NSF-GRFP: RESEARCH STATEMENT

DAVID E. HUFNAGEL

The main focus of this project is the affects of admixture on populations of Zea, the genus containing maize. Maize is the ideal system for this study not only because it is one of the most important world food crops **(citation)**, but also because it is spread over a wide area with many locally adapted populations. Previous studies suggest at least some of these adaptions have been conferred to maize during it's spread across the Americas following domestication in the Balsas River Valley in Mexico **(citation)**. Admixture can allow locally adapted haplotypes to be transferred between populations by a process of repeated hybridization and natural selection.

We have discovered three populations of hybrids between lowland *Zea mays ssp. parviglumis* (hereafter parviglumis) and highland *Zea mays ssp. mexicana* (hereafter mexicana) in the central plateau, northern balsas river valley and southern balsas river valley regions in Mexico. Together parviglumis and mexicana are called teosinte. These hybrid teosinte populations are believed to be in either modern or ancient hybrid zones. Hybrid zones exist on the borders between parapatric populations of different but closely related species where hybrids are easily formed regardless of whether there is a selective advantage to the hybrid phenotype. These hybrid populations have never been studied before and could potentially reveal information about the width and dynamics of hybrid zones as well as the history of maize and it's close relatives in the region near the 2 proposed regions of domestication in the Mexican Central Plateau and the Balsas River Valley **(2 citations)** .

To better understand these hybrid populations I have composed four questions:

(1) Where are these populations distributed across Central and Southern Mexico?
(2) Are these populations stable, locally adapted populations or simply a product of ongoing hybridization between neighboring Zea populations?
(3) What is the relationship of these hybrid populations with each other and their neighboring Zea populations?
(4) If the hybrid populations lie in a true hybrid zone what are the widths of those hybrid zones and how do they compare to the expected width?

I plan to use two datasets to answer these questions. The first is an existing Single Nucleotide Polymorphism (SNP) dataset that has previously been used in three publications **(3 citations)**. This dataset includes 983 SNPs and 2793 individuals from all species and subspecies of Zea all across the Americas as well as some members

of the genus Tripsicum to be used as an outgroup. This data will provide me an opportunity to compare these hybrid populations to a broad spectrum of individuals across many regions and taxa. I also believe that using previously published datasets to answer new questions is a great way to take advantage of the growing pool of underused data freely available on internet databases.

Additionally, I would like to generate Genotyping By Sequencing data (GBS) of hybrid individuals as well as members of nearby populations of teosinte to answer more in-depth questions about the history of these hybrid populations as well as the nature and degree of admixture amongst the hybrid populations and between the hybrid populations and their neighbors. In order to acquire the GBS data I will need to personally do some sampling in the regions where these hybrid populations reside. Due to the high crime rates in both Balsas River Valley regions and the recent Geurrero student massacre I will only collect samples in the Central Plateau region and will be therefore required to limit my study of the Balsas River Valley hybrid populations to what I can gather from the SNP dataset.

To identify the hybrids, I have analyzed the SNP dataset using STRUCTURE. STRUCTURE provides a q-value matrix representing the percent attribution of an individual to a specified number of groups. For the STRUCTURE analysis I used only Mexican samples of maize and teosinte and set the number of groups to 3. Samples of majority mexicana attribution with $\geq 25\%$ parviglumis attribution and samples of majority parviglumis attribution with $\geq 25\%$ mexicana attribution are considered to be hybrids. As these hybrid identifications are based on admixture proportions I plan to confirm them using Reich's F statistic. These tools will allow me to identify the individuals and therefore roughly determine the range of the hybrid populations.

To answer my second question about the origin and stability of the populations, one measure I plan to use is the relative fitness of the populations based on a common garden experiment in the Central Plateau. If these populations are stable, locally adapted populations that are fitting into a niche they should not only be present in the intermediate altitude, but also be more fit there and be less fit in higher and lower altitudes relative to mexicana and parviglumis respectively. I would also like to run the GBS data for the Central Plateau hybrids through HapMix so that I can make a histogram of the lengths of ancestry segments. As ancestry segments of a hybrid should break up over time due to recombination, in the case that these population are stable and resulted from one or more hybridization events close in time there should be a peak of ancestry segments near a specific length. If these hybrids are simply the result of ongoing hybridization the histogram should look roughly like an exponential decay graph as most ancestry segments would be highly broken up and some would be longer indicating many hybridization events including recent events.

To determine the relationship of these hybrids with each other and their ancestors I plan to determine the diversity of these populations using measures of heterozygosity as well as the differentiation of these populations using Wright's Fst. Together these will give us a rough idea of how closely related the individuals in the population are to each other and how closely related the populations are to each other. I also plan to use the D statistic from **person's paper (cite)** to determine whether these groups ancestries are sister to either parviglumis or mexicana, ancestral to them both or or are true hybrids showing equal clustering with both mexicana and parviglumis.

To build a tree of these hybrids and proximal teosinte populations I will use a new software called Treemix **(cite)**. Phylogenetic trees, while useful, are often an oversimplification of the relationships between populations which can include migrations between populations, hybridization between individuals and other events which cannot be represented on a traditional phylogenetic tree. Treemix improves on traditional phylogenetic trees by adding directed and weighted migration edges to improve the statistical likelihood of the tree. Ideally this will reveal the nature of admixture that created all three of these populations including whether they were originally more parviglumis or mexicana and whether some hybrid populations are derived from others. In Addition to a STRUCTURE analysis comparing these groups to each other and their neighboring teosinte it should paint a clear picture of the introgression history amongst these hybrid groups.

Another thing that these data may tell is if there is a cline of attribution to certain teosinte based on the altitudinal gradient and what the width of the hybrid zone is. **(hybrid zone talk)**

To successfully complete this project I will require money for my salary, for GBS and for travel to Mexico. I also depend on the assistance of my major professor Matthew Hufford as well as our collaborators in California, Jeff Ross-Ibara, and in Mexico. A benefit of my work being largely computational and using a previously published dataset is low data generation costs.

To reach out to the greater community I plan to participate in Iowa State University's GK12 program. Through the program I will be sharing my research with children from the Des Moines public school system. I hope that It will be easy to get these kids interested in my work as Iowa's agriculture is dominated by corn. The Des Moines public school system is the most diverse in the state of Iowa in terms of representing ethnic minorities. This is a great opportunity to get schoolchildren, including underprivileged groups, excited about STEM research.

I believe that I am uniquely capable of answering these questions because I have a solid academic foundation in genomics, experience in programming and computational analyses, experienced advisors and collaborators and the drive and curiosity to stay focused on the project.

[works cited]