Project Narrative

1 Training/Career Development Plan

Overview, Goals, Objectives

As environmental pressures mount from current management practices and a changing climate, global food security depends on knowledge of pathogen evolution [1, 2, 3]. As a first year postdoctoral scholar working in the intersection of plant pathology, fungal evolutionary biology, and data science at the University of Nebraska-Lincoln (UNL), I am committed to developing the professional skills, teaching competencies, and training needed to transition to a successful career. My **long-term career goal** is to obtain and develop within a faculty position with both research and teaching responsibilities, that will enable me contribute to plant pathology by both mentoring the future generation of scientists and pursuing my research interests on the evolution of clonal plant pathogens. My **long-term research goal** is to understand how clonal plant pathogen populations evolve in agricultural ecosystems.

The accomplish these long-term goals, it is necessary to further develop my professional skills through focused research training, mentoring, and development. **The goal of this FY 2017 AFRI ELI Postdoctoral Fellowship** application is to provide educational and research training to develop my skills needed to teach and successfully perform genomic scale-research. The **educational training** gained is expected to aid in my acquisition of a faculty position and provide the curriculum for a course that I can implement and teach within that position. The **research training** is expected to compliment my skills as a data scientist and provide me with the reproducible research practices, tools, and techniques necessary to perform evolutionary genomic data analysis. The research project seeks to understand local adaptation of the plant pathogenic fungus *Sclerotinia sclerotiorum* across thermal regions and provide me with results, data, and analyses that I can use to create future educational activities on reproducible research. These will be focused on the topic of reproducible and open research (Fig 1) in agricultural sciences, which will be taught as a special topics course in the Department of Plant Pathology at UNL. A summary of the course will additionally be published in the journal *Journal of College Science Teaching*. To accomplish this goal, I seek completion of the **following training/career development objectives**:

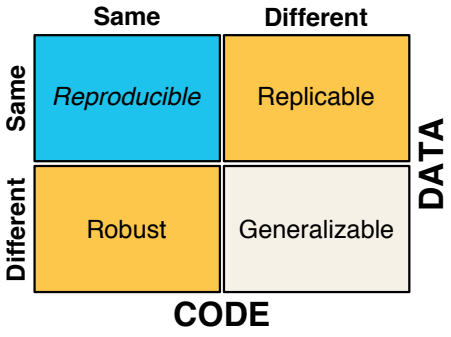

Figure 1: Conceptual definition of computational reproducible research (top left) adapted from Whitaker [4]. If the code and data are the same, they should produce the same results [5].

- **Training Objective 1:** Complete a research training plan for development of population genomic skills
- **Training Objective 2:** Complete a pedagogical training plan for the development of evidence-based education skills
- **Training Objective 3:** Complete a career development plan to facilitate my transition to a tenure-track faculty position

1.1 Training Objective 1: Research Training Plan

Previous Skills, Aptitudes, Training

My Ph. D. dissertation focused on developing open source tools for reproducible analysis of clonal pathogen populations [6]. I am also dedicated to the practice of open and reproducible research as there are many well-documented benefits, including the potential for making agricultural science a more diverse and inclusive field [4, 7]. My research has also largely focused on simple sequence repeat (SSR) data, and I have experience with genome assembly gained by creating a make-based genome assembly pipeline (`https://github.com/zkamvar/read-processing`). I have also received background training in population genetics, evolution, molecular techniques, microbiological techniques, and R programming. I also have experience with reproducible research techniques, methods development, cluster computing, population genetic analysis, software development, multivariate analysis, and sequence alignment.

Opportunities For Skills Development

I do not have substantial experience asking evolutionary questions of large-scale genomic data. I also do not yet have significant experience in coalescent analysis, Bayesian statistics, and phylogeographic analysis. Having these skills will allow me to be more versatile in my research.

Planned Activities

Attend a Workshop on Molecular Evolution at Woods Hole, Massachusetts: this workshop consists of lectures, discussions, and bioinformatic exercises that span contemporary topics in molecular evolution. Training topics specific to my deficiencies include, but are not limited to: phylogenetic analysis (maximum likelihood theory and practice, Bayesian analysis; hypothesis testing), population genetics analysis (coalescence theory, maximum likelihood and Bayesian estimation of population genetic parameters), molecular evolution (gene duplication and divergence, gene family organization), and comparative genomics (genome content, structure, and evolution).
DNA Extraction and Illumina Sequencing: Genomic DNA from 96 *S. sclerotiorum* isolates from China and the United States will be extracted under the supervision of Dr. Sydney E. Everhart and sent for library preparation and sequencing at the Beijing Genome Institute (BGI).
Genomic Data Analysis I will apply the skills gained from the workshop on molecular evolution to define an analysis plan to address the research objectives. Once approved by Dr. Sydney Everhart, the plan will be publicly deposited in the Open Science Framework as a time-stamped Pre-Registration of the project [8].

Expected Outcomes

The planned activities will give me the skills and techniques necessary for high-quality DNA extraction, whole genome sequencing, data storage, and analysis. The direct outcomes of this research will be an understanding of thermal adaptation of *S. sclerotiorum* to sub-tropical climates.

1.2 Training Objective 2: Pedagogical Training

Previous Skills, Aptitudes, Training

I have experience in assignment evaluation, laboratory section management, interactive laboratory instruction, K-12 English as a Foreign Language instruction, and short-form workshop development and instruction.

Opportunities For Skills Development

I do not have significant training active learning techniques, course creation and evaluation, developing goals for specific learning outcomes, and course portfolio development. These skills are essential for effective teaching and would make me an ideal candidate for a teaching and research position.

Planned Activities

Northstar Summer Institute for Scientific Teaching: I will attend this HHMI-funded program in Twin Cities, Minnesota, which will cover tools and techniques for scientific teaching. These include modes of assessment, backward design, creating and sustaining inclusive environments, and the use of current education literature for course improvement.

Course Development: Development of the reproducible research course will begin immediately by defining public data sets to use as examples and developing basic materials for reproducible research.

Peer Review of Teaching: I will join the Discipline-Based Education Research (DBER) community at UNL and present on the proposed course in the DBER seminar series for feedback from the community.

College of Agricultural Sciences and Natural Resources Workshop: I will attend the winter interim Teaching and Learning Workshop focused on retention strategies and learning technologies.

Special Topics Course in Plant Pathology at UNL: I will create a three credit-hour special topics course about reproducible research. This will be held in the department of Plant Pathology and offered to anyone in the agricultural sciences.

Expected Outcomes

Through the planned activities, I will have the skill-set to successfully create, teach, and evaluate a course at the college level. I will also be able to take this course with me to a future tenure-track position.

1.3 Training Objective 3: Career Development

Previous Skills, Aptitudes, Training

I have received previous training in manuscript preparation, peer-review process, research ethics, and outreach methods. I have significant experience in professional writing (CV, résumé, cover letters), written and oral communication of scientific research to professional and lay audiences, and leadership.

Opportunities For Skills Development

To further prepare my career, I require training and skills development in research grant writing, course development, grant budget management, mentorship, and inter-organizational leadership. These are all traits of successful tenure-track faculty.

Planned Activities

Professional Development I will receive training in writing grant proposals, impact statements, and grant-budget management will be provided under the guidance of Dr. Sydney E. Everhart.

Mentorship I will additionally receive training in mentorship of an undergraduate student hired (through the UCARE or IANR Undergraduate Research Scholars programs) for a project (to be

determined) with these data will additionally be provided under the guidance of Dr. Sydney E. Everhart.

Five-year Plan I will develop a five-year plan for future projects, funding sources, and jobs including applications to tenure-track faculty positions.

2 Mentoring Plan

The goal of this proposed integrated project is to provide the fellow with teaching and analytical skill sets that will enable a transition into an independent research and teaching career. Training will be received in three ways: 1) through the teaching and molecular biology workshops as described in the above section 2) through formal monthly evaluation of progress reports, and 3) through informal meetings as needed by the fellow. The monthly meetings will ensure that the fellow is held accountable for all project deliverables (Fig. 5). Both mentors have committed to work closely with the fellow on this project.

The primary mentor, Dr. Sydney E. Everhart, will provide mentoring in all research and career development aspects of this project. The collaborating mentor, Dr. Jenny Dauer, will provide mentoring in the development for the pedagogical skill set. The fellow will be evaluated on organization and completion of project deliverables as well as adaptation to any unforseen pitfalls.

Mentor's Former Mentees and their current positions

- Thomas Miorini, postdoctoral scholar UNL 2016-2017; presently postdoctoral scholar with Dr. Loren Giesler, Department of Plant Pathology, UNL
- Bimal Sajeewa Amaradasa, postdoctoral scholar UNL 2014-2016; presently postdoctoral scholar with Dr. Nick Dufault, Department of Plant Pathology, University of Florida
- Sarah Campbell, undergraduate lab assistant UNL 2014-2016; presently graduate student with Drs. Phil Brannen and Harald Scherm, Dept. of Plant Pathology, University of Georgia
- Morgan Thompsen, undergraduate lab assistant UNL 2016; presently working towards completion of her bachelor's degree in biological science for pre-nursing, UNL
- Josh Hanson, undergraduate lab assistant UNL 2014-2016; presently working towards completion of his bachelor's in Biological Systems Engineering, UNL

3 Project Plan

3.1 Introduction

Principles behind the concept of reproducible research suggest that if a researcher can recreate the results of another individual's study using the same data and analyses, then it is reproducible (Fig. 1) [9, 10, 5]. In theory, all scientific research should be reproducible, but in practice, this is not always the case. A recent survey revealed that up to 50% of researchers were unable to reproduce their own published results and that 90% indicated that there was a "reproducibility crisis" [11]. Despite major journals and funding agencies taking steps to implement incentives for transparency [12], **there is still a need for training on reproducibility across the biological science disciplines** [13]. In the last decade, computational tools have become available that make reproducibility more accessible [7, 14]. Moreover, there are many benefits for scientists who adopt open and reproducible methods including increased citation counts, a greater chance of receiving external funding, and improved relationships among peers potential collaborators [15, 16, 17].

The results produced from agricultural sciences research often have a direct effect on key agri-

cultural stakeholders. As a result, it imperative that research results be replicable. As genomic and microbiome data becomes more readily available, scientists with very little background in bioinformatics are conducting studies with no knowledge of how to process their data [16, 13]. To address the need for training on reproducible research and bioinformatics, **the long term goals** of the proposed project are to: **a)** develop a course on reproducible research for students in agricultural sciences using real-world examples and **b)** create an open and reproducible example of genomic research. The second part will be carried out by testing the **hypothesis** that, due to increased generational times, the plant pathogenic fungus, *Sclerotinia sclerotiorum* is locally adapted to subtropical climates. To complete these goals and test this hypothesis, we seek completion of the following **objectives**:

- **Objective #1:** Create educational materials for graduate students using agricultural data, covering topics from data validation and curation, method choice, version control, and open data science practices (Fig. 1, 2).
- **Objective #2:** Perform whole genome sequencing on 96 isolates of *S. sclerotiorum* collected hierarchically over eight subpopulations across North America and East Asia (Fig. 4).
- **Objective #3:** Perform population genetic, phylogeographic, and coalescent analyses to detect and test signatures of local adaptation to climate regions.
- **Objective #4:** Lead a semester-long special topics course on reproducible research in the department of Plant Pathology at the University of Nebraska-Lincoln.

Need For Proposed Research Project

Sclerotinia sclerotiorum (Ascomycota) is a cosmopolitan, haploid, plant pathogenic soil fungus with over 400 hosts worldwide [18] and causes up to $252M in losses per year on sunflower, soybeans, dry edible beans, canola, and pulse crops [19]. Reproduction is largely asexual, but outcrossing has been known to occur [18, 20]. This fungus can survive for several years in the soil in the form of melanized sclerotia (Fig. 3) [18]. Due to its wide host range and necrotrophism, management of this pathogen largely involves fungicide applications and, to a lesser extent, resistant cultivars [18]. Effective management strategies for any fungal pathogen depends on the knowledge of genetic diversity and adaptive potential [2].

Studies using both phenotypic traits and molecular markers have suggested that *S. sclerotiorum* may be locally adapted to thermal conditions [21, 22]. Numerous studies over the last 30 years around the world show increased diversity in subtropical regions, but Lehner & Mizubuti [23] show that this may be an artifact of low-resolution genetic markers. In the past 30 years, studies have been conducted either across continental or climate regions, but no studies have addressed if the differences observed between *S. sclerotiorum* populations is due to climate or region. In addition, no previous studies have evaluated the population genomic variation of this fungal pathogen, despite more than 50 population studies to date using simple sequence repeat (SSR) markers, which have been shown to poorly resolve haplotype diversity [24].

As *S. sclerotiorum* occurs globally, it is important to determine if populations are adapted to thermal gradients, which could affect dispersal under warming climate conditions [18, 23, 1]. Significant levels of differentiation have been found between populations on canola in the United States (US) and China [25], but these populations occur in different climate regions. To test the hypothesis of local adaptation to thermal environments, we plan to infer genetic diversity, and migration patterns with modern, reproducible population genetic and phylogeographic approaches on

whole genome sequence data.

3.2 Rationale and Significance

AFRI Foundational Priority for Plant Health and Production and Plant Products

The rationale of the proposed work is that **1)** producing educational materials on open and reproducible research will have long-term benefits for agricultural research by increasing access to education and emphasizing practices that will positively influence stakeholder confidence, **2)** the identification of patterns for adaptation in *S. sclerotiorum* will reduce inputs by influencing management decisions regarding movement of inoculum, and **3)** genomic data and analyses from 96 isolates of *S. sclerotiorum* will help drive novel research on this cosmopolitan pathogen.

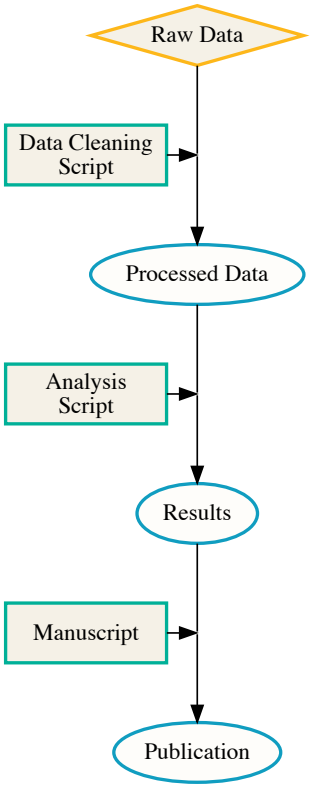

Figure 2: Diagram of an automated reproducible research workflow

The practice of open and reproducible research has been shown to increase the speed and accuracy of which scientific research is conducted [16, 14]. Surveys conducted in 2016 showed that 90% of researchers acknowledge a reproducibility crisis in science [11] and that the most pressing need is for training on data management and integration [13]. Major scientific journals have even implemented checklists [12] and badges [26] in an effort to incentivize reproducibility. Calls for conscious efforts to increase reproducibility have been around for many years [9, 10], but only recently has training been emphasized [17, 16, 14]. By training graduate students seeking masters or doctoral degrees on how to perform open and reproducible research, future agricultural science will be cost effective, innovative, and engaging. This is in line with the teaching mission of the University of Nebraska, where multidisciplinary graduate education is explicitly mentioned as a focal point.

Canola production is a growing industry in the United States, producing three billion pounds of seed grossing at $500 million in 2016 [27, 28]. Losses due to Sclerotinia Stem Rot are estimated to be one percent yield for every two percent disease incidence [29], and the midwest has seen an average incidence rate of 13% [30], which would mean an annual loss of $30 million. It is known that local climate can exacerbate the spread of inoculum and thus, understanding the evolutionary potential for *S. sclerotiorum* to adapt to a warming climate is essential for targeting future management decisions [20, 31, 32].

For over 15 years, population genetic analysis on *S. sclerotiorum* has been performed using the same set of SSR markers [33]. These markers, however, have been unable to accurately determine

Figure 3: Sclerotia of *S. sclerotiorum* (left). Sclerotia can germinate producing apothecia. Ascospores from apothecia serve as primary inoculum. Image and description modified from L. Del Rio Mendoza (`https://www.ag.ndsu.edu/sclerotinia/pictures.html`)

the number of unique genotypes in a sample [24, 23, 34]. Moreover, these markers may suffer from confirmation bias, only able to describe the diversity of *S. sclerotiorum* in the Midwestern United States [25]. Using whole genome sequences of representative isolates across continents allows us to capture a breadth of diversity, giving us a clear picture of the adaptive potential in *S. sclerotiorum* [2]. By understanding the potential for thermal adaptation in *S. sclerotiorum*, we can better predict how it will respond to changes in climate [1].

3.3 Approach

Objective #1: Development of Education Modules

The overall goal of the course is to help students develop thought processes that allow them to adapt to new tools and workflows. The objective of the course is for the students to be able to identify and utilize reproducible research practices including data management, version control, scripting, and archiving with the Open Science Framework [8]. The target audience for this course are graduate students in the agricultural sciences.

Methods - The fellow will use a combination of materials previously developed for other domains of research (i.e. Broman [35]) and small example data sets (i.e. those found in Sparks *et al.* [36]) for development of primary course materials. This course will be split up into four distinct sections, Philosophy of Open Science, Data Management and Ethics, Best Practices in Reproducible Research, and Research Dissemination. The students will be encouraged to use their own data or choose a publicly available data set (i.e. from `https://www.data.gov/`) to ask questions and test their own hypotheses. The course will be taught using the R programming language, but knowledge of the language will not be required [37]. All course materials with a website will be publicly available on GitHub (GitHub, Inc.) and archived on the CERN-funded Zenodo [38].

Expectations - By the end of the first year, the fellow expects to have a publicly available, fully developed course on reproducible research. This work will be licensed in the public domain, allowing other educators to freely use this material in their own courses.

Pitfalls and Limitations - GitHub currently allows open source projects to be hosted free of charge, but since they are a for-profit company, their policy could change at any time. If this risk presents itself, the data are archived at Zenodo and the fellow will host the materials on the Open Science Framework.

Objective #2: Whole Genome Sequencing of 96 *Sclerotinia sclerotiorum* Isolates

Sampling and Experimental Design - A total of 96 isolates split evenly across eight subpopulations of *S. sclerotiorum* will be gathered from canola fields in China and the USA representing temperate regions temperate regions (between $40°N$ and $66°N$) and subtropical regions (between $23.5°N$ and $40°N$) (Fig. 4). Isolates from China will be random samples from Gansu, Qinghai, Anhui, and Hunan provinces originating from Zhou *et al.* [39] and Attanayake *et al.* [25]. Isolates from USA will be from canola fields in North Dakota, Colorado, Georgia, and South Carolina [40, 41]. Prior to DNA extraction, all isolates will be grown on Potato Dextrose Agar (PDA) for 5 days at $24°C$ and transferring four 8mm plugs to an 100mL liquid PDA with shaking for four days at $24°C$ as performed in Derbyshire *et al.* [42].

DNA Extraction - Mycelia will be collected via vacuum filtration, discarding agar plugs. Tissue will be homogenized using liquid nitrogen in sterile mortar and pestles stored -$80°C$. Extraction

will be performed using the DNA Plant Maxi Kit (Qiagen) following manufacturer's instructions yielding into 1mL sterile Milli-Q Water (Millipore). Final DNA concentration will be checked using the Qubit Fluorametric quantification system (ThermoFisher) and quality will be assessed via 1% gel electrophoresis. DNA will be stored at -80°C.

Genome Sequencing and Alignment - Whole genomic DNA will be sent to the Beijing Genomics Institute (BGI) for Illumina (Illumina inc.) library preparation of 150bp paired-end libraries and sequencing. Libraries will be sequenced to $> 15\times$ depth with samples from all populations evenly split on two lanes of Illumina HiSeq 4000 sequencer. Data will be downloaded from BGI, validated by checking the MD5 hash sums, and read-only copies will be uploaded to the Open Science Framework [8]. On UNL's cluster computing infrastructure, reads will be mapped to the *S. sclerotiorum* reference genome using a workflow modified from `https://github.com/zkamvar/read-processing`, which uses Bowtie2 and Samtools to map reads and GATK for variant discovery [43, 44, 45, 42].

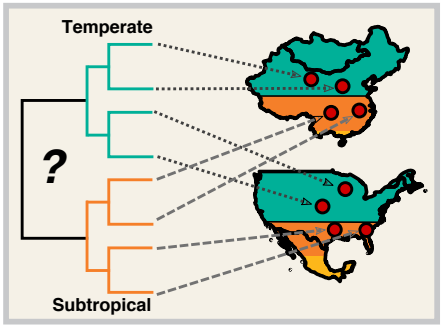

Figure 4: A hypothesis of the evolutionary history of *S. slcerotiorum* in temparate and subtropical regions between North America and East Asia

Expectations - Based on preliminary results from sequencing clonal strains of *S. sclerotiorum* after fungicide exposure we expect to see, on average, $\geq 98.5\%$ coverage with $\geq 8\times$ depth for all isolates. Because the reference genome of *S. sclerotiorum* is from the US, we additionally expect a greater number of variants in the isolates from China.

Pitfalls and Limitations - If the differentiation due to continent is sufficiently large, our alignments of the isolates from China may not efficiently align to the reference genome. We can monitor this by assessing mapping quality for all alignments and performing de-novo assembly on the isolates.

Objective #3: Assessment of Local Adaptation to Climate

Methods - We will use phylogenetic data from each gene independently to assess the evolutionary history of these populations using *Botrytis cinerea* as an outgroup [46]. To address phylogenetic incongruence, we will use newly developed multivariate methods for exploring the space of phylogenetic trees across the genome [47, 48]. Adaptive potential will be assessed using ratios of synonymous and non-synonymous mutation rates, classifying them into positive, neutral, or negative selection, and comparing them to classes of tree topologies (geo-centric, climate-centric, or random) via χ^2 test. Migration rates between populations will be assess using Approximate Bayesian Computation with coalescent simulations. To assess migration among climate and geographic regions, we will use ABCtoolbox [49]. Simulations will be carried out on each gene independently using the `msprime` simulator [50]. Mean nucleotide diversity (π) and mean number of alleles will be used as summary statistics for evaluating simulations. Partial least squares regression and Euclidean distance to observed data will be utilized as the rejection method.

Expectations - Based on initial data from Attanayake *et al.* [25], I expect a majority of genes will segregate for geographic differentiation, but I predict that the majority of the positive selection will be found in the climate-centric genes, whereas the others will be majority neutral.

Pitfalls and Limitations - If the samples are not true representatives of their populations, there

exists a possibility for false positives for a signature of adaptation to climate.

Objective #4: Leading and Evaluating Special Topics Course

In the beginning of the second year, a special topics course will be offered in the Plant Pathology Department at UNL targeted at graduate students in the agricultural sciences.

Methods - The course duration will be one full semester (16 weeks). Course instruction will be designed using an active learning approach [51]. A strict code of conduct emphasizing inclusiveness, accessibility, and anti-harassment will be enforced. By the end of the course, students will be able to define reproducible research, use data management plans, use version control software, and communicate and reproduce their research. Students will be expected to identify a project using their own data or data sets found in public repositories in the first three weeks of the course. To emphasize the importance of communication, the students will be asked to tailor final reports of their projects to a specific audience. Evaluation will be based on participation, in class quizzes, and the successful execution of the final project. Upon completion of the course, a description will be written and submitted to the peer-reviewed publication, *Journal of College Science Teaching*.

Expectations - Each student will have a completed project that is fully reproducible. Because of the emphasis on open science, we also expect at least half of the projects to be openly available. The training in reproducible research and communication has the long-term benefits of advancing scientific progress at a faster rate due in part due to verifiable workflows and increased confidence in communication.

Pitfalls and Limitations - The students will be self-selecting for those interested in reproducible research and open science. To avoid this, advertisements will be tailored to specific departments and will emphasize that no prior knowledge of reproducible research is required.

Hazards - No known hazards are associated with this project.

Timeline - The proposed objectives will be completed within the timeline presented in Figure 5.

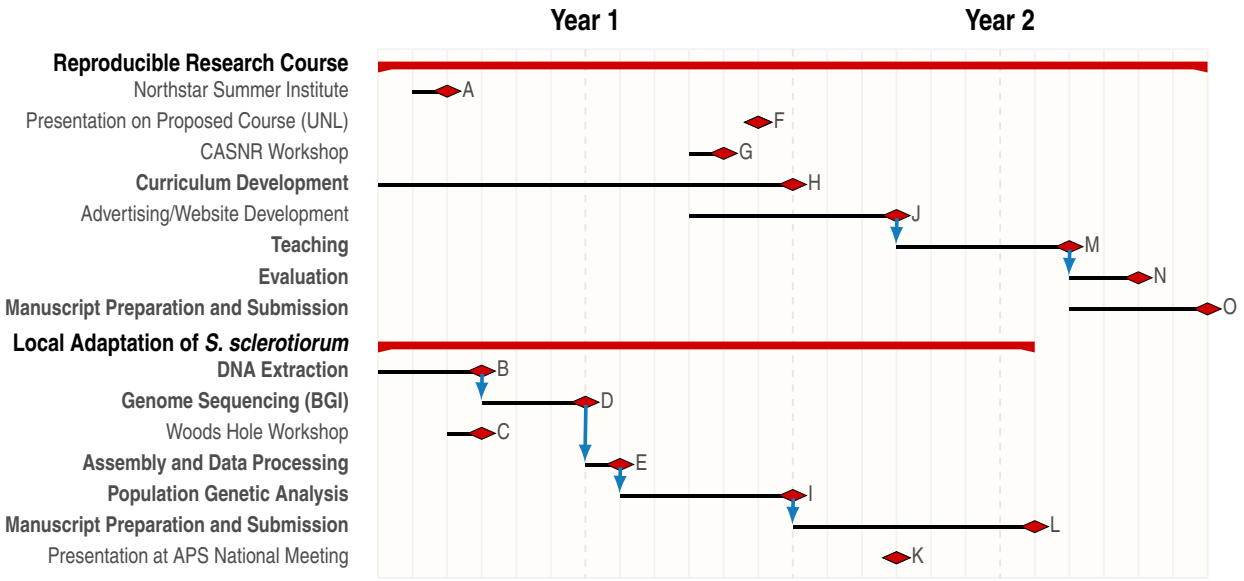

Figure 5: Project Timeline. The milestones **A-O** each represent generation of a measurable outcome used in the evaluation plan.

4 Evaluation Plan

4.1 Progress Evaluation

Monthly progress reports will be written and shared with mentors identifying achieved, in-progress, and outstanding goals as well as challenges and opportunities. During the first year, a quarterly course curriculum update and evaluation will be presented to the project mentors in group meetings. Course progress will be assessed by monthly observations from the project mentors and anonymous surveys from the students.

Milestones

- **6 months:** Genomes sequenced (**B, D**), workshops on teaching and molecular evolution attended (**A, C**), outline for each week of course created, preregistration of research uploaded to the Open Science Framework
- **12 months:** Genomes assembled (**E**), population genetic analysis performed (**I**), workshop on Teaching and Learning attended (**G**)lesson plans for each week of the course completed (**H**), special topics course registered, course website created, presentation at UNL Discipline-Based Education Research seminar held (**F**)
- **18 months:** Presentation at APS National Meeting on research (**K**), course for graduate students underway
- **24 months:** Course completed (**J, M, N**), materials freely available in the public domain, manuscripts submitted as preprints to the Open Science Framework and to both *Molecular Ecology* and *Journal of College Science Teaching* (**L, O**)

4.2 Dissemination Plan

Research Data and Scripts
Research data and scripts will be deposited in the Open Science Framework. All materials will be released with open licenses. Data will be released under the ODC Open Database License 1.0 and code will be released under the MIT License.
Course Materials
Course materials will be hosted on GitHub and archived on Zenodo. To promote sharing and reuse, all applicable materials will be released in the public domain.
Publications
Publications will first be submitted as preprints to the Open Science Framework preprint server. They will additionally be submitted for peer-review in the journals *Molecular Ecology* and *Journal of College Science Teaching*.
Presentations
Presentations will be held in both UNL and the American Phytopathological Society (APS) National Meeting. The presentation at UNL will be on the proposed course on reproducible research presented at the Discipline-Based Education Research seminar. The presentation at the APS National Meeting will be on my research on *S. sclerotiorum*.