Independent molecular basis of convergent highland adaptation in 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 Organismal Biology, Iowa State University, Ames, Iowa 50011, USA,
[‡‡]The Center for Population Biology and the Genome Center, University of California, Davis, California 95616 , USA

Revised manuscript for *Genetics*, August 28, 2014

ABSTRACT Convergent evolution occurs when multiple species/subpopulations adapt to similar environments via similar phenotypes. We investigate here the molecular basis of convergent adaptation in maize to highland climates in Mexico and South America using genome-wide SNP data. Taking advantage of archaeological data on the arrival of maize to the highlands, we infer demographic models for both populations, identifying evidence of a strong bottleneck and rapid expansion in South America. We use these models to then identify loci showing an excess of differentiation as a means of identifying putative targets of natural selection, and compare our results to expectations from recently developed theory on 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. *abstract could use some more exciting wording*

Introduction

Convergent evolution occurs when multiple species or populations exhibit similar phenotypic adaptations to similar environmental challenges (**???**). *after some discussion of this online, i'll admit Coop's choice is a good one, and let's follow Arendt2008 and call everything "convergent". thoughts? if so, some find/replace rewording will be needed throughout to get rid of our initial "parallel adaptation" and my later "repeated evolution"* Evolutionary genetic analysis of a wide range of species has provided evidence for multiple pathways of repeated 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 (**??**). Repeated 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 (**?**). Finally, repeated evolution may arise from natural selection acting on standing genetic variation in an ancestral population. In the three-spined stickleback, natural selection has repeatedly acted to reduce armor plating in independent colonizations of freshwater environments. Adaptation in these populations occurred from standing variation at the *Eda* locus in marine populations (**?**). We still know relatively little, however, about how common it is for repeated phenotypic evolution to be driven by common genetic changes or the relative frequencies of these different routes of repeated evolution. *another recent nice example, repeated gene loss in cyanogenic clovers* `http://rstb.royalsocietypublishing.org/content/369/1648/20130347` *or convergene evolution of electric shocks* `http://www.sciencemag.org/content/344/6191/1522`*; also feel like we need more intro here. this mentions a few ideas, but not some of the interesting questions or relevant points about genetic architecture, target size,*

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

the role of migration, etc.

Domesticated maize (*Zea mays* ssp. *mays*) provides an excellent opportunity to investigate the molecular basis of repeated 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) (**???**). After domestication, maize spread rapidly across the Americas, reaching the lowlands of South America and the high altitudes of the Mexican Central 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 **??**) 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 (**?**). *Sho: please check for hits in any UV related genes. see http://www.biomedcentral.com/1471-2229/12/92, http://www.reeis.usda.gov/web/crisprojectpages/0195696-maize-responses-to-uv-b-a-genomics-assessment.html, and http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3040.2005.01329.x/full as well as (**?**)*

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 repeated evolution. 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 from *mexicana* likely 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 the highlands of Mexico and South America show evidence of repeated evolution at the molecular level? How do observed patterns of repeated evolution 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 repeated evolution. We estimate unique demographic histories in the highlands of Mexico and South America, and find evidence supporting our theoretical predictions that adaptation should be largely independent. Our population genetic analysis also supports an adaptive role for gene flow from *mexicana* and highlights the contribution of standing variation to adaptation in both populations.

Materials and Methods

Materials and DNA extraction

We included one individual from each of 94 open-pollinated landrace maize accessions from high and low elevation sites in Mexico and S. America (Table **??**). Accessions were provided by the USDA germplasm repository or kindly donated by Major Goodman (North Carolina State University). Sampling locations are shown in Figure **??**A (see also Table **??**). Landraces sampled from altitudes $< 1,700$ 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 (**?**). The quality of DNA was ensured through inspection on a 2% agarose gel and quantification of the ratio of light absorbance at 260 and 280 nm using a NanoDrop spectrophotometer (Thermo Scientific, NanoDrop Products, Wilmington, DE, USA).

SNP data

We generated two complementary SNP data sets for the sampled maize landraces. The first set was generated using the Illumina MaizeSNP50 BeadChip platform, including 56,110 SNPs (**?**). 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. This array contains SNPs discovered in multiple ascertainment schemes (**?**); however, 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 inbred lines (40,594 SNPs; **?**).

The second data set was generated for a subset of 87 of the landrace accessions (Table **??**) utilizing high-throughput Illumina sequencing data via genotyping-by-sequencing (GBS; **?**). Genotypes were called using TASSEL-GBS (**?**) resulting in 2,848,284 SNPs with an average of 71.3% missing data per individual. *Sho, please fill in the XX's and add label references for supp tables throughout* done

To assess data quality, we compared genotypes at the 7,197 SNPs (229,937 genotypes, excluding missing data) that overlap

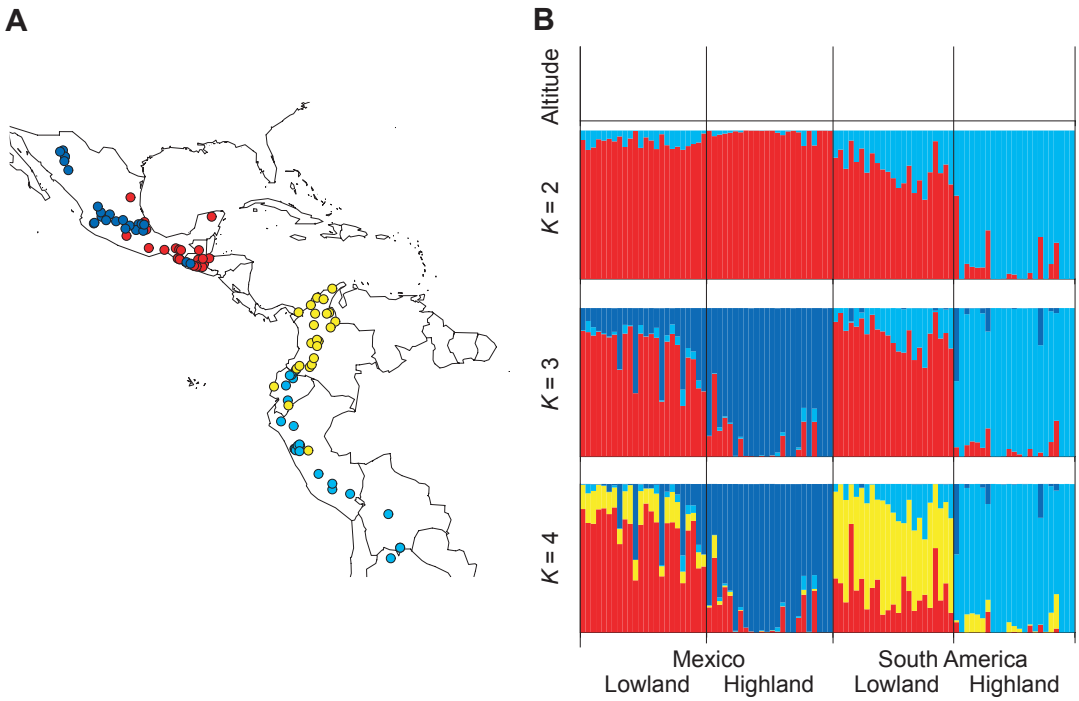


Figure 1 (A) Sampling locations of landraces. Red, blue, yellow and light blue dots represent Mexican lowland, Mexican highland, S. American lowland and S. American highland populations, respectively. (B) Results of STRUCTURE analysis of the maizeSNP50 SNPs with $K = 2 \sim 4$. The top panel shows the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

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 **??**) supports its utility in estimating allele frequencies.

We annotated SNPs using the filtered gene set from RefGen version 2 of the maize B73 genome sequence (**?**; 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 (**??**) 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 0.05 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 $K = 2$ through 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 **??**). Observed joint frequency distributions (JFDs) were calculated using the GBS data set due to its lower level of ascertainment bias. A subset of silent *would it be correct to change "silent" to say "synonymous and noncoding"?* ok 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 $P < 0.005$. Demographic parameters were inferred using the software $\delta a \delta i$ (**?**), which uses a diffusion method to calculate an expected JFD and evaluates the likelihood of the data using a multinomial assumption.

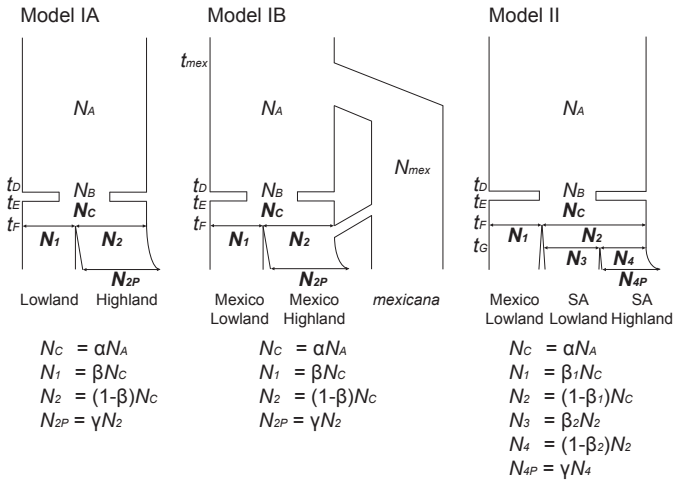


Figure 2 Demographic models of maize low- and highland populations. Parameters in bold were estimated in this study. See text for details.

***Model IA*:** This model is applied to the Mexican and S. American populations. We assume the ancestral diploid population representing *parviglumis* follows a standard Wright-Fisher model with constant size. The size of the ancestral population is denoted by N_A. At t_D generations ago, the bottleneck event begins at domestication, and at t_E generations ago, the bottleneck ends. The population size and duration of the bottleneck are denoted by N_B and $t_B = t_D - t_E$, respectively. The population size recovers to $N_C = \alpha N_A$ in the lowlands. Then, the highland population is differentiated from the lowland population at t_F generations ago. The size of the 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.

isn't this really a shrinking population in the lowlands, since $\beta N_C < N_C$? wouldn't we want instead for lowlands to stay at N_C and a new population branching off? how much do we worry about this? actually, our conclusion holds when I assumed the pop size of lowlands stays at N_C. However, the likelihood is a bit better in my original model.

***Model IB*:** We expand Model IA for the Mexico populations by incorporating admixture from the teosinte *mexicana* to the highland Mexican maize population. *do we say "Mexico population" or "Mexican" (and thus "South American") "population" throughout? as long as we're consistent probably OK either way.* vote to Mexican population The time of differentiation between *parviglumis* and *mexicana* occurs at t_{mex} generations ago. The *mexicana* population size is assumed to be constant at N_{mex}. At t_F generations ago, the Mexican highland population is derived from admixture of between the Mexican lowland population and a portion P_{mex} from the teosinte *mexicana* .

***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 set to 150,000 using estimates of the composite parameter $4N_A\mu \sim 0.018$ from *parviglumis* (**?????**) and an estimate of the mutation rate $\mu \sim 3 \times 10^{-8}$ (**?**) per site per generation. 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 IB (**?**), $t_F = 4,000$ for S. American populations in Model lA (**??**), 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).

*t_F for model II is listed as 4,000 and 6,000 above. 6,000 is the number that matches the lit best. is that what was used? if so, we should cite (**?**)* fixed

Differentiation between low- and highland populations

We used our inferred demographic model to generate a null distribution of F_{ST}. As implemented in $\delta a \delta i$ (**?**), we calculated an expected JFD given estimated demographic parameters and the sample sizes of highlands and lowlands. Then, we converted the JFD into the distribution of Fst values. The *P*-values of a SNP was calculated by $P(F_{ST\_E} \geq F_{ST\_O} | p \pm 0.05) = P(F_{ST\_E} \geq F_{ST\_O} \cap p \pm 0.05)/P(p \pm 0.05)$, where F_{ST_O} and F_{ST_E} are observed and expected F_{ST} values and p is the mean allele frequency of highlands and lowlands.

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 (**?**) 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 a JFD from 10^7 SNPs using the software ms (**?**). Then, we calculated *P*-values from the JDF 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.

we don't correct for allele frequency (or heterozygosity) in our Fst outlier analysis do we? if not, this is a problem I think. I did. No essential change

do we use all GBS data in the Fst outlier test, or just silent SNPs? we should be able to use nonsynonymous too, right? I used all.

Haplotype sharing test

We performed a pairwise haplotype sharing (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 from the Hapmap v2 data set (**?**). We ran fastphase with default parameter settings. PHS was calculated for an allele A at position x by

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

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 each allele was calculated as the proportion of alleles of the same frequency genome-wide that have a larger PHS value.

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

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

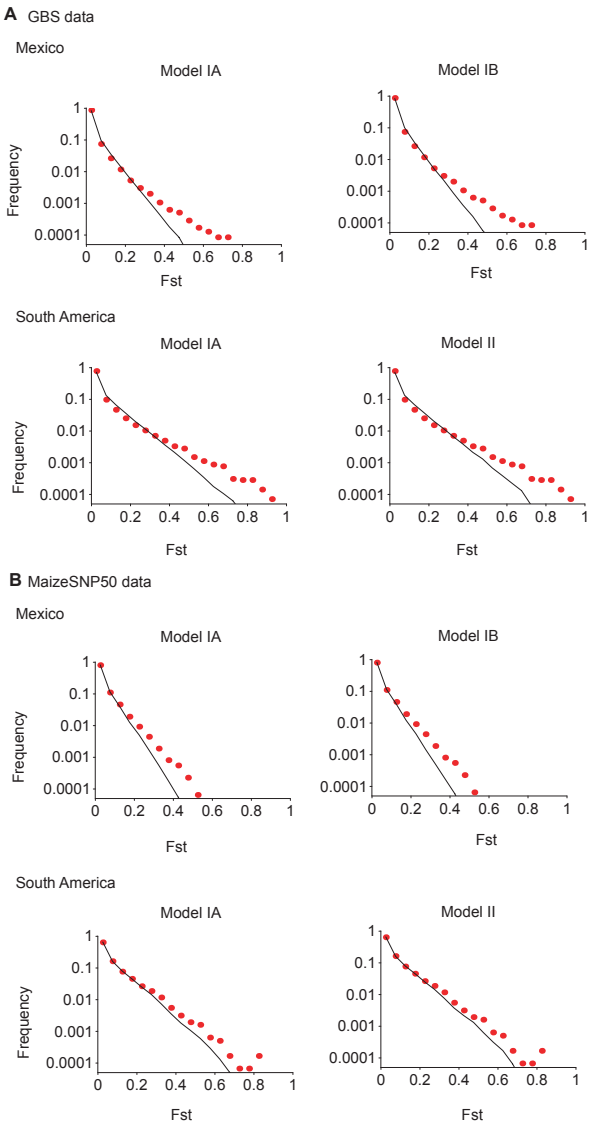


Figure 3 Observed and expected distributions of F_{ST} values in GBS (A) and MaizeSNP50 data (B). The *x*-axes represent F_{ST} values. The *y*-axes represent the frequency of SNPs with F_{ST} values within a bin of 0.05 size. Red dots and solid lines indicate observed and expected distributions.

mutation in a highland region (λ_{mut}), and the rate at which such an allele already present in the Mexican highlands would transit the intervening lowlands and fix in the Andean highlands (λ_{mig}). 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. Here we describe

the mathematical details; readers may skip to the results without loss of continuity.

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. To do this, we followed **?** in constructing a detailed demographic model for domesticated maize. 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 **??**.

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 **??**); 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 to occur. In other words, we can bound the rate of migration by

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

with N being 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

Our sample included 94 maize landraces from four distinct regions in the Americas: the lowlands of Mexico/Guatemala

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	–

Table 2 Inference of demographic parameters

Mexico	Model I		Model II	
	Likelihood	−5592.80	Likelihood	−4654.79
	α	0.92	α	1.5
	β	0.38	β	0.76
	γ	1	γ	1
South America	Model I		Model III	
	Likelihood	−3855.28	Likelihood	−8044.71
	α	0.52	α	1.0
	β	0.97	β_1	0.64
	γ	88	β_2	0.95
			γ	54

(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). We hereafter refer to the two SNP data sets as "MaizeSNP50" and "GBS". In total, we genotyped 91,779 SNPs after filtering by Hardy-Weinberg criteria with sample size $\geq$ 10 in each of the four populations (see Materials and Methods); 67,828 and 23,951 SNPs were generated by GBS and MaizeSNP50 respectively.

Population structure

We performed a STRUCTURE analysis (**??**) of our landrace sample, varying the number of groups from K = 2 to 6 (Figure **??**, 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

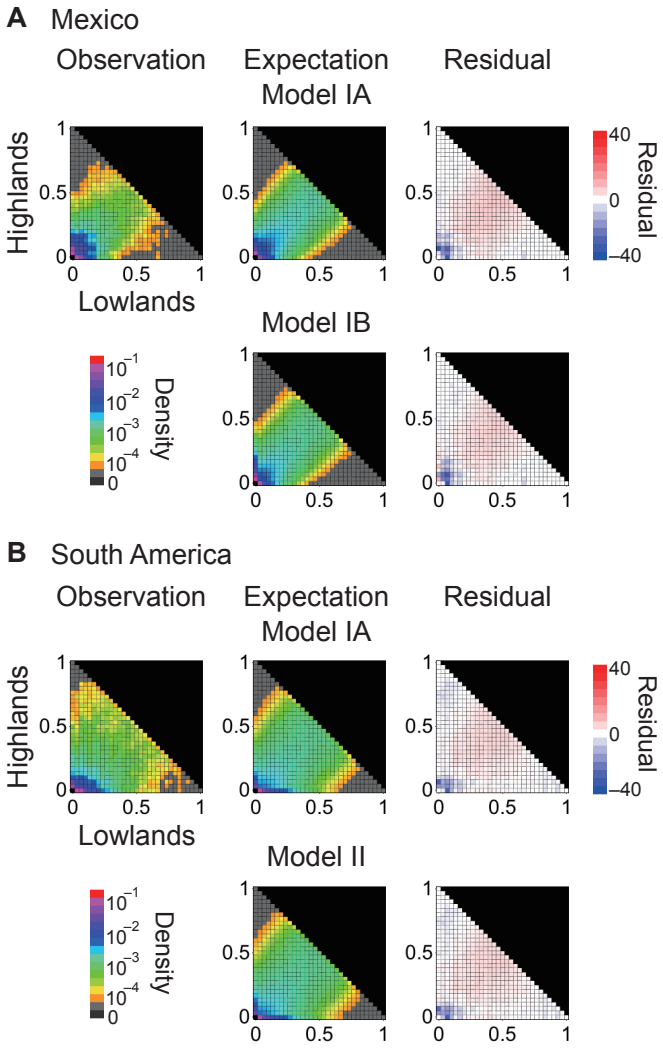


Figure 4 Observed and expected joint distributions of minor allele frequencies in low- and highland populations in (A) Mexico and (B) S. America. Residuals are calculated as $(\text{model} - \text{data})/\sqrt{\text{model}}$

differentiation may be the result of a stronger bottleneck during colonization of the S. American highlands.

Population differentiation under inferred demography

To provide a null expectation for allele frequency differentiation, we used the joint site frequency distribution (JFD) of lowland and highland populations to estimate parameters of two demographic models using the maximum likelihood method implemented in $\delta a \delta i$ (**?**). 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 **??**; 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 **??**). 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 **??**, we investigated the impact of varying the domestication bottleneck size (N_B). Surprisingly, N_B was estimated to be equal to N_C, the population size at the end of the bottleneck, and the likelihood of $N_B < N_C$ was much smaller than for alternative parameterizations (Table **??**). This result appears to contradict earlier work using sequences from coding regions to infer a maize domestication bottleneck (**???**). Consistent with **?**, our genome-wide SNP data show an excess of rare variants relative to expectations under **?**'s bottleneck model (Figure **??**), suggesting a domestication model involving a weaker bottleneck or more rapid population growth.

Comparisons of 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 **??**), and model choice – including the parameterization of the domestication bottleneck – had little impact on the distribution of estimated p-values (Figure S**??**). Thus, hereafter, we show the results under Model IB for Mexican populations and Model II for S. American populations. We chose $P < 0.01$ as an arbitrary cut-off for significant differentiation between low- and highland populations, and identified 687 SNPs in Mexico (687/76,989=0.89%) and 409 SNPs in South America (409/63,160=0.65%) as outliers (Figure **??**).

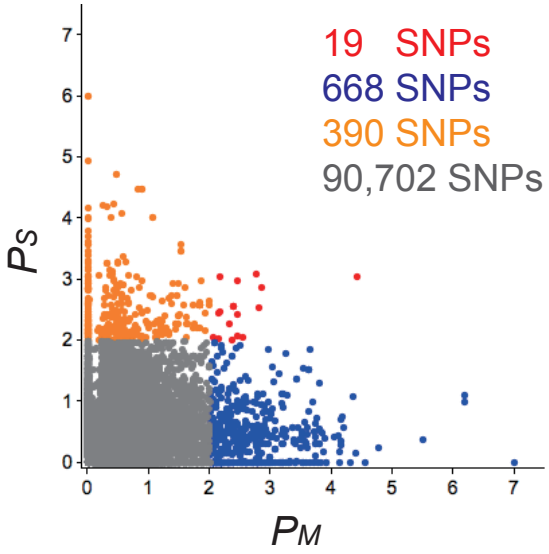


Figure 5 Scatter plot of $-\log_1 0P$-values of observed F_{ST} values based on simulation from estimated demographic models. P-values are shown for each SNP in both Mexico (Model IB; P_M on x-axis) and South America (Model II; P_S on y-axis). Red, blue, orange and gray dots represents SNPs showing significance 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 as the points.

Patterns of adaptation

***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* (See Supplementary Text for details). Consistent with predictions, we infer that differentiation at 72.3% (264) and 76.7% (230) of SNPs in Mexico and S. America is due to adaptation in the highlands after excluding 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 (Supplementary Text and Table **??**).

***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 Mexican lowland population more often than other SNPs (85.3% vs. 74.8% in Mexico, Fisher's exact test (FET) $P < 10^{-9}$ and 94.8% vs 87.4% in South America, $P < 10^{-4}$).

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* (78.3% vs. 72.2% in Mexico, FET $P < 0.01$ and 80.2% vs 72.8% in South America, $P < 0.01$). These results suggest that maize adaptation to high altitude has largely made use of standing genetic variation. Recent empirical examples of adaptation from standing variation (Reviewed in **??**) and detection of soft selective sweeps in Drosophila (**?**) and humans (**??**) suggest this may be a common form of adaptation. Selection from standing variation should be common when the scaled mutation rate (the product of the effective population size, mutation rate and target size), $\Theta \geq 1$, as long as the scaled selection coefficient

Table 3 Tanja's F_{CT} between *parviglumis* and *mexicana*

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

(product of the effective population size and selection coefficient) Ns is large enough (**?**). Estimates of θ from synonymous nucleotide diversity in maize ($\cong 0.014$ (*e.g.,* **???**)) then suggest adaptation from standing genetic variation would be likely for target sizes larger than a few hundred nucleotides.

***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 F_{ST}-outlier approach. We observed a significant excess of overlap between our putatively adaptive SNPs in Mexican maize and those identified in the **?** analysis (Table **??**; $P < 10^{-8}$ by FET). Similar to that investigation, we found that SNPs with significant F_{ST} P-values were enriched in intergenic regions than protein coding regions (60.0% vs. 47.9%, FET $P < 10^{-7}$ for Mexico; 62.0% vs. 47.8%, FET $P < 10^{-5}$ for S. America). Significant overlap was also observed between S. America and teosinte (Table **??**; $P < 10^{-3}$), but the proportion of SNPs was lower than 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 Mexican highlands. The proportion of significant SNPs in introgressed regions in Mexico is significantly higher than found in S. America (FET $P < 10^{-6}$). Outside introgressed regions, the Mexican and S. American populations did not show marked differences in the proportion of significant SNPs (Fisher's exact test, $P = 0.036$). These results combined with those from (**?**) suggest that SNPs in introgressed regions have indeed been under selection. *add sho's rsults about fst in and out of introgressed region*

***Genetic basis of convergent evolution*:**

While maize adaptation in Mexico and S. America are likely distinguished by unique histories of gene flow with wild relatives, convergent phenotypic evolution in these population may nonetheless have a shared genetic basis. Convergent evolution at the nucleotide level should be reflected in an excess of SNPs showing significant differentiation between low- and highland populations in both Mexico and S. America. We indeed see such an excess, with 19 SNPs showing F_{ST} P-values < 0.01 in both Mexico (P_M) and S. America (P_S), compared to ≈ 5 expected ($48,370 \times 0.01 \times 0.01 \approx 4.8$; χ^2-test, $P \ll 0.001$).

For 13 of 19 SNPs showing putative evidence of shared selection 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 (Supplemental Text). Surprisingly, SNPs identified as shared adaptive variants more frequently showed segregation patterns consistent with lowland (10 SNPs) rather than highland adaptation (2 SNPs). (1 SNP with an ambiguous pattern).

Our demographic models are not perfect; the observed proportion of rare variants was much higher than expected (Figure **??**). Nevertheless, this faultiness does not affect our conclusion. We performed a demographic model-free approach. We simply defined adaptive variants that showed top 1% highest Fst values given $p \pm 0.05$, where p is the mean allele frequency in highlands and lowlands. We detected no shared adaptive variant between Mexican and S. American populations by this approach.

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 defined 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 778 GUs. Of these, 14 GUs contained at least one SNP with a pattern of differentiation suggesting convergent evolution and 2 GUs contained both Mexico-specific and SA-specific significant SNPs. Overall, fewer GUs showed evidence of convergent evolution than expected by chance (permutation test; $P < 10^{-5}$), with 485 and 277 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.

this paragraph needs reworking and to jive with the below from peter The rarity of parallel adaptations in maize contrasts with data from humans (**?**) showing selection on the same genes in multiple pairs of tropical and temperate populations. However, in both maize and humans the majority of adaptive variants appear to have been derived from standing variation (**?**). One difference between these two species is effective population size: the effective population size of maize ($\sim 10^5$) is an order of magnitude larger than that in humans (**?**). Humans would therefore have less standing variation as a source of adaptation, resulting in the same variants being selected in multiple subpopulations (as long as s is sufficiently large and initial frequency is, for example, > 0.1). In contrast, maize could maintain a larger number of adaptive variants in the ancestral lowland popula-

tion. In this case, if genetic variants produce similar phenotypic effects, they may be selected to high frequency in independent highland regions at random. Or if Mexican and S. American highlands have slightly different climates, it is feasible that different variants are selected. The target size of mutations can also increase the variants for adaptation, but there are currently no data regarding mutational target size in maize versus humans. *(You are discussing adaptation from standing variation; what about from new mutation? This wouldn't require postulating lots of equivalent variants?)* In fact, adaptation from multiple standing variants that produce similar phenotypes has previously been observed in maize: the *grassy tillers1* (*gt1*) gene (**?**) contains two artificially selected mutations that reduce ear number. These mutations both segregate at low frequency in *parviglumis* but have been individually selected to high frequency in different populations of maize.

Comparison to theory

For a final point of comparison, we assessed the degree of convergence expected under a spatially explicit population genetic model. We estimated 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 convergent 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.

While convergent SNP changes seem unlikely from new mutation, the potential remains for gene flow between highland regions. 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 highland regions 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.

Our analysis 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 alleles 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, convergent evolution 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 underestimated the rate of very-long-distance sharing of seed, e.g. if sharing across hundreds of kilometers was common at some point.

Conclusions

WE NEED A CONCLUSION! 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 the Ross-Ibarra lab and Coop labs. This project was supported by Agriculture and Food Research Initiative Competitive Grant 2009-01864 from the USDAs National Institute of Food and Agriculture and funding from the National Science Foundation IOS-1238014.

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 a second order in s. The precise value can be found by defining the generating function $\Phi(u) = \mathbb{E}[u^X]$; the probability of local nonextinction p_* is the minimal solution to $\Phi(1-u) = 1-u$. (This can be seen because: $1-p_*$ is the probability that an individual's family dies out; this is equal to the probability that the families of all that individuals' children die out; since each child's family behaves independently, if the individual has x offspring, this is equal to $(1-p_*)^x$; and $\Phi(1-p_*) = \mathbb{E}[(1-p_*)^X]$.)

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

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

1.1 Maize model

The migration and reproduction dynamics we use are taken largely from **?**. 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; *what happens if we change mean ears per plant from 3 to 1?* 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 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 (**??**), 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 **??**. Note that the approximation $2s$ divided by the variance in offspring number is pretty darn close.

2 Adaptation by mutation

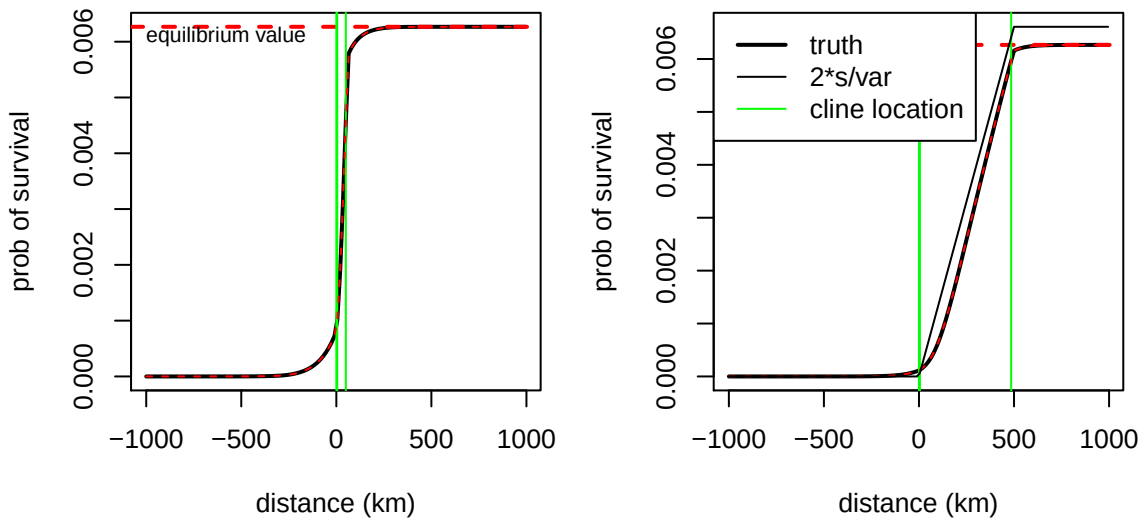
(just a placeholder for now; to be merged in)

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 **??**, 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 *insert citation to theory paper* , the rate of adaptation by diffusive migration is roughly

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

do we need to expland? not sure.

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

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

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.

Later we should separate all Text S1, figs and tables in this file if we submit PLoS G.

Supplemental Text

This needs some rewriting – Matt?

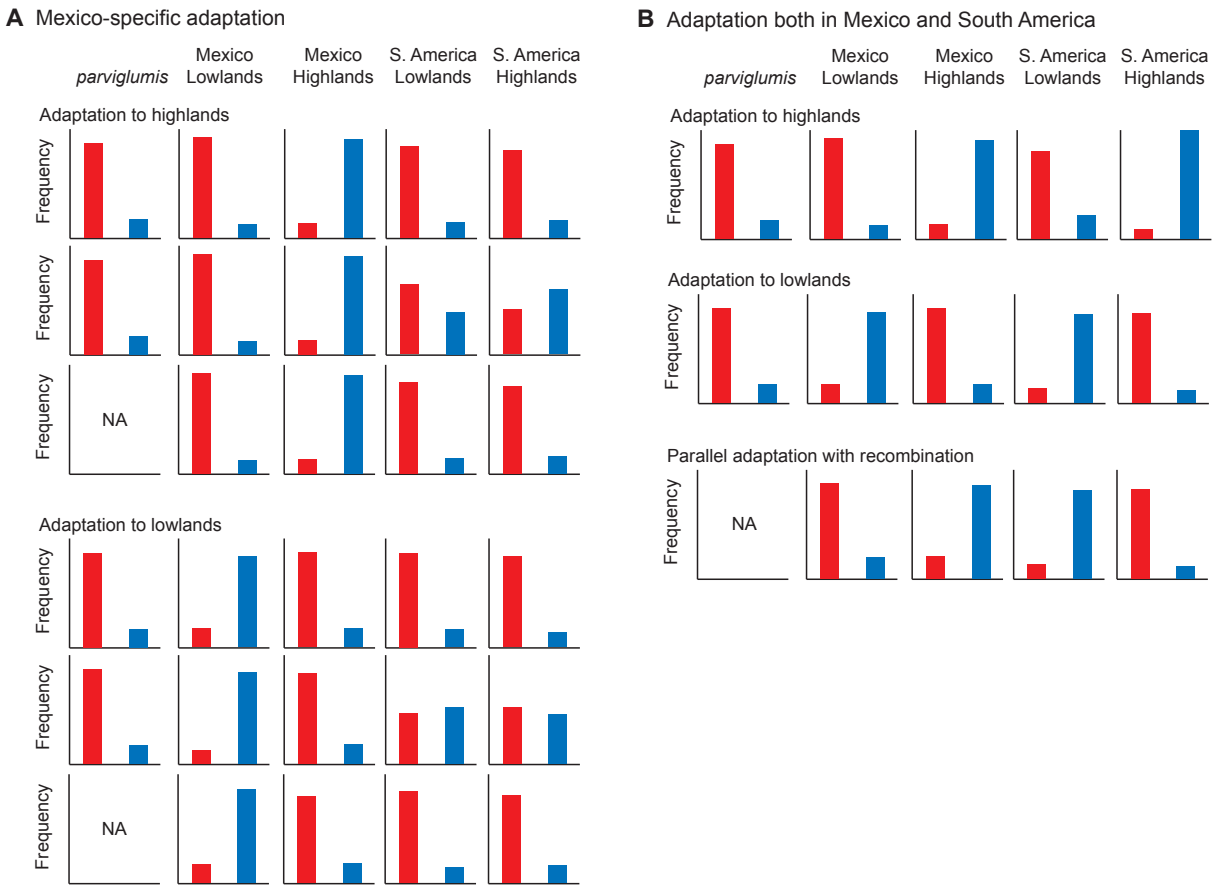
We classified the patterns of allelic differentiation among highland and lowland populations in Mexico and S. America together with the information of *parviglumis* in an *ad hoc* manner; the allelic differentiation pattern is consistent with highland or lowland adaptation scenario. In Figure **??**,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

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

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

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

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



SUPPLEMENTAL FIGURE 2 Illustration of allele frequency changes in maize and *parviglumis*. Red and blue bars represent the allele frequency of ancestral and derived, adaptive alleles, respectively. The allele frequencies in the five populations are shown: *parviglumis*, Mexican lowlands and highlands, and S. America lowlands and highlands. NA in *parviglumis* indicates that there is no SNP data in the site.

SUPPLEMENTAL TABLE 1 List of maize landraces used in this study

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

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

SUPPLEMENTAL TABLE 1 (continued)

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

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

SUPPLEMENTAL TABLE 2 Inference of demographic parameters

Mexico	Model I	
	Likelihood	-3052.34
	α	0.99
	β	0.42
	γ	1
	σ	1
South America	Model I	
	Likelihood	-2717.64
	α	0.51
	β	0.97
	γ	151
	σ	1

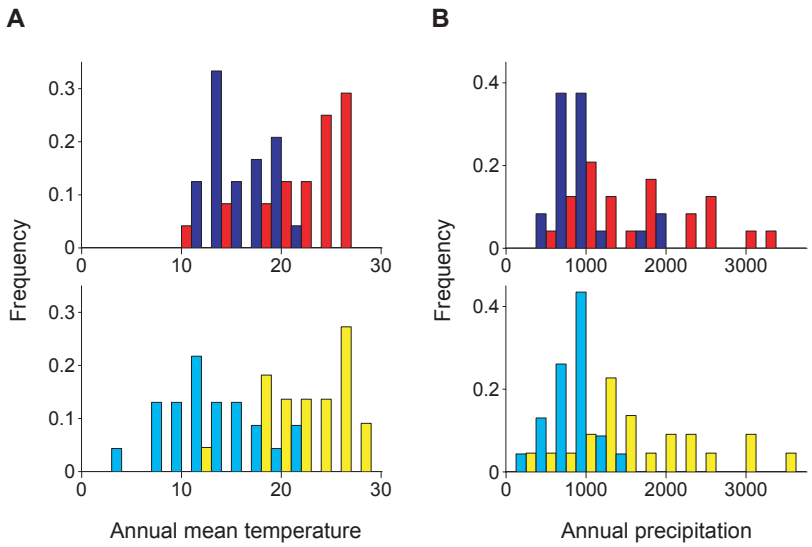
The description of α, β and γ is in Figure 3.
σ is a relative size of N_B to N_C ($N_B = \sigma N_C$).

SUPPLEMENTAL TABLE 3 Summary of PHS test

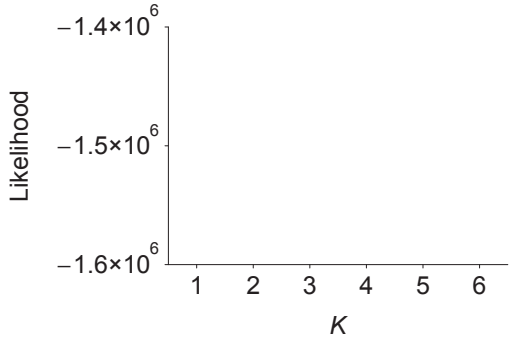
Population	Pattern of adaptation	No. SNPs	No. SNPs supported by PHS test
Mexico	Highland adaptation	264	172 (65.2%)
	Lowland adaptation	101	66 (65.3%)
S. America	Highland adaptation	164	230 (71.3%)
	Lowland adaptation	70	50 (71.4%)

SUPPLEMENTAL TABLE 4 ms command

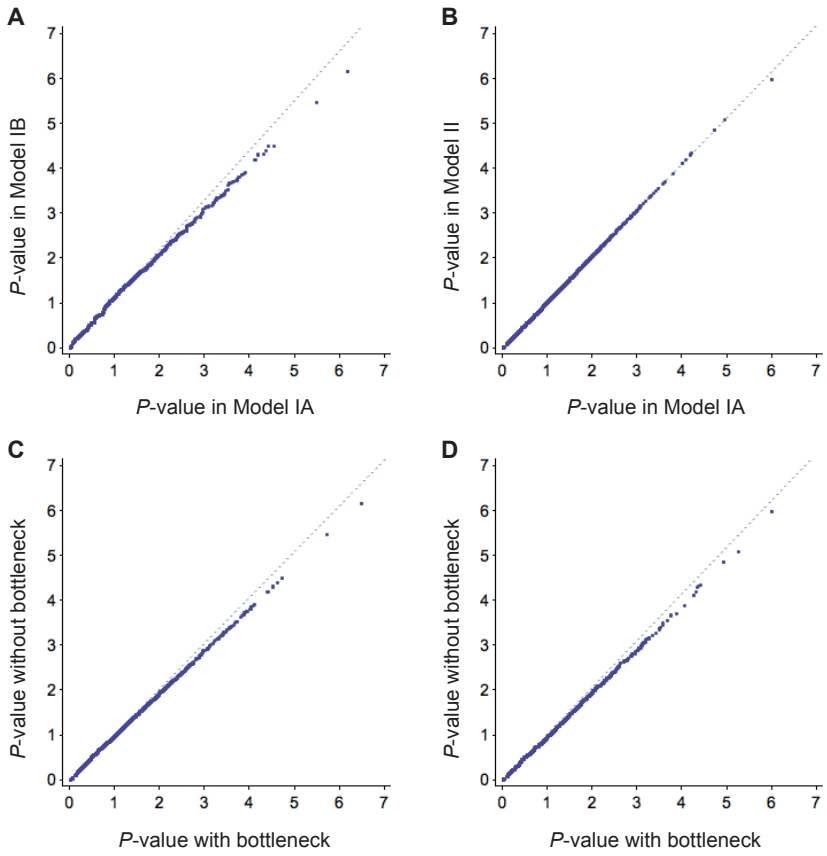
Model I for Mexico populations
Population 1: Mexico lowland population
Population 2: Mexico highland population
-I 2 n_{m1} n_{m2} -n 1 0.3496 -n 2 0.5704 -ej 0.01 2 1 -en 0.01 1 0.92 -en 0.0133 1 0.0163 -en 0.015 1 1.0
Model II for Mexico populations
Population 1: Mexico lowland population
Population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



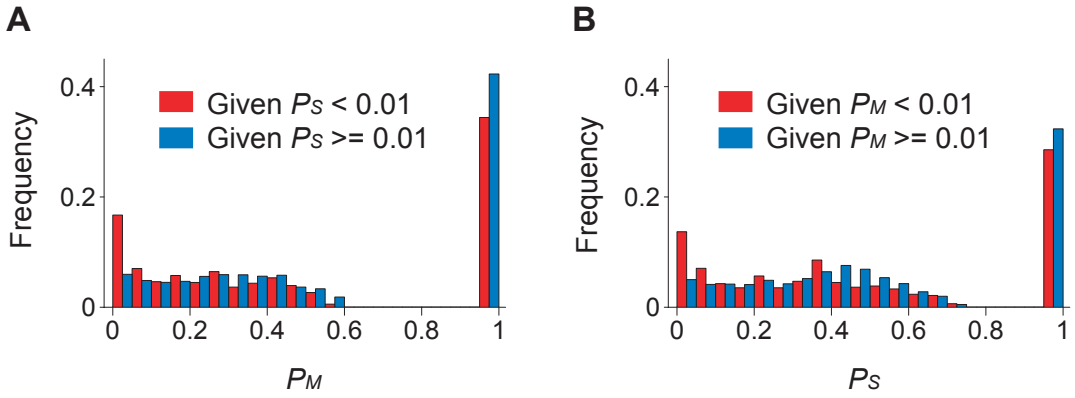
SUPPLEMENTAL FIGURE 1 Correlation of allele frequencies between GBS (*x*-axes) and MaizeSNP50 (*y*-axes) data. We used overlapped SNPs with $n \geq 40$ for both data sets. Correlation coefficient is 0.890 ($P < 10^{-5}$ by permutation test with 10^5 replications).



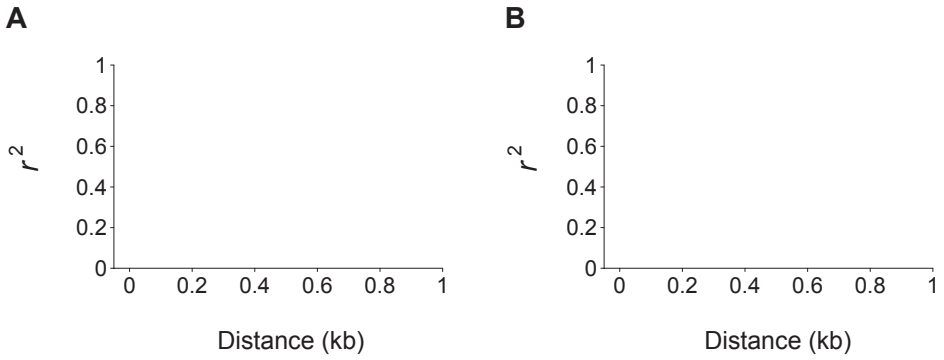
SUPPLEMENTAL FIGURE 2 Likelihood of STRUCTURE analysis given K. The *x*-axes represents K and the *y*-axes represents likelihood.



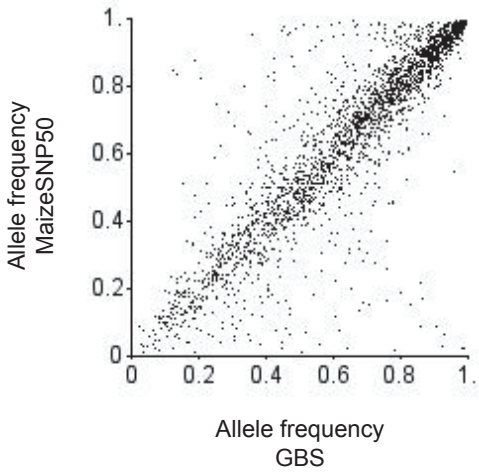
SUPPLEMENTAL FIGURE 3 Q-Q plot for $-\log_{10}$-scaled *P*-values of population differentiation between lowland and highland populations. (A) Model IA *v.s.* Model IB in Mexico, (B) Model IA *v.s.* Model II in S. America, (C) Model with *v.s.* without bottleneck in Mexico and (D) Model with *v.s.* without bottleneck in S. America.



SUPPLEMENTAL FIGURE 4 (A) Frequency distribution of P_M given $P_S < 0.01$ and $P_S \geq 0.01$. (B) Frequency distribution of P_S given $P_M < 0.01$ and $P_M \geq 0.01$.



SUPPLEMENTAL FIGURE 5 Pattern of decay of linkage equilibrium in Mexico (A) and South America (B). Red and blue dots represent low- and highland population, respectively. r^2 values were calculated as a statistics and averaged within 10-bp bins of distance between SNPs. The *x*- and *y*-axes represent distance between SNPs (kb) and average r^2 values.



SUPPLEMENTAL FIGURE 6 Correlation of allele frequencies between GBS (*x*-axes) and MaizeSNP50 (*y*-axes) data. We used overlapped SNPs with $n \geq 40$ for both data sets. Correlation coefficient is 0.890 ($P < 10^{-5}$ by permutation test with 10^5 replications).