Demography and selection since maize domestication

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

[*]University of California,Davis,[†]Iowa State University,[‡]Genome Center, UC Davis, 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. Previous research of the maize domestication process has relied upon observations drawn from genic DNA (several references here). Our data were generated through whole genome sequencing, which eliminated this constraint. Importantly, we observed substantial differences in patterns of diversity between genic and non-genic regions of the genome for both maize and teosinte. For maize, mean pairwise diversity (π) within genes was significantly lower than at positions at least 5kb away from genes (0.00668 within, 0.00691 away, p<2e-44). The same pattern of significantly greater diversity outside of genes was observed in teosinte, but to a larger extent. For teosinte we observed π within genes to be 0.0088 and at least 5kb from genes it was 0.1153 (p≈0). These observrvations suggest that genes are not evolving neutrally in maize or teosinte. Instead, some form of selection must be reducing diversity within genes. Additionally, these observations suggest that demographic inference, which relies on the assumption of neutrally evolving DNA, will be more accurate if it is based on observations from non-genic genetic material.

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). 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 called background selection [Charlesworth et al.(1993)Charlesworth, Morgan, and Charlesworth]. Purifying and background selection lead to lower diversity within genes and other functional sites relative to neutral regions (ref here). 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

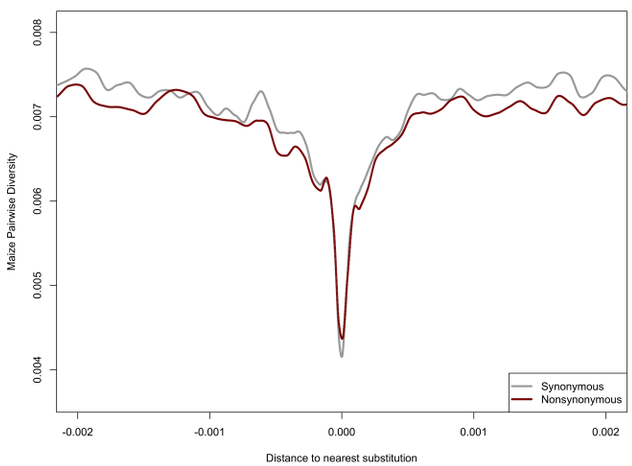

Fig. 1. Pairwise diversity surrounding synonymous and nonsynonymous substitutions in maize. Observe that the reduction of diversity surrounding nonsynonymous subsitutions is no more severe than the reduction of diversity surrounding synonymous substitutions. This equivalency suggests that hard sweeps are not a primary driver of maize diversity.

Significance

This work is insignificant ;-)

Reserved for Publication Footnotes

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 curve fitting B to estimate s, mu.

Demography of maize domestication. To explore whether differences in purifying selection between maize and teosinte can be explained purely by demographic processes, we first estimated the parameters of maize domestication using large-scale sequencing data involving 23 inbred maize landraces and 13 teosinte inbred lines included in the HapMap 2 panel [Chia et al.(2012)Chia, Song, Bradbury, Costich, de Leon, Doebley, Elshire, Gaut, Geller, Glaubitz, et al.] (supplemental table here). The maize lines were collected from across the Americas and the teosinte lines came from central Mexico. Before estimating demography, we compared the site frequency spectrum (SFS) in genic and non-genic regions, and observed substantial differences in the evolution of these classes of sites. For both maize and teosinte, the SFS within genes showed a dearth of low-frequency alleles and Tajima's D (reference here) was therefore shifted to more positive values (figure here). This is consistent with the aforementioned purifying selection and indicates that genic regions are not evolving neutrally. Therefore, for demographic modeling we restricted analysis to non-genic sites.

We used diffusion approximation as implemented in dadi [Gutenkunst et al.(2009)Gutenkunst, Hernandez, Williamson, and Bustamante] to find the domestication parameters that best explain the joint site frequency spectrum of maize and teosinte. The model we optimized began with an equilibrium ancestral population of size N_a splitting into separate maize and teosinte populations T_B generations in the past. Moving forward in time from the split, teosinte maintained the ancestral population size, while maize experienced an immediate effective population size change to N_b individuals, followed by exponential growth to size N_m. Additionally, since the split a M_{tm} individuals migrated from teosinte into maize and M_{mt} individuals migrated from maize to teosinte.

The most likely model suggested an ancestral for θ $(4N\mu)$ of 0.014734, which is similar to previous estimates (references here). Assuming a mutation rate of $\mu = 3 \times 10^{-8}$ (reference here), this suggests an effective population size of $N_a = 122,783$ individuals, a population split $T_B = 15,523$ generations in the past, a maize bottleneck effective populations size of $N_b = 6,455$ individuals (5.26% of N_a), a modern maize effective population size of $N_m = 366,973$ individuals, $M_{tm} = 1.35$ migrants per generation from teosinte to maize, and $M_{mt} = 1.72$ migrants per generation from maize to teosinte (figure here). We note that a population split over 15 thousand generations before present precedes estimates from archaeological data which suggests maize domestication began approximately 9,000 years before present (reference here). This could result from multiple generations per year, or in may reflect teosinte population structure that was present before domestication. Also, note that the genetic time of the population split must precede morphological changes that could be identified morphologically. Additionally, because recent expansion is most evidenced by rare alleles, and because these data provide low power to detect rare alleles, we expect that the estimate of $N_m = 366,973$, or $\sim 3N_a$, is likely an underestimate. (This is where evaluation using Kate's SFS should go).

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 [Chia et al.(2012)Chia, Song, Bradbury, Costich, de Leon, Doebley, Elshire, Gaut, Geller, Glaubitz, et al.] . 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 [Schnable et al.(2009)Schnable, Ware, Fulton, Stein, Wei, Pasternak, Liang, Zhang, Fulton, Graves, et al.] (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 [R Core Team(2014)], 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.

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 [Durinck et al.(2009)Durinck, Spellman, Birney, and Huber,Durinck et al.(2005)Durinck, Moreau, Kasprzyk, Davis, De Moor, Brazma, and Huber] of R [R Core Team(2014)]. 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 [Korneliussen et al.(2014)Korneliussen, Albrechtsen, and Nielsen]. 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 [Gutenkunst et al.(2009)Gutenkunst, Hernandez, Williamson, and Bustamante].

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 [Korneliussen et al.(2014)Korneliussen, Albrechtsen, and Nielsen]. 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 [R Core Team(2014)] 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 [McLaren et al.(2010)McLaren, Pritchard, Rios, Chen, Flicek, and Cunningham].

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 [Hernandez et al.(2011)Hernandez, Kelley, Elyashiv, Melton, Auton, McVean, Sella, Przeworski, et al.], 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 [Sattath et al.(2011)Sattath, Elyashiv, Kolodny, Rinott, and Sella], 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 [Korneliussen et al.(2014)Korneliussen, Albrechtsen, and Nielsen] using the same quality filters previously described. Then, the number of singletons in each non-overlapping 1kb window was calculated with R [R Core Team(2014)]. BiomaRt [Durinck et al.(2009)Durinck, Spellman, Birney, and Huber, Durinck et al.(2005)Durinck, Moreau, Kasprzyk, Davis, De Moor, Brazma, and Huber] 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.

Charlesworth et al.(1993)Charlesworth, Morgan, and Charlesworth. B. Charlesworth, M. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. Genetics, 134(4):1289–1303, 1993.

Chia et al.(2012)Chia, Song, Bradbury, Costich, de Leon, Doebley, Elshire, Gaut, Geller, Glaubitz, et al. J.-M. Chia, C. Song, P. J. Bradbury, D. Costich, N. de Leon, J. Doebley, R. J. Elshire, B. Gaut, L. Geller, J. C. Glaubitz, et al. Maize hapmap2 identifies extant variation from a genome in flux. Nature genetics, 44(7):803–807, 2012.

Durinck et al.(2005)Durinck, Moreau, Kasprzyk, Davis, De Moor, Brazma, and Huber. S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, 21(16):3439–3440, 2005.

Durinck et al.(2009)Durinck, Spellman, Birney, and Huber. S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. Nature protocols, 4(8):1184–1191, 2009.

Gutenkunst et al.(2009)Gutenkunst, Hernandez, Williamson, and Bustamante. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS genetics, 5(10):e1000695, 2009.

Hernandez et al.(2011)Hernandez, Kelley, Elyashiv, Melton, Auton, McVean, Sella, Przeworski, et al. R. D. Hernandez, J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, G. Sella, M. Przeworski, et al. Classic selective sweeps were rare in recent human evolution. science, 331(6019):920–924, 2011.

Korneliussen et al.(2014)Korneliussen, Albrechtsen, and Nielsen. T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. Angsd: analysis of next generation sequencing data. BMC bioinformatics, 15(1):356, 2014.

McLaren et al.(2010)McLaren, Pritchard, Rios, Chen, Flicek, and Cunningham. 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):2069–2070, 2010.

R Core Team(2014). R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL .

Sattath et al.(2011)Sattath, Elyashiv, Kolodny, Rinott, and Sella. S. Sattath, E. Elyashiv, O. Kolodny, Y. Rinott, and G. Sella. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in drosophila simulans. PLoS genetics, 7(2):e1001302, 2011.

Schnable et al.(2009)Schnable, Ware, Fulton, Stein, Wei, Pasternak, Liang, Zhang, Fulton, Graves, et al. P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, et al. The b73 maize genome: complexity, diversity, and dynamics. science, 326(5956):1112–1115, 2009.