2.2.Regulations

The use of science to generate harm has been seen throughout history. It has changed how scientists acquire and work with pathogenic and high-risk microorganisms, including biological toxins (Morse and Quigley 2020). To control and restrict the inappropriate and unauthorized use and manipulation of microorganisms that threaten humans, plants, animals, and the environment; the U.S. approved several laws and promulgated regulations (Morse 2015).

Regulations involve the collaboration, coordination, and communication among multiple agencies and organizations at different levels (local, state, federal, and international). Most U.S.

states have laws requiring reporting diseases with regulatory implications to officers, being the highest-level plant health official the State Plant Regulatory Official (SPRO). The SPRO and the SDA (State Department of Agriculture) have the authority to deploy a 90-day stop-movement order on plant materials and establish quarantine protocols. Overall, the federal plant regulatory authority is represented by the USDA APHIS Plant Protection Quarantine Unit (PPQ) (Fletcher et al. 2020; Lee et al. 2019; Morse, Budowle, and Schutzer 2020; Morse and Quigley 2020).

1.2.3.Select Agent Program

Over the last two decades, multiple events have changed the way scientists acquire and work with pathogenic organisms and biological toxins, driven by numerous biological attacks such as: 1) the release of the sarin nerve agent in Tokyo (1995), 2) the bombing on the Murrah Federal Building in Oklahoma City (1995), 3) and terrorist attacks involving anthrax in 2001 (Morse and Weirich 2011). These events created the need to limit unauthorized access to high-risk pathogenic microorganisms and biological toxins. Therefore the U.S. established legislation to oversight biological sciences by the Federal Government (Murrin 2018). The newly incorporated regulations were described into three Codes of Federal Regulations (7 C.F.R. Part 331, 9 C.F.R. Part 121, and 42 C.F.R. Part 73), which included a list of infectious agents and biological toxins, known as Select Agents, that can be used with terrorism purposes (National Research Council 2010).

The Federal Select Agent Program (FSAP) is formed by two agencies, the U.S. Department of Health and Human Services (HHS) and the U.S. Department of Agriculture (USDA). The first oversees the Select Agents that threaten public health and safety through the Centers for Disease Control and Prevention (CDC) – Division of Select Agents and Toxins (DSAT). On the other hand, the USDA, through the Animal and Plant Health Inspection Service (APHIS) – Agriculture Select Agent Services (AgSAS), supervise the Select Agents that can harm animal and plant health, as well as their products (Morse and Quigley 2020; Murrin 2018). When this review was done, the FSAP controlled 68 biological select agents and toxins listed at https://www.selectagents.gov.

2. Soybean production in the United States

2.1. Overview and importance of soybeans

2.1.1.Economic importance

Soybean (*Glycine max* [L]. Merr.) is a legume that is considered one of the most critical commodities globally since more than 6% of the total agricultural land is used to grow this crop, producing 71% and 29% of the entire plant-based protein meal and oil, respectively (Bateman et al. 2020; Lin et al. 2022). In 2022, soybean production in the U.S. recorded a total of 4.3 billion bushels, with an average yield of 50 bushels per acre and a total planted area of 86 million acres, which represented an annual worth of more than $60 billion (United States Department of Agriculture 2023), establishing soybean as the second most valuable crop in the U.S. just preceded by corn (Hartman et al. 2011).

Soybean seeds are appraised for their high-quality oil and protein content, where 70% of their value is due to their use as a meal for livestock and poultry (Roth et al. 2020). Additionally, soybean has multiple applications, such as alternative protein sources for human consumption and vegetable oil (Bateman et al. 2020). Therefore, its price is supported by domestic demand, the development of renewable diesel capacity, and the increasing demand for diverse diets, including plant-based protein meals (Meyer 2021).

2.1.2.Food safety

Soybean is one of the most used and cultivated crops worldwide since its seed has an estimated 20% oil and 40% protein, the highest protein content per unit land area of any crop

(Murithi et al. 2022). Even though most of its production is destined for livestock meals and vegetable oil, there is a growing rate of soybean seed used directly for human consumption (Watanabe, Losák, and Vollmann 2018). Due to their nutritional value and functional components, whole soybeans and their derivates are widely used in everyday foods and convenience products (Bryant et al. 2020). This crop plays a vital role in the future of the world's food safety because it has an intrinsic high temperature and drought tolerance and an important amount of protein content (Ahmadzadeh et al. 2018), which can ensure sustainable future food supply based on the current farmland areas and the climate change trends (Zhan et al. 2019).

There is a high variation in the soybean market, and due to the amount of acreage used for its cultivation and its economic and food safety importance, there is a common need to minimize yield losses. However, soybeans, as well as other crops, are susceptible to the attack of different pests and diseases that will decrease yield. For instance, in 2018, more than 500 million bushels were lost to soybean diseases (Juroszek et al. 2020; Roth et al. 2020).

2.2. Soybean plant diseases

The soybean yield is affected by plant diseases and pests, and the first can generate up to 20% of yield losses. However, some diseases may vary and produce a higher or lower impact (Faske et al. 2014). More than 200 pathogens are known to affect soybeans, and at least 35 can significantly impact the economy (Lin et al. 2022).

2.2.1.Common diseases

Soybean is affected by multiple and diverse plant pathogens, including bacteria, fungi, oomycetes, viruses, and nematodes (Faske et al. 2014). Bacterial blight, caused by *Pseudomonas savastanoi* pathovar *glycinea,* is the most common and ubiquitous bacterial disease of soybeans, generating significant yield reductions mainly when a susceptible cultivar is under high-stress pressure (Hartman et al. 2015). Anthracnose, Brown Spot, Sudden Death Syndrome, and Leaf

Blight are examples of soybean fungal diseases that are caused by *Colletotrichum truncatum, Septoria glycines*, *Fusarium virguliforme,* and *Cercospora kikuchii,* respectively (Boufleur et al. 2021; Kashiwa et al. 2021; Neves et al. 2022; Rodriguez et al. 2021). However, Soybean Rust, caused by *Phakopsora pachyrizi,* is the most important fungal disease due to its ability to generate up to 80% of yield losses (Hu et al. 2020). Soybean crops are also susceptible to the attack of different oomycetes such as *Pernospora manshurica* and *Phytophthora sojae*, causal agents of Downy Mildew and Phytophthora Root Rot, respectively (Madina et al. 2022; Taguchi-Shiobara et al. 2019). Regarding viral diseases, the two most common ones are Soybean Mosaic Virus (SVM) and Bean Pod Mottle Virus (BPMV), which can generate up to 94% of yield losses depending on the production system, variety, and location (Widyasari, Alazem, and Kim 2020; Zhou and Tzanetakis 2020). Finally, Soybean Cyst Nematode (SCN), caused by *Heterodera glycines,* is one of the most critical soybean pathogens due to the yield losses but also because the nematode cysts can survive up to ten years in the soil (Hartman et al. 2015; Lin et al. 2022; Roth et al. 2020). Over the years, there has been a notorious trend where disease pressure is increasing, and more yield losses have been evidenced (Bandara et al. 2020).

3. Red Leaf Blotch of soybeans

3.1. Symptoms on soybean

Red leaf blotch (RLB) of soybeans is a disease caused by the fungal pathogen *Coniothyrium glycines* (de Gruyter et al. 2013), formerly known as *Phoma glycinicola* (de Gruyter and Boerema 2002), *Dactuliochaeata glycines* (Leakey 1964), and *Pyrenochaeta glycines* (Stewart 1957). Soybeans are susceptible to RLB from the seedling stage to maturity, and the disease starts with irregular, circular, small, and dark leaf spots. The lesions are usually related to leaf veins and become more extensive, necrotic, and surrounded by a chlorotic halo (Hartman et al. 1987). During the enlargement of the lesions, sclerotia start to form beneath the leaf surface (Datnoff

1987). Small and heterogeneously distributed pycnidia may be visible to the naked eye within the blotches (Hartman et al. 1987; Stewart 1957). The disease can cause lesions in multiple soybean tissues, such as foliage, petioles, pods, and stems (Hartman et al. 1987). The most damaging outcome of the disease is leaf abscission, which generates up to 75 percent defoliation of some soybean varieties, reducing their photosynthetic abilities and undermining their yield (Levy et al. 1990; Stewart 1957).

High humidity and rain can influence and increase the severity of the disease; this can happen throughout the growing season from November to April (Hartman et al. 2011). Additionally, the condition is more severe during the late or mature stages of development. Regarding different soybean varieties, Juniper and Tunia showed lower disease severity (Datnoff 1987).

3.2. Impact on soybean production

The assessment of yield losses generated by RLB is not well described but is related to the severe defoliation it causes when the plant is severely affected (Murithi et al. 2022). Premature leaf abscission is correlated with a reduced photosynthetic metabolism and slow movement of active compounds through the plant, which in the end, affects the weight of the seed (Datnoff 1987). Since 1981, this disease has negatively impacted soybean production and its commercialization in multiple African countries (Levy et al. 1990). For instance, the yield reduction in Zambia is approximately 50% (Datnoff 1987). In 1989, Sinclair found that all the U.S. soybean cultivars are susceptible to the disease. No resistance was present across local and exotic varieties; this study evaluated more than five thousand soybean lines. Finally, Murithi et al. (2022) analyzed 59 soybean entries for RLB in four African countries (Ethiopia, Kenya, Uganda, and Zambia), concluding that there was a disease incidence of 100% and that the severity differed across locations. These findings suggest a partial resistance of soybean to RLB. However, more research must be done to verify the resistance level and find new sources of resistance.

3.3. Control measures

Since the first report of the disease in 1957, chemical and cultural methods to control RLB have been assessed. Different cultural approaches have been tested (crop rotations, plant spacing, and different tillage strategies); however, the results were inconclusive, and further studies are needed before any official recommendation. An analysis performed in 2022 found that the severity of RLB is linked to rainfall and wind speed (Murithi et al. 2022), and the management of these factors can be used to establish control strategies for the disease.

Regarding the chemical controls, fentin acetate and benomyl have been used to effectively control the disease increasing the yield by 13 and 23%, respectively (Hartman et al. 1987; Levy et al. 1990). Fungicide application seemed to enhance genetic resistance by breeding soybean cultivars with higher tolerance to the disease (Levy et al. 1990). Applying chemical treatments increases production costs significantly, which is not feasible for most growers and generates adverse environmental effects (Murithi et al. 2022).

4. *Coniothyrium glycines*

4.1. Taxonomy

Coniothyrium glycines [(R.B. Stewart) Verkley & Gruyter] taxonomy has gone through multiple revisions across the years, it started as *Pyrenochaeta glycines* (Stewart 1957), and then Leakey (1964) described its sclerotial stage as *Dactuliophora glycines.* Later,Datnoff (1987) revealed that *P. glycines* and *D.* glycines are the same organisms based on observations in herbarium specimens. Borema classified the organism within the *Phoma* genera and named it *Phoma glycinicola.* Finally, in 2013 de Gruyter et al. defined it as *Coniothyrium glycines* (Hartman et al. 2011) based on the greenish-yellow color of the conidia mass (Tooley 2017). The current taxonomy of *C. glycines* is described as follows (Schoch et al. 2020):

Kingdom: Fungi

Phylum: Ascomycota

Class: Dothideomycetes

Order: Pleosporales

Family: Coniothyriaceae

Genus: *Coniothyrium*

Species: *Coniothyrium glycines*

4.2. Global distribution

C. glycines was first reported in Ethiopia in 1953 at the Jimma Agricultural Experiment Station (Stewart, 1957). Since then, its incidence has been increasing mainly across countries in sub-Saharan Africa (Cameroon, the Democratic Republic of the Congo, Malawi, Mozambique, Nigeria, Rwanda, Tanzania, Uganda, Zambia, and Zimbabwe) (Datnoff 1987; Hartman and Murithi 2022). Additionally, there is a single report of the disease in Bolivia (A south American country) and India (an Asian country) (EPPO 2023). Even though its current distribution is limited to African countries, it can potentially become a significant soybean foliar disease (Tooley 2017). The current geographical distribution of *C. glycines* is shown in Figure II-1.

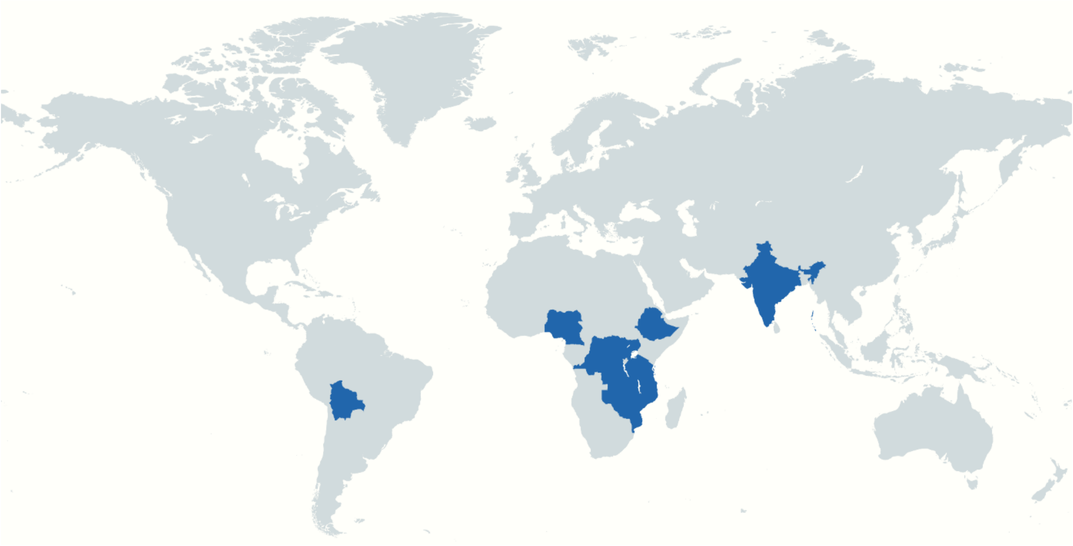


Figure II-1: Geographical distribution of *C. glycines* based on indexed reports. Source: EPPO (2023) & Hartman & Murithi (2022).

4.3. Biology and life cycle

The only known economically important crop that *C. glycines* infects is soybean. However, it has been detected in a wild perennial legume, *Neonotonia wightii*, which is likely the natural reservoir of the pathogen (Hartman et al. 1987; Hartman and Sinclair 1992; Stewart 1957). The general life cycle of *C. glycines* is not fully characterized, but in 1987 Hartman et al. suggested a model (Figure II-2) where sclerotia present in the soil can reach and infect the plant tissue by rain-splashes, wind, fomites (contaminated materials), and other biotic factors such as the movement of humans and animals (Hartman et al. 2011). Under suitable conditions, sclerotia germinate, forming mycelia or pycnidia, a process that can take up to seven days. Conidia are included within each pycnidium and serve as a secondary inoculum to reinfect the same plant. To complete the life cycle, pycnidia, and sclerotia return to the soil in the plant debris. Those fungal surviving structures can overwinter and serve as primary inoculum for the next growing season (Hartman et al. 1987). Under natural conditions, sclerotia can survive without a host for up to seven consecutive dry moths, and if they are kept at 5°C, they have a viability of 22% after 18 months (Hartman and Sinclair 1992).

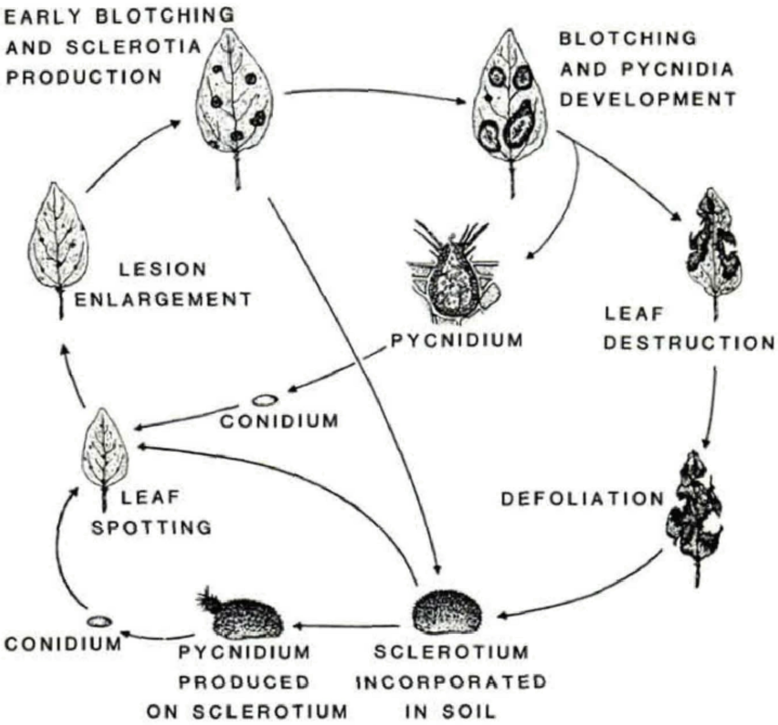

EARLY BLOTCHING AND SCLEROTIA PRODUCTION
BLOTCHING AND PYCNIDIA DEVELOPMENT
LESION ENLARGEMENT
PYCNIDIUM
LEAF DESTRUCTION
CONIDIUM
LEAF SPOTTING
DEFOLIATION
CONIDIUM
PYCNIDIUM PRODUCED ON SCLEROTIUM
SCLEROTIUM INCORPORATED IN SOIL

Figure II-2: ***C. glycines*** **life cycle during the development of red leaf blotch of soybeans. Source: Hartman et al. (1987).**

C. glycines pycnidia are spherical or flattened structures formed of brownish cells that become darker at the ostiole. Conidia or pycnospores are oval, straight, hyaline systems developed inside each pycnidium and serve as a secondary inoculum (Stewart, 1957). On the other hand, *C. glycines*, well-defined melanized sclerotia surrounded with setae, is unique among other *Coniothyrium* species and is used for field diagnosis (Hartman et al. 2011).

In a laboratory setting, *C. glycines* can be cultured in artificial mediums such as clarified V8 juice, cornmeal, and malt agars. Depending on the media used, isolates show different pigmentation and macroscopic characteristics. For instance, mycelia and pycnidia production are restricted in nutrient-deficient media (water agar) (Hartman et al. 1987).

4.4. Select Agent status

Permits and registrations are needed to handle *C. glycines*, which are provided and regulated by two authorities: the Plant Protection Act of 2000 and the Agricultural Bioterrorism Protection Act of 2002; both codified by the 7 C.F.R. (Code of Federal Regulations) Part 330, and Part 331, respectively (Hartman et al. 2011). These regulations are the same as those that manage the Federal Select Agent Program. PPQ (Plant Protection and Quarantine) permits are mandatory to possess, use, and transfer the select agent, *C. glycines*. However, diagnostic laboratories are exempt from these regulations if they destroy the cultures within seven days (Morse and Quigley 2020). Within Oklahoma State University, access to *C. glycines* is restricted and authorized just for people that went through a Security Risk Assessment (SRA) clearance by the Department of Justice (DOJ) and the Federal Bureau of Investigations (FBI) (Oklahoma State University 2023).

5. Traditional and novel plant disease identification

It is estimated that by 2050, food production must be increased by about 70% to meet the needs of the growing human population (Godfray et al. 2010). Currently, more than one billion people live in malnutrition conditions due to the poor food supply and lack of food nutrient values within their food (Fang and Ramasamy 2015). Among all the factors affecting agricultural productivity, damage caused by pests and plant pathogens is one of the most important. Disease detection and prevention are essential to minimize their negative impact on productivity and ensure agricultural sustainability (Kartikeyan and Shrivastava 2021; Ristaino et al. 2021).

Identifying a disease can be divided into two major categories, direct and indirect sample analysis. The first focuses on finding the target or any of its components within a sample, while the latter aims to detect secondary products associated with the presence of the target (Fang and Ramasamy 2015).

Traditional techniques used for the identification of plant diseases are based on the observation skills of growers, producers, master gardeners, farmers, and agronomists, that explore a sample with microscopes, lenses, or with their naked eye, trying to find characteristic structures such as fruiting bodies, mycelium, and spores (Gullino et al. 2017). Another traditional approach is the *in vitro* culturing of tentative causal organisms isolated from infected tissues and then purified to confirm causality by performing Koch's postulates (Sankaran et al. 2010). Both traditional methods require expertise and are time-consuming. Additionally, they have limitations when different organisms, species, cryptic species, sub-species, pathotypes, or other variants share the same morphological characteristics or cannot be cultured *in vitro* (Buja et al. 2021).

Advances in nucleic acid technologies have driven the molecular diagnosis of plant diseases, which currently are the best choice due to their speed, reliability, scaling capacities, and high specificity and sensitivity; hence, for some conditions, they are considered the "gold standard" (Gullino et al. 2017). Techniques such as PCR (Polymerase Chain Reaction), LAMP (Loop-mediated Isothermal Amplification), RPA (Recombinase Polymerase Amplification), RT-PCR (Retro-transcribed PCR), and qPCR (Quantitative PCR) are commonly used in diagnostic laboratories worldwide. However, they need bulky instruments, experienced personnel, and previous knowledge of the organism to be identified so primers and probes can be developed (Fang and Ramasamy 2015).

Next-generation sequencing (NGS) has been growing and improving during the last few years, reducing the costs of DNA sequencing and generating large amounts of data to be analyzed. When combined with bioinformatic pipelines, NGS is suitable for detecting new causal agents without previous knowledge of their identity (Gullino et al. 2017).

5.1. Sequencing

5.1.1.Background

After discovering the three-dimensional structure of DNA by Watson and Crick in 1953, the need to find what was encoded within that molecule drove science and technology (Heather and Chain 2016). However, it was not until 1977 that a breakthrough was made with the development of Sanger sequencing (first-generation sequencing) based on "chain termination" or dideoxy technique (Sanger, Nicklen, and Coulson 1977), which until now is one of the most used sequencing technologies due to its accuracy, robustness, and ease of use (Shendure et al. 2017). From its invention, Sanger sequencing underwent multiple improvements to the point that it was used in the Human Genome Project (Lander et al. 2001).

The second-generation sequencing took advantage of the amount of light produced by the formation of pyrophosphate after the reaction of ATP, hence known as pyrosequencing (Ronaghi, Uhlén, and Nyrén 1998). This sequencing strategy became the first primary successful next-generation sequencing (NGS) technology after being licensed to 454 Life Sciences and then purchased by Roche® (Heather and Chain 2016). Illumina followed the path of pyrosequencing and started as a second-generation sequencing, currently considered an NGS, that uses complementary oligonucleotides fixed to a flowcell that will generate a cluster of clonal populations after a solid phase PCR, a process known as "bridge amplification" (Bentley et al. 2008).

Third-generation sequencing or next-generation sequencing (NGS) is a combination of single-molecule sequencing (SMS) and real-time sequencing without the need for DNA amplification (Heather and Chain 2016). PacBio is an essential NGS technology that generates the same polymerase rate and produces long-read functional genome assemblies (Hu et al. 2021). Similarly, Oxford Nanopore Technologies (ONT) developed a long-read sequencing technique using nanopores. Each nucleotide is identified based on changes in electric signals while the DNA/RNA molecule moves through the pore (Eisenstein 2012).

The identification of organisms, including plant pathogens, has been aided by the analysis of nucleic acid sequences since they have revealed genetic differences that older methodologies have been unable to detect (Sjödin et al. 2013). Additionally, high throughput sequencing enabled the screening of many samples for the presence of target pathogens (Fletcher et al. 2020).

5.1.2.Next-Generation Sequencing (NGS)

Advances and constant improvements have been made to NGS platforms since their discovery. Nowadays, they are more accessible, cost-effective, and faster, allowing multiple organisms to be detected based on their genomic information (Gilchrist et al. 2015). The amount of data generated through NGS is challenging and becomes a bottleneck for further processing (Espitia-Navarro et al. 2020). It has been estimated that sequencer capabilities have been growing faster than computational growth, which is against Moore's Law (Stein 2010).

Compared to Sanger sequencing, NGS has a higher error rate. However, it is mitigated by the amount of data generated and the depth coverage, allowing the generation of consensus sequences with different qualities (Espitia-Navarro et al. 2020). NGS has been applied mainly to study single organisms. This approach has recently been extended to analyze whole populations of microorganisms and implemented in forensics decision-making pipelines (Gilchrist et al. 2015).

5.1.2.1. Illumina sequencing

Illumina's NGS technology uses a sequencing by synthesis (SBS) strategy with fluorescent reversible terminators. Before sequencing, there is an amplification step where a cluster of clonal molecules is generated in a process called "bridge amplification" PCR that strengthens the intensity of the signal (Hu et al. 2021). Sequencing is performed by incorporating a uniquely labeled reversible terminator into the nucleic acid chain, and a sensor captures the resulting signal. The terminator and dye are removed, allowing the integration of a newly labeled nucleotide (Shendure et al. 2017). Illumina generates short reads (< 300bp) that can be single- or paired-end, depending on the application and platform used. Additionally, it is the most accurate base-by-base sequencing platform on the market, with an accuracy of 99.9%. Illumina's disadvantages are its relatively long run time, equipment cost, haplotype phasing, and complications during *de novo* assemblies (Ansorge 2009; Heather and Chain 2016).

5.1.2.2. Oxford Nanopore sequencing

Oxford Nanopore Technologies (ONT) is a long-read NGS platform that moves a single nucleic acid molecule through a pore aided by ligation adaptors. The sequencing library is loaded into a flow cell harboring multiple nanopores embedded in a membrane. Sequencing captures the ion current's disruption while each nucleotide passes through the pores (Karst et al. 2021). The performance of this sequencing platform is related to the quality of the high molecular weight DNA used during the library preparation. Based on the base-calling algorithm, its accuracy ranges from 87% to 98% (Rang, Kloosterman, and de Ridder 2018). The advantages of ONT are its relatively low cost, portability, and field adaptability. Meanwhile, the main disadvantage of this sequencing platform remains its high-error rate (2-15%) (Hu et al. 2021).

5.1.3. Applications in plant diagnostics

NGS is considered the future for detecting plant pathogens, making it valuable for quarantine and certification purposes, especially in high-value crops such as fruits and woody plants (Gullino et al. 2017). These newer technologies allow gene discovery, *de novo* genome assemblies, pathogen discovery, and tracking plant pathogen movement at a relatively low cost. Combined with multiple bioinformatic pipelines, the data generated will elucidate important plant pathogens' origin, ancestry, and evolution (Ristaino et al. 2021).

The limitations of NGS are related to the amount of genetic material used for sequencing and the samples' purity. Also, the analysis of NGS data requires high computational resources. Moreover, there is a dependency on nucleic acid and genomic databases, which can be redundant or poorly curated (Loman et al. 2012). Detecting a pathogen through sequencing doesn't mean that it is the causal agent of a disease. It must be confirmed with cultures or Koch's postulates (Ristaino et al. 2021). Finally, these new technologies require new legal frameworks to be approved in court in case of a forensic investigation (Gilchrist et al. 2015).

5.1.3.1. E-probe Diagnosis Nucleic acid Analysis (EDNA)

NGS has revolutionized the detection of plant pathogens, disregarding the sample's purity. It can detect a target within a pure sample or multiple targets within a microbiome. However, bioinformatic processing is a bottleneck for data analysis and the computational resources needed (Hu et al. 2021). E-probe Diagnostic Nucleic Acid Analysis (EDNA) is a bioinformatic pipeline that eliminates the need for bioinformatic expertise and computational resources to detect one or multiple targets within raw sequencing data. The detection is performed by comparing e-probes, short sequences that represent an electronic fingerprint of the target, with the raw sequencing reads of a sample using BLAST (Basic Local Alignment Search Tool) (Espindola et al. 2018; Stobbe et al. 2013).

EDNA has been used and validated in several studies, mainly for detecting plant pathogens within metagenomic raw sequencing data. In 2014, Stobbe et al. validated EDNA using mock sequence databases to detect Bean golden yellow mosaic virus (BGYMV) and Plum pox virus (PPV). The following year, A. Espindola et al. (2015) used EDNA to detect two fungal pathogens (*Puccinia graminis f. sp. tritici* and *Phakopsora pachyrhizi*) and two oomycetes (*Phytophthora ramorum* and *Pythium ultimum*) within simulated metagenomic data. In 2020, Pena-Zuniga validated EDNA for detecting 117 plant viruses across three matrices (roses, cucurbits, and water). Recently, EDNA was used to detect several *Pythium* spp., *Phytophthora* spp., *Globisporangium* spp., *Hyaloperonospora* spp., and *Plasmopara* spp. within simulated and real sequencing reads (Espindola et al. 2022; Proano-Cuenca, Espíndola, and Garzon 2022).

5.1.3.2. MiFi®: Microbe Finder

MiFi® is a web application (https://bioinfo.okstate.edu/) that harbors EDNA bioinformatic pipeline and allows users to develop specific e-probes and test them against sequencing datasets. It is formed by two main components, MiProbe® and MiDetect®. The first one is used for the e-probe design process, where the user uploads the target(s) genomic information and the near neighbor(s) dataset and selects the desired e-probe length. MiProbe® performs a genomic comparison between both datasets and extracts unique e-probes for the target, which are stored in a database within MiFi®. Regarding MiDetect®, it allows the user to test the designed e-probes against any sequencing data by comparing them using BLAST and providing a positive or negative outcome based on a statistical test (Espindola and Cardwell 2021).

5.2. Validation metrics for diagnostic tests

To select a diagnostic tool or during its validation process, different metrics have been proposed to frame the performance characteristics of a test. These metrics aim to help understand an assay's reliability under different scenarios (Cardwell et al. 2018).

5.2.1.Analytical and diagnostic sensitivity

The analytical sensitivity of an assay is related to the ability to detect a low concentration of a given target in a sample and is expressed in terms of concentration. Hence, the higher the analytical sensitivity, the lower the detectable concentration (Saah and Hoover 1997). Some literature showed that the limit of detection (LOD) or the minimal detectable concentration are synonyms for analytical sensitivity (Shreffler and Huecker 2022; Šimundić 2009).

On the other hand, diagnostic sensitivity is the ability to detect a condition in a population, which means the percentage or proportion of individuals/samples with a specific disease. The diagnostic sensitivity is calculated by the number of true positives (have the condition and a positive test result) divided by the number of individuals with the disease, including the ones not detected by the test (Cardwell et al. 2018; Saah and Hoover 1997).

5.2.2. Analytical and diagnostic specificity

Analytical specificity refers to the ability of a test to identify exclusively a desired target instead of similar or related individuals (Shreffler and Huecker 2022). Meanwhile, diagnostic specificity stands for the ability of a test to correctly identify an individual/sample that lacks the target in question (Saah and Hoover 1997). This value is calculated by dividing the true negatives (individuals without the target that were negative in the assay) by the individuals without the condition (Cardwell et al. 2018; Šimundić 2009).

References

Ahmadzadeh, Hamidreza, Aimrun Wayayok, Alireza Massah, Ebrahim Amiri, Ahmad Fikri Abdullah, Jahanfar Daneshian, and C. B. S. Teh. 2018. “Impacts of Climate Change on Soybean Production under Different Treatments of Field Experiments Considering the Uncertainty of General Circulation Models.” *Agricultural Water Management* 205:63–71. doi: 10.1016/j.agwat.2018.04.023.

Alonso, Vasthi, Stephen Parnell, and Frank Van den Bosch. 2016. “Monitoring Invasive Pathogens in Plant Nurseries for Early-Detection and to Minimise the Probability of Escape.” *Journal of Theoretical Biology* 407:290–302. doi: 10.1016/j.jtbi.2016.07.041.

Anderson, Pamela K., Andrew A. Cunningham, Nikkita G. Patel, Francisco J. Morales, Paul R. Epstein, and Peter Daszak. 2004. “Emerging Infectious Diseases of Plants: Pathogen Pollution, Climate Change and Agrotechnology Drivers.” *Trends in Ecology & Evolution* 19(10):535–44. doi: 10.1016/j.tree.2004.07.021.

Ansorge, Wilhelm J. 2009. “Next-Generation DNA Sequencing Techniques.” *New Biotechnology* 25(4):195–203. doi: 10.1016/j.nbt.2008.12.009.

Avila-Quezada, Graciela Dolores, Jesus Fidencio Esquivel, Hilda Victoria Silva-Rojas, Santos Gerardo Leyva-Mir, Clemente de Jesús Garcia-Avila, Andrés Quezada-Salinas, Lorena Noriega-Orozco, Patricia Rivas-Valencia, Damaris Ojeda-Barrios, and Alicia Melgoza-Castillo. 2018. “Emerging Plant Diseases under a Changing Climate Scenario: Threats to Our Global Food Supply.” *Emirates Journal of Food and Agriculture* 30(6):443–50. doi: 10.9755/ejfa.2018.v30.i6.1715.

Bandara, Ananda Y., Dilooshi K. Weerasooriya, Carl A. Bradley, Tom W. Allen, and Paul D. Esker. 2020. "Dissecting the Economic Impact of Soybean Diseases in the United States over Two Decades." *PLOS ONE* 15(4):e0231141. doi: 10.1371/journal.pone.0231141.

Bateman, Nick R., Angus L. Catchot, Jeff Gore, Don R. Cook, Fred R. Musser, and J. Trent Irby. 2020. "Effects of Planting Date for Soybean Growth, Development, and Yield in the Southern USA." *Agronomy* 10(4):596. doi: 10.3390/agronomy10040596.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M. J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M. D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L.

Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung-Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W. Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie vandeVondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456(7218):53–59. doi: 10.1038/nature07517.

Borema, Gerhard. 2004. *Phoma Identification Manual: Differentiation of Specific and Infra-Specific Taxa in Culture*. Vol. 1. 1st ed. edited by G Borema and CABI. CABI.

Boufleur, Thais R., Maisa Ciampi-Guillardi, Ísis Tikami, Flávia Rogério, Michael R. Thon, Serenella A. Sukno, Nelson S. Massola Júnior, and Riccardo Baroncelli. 2021. "Soybean Anthracnose Caused by Colletotrichum Species: Current Status and Future Prospects." *Molecular Plant Pathology* 22(4):393–409. doi: 10.1111/mpp.13036.

Bryant, C. J., L. J. Krutz, D. B. Reynolds, M. A. Locke, B. R. Golden, T. Irby, R. W. Steinriede, G. D. Spencer, B. E. Mills, and C. W. Wood. 2020. "Conservation Soybean Production Systems in the Mid-southern USA: II. Replacing Subsoiling with Cover Crops." *Crop, Forage & Turfgrass Management* 6(1). doi: 10.1002/cft2.20058.

Buja, Ilaria, Erika Sabella, Anna Grazia Monteduro, Maria Serena Chiriacò, Luigi De Bellis, Andrea Luvisi, and Giuseppe Maruccio. 2021. "Advances in Plant Disease Detection and Monitoring: From Traditional Assays to In-Field Diagnostics." *Sensors* 21(6):2129. doi: 10.3390/s21062129.

Cardwell, K., D. Geoffrey, A. Flannery, J. Fletcher, D. Luster, M. Nakhla, A. Rice, P. Shiel, J. Stack, C. Walsh, and L. Levy. 2018. "Diagnostic Assay Validation Terminology." *Plant Health Instructor* 2018(1). doi: 10.1094/PHI-I-2018-0709-01.

Cook, G., J. Morisette, M. Remmenga, J. Russo, and K. Spiegel. 2022. "Surveillance Design after Initial Detection." Pp. 1–51 in *Biosecurity Toolbox*. Vol. 1.

Datnoff, L. 1987. "Effect of Red Leaf Blotch on Soybean Yields in Zambia." *The American Phytopathological Society* 71(2):1–4.

Eisenstein, Michael. 2012. "Oxford Nanopore Announcement Sets Sequencing Sector Abuzz." *Nature Biotechnology* 30(4):295–96. doi: 10.1038/nbt0412-295.

EPPO. 2023. "EPPO Global Database." *Global Database*. Retrieved February 19, 2023 (https://gd.eppo.int/).

Espindola, Andres, and Kitty Cardwell. 2021. “Microbe Finder (MiFi®): Implementation of an Interactive Pathogen Detection Tool in Metagenomic Sequence Data.” *Plants* 10(2):250. doi: 10.3390/plants10020250.

Espindola, Andres, Kitty Cardwell, Frank N. Martin, Peter R. Hoyt, Stephen M. Marek, William Schneider, and Carla D. Garzon. 2022. “A Step Towards Validation of High-Throughput Sequencing for the Identification of Plant Pathogenic Oomycetes.” *Phytopathology®* 112(9):1859–66. doi: 10.1094/PHYTO-11-21-0454-R.

Espindola, Andres, William Schneider, Kitty Cardwell, Yisel Carrillo, Peter Hoyt, Stephen Marek, Hassan Melouk, and Carla Garzon. 2018. “Inferring the Presence of Aflatoxin-Producing Aspergillus Flavus Strains Using RNA Sequencing and Electronic Probes as a Transcriptomic Screening Tool” edited by R. A. Wilson. *PLOS ONE* 13(10):e0198575. doi: 10.1371/journal.pone.0198575.

Espindola, Andres, William Schneider, Peter R. Hoyt, Stephen M. Marek, and Carla Garzon. 2015. “A New Approach for Detecting Fungal and Oomycete Plant Pathogens in next Generation Sequencing Metagenome Data Utilising Electronic Probes.” *International Journal Data Mining and Bioinformatics* 12(2):1–14.

Espitia-Navarro, Hector F., Lavanya Rishishwar, Leonard W. Mayer, and I. King Jordan. 2020. “Bioinformatics.” Pp. 267–82 in *Microbial Forensics*. Elsevier.

Fang, Yi, and Ramaraja Ramasamy. 2015. “Current and Prospective Methods for Plant Disease Detection.” *Biosensors* 5(3):537–61. doi: 10.3390/bios5030537.

FAO. 2007. *FAO Biosecurity Toolkit*. Vol. 1. 1st ed. edited by E. Boutrif and S. Pandey. Rome: FAO.

Faske, Travis, Terry Kirkpatrick, Jing Zhou, and Ioannis Tzanetakis. 2014. “Chapter 11: Soybean Diseases.” Pp. 1–18 in *Arkansas Soybean Production Handbook*. Vol. 4. University of Arkansas.

Fisher, Matthew C., Daniel. A. Henk, Cheryl J. Briggs, John S. Brownstein, Lawrence C. Madoff, Sarah L. McCraw, and Sarah J. Gurr. 2012. “Emerging Fungal Threats to Animal, Plant and Ecosystem Health.” *Nature* 484(7393):186–94. doi: 10.1038/nature10947.

Fletcher, Jacqueline, Neel G. Barnaby, James Burans, Ulrich Melcher, Douglas G. Luster, Forrest W. Nutter, Harald Scherm, David G. Schmale, Carla S. Thomas, and Francisco M. Ochoa Corona. 2020. “Forensic Plant Pathology.” Pp. 49–70 in *Microbial Forensics*. Elsevier.

Fones, Helen N., Daniel P. Bebber, Thomas M. Chaloner, William T. Kay, Gero Steinberg, and Sarah J. Gurr. 2020. “Threats to Global Food Security from Emerging Fungal and Oomycete Crop Pathogens.” *Nature Food* 1(6):332–42. doi: 10.1038/s43016-020-0075-0.

Gilchrist, Carol A., Stephen D. Turner, Margaret F. Riley, William A. Petri, and Erik L. Hewlett. 2015. “Whole-Genome Sequencing in Outbreak Analysis.” *Clinical Microbiology Reviews* 28(3):541–63. doi: 10.1128/CMR.00075-13.

Godfray, H. Charles J., John R. Beddington, Ian R. Crute, Lawrence Haddad, David Lawrence, James F. Muir, Jules Pretty, Sherman Robinson, Sandy M. Thomas, and Camilla Toulmin. 2010. “Food Security: The Challenge of Feeding 9 Billion People.” *Science* 327(5967):812–18. doi: 10.1126/science.1185383.

de Gruyter, J., and G. H. Boerema. 2002. “Contributions towards a Monograph of Phoma (Coelomycetes) VIII. Section Paraphoma: Taxa with Setose Pycnidia.” *Persoonia* 17:59–60.

de Gruyter, J., J. H. C. Woudenberg, M. M. Aveskamp, G. J. M. Verkley, J. Z. Groenewald, and P. W. Crous. 2013. “Redisposition of Phoma-like Anamorphs in Pleosporales.” *Studies in Mycology* 75:1–36. doi: 10.3114/sim0004.

Gullino, Maria Lodovica, James P. Stack, Jacqueline Fletcher, and John D. Mumford. 2017. *Practical Tools for Plant and Food Biosecurity*. Vol. 8. 1st ed. edited by M. L. Gullino, J. P. Stack, J. Fletcher, and J. D. Mumford. Cham: Springer International Publishing.

Hartman, G., L. Datnoff, C. Levy, J. Sinclair, D. Cole, and F. Javaheri. 1987. "Red Leaf Blotch of Soybeans." *Plant Disease* 113–18.

Hartman, G. L., and J. B. Sinclair. 1992. "Cultural Studies on Dactuliochaeta Glycines, the Causal Agent of Red Leaf Blotch of Soybeans." *Plant Disease* 76:847–52.

Hartman, G., and H. M. Murithi. 2022. "Coniothyrium Glycines (Red Leaf Blotch)." *CABI Compendium* CABI Compendium. doi: 10.1079/CABICOMPENDIUM.17687.

Hartman, Glen, James Haudenshield, Kent Smith, and Paul Tooley. 2011. *Recovery Plan for Red Leaf Blotch of Soybean Caused by Phoma Glycinicola*.

Hartman, Glen Lee, John Clark Rupe, Edward J. Sikora, Leslie Leigh Domier, Jeff A. Davis, and Kevin Lloyd Steffey. 2015. *Compendium of Soybean Diseases and Pests*. Vol. 1. The American Phytopathological Society.

Heather, James M., and Benjamin Chain. 2016. "The Sequence of Sequencers: The History of Sequencing DNA." *Genomics* 107(1):1–8. doi: 10.1016/j.ygeno.2015.11.003.

Hu, Dongfang, Zhi-Yuan Chen, Chunquan Zhang, and Mala Ganiger. 2020. "Reduction of Phakopsora Pachyrhizi Infection on Soybean through Host- and Spray-induced Gene Silencing." *Molecular Plant Pathology* 21(6):794–807. doi: 10.1111/mpp.12931.

Hu, Taishan, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. 2021. "Next-Generation Sequencing Technologies: An Overview." *Human Immunology* 82(11):801–11. doi: 10.1016/j.humimm.2021.02.012.

Juroszek, P., P. Racca, S. Link, J. Farhumand, and B. Kleinhenz. 2020. "Overview on the Review Articles Published during the Past 30 Years Relating to the Potential Climate

Change Effects on Plant Pathogens and Crop Disease Risks." *Plant Pathology* 69(2):179–93. doi: 10.1111/ppa.13119.

Karst, Søren M., Ryan M. Ziels, Rasmus H. Kirkegaard, Emil A. Sørensen, Daniel McDonald, Qiyun Zhu, Rob Knight, and Mads Albertsen. 2021. "High-Accuracy Long-Read Amplicon Sequences Using Unique Molecular Identifiers with Nanopore or PacBio Sequencing." *Nature Methods* 18(2):165–69. doi: 10.1038/s41592-020-01041-y.

Kartikeyan, Punitha, and Gyanesh Shrivastava. 2021. *Review on Emerging Trends in Detection of Plant Diseases Using Image Processing with Machine Learning*. Vol. 174.

Karunarathna, Samantha C., Sajeewa S. N. Maharachchikumbura, Hiran A. Ariyawansa, Belle Damodara Shenoy, and Rajesh Jeewon. 2021. "Editorial: Emerging Fungal Plant Pathogens." *Frontiers in Cellular and Infection Microbiology* 11.

Kashiwa, Takeshi, Miguel Angel Lavilla, Antonio Diaz Paleo, Antonio Juan Gerardo Ivancovich, and Naoki Yamanaka. 2021. " The Use of Detached Leaf Inoculation for Selecting Cercospora Kikuchii Resistance in Soybean Genotypes ." *PhytoFrontiers*^{TM} 1(4):250–57. doi: 10.1094/phytofr-01-21-0002-ta.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew

Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R.

Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409(6822):860–921. doi: 10.1038/35057062.

Leakey, C. L. A. 1964. “Dactuliophora, a New Genus of Mycelia Sterilia from Tropical Africa.” *Transactions of the British Mycological Society* 47(3):341-IN10. doi: 10.1016/S0007-1536(64)80006-1.

Lee, Steven B., De Etta K. Mills, Stephen A. Morse, Steven E. Schutzer, Bruce Budowle, and Paul Keim. 2019. “Education and Training in Microbial Forensics.” Pp. 473–95 in *Microbial Forensics*. Elsevier.

Levy, C., G. S. Mahuku ?, J. R. Tattersfield $ And, and Desiri~e L. Cole. 1990. *Method of Assessment of Red Leaf Blotch on Soybeans Used to Evaluate Cultivar Susceptibility and Chemical Control*.

Lin, Feng, Sushil Satish Chhapekar, Caio Canella Vieira, Marcos Paulo Da Silva, Alejandro Rojas, Dongho Lee, Nianxi Liu, Esteban Mariano Pardo, Yi Chen Lee, Zhimin Dong, Jose Baldin Pinheiro, Leonardo Daniel Ploper, John Rupe, Pengyin Chen, Dechun Wang, and Henry T. Nguyen. 2022. "Breeding for Disease Resistance in Soybean: A Global Perspective." *Theoretical and Applied Genetics* 135(11):3773–3872.

Loman, Nicholas J., Raju V Misra, Timothy J. Dallman, Chrystala Constantinidou, Saheer E. Gharbia, John Wain, and Mark J. Pallen. 2012. "Performance Comparison of Benchtop High-Throughput Sequencing Platforms." *Nature Biotechnology* 30(5):434–39. doi: 10.1038/nbt.2198.

Madden, L. V., and M. Wheelis. 2003. "The Threat of Plant Pathogens as Weapons against U.S. Crops." *Annual Review of Phytopathology* 41(1):155–76. doi: 10.1146/annurev.phyto.41.121902.102839.

Madina, Mst Hur, Parthasarathy Santhanam, Yanick Asselin, Rajdeep Jaswal, and Richard R. Bélanger. 2022. "Progress and Challenges in Elucidating the Functional Role of Effectors in the Soybean-Phytophthora Sojae Interaction." *Journal of Fungi* 9(1):12. doi: 10.3390/jof9010012.

Meyer, Seth. 2021. *The Outlook for U.S. Agriculture-2021. Building on Innovation: A Pathway to Resilience*.

Morse, Stephen A. 2015. "Pathogen Security-Help or Hindrance?" *Frontiers in Bioengineering and Biotechnology* 2. doi: 10.3389/fbioe.2014.00083.

Morse, Stephen A., Bruce Budowle, and Steven E. Schutzer. 2020. "Microbial Forensics." Pp. 497–500 in *Microbial Forensics*. Elsevier.

Morse, Stephen A., and Bernard R. Quigley. 2020. "Select Agent Regulations." Pp. 425–39 in *Microbial Forensics*. Elsevier.

Morse, Stephen A., and Elizabeth Weirich. 2011. “Select Agent Regulations.” Pp. 199–220 in *Microbial Forensics*. Elsevier.

Murithi, Harun M., Michelle Pawlowski, Tizazu Degu, Deresse Hunde, Molla Malede, Tonny Obua, Hapson Mushoriwa, Danny Coyne, Phinehas Tukamuhabwa, and Glen L. Hartman. 2022. “Evaluation of Soybean Entries in the Pan-African Trials for Response to Coniothyrium Glycines, the Cause of Red Leaf Blotch.” *Plant Disease* 106(2):535–40. doi: 10.1094/PDIS-05-21-1017-RE.

Murrin, Suzanne. 2018. *Entities Generally Met Federal Select Agent Program Internal Inspection Requirements But CDC Could Do More To Improve Effectiveness*.

National Research Council. 2010. *Sequence-Based Classification of Select Agents*. Vol. 1. 1st ed. Washington : The National Academic Press.

Neves, Danilo L., Aiqin Wang, Japheth D. Weems, Heather M. Kelly, Daren S. Mueller, Mark Farman, and Carl A. Bradley. 2022. “Identification of Septoria Glycines Isolates from Soybean with Resistance to Quinone Outside Inhibitor Fungicides.” *Plant Disease* 106(10):2631–37. doi: 10.1094/PDIS-08-21-1836-RE.

Oklahoma State University. 2023. “Oklahoma State University: Select Agent Training.” *Biosafety Office*. Retrieved February 20, 2023 (https://research.okstate.edu/research-compliance/ibc/select-agent-training.html).

Pena-Zuniga, Lizbeth Daniela. 2020. “EDNA-HOST: Detection of Global Plant Viromes Using High Throughput Sequencing.” PhD, Oklahoma State University, Stillwater.

Proano-Cuenca, Fernanda, Andrés S. Espíndola, and Carla Garzon. 2022. “Detection of Phytophthora, Pythium, Globisporangium, Hyaloperonospora and Plasmopara Species in High-Throughput Sequencing Data by in Silico and in Vitro Analysis Using Microbe Finder (MiFi®).” *PhytoFrontiers^{TM}* 1–73. doi: 10.1094/PHYTOFR-04-22-0039-FI.

Rang, Franka J., Wigard P. Kloosterman, and Jeroen de Ridder. 2018. “From Squiggle to Basepair: Computational Approaches for Improving Nanopore Sequencing Read Accuracy.” *Genome Biology* 19(1):90. doi: 10.1186/s13059-018-1462-9.

Ristaino, Jean B., Pamela K. Anderson, Daniel P. Bebber D, Kate A. Brauman E, Nik J. Cunniffe, Nina V Fedoroff, Cambria Finegold, Karen A. Garrett, Christopher A. Gilligan, Christopher M. Jones K, Michael D. Martin, Graham K. Macdonald, Patricia Neenan, Angela Records, David G. Schmale, Laura Tateosian, and Qingshan Wei. 2021. “The Persistent Threat of Emerging Plant Disease Pandemics to Global Food Security.” *PNAS* 118(23):1–9. doi: 10.1073/pnas.2022239118/-/DCSupplemental.

Rodriguez, Maria C., Francisco Sautua, Mercedes Scandiani, Marcelo Carmona, and Sebastián Asurmendi. 2021. “Current Recommendations and Novel Strategies for Sustainable Management of Soybean Sudden Death Syndrome.” *Pest Management Science* 77(10):4238–48. doi: 10.1002/ps.6458.

Ronaghi, Mostafa, Mathias Uhlén, and Pål Nyrén. 1998. “A Sequencing Method Based on Real-Time Pyrophosphate.” *Science* 281(5375):363–65. doi: 10.1126/science.281.5375.363.

Roth, Mitchell G., Richard W. Webster, Daren S. Mueller, Martin I. Chilvers, Travis R. Faske, Febina M. Mathew, Carl A. Bradley, John P. Damicone, Mehdi Kabbage, and Damon L. Smith. 2020. “Integrated Management of Important Soybean Pathogens of the United States in Changing Climate” edited by N. Walker. *Journal of Integrated Pest Management* 11(1). doi: 10.1093/jipm/pmaa013.

Saah, Alfred J., and Donald R. Hoover. 1997. “‘Sensitivity’ and ‘Specificity’ Reconsidered: The Meaning of These Terms in Analytical and Diagnostic Settings.” *Annual Internal Medicine* 126:91–94.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences* 74(12):5463–67. doi: 10.1073/pnas.74.12.5463.

Sankaran, Sindhuja, Ashish Mishra, Reza Ehsani, and Cristina Davis. 2010. “A Review of Advanced Techniques for Detecting Plant Diseases.” *Computers and Electronics in Agriculture* 72(1):1–13. doi: 10.1016/j.compag.2010.02.007.

Schoch, Conrad L., Stacy Ciufo, Mikhail Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O’Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P. Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi. 2020. “NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools.” *Database : The Journal of Biological Databases and Curation* 2020. doi: 10.1093/database/baaa062.

Shendure, Jay, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. 2017. “DNA Sequencing at 40: Past, Present and Future.” *Nature* 550(7676):345–53. doi: 10.1038/nature24286.

Shreffler, Jacob, and Martin R. Huecker. 2022. “Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios.” *StatPearls*.

Šimundić, Ana-Maria. 2009. “Measures of Diagnostic Accuracy: Basic Definitions.” *EJIFCC* 19(4):203.

Sinclair, J. B. 1989. “Threats to Soybean Production in the Tropics: Red Leaf Blotch and Leaf Rust.” *Plant Disease* 73(7):604–6.

Sjödin, Andreas, Tina Broman, Öjar Melefors, Gunnar Andersson, Birgitta Rasmusson, Rickard Knutsson, and Mats Forsman. 2013. “The Need for High-Quality Whole-Genome Sequence Databases in Microbial Forensics.” *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 11(S1):S78–86. doi: 10.1089/bsp.2013.0007.

Stein, Lincoln D. 2010. “The Case for Cloud Computing in Genome Informatics.” *Genome Biology* 11(5):207. doi: 10.1186/gb-2010-11-5-207.

Stewart, Robert B. 1957. “An Undescribed Species of Pyrenochaeta on Soybean.” *Mycologia* 49(1):115–17. doi: 10.1080/00275514.1957.12024619.

Stobbe, A. H., W. L. Schneider, P. R. Hoyt, and U. Melcher. 2014. “Screening Metagenomic Data for Viruses Using the E-Probe Diagnostic Nucleic Acid Assay.” *Phytopathology* 104(10):1125–29. doi: 10.1094/PHYTO-11-13-0310-R.

Stobbe, Anthony H., Jon Daniels, Andres S. Espindola, Ruchi Verma, Ulrich Melcher, Francisco Ochoa-Corona, Carla Garzon, Jacqueline Fletcher, and William Schneider. 2013. “E-Probe Diagnostic Nucleic Acid Analysis (EDNA): A Theoretical Approach for Handling of next Generation Sequencing Data for Diagnostics.” *Journal of Microbiological Methods* 94(3):356–66. doi: 10.1016/j.mimet.2013.07.002.

Taguchi-Shiobara, Fumio, Kenichiro Fujii, Takashi Sayama, Kaori Hirata, Shin Kato, Akio Kikuchi, Koji Takahashi, Masao Iwahashi, Chiaki Ikeda, Kazuma Kosuge, Katsunori Okano, Masahiro Hayasaka, Yasutaka Tsubokura, and Masao Ishimoto. 2019. “Mapping Versatile QTL for Soybean Downy Mildew Resistance.” *Theoretical and Applied Genetics* 132(4):959–68. doi: 10.1007/s00122-018-3251-y.

Tooley, Paul W. 2017. “Development of an Inoculation Technique and the Evaluation of Soybean Genotypes for Resistance to Coniothyrium Glycines.” *Plant Disease* 101(8):1411–16. doi: 10.1094/PDIS-09-16-1373-RE.

United States Department of Agriculture. 2023. *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2022 Summary*.

Waage, J. K., and J. D. Mumford. 2008. “Agricultural Biosecurity.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1492):863–76. doi: 10.1098/rstb.2007.2188.

Watanabe, Daisuke, Tomás Losák, and Johann Vollmann. 2018. “From Proteomics to Ionomics: Soybean Genetic Improvement for Better Food Safety.” *Genetika* 50(1):333–50. doi: 10.2298/GENSR1801333W.

Widyasari, Kristin, Mazen Alazem, and Kook-Hyung Kim. 2020. “Soybean Resistance to Soybean Mosaic Virus.” *Plants* 9(2):219. doi: 10.3390/plants9020219.

Zhan, Jie, Irena Twardowska, Siqi Wang, Shuhe Wei, Yanqiu Chen, and Mihajlov Ljupco. 2019. “Prospective Sustainable Production of Safe Food for Growing Population Based on the Soybean (Glycine Max L. Merr.) Crops under Cd Soil Contamination Stress.” *Journal of Cleaner Production* 212:22–36. doi: 10.1016/j.jclepro.2018.11.287.

Zhou, Jing, and Ioannis E. Tzanetakis. 2020. “Soybean Vein Necrosis Orthotospovirus Can Move Systemically in Soybean in the Presence of Bean Pod Mottle Virus.” *Virus Genes* 56(1):104–7. doi: 10.1007/s11262-019-01715-6.

CHAPTER III

MOLECULAR IDENTIFICATION AND MULTILOCUS PHYLOGENETIC ANALYSIS OF *CONIOTHYRIUM GLYCINES* ISOLATED FROM SOYBEAN FIELDS IN ZAMBIA AND ZIMBABWE

Abstract

Red leaf blotch (RLB) of soybeans, caused by the U.S. select agent *Coniothyrium glycines*, is a severe disease distributed across sub-Saharan Africa. Although *C. glycines* is a potential threat to agriculture, information about its biology and genetic diversity is limited. Morphological identification of the pathogen remains complicated due to the lack of distinctive features, and molecular methods are not commonly used. This study used two nuclear and two mitochondrial fungal molecular markers to identify fourteen samples from symptomatic soybean fields in Zambia and Zimbabwe and assess their phylogeny using a multilocus phylogenetic approach. All the samples were confirmed as *C. glycines* based on the percentage of identity and query coverage after comparing them with the GenBank nucleotide database using BLAST. The multilocus phylogenetic analyses based on the large ribosomal subunit (LSU), Internal transcribed spacer (ITS) region, β-tubulin gene (BTUB), and the translation elongation factor 1-alpha gene (TEF) revealed that isolates from matching locations form a monophyletic clade. In addition, results suggest the movement of the fungus across borders since some isolates from the two countries share a common ancestor. These findings provide a tool for *C. glycines* identification and improve the understanding of the pathogen biology and diversity, impacting its field diagnosis, detection, and biosecurity measures.

1. Introduction

Soybean [*Glycine max* (L.) Merr.] is one of the main crops worldwide since it contributes to 70% and 30% of the global supply of plant-based protein meal and plant-based oil, respectively (Lin et al. 2022). Even though 96% of the worldwide soybean production comes from ten countries, mainly Brazil, Argentina, and the U.S., African countries have the potential to become significant soybean producers. Africa's soybean production has been increasing since 1961, with an average rate of 7% per year (Cornelius and Goldsmith 2019). In 2017, Africa produced over three million metric tons of soybean (1% of global soybean production). South Africa, Nigeria, and Zambia are the top three producers on the continent (Murithi et al. 2022). The high instability of the soybean market, disease susceptibility, and food safety importance require stakeholders to minimize soybean yield losses (Lin et al. 2022; Roth et al. 2020). However, African soybean production is diminished by several diseases, including red leaf blotch (RLB) of soybeans, as a significant pathogen (Hartman et al. 1987; Hartman and Murithi 2022; Lin et al. 2022).

RLB is caused by the fungal pathogen *Coniothyrium glycines* (R.B. Stewart) Verkley & Gruyter, which was first reported in 1957 in Ethiopia, and since then, its incidence has been increasing within the Sub-Saharan African region (de Gruyter et al. 2013; Stewart 1957). Based on morphological and phylogenetic analyses, this fungus has undergone multiple taxonomic revisions over the years. First, it was classified in the genus *Pyrenochaeta* (Stewart 1957) and then as a member of the genus *Dactuliophora* (Leakey 1964). In 2002, it was renamed *Phoma glycinicola* (de Gruyter and Boerema 2002) and in 2013 as *Coniothyrium glycines* (de Gruyter et al. 2013). The constant changes in its taxonomy, combined with the lack of molecular and biological information, have complicated the development of diagnostic tools (Hartman et al. 2011; Hartman and Murithi 2022).

Identifying fungal plant pathogens at a species level based on morphology has been challenging throughout the years due to phenotypic plasticity, hybridization, cryptic speciation, and convergent evolution (Raja et al. 2017). Consequently, DNA sequence-based approaches have been used to differentiate species with similar or identical morphological traits (Hibbett and Taylor 2013). DNA barcoding is a commonly used technique that identifies species based on one or multiple molecular markers (Naranjo-Ortiz and Gabaldón 2019). This approach compares an unknown sequence with a reference sequence database, such as the GenBank of the National Center for Biotechnology Information (NCBI). The identification is performed based on the similarity between both sequences (Hibbett 2016).

Phylogenetic analysis is another approach to identify unknown samples by clustering them within an evolutionary context with other homologous sequences (Raja et al. 2017). Additionally, phylogeny aids the understanding of how species are connected through time by rebuilding their historical path of evolution (Kapli, Yang, and Telford 2020). These analyses have served not only for systematics but also to study the progression of morphological features, genes, ecology, and diversification in time (James et al. 2020). However, the main drawback of both approaches is that they depend highly on the quality and availability of DNA sequences and prior knowledge of the organisms (Kapli et al. 2020).

This chapter objective was to identify fourteen isolates from Zambia and Zimbabwe soybean fields by sequencing four DNA molecular markers: 1) ITS: Internal Transcribed Spacer, 2) LSU: Large Ribosomal Subunit, 3) BTUB: beta-tubulin gene, and 4) TEF: translation elongation factor 1-alpha. A multilocus phylogenetic analysis using four loci was made to infer the evolution and differentiation between the isolates.

2. Materials and methods

2.1. Isolates and pure culture

Fourteen *Coniothyrium glycines* isolates were received from the United States Department of Agriculture – Agricultural Research Service (USDA-ARS). Fungal cultures were isolated from infected soybean plants (*Glycine max* L. Merr.) in Zambia (ZB) and Zimbabwe (ZW) fields from 2001 to 2006 (Table III-1). The isolates growing on Potato Dextrose Agar (PDA) arrived at the Institute of Biosecurity and Microbial Forensics (IBMF) at Oklahoma State University.

Table III-1: C. glycines isolates collected from two African countries (Zambia and Zimbabwe) from 2001 to 2006. The name of the collector is shown if available.

Isolate	Source	Isolation	Collected	Isolate	Source	Isolation	Collected
IMI294986[a]	Zambia	03/2005	J.M. Waller	**Pg36**	Zambia	05/2005	-
Pg1[b]	Zimbabwe	04/2001	C. Levy	**Pg42**	Zambia	05/2005	-
Pg21	Zimbabwe	03/2005	C. Levy	**Pg43**	Zamia	05/2005	J. Tichagwa
Pg23	Zimbabwe	03/2005	-	**Pg44**	Zambia	05/2005	-
Pg31	Zambia	05/2005	J. Tichagwa	**Pg45**	Zambia	05/2005	-
Pg34	Zambia	05/2005	-	**RA1**	Zimbabwe	05/2006	-
Pg35	Zambia	05/2005	-	**RA106**	Zimbabwe	05/2006	-

[a]: Representative specimen CBS124455. [b]: Representative specimen CBS124141

Upon arrival, cultures were transferred to 2% water agar (WA) media with 100 ppm of ampicillin and streptomycin and incubated at room temperature (~22°C) in the dark. After one week, a single hypha from each plate was obtained following the single hyphal tip technique and placed over cV8 (clarified V8) agar, supplemented with 100 ppm of ampicillin and streptomycin. This process was performed in triplicate. Plates were incubated for two weeks at room temperature in the dark, and their macroscopic and microscopic characteristics were recorded. Pure cultures were established once all three plates showed the same phenotypic features.

Autoclaved filter paper placed over cV8 agar and sterile mineral oil tubes were used to achieve the long-term storage of *C. glycines* isolates. Three agar plugs were placed over the filter paper, and once a complete growth was reached, the filter paper was dried into a desiccator and stored at room temperature within Ziploc bags. Several agar plugs with mycelium and sclerotia were placed inside the tubes and stored at room temperature for the mineral oil. *C. glycines* working cultures were kept growing on cV8 agar supplemented with 100 ppm of ampicillin and streptomycin, and every 21-35 days, a new subculture was made.

2.2. DNA extraction

Fungal isolates were grown on cellophane overlaying cV8 agar for 15-21 days at room temperature in the dark. Mycelium and sclerotia were harvested by scraping the plates with a sterile metal spatula and stored at -80°C until used. DNeasy® Plant Mini kit (Qiagen, Hilden, Germany) was used to extract the DNA of each isolate, starting with 80-100 mg of mycelium/sclerotia and following the manufacturer's protocol. DNA concentration and quality were measured using a NanoDrop® ND-2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA).

2.3. DNA amplification and Sanger sequencing

Two non-coding regions (ITS: Internal Transcribed Spacer region, LSU: Large Ribosomal Subunit 28S) and two coding regions (BTUB: β-tubulin gene, TEF: Translation Elongation Factor 1-$\boldsymbol{\alpha}$) were amplified using previously reported PCR primers and conditions (Table III-2 and

Table III-3). For all loci, reactions were performed in 25uL containing 1.25uL of each 5uM primer, 12.5uL of 2X GoTaq Green Master Mix (Promega, Madison, WI, USA), 2uL of DNA (25ng/uL), and 8uL of nuclease free water.

Table III-2: Primers used for PCR amplification and sequencing of four loci: ITS (Internal Transcribed Spacer), LSU (Large ribosomal subunit 28S), BTUB ($\boldsymbol{\beta}$-tubulin gene), and TEF (Translation elongation factor 1-$\boldsymbol{\alpha}$).

Locus	Primer name	Primer sequence (5'-3')	Reference
ITS	V9G	TTACGTCCCTGCCCTTTGTA	De Hoog & Van den Ended (1998)
	ITS4	TCCTCCGCTTATTGATATGC	White, et al. (1990)
LSU	LR5	TCCTGAGGAAACTTCG	Vilgalys & Hester (1990)
	LROR	GTACCCGCTGAACTTAAGC	Rehner & Samuels (1994)
BTUB	T1	AACATGCGTGAGATTGTAAGT	O'Donnell & Cigelnik (1997)
	B-Sandy-R	GCRCGNGGVACRTACTTGTT	Stukenbrock, et al. (2012)
TEF	EF1-728F	CATCGAGAAGTTCGAGAAGG	Carbone & Kohn (1999)
	EF-2	GGARGTACCAGTSATCATGTT	O'Donnell, et al. (1998)

Table III-3: PCR conditions for the amplification of four loci: ITS (Internal Transcribed Spacer), LSU (Large ribosomal subunit 28S), BTUB (β-tubulin gene), and TEF (Translation elongation factor 1-α).

PCR Conditions					
Locus: ITS		**Locus: LSU**		**Loci: BTUB & TEF**	
Temperature – Time	**Cycles**	**Temperature – Time**	**Cycles**	**Temperature – Time**	**Cycles**
94°C – 5 min	X1	94°C – 5 min	X1	96°C – 2 min	X1
94°C – 30 s	X35	94°C – 45 s	X35	96°C – 45 s	X40
48°C – 30 s		48°C – 45 s		52°C – 30 s	
72°C – 1 min		72°C – 2 min		72°C – 90 s	
72°C – 7 min	X1	72°C – 7 min	X1	72°C – 2 min	X1
4°C – Hold		4°C – Hold		4°C – Hold	

After amplification, electrophoresis was performed using 5 μL of PCR product loaded into a 1.5% agarose gel (VWR Life Sciences) and stained with 3 μL of SYBR® Safe DNA gel stain (Invitrogen, Waltham, MA, USA). The electrophoresis was run at 95 V for one hour, and the gel was visualized using the Molecular Imager® Gel Doc XR+ (Bio-Rad, Hercules, CA, USA). PCR products that showed a clear band were treated before Sanger sequencing with the enzymatic purification kit illustra™ ExoProStar™ (Millipore® Sigma, Darmstadt, Germany), following the manufacturer's protocol. Sanger sequencing was performed with the same PCR primers (Table III-2) on an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems, Waltham, MA, USA) at the DNA and Protein Core Facility - Oklahoma State University.

2.4. Sequence identification

Unipro UGENE v40.1 (Okonechnikov, Golosova, and Fursov 2012) was used for the manual edition and to build the consensus sequences for each locus per isolate. Then, each isolate was identified by comparing the consensus sequence of each locus with the nucleotide database in NCBI using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990).

2.5. Maximum likelihood and Bayesian inference of multilocus phylogeny

Consensus sequences for each locus were aligned separately using the ClustalW algorithm with default parameters integrated into MEGA11: Molecular Evolutionary Genetics Analysis v11.0.1 (Tamura, Stecher, and Kumar 2021). Each alignment was manually examined and adjusted. Nucleotide substitution models for each alignment were defined using MEGA11, and the best one was selected based on the lowest Bayesian Information Criterion Score (BIC). Phylogenetic analyses were performed for each locus and a concatenated dataset, using Bayesian inference (BI) and maximum likelihood (ML) approaches. The concatenated dataset was built using SequenceMatrix v1.8.1 (Vaidya, Lohman, and Meier 2011).

BI analyses were performed using MrBayes v3.2.7 (Ronquist et al., 2012), where the best nucleotide substitution model for each locus was used, and the sampling and diagnostic frequency was set to 1000. The number of generations was assigned to a minimum of one million, where an assessment of convergence and quality of the MCMC process was performed based on the average standard deviation of split frequencies (STDEV), the potential scale reduction factors (PSRF), and the effective sampling size (ESS). The analyses stopped once a run reached values of STDEV < 0.01, PSRF ~ 1, and ESS > 100; if not, the number of generations were increased until those thresholds were met. By default, the burn-in was set to 25%, and the 50% majority rule estimated each branch's posterior probabilities (PP).

The ML analyses were performed using RaxML v8.2.12 (Stamatakis, 2014), where a rapid bootstrap analysis and searching for the best-scoring tree were performed. Each run started by setting a fixed seed number for both the rapid bootstrapping and the parsimony inferences. Additionally, 5,000 bootstraps were set to run with a GTRGAMMA model to obtain each branch's bootstrap (BS) support.

BI and ML output trees were visualized and edited using FigTree v1.4.4 (https://github.com/rambaut/figtree) and TreeGraph v2.15.0-887 (Stöver and Müller 2010). Both

ML and BI analyses were performed within the High-Performance Computing (HPC) Center at Oklahoma State University. Five organisms served as the outgroup taxa for both studies, and available reference nucleotide sequences of *Coniothyrium glycines* were also included (Table III-4).

Table III-4: Reference and outgroup isolates used in this study and their GenBank accession numbers.

Taxa	Strain	GenBank Accession Numbers			
		ITS	**LSU**	**BTUB**	**TEF**
***Coniothyrium glycines*[r]**	CBS 124141	KF251211.1	KF251714.1	KF252702.1	KF253167.1
***Coniothyrium glycines*[r]**	CBS 124455	JF740184.1	GQ387597.1	-	-
***Didymella exigua*[o]**	CBS 183.55	MH857436.1	MH868977.1	GU237525.1	KR184187.1
***Coniothyrium palmarum*[o]**	CBS 400.71	AY720708.1	EU754153.1	KT389792.1	DQ677903.1
***Pyrenochaeta nobilis*[o]**	CBS 407.76	MH860989.1	MH872759.1	KT389845.1	MF795880.1
***Phoma herbarum*[o]**	CBS 567.63	MH858359.1	MH869982.1	AY749027.1	-
***Leptosphaeria doliolum*[o]**	CBS 505.75	JF740205.1	GU301827.1	JF740144.1	GU349069.1

r: Reference nucleotide sequences. o: Outgroup nucleotide sequences.

3. Results

3.1. Molecular identification of isolates by Sanger sequencing

The extracted DNA had an average concentration of ~750 ng/uL, and their quality, based on the A260/280 and A260/230 ratios, ranged between 1.60-2.10 and 1.70-2.20, respectively. All the fourteen (14) isolates were confirmed as *Coniothyrium glycines* based on the four loci (Table III-5). The ITS amplicon had an average size of 670 bp and showed an overall identity percentage and query coverage of > 99.5% and > 80%, respectively. In most cases, there was a correlation between the origin of each isolate and the source of the reference isolate sequences it hit. Isolates obtained from Zambia (Pg31, Pg34, and Pg36) match the *C. glycines* Zimbabwe reference isolate. The LSU amplicon had an average size of 850 bp, and each isolate matched sequences of both reference isolates with the same alignment metrics (identity > 99% and query coverage > 99.8%). The average amplicon sizes for the BTUB and TEF were 320 bp and 480 bp, respectively. Unfortunately, for both loci, the available sequences in the NCBI nucleotide database belonged to one reference isolate (CBS124141) obtained from Zimbabwe. The isolates hit the BTUB with an identity > 97% and a query coverage > 50%, while they hit TEF with an identity > 99% and query coverage > 85%. The mentioned identity percentages and query coverages are associated alignment metrics between the query and the sequences of the reference *C. glycines* isolates available in the NCBI nucleotide database.

Table III-5: Molecular identification of *C. glycines* isolates based on four loci, ITS: Internal Transcribed Spacer, LSU: Large Ribosomal Subunit, BTUB: beta-tubulin gene, and TEF: transcribed elongation factor 1-alpha. The color reflects the reference isolate to which each isolate had the best hit.

Isolate	Source	ITS	LSU	BTUB[a]	TEF[a]
RA1	Zimbabwe	🟥	🟥🟦	🟥	🟥
RA106	Zimbabwe	🟥	🟥🟦	🟥	🟥
IMI294986	Zambia	🟦	🟥🟦	🟥	🟥
Pg1	Zimbabwe	🟥	🟥🟦	🟥	🟥
Pg21	Zimbabwe	🟥	🟥🟦	🟥	🟥
Pg23	Zimbabwe	🟥	🟥🟦	🟥	🟥
Pg31	Zambia	🟥	🟥🟦	🟥	🟥
Pg34	Zambia	🟥	🟥🟦	🟥	🟥
Pg35	Zambia	🟦	🟥🟦	🟥	🟥
Pg36	Zambia	🟥	🟥🟦	🟥	🟥
Pg42	Zambia	🟦	🟥🟦	🟥	🟥
Pg43	Zambia	🟦	🟥🟦	🟥	🟥
Pg44	Zambia	🟦	🟥🟦	🟥	🟥
Pg45	Zambia	🟦	🟥🟦	🟥	🟥

Blue: CBS124455 (Zambia). Red: CBS124141 (Zimbabwe). a: Only CBS124141 sequences are available.

The nucleotide sequences generated in this project were submitted to the NCBI GenBank nucleotide database with the accession numbers in Table III-6.

Table III-6: GenBank accession numbers for the sequences generated in the present chapter and used for molecular identification and phylogenetic analysis.

Isolate	ITS	LSU	BTUB	TEF
RA1	ON230304.1	ON231275.1	ON871592.1	ON871606.1
RA106	ON230317.1	ON231288.1	ON871605.1	ON871619.1
IMI294986	ON230316.1	ON231287.1	ON871604.1	ON871618.1
Pg1	ON230315.1	ON231286.1	ON871603.1	ON871617.1
Pg21	ON230306.1	ON231277.1	ON871594.1	ON871607.1
Pg23	ON230305.1	ON231276.1	ON871593.1	ON871608.1
Pg31	ON230307.1	ON231278.1	ON871595.1	ON871609.1
Pg34	ON230308.1	ON231279.1	ON871596.1	ON871610.1
Pg35	ON230312.1	ON231283.1	ON871600.1	ON871611.1
Pg36	ON230309.1	ON231280.1	ON871597.1	ON871612.1
Pg42	ON230310.1	ON231281.1	ON871598.1	ON871613.1
Pg43	ON230313.1	ON231284.1	ON871601.1	ON871614.1
Pg44	ON230314.1	ON231285.1	ON871602.1	ON871615.1

Isolate	ITS	LSU	BTUB	TEF
Pg45	ON230311.1	ON231282.1	ON871599.1	ON871616.1

3.2. Multilocus phylogenetic analysis

Each locus (ITS, LSU, BTUB, and TEF) was analyzed separately and then combined to perform a multilocus phylogenetic analysis. For the individual and the concatenated dataset, maximum likelihood (ML), and Bayesian inference (BI) approaches were used, obtaining the bootstrap branch support (BS) and posterior probabilities (PP), respectively. Based on the Multiple Sequence Alignment (MSA), the best nucleotide substitution model for the ITS, LSU, and BTUB loci was the Kimura two-parameter with Gamma Distribution (K2+G). In contrast, the best model for the TEF was the Kimura two-parameter (K2). The ITS-based tree grouped all the isolates with the reference sequences of *C. glycines* (BS=100 and PP=0.99) and showed the formation of three well-supported clades (BS > 85 and PP > 0.95). However, ITS-based phylogeny could not infer the origin of almost all the Zimbabwean isolates. Hence polytomies were formed, and the inner nodes of the tree were not well supported (Figure III-1).

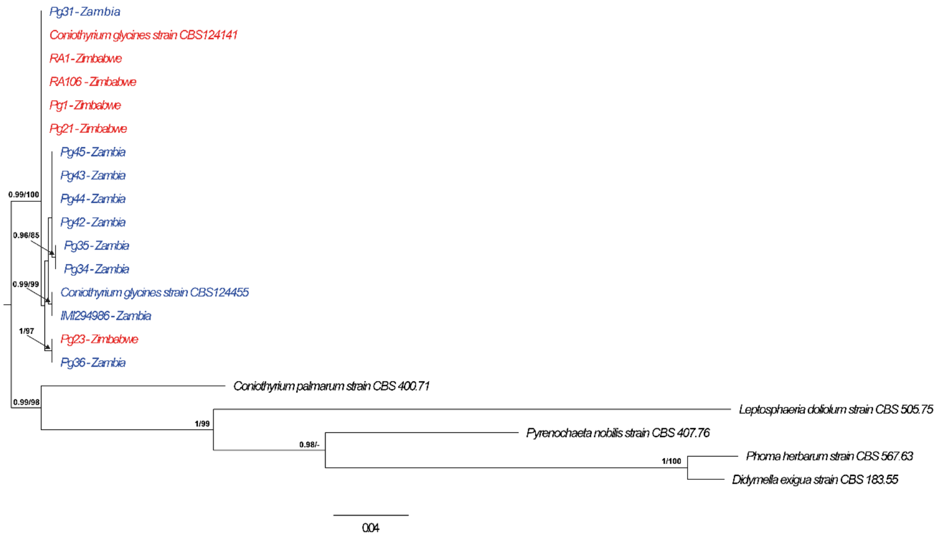

Pg31 - Zambia
Coniothyrium glycines strain CBS124141
RA1 - Zimbabwe
RA106 - Zimbabwe
Pg1 - Zimbabwe
Pg21 - Zimbabwe
Pg45 - Zambia
Pg43 - Zambia
Pg44 - Zambia
Pg42 - Zambia
Pg35 - Zambia
Pg34 - Zambia
Coniothyrium glycines strain CBS124455
IMI294986 - Zambia
Pg23 - Zimbabwe
Pg36 - Zambia
0.99/100
0.96/85
0.99/99
1/97
Coniothyrium palmarum strain CBS 400.71
0.99/98
Leptosphaeria doliolum strain CBS 505.75
1/99
Pyrenochaeta nobilis strain CBS 407.76
0.98/-
Phoma herbarum strain CBS 567.63
1/100
Didymella exigua strain CBS 183.55
0.04

Figure III-1: Phylogenetic tree of *C. glycines* isolates based on the ITS region. The tree's topology was built using K2+G nucleotide substitution model and ML approach. PP (left) was estimated with one million generations, and BS (right) with 5000 bootstraps. Only the PP >0.90 and BS>80 are displayed above each branch. Zambia and Zimbabwe isolates are colored blue and red, respectively.

Similarly, the LSU-based tree grouped all the analyzed isolates with *C. glycines* reference sequences (BS=100 and PP=1). Though there was not enough information in this locus to suggest a common ancestor between isolates, the tree's topology showed an overall polytomy without the formation of defined clades (Figure III-2).

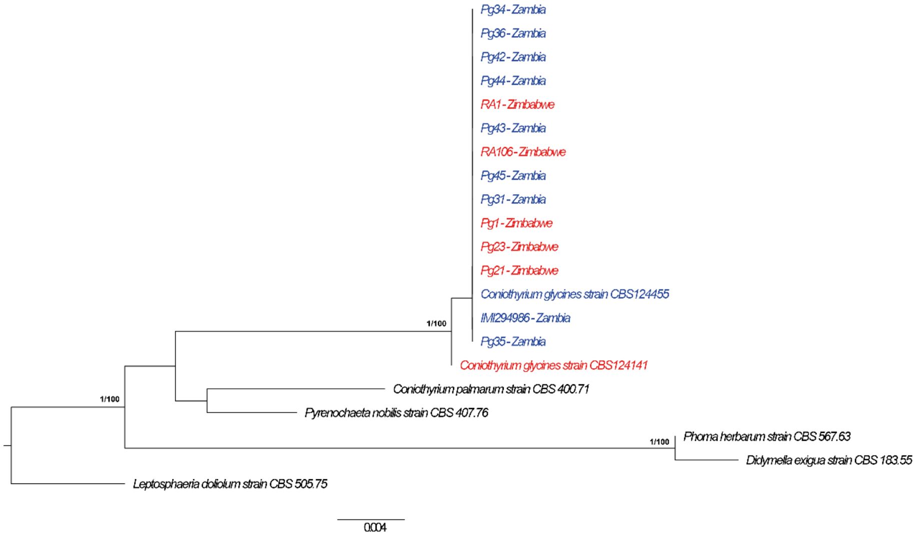

Pg34-Zambia
Pg36-Zambia
Pg42-Zambia
Pg44-Zambia
RA1-Zimbabwe
Pg43-Zambia
RA106-Zimbabwe
Pg45-Zambia
Pg31-Zambia
Pg1-Zimbabwe
Pg23-Zimbabwe
Pg21-Zimbabwe
Coniothyrium glycines strain CBS124455
IMI294986-Zambia
Pg35-Zambia
1/100
Coniothyrium glycines strain CBS124141
Coniothyrium palmarum strain CBS 400.71
1/100
Pyrenochaeta nobilis strain CBS 407.76
1/100
Phoma herbarum strain CBS 567.63
Didymella exigua strain CBS 183.55
Leptosphaeria doliolum strain CBS 505.75
0.004

Figure III-2: Phylogenetic tree of *C. glycines* isolates based on the LSU region. The tree's topology was built using K2+G nucleotide substitution model and ML approach. PP (left) was estimated with one million generations, and BS (right) with 5000 bootstraps. Only the PP >0.90 and BS>80 are displayed above each branch. Zambia and Zimbabwe isolates are colored blue and red, respectively.

Beta-tubulin gene (BTUB) phylogenetic analysis showed comparable results to the ITS and LSU analyses. Even though all the isolates grouped with the *C. glycines* reference sequence (BS=99 and PP=0.98), there was not well inner node support, and multiple polytomies were formed (Figure III-3). The TEF-based tree clustered all the samples (BS=100 and PP=1) with the *C. glycines* reference isolate and allowed the formation of four well-supported clades (BS>88 and PP>0.91). Some isolates from the exact geographical origin shared a common ancestor, but at the same time, some clades included isolates from both locations (Zambia and Zimbabwe). As with the previous trees, a polytomy was seen, and some inner nodes were not well-supported (Figure III-4).

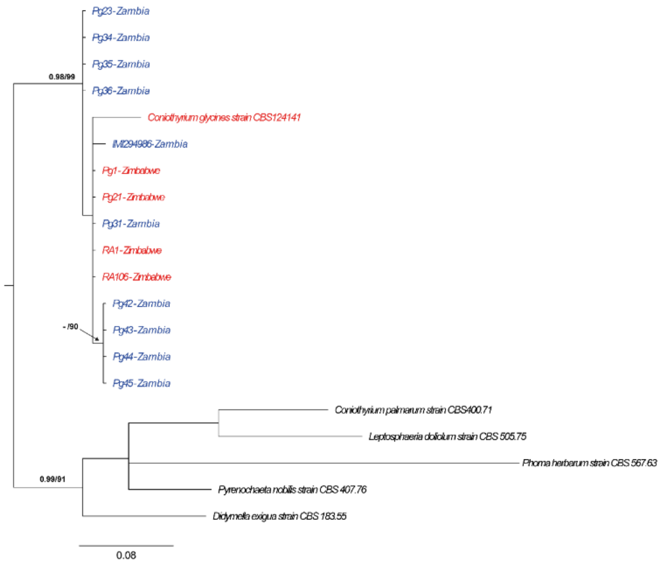

Pg23-Zambia
Pg34-Zambia
Pg35-Zambia
0.98/99
Pg36-Zambia
Coniothyrium glycines strain CBS124141
IMI294986-Zambia
Pg1-Zimbabwe
Pg21-Zimbabwe
Pg31-Zambia
RA1-Zimbabwe
RA106-Zimbabwe
Pg42-Zambia
- /90
Pg43-Zambia
Pg44-Zambia
Pg45-Zambia
Coniothyrium palmarum strain CBS400.71
Leptosphaeria doliolum strain CBS 505.75
Phoma herbarum strain CBS 567.63
0.99/91
Pyrenochaeta nobilis strain CBS 407.76
Didymella exigua strain CBS 183.55
0.08

Figure III-3: Phylogenetic tree of *C. glycines* isolates based on the BTUB gene. The tree's topology was built using K2+G nucleotide substitution model and BI approach. PP (left) was estimated with 1.5 million generations, and BS (right) with 5000 bootstraps. Only the PP >0.90 and BS>80 are displayed above each branch. Zambia and Zimbabwe isolates are colored blue and red, respectively.

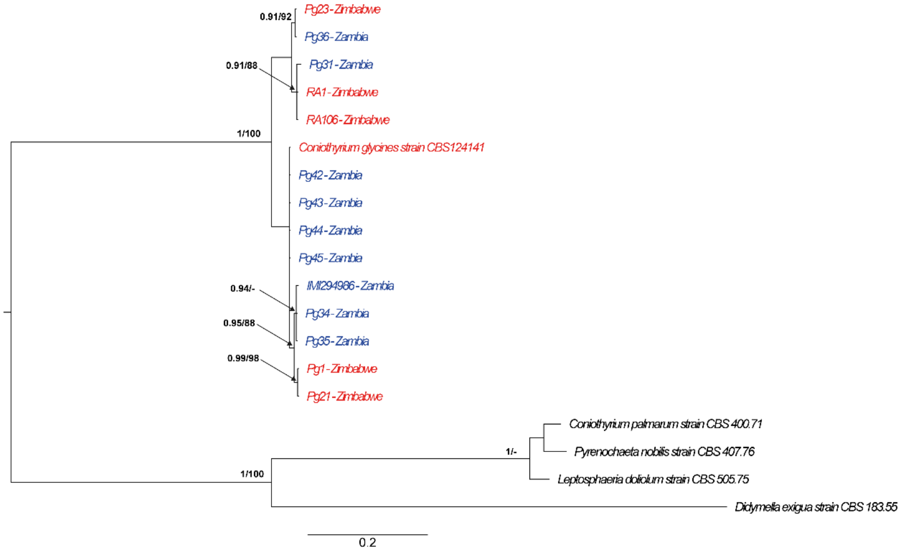


Figure III-4: Phylogenetic tree of *C. glycines* isolates based on the TEF gene. The tree's topology was built using the K2 nucleotide substitution model and BI approach. PP (left) was estimated with 1.5 million generations, and BS (right) with 5000 bootstraps. Only the PP >0.90 and BS>80 are displayed above each branch. Zambia and Zimbabwe isolates are colored blue and red, respectively.

Even though each locus could cluster *C. glycines* reference sequences and the target isolates with good branch support, they did not form well-supported nodes, especially inner nodes, and multiple polytomies were seen. Therefore, a multilocus phylogenetic analysis was performed using four loci (ITS+LSU+BTUB+TEF), where six well-supported clades were formed (Figure

III-5), and no polytomies were present. Based on the tree's topology and the fact that isolates from different countries share a common ancestor (Clades D and E), it is suggested that there has been a movement of *C. glycines* isolates between Zambia and Zimbabwe. On the other hand, it is suspected that *C. glycines* is evolving within each country and generating location-specific genotypes explained by the formation of clades where isolates from the exact origin share a common ancestor (Clades A, B, C, and F).

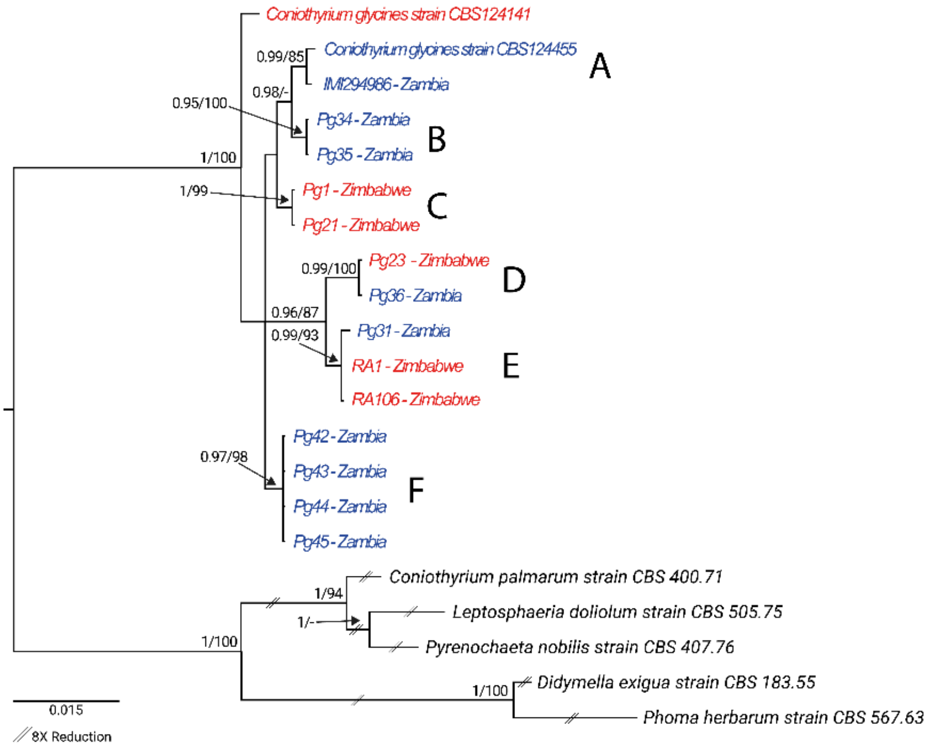


Figure III-5: Multilocus phylogenetic tree of *C. glycines* isolates based on four loci (ITS+LSU+BTUB+TEF). The tree's topology was built using nucleotide substitution model partitions for each locus and BI approach. PP (left) was estimated with 1.5 million generations, and BS (right) with 5000 bootstraps. Only the PP >0.90 and BS>80 are displayed above each branch. Zambia and Zimbabwe isolates are colored blue and red, respectively.

4. Discussion

Soybean is one of the most important crops worldwide due to their nutritional value and protein and oil commercialization (Bateman et al. 2020). It is estimated that soybean diseases can reduce the yield by up to 90% (Faske et al., 2014), and in particular red leaf blotch of soybeans (RLB), caused by the fungus *Coniothyrium glycines*, can decrease soybean production by up to 50% (Hartman et al. 1987, 2011; Murithi et al. 2022). Accurate identification of the pathogen is essential to ensure the stability of soybean production and establish adequate control and management strategies (Fang and Ramasamy 2015; Fletcher et al. 2020). Here fourteen isolates obtained from symptomatic fields in Zambia and Zimbabwe were identified as *Coniothyrium glycines* by sequencing four molecular markers and assessing their similarity with previously reported *C. glycines* sequences (de Gruyter et al. 2013).

The four loci used in this study have been employed alone or combined in multiple studies and are considered the most used molecular markers for fungal identification, being ITS the universal fungal barcode (Matute and Sepúlveda 2019; Raja et al. 2017). These loci are considered useful molecular markers due to their inter- and intra-specific variation, sequence length, and conserved flanking regions (Tekpinar and Kalmer 2019).

Obtained identity percentages were above 97% for all four loci between the unknown isolates and the available *C. glycines* sequences. Even though there is a lack of a universal threshold that indicates a reliable identification, multiple studies have used an arbitrary cut-off value for BLAST search sequence similarity that allows up to 3% of sequence divergence (>97% sequence similarity) (Hughes et al. 2013; O'Brien et al. 2005; Ryberg et al. 2008), which validates these findings. Regarding the query coverage, values above 80% were obtained with the ITS, LSU, and TEF loci, following the recommendations made by Raja et al. (2017). On the other hand, when using the BTUB, an average query coverage below 80% was reached. This lack of coverage is explained because the sequences of the *C. glycines* beta-tubulin gene in the GenBank nucleotide database have an average length of 201bp. The amplicon size of our isolates was around 320bp, meaning that different amplification primers were used, covering other portions of the gene.

DNA barcoding identifies unknown samples and unveils phylogenetic relationships with the same or other taxa (Tekpinar and Kalmer 2019). However, it is worth mentioning that each barcode has different evolution rates (Raja et al., 2017), and to address this phenomenon; the best nucleotide substitution model was selected for each locus. The Kimura 2-parameter with and without Gamma distribution was the best nucleotide substitution model for our four loci. This model is undoubtedly the most extensively employed to assess phylogenetic relationships and genetic differences (Nishimaki and Sato 2019).

According to Matute & Sepúlveda (2019), species can be identified within a phylogenetic framework as a cluster of individuals that differ significantly from other groups. It was found that with either a single locus or a concatenated dataset, unknown samples are clustered with *C. glycines* reference sequences and differed from the outgroup, supporting the previous results found with BLAST.

The single-locus phylogenetic analyses found that the ITS (Figure III-1) and the TEF (Figure III-4) loci provided the best topology with suitable branch supports and fewer polytomies. On the other hand, the LSU (Figure III-2) locus showed a complete polytomy, and no primary topology was evidenced. The phylogeny built with BTUB (Figure III-3) locus had an intermediate behavior, where polytomies were present, and a well-supported clade was formed. Overall, the single-locus phylogenetic analysis didn't provide a good differentiation and evolutionary organization within the isolates. The different topologies observed across loci are explained by the different evolution rates of each locus, the LSU being the slowest (lowest amount of variation) and the ITS being the fastest (highest variation) (Raja et al. 2017). Based on these results, the resolution of the analysis was increased by combining slowly evolving protein-coding genes (BTUB and TEF) with faster-evolving regions (ITS and LSU), as explained by Tekpinar & Kalmer (2019).

Concatenated multilocus phylogenetic analysis (ITS+LSU+BTUB+TEF) was able to form six well-supported clades (PP > 0.90 and BS > 80), and no polytomies were seen (Figure III-5). It was that found that species from matching locations form monophyletic clades (Figure III-5: A, B, C, and F), and that two clades (Figure III-5: D and E) harbored isolates from different locations but shared a common ancestor. According to these results, it is suggested there was a possible movement of contaminated material across the borders of Zambia and Zimbabwe and that geographical location is driving the development of specific genotypes. The “support rule” supports these assumptions, which states that a clade that forms monophyletic groups with a posterior probability higher than 0.90 and a bootstrap value greater than 70 tends to be reproductively isolated (Hillis and Bull 1993).

In conclusion, the identity of fourteen isolates from Zambia and Zimbabwe soybean-producing fields was confirmed. It was found that the percentage of identity or similarity between the *C. glycines* reference sequences and our samples was greater than 97%. The query coverage changed according to the primers used and the availability of sequences in the database. Additionally, this study confirmed the identity of those isolates using a multilocus phylogenetic analysis. This study suggest a development of location-specific genotypes and a movement of contaminated goods across the borders of Zambia and Zimbabwe. The results obtained in this chapter will aid in the knowledge and detection of *C. glycines* by increasing the number of publicly nucleotide available sequences and exploring the evolution and diversification of this fungal pathogen.

References

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215(3):403–10. doi: 10.1016/S0022-2836(05)80360-2.

Bateman, Nick R., Angus L. Catchot, Jeff Gore, Don R. Cook, Fred R. Musser, and J. Trent Irby. 2020. “Effects of Planting Date for Soybean Growth, Development, and Yield in the Southern USA.” *Agronomy* 10(4):596. doi: 10.3390/agronomy10040596.

Cornelius, Margaret, and Peter Goldsmith. 2019. *The State of Soybean in Africa: Soybean Yield in Africa*. Illinois.

Fang, Yi, and Ramaraja Ramasamy. 2015. “Current and Prospective Methods for Plant Disease Detection.” *Biosensors* 5(3):537–61. doi: 10.3390/bios5030537.

Faske, Travis, Terry Kirkpatrick, Jing Zhou, and Ioannis Tzanetakis. 2014. “Chapter 11: Soybean Diseases.” Pp. 1–18 in *Arkansas Soybean Production Handbook*. Vol. 4. University of Arkansas.

Fletcher, Jacqueline, Neel G. Barnaby, James Burans, Ulrich Melcher, Douglas G. Luster, Forrest W. Nutter, Harald Scherm, David G. Schmale, Carla S. Thomas, and Francisco M. Ochoa Corona. 2020. “Forensic Plant Pathology.” Pp. 49–70 in *Microbial Forensics*. Elsevier.

de Gruyter, J., and G. H. Boerema. 2002. “Contributions towards a Monograph of Phoma (Coelomycetes) VIII. Section Paraphoma: Taxa with Setose Pycnidia.” *Persoonia* 17:59–60.

de Gruyter, J., J. H. C. Woudenberg, M. M. Aveskamp, G. J. M. Verkley, J. Z. Groenewald, and P. W. Crous. 2013. “Redisposition of Phoma-like Anamorphs in Pleosporales.” *Studies in Mycology* 75:1–36. doi: 10.3114/sim0004.

Hartman, G., L. Datnoff, C. Levy, J. Sinclair, D. Cole, and F. Javaheri. 1987. “Red Leaf Blotch of Soybeans.” *Plant Disease* 113–18.

Hartman, G., and H. M. Murithi. 2022. “Coniothyrium Glycines (Red Leaf Blotch).” *CABI Compendium* CABI Compendium. doi: 10.1079/CABICOMPENDIUM.17687.

Hartman, Glen, James Haudenshield, Kent Smith, and Paul Tooley. 2011. *Recovery Plan for Red Leaf Blotch of Soybean Caused by Phoma Glycinicola*.

Hibbett, David. 2016. “The Invisible Dimension of Fungal Diversity.” *Science* 351(6278):1150–51. doi: 10.1126/science.aae0380.

Hibbett, David S., and John W. Taylor. 2013. “Fungal Systematics: Is a New Age of Enlightenment at Hand?” *Nature Reviews Microbiology* 11(2):129–33. doi: 10.1038/nrmicro2963.

Hillis, D. M., and J. J. Bull. 1993. “An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis.” *Systematic Biology* 42(2):182–92. doi: 10.1093/sysbio/42.2.182.

Hughes, Karen W., Ronald H. Petersen, D. Jean Lodge, Sarah E. Bergemann, Kendra Baumgartner, Rodham E. Tulloss, Edgar Lickey, and Joaquin Cifuentes. 2013. “Evolutionary Consequences of Putative Intra-and Interspecific Hybridization in Agaric Fungi.” *Mycologia* 105(6):1577–94. doi: 10.3852/13-041.

James, Timothy Y., Jason E. Stajich, Chris Todd Hittinger, and Antonis Rokas. 2020. “Toward a Fully Resolved Fungal Tree of Life.” *Annual Review of Microbiology*. doi: 10.1146/annurev-micro-022020.

Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. 2020. “Phylogenetic Tree Building in the Genomic Age.” *Nature Reviews Genetics* 21(7):428–44.

Leakey, C. L. A. 1964. “Dactuliophora, a New Genus of Mycelia Sterilia from Tropical Africa.” *Transactions of the British Mycological Society* 47(3):341-IN10. doi: 10.1016/S0007-1536(64)80006-1.

Lin, Feng, Sushil Satish Chhapekar, Caio Canella Vieira, Marcos Paulo Da Silva, Alejandro Rojas, Dongho Lee, Nianxi Liu, Esteban Mariano Pardo, Yi Chen Lee, Zhimin Dong, Jose Baldin Pinheiro, Leonardo Daniel Ploper, John Rupe, Pengyin Chen, Dechun Wang, and Henry T. Nguyen. 2022. “Breeding for Disease Resistance in Soybean: A Global Perspective.” *Theoretical and Applied Genetics* 135(11):3773–3872.

Matute, Daniel R., and Victoria E. Sepúlveda. 2019. “Fungal Species Boundaries in the Genomics Era.” *Fungal Genetics and Biology* 131.

Murithi, Harun M., Michelle Pawlowski, Tizazu Degu, Deresse Hunde, Molla Malede, Tonny Obua, Hapson Mushoriwa, Danny Coyne, Phinehas Tukamuhabwa, and Glen L. Hartman. 2022. “Evaluation of Soybean Entries in the Pan-African Trials for Response to Coniothyrium Glycines, the Cause of Red Leaf Blotch.” *Plant Disease* 106(2):535–40. doi: 10.1094/PDIS-05-21-1017-RE.

Naranjo-Ortiz, Miguel A., and Toni Gabaldón. 2019. “Fungal Evolution: Diversity, Taxonomy and Phylogeny of the Fungi.” *Biological Reviews* 94(6):2101–37. doi: 10.1111/brv.12550.

Nishimaki, Takuma, and Keiko Sato. 2019. “An Extension of the Kimura Two-Parameter Model to the Natural Evolutionary Process.” *Journal of Molecular Evolution* 87(1):60–67. doi: 10.1007/s00239-018-9885-1.

O’Brien, Heath E., Jeri Lynn Parrent, Jason A. Jackson, Jean-Marc Moncalvo, and Rytas Vilgalys. 2005. “Fungal Community Analysis by Large-Scale Sequencing of

Environmental Samples." *Applied and Environmental Microbiology* 71(9):5544–50. doi: 10.1128/AEM.71.9.5544-5550.2005.

Okonechnikov, Konstantin, Olga Golosova, and Mikhail Fursov. 2012. "Unipro UGENE: A Unified Bioinformatics Toolkit." *Bioinformatics* 28(8):1166–67. doi: 10.1093/bioinformatics/bts091.

Raja, Huzefa A., Andrew N. Miller, Cedric J. Pearce, and Nicholas H. Oberlies. 2017. "Fungal Identification Using Molecular Tools: A Primer for the Natural Products Research Community." *Journal of Natural Products* 80(3):756–70. doi: 10.1021/acs.jnatprod.6b01085.

Ronquist, Fredrik, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard, and John P. Huelsenbeck. 2012. "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space." *Systematic Biology* 61(3):539–42. doi: 10.1093/sysbio/sys029.

Roth, Mitchell G., Richard W. Webster, Daren S. Mueller, Martin I. Chilvers, Travis R. Faske, Febina M. Mathew, Carl A. Bradley, John P. Damicone, Mehdi Kabbage, and Damon L. Smith. 2020. "Integrated Management of Important Soybean Pathogens of the United States in Changing Climate" edited by N. Walker. *Journal of Integrated Pest Management* 11(1). doi: 10.1093/jipm/pmaa013.

Ryberg, Martin, R. Henrik Nilsson, Erik Kristiansson, Mats Töpel, Stig Jacobsson, and Ellen Larsson. 2008. "Mining Metadata from Unidentified ITS Sequences in GenBank: A Case Study in Inocybe (Basidiomycota)." *BMC Evolutionary Biology* 8(1):50. doi: 10.1186/1471-2148-8-50.

Stewart, Robert B. 1957. "An Undescribed Species of Pyrenochaeta on Soybean." *Mycologia* 49(1):115–17. doi: 10.1080/00275514.1957.12024619.

Stöver, Ben C., and Kai F. Müller. 2010. “TreeGraph 2: Combining and Visualizing Evidence from Different Phylogenetic Analyses.” *BMC Bioinformatics* 11(1):7. doi: 10.1186/1471-2105-11-7.

Tamura, Koichiro, Glen Stecher, and Sudhir Kumar. 2021. “MEGA11: Molecular Evolutionary Genetics Analysis Version 11.” *Molecular Biology and Evolution* 38(7):3022–27. doi: 10.1093/molbev/msab120.

Tekpinar, Ayten Dizkirici, and Aysenur Kalmer. 2019. “Utility of Various Molecular Markers in Fungal Identification and Phylogeny.” *Nova Hedwigia* 109(1):187–224. doi: 10.1127/nova_hedwigia/2019/0528.

Vaidya, Gaurav, David J. Lohman, and Rudolf Meier. 2011. “SequenceMatrix: Concatenation Software for the Fast Assembly of Multi-Gene Datasets with Character Set and Codon Information.” *Cladistics* 27(2):171–80. doi: 10.1111/j.1096-0031.2010.00329.x.

CHAPTER IV

DE NOVO WHOLE GENOME ASSEMBLY OF REPRESENTATIVE *CONIOTHYRIUM GLYCINES* ISOLATES FROM ILLUMINA AND OXFORD NANOPORE SEQUENCING

Abstract

Next-generation sequencing technologies (NGS) have evolved for increased sequencing speed and scalability and decreased costs, making them available for multiple applications. Different strategies are currently employed for whole genome sequencing (WGS), Illumina is the market's most accurate short-read sequencing platform, and Oxford Nanopore Technologies (ONT) is the most common nanopore-based long-read sequencing platform. A hybrid genome assembly combines short- and long-reads to accurately represent an organism's genomic information useful for plant-pathogen diagnosis. *Coniothyrium glycines* is a fungal pathogen that causes red leaf blotch (RLB) of soybeans, a severe disease not present in the U.S. Even though *C. glycines* is listed as a select agent; no molecular diagnostic tools are currently available. In this chapter, a *de novo* whole genome assembly of five representative *C. glycines* isolates (IMI294986, Pg1, Pg21, Pg43, and RA1) was performed using Illumina and ONT sequencing platforms and three bioinformatic assembly approaches (Flye, Flye + Pilon, and MaSuRCA) were tested. MaSuRCA assemblies had the highest N_{50} values (550-760 kb) and the lowest number of contigs (83-127). However, they also had the highest number of duplicate regions. Meanwhile, Flye + Pilon generated adequate assemblies with good N_{50} (80-825 kb) and few contigs (160-250). Assembly completeness of > 95% and > 89%, respectfully, was based on the BUSCO analysis with Fungal and Pleosporales single copy orthologue genes. These results will contribute to developing diagnostic tools and increase the knowledge of this fungus, facilitating future genomic research.

1. Introduction

Over the last decade, there has been a constant improvement in next-generation sequencing platforms (NGS), allowing high-speed sequencing, scalability, cost reduction, and high-throughput analysis (Hu et al. 2021; Jiao and Schneeberger 2017). Due to these advancements, NGS is used in multiple applications such as whole genome sequencing (WGS), whole-exome sequencing, variant calling, targeted sequencing, and transcriptome sequencing (Xuan et al. 2013). Different sequencing technologies have been developed over the years, driven by technological improvements and the continuous discovery of varying sequencing methods (Shendure et al. 2017).

Sequencing by synthesis is one of the oldest methodologies applied to discover the composition of genomic information. Illumina^{TM} Sequencing Technologies is the most popular approach (Pervez et al. 2022). Illumina sequencing platforms generate single- or paired-end short-reads (<500bp) with an accuracy of 99.9%, which is the most accurate base-by-base sequencing platform on the market (Bentley et al. 2008; Hu et al. 2021). However, performing a whole-genome assembly based on short reads is challenging, especially for complex organisms, due to the amount of data generated, lack of sufficient overlapping DNA regions, and a struggle to handle and interpret repetitive genomic regions (Shendure et al. 2017).

Nanopore sequencing is a novel approach that measures electric fields while single nucleic acid molecules pass through a nanopore. Oxford Nanopore Technologies (ONT) is the most used nanopore-based technology (Eisenstein 2012; Pervez et al. 2022). The nucleic acid sequence is generated by the interpretation of the unique changes in the electric fields by a base-calling algorithm (Rang, Kloosterman, and de Ridder 2018; Wick, Judd, and Holt 2019). ONT is a long-read sequencing platform that can produce reads bigger than 1 Mb. However, the main drawback of this technology is the high-error rate, which is constantly improving with newer chemistries and base calling algorithms (Hu et al. 2021; Karst et al. 2021). Recently, a novel approach to assembling genomes using both short- and long-reads, known as hybrid genome assembly, is being used and has revolutionized how genomic information is generated (Lu, Giordano, and Ning 2016). This approach uses long-reads to resolve large repetitive regions and build longer contigs. In contrast, the low-error rate short-reads are used to increase the overall accuracy of the assembly (Utturkar et al. 2014).

At the same time, there has been a proliferation and development of novel genome assembly bioinformatic software (Haridas et al. 2011). Flye is a genome assembler that uses long, error-prone reads. First, an assembly graph is built based on disjointigs (arbitrary concatenation of genomic segments), which are then resolved with more reads. Lastly, the path in the final graph allows the formation of accurate contigs (Kolmogorov et al. 2019). However, assemblies generated with high-error rate reads must be polished to increase their accuracy (Lu et al. 2016). Pilon is an integrated software tool that fills out and corrects sequences based on an internal local reassembly method with heuristic attributes (Walker et al. 2014). On the other hand, MaSuRCA (Maryland Super Read Cabog Assembler) is another commonly used bioinformatic pipeline that combines short and long sequencing reads by extending the first ones and merging them with the latter (Zimin et al. 2017).

Whole genome sequencing (WGS) allowed the reconstruction and representation of the genomic information of a biological organism (Shendure et al. 2017), becoming a powerful tool to assess population structure, phylogeography, molecular annotations, epigenetic studies, and the generation of reference genome sequences that can be used to evaluate gene expression or to create diagnostic tools (Brown 2021). Additionally, genomic information is believed to be the most specific fingerprint that can unequivocally differentiate tightly related organisms, disregarding their phenotypes, which is appropriate to microbial forensics objectives (Slezak, Allen, and Jaing 2020; Slezak, Hart, and Jaing 2020). WGS is used to diagnose several plant pathogens since it can reveal minimal genetic differences that older methodologies cannot detect.

This technique does not rely on previous knowledge of the pathogens (Chalupowicz et al. 2019; Sjödin et al. 2013).

Coniothyrium glycines is a fungal plant pathogen that causes the disease red leaf blotch of soybeans (RLB), which has not been detected in the U.S. (Hartman et al. 1987; Stewart 1957). *C. glycines* is listed as a Select Agent by the Federal Government due to its potential risk to the economy and agricultural stability. Nevertheless, no diagnostic tool or monitoring program is available (Hartman et al. 2011). The accessible molecular information of *C. glycines* is limited, and no genomes have been reported. In 2019, three draft genomes of *C. glycines* were published (Blagden et al., 2019). However, in 2023, it was discovered that those genomes were misidentified and belonged to *Epicoccum* spp. (Proano-Cuenca et al. 2023). The objective of this study was to perform a *de novo* whole genome assembly of five representative *Coniothyrium glycines* isolates using sequencing data from Illumina and ONT platforms was performed. The resulting assemblies will increase the current knowledge of the fungus, aid in the development of diagnostic tools, and facilitate future genomic research on this plant pathogen.

2. Materials and methods

2.1. High-molecular weight DNA extraction

A multilocus phylogenetic analysis was previously performed. Its topology showed the formation of six well-supported clades (A-F), where isolates from matching locations formed monophyletic relationships (clades A, B, C, and F). This analysis also suggested the movement of *C. glycines* across borders because isolates from different countries shared a common ancestor (clades D and E). Based on these results, five representative C. glycines isolates (IMI294986, Pg1, Pg21, Pg43, and RA1) were selected (Table IV-1). Isolates were grown on cellophane overlaying cV8 (clarified V8) agar for 15-21 days at room temperature in the dark. Mycelium and sclerotia were harvested by scraping the plates with a sterile metal spatula. 250 mg of fresh mycelium and sclerotia were disrupted using Precellys 24 Tissue Homogenizer (Bertin Instruments, Montigny-le-Bretonneux, France).

Table IV-1: Representative *C. glycines* isolates used to perform the *de novo* genome assemblies.

Parameter	IMI294986[a]	Pg1[b]	Pg21	Pg43	RA1
Clade	A, B	C	C	F	D, E
Source	Zambia	Zimbabwe	Zimbabwe	Zambia	Zimbabwe
Isolation	03/2005	04/2001	03/2005	05/2005	05/2006
Collected	J.M. Waller	C. Levy	C. Levy	J. Tichagwa	-

a: Representative specimen CBS124455. b: Representative specimen CBS124141.

DNA was extracted following the procedure provided by DNeasy® PowerSoil® Pro Kit (Qiagen, Germantown, MD, USA) manufacturer. To avoid DNA degradation due to mechanical forces, the tissue disruption was made two times at 4000 rpm for 20s, and buffer CD1 was added before this process. During the extraction, an extra centrifugation step at 12,000 rpm for one min was added to remove the entire washing buffer from the extraction column. Finally, the DNA was concentrated and purified by performing sodium acetate–ethanol precipitation. First, the DNA was mixed with fresh, ice-cold ethanol (1:3 v/v) and 3 M sodium acetate (10:1 v/v). The mixture was incubated for one hour at -20°C and centrifuged at 12,000 rpm for 30 min. Then, the precipitate was washed twice with 500 μL of 70% ethanol, and between washes, a centrifugation step at 12,000 rpm for 10 min was done. Finally, ethanol was removed by decanting and pipetting, and the DNA was resuspended in 1X TE (Tris-EDTA) buffer.

NanoDrop® ND-2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) was used to assess the purity of genomic DNA (gDNA) by measuring the absorbance ratios 260/280 and 260/230. The combination of both ratios represented the presence or absence of contaminants or carry-on compounds in the extracted DNA, being the most common proteins, phenols, and other chemical reagents usually found in extraction kit buffers. The extracted gDNA concentration was measured using the QuantiFluor® ONE dsDNA System with the Quantus™ fluorometer (Promega, Madison, WI, USA). Finally, ~100 ng of gDNA were loaded in a 0.8%

agarose gel, and an electrophoresis run was done at 60 V for two hours; the process was performed to evaluate the integrity of the gDNA.

2.2. Library preparation and Next-Generation Sequencing (ngs)

For each one of the five representative isolates, 500-1000 ng of gDNA were used to prepare the Oxford Nanopore Technologies (ONT) sequencing libraries following the protocol provided by the Genomic DNA by Ligation Kit (SQK-LSK110) (Oxford Nanopore Technologies, Oxford, UK) the protocol started by repairing and adding a dA-tailing module at the end of the gDNA fragments. Then, a purification step was done using AMPure XP beads, followed by the ligation of adapters, and a second purification step with AMPure XP beads and the Long Fragment Buffer (LFB). After each purification step, the concentration of the library was measured using the QuantiFluor® ONE dsDNA System with the Quantus™ fluorometer. Once the library was ready, it was kept on ice until loaded into the MinION™ Mk1C device. Long-read sequencing was performed using one FLO-Min106 R9.4 flow cell (Oxford Nanopore Technologies, Oxford, UK) per isolate with a maximum 90-hour run time. Before sequencing, a quality check was performed to calculate the available number of sequencing pores in each flow cell, and only the ones that reached more than 800 pores were used. The resulting FAST5 raw reads were then transformed into FASTQ sequence files using Guppy basecaller v6.1.2 (Wick et al. 2019).

Illumina sequencing was performed by the Center for Genomics and Proteomics at Oklahoma State University. The sequencing library preparation started with ~1000 ng of gDNA from each *C. glycines* isolate using the KAPA HyperPlus Kit (KAPA Biosystems, Roche, IN, USA) and sequenced on the Illumina^{TM} NextSeq® system (Illumina, CA, USA) using the NextSeq® 500/550 High Output Kit v2 (300 cycles, 2x150bp paired-end reads).

2.3. NGS reads quality check

NanoPlot v1.30.1 (https://github.com/wdecoster/NanoPlot) was used to assess the quality of the base called ONT reads by estimating the mean read length, the number of reads, the read length N_{50}, and the total of the base called bases. ONT sequencing adaptors were removed using Porechop v0.2.4-beta (https://github.com/rrwick/Porechop) and using Filtlong v0.2.1 (https://github.com/rrwick/Filtlong) base called reads with a quality below nine and length shorter than 1000 bp were deleted. After trimming and filtering, a second quality check using NanoPlot v1.30.1 was performed.

On the other hand, Illumina reads quality was estimated using FastQC v0.11.9 (https://github.com/s-andrews/FastQC), where the number of reads, mean read length, read Phred-Score and GC content were recorded. Finally, BBDuk v22.10 (https://sourceforge.net/projects/bbmap/) removed Illumina sequencing adaptors and reads with a Phred Quality Score below 30.

2.4. *De novo* whole genome hybrid assembly

De novo whole genome hybrid assembly was performed using three bioinformatic approaches; the first used the high-error rate long ONT reads to build a draft assembly, the second one corrected the draft assembly with the low-error rate short Illumina reads, and the third approach used the short Illumina reads to generate longer DNA fragments which then were merged using the longer ONT reads. Quality-checked, filtered, and trimmed reads were used in each procedure.

For the first approach (Flye), a draft genome was built with the ONT reads using Flye v2.9-b1174 (Kolmogorov et al. 2019) assembler. For the second approach (Flye + Pilon), Illumina short reads were mapped against the ONT draft genomes using BWA v0.7.17 (Li and Durbin 2009). After mapping, a SAM (Sequence Alignment/Map Format) file was generated for each isolate, later using SAMtools v1.10 (https://github.com/samtools/samtools), the mapping was viewed, sorted, and indexed, creating a BAM (Binary Alignment Map) file. BAM files for each isolate were used to polish each ONT draft genome using Pilon v1.24 (Walker et al. 2014). Finally, the third approach used the MaSuRCA v4.0.9 (Zimin et al. 2017) pipeline based on the Celera Assembler with the Best Overlap Graph (CABOG) and the quality-checked Illumina and ONT read as input.

2.5. Assemblies’ quality check and completeness

To select the best assembly for each isolate, all the approaches were compared using QUAST v5.0.2 (Mikheenko et al. 2018) and BUSCO (Benchmarking Universal Single-Copy Orthologue) v5.3.1 (Simão et al. 2015). The first one provided the essential genome quality metrics and statistics such as assembly size, number of contigs, N_{50}, %GC content, largest contig, and genome coverage. While the latter assessed the completeness of each assembly based on the universal single-copy orthologue genes using two pre-build databases, fungi_odb10 (fungal orthologue genes database) and pleosporales_obd10 (Pleosporales orthologue genes database). The assemblies with the lowest number of contigs, the biggest N_{50}, and the highest percentage of BUSCO completeness were considered the best. Finally, each isolate's de novo genome assembly was uploaded to the Genome NCBI database.

3. Results

3.1. ONT library preparation and Next-Generation Sequencing (NGS)

The DNA of five *C. glycines* isolates (IMI294986, Pg1, Pg21, Pg43, and RA1) was extracted using the Qiagen DNeasy® PowerSoil® Pro Kit following the manufacturer’s protocol. The DNA purity was estimated based on the absorbance ratios A260/280 and A260/230; for all five isolates, we obtained ratios between 1.67-1.90 and 1.80-2.10, respectively. The extracted dsDNA (double-stranded DNA) concentration ranged between 74 and 179 ng/uL (Table IV-2), which was suitable to start the ONT library preparation following the Genomic DNA by Ligation Kit (SQK-LSK110). After performing the first AMPure XP Beads purification step, we obtained dsDNA concentrations between 20 and 40 ng/uL (Table IV-2). Finally, the DNA Library (after the second purification step) had a dsDNA concentration higher than 35 ng/uL (Table IV-2).

Table IV-2: Concentration of dsDNA (double-stranded DNA) obtained from Quantus fluorometer. The dsDNA concentration was measured before the ONT library preparation, after the purification with the AMPure XP Beads (1st QC), and once the library was ready (DNA Library).

Samples	DNA Concentration [ng/uL]	Extraction yield [ng]	1st QC[c] [ng/uL]	DNA Library[d] [ng/uL]
IMI294986[a]	124	7,440	40	40
Pg1[b]	83	4,980	17	44
Pg21	80	4,800	20	44
Pg43	74	4,440	40	38
RA1	179	12,550	23	35

a: Reference isolate CBS124455. b: Reference isolate CBS124141. c: final volume 60uL. d: final volume 15uL.

The prepared DNA Library of each isolate was loaded into the FLO-MIN106 R9.4 flow cell (one flow cell per isolate) and sequenced with the MinION™ Mk1C device. On average, each flow cell had 1339 available pores for sequencing, and the sequencing ran for more than 72 hours. The number of gigabases produced differed among isolates, the highest 39.32 Gbp (Pg1) and the lowest 10.94 Gbp (Pg21). The number of reads generated ranged from 3.18 (Pg21) to 16.68 (RA1) million reads, and the N_{50} had an average of 5.09 kb. The gigabytes produced were between 116.55 GB (Pg21) and 385.54 GB (Pg1) (Table IV-3).

Table IV-3: ONT sequencing conditions using the MinION Mk1C sequencer device. The number of available pores, sequencing time, generated gigabases, produced read, generated gigabytes, and N50 are shown.

Samples	Available pores	Sequencing time [hours]	Generated gigabases [Gbp]	Generated reads [x10^6]	Generated gigabytes [GB]	N_{50} [kb]
IMI294986	1,508	72	36.38	12.4	365.82	5.23
Pg1	1,431	90	39.32	12.95	385.54	4.55
Pg21	1,268	72	10.94	3.18	116.55	6.53
Pg43	1,244	72	25.64	9.85	256.95	5.03
RA1	1,245	82	31.54	16.68	343.5	4.12

3.2. NGS reads quality check

ONT-generated reads underwent a quality check process, removing reads with a sequencing quality below nine and length below 1000 bp; sequencing adaptors were removed. After the quality check process, there was a reduction in the number of reads per isolate and an increase in the mean read quality and N_{50} metric. On the other hand, once Illumina reads were obtained from the Core Facility at Oklahoma State University, they went through a similar quality check process where sequencing adaptors were eliminated and only reads above Phred quality of 30 were kept (Table IV-4). Quality-checked reads for each platform were then used to assemble the genome of *C. glycines* isolates.

Table IV-4: ONT and Illumina sequencing quality metrics. Metrics were estimated using NanoPlot and FastQC for ONT and Illumina reads, respectively.

Samples	Sequencing platform	Total reads [x10^6]	Mean read length [bp]	Mean read quality	N_{50} [kb]	Total bases [Gbp]
IMI294986	ONT	4.84	5,148	13.6	5.72	24.99
	Illumina	111.3	140	32	-	-
Pg1	ONT	6.20	4,507	13.9	4.93	27.93
	Illumina	87.4	138	32	-	-
Pg21	ONT	1.30	6,172	14	7.00	8.03
	Illumina	100.3	135	35	-	-
Pg43	ONT	3.53	4,968	13.2	5.70	17.52
	Illumina	90.3	136	34	-	-
RA1	ONT	5.89	3,628	13.8	4.71	21.38
	Illumina	116.4	139	32	-	-

3.3. *De novo* whole genome hybrid assembly

Three assembly pipelines were assessed, and the best genome assembly was selected based on the genomic metrics and the completeness of each assembly. The first approach used the Flye assembler with ONT reads, the second employed Illumina reads to polish the previously assembled genome (Flye + Pilon), and the third employed both ONT and Illumina reads with the MaSuRCA assembler.

Genomic metrics assembly length, number of contigs, largest contig, GC content, and N50-didn’t change significantly between Flye and Flye + Pilon approaches (Table IV-5). However, there was an improvement in the completeness of the assemblies based on the BUSCO results of both used databases. With the Flye + Pilon approach, there was a more significant amount of complete and single-copy BUSCOs, and a lower amount of duplicated, missing, and fragmented BUSCOs (Table IV-6 and Table IV-7). On the other hand, the MaSuRCA approach generated the longest assemblies and contigs, with the lowest amount of contigs per assembly.

Regarding the GC content and N_{50}, there were no apparent changes between all three approaches (Table IV-5). Even though the assemblies generated using MaSuRCA were the longest with fewer contigs, there was a significant increase in duplicated BUSCOs in both fungal and Pleosporales databases (Table IV-6 and Table IV-7). Based on these results, combining ONT and Illumina reads following the Flye + Pilon approach generated the best assemblies. Nevertheless, for the IMI294986 isolate, the best assembly was generated with MaSuRCA since the number of contigs was significantly lower, and both the N_{50} and the BUSCO metrics were better.

Table IV-5: Comparison of genomic metrics between different genome assembly approaches. The metrics used are the assembly length, number of contigs, largest contig, GC content, and N50. Metrics were obtained using NanoPlot.

Isolate	Assembler	Assembly Length [Mbp]	# Contigs	Largest Contig [bp]	%GC	N_{50} [kb]
IMI294986	**Flye** ONT	30.50	782	515,828	50.69	79.92
	Flye + Pilon ONT + Illumina	30.52	782	516,117	50.70	79.96
	MaSuRCA Illumina + ONT	37.73	127	3,936,768	46.48	554.50
Pg1	**Flye** ONT	31.38	165	1,770,256	49.77	824.21
	Flye + Pilon ONT + Illumina	31.40	165	1,771,194	49.77	824.66

Isolate	Assembler	Assembly Length [Mbp]	# Contigs	Largest Contig [bp]	%GC	N_{50} [kb]
	MaSuRCA Illumina + ONT	35.85	101	3,513,556	47.26	674.21
Pg21	**Flye** ONT	32.39	244	1,421,129	49.03	612.20
	Flye + Pilon ONT + Illumina	32.40	244	1,421,523	49.03	612.44
	MaSuRCA Illumina + ONT	37.15	125	3,537,457	46.97	722.59
Pg43	**Flye** ONT	31.25	192	1,992,637	49.92	827.73
	Flye + Pilon ONT + Illumina	31.27	192	1,993,975	49.93	828.25
	MaSuRCA Illumina + ONT	36.12	98	3,471,694	47.08	683.86
RA1	**Flye** ONT	30.56	167	1,394,117	50.49	716.06
	Flye + Pilon ONT + Illumina	30.57	167	1,394,761	50.50	716.45
	MaSuRCA Illumina + ONT	36.10	83	2,780,443	46.95	762.68

Table IV-6: Genome completeness based on BUSCO analysis using the available fungal database harboring 758 single-copy orthologue genes.

Isolate	Assembler	C	CS	CD	F	M
IMI	Flye	701(92%)	688	13	25	32
	Flye + Pilon	720(95%)	707	13	11	27
	MaSuRCA	746(98%)	721	25	2	10
Pg1	Flye	721(95%)	720	1	19	18
	Flye + Pilon	746(98%)	745	1	2	10
	MaSuRCA	746(98%)	722	24	3	9
Pg21	Flye	732(97%)	731	1	15	11
	Flye + Pilon	748(99%)	747	1	2	8
	MaSuRCA	745(98%	705	40	2	11
Pg43	Flye	720(95%)	719	1	20	18
	Flye + Pilon	750(99%)	749	1	2	6
	MaSuRCA	745(98%)	723	22	2	11
RA1	Flye	719(95%)	718	1	23	16
	Flye + Pilon	743(98%)	742	1	2	13
	MaSuRCA	747(99%)	732	15	2	9

C: Complete. CS: Complete Single Copy. CD: Complete Duplicated. F: Fragmented. M: Missing

Table IV-7: Genome completeness based on BUSCO analysis using the Pleosporales database, which harbors 6641 single-copy orthologue genes.

Isolate	Assembler	C	CS	CD	F	M
IMI	Flye	5,685(86%)	5,570	115	174	782
	Flye + Pilon	5,903(89%)	5,784	119	68	670
	MaSuRCA	6,163(93%)	5,929	234	28	450
Pg1	Flye	5,934(89%)	5,925	9	138	569
	Flye + Pilon	6,169(93%)	6,162	7	34	438
	MaSuRCA	6,172(93%)	5,997	175	35	434
Pg21	Flye	6,030(91%)	6,021	9	111	500
	Flye + Pilon	6,187(93%)	6,180	7	39	415
	MaSuRCA	6,160(93%)	5,856	304	35	446
Pg43	Flye	5,844(88%)	5,837	7	203	594
	Flye + Pilon	6,180(93%)	6,174	6	35	426
	MaSuRCA	6,159(93%)	5,958	201	30	452
RA1	Flye	5,951(90%)	5,944	7	132	558
	Flye + Pilon	6,165(93%)	6,158	7	36	440
	MaSuRCA	6,170(93%)	6,050	120	30	441

C: Complete. CS: Complete Single Copy. CD: Complete Duplicated. F: Fragmented. M: Missing

The assemblies with the best quality metrics and genome completeness were uploaded to the GenBank Genome database. A resource announcement was published, the first genomic information available for *C. glycines* (Proano-Cuenca et al. 2023). Table IV-8 summarizes the metrics of the published genomes. The accession numbers for the genome assemblies are the following: GCA_025742395.1 (IMI294986), GCA_025742375.1 (Pg1), GCA_025742385.1 (Pg21), GCA_025742365.1 (Pg43), and GCA_025742355.1 (RA1).

Table IV-8: *C. glycines* final genome hybrid assembly's metrics. Metrics were estimated using Quast.

Attribute	Assemblies				
	IMI294986	**Pg1**	**Pg21**	**Pg43**	**RA1**
ONT mean read length (kbp)	5.15	4.98	4.51	6.17	3.63
Illumina mean read length (bp)	140	136	138	135	139
Assembly size (Mb)	37.73	31.27	31.40	32.40	30.57
Number of contigs	127	192	265	244	167
N_{50} (kbp)	554.50	828.25	824.66	612.45	716.45
%GC	46.48	49.77	49.03	49.93	50.50
Largest contig (bp)	3,936,768	1,993,975	1,771,194	1,421,523	1,394,761
Genome coverage (X)	1,562	1,442	1,830	1,130	1,837

4. Discussion

The yield, concentration, and purity (A260/280 and A260/230 ratios) of the extracted DNA were assessed using the NanoDrop 2000 and the Quantus Fluorometer. Uneven concentrations and yields amongst isolates were obtained, being the lowest at 74 ng/uL (yield: 4440 ng) and the highest at 179 ng/uL (yield: 12,550 ng), reaching the minimum amount of DNA (1,000 ng) needed to perform ONT sequencing using the Genomic DNA by Ligation Kit (SQK-LSK110). The different yield obtained in each DNA extraction is explained by the influence of multiple factors such as i) composition of the cell wall of the treated biological material (sclerotia vs. hyphae), ii) efficacy of the cell lysis using mechanical and chemical approaches, and iii) the initial condition of the sample (old vs. new tissue) (Frau et al. 2019). All the extracted DNA was considered pure since the A260/280, and A260/230 ratios ranged between 1.67-1.90 and 1.80-2.10, respectively. Obtained ratios were close to the recommended values of 1.80 and 2.0 for the A260/280 and A260/230 ratios, respectively. The minor deviations from the optimum values could be explained by the possible presence of carry-on compounds such as proteins, phenols, EDTA, guanidine salts, or carbohydrates, which are usually found in DNA extraction kits (Griffin et al. 2002; Jaudou et al. 2022). Finally, after the two purification steps with the AMPure XP Beads, an overall decrease in the final amount of extracted DNA of 75% for the first purification and 64% for the second was recorded. The reduction in the amount of DNA could be addressed by the fact that this clean-up process removes adapter dimers and short-length DNA fragments (< 300bp) (Quail, Swerdlow, and Turner 2009), suggesting that the extracted DNA was a mixture of long and short nucleic acid molecules.

ONT sequencing was done with the MinION™ Mk1C device and the FLO-Min106 R9.4 flow cell. The run-time of the sequencing process ranged between 72 and 90 hours, time which an average of 28.7±11.2 gigabases, 11.01±5.01 million reads (N_{50}=5.09±0.91 kb), and 293.67±110.51 gigabytes were generated. On the other hand, Illumina produced, on average, 100±12.67 million reads with an estimated size of 11 gigabytes. The results obtained are like the ones described by Delahaye & Nicolas (2021), T. Hu et al. (2021), and Pervez et al. (2022).

MaSuRCA was the easiest to use from all three assembly bioinformatic pipelines since it is optimized to use short- and long-reads at once, automating polishing steps. In contrast, the combination of Flye and Pilon required the user to map short reads to a draft assembly before the polishing stage, increasing the assembly run time. On the other hand, troubleshooting was easier to do with Flye and Pilon because of the autonomous nature of MaSuRCA, which reduces the user's control during the assembly process.

Regarding the performance of each bioinformatic approach, assembly metrics and completeness were assessed using QUAST and BUSCO, respectively. There was no significant difference in the N_{50} value and GC content across all assemblers. However, MaSuRCA generated the largest assembly with the fewest contigs, and the longest contigs were also found when MaSuRCA was used. There was no significant difference between Flye and Flye + Pilon in all the assembly metrics, meaning that the extra polishing step didn't modify the core structure of the draft assembly. The N_{50} metric, the number of contigs, and the genome coverage have been commonly used to assess the performance of different genome assembly pipelines. The first two metrics represent the contiguity of the assembly, and the third one provides an idea of the robustness of the assembly (Gavrielatos et al. 2021; Jiao and Schneeberger 2017; Lu et al. 2016). A more contiguous assembly is achieved with the lowest number of contigs and the highest N_{50} value (Hu et al., 2021), and the genome coverage is expected to be greater than 800X (Utturkar et al. 2014). Overall, MaSuRCA generated the most continuous assemblies based on the calculated metrics, and for all three approaches, the genome coverage was above the suggested threshold.

A complementary quantitative assessment of the genome assembly process was performed using BUSCO (Benchmarking Universal Single-Copy Orthologue). This tool quantifies the completeness of an assembly based on single-copy orthologue genes databases (Simão et al.

2015). We found that using Flye + Pilon instead of just Flye increased the completeness of the assemblies by reducing the amount of fragmented and missing genes. There was no significant difference in the completeness of the assemblies between Flye + Pilon and MaSuRCA. However, we found that MaSuRCA assemblies had more duplicate regions than the other two approaches. Gavrielatos et al. (2021) reported the same issue while using MaSuRCA to assemble the genome of *Drosophila* spp.

To conclude, even though MaSuRCA generated the most continuous assemblies based on their N_{50} and the number of contigs, the Flye assembler with the Pilon polishing tool (Flye + Pilon) generated the most complete assemblies based on the BUSCO results. Therefore, Flye + Pilon assemblies were published and uploaded to the GenBank NCBI Database, being the five generated assemblies the first ones to be announced for the fungal pathogen *C. glycines*. The published assemblies will contribute to the knowledge of the fungus, aid in the development of diagnostic tools, and facilitate future research on this plant pathogen.

References

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456(7218):53–59. doi: 10.1038/nature07517.

Blagden, Trenna, Andres Espindola, Kitty Cardwell, Alejandro Ortega-Beltran, and Ranajit Bandyopadhyay. 2019. "Draft Genome Sequences of Three Isolates of Coniothyrium Glycines, Causal Agent of Red Leaf Blotch of Soybean." *Microbiology Resource Announcements* 8(40):1–2. doi: 10.1128/MRA.

Brown, Amanda Claire. 2021. "Whole-Genome Sequencing of Mycobacterium Tuberculosis Directly from Sputum Samples." Pp. 459–80 in.

Chalupowicz, L., A. Dombrovsky, V. Gaba, N. Luria, M. Reuven, A. Beerman, O. Lachman, O. Dror, G. Nissan, and S. Manulis-Sasson. 2019. "Diagnosis of Plant Diseases Using the Nanopore Sequencing Platform." *Plant Pathology* 68(2):229–38. doi: 10.1111/ppa.12957.

Delahaye, Clara, and Jacques Nicolas. 2021. "Sequencing DNA with Nanopores: Troubles and Biases." *PLOS ONE* 16(10):e0257521. doi: 10.1371/journal.pone.0257521.

Eisenstein, Michael. 2012. "Oxford Nanopore Announcement Sets Sequencing Sector Abuzz." *Nature Biotechnology* 30(4):295–96. doi: 10.1038/nbt0412-295.

Frau, Alessandra, John G. Kenny, Luca Lenzi, Barry J. Campbell, Umer Z. Ijaz, Carrie A. Duckworth, Michael D. Burkitt, Neil Hall, Jim Anson, Alistair C. Darby, and Christopher S. J. Probert. 2019. "DNA Extraction and Amplicon Production Strategies Deeply

InfLuence the Outcome of Gut Mycobiome Studies." *Scientific Reports* 9(1):9328. doi: 10.1038/s41598-019-44974-x.

Gavrielatos, Marios, Konstantinos Kyriakidis, Demetrios Spandidos, and Ioannis Michalopoulos. 2021. "Benchmarking of next and Third Generation Sequencing Technologies and Their Associated Algorithms for de Novo Genome Assembly." *Molecular Medicine Reports* 23(4):251. doi: 10.3892/mmr.2021.11890.

Griffin, DW, Ca Kellogg, KK Peak, and Ea Shinn. 2002. "A Rapid and Efficient Assay for Extracting DNA from Fungi." *Letters in Applied Microbiology* 34:210–14.

Haridas, Sajeet, Colette Breuill, Joerg Bohlmann, and Tom Hsiang. 2011. "A Biologist's Guide to de Novo Genome Assembly Using next-Generation Sequence Data: A Test with Fungal Genomes." *Journal of Microbiological Methods* 86(3):368–75. doi: 10.1016/j.mimet.2011.06.019.

Hartman, G., L. Datnoff, C. Levy, J. Sinclair, D. Cole, and F. Javaheri. 1987. "Red Leaf Blotch of Soybeans." *Plant Disease* 113–18.

Hartman, Glen, James Haudenshield, Kent Smith, and Paul Tooley. 2011. *Recovery Plan for Red Leaf Blotch of Soybean Caused by Phoma Glycinicola*.

Hu, Taishan, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. 2021. "Next-Generation Sequencing Technologies: An Overview." *Human Immunology* 82(11):801–11. doi: 10.1016/j.humimm.2021.02.012.

Jaudou, Sandra, Mai-Lan Tran, Fabien Vorimore, Patrick Fach, and Sabine Delannoy. 2022. "Evaluation of High Molecular Weight DNA Extraction Methods for Long-Read Sequencing of Shiga Toxin-Producing Escherichia Coli." *PLOS ONE* 17(7):e0270751. doi: 10.1371/journal.pone.0270751.

Jiao, Wen-Biao, and Korbinian Schneeberger. 2017. "The Impact of Third Generation Genomic Technologies on Plant Genome Assembly." *Current Opinion in Plant Biology* 36:64–70. doi: 10.1016/j.pbi.2017.02.002.

Karst, Søren M., Ryan M. Ziels, Rasmus H. Kirkegaard, Emil A. Sørensen, Daniel McDonald, Qiyun Zhu, Rob Knight, and Mads Albertsen. 2021. "High-Accuracy Long-Read Amplicon Sequences Using Unique Molecular Identifiers with Nanopore or PacBio Sequencing." *Nature Methods* 18(2):165–69. doi: 10.1038/s41592-020-01041-y.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. "Assembly of Long, Error-Prone Reads Using Repeat Graphs." *Nature Biotechnology* 37(5):540–46. doi: 10.1038/s41587-019-0072-8.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25(14):1754–60. doi: 10.1093/bioinformatics/btp324.

Lu, Hengyun, Francesca Giordano, and Zemin Ning. 2016. "Oxford Nanopore MinION Sequencing and Genome Assembly." *Genomics, Proteomics & Bioinformatics* 14(5):265–79. doi: 10.1016/j.gpb.2016.05.004.

Mikheenko, Alla, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, and Alexey Gurevich. 2018. "Versatile Genome Assembly Evaluation with QUAST-LG." *Bioinformatics* 34(13):i142–50. doi: 10.1093/bioinformatics/bty266.

Pervez, Muhammad Tariq, Mirza Jawad ul Hasnain, Syed Hassan Abbas, Mahmoud F. Moustafa, Naeem Aslam, and Syed Shah Muhammad Shah. 2022. "A Comprehensive Review of Performance of Next-Generation Sequencing Platforms." *BioMed Research International* 2022:1–12. doi: 10.1155/2022/3457806.

Proano-Cuenca, Fernanda, Daniel Carrera-Lopez, Douglas Luster, Kurt Zeller, and Kitty Cardwell. 2023. "Genome Sequence Resources for Five Isolates of Coniothyrium Glycines, Causal Pathogen of Red Leaf Blotch of Soybeans." *PhytoFrontiers^{TM}* 1–15. doi: 10.1094/PHYTOFR-10-22-0113-A.

Quail, Michael A., Harold Swerdlow, and Daniel J. Turner. 2009. “Improved Protocols for the Illumina Genome Analyzer Sequencing System.” *Current Protocols in Human Genetics* 62(1). doi: 10.1002/0471142905.hg1802s62.

Rang, Franka J., Wigard P. Kloosterman, and Jeroen de Ridder. 2018. “From Squiggle to Basepair: Computational Approaches for Improving Nanopore Sequencing Read Accuracy.” *Genome Biology* 19(1):90. doi: 10.1186/s13059-018-1462-9.

Shendure, Jay, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. 2017. “DNA Sequencing at 40: Past, Present and Future.” *Nature* 550(7676):345–53. doi: 10.1038/nature24286.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31(19):3210–12. doi: 10.1093/bioinformatics/btv351.

Sjödin, Andreas, Tina Broman, Öjar Melefors, Gunnar Andersson, Birgitta Rasmusson, Rickard Knutsson, and Mats Forsman. 2013. “The Need for High-Quality Whole-Genome Sequence Databases in Microbial Forensics.” *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 11(S1):S78–86. doi: 10.1089/bsp.2013.0007.

Slezak, Tom, Jonathan Allen, and Crystal Jaing. 2020. “Genomics.” Pp. 283–97 in *Microbial Forensics*. Elsevier.

Slezak, Tom, Bradley Hart, and Crystal Jaing. 2020. “Design of Genomic Signatures for Pathogen Identification and Characterization.” Pp. 299–312 in *Microbial Forensics*. Elsevier.

Stewart, Robert B. 1957. “An Undescribed Species of Pyrenochaeta on Soybean.” *Mycologia* 49(1):115–17. doi: 10.1080/00275514.1957.12024619.

Utturkar, Sagar M., Dawn M. Klingeman, Miriam L. Land, Christopher W. Schadt, Mitchel J. Doktycz, Dale A. Pelletier, and Steven D. Brown. 2014. "Evaluation and Validation of de Novo and Hybrid Assembly Techniques to Derive High-Quality Genome Sequences." *Bioinformatics* 30(19):2709–16. doi: 10.1093/bioinformatics/btu391.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, and Ashlee M. Earl. 2014. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement." *PLoS ONE* 9(11):e112963. doi: 10.1371/journal.pone.0112963.

Wick, Ryan R., Louise M. Judd, and Kathryn E. Holt. 2019. "Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing." *Genome Biology* 20(1):129. doi: 10.1186/s13059-019-1727-y.

Xuan, Jiekun, Ying Yu, Tao Qing, Lei Guo, and Leming Shi. 2013. "Next-Generation Sequencing in the Clinic: Promises and Challenges." *Cancer Letters* 340(2):284–95. doi: 10.1016/j.canlet.2012.11.025.

Zimin, Aleksey V., Daniela Puiu, Ming-Cheng Luo, Tingting Zhu, Sergey Koren, Guillaume Marçais, James A. Yorke, Jan Dvořák, and Steven L. Salzberg. 2017. "Hybrid Assembly of the Large and Highly Repetitive Genome of Aegilops Tauschii, a Progenitor of Bread Wheat, with the MaSuRCA Mega-Reads Algorithm." *Genome Research* 27(5):787–92. doi: 10.1101/gr.213405.116

CHAPTER V

DEVELOPMENT AND VALIDATION OF E-PROBES FOR THE DETECTION OF *CONIOTHYRIUM GLYCINES* IN HIGH-THROUGHPUT SEQUENCING (HTS) SAMPLES

Abstract

E-probe Diagnostic Nucleic acid Analysis (EDNA) is a workflow that uses high-throughput sequencing (HTS) and metagenomic data for the detection of target organisms using e-probes, unique DNA sequences serving as electronic “fingerprints”. Red leaf blotch (RLB) of soybeans [*Glycine max* (L.) Merr.] is a severe disease caused by the fungal pathogen *Coniothyrium glycines* [R.B. Stewart]. Currently, there is no program monitoring the introduction of *C. glycines* to the U.S. due to the lack of reliable detection methods and limited knowledge about the pathogen. Therefore, in this chapter, species- and strain-collection location-specific e-probes were developed and validated for the detection and discrimination of *C. glycines* genotypes within simulated, spiked, and actual sequencing data. Two sequencing platforms (Oxford Nanopore Technologies-ONT and Illumina) were used for the simulated data and the ONT platform for the spiked and real sequencing data. To design the e-probes, all available genome sequences from five *C. glycines* isolates served as the inclusivity panel, using isolates from Zambia and Zimbabwe. The exclusivity panel comprised genomic sequences from the nearest related fungi to *C. glycines* and the soybean host plant. Designed e-probes were validated using *in silico, in vitro,* and *in vivo* approaches. The diagnostic performance metrics were calculated based on the results obtained from the MiFi® platform. Cross-reactivity of Zambia (ZB) strain-specific e-probes with Zimbabwe (ZW) strain-specific e-probes was evaluated and vice versa. 863 species-specific, 43 ZB strain-specific, and 17 ZW strain-specific e-probes were designed, each of 40 nucleotides in

length. During the *in silico* validation, the Limit of Detection (LOD) of the species-specific, ZB strain-specific, and ZW strain-specific e-probes using ONT data were 250, 4,000, and 10,000 reads. While more reads were required using Illumina data, with LODs of 4,500, 110,000, and 145,000 reads, respectively. Estimated LODs for each platform differed due to the fewer but longer reads in ONT versus Illumina. Thus, based on mean base pairs per simulated metagenome, MiFi® is more sensitive when using shorter reads. For the three e-probe sets, we saw that once the LOD is exceeded, diagnostic sensitivity and specificity reached values of 100%. The *in vitro* analysis suggested that the LOD based on the amount of *C. glycines* DNA in a pure host background (1,600 ng) was 2 ng, 150 ng, and 80 ng for the species-specific, ZB strain-specific, and ZW strain-specific e-probes, respectively. The ONT reads associated with the reported LODs were 289, 9,200, and 7,100, respectively. Finally, during the *in vivo* validation, we inoculated and infected detached soybean leaves with different *C.glycines* isolates. Even though we reached a diagnostic specificity of 100% for all the e-probe sets, the diagnostic sensitivity was 60%, 11%, and 0% for the species-specific, ZB strain-specific, and ZW strain-specific, respectively, which suggested that more reads are needed to reach the previously estimated LODs. These findings provide valuable guidelines for using HTS-based MiFi® for the detection and discrimination of *C. glycines* to improve international biosecurity measures. In this study, MiFi® was used for the first time to discriminate the same species from different origin locations, impacting the ability to perform accurate back and forward tracing during a microbial forensic analysis. Finally, limitations of this pipeline were identified that need to be addressed to improve the diagnostic performance of the designed e-probes.

1. **Introduction**

Based on its acreage and market value, soybean (Glycine max L. Merr) is the second most valuable crop in the U.S. In 2022, 86 million acres of soybean were planted, representing an average market value of $60 billion (Roth et al. 2020; United States Department of Agriculture 2023). Approximately 70% of the soybean production is used as livestock and poultry feed, and the remaining 30% is used for industrial procedures (Tooley 2017). Soybean producers are concerned about minimizing yield losses since multiple studies have shown that this crop is susceptible to several diseases caused by viruses, bacteria, fungi, nematodes, and oomycetes (Roth et al. 2020). Red leaf blotch of soybeans (RLB) is a severe disease that reduces yield by up to 60%, and it's caused by the fungal pathogen *Coniothyrium glycines* [(R.B. Stewart) Verkley & Gruyter] (de Gruyter et al. 2013; Hartman et al. 1987; Stewart 1957). RLB was first reported in 1953, and its incidence has been increasing mainly in African soybean-growth regions (Hartman and Murithi 2022; Murithi et al. 2022). Currently, the disease is not present within the U.S., and due to the possible economic and food safety impact it could cause if introduced, *C. glycines* is listed as a Select Agent by the Federal Select Agent Program (Hartman et al. 2011; Morse and Quigley 2020). Even though the fungus is a high-risk plant pathogen, there is limited molecular information available hence no diagnostic tool has been developed so far (Proano-Cuenca et al. 2023). Therefore, the development and validation of molecular diagnostic tools are needed to solidify biosecurity across U.S. borders and within soybean fields.

Science, technology, biology, and physiology developments have influenced and improved diagnostic strategies over the last years (Gullino et al. 2017). Initially, disease diagnosis and pathogen identification were based on visual inspections of a sample searching for unique phenotypic characteristics (Fletcher et al. 2020). This technique is still used today and is considered the 'gold standard' for diagnosing several plant diseases. However, morphology can't differentiate between closely related organisms, which makes it inadequate for diagnosis and regulatory or biosecurity issues (Engelthaler and Litvintseva 2020; Karunarathna et al. 2021). In those cases, serological or molecular technologies are used. Sequencing has become a standard tool for diagnosing multiple plant pathogens in the last decade, especially those with similar characteristics (Slezak, Allen, and Jaing 2020). One of the main drawbacks of sequencing is that it generates a large amount of data whose analysis requires expensive and specialized computing resources and bioinformatics expertise (Espindola and Cardwell 2021; Xuan et al. 2013).

E-probe Diagnostic Nucleic acid Analysis (EDNA) is one of the applications of next-generation sequencing (NGS) (Stobbe et al. 2013). This pipeline uses high-throughput sequencing (HTS) data and unique small DNA regions, known as e-probes, for fast and reliable detection of an organism within a sample. Microbe Finder Website (MiFi®) implements the EDNA pipeline but simplifies and avoids specialized computational resources (Espindola and Cardwell 2021). MiFi® generates a database of species-specific short DNA regions, known as e-probes, based on genome comparison between target, near taxonomically related organisms, and host genomes. The application of MiFi® for detecting and identifying one or multiple pathogens can decrease response time, which is crucial during an outbreak, especially of an exotic disease (Espindola et al. 2022; Pena-Zuniga 2020). Early diagnosis and identification of select agents is crucial for microbial forensics. It will serve as evidence to determine the pathogen origin, biosecurity pathways of movement, management strategies, and anthropological intent (Fletcher et al. 2020; Gullino et al. 2017). MiFi®-generated e-probes have been validated for multiple plant pathogens, including viruses, bacteria, fungi, and oomycetes (Blagden et al. 2016; Dang et al. 2022; Espindola et al. 2015, 2018, 2022; Pena-Zuniga 2020; Proano-Cuenca, Espíndola, and Garzon 2022); however, it has never been used for the detection of *C. glycines*.

In this chapter, species- and strain-collection location-specific e-probes were designed and validated for detecting and discriminating the select agent fungal pathogen *C. glycines* using *in silico, in vitro,* and *in vivo* approaches. Simulated and real sequencing data were used to evaluate and estimate the limit of detection (LOD), specificity, and sensitivity. Results indicate that MiFi®-generated e-probes will contribute to soybean industry and fill the diagnostic needs of control agencies. We propose that MiFi® should be incorporated within biosecurity testing pipelines at U.S. borders due to its reduced time of action and reliable diagnostic metrics.

2. Materials and methods

2.1. E-probe design and curation

Five available *C. glycines* genomes were used for the e-probe design process. Two assembled genomes belonged to Zambian isolates (IMI294986 and Pg43), and the remaining to Zimbabwean isolates (Pg1, Pg21, and RA1). These genomes were merged into three FASTA files. The first contained the genome of all the isolates (Zambian and Zimbabwean), the second the Zambian and the last one the Zimbabwean assemblies. The three resulting files will be addressed as All strains, Zambia strain-specific and Zimbabwe strain-specific, respectively. The soybean reference genome (GCF_000004515.6) was obtained from GenBank and will be known as the host. Finally, the near neighbor (NN) database (22 RefSeq and 9 GenBank genomes) was built based on the genomes of the closest taxonomically related organisms to *C. glycines* and common soybean fungal pathogens (Table V-1).

Table V-1: Genomes accession numbers and organisms used to build *C. glycines* near neighbor database. The near neighbors were selected based on taxonomy and common soybean fungal pathogens.

Accession Number	Organism	Accession Number	Organism
GCF000146915.1	*Parastagonospora nodorum*	**GCF000149985.1**	*Pyrenophora tritici-repentis*
GCF013036055.1	*Alternaria burnsii*	**GCF000240135.3**	*Fusarium graminearum*
GCF002742065.1	*Cercospora beticola*	**GCA022559915.1**	*Septoria petroselini*
GCA001599375.1	*Phoma herbarum*	**GCA004835665.1**	*Phoma sp.*
GCA002116315.1	*Epicoccum nigrum*	**GCA020272525.1**	*Epicoccum sorghinum*
GCA001644535.1	*Pyrenochaeta sp.*	**GCA020747015.1**	*Pyrenochaeta sp.*
GCF010015615.1	*Cucurbitaria berberidis*	**GCF020726555.1**	*Boeremia exigua*
GCF000523435.1	*Bipolaris zeicola*	**GCF004154835.1**	*Alternaria arborescens*
GCF004011695.1	*Ascochyta rabiei*	**GCF000354255.1**	*Bipolaris maydis*
GCF000338995.1	*Bipolaris sorokiniana*	**GCF001642055.1**	*Alternaria alternata*
GCF010093625.1	*Macroventuria anomochaeta*	**GCF000230375.1**	*Leptosphaeria maculans*
GCF019650295.1	*Cercospora kikuchii*	**GCF014235925.1**	*Colletotrichum truncatum*
GCA015266435.1	*Epicoccum latusicollum*	**GCA000359685.2**	*Pyrenochaeta sp.*
GCF907166805.1	*Alternaria atra*	**GCF020736505.1**	*Alternaria rosae*
GCF000523455.1	*Bipolaris oryzae*	**GCF000359705.1**	*Exserohilum turcica*

Accession Number	Organism	Accession Number	Organism
GCF010094145.1	*Didymella exigua*		

Once all the genomic information was gathered, the MiFi® website (https://bioinfo.okstate.edu/) was used to design e-probes (MiProbe® built-in feature) for each target following the pipeline proposed by Stobbe et al. (2013) and modified by Espindola and Cardwell (2021). Each target genome was uploaded to the MiFi® interface, and the exclusion panel was built based on the host and the near neighbors (NN) database. For the Zambian strain-specific and Zimbabwean strain-specific e-probe sets, we included in each exclusion panel Zimbabwean and Zambian isolates, respectively. The selected design parameters were a minimum match of 15 and a fixed e-probe length, where three e-probe lengths (40, 60, and 80 nucleotides) were tested for each e-probe set (Table V-2).

Table V-2: E-probe design considerations for each target genome, including the isolates aimed by the e-probes, the exclusion panel, and the different e-probe lengths assessed.

E-probe set name	Target genomes	Target isolates	Exclusion panel	E-probe length [nt]
All-probes	All strains	All (5) isolates	Host + NN	40, 60, 80
ZB-probes	Zambian strain-specific	IMI294986, Pg43	Host + NN + Zimbabwean isolates	40, 60, 80
ZW-probes	Zimbabwean strain-specific	Pg1, Pg21, RA1	Host + NN + Zambian isolates	40, 60, 80

The designed e-probes passed through a curation process where duplicate sequences and unspecific e-probes were removed. The removal of duplicate sequences was done by an in-house bash script (Table V-3), while the specificity of each e-probe was assessed by comparing them to the NCBI nucleotide database using BLAST. Then, with an in-house perl script, the BLAST output for each e-probe set was parsed, and the e-probes that made a non-specific hit with an e-value lower than the defined threshold (1e^{-9}) were eliminated. Generated e-probes were mapped to the *C. glycines* reference genomes to assess their distribution and specificity using minimap2 v2.14 (https://github.com/lh3/minimap2).

Table V-3: In-house bash script that removes the duplicate e-probe sequences. Parameters and file descriptions are detailed.

Removal of duplicate sequences script

```
#!/bin/bash

<Duplicate.txt awk '!seen[$0]++' -> Intermediate.txt

File="Intermediate.txt"
lines=$(grep "" -c Intermediate.txt)
length=$(seq 1 $lines)

for i in $length; do if [[ $(sed "$i'!d' Intermediate.txt) = ">"* ]] &&
[[ $(sed "$((i+1))'!d' Intermediate.txt) = ">"* ]];
then sed "$i'd' Intermediate.txt;
else echo $(sed "$i'!d' Intermediate.txt) >> Final.txt;
fi;
done
```

Parameters

Duplicate.txt: File name that contains the raw e-probe list with possible duplicate sequences.
Intermediate.txt: Temporary file that stores just the unique sequences.
Final.txt: Output file.

2.2. Mock HTS sample preparation

Mock samples of HTS data were constructed to test the detection performance of each set of designed e-probes and their behavior against different sequencing platforms, Illumina and Nanopore. The construction of the mock HTS data was performed by sampling randomly either Illumina or Nanopore raw reads generated during the assembly of *C. glycines* in the previous chapter and including the host (*Glycine max*) reads as a background. Soybean Nanopore reads were simulated using NanoSim v3.1.0 (https://github.com/bcgsc/NanoSim), and soybean Illumina reads were obtained from the Sequence Read Archive (SRA) SRR22107929.

The sampling process was done using an in-house bash script (Table V-4) based on the software seqkit v2.1.0 (https://github.com/shenwei356/seqkit), which randomly extracted the desired amount of *C. glycines* and host raw reads, and combined both in a single FASTQ file. For each e-probe set, nine different target concentrations with ten replicates were obtained. A fixed total number of reads (target + host) per titer was used, for ONT 250,000 reads and Illumina one million reads (Table V-5). The generated mock HTS datasets were uploaded to the MiFi® website for further analysis.

Table V-4: In-house bash script used for random sampling of host and target reads.

Bash script for random sampling and mock HTS sample preparation

```
#!/bin/bash
module load anaconda3/2019.10
source activate seqkit
cd WDIR
for index in TARGET READS
do
lth=$((SIZE-index))
mkdir ${index}
x=1
echo $lth
while [ $x -le REP]
do
cat \
HTPATH | seqkit sample -s $x -p 0.5 \
HTPATH | seqkit head -n $lth -o OUTHOST
cat \
TGPATH | seqkit sample -s $x -p 0.7 \
TGPATH | seqkit head -n $index -o OUTTARG
x=$(( $x +1 ))
done
done
echo "DONE"
```

Parameters

WDIR: Working directory path.
TARGET READS: List of the desired number of targets reads.
SIZE: Total number of desired reads, sum of host and target reads.
REP: Number of replicates per concentration.
HTPATH AND TGPATH: Host reads directory and target reads directory.
OUTHOST AND OUTTARG: Output filename for random sampled host and target reads.

Table V-5: Number of *C. glycines* reads per titer for each sequencing platform and designed e-probe sets. The percentage of target reads and the total number of reads within each metagenome are shown.

Titer	Nanopore (n=250,000 reads)						Illumina (n=1x10^6 reads)					
	All-probes	%	ZB-probes	%	ZW-probes	%	All-probes	%	ZB-probes	%	ZW-probes	%
1	10	0.004	10	0.004	10	0.004	300	0.03	300	0.03	300	0.03
2	20	0.008	100	0.040	100	0.040	600	0.06	3K	0.30	3K	0.30
3	40	0.016	500	0.200	1K	0.400	1.2K	0.12	16K	1.60	32K	3.20
4	60	0.024	1K	0.400	1.5K	0.600	2K	0.20	32K	3.20	48K	4.80
5	80	0.032	2K	0.800	2K	0.800	2.5K	0.25	72K	7.20	72K	7.20
6	100	0.040	3K	1.200	3K	1.200	3K	0.30	110K	11.00	110K	11.00
7	140	0.056	4.5K	1.800	4K	1.600	4.5K	0.45	145K	14.50	145K	14.50
8	180	0.072	5K	2.000	5K	2.000	5.5K	0.55	180K	18.00	180K	18.00
9	250	0.100	10K	4.000	10K	4.000	8K	0.80	360K	36.00	360K	36.00

ZB: Zambia. ZW: Zimbabwe. K: Thousand.

2.3. *In silico* e-probe validation using mock HTS data

The *in silico* validation of the designed e-probe sets started with detecting the target (*C. glycines*) within the different mock HTS datasets using MiDetect®, a built-in feature within the MiFi® website. MiDetect® parameters were set to an e-value of 1e^{-9} (recommended for eukaryotes) and a maximum of 250 hits (one hit represents a match between an e-probe and a sample read). The limit of detection (LOD), analytical sensitivity, and analytical specificity were calculated for each e-probe set and target read concentration. The analytical specificity was calculated just for the ZB-probes and ZW-probes. Analytical sensitivity was calculated using the formula S=TP/(TP+FN), where S: analytical sensitivity, TP: true positives, and FN: false negatives. On the other hand, the analytical specificity was calculated following the formula, Sp=TN/(TN+FP), where Sp: analytical specificity, TN: true negatives, and FP: false positives. Additionally, the *in silico* Limit of Detection (LOD) for each e-probe set was defined as the ratio between target and host reads (T/H) needed to consistently reach an analytical sensitivity of 100%. The number of target reads associated to each estimated LOD was also reported.

2.4. Estimation of the *in vitro* limit of detection (LOD)

To calculate the *in vitro* Limit of Detection (LOD), pure host (soybean) DNA was spiked with different concentrations of pure *C. glycines* DNA. Two fungal isolates were used separately, one from Zambia (IMI294986) (Table V-6) and one from Zimbabwe (Pg1) (Table V-7). The soybean and fungal DNA was extracted using the DNeasy® Plant Mini Kit (Qiagen, Germantown, MD, USA) following the manufacturer's protocol. DNA concentration and quality were assessed with NanoDrop® ND-2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and QuantiFluor® ONE dsDNA System with the Quantus™ fluorometer (Promega, Madison, WI, USA).

Table V-6: Nanograms (ng) of soybean (host) and *C. glycines* IMI294986 isolate DNA used to build spiked samples to estimate the *in vitro* LOD.

Dilution	Soybean [ng]	Target[a] [ng]	Volume [uL]	Concentration [ng/uL]
1	800	800	20	80
0.5	1,060	540	20	80
0.25	1,280	320	20	80
0.10	1,450	150	20	80
0.05	1,520	80	20	80
0.01	1,580	20	20	80
0.001	1,598	2	20	80

a: Amount of DNA from *C. glycines* IMI294986 isolate.

Table V-7: Nanograms (ng) of soybean (host) and *C. glycines* Pg1 isolate DNA used to build spiked samples to estimate the in vitro LOD.

Dilution	Soybean [ng]	Target[a] [ng]	Volume [uL]	Concentration [ng/uL]
0.25	1,280	320	20	80
0.10	1,450	150	20	80
0.05	1,520	80	20	80
0.01	1,580	20	20	80
0.001	1,598	2	20	80
0.0005	1,599	1	20	80
0.0001	1,599.8	0.2	20	80

a: Amount of DNA from *C. glycines* Pg1 isolate.

The spiked DNA pool was sequenced using the Rapid sequencing gDNA – barcoding kit (SQK-RBK004) from Oxford Nanopore Technologies (Oxford Nanopore Technologies, Oxford, UK). Two quality checks were performed based on the QuantiFluor® ONE dsDNA System during the library preparation to assess the DNA concentration. The first quality check was done after all the barcodes were pooled together, and the second one after the AMPure XP beads cleaning procedure. Once the library was ready, it was kept on ice until loaded into the MinION™ Mk1C sequencer. Sequencing was performed using one FLO-Min106 R9.4 flow cell (Oxford Nanopore Technologies, Oxford, UK) per experiment with a maximum 72-hour run time. Before sequencing, the number of available pores in each flow cell was checked, and only the ones that reached more than 800 pores were used. The resulting FAST5 raw reads were transformed into FASTQ sequence files using Guppy basecaller v6.1.2 (Wick, Judd, and Holt

2019). The resulting FASTQ files were combined into one file and uploaded to the MiFi® website to perform the detection of *C. glycines* using the three designed e-probe sets. To know the number of reads from the target within each spiked sample, raw FASTQ files were mapped to the five *C. glycines* reference genomes using minimap2 v2.14 (https://github.com/lh3/minimap2). Finally, the *in vitro* LOD was expressed in terms of the number of mapped reads, mean read length, and nanograms of the fungus needed to have a positive result.

2.5. Plant material and fungal isolate preparation

Soybean seeds were treated with a 5-minute wash, a 5-minute 4% bleach treatment, a 1-minute 70% ethanol treatment, and a tap water rinse. The seeds were then coated with Exceed® Superior Legume Inoculant (Hancock Seed Company, Dade City, FL). The plants were grown in 10 cm diameter plastic pots filled with Miracle-Gro Potting-Mix® and fertilized with 14-14-14 Osmocote (Scotts-Sierra Horticultural Products Co., Marysville, OH). Growth occurred in a Conviron A100 growth chamber (Conviron, Winnipeg, Canada) under a 16-hour photoperiod with 50% light-intensity fluorescent lamps. The temperature was 28°C during the day and 24°C at night, with a relative humidity of 80%.

Fourteen *Coniothyrium glycines* isolates, received from USDA-ARS and maintained on clarified V8 agar (cV8) at room temperature in the dark, were used for inoculum preparation. Ten to fifteen agar plugs (5 mm diameter) with mycelia and sclerotia were extracted from a fresh culture plate and placed within a 2 mL tube with 100 uL of sterile water, one big (6 mm), and three medium (3 mm) size glass beads. Tissue was homogenized thrice for 30 s at 4000 rpm by the Precellys 24 Tissue Homogenizer (Bertin Instruments, Montigny-le-Bretonneux, France). Once disrupted, 1 mL of sterile water was added and vortexed. The solution was then centrifuged for 10 min at 4000 rpm. The supernatant was discarded, and the pellet was resuspended in 1.5 mL of 0.02% (v/v) Tween20®. To estimate the concentration of each prepared inoculum, 100 uL of the stock solution (or dilution) was plated in fresh cV8 agar, and the CFU (colony-forming units) were recorded after five days. The concentration was calculated with the following formula: CFU/(V*DF), where CFU: Colony Forming Units, V: volume placed over the agar, and DF: dilution factor (if used).

2.6. Detached leaf inoculation and DNA extraction

V1 to V3 stage healthy soybean leaves were placed within a Petri dish containing a soaked Whatman-70mm filter paper (Cytiva Life Sciences, Marlborough, MA) to keep the humidity within the case during inoculation. One leaf per Petri dish was placed, and three leaflets from a single trifoliate were considered replicates of the inoculation. Over each leaf, 200uL of the prepared inoculum stock solution was set and using a sterile glass stick, the inoculum was dispersed, covering the whole leaf. Control leaves were inoculated with a solution derived from sterile agar. Inoculated leaflets were stored at room temperature in the dark for two days, and then they were kept at 24°C under a 16 h of daily photoperiod with 25% light intensity. The infection process was assessed every five days post-inoculation (dpi). Once leaves were infected (>75% symptomatic tissue), the DNA of ~250mg of symptomatic tissue was extracted with the Qiagen® DNeasy Plant Mini Kit (Qiagen, Germantown, MD, USA) following the manufacturer’s instructions. DNA quality and concentration were measured with a NanoDrop® ND-2000 spectrophotometer.

2.7. Confirmation of the infection with PCR and specific primers

The infection of *C. glycines* in soybean leaves was confirmed by PCR using two species-specific primer sets (LAC38 and LAC86) developed and validated by the USDA based on the *C. glycines* available genomic information. LAC38 primer set amplifies a nucleoside triphosphate hydrolase protein (NTHP), while the LAC86 primer set amplifies a laccase precursor protein (LPP). PCR primers and amplification conditions are detailed in Table V-8 and Table V-9, respectively. For both primers, reactions were performed in 25 uL containing 2 uL of each 5 uM primer, 12 uL of 2X GoTaq Green Master Mix (Promega, Madison, WI, USA), 2 uL of DNA (25 ng/μL), and 7 uL of nuclease free water.

Table V-8: Primer sets used to confirm the infection of soybean leaves with *C. glycines* isolates. Primer names and sequences are detailed.

Locus	Primer name	Primer sequence (5'-3')	Reference
NTHP	LAC38F	TTCGCAACAGCAACGTACTC	Provided by the USDA
	LAC38R	AAATGTACACTTTCCGCCGG	Provided by the USDA
LPP	LAC86F	CCAGTAGTTTCCGGCAGTCT	Provided by the USDA
	LAC86R	GGTGGGCACTGTACAGATCA	Provided by the USDA

Table V-9: PCR conditions for amplifying two *C. glycines*-specific primer sets. LAC38 and LAC36 primer set the same temperatures and times for both.

PCR Conditions. LAC38 & LAC36	
Temperature – Time	**Cycles**
95°C – 2 min	X1
95°C – 30 s	
58°C – 20 s	X35
72°C – 20 s	
72°C – 2 min	X1
4°C – Hold	

To visualize the amplicons, electrophoresis was performed where 5 uL of PCR product was loaded into a 1.5% agarose gel (VWR Life Sciences) stained with 3 uL of SYBR® Safe DNA gel stain (Invitrogen, Waltham, MA, USA). The electrophoresis was run for one hour at 95 V, and then the gel was visualized using the Molecular Imager® Gel Doc XR+ (Bio-Rad, Hercules, CA, USA).

2.8. Next-Generation Sequencing (NGS) of infected leaf tissue

Extracted DNA from each infected soybean leaf was used to prepare the sequencing library following the Rapid sequencing gDNA – barcoding kit (SQK-RBK004) from Oxford Nanopore Technologies (Oxford Nanopore Technologies, Oxford, UK). The DNA concentration was measured using the QuantiFluor® ONE dsDNA System after pooling all the used barcodes and after the AMPure XP beads cleaning procedure. MinION™ Mk1C sequencer device and one

FLO-Min106 R9.4 flow cell (Oxford Nanopore Technologies, Oxford, UK) were used for sequencing. The run time for each sequencing experiment was set to a maximum of 72 hours. The number of available pores in each flow cell was checked, and only the ones that reached more than 800 pores were used. Basecalling was performed using Guppy basecaller v6.1.2 (Wick et al. 2019), transforming FAST5 files to FASTQ files. FASTQ files were combined into a single file and uploaded to the MiFi® website to perform detection of *C. glycines* using the three designed e-probe sets. The number of reads from the target within each metagenome was evaluated by mapping the raw FASTQ files to the five *C. glycine* reference genomes using minimap2 v2.14 (https://github.com/lh3/minimap2).

2.9. *In vivo* e-probe validation using HTS data from infected leaves

The *in vivo* validation of the designed e-probe sets started with detecting the target (*C. glycines*) within each metagenomic sample using MiDetect®. Detection parameters were set to an e-value of 1e^{-9} (recommended for eukaryotes) with a maximum of 250 hits. The diagnostic sensitivity and specificity were calculated for each e-probe set. The diagnostic specificity was calculated just for the ZB-probes and ZW-probes. Diagnostic sensitivity was calculated using the formula S=TP/(TP+FN), where S: diagnostic sensitivity, TP: true positives, and FN: false negatives. On the other hand, the diagnostic specificity was calculated following the formula, Sp=TN/(TN+FP), where Sp: diagnostic specificity, TN: true negatives, and FP: false positives.

3. Results

3.1. E-probes for the detection of *C. glycines*

MiFi® was used to design e-probes to detect the select agent fungal pathogen *C. glycines*, clustering them into three detection groups: All-probes, ZB-probes, and ZW-probes, which are meant to detect all *C. glycines* strains, Zambian strains, and Zimbabwean strains, respectively.

We tested three e-probe lengths (40, 60, and 80 nucleotides) for each e-probe set, and we found that only 40 nucleotide-length e-probes were generated. The All-probes, ZB-probes, and ZW-probes included 863, 43, and 17 e-probes, respectively.

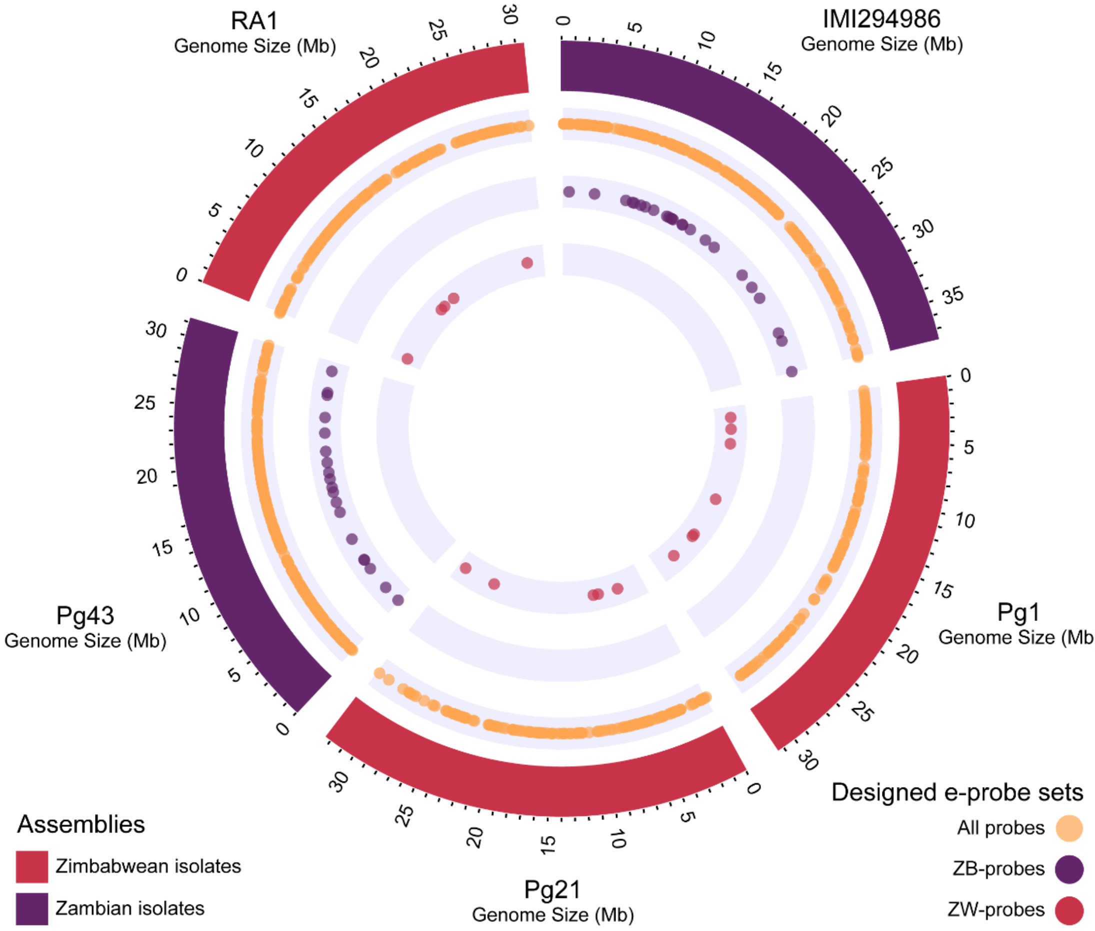


Figure V-1: Schematic representation of the distribution and specificity of the three designed e-probe sets. The outer circle represents the available *C. glycines* assemblies, colored in red and blue the Zambian and Zimbabwean isolates, respectively. The three inner circles represent the distribution of each e-probe set across the genomes.

The generated e-probes were mapped against the available *C. glycines* assemblies to evaluate their genomic distribution and specificity. We found that the All-probes set (n=863 e-probes) was distributed homogeneously across each assembly, the ZB-probes set (n=43 e-probes) was disseminated just across the Zambian assemblies but with lower coverage, and the ZW-probes set (n=17 e-probes) seemed to be clustered in specific regions of the Zimbabwean assemblies with even more inadequate coverage. Additionally, this analysis showed that no e-probes were mapped with non-target assemblies, suggesting a high specificity of the designed e-probes (Figure V-1).

3.2. *In silico* e-probe validation using mock HTS Illumina and ONT data

The *in silico* validation of the designed e-probe sets was performed by testing their diagnostic performance with MiFi®, using an e-value of $1e^{-9}$, against mock or simulated HTS Illumina and ONT sequencing data (Table V-5). We used three 40-nucleotide length e-probe sets, All-probes, ZB-probes, and ZW-probes. We tested them against simulated metagenomes that included All *C. glycines* strains, Zambian strain-specific and Zimbabwean strain-specific, respectively. The total reads were 250,000, and one million for the ONT and Illumina simulated data, respectively. Since ONT and Illumina sequencing platforms differ in the number of reads generated and their length, each prepared dilution harbors the same number of nucleotide base pairs of the target (*C. glycines*), disregarding the sequencing platform used. For instance, a concentration of 0.1% (250 target reads) of an ONT simulated data had the same number of target nucleotides as 0.8% (8000 target reads) Illumina simulated data.

The *in silico* limit of detection (LOD), using ONT data, for the All-, ZB-, and ZW-probe set was estimated to be 0.1% (250 target reads) (Table V-10), 1.6% (4,000 target reads) (Table V-11), and 4% (10,000 target reads) (Table V-12), respectively. When we used Illumina data, we found that the *in silico* LOD was 0.45% (4,500 target reads) (Table V-10), 11% (110,000 target reads) (Table V-11), and 14.5% (145,000 target reads) (Table V-12) for the All-, ZB-, and ZW-probe set, respectively.

For all the designed e-probe sets, we saw that once the LOD was reached, diagnostic sensitivity and specificity had a value of 100% (Table V-10, Table V-11, and Table V-12). We could not calculate the diagnostic specificity of the All-probe set. However, when tested against the NCBI nucleotide database, the e-probes didn’t match unspecific organisms.

Table V-10: ***In silico*** **validation of the All-probe set (n=863 e-probes) using simulated ONT and Illumina sequencing data. The table shows the number and concentration of target reads, hits, scores, and analytical sensitivity and specificity. The estimated LOD is highlighted.**

	ONT					Illumina				
Dilution	**Target reads**	**C [%]**	**Hits***	**Score* [x10^3]**	**S [%]**	**Target reads**	**C [%]**	**Hits***	**Score* [x10^3]**	**S [%]**
1	10	0.004	0.5	4.78	0	300	0.03	0.3	3.0	0
2	20	0.008	0.6	5.75	0	600	0.06	0.5	5.0	0
3	40	0.016	1.2	11.6	10	1,200	0.12	1.1	11.0	0
4	60	0.024	1.9	18.4	30	2,000	0.20	3.3	32.6	70
5	80	0.032	2.5	24.3	40	2,500	0.25	3.8	37.4	80
6	100	0.040	4.1	39.6	70	3,000	0.30	4.7	46.3	90
7	140	0.056	3.6	34.7	80	4,500	0.45	8.7	85.2	100
8	180	0.072	4.7	45.4	90	5,500	0.55	11.4	111.3	100
9	250	0.100	9.1	88.1	100	8,000	0.80	18.8	184.4	100

C: Concentration of target reads. S: Analytical Sensitivity. Score: Identity percentage x Query coverage. *: estimated average of ten replicates.

Table V-11: ***In silico*** **validation of the ZB-probe set (n=43 e-probes) using simulated ONT and Illumina sequencing data. The table shows the number and concentration of target reads, hits, scores, and analytical sensitivity and specificity. The estimated LOD is highlighted.**

	ONT						Illumina					
Dilution	**Target reads**	**C [%]**	**Hits***	**Score* [x10^3]**	**S [%]**	**Sp. [%]**	**Target reads**	**C [%]**	**Hits***	**Score* [x10^3]**	**S [%]**	**Sp. [%]**
1	10	0.004	0	0	0	100	300	0.03	0	0	0	100
2	100	0.040	0.4	3.9	0	100	3,000	0.30	0	0	0	100
3	500	0.200	1.4	13.5	0	100	16,000	1.60	3.1	30.3	30	100
4	1,000	0.400	1.7	16.4	20	100	32,000	3.20	4.8	46.4	40	100
5	2,000	0.800	3.5	34.2	60	100	72,000	7.20	9.8	94.0	60	100
6	3,000	1.20	6.2	60.6	80	100	110,000	11.0	14.7	141.5	100	100
7	4,000	1.60	8.7	84.9	100	100	145,000	14.5	18.9	182.8	100	100
8	5,000	2.00	10.9	106.0	100	100	180,000	18.0	25.9	251.9	100	100
9	10,000	4.00	20.8	203.2	100	100	360,000	36.0	43.6	426.7	100	100

C: Concentration of target reads. S: Analytical Sensitivity. Sp.: Analytical Specificity. Score: Identity percentage x Query coverage. *: estimated average of ten replicates.

Table V-12: ***In silico*** **validation of the ZW-probe set (n=17 e-probes) using simulated ONT and Illumina sequencing data. The table shows the number and concentration of target reads, hits, scores, and analytical sensitivity and specificity. The estimated LOD is highlighted.**

	ONT						Illumina					
Dilution	**Target reads**	**C [%]**	**Hits***	**Score* [x10^3]**	**S [%]**	**Sp. [%]**	**Target reads**	**C [%]**	**Hits***	**Score* [x10^3]**	**S [%]**	**Sp. [%]**
1	10	0.004	0	0	0	100	300	0.03	0	0	0	100
2	100	0.040	0	0	0	100	3,000	0.30	0	0	0	100
3	1,000	0.400	1.2	11.7	0	100	32,000	3.20	0.9	8.9	0	100
4	1,500	0.600	1.2	11.7	0	100	48,000	4.80	2.3	22.9	10	100
5	2,000	0.800	1.2	11.7	0	100	72,000	7.20	2.9	28.6	40	100
6	3,000	1.20	2.8	27.6	50	100	110,000	11.0	5.4	53.1	90	100
7	4,000	1.60	3.7	36.5	70	100	145,000	14.5	6.1	60.1	100	100
8	5,000	2.00	4.4	43.4	90	100	180,000	18.0	6.5	64.1	100	100
9	10,000	4.00	8.7	85.4	100	100	360,000	36.0	13.4	132.3	100	100

C: Concentration of target reads. S: Analytical Sensitivity. Sp.: Analytical Specificity. Score: Identity percentage x Query coverage. *: estimated average of ten replicates.

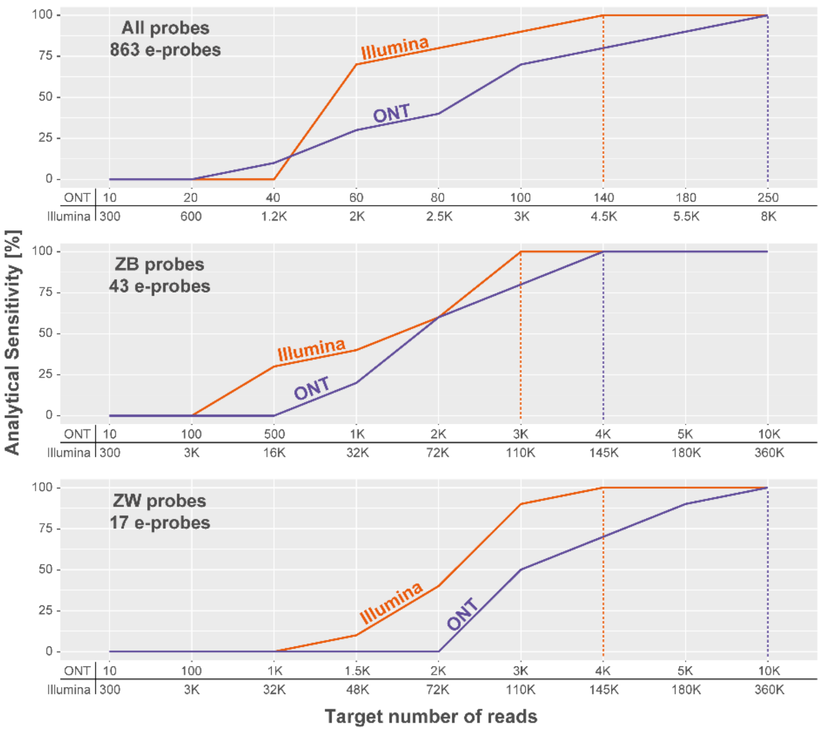


Figure V-2: Estimated analytical sensitivity for each dilution using two sequencing platforms (Illumina and ONT). The figure shows that we needed less amount of the target nucleic acid information with Illumina simulated data to reach 100% diagnostic sensitivity.

Comparing the diagnostic performance of the designed e-probe sets between sequencing platforms based on the mean target base pairs per simulated metagenome, all designed e-probes are more sensitive with Illumina sequencing data, which means that a lower amount of target nucleic acid information is needed for detection with the Illumina platform to reach 100% diagnostic sensitivity (Figure V-2).

3.3. Estimation of the *in vitro* limit of detection (LOD)

The *in vitro* limit of detection (LOD) was estimated by spiking pure soybean DNA with known concentrations of fungal DNA. *C. glycines* IMI294986 (Zambia) and Pg1 (Zimbabwe) isolates were used separately to estimate the LOD of ZB-probes and ZW-probes, respectively. To estimate the LOD of the All-probes set, we combined the results of both isolates. Spiked samples were then sequenced using the MinION^{TM} Mk1C device following the Rapid sequencing gDNA-barcoding kit from ONT.

There was a direct correlation between the amount of fungal DNA present in each sample and the number of mapped reads. We estimated that the LOD for the All-probe set was two ng (mapped reads: 289), for the ZB-probe set, it was 150 ng (mapped reads: 9,182), and for the ZW-probe set the LOD was 80 ng (mapped reads: 7,072). Our result showed that both ZB- and ZW-probes sets didn't generate any positive result, regardless of the concentration, when used against Zimbabwean and Zambian isolates, respectively (Table V-13) confirming their analytical specificity.

Table V-13: ***In vitro*** **estimation of the LOD using spiked metagenomes and ONT sequencing platform. The three designed e-probe sets are presented with their associated number of hits (H), score (S), and MiFi® result (R). The detection was performed using an e-value of 1e^{-9}.**

IMI294986 isolate (Zambia)				**All-probes**			**ZB-probes**			**ZW-probes**		
Ratio	**Target [ng]**	**Map reads**	**Mean read length (bp)**	**H**	**S**	**R**	**H**	**S**	**R**	**H**	**S**	**R**
1	800	113,107	1,503	2,180	21.3M	+	97	954K	+	1	9.5K	-
0.5	540	64,763	1,623	1,360	13.3M	+	67	660K	+	2	19K	-
0.25	320	16,561	1,623	426	4.2M	+	16	157K	+	0	0	-
0.10	150	9,182	1,674	199	1.9M	+	10	99K	+	0	0	-
0.05	80	3,686	1,610	97	949K	+	4	39K	-	0	0	-
0.01	20	1,901	1,740	29	283K	+	0	0	-	0	0	-
0.001	2	289	1,751	8	77K	+	0	0	-	0	0	-
Pg1 isolate (Zimbabwe)				**H**	**S**	**R**	**H**	**S**	**R**	**H**	**S**	**R**
0.25	320	31,633	1,844	606	5.9M	+	0	0	-	13	128K	+
0.10	150	18,848	2,043	385	3.8M	+	0	0	-	10	99K	+
0.05	80	7,072	1,886	129	1.2M	+	0	0	-	6	58K	+
0.01	20	3,457	2,001	65	635K	+	0	0	-	2	20K	-
0.001	2	436	1,983	7	69K	+	0	0	-	1	9.8K	-
0.0005	1	922	2,242	2	19K	-	0	0	-	0	0	-
0.0001	0.2	366	2,028	1	9.5K	-	0	0	-	0	0	-

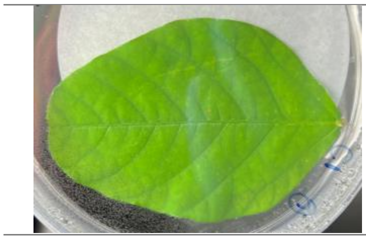
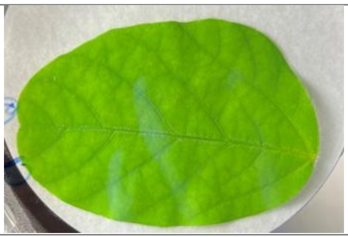
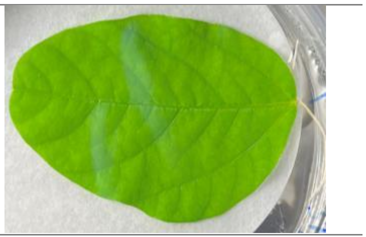
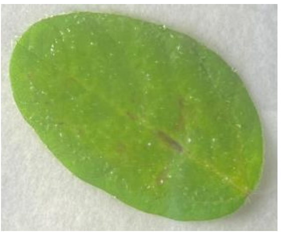
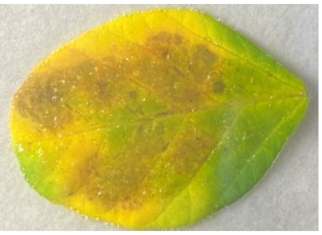
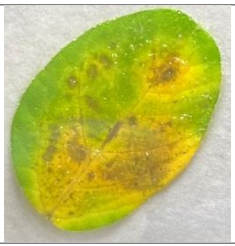
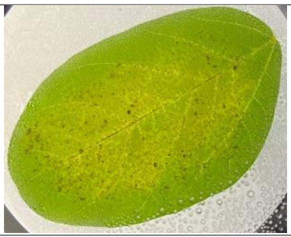
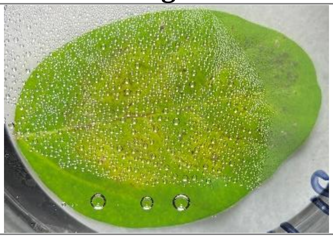
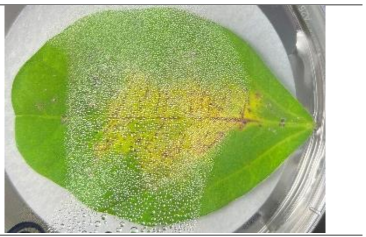
H: Number of hits. S: Score. R: MiFi® Result. M: Million. K: Thousand.

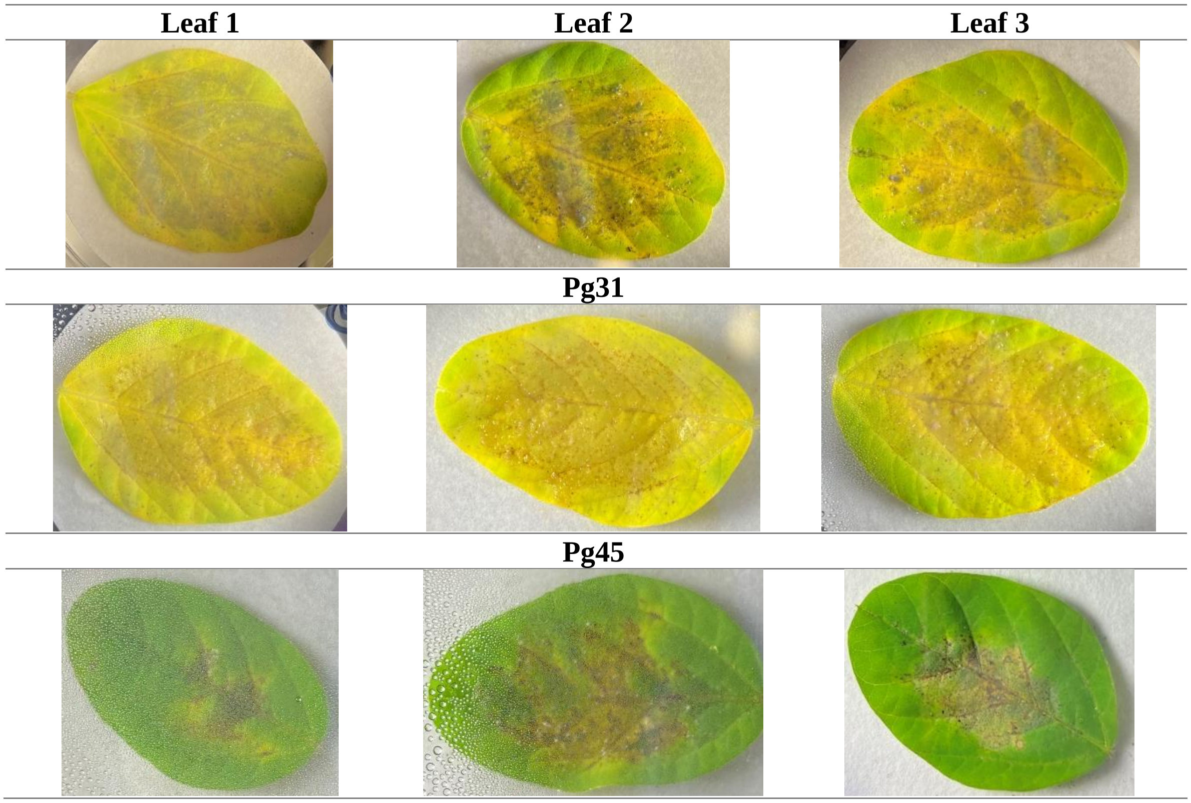
3.4. *In vivo* e-probe validation using HTS data from infected leaves

Symptoms started to develop two days post-inoculation, and at day five, more than 70% of the leaf surface was symptomatic, at which time DNA was extracted. There were 15 symptomatic leaves inoculated with five different *C. glycines* isolates (three leaves per isolate). DNA from negative control leaves was also extracted; these leaves showed no symptoms (Table V-14). The concentration of each inoculum preparation ranged between 1.2x10^{3} and 1.14x10^{5} CFU/ml. We didn’t see any correlation between the mapped reads and the inoculum concentration (Table V-15).

To confirm that the infected leaves were inoculated with *C. glycines* isolates, we performed a confirmatory PCR using species-specific primers (LAC38 and LAC86) developed by the USDA. All the infected leaves were positive, and the control leaves were negative for both reactions (Table V-15). Once the infection with *C. glycines* was confirmed, we sequenced each sample using the ONT sequencing platform and their gDNA-barcoding kit. Two flow cells with nine barcodes each were used. The first included the three replicates of IMI294986, Pg1, and control leaves. At the same time, the second flow cell included the three replicates of RA1, Pg31, and Pg43.

Table V-14: Inoculated leaves (five days post inoculation) from where the DNA was extracted. Control leaves didn't show any changes, while the rest developed red blotches, black leaf dots, and chlorotic halos.

Leaf 1	Leaf 2	Leaf 3
Control leaves		
		
IMI294986		
		
Pg1		
		
RA1		



We evaluated the diagnostic performance of each e-probe set and found that ZB- and ZW-probes sets could not detect the pathogen within the metagenomic samples. However, the All-probe set performed better, reaching a diagnostic sensitivity of 60%. The diagnostic specificity of all three e-probe sets reached 100% since they didn't hit the negative controls (Table V-15).

Table V-15: *In vivo* validation of three designed e-probe sets for detecting *C. glycines.* The concentration (CFU/ml) of the inoculum and the mapped reads and read length are included. Confirmatory PCR was performed using LAC38 (L38) and LAC86 (L86) primers. MiFi® results are summarized, including the number of hits (H), score (S), and result (R). The detection was performed using an e-value of 1e^{-9}.

Sample	CFU/ml	Map reads	Read length (bp)	L38	L86	All-probes			ZB-probes			ZW-probes		
						H	S	R	H	S	R	H	S	R
IMI-1	1.85x10^3	884	2,134	+	+	3	28K	+	0	0	-	0	0	-
IMI-2	1.85x10^3	1918	2,405	+	+	7	69K	+	0	0	-	0	0	-
IMI-3	1.85x10^3	468	895	+	+	1	10K	-	0	0	-	0	0	-
Pg1-1	1.20x10^3	597	1,165	+	+	0	0	-	0	0	-	0	0	-
Pg1-2	1.20x10^3	611	1,629	+	+	3	29K	+	0	0	-	0	0	-
Pg1-3	1.20x10^3	404	727	+	+	1	9.5K	-	0	0	-	0	0	-
RA1-1	3.80x10^4	262	381	+	+	1	9.5K	-	0	0	-	0	0	-
RA1-2	3.80x10^4	542	366	+	+	7	69K	+	0	0	-	0	0	-
RA1-3	3.80x10^4	161	431	+	+	3	29K	+	0	0	-	0	0	-
Pg31-1	1.14x10^5	242	597	+	+	3	30K	+	0	0	-	0	0	-
Pg31-2	1.14x10^5	719	1,351	+	+	17	165K	+	3	29K	+	0	0	-
Pg31-3	1.14x10^5	1,195	1,273	+	+	14	134K	+	1	9K	-	0	0	-
Pg45-1	5.20x10^4	128	919	+	+	4	3.8K	+	0	0	-	0	0	-
Pg45-2	5.20x10^4	56	1,915	+	+	0	0	-	0	0	-	0	0	-

Sample	CFU/ml	Map reads	Read length (bp)	L38	L86	All-probes			ZB-probes			ZW-probes		
						H	S	R	H	S	R	H	S	R
Pg45-3	5.20x10^4	13	455	+	+	1	10K	-	0	0	-	0	0	-
Control-1	NA	0	1,124	-	-	0	0	-	0	0	-	0	0	-
Control-2	NA	0	1,635	-	-	0	0	-	0	0	-	0	0	-
Control-3	NA	0	1,891	-	-	0	0	-	0	0	-	0	0	-

H: Number of hits, S: Score, R: MiFi® Result. M: Million, K: Thousand. L38: Confirmatory PCR using LAC38 primers. L86: Confirmatory PCR using LAC86 primers.

4. Discussion

Identifying and diagnosing plant pathogens through HTS previously required complex data analysis, large-scale alignments, and *de novo* assembly techniques requiring expensive computational resources and bioinformatic expertise (Fang and Ramasamy 2015; Fletcher et al. 2020; Melcher, Verma, and Schneider 2014). In 2013, E-probe Diagnostic Nucleic acid Analysis (EDNA) was proposed as a diagnostic tool that avoids intensive data analysis (Stobbe et al. 2013). This technique has evolved from a theoretical approach to implementation within the MiFi® website. MiFi® is an online tool that has been used to successfully detect and diagnose multiple plant pathogens from raw metagenomic samples (Espindola et al. 2015, 2018, 2022; Espindola and Cardwell 2021; Pena-Zuniga 2020; Proano-Cuenca et al. 2022; Visser, Burger, and Maree 2016). In this chapter, we used MiFi® for the development and validation of e-probes to detect the select agent *Coniothyrium glycines* and to differentiate isolates based on their origin country (Zambia and Zimbabwe). Three 40 nucleotide length e-probe sets were designed, one for each detection purpose (All-probes, ZB-probes, and ZW-probes). The All-, ZB-, and ZW-probe sets included 863, 43, and 17 e-probes, respectively. The reduction in e-probes is explained by the similarity between the exclusion and inclusion panels used for each set. E-probes are generated based on a genome comparison approach. Hence the more similar the sequences, the fewer unique e-probes will be found (Espindola and Cardwell 2021; Stobbe et al. 2013). Based on this and the multilocus phylogenetic analysis performed earlier in this research, we can suggest that Zimbabwean and Zambian genotypes are closely related. We saw that the number of generated e-probes impacted their distribution and coverage of the assembled genomes. The All-probe set was the most homogeneously distributed and with the highest coverage. It is worth mentioning that the designed e-probes address the currently available genomic information of *C. glycines*. If new assemblies are published, these e-probes must be updated to account for genomic variability.

The *in silico* validation of the designed e-probe sets was performed using simulated metagenomic data with nine different target concentrations and two sequencing platforms (ONT and Illumina). We found that the *in silico* LOD increases when the number of e-probes decreases. This result is consistent with previous studies, where designed e-probe sets with a greater number of e-probes require less amount of target concentrations to report a sample as positive (Blagden et al. 2016; Dang et al. 2022; Pena-Zuniga 2020; Proano-Cuenca et al. 2022; Stobbe et al. 2014). When we compared the diagnostic performance of the designed e-probe sets between sequencing platforms based on the mean target base pairs per simulated metagenome, we found that all our designed e-probes are more sensitive with Illumina sequencing data. This result can be explained by the higher accuracy of the Illumina reads (Quail, Swerdlow, and Turner 2009), which increases the score allowing more samples to be classified as positive by MiFi®. Another factor that could influence this result is that although each dilution has the same number of target nucleotides, Illumina, having more reads than ONT, can cover more assembly regions, increasing the chances of matching an e-probe.

The three e-probe sets had a diagnostic specificity of 100%, disregarding the pathogen concentration and the sequencing platform. Explained by the intensive curation process, each e-probe set went through. The curation started with selecting an adequate exclusion panel and removing duplicated and non-specific e-probes after a BLAST against the GenBank nucleotide database (Espindola et al. 2022; Pena-Zuniga 2020; Stobbe et al. 2013). We found that the diagnostic sensitivity changed depending on the target concentration in the metagenomic sample. If the concentration increases, so does the diagnostic sensitivity. Once the LOD was reached, all the designed e-probe sets had a diagnostic specificity of 100%.

We found that the *in vitro* LOD calculated with spiked metagenomic samples differed from the one calculated during the *in silico* validation. The disparity could be explained by the different mean read lengths between the ONT reads of the simulated and spiked metagenomic samples. We found that the LOD for the All-, ZB-, and ZW-probe was two ng, 150ng, and 80 ng, respectively. However, confirmatory PCR detected 0.2 ng of *C. glycines* within a 1,600 ng metagenomic sample. The difference between the LOD of MiFi®-based detection and PCR-based detection relies on the fact that PCR performs an amplification process of the target organism, which increases its concentration and facilitates the detection of the desired organism when in low titer (Sankaran et al. 2010).

Inoculated leaves with different *C. glycines* isolates showed symptoms similar to the ones reported in previous studies and are associated with the disease red leaf blotch of soybeans (RLB) (Hartman et al. 1987; Stewart 1957; Tooley 2017). Additionally, we found that each isolate had different pathogenicity and disease progression, which was already reported by Murithi et al. (2022). The DNA of symptomatic leaves was extracted and sequenced using the ONT sequencing platform. The resulting metagenomic reads were used to test the *in vivo* diagnostic performance of the designed e-probe sets using MiFi®. We found no correlation between the initial concentration of the inoculum and the mapped reads. This statement could be explained by the fact that not all the pathogen inoculated will invade the leaf (Engelthaler and Litvintseva 2020; Karunarathna et al. 2021) and that during the DNA extraction, there is always a general yield loss (Griffin et al. 2002; Jaudou et al. 2022), which will directly impact in the number of reads generated after sequencing. We found that the *in vivo* diagnostic specificity of all the designed e-probe sets was 100%, like the values obtained during the *in silico* and *in vitro* analyses. However, we found that the *in vivo* diagnostic sensitivity of the All-, ZB-, and ZW-probes was 60%, 6.6%, and 0%, respectively. The low diagnostic sensitivity achieved during the *in vivo* validation is explained by the fact that most of the samples didn’t reach the estimated LOD for each e-probe set. We suggest that a higher pathogen concentration is inoculated for future assays or that fewer barcodes are used during sequencing to increase the number of reads per sample.

This new approach to analyzing HTS data from metagenomic samples offers tremendous potential for improving the response to exotic pathogens, including Select agents, during the initial border detection and outbreaks. This is the first time that MiFi® has been used to discriminate the same species from different origin locations, impacting the ability to perform accurate back and forward tracing during a microbial forensic analysis. Future research has to be done to reassess the diagnostic sensitivity of the proposed e-probes and to address the limitations encountered in this study.

References

Blagden, Trenna, William Schneider, Ulrich Melcher, Jon Daniels, and Jacqueline Fletcher. 2016. “Adaptation and Validation of E-Probe Diagnostic Nucleic Acid Analysis for Detection of Escherichia Coli O157:H7 in Metagenomic Data from Complex Food Matrices.” *Journal of Food Protection* 79(4):574–81. doi: 10.4315/0362-028X.JFP-15-440.

Dang, Tyler, Huizi Wang, Andrés S. Espíndola, Joshua Habiger, Georgios Vidalakis, and Kitty Cardwell. 2022. “Development and Statistical Validation of E-Probe Diagnostic Nucleic Acid Analysis (EDNA) Detection Assays for the Detection of Citrus Pathogens from Raw High Throughput Sequencing Data.” *PhytoFrontiers*^{TM} 1–51. doi: 10.1094/PHYTOFR-05-22-0047-FI.

Engelthaler, David M., and Anastasia P. Litvintseva. 2020. “Genomic Epidemiology and Forensics of Fungal Pathogens.” Pp. 141–54 in *Microbial Forensics*. Elsevier.

Espindola, Andres, and Kitty Cardwell. 2021. “Microbe Finder (MiFi®): Implementation of an Interactive Pathogen Detection Tool in Metagenomic Sequence Data.” *Plants* 10(2):250. doi: 10.3390/plants10020250.

Espindola, Andres, Kitty Cardwell, Frank N. Martin, Peter R. Hoyt, Stephen M. Marek, William Schneider, and Carla D. Garzon. 2022. “A Step Towards Validation of High-Throughput Sequencing for the Identification of Plant Pathogenic Oomycetes.” *Phytopathology®* 112(9):1859–66. doi: 10.1094/PHYTO-11-21-0454-R.

Espindola, Andres, William Schneider, Kitty Cardwell, Yisel Carrillo, Peter Hoyt, Stephen Marek, Hassan Melouk, and Carla Garzon. 2018. “Inferring the Presence of Aflatoxin-Producing Aspergillus Flavus Strains Using RNA Sequencing and Electronic Probes as a

Transcriptomic Screening Tool" edited by R. A. Wilson. *PLOS ONE* 13(10):e0198575. doi: 10.1371/journal.pone.0198575.

Espindola, Andres, William Schneider, Peter R. Hoyt, Stephen M. Marek, and Carla Garzon. 2015. "A New Approach for Detecting Fungal and Oomycete Plant Pathogens in next Generation Sequencing Metagenome Data Utilising Electronic Probes." *International Journal Data Mining and Bioinformatics* 12(2):1–14.

Fang, Yi, and Ramaraja Ramasamy. 2015. "Current and Prospective Methods for Plant Disease Detection." *Biosensors* 5(3):537–61. doi: 10.3390/bios5030537.

Fletcher, Jacqueline, Neel G. Barnaby, James Burans, Ulrich Melcher, Douglas G. Luster, Forrest W. Nutter, Harald Scherm, David G. Schmale, Carla S. Thomas, and Francisco M. Ochoa Corona. 2020. "Forensic Plant Pathology." Pp. 49–70 in *Microbial Forensics*. Elsevier.

Griffin, DW, Ca Kellogg, KK Peak, and Ea Shinn. 2002. "A Rapid and Efficient Assay for Extracting DNA from Fungi." *Letters in Applied Microbiology* 34:210–14.

de Gruyter, J., J. H. C. Woudenberg, M. M. Aveskamp, G. J. M. Verkley, J. Z. Groenewald, and P. W. Crous. 2013. "Redisposition of Phoma-like Anamorphs in Pleosporales." *Studies in Mycology* 75:1–36. doi: 10.3114/sim0004.

Gullino, Maria Lodovica, James P. Stack, Jacqueline Fletcher, and John D. Mumford. 2017. *Practical Tools for Plant and Food Biosecurity*. Vol. 8. 1st ed. edited by M. L. Gullino, J. P. Stack, J. Fletcher, and J. D. Mumford. Cham: Springer International Publishing.

Hartman, G., L. Datnoff, C. Levy, J. Sinclair, D. Cole, and F. Javaheri. 1987. "Red Leaf Blotch of Soybeans." *Plant Disease* 113–18.

Hartman, G., and H. M. Murithi. 2022. "Coniothyrium Glycines (Red Leaf Blotch)." *CABI Compendium* CABI Compendium. doi: 10.1079/CABICOMPENDIUM.17687.

Hartman, Glen, James Haudenshield, Kent Smith, and Paul Tooley. 2011. *Recovery Plan for Red Leaf Blotch of Soybean Caused by Phoma Glycinicola*.

Jaudou, Sandra, Mai-Lan Tran, Fabien Vorimore, Patrick Fach, and Sabine Delannoy. 2022. “Evaluation of High Molecular Weight DNA Extraction Methods for Long-Read Sequencing of Shiga Toxin-Producing Escherichia Coli.” *PLOS ONE* 17(7):e0270751. doi: 10.1371/journal.pone.0270751.

Karunarathna, Samantha C., Sajeewa S. N. Maharachchikumbura, Hiran A. Ariyawansa, Belle Damodara Shenoy, and Rajesh Jeewon. 2021. “Editorial: Emerging Fungal Plant Pathogens.” *Frontiers in Cellular and Infection Microbiology* 11.

Melcher, Ulrich, Ruchi Verma, and William L. Schneider. 2014. “Metagenomic Search Strategies for Interactions among Plants and Multiple Microbes.” *Frontiers in Plant Science* 5(JUN).

Morse, Stephen A., and Bernard R. Quigley. 2020. “Select Agent Regulations.” Pp. 425–39 in *Microbial Forensics*. Elsevier.

Murithi, Harun M., Michelle Pawlowski, Tizazu Degu, Deresse Hunde, Molla Malede, Tonny Obua, Hapson Mushoriwa, Danny Coyne, Phinehas Tukamuhabwa, and Glen L. Hartman. 2022. “Evaluation of Soybean Entries in the Pan-African Trials for Response to Coniothyrium Glycines, the Cause of Red Leaf Blotch.” *Plant Disease* 106(2):535–40. doi: 10.1094/PDIS-05-21-1017-RE.

Pena-Zuniga, Lizbeth Daniela. 2020. “EDNA-HOST: Detection of Global Plant Viromes Using High Throughput Sequencing.” PhD, Oklahoma State University, Stillwater.

Proano-Cuenca, Fernanda, Daniel Carrera-Lopez, Douglas Luster, Kurt Zeller, and Kitty Cardwell. 2023. “Genome Sequence Resources for Five Isolates of Coniothyrium Glycines, Causal Pathogen of Red Leaf Blotch of Soybeans.” *PhytoFrontiers^{TM}* 1–15. doi: 10.1094/PHYTOFR-10-22-0113-A.

Proano-Cuenca, Fernanda, Andrés S. Espíndola, and Carla Garzon. 2022. “Detection of Phytophthora, Pythium, Globisporangium, Hyaloperonospora and Plasmopara Species in

High-Throughput Sequencing Data by in Silico and in Vitro Analysis Using Microbe Finder (MiFi®).” *PhytoFrontiers^{TM}* 1–73. doi: 10.1094/PHYTOFR-04-22-0039-FI.

Quail, Michael A., Harold Swerdlow, and Daniel J. Turner. 2009. “Improved Protocols for the Illumina Genome Analyzer Sequencing System.” *Current Protocols in Human Genetics* 62(1). doi: 10.1002/0471142905.hg1802s62.

Roth, Mitchell G., Richard W. Webster, Daren S. Mueller, Martin I. Chilvers, Travis R. Faske, Febina M. Mathew, Carl A. Bradley, John P. Damicone, Mehdi Kabbage, and Damon L. Smith. 2020. “Integrated Management of Important Soybean Pathogens of the United States in Changing Climate” edited by N. Walker. *Journal of Integrated Pest Management* 11(1). doi: 10.1093/jipm/pmaa013.

Sankaran, Sindhuja, Ashish Mishra, Reza Ehsani, and Cristina Davis. 2010. “A Review of Advanced Techniques for Detecting Plant Diseases.” *Computers and Electronics in Agriculture* 72(1):1–13. doi: 10.1016/j.compag.2010.02.007.

Slezak, Tom, Jonathan Allen, and Crystal Jaing. 2020. “Genomics.” Pp. 283–97 in *Microbial Forensics*. Elsevier.

Stewart, Robert B. 1957. “An Undescribed Species of Pyrenochaeta on Soybean.” *Mycologia* 49(1):115–17. doi: 10.1080/00275514.1957.12024619.

Stobbe, A. H., W. L. Schneider, P. R. Hoyt, and U. Melcher. 2014. “Screening Metagenomic Data for Viruses Using the E-Probe Diagnostic Nucleic Acid Assay.” *Phytopathology* 104(10):1125–29. doi: 10.1094/PHYTO-11-13-0310-R.

Stobbe, Anthony H., Jon Daniels, Andres S. Espindola, Ruchi Verma, Ulrich Melcher, Francisco Ochoa-Corona, Carla Garzon, Jacqueline Fletcher, and William Schneider. 2013. “E-Probe Diagnostic Nucleic Acid Analysis (EDNA): A Theoretical Approach for Handling of next Generation Sequencing Data for Diagnostics.” *Journal of Microbiological Methods* 94(3):356–66. doi: 10.1016/j.mimet.2013.07.002.

Tooley, Paul W. 2017. “Development of an Inoculation Technique and the Evaluation of Soybean Genotypes for Resistance to Coniothyrium Glycines.” *Plant Disease* 101(8):1411–16. doi: 10.1094/PDIS-09-16-1373-RE.

United States Department of Agriculture. 2023. *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2022 Summary*.

Visser, Marike, Johan T. Burger, and Hans J. Maree. 2016. “Targeted Virus Detection in Next-Generation Sequencing Data Using an Automated e-Probe Based Approach.” *Virology* 495:122–28. doi: 10.1016/j.virol.2016.05.008.

Wick, Ryan R., Louise M. Judd, and Kathryn E. Holt. 2019. “Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing.” *Genome Biology* 20(1):129. doi: 10.1186/s13059-019-1727-y.

Xuan, Jiekun, Ying Yu, Tao Qing, Lei Guo, and Leming Shi. 2013. “Next-Generation Sequencing in the Clinic: Promises and Challenges.” *Cancer Letters* 340(2):284–95. doi: 10.1016/j.canlet.2012.11.025.

VITA

Daniel Alexis Carrera Lopez

Candidate for the Degree of

Master of Science

Thesis: DETECTION OF THE SELECT AGENT CONIOTHYRIUM GLYCINES, CAUSAL PATHOGEN OF RED LEAF BLOTCH OF SOYBEANS USING HIGH-THROUGHPUT SEQUENCING DATA.

Major Field: Entomology and Plant Pathology

Biographical:

Education:
Completed the requirements for the Master of Science in Entomology and Plant Pathology at Oklahoma State University, Stillwater, Oklahoma in May 2023.

Completed the requirements for the Bachelor of Science in Biotechnology Engineering at Universidad de las Fuerzas Armadas ESPE, Sangolqui, Ecuador in July 2019.

Experience:
Science and Research and Development Leader, 360Life Technologies, Quito, Ecuador, 2020-2021.

Junior Researcher, Oklahoma State University, Stillwater, Oklahoma, September 2019 – December 2019.

Undergraduate Research Assistant, Oklahoma State University, Stillwater, Oklahoma, February 2019 – May 2019.

Professional Memberships:
American Phytopathological Society – APS from 2021.
International Society for Plant Pathology from 2021.