Project Description

I. Introduction and Background Section - Understanding the genetic factors that contribute to the reproductive isolation of nascent species is fundamental to evolutionary biology. The biological species concept frames this process as the accumulation of genetic variants that create barriers to gene flow between two populations leading to reproductive isolation[1].

In many cases, plant breeders seek to take advantage of useful alleles present in wild relatives of crop species. Hybrid incompatibilities or other phenomena triggered by wide crosses can limit our ability to utilize broad allelic diversity. A better understanding of the potential barriers to wide crosses will improve our ability to fully harness existing genetic diversity for crop improvement.

The uncoupling of selfish repetitive DNA, such as transposable elements (TEs), and the molecular machinery that epigenetically silences their proliferation is a potential cause of hybrid incompatibility. While the role of TEs in driving speciation is well-characterized[2-4], comparatively little attention has been given to how the interaction between repetitive DNA and their repressor systems may generate hybrid incompatibilities, in an epigenetic paradigm somewhat analogous to traditional Dobzhansky-Muller incompatibilities (DMIs)[5,6].

To address this, I will elucidate the mechanism underlying an instance of hybrid incompatibility between maize (*Zea mays* ssp. *mays*) and their wild relatives, the Mexican teosintes, (*Zea mays* ssp. *parviglumis*) sourced from near Valle de Bravo, Mexico[7]. No abnormal phenotype in F1s, the persistence of the incompatibility despite recurrent backcrossing to maize, and epigenetic changes in hybrids distinguish this case from others. A rapid spread of repetitive DNA, potentially due to the loss of silencing machinery in hybrids, is thought to be causal.

I propose leveraging an interdisciplinary approach that applies third-generation sequencing technology, population genetic analyses, and complex evolutionary modeling to investigate this paradigm. Specifically, I will:

Aim 1) Identify which repetitive DNA sequences expand in advanced generation backcrosses. Advances in long-read, whole-genome sequencing techniques will allow the fine-scale characterization of changes in repetitive DNA. I will sequence advanced generation hybrid backcrosses to maize that maintain sickliness but no direct teosinte ancestry to characterize the extent, placement, and consistency of repetitive DNA expansion in hybrids.

Aim 2) Describe the prevalence of repetitive DNA repressor systems in natural Bravo teosinte populations*. C*ollection of additional teosinte samples from Valle de Bravo will be performed to allow population-level exploration. Repetitive DNA repressor systems are likely to be under at least mild selective pressure to prevent the fitness costs associated with uncontrolled repetitive DNA expansion. These systems will be identified by scanning across DNA sequencing data generated from this collection for signatures associated with positive selection.

Aim 3) Generate a theoretical framework to model the effect of repetitive DNA expansions on population divergence. The PI will expand the traditional Dobzhansky-Muller incompatibility framework to determine the consequence of severing repetitive DNA from their repressor systems. The consequence of this on population divergence will be evaluated by modeling the level of introgression between populations and the relative fitness of hybrids.

The proposed work will greatly expand the understanding of how repetitive DNA and repressor systems contribute to establishing and maintaining population boundaries.

II. Research Objectives, Methods, and Significance -

This proposal, which utilizes interdisciplinary techniques including plant breeding, population genetics, and evolutionary modeling, can be subdivided into 3 distinct areas each with its own design and motivations.

Figure 1. provides a visual overview of all projects contained within the proposal

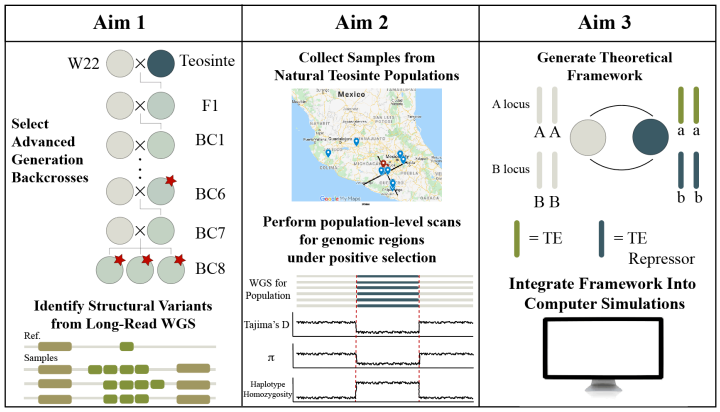

Figure 1. Summary of Projects Detailed in this Proposal

Aim 1: Identify which repetitive DNA sequences expand in advanced generation backcrosses

We hypothesize that the sickly phenotype is a consequence of a TE repressor system failing to function in hybrids, potentially due to a loss of this element entirely, that causes TEs to not only proliferate throughout the hybrid genome but also alter gene expression and methylation states.

Advanced generation backcrosses descending from the initial cross performed by Xue et al. that exhibit hybrid incompatibility can be genotyped and analyzed. Each generation is constructed by crossing males from the previous generation to female W22 maize and is classified by the cumulative number of backcrosses to W22. DNA will be extracted from the leaf tissue of four plants, one belonging to backcross generation six (BC6, n=1) and three belonging to backcross generation eight (BC8,n=3). The BC8 samples are direct descendants of the BC6 sample.

I will isolate long molecules for Nanopore Minion sequencing with Circulomics' Nanobind Plant Nuclei Big DNA Kit on extracted nuclei (following Workman et al.'s 2018 protocol)[8]. Resulting fast files will be aligned using minimap2[9] to the W22 reference genome (ncbi.nlm.nih.gov/assembly/GCA_001644905.2)[10]. Sniffles[11] and SURVIVOR[12] will be used for calling structural variants and comparing these across samples respectively.

Inserted sequences can be identified by calling structural variants between samples and the reference genome. These sequences can be analyzed for the structural components of specific classes of TEs by looking for long terminal repeats (LTRs) or terminal inverted repeat sequences (TIRs) by calling sequence similarity to known TEs. Comparing structural variants across the BC8 samples will be used to determine whether repetitive DNA proliferation continues across generations, and, if so, if there is consistency in where these elements insert.

Aim 2: Determine the prevalence of repetitive DNA repressor systems in natural Bravo teosinte populations

It is currently unknown how prevalent the genetic factors that cause, or protect against, this hybrid phenotype are within natural teosinte populations within Valle de Bravo, Mexico.

The initial project focused on descendants from two teosinte plant samples from this region. Other plants sampled from the surrounding area for other projects do not induce this hybrid decay phenotype. To truly ascertain the frequency of this phenotype and the associated genetic variants, additional samples must be collected.

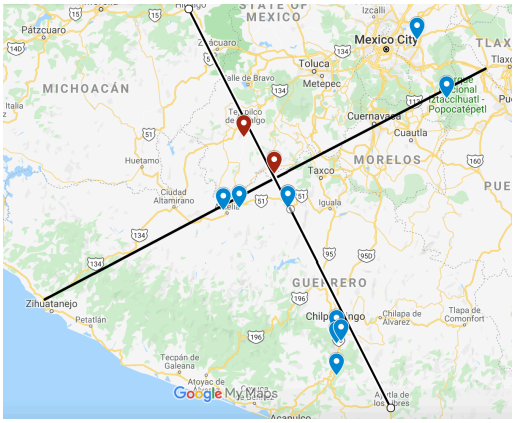

Figure 2. Sampling of teosinte native to the Valle de Bravo region will occur along black transects. Red and blue tags indicate sampling sites of teosinte that respectively either generated hybrid incompatibility or did not.

The PI will collect seeds from teosinte plants along the two transects (~225 miles long) denoted in Figure 2 roughly centered on the initial collection sites for the plants whose lineages exhibited hybrid decay.

Collection sites will be centered approximately 25 miles apart along each transect. All seeds will be collected from 25 plants per collection site.

Seeds will be tagged with location sampled, stored in an envelope, and shipped to UMN.

As hybrid decay is associated with a loss of fitness, genetic machinery that halts the spread of repetitive DNA is likely under selective pressure. Genomic regions under positive selection display several key characteristics compared to natural regions that can be exploited to identify the repressor system.

These characteristics include a reduction in genetic diversity, biases in the frequency distribution of single-nucleotide polymorphisms, and the presence of long haplotype blocks in regions under selection[13-16]. These regions can be identified by selection scans that traverse the genome and collate this information[17].

The PI will germinate 20 seedlings/collection site using standard conditions. Whole above-ground seedlings will be harvested and stored at -80°C before DNA extraction for whole-genome sequencing. High-quality DNA will be extracted, sequenced to a depth of 20X, and single-nucleotide polymorphism variants called using the protocol detailed in Xue et al[7]. PCR assays can also be performed to test for the presence or absence of the repetitive elements identified in Aim 1 as these elements may vary within natural samples.

Samples will be divided into subgroups using ADMIXTURE[18] to compare signatures across distinct groups. If repressor systems evolved to silence repetitive DNA expansion, the repressor systems should show signals of positive selection. Selection scans, that incorporate population genetic metrics such as levels of heterozygosity ($\prod$)[19], Tajima's D[20], and haplotype homozygosity[16,21], will be performed across the genomic data collected from seedlings to identify adaptive regions that should include the repressor system.

In parallel with this work, I will recapitulate the initial crossing scheme that triggered hybrid decay. 15 - 20 plants for collection sites near the intersection of sampling transects and 2-4 plants from the 4 edges of the transects will be planted to generate adult teosinte plants.

W22 maize will be regularly planted, given the differences in flowering time between maize and teosinte, in the hopes of generating reciprocal F1 hybrids. The F1s will then be backcrossed to the parental strains and to each other to generate BC1s and F2s. This will allow us to characterize how spread out this phenotype is among the entire region.

If successful, I will be able to fully characterize the transmission patterns of this hybrid incompatibility, as current work only tracks the transmission in backcrosses to maize. Adult tissue from any successful cross will be stored at -80°C. A variety of sequencing techniques,

including whole-genome sequencing, whole-transcriptome sequencing, and whole-genome bisulfite sequencing will be employed to ascertain which genetic factors are both transmitted and not transmitted and to track epigenetic changes in hybrids.

Aim 3: Generate a theoretical framework to model the effect of repetitive DNA expansions on population divergence

Previous theoretical work has demonstrated that selfish repetitive DNA can readily invade the genome of a species under a variety of conditions[22-24]. To protect itself, the host genome may evolve systems that halt the spread of repetitive DNA. While evolutionary theory has long focused on the spread of repetitive DNA within genomes and throughout populations[25-26], it has ignored the evolution of repressor systems and their role in reproductive isolation.

Traditional evolutionary models that explore DMIs establish two discrete populations with fixed genetic differences at two or more loci that, when unlinked via migration, generate reduced fitness in hybrids[27-29]. I will build upon this framework to explore the consequence of uncoupling repetitive DNA from their repressor system on driving population divergence. By modeling this scenario over time, the environmental and population genetic conditions necessary for these interactions to induce reproductive barriers can be characterized.

The simplest case considers a two-locus, two allele model. At the first locus, the mutant allele *a* represents a transposable element. The second locus houses a recessive mutant allele *b* that encodes silencing machinery that suppresses the transposition of allele *a.* When not coupled with two copies of the silencing allele *b***,** the *a* allele will increase in copy number with probability p_t. Any individual with more than one copy of the *a* allele will incur a fitness cost s_a unless silenced.

Two distinct populations, X and Y, with fixed differences at these two loci will be established. Population X has never been exposed to the transposable element or the silencing machinery, so it houses only *AABB* individuals. The transposable element has invaded Population Y, which in turn evolved the repressor system. As such, it houses only *aabb* individuals.

Migration is then allowed between X and Y at rate *m*, which can unlink the transposable element from its repressor system in hybrids. The amount of introgression between X and Y can be calculated over time to evaluate the extent to which these populations are isolated.

This framework can be simulated and evaluated using SLiM[30], arguably the most versatile tool for forward in-time genetic simulations. SLiM allows the incorporation of several key elements of the model, including migration between populations, epistatic interactions, and the tracking of allele frequencies over time. Crucially, SLiM allows the implementation of “true local ancestry” using marker mutations which allows the calculation of average ancestry for specific genomic regions. This will be used to calculate the level of introgression between populations.

Asymmetrical or sex-biased migration, scaling the fitness cost of transposable elements by the number of copies present in the genome, or modulating the transposition rate are all facets of the model that can be incorporated or adjusted to increase the complexity of the modeling framework. This allows the model to be altered both by new insights into the mechanism gained from experimental work or by the user based on specific points of interest.

Project Significance: The projects detailed within this proposal will greatly advance our understanding of not only the hybrid decay phenotype seen in maize x teosinte hybrids but will also expand evolutionary theory more broadly to characterize the role of TEs and their repressor systems on hybrid incompatibility.

The application of long-read, whole-genome sequencing detailed in Aim 1 leverages novel technology and software to identify the specific classes of TEs that trigger this phenotype and how they insert themselves throughout hybrid genomes.

Additional sampling of teosinte plants from Valle de Bravo can be used for two distinct purposes. Population genetic metrics can be applied to sequencing data generated from these plants to identify the repressor system that averts this decay in natural populations, while the remaining plants can be used in robust crossing schemes to investigate the transmission patterns of hybrid decay.

Analytical modeling and computer simulation will be used to develop theoretical frameworks for evaluating how disrupting TEs from their repressor systems impact hybrid incompatibility and, ultimately, speciation. These models will be refined as we uncover more about the system operation in these hybrids.

III. Training Objectives - My first objective is to gain experience in a novel system, namely *Zea mays,* but also in plants more broadly. Plants exhibit several unique instances of hybrid incompatibility[31], which make them well suited to investigate reproductive isolation and speciation. The ability to participate in fieldwork is another novel opportunity that will greatly expand my ability to engage in and supervise similar projects. Opportunities to develop teaching modules, detailed in section VI, will also further advance my experience instructing students.

My second objective is to expand my computational skill set by analyzing third-generation sequencing data. This skill is highly relevant for modern genomicists as long-read sequencing data can be used to develop high resolution genome assemblies, identify structural variation, and, when applied to transcriptomics, characterize full length isoforms[32].

IV. Sponsoring Scientists and Host Institutions

Dr. Yaniv Brandvain and Dr. Nathan Springer, both from the University of Minnesota (UMN), have been selected as co-sponsors for this project due to their unique experience with distinct areas of the proposal.

Dr. Brandvain has extensively utilized theoretical, comparative, and experimental analyses to study the roles of hybridization, genetic conflict, and geographic isolation play in generating and maintaining plant diversity. Dr. Springer has demonstrated experience investigating how variation in repetitive DNA content and epigenetic states induce phenotypic differences in maize. He also has prior work investigating this particular incompatibility in maize x teosinte.

UMN not only houses a robust community of researchers studying plant genetics, genetic conflict, and speciation but also runs several training programs for postdoctoral scholars to help advance their research, advising, teaching, and leadership skills. These skills will allow the PI to competitively seek a tenure-track position at a tier 1 research institution in the future.

V. Alignment with the National Plant Genome Initiative - Hybrid incompatibility likely played a significant role in plant diversification, with several major crop species showing some level of reproductive isolation from their progenitors[33]. Understanding the genetic underpinnings of these incompatibilities is therefore relevant to understanding the emergence of new plant strains and species. This will play an important role in advancing plant breeding techniques as it allows a more robust prediction of hybrid fitness.
Furthermore, the application of computational tools to tackle research questions in plant systems will only increase in relevance as sequencing costs decrease. Computational pipelines

that analyze large genomic datasets to identify loci influencing various phenotypes along with evolutionary models that will help refine experimental predications will be highly useful for researchers, and both aspects are represented within this project. All pipelines and codes will be shared publicly so that others may adapt these tools for their scientific needs.

VI. Broader Impacts - Computational biology, statistical genetics, and computer modeling are advancing research in plant biology. These new approaches allow us to make use of and interpret large omic datasets. Building on my PhD work in *Drosophila melanogaster* and human computational and population genomics, the training in plant biology I will obtain in this work will leave me well positioned to unite these (at times) divergent literatures/approaches, and advance studies across many taxa.

Outside of the formal classroom setting, I will both develop my skills as a teacher and actively train students in computational techniques by enrolling in the Carpentries (www.carpentries.org) which focuses on best practices in inclusive teaching of computational skills. After this training, I will lead three software carpentry workshops throughout my postdoctoral training.

These workshops are designed to build an inclusive community of researchers well-versed in open, reproducible computational research. My academic training and identity as a queer, woman of color, an identity historically underrepresented within this field, will make me an effective software carpentry trainer. My identity fits well with the Carpentries training goals as, by increasing the diversity of instructors, we can better attract, train, and inspire a strong and diverse generation of trainees.

I will also develop two distinct teaching modules - One for Dr. Brandvain's Biostatistics course (BIOL 3272) and one for Dr. Springer's genetics course (BIOL 4003).

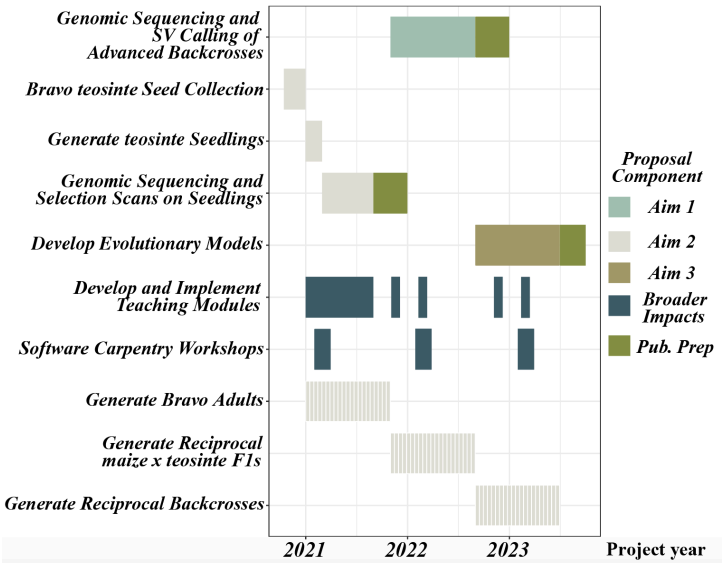

Figure 3. Timeline for Completing Detailed Projects

The first will be an "active learning" exercise that expands on the statistical methodology developed in Dr. Brandvain's course and connects it with advanced methods used to identify genetic variants underlying evolutionarily important phenotypes.

The second will bridge population genetic theory with SLiM simulations to show how evolutionary conditions drive patterns of genetic diversity.

The sum total of this work will help biostatistic and genetic students scale up and see the connection between their classes in small data sets to modern genomic techniques.

The anticipated timeline for all projects, including research aims and broader impacts, are detailed in Figure 3.