Independent molecular basis of convergent highland adaptation in maize


*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, §School of Plant Sciences, Section of Plant Breeding and Genetics, Cornell University, Ithaca, New York 14853-2703, USA, **Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853-2703, USA, ††United States Department of Agriculture Agricultural Research Service, Ithaca, NY 14853, USA, ‡‡The Center for Population Biology and the Genome Center, University of California, Davis, California 95616, USA, 1Present address: SOKENDAI (Graduate university for advanced studies), Hayama, Kanagawa 240-0193, Japan

June 10, 2015

ABSTRACT Convergent evolution is the independent evolution of similar traits in different species or lineages of the same species; this often is a result of adaptation to similar environments, a process referred to as convergent adaptation. We investigate here the molecular basis of convergent adaptation in maize to highland climates in Mesoamerica 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 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 highland 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 that lead to 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).

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 ori-
gin may be due to a complex genetic architecture or standing genetic variation, introgression from related populations may also play a 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). Beyond these examples, however, 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 (Zea mays ssp. mays) to high elevation environments 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 ~6,000 BP (Piperno 2006), and the Andean highlands by ~4,000 BP (Perry et al. 2006; Grobman et al. 2012). The transition from lowland to highland habitats spanned similar environmental gradients in Mesoamerica and S. 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 Mesoamerican and S. 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 and stem pigmentation (Wilkes 1977; Wellhausen et al. 1957), differences in tassel branch and ear husk number (Brewbaker 2014), and a changed 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 (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 independent origins for altitude-adapted traits.

Here we use genome-wide SNP data from Mesoamerican 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 Mesoamerica and S. 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 across highland populations at the nucleotide or gene level, a result we show is consistent with expectations from recent theoretical work on convergent adaptation (Ralph and Coop 2014a). Instead, our results support a role for 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 landrace maize accessions from high and low elevation sites in Mesoamerica 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 assessed through inspection on a 2% agarose gel and 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 (Galán et al. 2011). SNPs were clustered with the default algorithm of the GenomeStudio Genotyping Module v1.0
Figure 1 (A) Sampling locations of landraces. Red, blue, yellow and light blue dots represent Mesoamerican lowland, Mesoamerican highland, S. American lowland and S. American highland populations, respectively. (B) Results of STRUCTURE analysis of the maizeSNP50 SNPs with \( K = 2 \) ⇐\( 4 \). 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).

(Illumina Inc., San Diego, CA, USA) and then visually inspected and manually adjusted. These data are referred to as “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 Ref-Gen 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 only synonymous and noncoding SNPs from the MaizeSNP50 data due to its low error in identifying heterozygous genotypes. 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 Models of historical population size for lowland and highland populations. Parameters in bold were estimated in this study. See text for details.

We calculated the observed joint frequency distributions (JFDs) using only the GBS data set due to its lower level of ascertainment bias. A subset of synonymous and noncoding SNPs were utilized that had $\geq 15$ 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. 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 assuming multinomial sampling. We did not use the “full” model that incorporates all four populations because parameter estimation under this model is computationally infeasible.

Model IA This model is applied separately to both the Mesoamerican and the 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 $\beta$ such that the population is divided by $\beta N_C$ 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 $\gamma$ times larger than that at $t_F$.

Model IB We expand Model IA for the Mesoamerican populations by incorporating admixture from the teosinte mexicana to the highland Mesoamerican 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 Mesoamerican highland population is derived from admixture between the Mesoamerican lowland population and a portion $P_{mex}$ from the teosinte mexicana.

Model II The final model includes the Mesoamerican lowland, S. American lowland and highland populations. This model was used for simulating SNPs with ascertainment bias (see below). At time $t_F$, the Mesoamerican and S. American lowland populations are differentiated, and the sizes of populations after splitting are determined by $\beta_1$. At time $t_G$, the S. American lowland and highland populations are differentiated, and the sizes of populations at this time are determined by $\beta_2$. As in Model IA, the S. American highland population is assumed to experience population growth with the parameter $\gamma$.

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 Mesoamerican 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 ($\alpha$, $\beta$ and $\gamma$), 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 ($\alpha$, $\beta_1$, $\beta_2$ and $\gamma$).
**Population differentiation**

We used our inferred models of population size change to generate a null distribution of $F_{ST}$ from the expected JFD estimated in $\delta$-ability (Gutenkunst et al. 2009). 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 Mesoamerican lowland maize. We thus added two additional individuals to the Mesoamerican 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 Mesoamerican 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 m5 (Hudson 2002). Then, we calculated $P$-values from the JFD in the same way. We calculated $F_{ST}$ values for all SNPs that had $\geq 10$ 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

$$PHS_{x,A} = \sum_{i=1}^{p-1} \left( \sum_{j=i+1}^{p} \frac{Z_{ijx}}{p(p-2)} - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{Z_{ijx}}{n(n-2)} \right),$$  \hspace{1cm} (1)

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

$$Z_{ijx} = \frac{d_{ijx} - \bar{d}_{ij}}{\sigma_{ij}},$$  \hspace{1cm} (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 $i$ and $j$ are identical, and $\sigma_{ij}$ is the standard deviation of the distribution of distances. 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).

subsection*{Polarizing adaptive alleles} To polarize the ancestral state of alleles and help identify adaptive alleles, we retrieved 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). Alleles were called ancestral if they were at higher frequency in *parviglumis* or uncalled in *parviglumis* but at higher frequency in all populations but one.

For SNPs identified as putative outliers by our $F_{ST}$ approach, we then used patterns of allele frequency across populations to infer which allele was likely adaptive. For SNPs with a significant $F_{ST}$ only in Mesoamerica, for example, we characterized them as adaptive if they were at high frequency in one Mesoamerican population (lowland or highland) and low frequency in the other as well as low frequency in *parviglumis* and at most intermediate frequency (or low frequency if missing in *parviglumis*) in S. American populations. SNPs were inferred to show convergent adaptation if they were at high frequency in both highland (or lowland) populations, and at low frequency in the other two populations and *parviglumis*.

**Theoretical evaluation of convergent evolution**

We next asked whether the abundance and degree of coincidence of presumably adaptive high-$F_{ST}$ alleles seen in the SNP data is consistent with what is known about the population history of maize. There are three ways that adaptive alleles could be shared between highland populations: (a) by appearing in both locations as independent, de novo mutations; (b) by moving from one highland population to the other by migration; and (c) through convergent selective forces acting on shared standing variation. Here, we provide rough estimates of these rates, and develop in the Appendix more detailed, complementary models that build on the work in Ralph and Coop (2014a) and Ralph and Coop (2014b).

We chose to implement a fairly detailed demographic model. This is because much of the population genetics theory we use relies on universality results that reduce demographic models to two parameters: the dispersal distance (mean parent-offspring distance), and the variance in offspring number. However, these universality results do not hold if either distribution (dispersal or offspring) is sufficiently long-tailed; the detailed model allows us to both get a good idea of what part of parameter space we should focus on, and to verify that the approximation results we use are robust.

To assess the likely importance of (a) and (b), we first evaluate the rate at which we expect an allele that provides a selective advantage at higher elevation to arise by new mutation in
or near a highland region \((\lambda_{ma})\), and then use coalescent theory to show that even a highland-adapted allele that was neutral in the lowlands is unlikely to have had time to spread between highland populations under neutral gene flow. It may be more likely that 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); in the Appendix we find the rate at which such an allele already present in the Mesoamerican highlands would transmit the intervening lowlands and fix in the Andean highlands. The resulting values depend most strongly on the population density, the selection coefficient, and the rate at which seed is transported long distances and replanted. While long-distance dispersal is certainly possible, evidence from traditional seed systems in Mexico suggests even today it is rare: when farmers exchange seed (a minority of the time) \(\sim 90\%\) of seed lots come from \(< 10\) km away and from a site with altitudinal difference of \(< 50\) m, although farmers in highland locales exchange seeds over a greater range than average (Bellon et al. 2011). We checked the results by evaluating several choices of these parameters as well as with simulations, 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 \(p_e = 0.068\)), or 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 (the fitness of homozygotes is assumed to not affect the probability of establishment). Migration is mediated by seed exchange when fields are replanted from new stock, 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). The mean numbers of each category of offspring (seed/pollen, migrant/nonmigrant) 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 a mixture of Poissons. These means, and more details of the computations, are found in the Appendix. At the parameter values given, the dispersal distance (mean distance between parent and offspring) is \(\sigma = 3.5\) km, and the haploid variance in number of offspring \((\xi^2\), the variance in number of inherited copies of a chosen parental allele) is between 20 (for wild-type) and 30 (for \(s_b = 0.1\)). (Note that in a panmictic population, the offspring variance is approximately the ratio of census size to effective population size, \(\xi^2 \approx N/N_e\)).

**New mutations** The rate at which new mutations appear and fix in a highland population, which we denote \(\lambda_{mut}\), is approximately equal to the total population size of the highlands multiplied by the mutation rate per generation and 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 haploid variance in offspring number. This can be shown by expanding the generating function near 1, as in Fisher (1922) and Jagers (1975); see Lambert (2006) for more sophisticated models.

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 \(\rho\), the selective benefit \(s_b\) it provides, the mutation rate \(\mu\), and the variance in offspring number \(\xi^2\). In terms of these parameters, the rate of appearance is

\[
\lambda_{mut} = \frac{2\mu\rho A s_b}{\xi^2}.
\]  

**Geographic distribution** Throughout we work with populations distributed continuously across geography, with two regions of high elevation, the Mesoamerican and Andean highlands, separated by about 4,000 km. The value \(A\) in equation (3) is the total cultivated area in which the (highland-adapted) alleles in question are beneficial; for estimation of \(A\) in South America we overlaid raster layers of altitude (www.worldclim.org) and extent of maize cultivation (www.earthstat.org) and calculated the total area of maize cultivated above 1700 m using functions in the raster package for R (Hijmans and van Etten 2014).

Of course, the selective benefit of highland alleles is not discrete, but likely changes continuously with altitude, and it may be that the adaptive mutation occurs in a lowland area, subsequently migrating into the highlands. The calculation above does not account for these points, but the approximation is quite good, as verified by exact numerical calculation of the chance of fixation of a mutation as a function of the location where it first appears (see Figure A1); for theoretical treatment see Barton (1987).
Migration  It is harder to intuit a corresponding expression for the chance that an allele established by selection in one highland population moves to the other.

For maize in the Andean highlands to have inherited a highland-adapted allele from the Mesoamerican highlands, those Andean plants must be directly descended from highland Mesoamerican plants that lived more recently than the appearance of the adaptive allele. In other words, the ancestral lines along which the modern Andean plants have inherited the adaptive allele must trace back to the Mesoamerican 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 lineages; but this turns out to not affect the calculation much. Assuming demographic stationarity, the motion of each lineage can be modeled as a random walk, whose displacement after $m$ generations has variance $m \sigma^2$, and for large $m$ is approximately Gaussian. If we assume that lineages move independently, and $Z_n$ is the distance to the furthest of $n$ lineages, then $Z_n \leq \sqrt{m \sigma^2} (\sqrt{2 \log n} + \sqrt{2/\log n})$ with very high probability (Berman 1964).

Since this depends only on the logarithm of $n$, the number of lineages, the practical upshot of this is that the most distant lineage is very unlikely to be more than about 6 times more distant than the typical lineage, even among $10^7$ lineages. Lineages are not independent, but this only makes this calculation conservative.

Results

Samples and data

We sampled 94 maize landraces from four distinct regions in the Americas (Table S1; Figure 1): the lowlands of Mesoamerica (Mexico/Guatemala; $n = 24$) and northern S. America ($n = 23$) and the highlands of Mesoamerica ($n = 24$) 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 at least 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 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 δ0δ1 (Gutenkunst et al. 2009). All models incorporate a domestication bottleneck 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). We used published estimates of the strength of the domestication bottleneck (Eyre-Walker et al. 1998; Tenaillon et al. 2004; Wright et al. 2004).

---

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

<table>
<thead>
<tr>
<th></th>
<th>Mesoamerica Lowlands</th>
<th>Mesoamerica Highlands</th>
<th>S. America Lowlands</th>
<th>S. America Highlands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesoamerica Lowlands</td>
<td>–</td>
<td>0.0244</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Mesoamerica Highlands</td>
<td>0.0227</td>
<td>0.0343</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>S. America Lowlands</td>
<td>0.0466</td>
<td>0.0534</td>
<td>0.0442</td>
<td>–</td>
</tr>
<tr>
<td>S. America Highlands</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### Table 2 Estimated parameters of population size model

<table>
<thead>
<tr>
<th>Mesoamerica</th>
<th>Model IA</th>
<th>Model IB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>–5592.80</td>
<td>–4654.79</td>
</tr>
<tr>
<td>$N_C$</td>
<td>138,000</td>
<td>225,000</td>
</tr>
<tr>
<td>$N_1$</td>
<td>52,440</td>
<td>171,000</td>
</tr>
<tr>
<td>$N_2$</td>
<td>85,560</td>
<td>54,000</td>
</tr>
<tr>
<td>$N_{2P}$</td>
<td>85,560</td>
<td>54,000</td>
</tr>
<tr>
<td>S. America</td>
<td>Model IA</td>
<td>Model II</td>
</tr>
<tr>
<td>Likelihood</td>
<td>–3855.28</td>
<td>–8044.71</td>
</tr>
<tr>
<td>$N_C$</td>
<td>78,000</td>
<td>150,000</td>
</tr>
<tr>
<td>$N_1$</td>
<td>75,660</td>
<td>96,000</td>
</tr>
<tr>
<td>$N_2$</td>
<td>2,340</td>
<td>54,000</td>
</tr>
<tr>
<td>$N_{2P}$</td>
<td>205,920</td>
<td>51,300</td>
</tr>
<tr>
<td>$N_4$</td>
<td>2,700</td>
<td>145,800</td>
</tr>
</tbody>
</table>

### Migration

Migration It is harder to intuit a corresponding expression for the chance that an allele established by selection in one highland population moves to the other.

For maize in the Andean highlands to have inherited a highland-adapted allele from the Mesoamerican highlands, those Andean plants must be directly descended from highland Mesoamerican 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 the adaptive allele must trace back to the Mesoamerican 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 lineages; but this turns out to not affect the calculation much. Assuming demographic stationarity, the motion of each lineage can be modeled as a random walk, whose displacement after $m$ generations has variance $m \sigma^2$, and for large $m$ is approximately Gaussian. If we assume that lineages move independently, and $Z_n$ is the distance to the furthest of $n$ lineages, then $Z_n \leq \sqrt{m \sigma^2} (\sqrt{2 \log n} + \sqrt{2/\log n})$ with very high probability (Berman 1964).

Since this depends only on the logarithm of $n$, the number of lineages, the practical upshot of this is that the most distant lineage is very unlikely to be more than about 6 times more distant than the typical lineage, even among $10^7$ lineages. Lineages are not independent, but this only makes this calculation conservative.
Figure 3 Observed and expected joint distributions of minor allele frequencies in lowland and highland populations in (A) Mesoamerica and (B) S. America. Residuals are calculated as $(model - data)/\sqrt{model}$.

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 Mesoamerica (Model IB; $P_M$ on x-axis) and S. America (Model II; $P_S$ on y-axis). Red, blue, orange and gray dots represent SNPs showing significance in both Mesoamerica and S. America, only in Mesoamerica, only in S. America, or in neither region, respectively. The number of SNPs in each category is shown in the same color as the points.

Patterns of adaptation

Given the historical spread of maize from an origin in the lowlands, it is tempting to assume that the observation of signif-
icant 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 Mesoamerica and S. America as well as to *parviglumis* (see Methods). Consistent with predictions, we infer that differentiation at 72.3% (264) and 76.7% (230) of SNPs in Mesoamerica and S. America is due to adaptation in the highlands after excluding SNPs with ambiguous patterns likely due to recombination (Table S2).

As further evidence of selection, we asked whether alleles showing excess differentiation also exhibit longer haplotypes than expected. We calculated the empirical quantile of the pairwise haplotype score from Toomajian et al. (2006) for each putatively adaptive SNP as the proportion of all SNPs at a similar frequency with PHS scores greater than or equal to the PHS score observed at the focal SNP (Table S2). If $F_{ST}$ outliers have indeed been targeted by selection in a particular population, we expect this empirical quantile to be smaller (i.e., fewer random SNPs of similar frequency have as large a PHS score) than in other populations. Indeed, we find that SNPs identified as putatively adaptive in each of the four populations show smaller empirical PHS quantiles more often than the 50% expected by chance (Table S2).

Convergent evolution at the nucleotide level should be reflected in an excess of SNPs showing significant differentiation between lowland and highland populations in both Mesoamerica and S. America. Although the 19 SNPs showing $F_{ST}$-values < 0.01 in both Mesoamerica ($P_M$) and S. America ($P_S$) is statistically greater than the ≈ 5 expected ($48,370 \times 0.01 \times 0.01 \approx 4.8$; $\chi^2$-test, $P < 0.001$), it nonetheless represents a small fraction (≈ 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 simple outlier approach based using the 1% highest $F_{ST}$ values finds no shared adaptive SNPs between Mesoamerican 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 (see Methods). 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, excluding SNPs in an miRNA or second transcript. We classified SNPs showing significant $F_{ST}$ in Mesoamerica, S. America or in both regions into 778 genes. Of these, 485 and 277 genes showed Mesoamerica-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 Mesoamerica-specific and SA-specific significant SNPs. Overall, however, fewer genes showed evidence of convergent evolution than expected by chance (permutation test; $P < 10^{-5}$).

Finally, we tested whether genes showing evidence of selection in both highland populations were enriched for particular metabolic pathways using data on 481 metabolic pathways from the MaizeCyc database (ver. 2.2; Monaco et al. 2013). We found 92 pathways that include a selected gene from only one of the highland populations, but no significant excess of shared pathways: only 32 pathways included a selected gene in both populations ($P = 0.0961$; Table S3). Despite similar phenotypes and environments, we thus see little evidence for convergent evolution at the SNP, gene, and metabolic-pathway levels.

**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 2014a) to assess the degree of convergence expected under a spatially explicit population genetic model (see Materials and Methods). Using current estimates of maize cultivation in S. America, we find a 270,200 km$^2$ area in which maize is cultivated in ≥ 1% of the land area, for a total area of cultivation of ≈ 600,000 ha. At a planting density of $\rho \approx 20,000$ plants per hectare, this gives a total maize population of ≈ 12 billion. Assuming an offspring variance of $\xi^2 = 30$, we can then compute the waiting time $T_{mut} = 1/\lambda_{mut}$ for a new beneficial mutation to appear and fix. If we assume an average selection coefficient of $s_b = 10^{-5}$ for each mutation, a single-base mutation with mutation rate $\mu = 3 \times 10^{-8}$ (Clark et al. 2005) would take an expected 4,162 generations to appear and fix. Our estimate of the maize population size uses the land area currently under cultivation and is likely an overestimate; $T_{mut}$ scales linearly with the population size and lower estimates of $A$ will thus increase $T_{mut}$ proportionally. However, because $T_{mut}$ also 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 were to have an equivalent selective benefit of $s_b = 10^{-4}$, $T_{mut}$ would be reduced to 42 generations assuming constant $A$ over time.

Gene flow between highland regions could also generate patterns of shared adaptive SNPs. The coalescent calculations described above suggest that highland area today is unlikely to draw any ancestry from a region more than 6σ$\sqrt{m}$/km kilometers away from m generations ago in any part of the genome that is neutral in the lowlands. Our estimated dispersal of $\sigma = 3.5$km thus provides an estimate of 1,328km. The Mesoamerican and Andean highlands are approximately 4,000 km apart, and neutral alleles are therefore not expected to transit between the Mesoamerican and Andean highlands within 4,000 generations. Changing the typical distance over which farmers share seed by a factor of 10 would change this conclusion, but data
from field surveys do not lend support to such high dispersal distances (Bellon et al. 2011).

These results for neutral alleles put a lower bound on the time for deleterious alleles to transit as well, suggesting that we should not expect even weakly deleterious alleles (e.g., $s_m = 10^{-5}$) to have moved between highlands. We expect many of the alleles adaptive in the highlands to be deleterious in the lowlands, and analyze this case in more detail in the Appendix.

Taken together, these theoretical considerations suggest that any alleles beneficial in the highlands that are neutral or deleterious in the lowlands and shared by both the Mesoamerican and S. American highlands would have been present as standing variation in both populations, rather than passed between them.

**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 Mesoamerica and S. America tended to segregate in both the lowland population (85.3% vs. 74.8% in Mesoamerica (Fisher’s exact test $P < 10^{-8}$ and 94.8% vs 87.4% in S. America, $P < 10^{-4}$) and parviglumis (78.3% vs. 72.2% in Mesoamerica (Fisher’s exact test $P < 0.01$ and 80.2% vs 72.8% in S. America, $P < 0.01$) more often than other SNPs of similar mean allele frequency.

While maize in highland Mesoamerica grows in sympatry with the highland teosinte *Zea mexicana*, maize in S. America is outside the range of wild *Zea* species, leading to a marked difference in the potential for adaptive introgression from wild relatives. Pyhähärvi 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 (2–7%) of our putatively adaptive SNPs were identified by Pyhähärvi et al. (2013), though these numbers are still in excess of expectations (Fisher’s exact test $P < 10^{-3}$ for S. America and $P < 10^{-8}$ for Mesoamerica; Table S4). The proportion of putatively adaptive SNPs shared with teosinte was twice as high in Mesoamerica, however, leading us to evaluate the contribution of introgression from *mexicana* (Hufford et al. 2013) in patterning differences between S. American and Mesoamerican highlands.

The proportion of putatively adaptive SNPs in introgressed regions of the genome in highland maize in Mesoamerica 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 S5). 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 Mesoamerica as in S. America (23 compared to 9, one-tailed Z-test $P = 0.0027$). Excluding putatively adaptive SNPs, mean $F_{ST}$ between Mesoamerica 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 Mesoamerica is not simply due to neutral introgression of a divergent teosinte haplotype.

**Discussion**

Our analysis of diversity and population structure in maize landraces from Mesoamerica and S. America points to 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 evidence of a strong bottleneck followed by expansion in the highlands of S. America. We identified SNPs deviating from patterns of allele frequencies determined by our demographic model as loci putatively under selection for highland adaptation.

Though the rapid decay of linkage disequilibrium in maize (Figure S6) makes it likely we have identified only a subset of selected loci (Tiffin and Ross-Ibarra 2014), several lines of evidence suggest our results are likely representative of genome-wide patterns. SNPs identified as $F_{ST}$ outliers by our method show evidence of longer haplotypes and patterns of among-population allele frequency consistent with adaptation (Table S2). Consistent with previous work suggesting adaptive introgression from teosinte, the Mesoamerican highland population shares a larger proportion of SNPs identified as adaptive in teosinte (Pyhähärvi et al. 2013). We also see more $F_{ST}$ outliers Mesoamerica in regions introgressed from teosinte and which overlap with QTL for differences between *parviglumis* and *mexicana* (Lauter et al. 2004; Hufford et al. 2013). Finally, though our SNP data are enriched in low-copy genomic regions, our results are consistent with both GWAS in maize (Wallace et al. 2014) and local adaptation in teosinte (Pyhähärvi et al. 2013) in finding an excess of putatively adaptive SNPs in intergenic regions of the genome.

Although our data identify hundreds of loci that may have been targeted by natural selection in Mesoamerica and S. America, fewer than 1.8% of SNPs and 2.1% of genes show evidence for convergent evolution between the two highland populations. This relative lack of convergent evolution is concordant with recently developed theory (Ralph and Coop 2014a), which applied to this system suggests that convergent evolution involving identical nucleotide changes is unlikely to have occurred in the time since highland colonization through either recurrent mutation or migration across Central America via seed sharing. These results are generally robust to variation in most of the parameters, but are sensitive to gross misestimation of some of the parameters – for example if seed sharing
was common over distances of hundreds of kilometers. The modeling highlights 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 determined by only a few loci (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 90% of the putatively adaptive variants in Mesoamerica 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 $\theta$ (product of the effective population size, mutation rate and target size) is greater than 1, as long as the scaled selection coefficient $Ns$ (product of the effective population size and selection coefficient) is reasonably large (Herminsson and Pennings 2005). Estimates of $\theta$ from synonymous nucleotide diversity in maize (Tenaillon et al. 2004; Wright et al. 2005; Ross-Ibarra et al. 2009), suggest that 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 tillers*1 (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.

Both our empirical and theoretical results suggest that adaptation to high elevation probably occurred through some combination of selection on standing variation and independent *de novo* mutation at highly quantitative traits. Because cultivated maize has retained high levels of diversity, much of the ancestral variation present in the populations that founded each of the two highlands was likely shared, allowing for the possibility of shared signals due to selection on the same ancestral variants. However, initial frequencies of alleles present as standing variation will be highly stochastic, leading to a significant role of chance in which alleles are selected, as well as the strength of the signal of $F_{ST}$. This is particularly true for alleles likely to be adaptive in the highlands and thus weakly deleterious in lowland populations, as these should be rare in individual populations. Epistasis could make it even less likely that the same allele is shared between regions.

Overall, our results highlight the complexity of studying convergent evolution for quantitative traits in highly diverse species. Our future efforts will take advantage of reciprocal transplant experiments to identify specific phenotypes under selection. We are also developing mapping populations in both Mesoamerica and South America that should allow identification of genomic regions underlying phenotypes of interest and estimation of the proportion of adaptive variation shared between populations.

**Acknowledgements**

We appreciate the helpful comments of P. Morrell and members of the Ross-Ibarra lab and Coop lab. This project was supported by Agriculture and Food Research Initiative Competitive Grant 2009-01864 from the USDA National Institute of Food and Agriculture as well as funding from National Science Foundation grants IOS-1238014 (to JRI) and DBI-1262645 (to PLR).

**Literature Cited**


**Appendix**

**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 $\text{Var}[X] = \xi^2$, and $\mathbb{P}\{X = 0\} > 0$ (some chance of leaving no offspring), then the probability of local nonextinction $p_s$ is approximately $p_s \approx 2s/\xi^2$ 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 $p_s$ is the minimal solution to $\Phi(1 - u) = 1 - u$. (This can be seen because: $1 - p_s$ 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 $(1 - p_s)^x$; and $\Phi(1 - p_s) = \mathbb{E}[(1 - p_s)^X]$.)

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 $\Phi_i(u) = \mathbb{E}[\prod_j u_j^{X_j}]$, where $X_{ij}$ is the (random) number of offspring that an individual at $i$ produces at location $j$. Then $p_s = (p_{s1}, \ldots, p_{sn})$, the vector of probabilities that a new mutation at each location eventually fixes, is the minimal solution to $\Phi_i(1 - p_s) = 1 - p_s$, i.e. $\Phi_i(1 - p_s) = 1 - p_{si}$.

Here we consider a linear habitat, so that the selection coefficient at location $i$ is $s_i = \min(s_b, \max(-s_d, \alpha \ell_i))$. There does not seem to be a nice analytic expression for $p_s$ in this case, but since $1 - p_s$ is a fixed point of $\Phi$, the solution can be found by iteration: $1 - p_s = \lim_{n \to \infty} \Phi^n(u)$ 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 $p_r$), from partially new stock (a proportion $r_m$ with probability $p_m$), or entirely from the same field. Plants have an average of $\mu_E$ ears per plant, and ears have an average of $N/N_f$ kernels; so a plant has on average $\mu_E N^2/N_f$ kernels, and a field has on average $\mu_E N$ ears and $\mu_E N^2/N_f$ kernels. 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 $\mu_E$ 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 $p_c$ 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 being chosen. Furthermore, the field is used to completely (or partially) replant another’s field with chance $p_r/(1 - p_c)$ (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 $\lambda_P$, and the number of seed children is a mixture of $K$ independent Poissons with mean $(1 + s)N_f/N$ each, where $K$ is the random number of ears chosen to replant, which is itself Poisson with mean $\mu_K$. 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 $p_c$, $\lambda_{PM} = m_g(1 + s)$ and $\mu_K = \mu_{PL} = 0$. Otherwise, with probability $(1 - p_c)(1 - p_m)$, $\mu_{KL} = N_f/N$ and $\lambda_{PL} = (1 + s)(1 - m_g)$; and with probability $(1 - p_c)p_m$, $\mu_{KL} = (1 - r_m)N_f/N$ and $\lambda_{PL} = (1 - r_m)(1 + s)(1 - m_g)$. The migrant means are, with probability $(1 - p_c)p_m$, $\mu_{KM} = r_m N_f/N$ and $\lambda_{PM} = (1 + s)(r_m(1 - m_g) + m_g)$, and otherwise $\mu_{KM} = 0$ and $\lambda_{PM} = m_g(1 + s)$. 
The generating function

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

$$
\Phi(u) = p_e \phi(u; m_g(1+s)) + \{(1 - p_e)(1 - p_m)\phi(u; (1 + s)(1 - m_g))\phi(\phi(u; (1 + s)N/N_f); N_f/N) + (1 - p_e)p_m\phi(u; (1 + s)(1 - r_m)(1 - m_g))\phi(\phi(u; (1 + s)N/N_f); (1 - r_m)N_f/N)\}
$$

$$
\times \{(p_e/(1 - p_e))\phi(u; (1 + s))\phi(\phi(u; (1 + s)N_f/N); N_f/N) + p_m\phi(u; (1 + s)(1 - r_m)(1 - m_g) + m_g))
\times \phi(\phi(u; (1 + s)N_f/N); r_mN_f/N) + (1 - p_e)/(1 - p_e) - p_m)\phi(u; m_g(1+s))\}
$$

$$
= \phi(u; m_g(1+s))(p_e + \{(1 - p_e)(1 - p_m)\phi(u; (1 + s)(1 - m_g))\phi(\phi(u; (1 + s)N/N_f); N_f/N)
+ (1 - p_e)p_m\phi(u; (1 + s)(1 - r_m)(1 - m_g))\phi(\phi(u; (1 + s)N_f/N); (1 - r_m)N_f/N)\}
\times \{p_e/(1 - p_e)\phi(u; (1 + s)(1 - m_g))\phi(\phi(u; (1 + s)N_f/N); N_f/N) + p_m\phi(u; (1 + s)r_m(1 - m_g))
\times \phi(\phi(u; (1 + s)N_f/N); r_mN_f/N) + (1 - p_e)/(1 - p_e) - p_m)\})
$$

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

$$
(1 + s)(m_g + (1 - p_e)\{(1 - p_m)((1 - m_g) + 1) + p_m((1 - r_m)(1 - m_g) + (1 - r_m))\}
+ \{(1 - p_e)(1 - m_g) + 1 + p_m(1 - p_e)(r_m(1 - m_g) + r_m)\})
$$

$$
= (1 + s)(m_g + (1 - p_e)(2 - m_g)(1 - p_m r_m) + p_e(2 - m_g) + p_m r_m(1 - p_e)(2 - m_g))
$$

$$
= (1 + s)(m_g + (2 - m_g)((1 - p_e)(1 - p_m r_m) + p_e + p_m r_m(1 - p_e)))
$$

$$
= (1 + s)(m_g + (2 - m_g))
$$

$$
= 2(1 + s).
$$

### Table A1 Parameter estimates used in calculations, and other notation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete seed stock replacement prob</td>
<td>$p_e$ 0.068</td>
</tr>
<tr>
<td>pollen migration rate</td>
<td>$m_g$ 0.0083</td>
</tr>
<tr>
<td>number of plants per field</td>
<td>$N$ $10^5$</td>
</tr>
<tr>
<td>number of ears used to replant</td>
<td>$N_f$ 561</td>
</tr>
<tr>
<td>mean ears per plant</td>
<td>$\mu_E$ 3</td>
</tr>
<tr>
<td>partial stock replacement prob</td>
<td>$p_m$ 0.02</td>
</tr>
<tr>
<td>mean proportion stock replaced</td>
<td>$r_m$ 0.2</td>
</tr>
<tr>
<td>pollen migration distance</td>
<td>$\sigma_p$ 0 km</td>
</tr>
<tr>
<td>seed replacement distance</td>
<td>$\sigma_s$ 50 km</td>
</tr>
<tr>
<td>distance between demes</td>
<td>$\alpha$ 15 km</td>
</tr>
<tr>
<td>width of altitudinal cline</td>
<td>$w$ 62 km</td>
</tr>
<tr>
<td>deleterious selection coefficient</td>
<td>$s_d$ varies</td>
</tr>
<tr>
<td>beneficial selection coefficient</td>
<td>$s_b$ varies</td>
</tr>
<tr>
<td>slope of selection gradient</td>
<td>$\alpha$ $(s_d + s_b)/w$</td>
</tr>
<tr>
<td>variance in offspring number</td>
<td>$\xi^2$ varies</td>
</tr>
<tr>
<td>maize population density</td>
<td>$\rho$ $5 \times 10^3$</td>
</tr>
<tr>
<td>area of highland habitat</td>
<td>$A$ 270,000 km$^2$</td>
</tr>
<tr>
<td>mean dispersal distance</td>
<td>$\sigma$ 1.8 km</td>
</tr>
</tbody>
</table>
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 the exact expression does not aid the intuition.

**Migration and spatial structure**

To incorporate spatial structure, suppose that the locations $\ell_k$ 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 $\ell_i$ is Poisson($\lambda_i$), 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

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

(A8)

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

(A9)

We can then substitute this expression into equation (A1), 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 $\sigma_s$ 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).

**Dispersal distance**

The dispersal distance – the mean distance between parent and offspring – is equal to the chance of inter-village movement multiplied by the mean distance moved. This is

$$
\sigma = (p_e + (1 - p_e)p_m r_m)\sigma_s = 3.5864 \text{ km}
$$

(A10)

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 A1. Note that the approximation $2s$ divided by the variance in offspring number is quite close.

In the main text, we used a rough upper bound on the rate of migration that ignored correlations in migrants. As we show in Ralph and Coop (2014a), the rate of adaptation by diffusive migration is more precisely

$$
\lambda_{\text{mig}} = \frac{1}{2} \rho s_m \min(s_m, 2sb/\ell^2) \exp \left( -\frac{\sqrt{2s_m} R}{\sigma} \right).
$$

First note that for $10^{-1} < s_m < 10^{-4}$, 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 $\sigma \approx 3.5$ km, and far below the mean distance $\sigma$ to the field that a farmer replants seed from, when this happens, which we have as $\sigma_s = 50$ km. Taking $\sigma = 3.5$ km, we have that $7 \leq \sigma/\sqrt{2s_m} \leq 250$ km. A very conservative upper bound might be $\sigma \leq \sigma_s/10$ (if farmers replaced 10% of their seed with long-distance seed every year). At this upper bound, we would have $10 \leq \sigma/\sqrt{2s_m} \leq 350$ km, which is not very different. This makes the exponential term small since $R$ is on the order of thousands of kilometers.

Taking $\sigma = 3.5$ 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 $\lambda_{\text{mig}} \leq 10^{-8}$, 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.

**Migration rate of deleterious alleles**

In the main text we computed $\lambda_{\text{mig}}$, the rate at which new adaptive alleles appeared by mutation. A corresponding expression for the chance that an allele moves from one highland population to another is harder to intuit. This problem is studied in more depth in Ralph and Coop (2014a), under the assumption that the alleles are deleterious between the highlands. Since such
deleterious alleles are much less likely to transit than neutral ones, the analysis in the main text implies that gene flow is unlikely to have shared these alleles between highland regions. However, because spatially continuous models assuming selective effects are better understood than neutral ones, and we do expect a tradeoff between highland- and lowland-adaptation, it is useful to understand what happens in this case as well.

If an allele is beneficial at high elevation and fixed in the Mesoamerican highlands but is deleterious at low elevations, then at equilibrium it will be present at low frequency at migration-selection balance in nearby lowland populations (Haldane 1948; Slatkin 1973). 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, $\sigma$ 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 we assume the allele is deleterious in the lowlands, if $R$ is large there are likely none present; but if there are, the expected number is of order $1/s_m$ (Geiger 1999; Ralph and Coop 2014a). This therefore puts an upper bound on the rate of migration of

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

and 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 ($\approx 2s_b/\xi^2$) (discussed above) and the time to reach migration-selection balance (discussed in the main text); both of these omissions mean we underestimate $T_{\text{mig}}$.

**Results for gene flow of deleterious alleles:** From our demographic model we have estimated a mean dispersal distance of $\sigma \approx 3.5$ 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 7 and 250 kilometers. Since the Mesoamerican 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 $T_{\text{mig}} \approx 8 \times 10^4$ generations. While the exponential dependence on distance in equation (A11) means that shorter distances could be transited more quickly, the waiting time $T_{\text{mig}}$ is also strongly dependent on the magnitude of the deleterious selection coefficient: with

---

**FIGURE A1** 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.
$s_m = 10^{-4}$, $T_{	ext{mig}} \approx 25$ generations over a distance of 2,000 km, but increases to $\approx 10^8$ generations with a still weak selective cost of $s_m = 10^{-3}$.