nature genetics

A study of allelic diversity underlying flowering-time adaptation in maize landraces

J Alberto Romero Navarro[1], Martha Willcox[2], Juan Burgueño[2], Cinta Romay[3], Kelly Swarts[1], Samuel Trachsel[2], Ernesto Preciado[4], Arturo Terron[4], Humberto Vallejo Delgado[5], Victor Vidal[6], Alejandro Ortega[7], Armando Espinoza Banda[8], Noel Orlando Gómez Montiel[9], Ivan Ortiz-Monasterio[2], Félix San Vicente[2], Armando Guadarrama Espinoza[2], Gary Atlin[2], Peter Wenzl[2], Sarah Hearne[2] & Edward S Buckler[1,3,10]

Landraces (traditional varieties) of domesticated species preserve useful genetic variation, yet they remain untapped due to the genetic linkage between the few useful alleles and hundreds of undesirable alleles[1]. We integrated two approaches to characterize the diversity of 4,471 maize landraces. First, we mapped genomic regions controlling latitudinal and altitudinal adaptation and identified 1,498 genes. Second, we used F-one association mapping (FOAM) to map the genes that control flowering time, across 22 environments, and identified 1,005 genes. In total, we found that 61.4% of the single-nucleotide polymorphisms (SNPs) associated with altitude were also associated with flowering time. More than half of the SNPs associated with altitude were within large structural variants (inversions, centromeres and pericentromeric regions). The combined mapping results indicate that although floral regulatory network genes contribute substantially to field variation, over 90% of the contributing genes probably have indirect effects. Our dual strategy can be used to harness the landrace diversity of plants and animals.

Maize (*Zea mays* subsp. *mays*) is a model organism with a 100-year legacy of cytological, genetic and biomolecular characterization[2]. Maize displays high genetic diversity with low linkage disequilibrium (LD)[3,4], low population differentiation[5], prevalent migration[6] and occasional introgression from wild relatives[7–9]. More recently, experimental populations like the Nested Association Mapping (NAM) populations[10,11], and large association panels[4,12] have allowed mapping and deployment of useful alleles for several quantitative traits[13–16]. However, these panels' founder lines are inbred improved lines (with many from temperate regions) and capture only a modest fraction of available diversity. In contrast, maize landraces span numerous ecogeographic areas and harbor most of the diversity of the species. Recent studies have provided small-scale characterization of Mexican[17] and European landraces[18]. Nevertheless, most maize and other crop landraces remain largely uncharacterized by genomics.

Here we map the genes that control flowering with two distinct methods. First, each landrace is well adapted to their native environments, and we used that environmental information as the trait to identify genes driving large-scale adaptation. Second, we mapped flowering-time variation in controlled field experiments through a newly developed, rapid experimental design called F-one association mapping (FOAM) (**Fig. 1**). FOAM consists of sampling single individuals across numerous populations, which are then genotyped and crossed to one or a small number of common parents to derive F1 families. Subsequently, a genome-wide association study (GWAS) is performed using multi-trial F1 progeny evaluation. The major advantages of this design include: (i) the ability to capture thousands of alleles across populations, (ii) the ability to maintain the tractability of two alleles per locus per population, and (iii) ample replication of alleles, increasing the power and accuracy for genetic effect estimation. FOAM's main limitation is that nested evaluation of different subsets of F1 progeny by ecological zone limits the ability to accurately estimate genotype-by-environment interactions.

Our FOAM population used individuals from 4,471 accessions across 35 countries in the Americas (**Fig. 2**), which were grouped into three adaptation classes to account for adaptation to low, middle or high elevation. Similarly, common parents and evaluation sites were nested within the adaptation class (**Supplementary Fig. 1**)[19,20]. We used genotyping-by-sequencing[21] on landrace parents and found almost one million SNPs, and missing data was imputed using BEAGLE4 (ref. 22). Of 4,471 accessions, 3,552 yielded F1 families that contained both genotypic profiles and sufficient progeny, 3,633 that contained

[1]School of Integrative Plant Sciences, Section of Plant Breeding and Genetics, Cornell University, Ithaca, New York, USA. [2]International Maize and Wheat Improvement Center (CIMMYT), Texcoco, México. [3]Institute for Genomic Diversity, Ithaca, New York, USA. [4]Instituto Nacional de Investigaciones Forestales Agrícolas y Pecuarias (INIFAP), Campo Experimental Bajio, Celaya, México. [5]Instituto Nacional de Investigaciones Forestales Agrícolas y Pecuarias (INIFAP), Campo Experimental Uruapan, Uruapan, México. [6]Instituto Nacional de Investigaciones Forestales Agrícolas y Pecuarias (INIFAP), Campo Experimental Santiago Ixcuintla, Santiago Ixcuintla, México. [7]Instituto Nacional de Investigaciones Forestales Agrícolas y Pecuarias (INIFAP), Campo Experimental Norman E. Borlaug, Ciudad Obregón, México. [8]Universidad Autonoma Agraria Antonio Narro, Torreon, México. [9]Instituto Nacional de Investigaciones Forestales Agrícolas y Pecuarias (INIFAP), Campo Experimental Iguala, Iguala, México. [10]US Department of Agriculture–Agricultural Research Service (USDA–ARS), Ithaca, New York, USA. Correspondence should be addressed to S.H. (s.hearne@cgiar.org) or E.S.B. (esb33@cornell.edu).

Received 20 April 2016; accepted 10 January 2017; published online 6 February 2017; corrected online 20 February 2017 (details online); doi:10.1038/ng.3784

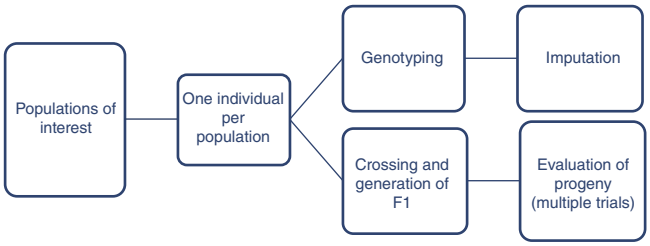

Figure 1 Experimental design. One individual from each of up to thousands of individuals is genotyped and used as the parent. Progeny are then evaluated for multiple years and locations to estimate the genetic contribution of the original individual, and phenotypic and genotypic data are used for genome-wide association analysis.

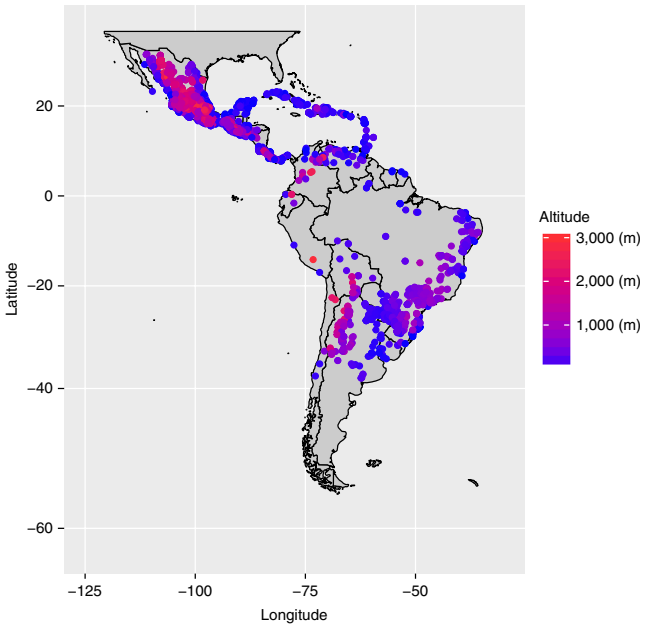

Figure 2 Geographic coordinates of original sampling sites of landrace accessions. Color gradient corresponds to altitude, with adaptation classes corresponding to low-elevation: <1,200 m above sea level and <30° N or 40° S; mid-elevation: between 1,200 and 1,900 m above sea level and <30° N or 40° S; or high-elevation: >1,900 m above sea level and <30° N or 40° S (n = 3,633). Map was drawn using Draw Geographical Maps, R package version 3.1.1.

passport information, which was used for mapping large-scale adaptation, and 2,603 that were present in both mapping studies.

We first explored the effects of recombination and geography-driven limited dispersal on the distribution of genetic diversity in the landrace parents. By using multidimensional scaling (MDS), we found that the first two axes only explained 8.7% of the variation, consistent with a low fixation index (F_{ST}) in landraces[5], with the main differentiation reflecting gradients across the northern and southern hemispheres and across Mexico (**Supplementary Fig. 2**). In addition, a Mantel test[23] showed a significant correlation between geographic and genetic distances (Pearson's $r = 0.46$, $P = 0.000999001$), with most of the association being driven by altitude. MDS shows that country of origin—not adaptation class—remains the main clustering factor among the landraces (**Supplementary Fig. 3**); landraces found at high elevation mainly corresponded to sampling locations across Mexico, and highland landraces show incomplete differentiation from middle and low elevation populations. This indicates that at their center of origin and diversity, maize landraces that are adapted to different elevations are not fully differentiated, with alleles segregating across adaptation classes. Recombination can also limit free segregation of alleles through the presence of genomic features that induce increased LD. To study recombination, we estimated an approximate LD statistic that, although limited in resolution, shows a distribution consistent with previous recombination estimates[24,25]—higher in gene-rich regions and lower around centromeres. Each chromosome displays a unique recombination landscape, with six high-LD regions (**Supplementary Fig. 4**) encompassing 6.1% of the physical genome but only accounting for 2.8% of the annotated coding genes. Taken together, these results suggest that geography (in the form of isolation by distance and altitude) and genome structure (through a complex recombination landscape) function together to shape the distribution of maize genetic variation.

Maize flowering time is crucial for local adaptation, and it is a complex trait controlled by hundreds of loci with small effects, many with multiple allelic series[4,14,26–31]. In many plant species, the genetic architecture underlying flowering time is key for adaptation to latitude and altitude[32–34]. Therefore, we used the altitude and latitude of the sampling location as traits to map local adaptation, and we chose significance thresholds to maximize the genic overlap rate with flowering time (**Supplementary Fig. 5**). For altitude, we observed that 58.4% of significant ($-\log_{10}(P$ value$) > 208.2$) SNPs corresponded to regions with higher LD. In particular, *Inv4m*, the 13-Mb adaptive introgression from highland teosinte into maize[8,35] was highly significant. For altitude we observed significance with the centromeres of chromosomes 2,5,6 and 8 and a large region upstream of the centromere on chromosome 3. Outside of these low-recombination regions, 366 genes showed significant association with altitude. For latitude, we observed that only 13.1% of the significant SNPs ($-\log_{10}(P$ value$) > 61.63$) were contained within low-recombination regions, particularly the centromere of chromosome 5. In total, across Latin America 1,498 genes showed significant association with latitude, of which 395 were shared with altitude. Minor-allele frequency distribution of significant SNPs indicates that many are shared across clades and landraces, which is very distinct from neutral distribution (**Fig. 3**). These 1,498 genes seem to be the main contributors to large-scale environmental adaptation of maize to altitude and latitude—key drivers of flowering time.

To study the genetic basis of flowering time, we conducted field evaluations on F1 progeny over 22 trials and 2 years in 13 locations across Mexico, with each trial containing a different subset of the collection to maximize the number of accessions evaluated (**Supplementary Table 1**). A GWAS was performed independently for each trial by using a mixed linear model (MLM). There was a 72% overlap between the significant ($-\log_{10}(P$ value$) > 18$) SNPs associated with male and female flowering, as expected from the overlapping genetic control[14]. There was a significant contribution of low-recombination regions in flowering-time variation, parallel to that of latitude and longitude, with 20-fold enrichment for significant SNPs at high-LD regions (Pearson's one-sided chi-squared test, $P < 2.2 \times 10^{-16}$). In particular, significant structural variants included the centromeres of chromosomes 3, 5 and 6, *Inv4m*, and a 6-Mb region on chromosome 3 beginning at 79 Mb. The 6-Mb region on chromosome 3 has similar segregation to that of *Inv4*, and its increased LD suggests that it might be an inversion. In NAM populations, this putative inversion and the centromere comprise a single quantitative trait locus (QTL) for flowering time[14]. For the centromere of chromosome 5, there were three distinct alleles segregating in the landraces, all of

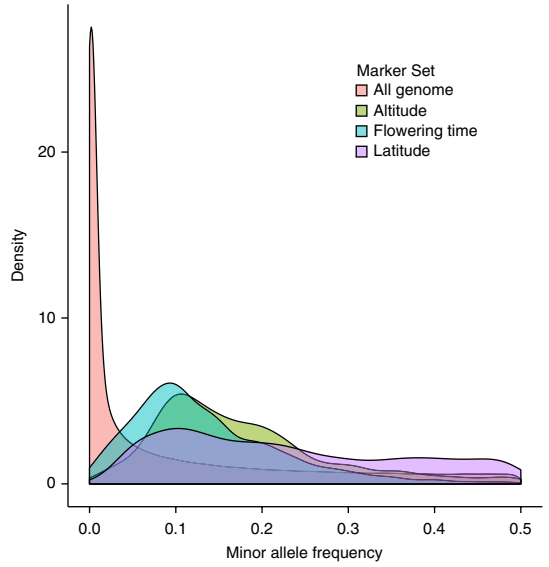

Figure 3 Minor allele frequency distributions. The genome-wide (n = 955,686) distribution of minor allele frequency shows a curve consistent with the expectation for a random mating population, with most of its density at low frequency; the median minor allele frequency was 1.8% genome wide. In contrast, the median minor allele frequencies of the SNPs associated with the various traits was significantly higher than those for the null distribution; median minor allele frequencies for flowering time (n = 5,818), altitude (n = 2,513) and latitude (n = 5,026) were 11%, 15% and 20%, respectively. For flowering time, although the density of the distribution was found at higher values than that for the null genome-wide distribution, a significant enrichment was observed just above 5%, which was the lower limit for SNPs to be considered in the GWAS models. A small fraction of flowering-time-associated SNPs overlap at high minor-allele-frequency SNPs with the SNPs associated with adaptation to altitude and latitude.

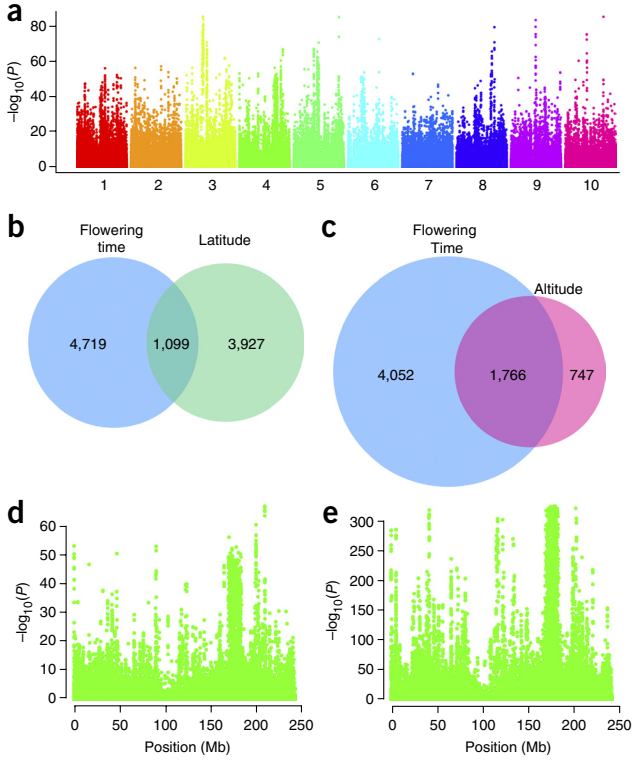

Figure 4 Significance for flowering time, and overlap between flowering time and latitude- and altitude-associated SNPs. (**a**) Manhattan plot for 'time to female flowering' (in days), The x axis shows the positions across the ten maize chromosomes (n = 502,601). The y axis represents the $-\log_{10}(P$ value) at each site (F-test; landrace accessions = 3,552; total landrace trial plots n = 18,797). (**b**,**c**) Venn diagrams showing the overlap between significant SNPs for flowering time and latitude (**b**) or altitude (**c**). Among the high-LD regions associated with both flowering time and altitude was *Inv4*, the adaptive introgression from highland teosinte to highland maize[8,35]. (**d**,**e**) Chromosome 4 Manhattan plots displaying GWAS P values for 'time to female flowering' (F-test; landrace accessions = 3,552; total landrace trial plots, n = 18,797) (**d**) and altitude (F-test; n = 3,633) (**e**). The region between 150 and 200 Mb with multiple contiguous significant SNPs corresponds to *Inv4m*, the adaptive introgression from highland teosinte to highland maize[8,35]. *Inv4m* was significantly associated with both flowering time and altitude.

which were present in the NAM population (**Supplementary Fig. 6**). The inverted allele of *Inv4m*, although absent from temperate plants, segregated at a high frequency in highland landraces (**Supplementary Fig. 7**), in which it had a very large additive effect, as it advanced flowering by 3 d, the largest effect for flowering time in maize to date. Both homozygous alleles from the putative chromosome 3 inversion segregated across our panel and NAM populations. Relative to that for *Inv4m*, this locus has a more modest effect on flowering time.

The phenotypic relevance of genomic structural variants in Mendelian and complex traits in species like *Drosophila*[36] has been known for nearly a century. In plants, chromosomal rearrangements can function as reproductive barriers in hybridization zones[37], underlie flowering time changes and contribute to local adaptation[38]. The observed significant association of structural variants with flowering time and local adaptation could be the product of heterosis, which in maize results in earlier flowering. At least one of the structural variants—the centromere of chromosome 5—has been reported to display a heterotic effect on yield[39], potentially the product of the complementation between alleles with deleterious mutations[24]. One of the alleles at the centromere of chromosome 5 segregates with a very low frequency across inbred lines, whereas the inversion of chromosome 4 is absent outside of the breeding material from the tropical highlands. The low frequency or complete absence of large-effect structural variants in improved lines could be the product of the process of selection during the development of inbred lines, in which favorable alleles were fixed, whereas other structural alleles were purged.

Outside of the structural variants, we observed 881 and 883 genes (~2.2% of genes) with significant association for days to female and male flowering, respectively (**Fig. 4** and **Supplementary Tables 2** and **3**). There was substantial enrichment for the candidate genes (Fisher's exact test $P = 4.3 \times 10^{-7}$), with association of 10 and 12 candidate genes with male and female flowering, respectively, which represented the circadian clock, photoperiod and gibberellin acid pathways (**Fig. 5**). The most significant hits corresponded to *VGT1* (refs. 31,40), one of the largest-known genotype-by-environment QTLs, and *ZCN8* (refs. 41,42), which encodes the maize florigen and homolog to FT in *Arabidopsis*. *ZmCCT*, the largest photoperiod QTL[30], was modestly significant for latitude, and significant only for days to female flowering, most likely due to sampling of the non-photoperiod-inducing accessions and of the trial locations. In maize, *dwarf8* has a cryptic association with flowering time[34]. We observed significance of *dwarf8* with latitude, altitude, and both male and female flowering, specifically in regions 50-kb upstream and 100-kb downstream of the coding region. This region displays divergent selection associated with

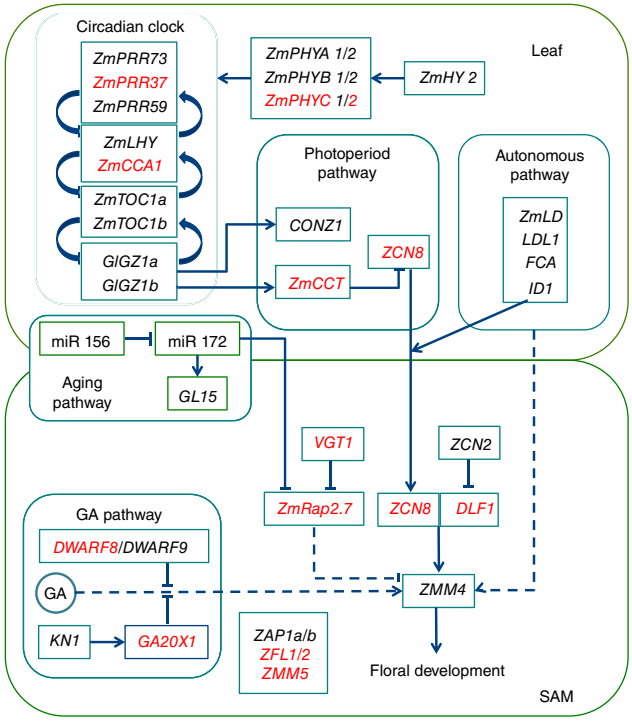

Figure 5 Flowering-time pathway, showing the genes involved in flowering time at the leaf and shoot apical meristem (SAM). Illustration is modified from that by Dong *et al.*[44]. The genes highlighted in red displayed significant association with flowering time in our study ($-\log_{10}(P) > 18$). miR, microRNA; GA, gibberellic acid.

climate adaptation[43]; in our landrace panel, adaptation and flowering time remained confounded.

Although the candidate and phase-change genes were enriched, most of the genes (**Supplementary Tables 4** and **5**) associated in this study may have an indirect effect on flowering-time variation through the action of their encoded protein products in upstream metabolism or in an interaction with the environment. Variation associated with flowering time displayed a significant effect with geography, with 61.4% and 19% SNP overlap with altitude and latitude, respectively (**Fig. 4**). The high level of overlap between these pathways was expected, but the stronger relationship with altitude suggested a key role of temperature and light quality on flowering time; the low level of overlap with latitude was likely due to trials and sampling of variation outside of short-day environments. However, we observed differences in the minor allele frequency distribution of the significant SNPs. Variation associated with altitude and latitude were enriched for high minor allele frequency polymorphisms (**Fig. 3**), which suggested that altitude and latitude associations were mostly due to globally adaptive SNPs, whereas flowering was a mix of high- and low-frequency mutations, likely adaptive variation and deleterious mutations (genetic load), respectively.

We assayed the potential for predicting flowering time in the landraces using either all of our high-density genetic markers or just the markers that were significantly associated with the trait. We performed genome-wide predictions, using gBLUP, independently for each trial. Across trials, the average fivefold cross-validated prediction accuracy was 0.45 for flowering time using either 30,000 markers or one SNP for each of the most significant genes (**Supplementary Fig. 8**). In contrast, for a similar number of random genes the prediction accuracy was only 0.22. The observed similar prediction accuracy between the top genes from GWAS to that of 30,000 random markers highlights the potential for using the significantly associated markers for breeding and for combining exotic beneficial alleles with current improved germplasm.

Landraces can be an incredible resource that can be used to adapt crops to the next century of climate change. Despite the tremendous diversity of landraces, genetic load prevents their efficient use without a genomic index. Our work lays out two complementary strategies for tapping landrace diversity. Geographic associations powerfully identify adaptive loci, which are common across populations and are unlikely to be deleterious given their high minor-allele frequency. Allele sharing is probably a consequence of outcrossing and extensive migration throughout Latin America in last several millennia. The limitation of this approach is that correlated traits and adaptations are being co-mapped. The FOAM GWAS helps differentiate the adaptive overlapping mutations from the potentially private deleterious mutations. These deleterious alleles have been the main limitation preventing breeders from exploiting landrace diversity. The strategy for tapping this diversity should use the overlapping genes and alleles of the two approaches, as these have proven to be adaptive and to target the trait of interest. Breeding could use standard genomic selection or genome editing. This provides an efficient strategy to use landrace diversity with the goal of helping to develop crops that adapt more efficiently to changing environments.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

J.A.R.N., M.W., J.B., S.T., E.P., A.T., H.V.D., V.V., A.O., A.E.B., N.O.G.M., I.O.-M., F.S.V., A.G.E., G.A., P.W. and S.H. were supported by La Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación (SAGARPA), Mexico under the MasAgro (Sustainable Modernization of Traditional Agriculture) initiative. J.A.R.N., C.R., K.S. and E.S.B. were supported by the US National Science Foundation (grant no. 1238014 and 0922493), and the USDA–ARS. We would like to thank ICAMEX and DuPont Pioneer Mexico for assistance in establishing the phenotypic trials.

AUTHOR CONTRIBUTIONS

J.A.R.N. conducted the GWAS analyses; M.W. coordinated the execution of the phenotypic trials, and the collection and curation of the phenotypic data; J.B. developed the phenotypic experimental designs, formulated models and determined landrace parent–environment estimates; C.R. assisted with the GWAS analysis and data interpretation; K.S. performed genotype imputation; S.T., E.P., A.T., H.V.D.,V.V., A.O., A.E.B., N.O.G.M., I.O.-M. and A.G.E. conducted the phenotypic trials; F.S.V. and A.G.E. developed the test-cross germplasm; G.A., P.W. and E.S.B. developed the project concept; S.H. coordinated the genotypic data collection, meta-data creation and passport data curation; and J.A.R.N., S.H. and E.S.B. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Warburton, M.L. *et al.* Genetic diversity in CIMMYT nontemperate maize germplasm: landraces, open pollinated varieties and inbred lines. *Crop Sci.* **48**, 617–624 (2008).
2. Wallace, J.G., Larsson, S.J. & Buckler, E.S. Entering the second century of maize quantitative genetics. *Heredity* **112**, 30–38 (2014).
3. Remington, D.L. *et al.* Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. USA* **98**, 11479–11484 (2001).

4. Romay, M.C. *et al.* Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* **14**, R55 (2013).
5. Hufford, M.B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
6. Mir, C. *et al.* Out of America: tracing the genetic footprints of the global diffusion of maize. *Theor. Appl. Genet.* **126**, 2671–2682 (2013).
7. van Heerwaarden, J. *et al.* Genetic signals of origin, spread and introgression in a large sample of maize landraces. *Proc. Natl. Acad. Sci. USA* **108**, 1088–1092 (2011).
8. Hufford, M.B. *et al.* The genomic signature of crop—wild introgression in maize. *PLoS Genet.* **9**, e1003477 (2013).
9. Warburton, M.L. *et al.* Gene flow among different teosinte taxa and into the domesticated maize gene pool. *Genet. Resour. Crop Evol.* **58**, 1243–1261 (2011).
10. McMullen, M.D. *et al.* Genetic properties of the maize nested-association-mapping population. *Science* **325**, 737–740 (2009).
11. Li, C. *et al.* Quantitative trait loci mapping for yield components and kernel-related traits in multiple connected RIL populations in maize. *Euphytica* **193**, 303–316 (2013).
12. Flint-Garcia, S.A. *et al.* Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* **44**, 1054–1064 (2005).
13. Peiffer, J.A. *et al.* The genetic architecture of maize height. *Genetics* **196**, 1337–1356 (2014).
14. Buckler, E.S. *et al.* The genetic architecture of maize flowering time. *Science* **325**, 714–718 (2009).
15. Harjes, C.E. *et al.* Natural genetic variation in lycopene ε-cyclase tapped for maize biofortification. *Science* **319**, 330–333 (2008).
16. Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested-association-mapping population. *Nat. Genet.* **43**, 159–162 (2011).
17. Arteaga, M.C. *et al.* Genomic variation in recently collected maize landraces from Mexico. *Genom. Data* **7**, 38–45 (2015).
18. Strigens, A., Schipprack, W., Reif, J.C. & Melchinger, A.E. Unlocking the genetic diversity of maize landraces with doubled haploids opens new avenues for breeding. *PLoS One* **8**, e57234 (2013).
19. Salhuana, W., Jones, Q. & Sevilla, R. The Latin American Maize Project: model for rescue and use of irreplaceable germplasm. *Diversity (Basel)* **7**, 40–42 (1991).
20. Pollak, L.M. The history and success of the public–private project on germplasm enhancement of maize (GEM). *Adv. Agron.* **78**, 45–87 (2003).
21. Elshire, R.J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high-diversity species. *PLoS One* **6**, e19379 (2011).
22. Browning, B.L. & Browning, S.R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
23. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).
24. Rodgers-Melnick, E. *et al.* Recombination in diverse maize is stable, predictable and associated with genetic load. *Proc. Natl. Acad. Sci. USA* **112**, 3823–3828 (2015).
25. Chia, J.-M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807 (2012).
26. Bertin, P., Madur, D., Combes, V., Dumas, F. & Brunel, D. Adaptation of maize to temperate climates: mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the *Vgt2* (*ZCN8*) locus. *PLoS One* **8**, e71377 (2013).
27. Ducrocq, S. *et al.* Key impact of *Vgt1* on flowering-time adaptation in maize: evidence from association-mapping and ecogeographical information. *Genetics* **178**, 2433–2437 (2008).
28. Hirsch, C.N. *et al.* Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* **26**, 121–135 (2014).
29. Chardon, F. *et al.* Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics* **168**, 2169–2185 (2004).
30. Hung, H.-Y. *et al. ZmCCT* and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. *Proc. Natl. Acad. Sci. USA* **109**, E1913–E1921 (2012).
31. Salvi, S. *et al.* Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc. Natl. Acad. Sci. USA* **104**, 11376–11381 (2007).
32. Ziska, L.H., Teramura, A.H. & Sullivan, J.H. Physiological sensitivity of plants along an elevational gradient to UV-B radiation. *Am. J. Bot.* **79**, 863–871 (1992).
33. Crimmins, T.M., Crimmins, M.A. & David Bertelsen, C. Complex responses to climate drivers in onset of spring flowering across a semi-arid elevation gradient. *J. Ecol.* **98**, 1042–1051 (2010).
34. Ziello, C., Estrella, N., Kostova, M., Koch, E. & Menzel, A. Influence of altitude on phenology of selected plant species in the Alpine region (1971–2000). *Clim. Res.* **39**, 227–234 (2009).
35. Pyhäjärvi, T., Hufford, M.B., Mezmouk, S. & Ross-Ibarra, J. Complex patterns of local adaptation in teosinte. *Genome Biol. Evol.* **5**, 1594–1609 (2013).
36. Dobzhansky, T. & Sturtevant, A.H. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics* **23**, 28–64 (1938).
37. Rieseberg, L.H., Whitton, J. & Gardner, K. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* **152**, 713–727 (1999).
38. Lowry, D.B. & Willis, J.H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).
39. Stuber, C.W., Lincoln, S.E., Wolff, D.W., Helentjaris, T. & Lander, E.S. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Genetics* **132**, 823–839 (1992).
40. Castelletti, S., Tuberosa, R., Pindo, M. & Salvi, S. A MITE transposon insertion is associated with differential methylation at the maize flowering time QTL *Vgt1*. *G3 (Bethesda)* **4**, 805–812 (2014).
41. Meng, X., Muszynski, M.G. & Danilevskaya, O.N. The FT-like *ZCN8* gene functions as a floral activator and is involved in photoperiod sensitivity in maize. *Plant Cell* **23**, 942–960 (2011).
42. Danilevskaya, O.N., Meng, X., Hou, Z., Ananiev, E.V. & Simmons, C.R. A genomic and expression compendium of the expanded *PEBP* gene family from maize. *Plant Physiol.* **146**, 250–264 (2008).
43. Camus-Kulandaivelu, L. *et al.* Patterns of molecular evolution associated with two selective sweeps in the *Tb1–Dwarf8* region in maize. *Genetics* **180**, 1107–1121 (2008).
44. Dong, Z. *et al.* A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. *PLoS One* **7**, e43450 (2012).

ONLINE METHODS

FOAM mating design and phenotypic evaluation. The mating design for the maize landrace FOAM population consisted of crossing each male landrace individual from the 4,500 accessions to single-cross hybrid females of matching altitude-adaptation class. The reason to cross landrace individuals to single-cross hybrids was to produce abundant progeny for multiple cycles of evaluation across years and locations. The reason for matching landrace and hybrid adaptation was to avoid the negative confounding effect of lack of adaptation during field experiments. For the FOAM approach, progeny is evaluated across multiple trials and locations to estimate the genetic effects corresponding to the landrace parent. Alternative approaches could be to self-pollinate the landrace individuals, cross them to a single or a few inbred lines representing the adaptation classes of landraces, or even generate doubled haploids[18]. Although those strategies could, in principle, result in higher statistical power, in outbreeding populations, such as maize landraces, this could lead to very few progenies for performing multiple evaluation cycles due to inbreeding depression, and several alleles could be lost due to linkage with deleterious lethal variants. In a pilot experiment, we observed several issues for using a doubled-haploid approach, first through asynchrony in flowering time between landraces and inducer lines, and later due to the very low rates of haploid induction (1–11%) and doubling rate (1.2–21.0%), partly due to the high rate of seedling mortality.

For our FOAM experiment, which was based on altitude adaptation, different subsets of progeny per trial were evaluated over 2 years in 13 locations across Mexico, with the main constraint for the number of progeny evaluated per location being field space. There were between 288 and 1,928 accessions per trial, with an average of 834. We used an augmented row–column design, which included systematic checks in field rows and columns[45], that allowed for accounting of the field effects in the estimation of genetic effects. For each trial, each experimental row contained between 9 and 25 progeny plants (**Supplementary Table 1**). For the FOAM strategy, this meant that each of the 10–25 individuals per landrace accession contained, at each locus, one of the landrace alleles plus the corresponding hybrid allele. In other words, the phenotypic effect of the two gametes of the landrace individual is observed 10–25 times per location. The replicated progeny evaluation across multiple locations means that the phenotypic effect of each landrace's alleles is observed tens to hundreds of times, allowing for the accurate estimation of their additive genetic effects. For our FOAM approach, over half of the accessions were replicated in five trials, with a maximum value of 13 trials per accession and a minimum of 1 (**Supplementary Fig. 9**). Given the absence of genotypic data for each segregating progeny, we have good power to estimate the sum of additive effects; however, we did not estimate or test the dominance or epistasis effects. Furthermore, the lack of balanced replication limits the ability to accurately estimate genotype-by-environment effect.

Flowering time was measured in each trial following the maize standard, i.e., the number of days from planting until half of the individuals within a plot displayed silks for female flowering or anthers in half of the central spike for male flowering.

Analysis of phenotypic data. To estimate the breeding values of the landrace accession parent, for each trial a mixed linear model was fitted using a restricted maximum-likelihood method, in ASREML (v. 3.0), using the progeny's calendar days to male or female flowering as a response variable. The models included fixed effects for checks, tester, and hybrid and a random effect of accession in a complete nested model. In addition, the model included the random effect of row and column using an autoregressive model of an order of 1 in row and columns to control experimental noise as a product of field variation. All random effects were considered independent from each other. The model used can be expressed as follows:

$$y_{ijkml} = \mu + \gamma_i + \lambda_j + \alpha_k + \beta_{l(k)} + \delta_{m(kl)} + \varepsilon_{ij}$$

where y_{ijkml} is the response variable, μ is the overall mean, γ_i is the effect of the i-th row, $\gamma_i \sim N(0,\sigma_1^2)$, λ_j is the effect of the j-th column, $\lambda_j \sim N\left(0,\sigma_2^2\right)$, α_k is the effect of the k-th group, $k = 1,\ldots,$ K, K + 1 (if $k \leq$ K the group is a check), the group K + 1 is the average of testers, $\beta_{l(k)}$ is the effect of the l-th tester in group K + 1, $\delta_{m(kl)}$ is the effect of the m-th accession in the tester k in group K + 1, $\delta_{m(kl)} \sim N(0,\sigma_{kl}^2)$, and ε_{ij} is the experimental error.

For the experimental error we assumed the following distribution:

$$\varepsilon \sim N(0,Ó) \text{ with } \Sigma = \Sigma_r \otimes \Sigma_c$$

where

$$Ó_r = \begin{bmatrix} 1 & \rho_r^1 & \rho_r^2 & \cdots & \rho_r^{r-2} & \rho_r^{r-1} \\ \rho_r^1 & 1 & \rho_r^1 & \cdots & \rho_r^{r-3} & \rho_r^{r-2} \\ \rho_r^2 & \rho_r^1 & 1 & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_r^{r-2} & \rho_r^{r-3} & \rho_r^{r-4} & \cdots & 1 & \rho_r^1 \\ \rho_r^{r-1} & \rho_r^{r-2} & \rho_r^{r-3} & \cdots & \rho_r^1 & 1 \end{bmatrix}$$

and

$$Ó_c = \begin{bmatrix} 1 & \rho_c^1 & \rho_c^2 & \cdots & \rho_c^{c-2} & \rho_c^{c-1} \\ \rho_c^1 & 1 & \rho_c^1 & \cdots & \rho_c^{c-3} & \rho_c^{c-2} \\ \rho_c^2 & \rho_c^1 & 1 & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_c^{c-2} & \rho_c^{c-3} & \rho_c^{c-4} & \cdots & 1 & \rho_c^1 \\ \rho_c^{c-1} & \rho_c^{c-2} & \rho_c^{c-3} & \cdots & \rho_c^1 & 1 \end{bmatrix}$$

Genotyping. Accessions that were used as male parents were genotyped using genotyping-by-sequencing (GBS)[21], with ApeKI as the restriction enzyme, to a replication level of ~96 individuals per sequencing plate. Approximately 8×10^9 sequencing reads were generated using an Illumina HiSeq for the landrace accessions, and sequence reads were analyzed jointly with another 40,000 maize lines as part of the GBS build 2.7 using TASSEL[46]. On average, the missing data per individual and per site was 0.5 (**Supplementary Fig. 10**). By comparing the distributions of depth of the called sites, the median number of reads per site was 2 (**Supplementary Fig. 11**). For association analyses, missing data were imputed using BEAGLE4 (ref. 22), which has been shown to yield the best current accuracies in maize heterozygous material[47]. We observed an imputation accuracy with an $R^2 = 0.68$, with no missing data after imputation. Following imputation, SNPs were filtered for minor allele frequency >1%, resulting in approximately 500,000 bi-allelic markers across the genome.

Diversity assessment. For the Mantel test[23], we calculated the pairwise Euclidean-distance matrix based on the geographical data from the accessions (latitude, longitude and altitude). We estimated and tested separate Euclidean-distance matrices for altitude, latitude and longitude, as well as one joint matrix. The genetic-distance matrix was estimated from a genome-wide random sample of 30,000 non-imputed markers using TASSEL. The correlation between the genetic matrix and either the joint or altitude Euclidean-distance matrices was 0.46, with P-value estimation based on 1,000 permutations. Latitude and altitude had correlations of 0.04 and 0.07 with genetic distance, respectively. Mantel tests were performed using the R library 'ade4' (ref. 48). MDS was performed on the genetic-distance matrix using the 'cmds' function in R.

Recombination. For estimation of LD, phased markers with accurate heterozygote calls are required. The distribution of depth of coverage (**Supplementary Fig. 11**) showed that most GBS markers in our panel had a depth of exactly 1, with half of the markers having a depth equal to or greater than 2. In the absence of phase or sufficient depth for all sites, we estimated an LD-like statistic using the non-imputed SNP markers. To account for lack of phase information and heterozygote under-calling, our LD-like statistic estimates the correlation between homozygous markers at 100-site non-overlapping

windows with the LD function on the software TASSEL. For GBS markers, we found that this was the smallest window size with informative correlations. We were interested in significant increases in LD that affected large regions on multiple individuals across populations. Therefore, we aggregated the correlations into 1-Mb regions by taking the median value. For comparing the LD and recombination values, we estimated the correlation at 1-Mb sliding windows between (i) the $\log_{10}$(median LD estimate), (ii) the log value for the median cross-over probabilities estimated using the American and Chinese NAM populations[24], and (iii) the log of the median population recombination rates (rho) estimated both for improved lines and landraces Hapmap2 project[24]. Our LD estimates displayed a negative correlation with gene density ($r = -0.57$) and NAM cross-over probability[8] ($r = -0.45$). We observed a modest negative correlation ($r = -0.33$) between our LD-like statistic and a population genetic estimate of historical recombination (rho)[24,25]. High-LD regions were defined based on the change in slope of the global median LD (**Supplementary Fig. 12**). High-LD regions, therefore, were those segments that had a median LD > 0.01. In total, there were 256 high-LD regions encompassing 7.8% of the genome. Of the candidate genes, only *PhyB1* (phyochrome B1), *Gl15* (Glossy15) and *ZCN13* were in the high-LD set and were, therefore, excluded from further gene-level analyses.

Flowering time genome-wide association and genomic prediction. Association analysis was performed in two steps for all trials using a linear mixed model[49,50]. For each trait (days to male and female flowering) two models were fitted, one with the trait 'best linear unbiased predictions' (BLUPs) as a response variable and another one with the standardized values of the same BLUPs. Although the use of cumulative heat units in the form of growing degree days can be used to standardize crop phenology data sets across locations, the standardization used consisted of subtracting the mean value of the trial and dividing by the corresponding s.d. This was done to assess the consistency of the results given the uneven variances for the trait across the various trials. Correlation between *P* values from both GWAS models was 0.84.

The first step models included the fixed effects for trial (categorical), population structure in the form of ten MDS weights (numerical) that together explained around 13% of the genetic variances and 10.6% of the phenotypic variances, and the effect of the hybrid used as parent for each accession's cross. The random effect of relatedness was added to both models in the form of a kinship matrix. The kinship matrix was estimated using the same subset of SNPs as the MDS weights. The mixed model was fit using the R package EMMREML. The vectors containing the residuals after fitting the first models were fitted in the second step models as a response variable for the single-marker analysis. Models were fitted using R, with the marker nested within the levels of the trial.

The model equation used was

$$Y_{ijk} = \mu + T_i + H_{ij} + Q_{ijk} + Z_u + \varepsilon_{ijk}$$

where Y_{ijk} is the response variable, μ is the overall mean, T_i is the effect of the i-th trial, H_{ij} is the effect of the j-th tester at each i-th trial, Q_{ijk} is the population-structure effect containing ten weights from MDS, Z_u, where u is a vector of size n (number of individuals) for unknown random polygenic effects, which have a distribution with mean of zero and covariance matrix of $G = 2K\sigma^2_a$ where K is the co-ancestry matrix with element k_{ij} ($i,j = 1,2,\ldots n$) calculated from 30,000 random SNPs, and ε_{ijk} is the vector containing the residual error.

In the second step of the association model, the residuals from the first model were fitted as a response variable in the following model

$$Y_i = S[t] + \varepsilon_i$$

Where Y_i is the residual from the previous model and S is the SNP effect that is nested within trial t. The model uses an F-test for the null hypothesis stating that the effect of each SNP is 0 in all trials. The alternative hypothesis is that the SNP has an effect on any trial. The reason for testing this hypothesis is that the effect of each SNP can, and often does, change on value and direction. This is a consequence of the segregation of alleles at different frequencies across all trials, as well as the change of phase between the tested SNPs with the causal polymorphisms. We observed significant deviation from the expected distribution of *P* values (**Supplementary Fig. 13**); therefore, to account for the false discovery rate, we only consider as significant the top 1% of the SNPs based on *P* value, which all had $-\log_{10}(P \text{ values}) > 18$. We reasoned that significance at candidate genes would depend on local LD and genotype coverage; therefore, a higher proportion of significant SNPs around candidate genes would be indicative of association at the gene itself rather than at the entire LD block or because of higher genotype coverage. On that account, we looked at significantly associating SNPs within a region 50-kb upstream and downstream of candidate genes, and assigned SNPs to the nearest genes using the R package GenomicRanges[51].

Genome-wide prediction was performed with the software GAPIT[52]. The models were run for each trial, and accuracy was measured by performing fivefold cross-validation in ten replicates for each trial. Two models were run for each trait and trial. One model used a kinship matrix that was estimated with one SNP for each of the 888 associated genomic regions, another model used 714 evenly distributed random SNPs, and a third model used 30,000 random SNPs for the estimation of the kinship matrix. All models included ten MDS weights to account for population structure.

Genome-wide association with altitude and latitude. We were interested in understanding the genomic regions that contributed both to flowering-time variation and to altitude and latitude adaptation. We performed genome-wide association using a generalized linear model with altitude and latitude as response variables and markers, filtered at 1% frequency, as explanatory variables. Consistent with other mapping studies using geography as a response variable in association studies, models with covariates for population structure in the form of principal-component weights, and mixed linear models including either only a kinship matrix or both the kinship and principal component weights, showed very limited association (**Supplementary Fig. 14**). This was mainly due to the high covariance between local adaptation and population structure, given that selection for local adaptation leads to population structure. This means that models accounting for local adaptation decrease the false-positive rate but also significantly increase the false-negative rate. To reduce the false-positive rate from the results of the generalized linear model and to establish a biologically meaningful significance threshold using additional independent information, we estimated the overlap rate using the most significant flowering time GWAS SNPs. The overlap rate was defined as the set of overlapping SNPs between the shared male and female top flowering-time SNPs and either altitude or latitude, divided by the union of the sets across significance thresholds. In other words, assuming that flowering-time-associating SNPs represented our current best candidate for true positives, the overlap rate was used to maximize true positives at *P*-value threshold values that minimized false negatives. Therefore, the overall rate was estimated for the percentiles ranging between 0.001 and 0.010. For example, for the first quantile threshold (0.001), the overlap rate for altitude corresponds to the number of SNPs at the top 0.001 quantile (around 500) that overlap with the top 5,000 flowering-time SNPs, divided by the sum of the flowering-time SNPs and the altitude SNPs at that quantile. The significance thresholds chosen from the overlap rate results (**Supplementary Fig. 5**) were the 0.005 percentile for altitude (the top 0.5% of the associating SNPs, $-\log_{10}(P \text{ values}) > 208.2$; **Supplementary Table 6**) and the 0.01 percentile for latitude (top 1% of associating SNPs $-\log_{10}(P \text{ values}) > 61.63$; **Supplementary Table 7**). It can be observed in **Supplementary Figure 5** that across the same quantile values, altitude has significantly higher overlap with flowering-time-associating SNPs, as compared to that with latitude, probably due to the landraces coming mostly from non-photoperiod-inducing locations. SNPs are associated with gene models based on distance using the R package GenomicRanges[51], and because GBS SNPs were enriched at gene regions, the average distance to gene model was 0. For each gene, the most significant SNP was chosen based on distance in base pairs. Heritability estimates were 0.88 for altitude and 0.85 for latitude, estimated using LDAK[53] with a single kinship matrix, estimated with all the imputed markers, and the matrix was estimated from the algorithm implemented in GCTA[54].

Analyses of structural variants. To infer the underlying haplotypes for the centromeres of chromosomes 3,5 and 6, as well as *Inv4* and the high-LD region

doi:10.1038/ng.3784

on chromosome 3, we first estimated a genetic distance matrix for each locus using the non-imputed markers. The distance matrices were then analyzed using MDS. In the complete absence of recombination, the dimensional reduction of the genetic-distance matrix yielded distinct clusters that corresponded to the homozygote alleles and the corresponding heterozygotes. The centromere of chromosome 5 segregates in the landraces with three distinct homozygous haplotypes and their corresponding heterozygote pairs. The region around the centromere of chromosome 6 was 12 Mb in size, and included the centromere and a large pericentromeric region that expanded out in both directions; it displayed a similar pattern to the centromere of chromosome 5. However, allele calls were not done due to incomplete clustering of homozygote and heterozygote classes, probably reflecting recombinant haplotypes. The centromere of chromosome 3 displayed a more complex pattern of distance than the other two associating centromeres, likely due to the presence of more than three segregating haplotypes. For *Inv4*, we observe two distinct alleles and the heterozygote. We observed that the allele was fixed in many of CIMMYT-improved lines (**Supplementary Table 8**), including those used as parents for the highland test crosses in the present experiment.

Expression across tissues. We used the transcription data from the maize atlas[55] for the following 11 tissues: embryo 16 d after pollination, endosperm 16 d after pollination, primary root 6 d after silking, tip of stage 2 leaf at the V5 plant stage, base of stage 2 leaf at the V5 plant stage, 13th leaf at the V9 stage, 13th leaf at the R2 stage, silk, anthers, immature cob at the V18 stage, 4th internode at the V9 stage, and the stem and shoot apical meristem at the V4 stage. We used the standardized expression values and estimated, for each gene, the tissue in which each gene had the highest expression. We then performed a chi-squared test comparing the global expression pattern for each tissue with either the the list of candidate genes or the list of all associating genes. *P* values were adjusted using the false discovery rate method, with candidate genes showing an enrichment at the immature cob with an adjusted *P* value of 2.42×10^{-18}.

Code availability. All of the code written is available upon pull request from github. The EMMREML package used for the mixed model is available at http://cran.r-project.org/web/packages/EMMREML/index.html. For the map in **Figure 2**, the package maps (maps: Draw Geographical Maps. R package version 3.1.1.) was used (https://CRAN.R-project.org/package=maps). The original Maps S code was written by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements were made by Thomas P Minka and Alex Deckmyn. (2016). **Figures 2** and **3** were produced using the R ggplot2 package[56] (http://ggplot2.org).

Data availability statement. The data that support the findings in the study are available from the following repositories: GBS non-imputed markers: http://hdl.handle.net/11529/10034; GBS imputed markers: http://hdl.handle.net/11529/10035; Phenotypic and passport data can be accessed upon registration at http://germinate.seedsofdiscovery.org/maize/.

45. Federer, W.T. & Crossa, J.I. 4 screening experimental designs for quantitative trait loci, association mapping, genotype-by-environment interaction and other investigations. *Front. Physiol.* **3**, 156 (2012).
46. Glaubitz, J.C. *et al.* TASSEL–GBS: a high-capacity genotyping-by-sequencing analysis pipeline. *PLoS One* **9**, e90346 (2014).
47. Swarts, K., Li, H., Romero Navarro, J.A. & An, D. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *Plant Genome* **7**, doi:10.3835/plantgenome2014.05.0023 (2014).
48. Dray, S. & Dufour, A.B. The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20 (2007).
49. Aulchenko, Y.S., de Koning, D.-J. & Haley, C. Genome-wide rapid association using mixed model and regression: a fast and simple method for genome-wide pedigree-based quantitative-trait-loci association analysis. *Genetics* **177**, 577–585 (2007).
50. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
51. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
52. Lipka, A.E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
53. Speed, D. & Balding, D.J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557 (2014).
54. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex-trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
55. Sekhon, R.S. *et al.* Genome-wide atlas of transcription during maize development. *Plant J.* **66**, 553–563 (2011).
56. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2009).

Corrigendum: A study of allelic diversity underlying flowering-time adaptation in maize landraces

J Alberto Romero Navarro, Martha Wilcox, Juan Burgueño, Cinta Romay, Kelly Swarts, Samuel Trachsel, Ernesto Preciado, Arturo Terron, Humberto Vallejo Delgado, Victor Vidal, Alejandro Ortega, Armando Espinoza Banda, Noel Orlando Gómez Montiel, Ivan Ortiz-Monasterio, Félix San Vicente, Armando Guadarrama Espinoza, Gary Atlin, Peter Wenzl, Sarah Hearne & Edward S Buckler
***Nat. Genet.*; doi:10.1038/ng.3784; corrected online 20 February 2017**

In the version of this article initially published online, the name of author Martha Willcox was misspelled as Martha Wilcox. The error has been corrected in the print, PDF and HTML versions of this article.