Demography and selection during maize domestication

Timothy M. Beissinger [*], **Li Wang** [†], **Arun Durvasula** [*], **Kate Crosby** [*], **and Jeffrey Ross-Ibrarra** [*] [†] [‡]

[*]University of California,Davis,[†]Iowa State University, and [‡]Center for population biology, UC Davis

Submitted to Proceedings of the National Academy of Sciences of the United States of America

This is the abstract. It should probably be somewhere around 200 words.

This is the beginning of the article. Notice the dropcap... that is a neat feature that PNAS likes, I think it looks pretty neat too! This is where the introduction will live.

Results

Patterns of variability differ between genic and nongenic regions of the genome. Text goes here

Hard sweeps do not shape maize (or teosinte) diversity. A mutation that is immediately beneficial and positively selected leaves a classical hard sweep signature in the genome, whereby genetic diversity surrounding that mutation is reduced as the haplotype where the mutation first arose increases in frequency to fixation. The prevalence of such hard sweeps was evaluated by comparing diversity surrounding non-synonymous and synonymous substitutions in maize and teosinte since divergence from tripsicum, roughly ZZZ years before present (ref here). For both taxa, no difference in diversity surrounding these classes of substitutions was identified (figure here). Additionally, diversity around maize substitutions not seen in teosinte was investigated. These sites have the potential to correspond to recent maize sweeps, occuring after the split from teosinte. Again, no difference was observed between diversity around synonymous and nonsynonymous substitutions (Figure here). Together, these observations suggest that hard sweeps are not a primary form of selection for either maize or teosinte.

Patterns of purifying and background selection can be explained by demography. Purifying selection refers to the situation where deleterious mutations arising in a population are continuously selected against. When this form of selection is operating, it can serve to reduce genetic variability at linked neutral sites, a phenomenon referred to as background selection [11]. Purifying and background selection lead to lower diversity within genes and other functional sites relative to neutral regions. We investigated purifying selection in maize and teosinte by evaluating the average magnitude of reduced diversity within genes and recovery away from genes in both taxa (figure here). When standardized by neutral levels (diversity far from genes), a stronger reduction of diversity and slower recovery was observed for teosinte than for maize, implying that purifying selection has left a more pronounced signature in the teosinte genome. This conflicted with our *a priori* hypothesis; we expected that strong artificial selection since domestication would have caused enhanced purifying selection for maize. We therefore conducted the same analysis based on singleton diversity. As a class, singleton alleles depict the most recent patterns of evolution, but also have the lowest effect on pairwise diversity. Therefore, unlike pairwise diversity, patterns of singleton diversity reflect recent patterns of evolution. When evaluating the data in this manner, an opposite relationship was observed. Maize singleton diversity was lower than teosinte singleton diversity near genes, and recovered more slowly (figure here), implying that in the recent past maize has been more influenced by purifying selection than teosinte. Together, these findings imply that demographic history has a strong influence on the effect of purifying selection. Historically, teosinte has had a larger population size than maize, and only recently has maize population size overcome that of teosinte. Since the efficacy of purifying selection scales with population size, these results likely reflect changes in N_e more than they reflect underlying changes in selection pressure.

This might be a good place to stick in a paragraph about integrating to estimate BGS based on Jeff's work.

Demography of maize domestication. Text goes here

Discussion

Discussion 1. Text goes here

Discussion 2. Text goes here

Materials and Methods

Plant materials. Accessions studied were selected from the Maize HapMap2 panel [2] . Principal component analysis was employed to ensure that closely related individuals were not included due to their potential to bias results (maybe a supplemental figure here). Ultimately, 23 maize inbreds derived from a diverse assortment of landraces were selected for inclusion. Thirteen teosinte inbred lines, all members of the subspecies Z. *mays* ssp. *parviglumis*, were utilized. Sequences were mapped to the maize B73 version 3 reference genome [3] (ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/dna/).

Interpolating genetic position. For many of the following analyses, physical position along a chromosome was a less relevant measure of map location than was genetic position. Therefore, physical positions were converted to genetic positions by interpolating from the NAM genetic map (REFERENCE), which provides a 1 cM resolution for physical to genetic conversion. Within R [5], physical positions with corresponding genetic positions in the NAM map were used as anchors. Physical positions in our dataset wihout corresponding genetic position were assigned genetic positions by scaling the anchored genetic positions according to the physical distance between the unlabeled position and the flanking anchors.

Significance

This work is insignificant ;-)

Reserved for Publication Footnotes

Estimating the site frequency spectrum. To estimate individual and joint site frequency spectra (SFS) for maize and teosinte from inbred lines, each inbred individual was treated as representing a single haplotype from its population. These were separately computed for genic and intergenic regions, as well as for the whole genome together. First, genic and intergenic regions were isolated using the biomaRt package [6, 7] of R [5]. Genic regions were defined as DNA between the start and stop position of a gene, while intergenic regions were required to be at least 5kb up- or down-stream from a gene start/stop. With regions defined, SFS were estimated with ANGSD [4]. Individual population SFS were estimated using all positions observed in at least 80% of the individuals in the population, and joint SFS were estimated using all positions observed in at least 80% of individuals in both populations. Individuals were assumed to be fully inbred (-doSaf 2), and subsequently allele frequencies were divided by two to indicate haplotype frequencies. Quality filters were employed such that reads with quality score below 30 and bases with quality score below 20 were discarded (-minMapq 30 and -minQ 20), as were reads that didn't map uniquely (-uniqueOnly 0). Quality scores around indels were adjusted as in Samtools (-baq 0). Genotype likelihoods were estimated using the samtools method (-GL 1). Major and minor alleles were inferred from the data (-doMaf 1). Because ANGSD cannot calculate a folded joint SFS, the maize reference genome was used for polarization and then unfolded spectra were folded using dadi [1].

Demographic inference.

Evaluating diversity around substitutions. To investigate diversity around substitutions, maize and teosinte pairwise diversity was first calculated in 1,000 kb non-overlapping windows using ANGSD [4]. This was performed separately for both maize and teosinte, using the same filters as employed for estimating the SFS. Next, SNPs and genotypes among maize, teosinte, and tripsicum were called. Tripsicum bam files were downloaded from (TRIPSICUM FILES), and then all SNPs with a p-value less than 1e-6 were called using ANGSD. Quality filters were as the same as before, and genotypes were only called when the posterior probability was above 0.95. From the set of called SNPs and genotypes, substitutions between maize and tripsicum, as well as between teosinte and tripsicum were identified using R [5] as all positions with no more than 20% missing data for which every maize or teosinte allele differed from the observed tripsicum allele. At each class of substitution, effects were estimated using the ensembl variant effects predictor [8].

For each diversity window with at least 100 bps observed, the distance from the window center to the nearest synonymous and nonsynonymous (missense) substitution was computed. Then, following the methods of [9], a loess curve was plotted for diversity values against the distance to the nearest synonymous or nonsynonymous substitution. A span of 0.01 was utilized. Unlike [10], we did not fit separate loess curves in the up- and down-stream directions, but instead fit single curves encompassing both directions.

Evaluating diversity around genes and conserved sequences. Two types of diversity surrounding genes were investigated. The first was pairwise diversity in 1kb windows, as described previously. The second was singleton diversity in 1kb windows. Singletons represent the rarest class of alleles that this dataset can identify, and collectively demonstrate the most recent patterns of evolution. Minor allele frequencies were estimated with ANGSD [4] using the same quality filters previously described. Then, the number of singletons in each non-overlapping 1kb window was calculated with R [5]. BiomaRt [6, 7] was then used to identify the center of each gene. Next, the distance from each diversity window to the nearest gene center was computed. Teosinte diversity is generally higher then maize diversity. Therefore, to enable comparisons between the reduction of diversity around genes in maize and teosinte, a neutral measure of pairwise and singleton diversity for each taxa was estimated according to mean pairwise and singleton diversity at windows greater than 0.01 cM from the nearest gene. Then, pairwise and singleton diversity at each window was standardized by dividing by the corresponding neutral measure. Separately for pairwise and singleton diversity in maize and teosinte, cubic smoothing splines were fit to describe diversity levels according to the distance to the nearest gene. Significant differences were assessed by taking 100 bootstrap samples and re-fitting the cubic smoothing spline to each. Then, the 2.5% and 97.5% quantiles of values along the bootstrapped splines were identified.

Simulations.

ACKNOWLEDGMENTS. Various thankyous will be in order.

1. **R. Gutenkunst, R. Hernandez,S. Williamson, and C. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data, PLoS Genetics., 5:10 (2009), e1000695.**
2. **J. Chia, C. Song, P. Bradbury, D. Costich, N. de Leon, and others, Maize HapMap2 identifies extant variation from a genome in flux, Nature Genetics., 44 (2012), 803-807.**
3. **P. Schnable, D. Ware, R. Fulton, J. Stein, F. Wei, The B73 maize genome:complexity, diversity, and dynamics, Science., 326:5956 (2009), 1112-1115.**
4. **T. Korneliussen, A. Albrechtsen, and R. Nielsen, ANGSD: Analysis of next generation sequencing data, BMC Bioinformatics., 15:356 (2014), 10.1186/s12859-014-0356-4.**
5. **R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing., Vienna, Austria.**
6. **S. Durinck, P.T. Soekknabm E. Birney, and W. Huber, Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt, Nature Protocols., 4 (2009), 1184-1191.**
7. **S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, and others, BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis, Bioinformatics, 21 (2005), 3439-3440.**
8. **W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor, Bioinformatics, 26:16 (2010), 2069-2070.**
9. **R.D. Hernandez, J.L. Kelley, E. Elyashiv, S.C. Melton, A. Auton, and others, Classic selective sweeps were rare in recent human evolution, Science, 331:6019 (2011), 920-924.**
10. **S. Sattath, E. Elyashiv, O. Kolodny, Y. Rinott, and G. Sella, Pervasive adaptative protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans, PLoS Genetics, 7:2 (2011), e1001302.**
11. **B. Charlesworth, M. T. Morgan, and D. Charlesworth, The Effect of Deleterious Mutations on Neutral Molecular Variation, Genetics, 134 (1993), 1289-1303.**