NSF-GRFP: RESEARCH STATEMENT

DAVID E. HUFNAGEL

Introduction: Hybridization is an underapreciated evolutionary force. Studies suggest most angiosperms have polyploidy somewhere in their history **(Masterson 1994)** which means that many of these organisms have hybrid ancestors. Hybridization improves adaptability through new combinations of alleles and increased heterozygosity **(cite plant that took over england)**. **(hybrid zones are interesting)**

Teosinte is a suitable study system for hybridization and hybrid zones for many reasons, one of which is its relation to maize. Although teosinte represents all wild relatives of maize in the genus Zea, for the purposes of this document parviglumis and mexicana in aggregate will be called teosinte. Maize is the world's most important food crop as well as a well studied system with many genetic resources available. The resources available for maize can be used to aid research in teosinte, which can in turn be used to improve our knowledge of maize. The genome size and complexity of teosinte is roughly average for angiosperms, making it a good representative of this clade **(Gregory 2007)**. Teosinte's two species diverged recently, yet there are clear signs of local adaption as well as ongoing hybridization between each other and with maize within hybrid zones **(Fukunaga 2005)**. It is also of interest that the habitats of parviglumis and mexicana are easily distinguished by their altitude and that the hybrids live in an intermediate altitude. Previous studies suggest some of teosinte's local adaptions have been conferred to maize during it's spread across the Americas following domestication **(cite)**

I have identified three zones of clustered Mexican hybrid populations between lowland *Zea mays ssp. parviglumis* (hereafter parviglumis) and highland *Zea mays ssp. mexicana* (hereafter mexicana) in the Central Plateau of Mexico (CPM) and in the northern and southern Balsas River Valley of Mexico (BRVM). To identify hybrids, I analyzed an existing Single Nucleotide Polymorphism (SNP) dataset that has previously been used in three publications **(Heerwaarden 2010, Heerwaarden 2011, Fang 2012)**. This dataset includes 983 SNPs and 2,793 individuals from all species and subspecies of Zea across the Americas as well as members of the genus *Tripsacum*. To identify hybrids, I have analyzed the SNP dataset using the program STRUCTURE, which 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 three. 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. **(Reich 2011)***will need to explain a bit more about this method and how it is complementary to STRUCTURE* Additionally, our collaborators in Mexico are currently gathering seed in all three zones so that we will have more individuals to analyze for this study.

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 zones have never been deeply studied before, although one hybrid population was mentioned and identified as admixed **(Pyhajarvi 2013)** and others have been briefly mentioned **(Fukunaga 2005)**. **(more hybrid zone talk)**

To better understand these hybrids I have formed 3 objectives:

(1) Determine how these hybrids are distributed across Mexico, the widths of the hybrid zones they reside in and how those widths compare to the expected widths.
(2) Explore whether these zones contain stable, locally adapted populations or are simply a product of ongoing hybridization between neighboring teosinte.
(3) Ascertain the relationship of these hybrids with each other and their neighboring teosinte.

Objective 1: In addition to the SNP dataset, 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 zones as well as the nature and degree of admixture amongst the hybrids and between the hybrid populations and their neighbors. One use for the GBS data would be sampling in a transect in the hybrid zones to investigate the proportion of hybrids and the genetic architecture of hybridization at various altitudes. It would also be interesting to see whether the genetic architecture vary amongst populations within and between zones. Understanding the genetic architecture could reveal whether there are conserved regions amongst different zones or at particular altitudes suggesting that they have been selected for within that niche. In order to acquire the GBS data I will likely need to do some sampling in the regions where these hybrid populations reside. **more in depth on GBS** Due to the high crime rates in both BRVM regions and the recent Guerrero student massacre I will only collect samples in the CPM, leaving sampling in more dangerous areas to our colleagues working in Mexico who have more local knowledge and experience working in these regions.

Another topic of interest is the observed and expected hybrid zone width. **(more hybrid zone talk)**

Objective 2: To answer my question about the origin and stability of the hybrid zones, one measure I plan to use is the relative fitness of the hybrids based on a common garden experiment in the CPM. If these zones contain 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 analyze the GBS data for the CPM hybrids with HapMix so 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 populations 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.

Objective 3: To determine the relationship of these hybrids with each other and their ancestors, I plan to determine the diversity of the zones using measures of heterozygosity as well as the differentiation of the zones using Wright's Fst. Together these will give us a

rough idea of within and between zone relatedness. I also plan to use the D statistic from **person's paper (cite)** to determine whether these hybrid's 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. Along with a STRUCTURE analysis of these groups and their neighboring teosinte, these analyses should paint a clear picture of the introgression history amongst these hybrid groups including whether their ancestry was 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* .

Resources Needed: 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.

Science Impacts:

Broader Impacts: 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 expect 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: **1: Fang** et al. (2012). *Genetics*, doi: 10.1534/genetics.112.138578. **1: Fukunaga** et al. (2005). *Genetics*, doi: 10.1534/genetics.104.031393. **2: Gregory** et al. (2007). *Nucleic Acids Research*, doi:10.1093/nar/gkl828. **3: Heerwaarden** et al. (2010). *Molecular Ecology*, 19(6) 1162-1173. **4: Heerwaarden** et al. (2011). *PNAS*, 108(3) 1088-1092. **5: Masterson** (1994). *Science*, 264(5157) 421-424. **6: Pyhajarvi** et al. (2013). *Genome Biology and Evolution*, 264(5157) 421-424. **7: Reich** et al. (2011). *The American Journal of Human Genetics*, 89 516-528.