Genome-wide SNP genotyping of DNA pools identifies untapped landraces and genomic regions that could enrich the maize breeding pool

Mariangela Arca, Brigitte Gouesnard, Tristan Mary-Huard, Marie-Christine Le Paslier, Cyril Bauland, Valérie Combes, Delphine Madur, Alain Charcosset, Stephane Nicolas

▶ **To cite this version:**

Mariangela Arca, Brigitte Gouesnard, Tristan Mary-Huard, Marie-Christine Le Paslier, Cyril Bauland, et al.. Genome-wide SNP genotyping of DNA pools identifies untapped landraces and genomic regions that could enrich the maize breeding pool. 2020. hal-02965049

HAL Id: hal-02965049
https://hal.inrae.fr/hal-02965049v1

Preprint submitted on 12 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Genome-wide SNP genotyping of DNA pools identifies untapped landraces and genomic regions that could enrich the maize breeding pool

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], Stéphane D. Nicolas[1]

Author's affiliations:

1 Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE - Le Moulon, 91190, Gif-sur-Yvette, France

2 Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34090 Montpellier, France

3 Université Paris-Saclay, INRAE, Etude du Polymorphisme des Génomes Végétaux, 91000, Evry, France

Corresponding author: stephane.nicolas@inrae.fr

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

ABSTRACT

Maize landraces preserved in genebanks have a large genetic diversity that is still poorly characterized and underexploited in modern breeding programs. Here, we genotyped DNA pools from 156 American and European landraces with a 50K SNP Illumina array to study the effect of both human selection and environmental adaptation on the genome-wide diversity of maize landraces. Genomic diversity of landraces varied strongly in different parts of the genome and with geographic origin. We detected selective footprints between landraces of different geographic origin in genes involved in the starch pathway *(Su1, Waxy1),* flowering time *(Zcn8, Vgt3, ZmCCT9)* and tolerance to abiotic and biotic stress (*ZmASR, NAC* and *dkg* genes). Landrace diversity was compared to that of (i) 327 inbred lines representing American and European diversity ("CK lines) and (ii) 103 new lines derived directly from landraces ("DH-SSD lines"). We observed limited diversity loss or selective sweep between landraces and CK lines, except in peri-centromeric regions. However, analysis of modified Roger's distance between landraces and the CK lines showed that most landraces were not closely related to CK lines. Assignment of CK lines to landraces using supervised analysis showed that only a few landraces, such as Reid's Yellow Dent, Lancaster Surecrop and Lacaune, strongly contributed to modern European and American breeding pools. Haplotype diversity of CK lines was more enriched by DH-SSD lines that derived from the landraces with no related lines and the lowest contribution to CK lines. Our approach opens an avenue for the identification of promising landraces for pre-breeding.

Keywords: *Zea mays*, gene banks, Landraces, Pre-breeding, DNA pooling, Genetic diversity, Selective footprints, Allelotyping

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

SIGNIFICANCE STATEMENTS

Maize landraces are a valuable 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. However, they are underutilized in modern breeding programs because they are poorly characterized, genetically heterogeneous and exhibit poor agronomic performance compared to elite hybrid material. In this study, we developed a high-throughput approach to identify landraces that could potentially enlarge the genetic diversity of modern breeding pools. We genotyped DNA pools from landraces using 50K array technology, which is widely used by breeders to characterize the genetic diversity of inbred lines. To identify landraces that could enrich the modern maize germplasm, we estimated their contribution to inbred lines using supervised analysis and a new measurement of genetic distance.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

INTRODUCTION

Plant genetic resources are the basic raw material for future genetic progress (1–4). Maize landraces are an interesting 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 (4–7). During the early twentieth century, landraces were used as parent material for the development of improved hybrid varieties to meet the needs of modern agriculture. During this transition from landraces to hybrids, many favorable alleles were probably lost as a result of their association with unfavorable alleles and/or genetic drift (8–11). Nowadays, modern breeding programs 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 (12–15). Landraces that did not contribute to this founding material may be expected to be useful for enriching modern maize diversity, particularly for traits that enhance adaptation to adverse environmental conditions (7). However, maize landraces are used to a very limited extent, if at all, in modern plant breeding programs because they are poorly characterized, genetically heterogeneous and exhibit poor agronomic performance compared to elite hybrid material (3, 6, 16–18). Therefore, understanding the genetic diversity of maize landraces and their relation to the maize elite pool is essential for better management of genetic resources and for genetic improvement through genome-wide association studies, genomic selection and the dissection of quantitative traits (2, 6, 7).

Maize was domesticated in the highlands of Central Mexico approximately 9,000 years ago (19, 20). It then diffused to South and North America (21, 22) and spread rapidly out from America (23). It is now cultivated in highly diverse climate zones ranging from 40°S to 50°N. In Europe, the presently accepted hypothesis is that maize was first introduced through Spain by Columbus, although other sources of maize that were pre-adapted to temperate climates have been important for adapting to northern European conditions (22, 24–29). 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 behavior, yield, nutritive value and resistance to biotic and abiotic stress, resulting in subsequent differentiation of the material (7, 27, 30).

In recent years, the genetic diversity of maize landraces, which are conserved *ex situ*, has been studied extensively using various types of molecular markers such as restriction fragment length polymorphisms (RFLPs) (8, 25–28, 31–34) and simple sequence repeats

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

(SSRs) (8, 23, 35, 36). Single nucleotide polymorphisms (SNPs) are now the marker of choice for various crop species such as maize (37), rice (38) and barley (39). 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 (40). 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 (37). It has been successfully used to analyze the diversity of inbred lines and landraces by genotyping a low number of plants per accession (13, 16, 41–44).

However, due to high within-accession diversity, the characterization of maize landraces should be carried out on a representative set of individuals (45). Despite recent technical advances, genotyping large numbers of individuals remains very expensive in the context of genetic characterization. As a result, DNA pooling (or allelotyping) has been actively developed as a valuable alternative strategy for collecting information on allele frequency from a group of individuals while significantly reducing the effort required for population studies using DNA markers (46, 47). In maize, DNA pooling has been successfully used to decipher the global genetic diversity of landraces using RFLP (32) and SSR markers (23, 27, 28, 48, 49). The recent development of SNP arrays in maize (37, 50), combined with DNA pooling, could 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 50K array within DNA pools based on the fluorescence intensity of the two alleles at each SNP (51). This new method accurately predicts allelic frequency, safeguards against the false detection of alleles and leads to little ascertainment bias for deciphering global genetic diversity (51).

In the present study, we applied this new 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; 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” (27) 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

RESULTS

Genetic diversity within maize landraces

Only 25 SNPs out of 23,412 were monomorphic in the landrace panel. The average total diversity (HT) was 0.338 ± 0.001. The distribution of minor allelic frequency (MAF) showed a deficit in rare alleles (MAF<0.05) compared to other frequency classes (Fig. S1).

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, Fig. S2, Table S1). All five geographic groups displayed both alleles for nearly all loci, with the exception of CAR which was monomorphic at 1,227 loci out of 23,387 (Fig. S3). The lowest and highest within-group HT was found in CAR (0.301) and CAM (0.328), respectively. Note that there was an excess of rare alleles in EUR, CAR and NAM but not in SAM and CAM (Fig. 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 (Hs) was on average 0.192 ± 0.001, (Table 1) and varied between 0.03 (Ger8 and Ger9) and 0.28 (Sp11) (Table S1). The CAM group displayed on average the highest diversity (0.219 ± 0.008), while the EUR group displayed the lowest (0.177 ± 0.002).

Genetic differentiation between landraces (FST) was 0.428 on average. FST within a geographic group varied between 0.314 (CAR) and 0.434 (EUR) (Table 1). Overall genetic differentiation between geographic groups was low (FST=0.05). FST between pairs of geographic groups varied between 0.016 (EUR and NAM) and 0.083 (NAM and CAR) (Table S2).

Relationship between maize landraces and population structure

The average modified Roger's distance (MRD) between landraces was 0.379. The lowest MRD between landraces was 0.158 (Chi12 and Chi9). It is slightly higher than the distance between two pools of independent individuals from a same population (0.092-0.120, (51)). The highest MRD was 0.552 (Ant1 and Ger8). The average MRD between populations from a same geographic group ranged from 0.320 (CAR) to 0.367 (EUR) (Table 1). The average MRD between populations belonging to two different geographic groups varied between 0.354 (CAM *vs* CAR) and 0.420 (NAM *vs* CAR) (Table S2).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

We investigated the relationship between maize landraces using Principal Coordinate Analysis (PcoA) and Ward hierarchical clustering based on MRD (Fig. 1). For both PcoA and Ward hierarchical clustering analysis, landraces mostly clustered according to their geographic proximity (Fig. 1, Fig. S4, Fig. 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) (Fig. 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, Fig. 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 Fig. 1B, Fig. S4). 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. Three clusters grouped together American and European landraces (cluster c, d and g on Fig. S4). 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 Fig. S4). Cluster “g” comprised 12 North American flint landraces from higher latitudes (>40°N) and 18 northeastern European landraces mainly from Germany (g on Fig. 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 ($r^2 = 0.05$, $P < 0.001$, Fig. S6A), NAM ($r^2 = 0.12$, $P < 0.001$, Fig. S6B) and CAM ($r2 = 0.0858$, $P = 0.02$, Fig. S6C).

We analyzed the genetic structure of 156 landraces using the ADMIXTURE program. Likelihood analysis indicated that the optimal number of genetic groupe was K=2, K=3 and K=7 (Fig. S7). We considered K=7 as the reference, as this value was consistent with the one obtained with 24 SSRs by Camus- Kulandaivelu et *al.* (27). Landraces from different geographic regions were assigned to different genetic groups, with a clear trend along latitude

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

and longitude. Fig. 2). Assignment to these groups was also highly consistent with PcoA and hierarchical clustering (Fig. 1, Fig. 2, Fig. S4, Fig. S5). The genetic structure obtained with SNP markers was highly consistent with that obtained with the 17 SSR markers; indeed, 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 at K=7 were that the Northern Flint landrace group obtained with SNPs is split in two with SSRs and the separate Pyrenean-Galician and Italian groups found with SNPs form a single group with SSRs.

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 ($\overline{HT_l}$ <0.069) across the entire landrace panel (Fig. 3A, Table S4). These regions were mainly located in the centromeric region of chromosomes 5 and 7. 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) (Fig. 3E to3I, Table S4). These regions were mostly located close to the centromeres but varied between geographic groups. In the centromeric region of chromosome 1, we observed (i) no loss of diversity for CAR and NAM and (ii) a depletion in genetic diversity for CAM, EUR and SAM. Conversely, we observed a strong depletion on chromosomes 3 and 4 in CAR landraces that was not observed in other geographic groups.

Outlier analysis of FST values among individual landraces identified 20 and 17 genomic regions displaying high differentiation ($\overline{FST_l}$>0.568) and low differentiation ($\overline{FST_l}$<0.235) between landraces, respectively (Fig. 3L, Table S4). Genetic differentiation was highest upstream of chromosome 6 (Sp10 in Table S5), in two regions upstream of chromosome 4 (Sp6 and Sp7in Table 2) and in one region on chromosome 3 (Sp3 in Table 2).

Outlier FST analysis between geographic groups identified 26 regions with high differentiation ($\overline{FST_g}$>0.150) and 8 regions with low differentiation ($\overline{FST_g}$<0.007) (Fig. 3J, Table S4); BAYESCAN identified 379 loci under divergent selection (Fig. 3J, Table S6 and S7, Fig. S8). The five genomic regions that were previously identified as being highly differentiated between landraces by outlier FST analysis were also detected by both FST outlier and BAYESCAN analyses between geographic groups. (Table 2). Only one highly differentiated genomic region was identified between landraces but not between all five geographic groups (Sp10 in Table S5) whereas 6 genomic regions were identified exclusively between the five geographic groups (Sg6, 7, 13, 15, 17, 18 and 20 in Table S5). These regions

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

displayed contrasted allelic patterns across geographic groups. Sp10 (11.7 Mbp – 15.3 Mbp on chromosome 6, $\overline{FST_g}$ = 0.08 and $\overline{FST_l}$ =0.65) had 9 SNPs that were close to fixation in CAM (HT<0.1), but were segregating in NAM (~0.4) and also to a lesser extent in EUR, CAR and SAM (HT~0.2). Sg2-Sp3 (84-85 Mbp on chromosome 3, $\overline{\overline{FST_g}}$ = 0.18 and $\overline{\overline{FST_l}}$ = 0.63) had 3 SNPs showing a continuous allelic frequency gradient between tropical and temperate landraces with one allele largely predominant in NAM and EU (~70%), minor in CAM (~30%) and absent in CAR (~0%). Sg4-Sp6 (40.3-41.8Mbp on chromosome 4, $\overline{\overline{FST_g}}$ = 0.27 and $\overline{\overline{FST_l}}$ = 0.63) had 4 SNPs that were nearly fixed in temperate landraces (NAM, EUR) and displaying intermediate frequencies in CAM. By contrast, the Sg5-Sp7 region ($\overline{\overline{FST_g}}$=0.16, $\overline{\overline{FST_l}}$ =0.63) displayed higher diversity in temperate (HT_{NAM} and HT_{EUR}~0.4) than in tropical landraces (HT_{CAM}~0.2 and HT_{CAR}~0.05) (Fig. 3 D, E, F, G, H). The outlier loci displaying the highest FST values within this region were located up to 10 kbp upstream of the *Su1* gene which is involved in the starch pathway.

Outlier FST analysis between pairs of geographic groups identified 214 and 41 regions displaying high and low differentiation, respectively (Fig. S9). BAYESCAN analysis identified 363 SNPs under selection between pair 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 (Fig. S9, Fig. S10). Putative functions could be assigned to 272 of the 536 (50.7%) outlier loci identified by BAYESCAN analysis of all five and pairs of geographic groups. These included known genes involved in adaptation to abiotic stress, flowering time or human uses (Table S8 and 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 (0.353 *vs* 0.338). Allelic frequency of loci and HT values in inbred lines and landraces were strongly correlated (r^2=0.89 and r2=0.80, respectively, Fig. S11). Overall genetic differentiation between landraces and inbred lines was limited (0.010 ± 0.066). Some regions were more diverse in landraces than in inbred lines, notably the peri-centromeric region of chromosomes 3 and 7, while the opposite was found in centromeric regions of chromosomes 1, 3, 4, 5 and 6 (Fig. 3B).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Comparison of landraces and inbred lines using the outlier FST approach identified 128 highly differentiated genomic regions (FST> 0.04) and 32 regions with an excess of similarity (FST<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 (Fig. 3K). BAYESCAN analysis of landraces *vs* inbred lines identified 61 loci (0.3%) that were significantly more differentiated than expected under the drift model (Fig. 3K, Table S10).

Relationship between inbred lines and landrace populations: genetic distances and supervised analysis

The average MRD 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 (MRD_{LI}) is displayed as a series of boxplots (Fig. 4A) listed in ascending order of landrace expected heterozygosity (Hs) (Fig. 4B). Landraces with a low genetic diversity generally showed a higher median and a wider range for MRD_{LI}, with some notable exceptions (*e.g.* Chi5, Per10, Par2, Par1, Bra4, Ecu17, Vir4 and Svt1 in Fig. 4). Accordingly, the median MRD_{LI} and the within landrace genetic diversity Hs were strongly negatively correlated (r = -0.978, t= -61.314, p-value < 2.2e-16) and displayed a linear relationship (Fig. S12). Considering a similar level of genetic diversity, some landraces were closely related to certain inbred lines, whereas other landraces were not (Fig. 4A and Fig. 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 quantitatively assigned 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. Among these landraces, 47 had at least one inbred line assigned with a probability >60%. 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 diversity panel (77.6% vs 58.3%, p-value=0.04). For the 208 inbred lines from more advanced breeding cycles, we identified a total of 66 landraces as the main assignment of at least one inbred line. Among these, temperate inbred lines were frequently assigned to Reid's

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

277 Yellow Dent and Lancaster Surecrop. Chandelle (one of the few tropical landraces in our
278 study) was identified as the most likely source for many tropical lines.

279 A few landraces contributed strongly to the whole diversity panel, with the 10 first
280 landraces cumulating half of the total contributions (Fig. 4C, Fig. S13A). 80% of lines were
281 assigned to these 10 landraces with a > 60% probability (Fig. S13B). Interestingly, the mean
282 contribution of landraces differed strongly between first cycle lines and more advanced lines
283 with a strong decrease (>1%) for 15 landraces and a strong increase (>1%) for 8 landraces
284 (Fig. S13C).

285 We tested whether the mean contribution of landraces and the MRD_{LI} distance
286 "normalized" by within landraces genetic diversity could be used as a criterion to identify
287 untapped sources of genetic diversity that could enrich the CK line panel. First, we selected
288 66 DH-SSD lines that were correctly assigned to 33 landraces from the landrace panel. We
289 then classified these 33 landraces according to: (i) their average contribution to CK lines (Fig.
290 5A) and (ii) the normalized MRD distance from their closest lines (Fig. 5C). For each class,
291 we estimated with 979 haplotype markers the average number of new haplotypes discovered
292 in the 66 DH-SSD lines compared to those existing in the CK lines. We discovered 66 new
293 haplotypes in the DH-SSD lines compared to 4,355 different haplotypes in the CK lines. The
294 number of new haplotypes discovered in DH-SSD lines ranged from 0 (Bul3) to 11 (Arg8).
295 The average number of new haplotypes was significantly higher for lines derived from
296 landraces with a low contribution than those with a high contribution (p-value = 0.008, Fig.
297 5B). It was also higher for landraces that were not close to any of the CK lines than for those
298 that were close to certain lines (p-value = 0.0004, Fig. 5D).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

DISCUSSION

Patterns of genetic diversity and population structure within landraces

The total expected heterozygosity observed in our study based on SNPs (0.338) was lower than the values reported previously for landraces of comparable origin that were analyzed with SSR markers (0.58 in (26), 0.63 in (27) , 0.62 in (28)) but comparable to those observed with SNPs in diversity inbred line panels (42, 43). 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 (43, 52). Trends in the partition of genetic diversity within and between landraces, and within and between geographic groups were similar to previous findings. The diversity of individual landraces represented on average 57% of the total genetic diversity, which was slightly lower than for RFLP markers (~66% in (23, 26)). This difference may be due to the counter-selection of SNP markers with low MAF during the design of 50K Illumina array (37), which may increase total diversity more than within diversity (53, 54). On the other hand, genetic structure analyses based on SNPs and 17 SSRs were highly congruent, which indicates that the ascertainment bias of prefixed PZE SNPs from the 50K Illumina chip used to study landraces is negligible (43, 51).

Each geographic group contained most of the overall landrace genetic diversity, ranging from 89% (CAR) to 97% (CAM). Central American and Mexican landraces displayed the highest diversity, which is consistent with their proximity to the center of maize domestication (13, 20). This confirms that genetic diversity was lost during the spread of maize away from its domestication center due to successive bottlenecks related to climatic adaptation and isolation by distance (7, 21, 22, 29, 55). 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 (21, 23, 29, 35). Our results nevertheless confirm that the bottleneck during the introduction of maize in Europe was certainly limited, as also shown by Brandebourg et *al.*, (29) with whole genome sequencing of 67 inbred lines from Europe and America. 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. The fact that some of these landraces

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

have been cultivated mostly in gardens may have decreased their effective population size (26). The genetic load could have been more or less purged depending on the severity and the duration of the bottleneck. This could explain the strong variation in success rate observed for deriving inbred lines from European Flint landraces by haplodiploidization (56, 57).

Genetic distance, Ward hierarchical clustering (Fig. 1B), principal component (Fig. 1A) and population structure (Fig. 2) analyses showed major trends in population structure. We confirmed the central position of Mexican and Caribbean landraces and a clear differentiation between North and South American landraces (Fig. 1 and 2). This is consistent with the domestication of maize in Mexico followed by southwards and northwards dispersion (22, 55). 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 (Fig. 1B, cluster d). Strong similarities between groups of northeastern American and northeastern European landraces (mostly from Germany, Poland and Austria) (Fig. 1B, cluster g) also supports an independent introduction of North American material that was pre-adapted to the northern European climate (21, 22, 26, 28, 29, 58). 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. (27, 29, 59). Interestingly, some southwestern Spanish landraces have elevated admixture with Italian Flint groups and are closely related to Italian landraces on the NJ tree (Fig. 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 (29, 60). 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.

Differentiation of landraces was greater in Europe than in Central America and the Caribbean, indicating that gene flow is lower in the latter two. Genetic and geographic distances were significantly correlated in NAM, EUR and CAM but not in SAM and CAR (Fig. S6), suggesting that isolation by distance played a role in shaping the genetic structure of

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

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 producing major gradients in environmental conditions (temperature, rainfall) (7, 30, 61). Indeed, Mexican landraces clustered according to both altitude and distance (Fig. 1B, Table S1) suggesting environmental adaptation (7, 30).

Genomic pattern of nucleotide variation in landraces

FST outlier and BAYESCAN analyses identified 13 genomic regions that showed high levels of differentiation between geographic groups and/or landraces (Table S5). The four highly differentiated genomic regions between landraces displayed contrasted patterns of allelic frequencies between geographic groups (Table 2, Table S5), suggesting different types of selection. The Sp10 region was found to be highly differentiated between landraces but not between the five geographical groups. It suggests that there was strong selection in some specific geographic areas but not across all geographic groups. This region contains genes associated with tolerance to high temperature and evaporative demand (62). The second genomic region (Sg4-Sp6: 7.8 Mbp – 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. This results is in agreement with Romero-Navaro et al. (55), who identified 5 SNPs in this region with allelic frequencies varying significantly with latitude in American landraces, and Brandeburg et al., (29), who identified two highly differentiated regions between Corn Belt Dent and Tropical first cycle lines. By contrast, the third genomic region (Sp5-Sg7; 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 included the *Su1* gene, which is involved in the starch pathway and is known to be under strong selective pressure (63–66). Romero-Navaro et al., (55) also found an association between allelic frequency variation at the *Su1* locus and both latitude and longitude. Futhermore, Brandeburg et al., (29) identified a strong selective sweep between Corn Belt Dent/Tropical and Northern Flint first cycle lines in the *Su1* gene. The fourth region (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. In agreement with this finding, Romero-Navaro et al., (55) identified in this region 22 and 4 SNPs with allelic

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

frequencies varying significantly with altitude and latitude, respectively. This region also carries a large 6 Mbp inversion that is putatively involved in flowering time variation (55).

BAYESCAN analysis between geographic groups identified several regions that were not identified by outlier FST analysis (Table S8 and S9). 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* (42, 67, 68); (ii) PZE-109070904 on chromosome 9 (115.7 Mbp) in *ZmCCT9* (69); (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 (62, 70). 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 (71); (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 (72); (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 (73). 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 have 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 (7, 22, 23, 29, 55, 74). 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 (7, 41, 74), or creating new genetic diversity by targeted mutation (7).

Identification of promising landraces to enlarge the modern genetic pool

Intensive selection to enhance agronomic performance can considerably reduce genetic diversity in crops (1). However, we found little difference in genetic diversity between landraces and inbred lines, which is consistent with the low genetic differentiation we

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

observed 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 maize improvement has not altered allele diversity over a very broad geographic scale. This observation is similar to findings in soybean (75) and wheat (76), which also showed a minor effect of crop improvement on diversity, suggesting that landraces have been and still are extensively used in the development of modern inbred lines in these crops. It is important to note however 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 (57).

Several factors could be responsible for the low 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 (77) 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, i.e. the initial modern inbred line breeding pools, selection of genetically diverse and complementary heterotic groups may have mitigated the loss of diversity (78). Furthermore, SNPs from 50K arrays were previously identified in 27 lines (79). 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) (57).

Despite the limited differences in overall diversity between landraces and inbred lines, two different approaches highlighted that the majority of landraces had made a limited contribution to recent breeding. We identified a number of landraces with a high median Hs value and a small MRD_{LI} distance range reflecting a lack of similarity 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 (12, 13, 55). Interestingly, we observed a large increase or decrease in the contribution of landraces between first cycle lines and more advanced lines

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

(Fig. S13C). This can be explained by the fact that some lines were extensively used to derive more advanced lines whereas others were not (12, 14, 15). 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 (Fig. 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 (Fig.5). By combining closely located SNPs, we were able to identify 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.

MATERIALS AND METHODS

Plant material

Landraces

A total of 156 different landrace populations (Table S1) were sampled from a panel of 413 landraces (Supplementary Information 1). These 156 landraces captured a large proportion of European and American diversity and have been analyzed in previous studies using RFLP (25, 31–34) and SSR markers (23, 27, 28). Each population was represented by either one or

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

two sets of 15 individual plants (for 146 and 10 populations, respectively), pooled equally as described in Reif *et al.* (48) and Dubreuil *et al.* (28) . The 166 DNA samples corresponding to the 156 landrace accessions were classified into five geographic groups (Table S1): Europe (EUR), North America (NAM), Central America and Mexico (CAM), the Caribbean (CAR) and South America (SAM).

Inbred lines

We analyzed 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 S11). These 442 lines were partitioned into three sets (the "Panel" column in Table S11):

1. "CK lines": a panel of 120 first cycle and 207 advanced lines (327 lines in total) representing American and European diversity (27, 42) including some key founders of modern breeding programs (*e.g.* F2, B73, C103).
2. "Parent Controlled Pools": a set of 12 lines used to build 4 series of 8 controlled DNA pools (see below).
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).

Controlled DNA Pools

To prepare the controlled DNA pools, two sets of three inbred lines were considered: EP1 – F2 – LO3 (European Flint inbred lines) and NYS302– EA1433 – M37W (Tropical inbred lines). For each set of parental lines, nine controlled pools were prepared by varying the proportion of each line in the mix, quantified by the number of leaf disks of equal size as *per* Dubreuil et al., (32). The genotype of each line and the proportion of parental lines in each pool were used to estimate allelic frequencies in the nine pools, and subsequently to calibrate the model for predicting allelic frequency (see (51) for more detail).

Genotyping and prediction of allelic frequencies in DNA pools

We used the 50K Illumina Infinium HD array (37) 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 and S11). For CK lines, we used the 50K genotyping data from Bouchet et al. (2013). 23,412 SNPs were filtered based on their suitability for diversity

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

analysis and their quality for predicting allelic frequency in DNA pools (Supplementary Information 2).

Allelic frequency of selected SNPs in DNA pools was estimated using the two-step procedure described in (51) 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 1,000 SNPs from 2 series of controlled pools (see (51) for more detail and equation 2 for the model).

We also used the genotyping data from 17 SSRs from 145 and 11 landraces obtained by Camus-Kulandaivelu et al. (27) and Mir et al. (23), respectively.

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) (80, 81).

Genetic differentiation (FST) was estimated between: individual landraces (FST_l), between the five landrace geographic groups (FST_g), between 10 pairs of geographic groups ($FST_{EUR\text{-}NAM}$, $FST_{EUR\text{-}CAM}$, $FST_{EUR\text{-}CAR}$, $FST_{EUR\text{-}SAM}$, $FST_{NAM\text{-}CAM}$, $FST_{NAM\text{-}CAR}$, $FST_{NAM\text{-}SAM}$, $FST_{CAM\text{-}CAR}$, $FST_{CAM\text{-}SAM}$, $FST_{CAR\text{-}SAM}$) and between landraces and inbred lines (FST_i). FST was estimated at each locus and across all loci as *per* (81, 82) (Supplementary Information 3).

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 analyze 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 4,095 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 FST 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 (83).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

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 FST, referred to as "Outlier FST analysis" (ii) the detection of loci under selection using the drift model implemented in the BAYESCAN software (84) (Supplementary Information 4).

Genetic structure and relationship between landraces

We estimated the genetic distance between all landraces using modified Roger's distance (MRD) (85) based on the allelic frequencies of 23,412 prefixed PZE SNPs. MRD was then averaged within and between geographic groups (Table 1, Table S2). We analyzed the relationship between genetic and geographic distances within each geographic group by plotting MRD against geographic distances. We tested this correlation using the Mantel test (86). Geographic distances were calculated using the latitude and longitude of each sampling site using the geosphere R package v. 1.5-10 (87).

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) (88), (ii) hierarchical clustering using either Ward or Neighbor-Joining algorithms implemented in the "hc" and "bionj" functions of the "ape" R package v 5.0 (89), 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 (90). Different methods were used to identify the most appropriate number of ancestral populations (K): Cross-validation error or difference between successive cross-validations (90) and Evanno's graphical methods (91). Since ADMIXTURE requires multi-locus genotypes of individual plants, we simulated the genotype of five individuals for each population for a subset of 2,500 independent SNPs to avoid artifacts of linkage disequilibrium (Supplementary Information 5).

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

1) A distance-based approach in which we estimated the modified Roger's distance between each landrace and the 327 CK lines (MDRLI) 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 S11). 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 analyzed 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 analyzed 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 50K 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 analyzed from this study.

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

To analyze 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

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 MRD_{med} and Hs. We defined MRD_{norm} as the orthogonal deviation of MRD_q (with q = 5% or 10% for MRD_{05} and MRD_{10}, respectively) from the linear regression:

$$MRD_{norm} = (MRD_{med} - MRD_q) \times \sin(\tan^{-1}(a)) \quad (1)$$

We used MRD_{norm} based on MRD_{10} to categorize the 33 landraces into three classes based on the percentile distribution of MRD_{norm}. Landraces with 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 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 "Diversity Zea" project and the French National Research Agencies with thein "Investissement d'Avenir Amaizing" project, (ANR-10-BTBR-01). We greatly acknowledge the French Maize Biological Ressource 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 and André Gallais for initiating these research programs. We also greatly acknowledge Pierre Dubreuil, Letizia Camus-Kulandaivelu, Cecile Rebourg, Céline Mir, Domenica Maniccaci who conducted previous studies on these landraces using the DNA pooling approach with SSR and RFLP markers. The Infinium genotyping work was supported by CEA-CNG. We thank Anne Boland and Marie-Thérèse Bihoreau and their staff. We acknowledge the EPGV group, Dominique Brunel, Aurélie Bérard and Aurélie Chauveau for discussion and management of Illumina genotyping.

Author contributions

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

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 andmaintained 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 analyzed the genetic diversity of the landrace panel; M.A and S.D.N analyzed 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

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. Note that data will be available at the two web links below upon the publication will have been accepted in a journal.

Conflicts of interest

No

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

REFERENCES

1. S. D. Tanksley, Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild. *Science* **277**, 1063–1066 (1997).

2. D. Hoisington, *et al.*, Plant genetic resources: What can they contribute toward increased crop productivity? *Proc. Natl. Acad. Sci.* **96**, 5937–5943 (1999).

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

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

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

6. M. Mascher, *et al.*, Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* **51**, 1076–1081 (2019).

7. D. J. Gates, *et al.*, "Single-gene resolution of locally adaptive genetic variation in Mexican maize" (Evolutionary Biology, 2019) https:/doi.org/10.1101/706739 (September 25, 2020).

8. J. C. Reif, *et al.*, Wheat genetic diversity trends during domestication and breeding. *Theor. Appl. Genet.* **110**, 859–864 (2005).

9. M. Yamasaki, *et al.*, A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**, 2859–2872 (2005).

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

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

12. J. T. Gerdes, W. F. Tracy, Pedigree Diversity within the Lancaster Surecrop Heterotic Group of Maize. *Crop Sci.* **33**, 334–337 (1993).

13. J. van Heerwaarden, *et al.*, Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl. Acad. Sci.* **108**, 1088–1092 (2011).

14. M. A. Mikel, Genetic Composition of Contemporary U.S. Commercial Dent Corn Germplasm. *Crop Sci.* **51**, 592–599 (2011).

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

16. A. Strigens, W. Schipprack, J. C. Reif, A. E. Melchinger, Unlocking the Genetic Diversity of Maize Landraces with Doubled Haploids Opens New Avenues for Breeding. *PLoS ONE* **8**, e57234 (2013).

17. P. C. Brauner, *et al.*, 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 (2019).

18. A. C. Hölker, *et al.*, European maize landraces made accessible for plant breeding and genome-based studies. *Theor. Appl. Genet.* **132**, 3333–3345 (2019).

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

20. Y. Matsuoka, *et al.*, A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci.* **99**, 6080–6084 (2002).

21. M. I. Tenaillon, A. Charcosset, A European perspective on maize history. *C. R. Biol.* (2011).

22. K. Swarts, *et al.*, Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. *Science* **357**, 512–515 (2017).

23. C. Mir, *et al.*, Out of America: tracing the genetic footprints of the global diffusion of maize. *Theor. Appl. Genet.* **126**, 2671–2682 (2013).

24. A. Brandolini, Maize. *Maize.* (1970).

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

26. C. Rebourg, *et al.*, Maize introduction into Europe: the history reviewed in the light of molecular data. *Theor Appl Genet* **106**, 895–903 (2003).

27. L. Camus-Kulandaivelu, Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. *Genetics* **172**, 2449–2463 (2006).

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

29. J.-T. Brandenburg, *et al.*, Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. *PLoS Genet.* **13**, e1006666 (2017).

30. L. Wang, *et al.*, "Molecular Parallelism Underlies Convergent Highland Adaptation of Maize Landraces" (Evolutionary Biology, 2020) https:/doi.org/10.1101/2020.07.31.227629 (September 25, 2020).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

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

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

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

34. P. Gauthier, *et al.*, RFLP diversity and relationships among traditional European maize populations. *Theor Appl Genet* **105**, 91–99 (2002).

35. Y. Vigouroux, *et al.*, An analysis of genetic diversity across the maize genome using microsatellites. *Genetics* **169**, 1617–1630 (2005).

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

37. M. W. Ganal, *et al.*, 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 (2011).

38. S. R. McCouch, *et al.*, Development of genome-wide SNP assays for rice. *Breed. Sci.* **60**, 524–535 (2010).

39. M. Moragues, *et al.*, Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theor. Appl. Genet.* **120**, 1525–1534 (2010).

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

41. M. B. Hufford, *et al.*, Comparative population genomics of maize domestication and improvement. *Nat. Genet.* (2012) https:/doi.org/10.1038/ng.2309 (July 11, 2012).

42. S. Bouchet, *et al.*, Adaptation of maize to temperate climates: mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the Vgt2 (ZCN8) locus. *PloS One* **8**, e71377 (2013).

43. E. Frascaroli, T. A. Schrag, A. E. Melchinger, 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 (2013).

44. M. C. Arteaga, *et al.*, Genomic variation in recently collected maize landraces from Mexico. *Genomics Data* **7**, 38–45 (2016).

45. M. H. Reyes-Valdés, *et al.*, Analysis and Optimization of Bulk DNA Sampling with Binary Scoring for Germplasm Characterization. *PLoS ONE* **8**, e79936 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

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

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

48. J. C. Reif, *et al.*, Grouping of accessions of Mexican races of maize revisited with SSR markers. *Theor. Appl. Genet.* **113**, 177–185 (2006).

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

50. S. Unterseer, *et al.*, A powerful tool for genome analysis in maize: development and evaluation of the high density 600k SNP genotyping array. *BMC Genomics* **15**, 823 (2014).

51. M. Arca, *et al.*, Deciphering the genetic diversity of landraces with high-throughput SNP genotyping of DNA bulks: methodology and application to the maize 50k array. *bioRxiv* (2020) https:/doi.org/10/gg8nxn (June 4, 2020).

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

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

54. A. Albrechtsen, F. C. Nielsen, R. Nielsen, Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Mol. Biol. Evol.* **27**, 2534–2547 (2010).

55. J. A. Romero Navarro, *et al.*, A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nat. Genet.* **49**, 476–480 (2017).

56. J. Böhm, W. Schipprack, H. F. Utz, A. E. Melchinger, Tapping the genetic diversity of landraces in allogamous crops with doubled haploid lines: a case study from European flint maize. *Theor. Appl. Genet.* **130**, 861–873 (2017).

57. L. Zeitler, J. Ross-Ibarra, M. G. Stetter, "Loss of diversity and accumulation of genetic load in doubled-haploid lines from European maize landraces" (Evolutionary Biology, 2019) https:/doi.org/10.1101/817791 (February 5, 2020).

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

59. Y. Diaw, *et al.*, "Genetic diversity of maize landraces from the South-West of France" (Genetics, 2020) https:/doi.org/10.1101/2020.08.17.253690 (September 3, 2020).

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

807 61. J. A. Aguirre-Liguori, *et al.*, Connecting genomic patterns of local adaptation and niche
808 suitability in teosintes. *Mol. Ecol.* **26**, 4226–4240 (2017).

809 62. E. Millet, *et al.*, Genome-wide analysis of yield in Europe: allelic effects as functions of
810 drought and heat scenarios. *Plant Physiol.*, pp.00621.2016 (2016).

811 63. P. Revilla, W. F. Tracy, Isozyme Variation and Phylogenetic Relationhips among Open-
812 Pollinated Sweet Corn Cultivars. *Crop Sci.* **35**, 219–227 (1995).

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

815 65. V. Jaenicke-Despres, *et al.*, Early allelic selection in maize as revealed by ancient DNA.
816 *Science* **302**, 1206–1208 (2003).

817 66. W. F. Tracy, S. R. Whitt, E. S. Buckler, Recurrent Mutation and Genome Evolution:
818 Example of and the Origin of Sweet Maize. *Crop Sci.* **46**, S-49 (2006).

819 67. M. C. Romay, *et al.*, Comprehensive genotyping of the USA national maize inbred seed
820 bank. *Genome Biol* **14**, R55 (2013).

821 68. B. Gouesnard, *et al.*, Genotyping-by-sequencing highlights original diversity patterns
822 within a European collection of 1191 maize flint lines, as compared to the maize USDA
823 genebank. *Theor. Appl. Genet.* (2017) https:/doi.org/10.1007/s00122-017-2949-6
824 (August 8, 2017).

825 69. C. Huang, *et al.*, ZmCCT9 enhances maize adaptation to higher latitudes. *Proc. Natl.*
826 *Acad. Sci.*, 201718058 (2017).

827 70. S. S. Negro, *et al.*, Genotyping-by-sequencing and SNP-arrays are complementary for
828 detecting quantitative trait loci by tagging different haplotypes in association studies.
829 *BMC Plant Biol.* **19**, 318 (2019).

830 71. L. Virlouvet, *et al.*, The ZmASR1 Protein Influences Branched-Chain Amino Acid
831 Biosynthesis and Maintains Kernel Yield in Maize under Water-Limited Conditions.
832 *Plant Physiol.* **157**, 917–936 (2011).

833 72. A. Yilmaz, *et al.*, GRASSIUS: A Platform for Comparative Regulatory Genomics across
834 the Grasses. *Plant Physiol.* **149**, 171–180 (2009).

835 73. Y. Gu, *et al.*, Genome-wide identification and abiotic stress responses of DGK gene
836 family in maize. *J. Plant Biochem. Biotechnol.* (2017) https:/doi.org/10/gd5758 (July 17,
837 2020).

838 74. B. Wang, *et al.*, Genome-wide selection and genetic improvement during modern maize
839 breeding. *Nat. Genet.* **52**, 565–571 (2020).

840 75. D. L. Hyten, *et al.*, Impacts of genetic bottlenecks on soybean genome diversity. *Proc.*
841 *Natl. Acad. Sci.* **103**, 16666–16671 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

76. C. R. Cavanagh, *et al.*, Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci.* **110**, 8057–8062 (2013).

77. S. I. Wright, *et al.*, The effects of artificial selection of the maize genome. *Science* **308**, 1310–1314 (2005).

78. Y. Jiao, *et al.*, Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815 (2012).

79. M. A. Gore, *et al.*, A first-generation haplotype map of maize. *Sci. Wash.* **326**, 1115–1117 (2009).

80. M. Nei, Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U A* **70**, 3321–3 (1973).

81. M. Nei, F-statistics and analysis of gene diversity in subdivided populations. *Ann Hum Genet* **41**, 225–33 (1977).

82. S. Wright, The genetical structure of populations. *Ann. Hum. Genet.* **15**, 323–354 (1949).

83. R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, 2013).

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

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

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

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

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

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

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

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

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	0.0192 ± 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 (FST_l) and between inbred lines (FST_i)	0.393 ± 0.001	0.341 ± 0.001	0.303 ± 0.001	0.275 ± 0.001	0.334 ± 0.001	0.405 ± 0.002	0.994

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

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 FST windows								Bayescan hits (Decisive) – Geographical						Frequency of allele B				
Name*	Chr	Start - Stop (Mbp)	SNP_w	FST_g	FST_l	HT_l	Hs	SNP_b	Marker name	Pos. (Mbp)	FST_b	Closest Gene	Dist. from gene (kbp)	Functionnal 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	0.33
									PZE-103058429	78.5	0.30	AC202959.3_FG001	0		0..30	0..34	0.72	1.00	0.67
									PZE-103058437	78.5	0.29	GRMZM2G112187	6		0.69	0.67	0.32	0.00	0.33
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	0.33
									PZE-107023081	84.9	0.29	GRMZM2G112579	5.6	Pectin lyase-like superfamily protein	0.69	0.65	0.32	0.00	0.33
									PZE-107023082	84.9	0.29		5.7		0.31	0.35	0.64	1.00	0.67
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	0.33
									PZE-104010477	7.6	0.31		0		0.97	0.95	0.24	0.83	0.67
									PZE-104010709	8.8	0.30	GRMZM2G119698	0	pectinesterase	0.06	0.06	0.79	0.29	0.33
									PZE-104010719	8.8	0.28	GRMZM2G702341	0.2		0.98	0.95	0.34	0.95	0.33
									PZE-104010855	9.4	0.27	GRMZM2G419836	0	Thioredoxin superfamily protein	0.98	0.96	0.42	0.80	0.33
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	0.67
									PZE-104033229	41.4	0.28	GRMZM2G138198	0	Pollen receptor-like kinase 4	0.49	0.66	0.06	0.00	0.33
									PZE-104033340	41.7	0.27	GRMZM2G174149	0	RNA pseudouridine synthase 3 mitochondrial	0.54	0.39	0.94	1.00	0.33
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	0.67
									PZE-106078990	134.8	0.24	GRMZM2G170646	0	GDSL esterase/lipase	0.50	0.63	0.07	0.01	0.33
									PZE-106079041	134.8	0.28		0.6		0.55	0.44	0.97	1.00	0.67
									PZE-106079060	134.9	0.25	GRMZM2G162702	0	Probable receptor-like protein kinase	0.57	0.49	0.96	1.00	0.33
									PZE-106079065	134.9	0.27		0		0.57	0.49	0.98	1.00	0.33
									PZE-106079127	135.0	0.29	GRMZM2G307720	0	TATA box-binding protein	0.49	0.31	0.92	1.00	0.33

* 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 FST windows and detected as being under selection by Bayescan, respectively; FST_g and FST_l indicate the average FST 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/).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

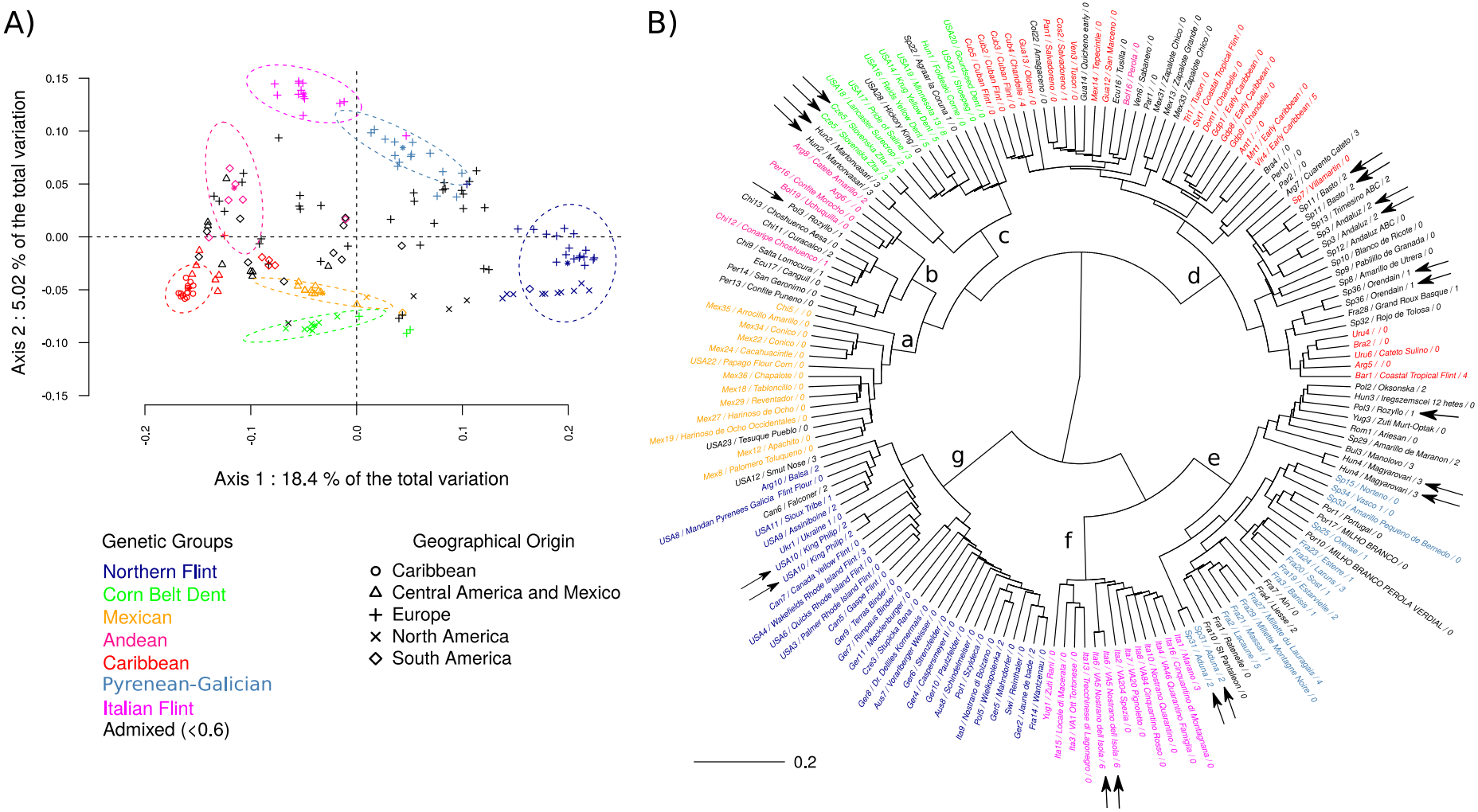


Fig. 1: Genetic relationship between 156 maize landraces based on their modified Roger's distance (MRD). A) Projection of the 166 DNA samples on the first two axes of the Principal Coordinate Analysis. Symbols indicate the geographic origin of landraces. B) Dendrogram obtained by Hierarchical clustering, 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. Colors 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 colored in black.

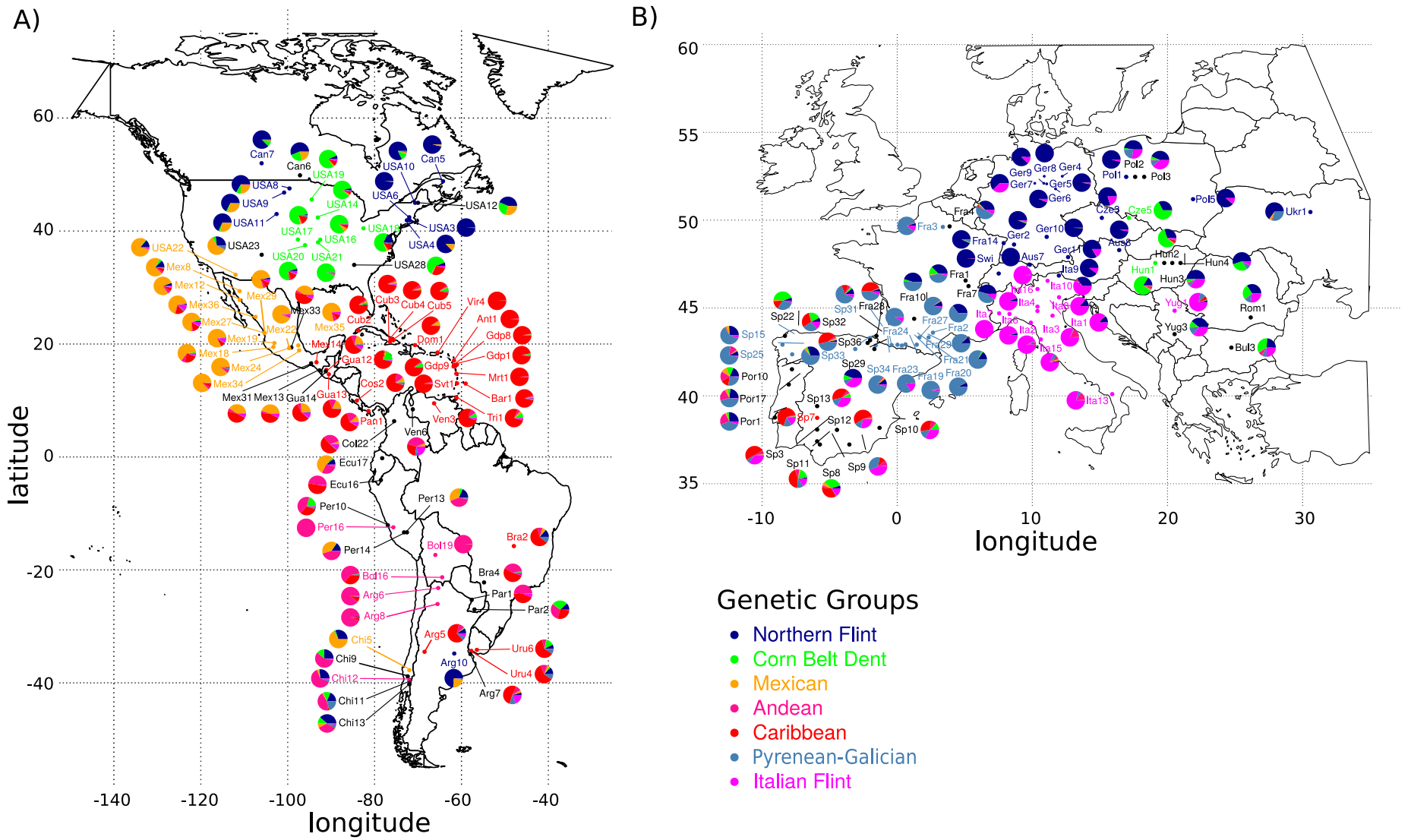
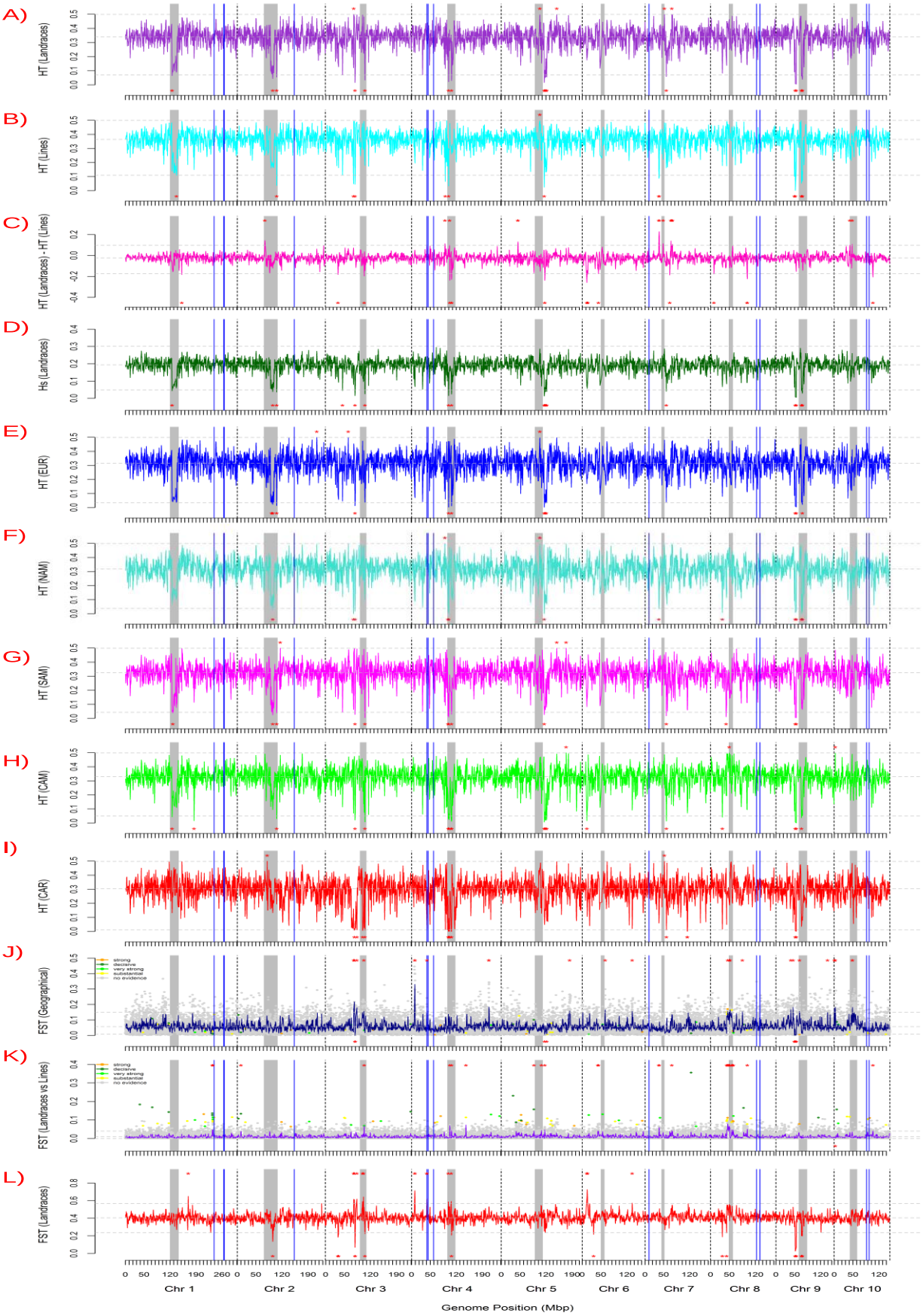


Fig. 2: Spatial genetic structure of American (A) and European (B) maize landraces. Population structure is based on ADMIXTURE analysis with K = 7. Each population is represented by a pie diagram whose composition indicates admixture coefficients. Population labels are colored according to their main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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.321018. this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Fig. 3: Variation in genetic diversity and differentiation along the maize genome. A) Total expected heterozygosity across landraces: HT (Landraces); B) total expected heterozygosity (HT) 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) FST between geographic groups of landraces: FST (Geographic); K) FST between landraces and inbred lines: FST (Landraces vs. Inbred lines); L) FST between landraces: FST (Landraces). Loci with decisive, very strong, strong, substantial, no evidence of selection using bayescan are colored in orange, dark green, light green, yellow and blue (J, K, L). Vertical gray bars correspond to centromere limits. Chromosome boundaries are indicated by vertical dashed lines. Horizontal dashed lines correspond to the mean, 5^{th} and 95^{th} 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 *Zmcc*

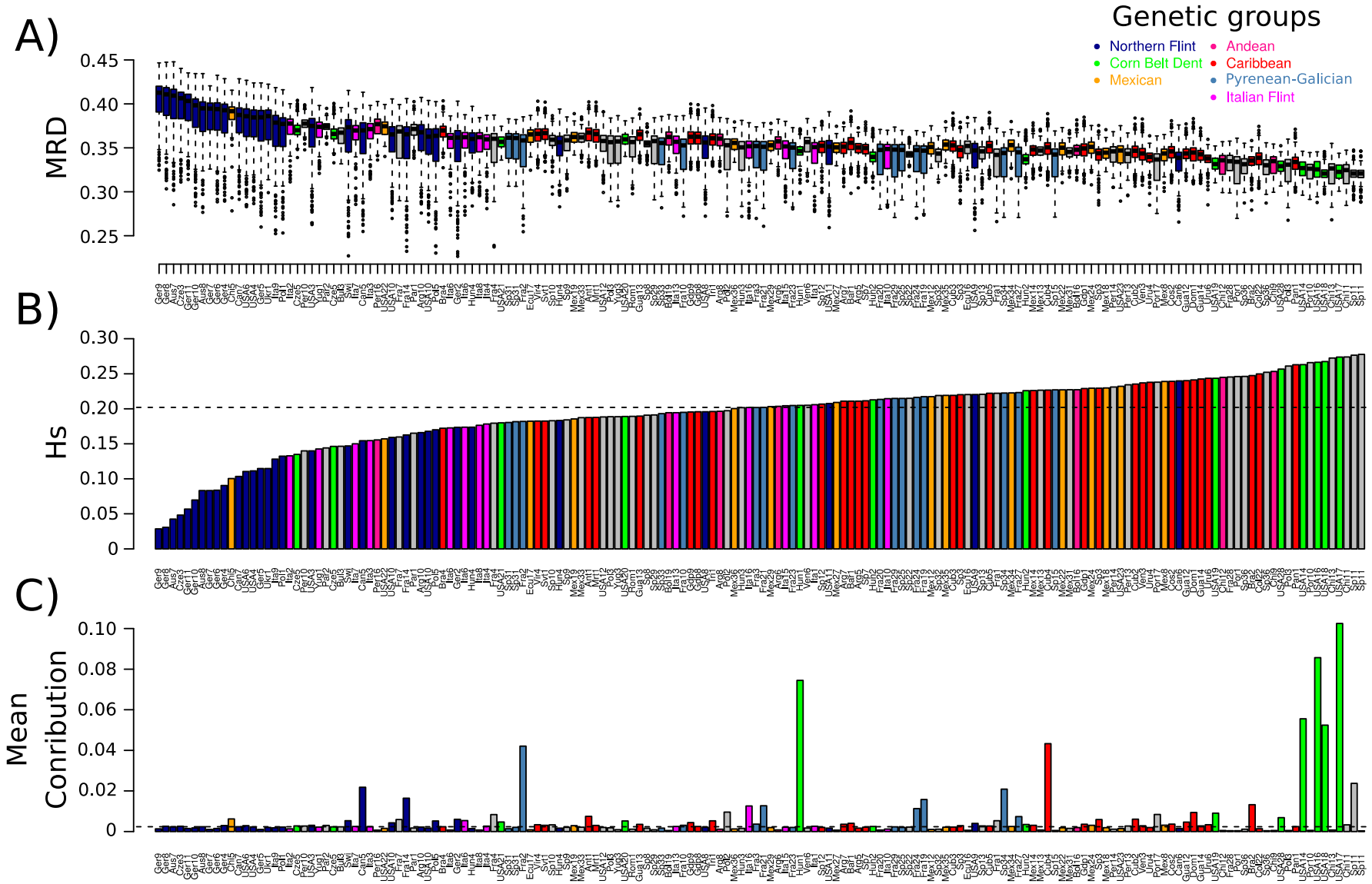


Fig. 4: Contribution of landraces to the panel of CK lines in relation to their genetic diversity. A) Box plot representation of pairwise modified Roger's distances (MRD) between individual landraces and inbred CK lines. Each box represents the interquartile range, the line within each box represents 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 colored based on the assignment of landraces to the seven genetic groups identified by ADMIXTURE (see bottom right for colors).

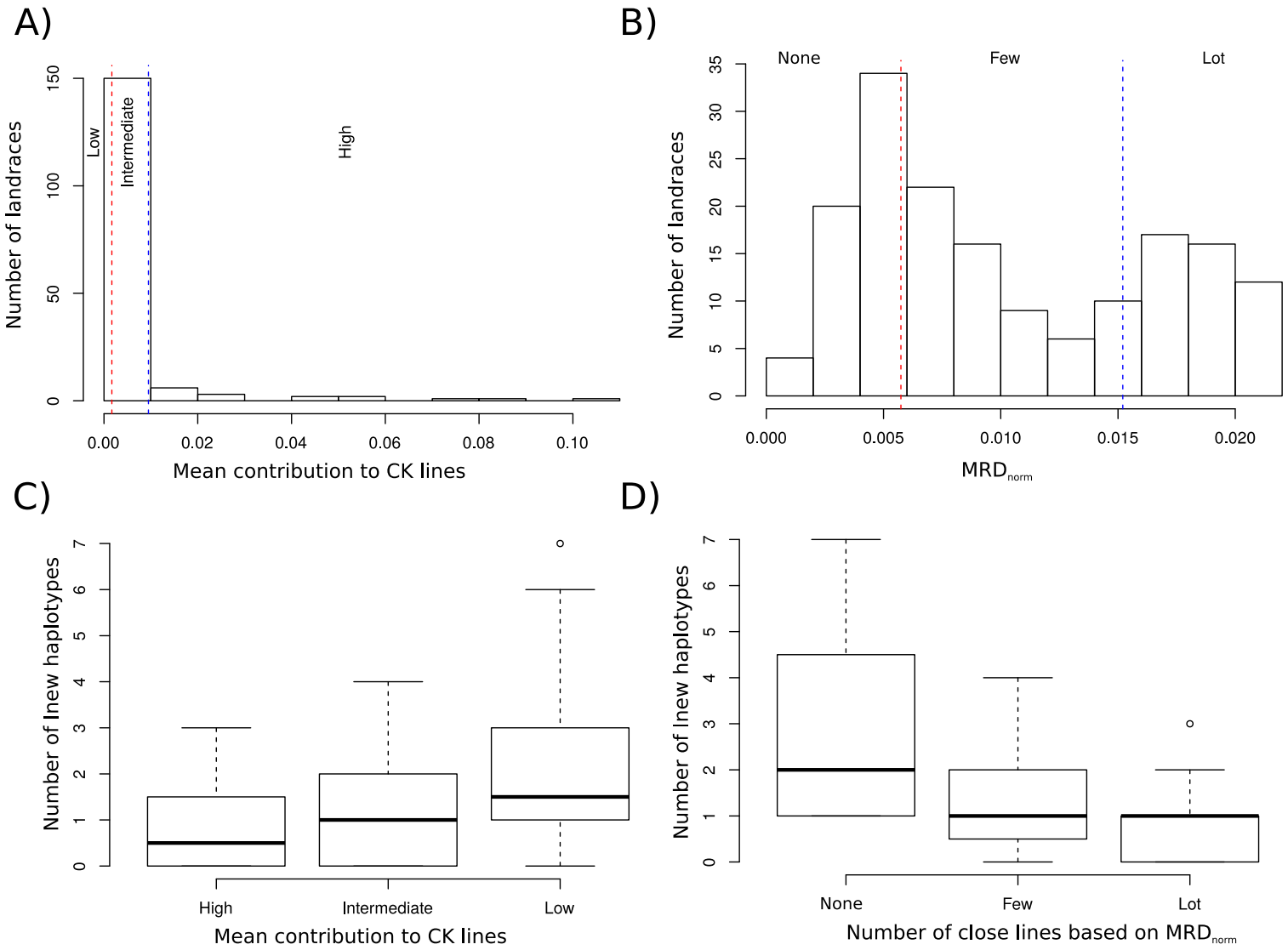


Fig. 5: Allelic enrichment of CK lines by new DH-SSD lines derived from landraces according their contribution and their genetic distance to CK lines. *Allelic enrichment was estimated by the number of new haplotypes discovered in the 66 new DH-SSD lines derived from 33 landraces, compared to the 327 CK lines (C, D) that 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 lue 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}.*