the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) **113,** 1109–1121 doi: 10.1111/tpj.16123

RESOURCE

A common resequencing-based genetic marker data set for global maize diversity

Marcin W. Grzybowski[1,2,3,*], Ravi V. Mural[1,2], Gen Xu[1,2], Jonathan Turkus[1,2], Jinliang Yang[1,2] and James C. Schnable[1,2,*]

[1]*Center for Plant Science Innovation, University of Nebraska–Lincoln, Lincoln, Nebraska, USA,*
[2]*Department of Agronomy and Horticulture, University of Nebraska–Lincoln, Lincoln, Nebraska, USA, and*
[3]*Department of Plant Molecular Ecophysiology, Institute of Plant Experimental Biology and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland*

Received 28 October 2022; revised 20 January 2023; accepted 23 January 2023; published online 27 January 2023.
*For correspondence (e-mails marcin.grzybowski@uw.edu.pl; schnable@unl.edu).

SUMMARY

Maize (*Zea mays* ssp. *mays*) populations exhibit vast ranges of genetic and phenotypic diversity. As sequencing costs have declined, an increasing number of projects have sought to measure genetic differences between and within maize populations using whole-genome resequencing strategies, identifying millions of segregating single-nucleotide polymorphisms (SNPs) and insertions/deletions (InDels). Unlike older genotyping strategies like microarrays and genotyping by sequencing, resequencing should, in principle, frequently identify and score common genetic variants. However, in practice, different projects frequently employ different analytical pipelines, often employ different reference genome assemblies and consistently filter for minor allele frequency within the study population. This constrains the potential to reuse and remix data on genetic diversity generated from different projects to address new biological questions in new ways. Here, we employ resequencing data from 1276 previously published maize samples and 239 newly resequenced maize samples to generate a single unified marker set of approximately 366 million segregating variants and approximately 46 million high-confidence variants scored across crop wild relatives, landraces as well as tropical and temperate lines from different breeding eras. We demonstrate that the new variant set provides increased power to identify known causal flowering-time genes using previously published trait data sets, as well as the potential to track changes in the frequency of functionally distinct alleles across the global distribution of modern maize.

Keywords: maize, genetic markers, GWAS, whole-genome resequencing, diversity panel, natural genetic variaton.

Linked article: This paper is the subject of a Research Highlight article. To view this Research Highlight article visit https://doi.org/10.1111/tpj.16163.

INTRODUCTION

The degree of DNA sequence diversity observed in *Zea mays* (maize) populations exceeds that of humans, most genetic model species and many wild plants (Buckler et al., 2006). This diversity includes not only small-scale variation – single-nucleotide polymorphisms (SNPs) and insertions/deletions (InDels) – but also copy-number and presence/absence variation (Swanson-Wagner et al., 2010). Scoring large populations of maize for common sets of segregating DNA sequence polymorphisms (markers) is a key step in a range of research approaches to identify targets of selection (Hufford et al., 2012; Wang et al., 2020), inferring past demographic events and geographic diffusion (Da Fonseca et al., 2015; Kistler et al., 2018; Swarts et al., 2017), and linking genotype to phenotype (Mural et al., 2022). Early approaches to scoring common sets of genetic markers across large maize populations have targeted thousands to hundreds of thousands of known markers, in the case of arrays (Ganal et al., 2011; Unterseer et al., 2014). Array-based genotyping allowed the wide reuse and combination of independent data sets generated using the same array platform and, in cases where

common probes were retained, between platforms. Reductions in the cost of DNA sequencing have enabled sequencing-based strategies to be combined with marker discovery and scoring in a single step (Elshire et al., 2011; Romay et al., 2013). This change reduced the substantial ascertainment bias present in many array-based genetic marker data sets. However, combining marker discovery and scoring into a single step created new barriers to combining data sets. It was not possible to target specific known markers to enable interoperability between genotyping platforms. Different approaches to reducing the proportion of the genotype sequenced targeted different subsets of the genome for sequencing. Even when the same region was sequenced in two studies, differences in allele frequency, SNP-calling software pipelines or stochastic distributions of read depths might result in the same marker being identified and scored in one data set and absent from another. Sequencing technology has continued to improve and so costs have continued to decline. Whole-genome resequencing is now economically viable even for populations of hundreds of maize genotypes. This removes the barrier of generating sequence data for largely non-overlapping sites that was present in earlier sequencing-based strategies. However, combining marker data sets across different studies remains challenging, as very different sets of markers will be discovered and pass quality filtering in different populations and/or when using different bioinformatics pipelines.

Identifying a common set of genetic variants in maize is challenging, and the optimal set of lines to use in defining a marker set is likely to depend upon the question of interest. Maize was domesticated from a wild progenitor teosinte (*Zea mays* ssp. *parviglumis*) 9000–10 000 years ago in south-west Mexico (Matsuoka et al., 2002; Piperno et al., 2009), with substantial gene flow from at least one other teosinte (*Zea mays* ssp. *mexicana*) (Chen et al., 2022; Van Heerwaarden et al., 2011). After domestication, maize spread across North and South America (Da Fonseca et al., 2015; Kistler et al., 2018; Swarts et al., 2017). Maize, almost certainly of Caribbean origin, was first cultivated in southern Europe in 1493 and was growing in Germany by 1539 (Tenaillon & Charcosset, 2011). By 1555, substantial maize cultivation was already being recorded in Henan, China (Ho, 1955). Therefore, maize was already cultivated on at least four continents in the mid 16th century. Tropical maize varieties that flower under short-day conditions, making them unsuitable for cultivation in regions with lethal frosts, retain many alleles and haplotypes not found in temperate populations (Hung et al., 2012). Breeding efforts in the USA, Europe and China focus on temperate-adapted cultivars that are less photoperiod sensitive than tropical maize. In the USA, hybrid production focuses on three heterotic groups (stiff stalk, non-stiff stalk and iodent), in Europe, many hybrids are generated from crosses between the flint and dent heterotic groups, whereas in China, the Huangzaosi group was also used alongside stiff stalk, non-stiff stalk and iodent (Wang et al., 2020). As a result, different research groups studying quantitative genetic variation, domestication, adaptation or crop improvement have selected different sets of inbred lines, open-pollinated landraces or maize wild relatives drawn from populations in different parts of the globe.

The maize HapMap2 project was motivated in part by understanding the changes in genetic diversity associated with maize domestication and improvement (Chia et al., 2012). The study identified more than 55 million total variants from an average of 4x depth of 103 samples, including 83 individuals representing domesticated maize and 20 individuals drawn from wild relative populations aligned with B73_RefGen_V1 (Chia et al., 2012; Hufford et al., 2012). A project focused on understanding the history and demography of the initial introduction of maize to Europe identified 22.3 million SNPs relative to the B73_RefGen_V2 genome by resequencing 67 maize samples originating in the Americas (n = 37) and Europe (n = 30), to an average depth of 18× (Brandenburg et al., 2017). Given the focus on the introduction of maize to Europe, this study focused primarily on maize lines originating in western (18) and central (11) Europe, with one line sourced from eastern Europe. Another study focused on the pre-Colombian demographic history of maize resequenced 35 maize landraces and wild relatives from the Americas to a median depth of 28× and identified 49.5 million SNPs via alignment with the B73_RefGen_V3 reference genome (Wang et al., 2017). A study of maize domestication and improvement in South America generated data from 49 living and archeological maize samples and generated a new SNP set by aligning data from these new samples and data from 70 published maize data sets with the B73_RefGen_V4 reference genome (Kistler et al., 2018). The resequencing of 521 diverse maize inbred lines to an average depth of 20× identified 11.5 million variants as part of an effort to link structural variation in the genome to changes in gene expression and phenotypic outcomes (Yang et al., 2019). A comparative analysis of phenotypic and genetic changes associated with the breeding effort in different temperate breeding programs generated resequencing data from 350 maize inbred lines from China (187) and the USA (163) sequenced to a median depth of 12× and identified more than 29 million genetic markers relative to the B73_RefGen_V3 reference genome (Wang et al., 2020). An effort to quantify SNP and transposon insertion diversity within an association panel used for genome-wide association studies identified approximately 2.4 million SNPs and 0.45 million segregating transposon associations across a panel of approximately 500 temperate adapted maize lines (Qiu et al., 2021; Renk et al., 2021). Finally, a recent study of

genus-wide genetic variation in maize identified approximately 65 million SNPs and approximately 8 million InDels by generating sequencing data to an average depth of 22× from 239 accessions of wild relatives (Chen et al., 2022; Gui et al., 2022), in combination with the diversity panel of approximately 500 maize lines sequenced by Yang et al. (2019). The largest scale of these efforts to date is likely the aggregate analysis of 1218 maize lines as part of the maize HapMap3 project representing global maize diversity; however, higher sequencing costs at the time of this study resulted in lines being resequenced to a median depth of 2× (Bukowski et al., 2018). Although different approaches have been used to identify genetic variants in these different studies, the most common alignment tool used has been BURROWS–WHEELER ALIGNMENT (BWA) (Li, 2013) and the most common variant caller used has been GENOME ANALYSIS TOOLKIT (GATK) (Poplin et al., 2018), for example see (Chen et al., 2022; Liang et al., 2021; Wang et al., 2017; Wang et al., 2020).

Here we sought to update and expand the reference set of segregating diversity in maize by incorporating published high-coverage resequencing data from maize lines originating on six continents, including resequencing data from lines relative to maize domestication and improvement, including wild relatives, tropical landraces and archeological maize samples, as well as maize wild relatives, and to further improve the resolution and mapping power for maize genome-wide association studies conducted in the temperate midwest through the resequencing of an additional 239 maize lines, including 228 lines from the Wisconsin Diversity panel not previously resequenced and 11 Eastern European lines. To ensure the greatest degree of reusability and forward compatibility, we employed the B73_RefGen_V5 maize reference genome (Hufford et al., 2021) and, in addition to raw and filtered SNP files, we are releasing GATK GenomicsDB datastores so that these same 1515 lines can be incorporated into future high-coverage maize resequencing efforts without the need to reprocess and realign the sequence data.

RESULTS AND DISCUSSION

Sequence variation across the genome of maize

Sequence data from 1276 maize individuals generated as part of eight different studies (Brandenburg et al., 2017; Bukowski et al., 2018; Chen et al., 2022; Chia et al., 2012; Kistler et al., 2018; Qiu et al., 2021; Unterseer et al., 2014; Wang et al., 2017; Wang et al., 2020) were retrieved from the European Nucleotide Archive. To this public data set, we added data from *de novo* resequencing of 228 maize inbred lines, which are part of the expanded Wisconsin Diversity Panel (Mazaheri et al., 2019) but were not resequenced as part of previous efforts (Bukowski et al., 2018; Qiu et al., 2021). A specific goal of this effort was to provide a high-density genome-wide set of markers for the Wisconsin Diversity Panel. In order to accomplish this goal, we included resequencing data of lower depth for 144 members of the Wisconsin Diversity Panel that were resequenced as part of maize HapMap3 (Bukowski et al., 2018). In addition, to increase the representation of undersampled maize subpopulations, we also chose to include sequence data from 70 other accessions included in HapMap3, primarily from tropical lines.

An average of 155 million reads were generated for each of these inbred lines, corresponding to an average sequencing depth of approximately 22×. Additionally, a set of 11 maize inbred lines from Poland, representative of Eastern Europe, a region only modestly represented among previous maize resequencing efforts, were resequenced here to an average depth of approximately 35×. Those lines were used in previous studies on the cold response in maize (Grzybowski et al., 2019; Sowiński et al., 2005). The total set of 1515 maize accessions included wild relatives, archeological samples, modern open-pollinated varieties and inbred lines from both public and private sector breeding efforts, representing maize lines originating in or developed over six continents (Table S1 and S3).

Aligning sequence data from each of these accessions with the maize B73_RefGen_5 reference genome (Hufford et al., 2021) and applying recommended filtering criteria from GATK resulted in the identification of 365 611 965 potential DNA sequence polymorphisms. This number is substantially higher than the approximately 83 million variants identified in the maize HapMap3 project, one of the largest surveys of maize genetic diversity conducted to date, incorporating data from 1218 maize accessions (Bukowski et al., 2018). However, it should be emphasized that HapMap3 utilized a different variant-calling pipeline and that the median sequencing depth of samples in that study was approximately 2×. A more recent study that examined the genetic differentiation of male and female heterotic groups in maize, using resequencing data from 1604 maize inbred lines, primarily from China and the USA, resequenced to an average depth of approximately 7.5×, identified roughly 242 million DNA sequence polymorphisms (Li et al., 2022).

Second-stage quality filtering (based on allele number, missing data rates, sequence depth and excess heterozygosity, see the Experimental procedures) resulted in a smaller set of 46 054 265 higher-confidence variants, including 43 296 332 SNPs and 2 757 933 InDels (Figure S1). The median total sequencing depth for higher-confidence variants was 17 365 (Figure S3), corresponding to an average sequence depth of 11.5 reads per site per individual. Concordance rates for SNP calls among the 26 nested association mapping (NAM) founder parents (Hufford et al., 2021) and SNP calls reported as part of the

de novo sequence assembly of these parents ranged from 92% to 99%, with a mean value of 98% (Table S2). Among these higher-confidence variable sites, the median accession was genotyped as heterozygous 2.8% of the time. However, the per-accession heterozygosity rates varied significantly across groups (Figure S4). Heterozygous calls were less common in centromeric regions (Figure S5). Groups expected to consist primarily of inbred lines, such as those classified as belonging to the stiff stalk, non-stiff stalk and iodent heterozygous groups typically exhibited per-accession heterozygosity values of <3%. Accessions classified as wild relatives frequently exhibited per-accession heterozygosity values of >10% (Figure S5; Table S1). Inbred lines with unexpectedly high heterozygosity were not removed from the final data set; however, they should be used with caution as these may represent contaminated or mislabeled samples.

Although many high-confidence SNPs (41%) and InDels (38%) were rare, defined here as a minor allele frequency of ≤5%, more than 26 million variants were common, defined as a minor allele frequency of >5% (25 154 632 SNPs and 1 704 190 InDels) (Figure 1b,d). Segregating SNPs were more common around pericentromeric regions (Figure 1a), whereas segregating InDels were more frequent on chromosome arms and less frequent in pericentromeric regions (Figure 2c). The relationship between distance from the centromere and SNP or InDel density was extremely weak but statistically significant for each chromosome (Figure S6), similar to the pattern of SNPs and InDels reported in *Sorghum bicolor* (Lozano et al., 2021). Linkage disequilibrium was typically elevated in pericentromeric regions, likely reflecting lower recombination rates in these regions (Figure 1e). The pattern of elevated linkage disequilibrium around the centromere was less prominent on chromosome 10, consistent with previous reports (Romero Navarro et al., 2017). Several other peaks of elevated linkage disequilibrium were observed that did not coincide with the known positions of maize centromeres. One potential explanation is that these peaks may represent large segregating structural variants (Crow et al., 2020); however, validating this hypothesis is beyond the scope of this article. Most high-confidence variants (57%) were located in the intragenic regions, defined as the regions ≥5 Kb from the closest annotated exon. Another 31% of variants were located in regions outside annotated genes but <5 Kb from the closest gene (Figure 1f). Among variants located between the annotated transcription start sites and transcription stop sites of genes, intronic variants were the most abundant (8.6%), followed by the 5′ and 3′ untranslated regions (UTRs) (0.7% and 0.9%) and the coding sequence (CDS) (0.4%). The highest density of variants per kilobase was observed in regions immediately upstream and downstream of the annotated genes, followed by the 5′- and 3′-UTRs (Figure 1g). Consistent with expectation, the CDS sequence contains on average the lowest density of variants.

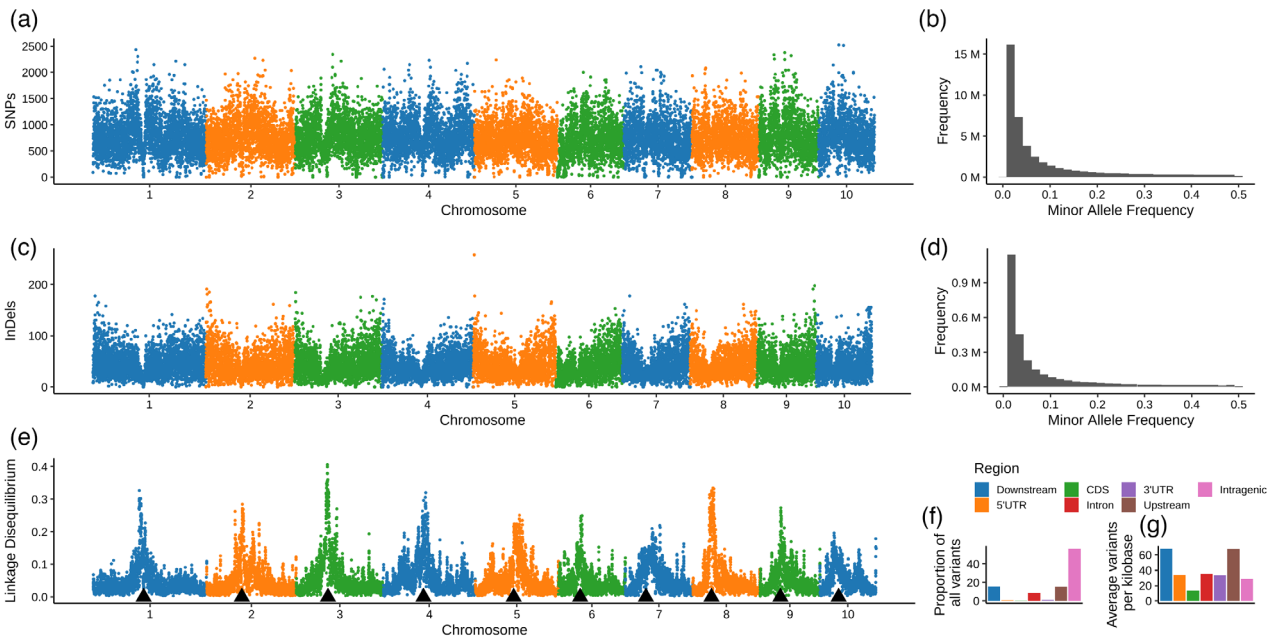

Figure 1. Properties of high-confidence maize genetic variants identified in this study. Distribution of high-confidence and common (MAF >5%, approx. 27 million) single-nucleotide polymorphisms (SNPs) (a) and insertions/deletions (InDels) (c) across each of the 10 maize chromosomes. For both (a) and (c) the genome was divided into non-overlapping 100-Kb windows and SNPs and InDels were counted in each window. Distribution of minor allele frequency of high-confidence (approx. 46 million) SNPs (b) and InDels (d). (e) Mean linkage disequilibrium (LD) value in 100-Kb window calculated with high confidence, and common (MAF >5%) SNPs. Black triangles indicate the centromere position on each chromosome. (f) Percentage and (g) average number per kilobase of variants across the major genic and intergenic regions calculated with the high-confidence variant set.

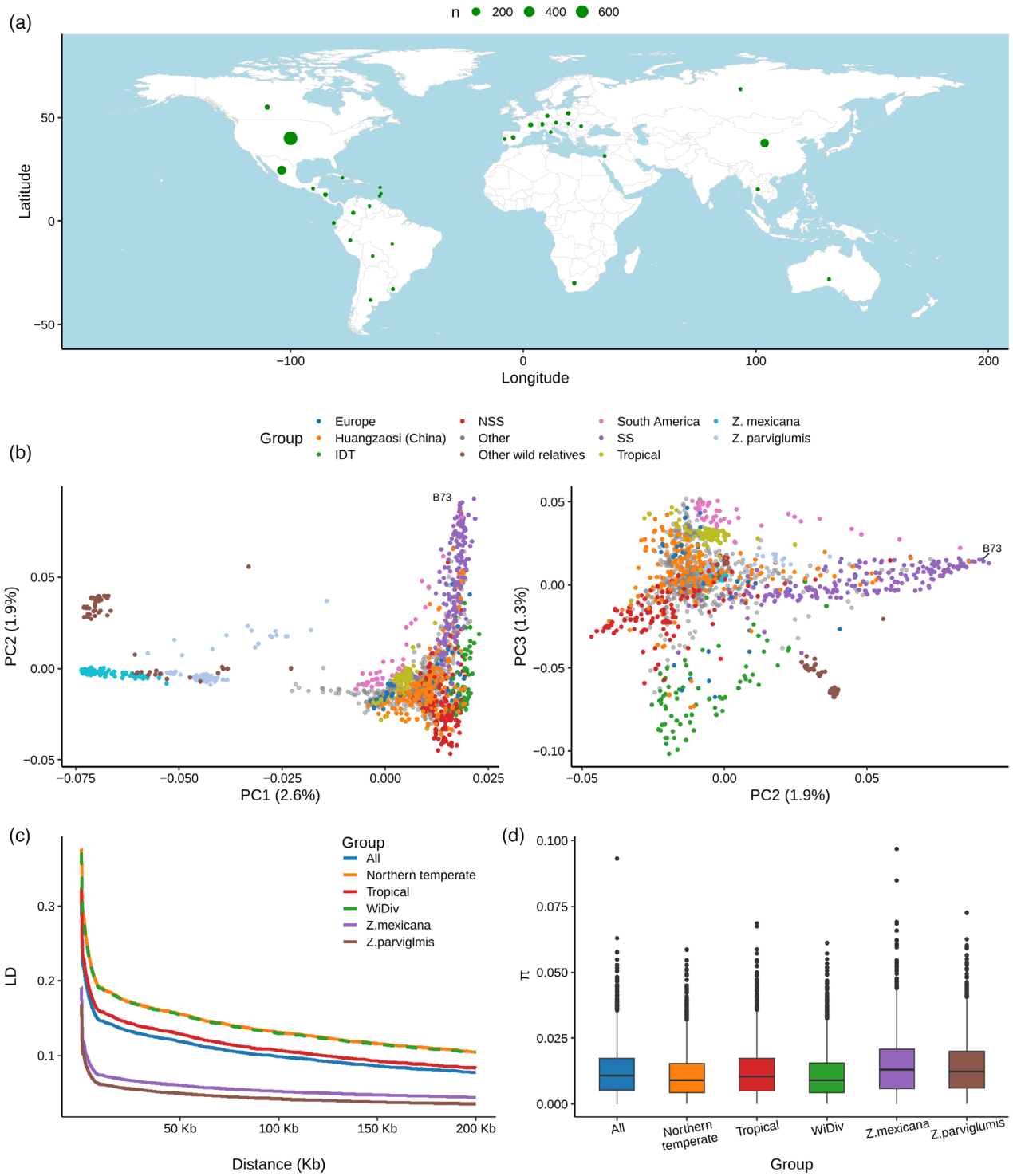

Figure 2. Geographical distribution, population structure, linkage disequilibrium (LD) patterns and nucleotide diversity in maize. (a) Geographical distribution of the country of origin for 1515 maize individuals. (b) First three principal components from PCA analysis of 1515 maize individuals. Each individual was assigned to different groups based on previous literature data. (c) Genome-wide averaged distance of LD decay for six maize groups. (d) Nucleotide diversity for six maize groups. High-confidence common (MAF >5%) variant sets were used for each analysis.

Intra- and interpopulation genetic variation

Of the 1515 maize samples used in this study, 966 were assigned to one of 10 groups through a combination of prior publication data and metadata associated with the United States Department of Agriculture (USDA) Germplasm Resources Information Network (GRIN) records. These 10 groups included three groups of wild relatives: *Z. mays* ssp. *mexicana* (n = 79, hereafter *mexicana*); *Z. mays* ssp. *parviglumis* (n = 84, hereafter *parviglumis*); and other wild relatives (n = 66). Among these samples, four groups were based on geographic origin – tropical (n = 86), South America (n = 48), China (n = 182) and Europe (n = 34) – and three groups were based on a combination of geographic origin and heterotic group – temperate North American stiff stalk (n = 191), non-stiff stalk (n = 127) and iodent (n = 69). The remaining 549 lines were classified as 'other' in the analyses below. Altogether, this set of lines comes from 35 countries across six continents (Figures 2a and S2). Lines tended to cluster based on the group assignment in analyses of population structure conducted using the genetic marker data generated in this study (Figures 2b and S7). The first principal component of variation for genetic marker data roughly corresponded to the division between domesticated maize and maize wild relatives (Figure 2b). The second principal component separates stiff stalk and non-stiff stalk heterotic groups by how closely or not, respectively, they are related to B73, the reference genotype for maize. Finally, the third principal component corresponds to latitudinal geographic distribution, with South American lines at one extreme, followed by tropical and wild populations, then Chinese, European and North American temperate populations, and other wild relatives (Figure 2b).

High-density genetic marker data is useful for both population genetic and quantitative genetic analyses (Mural et al., 2021). Many population genetic analyses require the measurement of plant traits. When trait data are collected in different environments, variance resulting from differences in genotype is confounded with variance resulting from different environments, reducing the statistical power to link genotype and phenotype. Growing and phenotyping large plant populations in common environments can more effectively isolate contributions of genetic variation to phenotypic variation, at least in that specific environment. However, this presents a challenge in capturing global genetic diversity in species such as maize, where different lines are adapted to different environments and may not even be able to successfully complete their lifecycles in environments to which they are not adapted. Efforts to establish common association panels for quantitative genetic analysis in maize, including the Maize Association Panel (MAP) (Flint-Garcia et al., 2005), Shoot Apical Meristem association panel (SAM) (Leiboff et al., 2015) and the Wisconsin Diversity Panel (WiDiv) (Hansey et al., 2011) have required researchers to prioritize the partially contradictory goals of maximizing genetic diversity while also selecting for a set of genotypes that can all grow and successfully complete their life cycles in a single common environment. Based on marker data for 798 genotypes from the WiDiv panel included in this study, linkage disequilibrium decays roughly as fast within the WiDiv panel as with the set of all northern temperate lines (1090 lines defined as all those excluding teosinte, tropical and South America lines), but mostly more slowly than the rate of linkage disequilibrium (LD) decay among all 1515 lines included in this study (Figure 2c). LD decayed fastest among the two maize wild relative populations with the largest number of samples: *mexicana* and *parviglumis*. The median value of π observed for randomly selected intervals in the maize genome within the WiDiv population was 0.0091, similar to the median observed for all temperate lines (π = 0.0090), but lower than the value when calculated for the population of all genotypes included in this study (π = 0.0108; Figure 2d). The difference in π for the overall population is likely to be driven by the inclusion of wild relatives in the overall population, as these populations exhibit elevated π values of *mexicana* (π = 0.0131) and *parviglumis* (π = 0.0125). A previous study indicated that 83% of nucleotide variation from *teosinte* was retained in maize landraces (Hufford et al., 2012). Here we found that WiDiv lines retain 72.8% of the nucleotide variation observed in *parviglumis*, which indicates that a substantial portion of genetic variation is still present in this set of northern temperate lines.

Accurately estimating nucleotide diversity requires accounting for monomorphic (invariant) sites and sequencing depth (Korunes & Samuk, 2021). However, generating and distributing variant call format (VCF) files with monomorphic sites for the entire genome becomes an intractable problem for large numbers of individuals because of the size of such files. Making the GATK GenomicsDB datastores from this project publicly available will allow researchers to generate VCF files for the region of interest and calculate accurate values of nucleotide diversity for this region in their population.

Greater utility of existing trait data as marker density increases

Community association panels are typically reused by many research groups working to study the genetic control of variation in different traits of interest. A recent literature study identified more than 160 distinct trait data sets scored across association panels for North American temperate maize between 2010 and 2020 (Mural et al., 2022). During the past 17 years, the density of publicly available markers for maize association panels has grown from 94

microsatellite markers (Flint-Garcia et al., 2005), to 1536 microarray-based SNP markers (Hansey et al., 2011), to hundreds of thousands of markers scored using genotyping by sequencing (Romay et al., 2013), to approximately one million markers scored using RNA-seq (Leiboff et al., 2015; Mazaheri et al., 2019) and now to typically include tens of millions of markers discovered and scored via whole-genome resequencing (Bukowski et al., 2018; Chen et al., 2022; Li et al., 2022; Qiu et al., 2021; Wang et al., 2020) or a combination of whole-genome resequencing for a subset of lines and imputation from lower density markers for additional lines (Mural et al., 2022; Sun et al., 2022).

We employed a previously published set of female flowering data (days to silking) generated for 752 temperate adapted maize inbreds (Mural et al., 2022) to assess the impact of increased marker density versus direct resequencing (this study) on the outcomes from genome-wide association studies in maize. When using approximately 400 000 markers discovered and scored using RNA-seq (minor allele freqency, MAF > 5% in 752 lines) (Mazaheri et al., 2019), a genome-wide association study identified one statistically significant signal corresponding to the cloned maize flowering-time gene *MADS69* (Liang et al., 2019; Figure 3a). A genome-wide association study conducted using the new, purely whole-genome resequencing-based marker data set generated in this study identified both *MADS69* and *ZCN8* (Figure 3b). The peak in the *ZCN8* region detected when conducting a genome-wide association study using resequencing-based markers was the result of significantly trait-associated non-genic SNPs (Figure S8). SNPs located within genes in this region are in weak LD with the significant variants identified in intergenic space, which is likely to prevent the

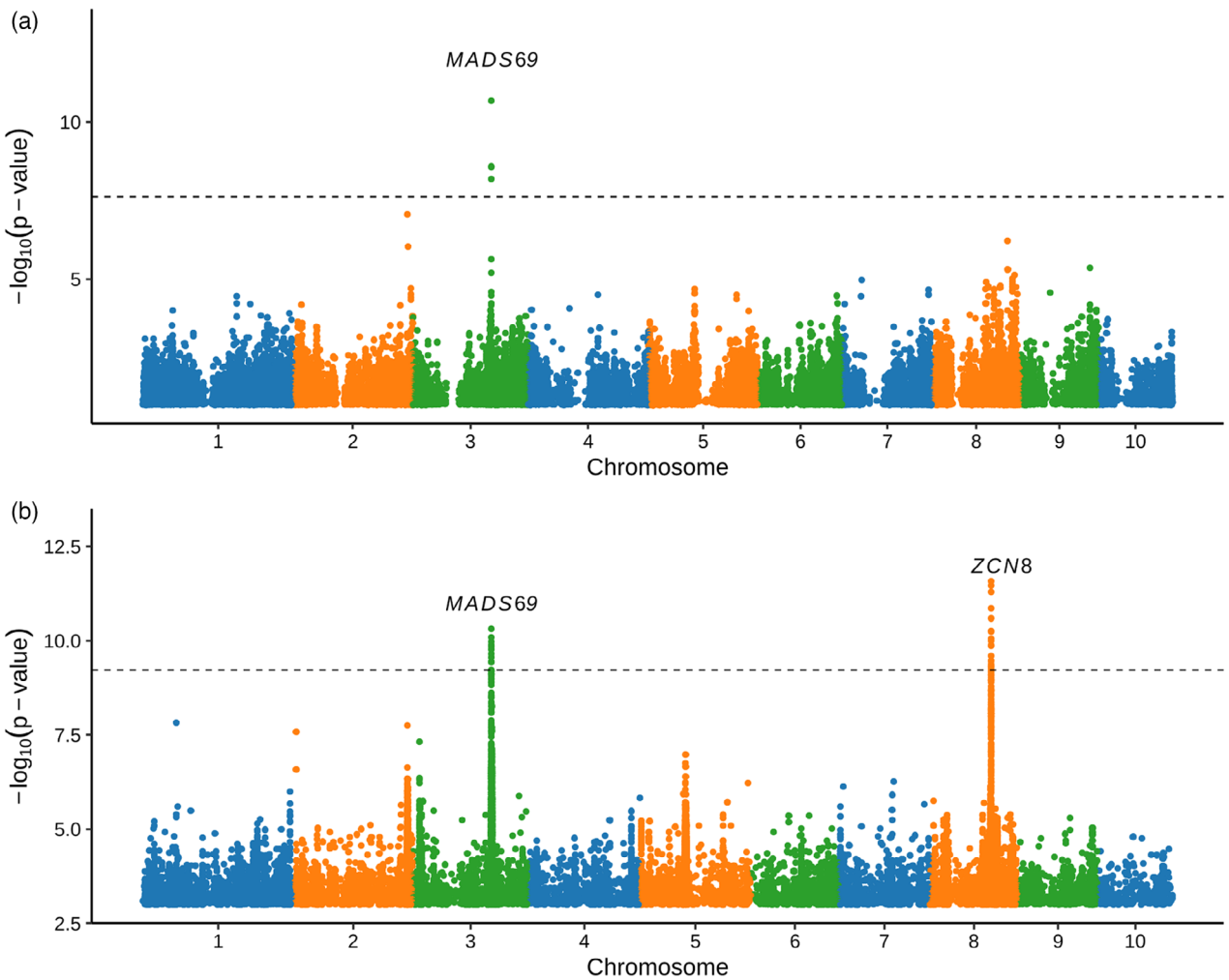

Figure 3. Identification of the candidate genes for flowering time (days to silking) through a genome-wide association study (GWAS). (a) Association between days to silking, as reported in Mural et al. (2022), and 428 487 segregating SNPs identified and genotyped using RNA-seq data in Mazaheri et al. (2019). (b) Association test for days to silking using the marker set defined in this study (n = 16 634 049). The horizontal dashed line on each plot indicates an α = 1% significance threshold after applying Bonferroni's correction assuming n number of variants in each data set as independent tests.

detection of this known flowering time gene using an RNA-seq-based genetic marker data set. Overall, the newly generated data set increases the power to detect causal genes and does not increase *P*-value inflation (Figure S9).

In addition to the total number of confident signals identified, an additional potential benefit of higher-density genetic marker data is the more precise localization of peaks to only one or several candidate genes. The peak corresponding to *MADS69* included 29 markers that were significant at a Bonferroni corrected *P*-value of 0.01. These markers span a region of 410 350 bp that includes three annotated genes. However, the peak SNP (e.g. the single SNP with the most significant *P*-value) was 8507 bases from *MADS69*, and *MADS69* was the closest gene to this SNP (Figure 4a). The peak corresponding to *ZCN8* included 35 significant markers that were significant at a Bonferroni corrected *P*-value of 0.01. These markers span a region of 349 944 bases that includes seven annotated genes. In this case the peak SNP was 18 912 bases from *ZCN8* and three genes separated *ZCN8* from the peak SNP (Figure 4b).

The diverse composition of the population used for genotyping in this study creates an opportunity to detect patterns of selection in the genome and track changes in favorable allele frequency of variants associated with traits of interest, during domestication, adaptation to a new environment, or genetic improvement during modern breeding. As flowering plays an important role in local adaptation, we attempted to evaluate patterns of selection around the two known flowering-time genes identified above. We observed a clear reduction in nucleotide diversity in the promoter of *MADS69* in tropical and temperate maize lines relative to *parviglumis* (Figure 4a), consistent with a previous report (Liang et al., 2019). The most significantly associated SNP for days to silking in the *MADS69*

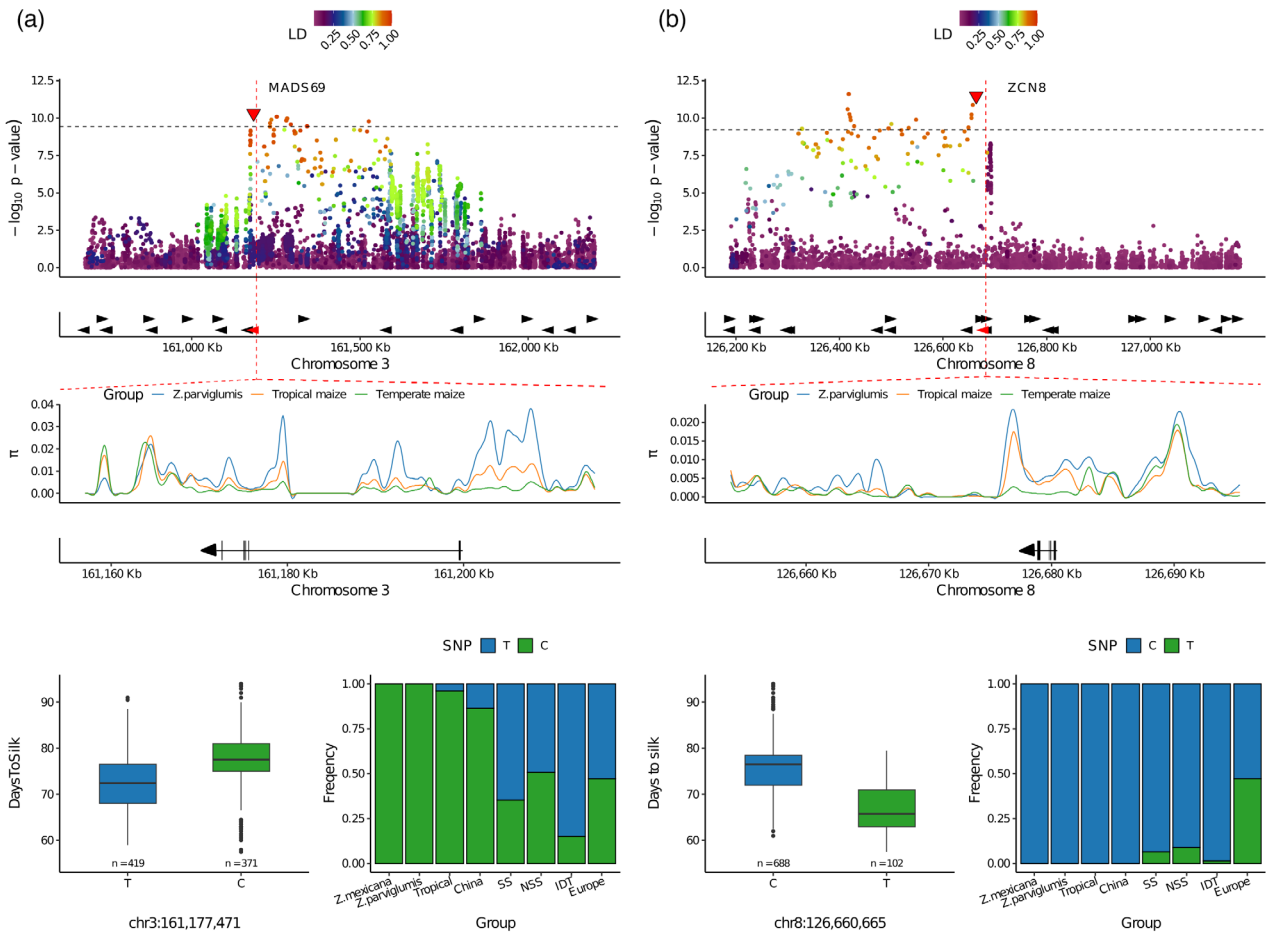

Figure 4. *MADS69* and *ZNC8* are associated with flowering time and were the target of selection. Top panel: zoom-in on genome-wide association study (GWAS) peaks around *MADS69* (a) and *ZNC8* (b). Linkage disequilibrium (LD) was calculated in each loci against the top associated SNP: chr3:161177471 and chr8:126660665 (marked as red triangles). The horizontal dashed line indicates the genome-wide Bonferroni correction level. Vertical red dashed lines mark the position of the gene of interest. Middle panel: nucleotide diversity in three maize groups. Gene bodies of *MADS69* and *ZNC8* were marked at the bottom. Vertical lines on gene bodies represent the specific positions of coding sequences (CDSs) for the genes. Bottom-left panels: allele effect of chr3:161177471 (a) and chr8:126660665 (b) on number of days to silking. Bottom-right panels: changes of allele frequency of chr3:161177471 (a) and chr8:126660665 (b) in eight maize groups. SS, stiff stalk; NSS, non-stiff stalk; IDT, iodent.

gene region was located at position 161 177 471. The reference allele (T) was associated with more rapid female silking relative to the alternate allele (C), with a mean difference of approximately 5 days (Figure 4a). In both *mexicana* and *parviglumis* populations only the slower flowering C allele was observed (Figure 4a). In lines classified as belonging to the tropical or Chinese populations, the C allele was predominant. In contrast, the T allele was the more common in the three North American populations (stiff stalk, non-stiff stalk and iodent). The T allele was particularly common among lines classified as belonging to the iodent heterotic group. Similarly, the T allele also made up the majority of genotype calls among the European maize lines included in this study. The large increase of frequency of shorter flowering T allele in temperate adapted lines is consistent with strong selection on *MADS69* during maize adaptation to temperate climates.

The second known flowering-time gene identified in this study was *ZCN8*, which has been previously shown to contribute to maize adaption to temperate climates and to have experienced a decline in nucleotide diversity in domesticated maize, relative to wild teosinte accessions, which is in parity with the conclusion that *ZCN8* was likely to have been a target of selection during maize domestication (Guo et al., 2018). However, the greater representation of different maize groups included in this study enabled the more specific identification of a decline in nucleotide diversity specifically between temperate and tropical domesticated maize populations, whereas tropical maize retained similar diversity to teosinte at this locus (Figure 4b). This result is consistent with selection on *ZNC8* occurring during adaptation to temperate conditions rather than during domestication.

The single most significant marker at the *ZCN8* locus was a C/T SNP at position 126 660 665 on chromosome 8. The T allele appears to be the derived allele and the median line homozygous for T at this position flowered 10.7 days earlier than the median line homozygous for the C allele. Although additional markers at the *ZCN8* locus that were not in LD (<0.2) with the most significant marker also exhibited a statistically significant association with flowering time, a haplotype-based model that incorporated information from the top not-in-linkage SNP (chr8:126689419) did not significantly improve the predictive ability for flowering time versus a single marker model. The rapid flowering allele at *ZCN8* was observed at extremely low frequencies in North American temperate germplasm. Almost all individuals homozygous for the rapid flowering allele originated in Europe (Figure 4b), demonstrating the importance of sampling broader global germplasm pools to gain greater power to identify functional variants primarily segregating only in individual geographic regions.

CONCLUSION

In summary, we performed a large-scale joint variant calling for 1515 maize individuals that included a wide range of maize accessions from multiple continents and eras, and discovered more than 46 million high-confidence sequence variants. In addition to releasing new sequence data for 239 new maize inbreds, we also release raw and filtered variant lists as well as processed GenomeDB files that will allow this SNP set to be further extended and expanded without the need to realign previously processed samples to the maize reference genome. We have shown that the new variant set accurately describes the population structure used in this study and improves power in genome-wide association studies relative to the previous state-of-the-art marker data sets for a large maize association panel.

EXPERIMENTAL PROCEDURES

Plant material and data sets

New resequencing data sets for maize are published regularly. The list of papers and accessions to include in this analysis was finalized on 30 June 2022 when the analysis commenced. Whole-genome resequencing data from 1515 total samples were used in this analysis, including 1276 previously published samples (Brandenburg et al., 2017; Bukowski et al., 2018; Chen et al., 2022; Chia et al., 2012; Kistler et al., 2018; Qiu et al., 2021; Unterseer et al., 2014; Wang et al., 2017, 2020) and 239 lines resequenced as part of this study. The origin and source of each sample included in this analysis are provided in Table S1.

Two hundred and twenty-eight inbred lines from the Wisconsin Diversity Panel (Mazaheri et al., 2019) were grown in a glasshouse setting (27–29°C during the day and 19–21°C at night, with 12 h of light/12 h of dark). After reaching V2, the youngest leaf was harvested onto ice and lyophilized for 2 days in a Flexi-Dry lyophilizer (FTS Systems Inc., now SP, https://www.scientificproducts.com). The lyophilized samples were ground to a fine powder in a Tissuelyzer II (Qiagen, https://www.qiagen.com), and DNA was extracted using the MagMAX Plant DNA Isolation Kit (ThermoFisher Scientific, https://www.thermofisher.com) with the help of a benchtop automated extraction instrument, KingFisher Flex (ThermoFisher Scientific). Raw DNA extracts were quantified using the Quant-iT dsDNA Broad Range Kit (Invitrogen, now ThermoFisher Scientific), and submitted to Psomagen (https://www.psomagen.com), where they were subject to an in-house quality assessment using TapeStation 4200 (Agilent, https://www.agilent.com). DNA samples from 29 lines did not meet the minimum DNA quality control standards for sequencing. An additional set of seeds from these lines were surface sterilized by washing them in a 5% v/v bleach solution for 10 mins, rinsed three times with sterile water, and then placed in centrifuge tubes with wetted paper and left in the dark at 23°C. Shortly after germinating (VE), the entire coleoptile was harvested, snap-frozen in liquid nitrogen and stored at −80°C. The tissue was then ground to a fine powder in a Tissuelyzer II (Qiagen) in the presence of dry ice in the pockets around tube holders. The DNA extraction was then performed utilizing the same procedure as was used on the original samples. Initial quality assessment of DNA samples, library preparation and sequencing was performed by Psomagen.

Libraries were prepared using the TruSeq DNA PCR-Free kit (Illumina, https://www.illumina.com). A NovaSeq6000 S4 (Illumina) sequencer was used to generate 150-bp paired-end reads.

Eleven Polish inbred lines were obtained from HR SMOLICE (https://www.hrsmolice.pl). Plants were grown in a phytotron chamber (24°C day/22°C night with 16 h of light/8 h of dark). Tissues for DNA extraction were harvested from the third fully developed leaf (V3 stage) and three individual plants were pooled into a single sample. Leaves were immediately flash-frozen in liquid nitrogen and tissue was ground in liquid nitrogen using a mortar and pestle. DNA extraction was performed with the DNeasy Plant Kit (Qiagen), according to the manufacturer's instructions. Genomic DNA for each genotype was submitted to Fasteris (https://www.fasteris.com) for whole-genome sequencing. For S160, S50676 and S68911 inbred lines, 100-bp paired-end reads, and for the remaining eight lines, 150-bp paired-end reads, were generated on a HiSeq X Ten sequencer (Illumina).

Creation of the global maize SNP set

After fastq files were downloaded from the European Nucleotide Archive (https://www.ebi.ac.uk/ena) or transferred from the sequencing provider, each file was cleaned using FASTP 0.23.2 with the default settings (Chen et al., 2018). Reads with >40% unqualified bases or with a quality value of <15 were removed. Cleaned fastq files were aligned with the B73_RefGen_V5 maize reference genome (Hufford et al., 2021) using SPEEDSEQ 0.1.2 (Chiang et al., 2015), which parallelizes BWA-MEM 0.7.10 (Li, 2013) for alignment, SAMBLASTER 0.1.22 for marking duplicated reads (Faust & Hall, 2014) and SAMBAMBA 0.5.9 for position sorting and BAM file indexing (Tarasov et al., 2015). SAMBLASTER defined duplicate read pairs as cases where two or more pairs of reads aligned with the same reference sequence on the same strand and with the same 5′ start position – or inferred 5′ start position if the alignment was clipped – for both forward reads and for both reverse reads. Unless otherwise stated, default parameters were used for each software package.

Individual gVCF files were generated for each maize pseudomolecule for each BAM file using the HaplotypeCaller tool provided by GATK 4.2.0.0 in diploid mode (Poplin et al., 2018). To enable extensive parallelization of variant calling, the maize genome was divided into 5-Mb windows for the creation of separate GenomicsDB datastores. During the project, GATK 4.2.6.1 was released and this update offered a reduction in the number of files stored in GenomicsDB datastores. Therefore, the GenomicsDBImport tool provided by GATK 4.2.6.1 was used for each genomic window to create the GenomicsDB datastore.

Joint variant calling was conducted using the GenotypeGVCFs tool provided by GATK 4.2.6.1, with default settings. To aid in additional parallelization, each 5-Mb GenomicsDB datastore was divided into five 1-Mb windows for variant calling.

Following GATK best practice recommendations, hard filters were applied to call variants. Variants were divided into SNPs and InDels for filtering. SNPs with QualByDepth < 2.0, FisherStrand > 60.0, RMSMappingQuality < 40.0, MappingQualityRankSumTest < −12.5 or ReadPosRankSumTest < −8.0 were removed. InDels with QualByDepth < 2.0, FisherStrand > 200.0 or ReadPosRankSumTest < −20.0 were also removed. After filtering, SNP and InDel variants were merged into single sorted VCF files for each chromosome using PICARD 2.9 (Pic, 2019). Finally, genotypes with depth < 2 were masked using the BCFTOOLS 1.10.2 plug-in SETGT (Danecek et al., 2021). All further VCF file manipulations were performed with BCFTOOLS 1.10.2 (Danecek et al., 2021).

Creation of the filtered and imputed maize SNP sets

The filtered and imputed variant set was generated by first removing variants where more than two alleles were observed in the population, variants with ≥50% missing data, variants with extremely low (<1515) or extremely high (>33 550) sequencing depth and variants with inbreeding coefficients ≥0, resulting in approximately 46 million variants. The inbreeding coefficient per variant was calculated as:

$$IC = 1 - \frac{H_{obs}}{H_{exp}}$$

where H_{obs} and H_{exp} are the observed and expected heterozygosity under Hardy–Weinberg equilibrium.Variants were phased and imputed using BEAGLE 5.0, with default settings (err = 0.0001; window = 40.0 cM; overlap = 4.0 cM; step = 0.1 cM; nsteps = 7) (Browning et al., 2018). In order to estimate imputation accuracy, 1% of variants on chromosome 10 were masked, and imputation was repeated with the same settings. Next, imputed and known variants were compared. The analysis was performed with TASSEL 5 (Bradbury et al., 2007). Imputation accuracy has been estimated to be 96.7%.

Population genetic analyses

Principal component analysis (PCA) was conducted with PLINK 1.9 (Purcell et al., 2007). Unimputed variants were filtered with MAF > 5% and with the fraction of missing data < 10%, leading to 19 205 674 markers, which were used for PCA. Individual genotypes were assigned to the population using published literature data (Brandenburg et al., 2017; Bukowski et al., 2018; Chen et al., 2022; Chia et al., 2012; Kistler et al., 2018; Qiu et al., 2021; Unterseer et al., 2014; Wang et al., 2020, 2017).

Measures of LD (r^2) were calculated for the entire population and predefined groups using POPLDDECAY 3.42 (Zhang et al., 2019) and the subset of unimputed SNPs with MAF > 0.05 and missing rate of <0.25. Local LD in a.

100-Kb window was calculated using GENOME-WIDE COMPLEX TRAIT ANALYSIS (GCTA), with default settings (Yang et al., 2011).

One hundred randomly selected 200-Kb windows from the maize genome (10 per chromosome) were selected for genome-wide nucleotide diversity (Nei & Li, 1979) analysis. Using the original GenomicsDB files described above, each genotype was called within each window using the -all-sites option in GATK, producing VCF files with records included for both variable and monomorphic (invariant) sites. All sites were hard-filtered as described above. Non-monomorphic sites were further filtered by excluding sites with more than two alleles or with a minor allele frequency of ≤1% prior to analysis. PIXY, which considers information on both missing sites and sequencing depth, was used to calculate nucleotide diversity (Korunes & Samuk, 2021). The same approach with recalling SNPs in the region including invariant sites was used to calculate patterns of nucleotide diversity around genes identified in the genome-wide association study.

Genome-wide association study (GWAS)

A published data set of female flowering time (days to silking) for 752 inbreds drawn from the Wisconsin Diversity Panel (Mazaheri et al., 2019), and grown in a replicated field study in Lincoln, NE, USA in 2020, was employed for genome-wide association (Mural et al., 2022). Two genetic marker sets for the same population of 752 maize inbreds were used to conduct the GWAS. The first set was created by filtering 752 maize inbreds with MAF > 5% from

899 784 variants called using RNA-seq relative to the B73_RefGen_V4 reference genome (Mazaheri et al., 2019). This leads to the creation of a set containing 428 487 variants. The second set was a set of 16 634 049 markers obtained by subsetting the filtered and imputed SNP set assembled in this study to include only those markers with an MAF > 5% among the 752 genotypes for which female flowering time phenotypes were available. In both cases, GWAS was conducted using the mixed linear model algorithm (Yu et al., 2006), as implemented in the RMVP (1.0.6) package in R (Yin et al., 2021). Both the kinship matrix, computed following the method described by VanRaden (2008), and the first five principal components of variation, calculated as described above, were included in the model. Calculations of local linkage disequilibrium were performed using PLINK 1.9 (Purcell et al., 2007).

All additional statistical analyses were conducted in R (R Core Team, 2022), with the extensive use of *data. table* (Dowle & Srinivasan, 2021) and *tidyverse* (Wickham et al., 2019) for data manipulation, and *tidyverse* and *patchwork* (Pedersen, 2020) for visualization.

ACKNOWLEDGEMENTS

This project was supported by US Department of Energy (DE-SC0020355), the National Science Foundation (OIA-1826781), USDA National Institute of Food and Agriculture (NIFA) under the AI Institute for Resilient Agriculture (2021-67021-35329), the Foundation for Food and Agriculture (602757) and the National Science Center (NCN), Poland (2012/05/B/NZ9/03407 and 2017/27/B/NZ9/00995). This project was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

CONFLICT OF INTEREST

JCS has equity interests in: Data2Bio, LLC; Dryland Genetics LLC; and EnGeniousAg LLC. He is a member of the scientific advisory board of GeneSeek and currently serves as a guest editor for *The Plant Cell*. The authors declare no other conflicts of interest associated with this work.

AUTHOR CONTRIBUTIONS

JCS and MWG conceived the study. JT conducted the experiments and generated the data. JY and GX provided advice and feedback on the design of the experiments and the analyses. MWG and RVM designed and conducted the analyses and visualized the results. MWG, RVM and JCS composed the initial draft of the article. All authors contributed to writing and editing the article and approved the final version for publication.

DATA AVAILABILITY STATEMENT

The additional resequencing data generated as part of this project have been deposited in the European Nucleotide Archive (ENA) under the study accession numbers: PRJEB56265, PRJEB56295 and PRJEB56320. Raw VCF files for the approximately 366 million variants identified in this study, imputed VCF files for the approximately 46 million filtered variations identified in this study and GATK GenomicsDBs files to enable new SNP calling with additional populations are available for download from MaizeGDB (https://www.maizegdb.org). In addition, imputed VCF files for the approximately 46 million filtered variations are also available from the Dryad repository: https://doi.org/10.5061/dryad.bnzs7h4f1. Code for the analysis is available from: https://github.com/mgrzy/Maize_Genetic_Variants_v5.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Schematic representation of the approach for variant calling.

Figure S2. Geographical distribution of the state of origin for 688 maize individuals of US origin.

Figure S3. Distribution of total aligned read depth for high-confidence variants.

Figure S4. Frequency of heterozygous genotype calls for each individual used in this study.

Figure S5. Example of three inbred lines with unexpectedly high heterozygosity.

Figure S6. Relationship between distance to centromere and density of SNPs and InDels.

Figure S7. Distribution of PCA values (1–3) assigned to the 1515 maize individuals.

Figure S8. Zoom-in on *ZCN8* region.

Figure S9. QQ plots of the two GWAS results shown in Figure 3.

Table S1. Summary of sequenced lines.

Table S2. Comparison of SNPs yielded in this study with those in Hufford et al. (2021).

Table S3. Number of accessions resequenced in the corresponding studies.

REFERENCES

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. & Buckler, E.S. (2007) Tassel: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Brandenburg, J.-T., Mary-Huard, T., Rigaill, G., Hearne, S.J., Corti, H., Joets, J. *et al.* (2017) Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. *PLoS Genetics*, **13**, e1006666.

Browning, B.L., Zhou, Y. & Browning, S.R. (2018) A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, **103**, 338–348.

Buckler, E.S., Gaut, B.S. & McMullen, M.D. (2006) Molecular and functional diversity of maize. *Current Opinion in Plant Biology*, **9**, 172–176.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z. *et al.* (2018) Construction of the third-generation Zea mays haplotype map. *Gigascience*, **7**, gix134.

Chen, L., Luo, J., Jin, M., Yang, N., Liu, X., Peng, Y. *et al.* (2022) Genome sequencing reveals evidence of adaptive variation in the genus Zea. *Nature Genetics*, **54**, 1736–1745.

Chen, S., Zhou, Y., Chen, Y. & Gu, J. (2018) fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, **34**, i884–i890.

Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., De Leon, N., Doebley, J. *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics*, **44**, 803–807.

Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P. *et al.* (2015) Speedseq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, **12**, 966–968.

Crow, T., Ta, J., Nojoomi, S., Aguilar-Rangel, M.R., Torres Rodríguez, J.V., Gates, D. *et al.* (2020) Gene regulatory effects of a large chromosomal inversion in highland maize. *PLoS Genetics*, **16**, e1009213.

Da Fonseca, R.R., Smith, B.D., Wales, N., Cappellini, E., Skoglund, P., Fumagalli, M. *et al.* (2015) The origin and evolution of maize in the southwestern United States. *Nature plants*, **1**, 1–5.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O. *et al.* (2021) Twelve years of samtools and bcftools. *Gigascience*, **10**, giab008.

Dowle, M. & Srinivasan, A. (2021) *Data.Table*: *extension of 'data.Frame'*. R package version 1.14.2.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Faust, G.G. & Hall, I.M. (2014) Samblaster: fast duplicate marking and structural variant read extraction. *Bioinformatics*, **30**, 2503–2505.

Flint-Garcia, S.A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S.M., Mitchell, S.E. *et al.* (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal*, **44**, 1054–1064.

Ganal, M.W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E.S., Charcosset, A. *et al.* (2011) A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One*, **6**, e28334.

Grzybowski, M., Adamczyk, J., Jonczyk, M., Sobkowiak, A., Szczepanik, J., Frankiewicz, K. *et al.* (2019) Increased photosensitivity at early growth as a possible mechanism of maize adaptation to cold springs. *Journal of Experimental Botany*, **70**, 2887–2904.

Gui, S., Wei, W., Jiang, C., Luo, J., Chen, L., Wu, S. *et al.* (2022) A pan-Zea genome map for enhancing maize improvement. *Genome Biology*, **23**, 1–22.

Guo, L., Wang, X., Zhao, M., Huang, C., Li, C., Li, D. *et al.* (2018) Stepwise *cis*-regulatory changes in *ZCN8* contribute to maize flowering-time adaptation. *Current Biology*, **28**, 3005–3015.

Hansey, C.N., Johnson, J.M., Sekhon, R.S., Kaeppler, S.M. & De Leon, N. (2011) Genetic diversity of a maize association population with restricted phenology. *Crop Science*, **51**, 704–715.

Ho, P.-T. (1955) The introduction of American food plants into China. *American Anthropologist*, **57**, 191–201.

Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J. *et al.* (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, **373**, 655–662.

Hufford, M.B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R.A. *et al.* (2012) Comparative population genomics of maize domestication and improvement. *Nature Genetics*, **44**, 808–811.

Hung, H.-Y., Shannon, L.M., Tian, F., Bradbury, P.J., Chen, C., Flint-Garcia, S.A. *et al.* (2012) *ZmCCT* and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, E1913–E1921.

Kistler, L., Maezumi, S.Y., Gregorio de Souza, J., Przelomska, N.A., Malaquias Costa, F., Smith, O. *et al.* (2018) Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science*, **362**, 1309–1313.

Korunes, K.L. & Samuk, K. (2021) Pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, **21**, 1359–1368.

Leiboff, S., Li, X., Hu, H.-C., Todt, N., Yang, J., Li, X. *et al.* (2015) Genetic control of morphometric diversity in the maize shoot apical meristem. *Nature Communications*, **6**, 1–10.

Li, C., Guan, H., Jing, X., Li, Y., Wang, B., Li, Y. *et al.* (2022) Genomic insights into historical improvement of heterotic groups during modern hybrid maize breeding. *Nature Plants*, **8**, 1–14.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Liang, Y., Liu, H.-J., Yan, J. & Tian, F. (2021) Natural variation in crops: realized understanding, continuing promise. *Annual Review of Plant Biology*, **72**, 10–1146.

Liang, Y., Liu, Q., Wang, X., Huang, C., Xu, G., Hey, S. *et al.* (2019) *ZmMADS69* functions as a flowering activator through the *ZmRap2.7-ZCN8* regulatory module and contributes to maize flowering time adaptation. *New Phytologist*, **221**, 2335–2347.

Lozano, R., Gazave, E., Dos Santos, J.P., Stetter, M.G., Valluru, R., Bandillo, N. *et al.* (2021) Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nature Plants*, **7**, 17–24.

Matsuoka, Y., Vigouroux, Y., Goodman, M.M., Sanchez, J., Buckler, E. & Doebley, J. (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6080–6084.

Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J.L., Burdo, B., Heckwolf, S. *et al.* (2019) Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology*, **19**, 1–17.

Mural, R.V., Grzybowski, M., Miao, C., Damke, A., Sapkota, S., Boyles, R.E. *et al.* (2021) Meta-analysis identifies pleiotropic loci controlling phenotypic trade-offs in sorghum. *Genetics*, **218**, iyab087.

Mural, R.V., Sun, G., Grzybowski, M., Tross, M.C., Jin, H., Smith, C. *et al.* (2022) Association mapping across a multitude of traits collected in diverse environments in maize. *GigaScience*, **11**, 1–15.

Nei, M. & Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, **76**, 5269–5273.

Pedersen, T.L. (2020) Patchwork: the composer of plots. R package version 1.1.1.

Pic. (2019) Picard toolkit. https://broadinstitute.github.io/picard/

Piperno, D.R., Ranere, A.J., Holst, I., Iriarte, J. & Dickau, R. (2009) Starch grain and phytolith evidence for early ninth millennium bp maize from the central Balsas River valley, Mexico. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 5019–5024.

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A. *et al.* (2018) Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**, 559–575.

Qiu, Y., O'Connor, C.H., Della Coletta, R., Renk, J.S., Monnahan, P.J., Noshay, J.M. *et al.* (2021) Whole-genome variation of transposable element insertions in a maize diversity panel. *G3*, **11**, jkab238.

R Core Team. (2022) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Renk, J.S., Gilbert, A.M., Hattery, T.J., O'Connor, C.H., Monnahan, P.J., Anderson, N. *et al.* (2021) Genetic control of kernel compositional variation in a maize diversity panel. *The Plant Genome*, **14**, e20115.

Romay, M.C., Millard, M.J., Glaubitz, J.C., Peiffer, J.A., Swarts, K.L., Casstevens, T.M. *et al.* (2013) Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, **14**, 1–18.

Romero Navarro, J.A., Willcox, M., Burgueño, J., Romay, C., Swarts, K., Trachsel, S. *et al.* (2017) A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nature Genetics*, **49**, 476–480.

Sowiński, P., Rudzińska-Langwald, A., Adamczyk, J., Kubica, I. & Fronk, J. (2005) Recovery of maize seedling growth, development and photosynthetic efficiency after initial growth at low temperature. *Journal of Plant Physiology*, **162**, 67–80.

Sun, G., Mural, R.V., Turkus, J.D. & Schnable, J.C. (2022) Quantitative resistance loci to southern rust mapped in a temperate maize diversity panel. *Phytopathology*, **112**, 579–587.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D. *et al.* (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research*, **20**, 1689–1699.

Swarts, K., Gutaker, R.M., Benz, B., Blake, M., Bukowski, R., Holland, J. *et al.* (2017) Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. *Science*, **357**, 512–515.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. & Prins, P. (2015) Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, **31**, 2032–2034.

Tenaillon, M.I. & Charcosset, A. (2011) A European perspective on maize history. *Comptes Rendus Biologies*, **334**, 221–228.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M. *et al.* (2014) A powerful tool for genome analysis in maize: development

and evaluation of the high density 600k SNP genotyping array. *BMC Genomics*, **15**, 1–15.

Van Heerwaarden, J., Doebley, J., Briggs, W.H., Glaubitz, J.C., Goodman, M.M., de Jesus Sanchez Gonzalez, J. *et al.* (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 1088–1092.

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science*, **91**, 4414–4423.

Wang, B., Lin, Z., Li, X., Zhao, Y., Zhao, B., Wu, G. *et al.* (2020) Genome-wide selection and genetic improvement during modern maize breeding. *Nature Genetics*, **52**, 565–571.

Wang, L., Beissinger, T.M., Lorant, A., Ross-Ibarra, C., Ross-Ibarra, J. & Hufford, M.B. (2017) The interplay of demography and selection during maize domestication and expansion. *Genome Biology*, **18**, 1–13.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R. *et al.* (2019) Welcome to the tidyverse. *Journal of Open Source Software*, **4**, 1686.

Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. (2011) Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, **88**, 76–82.

Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L. *et al.* (2019) Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nature Genetics*, **51**, 1052–1059.

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z. *et al.* (2021) Rmvp: a memoryefficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteomics & Bioinformatics*, **19**, 619–628.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203–208.

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. (2019) Popldecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, **35**, 1786–1788.