# Genotyping of DNA pools identifies untapped landraces and genomic regions to develop next-generation varieties

Mariangela Arca, Brigitte Gouesnard, Tristan Mary-huard, Marie-christine Le Paslier, Cyril Bauland, Valérie Combes, Delphine Madur, Alain Charcosset, Stéphane D Nicolas

Plant Biotechnology Journal Open Access

aab Association of Applied Biologists SEB

*Plant Biotechnology Journal* (2023), pp. 1–17 doi: 10.1111/pbi.14022

# Genotyping of DNA pools identifies untapped landraces and genomic regions to develop next-generation varieties

Mariangela Arca[1], Brigitte Gouesnard[2], Tristan Mary-Huard[1], Marie-Christine Le Paslier[3], Cyril Bauland[1], Valérie Combes[1], Delphine Madur[1], Alain Charcosset[1] and Stéphane D. Nicolas[1,*]

[1]*INRAE, CNRS, AgroParisTech, GQE – Le Moulon, Université Paris-Saclay, Gif-sur-Yvette, France*
[2]*UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France*
[3]*Université Paris-Saclay, INRAE, Etude du Polymorphisme des Génomes Végétaux, Evry, France*

*Received 22 January 2021;*
*accepted 18 January 2023.*

**Correspondence* (Tel +33 1 69 33 23 69; fax +33 1 69 33 23 40; email stephane.nicolas@inrae.fr)

## Summary

Landraces, that is, traditional varieties, have a large diversity that is underexploited in modern breeding. A novel DNA pooling strategy was implemented to identify promising landraces and genomic regions to enlarge the genetic diversity of modern varieties. As proof of concept, DNA pools from 156 American and European maize landraces representing 2340 individuals were genotyped with an SNP array to assess their genome-wide diversity. They were compared to elite cultivars produced across the 20th century, represented by 327 inbred lines. Detection of selective footprints between landraces of different geographic origin identified genes involved in environmental adaptation (flowering times, growth) and tolerance to abiotic and biotic stress (drought, cold, salinity). Promising landraces were identified by developing two novel indicators that estimate their contribution to the genome of inbred lines: (i) a modified Roger's distance standardized by gene diversity and (ii) the assignation of lines to landraces using supervised analysis. It showed that most landraces do not have closely related lines and that only 10 landraces, including famous landraces as Reid's Yellow Dent, Lancaster Surecrop and Lacaune, cumulated half of the total contribution to inbred lines. Comparison of ancestral lines directly derived from landraces with lines from more advanced breeding cycles showed a decrease in the number of landraces with a large contribution. New inbred lines derived from landraces with limited contributions enriched more the haplotype diversity of reference inbred lines than those with a high contribution. Our approach opens an avenue for the identification of promising landraces for pre-breeding.

**Keywords:** genebanks, genetic diversity, pre-breeding, open-pollinated variety, selective footprints, *Zea mays* L.

## Introduction

Plant genetic resources are the basic raw material for future genetic progress (Hoisington *et al*., 1999; Kilian and Graner, 2012; McCouch *et al*., 2012; Tanksley, 1997). Landraces, that is, traditional varieties, are expected to be a major source of genetic diversity for addressing the challenges of climate change and the requirements of low-input agriculture, as they have been long selected to be well adapted to local agro-climatic conditions and human uses (Fernie *et al*., 2006; Gates *et al*., 2019; Mascher *et al*., 2019; McCouch *et al*., 2012). However, landraces are used to a very limited extent, if at all, in modern plant breeding programmes because they are poorly characterized, genetically heterogeneous and generally exhibit poor agronomic performance compared to elite material (Brauner *et al*., 2019; Hölker *et al*., 2019; Kilian and Graner, 2012; Mascher *et al*., 2019; Strigens *et al*., 2013). Therefore, understanding their genetic diversity and relationship to the elite breeding pool is essential for a better management of genetic resources and for genetic improvement (Gates *et al*., 2019; Hoisington *et al*., 1999; Mascher *et al*., 2019). In maize, less than 5% of maize genetic variability has been exploited in elite breeding pools (Hoisington *et al*., 1999). During the early twentieth century, maize landraces were used as parent material for the development of improved hybrid varieties to meet the needs of modern agriculture. During this transition from maize landraces to hybrids, many favourable alleles were probably lost as a result of their association with unfavourable alleles and/or genetic drift (Buckler *et al*., 2006; Reif *et al*., 2005; Yamasaki *et al*., 2005, 2007). Nowadays, modern breeding programmes tend to focus on breeding populations that can be traced back to a few ancestral inbred lines derived from landraces at the start of the hybrid era (Coffman *et al*., 2020; Gerdes and Tracy, 1993; Mikel, 2011; van Heerwaarden *et al*., 2011). Maize landraces that did not contribute to this founding material are expected to be useful for enriching modern maize diversity, particularly for traits that enhance adaptation to adverse environmental conditions (Gates *et al*., 2019).

Maize was domesticated in the highlands of Central Mexico approximately 9000 years ago (Beadle, 1939; Matsuoka *et al*., 2002). It then diffused to South and North America (Swarts *et al*., 2017; Tenaillon and Charcosset, 2011) and spread rapidly out from America (Brandenburg *et al*., 2017; Brandolini, 1970; Camus-Kulandaivelu *et al*., 2006; Dubreuil *et al*., 2006; Mir *et al*., 2013; Rebourg *et al*., 2001, 2003; Swarts *et al*., 2017). It is

Please cite this article as: Arca, M., Gouesnard, B., Mary-Huard, T., Le Paslier, M-C., Bauland, C., Combes, V., Madur, D., Charcosset, A. and Nicolas, S D. (2023) Genotyping of DNA pools identifies untapped landraces and genomic regions to develop next-generation varieties. *Plant Biotechnol J*., https://doi.org/10.1111/pbi.14022.

now cultivated in highly diverse climate zones ranging from 40°S to 50°N. After being introduced in different parts of the world, maize landraces were then selected by farmers to improve their adaptation to specific environments, leading to changes in flowering behaviour, yield, nutritive value and resistance to biotic and abiotic stress, resulting in subsequent differentiation of the varieties (Camus-Kulandaivelu *et al*., 2006; Castelletti *et al*., 2020; Gates *et al*., 2019; Wang, Josephs, *et al*., 2021).

In recent years, the genetic diversity of maize landraces has been studied extensively using various types of molecular markers such as restriction fragment length polymorphisms (RFLPs) (Camus-Kulandaivelu *et al*., 2006; Dubreuil *et al*., 1999, 2006; Dubreuil and Charcosset, 1998; Gauthier *et al*., 2002; Rebourg *et al*., 1999, 2001, 2003; Reif *et al*., 2005) and simple sequence repeats (SSRs) (Eschholz *et al*., 2010; Mir *et al*., 2013; Reif *et al*., 2005; Vigouroux *et al*., 2005). Single-nucleotide polymorphisms (SNPs) are now the marker of choice for various crop species such as maize (Ganal *et al*., 2011), rice (McCouch *et al*., 2010) and barley (Moragues *et al*., 2010). They are the most abundant class of sequence variation in the genome, are co-dominantly inherited, genetically stable, easily automated and, thus, suitable for high-throughput automated analysis (Rafalski, 2002). Unlike SSRs, allele coding can be easily standardized across laboratories and the cost of genotyping is very low, which is a major advantage for characterizing genetic resources. A maize array with approx. 50 000 SNP markers has been available since 2010 (Ganal *et al*., 2011). It has been successfully used to analyse the diversity of inbred lines and landraces by genotyping a low number of plants per accession (Arteaga *et al*., 2016; Bouchet *et al*., 2013; Frascaroli *et al*., 2013; Hufford *et al*., 2012; Strigens *et al*., 2013; van Heerwaarden *et al*., 2011).

However, due to high within-accession diversity, the characterization of each maize landrace should be carried out on a representative set of individuals (Reyes-Valdés *et al*., 2013). Despite recent technical advances, genotyping large numbers of individuals remains very expensive in the context of genetic resources characterization. As a result, DNA pooling has been actively developed as a valuable alternative strategy for collecting information on allele frequency from a group of individuals while significantly reducing the genotyping effort (Schlötterer *et al*., 2014; Sham *et al*., 2002). In maize, DNA pooling has been successfully used to decipher the global genetic diversity of landraces using RFLP (Dubreuil *et al*., 1999) and SSR markers (Camus-Kulandaivelu *et al*., 2006; Dubreuil *et al*., 2006; Mir *et al*., 2013; Reif *et al*., 2006; Yao *et al*., 2007). The recent development of SNP arrays in maize (Ganal *et al*., 2011; Unterseer *et al*., 2014), combined with DNA pooling, should be useful for characterizing the genetic diversity of maize landraces at a fine genomic scale. In a previous study, we developed a new method for predicting the allelic frequency of each SNP from a maize Illumina 50 K array within DNA pools based on the fluorescence intensity of the two alleles at each SNP (Arca *et al*., 2021). This new method accurately predicts allelic frequency, safeguards against the false detection of alleles. Additionally, structure results and genetic distance obtained with 50 K array were highly congruent with those obtained with SSR in previous studies indicating little consequences of ascertainment bias for deciphering global genetic diversity organization (Arca *et al*., 2021).

In the present study, we applied this recent method on a pilot scale to: (i) investigate the genome-wide diversity and genetic structure of 156 maize landraces that are representative of European and American diversity and that represented 2340 individuals; (ii) compare the diversity of these landraces to that of a panel of 327 inbred lines that represent the diversity presently used in North-American and European breeding, the 'CK lines' (Camus-Kulandaivelu *et al*., 2006) and 103 new inbred lines derived from landraces, the 'DH-SSD lines'; and (iii) identify the landraces that could potentially broaden the genetic diversity of the CK lines.

## Results

### Genetic diversity within maize landraces

Only 25 SNPs out of 23 412 were monomorphic in the landrace panel represented by 2340 individuals (15 individuals per landraces accessions). The average total diversity estimated with SNP ($Ht_{SNP}$) was 0.338 ± 0.001. Ht estimated with 17 SSRs was 1.8 time higher than $Ht_{SNP}$ ($Ht_{SSR}$ = 0.61 ± 0.118). The distribution of minor allelic frequency of SNP (MAF) showed a deficit in rare alleles (MAF <0.05) compared to other frequency classes (Figure S1). It suggests an ascertainment bias towards the selection of SNPs with intermediate allele frequencies when defining the 50 K Illumina array (Ganal *et al*., 2011).

In order to compare the genetic diversity of populations from different regions, we classified the 156 landraces into five geographic groups: Europe (EUR), North America (NAM), Central America and Mexico (CAM), the Caribbean (CAR) and South America (SAM) (Table 1, Figure S2, Table S1). All five geographic groups displayed both alleles for nearly all loci, with the exception of CAR being monomorphic at 1227 loci out of 23 387 (Figure S3). The lowest and highest within-group diversity (Ht) were observed for CAR ($Ht_{SNP}$ = 0.301) and CAM ($Ht_{SNP}$ = 0.328), respectively. Note that there were more rare alleles in EUR, CAR and NAM than in SAM and CAM (Figure S1).

The average number of alleles per locus and per landrace within the entire landrace panel was 1.629 ± 0.003 and ranged from 1.098 (Ger8) to 1.882 (Sp11). Gene diversity within landraces estimated with SNP ($Hs_{SNP}$) was on average 0.192 ± 0.001, (Table 1) and varied between 0.03 (Ger8 and Ger9) and 0.28 (Sp11) (Table S1). Hs estimated with SSRs ($Hs_{SSR}$) was highly and linearly correlated with $Hs_{SNP}$ ($r^2$ = 0.73) and was on average two times higher than $Hs_{SNP}$ (Table S1). The CAM group displayed on average the highest diversity ($Hs_{SNP}$ = 0.219 ± 0.008; $Hs_{SSR}$ = 0.446), while the EUR group displayed the lowest ($Hs_{SNP}$ 0.177 ± 0.002; $Hs_{SSR}$ = 0.368).

Genetic differentiation between landraces estimated with SNPs ($Gst_{SNP}$) was 0.432 on average, higher than differentiation estimated with SSRs ($Gst_{SSR}$ = 0.369). $Gst_{SNP}$ within a geographic group varied between 0.332 (CAR) and 0.436 (EUR) (Table 1). Overall genetic differentiation between geographic groups was low whatever marker type ($Gst_{SNP}$ = 0.04 and $GST_{SSR}$ = 0.07). $Gst_{SNP}$ between pairs of geographic groups varied between 0.017 (EUR and NAM) and 0.099 (NAM and CAR) and was on average 1.5 times lower than Gst estimated with SSR (Table S2).

### Relationship between maize landraces and population structure

The average modified Roger's distance estimated with SNPs ($MRD_{SNP}$) between landraces was 0.379 which was lower than that estimated with SSR ($MRD_{SSR}$ = 0.461). $MRD_{SNP}$ between landraces was highly correlated with $MRD_{SSR}$ ($r$ = 0.78). The lowest $MRD_{SNP}$ between landraces was 0.158 (Chi12 and Chi9), that was slightly higher than the distance between two pools of

**Table 1** Genetic diversity within the five geographic groups of landraces, the entire landrace panel and the CK line panel.

| | Europe (EUR) mean ± s.d. | North America (NAM) mean ± s.d. | Central America and Mexico (CAM) mean ± s.d. | Caribbean (CAR) mean ± s.d. | South America (SAM) mean ± s.d. | Landrace Panel (LP) mean ± s.d. | CK line Panel (IL) mean ± s.d. |
|---|---|---|---|---|---|---|---|
| Number of populations/inbred lines | 83 | 22 | 25 | 14 | 22 | 166 | 327 |
| Allele number (A) group level | 1.996 ± 0.001 | 1.989 ± 0.005 | 1.990 ± 0.004 | 1.947 ± 0.017 | 1.992 ± 0.004 | 1.999 ± 0.000 | 1.989 ± 0.001 |
| Allele number (A) average within pop/line | 1.584 ± 0.005 | 1.649 ± 0.021 | 1.701 ± 0.018 | 1.662 ± 0.034 | 1.671 ± 0.021 | 1.629 ± 0.003 | 1.004 ± 0.000 |
| Minor Allele Frequency (MAF) group level | 0.235 ± 0.001 | 0.235 ± 0.006 | 0.244 ± 0.006 | 0.223 ± 0.011 | 0.240 ± 0.007 | 0.253 ± 0.001 | 0.265 ± 0.001 |
| Minor Allele Frequency (MAF) average within pop/line | 0.128 ± 0.001 | 0.141 ± 0.002 | 0.159 ± 0.001 | 0.150 ± 0.001 | 0.149 ± 0.001 | 0.139 ± 0.000 | 0.002 ± 0.000 |
| Total expected heterozygosity across groups (Ht) | 0.314 ± 0.002 | 0.317 ± 0.007 | 0.328 ± 0.006 | 0.301 ± 0.012 | 0.323 ± 0.007 | 0.338 ± 0.001 | 0.353 ± 0.001 |
| Expected heterozygosity (Hs) average of within pop/line | 0.177 ± 0.002 | 0.195 ± 0.009 | 0.219 ± 0.008 | 0.206 ± 0.014 | 0.205 ± 0.009 | 0.192 ± 0.001 | 0.002 ± 0.000 |
| Modified Roger's distance between landraces/inbred lines (MRD) | 0.367 ± 0.061 | 0.351 ± 0.063 | 0.336 ± 0.033 | 0.320 ± 0.026 | 0.346 ± 0.068 | 0.379 ± 0.059 | 0.580 ± 0.024 |
| Differentiation between landraces ($Gst_l$) and between inbred lines ($Gst_i$) | 0.436 ± 0.001 | 0.384 ± 0.001 | 0.332 ± 0.001 | 0.315 ± 0.001 | 0.365 ± 0.001 | 0.432 ± 0.002 | 0.994 |

independent individuals from a same population (0.087–0.120 in Arca *et al*. (2021)). The highest $MRD_{SNP}$ was 0.552 (Ant1 and Ger8). The average $MRD_{SNP}$ between populations from a same geographic group ranged from 0.320 (CAR) to 0.367 (EUR) (Table 1). The average $MRD_{SNP}$ between populations belonging to two different geographic groups varied between 0.354 (CAM vs. CAR) and 0.420 (NAM vs. CAR) which was on average 1.2 fold lower than $MRD_{SSR}$ (Table S2).

We investigated the relationship between maize landraces using Principal Coordinate Analysis (PcoA) and Ward hierarchical clustering based on MRD (Figure 1). For both, landraces mostly clustered according to their geographic proximity (Figure 1; Figure S4; Figure S5). The first axis (PC1, 18.4% of the total variation) discriminated (i) temperate landraces belonging to the Northern Flint cluster (from northern Europe and North America) from (ii) tropical and subtropical landraces (from the Caribbean and South and Central America) (Figure 1a). The second axis (PC2, 5% of the total variation) discriminated (i) North American (Corn Belt Dent cluster), Central American and Mexican populations (Mexican cluster) from (ii) Italian (Italian Flint cluster), and Spanish and French populations (Pyrenean-Galician cluster). Ward hierarchical clustering showed that at the highest level ($k = 2$, Figure 1b), 62 of the 83 European landraces clustered together (European cluster) while 70 of the 83 American landraces clustered together (American cluster). At a deeper level ($k = 7$), we distinguished 4 clusters of American or European landraces, each originating from a geographic area with homogeneous agro-climatic conditions (cluster a, b, e and f in Figure 1b and Figure S4). Three clusters grouped together American and European landraces (cluster c, d and g in Figure 1b and Figure S4). Using a pairwise Mantel test for each geographic area, we observed a low but significant correlation between the genetic distance and geographic distance matrices for EUR ($r^2 = 0.05$, $P < 0.001$, Figure S6A), NAM ($r^2 = 0.12$, $P < 0.001$, Figure S6B) and CAM ($r^2 = 0.0858$, $P = 0.02$, Figure S6C).

We analysed the genetic structure of 156 landraces using the ADMIXTURE program. Likelihood analysis indicated that the optimal number of genetic groups were $K = 2$, $K = 3$ and $K = 7$ (Figure S7). The genetic structure obtained with SNP markers was highly consistent with that obtained with the 17 SSR markers since 72% ($K = 7$) to 100% ($K = 3$) of landraces were assigned to the same group by both types of markers (Table S3). The main differences between the SSR and SNP results were observed at $K = 7$ when the Northern Flint landrace group obtained with SNPs was split into two with SSRs, while Pyrenean-Galician and Italian groups which were separated with SNPs formed a single group with SSRs. We considered $K = 7$ as the reference, as this value was consistent with the one obtained with 24 SSRs by Camus-Kulandaivelu *et al*. (2006). Assignation of landraces to the different genetic groups was consistent with geographic origin, with a clear trend along latitude and longitude (Figure 2). Assignment to these groups was also highly consistent with PcoA and hierarchical clustering (Figure 1; Figure 2; Figure S4; Figure S5).

## Scanning the maize landrace genomes for regions under selection

Using a sliding window approach, we identified 14 regions with windows containing at least two SNPs with extremely low genetic diversity ($\overline{\mathrm{H}t_l} < 0.069$) across the entire landrace panel (Figure 3a; Table S4). Genomic regions showing low diversity within geographic groups were most abundant in CAR (67), followed by EUR (56), CAM (39), SAM (36) and NAM (26) (Figure 3e–i; Table S4). These regions were mostly located close to the centromeres but varied between geographic groups.

Outlier analysis of Gst values among individual landraces identified 20 and 17 genomic regions displaying high differentiation ($\overline{Gst_l} > 0.568$) and low differentiation ($\overline{Gst_l} < 0.235$) between landraces, respectively (Figure 3l, Table S5). Genetic differentiation was highest at the beginning of chromosome 6

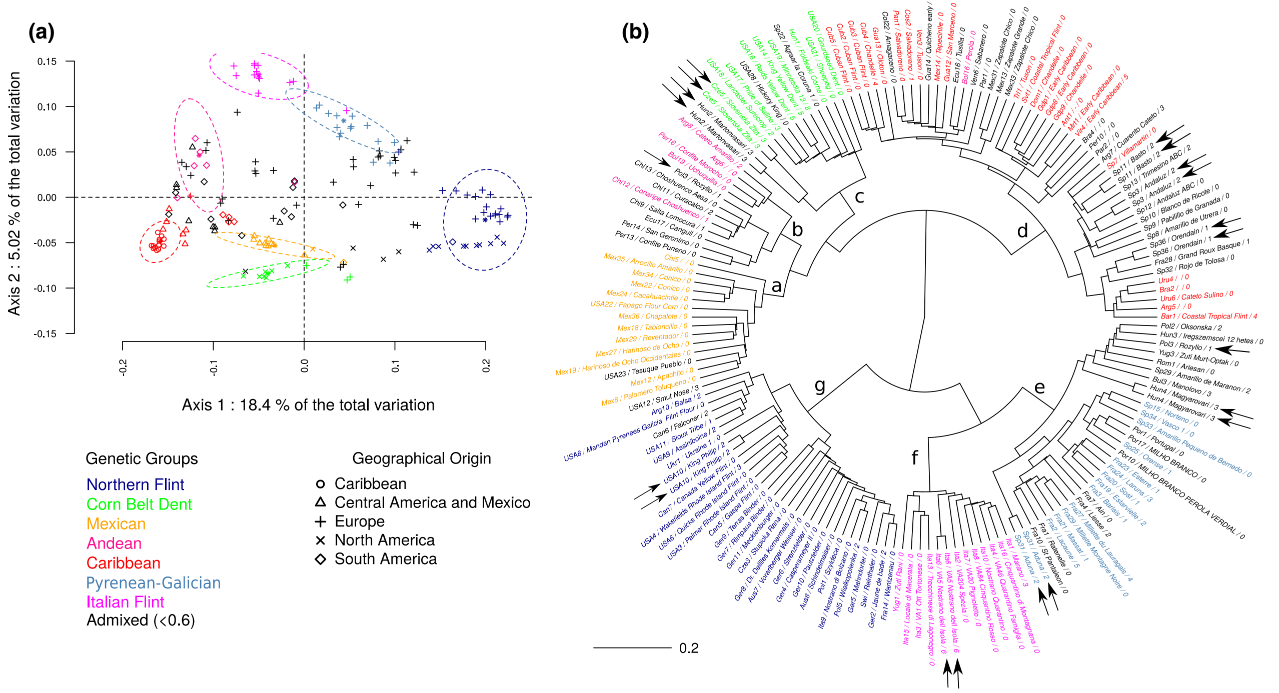

**Figure 1** Genetic relationship between 156 maize landraces based on their modified Roger's distance (MRD). (a) Projection of the 166 DNA samples on the first two axes of the Principal Coordinate Analysis. Symbols indicate the geographic origin of landraces. (b) Dendrogram obtained by Hierarchical clustering, using Ward's algorithm. Labels indicate for each landrace their abbreviation code, common names and number of first cycle inbred lines they contributed to, respectively. Black arrows indicate the 10 landraces with duplicated DNA samples. Colours indicate the assignment of landraces to the seven genetic groups defined by ADMIXTURE. Landraces with an assignment probability below 0.6 were considered admixed and coloured in black. Cluster 'a' grouped 15 landraces that originated mainly in Mexico and southwestern USA. Cluster 'b' comprised 10 South American landraces that originated along the Andean Mountains. Cluster 'e' grouped 31 European landraces that originated either along the Pyrenean Mountains or in Central Eastern Europe. Cluster 'f' grouped mainly Italian Flint landraces. Cluster 'c' comprised 14 dent landraces that originated mainly from Eastern European landraces and the US Corn Belt. Cluster 'd' grouped 65 landraces mostly from southern Spain (latitude <40°N), southwestern France and from the Caribbean Islands and countries bordering the Caribbean Sea (d1, d2 and d3 on Figure S4). Cluster 'g' comprised 12 North American flint landraces from higher latitudes (>40°N) and 18 northeastern European landraces mainly from Germany.

(Sp10 in Table S5), in two regions at the beginning of chromosome 4 (Sp6 and Sp7 in Table 2), in two regions of chromosome 3 (Sp3 and Sp5 in Table 2) and in one region on chromosome 1 (Sp1 in Table 2).

Outlier Gst analysis considering jointly all geographic groups identified 26 regions with high differentiation ($\overline{Gst_g} > 0.150$) and 8 regions with low differentiation ($\overline{Gst_g} < 0.007$) (Figure 3j; Table S4); BAYESCAN identified 377 loci under divergent selection (Figure 3j; Tables S6 and S7, Figure S8). Six genomic regions were identified between landraces but not between all five geographic groups whereas six genomic regions were identified exclusively between the five geographic groups (Table S5). Among the 11 highest differentiated genomic regions between landraces with at least two SNPs, only five were also detected between the five geographic groups by both Gst outlier and BAYESCAN analyses (Table 2). These regions displayed contrasted allelic gradients across geographic groups (Figure 3d–h; Table S5).

Outlier Gst analysis between pairs of geographic groups identified 214 and 41 regions displaying high and low differentiation, respectively (Figure S9). BAYESCAN analysis identified 363 SNPs under selection considering pairs of geographic groups, including 167 new SNPs that were not previously identified between all five geographic groups (Table S8). The new highly differentiated regions identified by BAYESCAN were mostly specific to a single pair of geographic groups (Figure S9; Figure S10). Putative functions could be assigned to 272 of the 536 (50.7%) outlier loci identified by BAYESCAN analysis of all and pairs of geographic groups. These included known genes involved in adaptation to abiotic stress, flowering time or human uses (Table S8; Table S9).

## Genome-wide comparison of diversity between landraces and inbred lines

The panel of CK lines contained more monomorphic SNPs than landraces (263 vs. 25) but still captured 99% of the alleles present within the landrace panel. Ht was slightly higher in inbred lines than in landraces for SNPs (0.353 vs. 0.338) and SSR (0.611 vs. 0.593). Allelic frequencies and Ht values of loci in inbred lines and landraces were strongly correlated for SNPs ($r^2$ = 0.89 and $r^2$ = 0.80, respectively, Figure S11) and SSR ($r^2$ = 0.81 for Ht). Overall, genetic differentiation between landraces and inbred lines was limited with SNPs (0.010 ± 0.066). Some regions were more diverse in landraces than in inbred lines, notably the pericentromeric region of chromosomes 3 and 7, while the opposite was found in centromeric regions of chromosomes 1, 3, 4, 5 and 6 (Figure 3a–c).

Comparison of landraces and inbred lines using the outlier Gst approach identified 128 highly differentiated genomic regions (Gst >0.04) and 32 regions with an excess of similarity (Gst <

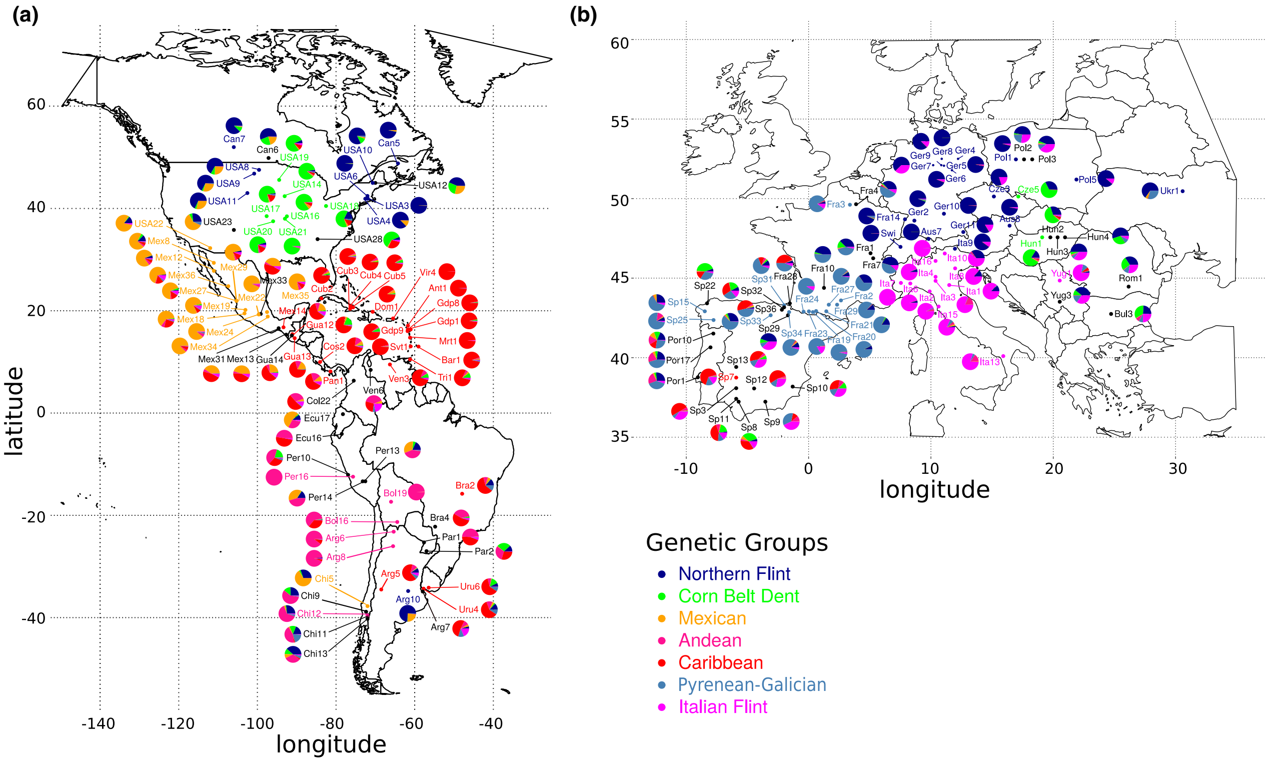

**Figure 2** Spatial genetic structure of American (a) and European (b) maize landraces. Population structure is based on ADMIXTURE analysis with $K = 7$. Each population is represented by a pie diagram whose composition indicates admixture coefficients. Population labels are coloured according to their main assignment (>0.6), and are black if the landrace is admixed.

4.21e-05). While highly differentiated regions were mainly located on chromosomes 3, 4, 8, 9 and 10, weakly differentiated regions were mainly located on chromosomes 3, 5 and 9 (Figure 3k). BAYESCAN analysis of landraces vs. inbred lines identified 61 loci (0.3%) that were significantly more differentiated than expected under the drift model (Figure 3k; Table S10).

## Relationship between inbred lines and landrace populations: Genetic distances and supervised analysis

The average MRD estimated with SNPs ($\text{MRD}_{\text{SNP}}$) between landraces and CK lines was 0.499 (±0.034), which is greater than between landraces (0.379 ± 0.059) and less than between lines (0.590 ± 0.024). The distribution of MRD genetic distances between a given landrace and CK lines ($\text{MRD}_{\text{LI}}$) is displayed as a series of boxplots (Figure 4a) listed in ascending order of landrace expected heterozygosity (Hs) (Figure 4b). Landraces with a low genetic diversity generally showed a higher median and a wider range for $\text{MRD}_{\text{LI}}$ (Figure 4). Accordingly, the median $\text{MRD}_{\text{LI}}$ and the within landrace genetic diversity Hs were strongly negatively correlated ($r = -0.978$, $t = -61.314$, $P$-value $<2.2\text{e-}16$) and displayed a linear relationship (Figure S12). Considering a similar level of genetic diversity, some landraces were closely related to certain inbred lines, whereas other landraces were not (e.g. Chi5, Per10, Par2, Par1, Bra4, Ecu17, Vir4 and Svt1 in Figure 4a and Figure S12).

In order to identify the source material of modern varieties, and *a contrario* the landraces that did not contribute much to these varieties, we assigned quantitatively 442 inbred lines to 166 landraces using a supervised analysis (Table S11). The 234 first cycle inbred lines (*i.e.* directly derived from a single landrace) were assigned to a total of 60 landraces. For first cycle inbred lines of known pedigree and whose ancestral landrace is included in our study (a total of 121 lines and 50 landraces), we noted a very good match between pedigree and main assignment (71.9% of cases). Among these 121 lines, DH-SSD lines, which were derived recently from landraces, were more frequently assigned to their population of origin than lines from the CK panel (77.6% vs. 58.3%, $P$-value = 0.04). The 208 lines from more advanced breeding cycles were assigned to a total of 66 landraces.

Few landraces contributed strongly to the whole diversity panel, with the 10 first landraces cumulating half of the total contributions (Figure 4c, Figure S13A). 80% of lines were assigned to these 10 landraces with a >60% probability (Figure S13B). Among these, temperate inbred lines were frequently assigned to Reid's Yellow Dent and Lancaster Surecrop. Chandelle (one of the few tropical landraces in our study) was identified as the most likely source for many tropical lines. Interestingly, the mean contribution of landraces differed strongly between first cycle lines and more advanced lines with a strong decrease (>1%) for 15 landraces and a strong increase (>1%) for 8 landraces (Figure S13C).

We tested whether the mean contribution of landraces and the $\text{MRD}_{\text{LI}}$ distance 'normalized' by within landraces genetic diversity could be used as a criterion to identify untapped sources of genetic diversity that could enrich the allelic diversity of the CK line panel. First, we selected 66 DH-SSD lines that were correctly assigned to 33 landraces from the landrace panel. We then classified these 33 landraces according to: (i) their average contribution to CK lines (Figure 5a) and (ii) the normalized MRD distance from their closest lines (Figure 5c). For each class, we

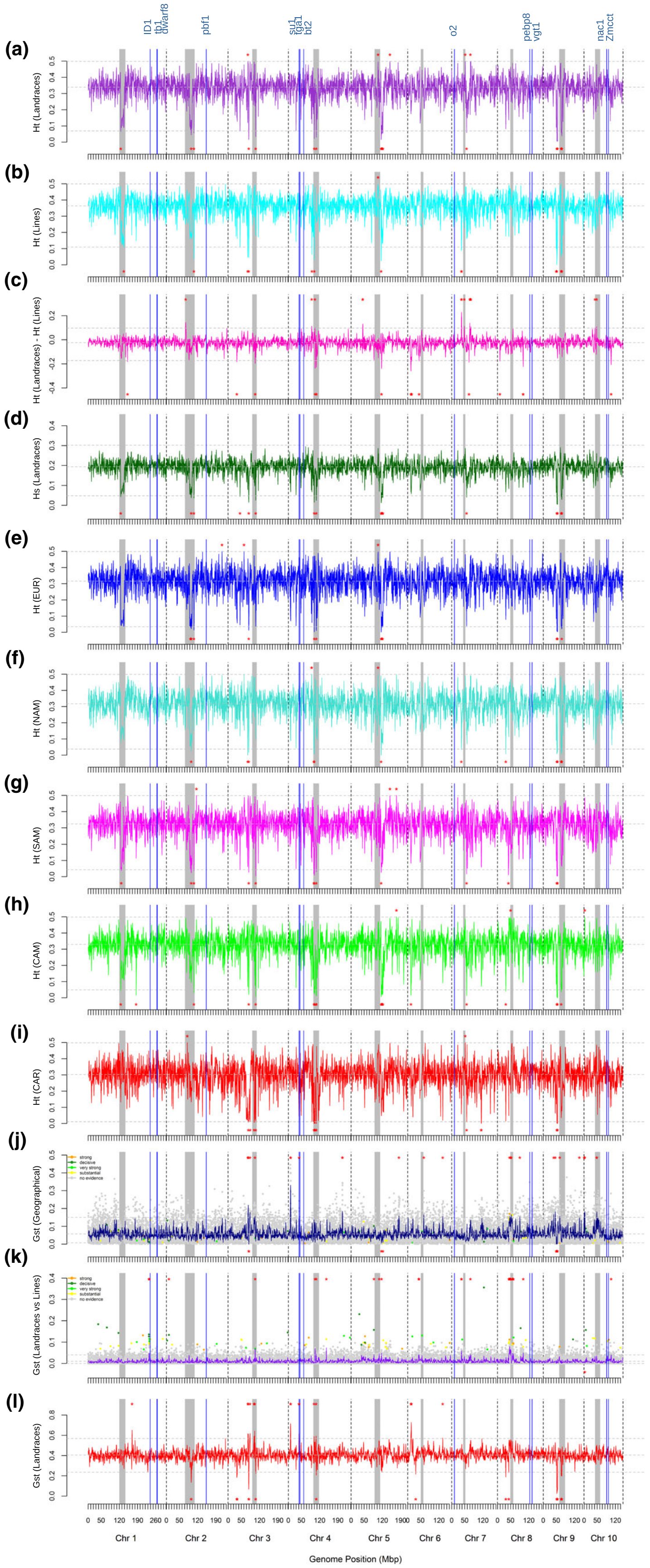
ID1
tb1
dwarf8
pbf1
su1
tga1
bt2
o2
pebp8
vgt1
nac1
Zmcct
(a)
Ht (Landraces)
(b)
Ht (Lines)
(c)
Ht (Landraces) - Ht (Lines)
(d)
Hs (Landraces)
(e)
Ht (EUR)
(f)
Ht (NAM)
(g)
Ht (SAM)
(h)
Ht (CAM)
(i)
Ht (CAR)
(j)
Gst (Geographical)
strong
decisive
very strong
substantial
no evidence
(k)
Gst (Landraces vs Lines)
strong
decisive
very strong
substantial
no evidence
(l)
Gst (Landraces)
0 50 120 190 260 0 50 120 190 0 50 120 190 0 50 120 190 0 50 120 1900 50 120 0 50 120 0 50 120 0 50 120 0 50 120
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Genome Position (Mbp)

**Figure 3** Variation in genetic diversity and differentiation along the maize genome. (a) Total expected heterozygosity across landraces: Ht (Landraces); (b) total expected heterozygosity across inbred lines: Ht (Lines); (c) difference between the total expected heterozygosity across landraces and across inbred lines: Ht (Landraces) – Ht (Lines); (d) mean expected heterozygosity within landraces: Hs (Landraces); total expected heterozygosity across landraces from (e) Europe: Ht (EUR)), (f) North America: Ht (NAM), (g) South America: Ht (SAM); (h) Central America and Mexico: Ht (CAM), (i) the Caribbean: Ht (CAR), (j) Gst between geographic groups of landraces: Gst (Geographic); (k) Differentiation between landraces and inbred lines: Gst (Landraces vs. Inbred lines); (l) Differentiation between landraces: Gst (Landraces). Loci with decisive, very strong, strong, substantial, no evidence of selection using BAYESCAN are coloured in orange, dark green, light green, yellow and blue (j-l). Vertical grey bars correspond to centromere limits. Chromosome boundaries are indicated by vertical dashed lines. Horizontal dashed lines correspond to the mean, 5th and 95th percentile of each parameter. Outlier regions are indicated by red asterisks (>95% at the top, <5% at the bottom). Vertical blue lines indicate the location of the genes *ID1, tb1, pbf1, su1, tga1, bt2, o2, pebp8, vgt1, nac1* and *Zmcct*.

estimated with 979 haplotype markers the average number of new haplotypes discovered in the 66 DH-SSD lines compared to those existing in the CK lines. We discovered 66 new haplotypes in the DH-SSD lines compared to 4355 different haplotypes in the CK lines. The number of new haplotypes discovered in DH-SSD lines ranged from 0 (Bul3) to 11 (Arg8). The average number of new haplotypes was significantly higher for lines derived from landraces with a low contribution than for those with a high contribution (*P*-value = 0.008, Figure 5b). It was also higher for landraces that were not close to any of the CK lines than for those that were close to certain lines (*P*-value = 0.0004, Figure 5d).

## Discussion

### Patterns of genetic diversity and population structure within landraces

We applied a DNA bulk approach with 50 K maize Illumina array developed by Arca *et al*., (2021) to decipher genetic diversity of a worldwide panel of 156 American and European landraces represented by 2340 individuals and compare it to 327 inbred lines. Compared to sequencing approach or SSR markers system, this approach is affordable (20–50€ *per* landraces including DNA extraction), high-throughput, labour-efficient, does not require strong bioinformatic and biological molecular skills and facilities. This approach produce genotyping data with very few missing genotyping data and low error rate and are easy to standardize across laboratories (see Arca *et al*. (2021) for more detailed discussion). For these reasons, this method can be easily and rapidly implemented and applied in a decentralized way in genebanks, academic laboratories and breeding companies. Our approach can also take advantage of huge number of maize inbred lines that have been already genotyped by 50 K arrays by breeding companies, academic laboratories and genebanks. However, using 50 K maize illumina array could lead to some ascertainment bias in diversity analysis.

The total expected heterozygosity (Ht) observed in our study based on SNPs (0.338) was lower than the values reported previously for landraces of comparable origin that were analysed with SSR markers (0.58 in Rebourg *et al*. (2003), 0.63 in Camus-Kulandaivelu *et al*. (2006), 0.62 in Dubreuil *et al*. (2006), 0.61 in this study). Similarly, the within genetic diversity of individual landraces (Hs) estimated with SNPs ($Hs_{SNP}$) is 1.6 times lower than Hs based on SSRs ($Hs_{SSR}$ = 0.385 vs. $Hs_{SNP}$ = 0.192; Table S1). Both estimates are nevertheless highly and linearly correlated among landraces ($r^2$ = 0.73). These differences can be primarily explained by the fact that SNP markers are typically bi-allelic, whereas SSR markers are multi-allelic, which has the potential to increase gene diversity (Frascaroli *et al*., 2013; Hamblin *et al*., 2007). The diversity of individual landraces estimated with SNPs represented on average 57% of the total genetic diversity, which was slightly lower than for SSR markers in this study (63%) or in previous studies with SSR and RFLP markers (~66% in Mir *et al*. (2013) and Rebourg *et al*. (2003)). Correlatively, the genetic differentiation between individual landraces estimated with SNPs ($Gst_{SNP}$ = 0.432) was slightly higher than that estimated with SSRs in this study ($Gst_{SSR}$ = 0.369). It was also higher than in the previous studies using SSR and RFLP markers (Gst = 0.343 with RFLPs in Mir *et al*., 2013, Gst = 0.313 with SSRs in Rebourg *et al*., 2003). This difference may be due to the counter-selection of SNP markers with low MAF during the design of 50 K Illumina array (Ganal *et al*., 2011), which may increase total diversity more than within diversity (Albrechtsen *et al*., 2010; Clark *et al*., 2005). The selection of the most common variants between lines from different geographical origins during array design could also explain the lower global genetic differentiation between landraces from different geographical origins as compared to SSR markers and to sequencing data (Brandenburg *et al*., 2017). This could lead to underestimate the number of genomic region under selection between geographical groups. On the other hand, genetic structure analyses of landraces based on SNPs were highly congruent with analyses based on SSRs in previous studies (Camus-Kulandaivelu *et al*., 2006; Mir *et al*., 2013). The proportion of landraces assigned to same group by SSRs and SNPs was 98%, 100%, 87%, 81%, 79%, 72% for *K* = 2, 3, 4, 5, 6, 7, respectively. Additionally, MRD between 156 landraces based on SNPs and SSRs were highly and linearly correlated (*r* = 0.78) as previously shown in Arca *et al*., (2021). As SSRs were free of ascertainment bias, it indicates that the ascertainment bias of prefixed PZE SNPs from the 50 K Illumina chip used to study landraces has negligible consequences, notably on genetic distance and structure analysis (Arca *et al*., 2021; Frascaroli *et al*., 2013).

Each geographic group contained most of the alleles present in the overall landrace panel, proportion of polymorphic loci ranging from 89% (CAR) to 97% (CAM). Central American and Mexican landraces displayed the highest diversity, which is consistent with their proximity to the centre of maize domestication (Matsuoka *et al*., 2002; van Heerwaarden *et al*., 2011). This confirms that genetic diversity was lost during the spread of maize away from its domestication centre due to successive bottlenecks related to climatic adaptation and isolation by distance (Brandenburg *et al*., 2017; Gates *et al*., 2019; Romero Navarro *et al*., 2017; Swarts *et al*., 2017; Tenaillon and Charcosset, 2011). This loss of genetic diversity is consistent with the scenario of maize diffusion with (i) less genetic diversity in European than in North and South American landraces, and (ii) more diversity in South America than in North America, where maize was introduced more recently

**Table 2** Genomic regions identified as being highly differentiated between landraces and geographic groups. Only SNPs that were detected by BAYESCAN with decisive evidence of selection and Outlier Fst windows carrying at least two SNPs are listed.

| Outlier Gst windows | | | | | | | | BAYESCAN hits (Decisive) – Geographical | | | | | | | Frequency of allele B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Name[†] | Chr | Start – Stop (Mbp) | $SNP_w$ | $Gst_g$ | $Gst_l$ | $Ht_l$ | Hs | $SNP_b$ | Marker name | Pos. (Mbp) | $Fst_b$ | Closest Gene | Dist. from gene (kbp) | Functional annotation | EUR | NAM | CAM | CAR | SAM |
| Sg1, Sp2 | 3 | 77.5–79 | 4 | 0.15 | 0.54 | 0.39 | 0.15 | 4 | PZE-103058385 | 78.2 | 0.26 | GRMZM2G584078 | 4 | | 0.76 | 0.73 | 0.39 | 0.00 | 0.53 |
| | | | | | | | | | PZE-103058429 | 78.5 | 0.30 | AC202959.3_FG001 | 0 | | 0.30 | 0.34 | 0.72 | 1.00 | 0.54 |
| | | | | | | | | | PZE-103058437 | 78.5 | 0.29 | GRMZM2G112187 | 6 | | 0.69 | 0.67 | 0.32 | 0.00 | 0.43 |
| Sg2, Sp3 | 3 | 84–85 | 3 | 0.18 | 0.63 | 0.50 | 0.19 | 3 | PZE-103059206 | 82.1 | 0.26 | GRMZM2G154496 | 0 | | 0.71 | 0.67 | 0.33 | 0.00 | 0.45 |
| | | | | | | | | | PZE-107023081 | 84.9 | 0.29 | GRMZM2G112579 | 5.6 | Pectin lyase-like superfamily protein | 0.69 | 0.65 | 0.32 | 0.00 | 0.44 |
| | | | | | | | | | PZE-107023082 | 84.9 | 0.29 | | 5.7 | | 0.31 | 0.35 | 0.64 | 1.00 | 0.52 |
| Sg4, Sp6 | 4 | 7.8–9.4 | 7 | 0.27 | 0.63 | 0.23 | 0.07 | 6 | PZE-104010475 | 7.6 | 0.30 | GRMZM2G012821 | 0 | F-box protein | 0.04 | 0.06 | 0.77 | 0.19 | 0.20 |
| | | | | | | | | | PZE-104010477 | 7.6 | 0.31 | | 0 | | 0.97 | 0.95 | 0.24 | 0.83 | 0.81 |
| | | | | | | | | | PZE-104010709 | 8.8 | 0.30 | GRMZM2G119698 | 0 | pectinesterase | 0.06 | 0.06 | 0.79 | 0.29 | 0.22 |
| | | | | | | | | | PZE-104010719 | 8.8 | 0.28 | GRMZM2G702341 | 0.2 | | 0.98 | 0.95 | 0.34 | 0.95 | 0.84 |
| | | | | | | | | | PZE-104010855 | 9.4 | 0.27 | GRMZM2G419836 | 0 | Thioredoxin superfamily protein | 0.98 | 0.96 | 0.42 | 0.80 | 0.94 |
| Sg5, Sp7 | 4 | 40.9–41.9 | 7 | 0.16 | 0.63 | 0.44 | 0.16 | 4 | PZE-104033199 | 41.2 | 0.26 | GRMZM5G889780 | 13.7 | | 0.43 | 0.28 | 0.90 | 1.00 | 0.79 |
| | | | | | | | | | PZE-104033229 | 41.4 | 0.28 | GRMZM2G138198 | 0 | Pollen receptor-like kinase 4 | 0.49 | 0.66 | 0.06 | 0.00 | 0.22 |
| | | | | | | | | | PZE-104033340 | 41.7 | 0.27 | GRMZM2G174149 | 0 | RNA pseudouridine synthase 3 mitochondrial | 0.54 | 0.39 | 0.94 | 1.00 | 0.82 |
| Sg9, Sp11 | 6 | 134.3–135.3 | 15 | 0.16 | 0.58 | 0.41 | 0.17 | 6 | PZE-106078726 | 134.5 | 0.25 | GRMZM2G055678 | 0 | Proline-rich receptor-like protein kinase PERK1 | 0.51 | 0.34 | 0.96 | 0.99 | 0.77 |
| | | | | | | | | | PZE-106078990 | 134.8 | 0.24 | GRMZM2G170646 | 0 | GDSL esterase/lipase | 0.50 | 0.63 | 0.07 | 0.01 | 0.25 |
| | | | | | | | | | PZE-106079041 | 134.8 | 0.28 | | 0.6 | | 0.55 | 0.44 | 0.97 | 1.00 | 0.80 |
| | | | | | | | | | PZE-106079060 | 134.9 | 0.25 | GRMZM2G162702 | 0 | Probable receptor-like protein kinase | 0.57 | 0.49 | 0.96 | 1.00 | 0.82 |
| | | | | | | | | | PZE-106079065 | 134.9 | 0.27 | | 0 | | 0.57 | 0.49 | 0.98 | 1.00 | 0.81 |
| | | | | | | | | | PZE-106079127 | 135.0 | 0.29 | GRMZM2G307720 | 0 | TATA box-binding protein | 0.49 | 0.31 | 0.92 | 1.00 | 0.72 |

[†] Sg and Sp indicate highly differentiated genomic regions between geographic groups and landraces, respectively; $SNP_w$ and $SNP_b$ indicate the number of SNPs within Outlier Gst windows and detected as being under selection by BAYESCAN, respectively; $Gst_g$ and $Gst_l$ indicate the average Gst across all loci in the window, and between geographic groups and landraces, respectively. $Fst_b$ indicates the Fst estimated by BAYESCAN for markers under selection between geographic groups. Distance from gene ('Dist. from gene') was based on the closest start or stop codon of the gene, 0 indicates that the SNP is within the gene. Functional annotation was retrieved from Gramene (https://www.gramene.org/).

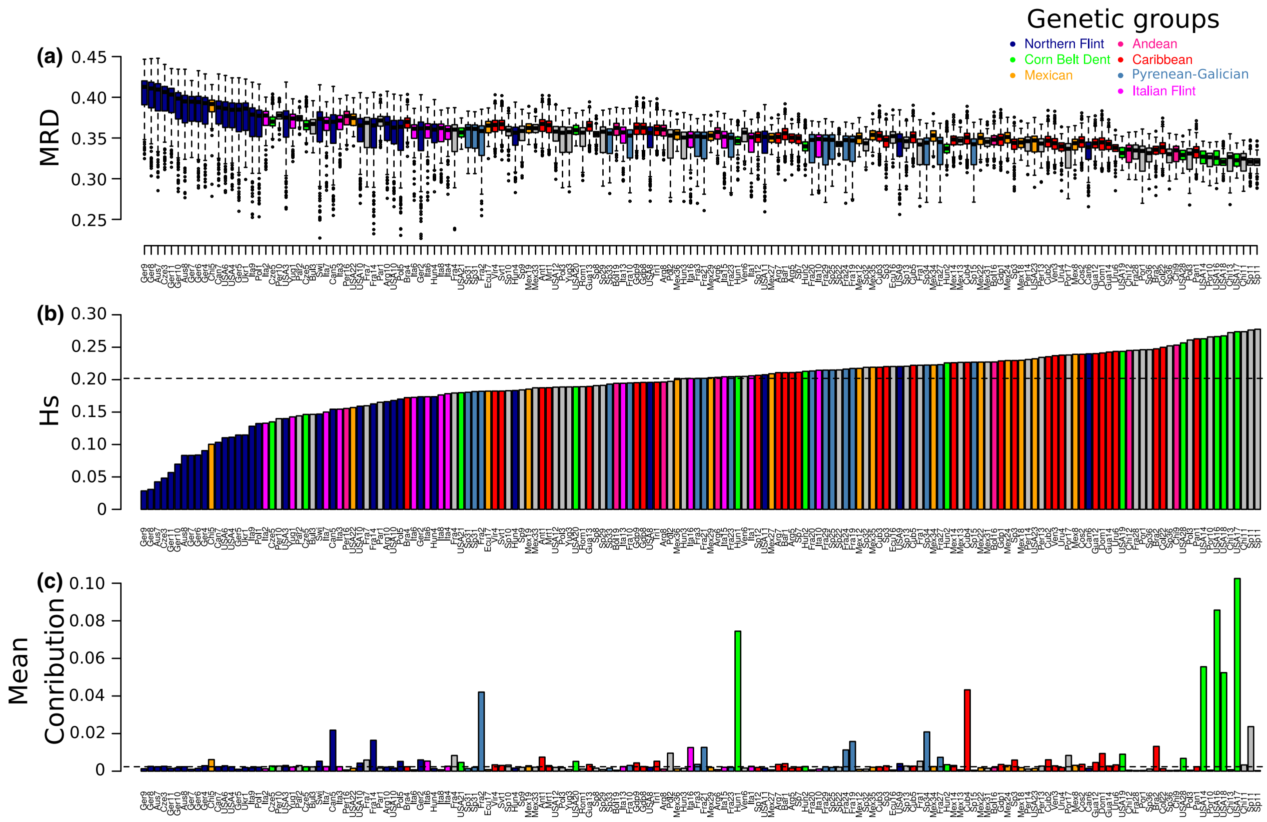

**Figure 4** Contribution of landraces to the panel of CK lines in relation to their genetic diversity. (a) Box plot representation of pairwise modified Roger's distances (MRD) between individual landraces and CK lines. Each box represents the interquartile range, the line within each box represents the median value and the error bars encompass 95% of values for each landrace. Circles represent outliers. (b) Within population genetic diversity (Hs) (c) Average contribution of the 166 landraces to the panel of CK lines estimated by supervised analysis with ADMIXTURE. Landraces are ranked in ascending order of Hs in the three figures. Boxplot and barplots are coloured based on the assignment of landraces to the seven genetic groups identified by ADMIXTURE (see bottom right for colours).

(Brandenburg *et al.*, 2017; Mir *et al.*, 2013; Tenaillon and Charcosset, 2011; Vigouroux *et al.*, 2005). Our results nevertheless confirm that the bottleneck during the introduction of maize into Europe was certainly limited, as also shown by Brandenburg *et al.* (2017). Some northern European landraces originating from Germany and Austria have extremely low genetic diversity (Hs < 0.10), with more than 70% of loci being fixed, suggesting a strong bottleneck. It relates to the fact that some of these landraces have been cultivated mostly in gardens which may have decreased their effective population size (Rebourg *et al.*, 2003).

Genetic distance, Ward hierarchical clustering (Figure 1b), principal component (Figure 1a) and population structure (Figure 2) confirmed the central position of Mexican and Caribbean landraces and a clear differentiation between North and South American landraces (Figure 1; Figure 2). This is consistent with the domestication of maize in Mexico followed by southwards and northwards dispersion (Romero Navarro *et al.*, 2017; Swarts *et al.*, 2017). The similarity between landraces from southern Spain and the Caribbean confirms the historical data on the introduction of maize in the south of Spain by Columbus in 1493 after his first trip to the Caribbean (Figure 1b, cluster d). Strong similarities between groups of northeastern American and northeastern European landraces (mostly from Germany, Poland and Austria) (Figure 1b, cluster g) also supports an independent introduction of North American material that was pre-adapted to the northern European climate (Brandenburg *et al.*, 2017; Dubreuil *et al.*, 2006; Dubreuil and Charcosset, 1999; Rebourg *et al.*, 2003; Swarts *et al.*, 2017; Tenaillon and Charcosset, 2011). Some landraces from northern Spain and southwestern France, located along the Pyrenean Mountains, were admixed either with Caribbean or Northern Flint. This result supports the hypothesis that new Pyrenean-Galicia Flint groups originated from hybridization between Caribbean and Northern Flint material that were introduced in southern Spain and northern Europe, respectively (Brandenburg *et al.*, 2017; Camus-Kulandaivelu *et al.*, 2006; Diaw *et al.*, 2020). Interestingly, some southwestern Spanish landraces have elevated admixture with Italian Flint groups and are closely related to Italian landraces on the NJ tree (Figure S5), while northern Spanish landraces (latitude >42°N) do not. These results support the hypothesis that Italian landraces are probably derived from an ancestor from southern Spain (Brandenburg *et al.*, 2017; Revilla *et al.*, 1998). Our results also highlighted a new putative hybridization event in Central Eastern Europe. Central Eastern European landraces were close to Italian Flint landraces on the Ward cluster tree and one northern Italian Flint landrace (Nostrano Quarantino) was admixed with Italian Flint (~30%–40%) and Northern Flint (~30%–50%). This suggests that Italian Flint landraces certainly spread in Central Eastern Europe, where they intermated with Northern Flint landraces.

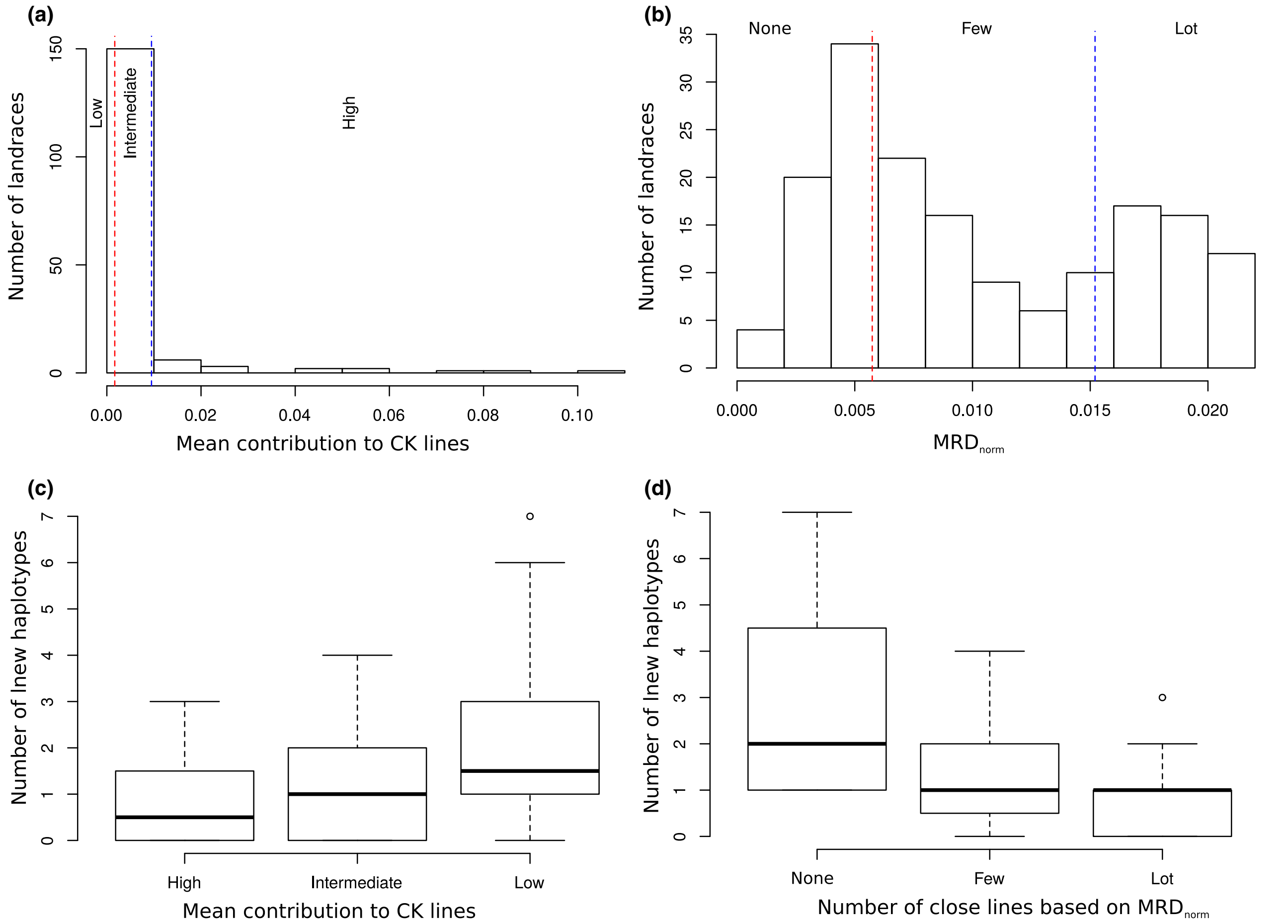

**Figure 5** Allelic enrichment of CK lines by new DH-SSD lines derived from landraces according their contribution and their genetic distance to CK lines. Allelic enrichment was estimated by the number of new haplotypes discovered in the 66 new DH-SSD lines derived from 33 landraces, compared to the 327 CK lines (c, d). These 33 landraces are classified in 3 classes according to the distribution of (a) the average contribution to CK line panel using supervised analysis and (b) the normalized MRD ($MRD_{norm}$) of the 10% closest CK lines with each landrace. Red and blue vertical dotted lines delineate the limits of three landrace classes displaying (a) low, intermediate and high contribution; (b) the presence of none, few and many closely related lines based on $MRD_{norm}$.

Differentiation of landraces was greater in Europe than in Central America and the Caribbean, indicating that gene flow is higher in the latter two. Genetic and geographic distances were weakly but significantly correlated in NAM, EUR and CAM but not in SAM and CAR ($r^2$ ~ 0.1 in Figure S6). It suggests that isolation by distance could play a role in shaping the genetic structure of maize landraces in these regions, albeit to a variable degree. In the case of CAM, the effect of isolation by distance is partially blurred by variation in altitude. Indeed, Mexican landraces clustered according to both altitude and distance (Figure 1b; Table S1) suggesting a role of environmental adaptation (Gates *et al*., 2019; Wang, Josephs, *et al*., 2021). Altitude is related to several environmental factors (temperature, rainfall) which change over short geographical distances and certainly contributed to genetic differentiation between landraces within several geographical groups (Aguirre-Liguori *et al*., 2017; Gates *et al*., 2019; Wang, Josephs, *et al*., 2021). This is expected to blur the effect of isolation by distance and likely explains the low correlation between geographical and genetic distances (Aguirre-Liguori *et al*., 2017; Gates *et al*., 2019; Wang, Josephs, *et al*., 2021).

## Genomic pattern of nucleotide variation in landraces

Gst outlier and BAYESCAN analyses identified 26 genomic regions that showed high levels of differentiation between geographic groups and/or landraces (Table 2; Table S5). 16 out 26 genomics regions were previously identified in American Landraces by Romero Navarro *et al*. (2017) as associated with male or female flowering time (7), with variation of latitude (11) or altitude (6) (Table S5). 21 out 26 genomic regions were previously identified in 67 American and European first cycle lines by Brandenburg *et al*. (2017) as differentially selected between geographical groups (11), associated with excess of heterozygosity (17), latitude (0) or longitude (1) (Table S5). Interestingly, we identified four new genomic regions including the two highest differentiated regions between landraces (Sp1 and Sp10), although marker density in our study was, respectively, 40 and 1000 fold lower than in Romero Navarro *et al*. (2017) and Brandenburg *et al*. (2017), respectively. The Sp1 and Sp10 regions was not differentiated between the five geographical groups suggesting that strong selection occurred in some geographic areas sharing similar agro-climatic condition gradients. For instance, Sp10 contains genes associated with tolerance to high temperature and evaporative demand (Millet *et al*., 2016). The other most differentiated genomic regions between geographical groups and landraces have been previously identified either by Romero Navarro *et al*. (2017) or Brandenburg *et al*. (2017) (Table S5). The highest differentiated genomic region between geographical groups (Sg4-Sp6: 7.8–9.3 Mbp on chromosome 4) was nearly fixed in temperate landraces (NAM, EUR), whereas it showed intermediate frequencies in CAM, suggesting a strong directional selection effect during the spread from Mexico to North America. Accordingly, Romero Navarro *et al*. (2017) identified 4 SNPs in this region with allelic frequencies varying significantly with latitude in American landraces, and Brandenburg *et al*. (2017) identified two highly differentiated regions between Corn Belt Dent and Tropical first cycle lines. By contrast, the genomic region (Sg5–Sp7; 40–41.9 Mbp on chromosome 4) displayed higher genetic diversity in temperate landraces (NAM, EUR) than in tropical landraces (CAM, CAR) suggesting strong diversifying selection in EU and NAM. This region includes the *Su1* gene, which is involved in the starch pathway and is known to be under strong selective pressure (Jaenicke-Despres *et al*., 2003; Revilla and Tracy, 1995; Tracy *et al*., 2006; Whitt *et al*., 2002). Romero Navarro *et al*. (2017) also found an association between allelic frequency variation at the *Su1* locus and both latitude and longitude. Furthermore, we identified a strong selective sweep between Corn Belt Dent/Tropical and Northern Flint first cycle lines in the *Su1* gene. The second genomic region with highest differentiation between geographical groups (Sg2–Sp3; 84–85 Mbp on chromosome 3) showed a continuous gradient of allelic frequencies between tropical and temperate landraces suggesting strong directional selection for adaptation either to temperate or tropical climates. Accordingly, Romero Navarro *et al*. (2017) identified in this region 22 and 4 SNPs with allelic frequencies varying significantly with altitude and latitude, respectively.

BAYESCAN analysis between geographic groups identified several regions that were not identified by outlier Gst analysis (Table S7; Table S8). Notably, we identified several loci under strong selection that were close to genes known to be involved in flowering time variation: (i) PZE-108070380 on chromosome 8 (123.5 Mbp) localized 5 kbp upstream of *Zcn8* (Bouchet *et al*., 2013; Gouesnard *et al*., 2017; Romay *et al*., 2013); (ii) PZE-109070904 on chromosome 9 (115.7 Mbp) in *ZmCCT9* (Huang *et al*., 2017); (iii) two loci on chromosome 3 PZE-103098664 (158.9 Mbp) and PZE-103098863 (159.17 Mbp) close to *Vgt3*, a major loci that is strongly associated with flowering time variation in temperate maize (Millet *et al*., 2016; Negro *et al*., 2019). We also identified several genes/genomic regions that are putatively involved in adaptation to abiotic stress: (i) PZE-102108435 on chromosome 10 that is 10 kbp upstream of *ZmASR2* which is involved in abscisic stress ripening (Virlouvet *et al*., 2011); (ii) PZE-104128228 on chromosome 4 in the *nactf125* gene (within Sg6 in Table S5), PZE-102051809 in the *nactf36* gene (chromosome 1) and PZE-107058109 in the *nactf14* gene (chromosome 7), all of which belong to the NAC protein family, which encodes plant transcription factors involved in biotic and abiotic stress responses (Yilmaz *et al*., 2009); (iii) two diaglycerol kinases (*dgk2* and *dgk3*) that exhibit differential expression patterns in response to abiotic stress including cold, salinity and drought and are upregulated in cold conditions (Gu *et al*., 2017). Finally, we identified several genomic regions carrying genes involved in the hormonal systems regulating growth, cell division and proliferation such as giberellin2-oxydase9 (*ZmGA2ox9*, GRMZM2G152354), phytosulfakine (GRMZM2G031317) or in the starch pathway (*Su1*, *waxy1*, *dull endosperm1*).

The detection of genomic regions and loci under selection has therefore allowed the identification of genes that underlie the adaption of maize to diverse agro-climatic conditions and/or human uses during the spread of landraces from America (Brandenburg *et al*., 2017; Gates *et al*., 2019; Mir *et al*., 2013; Romero Navarro *et al*., 2017; Swarts *et al*., 2017; Wang, Lin, *et al*., 2020). These genomic regions could be useful for mining new alleles from landraces, retrieving some of the genetic diversity that was lost by genetic drag linked to genes close to those under selection (Gates *et al*., 2019; Hufford *et al*., 2012; Wang, Lin, *et al*., 2020), or creating new genetic diversity by targeted mutation (Gates *et al*., 2019).

## Identification of promising landraces to enlarge the modern genetic pool

Intensive selection to enhance agronomic performance can considerably reduce genetic diversity in crops (Tanksley, 1997). However, we found little difference in genetic diversity and a low genetic differentiation between landraces and inbred lines. This suggests that the genomic diversity (inferred from SNPs) present in landraces was retained in our panel of CK lines and that selection during the first steps of modern maize improvement has not altered allele diversity over a very broad geographic scale. This observation is similar to findings in soybean (Hyten *et al*., 2006) and wheat (Cavanagh *et al*., 2013), which also showed a minor effect of crop improvement on diversity. It is important to note that our line panel included many old lines that have made only a limited contribution, if any, to commercial F1 hybrids or recent breeding pools. Our panel therefore certainly overestimates the genetic diversity present in the germplasm of modern breeding inbred lines (Zeitler *et al*., 2020).

Several factors could be responsible for the low apparent genetic erosion accompanying the transition from landraces to inbred lines. A first hypothesis is that selection during modern maize breeding targeted only a small number of genes (Wright *et al*., 2005) and therefore affected genetic diversity and allelic frequency only in the genomic regions flanking the genes under selection. Another hypothesis is that, even if only a limited number of landraces were used as parents of first cycle lines, that is, the initial modern inbred line breeding pools, selection of genetically diverse and complementary heterotic groups may have mitigated the loss of diversity (Jiao *et al*., 2012). Furthermore, SNPs from 50 K arrays were previously identified in 27 lines (Gore *et al*., 2009). These SNPs may not reflect well the total genetic diversity of landraces, as certain specific landrace haplotypes may not have been transmitted to first cycle lines due to their deleterious effect at the homozygous state (inbreeding depression) or gamete sampling (drift) (Zeitler *et al*., 2020).

Despite the limited differences in overall diversity between landraces and inbred lines, two different approaches highlighted that the majority of landraces have made a limited contribution to recent breeding. Several landraces have a high median Hs value and a small $MRD_{LI}$ distance range reflecting a lack of similarity to any inbred line. These landraces probably did not contribute to the modern maize germplasm. Indeed, supervised analysis showed that inbred lines from our diversity panel could be traced back to a few landraces and that the first 10 landraces cumulated half of the total contribution to the diversity panel. Most of these landraces (Reid's Yellow Dent, Lancaster Surecrop and Krug

Yellow Dent for the dent genetic group, Lacaune and Gaspe Flint for the flint genetic group and Chandelle for Tropical lines) were previously identified as the source of the modern maize breeding germplasm (Gerdes and Tracy, 1993; Romero Navarro *et al*., 2017; van Heerwaarden *et al*., 2011). Interestingly, we observed a large increase or decrease in the contribution of landraces between first cycle lines and more advanced lines (Figure S13C). This can be explained by the fact that some lines were extensively used to derive more advanced lines whereas others were not (Coffman *et al*., 2020; Gerdes and Tracy, 1993; Mikel, 2011). Interestingly, DH-SSD lines that were recently derived from landraces were assigned more frequently (and with higher probability) to their population of origin than older lines that were maintained for a long time in gene banks. This suggests that some landraces could have evolved since contributing to inbred lines from the diversity panel or that the pedigree of these lines was erroneous. Our results suggest that we could use supervised analyses to curate the landrace collection and the pedigree of first cycle lines.

In order to identify landraces that differ the most from inbred lines, we developed an indicator of genetic distance from inbred lines which was normalized by their genetic diversity (Figure S12). By classifying landraces according to (i) this normalized distance and (ii) their average contribution to reference inbred lines, we were able to identify landraces that have the greatest potential to broaden the genetic diversity of these lines (Figure 5). By combining closely located SNPs, we identified novel haplotypes in the DH-SSD lines, which were absent in the CK panel, even though both alleles were present in landraces and the inbred line panel. The number of new haplotypes was significantly higher for DH-SSD lines created from landraces classified as genetically distant from the modern germplasm according to the criteria described previously, which confirms their relevance when choosing landraces for diversity enhancement. This strategy to identify untapped landraces in modern breeding germplasm can be easily extended to other plant species, other material (hybrids, private germplasm) and other technologies (sequencing). Additionally, this strategy can be focused on some genomic region to identify new alleles of interest. Our strategy opens an avenue to identify valuable landraces and genomic regions for prebreeding.

## Experimental procedures

### Plant material

#### *Landraces*

A total of 156 different landrace populations (Table S1) were sampled from a panel of 413 landraces (Appendix S1). These 156 landraces were represented by a total of 2340 individual plants and captured a large proportion of European and American diversity and have been analysed in previous studies using RFLP (Dubreuil *et al*., 1999; Dubreuil and Charcosset, 1998; Gauthier *et al*., 2002; Rebourg *et al*., 1999, 2001) and SSR markers (Camus-Kulandaivelu *et al*., 2006; Dubreuil *et al*., 2006; Mir *et al*., 2013). Each landrace population was represented by either one or two sets of 15 individual plants (for 146 and 10 populations, respectively), pooled equally as described in Reif *et al*. (2006) and Dubreuil *et al*. (2006). Plants were germinated directly from genebank seeds, with no prior self-pollination. A DNA pool was obtained for each landrace by mixing equal amount of leaves from 15 individual plants prior to DNA extraction. The 166 DNA samples corresponding to the 156 landrace accessions were classified into five geographic groups (Table S1).

#### *Inbred lines*

We analysed 234 inbred lines that were derived directly by single seed descent or by haplodiploidization of landraces, referred to as 'first cycle lines', and 208 lines that were derived from a more advanced cycle of breeding, referred to as 'advanced lines' (Table S10). These 442 lines were partitioned into three sets (the 'Panel' column in Table S10):

1. 'CK lines': a panel of 120 first cycle and 207 advanced lines (327 lines in total) representing American and European diversity (Bouchet *et al*., 2013; Camus-Kulandaivelu *et al*., 2006) including some key founders of modern breeding programmes (*e.g*. F2, B73, C103).
2. 'Parent Controlled Pools': a set of 12 lines used to build 4 series of 8 controlled DNA pools. Two series were used to assess the accuracy of our genotyping method and to calibrate the model for predicting allelic frequency (see Arca *et al*. (2021) for more detail).
3. 'DH-SSD lines': a set of 45 single seed descent (SSD) and 58 double haploid (DH) lines derived recently from 48 landraces (first cycle lines).

### Genotyping and prediction of allelic frequencies in DNA pools

We used the 50 K Illumina Infinium HD array (Ganal *et al*., 2011) to genotype (i) landraces, (ii) controlled DNA pools, (iii) the DH-SSD inbred lines and (iv) the parental lines of the controlled DNA pools (Table S1; Table S10). For CK lines, we used the 50 K genotyping data from Bouchet *et al*. (2013). 23 412 SNPs were filtered based on their suitability for diversity analysis and their quality for predicting allelic frequency in DNA pools (Appendix S2). We also used genotyping of 17 SSRs from Mir et al.[23] for landraces and Camus-Kulandaivelu et al.,[27] for inbred lines to evaluate ascertainment bias due to array design on diversity parameters estimated by SNPs.

Allelic frequency of selected SNPs in DNA pools was estimated using the two-step procedure described in Arca *et al*. (2021) based on the fluorescence intensity ratio (FIR) of alleles A and B for each SNP. First, we tested whether SNPs were monomorphic or polymorphic. For SNPs that were considered to be polymorphic, we then estimated the allelic frequency of the B allele using a generalized linear model calibrated on FIR data from 1000 SNPs from 2 series of controlled pools (see Arca *et al*. (2021) for more detail and equation 2 for the model). This two-step approach led to a global mean absolute error of 3% and was more conservative for SNP fixed or close to fixation than for SNP with balanced allelic frequency (Arca *et al*., 2021). Threshold to reject hypothesis that landraces were monomorphic was set to 5%, indicating that 5% of landraces are expected to be declared polymorphic, whereas they are monomorphic.

### Diversity analyses

#### *Estimation of genetic diversity parameters*

For each landrace, each geographic group, all landraces combined and the panel of inbred lines, we determined for each locus: the mean allele number (A), the Minor Allele Frequency (MAF) and the expected heterozygosity (H) (Nei, 1973, 1977). Considering that all landraces were represented by 15 different plants (30 gametes), we did not apply a correction for the number of individuals in estimating these parameters because it would conduct to only a small increase for diversity parameters (3.4% according to Nei and Chesser, 1983).

Genetic differentiation (Gst) was estimated using 23 412 SNPs ans 17 SSRs according to Nei (1973) between: individual landraces ($Gst_l$), between the five landrace geographic groups ($Gst_g$), between 10 pairs of geographic groups ($Gst_{EUR\text{-}NAM}$, $Gst_{EUR\text{-}CAM}$, $Gst_{EUR\text{-}CAR}$, $Gst_{EUR\text{-}SAM}$, $Gst_{NAM\text{-}CAM}$, $Gst_{NAM\text{-}CAR}$, $Gst_{NAM\text{-}SAM}$, $Gst_{CAM\text{-}CAR}$, $Gst_{CAM\text{-}SAM}$, $Gst_{CAR\text{-}SAM}$) and between landraces and inbred lines ($Gst_i$). Gst was estimated at each locus and across all loci as *per*[80,81,82] (Appendix S3).

### *Genome-wide diversity analysis and scans for identifying selection signatures*

We used a sliding window of 1 Mbp, shifting by 500 kbp at each step along the genome, to analyse the genome-wide variation in genetic diversity and differentiation between landraces, between geographic groups and between landraces and inbred lines. The maize genome was divided into 4095 overlapping windows containing an average of 11.3 ± 5.2 SNPs. We computed the average value for the parameters described above for all loci in a given window. Outlier regions for H and Gst were identified based on the distribution of these parameters for individual loci over the entire genome using the 5th and 95th percentile (below 5% and above 95%) as thresholds (Table S4). All statistics were computed using *ad hoc* scripts in the R language v 3.0.3 (R Core Team, 2013).

Genomic scans were carried out to detect the genomic signature of selection between landraces, between the five geographic groups and between landraces and inbred lines using two approaches: (i) the detection of 1 Mbp regions that were outliers for Gst, referred to as 'Outlier Gst analysis' (ii) the detection of loci under selection using the drift model implemented in the BAYESCAN software (Foll and Gaggiotti, 2005; Appendix S4).

### *Genetic structure and relationship between landraces*

We estimated the genetic distance between all landraces using modified Roger's distance (MRD) (Rogers, 1972) based on the allelic frequencies of 23 412 prefixed PZE SNPs ($MRD_{SNP}$) and 17 SSR ($MRD_{SSR}$). We used Mantel test to test the correlation between genetic and geographic distances within each geographic group. (Smouse *et al*., 1986). Geographic distances were calculated using the latitude and longitude of each sampling site using the geosphere R package v. 1.5–10 (Hijmans, 2019).

To decipher the structure of genetic diversity within our panel of landraces from 23 412 filtered SNPs, we used two approaches:

1. A distance-based approach in which MRDs between the 166 landraces were used to perform (i) a principal coordinate analysis (PCoA) (Gower, 1966), (ii) hierarchical clustering using either Ward or Neighbour-Joining algorithms implemented in the 'hc' and 'bionj' functions of the 'ape' R package v 5.0 (Paradis and Schliep, 2019), respectively.
2. A Bayesian multi-locus approach, implemented in the ADMIXTURE software, to assign probabilistically each landrace to K ancestral populations assumed to be in Hardy–Weinberg Equilibrium (Alexander *et al*., 2009). Different methods were used to identify the most appropriate number of ancestral populations (K): Cross-validation error or difference between successive cross-validations (Alexander *et al*., 2009) and Evanno's graphical methods (Evanno *et al*., 2005). Since ADMIXTURE requires multi-locus genotypes of individual plants, we simulated the genotype of five individuals for each population for a subset of 2500 independent SNPs to avoid artefacts of linkage disequilibrium (Appendix S5).

## Contribution of populations to inbred lines using supervised analysis and modified Roger's distance

To analyse the contribution of landraces to the modern breeding germplasm, we used two different approaches:

1. A distance-based approach in which we estimated the modified Roger's distance between each landrace and the 327 CK lines ($MDR_{LI}$) in order to determine whether they are related or not.
2. A Bayesian supervised approach implemented in ADMIXTURE in which the 442 inbred lines were assigned probabilistically to the 166 landrace populations in order to identify the most likely source population of each inbred line (Table S10). For each landrace, we estimated (i) its average contribution to CK lines by averaging the assignment probability over 327 lines and (ii) the number of inbred lines mainly assigned to this landrace, with an assignment probability >60%. We also analysed the evolution of the contribution of landraces across breeding cycles by comparing contributions to (i) first cycle lines and (ii) advanced lines from the CK line panel. To check the accuracy of the assignment method, we estimated the percentage of first cycle lines that were correctly assigned to their parental landrace as known from their pedigree and analysed in our study (121 of the 234 first cycle lines, known to be derived from 50 landraces). We tested if this percentage was different between CK lines and DH-SSD lines using a Kruskal-Wallis chi-squared test. To represent each landrace, we used the same five simulated individuals as in the structure analysis.

## Identification of landraces that could enrich the modern breeding germplasm

We assessed whether the mean contribution of landraces and their $MRD_{LI}$ distribution parameters could be used as criteria to identify landraces that could enrich the modern breeding germplasm. To this end, allelic diversity was estimated in the two inbred panels (DH-SSD and CK lines) for 979 haplotypes. These haplotype markers were defined by genotyping triplets of adjacent SNPs from 50 K arrays that were less than 2 kbp apart. We estimated the average number of new haplotypes discovered in the DH-SSD lines compared to those in the 327 CK lines. To avoid noise due to seedlot error during DH-SSD line production, we selected 66 DH-SSD lines that were correctly assigned to 33 landraces analysed from this study.

To analyse the effect of mean contribution, we classified these 33 landraces into three classes: low, intermediate, and high contribution based on the 30th and 90th percentile of the distribution of mean landrace contribution to CK lines.

To analyse the usefulness of $MRD_{LI}$, we took into account the negative correlation between $MRD_{LI}$ and within-gene diversity of landraces (Hs), which could strongly bias against landraces with the lowest within diversity. For each landrace, we defined a 'normalized' MRD distance ($MRD_{norm}$) based on the absolute difference between (i) the median $MRD_{LI}$ between a landrace and lines of CK panel ($MRD_{med}$) and (ii) the $MRD_{LI}$ from the closest lines ($MRD_q$) defined by the 5th (MRD05) and 10th (MRD10) percentile of $MRD_{LI}$, corresponding to the 5% and 10% closest lines. In order to correct the bias due to Hs, we used the linear regression coefficient ‘a’ between $\mathrm{MRD_{med}}$ and Hs. We defined $\mathrm{MRD_{norm}}$ as the orthogonal deviation of $\mathrm{MRD_q}$ (with $q = 5\%$ or 10% for $\mathrm{MRD_{05}}$ and $\mathrm{MRD_{10}}$, respectively) from the linear regression:

$$\mathrm{MRD_{norm}} = \left(\mathrm{MRD_{med}} - \mathrm{MRD_q}\right) \times \sin\left(\tan^{-1}(a)\right) \qquad (1)$$

We used $\mathrm{MRD_{norm}}$ based on $\mathrm{MRD_{10}}$ to categorize the 33 landraces into three classes based on the percentile distribution of $\mathrm{MRD_{norm}}$. Landraces with $\mathrm{MRD_{norm}}$ below 30%, between 30% and 70% quantile and above 70% were considered to have none, few or many derived lines, respectively.

Finally, we performed a variance analysis to test the effect of mean contribution and $\mathrm{MRD_{norm}}$ on the number of new haplotypes discovered in the DH-SSD lines.

## Acknowledgements

This study was funded by the ‘Association pour l’étude et l’amélioration du maïs’ (PROmais) within the ‘DiversitéZea’ project and the French National Research Agencies within the ‘Investissement d’Avenir Amaizing’ project, (ANR-10-BTBR-01). We greatly acknowledge the French Maize Biological Resource Center, PROmais, and the INRAE experimental units of St Martin de Hinx and Mauguio for collecting and maintaining the collection of landraces and inbred lines. We greatly acknowledge colleagues who initially collected these landraces, André Gallais for initiating these research programmes, and Pierre Dubreuil, Letizia Camus-Kulandaivelu, Cecile Rebourg, Céline Mir, Domenica Maniccaci who conducted previous analysis using DNA pooling approach. We thank CEA-CNG (A. Boland, M-T Bihoreau) and EPGV group (D. Brunel, A. Bérard, A Chauveau) for discussion and management of Illumina genotyping.

## Conflict of interest statement

None declared.

## Author contributions

S.D.N, A.C and B.G designed and supervised the study and selected the plant material; M.A, S.D.N, A.C, B.G drafted and corrected the manuscript; D.M, V.C and M-C.L-P extracted DNA and managed the genotyping of landraces and inbred lines; C.B, B.G and A.C collected and maintained the collection of landraces and inbred lines; S.D.N, M.A, A.C and T.M-H developed the statistical methods for predicting allelic frequency from fluorescence data; M.A, B.G and S.D.N analysed the genetic diversity of the landrace panel; M.A and S.D.N analysed the selective sweep; MA, S.D.N and A.C investigated the relationship between landraces and inbred lines; S.D.N developed the normalized distance measure and performed the analysis of diversity enrichment.

## Data availability statement

Fluorescence Intensity Data from 166 DNA samples of landraces used for predicting allelic frequency and modified Roger’s distance matrix are available at https://doi.org/10.15454/D4JTKB. To predict allelic frequency in 166 DNA pools, we calibrated our two-step model with fluorescence intensity data of 327 inbred lines (for calibrating the fixation test) and two series of controlled pools (for calibrating logistic regression) with R script that are available at the following address: https://doi.org/10.15454/GANJ7J.

## References

Aguirre-Liguori, J.A., Tenaillon, M.I., Vázquez-Lobo, A., Gaut, B.S., Jaramillo-Correa, J.P., Montes-Hernandez, S. *et al*. (2017) Connecting genomic patterns of local adaptation and niche suitability in teosintes. *Mol. Ecol.* **26**, 4226–4240.

Albrechtsen, A., Nielsen, F.C. and Nielsen, R. (2010) Ascertainment Biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* **27**, 2534–2547.

Alexander, D.H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664.

Arca, M., Mary-Huard, T., Gouesnard, B., Bérard, A., Bauland, C., Combes, V. *et al*. (2021) Deciphering the genetic diversity of landraces with high-throughput SNP genotyping of DNA bulks: methodology and application to the maize 50k array. *Front. Plant Sci.* **11**, 568699.

Arteaga, M.C., Moreno-Letelier, A., Mastretta-Yanes, A., Vázquez-Lobo, A., Breña-Ochoa, A., Moreno-Estrada, A. *et al*. (2016) Genomic variation in recently collected maize landraces from Mexico. *Genomics Data*, **7**, 38–45.

Beadle, G.W. (1939) Teosinte and the origin of maize. *J. Hered.* **30**, 245–247.

Bouchet, S., Servin, B., Bertin, P., Madur, D., Combes, V., Dumas, F. *et al*. (2013) 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.

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 Genet.* **13**, e1006666.

Brandolini, A. (1970) *Maize*.

Brauner, P.C., Schipprack, W., Utz, H.F., Bauer, E., Mayer, M., Schön, C.-C. and Melchinger, A.E. (2019) Testcross performance of doubled haploid lines from European flint maize landraces is promising for broadening the genetic base of elite germplasm. *Theor. Appl. Genet.* **132**, 1897–1908.

Buckler, E.S., Gaut, B.S. and McMullen, M.D. (2006) Molecular and functional diversity of maize. *Curr. Opin. Plant Biol.* **9**, 172–176.

Camus-Kulandaivelu, L., Veyrieras, J.-B., Madur, D., Combes, V., Fourmann, M., Barraud, S. *et al*. (2006) Maize adaptation to temperate climate: relationship between population structure and polymorphism in the Dwarf8 gene. *Genetics*, **172**, 2449–2463.

Castelletti, S., Coupel-Ledru, A., Granato, I., Palaffre, C., Cabrera-Bosquet, L., Tonelli, C. *et al*. (2020) Maize adaptation across temperate climates was obtained via expression of two florigen genes. *PLOS Genet.* **16**, e1008882.

Cavanagh, C.R., Chao, S., Wang, S., Huang, B.E., Stephen, S., Kiani, S. *et al*. (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci. USA*, **110**, 8057–8062.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. and Nielsen, R. (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502.

Coffman, S.M., Hufford, M.B., Andorf, C.M. and Lübberstedt, T. (2020) Haplotype structure in commercial maize breeding programs in relation to key founder lines. *Theor. Appl. Genet.* **133**, 547–561.

Diaw, Y., Tollon-Cordet, C., Charcosset, A., Nicolas, S., Madur, D., Ronfort, J. *et al*. (2020) *Genetic diversity of maize landraces from the South-West of France. BioRxiv*.

Dubreuil, P. and Charcosset, A. (1998) Genetic diversity within and among maize populations: a comparison between isozyme and nuclear RFLP loci. *Theor. Appl. Genet.* **96**, 577–587.

Dubreuil, P. and Charcosset, A. (1999) Relationships among maize inbred lines and populations from European and North-American origins as estimated using RFLP markers. *Theor. Appl. Genet.* **99**, 473–480.

Dubreuil, P., Rebourg, C., Merlino, M. and Charcosset, A. (1999) Evaluation of a DNA pooled-sampling strategy for estimating the RFLP diversity of maize populations. *Plant Mol. Biol. Rep.* **17**, 123–138.

Dubreuil, P., Warburton, M., Chastanet, M., Hoisington, D. and Charcosset, A. (2006) More on the introduction of temperate maize into Europe: large-scale bulk SSR genotyping and new historical elements. *Maydica*, **51**, 281–291.

Eschholz, T.W., Stamp, P., Peter, R., Leipner, J. and Hund, A. (2010) Genetic structure and history of Swiss maize (*Zea mays* L. ssp. mays) landraces. *Genet. Resour. Crop Evol.* **57**, 71–84.

Evanno, G., Regnaut, S. and Goudet, J. (2005) Detecting the number of clusters of individuals using the software Structure: a simulation study. *Mol. Ecol.* **14**, 2611–2620.

Fernie, A.R., Tadmor, Y. and Zamir, D. (2006) Natural genetic variation for improving crop quality. *Curr. Opin. Plant Biol.* **9**, 196–202.

Foll, M. and Gaggiotti, O.E. (2005) Colonise: a computer program to study colonization processes in metapopulations. *Mol. Ecol. Notes*, **5**, 705–707.

Frascaroli, E., Schrag, T.A. and Melchinger, A.E. (2013) Genetic diversity analysis of elite European maize (*Zea mays* L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. *Theor. Appl. Genet.* **126**, 133–141.

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.

Gates, D.J., Runcie, D., Janzen, G.M., Navarro, A.R., Willcox, M., Sonder, K. *et al*. (2019) Single-gene resolution of locally adaptive genetic variation in Mexican maize. *Evolution. Biol.* https://doi.org/10.1101/706739

Gauthier, P., Gouesnard, B., Dallard, J., Redaelli, R., Rebourg, C., Charcosset, A. and Boyat, A. (2002) RFLP diversity and relationships among traditional European maize populations. *Theor. Appl. Genet.* **105**, 91–99.

Gerdes, J.T. and Tracy, W.F. (1993) Pedigree diversity within the lancaster surecrop heterotic group of maize. *Crop Sci.* **33**, 334–337.

Gore, M.A., Chia, J.M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L. *et al*. (2009) A first-generation haplotype map of maize. *Sci. Wash.* **326**, 1115–1117.

Gouesnard, B., Negro, S., Laffray, A., Glaubitz, J., Melchinger, A., Revilla, P. *et al*. (2017) Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank. *Theor. Appl. Genet.* **130**, 2165–2189.

Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Gu, Y., Zhao, C., He, L., Yan, B., Dong, J., Li, Z. *et al*. (2017) Genome-wide identification and abiotic stress responses of DGK gene family in maize. *J. Plant Biochem. Biotechnol.* **27**, 156–166.

Hamblin, M.T., Warburton, M.L. and Buckler, E.S. (2007) Empirical Comparison of Simple Sequence Repeats and Single Nucleotide Polymorphisms in Assessment of Maize Diversity and Relatedness. *PLoS ONE*, **2**, e1367.

van Heerwaarden, J., Doebley, J., Briggs, W.H., Glaubitz, J.C., Goodman, M.M., de Jesus Sanchez Gonzalez, J. and Ross-Ibarra, J. (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl. Acad. Sci. USA*, **108**, 1088–1092.

Hijmans, R.J. (2019) *geosphere: Spherical Trigonomery. R package version 1.5-10*.

Hoisington, D., Khairallah, M., Reeves, T., Ribaut, J.-M., Skovmand, B., Taba, S. and Warburton, M. (1999) Plant genetic resources: What can they contribute toward increased crop productivity? *Proc. Natl. Acad. Sci. USA*, **96**, 5937–5943.

Hölker, A.C., Mayer, M., Presterl, T., Bolduan, T., Bauer, E., Ordas, B. *et al*. (2019) European maize landraces made accessible for plant breeding and genome-based studies. *Theor. Appl. Genet.* **132**, 3333–3345.

Huang, C., Sun, H., Xu, D., Chen, Q., Liang, Y., Wang, X. *et al*. (2017) ZmCCT9 enhances maize adaptation to higher latitudes. *Proc. Natl. Acad. Sci. USA*, **115**(2), E334–E341.

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. *Nat. Genet.* **44**, 808–811.

Hyten, D.L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R.L., Costa, J.M. *et al*. (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. USA*, **103**, 16666–16671.

Jaenicke-Despres, V., Buckler, E.S., Smith, B.D., Gilbert, M.T.P., Cooper, A., Doebley, J. and Pääbo, S. (2003) Early allelic selection in maize as revealed by ancient DNA. *Science*, **302**, 1206–1208.

Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J. *et al*. (2012) Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815.

Kilian, B. and Graner, A. (2012) NGS technologies for analyzing germplasm diversity in genebanks. *Brief. Funct. Genomics*, **11**, 38–50.

Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J.C. and Stein, N. (2019) Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* **51**, 1076–1081.

Matsuoka, Y., Vigouroux, Y., Goodman, M.M., Sanchez, J., Buckler, E. and Doebley, J. (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. USA*, **99**, 6080–6084.

McCouch, S.R., Zhao, K., Wright, M., Tung, C.-W., Ebana, K., Thomson, M. *et al*. (2010) Development of genome-wide SNP assays for rice. *Breed. Sci.* **60**, 524–535.

McCouch, S.R., McNally, K.L., Wang, W. and Hamilton, R.S. (2012) Genomics of gene banks: a case study in rice. *Am. J. Bot.* **99**, 407–423.

Mikel, M.A. (2011) Genetic composition of contemporary U.S. commercial dent corn germplasm. *Crop Sci.* **51**, 592–599.

Millet, E., Welcker, C., Kruijer, W., Negro, S., Nicolas, S., Praud, S. *et al*. (2016) Genome-wide analysis of yield in Europe: allelic effects as functions of drought and heat scenarios. *Plant Physiol.* **172**, 749–764.

Mir, C., Zerjal, T., Combes, V., Dumas, F., Madur, D., Bedoya, C. *et al*. (2013) Out of America: tracing the genetic footprints of the global diffusion of maize. *Theor. Appl. Genet.* **126**, 2671–2682.

Moragues, M., Comadran, J., Waugh, R., Milne, I., Flavell, A.J. and Russell, J.R. (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theor. Appl. Genet.* **120**, 1525–1534.

Negro, S.S., Millet, E.J., Madur, D., Bauland, C., Combes, V., Welcker, C. *et al*. (2019) Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biol.* **19**, 318.

Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA*, **70**, 3321–3323.

Nei, M. (1977) F-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.* **41**, 225–233.

Nei, M. and Chesser, R.K. (1983) Estimation of fixation indices and gene diversities. *Annals of Human Genetics*, **47**, 253–259. https://doi.org/10.1111/j.1469-1809.1983.tb00993.x.

Paradis, E. and Schliep, K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.

R Core Team. (2013) *R: A language and environment for statistical computing*. Vienna: Foundation for Statistical Computing. https://www.R-project.org/

Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **5**, 94–100.

Rebourg, C., Dubreuil, P. and Charcosset, A. (1999) Genetic diversity among maize populations: bulk RFLP analysis of 65 accessions. *Maydica*, **44**, 237–249.

Rebourg, C., Gouesnard, B. and Charcosset, A. (2001) Large scale molecular analysis of traditional European maize populations. Relationships with morphological variation. *Heredity*, **86**, 574–587.

Rebourg, C., Chastanet, M., Gouesnard, B., Welcker, C., Dubreuil, P. and Charcosset, A. (2003) Maize introduction into Europe: the history reviewed in the light of molecular data. *Theor. Appl. Genet.* **106**, 895–903.

Reif, J.C., Zhang, P., Dreisigacker, S., Warburton, M.L., van Ginkel, M., Hoisington, D. *et al*. (2005) Wheat genetic diversity trends during domestication and breeding. *Theor. Appl. Genet.* **110**, 859–864.

Reif, J.C., Warburton, M.L., Xia, X.C., Hoisington, D.A., Crossa, J., Taba, S. *et al*. (2006) Grouping of accessions of Mexican races of maize revisited with SSR markers. *Theor. Appl. Genet.* **113**, 177–185.

Revilla, P. and Tracy, W.F. (1995) Isozyme variation and phylogenetic relationhips among open-pollinated sweet corn cultivars. *Crop Sci.* **35**, 219–227.

Revilla, P., Soengas, P., Malvar, R.A., Cartea, M.E. and Ordás, A. (1998) Isozyme variation and historical relationships among the maize races of Spain. *Maydica*, **43**, 175–182.

Reyes-Valdés, M.H., Santacruz-Varela, A., Mart́ınez, O., Simpson, J., Hayano-Kanashiro, C. and Cortés-Romero, C. (2013) Analysis and optimization of bulk DNA sampling with binary scoring for germplasm characterization. *PLoS ONE*, **8**, e79936.

Rogers, J.S. (1972) Measures of genetic similarity and genetic distance. *Stud. Genet.* **7**, 145–153.

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 Biol.* **14**, R55.

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. *Nat. Genet.* **49**, 476–480.

Schlötterer, C., Tobler, R., Kofler, R. and Nolte, V. (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**, 749–763.

Sham, P., Bader, J.S., Craig, I., O'Donovan, M. and Owen, M. (2002) DNA Pooling: a tool for large-scale association studies. *Nat. Rev. Genet.* **3**, 862–871.

Smouse, P.E., Long, J.C. and Sokal, R.R. (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool.* **35**, 627–632.

Strigens, A., Schipprack, W., Reif, J.C. and Melchinger, A.E. (2013) Unlocking the genetic diversity of maize landraces with doubled haploids opens new avenues for breeding. *PLoS ONE*, **8**, e57234.

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.

Tanksley, S.D. (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science*, **277**, 1063–1066.

Tenaillon, M.I. and Charcosset, A. (2011) A European perspective on maize history. *C. R. Biol.* **334**, 221–228.

Tracy, W.F., Whitt, S.R. and Buckler, E.S. (2006) Recurrent mutation and genome evolution: example of and the origin of sweet maize. *Crop Sci.* **46**, S-49.

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**, 823.

Vigouroux, Y., Mitchell, S., Matsuoka, Y., Hamblin, M., Kresovich, S., Smith, J.S.C. *et al*. (2005) An analysis of genetic diversity across the maize genome using microsatellites. *Genetics*, **169**, 1617–1630.

Virlouvet, L., Jacquemot, M.-P., Gerentes, D., Corti, H., Bouton, S., Gilard, F. *et al*. (2011) The ZmASR1 protein influences branched-chain amino acid biosynthesis and maintains kernel yield in maize under water-limited conditions. *Plant Physiol.* **157**, 917–936.

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. *Nat. Genet.* **52**, 565–571.

Wang, L., Josephs, E.B., Lee, K.M., Roberts, L.M., Rellán-Álvarez, R., Ross-Ibarra, J. and Hufford, M.B. (2021) Molecular parallelism underlies convergent highland adaptation of maize landraces. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msab119

Whitt, S., Wilson, L., Tenaillon, M., Gaut, B. and Buckler, E. (2002) Genetic diversity and selection in the maize starch pathway. *Proc. Natl. Acad. Sci. USA*, **99**, 12959.

Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D. and Gaut, B.S. (2005) The effects of artificial selection of the maize genome. *Science*, **308**, 1310–1314.

Yamasaki, M., Tenaillon, M.I., Bi, I.V., Schroeder, S.G., Sanchez-Villeda, H., Doebley, J.F. *et al*. (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell*, **17**, 2859–2872.

Yamasaki, M., Wright, S.I. and McMullen, M.D. (2007) Genomic screening for artificial selection during domestication and improvement in maize. *Ann. Bot.* **100**, 967–973.

Yao, Q., Yang, K., Pan, G. and Rong, T. (2007) Genetic diversity of maize (*Zea mays* L.) landraces from Southwest China based on SSR data. *J. Genet. Genomics*, **34**, 851–860.

Yilmaz, A., Nishiyama, M.Y., Fuentes, B.G., Souza, G.M., Janies, D., Gray, J. and Grotewold, E. (2009) GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol.* **149**, 171–180.

Zeitler, L., Ross-Ibarra, J. and Stetter, M.G. (2020) Selective loss of diversity in doubled-haploid lines from European Maize landrace. *G3 Genes* **10**, 2497–2506. https://doi.org/10.1534/g3.120.401196

## Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Figure S1** Distribution of the minimum allele frequency (MAF) across the entire landrace panel (Whole) and within the five geographic groups.
**Figure S2** Geographic distribution of the 156 landrace accessions. Landraces are coloured according to their geographic origin.
**Figure S3** Distribution of rare alleles in the five geographic groups of maize landraces.
**Figure S4** Geographic distribution of landrace clusters obtained by hierarchical clustering based on modified Roger's distance and Ward's method.
**Figure S5** Dendrogram of the 166 DNA samples corresponding to 156 landrace accessions based on Ward (A) and Neighbour-Joining (B) hierarchical clustering.
**Figure S6** Relationship between modified Roger's genetic distances and geographic distances between landraces within the five geographic groups.
**Figure S7** Determination of the *K* value by ADMIXTURE analysis performed across the landrace panel for 2500 Panzea markers.
**Figure S8** Genomic scan performed by BAYESCAN to identify outlier loci within the landrace panel
**Figure S9** Variation in the level of genetic differentiation between pairs of geographic groups of landraces along the maize genome.
**Figure S10** Distribution of outlier loci in the five geographic groups of maize landraces.
**Figure S11** Relationship between the mean frequency of allele B (A) and expected heterozygosity (HT, B) in panels of 166 landraces and 327 inbred lines for 23 412 SNPs.
**Figure S12** Relationship between landrace genetic diversity (Hs) and their modified Roger's distance (MRD), and the inbred lines from the panel of CK lines.
**Figure S13** Number of assigned lines, average contribution and its change across breeding cycles of 21 landraces with the highest average contribution to the 327 CK lines from the diversity panel.

**Table S1** Description of the 156 landraces: geographic origin, genetic diversity, genetic group, mean contribution and number of assigned lines within the panel of CK lines.
**Table S2** Pairwise genetic differentiation (Gst) and Modified Roger's Distance (MRD) among the five geographic groups estimated with 23 412 SNPs and 17 SSRs.
**Table S3** Number of landraces assigned to each cluster in the two Bayesian approaches for $K = 2$ to $K = 7$ with SSR and SNP markers.
**Table S4** Mean and quantile values of the genetic diversity parameters (Hs, Gst, Ht) between landraces, geographic groups, and inbred lines.
**Table S5** Highly differentiated genomic regions between landraces and the five geographic groups identified by outlier Gst analysis and BAYESCAN analysis.
**Table S6** Number of SNPs under selection between geographic groups identified by BAYESCAN analysis according to their genome position.

**Table S7** List of 379 loci under selection between the five geographic groups identified by BAYESCAN.

**Table S8** List of 505 loci identified under selection between the 10 pairs of geographical groups identified by BAYESCAN.

**Table S9** Number of SNPs under selection between landraces and inbred lines identified by BAYESCAN analysis according to their genome position.

**Table S10** List of 49 loci under selection between landraces and inbred lines identified by BAYESCAN.

**Table S11** Quantitative assignment of 442 inbred lines to 166 landraces using supervised analysis implemented in the Admixture software.

**Appendix S1** Selection and sampling of landraces

**Appendix S2** Filtering SNPs according to their suitability for diversity analysis and their quality for predicting allelic frequency.

**Appendix S3** Estimation of 'within group' and 'across group' diversity parameters and Gst.

**Appendix S4** Classification of SNP under selection by BAYESCAN.

**Appendix S5** Simulation of individuals for structure analysis of landraces.