

# Molecular Parallelism Underlies Convergent Highland Adaptation of Maize Landraces

Li Wang<sup>1,2,3</sup>, Emily B. Josephs<sup>3</sup>, Kristin M. Lee<sup>3,4</sup>, Lucas M. Roberts<sup>2</sup>, Rubén Rellán-Álvarez<sup>5,6</sup>, Jeffrey Ross-Ibarra<sup>3,7</sup> and Matthew B. Hufford<sup>2</sup>

<sup>1</sup>Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

<sup>2</sup>Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA

<sup>3</sup>Department of Evolution and Ecology, University of California Davis, Davis, CA, USA

<sup>4</sup>Department of Plant Biology, Michigan State University, East Lansing, MI, USA

<sup>5</sup>Langebio, Libramiento Norte Carretera Leon Km 9.6, 36821 Irapuato, Gto., Mexico

<sup>6</sup>Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, NC, USA

<sup>7</sup>Genome Center and Center for Population Biology and Genome Center, University of California Davis, Davis, CA, USA

## Abstract

Convergent phenotypic evolution provides some of the strongest evidence for adaptation. However, the extent to which recurrent phenotypic adaptation has arisen via parallelism at the molecular level remains unresolved, as does the evolutionary origin of alleles underlying such adaptation. Here, we investigate genetic mechanisms of convergent highland adaptation in maize landrace populations and evaluate the genetic sources of recurrently selected alleles. Population branch excess statistics reveal strong evidence of parallel adaptation at the level of individual SNPs, genes and pathways in four independent highland maize populations, even though most SNPs show unique patterns of local adaptation. The majority of selected SNPs originated via migration from a single population, most likely in the Mesoamerican highlands. Polygenic adaptation analyses of quantitative traits reveal that alleles affecting flowering time are significantly associated with elevation, indicating the flowering time pathway was targeted by highland adaptation. In addition, repeatedly selected genes were significantly enriched in the flowering time pathway, indicating their significance in adapting to highland conditions. Overall,

our study system represents a promising model to study convergent evolution in plants with potential applications to crop adaptation across environmental gradients.

# Introduction

Convergent adaptation of populations to similar environments provides compelling evidence that natural selection, not neutral processes, shape trait variation. Similar phenotypes often arise independently in multiple species or populations exposed to the same evolutionary pressure [1, 2]. For example, species from four orders of insects, spanning 300 million years of divergence, have independently evolved tolerance to toxic compounds in milkweed and dogbane plant species [3, 4]. Likewise, reduction-of-function alleles at the FLC locus have evolved multiple times independently in *Capsella rubella* populations, conferring variation in flowering time [5]. Pool et al.[6] investigated cold adaptation across three pairs of highland and lowland *Drosophila melanogaster* populations, finding strong evidence for alleles that were repeatedly selected during highland colonization.

But while convergent phenotypes have been observed across the tree of life, in many cases their underlying genetic basis (*e.g.*, molecular parallelism in which the same nucleotides, genes, or pathways are targeted versus convergence through independent molecular means) is unknown [7, 8]. Study of the genetics of convergence can help shed light on fundamental questions in evolutionary biology, including whether natural selection is constrained and repeatable or instead characterized by many molecular paths to similar phenotypes. The answer to these questions may depend, to a certain extent, on the trait itself. For example, the genetic architecture of a phenotype is a primary determinant of whether trait convergence results from parallel selection on orthologous loci [9]. Such convergence is more likely for simple traits where only a few loci contribute or when loci are subject to antagonistic pleiotropy and less likely for highly polygenic traits such as biomass or height [7, 9, 10].

Convergence through molecular parallelism, when it occurs, can originate in a number of ways. Genetic variation underlying convergent traits can arise as independent mutations, be derived from existing standing variation in a shared ancestral population, or be transferred between populations via migration [11]. Empirical studies have documented each of these modes of parallel adaptation. For instance, independent evolution of C4 photosynthesis in grasses [12] and sedges [13] has involved multiple *de novo* mutations in the phosphoenolpyruvate carboxylase (PEPC) gene, a key C4 enzyme. In contrast, parallel adaptation in rabbit populations resistant to the myxoma virus in Australia, France, and the United Kingdom is achieved through selection on standing variation in immunity-related genes [2]. Finally, convergence in patterns of wing coloration in butterfly species, an example of Müllerian mimicry, has resulted from gene flow across species [14, 15].

Given its history, maize is an ideal model system to study parallel adaptation. Maize was first domesticated in the warm lowlands of the Balsas River Valley approximately 9,000 years ago [16, 17], and subsequently spread to several independent highland regions, first to the the Mexican Central Plateau, and then to the highlands of the southwestern United States, Guatemala and the Andes [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. While highland regions colonized by maize are far from identical, commonalities include a shorter growing season, low temperature, low partial pressure of atmospheric gases, and high ultraviolet radiation [30, 31]. Highland individuals exhibit several traits that are thought to be

adaptive under these conditions, including highly pigmented and hairy leaf sheaths [32]. Common garden experiments have also demonstrated that highland maize flowers substantially earlier than lowland material [33, 34, 35].

The genetic basis of highland phenotypes across these populations is largely unknown, however. Earlier population genetic analysis of maize in Mexico and the Andes identified a small percent of SNPs showing evidence of selection in both populations, but concluded that maize highland adaptation arose largely independently in the two populations [36]. One contributor to the differences between these populations has been gene flow from wild relatives: adaptive introgression from the wild congener teosinte (*Zea mays* ssp. *mexicana*, hereafter *mexicana*) is thought to have played an important role during highland adaptation in Mexico [37], but maize has no wild relatives in South America and simulations suggested that long-distance gene flow between the populations is unlikely [36]. Nonetheless, this earlier work was limited to two highland populations and relied on limited genotyping data. Subsequent investigation using whole genome sequencing, for example, has discovered alleles introgressed from teosinte in both the Guatemalan and southwestern US highlands outside of the distribution of *mexicana* [38], suggesting that migration from the Mexican highlands may not be as implausible as previously thought.

Here, we evaluate evidence for trait convergence through parallel molecular means using resequencing data from four highland populations of maize (Southwestern US, Mexico, Guatemala and the Andes) and two lowland populations (Mexico and South America). We assess the prevalence of molecular parallelism at the SNP, gene, and pathway levels, comparing models of *de novo* mutation, standing variation, and gene flow. Finally, our analysis of parallelism in pathways focuses on the well-characterized flowering time genes in maize and assesses signatures of polygenic adaptation for this trait.

## Results

### Description of samples and data

Four independent highland populations (the Southwestern US, Mexico, Guatemala and the Andes) and two reference lowland populations (Mexico and South America) were sampled to investigate highland adaptation in maize landraces (fig. 1). SNPs called from high-depth whole genome re-sequencing data were obtained from our previous study [38]. After filtering (see Materials and Methods), a total of 1,567,351 SNPs across 35 samples were retained for further analyses. We also characterized how environmental factors varied across our highland and lowland populations. A principal component analysis of 19 bioclim environmental variables (<http://www.worldclim.org/>) revealed that highland and lowland accessions were differentiated along PC1 (comprising 51.8% of variation); while PC2 (comprising 22.8% of variation) reflected latitudinal differences across highland populations (supplementary fig. S1, Supplementary Material online). By plotting the loadings from the 19 bioclim variables, we found PC1 was primarily polarized by temperature seasonality and PC2 by precipitation (supplementary fig. S1, Supplementary Material online).

### Convergence through parallelism at the SNP level

In order to detect loci targeted by selection in highland populations, we utilized the Population Branch Excess (*PBE*) statistic [6], which characterizes changes

in allele frequencies in a focal population relative to two independent “outgroup” populations [6]. *PBE* values were calculated for each SNP site and the top 5% were considered outliers and potential targets of selection.

To gauge the extent of convergence across our highland populations at the SNP level, we evaluated whether shared *PBE* outliers in pairs of populations and larger groups were more common than expected by chance. Significant overlap was observed in all pairs of highland populations ( $p \approx 0$ , hypergeometric statistic test; supplementary table S1; supplementary table S2, Supplementary Material online). The Mexican Highland population showed the most substantial overlap in outlier SNPs with the Guatemalan Highlands (7.5-fold enrichment) and the Southwestern US Highlands (6-fold enrichment). While significantly more than expected, shared outliers between the Andes and all other highland populations were not nearly as extensive (between 2- and 4-fold enrichment), consistent with reduced parallelism at the SNP level between these isolated regions. The intersection of outlier SNPs across all highland populations was also significantly enriched (supplementary fig. S2, Supplementary Material online). Enrichment of overlapping selected sites was also confirmed in a more stringent tail of the top 1% of *PBE* values with qualitatively similar results (supplementary fig. S3, Supplementary Material online). While we did observe significant enrichment for parallelism at the SNP level, it is important to note that the majority of selected SNPs (from 59.7% to 68.1%) showed signatures of adaptation in a single highland population.

Our analysis of convergence at the SNP level using *PBE* revealed some particularly compelling candidate loci. For example, a SNP within PIF3.1 (phytochrome-interacting factor) exhibited the highest *PBE* value in the Mexican Highland population and was also detected as a target of selection in all other highland populations (fig. 2A). A non-synonymous, derived allele at this locus is fixed across all highland populations (fig. 2C). In addition, SNPs within GRMZM2G078118, a gene included in the jasmonic acid biosynthesis pathway, were detected as outliers in all Mesoamerican highland populations (SW US, Mexican Central Plateau, Guatemalan Highlands), but not the Andes (fig. 2BD).

While our *PBE* analysis reveals that selection targeted many of the same SNPs for adaptation across highland populations, this test alone does not clarify whether the same specific alleles were adaptive. We term outlier SNPs with the same allele elevated to high frequency in highland population pairs “co-directional”, and SNPs with different alleles as “anti-directional”. Predominantly, shared outliers in highland population pairs show co-directional allele frequency shifts (> 95% in all population pairs with the exception of the Southwestern US/Andes comparison, which was 87.6%). These results provide stronger evidence of molecular parallelism at the SNP level (fig. 3A). In contrast, random samples of SNPs that do not show evidence of selection exhibit much reduced signals of co-directional change. While random SNP samples between the Andes and other highland populations show roughly equal proportions of co- and anti-directionality, substantially more co-directional SNPs are observed in comparisons of Mesoamerican populations, perhaps reflecting their more closely shared histories (fig. 3A). However, the strong excess of co-directional changes in candidate SNPs may also be influenced by ancestral frequencies in lowland populations. For example, neutral drift has a greater likelihood of simultaneously fixing the same allele in multiple highland populations if the allele is already at high frequency in the lowlands, or, similarly, removing alleles at low frequency in the lowlands. In order to address this bias, we first approximated the ancestral allele frequencies of outlier SNPs (based on the two-dimensional site frequency spectrum of Mexican lowland maize and the wild relative

*Zea mays* ssp. *parviglumis*) in a subset of the neutral SNP set, and recalculated the number of co-directional and anti-directional SNPs. The same pattern was revealed (supplementary fig. S4, Supplementary Material online). Second, we sub-sampled outlier SNPs with a high minor allele frequency (*i.e.*, between 0.3 and 0.5) and re-calculated the ratio of co-directional versus anti-directional SNPs. We found anti-directional SNPs were slightly increased in this subset, but co-directional SNPs were still far more numerous (fig. 3A). Taken together, the excess of co-directional SNPs in the outlier set was not biased by ancestral allele frequencies and the directionality of allele frequency changes in shared outlier SNPs provides strong evidence of molecular parallelism at the SNP level among highland maize populations.

## The source of shared adaptation among highland populations

To determine whether shared *PBE* outlier SNPs had independent origins in highland populations, we applied DMC, a composite-likelihood-based inference method that distinguishes among possible mechanisms [11]. We controlled for possible effects of linkage disequilibrium by subsampling outliers to one SNP per 2-kb window. First, we examined SNPs selected in all four independent highland populations. We found the vast majority of these common outliers (408 among 428) had the highest composite likelihood under models where there was a single origin of the beneficial allele. Within this subset, 49.1% and 46.3% of outliers had the standing variant model from a source population and the migration model as best fits, respectively (fig. 3B). The standing variant model from a source population (outlined in detail in Appendix A.4 of [11]) specifies that a beneficial allele originates in a single source population where it spreads via gene flow to other adapted populations where it may segregate for a standing time  $t$  prior to the onset of selection. For the common outliers identified under the migration model, the main migration sources were the Mexican (43.9%) and Guatemalan (41.9%) Highland populations. This result suggests that, for many shared outlier SNPs, gene flow among highland populations may have been an important source of adaptive variation.

We further explored the source of shared adaptive variation across pairwise samples of highland populations. As seen in four-wise comparisons, the standing variant model from a source population (49.4% to 69.9%; mean 56.7%) and the migration model (12.7% to 47.4%; mean 36.6%) were most commonly supported (fig. 3B). It is noteworthy that the parallelism level among Mesoamerican highland populations was much stronger than observed between the Andes and other highland populations (fig. 3B). Not only do fewer loci show evidence of selection in pairs of populations including the Andes, but repeatedly selected loci are more likely to have arisen independently. The mean value of an independent source was 6.2% in Andean comparisons, in contrast to a mean of 2.0% within Mesoamerican highland population comparisons (fig. 3B). Previous studies have suggested that migration between the Andes and Mesoamerican populations is unlikely [36]. We therefore further evaluated all repeatedly selected SNPs by calculating the divergence statistic,  $F_{ST}$ , between pairwise highland populations to assess how this varies based on the assigned model. Repeatedly selected SNPs consistent with a migration model exhibited weaker divergence (lower  $F_{ST}$ ;  $p \approx 0$ ; supplementary fig. S5, Supplementary Material online), an additional indicator of gene flow. Moreover, we calculated the  $\hat{f}_{dM}$  statistic (Malinsky et al. 2019) in 10-kb non-overlapping windows across the genome to further assess the signal of gene flow between the Andes and Mesoamerican highland populations. In a (((P1,P2)P3),O) shaped phylogenetic tree, the  $\hat{f}_{dM}$  statistic gives positive values for introgression between P3 and P2 and nega-



tive values for introgression between P3 and P1. We found SNPs consistent with the migration model demonstrated more positive  $\hat{f}_{dM}$  values, providing additional support for gene flow between the Andes and Mesoamerican highland populations (supplementary fig. S6, Supplementary Material online). In summary, while the proportion of dually selected SNPs between the Andes and Mesoamerican highland populations (from 7.8% to 11.2%) was consistent with previous work (7% - 8%) [36], the two most supported source models were standing variation and migration, suggesting a more prominent role for migration than previously hypothesized [36].

## Molecular parallelism at the genic level

We next evaluated the extent of molecular parallelism at the gene level. A gene was classified as an outlier if it contained *PBE* outlier SNPs within the gene or 10 kb upstream or downstream (supplementary table S3, Supplementary Material online). 1651 candidate genes were observed in the Mexican Highlands, among which 360 (21.8%) contained or were nearby outlier SNPs which were shared with at least one other highland population. The percentage dropped to 18.9% in the Guatemalan Highlands, 10.0% in the Southwestern US Highlands and 9.4% in the Andes (supplementary table S4, Supplementary Material online). We utilized the R package SuperExactTest [39] to evaluate the two- to four-degree intersection of outlier genes among highland populations. As observed in results at the SNP level, significant overlap was observed for each comparison (supplementary fig. S7, supplementary fig. S8, Supplementary Material online). In many instances, common outlier genes involved selection on different SNPs across highland populations. For example, 349 (21.1%) of the 1651 selected genes in the Mexican Highlands were in common with other highland populations, but showed evidence of selection on distinct SNPs. The percentage was 18.8% in Guatemala, 24.7% in the Southwestern US, and 18.5% in the Andes (supplementary table S4, Supplementary Material online).

## Constraint of the adaptation target size leads to molecular parallelism

There are two explanations for the extent of molecular parallelism we have observed in genes selected in the highlands [9]. First, it is possible that adaptation is constrained by the number of genes that can affect a trait, effectively placing physiological limits on the routes adaptation can take. Alternatively, many different genes may have the potential to affect the trait, but deleterious pleiotropic effects of variation in many genes may prevent them from playing a role in adaptation. We will refer to these possibilities as “physiological” and “pleiotropic” constraint. Consistent with our analysis above, Yeaman [9]’s  $C_{\chi^2}$  statistic finds strong evidence that selection at the same loci is shared among populations more often than can be explained by chance ( $C_{\chi^2}$  from 19.6 to 40.1 with a mean of 28.05, permutation  $p < 0.0001$ ). To evaluate the role of physiological constraint in parallel adaptation, we tested an alternative null model which assumes the set of genes selected in at least one population represents all genes that can possibly affect the trait. We then tested whether there was still evidence of excessive sharing after accounting for physiological constraint [9]. We found  $C_{\chi^2}$  was less than 0 in all highland pairs except the Mexican and Guatemalan Highlands pair ( $C_{\chi^2} = 1.9, p = 0.025$ ). These results indicate that repeated selection of genes among most highland pairs may simply be due to physiological constraint in the number of loci that can contribute

to variation in a trait. To explore this possibility further, we assessed the prevalence of molecular parallelism in the flowering time pathway, a well-characterized gene network known to be of adaptive importance in highland environments.

## Molecular parallelism within the flowering time pathway

Flowering time in maize is known to be a highly quantitative trait and an example of polygenic adaptation [40, 41, 42, 43]. As such, we may expect that highland adaptation for flowering is the result of subtle, coordinated allele frequency shifts across many loci. In evaluating flowering time genes we tested: 1) the strength of selection on loci associated with flowering time relative to expectations based on drift, and 2) the intersection of selected genes in the pathway in multiple highland regions. To first identify whether flowering time was targeted during highland adaptation, we conducted a genome-wide association study (GWAS) of 29 traits measured in short-day conditions in Ponce, Puerto Rico and Homestead, Florida [44] using a maize association mapping panel comprised of tropical and temperate inbred lines [45]. We then summarized the frequency of multiple alleles affecting a trait of interest with a polygenic score and tested for associations between polygenic score and elevation of origin of our landraces, while controlling for variation in relatedness between samples following Josephs et al. [43]. A significant result implies that allele frequencies that increase trait value are more common at one end of the elevation gradient (either high or low) than expected due to drift. Seven traits showed a significant relationship between their conditional polygenic score and elevation with a false discovery rate  $< 0.1$ . These traits were primarily related to flowering time: Days to Silk, Growing Degree Days to Silk, and Ear Height from '06 Puerto Rico, Growing Degree Days to Silk, Growing Degree Days to Tassel, and Ear Height in '06 Florida, and Growing Degree Days to Silk in '07 Florida (fig. 4).

Given this support for selection on flowering time during highland adaptation, we explored the overlap between our selected genes and known flowering time candidates. First, we utilized a list of 904 maize flowering time candidate genes aggregated by Li et al. [46] (supplementary table S5, Supplementary Material online). Within this list, we found a substantial excess of flowering time genes targeted by selection among groups of two, three, and four highland populations ( $p < 2.87e - 5$ , hypergeometric test; fig. 5A, supplementary table S6, Supplementary Material online). Outlier SNPs within these genes also showed strong co-directional changes in allele frequency across highland populations (91.8% – 100%) (supplementary fig. S9, Supplementary Material online). As observed in our genome-wide analysis, candidate sharing at flowering time loci was stronger among Mesoamerican highland populations than between comparisons with the Andes (fig. 5). We also applied the  $C_{\chi^2}$  statistic to this list of genes and confirmed enriched sharing of selected genes among highland populations (mean  $C_{\chi^2}$  6.618, permutation  $p < 0.0001$ ). Our finding of an excess of shared candidates across regions within a large set of known genes that affect flowering time, suggests physiological constraint alone cannot explain patterns of convergence, but instead that variation at only a subset of genes can affect flowering without deleterious pleiotropic consequences.

We then narrowed our focus to a subset of core flowering time genes [47] with highly characterized functional roles (fig. 5C). Thirty-two of the 39 core flowering time genes were selected in at least one highland population, which indicated over-representation of flowering time genes among all selected genes ( $p = 0.005$ , hypergeometric test). While common outlier genes between the Andes and other highland populations were not significantly enriched in this smaller list ( $p > 0.074$ ,

hypergeometric test; supplementary table S6, Supplementary Material online), enrichment was observed among all other two and three highland-population comparisons ( $p < 0.042$ , hypergeometric test; supplementary table S6, Supplementary Material online). These results were confirmed through calculation of the  $C_{\chi^2}$  statistic (Mesoamerican highland population comparisons, mean  $C_{\chi^2}$  2.32, permutation  $p < 0.01$ ; comparisons to the Andes, mean  $C_{\chi^2}$  1.575, permutation  $p$  from 0.015 to 0.054).

The *PBE* values of outlier SNPs within or in the 10-kb flanking regions of flowering time genes were in the extreme tail of the genome-wide distribution (fig. 5B), with some of the best-characterized flowering time genes in maize clearly targeted by selection in multiple highland populations (fig. 5C). For example, ZCN8 (GRMZM2G179264), which is orthologous to FLOWERING LOCUS T (FT) in *Arabidopsis*, has been shown to play a significant role in maize adaptation to high latitudes [48] and was selected in all four highland populations in our study. Floral transition is regulated by several MADS-box genes, particularly ZMM4 (GRMZM2G032339), which is under selection in all but the Southwestern US Highlands. VGT1 (GRMZM2G700665; vegetative to generative transition 1) was identified as an outlier in all four highland populations, with the same non-synonymous SNP (8:131578990, indicating locus 131578990 on chromosome 8) targeted in the Mexican and Guatemalan Highlands (supplementary fig. S10, Supplementary Material online). In contrast, in the Southwestern US Highlands, the selected SNP (8:131580179) in VGT1 was located in the 3' UTR region and, in the Andes, the selected SNP (8:131579463) was a synonymous mutation near the 3' UTR (supplementary fig. S10A,B,C, Supplementary Material online). Two CONSTANS (CO) genes, CONZ1 and ZmCCT, were selected in both the Mexican and Guatemalan Highland populations. We further found Gigantea2 (GRMZM5G844173) was selected in all but the Guatemalan Highlands and FKF2 (GRMZM2G106363) in all but the Mexican Highlands. Among circadian clock genes, LUX (GRMZM2G067702)(supplementary fig. S10D, Supplementary Material online) was selected in all four highland populations and ZmPRR59 was an outlier in the three Mesoamerican highland populations. In addition, the zmZTL $\alpha$  - F box protein ZEITLUPE (GRMZM2G113244), RVE2 (GRMZM2G145041) and PRR5 (GRMZM2G179024) were selected in all four highland populations. Finally, the light transduction gene PHYA1 (GRMZM2G157727) was targeted by selection in all but the Southwestern US highland populations.

## Discussion

The prevalence of parallel evolution is a long-standing question in the field of evolution. Through a comprehensive evaluation of parallel adaptation among independent maize highland populations, we found that, while most adaptation is independent, parallelism is more common than expected. We observed a particularly high level of parallelism among Mesoamerican highland populations relative to comparisons including Andean maize. The vast majority of parallel adaptive alleles we discovered have risen in frequency in the highlands from migration and standing variation, and only a small percentage represent *de-novo* mutations. However, the proportion of repeatedly selected SNPs from *de-novo* mutations is highest in pairs of populations that include the Andes. Our analyses further reveal that adaptive routes from genotype to phenotype have been canalized by both physiological (*i.e.*, few genes contributing to a trait) and pleiotropic (*i.e.*, a subset of causal genes can contribute to trait variation without negative pleiotropic consequences)



processes. For example, we observe that the flowering time pathway is saturated with selected genes potentially contributing to highland adaptation, with many of these loci showing evidence of selection in multiple highland regions, particularly in Mesoamerica.

Approximately one quarter (mean 24.0%) of highland adaptation SNPs showed a pattern of molecular parallelism in Mesoamerica. A smaller percentage also showed this pattern in Andean comparisons (mean 10.7%), but this represents a slightly higher proportion than previously reported (7% - 8% between the Andes and the Mexican Highlands) [36]. We have also expanded upon previous studies of parallel highland maize adaptation [36] in a number of ways. Our study has broadened the geographic scope of sampling for highland maize populations and is the first such study using high-depth, whole-genome resequencing data, which allowed us to attain a more complete picture of parallelism. Additionally, our use of the PBE statistic allowed for detection of selection specifically in the highlands and reduced the extent of false positives, a limitation of studies that focus solely on population differentiation outliers. Our polygenic adaptation analyses also helped to clarify specific phenotypes targeted by selection during adaptation, showing clear evidence of selection on flowering time. We have also uncovered a pervasive role of migration in transferring adaptive alleles across highland regions. In fact, our analyses revealed a moderate proportion of parallel adaptation through migration even between the Andes and other highland populations, in contrast to previous studies that appeared to exclude this possibility [36] based on simulations. This finding is consistent with a recent study which showed that some traits, for example flowering time, may not demonstrate a GxE effect [42], allowing adaptive alleles to move between regions through a matrix of habitat (*e.g.*, the lowlands) where they do not clearly confer an adaptive advantage.

The extensive parallelism in maize highland adaptation has likely been affected by multiple factors and is consistent with expectations based on population genetic theory. First, a high level of genetic diversity in the species provides substantial standing genetic variation and potentially beneficial mutations. Similarly, Zhao et al. [49] showed that the more genetically diverse species *Drosophila hydei* contained more adaptive alleles than *D. melanogaster*. Second, the recent timing of expansion and divergence in maize has likely affected the extent of parallelism. The probability of parallelism will decrease when populations diverge over a greater period of time, as less shared standing variation would be expected to be the source of adaptation. Increased divergence time would also allow for novel beneficial mutations to arise, potentially in distinct genes and pathways [50]. For example, Preite et al. [51] demonstrated a moderate level of molecular parallelism in adaptation to calamine metalliferous soils within *Arabidopsis* species, but less parallelism between species. In addition, a recent study [52] discovered that the degree of molecular parallelism decreases with increasing divergence between lineages in *Arabidopsis* alpine species. Third, population genetic modeling predicts that traits with pleiotropic constraint are more likely to demonstrate molecular parallelism [53]. Despite the quantitative nature of flowering time in maize, the  $C_{\chi^2}$  statistic still showed evidence of genetic constraint and convergence even after limiting our null to genes in the known flowering time pathway. The genetic parallelism could also be attributed to the high effective population size of the species and extensive migration between highland regions. Similarly, a recent study found molecular parallelism for cold tolerance between distantly related species (lodgepole pine and interior spruce), indicating certain key genes playing crucial roles in cold adaptation [54]. Finally, theory predicts that gene flow may have contrasting impacts on parallel adaptation. Gene

flow between differentially adapted populations may introduce maladaptive alleles, counteracting the effects of local selection [53]. For example, Holliday et al. [55] found islands of divergence in *Populus trichocarpa* populations across altitudinal clines and proposed that coadapted genes in strong linkage may buffer against genetic introgression from maladapted individuals. Our results, consistent with earlier findings for maize in the highlands of Mexico [36, 37], argue instead for an important role for adaptive introgression in increasing the likelihood of parallelism. Similarly, adaptive introgression at the EPAS1 locus has been shown to underlie convergent highland adaptation in both humans [56] and dogs [57].

In summary, the emerging story of parallel highland adaptation in maize is consistent with existing theory. Gene flow is common across maize populations, but does not appear to swamp locally adaptive alleles. Rather, highland adaptation appears to have initially been facilitated in the Central Plateau of Mexico by introgression from the locally adapted wild species, *mexicana* [37]. Subsequently, gene flow from the Mexican Highlands appears to have contributed to highland adaptation in other regions of the Americas [38], consistent with our finding of migration as the major source of molecular parallelism. Novel alleles from *mexicana* augmented the extensive standing variation already found within the species [26], providing a pool of adaptive variants that appear to have independently risen in frequency in multiple highland regions.

Similar adaptive processes (*e.g.*, gene flow with newly encountered, locally adapted populations and species, repeated selection on standing variation in independent but similar habitats) likely occurred as other crops expanded from their centers of origin [58]. However, crops that were domesticated from wild species with small effective population sizes, with more pronounced domestication bottlenecks, or without widespread, locally adapted congeners may have lacked the adaptive potential necessary to achieve the broad and varied distribution of maize. The success of invasive species has also been linked to hybridization [59] and high levels of standing variation [60]. A thorough understanding of adaptive processes during rapid expansion across diverse habitats may therefore inform invasive species mitigation strategies, assisted migration in the face of climate change, and crop breeding for tolerance of extreme environmental conditions. Highland adaptation in maize is a clear example in which adaptation has drawn substantially from a shared pool of variants, with the same alleles, genes, and pathways contributing to the evolutionary success of the species across continents and millennia.

## Materials and Methods

### Samples and Data

Samples and SNPs were from our previous publication [38]. SNPs were removed when located in a known inversion, large introgression regions from *mexicana* and centromeric regions, and later filtered to retain only those with sequencing depth greater than 10X and minor allele frequency greater than 0.05 (supplementary fig. S11, Supplementary Material online). Nineteen environmental variables were obtained from the WorldClim data set (<http://www.worldclim.org/>) and were utilized to perform a Principal Component Analysis (PCA) with the vegan package in R [61]. We employed Variant Effect Predictor (VEP) from Ensembl [62] to further annotate genomic location and functional consequences of SNPs.

## PBS and PBE calculation

Compared to the commonly used statistic to estimate levels of population differentiation -  $F_{ST}$ ,  $PBS$  calculates allele frequency change specific to one focal population by adding pairwise  $F_{ST}$  values involving a third "outgroup" population [6].

$$PBS = \frac{T^{HL_1} + T^{HL_2} - T^{L_1L_2}}{2} \quad (1)$$

$$T = -\log(1 - F_{ST}) \quad (2)$$

$PBE$  is a recently described derivative of  $PBS$ , which overcomes the limitation of  $PBS$  values being high when all populations have long branches.  $PBE$  instead quantifies the difference of the observed and the expected  $PBS$  values [63].

$$PBE = PBS_{obs} - PBS_{exp} = PBS - [T^{L_1L_2} * \frac{PBS_{med}}{T_{med}^{L_1L_2}}] \quad (3)$$

Here,  $T^{L_1L_2}$  quantifies genetic differentiation of the two non-focal populations.  $PBE$  is expected to be strongly positive when selection acts specifically on the focal population.

We calculated  $PBE$  for filtered SNPs among one highland population, its corresponding lowland population (the Mexican Lowland population was paired with the Mesoamerican Highland populations; the South American Lowland population was paired with the Andes population) and the *parviglumis* Palmar Chico population (supplementary fig. S11, Supplementary Material online).

Pairwise  $F_{ST}$  values were calculated in vcfTools [64] and custom R scripts were used to compute  $PBS$  and  $PBE$  values of the focal highland population. SNPs with  $PBE$  values higher than the 95% quantile of its distribution were regarded as outliers and qualitative results were confirmed in the 99% quantile. The R package "SuperExactTest" [39] was utilized to evaluate if the overlap of outlier SNPs between pairwise highland populations was enriched.

In addition, we also evaluated if the same allele is elevated to high frequency among outlier SNPs between pairwise highland populations (supplementary fig. S12, Supplementary Material online). If there is no allele frequency change between the highland and its lowland counterpart, we categorized the SNPs as "non-directional". When a SNP was non-directional in both highland populations being compared, it was removed from the analyses. When the same allele was at elevated frequency in both highland populations compared to their lowland counterparts, the SNP was categorized as "co-directional". When different alleles were at high frequency across two highland populations, the SNP was categorized as "anti-directional". We then compared the proportion of co-directional SNPs between the outlier and neutral SNPs.

In order to exclude the bias of genetic drift on the directionality of changes in allele frequency, we approximated the two-dimensional site frequency spectrum (2dsfs) of outlier SNPs using a subset of the neutral SNP set. We divided the non-reference allele frequency in Mexican Lowland maize and *parviglumis* into ten equal bins in both the outlier and neutral SNP set. Then we matched neutral SNPs with the same ancestral 2dsfs (representing the combination of allele frequencies in both the Mexican Lowland and *parviglumis* populations) with the outlier SNP set and sampled the same amount of SNPs in each allele frequency bin. Last, we tabulated the directionality of the randomly sampled neutral SNPs to check how ancestral allele frequency influenced the proportion of co-directional SNPs.

## Composite likelihood calculations to distinguish the mode of parallelism

Following the method developed and applied in [11], we calculated the composite likelihood under four models for the outlier regions shared among all four highland populations, after thinning for linkage. The four models tested were as follows: 1) neutrality (no selection), 2) independent mutations of the beneficial allele in four highland populations, 3) a single origin of the beneficial allele in one single highland population and spread via gene flow into the other population during the sweep, 4) standing variation present in the ancestor of the four highland populations and spread via gene flow into other populations some time  $t$  before selection occurred. The details of these four models can be found in [11]. For each outlier, we obtained SNPs in the surrounding 20-kb window and calculated the composite likelihood under each of the four models for this dataset. We compared composite likelihoods under the models for each outlier and obtained maximum composite-likelihood-parameter estimates, specifically for the standing time  $t$  in the fourth model, to assess the timing of gene flow. Similarly, we also evaluated the composite likelihood values under four models for each set of dually selected SNPs in pairs of highland populations.

## Polygenic adaptation for quantitative traits

We followed the methods developed by Berg et al. [65] and Josephs et al. [66] to detect polygenic adaptation in the highland populations. Briefly, these methods detect adaptive divergence for a trait by 1) finding loci associated with that trait in a GWAS, 2) summarizing the allele frequencies and effect sizes at these loci in the populations of interest using a polygenic score, and 3) testing to see if the association between these polygenic scores and an environmental character (in this case, elevation) are greater than could be explained by drift.

To identify loci associated with traits that could be under selection, we conducted genome-wide association studies using a maize panel developed for GWAS, often referred to as “the 282” or “the Major Goodman panel” [45]. Single nucleotide polymorphisms (SNPs) for 263 individuals in the Major Goodman panel were called from whole genome sequencing data from Bukowski et al. [67], removing individuals with genotype calls for <70% of polymorphic sites. Of all the SNPs called in Bukowski et al. [67], we only used those that were also polymorphic in the maize landraces and that had a MAF > 0.01 and were missing data for <5 % of individuals, leaving us with approximately 5 million SNPs per test. We used trait measurements made in three short-day common garden experiments, from Florida in 2006 and 2007 and Puerto Rico in 2006, from Hung et al. [44] where there were data available for at least 80% of the 263 individuals, leaving us with 29 trait-environment combinations. We conducted GWAS using GEMMA with default parameters [68], with a standardized kinship matrix to control for population structure.

We used the GWAS hits to generate polygenic scores for each individual in the landrace panel using all loci with a p value < 0.1 that had been pruned down to the strongest hit per 1 cM window using a linkage map from Ogut et al. [69], leaving us with an average of 216 SNPs per trait (range = 147-289). If there are  $M$  loci associated with a trait, each with effect size  $\alpha_m$ ,  $p_{im}$  is the allele frequency of the  $m^{th}$  allele in line  $i$ , which will either be 0, 0.5, or 1, then we can calculate the

polygenic score for the  $i^{th}$  line as  $Z_i$  where

$$Z_i = 2 \sum_{m=1}^M \alpha_m p_{im} \quad (4)$$

Shared population structure between the GWAS panel and the test panel can lead to false positive signals of selection [66]. To test for shared structure between the GWAS panel used here and the landraces, we constructed a joint kinship matrix between all samples in both panels, following the same methods used in Josephs et al. [66]. Plotting the first two eigenvectors of this joint kinship matrix (the principal components) show that there is likely shared population structure between the landraces and the GWAS panel (supplementary fig. S13A, Supplementary Material online). We constructed a conditional kinship matrix that estimates the amount of relatedness between landraces conditional on relatedness captured in the GWAS panel, and found that this new conditional matrix explains less variation than the original matrix made from landraces alone (supplementary fig. S13B, Supplementary Material online). This finding is further evidence that there is shared population structure between the two panels.

In light of this shared structure, we conducted a conditional test, following Josephs et al. [66]. Let  $\vec{Z}_1$  be the vector of  $Z_i$  for the 31 highland and lowland landraces discussed in this paper and  $\vec{Z}_2$  is the vector of  $Z_i$  for the individuals used in the GWAS. We model the combined vector of polygenic scores in both panels as

$$\begin{pmatrix} \vec{Z}_1 \\ \vec{Z}_2 \end{pmatrix} \sim MVN \left( \begin{pmatrix} \mu \\ \mu \end{pmatrix}, V_A \begin{pmatrix} K_{11} & K_{12} \\ K_{12}^T & K_{22} \end{pmatrix} \right). \quad (5)$$

where,  $\mu$  is the mean of the combined vector  $[X_1, X_2]$ ,  $K_{11}$  and  $K_{22}$  are the kinship matrices of the genotyping and GWAS panels, and  $K_{12}$  is the set of relatedness coefficients between lines in the genotyping panel (rows) and GWAS panel (columns).

The conditional multivariate null model for our polygenic scores in the landraces conditional on the GWAS panel is then

$$\vec{Z}_1 | \vec{Z}_2 \sim MVN(\vec{\mu}', 2V_A K'), \quad (6)$$

where  $\vec{\mu}'$  is a vector of conditional means with an entry for each sample in the genotyping panel:

$$\vec{\mu}' = \mu + K_{12} K_{22}^{-1} (\vec{Z}_2 - \mu) \quad (7)$$

and  $K'$  is the relatedness matrix for the genotyping panel conditional on the matrix of the GWAS panel,

$$K' = K_{11} - K_{12} K_{22}^{-1} K_{12}^T. \quad (8)$$

To test for selection, we calculate the difference between observed polygenic scores and conditional means as  $\vec{Z}' = \vec{Z}_1 - \vec{\mu}'$ . Higher values of  $\vec{Z}'$  indicate that the observed polygenic score in a landrace individual is greater than would be expected based on relatedness between that landrace individual and individuals in the GWAS panel.

After calculating  $\vec{Z}'$  for 29 environment-trait combinations, we test for a correlation between  $\vec{Z}'$  and the elevation of origin of each landrace beyond what would be expected due to neutral drift using the methods described in [65]. We transformed both  $\vec{Z}'$  and the mean-centered vector of elevations for each landrace by the Cholesky decomposition of the kinship matrix ( $C$ ) using the following equation (shown here for  $\vec{Z}'$ )

$$\vec{X} = \frac{1}{\sqrt{2V_A}} C^{-1}(\vec{Z}') \quad (9)$$



We estimate  $V_A$  using the allele frequencies and effect sizes of the GWAS loci, using the following formula. If  $\bar{p}_m$  is the allele frequency of the  $m^{th}$  locus across all individuals and  $\alpha_m$  is the effect size of that locus,

$$V_A = \sum_{m=1}^M \alpha_m^2 \bar{p}_m (1 - \bar{p}_m) \quad (10)$$

We then test for a linear relationship between  $\vec{X}$  and the similarly transformed vector of elevations using the `lm` function in R [70] and control for the 29 tests done using Qvalue [71].

## Acknowledgments

This study was supported by Shenzhen Science and Technology Program (Grant No. KQTD2016113010482651), special funds for science technology innovation and industrial development of Shenzhen Dapeng New District (Grant Numbers RC201901-05 and PT201901-19), the US Department of Agriculture (USDA #2009-65300-05668), the National Science Foundation (NSF IOS #1546719; NSF IOS #1523733 to EB), USDA Hatch project (CA-D-PLS-2066-H), and startup funds from Iowa State University. We thank Graham Coop and members of the Hufford Lab for helpful comments and suggestions.

## References

- [1] Kalina TJ Davies, Nigel C Bennett, Chris G Faulkes, and Stephen J Rossiter. Limited evidence for parallel molecular adaptations associated with the subterranean niche in mammals: a comparative study of three superorders. *Mol. Biol. Evol.*, 35(10):2544–2559, 2018.
- [2] Joel M Alves, Miguel Carneiro, Jade Y Cheng, Ana Lemos de Matos, Mas-mudur M Rahman, Liisa Loog, Paula F Campos, Nathan Wales, Anders Eriksson, Andrea Manica, et al. Parallel adaptation of rabbit populations to myxoma virus. *Science*, 363(6433):1319–1326, 2019.
- [3] Susanne Dobler, Safaa Dalla, Vera Wagschal, and Anurag A Agrawal. Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the na, k-atpase. *PNAS*, 109(32):13040–13045, 2012.
- [4] Ying Zhen, Matthew L Aardema, Edgar M Medina, Molly Schumer, and Peter Andolfatto. Parallel molecular evolution in an herbivore community. *Science*, 337(6102):1634–1637, 2012.
- [5] Li Yang, Hui-Na Wang, Xing-Hui Hou, Yu-Pan Zou, Ting-Shen Han, Xiao-Min Niu, Jie Zhang, Zhong Zhao, Marco Todesco, Sureshkumar Balasubramanian, et al. Parallel evolution of common allelic variants confers flowering diversity in *capsella rubella*. *Plant Cell*, 30(6):1322–1336, 2018.
- [6] John E. Pool, Dylan T. Braun, and Justin B. Lack. Parallel Evolution of Cold Tolerance within *Drosophila melanogaster*. *Mol. Biol. Evol.*, 34(2):349–360, 10 2016.

- [7] Erica Bree Rosenblum, Christine E Parent, and Erin E Brandt. The molecular basis of phenotypic convergence. *Annu. Rev. Ecol. Evol. Syst.*, 45:203–226, 2014.
- [8] Jay F Storz. Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.*, 17(4):239, 2016.
- [9] Sam Yeaman, Aleeza C Gerstein, Kathryn A Hodgins, and Michael C Whitlock. Quantifying how constraints limit the diversity of viable routes to adaptation. *PLoS Genet.*, 14:e1007717, 2018.
- [10] H. Allen Orr. The probability of parallel adaptation. *Evolution*, 59(1):216–220, 2005.
- [11] Kristin M Lee and Graham Coop. Distinguishing among modes of convergent adaptation using population genomic data. *Genetics*, 207(4):1591–1619, 2017.
- [12] Pascal-Antoine Christin, Nicolas Salamin, Vincent Savolainen, Melvin R Duvall, and Guillaume Besnard. C4 photosynthesis evolved in grasses via parallel adaptive genetic changes. *Curr. Biol.*, 17(14):1241–1247, 2007.
- [13] Guillaume Besnard, A Muthama Muasya, Flavien Russier, Eric H Roalson, Nicolas Salamin, and Pascal-Antoine Christin. Phylogenomics of c4 photosynthesis in sedges (cyperaceae): multiple appearances and genetic convergence. *Mol. Biol. Evol.*, 26(8):1909–1919, 2009.
- [14] Kanchon K Dasmahapatra, James R Walters, Adriana D Briscoe, John W Davey, Annabel Whibley, Nicola J Nadeau, Aleksey V Zimin, Daniel ST Hughes, Laura C Ferguson, Simon H Martin, et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405):94, 2012.
- [15] Nathaniel B Edelman, Paul B Frandsen, Michael Miyagi, Bernardo Clavijo, John Davey, Rebecca B Dikow, Gonzalo García-Accinelli, Steven M Van Belleghem, Nick Patterson, Daniel E Neafsey, et al. Genomic architecture and introgression shape a butterfly radiation. *Science*, 366(6465):594–599, 2019.
- [16] Yoshihiro Matsuoka, Yves Vigouroux, Major M. Goodman, Jesus Sanchez G., Edward Buckler, and John Doebley. A single domestication for maize shown by multilocus microsatellite genotyping. *PNAS*, 99:6080–6084, 2002.
- [17] Dolores R Piperno, J Enrique Moreno, José Iriarte, Irene Holst, Matthew Lachniet, John G Jones, Anthony J Ranere, and Ronald Castanzo. Late pleistocene and holocene environmental history of the iguala valley, central balsas watershed of mexico. *PNAS*, 104(29):11874–11881, 2007.
- [18] Dolores R Piperno, Anthony J Ranere, Irene Holst, Jose Iriarte, and Ruth Dickau. Starch grain and phytolith evidence for early ninth millennium bp maize from the central balsas river valley, mexico. *PNAS*, 106(13):5019–5024, 2009.
- [19] Ruth Dickau, Anthony J. Ranere, and Richard G. Cooke. Starch grain evidence for the preceramic dispersals of maize and root crops into tropical dry and humid forests of panama. *PNAS*, 104(9):3651–3656, 2007.

- [20] Alexander Grobman, Duccio Bonavia, Tom D. Dillehay, Dolores R. Piperno, José Iriarte, and Irene Holst. Preceramic maize from paredones and huaca prieta, peru. *PNAS*, 109(5):1755–1759, 2012.
- [21] M.B. Bush, A. Correa-Metrio, C.H. McMichael, S. Sully, C.R. Shadik, B.G. Valencia, T. Guilderson, M. Steinitz-Kannan, and J.T. Overpeck. A 6900-year history of landscape modification by humans in lowland amazonia. *Quat. Sci. Rev.*, 141:52 – 64, 2016.
- [22] D. R. Piperno and K. V. Flannery. The earliest archaeological maize (*zea mays* l.) from highland mexico: New accelerator mass spectrometry dates and their implications. *PNAS*, 98(4):2101–2103, 2001.
- [23] William L Merrill, Robert J Hard, Jonathan B Mabry, Gayle J Fritz, Karen R Adams, John R Roney, and Art C MacWilliams. The diffusion of maize to the southwestern united states and its impact. *PNAS*, 106(50):21019–21026, 2009.
- [24] J. Stephen Athens, Jerome V. Ward, Deborah M. Pearsall, Karol Chandler-Ezell, Dean W. Blinn, and Alex E. Morrison. Early prehistoric maize in northern highland ecuador. *Lat. Am. Antiq.*, 27(1):321, 2016.
- [25] Thomas E. Emerson, Kristin M. Hedman, Mary L. Simon, Mathew A. Fort, and Kelsey E. Witt. Isotopic confirmation of the timing and intensity of maize consumption in greater cahokia. *Am. Antiq.*, 85(2):241262, 2020.
- [26] Matthew B Hufford, Xun Xu, Joost Van Heerwaarden, Tanja Pyhäjärvi, Jer-Ming Chia, Reed A Cartwright, Robert J Elshire, Jeffrey C Glaubitz, Kate E Guill, Shawn M Kaeppler, et al. Comparative population genomics of maize domestication and improvement. *Nat. Genet.*, 44:808–811, 2012.
- [27] Rute R da Fonseca, Bruce D Smith, Nathan Wales, Enrico Cappellini, Pontus Skoglund, Matteo Fumagalli, José Alfredo Samaniego, Christian Carøe, María C Ávila-Arcos, David E Hufnagel, et al. The origin and evolution of maize in the american southwest. *Nat. Plants*, 1:14003, 2015.
- [28] Baobao Wang, Zechuan Lin, Xin Li, Yongping Zhao, Binbin Zhao, Guangxia Wu, Xiaojing Ma, Hai Wang, Yurong Xie, Quanquan Li, et al. Genome-wide selection and genetic improvement during modern maize breeding. *Nat. Genet.*, pages 1–7, 2020.
- [29] Jean-Tristan Brandenburg, Tristan Mary-Huard, Guillem Rigau, Sarah J Hearne, Hélène Corti, Johann Joets, Clementine Vitte, Alain Charcosset, Stéphane D Nicolas, and Maud I Tenaillon. Independent introductions and admixtures have contributed to adaptation of european maize and its american counterparts. *PLoS genet.*, 13(3):e1006666, 2017.
- [30] Christian Korner. The use of ‘altitude’ in ecological research. *Trends Ecol. Evol.*, 22(11):569 – 574, 2007.
- [31] Barry H. Lomax, Wesley T. Fraser, Guy Harrington, Stephen Blackmore, Mark A. Sephton, and Nigel B.W. Harris. A novel palaeoaltimetry proxy based on spore and pollen wall chemistry. *Earth Planet. Sci. Lett.*, 353-354:22 – 28, 2012.

- [32] Nick Lauter, Charles Gustus, Anna Westerbergh, and John Doebley. The inheritance and evolution of leaf pigmentation and pubescence in teosinte. *Genetics*, 167(4):1949–1959, 2004.
- [33] HA Eagles and JE Lothrop. Highland maize from central mexicoits origin, characteristics, and use in breeding programs. *Crop Sci.*, 34(1):11–19, 1994.
- [34] C Jiang, GO Edmeades, I Armstead, HR Lafitte, MD Hayward, and D Hoisington. Genetic analysis of adaptation differences between highland and lowland tropical maize using molecular markers. *Theor. Appl. Genet.*, 99(7-8):1106–1119, 1999.
- [35] JG Rodriguez, G Sánchez, BM Baltazar, LL De la Cruz, F Santacruz-Ruvalcaba, PJ Ron, and JB Schoper. Characterization of floral morphology and synchrony among zea species in mexico. *Maydica*, 2006.
- [36] Shohei Takuno, Peter Ralph, Kelly Swarts, Rob J Elshire, Jeffrey C Glaubitz, Edward S Buckler, Matthew B Hufford, and Jeffrey Ross-Ibarra. Independent molecular basis of convergent highland adaptation in maize. *Genetics*, 200(4):1297–1312, 2015.
- [37] Matthew B Hufford, Pesach Lubinsky, Tanja Pyhäjärvi, Michael T Devenogeno, Norman C Ellstrand, and Jeffrey Ross-Ibarra. The genomic signature of crop-wild introgression in maize. *PLoS Genet.*, 9(5):e1003477, 2013.
- [38] Li Wang, Timothy M Beissinger, Anne Lorant, Claudia Ross-Ibarra, Jeffrey Ross-Ibarra, and Matthew B Hufford. The interplay of demography and selection during maize domestication and expansion. *Genome Biol.*, 18(1):215, 2017.
- [39] Minghui Wang, Yongzhong Zhao, and Bin Zhang. *SuperExactTest: Exact Test and Visualization of Multi-Set Intersections*, 2018. R package version 1.0.0.
- [40] Edward S Buckler, James B Holland, Peter J Bradbury, Charlotte B Acharya, Patrick J Brown, Chris Browne, Elhan Ersoz, Sherry Flint-Garcia, Arturo Garcia, Jeffrey C Glaubitz, et al. The genetic architecture of maize flowering time. *Science*, 325(5941):714–718, 2009.
- [41] J Alberto Romero Navarro, Martha Willcox, Juan Burgueño, Cinta Romain, Kelly Swarts, Samuel Trachsel, Ernesto Preciado, Arturo Terron, Humberto Vallejo Delgado, Victor Vidal, et al. A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nat. Genet.*, 49(3):476–480, 2017.
- [42] Daniel J Gates, Dan Runcie, Garret M Janzen, Alberto Romero Navarro, Martha Willcox, Kai Sonder, Samantha Snodgrass, Fausto Rodríguez-Zapata, Ruairidh JH Sawers, Rubén Rellán-Álvarez, et al. Single-gene resolution of locally adaptive genetic variation in mexican maize. *bioRxiv*, page 706739, 2019.
- [43] Emily B Josephs, Jeremy J Berg, Jeffrey Ross-Ibarra, and Graham Coop. Detecting adaptive differentiation in structured populations with genomic data and common gardens. *Genetics*, 211(3):989–1004, 2019.

- [44] HY Hung, Christopher Browne, Katherine Guill, Nathan Coles, Magen Eller, Arturo Garcia, Nicholas Lepak, S Melia-Hancock, M Oropeza-Rosas, S Salvo, et al. The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity*, 108(5):490, 2012.
- [45] Sherry A Flint-Garcia, Anne-Céline Thuillet, Jianming Yu, Gael Pressoir, Susan M Romero, Sharon E Mitchell, John Doebley, Stephen Kresovich, Major M Goodman, and Edward S Buckler. Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.*, 44(6):1054–1064, 2005.
- [46] Yong-xiang Li, Chunhui Li, Peter J Bradbury, Xiaolei Liu, Fei Lu, Cinta M Romay, Jeffrey C Glaubitz, Xun Wu, Bo Peng, Yunsu Shi, et al. Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. *Plant J.*, 86(5):391–402, 2016.
- [47] Zhanshan Dong, Olga Danilevskaya, Tabare Abadie, Carlos Messina, Nathan Coles, and Mark Cooper. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. *PLoS One*, 7(8):e43450, 2012.
- [48] Li Guo, Xuehan Wang, Min Zhao, Cheng Huang, Cong Li, Dan Li, Chin Jian Yang, Alessandra M York, Wei Xue, Guanghui Xu, et al. Stepwise cis-regulatory changes in *zcn8* contribute to maize flowering-time adaptation. *Curr. Biol.*, 28(18):3005–3015, 2018.
- [49] Li Zhao and David J Begun. Genomics of parallel adaptation at two timescales in drosophila. *PLoS Genet.*, 13(10):e1007016, 2017.
- [50] Gina L Conte, Matthew E Arnegard, Catherine L Peichel, and Dolph Schluter. The probability of genetic parallelism and convergence in natural populations. *P. Roy. Soc. B. - Biol. Sci.*, 279(1749):5039–5047, 2012.
- [51] Veronica Preite, Christian Sailer, Lara Syllwasschy, Sian Bray, Hassan Ahmadi, Ute Krämer, and Levi Yant. Convergent evolution in *arabidopsis halleri* and *arabidopsis arenosa* on calamine metalliferous soils. *Philos. Trans. R. SOC. B*, 374(1777):20180243, 2019.
- [52] Magdalena Bohutinska, Jakub Vlcek, Sivan Yair, Benjamin Leanen, Veronika Konecna, Marco Fracassetti, Tanja Slotte, and Filip Kolar. Genomic basis of parallel adaptation varies with divergence in *arabidopsis* and its relatives. *bioRxiv*, 2020.
- [53] Sam Yeaman. Local adaptation by alleles of small effect. *Am. Nat.*, 186(S1):S74–S89, 2015.
- [54] Sam Yeaman, Kathryn A Hodgins, Katie E Lotterhos, Haktan Suren, Simon Nadeau, Jon C Degner, Kristin A Nurkowski, Pia Smets, Tongli Wang, Laura K Gray, et al. Convergent local adaptation to climate in distantly related conifers. *Science*, 353(6306):1431–1433, 2016.



- [55] Jason A Holliday, Lecong Zhou, Rajesh Bawa, Man Zhang, and Regis W Oubida. Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in *p. opulus trichocarpa*. *New phytol.*, 209(3):1240–1251, 2016.
- [56] Emilia Huerta-Sánchez, Xin Jin, Zhuoma Bianba, Benjamin M Peter, Nicolas Vinckenbosch, Yu Liang, Xin Yi, Mingze He, Mehmet Somel, Peixiang Ni, et al. Altitude adaptation in tibetans caused by introgression of denisovan-like dna. *Nature*, 512(7513):194–197, 2014.
- [57] Yan Li, Dong-Dong Wu, Adam R Boyko, Guo-Dong Wang, Shi-Fang Wu, David M Irwin, and Ya-Ping Zhang. Population variation revealed high-altitude adaptation of tibetan mastiffs. *Mol. Biol. Evol.*, 31(5):1200–1205, 2014.
- [58] Garrett M Janzen, Li Wang, and Matthew B Hufford. The extent of adaptive wild introgression in crops. *New phytol.*, 221(3):1279–1288, 2019.
- [59] Kurt C Heim, Thomas E McMahon, Brian D Ertel, and Todd M Koel. Leveraging public harvest to reduce invasive hybridization in yellowstone national park: Field identification and harvest of cutthroat x rainbow trout hybrids. *Biol. invasions*, 2020.
- [60] Makiko Mimura, K Ono, Koichi Goka, and T Hara. Standing variation boosted by multiple sources of introduction contributes to the success of the introduced species, lotus corniculatus. *Biol. invasions*, 15(12):2743–2754, 2013.
- [61] Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O’Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, and Helene Wagner. *vegan: Community Ecology Package*, 2016. R package version 2.4-0.
- [62] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biol.*, 17(1):122, 2016.
- [63] Amir Yassin, Vincent Debat, Héloïse Bastide, Nelly Gidaszewski, Jean R David, and John E Pool. Recurrent specialization on a toxic fruit in an island drosophila population. *PNAS*, 113(17):4771–4776, 2016.
- [64] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.
- [65] Jeremy J Berg and Graham Coop. A population genetic signal of polygenic adaptation. *PLoS Genet.*, 10(8):e1004412, 2014.
- [66] Emily B. Josephs. Determining the evolutionary forces shaping GxE. *New Phytol.*, 219(1):31–36, 2018.
- [67] Robert Bukowski, Xiaosen Guo, Yanli Lu, Cheng Zou, Bing He, Zhengqin Rong, Bo Wang, Dawen Xu, Bicheng Yang, Chuanxiao Xie, et al. Construction of the third-generation zea mays haplotype map. *GigaScience*, 7(4):gix134, 2017.

- [68] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, 44(7):821, 2012.
- [69] Funda Ogut, Yang Bian, Peter J Bradbury, and James B Holland. Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. *Heredity*, 114(6):552, 2015.
- [70] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.
- [71] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *PNAS*, 100(16):9440–9445, 2003.

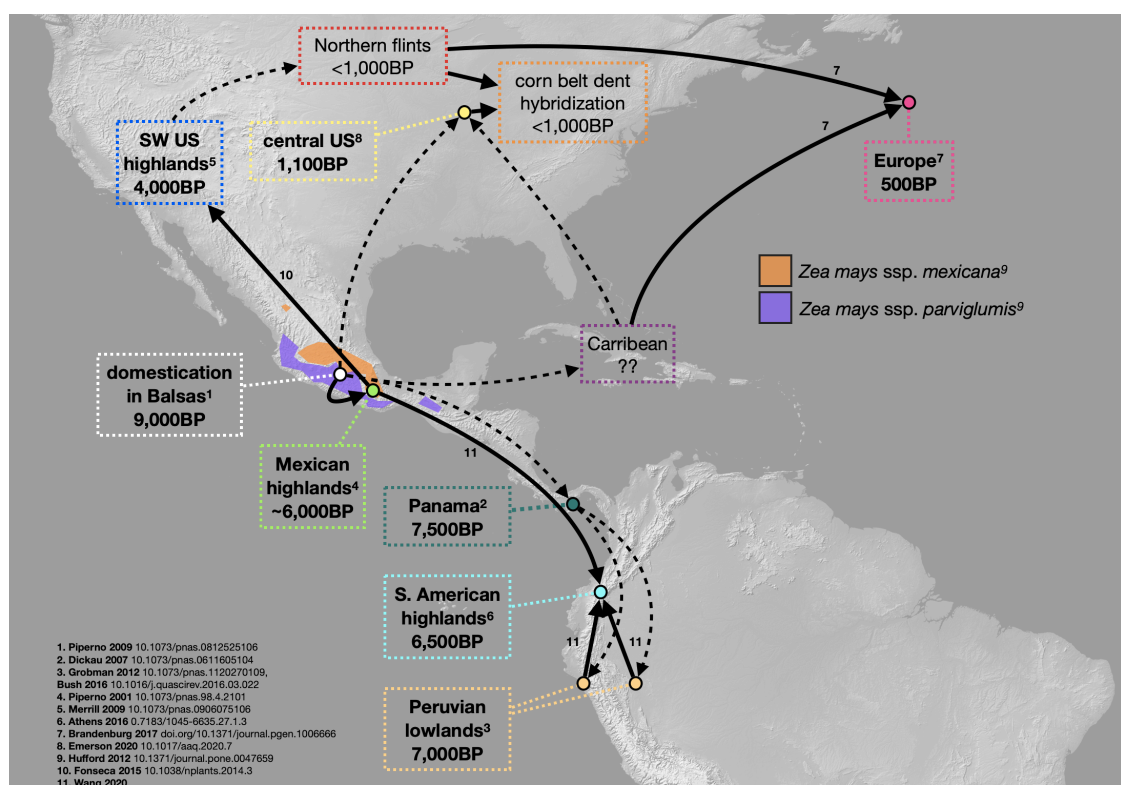


Figure 1: Sampling locations and expansion route of maize landraces. Domestication and expansion times for maize populations are from published articles [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29].

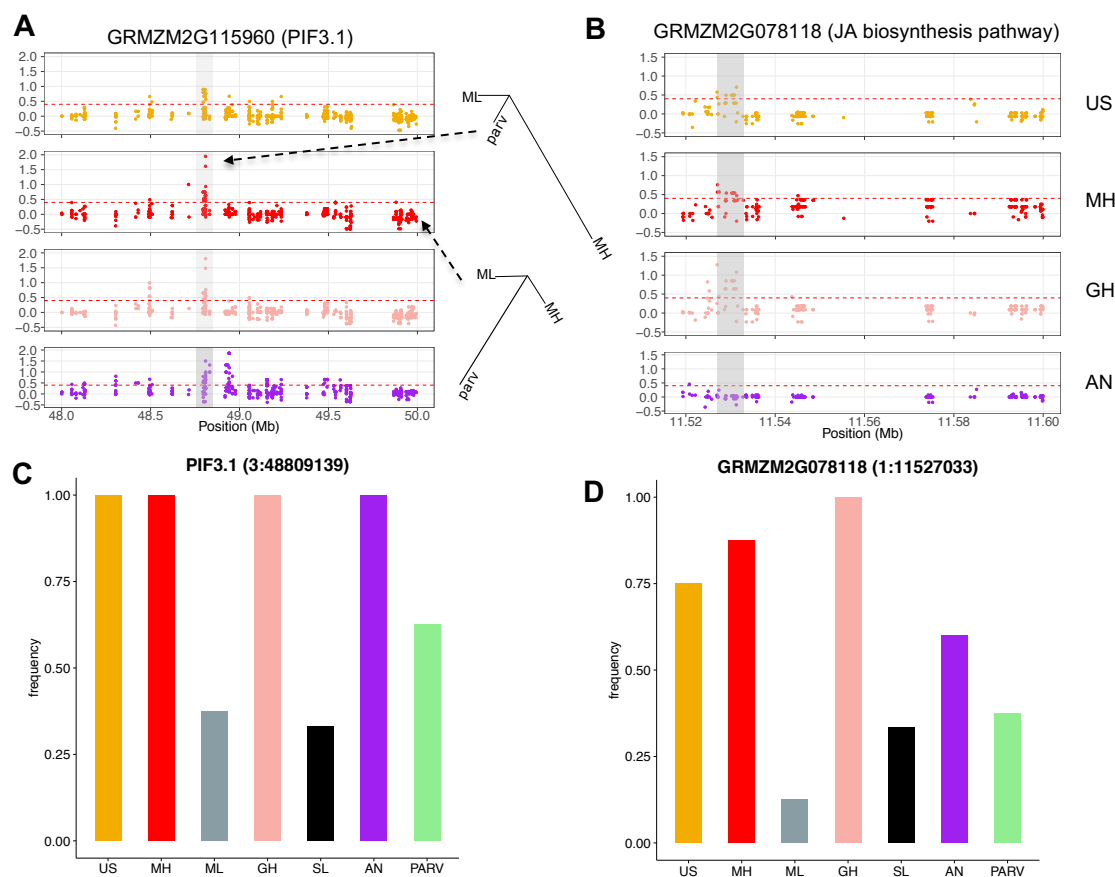


Figure 2: Patterns of parallel adaptation at two candidate loci. A. Distribution of *PBE* values in a 2-Mb region around the gene PIF3.1. The branch length of inset trees was based on *PBE* values and indicates the difference between a selected and unselected SNP. B. Distribution of *PBE* values in a 0.5 Mb region around the gene GRMZM2G078118 involved in jasmonic acid biosynthesis. C. Barplot of the reference allele frequency of one SNP located in PIF3.1. D. Barplot of the non-reference allele frequency of one SNP located in GRMZM2G078118.

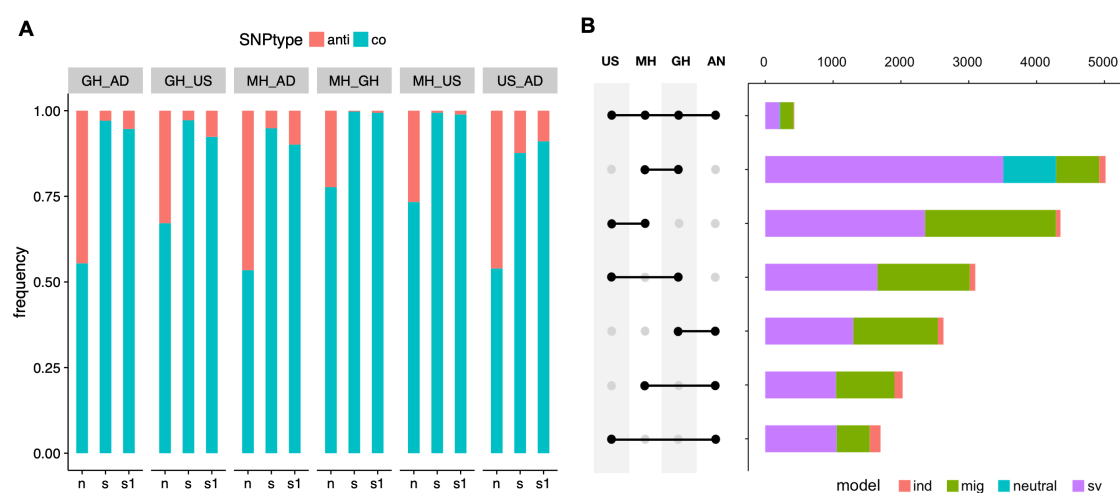


Figure 3: Patterns of parallel adaptation across loci. A. Distribution of co-directional and anti-directional SNPs in neutral and outlier SNP sets. n: neutral SNPs; s: common outliers; s1: a subset of common outliers with MAF between 0.3 – 0.5. B. Number of best fit models for genetic source of repeatedly selected SNPs. ind: independent *de-novo* mutation; mig: migration; sv: standing variation. Abbreviations for populations: AN, Andes; GH, Guatemalan Highlands; MH, Mexican Highlands; US, Southwestern US Highlands.



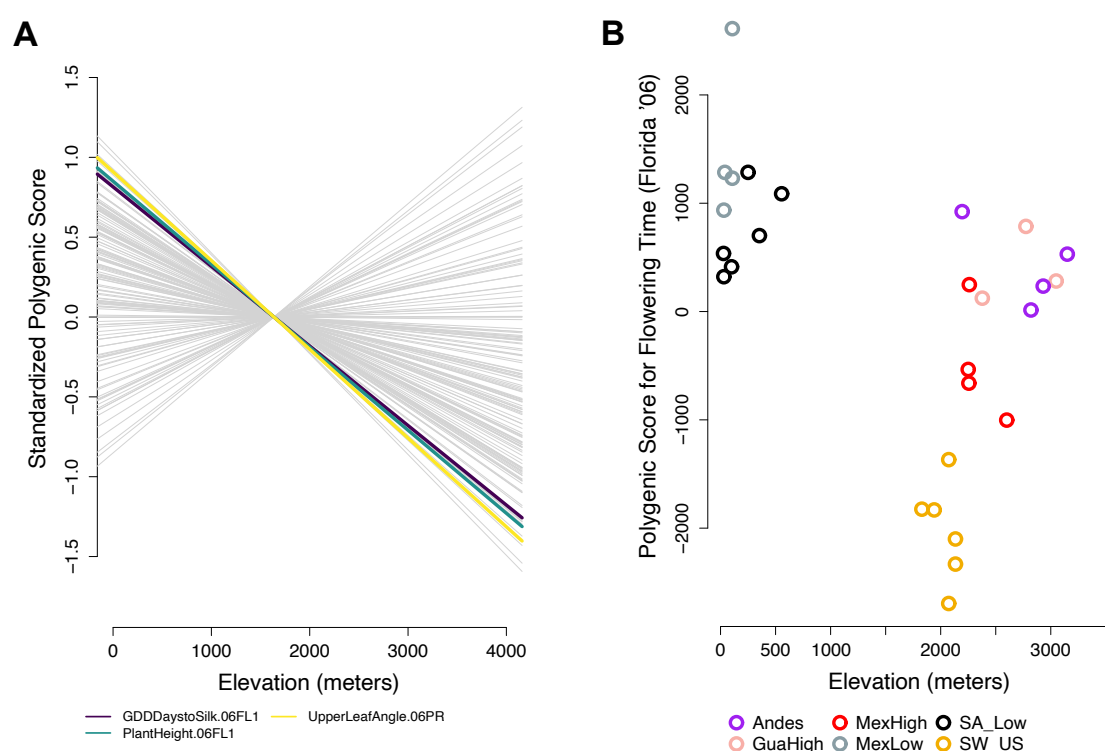


Figure 4: Polygenic adaptation along elevation in the landraces. A. A linear regression between polygenic score for all 29 trait-environment combinations tested and elevation of origin. The lines for the seven traits that showed significant signals of polygenic adaptation are colored and all other traits are shown in gray. B. Polygenic score for days to silk in Florida (2006) for all landraces is negatively correlated with elevation of origin

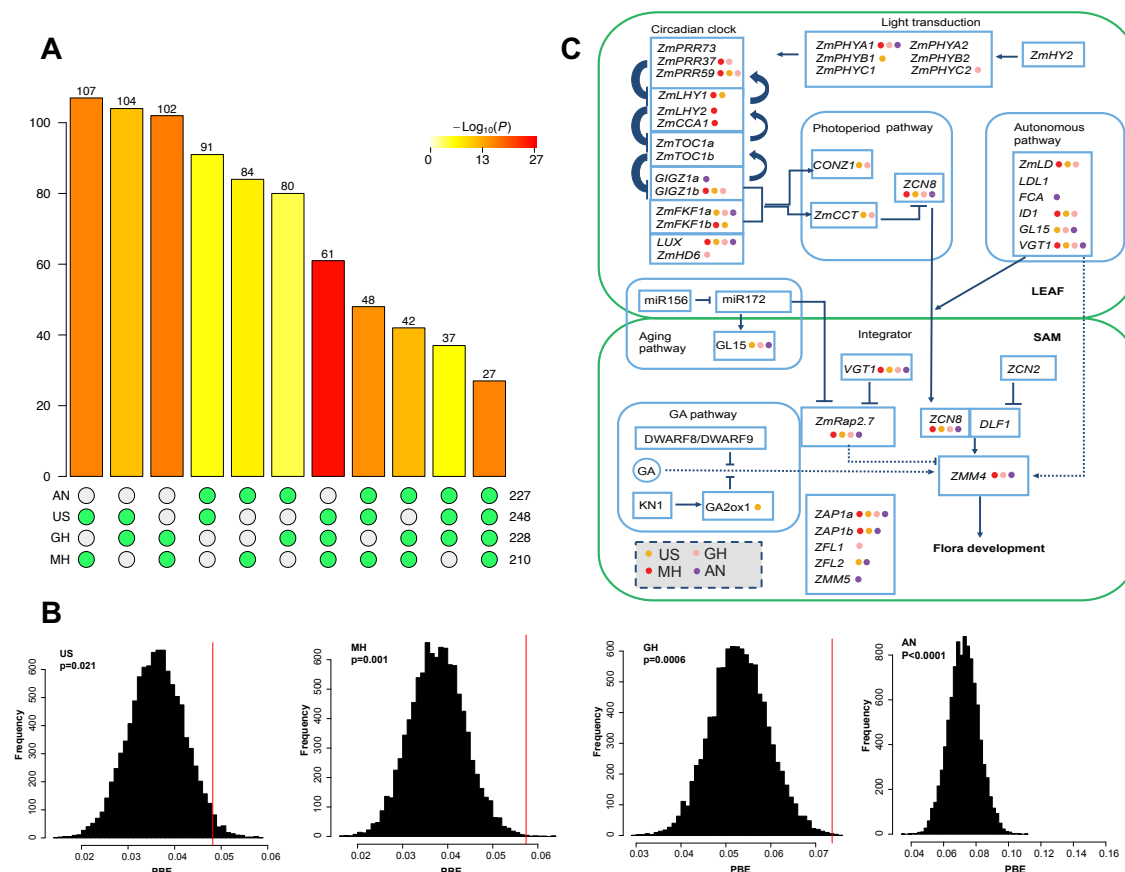


Figure 5: Convergent adaptation in the flowering time pathway. A. Intersection of flowering time outlier genes among four highland populations. B. Distribution of mean *PBE* values of SNPs located within and in the 10-kb flanking regions of core flowering time genes [47] (the red line) against genomic background (the black bars). C. Diagram showing selected genes in the flowering time network. Colored dots indicate the population(s) in which selection was detected.

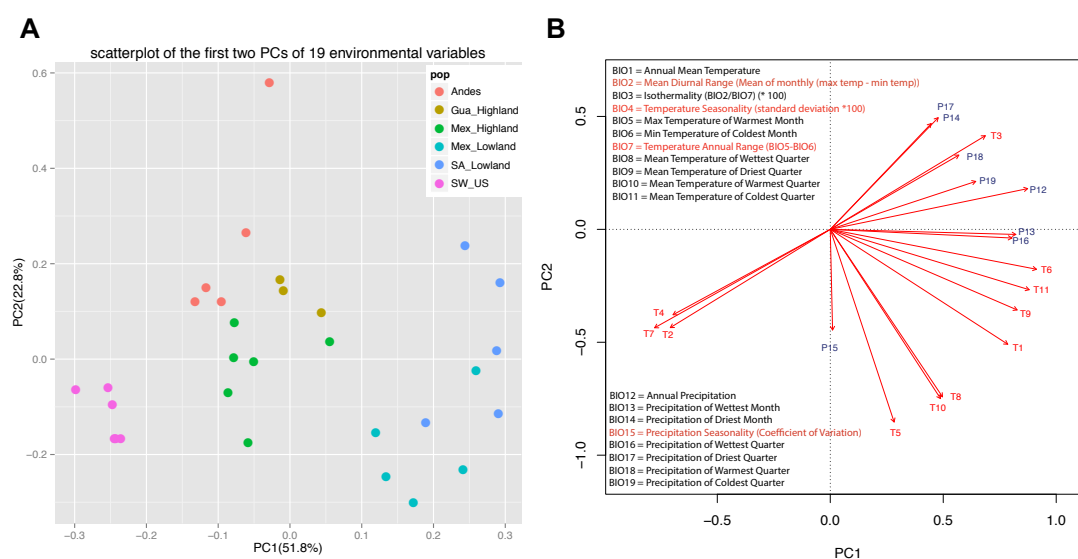


Figure S1: Environment of maize populations. A. PCA of 19 bioclim environmental variables for our samples. B. Projection of the 19 environmental variables on the first two PCs. Abbreviations for populations: GuaHigh, Guatemalan Highlands; MexHigh, Mexican Highlands; MexLow, Mexican Lowlands; SA.Low, South American Lowlands; SW\_US, Southwestern US Highlands.

## Supporting Information

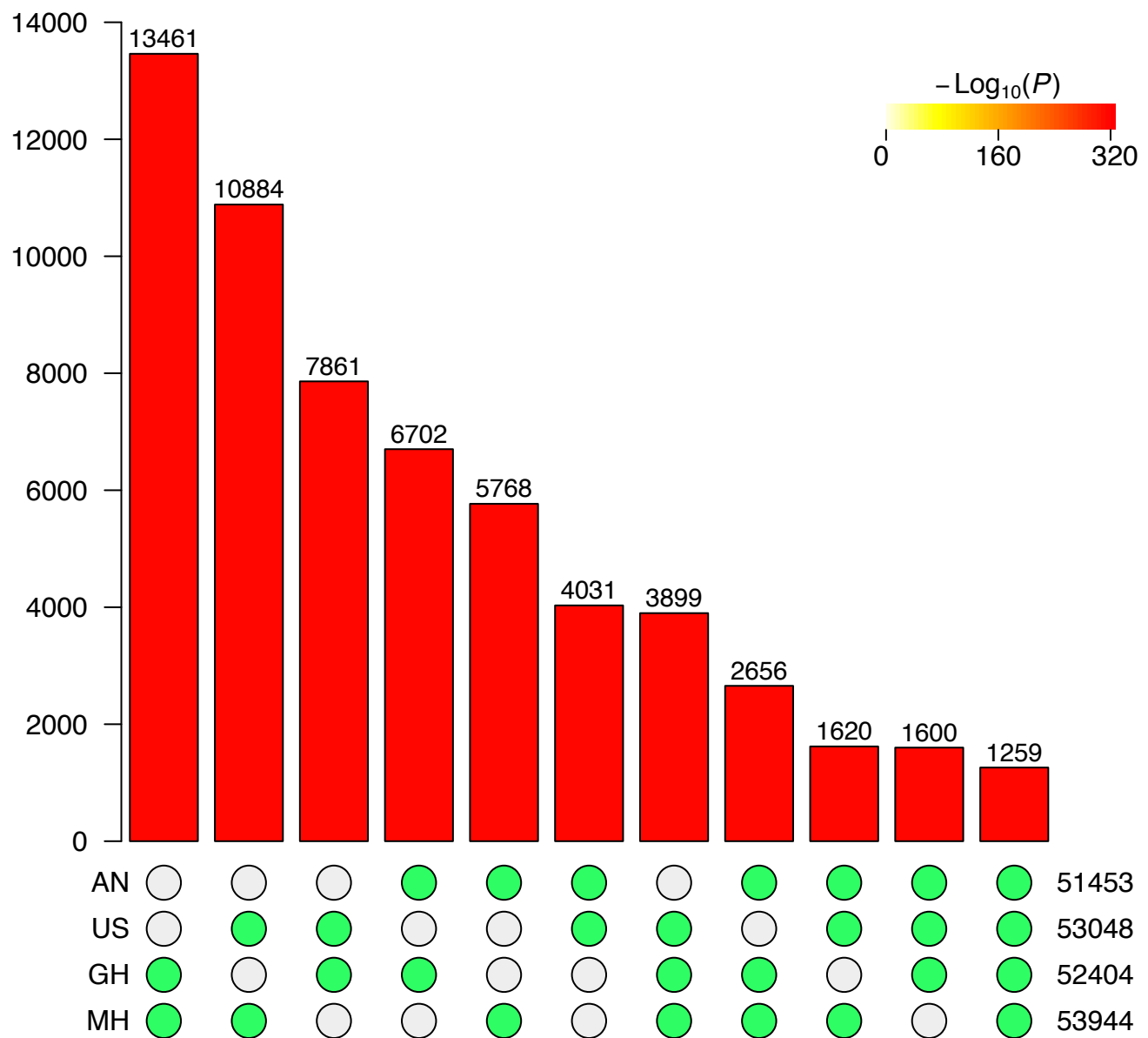


Figure S2: Intersection of outlier SNPs among four highland populations.

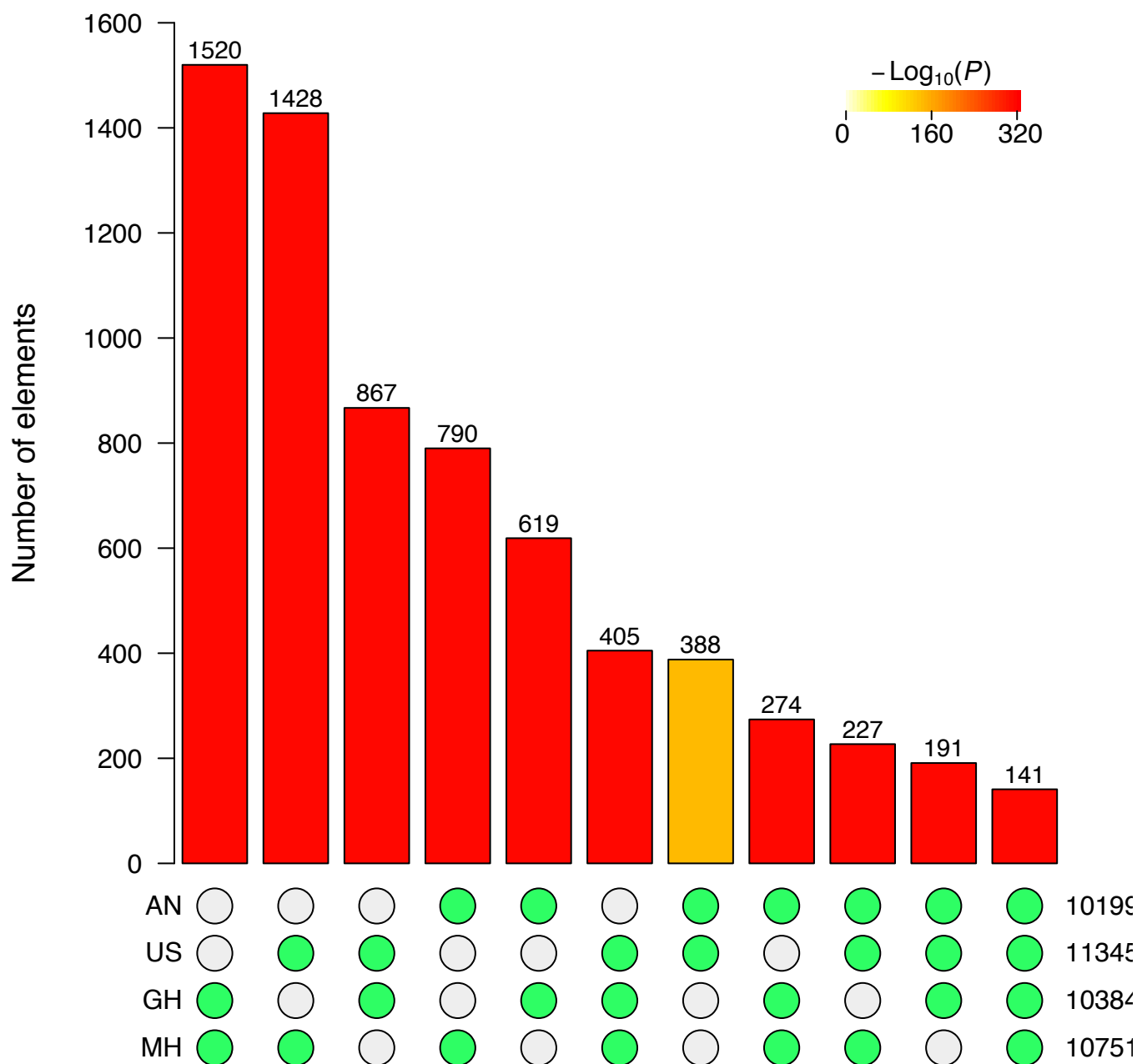


Figure S3: Intersection of outlier SNPs (1% top PBE values) among four highland populations.



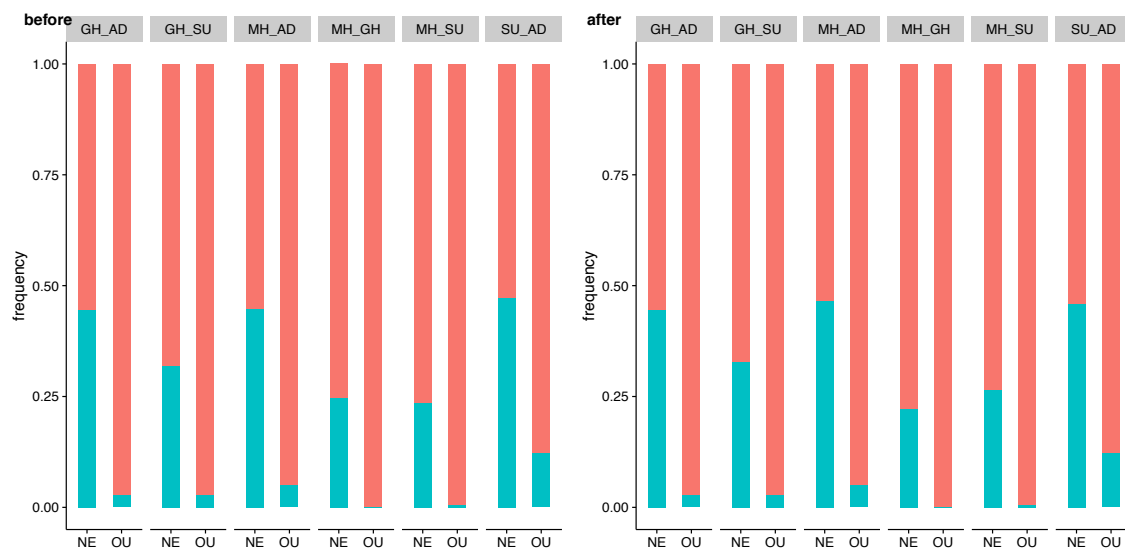


Figure S4: Ratio of co-directional and anti-directional SNPs in common neutral (NE) and outlier (OU) SNP set before and after correcting for 2dsfs.

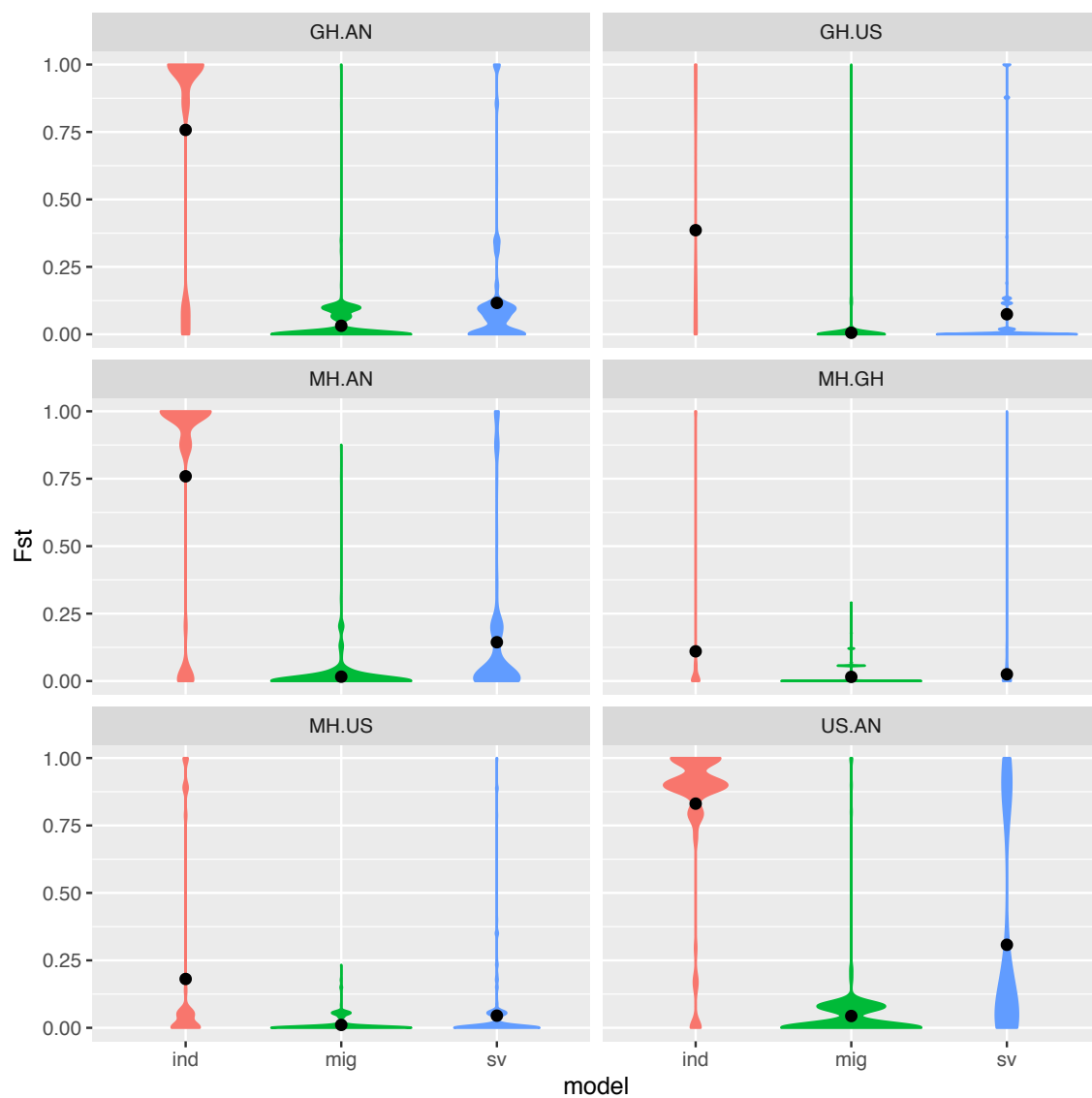


Figure S5: Comparison of  $F_{st}$  between pairs of highland populations for repeatedly selected SNPs assigned to different source models.

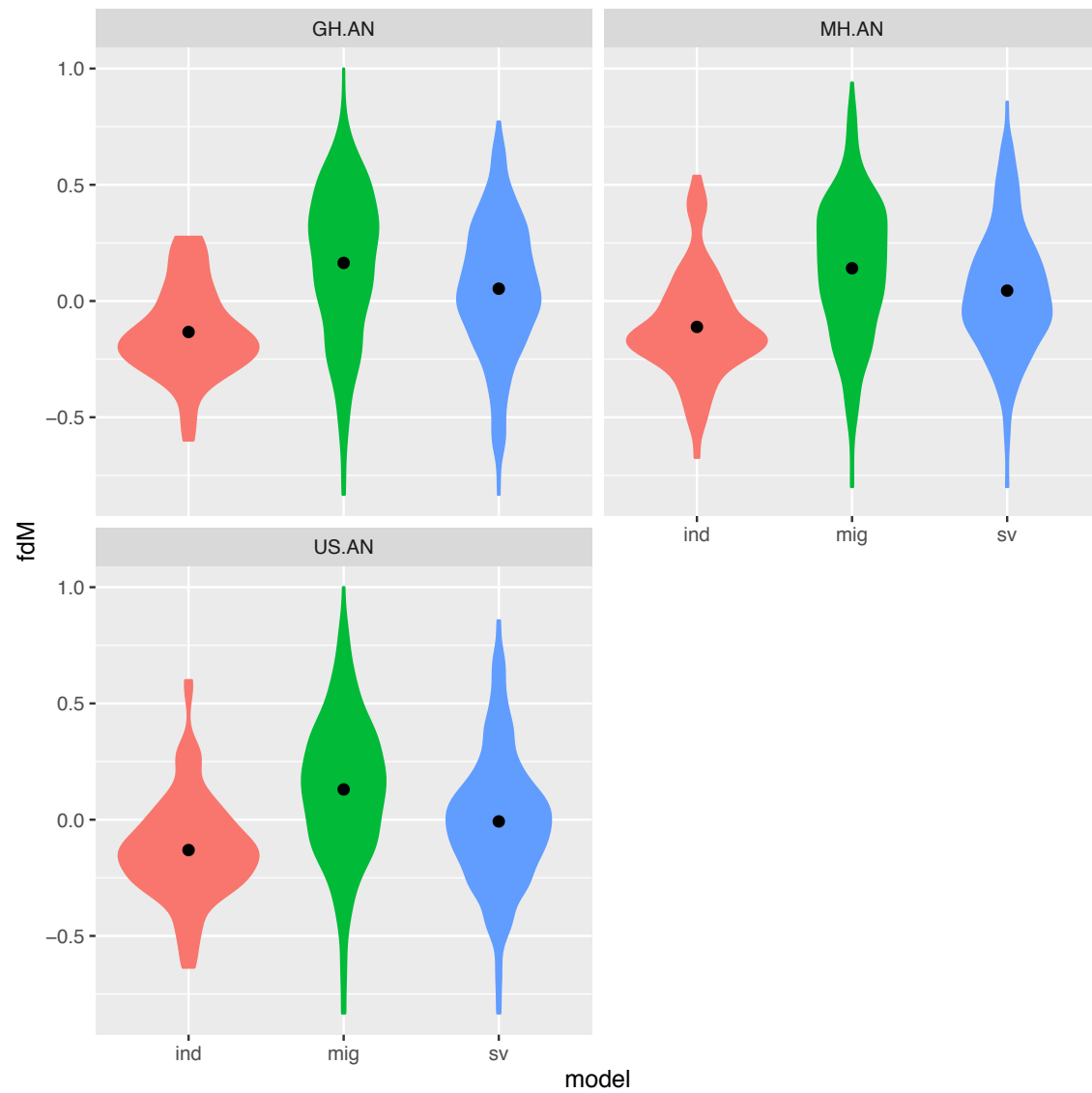


Figure S6: Comparison of  $\hat{f}_{dM}$  between pairs of highland populations for repeatedly selected SNPs assigned to different source models.

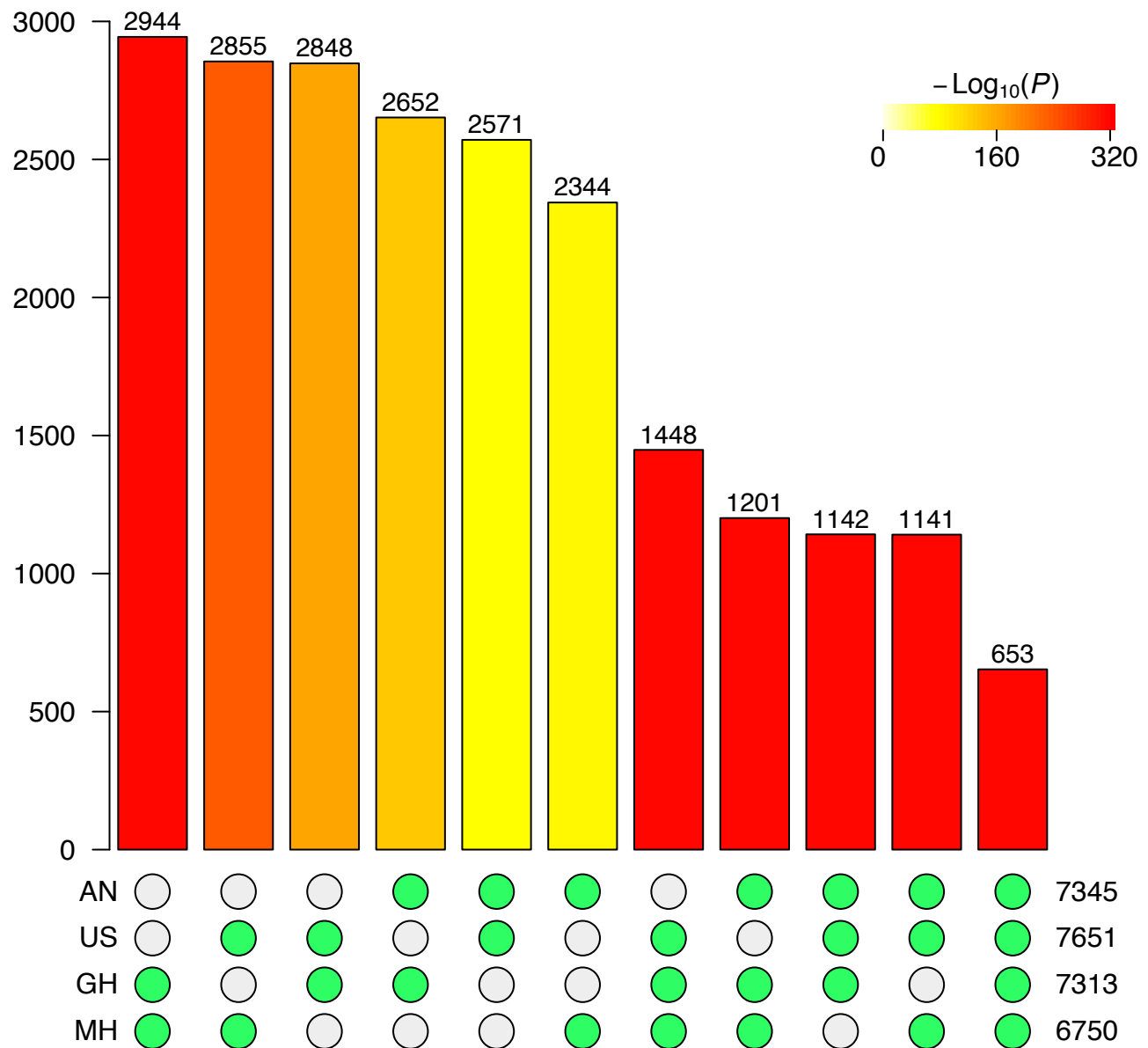


Figure S7: Intersection of outlier genes (based on SNPs with the top 5% PBE values) among four highland populations.

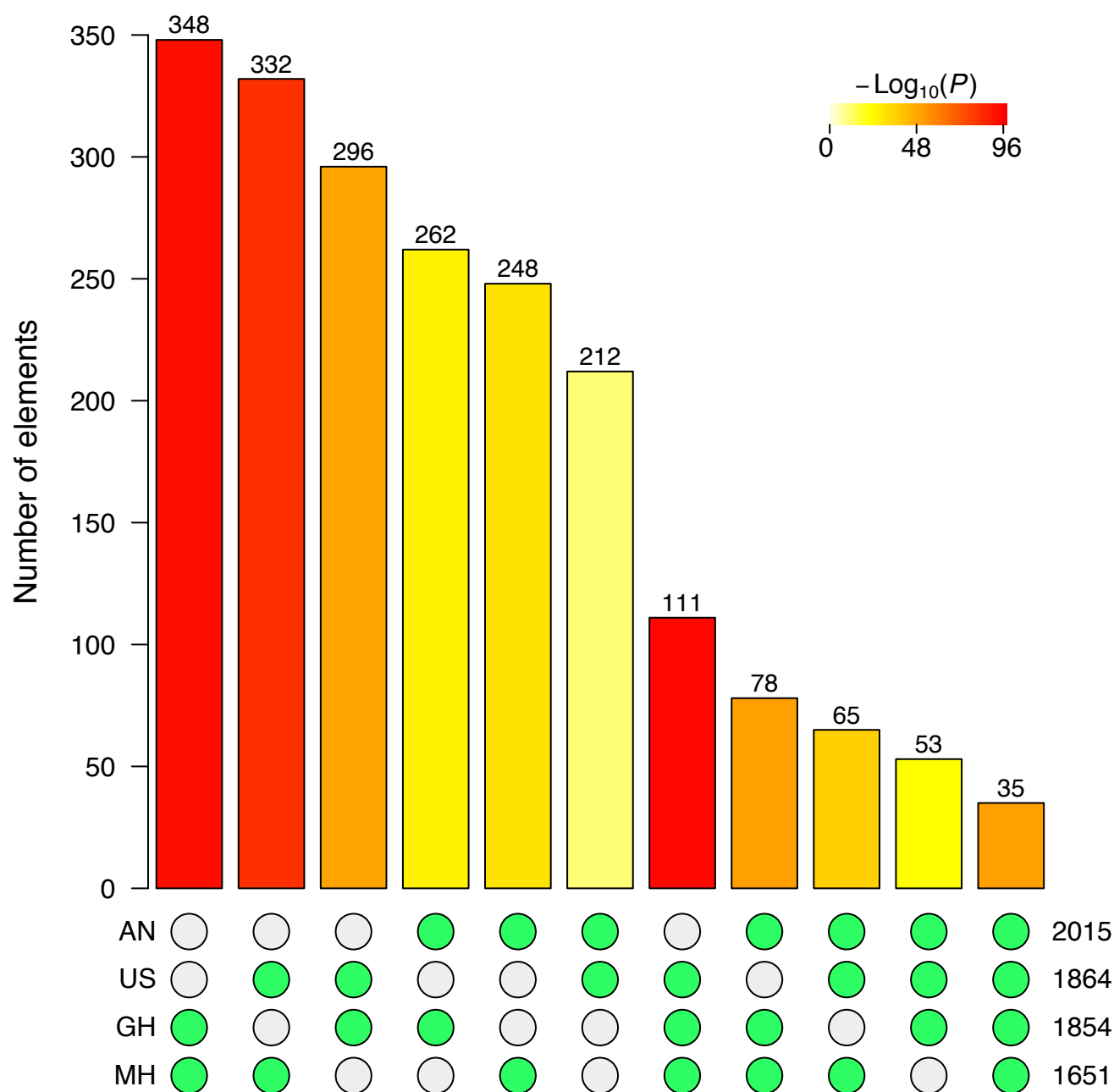


Figure S8: Intersection of outlier genes (based on SNPs with the top 1% PBE values) among four highland populations.

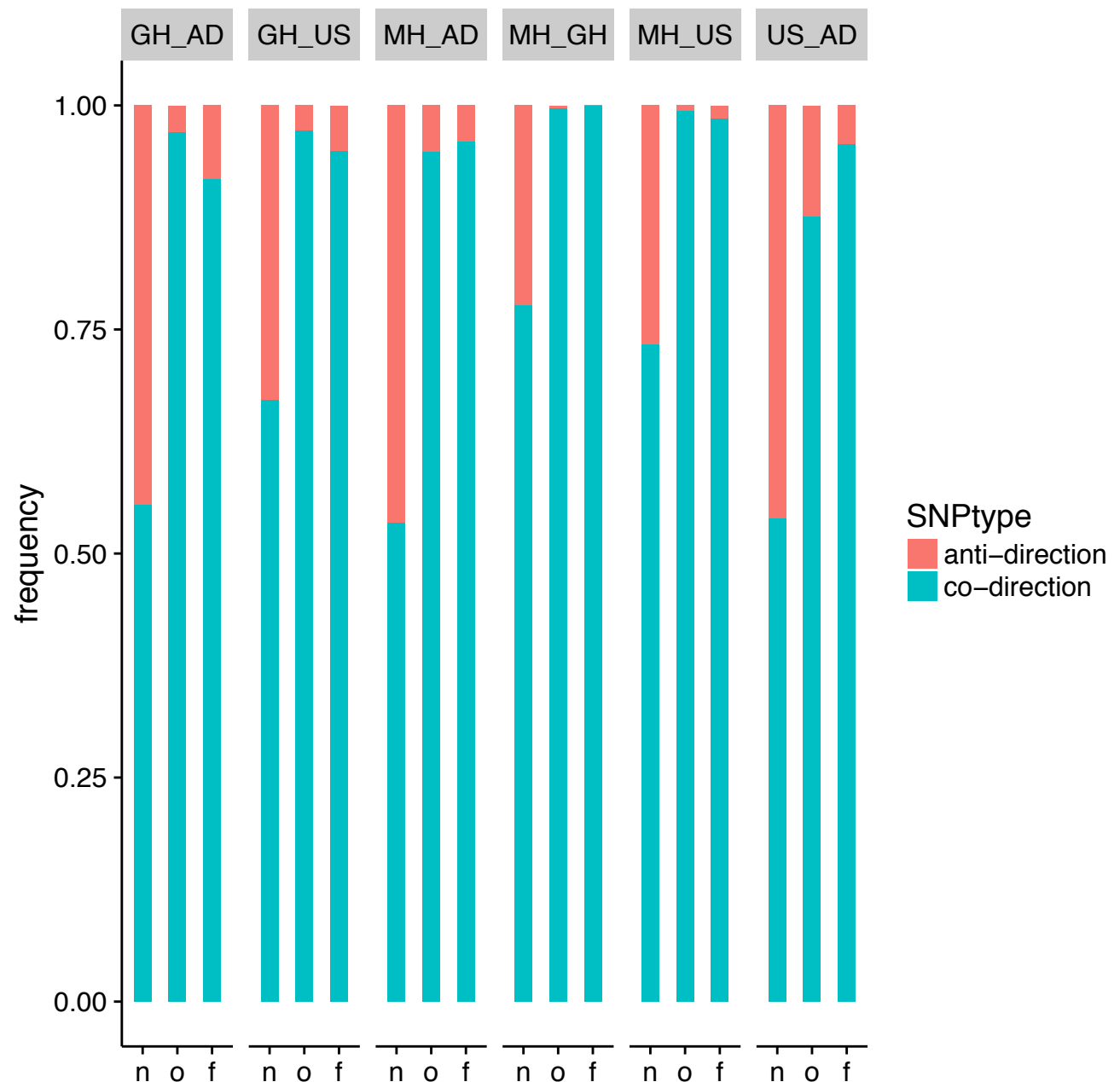


Figure S9: A barplot showing the dominance of co-directional SNPs in flowering time metabolism pathways. n: neutral SNPs; o: outlier SNPs; f: outlier SNPs within or flanking genes in the flowering time pathway.



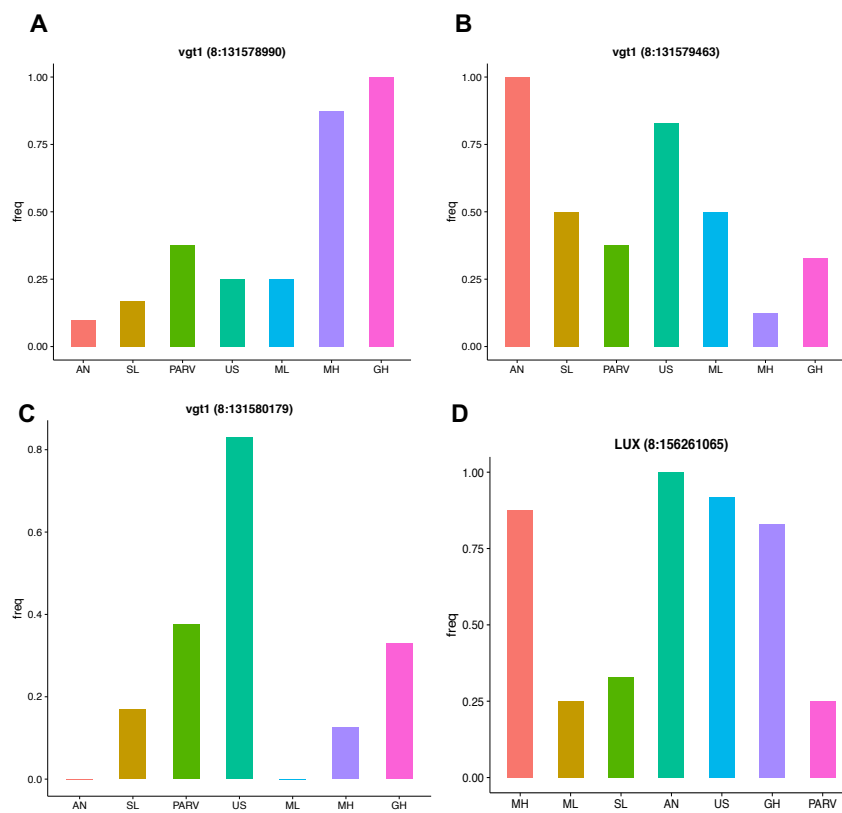


Figure S10: SNP frequency change in vgt1 (A, B, C) and LUX (D) among populations.

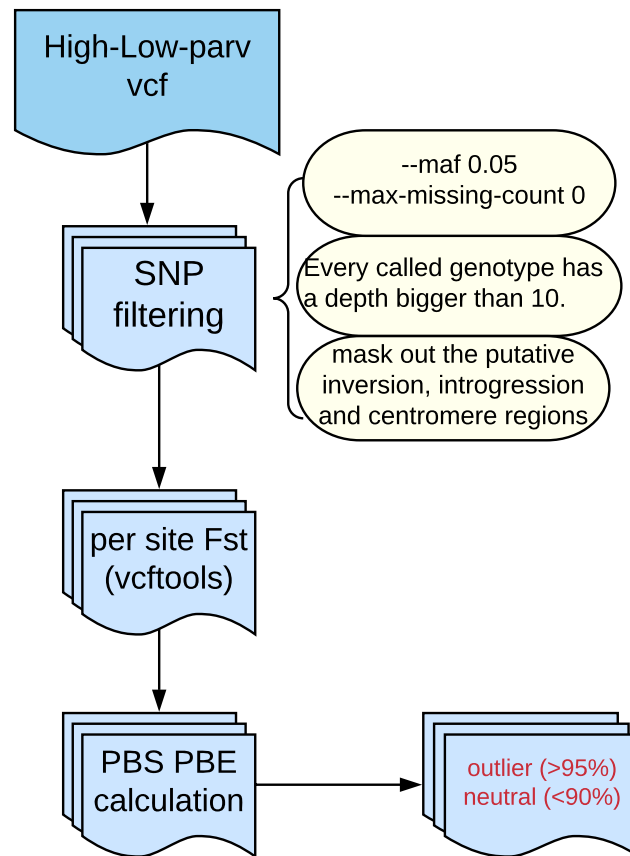


Figure S11: The analyses pipeline for *PBE* calculation.

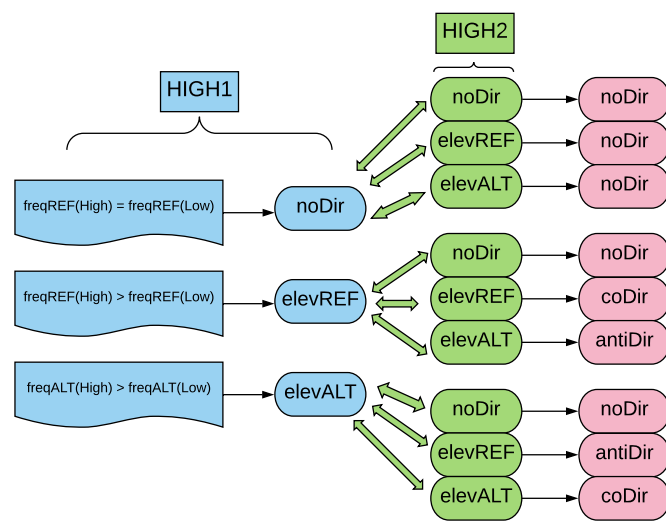


Figure S12: The analyses pipeline for determining co-directional and anti-directional SNPs.

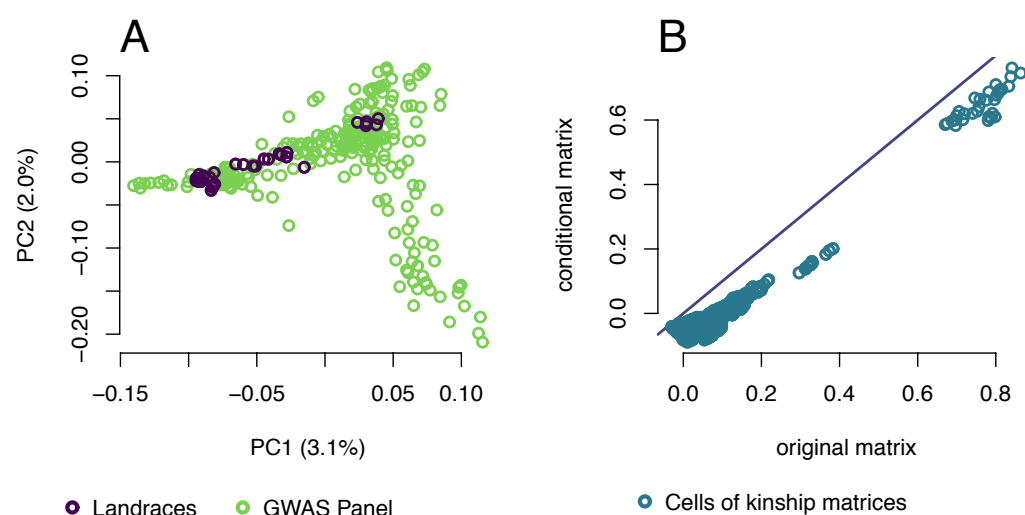


Figure S13: Evidence for shared population structure between the GWAS panel and the landraces. A. Individuals in both panels plotted along the first two PCs of the joint kinship matrix for the GWAS panel and the landraces. B. A comparison of all cells in the original kinship matrix for the landraces (X axis) and the conditional matrix accounting for relatedness with the GWAS panel (Y axis). The  $y=x$  line is plotted for reference.