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. *I think you need a more general introduction here. What doe the literature say about the significance of hybridization and introgression in the evolution of species? How are domesticated species and their wild relatives unique opportunities to study hybridization and introgression given their history and relationship to each other?* 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. *you'll need to be careful to distinguish between maize and teosinte...are we looking at local adaptation and gene flow in maize or in teosinte?* Previous studies suggest some of these adaptions have been conferred to maize during it's spread across the Americas following domestication in either the Central Plateau of Mexico (CPM) or the Balsas River Valley of Mexico (BRVM) **(2 citations)**. *Matsuoka et al. 2002 pretty clearly demonstrated a single origin in the Balsas* Admixture can allow locally adapted haplotypes to be transferred between populations by a process of repeated hybridization and natural selection. *will need to add citation(s) here and consider whether we're talking about adaptive introgression or hybrid zone dynamics*

I have discovered three populations of Mexican hybrids between lowland *Zea mays ssp. parviglumis* (hereafter parviglumis) and highland *Zea mays ssp. mexicana* (hereafter mexicana) in the CPM and in the northern and southern BRVM. *this sounds a bit like you found these by trekking through Mexico rather than identifying them analytically...I'd be a little more specific* Although teosinte represents all maize wild relatives in the genus Zea, for the purposes of this document parviglumis and mexicana in aggregate will be called teosinte. These hybrid teosinte populations are believed to be in either modern or ancient hybrid zones. *there's really very little information out there to indicate the age of hybridization, but this is certainly something you could do as part of your project!* 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 regarding the history of maize and it's close relatives near the 2 proposed regions of domestication in the CNP *again, CNP isn't thought to be a center of origin* and the BRVM **(2 citations)** .

To better understand these hybrid populations I plan to investigate four questions:

(1) Where are these populations distributed across Mexico?
(2) Are these populations stable, locally adapted populations or simply a product of ongoing hybridization between neighboring teosinte?
(3) What is the relationship of these hybrid populations with each other and their neighboring teosinte?
(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 widths?

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 2,793 individuals from all species and subspecies of Zea across the Americas as well as some members of the genus *Tripsacum*.

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 *will need to clearly lay out how higher density data will allow you to answer additional 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 BRVM regions and the recent Guerrero student massacre I will only collect samples in the CPM.

To identify the hybrids, I have analyzed the SNP dataset using the program 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 attribution to one teosinte with $\geq 25\%$ attribution to the other are considered to be hybrids. As these hybrid identifications are based on admixture proportions I plan to confirm them using Reich's F statistic. *will need to explain a bit more about this method and how it is complementary to STRUCTURE* 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 CPM. 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 CPM hybrids through HapMix *more specifics here as well* so that I can make a histogram of the lengths of ancestry segments. As ancestry segments of a hybrid individual should break up over time due to recombination, in the case that these population 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 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.

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 teosinte, ancestral to both or or are true hybrids showing equal clustering with each teosinte.

To build a tree of these hybrids and proximal teosinte populations I will use a new software called Treemix. Phylogenetic trees, while useful, are often an oversimplification of the relationships between populations.Treemix improves on these trees by adding directed and weighted migration edges to improve the statistical likelihood of the tree. In Addition to a STRUCTURE analysis of these groups to each other and their neighboring teosinte it

should paint a clear picture of the introgression history amongst these hybrid groups including whether they were originally more parviglumis or mexicana and whether some hybrid populations are derived from others. *you have a good start toward framing your analyses in the questions, but we'll need to fit them in more fluidly in an evolutionary narrative* .

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. **(more hybrid zone talk)**

To successfully complete this project I will require funding for my salary, for GBS and for travel to Mexico. I will also require the assistance of my major professor Matthew Hufford as well as our collaborators at UC Davis 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 my work will interest these students 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 program is therefore 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 experience in programming and computational analyses, experienced advisors and collaborators, a solid academic foundation in genomics and the drive and curiosity to stay focused on the project.

Works Cited: [works cited]