# Genetic Relationships and Genome Selection Signatures Between Soybean Cultivars from Brazil and United States After Decades of Breeding

**João Vitor Maldonado dos Santos** ( joaomaldonado@tmg.agr.br )
Tropical Melhoramento & Genética
**Gustavo Cesar Sant'Ana**
Tropical Melhoramento & Genética
**Philip Traldi Wysmierski**
Tropical Melhoramento & Genética
**Matheus Henrique Todeschini**
Tropical Melhoramento & Genética
**Alexandre Garcia**
Tropical Melhoramento & Genética
**Anderson Rotter Meda**
Tropical Melhoramento & Genética

**Research Article**

**Keywords:** Glycine max, SNP, Germplasm, Population structure, Diversity, Genomic signatures.

**Posted Date:** December 8th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-1111240/v1

# Abstract

Soybeans are one of the most important crops worldwide. Brazil and the United States (US) are the world's two biggest producers of this legume. The increase of publicly available DNA sequencing data as well as high-density genotyping data of multiple soybean germplasms has made it possible to understand the genetic relationships and identify genomics regions that underwent selection pressure during soy domestication and breeding. In this study, we analyzed the genetic relationships between Brazilian (N=235) and US soybean cultivars (N=675) released in different decades, and screened for genomic signatures between Brazilian and US cultivars. The population structure analysis demonstrated that the Brazilian germplasm has a narrower genetic base than the US germplasm. The US cultivars was grouped according to the maturity groups, while Brazilian cultivars were separated according to decades of releases. We found 73 SNPs that differentiate Brazilian and US soybean germplasm. Maturity-associated SNPs showed high allelic frequency differences between Brazil and US accessions. Other important loci were identified separating cultivars released before and after 1996 in Brazil. Our data showed important genomic regions under selection during decades of soybean breeding in Brazil and the US that should be targeted to adapt lines from different origins in these countries.


# Introduction

Soybean [*Glycine max* (L.) Merrill] is one of the most important crops worldwide. It contributes to oil production as well as human and animal diets[1]. Brazil and United States (US) are responsible for more than 60% of the world production in the 2018-2019 growing season. Brazil had a soybean production of 119.0 million tons harvested from 35.9 million hectares of cultivated area, and the US was responsible for 120.5 million tons harvested from 35.45 million hectares of cultivated area[2]. The Brazilian soybean-breeding programs has a relatively brief history, with US germplasm introduction starting in the 1940s and becoming economically important after 1970s with the discovery of genes related to long juvenile period trait[3]. The US soybean-breeding program has an older history than the Brazilian breeding program. Soybeans were introduced in the US at the beginning of the 1900s, but it only became important as an oilseed crop after World War II[4].

Despite advances in the soybean-breeding program in germplasm improvement, some important factors limit crop production. One of the biggest challenges is the narrow genetic base observed in soybean germplasm. According to a pedigree analysis, the US genetic base was basically generated by 35 soybean genotypes[5]. In another study, a similar analysis found that Roanoke, S-100, CNS, and Tokyo contributed to 55.3% of the Brazilian genetic base. Furthermore, Brazilian and US germplasms shared six ancestors: CNS, S-100, Roanoke, Tokyo, PI 54610, and PI 548318[6]. A genomic study with 28 Brazilian soybean accessions suggested that the genetic base remains narrow despite some diversified genomic regions[7].

Next-generation sequencing methods have become an important tool to increase soybean genome knowledge[8,9]. The first reference genome for *Glycine max* was assembled in Williams 82 cultivar with 46,430 protein-coding genes distributed on 20 chromosomes with approximately 978-megabases (Mb) in total size[10]. Recently, two other reference genomes were generated in *G. max*.: the Chinese accession Zhonghuang

13 (with a reference genome with 1,025 Mb of total size, and 52,051 protein-coding genes)[11] and cultivar Lee (with approximately 1,015-Mb of total size)[12]. The existence of reference genomes in soybean facilitated the publication of a large number of studies associated with diversity and population analysis, allelic variation discovery, and genome-wide association studies (GWAS)[8].

In this context, the objectives of this study were to analyze genetic kinship relationships between Brazilian and US soybean cultivars from different maturity groups and release dates as well as to identify genome selection signatures between and within Brazilian and North-American cultivars.

## Results

# Different structures were detected between the Brazilian and US genetic bases

Principal component analysis revealed that most Brazilian cultivars (red circle) were grouped with a subgroup of US cultivars (green circle). Most of them belonged to maturity groups VI, VII, VIII and IX (Figure 1A). Based on the Evanno criterion (Figure 1B), the structure results based on four groups (K = 4) showed a high ΔK value (312.35), but the upper-most level of the structure was in two groups (K = 2; ΔK = 1885.43).

Considering K = 2 (Figure 1C), the Brazilian cultivars jointly presented an assignment to the Q1 group (green) equal to 86.7% which was much higher than that observed for the US cultivars (43.9%). Considering K = 4 (Figure 1D), the Brazilian cultivars jointly presented an assignment to the Q2 group (red) of only 4.7% while the US cultivars jointly presented an assignment to the Q2 group of 27.4%. The Q1 group (green) has a lower assignment in Brazilian cultivars than US accessions (11.1%, and 30.1%, respectively). As expected, these results confirm that the set of Brazilian cultivars has a narrower genetic base compared to US cultivars.

# Large genetic divergences in United States and Brazil soybean germplasm were observed according to their maturity groups

When we compared the cultivars between maturity groups, we observed a clear differentiation between early and late groups (Table 1). The highest genetic distance (0.42) was observed between cultivars with MG 00-0 and MG VIII-IX.

Table 1
Summary of the genomic regions with high $F_{ST}$ values between Brazilian and US germplasms.

| Chr.[a] | Start (Mbp)[b] | End (Mbp)[c] | SNP[d] | $F_{ST}$ | | Tajima's D[g] | | | π ($10E^{-05}$)[h] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | (High)[e] | (Reg.)[f] | ALL | BR | US | BR | US | US/BR[i] |
| 1 | 48.70 | 48.80 | 5 | 0.45 | 0.47 | 4.07 | 2.45 | 2.47 | 1.76 | 1.60 | 0.91 |
| 4 | 50.20 | 50.30 | 7 | 0.44 | 0.19 | 4.12 | 2.12 | 3.93 | 1.89 | 2.86 | 1.51 |
| 6 | 0.60 | 0.70 | 6 | 0.40 | 0.32 | 4.20 | 1.42 | 3.53 | 1.58 | 2.38 | 1.50 |
| 6 | 46.90 | 47.00 | 8 | 0.41 | 0.29 | 4.24 | 1.70 | 3.55 | 2.18 | 2.92 | 1.34 |
| 6 | 47.30 | 47.40 | 4 | 0.40 | 0.42 | 4.19 | -0.03 | 3.83 | 0.58 | 1.80 | 3.10 |
| 6 | 47.40 | 47.50 | 9 | 0.41 | 0.39 | 5.58 | 0.37 | 5.08 | 1.73 | 4.08 | 2.36 |
| 6 | 47.50 | 47.60 | 4 | 0.49 | 0.35 | 3.35 | 1.16 | 2.84 | 0.81 | 1.47 | 1.82 |
| 6 | 47.70 | 47.80 | 16 | 0.46 | 0.22 | 5.42 | 2.30 | 5.23 | 3.85 | 6.81 | 1.77 |
| 6 | 47.80 | 47.90 | 15 | 0.40 | 0.29 | 5.87 | 2.98 | 4.97 | 5.05 | 6.20 | 1.23 |
| 6 | 48.10 | 48.20 | 20 | 0.44 | 0.17 | 5.82 | 2.64 | 5.61 | 6.10 | 8.63 | 1.42 |
| 6 | 48.40 | 48.50 | 4 | 0.47 | 0.15 | 1.94 | 1.15 | 1.72 | 0.80 | 1.08 | 1.35 |
| 7 | 6.30 | 6.40 | 6 | 0.44 | 0.16 | 1.32 | 2.34 | 0.78 | 1.63 | 0.90 | 0.55 |
| 9 | 41.50 | 41.60 | 7 | 0.40 | 0.17 | 4.34 | 1.82 | 4.55 | 1.52 | 3.13 | 2.06 |
| 10 | 44.20 | 44.30 | 6 | 0.52 | 0.23 | 2.95 | 2.61 | 2.00 | 2.13 | 1.63 | 0.77 |
| 10 | 44.40 | 44.50 | 7 | 0.44 | 0.16 | 3.84 | 3.05 | 2.90 | 2.66 | 2.58 | 0.97 |
| 12 | 6.10 | 6.20 | 9 | 0.46 | 0.10 | 4.99 | 3.92 | 5.22 | 3.83 | 3.83 | 1.00 |
| 16 | 3.00 | 3.10 | 12 | 0.42 | 0.09 | 1.74 | 2.25 | 1.24 | 3.27 | 2.26 | 0.69 |
| 16 | 29.40 | 29.50 | 10 | 0.45 | 0.12 | 4.63 | 3.96 | 4.24 | 3.86 | 4.01 | 1.04 |
| 16 | 30.70 | 30.80 | 6 | 0.41 | 0.30 | 2.21 | 2.96 | 0.97 | 2.30 | 1.03 | 0.45 |
| 16 | 31.10 | 31.20 | 6 | 0.51 | 0.27 | 3.38 | 0.55 | 3.18 | 0.98 | 2.20 | 2.24 |
| 18 | 48.60 | 48.70 | 5 | 0.42 | 0.32 | 2.76 | 4.00 | 1.20 | 2.45 | 1.12 | 0.46 |
| 18 | 57.20 | 57.30 | 9 | 0.46 | 0.17 | 2.76 | 3.42 | 2.03 | 3.21 | 2.13 | 0.66 |
| 19 | 0.90 | 1.00 | 7 | 0.40 | 0.11 | 2.65 | 3.40 | 2.12 | 2.15 | 1.97 | 0.92 |

**a**: Soybean chromosome; **b**: start position of the genomic region with high $F_{ST}$; **c**: end position of the genomic region with high $F_{ST}$ values; **d**: total of SNPs observed in this interval; **e**: the highest $F_{ST}$ value observed in a SNP of this interval; **f**: the genomic region average $F_{ST}$; **g**: Tajima's D coefficient for all (ALL), Brazilian (BR), and United States (US) germplasms; **h**: nucleotide diversity values for all (ALL), Brazilian (BR), and United States (US) germplasms; **i**: nucleotide diversity ratio between the populations.

| Chr.[a] | Start (Mbp)[b] | End (Mbp)[c] | SNP[d] | $F_{ST}$ | | Tajima's D[g] | | | π ($10E^{-05}$)[h] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | (High)[e] | (Reg.)[f] | ALL | BR | US | BR | US | US/BR[i] |
| 19 | 3.00 | 3.10 | 5 | 0.42 | 0.39 | 2.21 | 4.08 | 0.34 | 2.45 | 0.76 | 0.31 |
| 19 | 3.10 | 3.20 | 4 | 0.40 | 0.42 | 2.84 | 3.23 | 1.25 | 1.78 | 0.94 | 0.53 |
| 19 | 3.40 | 3.50 | 4 | 0.40 | 0.42 | 2.84 | 3.23 | 1.25 | 2.24 | 1.31 | 0.58 |

**a**: Soybean chromosome; **b**: start position of the genomic region with high $F_{ST}$; **c**: end position of the genomic region with high $F_{ST}$ values; **d**: total of SNPs observed in this interval; **e**: the highest $F_{ST}$ value observed in a SNP of this interval; **f**: the genomic region average $F_{ST}$; **g**: Tajima's D coefficient for all (ALL), Brazilian (BR), and United States (US) germplasms; **h**: nucleotide diversity values for all (ALL), Brazilian (BR), and United States (US) germplasms; **i**: nucleotide diversity ratio between the populations.

To examine the influence of maturity groups on population structure, we next analyzed the average assignment coefficients (K=4) of Brazilian and US cultivars for each maturity group (**Supplementary Figure S1**). Brazilian cultivars from maturity group V presented Q1, Q2, Q3, and Q4 equal to 30.4%, 1.9%, 32.1, and 32.0%, respectively; US cultivars from this same maturity group (V) presented means of Q1, Q2, Q3, and Q4 equal to 9.2%, 8.2%, 65.1%, and 17.6%, respectively. This result indicates that, although belonging to the same maturity group, the Brazilian cultivar group presents considerably different allelic frequencies than the US cultivar group especially for Q3 and Q4. US cultivars belonging to earlier maturity groups (00, 0, I, and II) had significantly higher mean assignment coefficient to Q2 group (red) compared to the other later maturity groups (V=8.2%, VI=8.1%, VIII=5.0%, and IX=13.6%). In the case of Brazilian cultivars, the average assignment coefficients for Q2 were much lower (V=1.9%, VI=4.2%, VII=5.6%, VIII=4.9% and IX=4.9%). These results demonstrate an important allelic pool that distinguishes early to late materials present in Q2.

In general, the Brazilian germplasm showed few differences between maturity groups (Table 1 and Figure 2A). This was also observed when we generated a population structure analysis exclusively with their cultivars (Figure 2C**)**. In contrast, the US germplasm showed a high variation of the genetic distance when we analyzed their maturity groups (Table 1) with a clear clustering of the cultivars **(**Figure 2B), which is more obvious when we observed their exclusive population structure analysis (Figure 2D). The results show that early materials tend to be genetically distant from the late cultivars in the US. The maturity groups from the southern-breeding program of the US (V, VI, VII, VIII, and IX) tend to be less genetically divergent versus northern groups (00, 0, I, II, III, and IV). This agrees with previous studies indicating distinct Northern and Southern genetic pools in the US[5]. There is a low divergence among US soybean cultivars from maturity groups higher than V (Figure 2B). In contrast, cultivars from groups 00 and 0 were more genetically distant from materials of the MG III and IV when compared to early materials. Maturity groups I-II showed as an intermediate group between 00-0 and III-IV. The population structure analysis showed a high influence of the Q2 in cultivars with MG 00-II. For cultivars with MG III and IV, we observed an increase of Q1. Finally, there is a high influence of Q3 in cultivars with maturity groups higher than V, which agrees with the genetic distance data.

# Meaningful genetic change of the Brazilian soybean germplasm occurred in modern materials

The results demonstrate that both genetic bases had few increases in genetic distance among modern materials (releases after 2000) when compared to cultivars from 1950 to 1970s (**Table 2**). According to the IBS genetic distance mean, the Brazilian genetic base was more diverse along the decades compared to US germplasm especially when we compared cultivars released before 1970s and after 2000s.

Average assignment coefficients (Q1, Q2, Q3, and Q4) from structure results were calculated for both germplasm pools. All accessions were sorted according to their origin and release decade (Figure 3). We observed high genomic modifications along the decades in the Brazilian germplasm. Modern materials (2000-2010) had Q1, Q2, Q3, and Q4 values of 36.8%, 2.3%, 31.7%, and 26.0%, respectively, while old accessions (1950-1960s) had means of Q1, Q2, Q3, and Q4 equal to 1.6%, 6.6%, 7.0%, and 84.7%, respectively. The Q4 had a high decrease since 1990s whereas Q1 and Q3 had a high increase at the same period. For the US genetic base, we observed an increase of Q3 and a decrease of Q2 over time. Old cultivars had Q1, Q2, Q3, and Q4 values of 36.0%, 33.7%, 12.3%, and 18.1%, respectively, while modern cultivars had Q1, Q2, Q3, and Q4 of 24.3%, 17.5%, 40.3%, and 17.8%, respectively.

Modification during the 1990s became more evident upon analysis of the PCA and structure results of the Brazilian genetic base considering the release decades (Figure 4A **and C**). We observed an increase in the influence of the Q2 in modern materials (2000-2010) when we compared the results to old materials (1950-1970). In contrast, the US genetic base showed few variations over time according to the average of genetic distance (**Table 2**), PCA, and the exclusive population structure analysis (Figure 4B **and D**). These results suggest a large influence of new alleles into Brazilian germplasm after the 1990s.

# Maturity genes under selection between Brazilian and United States cultivars

Seventy-two SNPs with $F_{ST} \geq 0.4$ between Brazilian and United States cultivars were identified (**Supplementary Table S1**). These SNPs are located on chromosomes 1, 4, 6, 7, 9, 10, 12, 16, 18, and 19 (**Supplementary Figure S2**). Twenty-six 100-Kbp genomic regions with a high degree of diversification between Brazilian and US genetic basis were also found (**Table 3**). The results for Tajima's D showed that these regions had balancing events that maintained the diversity of their bases. Two regions on chromosome 6 (47.3 – 47.4 Mbp and 47.3 - 47.4 Mbp) and another on chromosome 16 (31.10 - 31.20 Mbp) had few variations in Brazilian accessions (**Supplementary Table S2**). In contrast, the allele distribution for most of the SNPs present in these genomic regions in US germplasm were higher compared to Brazilian germplasm. An opposite scenario was observed for the other three regions located on chromosomes 7 (6.30 – 6.40 Mbp), 16 (30.70 – 30.80), and 19 (3.00 – 3.10) (**Supplementary Table S2**). The allele variance was higher in the Brazilian genetic base than US germplasm for these three intervals.

Some SNPs had a large impact on the differentiation of Brazil and US genetic bases. These were located close to three important soybean maturity loci: *E1* (Chr06: 20,207,077 to 20,207,940 bp), *E2* (Chr10: 45,294,735 to 45,316,121 bp) and *FT2a* (Chr16: 31,109,999 to 31,114,963)[13–15] (Figure 5). For the SNPs ss715607350 (Chr10: 44,224,500), ss715607351 (Chr10: 44,231,253), and ss715624321 (Chr16: 30,708,368), we found that the alternative allele was barely present in US germplasm whereas the Brazilian genetic base had an equal distribution between reference and alternative alleles. When we examined the

SNPs ss715624371 (Chr16: 31,134,540) and ss715624379 (Chr16: 31,181,902), the frequency of the alternative allele remains low in the US germplasm. However, the alternative alleles of these two SNPs were present in more than 78% of the Brazilian accessions in contrast to the previous three SNPs. Finally, the alternative allele for SNPs ss715593836 (Chr06: 20,019,602) and ss715593843 (Chr06: 20,353,073) were extremely rare in Brazilian germplasm with only 2% of the accessions carrying them. In contrast, the US germplasm had an equal distribution of reference and alternative alleles in their accessions. However, all accessions with the alternative alleles belonged to MGs lower than VI with less than five cultivars in MG V.

Ten SNPs were identified related to the gene's modifier mutations present in Brazilian and US germplasm; these were distributed on chromosomes 4, 6, 10, 12, 16, and 19 (**Supplementary Table S3**). These SNPs had different allele frequency and could distinguish both genetic bases. Six modifications had a clear influence on the maturity of the accessions whereas two of these had a large influence in some decades of breeding (**Supplementary Figure S3)**. The SNP ss715593833 had similar haplotype of the two SNPs described close to *E1* loci (ss715593836 and ss715593843) due the LD among them. At the end of this chromosome, we also observed another three relevant SNPs in LD: ss715594746, ss715594787, and ss715594990. In the US germplasm, we observed a decrease in the alternative allele in accessions with MG values below to IV. We detected other relevant modifications on chromosome 12 for SNPs ss715613204 and ss715613207. Both SNPs had a minor allele frequency higher than 0.35 in Brazilian germplasm with an increase in the alternative allele in materials with MGs higher than VII. In contrast, alternative alleles for both SNPs were barely present in the US germplasm except for accessions with MG higher than VII.

There were 312 genomic regions that differentiate north (00 – IV MG) and south (V – IX MG) cultivar groups (**Supplementary Table S4**). Some important regions were observed to be less diverse in northern accessions whereas the nucleotide diversity remains present in southern cultivars. The genomic region close to the *Dt1* gene is one example of these specific regions. We compared the SNPs observed in the genomic region close to the *Dt1* gene (Chr19: 45.20 - 45.30 Mbp) with the growth habit phenotype data available for 284 lines at the USDA website (www.ars-grin.gov). The phenotypic data suggested that these SNPs were associated with trait growth habit. Moreover, our diversity analysis demonstrated a putative selective sweep for the *Dt1* gene in the northern germplasm, which has the dominant loci fixed for *Dt1*; the southern lines tend to be more diverse compared to the northern US cultivars (**Supplementary Table S5**). In contrast, other genomic regions have lower nucleotide diversity in southern accessions compared to the northern accessions. An important disease resistance cluster gene was observed on chromosome 13 bearing four loci: *Rsv1*, *Rpv1*, *Rpg1*, and *Rps3*[16–19]. In this interval, we observed two genomic regions (29.70 – 29.80 Mbp and 31.90 – 32.00 Mbp) under putative selective sweeps in the southern germplasm (**Supplementary Table S6**).

Besides these regions, 1,401 SNPs with $F_{ST}$ values higher than 0.40 between northern and southern US cultivars were also identified (**Supplementary Table S7**). In addition, there were 23 SNPs with $F_{ST}$ values higher than 0.70 spread on chromosomes 1, 3, 6, and 19. Seven of them were located close to another important soybean locus: *E1* (involved in soybean maturity control) (**Supplementary Table S8**). These SNPs clearly differentiate northern and southern US cultivars with the reference allele fixed in northern materials, and the alternative alleles into southern accessions. Gene modification in US germplasms were also detected

in our study. One hundred twenty-six SNPs were identified in $F_{ST}$ analysis modifying 125 genes (**Supplementary Table S9**).

Finally, we detected 1,557 SNPs with $F_{ST}$ values higher than 0.40 between super-early cultivars (00 – 0 MG) and early cultivars (III – IV MG) (**Supplementary Table S10**). Seventeen SNPs had $F_{ST}$ values higher than 0.70 spread on chromosomes 4, 7, 8, and 10. The SNPs identified on chromosome 10 were close to the *E2* loci. We also detected 168 SNPs associated with modifications in 164 genes (**Supplementary Table S11**).

# Genetic diversity over time was higher in Brazilian modern cultivars than founder lines

We observed two large SNPs differences in allelic frequencies on Brazilian germplasm (**Supplementary Figure S4**). On chromosome 4, the SNP ss715588874 (50,545,890 bp) had a decrease of the allele "A" in materials released after 2000 with only nine of the 45 Brazilian materials with this allele. Similar situations were observed on chromosome 19 for ss715633722 (3,180,152 bp) with half of the modern accessions having the presence of allele C. Both SNPs had similar distribution according to their decades in the US genetic base with a large influence of reference alleles.

We also observed important results associated with the Brazilian genetic base. There were 126 genomic regions spread on almost all soybean chromosomes. The only exception was chromosome 20 (**Supplementary Table S12**). Our analysis between cultivars released before and after 1996 identified 30 putative regions under breeding sweep events. Thirteen regions had a decrease in diversity in modern cultivars according to Tajima's D and π results. Two genomic regions were observed close to important disease resistance loci: one on chromosome 13 (30.30 – 30.40 Mbp) close to an important resistance gene cluster (with *Rsv1*, *Rpv1*, *Rpg1*, and *Rps3*)[16–19] and another on chromosome 14 (1.70 – 1.80 Mbp) with a southern stem canker resistance loci[20,21]. In contrast, thirty-one genomic regions had an increase in diversity in modern materials, which suggested putative introgression events in these accessions. Two genomic regions were observed on chromosome 2 (40.90 – 40.10 Mbp) and 9 (40.30 - 40.40 Mbp). These were previously reported to have an association with ureide content and iron nutrient content, respectively[22,23].

Besides these regions, there were also 409 SNPs with $F_{ST}$ values higher than 0.40, distributed across all soybean chromosomes. There were 73 SNPs with $F_{ST}$ values higher than 0.70 (**Supplementary Table S13**). Some of these SNPs were also reported to be associated with important soybean traits such as plant height, seed mass, water use efficiency, nutrient content, and ureide content[22–26].

We also identified gene modifications with a high impact on the Brazilian genetic base when we compared cultivars according to their release decade. Of the 409 SNPs identified in $F_{ST}$ analysis, we observed 40 SNPs causing modifications in 39 soybean genes (**Supplementary Table S14**). Three SNPs with $F_{ST}$ values higher than 0.70 were associated with non-synonymous modifications: ss715588896 (*Glyma.04G239600* – a snoaL-like polyketide cyclase), ss715607653 (*Glyma.10g051900* – a gene with a methyltransferase domain), and ss715632020 (*Glyma.18G256700* – a PQQ enzyme repeat).

# Discussion

These data suggest that soybeans were domesticated in China from its annual wild ancestor [*Glycine soja* (Sieb. and Zucc.)] more than 5,000 years ago [27]. US soybean history began in colonial times as a forage crop, but breeding programs began in the early 1900s. During 1940s and 1950s, US breeding programs grew in importance and aimed to change plant architecture, maturity, seed quality, and yield. Most of the cultivated soybean came from the public sector until the early 1980s when private companies became an important and leading source of soybean genetics in US[28–30].

The US soybean breeding history is longer than the Brazilian breeding history. The first report of soybeans in Brazil was from 1882 in Bahia state, but the first released cultivars were from the 1950s in Sao Paulo and Rio Grande do Sul states. The national public and private institutes were responsible for most of the cultivars released in Brazil until the 1990s. As soybean production in Brazil became more relevant—along with a more favorable scenario of intellectual property rights—the private sector for cultivar development took a further step in expanding the soybean breeding programs in the country[31].

Here, we compared Brazil and US germplasm across decades and identified four genetic groups in the population structure analysis. When we compared Brazilian population structure, we found that the Q1 genetic group had a large influence in modern materials. Q1 was evenly distributed in the US germplasm across decades. These results might indicate that similar alleles from US germplasm were incorporated into modern Brazilian cultivars. Furthermore, modern cultivars from both germplasms had similar assignments for Q1, Q3, and Q4, which might represent allele introgressions into Brazilian germplasm though soybean-breeding programs. The emergence of new companies brought new lines from other germplasm pools, which might explain the meaningful change in the modern Brazilian cultivars versus those released before 1990.

In contrast, the US genetic base did not show large modifications between decades according to the population structure results. However, when we analyzed the US germplasm according to their maturity groups, it was possible to identify three clusters among the cultivars. The first group was represented by early cultivars (MG = 00, 0, I, and II) with a large influence of Q2 in this germplasm pool; Q3 and Q4 were barely present. The second group was formed by cultivars with MG III, and IV with Q1 having a large influence on the US soybean germplasm. The third group was comprised of cultivars with MG higher than V: This group had a large influence of Q3 in the germplasm. These results indicate that the US genetic base has a large influence of maturity genes in the germplasm. Similar results were observed in another study that analyzed 579 soybeans from the US and Canada. These were clustered into the same three groups that we identified[32]. Our analysis showed an increase of 230 cultivars from other panels, but there was no modification in the genetic structure of the US germplasm even with the addition of new samples.

The comparison between the Brazilian and US genetic bases identified 72 SNPs with high $F_{ST}$ values in 11 chromosomes. Some of these SNPs were located on three known maturity loci: *E1*, *E2*, and *FT2a*. All of these maturity loci have a large impact on soybean maturity. The *E1* locus was previously cloned and identified as a transcription factor with a region distantly related to B3 domain (*Glyma.06g207800*)[15]. A map-based cloning strategy was used to show that the *E2* locus was homologous to the cloned Arabidopsis GIGANTEA

protein (Glyma.10g221500)[13]. *FT2a* (Glyma.16g150700), previously described as *E9* locus, has been associated with flowering control and soybean adaptation to different photoperiodic environments in other studies[14,33]. Previous studies proposed that *E1* acts as a repressor and has an important role in controlling photoperiodic expression patterns of *FT2a* loci[34,35]. *E2* recessive alleles could not suppress the *FT2a* loci expression, which directly impacts soybean flowering with early plants[13].

We previously found identified that the *E1* recessive allele was predominant in northern germplasms, and the *E2* recessive allele were not present in southern germplasms (MG higher than V) [30]. US founder lines with MG lower than I had a unique influence of *E2* loci on their background compared to the founder lines with MG values higher than III [32]. In Canada, soybean cultivars were concentrated on MGs lower than II. Here, the *e2* recessive allele was under selection in Guelph cultivars and fixed in Ridgetown accessions[36]. Large $F_{ST}$ values were also observed when Chinese germplasms compared to the US and Canada genetic bases[9]. Our results were associated with previous studies and suggest that these three loci play different roles in Brazil and US germplasm. One explanation for this finding might be associated with the large number of US cultivars with MG values lower than V. This increases the need for genes controlling maturity. Brazilian accessions were cultivated only in MG higher than V, which decreases the need for cultivars with recessive maturity *E* loci for adaptation in most parts of the country. This scenario is different from the US that has a large planted area in MG lower than V. However, SNPs close to FT2a loci were barely distributed in the US germplasm. These data demonstrate that maturity loci have different roles in both germplasms.

The analysis between Brazilian and US germplasm also revealed eight SNPs with high $F_{ST}$ values. Five of them were previously associated with four important soybean-group traits: yield, maturity, water-use efficiency, and shoot-nutrient concentration[22,24–26,37–39]. Interestingly, US germplasms fixed all SNPs with high $F_{ST}$ as detected in our study. These were reportedly associated with yield, maturity, and shoot-nutrient content except for ss715593829 (shoot-potassium content and water-use efficiency). This has an equal distribution of the alleles. On the contrary, the Brazilian genetic base fixed the allele T (reference allele) for ss715593829 but has an equal allele distribution for ss715588874 (seed weight), ss715613207 (seed weight and yield), and ss715624268 (maturity). Finally, we found that the alternative allele for SNP ss715624371 that is related to maturity was fixed in Brazilian accessions. Thus, the genotypic differences detected among the SNPs with high $F_{ST}$ values observed here might represent the geographical and adaptive modifications present in Brazilian and US soybean germplasms.

The US germplasm concentrated its diversity into differences among maturity loci. Our data demonstrates that *E1* has a major role in differentiating northern (00 – IV) and southern (V – IX) germplasms. Similar results were observed in a previous study [30]. We further observed that the *E2* locus has a large impact in differentiating early and super-early cultivars similar to prior work[30,32,36]. Other important loci that differentiate the US germplasm were observed in our data such as the *Dt1* locus that appears to have fixed the dominant allele in northern cultivars. Our results represent breeding efforts to improve soybean cultivars to most US regions.

Historically, the Brazilian soybean accessions have gone through several modifications. Concerning morphological traits, modern Brazilian soybeans tend to be earlier, more productive, shorter, with a low number of number of branches per plant, and lower lodging score than old cultivars[40]. Moreover, modern Brazilian cultivars remove more nutrients from the soil versus older accessions (except for calcium and sulfur). There was a meaningful impact for magnesium and nitrogen in grain nutrient concentration within a 10-year perspective. High-yield Brazilian modern cultivars could remove more potassium (21.4%) and less nitrogen (4.3%) versus older varieties[41]. Our data identified 126 genomic regions that differentiate older and modern cultivars. Similar results for regions on chromosomes 7, 17, and 18 were described previously in the Brazilian germplasm[7]. Our data also identified 409 SNPs with $F_{ST}$ values higher than 0.40 versus cultivars released before 1996 and after 1996. There were 14 SNPs previously reported in other studies that were related to maturity, seed mass, water-use efficiency, plant height, ureide content, and shoot-nutrient content (**Supplementary Table S13**)[22–26]. Four SNPs (ss715582676, ss715582689, ss715603946, and ss715603949), were putative introgressed genomic regions in modern materials. They were associated with ureide and shoot-iron content. These results are associated with other studies and indicated that modern materials incorporated nutrient absorption alleles associated with new architecture, maturity, and yield genes. In turn, these features impact modern Brazilian cultivar diversity.

There were some important gene introgressions into the Brazilian germplasm (these diseases can cause large losses). Southern stem canker was an important and historical soybean disease responsible for losses of 1.8 million metric tons in Brazil in 1994 alone[42]. A massive introgression of resistance genes to control this pathogen was necessary. We found some phenotypic results from 43 Brazilian accessions used in another study[21]. Most of the materials released after 1996 were reported to be resistant to *Diaporthe aspalathi* while there was phenotypic variation among old cultivars. We analyzed the mapping loci region associated with southern stem canker resistance[21] and observed eight SNPs with $F_{ST}$ values of 0.56, which had a perfect correlation between phenotypic and genotypic data (**Supplementary Table S15**). Moreover, ss715617869 (Chr14:1,731,256) and ss715617951 (Chr14:1,938,019) were also associated with southern stem canker in another study[20]. Our results showed that this region passed to a high contraction with a decrease in diversity in modern materials versus older materials (Figure 6). This suggests a selective sweep region that breeders incorporated into modern Brazilian seed lines.

In summary, we identified factors that differentiate the Brazil and US germplasms. Maturity loci play a more important role in the US germplasm compared to Brazil due to the large number of MGs in the US. There is a clear influence of major *E* loci on the MGs of the US germplasm. In contrast, the Brazilian genetic base appears to have more influence from the incorporation of new lines from other origins such as US and Argentina. The population structure analysis suggested a major change in Brazilian germplasms after 1996. The $F_{ST}$ demonstrated that some regions are related to adaptive, maturity, and productivity traits that might have been influenced by this change. We also observed important genomic regions that were under selection such as southern stem canker loci that demonstrate the importance of breeding programs to solve the impact of pathogens on crop productivity. Our study generated more information regarding the soybean adaptation of the world's two major soybean producers. Finally, these results offer new insights into the

genomic regions that should be the focus of breeding programs to adapt new lines and generate competitive cultivars.

# Methods

### Soybean genotypic data

This work used 230 Brazilian cultivars and 675 US cultivars from different maturity groups and release decades (**Supplementary Table S16**). These materials were previously genotyped with the SoySNP50K panel as described previously[43]. We also extracted public information from other cultivars as described [7,44–46]. The entire dataset was obtained from the Soybase website[46]. To obtain a consensus genotypic information, we only selected SNPs in SoySNP50K. The SNPs used in this study were referenced to version 2 of the soybean genome (Glyma.Wm82.a2 – Gmax2.0)[10], and only biallelic variation was maintained in the final panel. The list containing all cultivars used in this study are shown in S**upplementary Table 1**. SNPs with minor allele frequency (MAF) and call rates (CR) higher than 0.05, and 0.8, respectively, were removed.

### Population structure analysis

In the original panel, we removed SNPs with linkage disequilibrium higher than 0.30 via plink 1.09 software with the "–indep-pairwise" option.[47] This step removed the allele variation with linkage disequilibrium and used 1,798 SNPs for analysis. The structure software[48] was used to generate the analysis with a 100,000 burn-in period, and 100,000 Markov Chain Monte Carlo (MCMC) repetitions for K from 1 to 10. Ten runs were performed for each analyzed K,  and we used Structure Harvester to define the two best delta K values based on the Evanno criterion[49]. We used STRUCTURE PLOT software to generated all the structure bar plots[50]. The same SNPs were used for principal component analysis (PCA) between Brazilian and US genetic basis using TASSEL 5.0 software[51].

### Distance matrix analysis between Brazilian and US genetic bases

To compare the genetic divergence in Brazilian and US germplasms, we created an identity-by-state (IBS) genetic distance matrix using TASSEL 5.0 software[51] We removed alleles with a minor allele frequency (MAF) lower than 0.05. We separated the cultivars according to their geographic origin, maturity groups, and decade of release.

### Genetic diversity analysis

We grouped the cultivars according to their location, maturity groups, and release date. We used vcftools software for each analysis[52]. We used the population fixation index coefficient ($F_{ST}$), nucleotide diversity coefficient (π), and the Tajima's D coefficient to detect genomic regions under selection. We performed three analyses: a) Brazilian accessions vs US accessions; b) among Brazilian cultivars; and c) among US cultivars. For each analysis, we generated the $F_{ST}$ per SNP, and 100-kbp sliding window size for π, Tajima's D, and $F_{ST}$.

### Genetic annotation of the genomic regions under selection

We used SnpEff and SnpSift programs to identify the possible allelic variation observed for each SNP identified in diversity studies[53]. The SnpEff software was used for annotation of the vcf file. We used the SnpSift program with the perl script vcfEffOnePerLine.pl to generate a matrix table with one effect per line. We only considered SNPs modifications that were influenced directly in genes such as start and stop codons, splice site, and exons.

# Declarations


### Acknowledgements

We thank the Tropical Melhoramento & Genética (TMG) company for financial and material support of this study.


### Author contributions

J.V.M.S., G.C.S., and A.R.M. conceived and planned the study; J.V.M.S. and G.C.S. performed the bioinformatics analysis and data interpretation of the study; J.V.M.S. and G.C.S. wrote the manuscript; P.W.T. and M.H.T. edited and revised the intellectual content of the manuscript; and A.R.M. and A.G. led the project and revised the final manuscript. All authors read and approved the final manuscript.

### Competing interests

The author(s) declare no competing interests.

# References


1. Liu, K. S. Chemistry and nutritional value of soybean components. in *Soybeans: Chemistry, Technology, and Utilization* (ed. Liu, K. S.) 25–113 (Aspen Publishers, 1999).
2. Companhia Nacional de Abastecimento. Séries Históricas de Área Plantada, Produtividade e Produção, Relativas às Safras 1976/77 a 2019/20 de Grãos, 2001 a 2020 de Café, 2005/06 a 2019/20 de Cana-de-Açúcar. http://www.conab.gov.br/conteudos.php?a=1252& (2020).
3. Embrapa Soja. EMBRAPA SOJA. História: Histórico no Brasil. https://www.embrapa.br/en/soja/cultivos/soja1/historia (2014).
4. Hartwig, E. E. Growth and reproductive characteristics of soybeans [*Glycine max* (L.) Merr.] grown under short-day conditions. *Trop. Sci.* **12**, 47–53 (1970).
5. Gizlice, Z., Carter, T. E. & Burton, J. W. Genetic Base for North American Public Soybean Cultivars Released between 1947 and 1988. *Crop Sci.* **34**, 1143–1151 (1994).
6. Wysmierski, P. T. & Vello, N. A. The genetic base of Brazilian soybean cultivars : evolution over time and breeding implications. *Genet. Mol. Biol.* **36**, 547–555 (2013).
7. Maldonado dos Santos, J. V. *et al.* Evaluation of genetic variation among Brazilian soybean cultivars through genome resequencing. *BMC Genomics* **17**, 110 (2016).

8. Lam, H.-M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–9 (2010).
9. Zhou, Z. *et al.* Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* (2015) doi:10.1038/nbt.3096.
10. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–83 (2010).
11. Shen, Y. *et al.* De novo assembly of a Chinese soybean genome. *Sci. China Life Sci.* **61**, 871–884 (2018).
12. Valliyodan, B. *et al.* Construction and comparison of three reference-quality genome assemblies for soybean. *Plant J.* **100**, 1066–1082 (2019).
13. Watanabe, S. *et al.* A map-based cloning strategy employing a residual heterozygous line reveals that the GIGANTEA gene is involved in soybean maturity and flowering. *Genetics* **188**, 395–407 (2011).
14. Zhao, C. *et al.* A recessive allele for delayed flowering at the soybean maturity locus E9 is a leaky allele of FT2a, a FLOWERING LOCUS T ortholog. *BMC Plant Biol.* **16**, 1–15 (2016).
15. Xia, Z. *et al.* Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. *Proc. Natl. Acad. Sci.* **109**, E2155–E2164 (2012).
16. Diers, B. W., Mansur, L., Imsande, J. & Shoemaker, R. C. Mapping Phytophthora Resistance Loci in Soybean with Restriction Fragment Length Polymorphism Markers. *Crop Sci.* **32**, 377–383 (1992).
17. Ashfield, T. *et al.* Rpg1, a soybean gene effective against races of bacterial blight, maps to a cluster of previously identified disease resistance genes. *Theor. Appl. Genet.* **96**, 1013–1021 (1998).
18. Gore, M. A. *et al.* Mapping tightly linked genes controlling potyvirus infection at the Rsv1 and Rpv1 region in soybean. *Genome* **45**, 592–599 (2002).
19. Roane, C. W., Tolin, S. A. & Buss, G. R. Inheritance of reaction to two viruses in the soybean cross 'York' × 'Lee 68'. *J. Hered.* **74**, 289–291 (1993).
20. Chang, H., Lipka, A. E., Domier, L. L. & Hartman, G. L. Characterization of Disease Resistance Loci in the USDA Soybean Germplasm Collection Using Genome-Wide Association Studies. *Genet. Resist.* **106**, 1139–1151 (2016).
21. Maldonado Dos Santos, J. V. *et al.* Association mapping of a locus that confers southern stem canker resistance in soybean and SNP marker development. *BMC Genomics* **20**, (2019).
22. Dhanapal, A. P., Ray, J. D., Smith, J. R., Purcell, L. C. & Fritschi, F. B. Identification of novel genomic loci associated with soybean shoot tissue macro and micronutrient concentrations. *Plant Genome* **11**, 170066 (2018).
23. Ray, J. D. *et al.* Genome-wide association study of ureide concentration in diverse maturity group IV soybean [Glycine max (L.) Merr.] accessions. *G3 Genes, Genomes, Genet.* **5**, 2391–2403 (2015).
24. Zhang, J. *et al.* Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. *BMC Genomics* **16**, 1–11 (2015).
25. Diers, B. W. *et al.* Genetic architecture of soybean yield and agronomic traits. *G3 Genes, Genomes, Genet.* **8**, 3367–3375 (2018).
26. Kaler, A. S. *et al.* Genome-wide association mapping of carbon isotope and oxygen isotope ratios in diverse soybean genotypes. *Crop Sci.* **57**, 3085–3100 (2017).

27. Hymowiltz, T. Speciation and cytogenetics. in *Soybeans: Improvement, Production, and Uses. Soybeans: Improvement, Production, and Uses.* (eds. Boerma, H. R. & Specht, J. E.) 97–136 (American Society of Agronomy, 2004).
28. Anderson, E. J. *et al.* Soybean [Glycine max (L.) Merr.] Breeding: history, improvement, production and future opportunities. in *Advances in Plant Breeding Strategies : Legumes* (eds. Al-khayri, J. M., Mohan, S. & Dennis, J.) vol. 7 431–516 (Springer Nature, 2019).
29. Specht, J. E. *et al.* Soybean. *Yield Gains Major U.S. F. Crop.* **59901**, 311–355 (2015).
30. Wolfgang, G. & An, Y. qiang C. Genetic separation of southern and northern soybean breeding programs in North America and their associated allelic variation at four maturity loci. *Mol. Breed.* **37**, 1–9 (2017).
31. Silva, F. C. dos S. *et al.* Economic importance and evolution of breeding. in *Soybean Breeding* (eds. Silva, F. L. da, Borem, A., Sediyama, T. & Ludke, W. H.) 1–16 (2017).
32. Vaughn, J. N. & Li, Z. Genomic signatures of North American soybean improvement inform diversity enrichment strategies and clarify the impact of hybridization. *G3* **6**, 2693–2705 (2016).
33. Kong, F. *et al.* Two coordinately regulated homologs of FLOWERING LOCUS T are involved in the control of photoperiodic flowering in Soybean. *Plant Physiol.* **154**, 1220–1231 (2010).
34. Lu, S. *et al.* Natural variation at the soybean J locus improves adaptation to the tropics and enhances yield. *Nat. Genet.* **49**, 773–779 (2017).
35. Xu, M. *et al.* The soybean-specific maturity gene E1 family of floral repressors controls night-break responses through down-regulation of FLOWERING LOCUS T orthologs. *Plant Physiol.* **168**, 1735–1746 (2015).
36. Bruce, R. W., Torkamaneh, D., Grainger, C., Belzile, F. & Eskandari, M. Genome - wide genetic diversity is maintained through decades of soybean breeding in Canada. *Theor. Appl. Genet.* 3089–3100 (2019) doi:10.1007/s00122-019-03408-y.
37. Zhang, J., Song, Q., Cregan, P. B. & Liang, G. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* **129**, 117–130 (2016).
38. Mao, T. *et al.* Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. *BMC Genomics* **18**, 1–17 (2017).
39. Contreras-Soto, R. I. *et al.* A genome-wide association study for agronomic traits in soybean using SNP Markers and SNP-Based Haplotype Analysis. *PLoS One* **12**, 1–22 (2017).
40. Todeschini, M. H. *et al.* Soybean genetic progress in South Brazil: physiological, phenological and agronomic traits. *Euphytica* **215**, (2019).
41. Esper Neto, M. *et al.* Nutrient removal by grain in modern soybean varieties. *Front. Plant Sci.* **12**, 1–14 (2021).
42. Wrather, J. A. *et al.* Special report soybean disease loss estimates for the top 10 soybean producing countries in 1994. *Plant Dis.* **81**, 107–110 (1997).
43. Song, Q. *et al.* Development and evaluation of SoySNP50K a high-density genotyping array for soybean. *PLoS One* **8**, 1–12 (2013).

44. Valliyodan, B. *et al.* Landscape of genomic diversity and trait discovery in soybean. *Sci Rep* **6**, 23598 (2016).
45. Torkamaneh, D., Laroche, J., Valliyodan, B. & Donoughue, L. O. Soybean haplotype map ( GmHapMap ): a universal resource for soybean translational and functional genomics. *bioRxiv* 1–33 (2019).
46. Grant, D., Nelson, R. T., Cannon, S. B. & Shoemaker, R. C. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* **38**, 843–846 (2009).
47. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
48. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
49. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361 (2012).
50. Ramasamy, R. K., Ramasamy, S., Bindroo, B. B. & Naik, V. G. STRUCTURE PLOT: A program for drawing elegant STRUCTURE bar plots in user friendly interface. *Springerplus* **3**, 1–3 (2014).
51. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
52. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8 (2011).
53. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118 ; iso-2; iso-3. *Landes Biosci.* **6**, 80–92 (2012).

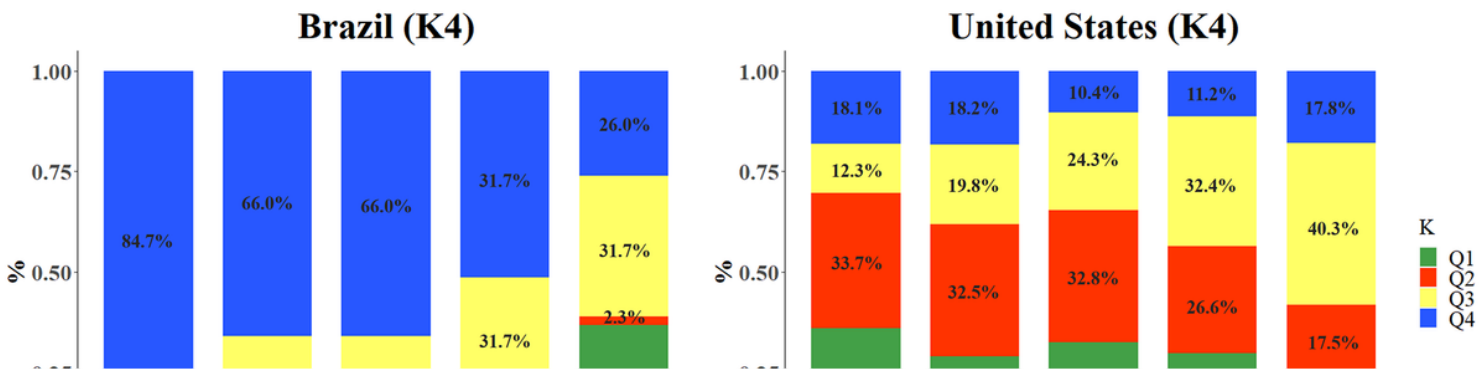
# Figures

### Figure 1

Population structure analysis between Brazilian and US germplasms. A) Principal component analysis of Brazilian and US soybean cultivars based on SNP markers; B) Delta K as a function of the number of groups (K); C) assignments coefficients of individual cultivars (bar plots) considering K = 2, and D) considering K = 4.

### Figure 2

Population structure analysis of Brazilian and US cultivars according to their maturity groups. Principal component analysis (PCA) within Brazilian (A) and US (B) germplasms for each maturity group; population structure of the Brazilian (C) and the US (D) genetic basis arranged according to their maturity groups.




## Figure 3

Mean assignment coefficients of the Brazilian and US cultivars belonging to the different release decades (1950 to 2010) and STRUCTURE groups (Q1, Q2, Q3, and Q4) considering K = 4.

## Figure 4

Population structure of Brazilian and US cultivars according to their release decade. Principal component analysis (PCA) within Brazilian (A) and US (B) germplasm for each decade; Population structure of the Brazilian (C) and the US (D) genetic bases arranged according to their release decade.

## Figure 5

The allele frequency distribution for SNPs close to loci (A) E1 (chromosome 6), (B) E2 (chromosome 10), and (C) FT2a (chromosome 16) in Brazilian and US germplasms.

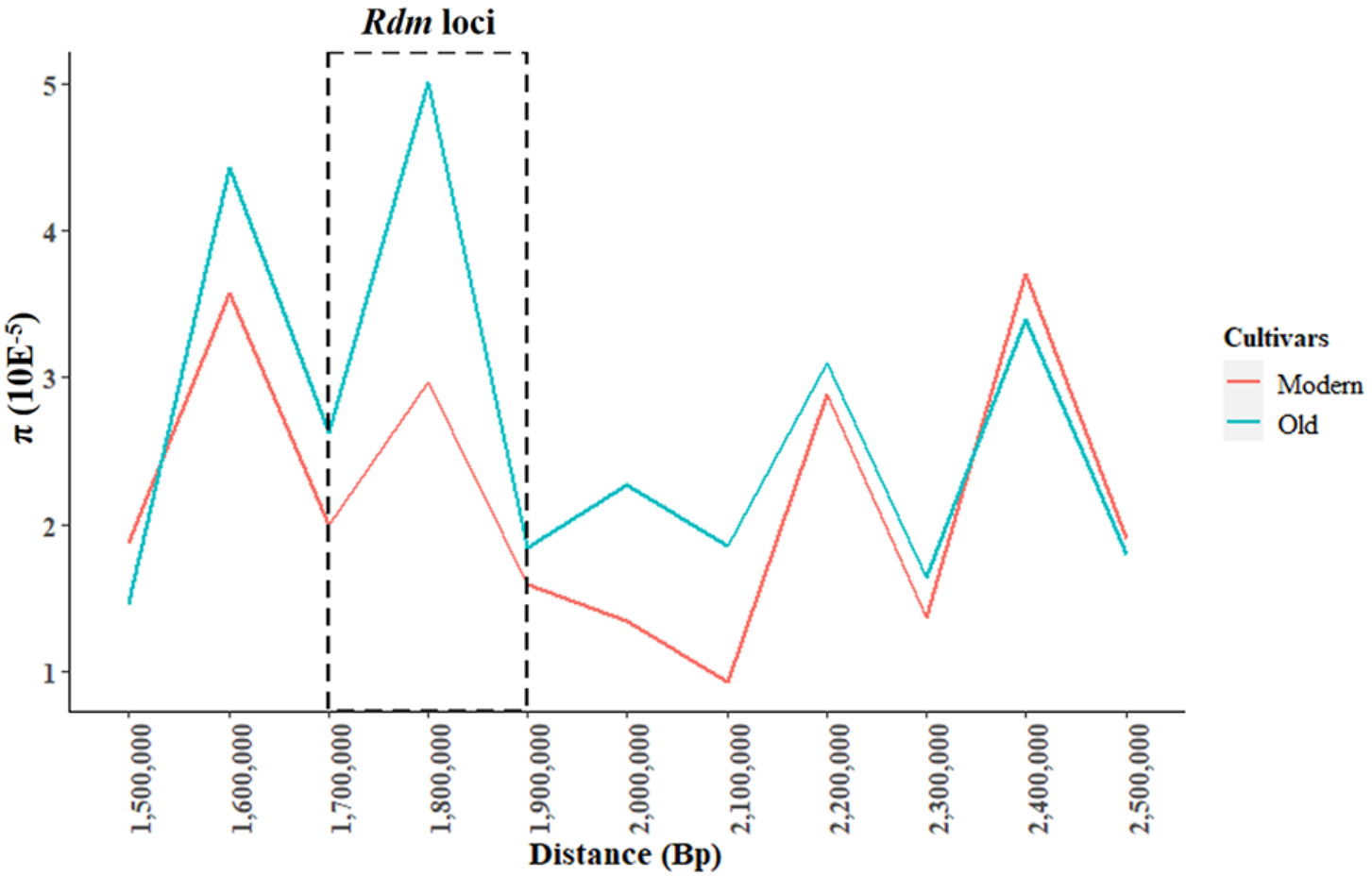



**Figure 6**

Nucleotide diversity (π) between modern and old cultivars in the southern stem canker resistance loci (Rdm).

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigureS14.pdf
- SupplementaryTableS1S18.xlsx