

# Genomic Analysis Of Spanish Wheat Landraces Reveals Their Variability And Potential For Breeding

**Laura Pascual**

Universidad Politécnica de Madrid

**Magdalena Ruiz**

Plant Genetic Resources Center. INIA

**Matilde López-Fernández**

Universidad Politécnica de Madrid

**Helena Pérez-Peña**

Universidad Politécnica de Madrid

**Elena Benavente**

Universidad Politécnica de Madrid

**José Francisco Vázquez**

Universidad Politécnica de Madrid

**Carolina Sansaloni**

Centro Internacional de Mejoramiento de Maiz y Trigo

**Patricia Giraldo** (✉ [patricia.giraldo@upm.es](mailto:patricia.giraldo@upm.es))

Universidad Politecnica de Madrid <https://orcid.org/0000-0003-4369-1078>

---

## Research article

**Keywords:** Wheat improvement, local germplasm, GBS, DArTseq markers, SNP, genetic diversity, population structure.

**Posted Date:** January 23rd, 2020

**DOI:** <https://doi.org/10.21203/rs.2.9775/v4>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published on February 4th, 2020. See the published version at <https://doi.org/10.1186/s12864-020-6536-x>.

# Abstract

**Background** One of the main goals of the plant breeding in the 21<sup>st</sup> century is the development of crop cultivars that can maintain current yields in unfavorable environments. Landraces that have been grown under varying local conditions include genetic diversity that will be essential to achieve this objective. The Center of Plant Genetic Resources of the Spanish Institute for Agriculture Research maintains a broad collection of wheat landraces. These accessions, which are locally adapted to diverse eco-climatic conditions, represent highly valuable materials for breeding. However, their efficient use requires an exhaustive genetic characterization. The overall aim of this study was to assess the diversity and population structure of a selected set of 380 Spanish landraces and 52 reference varieties of bread and durum wheat by high-throughput genotyping.

**Results** The DArTseq GBS approach generated 10K SNPs and 40K high-quality DArT markers, which were located against the currently available bread and durum wheat reference genomes. The markers with known locations were distributed across all chromosomes with relatively well-balanced genome-wide coverage. The genetic analysis showed that the Spanish wheat landraces were clustered in different groups, thus representing genetic pools providing a range of allelic variation. The subspecies had a major impact on the population structure of the durum wheat landraces, with three distinct clusters that corresponded to subsps. durum, turgidum and dicoccon being identified. The population structure of bread wheat landraces was mainly biased by geographic origin.

**Conclusions** The results showed broader genetic diversity in the landraces compared to a reference set that included commercial varieties and higher divergence between the landraces and the reference set in durum wheat than in bread wheat. The analyses revealed genomic regions whose patterns of variation were markedly different in the landraces and reference varieties, indicating loci that have been under selection during crop improvement, which could help to target breeding efforts. The results obtained from this work will provide a basis for future genome-wide association studies.

## Background

Wheat is a cereal that belongs to the Poaceae family. Wheat occupies a central place in human nutrition providing 20% of the daily protein and food calories of the human population. Currently cultivated wheat originated from natural hybridization events between different species (1). Roughly 90 to 95% of the wheat produced in the world is common, or bread wheat (*Triticum aestivum* L.;  $2n = 6x = 42$ , 17Gb, AABBDD genomes). The remainder of the world's wheat production includes about 35–40 million tons of durum wheat (*T. turgidum* var. *durum*;  $2n = 4x = 28$ , 13 Gb, AABB genomes), which is cultivated mainly in the Mediterranean region (<http://www.fao.org/faostat/en/>).

Advances in molecular biology and high-throughput genotyping technologies have significantly impacted the field of molecular plant breeding, leading to shift from a phenotype-based to a genotype-based selection (2). The integrated use of genomic and molecular tools in conventional phenotype selection

programs has allowed the development of new breeding strategies such as marker assisted selection (MAS) and genomic selection. However, in wheat, the large complex genome with an over 85% repetitive DNA content has hampered the application of these molecular breeding approaches compared to their use in other crops, as the presence of two or three separate but closely related subgenomes hinders the analysis of homoeologous gene sequences (3). The recently published durum and bread wheat reference genomes (4,5) provide high-quality data that will help to physically locate thousands of scattered molecular markers, thus facilitating the identification of key genes by genome-wide association studies (GWAS) that will be highly valuable for MAS in wheat breeding programs (6).

The successful genomics-assisted breeding of any crop will be enhanced by a thorough understanding of the species' genetic diversity. As is the case in other crops, genetic diversity of wheat has declined as a consequence of bottlenecks encountered during polyploidization and domestication (7,8). Modern plant breeding practices, in which only a small number of elite cultivars are included in crossing programs, have further narrowed the genetic base of wheat throughout the world, limiting the pool of alleles in which to search for new traits of agronomic interest. This has promoted wide crossing programs carried out since the 1980s at different centers of wheat research such as CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo). Indeed, by 1990, CIMMYT breeders began to successfully increase wheat productivity and genetic diversity through the introgression of various novel wheat materials. However, the genetic diversity represented by current wheat cultivars needs to be further increased to face novel threats, such as climate change, which demands an enlarged pool of alleles. Fortunately, an enormous number of genetically different, locally well-adapted wheat landraces were generated through natural or farmer-mediated selection in the previous century. Because future gains in yield potential will surely require the exploitation of these largely untapped sources of genetic diversity (9), deep knowledge of their genetic/genomic diversity is highly valuable to address the forthcoming plant breeding challenges (10).

A large number of studies have been performed to estimate genetic diversity by employing different methodologies in diverse plant species (11,12), including wheat (13). It is accepted that molecular markers are the best option for genetic variation studies. Among these markers, single nucleotide polymorphisms (SNPs), whose detection has been enormously facilitated by high-throughput technologies such as SNP arrays (14) or genotyping-by-sequencing (GBS) (15), are the most frequently used for genome-wide diversity studies.

The assessment of genome-wide diversity by GBS provides robust estimates of diversity and has been increasingly adopted as a fast, high-throughput cost-effective tool for whole-genome genetic diversity analysis in large germplasm sets (16). The DArTseq (Diversity Array Technology sequence) markers, based on GBS (17), efficiently target low-copy-number sequences via a complexity reduction method and have been successfully applied for genetic diversity studies in different species (18-21). Moreover, DArTseq provides data at an affordable cost, especially in complex polyploid species such as wheat (<https://www.diversityarrays.com>), where it has been extensively used (20,22,23). It is indeed the method employed by CIMMYT to build the most comprehensive genotype datasets for genetic resources in wheat (<https://seedsofdiscovery.org/about/genotyping-platform/>).

High-throughput genotyping also provides essential information for the design of high-power GWAS, which enable the identification of agriculturally important genes and facilitate their transfer from wild or local germplasm into modern cultivars through marker-assisted selection and marker-assisted breeding and/or genomic selection. For GWAS analysis, the optimum diverse panel must be genotyped with a set of molecular markers covering as much of the genome of the species as possible (24), but the population structure needs to be investigated to avoid false associations between phenotypes and markers (25).

The Spanish wheat landraces conserved at the National Plant Genetic Resources Center (CRF-INIA) and maintained in the national collection were collected in the first half of the 20<sup>th</sup> century. Several studies have shown the great variability of the Spanish durum wheat accessions compared to other germplasm collections (26-29). However, no genetic description of the bread wheat landraces has been reported, and the high-throughput genomic characterization of the durum wheat landraces remains to be fully realized.

The aim of the present study was to characterize two collections of durum and bread wheat landraces from CRF-INIA by using the DArTseq-GBS approach. The specific objectives of the present investigation were: (1) to assess the genomic diversity of a set of durum wheat accessions comprising 191 Spanish landraces and 23 reference varieties, (2) to assess the genomic diversity of a set of bread wheat accessions comprising 189 Spanish landraces and 29 reference varieties, and (3) to compare the genetic diversity of landraces and modern cultivars in both wheat species.

## Results

### Wheat genotyping

We characterized a set of 380 landraces and 52 reference varieties at genomic level (Additional file 1). The DArTseq approach allowed us to detect approximately 100 K DArTs (presence/absence markers) and 50K SNPs in each analyzed species.

In tetraploid durum wheat (214 accessions), a total of 98,983 DArTseq markers and 51,751 SNP markers were obtained (Additional files 2 and 3). When the markers were located in the *T. turgidum* reference genome, they were distributed throughout the genome (Table 1). According to the raw data, approximately 58% of the DArTs, and 37% of the SNPs were not located in the *T. turgidum* reference genome. After filtering to obtain highly informative markers, 38,700 DArTs and 9,324 SNPs were selected for further analysis. In this set of markers, the percentage of located markers was similar 66% in SNPs and 45% in DArTs). As shown in Table 1, the filters applied did not affect the marker distribution within the genome. The A and B genomes presented a comparable number of markers, both before and after filtering. Chromosome 4B exhibited the lowest density of both types of markers.

For hexaploid bread wheat (218 accessions), a slightly higher number of markers, including 130,899 DArTseq markers and 58,660 SNPs, was generated (Additional files 4 and 5). As in durum wheat, the markers were detected throughout the whole genome; around 64% of raw DArTs and 41% of raw SNPs were not located in the bread wheat reference genome, a percentage similar to that of the durum wheat

(Table 1). After filtering, 44,241 DArTs and 8,238 SNPs were selected for further analysis. In this set of markers, the percentage of located markers was similar to the percentage obtained in the raw data (36% in DArTs and 57% in SNPs). The D genome presented a reduced amount of markers compared to A and B genomes. Regarding these latter, chromosome 4B exhibited again the lowest density of both types of markers.

**Table 1.** Numbers of SNP and DArT markers identified in *T. turgidum* and *T. aestivum* accessions. The total numbers of markers before and after filtering, and their distribution within the genomes and chromosomes are presented. NA, no data available, as D genome is not present in *T. turgidum*

	<i>Triticum turgidum</i>				<i>Triticum aestivum</i>			
	Raw DArTs	Filtered DArTs	Raw SNPs	Filtered SNPs	Raw DArTs	Filtered DArTs	Raw SNPs	Filtered SNPs
<b>Total</b>	98983	38700	51751	9324	130899	44241	58660	8238
<b>Located</b>	41429	17442	32811	6192	46665	16090	34497	4738
<b>Not located</b>	57554	21258	18940	3132	84234	28151	24163	3500
<b>Genome A</b>	19307	7907	15719	2957	16127	5957	12762	1958
<b>Genome B</b>	22122	9535	17092	3235	17754	7000	13636	1963
<b>Genome D</b>	NA	NA	NA	NA	12784	3133	8099	817
<b>1A</b>	2047	863	1758	378	1841	723	1594	285
<b>1B</b>	2972	1311	2341	505	2431	952	1888	278
<b>1D</b>	NA	NA	NA	NA	1797	384	1046	101
<b>2A</b>	3079	1173	2581	519	2549	891	2072	302
<b>2B</b>	4017	1661	3117	553	3188	1272	2500	345
<b>2D</b>	NA	NA	NA	NA	2501	773	1627	156
<b>3A</b>	2757	1057	2252	409	2258	687	1923	316
<b>3B</b>	3604	1570	2795	527	2763	1076	2256	331
<b>3D</b>	NA	NA	NA	NA	1877	406	1279	106
<b>4A</b>	2587	1050	1805	295	2233	877	1453	167
<b>4B</b>	1763	743	1328	295	1492	491	1028	145
<b>4D</b>	NA	NA	NA	NA	1033	170	515	71
<b>5A</b>	2603	998	2224	486	2135	701	1749	330
<b>5B</b>	3054	1197	2332	462	2572	1007	1948	299
<b>5D</b>	NA	NA	NA	NA	1731	383	1068	119
<b>6A</b>	2364	1027	1819	306	1909	777	1385	203
<b>6B</b>	3360	1541	2567	411	2467	1027	1970	283
<b>6D</b>	NA	NA	NA	NA	1536	432	1013	118
<b>7A</b>	3870	1739	3280	564	3202	1301	2586	355
<b>7B</b>	3352	1512	2612	482	2841	1175	2046	282
<b>7D</b>	NA	NA	NA	NA	2309	585	1551	146

The marker distribution along the chromosomes was comparable in bread and durum wheat for all the A and B genome chromosomes. In general, a higher density of markers was found at both chromosome

ends, as illustrated for chromosome 2A in Figure 1. The same pattern was observed for the bread wheat D genome chromosomes.

In both species, the distribution of PIC (polymorphic index content) values for the DArT and SNP data was asymmetrical and skewed towards the lower values (Figure 2). In durum wheat, 82% of the DArTs and 75% of the SNP markers showed a PIC value >0.2. In bread wheat, the corresponding values were 76% and 70% for DArTs and SNPs, respectively. For both species and types of markers, the average PIC values were between 0.30 and 0.35.

### Genetic structure of the durum wheat collection

fastSTRUCTURE runs with 38,700 DArT markers divided the tetraploid wheat landraces into seven populations (K=7) (Figure 3A). All but one (BGE021775) of the 14 accessions belonging to subsp. *dicoccon* were grouped in Pop5, and all 37 of the subsp. *turgidum* accessions were clustered in Pop3. The landraces in both populations came mostly from the north of Spain (Figure 3B). All of the 140 subsp. *durum* landraces except for one (BGE013103), which was classified into Pop3, were distributed among five populations (Pop1, Pop2, Pop4, Pop6 and Pop7), containing between 10 and 80 accessions (Additional file 1). Pop6 exhibited the highest number of accessions, showed the greatest degree of admixture and was the population with the most diverse eco-geographical origin (Figure 3). However, some subsp. *durum* populations showed a narrower geographic distribution (Additional file 1). That is, the landraces in Pop1 originated mostly from eastern Spain, whereas those in Pop2 came from the southwestern provinces. Pop4 included landraces from the South and East of Spain, and from the Canary Islands.

Genetic diversity parameters were calculated for the fastSTRUCTURE populations based on the SNP data (Table 2). The population with the highest genetic diversity value (Hs, Nei's diversity index) was Pop6 (0.272), and the population with the lowest value was Pop2 (0.048). For the whole landrace collection, the  $D_{est}$  (Population differentiation index) value, a measure of population differentiation in collections with several populations, was 0.22. The  $F_{ST}$  values, which are related to genetic differentiation between populations, ranged from 0.743 (Pop2 vs Pop5) to 0.226 (Pop1 vs Pop6). Pop6 showed the least genetic differentiation from the rest of the populations, including those of subsp. *turgidum* (Pop3) and *dicoccon* (Pop5) (Table 2). When the diversity between the three subspecies was estimated, regardless of the structured populations, the *dicoccon* and *durum* landraces showed the highest value of genetic differentiation between the subspecies ( $F_{ST} = 0.42$ ) whereas subsp. *turgidum* showed lower values compared to either *durum* and *dicoccon* ( $F_{ST} = 0.31$  and 0.38, respectively).

**Table 2.** Genetic diversity within populations (Hs) and  $F_{ST}$  values between populations of *T. turgidum* landraces assessed with SNPs.

<b>Hs</b>	0.186	0.048	0.253	0.089	0.169	0.272	0.104
<b>F<sub>ST</sub></b>	<b>Pop1</b>	<b>Pop2</b>	<b>Pop3</b>	<b>Pop4</b>	<b>Pop5</b>	<b>Pop6</b>	<b>Pop7</b>
<b>Pop7</b>	0.476	0.709	0.474	0.653	0.692	0.238	-
<b>Pop6</b>	0.226	0.297	0.311	0.264	0.452		
<b>Pop5</b>	0.573	0.743	0.393	0.725			
<b>Pop4</b>	0.546	0.730	0.532				
<b>Pop3</b>	0.371	0.526					
<b>Pop2</b>	0.575						

When we analyzed the distribution of Hs values across the genome, we detected some genomic diversity patterns that were population- specific (Additional file 6). For example, Pop2 and Pop7 presented a region of low diversity in the central part of chromosome 2A, while for chromosome 2B we only detected a region of low diversity in Pop5. On the other hand, a similar analysis contrasting the Hs values across the genome in the three durum wheat subspecies showed some common low diversity regions in *turgidum* and *dicoccon* (e.g., chromosomes 1A and 2A), while *durum* showed higher diversity values across the genome (Additional file 7).

We also explored the genomic structure of the durum wheat collection, including the landraces and reference varieties, through a principal coordinate analysis (PCoA) based on the 9,324 filtered SNP markers. The first two principal coordinates explained 21% of the total variation. Three discrete groups corresponding to the three subspecies could be clearly identified (Figure 4A). This was in agreement with the results obtained with fastSTRUCTURE where the three subspecies were grouped into different populations. A fourth group corresponding to the reference varieties also appeared to be clearly separated from the subsp. *durum* landraces. The subsp. *durum* accessions were differentiated from the others by PCo1, but the difference between *turgidum* and *dicoccon* was due to PCo2, demonstrating that different sets of markers are responsible of the genetic divergence among subspecies, as detected in the Hs analysis. Some landraces of subsp. *durum* were located close to subsp. *dicoccon* and *turgidum*. These landraces were from Pop6 and some of them (e.g., BGE019290) come from the North of Spain (Additional file 1).

We further investigated the allelic variability of a functional marker involved in wheat adaptability, the vernalization gene *Vrn-A1*, in relation to the population structure. Three different alleles were identified in the collection: the winter-type allele *vrn-A1*, and two alleles related to the spring growth habit, *Vrn-A1b* and *Vrn-A1c*. The representation of allelic variation in the PCoA showed that most of the accessions carrying the winter-type allele, *vrn-A1*, were grouped together and corresponded to *dicoccon* accessions (Figure 4A). Most of the reference cultivars and subsp. *durum* accessions carried the *Vrn-A1c* allele, and almost all of the subsp. *turgidum* accessions carried the *Vrn-A1b* allele. When analyzed within the population structure, all but one *durum* accession from Pop2, Pop7, and Pop1 presented the *Vrn-A1c* allele, which was also identified in the 80% of the durum wheat landraces clustered in Pop 6. In Pop3 and Pop4, the most frequent allele was *Vrn-A1b*. According to passport data (see Additional file 1), the accessions with the winter-type allele *vrn-A1* came mostly from the North of Spain.

Allelic variation was also studied for the HMW-GS (High Molecular Weight Glutenin Subunits) loci *Glu-A1* and *Glu-B1*, but no relationship with the population structure could be observed (data not shown).

As the subspecies was the main discriminant factor in the global PCoA, we decided to perform the analysis excluding the *turgidum* and *dicoccon* accessions to gain insight into the variability within subsp. *durum* (Figure 4B). The populations identified in the previous analysis with DArTs (Figure 3A) were similarly grouped in the SNP-based PCoA, with Pop6 again being the population showing the greatest dispersion due to its higher intrapopulation variability (Figure 3A). The only *durum* accession clustered in Pop3 (BGE013103) appeared to be located close to the Pop1 landraces in this case (Figure 4B). This local variety can be identified in Figure 3A at the edge of Pop3, showing admixture with Pop1, which suggests that it could present a hybrid genotype between *durum* and *turgidum*.

Pop1, Pop2, and Pop7 were the *durum* populations that were most differentiated from the reference set. On the other hand, Pop4 was closest to the reference group. This population included old local varieties such as 'Ledesma' and 'Lebrija', obtained from crosses between 'Senatore Capelli' and Spanish landraces. One landrace from Pop6 of subsp. *durum* (BGE026954) was grouped together with the reference varieties. This accession, collected at the end of the 1990, is characterized by early-maturity and short plants (22), which suggests that it is probably not a true landrace.

### Genetic structure of the bread wheat collection

fastSTRUCTURE runs with 44,241 DArT markers divided the hexaploid wheat landrace accessions into four populations (K=4). Compared to durum wheat landraces, a higher level of admixture was detected in the bread wheat populations, especially within Pop2, which was the largest population, containing 112 accessions (Figure 5A, Additional file 1). The landraces from Pop1 came from central Spain, and the landraces from Pop4 came from the west, including the Canary Islands. Pop2 and Pop3 showed more diverse eco-geographical origins (Figure 5B, Additional file 1).

As shown in Table 3, the population with the highest Hs was Pop2 (0.277), and the population with the lowest value was Pop3 (0.101). In the whole landrace collection, the  $D_{est}$  value was 0.17, which was lower than the differentiation found in the durum collection (0.22). The  $F_{ST}$  values between populations ranged from 0.169 (Pop1 vs Pop2) to 0.573 (Pop3 vs Pop4). According to the  $F_{ST}$  values, Pop4 was the most differentiated population (Table 3). Regarding the Hs distribution across the genome, we detected some low-diversity regions specific to certain populations (Additional file 8). For instance, Pop3 and Pop4 showed low-diversity regions on chromosomes 1A and 7A, while Pop1 and Pop3 showed low diversity on chromosomes 3A and 2B. This suggests that different genomic regions are responsible for the divergence among populations.

**Table 3.** Genetic diversity within populations (Hs) and  $F_{ST}$  values between populations of *T. aestivum* landraces assessed with SNPs.

<b>Hs</b>	0.176	0.277	0.101	0.188
<b>F<sub>ST</sub></b>	<b>Pop1</b>	<b>Pop2</b>	<b>Pop3</b>	<b>Pop4</b>
<b>Pop4</b>	0.506	0.346	0.573	-
<b>Pop3</b>	0.483	0.265		
<b>Pop2</b>	0.169			

The relationships among the bread wheat accessions were also assessed by PCoA, based on 8,238 SNPs in the whole bread wheat collection. The total amount of genetic variation explained by the first two principal coordinates was 19.2%. The first two coordinates clearly separated the Pop4 (by PCo1), which formed the most distant group, from the other three populations, which appeared to be distributed along PCo2 (Figure 6A). Some degree of overlap was shown between Pop2, distributed along PCo2, and both Pop1 and Pop3, located at the upper extreme of PCo2. These results are in agreement with the higher degree of admixture in Pop2 revealed by fastSTRUCTURE analysis (Figure 5A). The reference varieties were located within a quite limited space but overlapped with some Pop2 landraces.

As in durum wheat, we investigated the allelic variability of the *Vrn-A1* gene in relation to the bread wheat population structure. Three different alleles were identified in the collection: the winter-type allele *vrn-A1* and two alleles that are generally related to a spring growth habit, *Vrn-A1a* and *Vrn-A1b*. The accessions from Pop4 were characterized by almost exclusively presenting the *Vrn-A1b* allele. The accessions presenting the other two alleles could be differentiated along the PCo2 axis, whose upper portion corresponded to spring-type landraces, whereas the reference varieties and winter-type landraces were included in the lower portion (Figure 6A).

The representation of *Glu-B1* alleles specific to Iberian landraces in the PCoA showed that the accessions that carried the *Glu-B1f* allele (HMW-GS 13+16) were clustered on the right side of the PCoA and corresponded to most of the Pop4 samples. Again, an interesting tendency could be observed along the PCo2 axis: accessions with the *Glu-B1e* allele (HMW-GS 20x+20y) appeared to be grouped in the upper portion, while the lower portion corresponded to accessions with *Glu-B1u* or *Glu-B1a1* alleles (HMW-GS 7+8 or 70E+8) (Figure 6B). The allelic variation was also studied for HMW-GS loci *Glu-A1* and *Glu-D1* but no relationship with the population structure could be observed (data not shown).

The overlap between the landraces and reference varieties was more remarkable in bread wheat than in durum wheat. In fact, the reference varieties fully overlapped with the Pop2 accessions. Some of the landraces located closer to the reference varieties were collected in the 1990s (e.g., BGE025410), and according to their early flowering and lower height phenotypes ([http://webx.inia.es/web\\_coleccionescrf/CaracterizacionCRFeng.asp](http://webx.inia.es/web_coleccionescrf/CaracterizacionCRFeng.asp)), they may not be real landraces.

### **Divergence between landraces and reference varieties**

As noted above, PCoA showed clear genetic divergence between the landraces and reference varieties of durum wheat, whereas such divergence was not as evident in bread wheat, in which the landraces and

reference varieties overlapped (Figures 4 and 6). Regarding the overall genetic diversity, the landraces showed higher diversity than the reference varieties in both species (Table 4), with the difference being greater in durum wheat.

The divergence between the reference varieties and landraces varied among the different populations identified by fastSTRUCTURE. The durum wheat reference varieties showed the highest differentiation from the landraces in Pop5 (*dicoccon*) and Pop2 (*durum*), and the lowest differentiation from the *durum* landraces in Pop4 and Pop6 ( $F_{ST}$  values of 0.586, 0.555, 0.372 and 0.214, respectively). In bread wheat, the reference varieties showed the highest and lowest differentiation from the landraces in Pop4 and Pop2, respectively ( $F_{ST}$  values of 0.416 and 0.08, respectively).

To analyze the degree of allele fixation in the reference varieties, we studied the presence of monomorphic markers. Approximately, 40% of the SNP markers were fixed in the durum wheat reference varieties, and the number of monomorphic markers (3,791) was comparable to that found in subsp. *dicoccon* and *turgidum* (4,079 and 2,119 respectively) and much higher than that found in subsp. *durum* (478 fixed markers). In the bread wheat reference accessions, the number of fixed SNP markers was 1,771 (21%) clearly lower than the number obtained in durum wheat (Table 4).

**Table 4.** Overall genetic diversity ( $H_s$ ) and number of monomorphic SNP markers in the set of reference varieties compared to landraces.

	<i>Triticum turgidum</i>		<i>Triticum aestivum</i>	
	Reference	Landraces	Reference	Landraces
<b><math>H_s</math></b>	0.196	0.323	0.250	0.300
<b>Monomorphic markers</b>	3791	24	1771	2

We also analyzed the  $H_s$  distribution across the genome in the landraces and reference varieties (Figure 7). We called low-diversity regions in the reference varieties with respect to the landraces, as they might be regions fixed by breeding efforts. For this particular analysis, we employed a different set of markers, as described in material and methods section, which include only SNPs located in the reference genomes. In durum wheat, we detected 20 of these genomic regions (Figure 7A). Chromosomes 1A, 1B, 2B, 4B, 5A and 6B did not present low-diversity regions, while the rest of chromosomes presented at least one fixed region. The largest wheat chromosome (3B) presented the greatest number of low diversity regions, with 5 regions spanning a total of 39 Mb. In turn, chromosome 2A presented the widest region of low diversity (37Mb). In summary, the identified low-diversity regions spanned 218 MB. In bread wheat, a similar number (24) of lower-diversity regions were called (Figure 7B). Interestingly, the D genome included only two low-diversity regions. In the A genome, we detected 12 such regions located on chromosomes 1A, 2A, 3A, 4A and 7A, while in the B genome, we observed low-diversity regions on all the chromosomes except for 3B. The widest fixed region was located on chromosome 2B, spanning 121Mb, and chromosome 2A contained the greatest number of regions, with 4 regions. In bread wheat, the low-diversity regions

spanned a total of 601 MB and included the semidwarfing gene *Rht-D1*, a key gene introduced in wheat cultivars during the 20th that is located on chromosome 4D, and *Vrn-B1*, located on 5B and related with growth habit.

## Discussion

Modern bread and durum wheat varieties have resulted from breeding programs that have mainly been focused on achieving high yields in conventional agricultural systems. The unpredictable effects on crop performance under future climate scenarios and the increasing concern for the environmental consequences of crop management practices, which are commonly based on excessive irrigation and fertilizer application, are reorienting breeders' aims toward the development of new wheat cultivars that can maintain high yields under uncertain climatic conditions and in a more sustainable manner. For this purpose, genetic resources maintained in wheat genebanks, which have generally been underused, need to be thoroughly characterized so that the useful genetic variation that is present is transferred to modern elite gene pools. Several recent international initiatives are characterizing local wheat germplasm collections ([www.seedsofdiscovery.org](http://www.seedsofdiscovery.org); [www.divsek.org](http://www.divsek.org)). However, Spanish landraces, which are locally adapted to a wide diversity of eco-climatic conditions and mostly cultivated under rainfed low-input management systems, are absent or only represented by a small number of accessions in these studies. We decided to explore the genetic diversity of these valuable materials, by performing an analysis at the genomic level of 380 selected landraces representing the genetic variability of the wheat collection maintained at the Spanish Plant Genetic Resources Center, which is composed of more than 1,600 accessions.

### Wheat high-throughput genotyping

This is the first report of the high-throughput genotyping by GBS of Spanish wheat germplasm. The GBS method has the potential to provide robust in-depth genomic diversity estimates. Moreover, this approach may reveal new alleles that might present high value for prebreeding (30-32). The DArTseq-GBS methodology has been successfully applied in wheat species, in which a standard GBS approach, requiring genome resequencing, is still challenging due to the extremely large complex genome (10).

In our study, 50K SNP and 100K DArT markers were analyzed in each of the wheat species, from which we were able to select approximately 10K SNP and 40K DArT high quality markers. The availability of the reference bread wheat genome (4) and durum wheat genome (5) allowed the marker location to be performed in both species. The results showed that markers were distributed throughout the genome, although D genome markers were markedly less abundant than A and B genome markers in bread wheat. Several previous studies have shown the reduced diversity of the wheat D genome, which has been explained by the close genetic distance between the *Ae. tauschii* parents involved in the formation of hexaploid wheat (e.g. (33,34)). Among homoeologous sets, the group 4 chromosomes were the least covered, especially 4B and 4D. No satisfactory hypothesis has yet been proposed to explain the relative

scarcity of markers consistently mapped on 4B compared to 4A (35,36), despite the latter being involved in several structural rearrangements in wheat (37).

The informativeness of the markers was assessed from their PIC values, and the distribution and average PIC value were found to be comparable to those previously reported in wheat (20,38-41).

### **Population structure of Spanish wheat landraces**

When a large germplasm collection is used for the identification of useful alleles or traits, knowledge of its genetic structure is highly recommended to optimize the search approach, which can then be focused on a smaller sample. The study of population structure is also important for genomic studies, as it is a mandatory prerequisite for successfully performing further GWAS analyses. In our case, since future GWAS analyses will be based on the high-quality SNPs identified herein, we decided to assess the structure with a different set of markers, DArTs. Population structure was assessed with the fastSTRUCTURE algorithm, which was developed by the authors of the classical STRUCTURE software but shows faster runs and provides comparable ancestry estimates and prediction accuracies (42). In both species, the analysis of population structure was complemented with PCoA performed with the SNP dataset. As expected, the results were in close agreement, but some additional information about the subgroups and their relationship to phenotypic traits could be extracted.

#### Durum wheat landraces

The Spanish durum wheat (*Triticum turgidum* L.) landraces belong to three main interfertile subspecies (*dicoccon*, *turgidum* and *durum*). The subsp. *dicoccon*, also known as emmer wheat, is a hulled wheat that is only grown in the North of Spain and represents the feral situation of this crop. Subsp. *durum* is the most widely cultivated and well adapted to the dry-summer conditions of the South. Subsp. *turgidum*, which is less common and grown in colder areas than *durum*, mostly consists of winter wheat (43). In a previous study, SSRs (Single Sequence Repeats) were used to assess the genetic structure of the collection of durum wheat landraces analyzed here and 9 populations were established. Some of the populations included more than one subspecies and several genotypes could not be classified into any population (27). In the present work, we identified seven populations, all of which were composed of landraces from one subspecies with only two exceptions (Figure 3). The discrepancies between this and the earlier work can be explained by the different types and numbers of markers employed for assessing the population structure in both studies (40K DArTs vs 39 SSRs).

Analyses comprising other durum wheat materials have not been able to separate the *durum* and *turgidum* subspecies (44,45). However, we clearly differentiated the three subspecies. This demonstrates the analytical power of DArT and SNP markers for taxonomical identification and supports the classification of Mac Key (46,47), with *turgidum* and *durum* as separate subtaxa. In our study, subsp. *dicoccon* was closer to *turgidum* than to *durum* (see  $F_{ST}$  values, Table 2), indicating that the two subspecies share a common allele pool. As the two subspecies were grown in similar environmental conditions, this closeness might be due to similar selective pressures during local adaptation. The

analysis also allowed the identification of accessions that could represent admixtures between subspecies, such as BGE021775, a *dicoccon* landrace located closer to *durum*, or BGE04564, a *durum* landrace grouped with the *turgidum* landraces. It has been reported that such admixture is not unusual in ancient local forms of durum wheat (47).

Vernalization genes are the main determinants of the growth habit (i.e., winter or spring) in temperate cereals, and by affecting the vegetative to reproductive transition, these genes are involved in the ability of wheat plants to adapt to a wide range of environments (48). The evolution of spring-habit cultivars from winter-habit accessions played a key role in the postdomestication spread of wheat. However, studies on the major vernalization gene *VRN1* have been mostly limited to hexaploid wheat species and very few reports from tetraploid species can be found in the literature (48,49). None of the durum wheat accessions characterized here showed the *Vrn-A1a* allele described in spring-habit hexaploid wheat varieties, which is in accordance with what has been found in other studies (50,51). All of the available data seem to indicate that this allele appeared during wheat evolution after the last polyploidization event. The *Vrn-A1c* allele has been described as the most frequent determinant of a spring habit in tetraploid wheat varieties (50,52,53) but has been described as rare in emmer wheat varieties (44). In our study, 78% of the *durum* accessions presented this allele, but almost all *turgidum* accessions (83%) carried *Vrn-A1b*, and 7 out of the 11 *dicoccon* landraces characterized for *Vrn-A1* carried the winter *vrn-A1* allele (Figure 4). The presence of the spring-habit associated alleles *Vrn-A1b* and *Vrn-A1c* in the remaining four *dicoccon* accessions is remarkable. Emmer wheat varieties are traditionally cultivated in cool mountainous regions, where vernalization seems to be an unavoidable requirement. It may explain why few Spanish spring emmer wheat varieties have actually been described (43). The spring type might have evolved from previous winter types as an adaptation to warmer conditions. Under the predicted climate change scenario, temperature warming may prevent the fulfilment of the requirements for vernalization in current temperate zones, thus having a negative global impact on winter wheat yields. The identification of genotypes with reduced vernalization requirements among germplasm adapted to cool zones could therefore be relevant for improving adaptability to changing eco-climatic conditions.

Within our collection, a restricted geographic distribution exists for the Spanish landraces belonging to the less represented subspecies *dicoccon* and, to some extent, for *turgidum* landraces. The clustering of these two groups of accessions, each in a single population, may reflect similar environmental conditions in their respective geographic origins. The landraces of subsp. *durum*, which were structured into five distinct populations, showed higher variability and greater complexity, including different phylogenetic groups (27). The identified clusters seemed to be influenced by the accessions' origin relatively little, although some geographic areas were predominant in some of the populations. The gene flow between regions via germplasm exchanges and local preferences towards a given agrotypic might be as significant as ecological conditions in determining the distribution of genetic diversity in this subspecies in Spain. The exchange of seeds by farmers has been noted as one of the likely explanations for the low or absent influence of geographic origin on the genetic structure of durum wheat landraces in Iran, the Central Fertile Crescent and Ethiopia (20, 54, 55).

## Bread wheat landraces

The great majority of the Spanish bread wheat landraces conserved at CRF-INIA belong to *Triticum aestivum* subsp. *vulgare*, and our bread wheat set was therefore composed exclusively of this subspecies. The population stratification of the bread wheat panel identified four groups of landraces with high divergence according to the obtained  $F_{ST}$  values. This clustering reflected the geographic origin of the accessions better than in the subsp. *durum*. The genetic differentiation estimated with  $D_{est}$  was lower in the bread landrace collection than in the whole durum wheat collection, in agreement with the lower level of stratification observed. Few studies have simultaneously addressed the variability of hexaploid and tetraploid wheat varieties but higher genetic diversity in durum than in bread wheat has been previously reported in landraces from other countries (e.g., (56)). Several studies support the occurrence of a limited number of independent crosses between the diploid and tetraploid progenitors of *T. aestivum*, where the resulting loss of diversity during the initial polyploidization step presumably caused a severe population bottleneck in hexaploid bread wheat (33,57).

One of the four groups detected (Pop4) was clearly more genetically distant. This group included landraces from western Spain, where there is a prevalence of acidic or neutral soils (58). Most of the accessions from this population show spring growth habit and carry the *f* allele at the *Glu-B1* locus. The *Glu-B1f* allele presents a low frequency in worldwide collections but has been previously described as being characteristic of Iberian landraces, which is also the case for the *Glu-B1e* allele (59). The latter was predominant in Pop1 and Pop3, and was also present in some landraces of Pop2 that were closely grouped by PCoA. The *Glu-B1e* allele is related to poor rheological properties in bread wheat, but the *f* allele has been associated with good dough quality (60). The presence of this variant in a discrete group of more differentiated landraces supports their common origin.

In our study, the winter-type allele *vrn-A1* was the most common *Vrn-A1* allele found in the *T. aestivum* reference cultivars (22 out of 29 varieties). However, the most frequent of these alleles within the bread wheat landraces was the *Vrn-A1a* allele, which was absent in the durum wheat. This *Vrn-A1* allele, which results in complete insensitivity to vernalization, is recognized as the spring-habit allele with the greatest effect among all such alleles described (61). Even if the growth habit shows just a limited impact on genetic differentiation, spring and winter bread wheat accessions are frequently separated by discriminant analysis of principal components (62,63). Two discrete groups regarding growth habit were not clearly defined in our PCoA, but the allelic variability of *Vrn-A1* showed some relationship to population structure and a somewhat biased tendency along the second PCo axis (Figure 6A). Regarding this matter, it must be kept in mind that the vernalization response is a complex process under polygenic control (64). *Vrn-A1* has been described as the main genotypic determinant of the vernalization requirements of temperate crops, but there are other genes, such as *Vrn-B1* and *Vrn-D1*, whose allelic variability has not yet been characterized in these Spanish wheat landrace collections.

This is the first time that the bread wheat resources maintained in the CRF-INIA Spanish national genebank have been characterized at the genetic level and compared to durum wheat accessions. This

deep knowledge of their genetic structure represents the starting point for the development of a core collection of Spanish *T. aestivum* wheat varieties, which will allow the efficient management and use of this valuable gene pool.

### **Relationship between landraces and modern cultivars**

In diversity studies, wheat landraces usually cluster in a separate group from elite cultivars (54,55,63,65) but some degree of mixture has also been found (39,62). Concerning the durum wheat materials examined here, the great genetic divergence between the bulk of the landraces and the reference set is remarkable (Figure 4A). Moreover, some of the *durum* landraces were located closer to *turgidum* and *dicoccon* accessions than to the reference varieties (which all belong to subsp. *durum*). However, high relatedness to the reference varieties was detected for a reduced group of *durum* landraces. This group included 'Caravaca' (BGE002869), a Spanish landrace used by CIMMYT in the development of some modern cultivars (see the Genetic Resources Information System for Wheat and Triticale of CIMMYT at <http://www.wheatpedigree.net/>). It can also be noted that other studies have reported a close relationship between Spanish and North African landraces (39,44) and that the reference set included two old cultivars that were commonly cultivated in Spain in the past ('Senatore Capelli' and 'Bidi-17'), both of which exhibit a North African origin. Nevertheless, our results supported little involvement of Spanish landraces in the development of the modern durum wheat varieties grown in Spain at present.

In bread wheat, the situation was quite different, and the mixing between the landraces and reference varieties was much higher, especially for some accessions in Pop2 (Figure 6). It is likely that some of these accessions are not true local landraces but, rather, old improved cultivars that were wrongly classified. The clustering of the landraces and reference varieties could also indicate a pedigree relationship. Hence, it is possible that some of the landraces characterized in our study were among the unidentified "Mediterranean" local varieties utilized by the early breeders as starting material to develop pure lines that were further involved in cross-breeding (see <http://www.wheatpedigree.net/>). Some other varieties such as 'Richela Blanca', 'Montnegre' or 'Ardito', are related to old Italian material (43,66), which provides another feasible genealogical link between the sets of landraces and reference bread wheat materials analyzed.

The overall genetic diversity of the reference cultivars was much lower than that of the landraces in both species (Table 2; Figures 4 and 6). Genomic regions showing patterns of variation that differ between landraces and varieties can aid in the identification of loci under selection during crop improvement, which will help to better target future breeding efforts (67). Our analysis allowed the identification of several such genomic regions by studying the distribution of genetic diversity across the reference genomes (Figure 7). The number of regions that have presumably been fixed by selective breeding was higher in bread wheat than in durum wheat. As expected, some of the chromosomes including fixed regions harbored genes related to agronomically important phenotypes, such as *Rht-D1*, associated with dwarf phenotype, and *Vrn-B1*, associated with vernalization response, on bread wheat chromosomes 4D and 5B, respectively (68,69). In turn, the fixed region on bread wheat chromosome 1A may be related to

the presence in this chromosome of major determinants of bread quality, such as the HMW-GS encoding the *Glu-1A* locus. Coupling this analysis with future GWA studies will help to identify the traits underlying each of the fixed regions detected in this work.

## Conclusions

The replacement of local landraces by high-yielding wheat varieties that began at the time of the Green Revolution has led to a loss of genetic variation in crop wheat varieties. This depletion has now encouraged the use of genetic resources in wheat breeding programs, but the genetic variability of these resources needs to be exhaustively characterized for their efficient use. The present study successfully used DArTseq technology for evaluating the diversity within and between two landrace collections of bread and durum wheat and for assessing the genetic relationships between each of these collections and a reference set of modern wheat cultivars. The study of genome-wide diversity provides a resource for the design of high-power GWAS experiments, which will help to achieve the overarching goal of improving wheat for cultivation in different environments, ecosystems and stress situations. The collections of Spanish landraces characterized in the present study were clearly clustered into different groups, representing different gene pools capable of providing different sources of genes for plant breeding. The investigated panel of genotypes showed an outstanding degree of diversity compared to the reference counterparts. It therefore clearly represents a strategic platform and a valuable genetic resource that must be further studied to ensure not only its efficient conservation and management but also its useful exploitation in breeding programs.

## Methods

### Plant material

The plant material analyzed in the present study comprised 432 selected accessions (detailed information is presented in supplementary Additional file 1). This sample included 191 durum and 189 bread wheat landraces and old local cultivars (hereafter referred to as landraces), representing different ecological and geographical areas of Spain (27,70). Among the durum wheat landraces, 140, 37 and 14 landraces corresponded to the subspecies *durum*, *turgidum* and *dicoccon*, respectively, while all the bread wheat landraces belonged to the subspecies *vulgare*. All these accessions were homozygous lines derived from genebank accessions provided by CRF-INIA. Additionally, the study included a set of 23 improved varieties of durum wheat and 29 improved varieties of bread wheat (hereafter referred to as reference varieties), which comprised the cultivars most widely grown in Spain during the last 50 years plus some varieties widely used in wheat research, such as tetraploid 'Langdon' and hexaploid 'Chinese Spring'.

### DNA isolation and genotyping analysis

For each accession, genomic DNA was isolated from the young leaves of a single plant using the CTAB method (71). Samples were genotyped using DArTseq GBS technology at Diversity Arrays Technology Pvt, Ltd. (Canberra, Australia) for the durum wheat accessions and SAGA (Genetic Analysis Service for Agriculture, Mexico City, Mexico) for the bread wheat accessions. (72,73)

A complexity reduction method including two enzymes (PstI and HpaII) was used to create a genome representation of the set of samples. PstI-RE site specific adapter was tagged with 96 different barcodes enabling multiplexing a 96-well microtiter plate with equimolar amounts of amplification products in order to run within a single lane on Illumina HiSeq2500 instrument (Illumina Inc., San Diego, CA). The successful amplified fragments were sequenced up to 77 bases, generating approximately 500,000 unique reads per sample. Thereafter the FASTQ files (full reads of 77bp) were quality filtered using a Phred quality score of 30, which represent a 90% of base call accuracy for at least 50% of the bases. More stringent filtering was also performed on barcode sequences using a Phred quality score of 10, which represent 99.9% of base call accuracy for at least 75% of the bases. A proprietary analytical pipeline developed by DArT P/L was used to generate allele calls for SNP and DArT markers.

After this process the genotyping services provided two different sets of markers. The DArT markers were scored as binary data (0/1) indicating the presence or absence of a marker in each accession, and the SNP markers were scored as 0/1/2 indicating the presence of the reference allele in homozygosity, the alternative allele in homozygosity or a heterozygous genotype, respectively. The raw data are available upon request to the corresponding author. To locate the markers in the durum and bread wheat reference genomes, the markers sequences were subjected to BLAST searches against the currently available *Triticum aestivum* genome IWGSC Refseq v1.0 (4) for bread wheat markers and *Triticum turgidum* genome Svevo v1.0 (5) for durum wheat markers. A marker was located according to the following criteria BLAST E-value < 5e-10 and sequence identity >90%.

For comparison with the population structure based on GBS-DArTseq markers, we investigated the allelic variability of functional markers in the *Vrn-A1* gene, one of the most determinant loci involved in the transition from vegetative to reproductive growth (74,75). It has been described that carrying a dominant allele at the *Vrn-A1* locus is sufficient to confer a spring growth habit (53). Three alleles (*Vrn-A1a*, *Vrn-A1b* and *Vrn-A1c*) were characterized by PCR according to (51,52,68) and following the protocols described at <https://maswheat.ucdavis.edu/protocols/Vrn/index.htm>.

Additionally, the panel of accessions was genotyped for *Glu-1* homoeoloci. These complex loci encode the HMW-GS, which are the major determinants of dough quality in wheat. For this purpose, endosperm proteins were extracted from single seeds and fractionated via sodium dodecyl sulfate polyacrylamide gel electrophoresis (12% polyacrylamide gels) according to Payne et al. (1980). HMW-GS allele classification was performed according to the Catalogue of Gene Symbols for Wheat 2013 (76).

## Data analysis

Prior to any further analysis, the set of SNP and DART markers was filtered employing homemade R scripts (77), which are available upon request to the corresponding author. For DART markers we selected high quality markers following a step-by-step filtering strategy. First, when several markers presented the same allelic profile, all of the markers but the one with the least missing data were removed. Then, the markers that presented more than 10% missing data or were monomorphic (Minimum Allele Frequency, MAF <0.05) were excluded.

For the SNP markers, prior to any filtering step we analyzed the presence of heterozygous genotypes. As we have previously described, genotyping was conducted on homozygous lines and thus we did not expect any heterozygous genotypes. When the genotypic values for a marker was only 0 and 2, we considered the heterozygous calling (2) to be an error caused by the presence of the SNP marker flanking sequence in homoeologous genomes. In this case, the genotypes scored as 2 were recoded as 1. The same procedure was followed when only genotypic values of 1 and 2 were present in a marker, but in those cases, genotypes scored as 2 were recoded as 0. Finally, when the genotypic values for a marker included 0, 1 and 2 we recoded the genotypes scored as 2 as missing data. After this analysis we selected high quality SNP markers following the filtering strategy described for the DART markers. First, when several markers presented the same allelic profile, all but the one with the less missing data were removed. Then, markers with more than 10% missing data or monomorphic (MAF <0.05) were excluded.

The genetic substructure within the durum and bread wheat landrace collections was investigated using the fastSTRUCTURE algorithm (42) and the DArT marker dataset (including all the available DArTs after filtering, located or not in the reference genomes). Default parameters and K values from 1 to 15 were tested. The appropriate number of components that explained the structure in the dataset was determined using the chooseK.py function (42). The results for the identified optimal values of K were visualized using DISTRUCT (78). Individual accessions were assigned to the population with the highest proportional membership.

The genetic similarity based on the SNP data within the full sets of accessions (landraces plus reference varieties) was analyzed by principal coordinates analysis (PCoA) using the gl.pcoa function from the dartR R package (79) (including all the available SNPs after filtering, located or not in the reference genomes).

The gene diversity ( $H_s$ ) within populations, landraces and reference varieties was calculated based on the SNP dataset according to (80) with the basic.stats function from the hierfstat R package (81).  $D_{est}$ , a measure of population differentiation in collections with several populations, was calculated as defined by (82) with this same function. The genetic differentiation between populations was analyzed by estimating the pairwise fixation index ( $F_{st}$ ) according to (83) with the stampFst function from R the StAMPP package (84) using the SNP dataset (including all the available SNPs after filtering, located or not in the reference genomes).

Fixed genomic regions in the reference varieties were identified by performing a scan of the  $H_s$  values along the different chromosomes.  $H_s$  was estimated as described previously (80). However, for this analysis we obtained a new SNP dataset by avoiding the first filtering step, thus when several markers presented the same allelic profile we kept all of them and only filter out the markers that presented more than 10% missing data or were monomorphic ( $MAF < 0.05$ ). Finally, for this particular analysis only markers with known location in the reference genomes were employed. A region was considered to be “fixed” when it contained at least 5 consecutive markers with an  $H_s$  equal to 0 in the reference varieties and at least 5 markers with an  $H_s > 0.1$  in the landraces, and spanned more than 5 Mb.

## Abbreviations

**CIMMYT:** Centro internacional de Mejoramiento de Maíz Y Trigo

**CRF-INIA:** National Plant Genetic Resources Center

**DArT:** Diversity Array Technology

**DArTseq:** Diversity Array Technology sequence

**$D_{est}$ :** Jost’s Population Differentiation index

**$F_{ST}$ :** Wright’s Fixation index

**GBS:** Genotyping By Sequencing

**GWAS:** Genome Wide Association Studies

**HMW-GS:** High Molecular Weight Glutenin Subunits

**$H_s$ :** Nei’s diversity index

**MAF:** Minimum Allele Frequency

**MAS:** Marker Assisted Selection

**PCoA:** Principal Coordinates Analysis

**PIC:** Polymorphic Information Content

**SNP:** Single Nucleotide Polymorphism

**SSRs:** Single Sequence Repeats

## Declarations

## **ETHICS APPROVAL AND CONSENT TO PARTICIPATE**

Not applicable

## **CONSENT FOR PUBLICATION**

Not applicable

## **AVAILABILITY OF DATA AND MATERIALS**

Plant material and raw data are available upon request to the corresponding author.

## **COMPETING INTERESTS**

The authors declare that they have no competing interests.

## **FUNDING**

This work was supported by the Ministry of Economy, Industry and Competitiveness projects AGL2016-77149-C2-1P and RFP2015-00008-C04-01, the Comunidad Autónoma de Madrid project P2018/BAA-4330, and the Universidad Politécnica de Madrid project VJIDOCUPM18LPB. ML is a recipient of a predoctoral fellowship from the Programa Propio of the Universidad Politécnica de Madrid.

## **AUTHORS' CONTRIBUTIONS**

PG and MR: conceived and designed the study; EB, MLF, LP, MR and JFV: characterized the panel; MLF, LP, HPP and CS: performed the bioinformatics analyses; PG, LP and MR: valorized the results and wrote the manuscript; all authors have read and approved the final manuscript.

## **ACKNOWLEDGMENTS**

The authors are grateful to Dr. JM Carrillo for helpful comments during the preparation of the original and revised versions of the manuscript.

## **References**

- (1) Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Jakobsen KS, et al. Ancient hybridizations among the ancestral genomes of bread wheat. *Science* 2014;345(6194):1250092.
- (2) Leng P, Lübberstedt T, Xu M. Genomics-assisted breeding—a revolutionary strategy for crop improvement. *J Integr Agric* 2017;16(12):2674-2685.
- (3) Mochida K, Yamazaki Y, Ogihara Y. Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Mol Genet Genomics* 2004;270(5):371-377.

- (4) Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 2018;361(6403):eaar7191.
- (5) Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, et al. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet* 2019;51(5):885.
- (6) Rasheed A, Xia X. From markers to genome-based breeding in wheat. *Theor Appl Genet* 2019;132(3):767-784.
- (7) Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C, et al. Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol Biol Evol* 2007;24(7):1506-1517.
- (8) Reif JC, Zhang P, Dreisigacker S, Warburton ML, van Ginkel M, Hoisington D, et al. Wheat genetic diversity trends during domestication and breeding. *Theor Appl Genet* 2005;110(5):859-864.
- (9) Lopes MS, El-Basyoni I, Baenziger PS, Singh S, Royo C, Ozbek K, et al. Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *J Exp Bot* 2015;66(12):3477-3486.
- (10) Sehgal D, Vikram P, Sansaloni CP, Ortiz C, Pierre CS, Payne T, et al. Exploring and mobilizing the gene bank biodiversity for wheat improvement. *PLoS One* 2015;10(7):e0132112.
- (11) Idrees M, Irshad M. Molecular markers in plants for analysis of genetic diversity: a review. *European academic research* 2014;2(1):1513-1540.
- (12) Govindaraj M, Vetriventhan M, Srinivasan M. Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genet Res Int*;2015.
- (13) Khan MK, Pandey A, Choudhary S, Hakki EE, Akkaya MS, Thomas G. From RFLP to DArT: molecular tools for wheat (*Triticum* spp.) diversity analysis. *Genet Resour Crop Evol* 2014;61(5):1001-1032.
- (14) Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, et al. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant biotechnol J* 2014;12(6):787-796.
- (15) Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 2011;6(5):e19379.
- (16) Holtz Y, Ardisson M, Ranwez V, Besnard A, Leroy P, Poux G, et al. Genotyping by sequencing using specific allelic capture to build a high-density genetic map of durum wheat. *PloS One* 2016;11(5):e0154609.
- (17) Cruz VMV, Kilian A, Dierig DA. Development of DArT marker platforms and genetic diversity assessment of the US collection of the new oilseed crop lesquerella and related species. *PLoS One*

- (18) Valdisser PA, Pereira WJ, Almeida Filho JE, Müller BS, Coelho GR, de Menezes IP, et al. In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping. *BMC Genomics* 2017;18(1):423.
- (19) Yang X, Ren R, Ray R, Xu J, Li P, Zhang M, et al. Genetic diversity and population structure of core watermelon (*Citrullus lanatus*) genotypes using DArTseq-based SNPs. *Plant Genetic Resources* 2016;14(3):226-233.
- (20) Baloch FS, Alsaleh A, Shahid MQ, Çiftçi V, Sáenz de Miera L, Aasim M, et al. A whole genome DArTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent. *PLoS One* 2017;12(1):e0167821.
- (21) Robbana C, Kehel Z, Sansaloni C, Bassi F, Amri A. Genome-Wide genetic diversity and population structure of tunisian durum wheat landraces based on DArTseq technology. *Int J Mol Sci* 2019;20(6):1352.
- (22) Edet OU, Gorafi YS, Nasuda S, Tsujimoto H. DArTseq-based analysis of genomic relationships among species of tribe Triticeae. *Sci Rep* 2018;8.
- (23) Rasheed A, Mujeeb-Kazi A, Ogonnaya FC, He Z, Rajaram S. Wheat genetic resources in the post-genomics era: promise and challenges. *Ann Bot* 2017;121(4):603-616.
- (24) Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* 2013;9(1):29.
- (25) Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, et al. An Arabidopsis example of association mapping in structured samples. *PLoS Genetics* 2007;3(1):e4.
- (26) Pfluger LA, Martin LM, Alvarez JB. Variation in the HMW and LMW glutenin subunits from Spanish accessions of emmer wheat (*Triticum turgidum* ssp. *dicoccum* Schrank). *Theor Appl Genet* 2001;102(5):767-772.
- (27) Ruiz M, Giraldo P, Royo C, Villegas D, Aranzana MJ, Carrillo JM. Diversity and genetic structure of a collection of Spanish durum wheat landraces. *Crop Sci* 2012;52(5):2262-2275.
- (28) Aguiriano E, Ruiz M, Fité R, Carrillo JM. Analysis of genetic variability in a sample of the durum wheat (*Triticum durum* Desf.) Spanish collection based on gliadin markers. *Genet Resour Crop Evol* 2006;53(8):1543-1552.
- (29) Moragues M, Moralejo M, Sorrells ME, Royo C. Dispersal of durum wheat [*Triticum turgidum* L. ssp. *turgidum* convar. *durum* (Desf.) MacKey] landraces across the Mediterranean basin assessed by AFLPs and microsatellites. *Genet Resour Crop Evol* 2007;54(5):1133-1144.

- (30) Kilian B, Graner A. NGS technologies for analyzing germplasm diversity in genebanks. *Brief Funct Genomics* 2012;11(1):38-50.
- (31) Heslot N, Rutkoski J, Poland J, Jannink JL, Sorrells ME. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PloS One* 2013 September 05;8(9):e74612.
- (32) Manickavelu A, Jighly A, Ban T. Molecular evaluation of orphan Afghan common wheat (*Triticum aestivum* L.) landraces collected by Dr. Kihara using single nucleotide polymorphic markers. *BMC Plant Biology* 2014;14(1):320.
- (33) Dvorak J, Luo M, Yang Z, Zhang H. The structure of the *Aegilops tauschii* genepool and the evolution of hexaploid wheat. *Theor Appl Genet* 1998;97(4):657-670.
- (34) Lagudah ES, Appels R, Brown A, McNeil D. The molecular–genetic analysis of *Triticum tauschii*, the D-genome donor to hexaploid wheat. *Genome* 1991;34(3):375-386.
- (35) Qi LL, Echalié B, Chao S, Lazo GR, Butler GE, Anderson OD, et al. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* 2004;168(2):701-712.
- (36) Miftahudin KR, Ma X, Mahmoud AA, Layton J, Milla MR, Chikmawati T, et al. Analysis of expressed sequence tag loci on wheat chromosome group 4. *Genetics* 2004;168(2):651.
- (37) Devos KM, Dubcovsky J, Dvorak J, Chinoy CN, Gale MD. Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. *Theor Appl Genet* 1995;91(2):282-288.
- (38) Cubizolles N, Rey E, Choulet F, Rimbart H, Laugier C, Balfourier F, et al. Exploiting the repetitive fraction of the wheat genome for high-throughput single-nucleotide polymorphism discovery and genotyping. *The plant genome* 2016;9(1).
- (39) Kabbaj H, Sall AT, Al-Abdallat A, Geleta M, Amri A, Filali-Maltouf A, et al. Genetic diversity within a global panel of durum wheat (*Triticum durum*) landraces and modern germplasm reveals the history of alleles exchange. *Front Plant Sci* 2017;8:1277.
- (40) Rimbart H, Darrier B, Navarro J, Kitt J, Choulet F, Leveugle M, et al. High throughput SNP discovery and genotyping in hexaploid wheat. *PloS One* 2018;13(1):e0186329.
- (41) Novoselovic D, Bentley AR, Simek R, Dvojkovic K, Sorrells ME, Gosman N, et al. Characterizing Croatian wheat germplasm diversity and structure in a European context by DArT markers. *Frontiers in plant science* 2016;7:184.
- (42) Raj A, Stephens M, Pritchard JK. Variational inference of population structure in large SNP datasets. *Genetics* 2014;197:573-589

- (43) Gadea M. Trigos españoles. Instituto Nacional de Investigaciones Agronómicas; 1954.
- (44) Oliveira HR, Campana MG, Jones H, Hunt HV, Leigh F, Redhouse DI, et al. Tetraploid wheat landraces in the Mediterranean basin: taxonomy, evolution and genetic diversity. PLoS One 2012;7(5):e37063.
- (45) Sahri A, Chentoufi L, Arbaoui M, Ardisson M, Belqadi L, Birouk A, et al. Towards a comprehensive characterization of durum wheat landraces in Moroccan traditional agrosystems: analysing genetic diversity in the light of geography, farmers' taxonomy and tetraploid wheat domestication history. BMC Evo Biol 2014;14(1):264.
- (46) Mac Key J. Wheat: its concept, evolution, and taxonomy. Durum Wheat Breeding: CRC Press; 2005. p. 35-94.
- (47) Mac Key J. Species relationship in Triticum. Hereditas 1966;2:237-276.
- (48) Basualdo J, Díaz ML, Cuppari S, Cardone S, Soresi D, Camargo GP, et al. Allelic variation and differential expression of VRN-A1 in durum wheat genotypes varying in the vernalization response. Plant Breed 2015;134(5):520-528.
- (49) Konopatskaia I, Vavilova V, Kondratenko EY, Blinov A, Goncharov NP. VRN1 genes variability in tetraploid wheat species with a spring growth habit. BMC Plant Biology 2016;16(3):244.
- (50) Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishna W, SanMiguel P, et al. The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. Science 2004 March 12;303(5664):1640-1644.
- (51) Shcherban A, Emtseva M, Efremova T. Molecular genetical characterization of vernalization genes *Vrn-A1*, *Vrn-B1* and *Vrn-D1* in spring wheat germplasm from Russia and adjacent regions. Cereal Res Commun 2012;40(3):351-361.
- (52) Fu D, Szűcs P, Yan L, Helguera M, Skinner JS, Von Zitzewitz J, et al. Large deletions within the first intron in VRN-1 are associated with spring growth habit in barley and wheat. Mol Genet Genomics 2005;273(1):54-65.
- (53) Chu CG, Tan CT, Yu GT, Zhong S, Xu SS, Yan L. A novel retrotransposon inserted in the dominant *Vrn-B1* allele confers spring growth habit in tetraploid wheat (*Triticum turgidum* L.) G3 2011 Dec;1(7):637-645.
- (54) Fayaz F, Sarbarzeh MA, Talebi R, Azadi A. Genetic diversity and molecular characterization of iranian durum wheat landraces (*Triticum turgidum durum* (Desf.) Husn.) Using DArT Markers. Biochem Genet 2018:1-19.
- (55) Mengistu DK, Kidane YG, Catellani M, Frascaroli E, Fadda C, Pé ME, et al. High-density molecular characterization and association mapping in Ethiopian durum wheat landraces reveals high diversity and potential for wheat breeding. Plant Biotechnol J 2016;14(9):1800-1812.

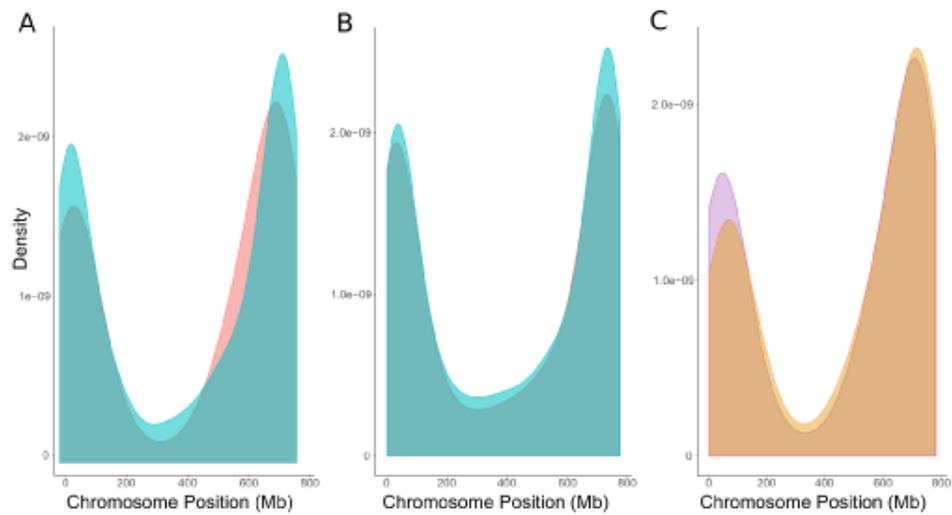
- (56) Zhang P, Dreisigacker S, Buerkert A, Alkhanjari S, Melchinger AE, Warburton ML. Genetic diversity and relationships of wheat landraces from Oman investigated with SSR markers. *Genet Resour Crop Evol* 2006;53(7):1351-1360.
- (57) Talbert LE, Smith LY, Blake NK. More than one origin of hexaploid wheat is indicated by sequence comparison of low-copy DNA. *Genome* 1998;41(3):402-407.
- (58) López-Arias M, Grau-Corbí JM. Metales pesados, materia orgánica y otros parámetros de la capa superficial de los suelos agrícolas y de pastos de la España peninsular. II. Resultados globales. INIA, MMA, MEC, MAPA. Madrid. 2005.
- (59) Rodríguez-Quijano M, Vázquez JF, Carrillo JM. Variation of high molecular weight glutenin subunits in Spanish landraces of *Triticum aestivum* ssp. *vulgare* and ssp. *spelta*. *Journal of Genetics & Breeding* 1990;44(2):121-126.
- (60) Giraldo P, Rodríguez-Quijano M, Simon C, Vázquez JF, Carrillo JM. Allelic variation in HMW glutenins in Spanish wheat landraces and their relationship with bread quality. *Span J Agric Res* 2010;8(4):1012-1023.
- (61) Iqbal M, Navabi A, Yang R, Salmon DF, Spaner D. The effect of vernalization genes on earliness and related agronomic traits of spring wheat in northern growing regions. *Crop Sci* 2007;47(3):1031-1039.
- (62) Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, et al. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc Natl Acad Sci USA* 2013;110(20):8057-8062.
- (63) Muller T, Schierscher-Viret B, Fossati D, Brabant C, Schori A, Keller B, et al. Unlocking the diversity of genebanks: whole-genome marker analysis of Swiss bread wheat and spelt. *Theor Appl Genet* 2018 Feb;131(2):407-416.
- (64) Brambilla V, Gomez-Ariza J, Cerise M, Fornara F. The importance of being on time: regulatory networks controlling photoperiodic flowering in cereals. [Front Plant Sci](#). 2017;8:665.
- (65) Soriano JM, Villegas D, Aranzana MJ, Garcia Del Moral, L. F., Royo C. Genetic structure of modern durum wheat cultivars and mediterranean landraces matches with their agronomic performance. *PLoS One* 2016 Aug 11;11(8):e0160983.
- (66) Andreu JP. Las innovaciones biológicas en la agricultura española antes de 1936: el caso del trigo. *Agricultura y Sociedad*, 1998:163-164.
- (67) Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* 2012;30(1):105.

- (68) Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J. Positional cloning of the wheat vernalization gene VRN1. *Proc Natl Acad Sci USA* 2003;100(10):6263-6268.
- (69) Hedden P. The genes of the Green Revolution. *Trends Genet.* 2003;19(1):5-9.
- (70) Bread wheat (*Triticum aestivum* L.) core collection based in Spanish landraces. Proceedings 8th International wheat conference; St. Petersburg, Russia. NI Vavilov Research Institute of Plant Industry 2010.
- (71) Doyle JJ, Doyle JL, Hortoriun LB. Isolation of plant DNA from fresh tissue. *Focus* 1990;12:13-15.
- (72) Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, et al. Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods Mol Biol.* 2012;888:67-89
- (73) Sansaloni C, Petrol C, Laccoud D, Carling J, Deterinh F, Grattapaglia D, Kilian A. Diversity Arrays Technology (DARt) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. *BMC proceedings: BioMed Central*; 2011.
- (74) Chen F, Gao M, Zhang J, Zuo A, Shang X, Cui D. Molecular characterization of vernalization and response genes in bread wheat from the Yellow and Huai Valley of China. *BMC Plant Biol* 2013;13(1):199.
- (75) Zhang XK, Xiao YG, Zhang Y, Xia XC, Dubcovsky J, He ZH. Allelic variation at the vernalization genes *Vrn-A1*, *Vrn-B1*, *Vrn-D1*, and *Vrn-B3* in Chinese wheat cultivars and their association with growth habit. *Crop Sci* 2008;48(2):458-470.
- (76) McIntosh RA, Yamazaki Y, Devos KM, Dubcovsky J, Rogers WJ, Appels R. Catalogue of gene symbols for wheat. *Wheat Information Service* 2013;97:27-37.
- (77) R Core Team. R: A language and environment for statistical computing. 2014.
- (78) Rosenberg NA. DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* 2004;4(1):137-138.
- (79) Gruber B, Unmack PJ, Berry OF, Georges A. DARTR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources* 2018;18(3):691-699.
- (80) Nei M. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* 1973;70(12):3321.
- (81) Goudet J. Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* 2005;5(1):184-186.
- (82) Jost L. GST and its relatives do not measure differentiation. *Mol Ecol* 2008;17(18):4015-4026.

(83) Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution* 1984;38(6):1358-1370.

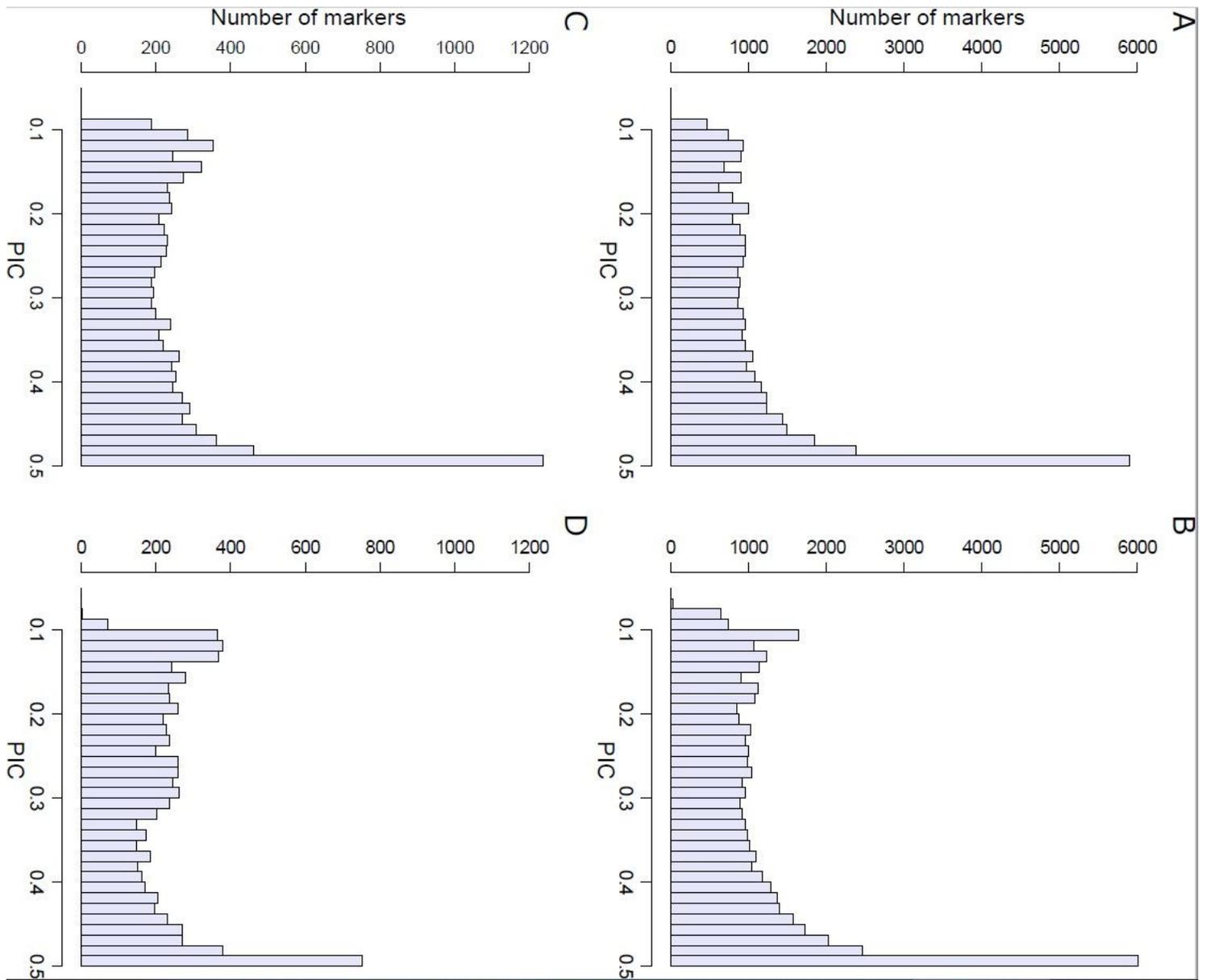
(84) Pembleton LW, Pembleton ML. Package 'StAMP'. 2017.

## Figures



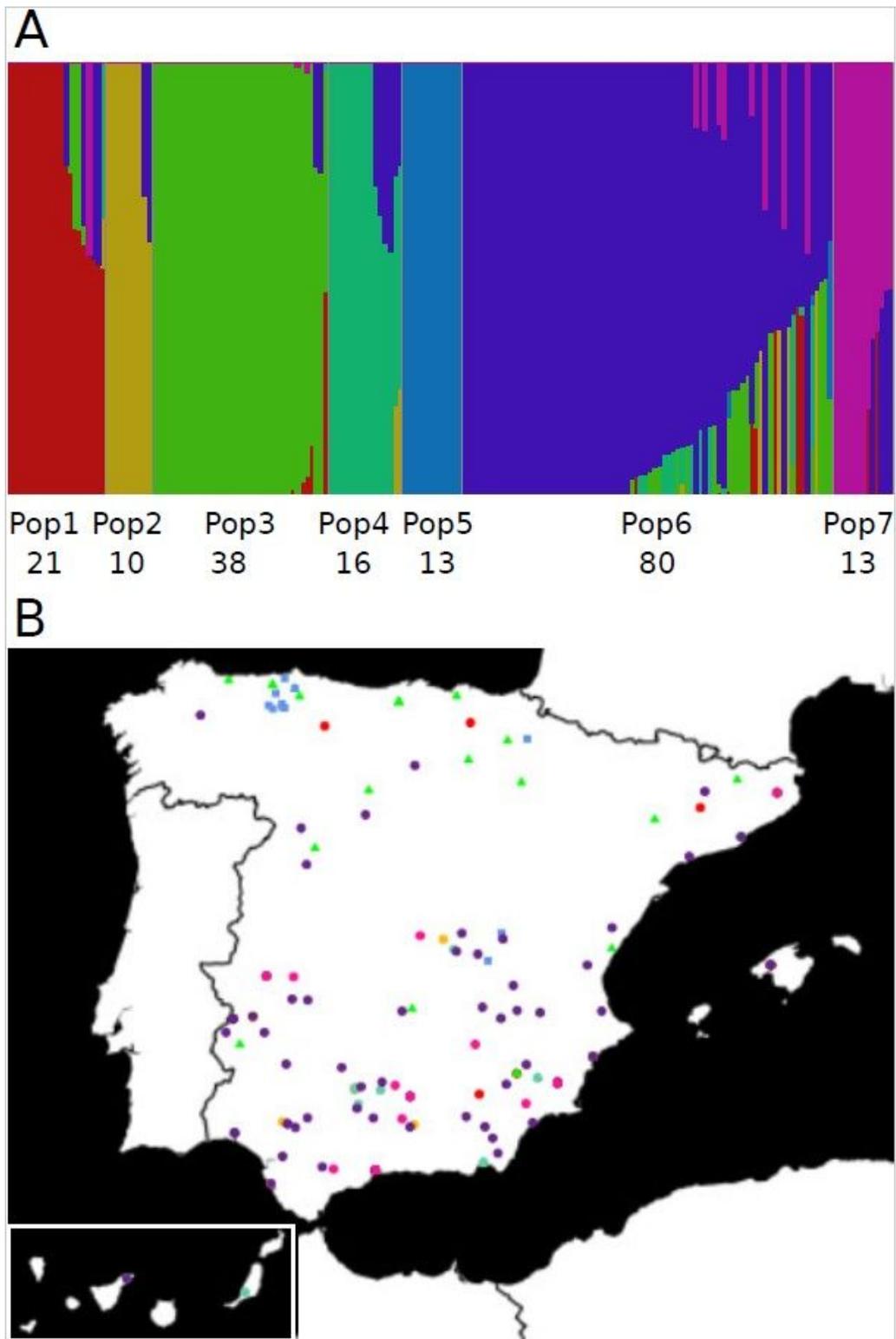
**Figure 1**

Marker density along chromosome 2A. (A) *T. turgidum* raw (blue) and filtered (red) SNP markers. (B) *T. turgidum* raw (blue) and filtered (red) DArT markers. (C) Filtered SNPs in *T. turgidum* (purple) and *T. aestivum* (yellow).



**Figure 2**

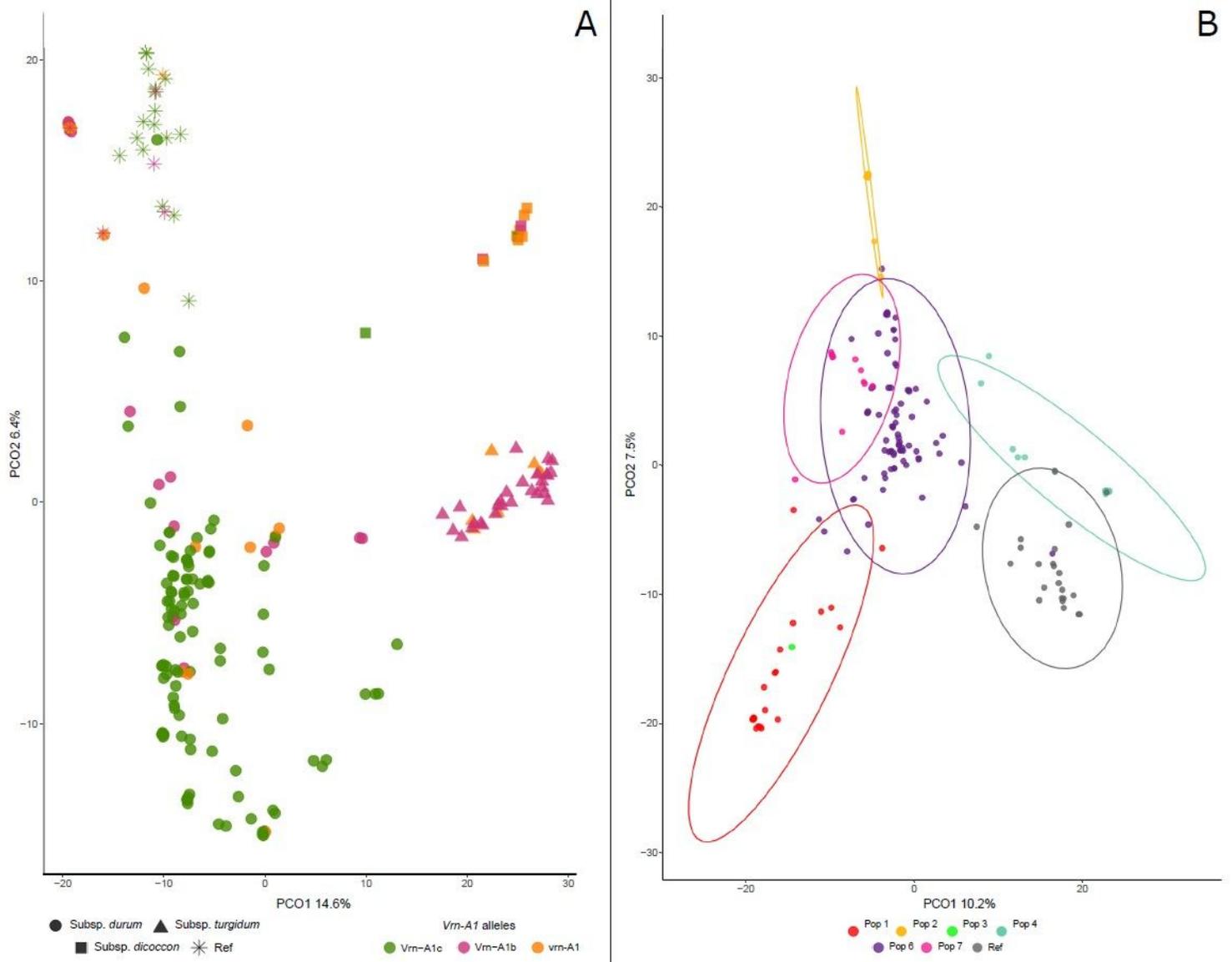
Average PIC distribution in filtered markers. (A) *T. turgidum* DArTs. (B) *T. aestivum* DArTs. (C) *T. turgidum* SNPs. (D) *T. aestivum* SNPs.



**Figure 3**

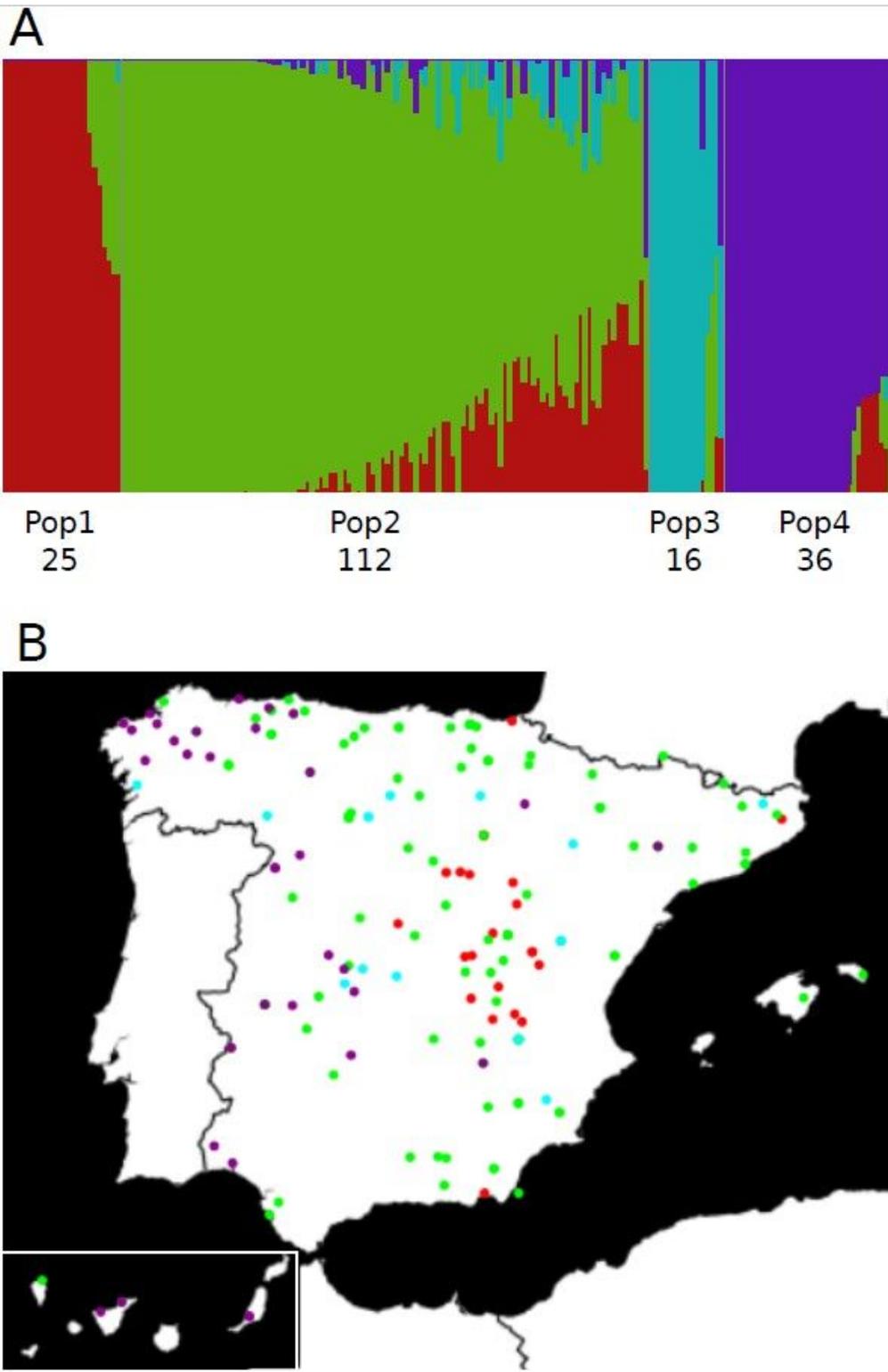
(A) *T. turgidum* STRUcTURE plot based on DArT markers. The number below the Pop indicates the number of accessions clustered in each population. (B) Collection sites of the different *T. turgidum* accessions, colored according to their STRUcTURE population assignment. When GPS coordinate data were not available, the coordinates of the capital of the province of origin were used. *T. turgidum* subsp.

durum landraces are shown with circles, subsp. dicoccon with squares and subsp. turgidum with triangles.



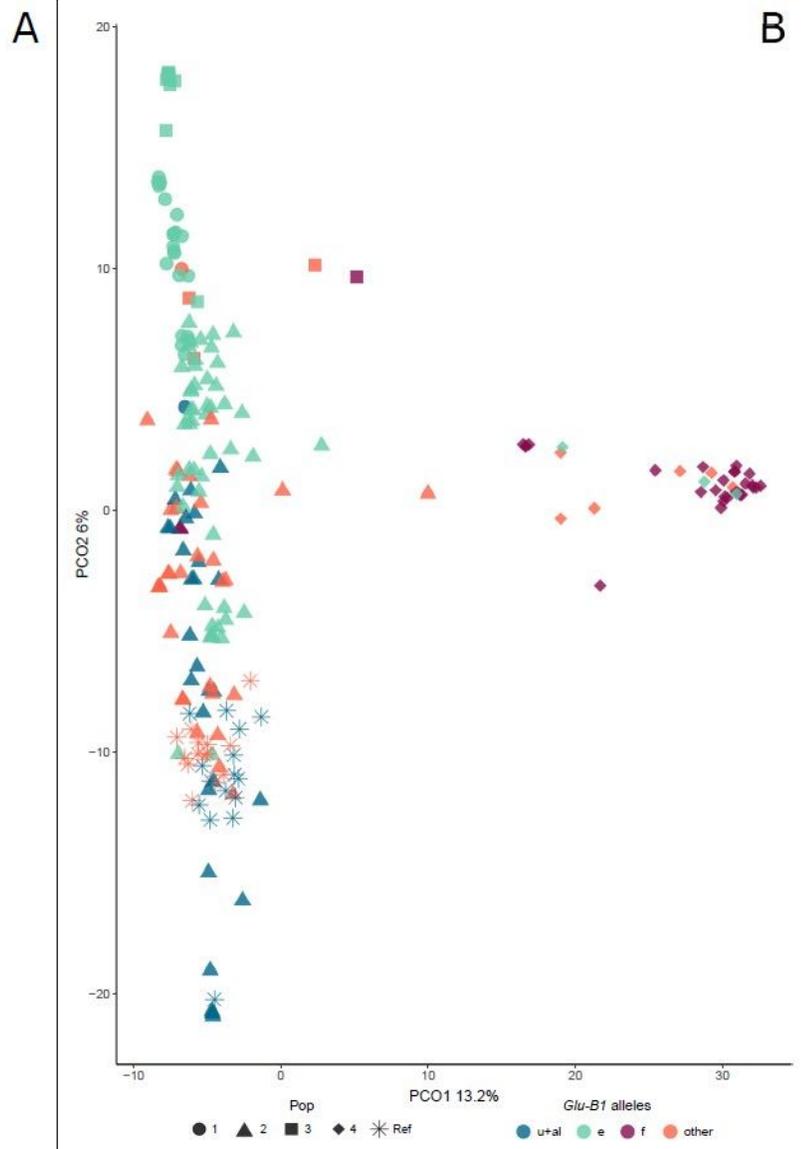
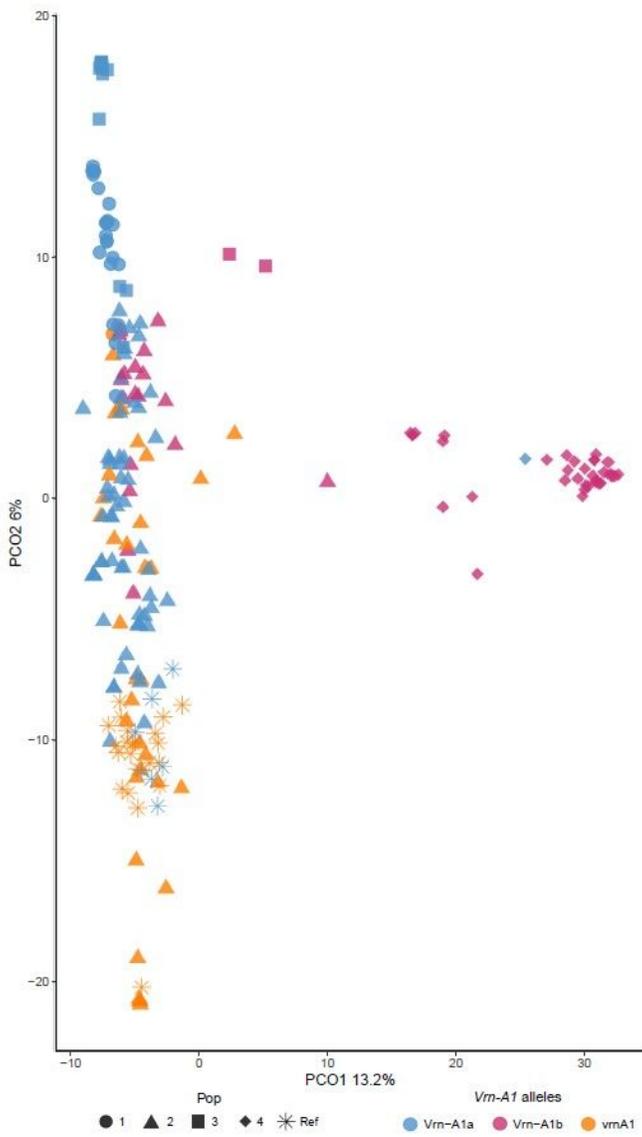
**Figure 4**

Cluster analysis of *T. turgidum* accessions using PCoA. Accessions from subsp. durum are shown with circles, subsp. dicoccon with squares, subsp. turgidum with triangles and the reference varieties with asterisks. (A) Graphical representation of PCo1 and PCo2 for the whole collection of durum wheat. Accessions are colored according to their Vrn-A1 alleles. (B) Graphical representation of PCo1 and PCo2 for subsp. durum accessions, which are colored according to their STRUcTURE population assignment.



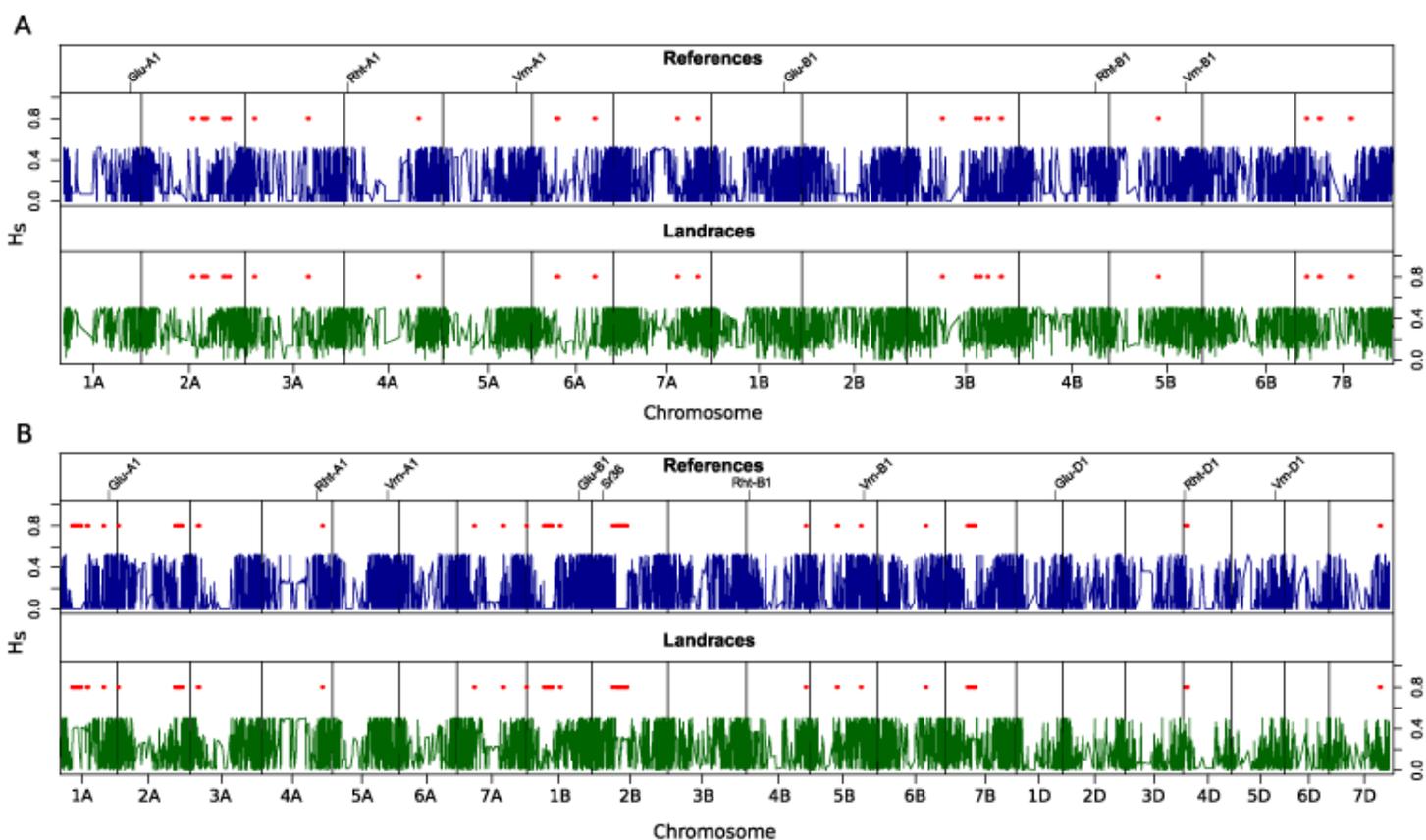
**Figure 5**

(A) *T. aestivum* STRUCTURE plot based on DArT markers. The number below the Pop indicates the number of accessions clustered in each population. (B) Collection sites of the different *T. aestivum* accessions, colored by their STRUCTURE population assignment. When GPS coordinate data were not available, the coordinate of the capital of the province of origin were used.



**Figure 6**

Cluster analysis of *T. aestivum* accessions using PCoA. Accessions from STRUCTURE Pop1 are shown with circles, Pop2 with triangles, Pop3 with squares, Pop 4 with rhombi, and the reference varieties with asterisks. (A) Graphical representation of PCo1 and PCo2 for the whole collection of bread wheat. Accessions are colored according to their *Vrn-A1* allele. (B) Accessions are colored according to their *Glu-B1* allele.



**Figure 7**

Genetic diversity ( $H_s$ ) distribution across the genome in the reference varieties and landraces. (A) *T. durum*. (B) *T. aestivum*. Fixed regions in the reference materials are indicated by red bars. The position of ten key genes in the genome is indicated.

## Supplementary Files

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

- [Additionalfile4.csv](#)
- [Additionalfile3.csv](#)
- [Additionalfile5.csv](#)
- [Additionalfile7.pdf](#)
- [Additionalfile6.pdf](#)

- [Additionalfile8.pdf](#)
- [Additionalfile2.csv](#)
- [Additionalfile1.xlsx](#)