The molecular basis of parallel adaptation to highland climate in domesticated maize

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

[*]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,
[**]US Department of Agriculture – Agriculture Research Service (USDA-ARS) address,
[††]Department of Ecology, Evolution, and Organ Biology, Iowa State University, Ames, Iowa 50011, USA,
[‡‡]The Center for Population Biology and the Genome Center, University of California, Davis, California 95616 , USA

Revised manuscript for *Genetics*, February 7, 2014

ABSTRACT Parallel adaptation is defined as the independent evolution of multiple species/subpopulations to similar environments via adaptive mutations in the same locus. We investigate here the molecular basis of maize adaptation 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 parallel adaptation. In spite of similar morphologies, we see limited evidence of selection on quantitative traits, and, consistent with predictions across a wide array of parameter space, we see few SNPs showing signs of parallel adaptation. 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. We discuss the significance of these results in the context of the molecular basis of adaptation to new environments.

Introduction

Parallel adaptation is a process in which multiple species or populations independently adapt to distinct regions with similar environments via mutations at the same loci (**???**). Evolutionary genetic analysis of a wide range of species has provided evidence for multiple pathways to parallel adaptation. 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 (**??**). Parallel adaptation can also be achieved when different mutations arise within the same locus and 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 (**?**). Finally, parallel adaptation 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. In most cases, adaptation in these populations took advantage of standing variation at the *Eda* locus in marine populations (**?**).

We still know relatively little, however, about how common it is for parallel phenotypic evolution to be driven by parallel genetic changes or the relative frequencies of these different routes of parallel adaptation. Domesticated maize (*Zea mays* ssp. *mays*) provides an excellent opportunity to investigate the molecular basis of parallel adaptation. 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) (**???**). After domestication, maize spread rapidly across the Americas, reaching the lowlands of South America and the high altitudes of the Mexican Central

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

Plateau by ∼6,000 BP (**?**), and the Andean highlands ∼2,000 years later (**??**). 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 commonly, if not universally, accompany highland adaptation including reduced temperature, increased ultraviolet radiation, and reduced partial pressure of atmospheric gases (**?**).

Common garden experiments in Mexico reveal that highland maize has successfully adapted to highland conditions (**?**), and phenotypic comparisons between Mexican and South American populations are suggestive of parallel adaptation. Landraces from both populations share a number of phenotypes not found in lowland populations, including dense macrohairs (**??**), stem pigmentation (**??**), and biochemical response to UV radiation (**?**). Genetic analyses of maize landraces from across the Americas indicate that the two highland populations are independently derived from their respective lowland populations (**??**), so observed patterns of phenotypic similarity are not simply due to recent shared ancestry.

Although there are no wild relatives of maize in South America, the teosinte *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 (**??**). Phenotypic differences between *mexicana* and *parviglumis* mirror those between highland and lowland maize (**?**) and population genetic analyses of the two subspecies reveal evidence of natural selection associated with altitudinal differences between *mexicana* and *parviglumis* (**?**). Landraces in the highlands of Mexico are often found in sympatry with *mexicana*, and gene flow between the two is thought to have contributed to maize adaptation to the highlands (**?**).

In this paper we set out to address a number of questions regarding highland adaptation in maize: What is the genetic architecture of highland adaptation? Do maize populations in South America show evidence of parallel adaptation when compared with highland maize from Mexico? How do observed patterns of parallel adaptation compare to theoretical expectations? We make use of SNP genotyping to characterize patterns of natural selection in highland maize and compare our results to expectations from theoretical models of parallel adaptation. We find evidence supporting unique demographic histories in the highlands of Mexico and South America, a predominance of independent versus parallel adaptation, and contributions of standing variation and introgression from wild relatives to highland adaptation in maize.

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 (see also Table S1). Landraces sampled from altitudes $<$1,700 m were considered lowland, while accessions from $>$1,700 m were considered highland (Table S1). 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 (**?**). The quality of DNA was ensured using methods described in **?**.

SNP data

We used the maize B73 genome sequence RefGen version 2 as a reference (**?**). The filtered gene set (release 5b.60) was retrieved from MaizeSequence.org for SNP annotations. We excluded genes annotated as transposable elements (84) and pseudogenes (323) from the filtered gene set, resulting in a total of 38,842 genes.

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 (**?**). SNPs were clustered with the default algorithm of the GenomeStudio Genotyping Module v1.0 (Illumina Inc., San Diego, CA, USA). Clustering for each SNP was then visually inspected and manually adjusted. These data are referred to as "MaizeSNP50" hereafter. MaizeSNP50 data have high reproducibility and a low proportion of missing data but are subject to ascertainment bias. This array contains SNPs discovered in five ascertainment schemes (**?**); however, the vast majority of SNPs come from two panels: the Syngenta set (14,810 SNPs), derived from polymorphisms distinguishing the parents of the IBM population (the maize lines B73 and Mo17), and the Panzea set, including 40,594 SNPs identified during sequencing of the 25 parents of the NAM population.

The second data set was generated utilizing high-throughput Illumina sequencing data in a method referred to as genotyping-by-sequencing (GBS). Kelly is up for details. Average coverage was relatively low (XXX), likely resulting in heterozygotes often being miscalled as homozygotes. However, this data set is relatively free from ascertainment bias. GBS data were obtained for a subset of 87 of the landrace accessions (Table S1).

To assess data quality, we compared genotypes at the 7,197 SNPs that overlap between the MaizeSNP50 and GBS data sets. Excluding missing data, we compared 229,937 genotypes. While only 0.8% of 173,670 homozygous loci in the MaizeSNP50 data set differed from GBS genotypes, 88.6% of 56,267 MaizeSNP50 heterozygotes had different genotypes in the GBS data, being homozygous in nearly all cases. Despite an extremely high heterozygote error rate, our GBS data should

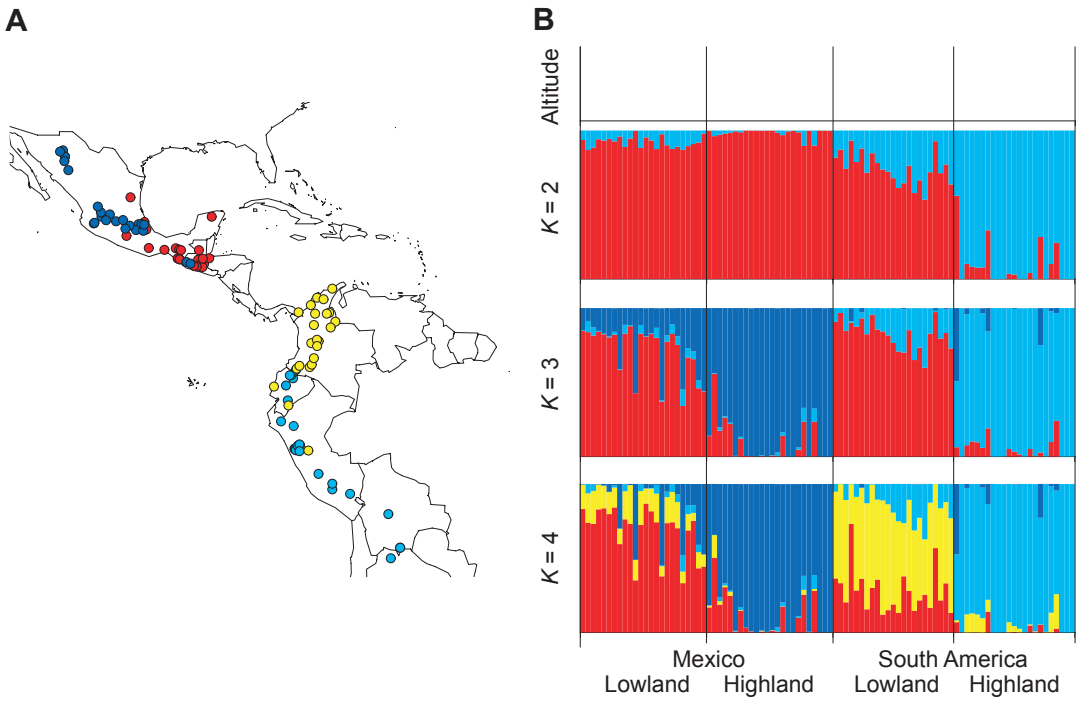


Figure 1 (A) Sampling locations of landraces. Red, blue, yellow and light blue dots represent Mexico lowland, Mexico highland, S. America lowland and S. America highland populations, respectively. (B) Results of STRUCTURE analysis of the maizeSNP50 SNPs with $K = 2 \sim 4$. The top panel shows the altitude,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

be informative given the high correlation in allele frequencies between data sets ($r = 0.89$; Figure SX change after finishing intro and results) and a lack of major allele or reference bias (data not shown).

Structure analysis

We performed a STRUCTURE analysis (**??**) using synonymous and noncoding SNPs from the MaizeSNP50 data. We assumed free recombination between SNPs without missing data and randomly pruned SNPs closer than 10 kb (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) based on a *G*-test was smaller than 0.5% using all individuals. 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 $K = 2 \sim 6$ populations, a burn-in length of 50,000 iterations and a run length of 100,000 iterations. Results across replicates were nearly identical.

Demographic inference

We tested three demographic models in which maize was differentiated into high- 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 silent SNPs were utilized that had ≥ 15 individuals without missing data in both low- 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 silent 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 $P < 0.05$. Demographic parameters were inferred using the software dadi (**?**), 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 Mexico and S. America 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

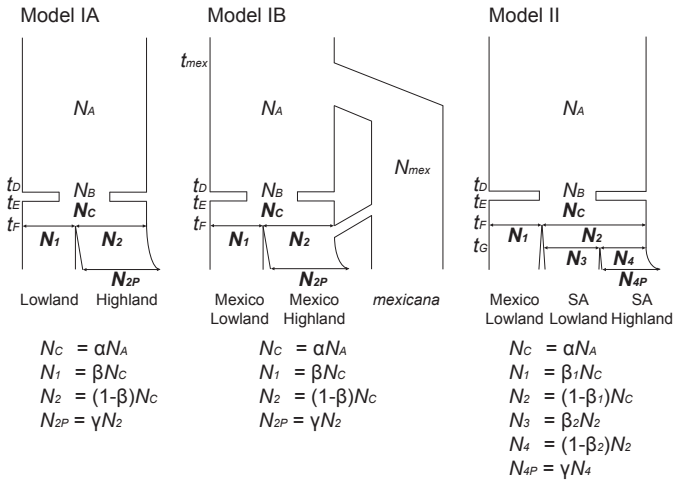


Figure 2 Demographic models of maize low- and highland populations. Parameters provided in bold were estimated in this study. See text for details.

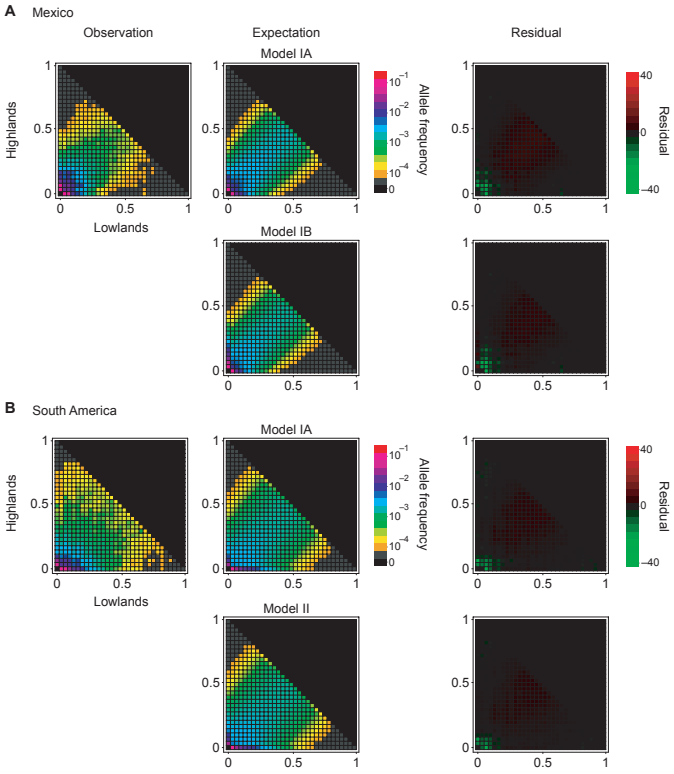


Figure 3 Observed and expected joint distributions of minor allele frequencies in low- and highland populations in (A) Mexico and (B) S. America.

population at t_F generations ago. The size of the low- and highland populations at time t_F is determined by a parameter, β such that the population is divided by βN_C and $(1-\beta)N_C$. 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 Mexico populations. We incorporate admixture from *mexicana* to the highland Mexican maize population. The time of differentiation between *parviglumis* and *mexicana* occurs at t_{mex} generations ago. The size of the *mexicana* population is denoted by N_{mex} and this size is assumed to be constant. At t_F generations ago, the Mexico highland population is derived from the Mexico lowland population and admixture with *mexicana*. The proportion of admixture with *mexicana* is denoted by P_{mex}.

***Model II*:** The final model is for 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, SA 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 determined using estimates of the composite parameter $4N_A\mu$ and a separate estimate of mutation rate, μ, per site per generation. $4N_A\mu$ was estimated in *parviglumis* to be $\sim$0.018 (**?????**). The mutation rate in maize has been estimated to be $2.9 \sim 3.3 \times 10^{-8}$, so we assume $\mu = 3 \times 10^{-8}$ (**?**). Thus, N_A was set to $0.018/4/(3 \times 10^{-8}) = 150,000$. The severity of the domestication bottleneck is represented by $k = N_B/t_B$ (**??**), and following **?** we assumed $k = 2.45$ and $t_B = 1,000$ generations. Taking into account archaeological evidence (**?**), 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 II (**?**), $t_F = 4,000$ for S. American populations in Model lI (**?**), and $t_{mex} = 60,000$, $N_{mex} = 160,000$ (**?**), and $P_{mex} = 0.2$ (**?**) 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 $t_G = 4,000$ (**??**) and estimated the remaining four parameters (α, β_1, β_2 and γ).

Differentiation between low- and highland populations

We used our inferred demographic model to generate a null distribution of F_{ST} via simulation using the software ms (**?**). The command line options for ms are provided in Table S4. Generating the null distribution of differentiation for the MaizeSNP50 data requires accounting for ascertainment bias. Evaluation of genetic clustering in our data (not shown) coin-

cides with previous work (**?**) in suggesting that the two lines most important in the ascertainment panel 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 low- and highland populations because the Mexican lowland population was not incorporated into this model. For each combination of sample sizes in low- and highland populations, we generated 10^7 F_{ST} values and used these as a null distribution in order to evaluate our data. 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 scoring test

We performed a pairwise haplotype scoring (PHS) test to detect further evidence of selection, following **?**. 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 (**?**). As a reference for phasing, we used data (excluding heterozygous SNPs) from an Americas-wide sample of 23 partially inbred landraces that were included in the Hapmap v2 data set (**?**). fastphase was run with default parameter settings. PHS was calculated for an allele A at position x by

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

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

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

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 an allele A with frequency p at position x was calculated such that Pr($PHS_{x_A} \leq PHS_{null|p}$), where $PHS_{null|p}$ are the PHS values for all alleles with frequency p across the genome. *(Do you mean that the P-value for allele x is the proportion of alleles of the same frequency genome-wide that have a larger PHS value? (and, should be PHS, not PHA?))* You are right. PHA is a typo. Corrected.

Genetic distances were obtained from the IBM population (**?**). While genetic positions were initially only available for the MaizeSNP50 data set, we fit genetic (cM) and physical (bp) distances to a tenth degree polynomial curve, and calculated cM for all SNPs.

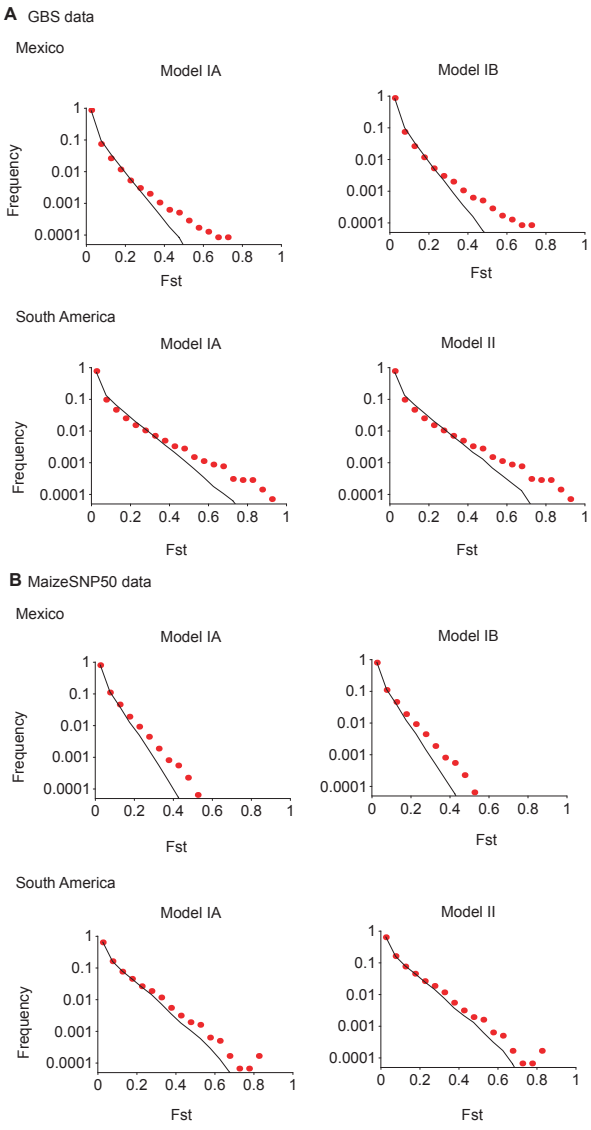


Figure 4 Observed and expected distributions of F_{ST} values in GBS (A) and MaizeSNP50 data (B). The *x*-axes represent F_{ST} values. The *y*-axes represent the frequency of SNPs with F_{ST} values within a bin of 0.05 size. Red dots and solid lines indicate observed and expected distributions.

Theoretical evaluation of parallel adaptation

We suggest below that many of the high-F_{ST} alleles are locally adaptive, and the degree of coincidence between highland regions informs us about whether these adaptations occurred in parallel, or if alleles were transmitted between the two by migration. To see if the abundance and degree of coincidence 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 altitude 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 tran-

sit the intervening lowlands and fix in the Andean highlands (λ_{mig}). In each case we assume alleles adapted in the highlands are slightly deleterious at lower altitude. This assumption is consistent with empirical findings in reciprocal transplants of highland and lowland maize in Mexico (**?**). These numbers depend most strongly on the population density, the selection coefficient, and the rate at which seed is transported long distances and replanted. We evaluated these rates using new and existing theory, and validated by simulation.

To calculate the rate at which new mutations appear and fix in a highland population, λ_{mut}, we multiplied the total population size of the highlands by the mutation rate per generation ======= In each case we assume alleles adapted in the highlands are slightly deleterious at lower altitude (but see later discussion). This assumption is consistent with empirical findings in reciprocal transplants of highland and lowland maize in Mexico (**?**). These numbers depend most strongly on the population density, the selection coefficient, and the rate at which seed is transported long distances and replanted. Here we describe the mathematical details; readers may skip to the results without loss of continuity.

To do this, we followed **?** in constructing a detailed demographic model for domesticated maize. *(this is long-ish? could perhaps summarize this more/better?)* 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 entirely from the same field otherwise. 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 (so 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 1.

(got a better term than "pollen-parent kernels"??)

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.

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 latter 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 (**?**). 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 **?** or **?**.

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

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

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 (**?**). If an allele is beneficial at high elevation, and fixed in the Mexican highlands, but deleterious at low elevations, then it will be present at low frequency in nearby lowland populations, maintained at migration-selection balance (**?**). 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 in that population. 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 there are some present. Since the latter is at least 1, the chance there are any present in a given generation is no more than $(N/2)\exp(-R\sqrt{2s_m}/\sigma)$, and so this puts an upper bound on λ_{mig}. Therefore, we would need to wait $T_{\text{mig}} = (2/N)\exp(R\sqrt{2s_m}/\sigma)$ generations for a rare such excursion

Table 1 Silent site F_{ST} from GBS SNPs

		Mexico		South America	
		Lowlands	Highlands	Lowlands	Highlands
Mexico	Lowlands	–			
	Highlands	0.0244	–		
SA	Lowlands	0.0227	0.0343	–	
	Highlands	0.0466	0.0534	0.0442	–

to occur. In other words, we can bound the rate of migration by

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

with N the total size of the unadapted highland population, and R the distance from the adapted to the yet-unadapted highland populations. This also 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.

To obtain specific predictions, we then computed λ_{mut} and λ_{mig} at various parameter values. We also checked these with simulations and more detailed computations, described in the Appendix.

Results and Discussion

Samples and data

for PLoS G We generated a large set of SNPs in 94 maize accessions that were sampled from the Americas (Table S1; Materials and Methods). These accessions were sampled from the four populations: 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 ($n = 23$). The SNPs were generated by the two high-throughput methods: genotyping-by-sequencing (GBS) method and MaizeSNP50 Beadchip platform (Materials and Methods). We hereafter refer to the two SNP data sets as "GBS" and "MaizeSNP50". We used a part of samples ($n = 87$) for generating GBS data, which is due to computational limitation?. In total, we obtained 91,779 SNPs after filtering by Hardy-Weinberg criteria with sample size ≥ 10 in all the four populations (see Materials and Methods), 67,828 and 23,951 of which were generated by GBS and MaizeSNP50, respectively. We performed population genomic analyses using these data sets to reveal the molecular basis of parallel highland adaptation in maize.

Our sample included 94 maize landraces from four distinct regions in the Americas: 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 (n=94) and a method referred to as genotyping-by-sequencing (GBS; n=87).

Population structure

We performed a STRUCTURE analysis (**??**) of our landrace sample, varying the number of groups from $K = 2 \sim 6$ (Figure 1, Figure S2). 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 (**?**), 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), and the higher F_{ST} values observed in S. America are consistent with decreased admixture seen in STRUCTURE. Archaeological evidence supports a more recent colonization of the highlands in S. America (**???**), suggesting that the observed differentiation may be the result of a stronger bottleneck during colonization of the S. American highlands.

Population differentiation under inferred demography

Modified for PLoS G 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 (Figure 2; see Materials and Methods for details) using the maximum likelihood method implemented in dadi (**?**). All models incorporate domestication bottleneck (**?**) and population differentiation between lowland and highland populations. Model IA is common between Mexican and S. America populations. Model IB is specific to Mexican populations, into which we incorporated the admixture from *mexicana* to highlands. Model II is specific to S. America, which includes S. American populations and Mexican lowlands. We introduced the last model to simulate SNPs with ascertainment bias because the discovery panel would be originated from Mexican lowlands (see Materials and Methods for details).

Estimated parameter values are listed in Table **??**; 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 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 **??**), consistent with analyses suggesting that introgression from *mexicana* played a significant role during the spread of maize into the Mexican highlands (**?**).

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

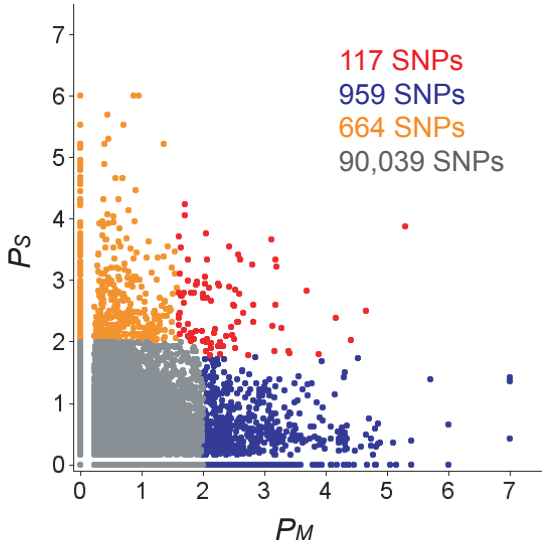


Figure 5 Scatter plot of $-\log_1 0P$-values of observed F_{ST} values, based on simulation from *(which model)* , in Mexico (P_M on x-axes) and South America (P_S on y-axes). Red, blue, orange and gray dots represents SNPs showing significance in both Mexico and South America, only in Mexico, only in South America, respectively (see text for details). The number of SNPs in each category is shown in the same color of dots.

of $N_B < N_C$ was much smaller than for alternative parameterizations (Table **??**, Table S2)

This result appears to contradict earlier work using sequences from coding regions to infer a maize domestication bottleneck (**???**). One explanation for this discrepancy may be the action of purifying selection in coding regions, which could act to retard the recovery of diversity and lead to estimates of a stronger bottleneck **?**. Consistent with **?**, our genome-wide SNP data show an excess of rare variants relative to expectations under **?**'s bottleneck model (Figure 3), suggesting a domestication model involving a weaker bottleneck or more rapid population growth.

Comparing our empirical F_{ST} values to the null expectation simulated under our demographic models allowed us to identify significantly differentiated SNPs between low- and highland populations. In all cases, observed F_{ST} values were quite similar to those generated under our null models (Figure 4A), and model choice – including the parameterization of the domestication bottleneck – had little impact on the distribution of estimated p-values (Figure S4). We chose $P < 0.01$ as an arbitrary cut-off for significant differentiation between low- and highland populations, and identified 1,040 SNPs in Mexico (1,040/76,989=0.0135) and 756 SNPs in South America (756/63,160=0.0120) as outliers.

Patterns of adaptation

***Adaptation via mutation versus standing variation*:**

In order to characterize patterns of adaptation, we first determined whether SNPs showing high differentiation between the lowlands and the highlands arose primarily through new mutations or standing genetic variation. We found that these putatively adaptive variants in both Mexico and South America tended to segregate in lowland populations more often than other SNPs (84.6% vs. 74.8% in Mexico, Fisher's exact test (FET) $P < 10^{-11}$ and 87.3% vs 81.8% in South America, $P < 10^{-3}$). *("segregate in" – are these percentages of SNPs that are polymorphic (as opposed to fixed for either allele) in the relevant population?)* yes. No fixed allele between Mexico and S. America *(I don't think I understand. I'd expect SNPs with high frequency differences between high and lowland to have different chances of polymorphism in lowland populations, somehow. Why is this meaningful?)* We extended this analysis to standing variation in *parviglumis* by retrieving SNP data from 14 *parviglumis* inbred lines included in the Hapmap v2 data set, using only SNPs with $n \geq 10$ (**??**). Again we found that putatively adaptive variants were more likely to be polymorphic in *parviglumis* (81.1% vs. 72.1% in Mexico, FET $P < 10^{-6}$ and 81.2% vs 72.7% in South America, $P < 10^{-4}$).

add para These results suggest that maize adaptation to high altitudes has largely made use of standing genetic variation. Recently, examples of adaptation from standing variation have been increased (Reviewed in **??**), and genome scan studies indicated that soft sweeps are plentiful in Drosophila (**?**) and human (**??**). Our result is consistent with this view. In the case of maize, the divergence events occurred very recently (**???**). Thus, it would be unlikely that adaptation via *de novo* mutations frequently occur, which requires a longer waiting time until new beneficial mutations are arisen. Theoretically, the situation of maize would be consistent with standing variation scenario. Selection from standing variation is expected to be common when the scaled mutation rate (the product of effective population size, mutation rate and target size), $\Theta \geq 1$, as long as the scaled selection coefficient (product of effective population size and seleciton coefficient) Ns is large enough (**?**). Maize should have large Θ and Ns because the effective population size is relatively large (synonymous nucleotide diversity = 0.014, (*e.g.,* **???**)).

***Highland versus lowland adaptation*:**

Given the historical spread of maize from an origin in the lowlands, it is tempting to assume that significant population differentiation 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* (Text S1). Consistent with predictions, we infer that differentiation at 67.5% (414) and 75.9% (453) of SNPs in Mexico and S. America is due to adaptation in the highlands, with only 32.5% (199) and 24.1% (144) of SNPs inferred to be due to lowland adaptation after excluding the SNPs with ambiguous patterns (probably due to recombination). The majority of these SNPs show patterns of haplotype variation (by PHS test) consistent with our inference (Text S1 and Table S3).

***Adaptation through introgression*:**

A marked difference between highland adaptation of maize in Mexico and S. America is the potential for adaptation through introgression from wild relatives. While maize in Mexico grows in sympatry with both the lowland taxon *parviglumis* and the highland taxon *mexicana*, maize in South America is outside the range of wild *Zea* species. (**?**) recently assessed the potential for local adaptation in *parviglumis* and *mexicana* populations, characterizing differentiation between these subspecies using an Fst-outlier approach. We observed a significant excess of overlap between our putatively adaptive SNPs in Mexican maize and those identified in the (**?**) analysis (Table **??**; $P < 0.01$ by FET). Similar to that paper, we also find that SNPs with significant F_{ST} P-values are enriched in intergenic regions compared to non-significiant SNPs (51.3% vs. 44.2%; FET $P < 10^{-8}$). Significant overlap was also observed between significant SNPs in S. America and teosinte ($P < 0.01$), but the proportion of SNPs was lower than observed in Mexico. These data suggest that adaptations in Mexican maize may have been obtained through gene flow with wild relatives. To more fully explore this hypothesis we evaluated our data in light of introgression identified by (**?**) from *mexicana* into maize in the Mexico highlands. The proportion of significant SNPs in introgressed regions in Mexico is significantly higher than found in S. America (FET $P \ll 0.001$). Outside introgressed regions, the Mexican and S. American populations did not show marked differences in the proportion of significant SNPs (Fisher's exact test, $P > 0.7$). These results combined with those from (**?**) suggest that SNPs in introgressed regions have indeed been under selection.

***No evidence for parallel adaptation*:**

While maize adaptation in Mexico and S. America are likely distinguished by unique histories of gene flow with wild relatives, the potential remains for parallel adaptation in these two regions. SNPs showing significant differentiation between low- and highland populations in both Mexico and S. America are likely candidates for parallel adaptation. We identify 56 SNPs with F_{ST} P-values in Mexico (P_M) and S. America (P_S) both < 0.01. This number was significantly larger than the random expectation ($48,370 \times 0.01 \times 0.01 \approx 4.8$; χ^2-test, $P \ll 0.001$). Furthermore, the distribution of P_M in the 712 SNPs with $P_S < 0.01$ was highly skewed toward zero (Figure S4A), and a similar tendency was observed in P_S given $P_M < 0.01$ (935 SNPs; Figure S4AB). Thus, we converted the P-values in one population given $P < 0.01$ in the other population into q-values. At a false discovery rate of 0.2 we found 117 SNPs with $P_M < 0.01 \cap P_S < 0.0169$ or $P_M < 0.0247 \cap P_S < 0.01$, and these SNPs were considered our candidates for parallel adaptation. We found 959 SNPs showing significant population differentiation only in Mexico ($P_M < 0.01 \cap P_S > 0.0169$) and 664 SNPs only in South America ($P_M > 0.0247 \cap P_S < 0.01$). The scatter plot of P_M and P_S is shown in Figure 5. For a subset of 67 of the SNPs showing putative evidence of parallel adaptation we also had data from *parviglumis* and were able to infer based on patterns of segregation whether these SNPs were potentially adaptive under lowland or highland conditions (Text S1). Surprisingly, SNPs identified as targets of parallel adaptation in Mexico and South America more frequently show segregation patterns consistent with lowland adaptation (62 SNPs) than highland adaptation (5 SNPs).

In addition to evaluating parallel adaptation at the SNP level, we investigated how often different SNPs in the same gene may have been targeted by selection. To search for this pattern, we define a "genetic unit" or GU as all SNPs within 10kb of a transcript. SNPs in an miRNA or second transcript within 10kb of the transcript of interest were excluded. We classified SNPs showing significance in Mexico, S. America or in both regions into 1,277 GUs. Of these, 95 GUs contained at least one SNP with a pattern of differentiation suggesting parallel adaptation, whereas only 12 GUs contained both Mexico-specific and SA-specific significant SNPs. Overall, fewer GUs showed evidence of parallel adaptation than expected by chance (permutation test; $P < 10^{-5}$), with more than 700 and 470 GUs showing Mexico-specific and SA-specific significant SNPs, respectively. Despite similar phenotypes and environments, we thus see little evidence for parallel adaptation at either the SNP or the gene (GU) level.

Need your comments!! Our result of few parallel adaptation in maize contrasts with data from humans (**?**) showing frequent evidence of selection on the same genes in multiple pairs of tropical and temperate human populations. It is common in maize and human that the majority of adaptive variants to high altitudes would be derived from standing variation (**?**). On the other hand, one difference between these two species is an effective population size. The effective size of maize (on the order of 10^5) is an order of magnitude larger than that in human (**?**). Human would have less potential to possess genetic variants as a source of adaptation, so it could be more likely that the same variants are selected in multiple subpopulations (as long as s is enough large and initial frequency is, for example, > 0.1). One the other hand, maize would maintain a larger amount of variants in the ancestral lowland population. In this case, if genetic variants have almost the same phenotypic effect, the each highland population might pick up one of them randomly. Or if Mexican and S. American highlands have slightly different climate condition, it is feasible that different variants are selected. The target size of mutations can also increase the variants for adaptation, but there is no data of this size in maize and human. *(You are discussing adaptation from standing variation; what about from new mutation? This wouldn't require postulating lots of equivalent variants?)*

Indeed, there are lines of evidence of adaptation from multiple standing variants in maize. One example is *grassy tillers 1* (*gt1*) genes in maize (**?**). There are at least two artificially selected mutations on this gene that reduces the number of ears and would be beneficial for the domesticated maize (easy to harvest). *parviglumis* possesses the two mutations as stand-

ing variation with low frequency, and both were spread across maize by selection after domestication.

Because linkage disequilibrium in maize decays rapidly (**??**) as in our data (Figure S5), it is plausible that a number of hard sweeps – strong selection on new mutations – would be missed by our data, but several lines of evidence suggest to us that this is unlikely.

Comparison to theory

For a final point of comparison, we assessed the degree of parallelism expected under a spatially explicit population genetic model. We estimate the (maize) population density ρ of the highlands to be around (0.5 people/km^2) $\times$ (0.5 ha field/person) $\times$ (2×10^4 plants per field ha) = 5,000 plants per km^2. The area of the Andean highlands is around $A = 500$km^2, leading to a total population of $A\rho = 2.5 \times 10^6$. Combined with an offspring variance of $\xi^2 = 30$, we can compute the rate λ_{mut} at which newly adapted alleles arise in the population. We observe that even if there is strong selection for an allele at high elevation ($s_b = 0.1$), a single-base mutation with mutation rate $\mu = 10^{-8}$ would still take at least 6,000 generations to appear and fix. On the other hand, a kilobase-sized target with mutation rate $\mu = 10^{-5}$ with this selection coefficient would appear and begin to fix in only 6 generations, while more weakly selected alleles with s_b of 10^{-2} or 10^{-3} would take hundreds or thousands of generations, respectively. (Note that the time scales linearly with the selection coefficient: at these values $T_{\text{mut}} = 1/\lambda_{\text{mut}} \approx \mu s_b \times 1.6 \times 10^5$.) Therefore, we might expect to see parallel changes of similar effect at the level of genes (e.g. disabling mutations), but would not expect to see adaptive SNPs that arise through independent mutation in the two populations.

Parallel SNP changes seem unlikely from new mutation; what about gene flow between the highlands? From the demographic model above we have estimated that $\sigma \approx 1.8$ kilometers per generation, so with $10^{-1} \geq s_m \geq 10^{-4}$ 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 selective cost $s_m = 10^{-3}$ in the lowlands, to transit between the two highlands is $T_{\text{mig}} \approx 5 \times 10^{34}$ generations. In other words, from these calculations it is almost impossible that an allele that is deleterious at low elevation with $s_m = 10^{-3}$ would ever transit from the Mexican to the Andean highlands. If the selection against the allele is even weaker ($s_m = 10^{-4}$) it is still expected to take $T_{\text{mig}} = 1.8 \times 10^8$ generations. However, shorter distances could be transited by very weakly deleterious alleles – if the distance between highland patches R is 1,000 km (or if σ is four times larger) then with $s_m = 10^{-4}$ the time T_{mig} is about 1.6 generations – so, adaptation by migration is certain in the known timeframe of maize diffusion. This is strongly dependent on the magnitude of the deleterious selection coefficient: for example, with $s_m = 10^{-3}$, T_{mig} is 2.3×10^6 generations.

This suggests that even when highland-adaptive mutations are weakly deleterious in the lowlands, gene flow will not result in shared adaptations. The situation where these are neutral in the lowlands 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 (**?**). 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 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 $\mathbb{P}\{Z_n/\sqrt{m\sigma^2} \leq x/\sqrt{2\log n} + \sqrt{2\log n} - (1/2)(\log\log n + \log 4\pi)/\sqrt{2\log n}\} \approx \exp(-e^{-x})$ (**?**). With $n = 1000$, the typical distance to the furthest displacement after $m = 1000$ generations is $\sqrt{2\sigma^2 m \log n} \approx 212$km; after $m = 6000$ generations it is ≈ 518km. In either case, the chance that the maximum is larger than 1,000km after 6,000 generations is well less than 10^{-4}. Of course, this is under an equilibrium population model; and maize reached the Andean highlands only around 4,000 years ago. Nonetheless, this suggests that even highland-adapted allele that are merely neutral in the lowlands would have difficulty moving between the Mexican and Andean highlands in a few thousand generations.

Based on our spatially explicit population genetic model, parallel adaptation involving identical nucleotide changes is quite unlikely under either scenarios of independent mutation or transit of Central America by undirected (diffusive) sharing of seed. However, independent mutations could be expected in kilobase-sized targets, suggesting there might be signal for genes that share adaptive changes. These 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.

Conclusions

1. We successfully inferred demography and detected the candidates of adaptive loci to highland climates in Mexico and South America by utilizing GBS and 55-k chip.

2. The main conclusion is parallel adaptation is rare in maize highland adaptation.

Acknowledgements

We appreciate the helpful comments of P. Morrell and the members of Ross-Ibarra lab and Coop labs.

1 Details of the demographic model

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_* is approximately $p_* \approx 2s/\xi^2$ to 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.

1.1 Maize model

Specifically, the migration and reproduction dynamics we use are as follows. 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. 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 in sight 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

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

1.2 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.

1.3 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).

1.4 Dispersal distance

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.

1.5 Results

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 pretty darn close.

2 Adaptation by mutation

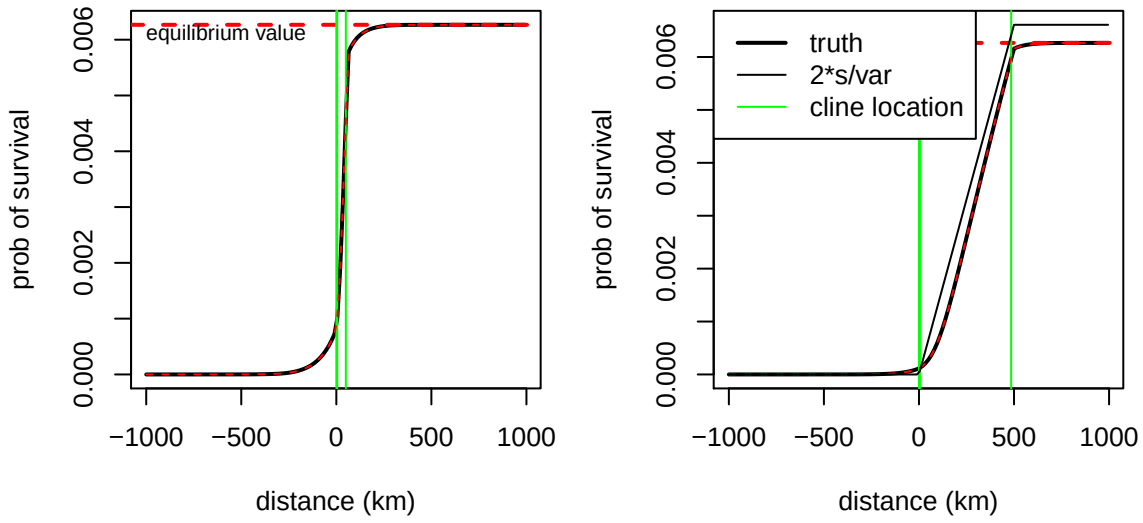
(just a placeholder for now; to be merged in)

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 altitudial gradient)) $\times$ (population density) $\times$ (prob of establishment at high altitude).

Let A denote (area of high altitude) plus (1/2 area of altitudial gradient). The population density ρ is roughly 0.5–5 people per km^2 $\times$ (0.5 ha field/person) $\times$ (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

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



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

At the values above, with $.1 \leq s_b \leq .001$, 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.

3 Adaptation by migration

As we show in the theory paper, the rate of adaptation by diffusive migration is roughly

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

We can talk about this more and give some simulations. But for now, let's interpret.

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 fallws 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.

4 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.