Independent molecular basis of convergent highland adaptation in maize

Shohei Takuno[*,1], **Peter Ralph**[†,‡], **Sofiane Mezmouk**[*], **Kelly Swarts**[§], **Rob J. Elshire**[§], **Jeffrey C. Glaubitz**[§], **Edward S. Buckler**[§,**], **Matthew B. Hufford**[*,††], **and Jeffrey Ross-Ibarra**[*,‡‡,2]

[*]Department of Plant Sciences, University of California, Davis, California 95616, USA,
[†]Department of Evolution and Ecology, University of California, Davis, California 95616, USA,
[‡]Molecular and Computational Biology, University of Southern California, Los Angeles,California 90089-0371, USA,
[§]Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853-2703, USA,
[**]United States Department of Agriculture Agricultural Research Service, Ithaca, NY 14853,
[††]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011, USA,
[‡‡]The Center for Population Biology and the Genome Center, University of California, Davis, California 95616 , USA,
[1] Present address: SOKENDAI (Graduate university for advanced studies), Hayama, Kanagawa 240-0193, Japan

December 16, 2014

ABSTRACT Convergent evolution occurs when multiple species/subpopulations adapt to similar environments via similar phenotypes. We investigate here the molecular basis of convergent adaptation in maize to highland climates in Mexico and South America using genome-wide SNP data. Taking advantage of archaeological data on the arrival of maize to the highlands, we infer demographic models for both populations, identifying evidence of a strong bottleneck and rapid expansion in South America. We use these models to then identify loci showing an excess of differentiation as a means of identifying putative targets of natural selection, and compare our results to expectations from recently developed theory on convergent adaptation. Consistent with predictions across a wide array of parameter space, we see limited evidence for convergent evolution at the nucleotide level in spite of strong similarities in overall phenotypes. Instead, we show that selection appears to have predominantly acted on standing genetic variation, and that introgression from wild teosinte populations appears to have played a role in adaptation in Mexican maize.

Introduction

Convergent evolution occurs when multiple species or populations exhibit similar phenotypic adaptations to comparable environmental challenges (Wood *et al.* 2005; Arendt and Reznick 2008; Elmer and Meyer 2011). Evolutionary genetic analysis of a wide range of species has provided evidence for multiple pathways of convergent evolution. One such route occurs when identical mutations arise independently and fix via natural selection in multiple populations. In humans, for example, malaria resistance due to mutations from Glu to Val at the sixth codon of the β-globin gene has arisen independently on multiple unique haplotypes (Currat *et al.* 2002; Kwiatkowski 2005). Convergent evolution can also be achieved when different mutations arise within the same locus yet produce similar phenotypic effects. Grain fragrance in rice appears to have evolved along these lines, as populations across East Asia have similar fragrances resulting from at least eight distinct loss-of-function alleles in the *BADH2* gene (Kovach *et al.* 2009). Finally, convergent evolution may arise from natural selection acting on standing genetic variation in an ancestral population. In the three-spined stickleback, natural selection has repeatedly acted to reduce armor plating in independent colonizations of freshwater environments. Adaptation in these populations occurred both from new mutations as well as standing variation at the *Eda* locus in marine populations (Colosimo *et al.* 2005).

Not all convergent phenotypic evolution is the result of convergent evolution at the molecular level, however. Recent studies of adaptation to high elevation in humans, for example, reveal that the genes involved in highland adaptation are largely distinct among Tibetan, Andean and Ethiopian populations (Bigham *et al.* 2010; Scheinfeldt *et al.* 2012; Alkorta-Aranburu *et al.* 2012). While observations of independent origin may be due to a complex genetic architecture or standing genetic variation, introgression from related populations may also play a

[2]Corresponding author: Department of Plant Sciences, University of California, Davis, California 95616, USA. E-mail: rossibarra@ucdavis.edu

role. In Tibetan populations, the adaptive allele at the *EPAS1* locus appears to have arisen via introgression from Denisovans, a related hominid group (Huerta-Sánchez *et al.* 2014). Overall, we still know relatively little about how convergent phenotypic evolution is driven by common genetic changes or the relative frequencies of these different routes of convergent evolution.

The adaptation of maize to high elevation environments (*Zea mays* ssp. *mays*) provides an excellent opportunity to investigate the molecular basis of convergent evolution. Maize was domesticated from the wild teosinte *Zea mays* ssp. *parviglumis* (hereafter *parviglumis*) in the lowlands of southwest Mexico ∼9,000 years before present (BP) (Matsuoka *et al.* 2002; Piperno *et al.* 2009; van Heerwaarden *et al.* 2011). After domestication, maize spread rapidly across the Americas, reaching the lowlands of South America and the high elevations of the Mexican Central Plateau by $\sim 6,000$ BP (Piperno 2006), and the Andean highlands by $\sim 4,000$ BP (Perry *et al.* 2006; Grobman *et al.* 2012). The transition from lowland to highland habitats spanned similar environmental gradients in Mexico and South America (Figure S1) and presented a host of novel challenges that often accompany highland adaptation including reduced temperature, increased ultraviolet radiation, and reduced partial pressure of atmospheric gases (Körner 2007).

Common garden experiments in Mexico reveal that highland maize has successfully adapted to high elevation conditions (Mercer *et al.* 2008), and phenotypic comparisons between Mexican and South American populations are suggestive of convergent evolution. Maize landraces (open-pollinated traditional varieties) from both populations share a number of phenotypes not found in lowland populations, including dense macrohairs (Wilkes 1977; Wellhausen *et al.* 1957), stem pigmentation (Wilkes 1977; Wellhausen *et al.* 1957), differences in tassel branch and ear husk number (Brewbaker 2014), and biochemical response to UV radiation (Casati and Walbot 2005). In spite of these shared phenotypes, genetic analyses of maize landraces from across the Americas indicate that the two highland populations are independently derived from their respective lowland populations (Vigouroux *et al.* 2008; van Heerwaarden *et al.* 2011), suggesting that observed patterns of phenotypic similarity are not simply due to recent shared ancestry.

In addition to convergent evolution between maize landraces, a number of lines of evidence suggest convergent evolution in the related wild teosintes. *Zea mays* ssp. *mexicana* (hereafter *mexicana*) is native to the highlands of central Mexico, where it is thought to have occurred since at least the last glacial maximum (Ross-Ibarra *et al.* 2009; Hufford *et al.* 2012a). Phenotypic differences between *mexicana* and the lowland *parviglumis* mirror those between highland and lowland maize (Lauter *et al.* 2004), and population genetic analyses of the two subspecies reveal evidence of natural selection associated with altitudinal differences between *mexicana* and *parviglumis* (Pyhäjärvi *et al.* 2013; Fang *et al.* 2012). Landraces in the highlands of Mexico are often found in sympatry with *mexicana* and gene flow from *mexicana* likely contributed to maize adaptation to the highlands (Hufford *et al.* 2013). No wild *Zea* occur in S. America, and S. American landraces show no evidence of gene flow from Mexican teosinte (van Heerwaarden *et al.* 2011), further suggesting an independent origin of convergent phenotypic adaptation.

Here we use genome-wide SNP data from Mexican and S. American landraces to investigate the evidence for convergent evolution to highland environments at the molecular level. We estimate demographic histories for maize in the highlands of Mexico and South America, then use these models to identify loci that may have been the target of selection in each population. We find a large number of sites showing evidence of selection, consistent with a complex genetic architecture involving many phenotypes and numerous loci. We see little evidence for shared selection at the nucleotide or gene level, a result we show is consistent with expectations from recent theoretical work on convergent adaptation (Ralph and Coop 2014). Instead, our results support a role of adaptive introgression from teosinte in Mexico and highlight the contribution of standing variation to adaptation in both populations.

Materials and Methods

Materials and DNA extraction

We included one individual from each of 94 open-pollinated landrace maize accessions from high and low elevation sites in Mexico and S. America (Table S1). Accessions were provided by the USDA germplasm repository or kindly donated by Major Goodman (North Carolina State University). Sampling locations are shown in Figure 1A. Landraces sampled from elevations $< 1,700$ m were considered lowland, while accessions from $> 1,700$ m were considered highland. Seeds were germinated on filter paper following fungicide treatment and grown in standard potting mix. Leaf tips were harvested from plants at the five leaf stage. Following storage at $-80°$C overnight, leaf tips were lyophilized for 48 hours. Tissue was then homogenized with a Mini-Beadbeater-8 (BioSpec Products, Inc., Bartlesville, OK, USA). DNA was extracted using a modified CTAB protocol (Saghai-Maroof *et al.* 1984). The quality of DNA was ensured through inspection on a 2% agarose gel and quantification of the ratio of light absorbance at 260 and 280 nm using a NanoDrop spectrophotometer (Thermo Scientific, NanoDrop Products, Wilmington, DE, USA).

SNP data

We generated two complementary SNP data sets for the sampled maize landraces. The first set was generated using the Illumina MaizeSNP50 BeadChip platform, including 56,110 SNPs (Ganal *et al.* 2011). SNPs were clustered with the default algorithm of the GenomeStudio Genotyping Module v1.0 (Illumina Inc., San Diego, CA, USA) and then visually inspected and manually adjusted. These data are referred to as

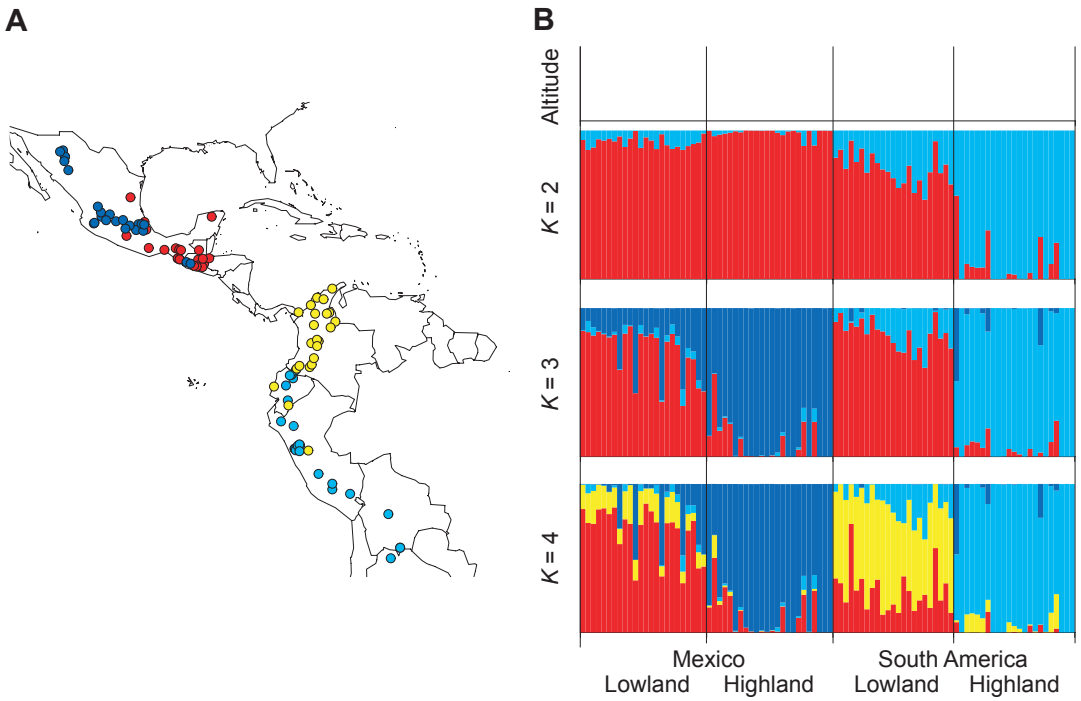

Figure 1 (A) Sampling locations of landraces. Red, blue, yellow and light blue dots represent Mexican lowland, Mexican highland, S. American lowland and S. American highland populations, respectively. (B) Results of STRUCTURE analysis of the maizeSNP50 SNPs with $K = 2 \sim 4$. The top panel shows the elevation, ranging from 0 to 4,000 m on the *y*-axes. The colors in $K = 4$ correspond to those in panel (A).

"MaizeSNP50" hereafter. This array contains SNPs discovered in multiple ascertainment schemes (Ganal *et al.* 2011), but the vast majority of SNPs come from polymorphisms distinguishing the maize inbred lines B73 and Mo17 (14,810 SNPs) or identified from sequencing 25 diverse maize inbred lines (40,594 SNPs; Gore *et al.* 2009).

The second data set was generated for a subset of 87 of the landrace accessions (Table S1) utilizing high-throughput Illumina sequencing data via genotyping-by-sequencing (GBS; Elshire *et al.* 2011). Genotypes were called using TASSEL-GBS (Glaubitz *et al.* 2014) resulting in 2,848,284 SNPs with an average of 71.3% missing data per individual.

To assess data quality, we compared genotypes at the 7,197 SNPs (229,937 genotypes, excluding missing data) that overlap between the MaizeSNP50 and GBS data sets. While only 0.8% of 173,670 comparisons involving homozygous MaizeSNP50 genotypes differed in the GBS data, 88.6% of 56,267 comparisons with MaizeSNP50 heterozygotes differed, nearly always being reported as a homozygote in GBS. Despite this high heterozygote error rate, the high correlation in allele frequencies between data sets ($r = 0.89$; Figure S2) supports the utility of the GBS data set for estimating allele frequencies.

We annotated SNPs using the filtered gene set from RefGen version 2 of the maize B73 genome sequence (Schnable *et al.* 2009; release 5b.60) from maizesequence.org. We excluded genes annotated as transposable elements (84) and pseudogenes (323) from the filtered gene set, resulting in a total of 38,842 genes.

Structure analysis

We performed a STRUCTURE analysis (Pritchard *et al.* 2000; Falush *et al.* 2003) using synonymous and noncoding SNPs from the MaizeSNP50 data. We randomly pruned SNPs closer than 10 kb and assumed free recombination between the remaining SNPs. Alternative distances were tried with nearly identical results. We excluded SNPs in which the number of heterozygous individuals exceeded homozygotes and where the *P*-value for departure from Hardy-Weinberg Equilibrium (HWE) using all individuals was smaller than 0.05 based on a *G*-test. Following these data thinning measures, 17,013 biallelic SNPs remained. We conducted three replicate runs of STRUCTURE using the correlated allele frequency model with admixture for $K = 2$ through $K = 6$ populations, a burn-in length of 50,000 iterations and a run length of 100,000 iterations. Results across replicates were nearly identical.

Historical population size

We tested three models in which maize was differentiated into highland and lowland populations subsequent to domestication (Figure 2). Observed joint frequency distributions (JFDs) were calculated using the GBS data set due to its lower level of ascertainment bias. A subset of synonymous and noncoding SNPs

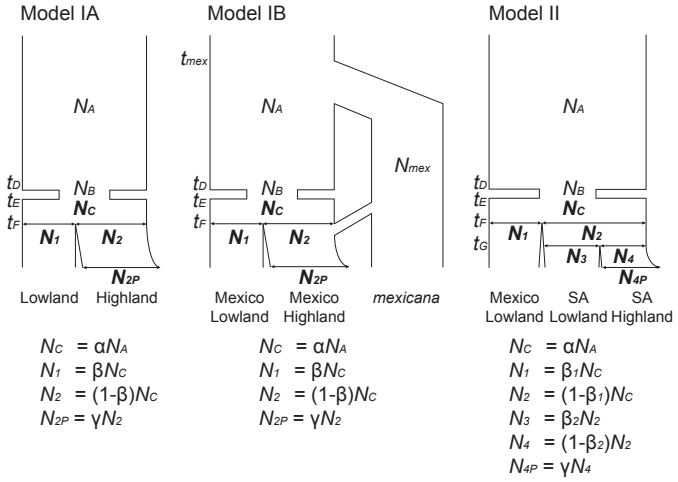

Figure 2 Models of historical population size for lowland and highland populations. Parameters in bold were estimated in this study. See text for details.

were utilized that had ≥ 15 individuals without missing data in both lowland and highland populations and did not violate HWE. A HWE cut-off of $P < 0.005$ was used for each subpopulation due to our under-calling of heterozygotes. In total, we included 18,745 synonymous and noncoding SNPs for the Mexican populations in Models IA and IB, 14,508 for the S. American populations in Model I and 11,305 for the Mexican lowland population and the S. American populations in Model II. We obtained similar results under more or less stringent thresholds for significance ($P < 0.05 \sim 0.0005$; data not shown), though the number of SNPs was very small at $P < 0.05$. Parameters were inferred with the software $\delta a \delta i$ (Gutenkunst *et al.* 2009), which uses a diffusion method to calculate an expected JFD and evaluates the likelihood of the data using a multinomial assumption.

***Model IA*:** This model is applied to the Mexican and S. American populations. We assume the ancestral diploid population representing *parviglumis* follows a standard Wright-Fisher model with constant size. The size of the ancestral population is denoted by N_A. At t_D generations ago, the bottleneck event begins at domestication, and at t_E generations ago, the bottleneck ends. The population size and duration of the bottleneck are denoted by N_B and $t_B = t_D - t_E$, respectively. The population size recovers to $N_C = \alpha N_A$ in the lowlands. Then, the highland population is differentiated from the lowland population at t_F generations ago. The size of the lowland and highland populations at time t_F is determined by a parameter β such that the population is divided by βN_C and $(1 - \beta)N_C$; our conclusions hold if we force lowland population size to remain at N_C (data not shown). We assume that the population size in the lowlands is constant but that the highland population experiences exponential expansion after divergence: its current population size is γ times larger than that at t_F.

***Model IB*:** We expand Model IA for the Mexican populations by incorporating admixture from the teosinte *mexicana* to the highland Mexican maize population. The time of differentiation between *parviglumis* and *mexicana* occurs at t_{mex} generations ago. The *mexicana* population size is assumed to be constant at N_{mex}. At t_F generations ago, the Mexican highland population is derived from admixture between the Mexican lowland population and a portion P_{mex} from the teosinte *mexicana*.

***Model II*:** The final model includes the Mexican lowland, S. American lowland and highland populations. This model was used for simulating SNPs with ascertainment bias (see below). At time t_F, the Mexican and S. American lowland populations are differentiated, and the sizes of populations after splitting are determined by β_1. At time t_G, the S. American lowland and highland populations are differentiated, and the sizes of populations at this time are determined by β_2. As in Model IA, the S. American highland population is assumed to experience population growth with the parameter γ.

Estimates of a number of our model parameters were available from previous work. N_A was set to 150,000 using estimates of the composite parameter $4N_A\mu \sim 0.018$ from *parviglumis* (Eyre-Walker *et al.* 1998; Tenaillon *et al.* 2001, 2004; Wright *et al.* 2005; Ross-Ibarra *et al.* 2009) and an estimate of the mutation rate $\mu \sim 3 \times 10^{-8}$ (Clark *et al.* 2005) per site per generation. The severity of the domestication bottleneck is represented by $k = N_B/t_B$ (Eyre-Walker *et al.* 1998; Wright *et al.* 2005), and following Wright *et al.* (2005) we assumed $k = 2.45$ and $t_B = 1,000$ generations. Taking into account archaeological evidence (Piperno *et al.* 2009), we assume $t_D = 9,000$ and $t_E = 8,000$. We further assumed $t_F = 6,000$ for Mexican populations in Models IA and IB (Piperno 2006), $t_F = 4,000$ for S. American populations in Model IA (Perry *et al.* 2006; Grobman *et al.* 2012), and $t_{mex} = 60,000$, $N_{mex} = 160,000$ (Ross-Ibarra *et al.* 2009), and $P_{mex} = 0.2$ (van Heerwaarden *et al.* 2011) for Model IB. For both Models IA and IB, we inferred three parameters (α, β and γ), and, for Model II, we fixed $t_F = 6,000$ and $t_G = 4,000$ (Piperno 2006; Perry *et al.* 2006; Grobman *et al.* 2012) and estimated the remaining four parameters (α, β_1, β_2 and γ).

Population differentiation

We used our inferred models of population size change to generate a null distribution of F_{ST}. As implemented in $\delta a \delta i$ (Gutenkunst *et al.* 2009), we calculated an expected JFD given estimated model parameters and the sample sizes from our highland and lowland populations. Then, we converted the JFD into the distribution of F_{ST} values. The *P*-value of a SNP was calculated by $P(F_{ST\_E} \geq F_{ST\_O} | p \pm 0.025) = P(F_{ST\_E} \geq$

$F_{ST\_O} \cap p \pm 0.025)/P(p \pm 0.025)$, where F_{ST_O} and F_{ST_E} are observed and expected F_{ST} values and $p \pm 0.025$ is the set of loci with mean allele frequency across both highland and lowland populations within 0.025 of the SNP in question.

Generating the null distribution of differentiation for the MaizeSNP50 data requires accounting for ascertainment bias. Evaluation of genetic clustering in our data (not shown) coincides with previous work (Hufford *et al.* 2012b) in suggesting that the two inbred lines most important in the ascertainment panel (B73 and Mo17) are most closely related to Mexican lowland maize. We thus added two additional individuals to the Mexican lowland population and generated our null distribution using only SNPs for which the two individuals had different alleles. For model IA in S. America we added two individuals at time t_F to the ancestral population of the S. American lowland and highland populations because the Mexican lowland population was not incorporated into this model. For each combination of sample sizes in lowland and highland populations, we generated a JFD from 10^7 SNPs using the software ms (Hudson 2002). Then, we calculated *P*-values from the JFD in the same way. We calculated F_{ST} values for all SNPs that had ≥ 10 individuals with no missing data in all four populations and showed no departure from HWE at the 0.5% (GBS) or 5% (MaizeSNP50) level.

Haplotype sharing test

We performed a pairwise haplotype sharing (PHS) test to detect further evidence of selection, following Toomajian *et al.* (2006). To conduct this test, we first imputed and phased the combined SNP data (both GBS and MaizeSNP50) using the fastPHASE software version 1.4.0 (Scheet and Stephens 2006). As a reference for phasing, we used data (excluding heterozygous SNPs) from an Americas-wide sample of 23 partially inbred landraces from the Hapmap v2 data set (Chia *et al.* 2012). We ran fastPHASE with default parameter settings. PHS was calculated for an allele *A* at position x by

$$PHS_{x_A} = \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \frac{Z_{ijx}}{\binom{p}{2}} - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{Z_{ijx}}{\binom{n}{2}} \ , \qquad (1)$$

where n is the sample size of haploids, p is the number of haploids carrying the allele A at position x, and

$$Z_{ijx} = \frac{d_{ijx} - \bar{d_{ij}}}{\sigma_{ij}} \ , \qquad (2)$$

where d_{ijx} is the genetic distance over which individuals i and j are identical surrounding position x, $\bar{d_{ij}}$ is the genome-wide mean of distances over which individuals are identical, and σ_{ij} is the standard deviation of the distribution of distances. The *P*-value for each allele was calculated as the proportion of alleles of the same frequency genome-wide that have a larger PHS value.

Genetic distances were obtained for the MaizeSNP50 data (Ganal *et al.* 2011) and fit using a tenth degree polynomial curve to all SNPs (data not shown).

Theoretical evaluation of convergent evolution

We build on results from Ralph and Coop (2014) to assess whether the abundance and degree of coincidence of presumably adaptive high-F_{ST} alleles is consistent with what is known about the population history of maize, we evaluated the rate at which we expect an allele that provides a selective advantage at higher elevation to arise by new mutation in a highland region (λ_{mut}), and the rate at which such an allele already present in the Mexican highlands would transit the intervening lowlands and fix in the Andean highlands (λ_{mig}). We assume alleles adapted in the highlands are slightly deleterious at lower elevation, consistent with empirical findings in reciprocal transplant experiments in Mexico (Mercer *et al.* 2008). These numbers depend most strongly on the population density, the selection coefficient, and the rate at which seed is transported long distances and replanted. To obtain specific predictions, we computed λ_{mut} and λ_{mig} at various parameter values. We also checked these with simulations and more detailed computations, described in the Appendix. Here we describe the mathematical details; readers may skip to the results without loss of continuity.

***Demographic model*:** Throughout, we followed van Heerwaarden *et al.* (2010) in constructing a detailed demographic model for domesticated maize. We assume fields of $N = 10^5$ plants are replanted each year from $N_f = 561$ ears, either from completely new stock (with probability $p_e = 0.068$), from partially new stock (a proportion $r_m = 0.2$ with probability $p_m = 0.02$), or otherwise entirely from the same field. Each plant is seed parent to all kernels of its own ears, but can be pollen parent to kernels in many other ears; a proportion $m_g = 0.0083$ of the pollen-parent kernels are in other fields. Wild-type plants have an average of $\mu_E = 3$ ears per plant, and ears have an average of N/N_f kernels; each of these numbers are Poisson distributed. The mean number of pollen-parent kernels, and the mean number of kernels per ear, is assumed to be $(1 + s_b)$ times larger for individuals heterozygous for the selected allele. Migration is mediated by seed exchange – when fields are replanted, the seed is chosen from a random distance away with mean $\sigma_s = 50$km, but plants only pollinate other plants belonging to the same village (distance 0). Each individual can have offspring in three categories: local seed, local pollen, and migrant seed; the mean numbers of each of these are determined by the condition that the population is stable (i.e. wild-type, diploid individuals have on average 2 offspring) except that heterozygotes have on average $(1 + s_b)$ offspring that carry the selected allele. Each ear has a small chance of being chosen for replanting, so the number of ears replanted of a given individual is Poisson, and assuming that

pollen is well-mixed, the number of pollen-parent kernels is Poisson as well. Each of these numbers of offspring has a mean that depends on whether the field is replanted with new stock, and whether ears are chosen from this field to replant other fields, so the total number of offspring is actually a mixture of Poissons; these means, and more details of the computations, are found in Appendix **??**. At these parameter values, we compute that the variance in number of offspring, ξ^2, is between 20 (for wild-type) and 30 (for $s_b = 0.1$), and the dispersal distance (mean distance between parent and offspring) is $\sigma = 1.8$km.

***New mutations*:** The rate at which new mutations appear and fix in a highland population, which we denote λ_{mut}, is equal to the total population size of the highlands multiplied by the mutation rate per generation and by the chance that a single such mutation successfully fixes (i.e. is not lost to drift). The probability that a single new mutant allele providing benefit s_b to heterozygotes at high elevation will fix locally in the high elevation population is approximately $2s_b$ divided by the variance in offspring number (Jagers 1975). The calculation above is not quite correct, as it neglects migration across the altitudinal gradient, but exact numerical calculation of the chance of fixation of a mutation as a function of the location where it first appears indicates that the approximation is quite good (see Figure 1); for theoretical treatment see Pollak (1966) or Barton (1987).

Concretely, the probability that a new mutation destined for fixation will arise in a patch of high-elevation habitat of area A in a given generation is a function of the density of maize per unit area ρ, the selective benefit s_b it provides, the mutation rate μ, and the variance in offspring number ξ^2. In terms of these parameters, the rate of appearance is

$$\lambda_{\text{mut}} = \frac{2\mu\rho A s_b}{\xi^2}. \tag{3}$$

***Migration*:** A corresponding expression for the chance that an allele moves from one highland population to another is harder to intuit, and is addressed in more depth in Ralph and Coop (2014). If an allele is beneficial at high elevation and fixed in the Mexican highlands but is deleterious at low elevations, then at equilibrium it will be present at low frequency at migration-selection balance (Slatkin 1973) in nearby lowland populations. This equilibrium frequency decays exponentially with distance, so that the highland allele is present at distance R from the highlands at frequency $C\exp(-R\sqrt{2s_m}/\sigma)$, where s_m is the deleterious selection coefficient for the allele in low elevation, σ is the mean dispersal distance, and C is a constant depending on geography ($C \approx 1/2$ is close). Multiplying this frequency by a population size gets the predicted number (average density across a large number of generations) of individuals carrying the allele. Therefore, in a lowland population of size N at distance R from the highlands, $(N/2)\exp(-R\sqrt{2s_m}/\sigma)$ is equal to the probability that there are any highland alleles present, multiplied by the expected number of these given that some are present. Since the latter is at least 1, this puts an upper bound on the rate of migration

$$\lambda_{\text{mig}} \leq (N/2)\exp(-R\sqrt{2s_m}/\sigma), \tag{4}$$

and we we would need to wait $T_{\text{mig}} = 1/\lambda_{\text{mig}}$ generations for a rare such excursion to occur. This calculation omits the probability that such an allele fixes ($\approx 2s_b/\xi^2$), but since such alleles arrive by migration, this omission is unlikely a large effect and is conservative. Modeling of highland alleles which are neutral in the lowlands is more difficult, but approximations are possible (see Supplemental Methods).

Table 1 F_{ST} of synonymous and noncoding GBS SNPs

		Mexico		S. America	
		Lowlands	Highlands	Lowlands	Highlands
Mexico	Lowlands	–			
	Highlands	0.0244	–		
S. America	Lowlands	0.0227	0.0343	–	
	Highlands	0.0466	0.0534	0.0442	–

Table 2 Estimated parameters of population size model

Mexico	Model IA		Model IB	
	Likelihood	−5592.80	Likelihood	−4654.79
	N_C	138,000	N_C	225,000
	N_1	52,440	N_1	171,000
	N_2	85,560	N_2	54,000
	N_{2P}	85,560	N_{2P}	54,000
S. America	Model IA		Model II	
	Likelihood	−3855.28	Likelihood	−8044.71
	N_C	78,000	N_C	150,000
	N_1	75,660	N_1	96,000
	N_2	2,340	N_2	54,000
	N_{2P}	205,920	N_3	51,300
			N_4	2,700
			N_{4P}	145,800

Results

Samples and data

We sampled 94 maize landraces from four distinct regions in the Americas (Table S1): the lowlands of Mexico/Guatemala ($n = 24$) and northern South America ($n = 23$) and the highlands of the Mexican Central Plateau ($n = 24$) and the Andes ($n = 23$). Samples were genotyped using the MaizeSNP50 Beadchip platform ("MaizeSNP50"; $n = 94$) and genotyping-by-sequencing ("GBS"; $n = 87$). After filtering for Hardy-Weinberg genotype frequencies and minimum sample size $\geq$

10 in each of the four populations (see Materials and Methods) 91,779 SNPs remained, including 67,828 and 23,951 SNPs from GBS and MaizeSNP50 respectively.

Population structure

We performed a STRUCTURE analysis (Pritchard *et al.* 2000; Falush *et al.* 2003) of our landrace samples, varying the number of groups from $K = 2$ to 6 (Figure 1, Figure S3). Most landraces were assigned to groups consistent with *a priori* population definitions, but admixture between highland and lowland populations was evident at intermediate elevations ($\sim$ 1700m). Consistent with previously described scenarios for maize diffusion (Piperno 2006), we find evidence of shared ancestry between lowland Mexican maize and both Mexican highland and S. American lowland populations. Pairwise F_{ST} among populations reveals low overall differentiation (Table 1), and the higher F_{ST} values observed in S. America are consistent with the decreased admixture seen in STRUCTURE. Archaeological evidence supports a more recent colonization of the highlands in S. America (Piperno 2006; Perry *et al.* 2006; Grobman *et al.* 2012), suggesting that the observed differentiation may be the result of a stronger bottleneck during colonization of the S. American highlands.

Population differentiation

To provide a null expectation for allele frequency differentiation, we used the joint site frequency distribution (JFD) of lowland and highland populations to estimate parameters of two demographic models using the maximum likelihood method implemented in $\delta a \delta i$ (Gutenkunst *et al.* 2009). All models incorporate a domestication bottleneck (Wright *et al.* 2005) and population differentiation between lowland and highland populations, but differ in their consideration of admixture and ascertainment bias (Figure 2; see Materials and Methods for details).

Estimated parameter values are listed in Table 2; while the observed and expected JFDs were quite similar for both models, residuals indicated an excess of rare variants in the observed JFDs in all cases (Figure 3). Under both models IA and IB, we found expansion in the highland population in Mexico to be unlikely, but a strong bottleneck followed by population expansion is supported in S. American highland maize in both models IA and II. The likelihood value of model IB was higher than the likelihood of model IA by 850 units of log-likelihood (Table 2), consistent with analyses suggesting that introgression from *mexicana* played a significant role during the spread of maize into the Mexican highlands (Hufford *et al.* 2013).

In addition to the parameters listed in Figure 2, we investigated the impact of varying the domestication bottleneck size (N_B). Surprisingly, N_B was estimated to be equal to N_C, the population size at the end of the bottleneck, and the likelihood of $N_B < N_C$ was much smaller than for alternative parameterizations (Table 2, S2).

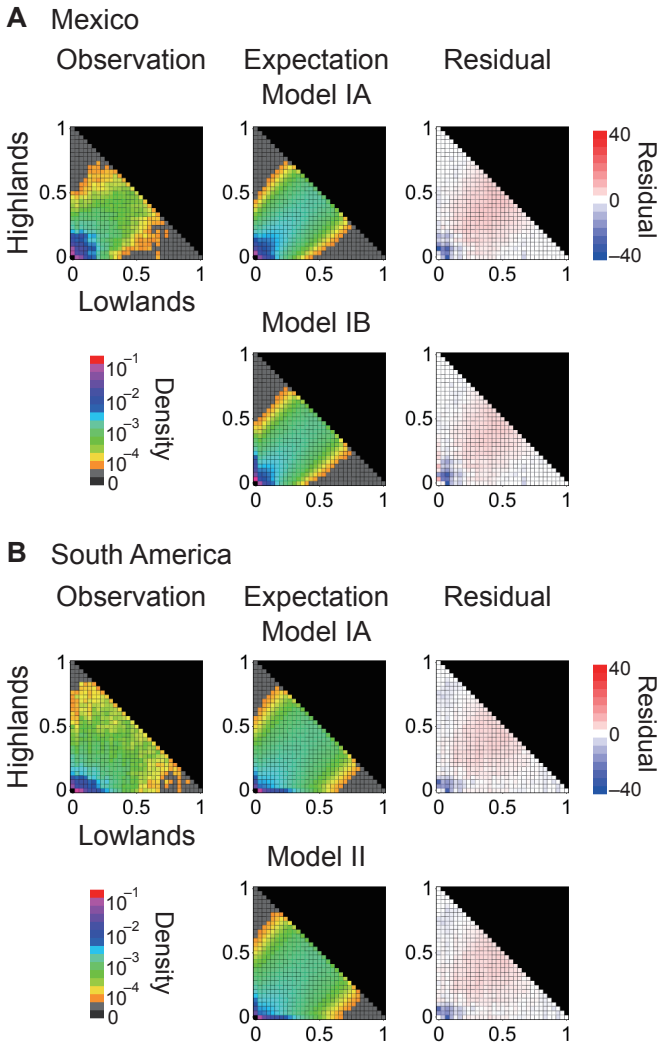

Figure 3 Observed and expected joint distributions of minor allele frequencies in lowland and highland populations in (A) Mexico and (B) S. America. Residuals are calculated as $(\text{model} - \text{data})/\sqrt{\text{model}}$

Comparisons of our empirical F_{ST} values to the null expectation simulated under our demographic models allowed us to identify significantly differentiated SNPs between lowland and highland populations. In all cases, observed F_{ST} values were quite similar to those generated under our null models (Figure S4), and model choice – including the parameterization of the domestication bottleneck – had little impact on the distribution of estimated *P*-values (Figure S5). We show results under Model IB for Mexican populations and Model II for S. American populations. We chose $P < 0.01$ as an arbitrary cut-off for significant differentiation between lowland and highland populations, and identified 687 SNPs in Mexico (687/76,989=0.89%) and 409 SNPs in South America (409/63,160=0.65%) as significant outliers (Figure 4). Differ-

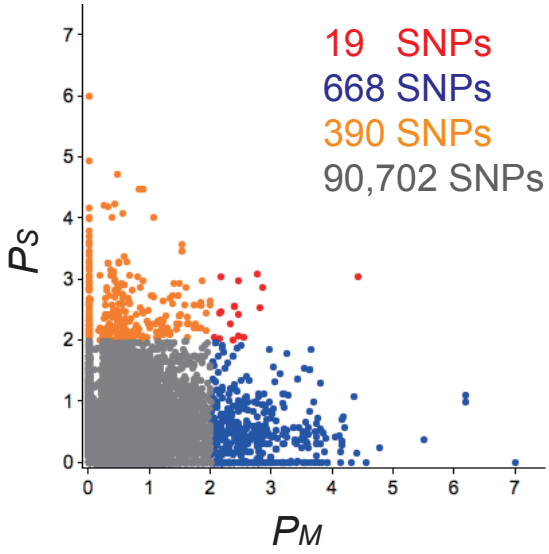

Figure 4 Scatter plot of $-\log_{10} P$-values of observed F_{ST} values based on simulation from estimated demographic models. P-values are shown for each SNP in both Mexico (Model IB; P_M on x-axis) and S. America (Model II; P_S on y-axis). Red, blue, orange and gray dots represents SNPs showing significance in both Mexico and S. America, only in Mexico, only in S. America, respectively (see text for details). The number of SNPs in each category is shown in the same color as the points.

ent cutoff values (0.05, 0.001) gave qualitatively identical results (data not shown). SNPs with significant F_{ST} P-values were enriched in intergenic regions rather than protein coding regions (60.0% vs. 47.9%, Fisher's Exact Test $P < 10^{-7}$ for Mexico; 62.0% vs. 47.8%, FET $P < 10^{-5}$ for S. America). Different cutoff values (0.05, 0.001) gave qualitatively identical results (data not shown).

Patterns of adaptation

Given the historical spread of maize from an origin in the lowlands, it is tempting to assume that the observation of significant population differentiation at a SNP should be primarily due to an increase in frequency of adaptive alleles in the highlands. To test this hypothesis, we sought to identify the adaptive allele at each locus using comparisons between Mexico and S. America as well as to *parviglumis* (See Supplementary Text for details). Consistent with predictions, we infer that differentiation at 72.3% (264) and 76.7% (230) of SNPs in Mexico and S. America is due to adaptation in the highlands after excluding SNPs with ambiguous patterns likely due to recombination. The majority of these SNPs show patterns of haplotype variation (by the PHS test) consistent with our inference of selection (Supplementary Text and Table S3).

Convergent evolution at the nucleotide level should be reflected in an excess of SNPs showing significant differentiation between lowland and highland populations in both Mexico and S. America. Although the 19 SNPs showing F_{ST} P-values < 0.01 in both Mexico (P_M) and S. America (P_S) is statistically greater than the ≈ 5 expected ($48,370 \times 0.01 \times 0.01 \approx 4.8$; χ^2-test, $P \ll 0.001$), it nonetheless represents a small fraction ($\approx 7 - 8\%$) of all SNPs showing evidence of selection. This paucity of shared selected SNPs does not appear to be due to our demographic model: a model-free approach based on the top 1% highest F_{ST} values finds no shared adaptive SNPs between Mexican and S. American highland populations. For 13 of 19 SNPs showing putative evidence of shared selection we could use data from *parviglumis* to infer whether these SNPs were likely selected in lowland or highland conditions (Supplemental Text). Surprisingly, SNPs identified as shared adaptive variants more frequently showed segregation patterns consistent with lowland (10 SNPs) rather than highland adaptation (2 SNPs).

We also investigated how often different SNPs in the same gene may have been targeted by selection. To search for this pattern, we considered all SNPs within 10kb of a transcript as part of the same gene, though SNPs in an miRNA or second transcript within 10kb of the transcript of interest were excluded. We classified SNPs showing significant F_{ST} in Mexico, S. America or in both regions into 778 genes. Of these, 485 and 277 genes showed Mexico-specific and SA-specific significant SNPs, while 14 genes contained at least one SNP with a pattern of differentiation suggesting convergent evolution and 2 genes contained both Mexico-specific and SA-specific significant SNPs. Overall, however, fewer genes showed evidence of convergent evolution than expected by chance (permutation test; $P < 10^{-5}$). Despite similar phenotypes and environments, we thus see little evidence for convergent evolution at either the SNP or the gene level.

Comparison to theory

Given the limited empirical evidence for convergent evolution at the molecular level, we took advantage of recent theoretical efforts (Ralph and Coop 2014) to assess the degree of convergence expected under a spatially explicit population genetic model (see Materials and Methods). Our modeling estimates assume a maize population density ρ of the highlands to be around (0.5 ha field/person) $\times$ (0.5 people/km^2) $\times$ (2×10^4 plants per ha field) $=$ 5,000 plants per km^2. The area of the Andean highlands is around $A = 500$km^2, leading to a total population of $A\rho = 2.5 \times 10^6$. *this area seems way too small* *agreed* Assuming an offspring variance of $\xi^2 = 30$, we can then compute the waiting time $T_{\text{mut}} = 1/\lambda_{\text{mut}}$ for a new beneficial mutation to appear and fix. We observe that even if there is relatively strong selection for an allele at high elevation ($s_b = 0.01$), a single-base mutation with mutation rate $\mu = 10^{-8}$ would take at least 60,000 generations to appear and fix. Because T_{mut} scales approximately linearly with both the selection coefficient and the mutation rate, strong selection and the existence of multiple equivalent mutable sites could reduce this time. For example,

if any one of 10 sites within a gene could have equivalent strong selective benefit ($s_b = 0.1$), T_{mut} would be reduced to 600 generations. *Peter please check this paragraph to make sure numbers and text look OK.*

Gene flow between highland regions could also generate patterns of shared adaptive SNPs. From our demographic model we have estimated a mean dispersal distance of $\sigma \approx 1.8$ kilometers per generation. With selection against the highland allele in low elevations $10^{-1} \geq s_m \geq 10^{-4}$, the distance $\sigma/\sqrt{2s_m}$ over which the frequency of a highland-adaptive, lowland-deleterious allele decays into the lowlands is still short: between 4 and 150 kilometers. Since the Mexican and Andean highlands are around 4,000 km apart, the time needed for a rare allele with weak selective cost $s_m = 10^{-4}$ in the lowlands to transit between the two highland regions is $T_{\text{mig}} \approx 4 \times 10^{10}$ generations. However, shorter distances could be transited more quickly – if the distance between highland patches R is 1,000 km (or if σ is four times larger) then the same allele would be expected to transit between populations in approximately 2 generations *is this correct? 4000km in* $\approx 4 \times 10^{10}$ *and 1000km in 2 generations?* . The waiting time T_{mig} is strongly dependent on the magnitude of the deleterious selection coefficient, however: with $s_m = 10^{-3}$, for example T_{mig} is 1.6×10^7 generations over 1,000km. *Peter please check* T_{mig} *here is correct. Original sentence (see .tex) didn't give numbers for waiting time that jived with what I get from R code in .tex*

Finally, a coalescent approach allows us to also estimate the distance travelled for a beneficial highland allele which is neutral in lowland environments. After $m = 1,000$ generations, there are approximately $n = 1,000$ lineages remaining and of these the furthest has travelled $\sqrt{2\sigma^2 m \log n} \approx 212$km from the highlands. The distance travelled, however, does not scale linearly with time, such that at $m = 6000$ generations the furthest alleles is ≈ 518km from its highland origin. *Peter I'm confused here. If the above is correct it argues a neutral allele will travel a* shorter *distance than a deleterious allele? 1000km would take neutral allele ¿6K generations but in the previous paragraph a deleterious allele with* $s_m = 10^{-4}$ *does it in 2 generations. NOTE: I think i may have messed up here, as i think the supp text doesn't jive with the numbers here. please take a look*

Alternative routes of adaptation

The lack of both empirical and theoretical support for convergent adaptation at SNPs or genes led us to investigate alternative patterns of adaptation.

We first sought to understand whether SNPs showing high differentiation between the lowlands and the highlands arose primarily via new mutations or were selected from standing genetic variation. We found that putatively adaptive variants identified in both Mexico and South America tended to segregate in the lowland population more often than other SNPs of similar mean allele frequency (85.3% vs. 74.8% in Mexico, FET $P < 10^{-9}$ and 94.8% vs 87.4% in South America, $P < 10^{-4}$). We extended this analysis by retrieving SNP data from 14 *parviglumis* inbred lines included in the Hapmap v2 data set, using only SNPs with $n \geq 10$ (Chia *et al.* 2012; Hufford *et al.* 2012b). Again we found that putatively adaptive variants were more likely to be polymorphic in *parviglumis* (78.3% vs. 72.2% in Mexico, FET $P < 0.01$ and 80.2% vs 72.8% in South America, $P < 0.01$).

While maize in highland Mexico grows in sympatry with the highland teosinte *mexicana*, maize in South America is outside the range of wild *Zea* species, leading to a marked difference in the potential for adaptive introgression from wild relatives. Pyhäjärvi *et al.* (2013) recently investigated local adaptation in *parviglumis* and *mexicana* populations, characterizing differentiation between these subspecies using an outlier approach. Genome-wide, only a small proportion of ($\sim 2 - 7\%$) of our putatively adaptive SNPs were identified by Pyhäjärvi *et al.* (2013), though these numbers are still in excess of expectations (FET $P < 10^{-3}$ for S. America and $P < 10^{-8}$ for Mexico; Table S4). The proportion of putatively adaptive SNPs shared with teosinte was twice as high in Mexico, however, leading us to evaluate our results in light of introgression identified by Hufford *et al.* (2013) from *mexicana* into maize in the Mexican highlands.

The proportion of putatively adaptive SNPs in introgressed regions of the genome in highland maize in Mexico was nearly four times higher than found in S. America (FET $P < 10^{-11}$), while differences outside introgressed regions were much smaller (7.5% vs. 6.2%; Table S6). Furthermore, of the 77 regions identified as introgressed in (Hufford *et al.* 2013), more than twice as many contain at least one F_{ST} outlier in Mexico as in S. America (23 compared to 9, one-tailed Z-test $P = 0.0027$). Excluding putatively adaptive SNPs, mean F_{ST} between Mexico and S. America is only slightly higher in introgressed regions (0.032) than across the rest of the genome (0.020), suggesting the enrichment of high F_{ST} SNPs seen in Mexico is not simply due to neutral introgression of a divergent teosinte haplotype. S6

Discussion

Our analysis of diversity and population structure in maize landraces from Mexico and S. America points to an an independent origin of S. American highland maize, in line with earlier archaeological (Piperno 2006; Perry *et al.* 2006; Grobman *et al.* 2012) and genetic (van Heerwaarden *et al.* 2011) work. We use our genetic data to fit a model of historical population size change, and find no evidence of a bottleneck in Mexico but a strong bottleneck followed by expansion in the highlands of S. America. Surprisingly, our models showed no support for a maize domestication bottleneck, apparently contradicting earlier work (Eyre-Walker *et al.* 1998; Tenaillon *et al.* 2004; Wright *et al.* 2005). One factor contributing to these differences is the set of loci sampled. Previous efforts focused on data exclusively from protein-coding regions, while our data set includes a large number of noncoding variants. Diversity differences between maize and teosinte are greatest in protein-

coding regions (Hufford *et al.* 2012b), presumably due to the effects of background selection (), and demographic estimates using only protein-coding loci should thus overestimates the strength of a domestication bottleneck. While a more detailed comparison with data from teosinte will be required to validate these results, they nonetheless suggest the value of a reassessment of the combined impacts of demography and selection on genome-wide patterns of diversity during maize domestication.

We identified SNPs deviating from patterns of allele frequencies determined by our demographic model as loci putatively under selection for highland adaptation. These conclusions are supported by evidence of haplotype differentiation *supp.* and the directionality of allele frequency change *supp.* . Consistent with results from both GWAS (**?**) and local adaptation in teosinte (Pyhäjärvi *et al.* 2013), we find that putatively adaptive SNPs are enriched in intergenic regions of the genome, further suggesting an important role for regulatory variation in maize evolution.

Although our data suggest that hundreds of loci were targeted by natural selection in Mexico and South America, fewer than $X\%$ of SNPs and $X\%$ of genes show evidence for convergent evolution between the two highland populations. To evaluate the significance of our empirical observations given the biology and colonization history of highland maize, we applied recently developed models of convergent evolution (Ralph and Coop 2014). Our modeling results suggest that convergent evolution involving identical nucleotide changes is quite unlikely due to recurrent mutation or migration across Central America via seed sharing. These results are generally robust to variation in most of the parameters *fair to say?* , but are sensitive to gross miss-estimation of some of the parameters – for example if seed sharing was common over distances of hundreds of kilometers. The modeling does highlight that our outlier approach may not detect traits undergoing convergent evolution if the genetic architecture of the trait is such that mutation at a large number of nucleotides would have equivalent effects on fitness (i.e. adaptive traits have a large mutational target). While QTL analysis suggests that some of the traits suggested to be adaptive in highland conditions may be oligogenic (Lauter *et al.* 2004), others such as flowering time (Buckler *et al.* 2009) are likely to be the result of a large number of loci, each with small and perhaps similar effects on phenotype. Future quantitative genetic analysis of highland traits using genome-wide association methods may prove useful in searching for the signal of selection on such highly quantitative traits.

Our observation of little convergent evolution is also consistent with the possibility that much of the adaptation to highland environments made use of standing genetic variation in lowland populations. Indeed, we find that as much as $X\%$ of the putatively adaptive variants in Mexico and South America are segregating in lowland populations, and the vast majority are also segregating in teosinte. Selection from standing variation should be common when the scaled mutation rate Θ(product of the effective population size, mutation rate and target size) is ≥ 1 as long as the scaled selection coefficient (product of the effective population size and selection coefficient) Ns reasonably large (Hermisson and Pennings 2005). Estimates of θ from synonymous nucleotide diversity in maize ($\cong 0.014$ (*e.g.,* Tenaillon *et al.* 2004; Wright *et al.* 2005; Ross-Ibarra *et al.* 2009)) then suggest adaptation from standing genetic variation may be likely for target sizes larger than a few hundred nucleotides. In maize, such a scenario has been recently shown for the locus *grassy tillers1* (Wills *et al.* 2013), at which adaptive variants in both an upstream control region and the 3' UTR are segregating in teosinte but show evidence of recent selection in maize, presumably due to the effects of this locus on branching and ear number.

Finally, although we evaluated a genome-wide sample of more than 90,000 SNPs, this sampling is likely insufficient to capture all of the signals of selection across the genome. Linkage disequilibrium in mays decays rapidly (**?**), reaching a plateau in only a few hundred bp (Figure S6) and a much greater density of SNPs would be needed to effectively identify the majority of selective sweeps in the history of these populations (Tiffin and Ross-Ibarra 2014). SNP density alone does not explain the lack of convergent evolution seen at SNPs showing evidence of selection, however. Our genomic sampling may have thus identified only a subset of all loci targeted by natural selection, but there is no reason to believe that the percentage of selected loci showing convergent selection should change with higher genotyping density.

Acknowledgements

We appreciate the helpful comments of P. Morrell and members of the Ross-Ibarra lab and Coop labs. This project was supported by Agriculture and Food Research Initiative Competitive Grant 2009-01864 from the USDA National Institute of Food and Agriculture and funding from the National Science Foundation IOS-1238014.

Literature Cited

Alkorta-Aranburu, G., C. M. Beall, D. B. Witonsky, A. Gebremedhin, J. K. Pritchard, *et al.*, 2012 The genetic architecture of adaptations to high altitude in Ethiopia. PLoS Genet. 8: e1003110.

Arendt, J., and D. Reznick, 2008 Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol. Evol. 23: 26–32.

Barton, N. H., 1987 The probability of establishment of an advantageous mutant in a subdivided population. Genet. Res. 50: 35–40.

Berman, S. M., 1964 Limit theorems for the maximum term in stationary sequences. Ann. Math. Statist. 35: 502–516.

Bigham, A., M. Bauchet, D. Pinto, X. Mao, J. M. Akey, *et al.*, 2010 Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet. 6: e1001116.

Brewbaker, J. L., 2014 Diversity and genetics of tassel branch numbers in maize. Crop Science.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown, *et al.*, 2009 The genetic architecture of maize flowering time. Science 325: 714–718.

Casati, P., and V. Walbot, 2005 Differential accumulation of maysin and rhamnosylisoorientin in leaves of high-altitude landraces of maize after UV-B exposure. Plant, Cell & Environment 28: 788–799.

Chia, J. M., C. Song, P. J. Bradbury, D. Costich, N. de Leon, *et al.*, 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44: 803–807.

Clark, R. M., S. Tavaré and J. Doebley, 2005 Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Mol. Biol. Evol. 22: 2304–2312.

Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villarreal Jr., M. Dickson, *et al.*, 2005 Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science 307: 1928–1933.

Currat, M., G. Trabuchet, D. Rees, P. Perrin, R. M. Harding, *et al.*, 2002 Molecular analysis of the β-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the β^s senegal mutation. Am. J. Hum. Genet. 70: 207–223.

Elmer, K. R., and A. Meyer, 2011 Adaptation in the age of ecological genomics: insights from parallelism and convergence. Trends Ecol. Evol. 26: 298–306.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6: e19379.

Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman and B. S. Gaut, 1998 Investigation of the bottleneck leading to the domestication of maize. Proc. Natl. Acad. Sci. USA 95: 4441–4446.

Falush, D., M. Stephens and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Fang, Z., T. Pyhäjärvi, A. L. Weber, R. K. Dawe, J. C. Glaubitz, *et al.*, 2012 Megabase-scale inversion polymorphism in the wild ancestor of maize. Genetics 191: 883–894.

Ganal, M. W., G. Durstewitz, A. Polley, A. Bérard, E. S. Buckler, *et al.*, 2011 A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS One 6: e28334.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire, *et al.*, 2014 TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9: e90346.

Gore, M. A., J. M. Chia, R. J. Elshire, Q. Sun, E. S. Ersoz, *et al.*, 2009 A first-generation haplotype map of maize. Science 326: 1115–1117.

Grobman, A., D. Bonavia, T. D. Dillehay, D. R. Piperno, J. Iriarte, *et al.*, 2012 Preceramic maize from Paredones and Huaca Prieta, Peru. Proc. Natl. Acad. Sci. USA 109: 1755–1759.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5: e1000695.

Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335–2352.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Huerta-Sánchez, E., X. Jin, Z. Bianba, B. M. Peter, N. Vinckenbosch, *et al.*, 2014 Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512: 194–197.

Hufford, M. B., P. Lubinsky, T. Pyhäjärvi, M. T. Devengenzo, N. C. Ellstrand, *et al.*, 2013 The genomic signature of crop-wild introgression in maize. PLoS Genet. 9: e1003477.

Hufford, M. B., E. Martinez-Meyer, B. S. Gaut, L. E. Eguiarte and M. I. Tenaillon, 2012a Past and present distributions of wild and domesticated *Zea mays*: a chance to revisit maize history. PLoS One 7: e47659.

Hufford, M. B., X. Xu, J. van Heerwaarden, T. Pyhäjärvi, J. M. Chia, *et al.*, 2012b Comparative population genomics of maize domestication and improvement. Nat. Genet. 44: 808–811.

Jagers, P., 1975 *Branching processes with biological applications*. Wiley-Interscience [John Wiley & Sons], London Wiley Series in Probability and Mathematical Statistics—Applied Probability and Statistics.

Körner, C., 2007 The use of 'altitude' in ecological research. Trends Ecol. Evol. 22: 569–574.

Kovach, M. J., M. N. Calingacion, M. A. Fitzgerald and S. R. McCouch, 2009 The origin and evolution of fragrance in rice (*Oryza sativa* L.). Proc. Natl. Acad. Sci. USA 106: 14444–14449.

Kwiatkowski, D. P., 2005 How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. 77: 171–192.

Lauter, N., C. Gustus, A. Westerbergh and J. Doebley, 2004 The inheritance and evolution of leaf pigmentation and pubescence in teosinte. Genetics 167: 1949–1959.

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, J. Sanchez G, E. Buckler, *et al.*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. USA 99: 6080–6084.

Mercer, K., A. Martínez-Vásquez and H. R. Perales, 2008 Asymmetrical local adaptation of maize landraces along an altitudinal gradient. Evolutionary Applications 1: 489–500.

Perry, L., D. H. Sandweiss, D. R. Piperno, K. Rademaker, M. A. Malpass, *et al.*, 2006 Early maize agriculture and interzonal interaction in southern Peru. Nature 440: 76–79.

Piperno, D. R., 2006 Quaternary environmental history and agricultural impact on vegetation in Central America. Annals of the Missouri Botanical Garden 93: 274–296.

Piperno, D. R., A. J. Ranere, I. Holst, J. Iriarte and R. Dickau, 2009 Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. Proc. Natl. Acad. Sci. USA 106: 5019–5024.

Pollak, E., 1966 On the survival of a gene in a subdivided population. Journal of Applied Probability 3: 142–155.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Pyhäjärvi, T., M. B. Hufford, S. Mezmouk and J. Ross-Ibarra, 2013 Complex patterns of local adaptation in teosinte. Genome Biol. Evol. 5: 1594–1609.

Ralph, P. L., and G. Coop, 2014 Convergent evolution during local adaptation to patchy landscapes. bioRxiv p. 006940.

Ross-Ibarra, J., M. Tenaillon and B. S. Gaut, 2009 Historical divergence and gene flow in the genus Zea. Genetics 181: 1399–1413.

Saghai-Maroof, M. A., K. M. Soliman, R. A. Jorgensen and R. W. Allard, 1984 Ribosomal DNA spacer-length polymorphisms in barley - Mendelian inheritance, chromosomal location, and population-dynamics. Proc. Natl. Acad. Sci. USA 81: 8014–8018.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78: 629–644.

Scheinfeldt, L. B., S. Soi, S. Thompson, A. Ranciaro, D. Woldemeskel, *et al.*, 2012 Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol. 13: R1.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, *et al.*, 2009 The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115.

Slatkin, M., 1973 Gene flow and selection in a cline. Genetics 75: 733–756.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley, *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA 98: 9161–9166.

Tenaillon, M. I., J. U'Ren, O. Tenaillon and B. S. Gaut, 2004 Selection versus demography: a multilocus investigation of the domestication process in maize. Mol. Biol. Evol. 21: 1214–1225.

Tiffin, P., and J. Ross-Ibarra, 2014 Advances and limits of using population genetics to understand local adaptation. Trends in Ecology & Evolution.

Toomajian, C., T. T. Hu, M. J. Aranzana, C. Lister, C. Tang, *et al.*, 2006 A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. PLoS Biol. 4: e137.

van Heerwaarden, J., J. Doebley, W. H. Briggs, J. C. Glaubitz, M. M. Goodman, *et al.*, 2011 Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. USA 108: 1088–1092.

van Heerwaarden, J., F. A. van Eeuwijk and J. Ross-Ibarra, 2010 Genetic diversity in a crop metapopulation. Heredity 104: 28–39.

Vigouroux, Y., J. C. Glaubitz, Y. Matsuoka, M. M. Goodman, D. Jesés Sánchez G, *et al.*, 2008 Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. Am. J. Bot. 95: 1240–1253.

Wakeley, J., 2005 *Coalescent Theory, an Introduction*. Roberts and Company, Greenwood Village, CO.

Wellhausen, E. J., A. O. Fuentes, A. H. Corzo and P. C. Mangelsdorf, 1957 *Races of Maize in Central America*. National Academy of Science, National Research Council, Washington, D. C.

Wilkes, H. G., 1977 Hybridization of maize and teosinte, in mexico and guatemala and improvement of maize. Eco. Bot. 31: 254–293.

Wills, D. M., C. J. Whipple, S. Takuno, L. E. Kursel, L. M. Shannon, *et al.*, 2013 From many, one: genetic control of prolificacy during maize domestication. PLoS Genet. 9: e1003604.

Wood, T. E., J. M. Burke and L. H. Rieseberg, 2005 Parallel genotypic adaptation: when evolution repeats itself. Genetica 123: 157–170.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley, *et al.*, 2005 The effects of artificial selection on the maize genome. Science 308: 1310–1314.

Directionality of adaptation

We classified the patterns of allelic differentiation among highland and lowland populations in Mexico and S. America together with the information of *parviglumis* in an *ad hoc* manner; the allelic differentiation pattern is consistent with highland or lowland adaptation scenario. In Figure I, we illustrate the frequency of putative ancestral and derived alleles in the five populations, drawn by red and blue, respectively.

First, we focus on the SNPs with the signature of adaptation only in Mexican populations (Figure IA). The first and second rows shows the typical patterns of highland adaptation with *parviglumis* data available. We simply assume that the allele in higher frequency in *parviglumis* is ancestral. *we have decent Tripsacum data now, should we go back and re-assess this?*

Both rows show the consistent pattern to highland adaptation in Mexico because the frequency of the putative derived allele in Mexican highlands is highly differentiated from those in both *parviglumis* and Mexican lowlands. The patterns in S. America are different between the first and second rows. However, we do not take the patterns in S. American populations into account because there is no adaptive signature in S. American. On the other hand, we should consider the allelic pattern in S. America in the case of the third row; we cannot utilize the information of *parviglumis*. It is impossible to infer the ancestral allele, so we assume the pattern is consistent with highland adaptation if one allele is in higher frequency in Mexican lowlands and S. American populations and the others is in higher frequency in Mexican highlands. We classified the SNPs into lowland adaptation in the same way (from fourth to sixth rows in Figure IA).

Next, we consider the SNPs with the signatures of adaptation in both Mexico and S. America (Figure IB). The pattern in the first row is consistent with parallel highland adaptation, whereas the second row shows parallel lowland adaptation. We cannot infer lowland or highland adaptation without the outgroup, so we ignore such SNPs. The pattern in the third row is the special case: the allele frequency is similar between Mexican lowlands and S. American highlands and similar between Mexican highlands and S. American lowlands. This pattern could be explained by that the SNP is linked to a read adaptive SNP and recombination breaks down the linkage between them.

Finally, we tested whether PHS test supports highland and lowland adaptation scenario. Consider the case of highland adaptation. We assumed that the putative derived allele is adaptive in highlands and checked whether the haplotype length is longer in highlands than that in lowlands. However, haplotype length cannot be compared directly because the derived allele frequency is different between highlands and lowlands. Thus, we compared the *P*-values of PHS test as a indicator of haplotype length given allele frequency (Pr($PHS_{xA} \leq PHS_{null|p}$ in Materials and Methods). We just say that the PHS test is consistent if the *P*-value in highlands is smaller than the *P*-value in lowlands (haplotype length is longer as *P*-value is smaller). The result is summarized in Table S3.

Demographic modeling

Throughout we use in many ways the *branching process approximation* – if an allele is locally rare, then for at least a few generations, the fates of each offspring are nearly independent. So, if the allele is locally deleterious, the total numbers of that allele behave as a subcritical branching process, destined for ultimate extinction. On the other hand, if the allele is advantageous, it will either die out or become locally common, with its fate determined in the first few generations. If the number of offspring of an individual with this allele is the random variable X, with mean $\mathbb{E}[X] = 1 + s$ (selective advantage $s > 0$), variance $\mathrm{Var}[X] = \xi^2$, and $\mathbb{P}\{X = 0\} > 0$ (some chance of leaving no offspring), then the probability of local nonextinction p_* is approximately $p_* \approx 2s/\xi^2$ to a second order in s. The precise value can be found by defining the generating function $\Phi(u) = \mathbb{E}[u^X]$; the probability of local nonextinction p_* is the minimal solution to $\Phi(1-u) = 1-u$. (This can be seen because: $1-p_*$ is the probability that an individual's family dies out; this is equal to the probability that the families of all that individuals' children die out; since each child's family behaves independently, if the individual has x offspring, this is equal to $(1-p_*)^x$; and $\Phi(1-p_*) = \mathbb{E}[(1-p_*)^X]$.)

If the selective advantage (s) depends on geographic location, a similar fact holds: index spatial location by $i \in 1, \ldots, n$, and for $u = (u_1, u_2, \ldots, u_n)$ define the functions $\Phi_i(u) = \mathbb{E}[\prod_j u_j^{X_{ij}}]$, where X_{ij} is the (random) number of offspring that an individual at i produces at location j. Then $p_* = (p_{*1}, \ldots, p_{*n})$, the vector of probabilities that a new mutation at each location eventually fixes, is the minimal solution to $\Phi(1-p_*) = 1-p_*$, i.e. $\Phi_i(1-p_*) = 1-p_{*i}$.

Here we consider a linear habitat, so that the selection coefficient at location ℓ_i is $s_i = \min(s_b, \max(-s_d, \alpha\ell_i))$. There does not seem to be a nice analytic expression for p_* in this case, but since $1-p_*$ is a fixed point of Φ, the solution can be found by iteration: $1-p_* = \lim_{n\to\infty} \Phi^n(u)$ for an appropriate starting point u.

Maize model

The migration and reproduction dynamics we use are taken largely from van Heerwaarden *et al.* (2010). On a large scale, fields of N plants are replanted each year from N_f ears, either from completely new stock (with probability p_e), from partially new stock (a proportion r_m with probability p_m), or entirely from the same field. Plants have an average of μ_E ears per plant, and ears have an average of N/N_f kernels; so a plant has on average $\mu_E N/N_f$ kernels, and a field has on average $\mu_E N$ ears and $\mu_E N^2/N_f$ kernels. *what happens if we change mean ears per plant from 3 to 1?* We suppose that a plant with the selected allele is pollen parent to $(1+s)\mu_E N/N_f$ kernels, and also seed parent to $(1+s)\mu_E N/N_f$ kernels, still in μ_E ears. The number of offspring a plant has depends on how many of its offspring kernels get replanted. Some proportion m_g of the pollen-parent kernels are in other fields, and may be replanted; but with probability p_e no other kernels (i.e. those in the same field) are replanted. Otherwise, with probability $1-p_m$ the farmer chooses N_f of the ears from this field to replant (or, $(1-r_m)N_f$ of them, with probability p_m); this results in a mean number N_f/N (or, $(1-r_m)N_f/N$) of the plant's ears of seed children being chosen, and a mean number $1+s$ of the plant's pollen children kernels being chosen. Furthermore, the field is used to completely (or partially) replant another's field with chance $p_e/(1-p_e)$ (or p_m); resulting in another N_f/N (or $r_m N_f/N$) ears and $1+s$ (or $r_m(1+s)$) pollen children being replanted elsewhere. Here we have assumed that pollen is well-mixed within a field, and that the selected allele is locally rare. Finally, we must divide all these offspring numbers by 2, since we look at the offspring carrying a particular haplotype, not of the diploid plant's genome.

The above gives mean values; to get a probability model we assume that every count is Poisson. In other words, we suppose that the number of pollen children is Poisson with random mean λ_P, and the number of seed children is a mixture of K independent Poissons with mean $(1+s)N/N_f$ each, where K is the random number of ears chosen to replant, which is itself Poisson with mean μ_K. By Poisson additivity, the numbers of local and migrant offspring are Poisson, with means $\lambda_P = \lambda_{PL} + \lambda_{PM}$ and $\mu_K = \mu_{KL} + \mu_{KM}$ respectively. With probability p_e, $\lambda_{PM} = m_g(1+s)$ and $\mu_K = \lambda_{PL} = 0$. Otherwise, with probability $(1-p_e)(1-p_m)$, $\mu_{KL} = N_f/N$ and $\lambda_{PL} = (1+s)(1-m_g)$; and with probability $(1-p_e)p_m$, $\mu_{KL} = (1-r_m)N_f/N$ and $\lambda_{PL} = (1-r_m)(1+s)(1-m_g)$. The migrant means are, with probability $(1-p_e)p_e/(1-p_e) = p_e$, $\mu_{KM} = N_f/N$ and $\lambda_{PM} = 1+s$; while with probability $(1-p_e)p_m$, $\mu_{KM} = r_m N_f/N$ and $\lambda_{PM} = (1+s)(r_m(1-m_g)+m_g)$, and otherwise $\mu_{KM} = 0$ and $\lambda_{PM} = m_g(1+s)$.

complete seed stock replacement prob	p_e	0.068
pollen migration rate	m_g	0.0083
number of plants per field	N	10^5
number of ears used to replant	N_f	561
mean ears per plant	μ_E	3
partial stock replacement prob	p_m	0.02
mean proportion stock replaced	r_m	0.2
pollen migration distance	σ_p	0 km
seed replacement distance	σ_s	50 km
distance between demes	a	15 km
width of altitudinal cline	w	62km
deleterious selection coefficient	s_d	varies
beneficial selection coefficient	s_b	varies
slope of selection gradient	α	$(s_d + s_b)/w$
variance in offspring number	ξ^2	varies
maize population density	ρ	5×10^3
area of highland habitat	A	500 km^2
mean dispersal distance	σ	1.8 km

TABLE 1 Parameter estimates used in calculations, and other notation.

Math

The generating function of a Poisson with mean λ is $\phi(u;\lambda) = \exp(\lambda(u-1))$, and the generating function of a Poisson(μ) sum of Poisson(λ) values is $\phi(\phi(u;\lambda);\mu)$. Therefore, the generating function for the diploid process, ignoring spatial structure, is

$$\begin{aligned}
\Phi(u) = {} & p_e\phi(u; m_g(1+s)) \\
& + \{(1-p_e)(1-p_m)\phi(u;(1+s)(1-m_g))\phi(\phi(u;(1+s)N/N_f);N_f/N) \\
& \quad + (1-p_e)p_m\phi(u;(1+s)(1-r_m)(1-m_g))\phi(\phi(u;(1+s)N/N_f);(1-r_m)N_f/N)\} \\
& \quad \times \{p_e/(1-p_e)\phi(u;1+s)\phi(\phi(u;(1+s)N_f/N);N_f/N) \\
& \quad\quad + p_m\phi(u;(1+s)(r_m(1-p_e)(1-m_g)+m_g)) \\
& \quad\quad\quad \times \phi(\phi(u;(1+s)N/N_f);r_mN_f/N) \\
& \quad\quad + (1-p_e/(1-p_e)-p_m)\phi(u;m_g(1+s))\}
\end{aligned} \tag{1}$$

$$\begin{aligned}
= {} & \phi(u;m_g(1+s))\big(\, p_e \\
& + \{(1-p_e)(1-p_m)\phi(u;(1+s)(1-m_g))\phi(\phi(u;(1+s)N/N_f);N_f/N) \\
& \quad + (1-p_e)p_m\phi(u;(1+s)(1-r_m)(1-m_g))\phi(\phi(u;(1+s)N/N_f);(1-r_m)N_f/N)\} \\
& \quad \times \{p_e/(1-p_e)\phi(u;(1+s)(1-m_g))\phi(\phi(u;(1+s)N_f/N);N_f/N) \\
& \quad\quad + p_m\phi(u;(1+s)r_m(1-m_g)) \\
& \quad\quad\quad \times \phi(\phi(u;(1+s)N/N_f);r_mN_f/N) \\
& \quad\quad + (1-p_e/(1-p_e)-p_m)\}\,\big)
\end{aligned} \tag{2}$$

To get the generating function for a haploid, replace every instance of $1+s$ by $(1+s)/2$.

As a quick check, the mean total number of offspring of a diploid is

$$\begin{aligned}
& (1+s)\big(m_g + (1-p_e)\{(1-p_m)((1-m_g)+1) + p_m((1-r_m)(1-m_g)+(1-r_m))\} \\
& \qquad + \{p_e((1-m_g)+1) + p_m(1-p_e)(r_m(1-m_g)+r_m)\}\,\big)
\end{aligned} \tag{3}$$

$$= (1+s)\big(m_g + (1-p_e)(2-m_g)(1-p_mr_m) + (p_e(2-m_g) + p_mr_m(1-p_e)(2-m_g))\big) \tag{4}$$

$$= (1+s)\big(m_g + (2-m_g)((1-p_e)(1-p_mr_m) + p_e + p_mr_m(1-p_e))\big) \tag{5}$$

$$= (1+s)\big(m_g + (2-m_g)\big) \tag{6}$$

$$= 2(1+s). \tag{7}$$

(CHECK)

We show numerically later that the probability of establishment is very close to $2s$ over the variance in reproductive number (as expected). It is possible to write down an expression for the variance, but it's a big, ugly one that doesn't lend itself to intuition.

Migration and spatial structure

To incorporate spatial structure, suppose that the locations ℓ_k are arranged in a regular grid, so that $\ell_k = ak$. Recall that s_k is the selection coefficient at location k. If the total number of offspring produced by an individual at ℓ_i is Poisson(λ_i), with each offspring independently migrating to location j with probability m_{ij}, then the number of offspring at j is Poisson($m_{ij}\lambda_i$), and so the generating function is

$$\phi(u;\lambda,m) = \prod_j \exp(\lambda_i m_{ij}(u_j - 1)) \tag{8}$$

$$= \exp\left\{\lambda_i\left(\left(\sum_j m_{ij}u_j\right) - 1\right)\right\}. \tag{9}$$

We can then substitute this expression into equation (1), with appropriate migration kernels for pollen and seed dispersal.

For migration, we need migration rates and migration distances for both wind-blown pollen and for farmer seed exchange. The rates are parameterized as above; we need the typical dispersal distances, however. One option is to say that the typical distance between villages is d_v, and that villages are discrete demes, so that pollen stays within the deme (pollen migration distance 0) and seed is exchanged with others from nearby villages; on average σ_s distance away in a random direction. The number of villages away the seed comes from could be geometric (including the possibility of coming from the same village).

The dispersal distance – the mean distance between parent and offspring – is the average of the pollen and seed mean dispersal distances. With the above assumptions, the pollen dispersal distance is zero, and the seed dispersal distance is the chance of inter-village movement multiplied by the mean distance moved. This is

$$\sigma = \frac{1}{2}(p_e + (1-p_e)p_m r_m)\sigma_s = 1.7932\text{km} \tag{10}$$

at the parameter values above.

Iterating the generating function above finds the probability of establishment as a function of distance along the cline. This is shown in figure 1. Note that the approximation $2s$ divided by the variance in offspring number is quite close.

As we show in Ralph and Coop (2014), the rate of adaptation by diffusive migration is roughly

$$\lambda_{\text{mig}} = \rho\frac{s_b\sqrt{2s_m}}{2\xi^2}\exp\left(-\frac{\sqrt{2s_m}R}{\sigma}\right).$$

do we need to expland? not sure.

First note that for $10^{-1} \leq s_m \leq 10^{-4}$, the value $1/\sqrt{2s_m}$ is between 2 and 70 – so the exponential decay of the chance of migration falls off on a scale of between 2 and 70 times the dispersal distance. Above we have estimated the dispersal distance to be $\sigma \approx 2$ km, and far below the mean distance σ_s to the field that a farmer replants seed from, when this happens, which we have as $\sigma_s = 50$ km. Taking $\sigma = 2$ km, we have that $4 \leq \sigma/\sqrt{2s_m} \leq 150$ km. A very conservative upper bound might be $\sigma \leq \sigma_s/20$ (if farmers replaced 10% of their seed with long-distance seed every year). At this upper bound, we would have $5 \leq \sigma/\sqrt{2s_m} \leq 175$ km, which is not very different. This makes the exponential term very small since R is on the order of 1,000 km.

Taking $\sigma = 2$ km, we then compute that if $s_m = 10^{-4}$ (very weak selection in the lowlands), then for $R = 1,000$ km, the migration rate is $\lambda_{\text{mig}} \leq 10^{-5}$, i.e. it would take on the order of 100,000 generations (years) to get a successful migrant only 1,000 km away, under this model of undirected, diffusive dispersal. For larger s_m, the migration rate is much smaller.

If highland alleles are neutral in the lowlands the situation is more difficult to model, but we can make some informed guesses. For maize in the Andean highlands to have inherited a highland-adapted allele from the Mexican highlands, those Andean plants must be directly descended from highland Mexican plants that lived more recently than the appearance of the adaptive allele. In other words, the ancestral lineages along which the modern Andean plants have inherited at that locus must trace back to the Mexican highlands. If the allele is neutral in the lowlands, we can treat the movement of these lineages as a neutral process, using the framework of coalescent theory (Wakeley 2005). To do this, we need to follow *all* of the $N \approx 2.5 \times 10^6$ lineages backwards; these quickly coalesce to fewer m lineages in approximately $\sum_{k=m}^{N} \frac{2N}{\xi^2 k(k+1)} \approx 1.25 \times 10^5/m$ generations, leaving

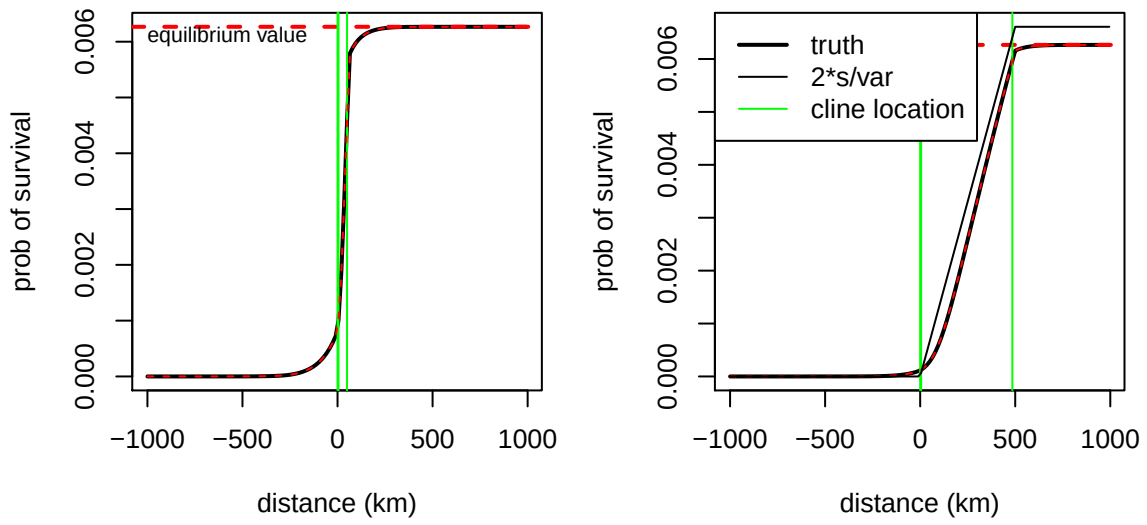

FIGURE 1 *(make this look better)* Probability of establishment, as a function of distance along and around an altitudinal cline, whose boundaries are marked by the green lines. **(A)** The parameters above; with cline width 62km; **(B)** the same, except with cline width 500km.

about 1000 lineages after 100 generations that are spread over a larger area. The displacement of a lineage after m generations has variance $m\sigma^2$ and is approximately Gaussian. If we assume that n lineages are independent, and Z_n is the distance to the furthest lineage, then $\mathbb{P}\{Z_n/\sqrt{m\sigma^2} \le x/\sqrt{2\log n} + \sqrt{2\log n} - (1/2)(\log\log n + \log 4\pi)/\sqrt{2\log n}\} \approx \exp(-e^{-x})$ (Berman 1964). *Peter are the braces in the right place? seems to me naively it should be* $\mathbb{P}\{Z_n/\sqrt{m\sigma^2} \le x\} = \sqrt{2\log n} + \ldots$ *??*

1 Adaptation by mutation

(just a placeholder for now; to be merged in)

First, we'd like to compute how difficult is it for the beneficial adaptation to arise by new mutation. The rate of appearance of mutant alleles is a Poisson process, and we can assume that each is successful or not independently, so the time until the new mutant appears and fixes is exponentially distributed, with rate equal to the mutation rate multiplied by the probability of establishment integrated over the population. Referring to figure 1, we see that this is pretty close to ((area of high altitude) + (1/2 area of altitudinal gradient)) × (population density) × (prob of establishment at high altitude).

Let A denote (area of high altitude) plus (1/2 area of altitudinal gradient). The population density ρ is roughly 0.5–5 people per km^2× (0.5 ha field/person) × (2×10^4 plants per field ha) = (5000–50000 plants per km^2). As a check, the other set of numbers was "one village per 15 km"; i.e. per square with 15km on a side, which is 0.444 people per km^2.

Since the probability of establishment at high altitude is approximately $2s_b/\xi^2$, with ξ^2 the variance in offspring number, the rate of appearance is just

$$\lambda_{\text{mut}} = 2\rho A s_b \mu/\xi^2.$$

At the values above, with $.1 \le s_b \le .001$, the factor $2\rho A s_b/\xi^2$ multiplying the mutation rate varies between 10^2 and 10^5, implying that a single-base mutation with $\mu = 10^{-8}$ would have to wait between 10^4 and 10^6 generations to fix, but a mutation with a larger target, say $\mu = 10^{-5}$, would fix in tens to thousands of generations, depending on the selection coefficient.

2 Adaptation by migration

3 Conclusion

It seems unlikely that any alleles that are adaptive in the highlands and deleterious at all in the lowlands would have transited central America by undirected (diffusive) sharing of seed. The conclusions could change if we drastically underestimate the rate

of very long distance sharing of seed, e.g. if sharing across hundreds of kilometers was common at some point.

Both calculations are very pessimistic about the chance of shared single-base changes through either migration or independent mutation. However, independent mutations could be expected in kilobase-size targets, suggesting there might be signal for genes that share adaptive changes.

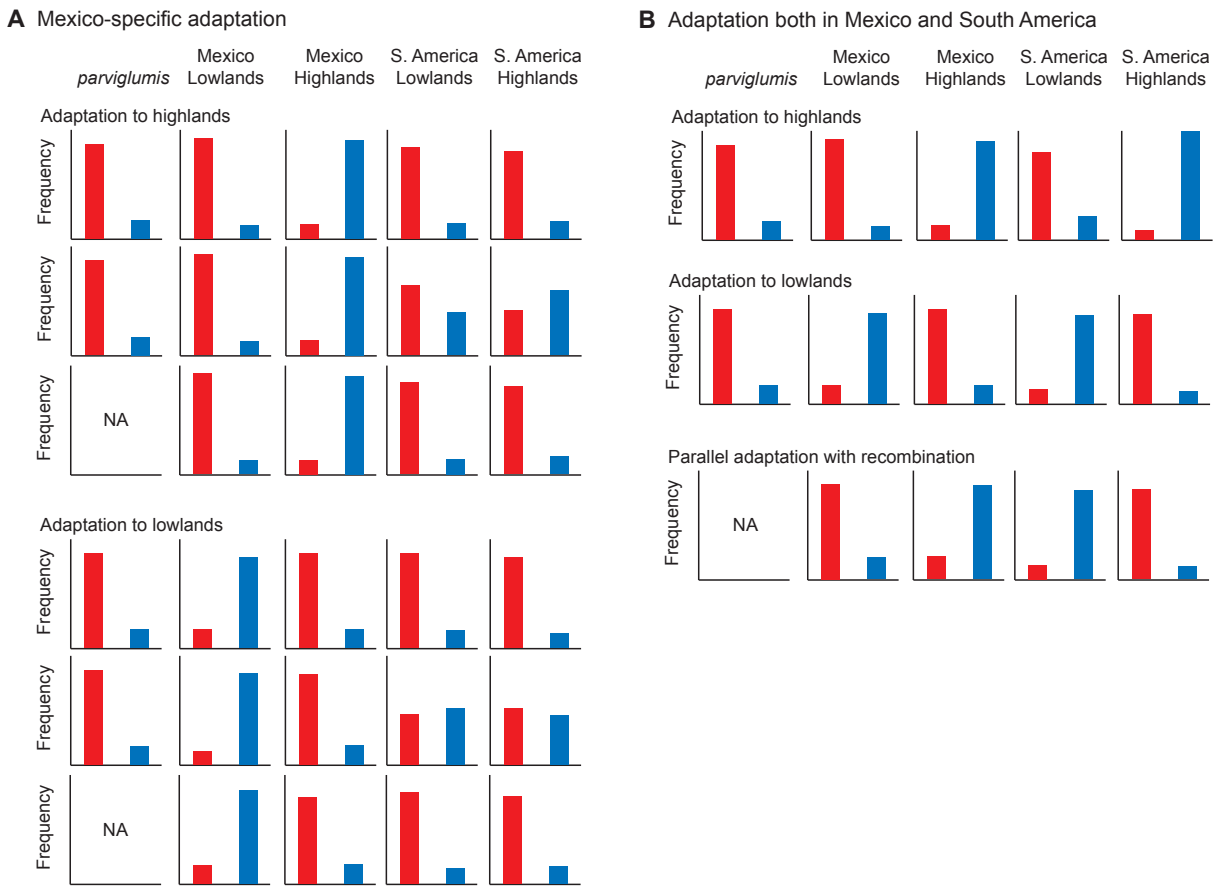

FIGURE I Illustration of allele frequency changes in maize and *parviglumis*. Red and blue bars represent the allele frequency of ancestral and derived, adaptive alleles, respectively. The allele frequencies in the five populations are shown: *parviglumis*, Mexican lowlands and highlands, and S. America lowlands and highlands. NA in *parviglumis* indicates that there is no SNP data in the site.

TABLE S1 List of maize landraces used in this study

ID[a]	USDA ID	Population	Landrace	Locality	Latitude	Longitude	Elevation	Origin
RIMMA0409	PI 478968	Mexico	Tepecintle	Chiapas, Mexico	15.4	-92.9	107	USDA
RIMMA0410	PI 478970	Lowland	Vandeno	Chiapas, Mexico	15.4	-92.9	107	USDA
RIMMA0433	PI 490825		Nal Tel ATB	Chiquimula, Guatemala	14.7	-89.5	457	USDA
RIMMA0441	PI 515538		Coscomatepec	Veracruz, Mexico	19.2	-97.0	1320	USDA
RIMMA0615	PI 628480		Tuxpeno	Puebla, Mexico	20.1	-97.2	152	USDA
RIMMA0619	PI 645772		Pepitilla	Guerrero, Mexico	18.4	-99.5	747	USDA
RIMMA0628	PI 646017		Tuxpeno Norteno	Tamaulipas, Mexico	23.3	-99.0	300	USDA
RIMMA0696	Ames 28568		Tuxpeno	El Progreso, Guatemala	16.5	-90.2	30	Goodman
RIMMA0700	NSL 291626		Olotillo	Chiapas, Mexico	16.8	-93.2	579	Goodman
RIMMA0701	PI 484808		Olotillo	Chiapas, Mexico	16.6	-92.7	686	Goodman
RIMMA0702	Ames 28534		Negro de Tierra Caliente	Sacatepequez, Guatemala	14.5	-90.8	1052	Goodman
RIMMA0703	NSL 283390		Nal Tel	Yucatan, Mexico	20.8	-88.5	30	Goodman
RIMMA0709	Ames 28452		Tehua	Chiapas, Mexico	16.5	-92.5	747	Goodman
RIMMA0710	PI 478988		Tepecintle	Chiapas, Mexico	15.3	-92.6	91	Goodman
RIMMA0712	NSL 291696 CYMT		Oloton	Baja Verapaz, Guatemala	15.3	-90.3	1220	Goodman
RIMMA0716	Ames 28459		Zapalote Grande	Chiapas, Mexico	15.3	-92.7	91	Goodman
RIMMA0720	PI 489372		Negro de Tierra Caliente	Guatemala	15.5	-88.9	39	Goodman
RIMMA0721	Ames 28485		Nal Tel ATB	Chiquimula, Guatemala	14.6	-90.1	915	Goodman
RIMMA0722	Ames 28564		Dzit Bacal	Jutiapa, Guatemala	14.3	-89.7	737	Goodman
RIMMA0727	Ames 28555		Comiteco	Guatemala	14.4	-90.5	1151	Goodman
RIMMA0729	PI 504090		Tepecintle	Guatemala	15.4	-89.7	122	Goodman
RIMMA0730	Ames 28517		Quicheno Late	Sacatepequez, Guatemala	14.5	-90.8	1067	Goodman
RIMMA0731	PI 484137		Bolita	Oaxaca, Mexico	16.8	-96.7	1520	Goodman
RIMMA0733	PI 479054		Zapalote Chico	Oaxaca, Mexico	16.6	-94.6	107	Goodman
RIMMA0416	PI 484428	Mexico	Cristalino de Chihuahua	Chihuahua, Mexico	29.4	-107.8	2140	NA
RIMMA0417	PI 484431	Highland	Azul	Chihuahua, Mexico	28.6	-107.5	2040	USDA
RIMMA0418	PI 484476		Gordo	Chihuahua, Mexico	28.6	-107.5	2040	USDA
RIMMA0421	PI 484595		Conico	Puebla, Mexico	19.9	-98.0	2250	USDA
RIMMA0422	PI 485071		Elotes Conicos	Puebla, Mexico	19.1	-98.3	2200	USDA
RIMMA0423	PI 485116		Cristalino de Chihuahua	Chihuahua, Mexico	29.2	-108.1	2095	NA
RIMMA0424	PI 485120		Apachito	Chihuahua, Mexico	28.0	-107.6	2400	USDA
RIMMA0425	PI 485128		Palomero Tipo Chihuahua	Chihuahua, Mexico	26.8	-107.1	2130	USDA
RIMMA0614	PI 628445		Mountain Yellow	Jalisco, Mexico	20.0	-103.8	2060	USDA
RIMMA0616	PI 629202		Zamorano Amarillo	Jalisco, Mexico	20.8	-102.8	1800	USDA
RIMMA0620	PI 645786		Celaya	Guanajuato, Mexico	20.2	-100.9	1799	USDA
RIMMA0621	PI 645804		Zamorano Amarillo	Guanajuato, Mexico	21.1	-101.7	1870	USDA
RIMMA0623	PI 645841		Palomero de Jalisco	Jalisco, Mexico	20.0	-103.7	2520	USDA
RIMMA0625	PI 645984		Cacahuacintle	Puebla, Mexico	19.0	-97.4	2600	USDA
RIMMA0626	PI 645993		Arrocillo Amarillo	Puebla, Mexico	19.9	-97.6	2260	USDA
RIMMA0630	PI 646069		Arrocillo Amarillo	Veracruz, Mexico	19.8	-97.3	2220	USDA
RIMMA0670	Ames 28508		San Marceno	San Marcos, Guatemala	15.0	-91.8	2378	Goodman
RIMMA0671	Ames 28538		Salpor Tardio	Solola, Guatemala	14.8	-91.3	2477	Goodman
RIMMA0672	PI 483613		Chalqueno	Mexico, Mexico	19.7	-99.1	2256	Goodman
RIMMA0674	PI 483617		Toluca	Mexico, Mexico	19.3	-99.7	2652	Goodman
RIMMA0677	Ames 28476		Conico Norteno	Zacatecas, Mexico	21.4	-102.9	1951	Goodman
RIMMA0680	Ames 28448		Tabloncillo	Jalisco, Mexico	20.4	-102.2	1890	Goodman
RIMMA0682	PI 484571		Tablilla de Ocho	Jalisco, Mexico	22.1	-103.2	1700	Goodman
RIMMA0687	Ames 28473		Conico Norteno	Queretaro, Mexico	20.4	-100.0	1921	Goodman

[a] GBS data are available for the accessions in bold font.

TABLE S1 (continued)

ID	USDA ID	Population	Landrace	Locality	Latitude	Longitude	Elevation (m)	Origin
RIMMA0388	PI 443820	South America	Amagaceno	Antioquia, Colombia	6.9	-75.3	1500	USDA
RIMMA0389	PI 444005	Lowland	Costeno	Atlantico, Colombia	10.4	-74.9	7	USDA
RIMMA0390	PI 444254		Comun	Caldas, Colombia	4.5	-75.6	353	USDA
RIMMA0391	PI 444296		Andaqui	Caqueta, Colombia	1.4	-75.8	700	USDA
RIMMA0392	PI 444309		Andaqui	Caqueta, Colombia	1.8	-75.6	555	USDA
RIMMA0393	PI 444473		Costeno	Cordoba, Colombia	8.3	-75.2	100	USDA
RIMMA0394	PI 444621		Pira	Cundinamarca, Colombia	4.8	-74.7	1000	USDA
RIMMA0395	PI 444731		Negrito	Choco, Colombia	8.5	-77.3	30	USDA
RIMMA0396	PI 444834		Caqueteno	Huila, Colombia	2.6	-75.3	1100	USDA
RIMMA0397	PI 444897		Negrito	Magdalena, Colombia	11.6	-72.9	50	USDA
RIMMA0398	PI 444923		Puya	Magdalena, Colombia	9.4	-75.7	27	USDA
RIMMA0399	PI 444954		Cariaco	Magdalena, Colombia	10.2	-74.1	250	USDA
RIMMA0403	PI 445163		Pira Naranja	Narino, Colombia	1.3	-77.5	1000	USDA
RIMMA0404	PI 445322		Puya Grande	Norte de Santander, Colombia	7.3	-72.5	1500	USDA
RIMMA0405	PI 445355		Puya	Norte de Santander, Colombia	8.4	-73.3	1100	USDA
RIMMA0406	PI 445514		Yucatan	Tolima, Colombia	5.0	-74.9	450	USDA
RIMMA0407	PI 445528		Pira	Tolima, Colombia	4.2	-74.9	450	USDA
RIMMA0428	PI 485354		Aleman	Huanuco, Peru	-9.3	-76.0	700	NA
RIMMA0462	PI 445073		Amagaceno	Narino, Colombia	1.6	-77.2	1700	USDA
RIMMA0690	PI 444946		Puya	Magdalena, Colombia	8.3	-73.6	250	Goodman
RIMMA0691	PI 445391		Cacao	Santander, Colombia	6.6	-73.1	1098	NA
RIMMA0707	PI 487930		Tuxpeno	Ecuador	-1.1	-80.5	30	Goodman
RIMMA0708	PI 488376		Yunquillano F Andaqui	Ecuador	-3.5	-78.6	1098	Goodman
RIMMA0426	PI 485151	South America	Rabo de Zorro	Ancash, Peru	-9.1	-77.8	2500	NA
RIMMA0430	PI 485362	Highland	Sarco	Ancash, Peru	-9.2	-77.7	2585	NA
RIMMA0431	PI 485363	(Andean)	Perlilla	Huanuco, Peru	-8.7	-77.1	2900	NA
RIMMA0436	PI 514723		Morocho Cajabambino	Amazonas, Peru	-6.2	-77.9	2200	NA
RIMMA0437	PI 514752		Ancashino	Ancash, Peru	-9.3	-77.6	2688	NA
RIMMA0438	PI 514809		Maranon	Ancash, Peru	-8.7	-77.4	2820	NA
RIMMA0439	PI 514969		Maranon	La Libertad, Peru	-8.5	-77.2	2900	NA
RIMMA0464	PI 571438		Chullpi	Huancavelica, Peru	-12.3	-74.7	1800	USDA
RIMMA0465	PI 571457		Huarmaca	Piura, Peru	-5.6	-79.5	2300	USDA
RIMMA0466	PI 571577		Confite Puneno	Apurimac, Peru	-14.3	-72.9	3600	USDA
RIMMA0467	PI 571871		Paro	Apurimac, Peru	-13.6	-72.9	2800	USDA
RIMMA0468	PI 571960		Sarco	Ancash, Peru	-9.4	-77.2	3150	USDA
RIMMA0473	PI 445114		Sabanero	Narino, Colombia	1.1	-77.6	3104	USDA
RIMMA0656	Ames 28799		Culli	Jujuy, Argentina	-23.2	-65.4	2287	Goodman
RIMMA0657	NSL 286594		Chake Sara	Bolivia	-17.5	-65.7	2201	Goodman
RIMMA0658	NSL 286812		Uchuquilla	Bolivia	-21.8	-64.1	1948	Goodman
RIMMA0661	PI 488066		Chillo	Ecuador	-2.9	-78.7	2195	Goodman
RIMMA0662	NSL 287008		Cuzco	Ecuador	0.0	-78.0	2195	Goodman
RIMMA0663	PI 488102		Mishca	Ecuador	0.4	-78.2	2067	Goodman
RIMMA0664	PI 488113		Blanco Blandito	Ecuador	0.4	-78.4	2122	Goodman
RIMMA0665	PI 489324		Racimo de Uva	Ecuador	-0.9	-78.9	2931	Goodman
RIMMA0667	Ames 28737		Patillo	Chuquisaca, Bolivia	-21.8	-64.1	2201	NA
RIMMA0668	Ames 28668		Granada	Puno, Peru	-14.9	-70.6	3925	Goodman

[a] GBS data are available for the accessions in bold font.

TABLE S2 Inference of demographic parameters

Mexico	Model IA
Likelihood	-3052.34
N_B	148,500
N_C	148,500
N_1	62,370
N_2	86,130
N_{2P}	86,130
South America	**Model IA**
Likelihood	-2717.64
N_B	76,500
N_C	76,500
N_1	74,205
N_2	2,295
N_{2P}	346,545

The description of α, β and γ is in Figure 3.
σ is a relative size of N_B to N_C ($N_B = \sigma N_C$).

TABLE S3 Summary of PHS test

Population	Pattern of adaptation	No. SNPs	No. SNPs supported by PHS test
Mexico	Highland adaptation	264	172 (65.2%)
	Lowland adaptation	101	66 (65.3%)
S. America	Highland adaptation	164	230 (71.3%)
	Lowland adaptation	70	50 (71.4%)

TABLE S4 F_{CT} between *parviglumis* and *mexicana*

Mexico	Number of SNPs		
	Significant	NS	Proportion
Significant F_{CT}	25	337	0.077
NS	299	18,493	0.018
S. America	**Number of SNPs**		
	Significant	NS	Proportion
Significant F_{CT}	10	327	0.070
NS	133	17,518	0.018

TABLE S5 ms command

Model I for Mexico populations
Population 1: Mexico lowland population
Population 2: Mexico highland population
-I 2 n_{m1} n_{m2} -n 1 0.3496 -n 2 0.5704 -ej 0.01 2 1 -en 0.01 1 0.92 -en 0.0133 1 0.0163 -en 0.015 1 1.0
Model II for Mexico populations
Population 1: Mexico lowland population
Population 2: Mexico highland population
Population 3: *mexicane* population
-I 2 n_{m1} n_{m2} -n 1 1.14 -n 2 0.36 -es 0.01 2 0.8 -en 0.01 3 1.0667 -ej 0.01 2 1 -en 0.01 1 1.5 -en 0.0133 1 0.0163 -en 0.015 1 1.0 -ej 0.1 3 1
Model I for SA populations
Population 1: SA lowland population
Population 2: SA highland population
-I 2 n_{s1} n_{s2} -n 1 0.5044 -n 2 1.3728 -g 2 671.60 -ej 0.006667 2 1 -eg 0.006667 2 0.0 -en 0.00667 1 0.52 -en 0.01333 1 0.0163 -en 0.015 1 1.0
Model III for SA populations
Population 1: Mexico lowland population
Population 2: SA lowland population
Population 3: SA highland population
-I 3 n_{m1} n_{s1} n_{s2} -n 1 0.64 -n 2 0.342 -n 3 0.972 -g 3 598.35 -ej 0.006667 3 2 -eg 0.006667 3 0.0 -en 0.006667 2 0.36 -ej 0.01 2 1
-en 0.01 1 1 -en 0.0133 1 0.0163 -en 0.015 1 1.0
Sample size of Mexico lowland, Mexico highland, SA lowland and SA highland populations are denoted by n_{m1}, n_{m2}, n_{s1} and n_{s2}, respectively.

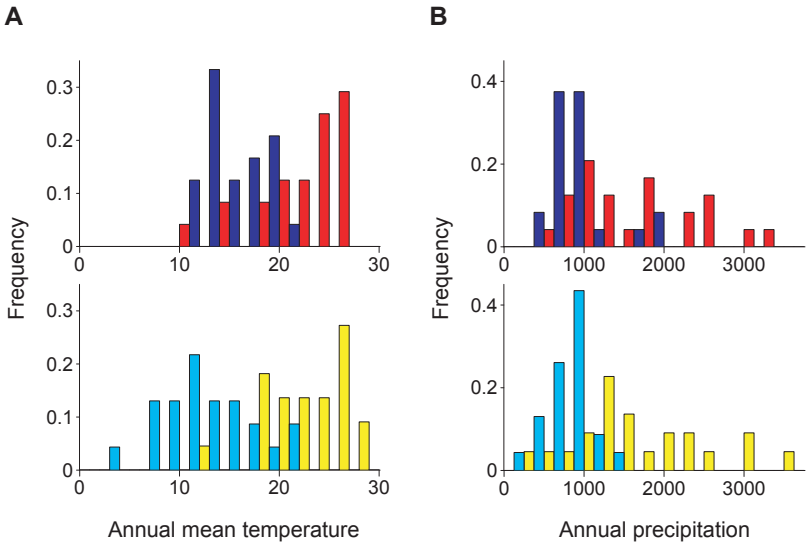

FIGURE S1 Correlation of allele frequencies between GBS (*x*-axes) and MaizeSNP50 (*y*-axes) data. We used overlapped SNPs with $n \geq 40$ for both data sets. Correlation coefficient is 0.890 ($P < 10^{-5}$ by permutation test with 10^5 replications).

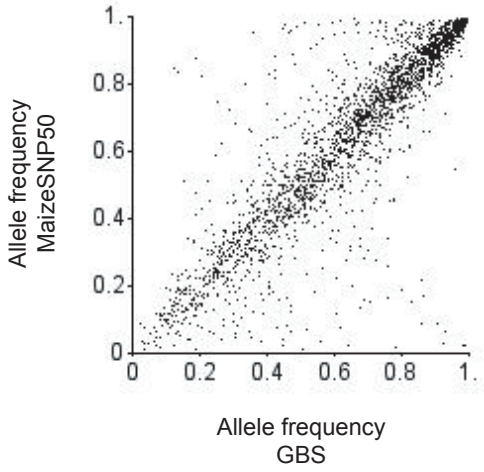

FIGURE S2 Correlation of allele frequencies between GBS (*x*-axes) and MaizeSNP50 (*y*-axes) data. We used overlapped SNPs with $n \geq 40$ for both data sets. Correlation coefficient is 0.890 ($P < 10^{-5}$ by permutation test with 10^5 replications).

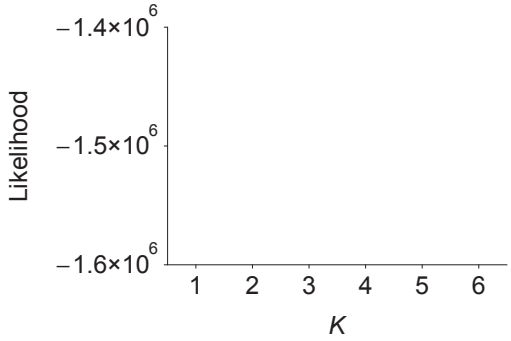

FIGURE S3 Likelihood of STRUCTURE analysis given K. The *x*-axes represents K and the *y*-axes represents likelihood.

TABLE S6 F_{ST} outlier SNPs and *mexicana* introgression

Introgression status	Population	F_{ST} outlier SNPs	all other SNPs
Introgressed	Mexico	114	1953
	S. America	26	1721
Not introgressed	Mexico	558	73892
	S. America	379	60666

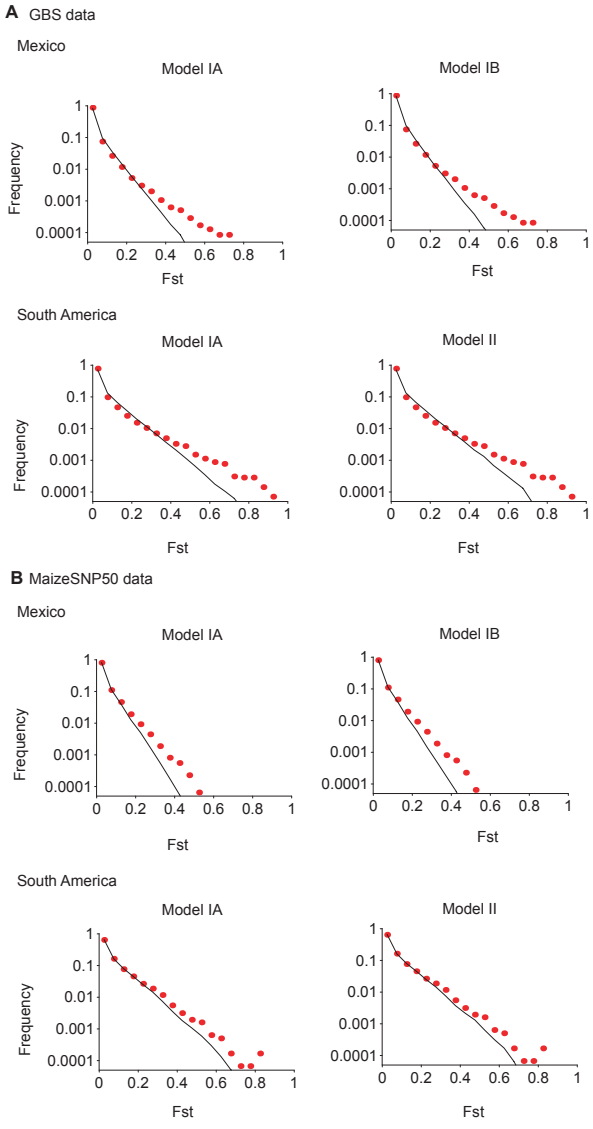

FIGURE S4 Observed and expected distributions of F_{ST} values in GBS (A) and MaizeSNP50 data (B). The *x*-axes represent F_{ST} values. The *y*-axes represent the frequency of SNPs with F_{ST} values within a bin of 0.05 size. Red dots and solid lines indicate observed and expected distributions.

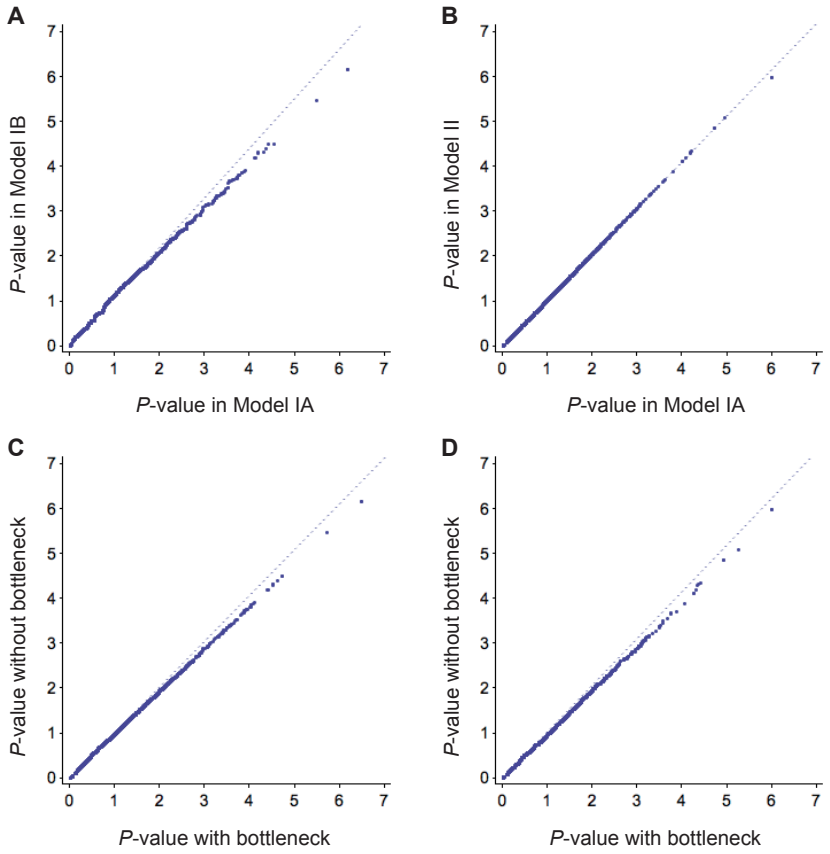

FIGURE S5 Q-Q plot for $-\log_{10}$-scaled *P*-values of population differentiation between lowland and highland populations. (A) Model IA *v.s.* Model IB in Mexico, (B) Model IA *v.s.* Model II in S. America, (C) Model with *v.s.* without bottleneck in Mexico and (D) Model with *v.s.* without bottleneck in S. America.

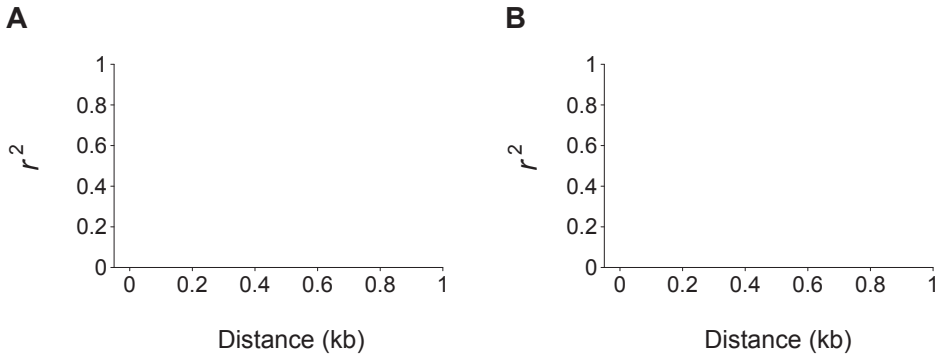

FIGURE S6 Pattern of decay of linkage equilibrium in Mexico (A) and South America (B). Red and blue dots represent low- and highland population, respectively. r^2 values were calculated as a statistics and averaged within 10-bp bins of distance between SNPs. The *x*- and *y*-axes represent distance between SNPs (kb) and average r^2 values.