What is your idea? Plant disease epidemics cost farmers and consumers billions of dollars every year, can cause famines and often lead to societal and even armed conflict. But with only a few exceptions, we lack crop disease early warning systems. Meanwhile there are plenty of microbial genetic information already available in public databases. We propose a game-changing monitoring and early warning system for major food security crops based on increasingly accumulating genetic data. **We will harness this information and integrate it with geographic data to automatically map the distribution, spread and evolution of phytopathogens in an open platform that is dynamically updated, as new sequence data becomes available.**

Why is the idea an unconventional or creative approach to the problem outlined in the topic? We argue that there is great potential for genetic data to disrupt current conventional approaches for monitoring pests and diseases. Based on identifying symptoms, the conventional approach to monitoring plant disease outbreaks is very often unreliable, time-consuming and costly. Symptom expression may come too late, after the disease spreads out-of-control and may suggest any number of possible diseases. But genetic identification is fast and the gold standard for pathogen identification. Next Generation Sequencing (NGS) technologies applied to field diagnostics are massively increasing the amounts of data (DNA sequences) that becomes available. We foresee that the availability of genetic data – ***not symptoms*** – will turn out to be the most relevant for early warning of the introduction and potential spread of phytopathogens. Massive sequence information is ever growing and available through different public databases, including the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank. These three organizations exchange data on a daily basis and have at this moment more than 70 million sequences corresponding to bacteria, fungi and viruses. For the first time ever, we will harness big sequence data for mapping and monitoring phytopathogens in near real-time.

Describe the hypothesis for the proposal and why it is expected to succeed. Our hypothesis is that there is sufficient sequence information of phytopathogens stored in public databases that can provide a wider, faster and automated monitoring. Such sequence information will continue growing dramatically, as novel NGS technologies take over classic time-consuming diagnostics tools based on biological isolation, PCR-based approaches and phenotypic evaluation of disease response in specific plant genotypes. We have successfully conducted preliminary tests and identified phytopathogen sequences in unpublished CIAT-datasets, constructed the phylogenetic tree and global maps for Cassava Mosaic Disease (CMD)-associated viruses, a disease that accounts for more than 30% cassava yield losses in Africa. We foresee this approach becoming an early warning system that will allow us to deliver disease alerts to large numbers of smallholder farmers, for example via SMS or through timely communication to National Plant Protection Organizations (NPPOs). With the coming revolution in portable NGS technologies and open access, our initiative will be able to capture and bring the huge DNA sequence resource to the masses.

How will you pilot it? The benefits of an early warning system would include a speed up in the deployment of long-term efforts to improve crop resistance. A proof-of-concept will be carried out for two key diseases affecting tropical crops worldwide, CMD-associated viruses and wheat rust associated fungi. Both pathogens have been extensively studied through conventional diagnostics and they are on the move. The pilot will use existing genetic data from public sequence databases. This will further be complemented by analyzing and

incorporating unpublished sequence data from CIAT plant pathology groups in all regions where CIAT operates. Ground truthing of the results of the automated diagnostics will be done by carrying out targeted surveys of identified disease hotspots. Preventive and confirmatory monitoring will be organized using CRISPR-Dx and NGS technologies, already being standardized in our laboratory.

Describe the implementation plan, including any new technologies or tools that will be developed. We will develop a platform for (1) the automation of genetic information display in form of a dynamic phylogenetic tree and (2) dynamic generation of global maps showing diffusion of phytopathogens. The dynamic phylogenetic tree will be produced in an incremental way; newly available genetic information will be added to the tree without the need of recalculating the whole new phylogeny. The production of the phylogenetic tree will be done utilizing existing machine learning algorithm and techniques, as well as developing and implementing new ones. We will use latitude and longitude coordinates and country information from sequence annotations to automate the geo-referencing. Any new phytopathogen sequence will automatically update maps available on the platform. The phylogenetic trees, geotagged data and other analyzed data will be visualized on the platform using modern, mobile friendly and open source visualization tools such as d3.js, leaflet.js and Three.js. The visualizations will be crafted to provide actionable insights from the analyzed genetic information.

Explain how the work will be performed within the budget (USD$100,000) and time (12 months) allowed? Researchers from CIAT with input from Penn State University's GeoVISTA center will create a technological platform that will filter out, visualize and map genetic data for phytopathogen monitoring. The platform will be hosted on a cloud server and will be tested in target regions using our existing CGIAR network of collaborators in Latin America, Africa and Asia. Target locations will be selected to run sequencing in the field using available NGS technology. We will do this to test real-time uploading and display of sequence information in the in form of dynamic maps and phylogenetic trees. Results of the one year pilot will be published in a reputable scientific journal in Gold Open Access.

What essential data will be generated during this pilot? We will generate spatial data and phylogenetic analysis data that will be automatically updated. Not only that, but we expect that plant pathologists in general will be able to submit their data and have it analyzed automatically. The platform will also automatically create animations of data visualizations that can be used to effectively explain the spread and severity of phytopathogens. We are already developing the two pillars of the platform described above – the phylogenetic tree and global dynamic maps.

If the pilot is successful, what are the next steps? If piloting is successful, one of our first next steps is to engage the plant pathology community and NPPOs in data sharing and preventive monitoring in targeted locations by extending the use of NGS-based diagnostics. We will also include databases of generic plant transcriptome projects, where (as we will show) there is plenty of overlooked phytopathogen information. Our results will complement monitoring technology based on smartphones, satellites, and drones which are limited to symptom expression and cannot solve common biological problems such as mixed pathogen infections, symptomless/early infections and masking of symptoms. We think that our genetic digital monitoring approach will be ground-breaking and easily applied to any other sequence-characterized phytopathogen, demonstrating how open big data and source code can have a practical and global impact for food security.