

Supporting Information

Kraft et al. 10.1073/pnas.1308933111

SI Materials and Methods

Species Distribution Modeling. The species distribution modeling (SDM) algorithm MaxEnt (1) was used as follows. The models were fitted with current climate data (2) and evaluated using fivefold cross-validation and the area under the curve (cAUC) statistic (3), which is a bias-corrected version of the AUC area under the receiver operating characteristic (ROC) curve statistic. Bias-corrected AUC goes from 1 (perfect discrimination between occupied and unoccupied places) to 0.5 (discrimination not better than random). The MaxEnt model was then used to predict climatic suitability during mid-Holocene conditions according to projections by nine global climate models (GCMs): bccsm1-1, CCSM4, CNRM-CM5, HadGEM2-CC, HadGEM2-ES, IPSL-CM5A-LR, MPI-ESM-P, MRI-CGCM3, and MIROC-ESM (Fig. S3) obtained from the Coupled Model Intercomparison Project (CMIP5). The GCM data were downscaled to a spatial resolution of 30 arc-seconds (~1 km²) by computing the difference between the average climate for modeled past climate conditions and the current climate computed by the same GCM. We then used smooth splines to interpolate these differences to a higher spatial resolution. Finally, we applied these differences to a high-resolution estimate of the current climate such that all datasets were bias corrected in the same manner (3). For each MaxEnt prediction, we computed the average score across GCMs. We standardized these by dividing by the maximum value such that the scores were between 0 and 1.

Paleobiolinguistics. Most of the reconstructed terms for chili pepper presented in Table 1 were extracted from existing literature: Totonacan (4), Mayan (5), Zapotecan (6), General Aztec (7), Popolocan-Zapotecan and Amuzgo-Mixtecan (8), Chinantecan (9), Mixe-Zoquean (10), Misumalpan (11), Sonoran (12), and Otopamean (13). Reconstructions for Otomanguean, Eastern Otomanguean, Mixtecan, and Popolocan are from Rensch (14). Rensch does not provide referents for his reconstructed terms, but these can be unambiguously inferred from the referents supplied for words supporting reconstructions. Kaufman (8) reconstructs a number of plant terms for Proto-Otomanguean, but does not follow Rensch (14) in reconstructing a term for chili pepper (whereas Kaufman does reconstruct *Capsicum* terms for Popolocan-Zapotecan and Amuzgo-Mixtecan). Evidence Rensch provides for his Proto-Otomanguean reconstruction (*?ki) appears to be robust. This entails in part reconstructed terms for *Capsicum* from protolanguages of the two major branches of the family (Western Otomanguean and Eastern Otomanguean) that show close phonological similarity, respectively, Proto-Chiapanc-Mangue *ni-ⁿgi? and Proto-Zapotecan *ki:?ⁿa? (where elements -ⁿgi? and ki?- are being compared). Proto-languages of Table 1 for which chili terms are “not reconstructable” (NR) are identified. NR is a designation used when terms for chili pepper are present

in all or most languages of a genetic group, but, nonetheless, are not cognate and, hence, do not attest to a term in their common ancestral language (protolanguage). NR, then, never indicates nonreconstructibility because of insufficient or missing data.

Protolanguage dates (Table 1) were calculated through use of Automated Similarity Judgment Program (ASJP) chronology (15), a computational dating approach based on the lexical similarity of languages and a set of 52 calibration dates for protolanguage breakups documented through historical, epigraphic, and archaeological records. The discrepancies between ASJP-estimated dates and the 52 calibration dates are on average 29% as large as the estimated dates themselves, a figure that does not differ much among language families of the world. This simple average can be understood as indicating the margin of error of a given date. To calculate distances from the center of phylogenetic diversity of Otomanguean languages, we computed the inverse distance from this location for an area including all of Mexico and parts of surrounding countries (16). We scaled this distance by dividing by the maximum value and quadrupling the result.

In Table 1, the phonetic representation is based on the following rules. *: the word is a reconstructed form rather than a word from a historically recorded language. Most symbols are pronounced as in English, e.g., k as in *kettle*, s as in *sea*, p as in *pill*. In addition, ç as ch in *chair*; θ as th in *thee*; ñ as ñ in Spanish; ⁿg is like g in *go* with an n immediately preceding it. ? is the glottal stop. The symbol “:” immediately following a vowel indicates that the vowel is stretched out; superscript numbers indicate tone level, e.g., rising, falling, or mid. Superscript HL indicates high tone followed by low tone. H represents an unknown laryngeal sound, either an h or ? . i' is a laryngealized version of the vowel i. Letters surrounded by parentheses occur optionally.

Genetic Distance Analysis. The following 17 microsatellite markers were used in this study: AF222989, AA840737, CM0005, HPMS1.148, HPMS1.106, HPMS1.143, HPMS1.165, HPMS1.227, HPMS1.281, HPMS2.18, HPMS2.41, HPMS2.45, HPMS2.9, and HPMSAT2-20 (17); CAMS075, CAMS321, and CAMS327 (18) (Dataset S1, *Microsat info*). Using the R package “gdistanalyst” (19), we computed the negative logarithm of the proportion of shared alleles (PSA) between each wild accession and all 49 domesticated accessions and computed the mean score for each wild chili accession. We then determined the minimum score (i.e., most related to the domesticated accessions) for generalized locations and spatially interpolated these scores with a thin plate spline (function Tps from the R package “fields”) (20) to produce, for each grid cell on our raster, an estimated genetic similarity between wild chilies (if they occur there) and the domesticated chilies in our sample (wherever they occur). We used 1 minus this score as a measure of the genetic similarity between wild chili pepper and the group of domesticated peppers in a place (grid cell).

1. Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecol Modell* 190(3–4):231–259.
2. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25(15):1965–1978.
3. Hijmans RJ (2012) Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model. *Ecology* 93(3):679–688.
4. Brown CH, Beck D, Kondrak G, Watters JK, Wichmann S (2011) Totozoquean. *International Journal of American Linguistics* 77(3):323–372.
5. Brown CH, Wichmann S (2004) Proto-Mayan syllable nuclei. *International Journal of American Linguistics* 70:128–186.
6. Campbell E (2013) The internal diversification and subgrouping of Chatino. *International Journal of American Linguistics* 79(3):395–420.
7. Campbell L, Langacker RW (1978) Proto-Aztec Vowels: Part III. *International Journal of American Linguistics* 44(4):262–279.
8. Kaufman TS (1990) Early Otomanguean homeland and cultures: Some premature hypotheses. (University of Pittsburgh Working Papers in Linguistics 1), pp 91–136.
9. Rensch CR (1989) *An Etymological Dictionary of the Chinantec Languages* (Summer Institute of Linguistics, Arlington, TX).
10. Wichmann S (1995) *The Relationship Among the Mixe-Zoquean Languages of Mexico* (University of Utah Press, Salt Lake City).
11. Constenla-Umaña A (1987) Elementos de fonología comparada de las lenguas Misumalpas. [Elements of comparative phonology of Misumalpa languages]. *Filología y Lingüística* 13:129–161.
12. Stubbs BD (2011) *Uto-Aztecan: A Comparative Vocabulary* (Shumway Family History Services, Flower Mound, TX).

