NSF-GRFP: RESEARCH STATEMENT

DAVID E. HUFNAGEL

Introduction: Hybridization is an underappreciated 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 an ideal system for studying hybridization and hybrid zones. Although teosinte represents all wild relatives of maize in the genus Zea, for the purposes of this document *Zea mays ssp. parviglumis* (hereafter parviglumis) and *Zea mays ssp. mexicana* (hereafter mexicana) in aggregate will be called teosinte. One reason teosinte is a good study system is that the habitats of *parviglumis* and *mexicana* are easily distinguished by their altitude with hybrids inhabiting intermediate altitudes between the two subspecies. 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 subspecies diverged recently, yet there are clear signs of local adaptation as well as ongoing hybridization between each other and with maize within hybrid zones (Fukunaga 2005). Previous studies suggest some of teosinte's local adaptations have been conferred to maize during it's spread across the Americas following domestication **(cite)**. Maize is the world's most important food crop as well as a well studied system with many available genetic resources. These resources can be used to aid research in teosinte, which can in turn be used to improve our knowledge of maize.

I have identified three zones of clustered Mexican hybrid populations between lowland *parviglumis* and highland *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 available Single Nucleotide Polymorphism (SNP) dataset (Heerwaarden 2010, Heerwaarden 2011, Fang 2012) including 983 SNPs genotyped in 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. This program uses a model-based approach to infer population structure and assign individuals to populations using multilocus genotype data (Pritchard 2000). **this was derived from the first line in the paper's abstract, is that okay?** Samples of majority attribution to one teosinte with a significant ($\geq 25\%$) attribution to the other are considered to be hybrids. I plan to confirm these identifications using Reich's f_3 ancestry estimation method. (Reich 2011)

Hybrid zones often form on the borders between parapatric populations of closely related populations 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). My expectation is that the width of these hybrid zones is largely determined by the steepness of the altitudinal gradient because the niches of *parviglumis* and *mexicana* are heavily dependent on altitude. **(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) Ascertain the relationship of these hybrids with each other and their neighboring teosinte.
(3) Explore whether these zones contain stable, locally adapted populations or are simply a product of ongoing hybridization between neighboring teosinte as in a tension zone.

Objective 1: How extensive are teosinte hybrid zones? In addition to the SNP dataset that is already available, 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. While we could use the same SNP-chip used to obtain our SNP dataset on new samples, GBS provides much higher resolution data at a lower cost. In order to acquire the GBS data I will likely need to 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, leaving sampling in more dangerous areas to our colleagues working in Mexico who have more local knowledge and experience working in these regions.

One use for the GBS data would be sampling in a transect of the hybrid zones to investigate the proportion of hybrids at various altitudes. 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. Another topic of interest is the observed and expected hybrid zone width. **(more hybrid zone talk)**

Objective 2: Is the genomic architecture of hybridization conserved? 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 (Durand 2011) to determine whether these hybrids are sister to either teosinte, ancestral to both or are true hybrids showing equal clustering with each teosinte. It would also be interesting to use the GBS data 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.

Objective 3: Is the hybrid phenotype adaptive? 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 use HapMix (Price 2009) with our GBS data to build a histogram of the lengths of ancestry segments. As ancestry segments of a hybrid individual should break up over time due to recombination, if 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.

To build a tree of these hybrids and proximal teosinte populations I plan to use a software called Treemix (Pickrell 2012). 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.

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 an available dataset is low data generation costs.

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.

Works cited: **1: Durand** et al. (2012). *Molecular Biology and Evolution*, doi: 10.1093/molbev/msr048. **2: Fang** et al. (2012). *Genetics*, doi: 10.1534/genetics.112.138578. **3: Fukunaga** et al. (2005). *Genetics*, doi: 10.1534/genetics.104.031393. **4: Gregory** et al. (2007). *Nucleic Acids Research*, doi:10.1093/nar/gkl828. **5: Heerwaarden** et al. (2010). *Molecular Ecology*, 19(6) 1162-1173. **6: Heerwaarden** et al. (2011). *PNAS*, 108(3) 1088-1092. **7: Masterson** (1994). *Science*, 264(5157) 421-424. **8: Pickrell** (2012). *PLOS Genetics*, DOI: 10.1371/journal.pgen.1002967. **9: Price** (2009). *PLOS Genetics*, DOI: 10.1371/journal.pgen.1000519. **10: Pritchard** (2000). *the Genetics Society of America*, 155: 945-959. **11: Pyhajarvi** et al. (2013). *Genome Biology and Evolution*, 264(5157) 421-424. **12: Reich** et al. (2011). *The American Journal of Human Genetics*, 89 516-528.