Genetic association with high-resolution climate data reveals selection footprints in the genomes of barley landraces across the Iberian Peninsula

Running title: Climate driven selection footprints in barley

Keywords: agriculture, agroclimatic indices, genome-wide association analysis, selection footprint, barley

Authors: Bruno Contreras-Moreira¹²*, Roberto Serrano-Notivoli¹*, Naheif E. Mohamed¹³, Carlos P. Cantalapiedra¹, Santiago Beguería¹, Ana M. Casas¹, Ernesto Igartua¹

¹Estación Experimental de Aula Dei (EEAD-CSIC), Av. Montañana 1005, 50059 Zaragoza, Spain.

²Fundación ARAID, Av. de Ranillas 1-D, planta 2ª, oficina B, 50018 Zaragoza, Spain.

³currently, Agronomy Department, Faculty of Agriculture, Sohag University, Sohag, Egypt.

* These authors contributed equally to this work

Correspondence to:

Ernesto Igartua, Estación Experimental de Aula Dei-CSIC, Av Montañana 1005, 50059 Zaragoza, Spain. Phone: +34 976716089. Fax: +34 976716145. E-mail: igartua@eead.csic.es

Orcid ID: 0000-0003-2938-1719

This article contains 6 colour figures, 4 tables and 8119 words in the main text.

There are 2 additional files, which summarize supplemental data.
ABSTRACT

Landraces are local populations of crop plants adapted to a particular environment. Extant landraces are surviving genetic archives, keeping signatures of the selection processes experienced by them until settling in their current niches. This study intends to establish relationships between genetic diversity of barley (*Hordeum vulgare* L.) landraces collected in Spain and the climate of their collection sites. A high-resolution climatic dataset (5x5 km spatial, 1-day temporal grid) was computed from over 2,000 temperature and 7,000 precipitation stations across peninsular Spain. This dataset, spanning the period 1981-2010, was used to derive agroclimatic variables meaningful for cereal production at the collection sites of 135 barley landraces. Variables summarize temperature, precipitation, evapotranspiration, potential vernalization and frost probability at different times of the year and time scales (season and month). SNP genotyping of the landraces was carried out combining Illumina Infinium assays and genotyping-by-sequencing, yielding 9,920 biallelic markers (7,479 with position on the barley reference genome). The association of these SNPs with agroclimatic variables was analysed at two levels of genetic diversity, with and without taking into account population structure. The whole datasets and analysis pipelines are documented and available at [https://eead-csic-compbio.github.io/barley-agroclimatic-association](https://eead-csic-compbio.github.io/barley-agroclimatic-association). We found differential adaptation of the germplasm groups identified to be dominated by reactions to cold temperature and late-season frost occurrence, as well as to water availability. Several significant associations pointing at specific adaptations to agroclimatic features related to temperature and water availability were observed, and candidate genes underlying some of the main regions are proposed.
INTRODUCTION

Landraces are populations of crop plants adapted to a particular environment, through a long history of cultivation by local farmers (Zeven, 1998). Landraces are valuable materials for phylogeographic studies (see review in Newton et al., 2010). They are supposed to bear the genetic signatures of adaptation to the environments in which they were developed, including human preferences. These selection footprints can be investigated with genomics tools. The study of patterns in the genomes of extant landraces in relation to the climate of their collection sites can indicate which adaptive processes were responsible for their distribution (Jones et al., 2008). However, extant landraces will provide a partial evolutionary history of crops (Fuller et al., 2011), and the hypotheses emerging from their study should ultimately be put to test against archaeological data. Barley, (*Hordeum vulgare* L.) is one of the main cereals, 4th in the World, after maize, rice and wheat. It is a diploid species (n=7) with a very large genome (about 5 Gb), and a recently published high quality genome sequence (Mascher et al. 2017). Its adaptive history can shed light on the context of the Neolithic expansion, because barley was part of the Neolithic package of crops that spread over Europe and the Mediterranean basin 10k to 7k YBP.

Spanish landraces are an appropriate subject to study barley adaptation. On the one hand, the Iberian Peninsula displays a rather wide variety of climates (De Castro et al., 2005), presenting a diversity of ecological habitats (Myers et al. 2000; Manzano-Piedras et al. 2014). On the other hand, it has received a wide diversity of barley germplasm since the Neolithic up until the middle Ages (Fischbeck, 2002; Komatsuda et al., 2007). Ancient and later arrivals of plant materials encountered this variety of climates and surely underwent a process of selection and adaptation. First, the germplasm groups arriving in the Peninsula prevailed in
areas where they found appropriate niches. These groups then hybridized with each other to some extent in boundary regions, and evolved locally through mutations (Casas et al., 2018) and recombination, therefore producing new alleles and new allelic combinations that may have provided more ground for adaptation and selection. In fact, barley landraces from Spain are far from homogeneous. At least four different germplasm groups were identified (Yahiaoui et al., 2008), pointing at different routes of entry, in a parallel process to the one proposed for wheat (Moragues et al., 2006a,b). Our hypothesis is that the distribution of these groups in the Iberian Peninsula followed routes characterized by yet unknown environmental features, and settled according to suitability to environmental niches.

The Spanish Barley Core Collection (SBCC) was compiled as a balanced representation of the crop cultivated in the country until the second half of the 20th century (Igartua et al., 1998). It has been studied extensively, showing distinct agronomic and genetic features that highlight its interest as research tool for pre-breeding and gene mining (Igartua et al., 1998). Some genotypes of this collection demonstrated good agronomic performance in stressed environments (Yahiaoui et al., 2014), even out-yielding modern cultivars, probably due to their prolonged adaptation to Mediterranean climates.

This work investigates the association of genetic markers with a set of environmental features, including agroclimatic indices (measures or indicators of an aspect of the climate that has specific agricultural significance), chosen based on their relevance for agriculture of winter cereals. With this aim, we used the SBCC and a high-resolution climatic dataset specifically assembled for this purpose. We assessed this relationship at two levels, with and without considering population structure, with two different and complementary objectives. Life history traits like growth cycle duration and morphology of reproductive organs are strongly
affected by natural and artificial selection, and can be easily confounded with population structure (Fournier-Level et al., 2013, Leinonen et al., 2013). Therefore, analyses that remove population structure could also remove genetic variation related to life history traits (Russell et al., 2016), difficulting the detection of some of the genetic factors responsible for adaptation. To circumvent this, we first tested the relationship of the distribution of germplasm groups detected through population structure analysis and their genetic polymorphisms to agroclimatic variables. We expected that these analyses would point at genomic regions that have driven the adaptation and differentiation of germplasm groups reflecting their history of evolution in distinct agro-ecological areas. Second, we searched for associations taking into account population structure, with the goal of unravelling polymorphisms that appear linked to regions active in adaptation at a finer scale, either within germplasm groups or after local group admixture.

We believe the processes that have shaped the diversity of Spanish barleys are typical of the expansion of crops outside their centres of origin, and that these analyses provide a case in point for the usefulness of environmental association analysis (EAA) to shed light on adaptation processes affecting crop landraces.

MATERIALS AND METHODS

Spanish Barley Core Collection

This collection (SBCC, http://www.eead.csic.es/barley/index.php?lng=1) was assembled as a representative set of the barleys cultivated in Spain prior to the introduction of modern
cultivars. These landraces had passport data, stating geographic coordinates (longitude, latitude and altitude) of collection sites, and were systematically selected (Igartua et al., 1998) from a set of nearly 2,000 local accessions held at the Spanish National Bank of Phytogenetic Resources (CRF-INIA). Initially, a set of 159 inbred lines was derived after three generations of selfing starting from single spikes collected from the same number of landraces. This collection was later reduced to 140 inbred lines, due to some seed losses and to the removal of 17 duplicates detected with molecular markers. Considering only mainland accessions (Balearic and Canary islands excluded), a set of 135 landraces remained, comprising 11 two-rowed types, and 124 six-rowed ones (Table S1).

Genotyping

Single Nucleotide Polymorphism (SNP) genotyping was carried out with the Illumina Infinium iSelect 9k SNP chip (Comadran et al., 2012), with 7,864 SNP assays, processed at TraitGenetics GmbH, (Gatersleben, Germany). Additionally, the collection was genotyped at Diversity Arrays Technology (Yarralumla, Australia), with the DArTseq technology (Kilian et al., 2012). This system combines complexity reduction methods with next-generation sequencing platforms, targeting primarily genic regions. It produces two types of markers, classical SNP and presence/absence variation, also named SilicoDArTs. Only the SNPs were considered in this study. After merging Infinium and DArTseq markers, the resulting number of SNPs was 9,920 (6,509 Infinium, 3,411 DArTseq). These were further reduced to 8,457 biallelic loci after removing those with more than 10% missing data (Table S2).

Physical positions of these markers were queried by aligning their sequences to the Barley RefSeq v1.0 genome assembly (Mascher et al., 2017) with default parameters in BARLEYMAP.
The reported end coordinates for each marker were recorded. Markers matching unmapped contigs (chrUn) or with multiple mappings, spanning different chromosomes or more than 200 kb apart, were discarded. After quality checks, 7,479 markers (5,261 Infinium, 2,218 DArTseq) were assigned unique physical positions, and their respective genetic positions (in cM) taken from the closest marker in the POPSEQ2017 map (Beier et al., 2017). The genetic positions of markers SCRI_RS_224392, SCRI_RS_187102, SCRI_RS_167463, BOPA2_12_30895, and BOPA2_12_30894 were interpolated based on their physical distances. Map positions are available in Table S3.

**Calculation of agroclimatic variables**

The raw climatic dataset was provided by the Spanish meteorological agency (AEMET). Daily data from 2,087 observatories of temperature and 6,952 of precipitation, evenly distributed over mainland Spain, were used (Fig. S1). The dataset spanned the most recent standard period of the World Meteorological Organization (WMO, 1981 to 2010, and all the observatories provided more than 10 years of data. A thorough reconstruction procedure was applied to the original precipitation and temperature data by using the reddPrec R package (Serrano-Notivoli et al., 2017a), and following the methodology described in Serrano-Notivoli et al. (2017b). This included: i) an exhaustive quality control to remove anomalous or suspect data; and ii) the imputation of new values to all the missing data, to have serially-complete data series covering the whole study period. This method, originally developed for daily precipitation, was adapted to daily temperature data by applying the quality control through the RClimDex v.1.1 software, developed by the WMO (Zhang and Yang, 2004). The reconstructed data series were used to create, using again the reddPrec R package, three
gridded datasets of daily maximum and minimum temperature and precipitation in the 1981-2010 period, covering the Iberian Peninsula with a spatial resolution of 5x5 km. An example of the daily grids is provided in Fig. S2.

Daily gridded data were then used to compute a set of 147 agroclimatic variables (ACV) that are related to the development of winter cereals, defined in Table 1. Some of them (monthly and seasonal pcp, tmed, tmax and tmin) were derived by temporal aggregation of the daily grids of precipitation, mean, maximum and minimum temperature. Daily data were also used to compute other variables such as the thermal amplitude, tamp, which is the (monthly, seasonal) mean daily difference between tmax and tmin; the number of frost days, frost, or days where tmin < 0 °C; the late-season frost probability, pfrost, which is the average first day in the year where the probability of tmin < 0 °C is lower or equal to 10%, that is, corresponding to a mean return period of one in ten years. Daily data was also used for computing monthly potential vernalization, verna, which accounts for the required exposure to cold temperatures for winter cereal to start flowering. Vernalization was computed at the daily level based on the maximum and minimum temperatures, assuming that temperature describes a sine curve during the day, as done in the CERES-Wheat model (Ritchie, 1991), with temperature thresholds modified for barley as in Ciudad (2002), following a personal communication from Dr. Roger B. Austin. The number of vernalization days was computed assuming that vernalization becomes effective at 0 °C, then increases linearly until 4 °C, and decreases linearly between 8 and 15 °C. No interactions between daily temperature and cumulative degree days or length of the photoperiod were considered as these need to be calibrated for each cultivar, so we called this variable ‘potential vernalization’ as it depends only on climate. In addition, the mean number of days since an average sowing day, estimated as November
178 15th, required to accumulate 10, 20, 30 and 40 potential vernalization days, verna Nd, were
179 also computed. November 15th was considered as a typical sowing date for all Spain, according
180 to the authors’ experience, and to the data reported by Supit and Wagner (1999), who found
181 that 59% of barley sowings all over Spain had occurred by November 10th. We added five more
days to account for seed imbibition, as vernalization acts on active tissues. Reference
182 evapotranspiration data, according to the Penman-Monteith equation as explained in the
183 FAO56 manual (Allen et al., 1998), ETo, were obtained from a gridded dataset for Spain
184 (Vicente-Serrano et al., 2017; Tomas-Burguera et al., 2017). The original data was down
185 sampled from the initial resolution of 1.1 x 1.1 to 5 x 5 km by averaging the values, and the
186 original weekly resolution was aggregated to monthly values. The time period 1981-2010 was
187 selected in accordance to the rest of the climatic data. In addition to monthly, seasonal and
188 annual mean values, ETo was further used to compute the climatic water balance, bal, as the
difference between the cumulative precipitation and the cumulative ETo. The complete
189 climatic dataset is available in Table S4.
190 For each climatic variable with the exception of pfrost and verna Nd, monthly, seasonal and
191 annual averages were calculated based on daily data. Monthly values are indicated after the
192 variable acronym with suffixes _jan to _dec, while seasonal and annual values are denoted by
193 _spr (spring, covering from March to April), _aut (September to November), _win (December
194 to February) and _annual. Summer aggregates and the months between July and October are
195 not expected to have influence on barley growth and, hence, were excluded from further
196 analysis. Additionally, three geographical variables were included: lon, lat and alt, which stand
197 for latitude, longitude and elevation. Latitude and longitude were extracted directly from the
grid structure, and elevation data was obtained from the GTOPO30 digital elevation model developed by USGS (LP DAAC, 1996).

In addition to the environmental variables, 12 dummy variables were generated and included in the dataset. Dummy variables are randomly generated synthetic variables that are incorporated into the analysis in order to test the robustness of the results with respect to type I errors (false positives), since it is known in advance that there is no true relationship between these variables and the dependent variable. Thus, an inflated (unexpectedly high) number of false positives with the dummy variables would raise a warning against the results obtained regarding the true independent variables. In our case the dummy variables were spatially correlated random fields (grids) generated by unconditional Gaussian simulation. The use of spatially correlated random fields instead of pure random (uncorrelated) variables corrects for the inflation of type I errors when this effect is not considered in the analysis of spatial environmental variables (Beguería and Pueyo, 2009). We computed 12 dummy variables by the unconditional Gaussian simulation algorithm using the gstat R package (Pebesma, 2004; and example code in Beguería, 2010), and extracted the values at the landrace collection sites. The degree of spatial correlation in the resulting grids is controlled by a semi-variogram model, which is used for computing the interpolation weights and, depending on its parameterization, can vary from a totally random spatial field to a smoothly varying one (Cressie, 1993). In this case, the parameters of the semi-variogram model used were chosen to ensure a degree of spatial smoothing similar to that of the climatic variables (Fig. S3).
Selection of agroclimatic variables

Exploratory analysis of the agroclimatic variables revealed substantial covariance, as shown in Fig. S4. Consequently, the variables were subjected to hierarchical cluster analysis in order to detect groups of similar variables. The Ward's D2 algorithm (Murtagh and Legendre, 2014) implemented in the R function hclust (R Core Team, 2017) was applied to the Euclidean distance matrix of the variables scaled and centred using the R function scale (R Core Team, 2017). The resulting dendrogram was then cut into 10 clusters (Fig. S5) and one or, at most, two variables representative for each group were then selected, considering periods that matched the growth phases and the occurrence of the main growth milestones of barley, as described in Slafer and Rawson (1994) and Sreenivasulu and Schnurbusch (2012). The duration and dates for these phases and events were estimated by the authors, assuming an average autumn sowing for the Iberian Peninsula. The following 20 variables that were kept for further analyses: lon, lat, alt, pcp_aut, pcp_win, pcp_mar_apr, pcp_may_jun, eto_spr, bal_aut, bal_win, bal_mar_apr_may, bal_jun, temp_win, temp_spr, verna30d, verna_jan_feb, verna_mar_apr, frost_jan_feb, frost_apr_may, and pfrost (Fig. 1). In some cases, multi-month variables were computed by summing the values of the corresponding months, which belonged in the same cluster (for instance, pcp_mar_apr corresponds to the aggregated precipitation of March and April). Examples of these variables are shown in Fig. S6.

Additionally, all the variables analyzed were subjected to a principal component analysis. The R function prcomp (R Core Team, 2017) was applied to the covariance matrix of the variables, after scaling and centering as explained above. The first three principal components (PC1-3), which account for 55%, 17% and 10% of the variance, respectively, were extracted and treated as environmental variables for further analysis. The first principal component of the
environmental variables (PC1) was positively correlated to vernalization, number of frost days, late frost probability and altitude, and negatively correlated to winter and spring temperature (Fig. S7). The second component (PC2) was positively correlated to autumn, winter and spring precipitation and climatic water balance, and negatively correlated with the temperature amplitude in autumn, winter and spring (Fig. S8). The third component was positively correlated with spring temperature amplitude and with the winter climatic water balance, and negatively correlated with the winter potential evapotranspiration and precipitation and water balance in June (Fig. S9). Spatial distribution of PC1 presented high values on the mountain ranges and in the northern half of the Iberian Peninsula. PC2 had high values on the northern and north-western rims of the Iberian Peninsula, as well as some other areas of Atlantic influence. PC3 distribution clearly identified the influence of the Mediterranean in the winter and spring climatology of the Iberian Peninsula (Fig. S10).

**Genome-wide association between biallelic SNPs and agroclimatic variables**

This step was performed with two different software packages that use different approaches, to minimize false positives. Seventeen selected agroclimatic variables, 3 geographic variables, plus 12 dummy variables, were standardized and formatted to be used as input ENVIRONFILE for software Bayenv2 ([https://gcbias.org/bayenv](https://gcbias.org/bayenv), version tguenther-bayenv2_public-8e4039f64d61, Günther and Coop, 2013), and software LFMM_v1.5 (Frichot et al., 2013). For both analyses, we treated each accession as being sampled from a different subpopulation, as in Russell et al. (2016).

In Bayenv2, matrices of covariance between SNP genotypes were computed, to account for the background similarity among landraces. Instead of using all SNPs, a set of 711 non-
redundant markers with linkage disequilibrium $r^2 < 0.2$ (computed on a window of 5 neighbors at each side) and unique genetic positions was shortlisted for this task (Table S5), as recommended by the software developer. Ten runs of Bayenv2, with different random seeds and 100K iterations each, were performed and the average final matrix computed, named $SBCCmatrix_nr_mean.txt$. Fig. S11 shows that this matrix reproduces the known population structure of these barleys. This matrix was used for the conventional Bayenv2 analysis (covariance model). An identity matrix was also formatted to model a null covariance matrix and named $SBCCmatrix_null.txt$, to be used for the analysis disregarding population structure (null model).

A SNPSFILE containing allele counts across 135 barley landraces was formatted, where each SNP is represented by two lines in the file, with the counts of allele 1 on the first line and the counts for allele 2 on the second. The resulting file was named $SBCC_9K_SNPs.tsv$ and used for association mapping with a custom Perl script that parallelized Bayenv2 jobs with the following parameters: `--threads --input-file $SBCC_9K_SNPs.tsv --pairwise --environment-file $SBCC_environfile.tsv --n 21 --mutations $SBCCmatrix_nr_mean.txt --k 100000 --covariance --null`. Five replicates per model (null and covariance), with different random seed each and 100K iterations, were ran and the median Bayes factors (BF) and Spearman correlations (rho) were computed for each marker (Fig. S12). For both models, we report the consensus set of SNPs within the top 1% of Bayes factors distribution that were also in the top 1% of absolute correlations in each of the five runs, for the combined distribution of results for the 20 agroclimatic variables. Additionally, we report thresholds based on the distribution of BF values calculated for the 12 dummy variables, i.e., a true null distribution. The threshold was set at percentile 99.99 of this distribution.
A previous version of LFMM was seen to be highly sensitive to the presence of missing data. Therefore, missing data were imputed using R package limkin, which provides an imputation routine particularly suited to homozygous individuals (Xu et al., 2015). After imputation and filtering for MAF>0.05, the remaining markers were stored in data set SBCC_9K_LFMM.imputed.tr.tsv (n=6,128). The program was run 5 times with 50K cycles, and the same number for the burn-in period. Several sets of latent factors (K) were tested, from 4 (as this is the optimal number of subpopulations detected by STRUCTURE) to 8. We determined that the optimum value, according to the profiles of the histograms of combined adjusted P-values was K=6. The correlation z-scores obtained from the 5 independent runs were combined using the Fisher-Stouffer approach, recommended by the authors, and a FDR threshold (Q=0.01) was calculated for each variable.

**Population differentiation**

Genotypes were classified into genetic clusters using the admixture model of the software package STRUCTURE v.2.3.4 (Falush et al., 2003). Data for 8,457 polymorphic SNPs were used to run STRUCTURE 6 times, setting the number of populations (K) from 1 to 6. For each run, burn-in time and replication numbers were set to 10,000 and 20,000 Monte Carlo Markov Chain (MCMC) iterations, respectively. Evanno’s ΔK (Evanno et al., 2005), as implemented in Structure Harvester (Earl and vonHoldt, 2012), was used to estimate the optimal number of subpopulations (Fig. S13). Then, the program was run one more time using a burn-in period of 100,000 and 100,000 MCMC iterations to estimate membership probability. Additionally, genotypes were classified using factorial analysis with DARwin 6.0.4 (Perrier and Jacquemoud-Collet, 2006), producing similar results as STRUCTURE (Fig. S14).
Following the population differentiation analyses (see Supplementary File 1: SBCC_landraces), landraces were allocated to 4 clusters (Fig. 2, Figs. S13 and S14). Expected heterozygosity per locus (H) and differentiation between populations (Fst) per marker were calculated with Arlequin 3.5 (Excoffier and Lischer, 2010). Bayenv2 and BayPass v2.1 (Gautier, 2015) were used to compute XtX, a statistic analogous to F_{ST} that can identify loci that are more differentiated than expected under pure drift among populations (Günther and Coop, 2013). XtX is more robust than Fst regarding differences in population sizes and independence from underlying genetic structure, as it explicitly accounts for the covariance structure among populations’ allele frequencies (Günther and Coop, 2013). With Bayenv2, three replicates were performed and the average XtX value taken for each SNP marker. The parameters were:

```
-X -t -i SBCC_9K_subpops.tsv -p 4 -m SBCC_nr_subpops_matrix_mean.txt -k 200000 -e envfile.dummy -n 1 -c (Fig. S15).
```

BayPass extends on the model used by Bayenv2, generating a set of theoretically neutral SNPs to help in the inference of significance thresholds for XtX. For this purpose, markers with less than 10% missing data and MAF≥0.05 from the 9920_SNPs_SBCC_50K.tsv file were selected (n=8,457). Linkage disequilibrium between neighboring markers was calculated in R using package LDcorSV (Mangin et al., 2012). We report r_{S}^{2}, which incorporates into the calculation the information about the origins of each individual, i.e., the values of the Q matrix produced by STRUCTURE. This calculation corrects for biases induced by population structure. For each SNP, we calculated r_{S}^{2} values with four SNPs to each side, and the average value is reported. Heterozygosities, XtX and r_{S}^{2} are reported for each single SNP, and also in 4 cM wide sliding windows, calculated with a purpose-made Perl script.
The association of population differentiation with agroclimatic variables was explored further using redundancy analysis. This technique is widely used to test whether the variation in one set of (independent) variables explains the variation in another set of (dependent) variables. We followed an approach similar to the one reported by Leamy et al. (2016). The genetic differences among the SBCC lines, assessed by the four vectors of probabilities of belonging to the four genetic groups (Q) identified by the STRUCTURE analysis were considered as the dependent variables. The complete set of variables, or the reduced set of 17 selected agroclimatic variables (excluding latitude, longitude, altitude and the dummy variables) comprised the independent sets. An independent assessment of the impact of the environment and geography on genetic differentiation among the landraces was assessed by comparing two partial RDA models. One included (besides the matrix with the Q values) all 17 agroclimatic variables, and the other the three geographic variables (latitude, longitude, altitude), in each case adjusted for the other set. By comparing the two models, the common and independent contributions of agroclimatic and geographic effects to the distribution of the genetic groups could be estimated, following the same procedure performed by Lasky et al. (2015). RDA was performed with the vegan package in R (Dixon, 2003). We used a permutational ANOVA-like test on redundancy-analysis fitted data (function anova.cca) to test the significance of the effect of agroclimatic and geographic variables on the distribution of the four genetic groups.

RESULTS

Germplasm groups
Four groups of accessions (comprising 15, 10, 48 and 62 individuals) were identified by STRUCTURE analysis (Fig. 2, Table S1). These groups, with minor variations, corresponded to the populations already identified by Yahiaoui et al. (2008) using SSR markers. Group 1 included 15 six-rowed accessions, related to European winter and spring barleys; group 2 consisted of 10 two-rowed barleys, rather close to spring European 2-rowed types; groups 3 (48 accessions) and 4 (62 accessions), all six-rowed types except one two-rowed in group 3, were widely distributed over the entire Peninsula (Fig. 2), predominantly in inland northern-central regions (group 3) and southern-coastal regions (group 4). The first two groups are closer to other European cultivars, whereas the last two are genetically more distant from European cultivars, as pointed out in Yahiaoui et al. (2008). Fst, a measure of the differences of allelic frequencies between populations, was calculated for the four germplasm groups (Table 2). Differentiation between groups was minimum between groups 3 and 4, and maximum between these and group 2 (2-rowed accessions). The Fst values between group 1 and the rest were intermediate, indicating a central position between them. H, averaged over all SNPs was rather low at the three predominantly 6-rowed groups (1, 3, 4), and higher at the 2-rowed group (2).

Another measure of population differentiation, XtX, was used to search for patterns of genetic differentiation possibly related with the presence of selection footprints. BayPass provided a significance threshold for XtX at 9.56 (i.e., XtX values above it indicate population differentiation above what could be expected for neutral markers). Heterozygosity and LD were examined in high XtX areas, to search for regions that hinted at the presence of selection footprints. The values for XtX calculated with BayPass were lower in general than those calculated with Bayenv2. As BayPass values seem more conservative, and allow the calculation
of a significance threshold, we will present only those. Moreover, the XtX scores calculated
with both programs gave close results ($r=0.83$, Fig. S16). Peaks of the 4 cM sliding windows
scan mark the regions with largest allelic differences across populations, indicating the most
likely regions around genes that may have acted as drivers of differentiation between the
groups (Fig. 3). In barley, genes that govern growth cycle duration (known as flowering time
genes), and spike-type usually diverge among populations (Muñoz-Amatriain et al., 2014).
Although it is no proof of association, it is worth noting that some of these genes fall within
the main regions distinguishing the germplasm groups (Fig. 3, Table S6). The rightmost XtX
peak on 1HL (85.16-98.40 cM, 497-522 Mb) contains the $HvFT3$ ($PpdH2$) gene (514 Mb),
among others. There were significant XtX values on the long arm of chromosome 2H. The most
conspicuous, at 94 cM (226 cumulative cM, ccM), 697-700 Mb, also presented some high LD
values (at 698 Mb), and low heterozygosity. Also in 2H, there was a cluster of high LD values
around the position of gene $HvCEN$ (51.81 cM, 184 ccM, 523 Mb, Tables S6, S7), accompanied
by moderately high XtX values, although not significant. Chromosome 3H showed one of the
main selection footprints, at 46.61-47.20 cM (306.3-306.9 ccM, 238-411 Mb in Fig. 3, Fig. S17,
respectively) covered most of the pericentromeric region and part of both chromosomal arms,
with high XtX, LD and heterozygosity values. This footprint was caused by the contrast
between all four groups, except groups 3 and 4, which were quite similar (Fig. S18a). A few
more XtX significant values were present at the end of 3HL. On 4H, a clear XtX signal was visible
at 29.72 cM, 16-19 Mb, coincident with the position of spike-type gene $int-c$, among other
genes. The highest XtX peak was in chromosome 5H, in a very wide region (33.84-34.43 cM,
582.5-583.1 ccM, Fig. 3, 72-348 Mb, Fig. S17) containing flowering time gene $HvTFL1$ (322 Mb,
Table S6), caused by a sharp contrast between group 4 and the rest (Fig. S18b). The highest
XtX values together with high heterozygosity and LD values occurred flanking the centromere, and extending into both arms (Table S7, Fig. S17). Other significant XtX values and high LD values were scattered on the distal part of the long arm.

**Agroclimatic variables related to germplasm group differentiation**

The redundancy analysis calculated with all the agroclimatic variables distributed the four genetic groups in a triangle shape (Fig. 4) over the triplot representing the first two axes. The first axis separated the two main six-rowed groups (3 and 4), and was related to temperature variables, with colder places on the left and warmer places on the right. The second axis separated groups 3 and 4 from groups 1 and 2. This axis was related to water availability variables, with groups 1 and 2 occurring in regions that are more humid. The analysis with the reduced (17) set of agroclimatic variables produced a very similar result (available in https://eead-csic-compbio.github.io/barley-agroclimatic-association/HOWTORDA.html). The whole set of agroclimatic variables explained close to 80% of the distribution of the germplasm groups, although this result was not significant, given the high number of variables involved. The reduced set of 17 variables explained a significant 37% (28% adjusted $R^2$, $p<0.001$) of the distribution of genetic groups, of which 22% (21 % adjusted $R^2$) was in common with spatial (or geographic) variables (latitude, longitude, altitude), and 7% adjusted $R^2$, still significant ($p=0.005$), was unique (Fig. S19). The spatial variables explained no unique variance after including the agroclimatic variables in the model. The most significant variables were selected via multiple regression. Only two variables entered into the model before the first dummy variable. These variables were pfrost and bal_jun, the first one related to temperature, the second one to water availability during the grain filling period. Together, they explained 24%
of the genetic groups’ distribution, with a 4% of unique variance, still significant \( p=0.002 \).

Spatial variables explained uniquely a non-significant \( p=0.058 \) 1.6% of the variance (Fig. S19).

In a different and complementary approach, we ran a Bayenv2 analysis for agroclimatic variables without taking into account population structure (null model). Under this model (Table 3), we found 905 marker-variable associations above the selected BF and rho thresholds (top 1% for both parameters). Overall, variables related to frost showed the largest number of associations, followed by variables related to geography (longitude, latitude, altitude), vernalization and, variables related to water availability (Table 3, Fig. S20). The sum of number of frost days in the months of January and February \( (\text{frost\_jan\_feb}) \), and the Julian date in which the probability of frost becomes lower than 10% \( (\text{pfrost}) \) were clearly the two variables with the highest number of SNPs associated. We are aware that many of these associations are false positives, caused by population structure, but the main purpose of this part of the study is to find the environmental drivers of population differentiation. The history of cultivation of barley in Spain points at climate adaptation, more than adaptation to different agricultural systems, as drivers of its geographic distribution, because most Spanish barleys were predominantly sown in autumn, in dryland conditions.

To gain more insight in this direction, we compared population differentiation \( (XtX) \) values at SNP level with the BF resulting from the Bayenv2 null analyses. The expectation was that the correlation between these two sets of variables would highlight the environmental traits most likely related with genetic group distribution. Pearson correlations between BF and XtX values were low overall, due to the occurrence of large number of SNPs with BF close to zero. The correlation coefficients for 12 dummy variables provided a baseline for comparison, varying between -0.07 and 0.20. For the null model (Table 3), non-zero correlation scores were mostly
driven by coincidences of high BF values with some XtX peaks, at the same genomic regions for several agroclimatic variables. This was partially expected, due to collinearity between agroclimatic variables. Correlation coefficients between BF and XtX had values clearly above the dummy variables average, mostly for variables related to temperature (vernalization, frost) and water availability. It is remarkable that latitude and longitude, which presented a large number of SNPs with large BF, were not much more related to XtX than the set of dummy variables and, therefore, were probably not related to population differentiation.

There was coincidence of position between some XtX peaks and accumulation of large BFs for agroclimatic variables, particularly for frost, which were the only variables with BF values over 100 (Fig. 5), and vernalization. Some regions presented markers with large BF scores for frost variables, at cM 70 on 2H (202.5 ccM), and at 616, 627, 762, 908, and 962 ccM, but did not show significant XtX values. Two regions, however, presented the largest number of markers with large BF values for frost variables, and the highest number of significant XtX values, on 3H (47 cM, 307 ccM), and 5H (34 cM, 583 ccM), indicating the highest probability for harboring genes relevant for population differentiation due to a differential response to temperature. At 92 cM on 5H (640 ccM), there was another coincidence of significant XtX with large BF for frost variables.

Genomic regions associated with distribution of agroclimatic variables

The Bayenv2 covariance model and LFMM analyses found fewer associations than the null Bayenv2 model, as expected (Fig. S21), and more associations with latitude and longitude than with any other variable (Table S8). There was good agreement overall between the results of the two analyses. Correlation coefficients between the BFs and the -log of the P-values
produced by LFMM varied between 0.48 and 0.73 for each variable (average of 0.63). In these analyses, removing population covariance also entailed removing associations of the agroclimatic variables most related to germplasm group adaptation, to a larger extent than for geographic variables, particularly longitude and latitude. Associations of agroclimatic variables were reduced from 673 (Table 3) to 57 (Table 4), whereas associations with geographic variables fell from 232 (Table 3) to 54 (Table S8). There was only one association per PC, meaning that most of the variation explained by PCs was removed together with population structure. This fact further supports that population differentiation in Spanish barleys related to agroclimatic variables and adaptation more than to spatial divergence. In these analyses, frost-, vernalization- and water-related variables appeared related to a similar degree with genomic regions (Table 4). Thirty-six SNPs presented 57 associations to agroclimatic variables, using the stringent Bayenv2 and LFMM thresholds according to the threshold combining BF and rho values. An even more stringent threshold was calculated for the Bayenv analyses, by taking as threshold the 99.99 percentile of the distribution of BF scores for the 12 dummy variables. This BF score was close to 10, coinciding with the lower limit for a “strong” evidence indicated by Bayes factors, according to the scale of Jeffreys (1961). Markers presented in Table 4 were significant in at least one analysis, and were beyond percentile 99 for the test statistic score for the other one. Twenty three genomic regions were identified, 10 for variables related to water availability (including PC2 and PC3), 11 to temperature (frost, vernalization, temperature amplitude), and two to both. A stretch of 1Mb of the barley genome, to each side of each significant SNP, was examined to search for potential candidate genes consistent with the associations detected (Table 4), using the application BARLEYMAP (Cantalapiedra et al., 2015). In all cases but one, LD decayed to
background levels along the 1Mb region. A variable number of high confidence genes per region was found. It is speculative to point at candidates without further experimental proof. It is remarkable, however, that the two SNPs with highest associations to frost\_jan\_feb at 127 cM on 7H fell within and adjacent to genes HORVU7Hr1G118240 and HORVU7Hr1G118260, respectively, which encode polyamine oxidases. It is noteworthy to mention that this region was uniquely related to frost variables, and nothing else. There were associations of markers BK\_23 and BOPA1\_2208-279 with frost and vernalization variables (Fig. 6). Also, two more markers in the same region were the only ones associated with altitude (which could be a surrogate for temperature. These two markers are within the well-known cluster of cold acclimation CBF genes, specifically inside CBF4.

**DISCUSSION**

Recent studies have attempted to reveal associations between environmental traits and genetic polymorphisms. Although the statistical techniques used were developed with a focus on natural populations (Günther and Coop, 2013, Günther et al., 2016), they are useful to detect meaningful associations in crops as well, using landraces (Abebe et al., 2015, Lasky et al., 2015, Russell et al., 2016, Zhong et al., 2017). These studies, and the present one, are a case in point for the usefulness of GEA approaches to study crop adaptation. The scarcity of studies in this area could be due to the lack of collections of germplasm with appropriate geographic coverage, reliable passport data, and undisputed genetic lineage. The SBCC is an excellent material in this respect. The apparently limited geographic scope of the collection is
compensated by the wide range of climatic conditions occurring in the Iberian Peninsula, and the wide genetic diversity at play.

The climate variables were derived from an extremely fine grid of weather stations and for a period of 30 years. This density of data is almost unprecedented in this kind of studies, and allows for a precise estimation of relevant agroclimatic variables at the places of collection of the accessions. The use of daily data, instead of more common monthly or seasonal values, allowed for the computation of relevant agroclimatic indices that cannot be derived from coarser data. These include variables highly relevant in the seasonal development of barley, such as vernalization indices or variables related to the risk of frost occurrence.

In the null model analysis, i.e., without considering genetic covariance among individuals (population structure), the associations found point to genomic regions truly related to adaptation to agroclimatic variables, but also detect false positives. The patterns of relationships observed, however, offer insights on which agroclimatic features were the main drivers of genetic differentiation of these barley germplasm groups. The highest number of associations were related to temperature, followed by water-related variables. This result indicates the paramount importance of temperature adaptation in the dissemination of germplasm groups arriving and settling in the Iberian Peninsula. Variables related to the frequency of frosts, both in the winter and early spring, gave the highest associations to genetic markers which, simultaneously, presented the highest relationship with population differentiation, followed by variables describing vernalization potential. Therefore, winter temperatures and, to a lesser extent, water availability, played main roles in the distribution of barley germplasm groups arriving in the Iberian Peninsula.
Barley cultivars (like wheat) are roughly divided in spring and winter types. Winter barleys must combine the presence of gene \( VrnH2 \) with an appropriate \( VrnH1 \) allele (von Zitzewitz et al., 2005), to induce a vernalization requirement. Previously, we found a different \( VrnH1 \) allele in each of the two largest germplasm groups of Spanish barleys (Casao et al. 2011a). Eighty four percent of Spanish landraces are winter-types, with a vernalization requirement (Casao et al., 2011b), that differs according to the \( VrnH1 \) allele present (Casao et al., 2011b).

Therefore, a role of \( VrnH1 \) as the driver of genetic group differentiation related to winter temperatures was expected. However, we did not find signals of population differentiation or association to agroclimatic variables around this gene (Fig. 3). The only distinctive feature was the presence of some of the lowest values for \( H \) across the genome. This would be consistent with the existence of selection pressure towards winter alleles in \( VrnH1 \) that could predate group differentiation.

From these analyses, it is not realistic to pinpoint the exact location of the genes responsible for these adaptations of the germplasm groups. It is also worth mentioning that geographical features like latitude and longitude presented a high number of associations with agroclimatic variables but, unlike frost and vernalization variables, had very low to negligible correlations with population differentiation. This distinction is important. Genetic diversity of landraces is expected to show some degree of spatial autocorrelation due to population movements and gene flow. If initial arrivals of the crop, or later movements within the country followed roughly East-West or North-South directions, associations of some markers with latitude and longitude are not surprising. Nevertheless, the fact that these associations were poorly related with population differentiation means that they bear little meaning as drivers of population adaptation.
In the models including population covariance, there were similar number of significant associations with variables related to water availability and temperature (Table 4), besides another twenty-two related to longitude, latitude, and altitude. This indicates that the evolution that occurred locally, once the different groups arrived in the Peninsula, was influenced as much by low temperature-related variables as by water availability variables. Therefore, this germplasm is a potential source of genes and alleles for adaptation to water and temperature features. Barley first arrived in the area around 7.5k YBP (Zapata et al., 2004), meaning that there has been enough time for admixture, local evolution and adaptation of barley. The fact that barley is essentially self-fertilizing does not rule out the relevance of hybridization as a factor to promote gene shuffling. Most experimental evidence of outcrossing rates in cultivated barley yields estimates of around 1%, depending on floral morphology and environmental conditions (Abdel-Ghani et al., 2004, and references therein). Indeed, there is enough evidence confirming that hybridization events have been important in the evolution of many crops, even self-fertilizing ones (Fuller et al., 2011). Intermediate Q values (probabilities of memberships to germplasm groups) of some or the barley accessions (Table S1) strongly indicate the occurrence of partial admixture in Spanish barleys.

Two regions stood out as harboring the most SNPs showing signs of recent selection footprints. Both regions, on 3H (238-411 Mb) and 5H (72-348 Mb) were narrow in genetic distance, but large in physical distance. The region on 5H apparently overlaps with one of the regions highlighted by Fang et al. (2014) as the most important for population differentiation and environmental adaptation in wild barley. The regions on 5H are actually not the same in both studies. As stated by Fang et al. (2014), their 5H region goes from 47 to 52 cM, in their map, pointing at markers that correspond to 371-448 Mb in the current reference genome,
and is located to the right of the centromere, according to the sequence published by Mascher et al. (2017). Therefore, it falls just to the right of our high XtX region. However, looking closer at the results of Fang et al (2014), their high Fst region on 5H includes another 3 markers with high Fst values, located at 38.78-41.45 cM (50 to 107 Mb in the current genome). This last region, to the left of the centromere, overlaps with our high XtX region. The maximum XtX values of the whole genome are located between 72 and 348 Mb, coinciding with some of the largest BF values for frost variables, between 66-354 Mb. This region on 5H actually covers about 41% of the chromosome in physical distance and, under closer inspection (Table S7), two regions are visible, one to each side of the proximal/pericentromeric region. Regarding the other region highlighted by Fang et al. (2014) on 2H, it spanned from 427 to 550 Mb, according to the current reference genome. In that area, there was a high LD signal in our data, coincident with moderately high XtX values, but below the significance threshold (Table S7). One possibility to explain the occurrence of these high LD regions differing among populations would be a suppressed recombination due to inversions that capture locally adapted alleles when two populations are hybridizing (Kirkpatrick and Barton, 2006).

A possible picture, compatible with all these previous data, is the occurrence of genetic differentiation driven by environmental adaptation of barley stocks moving along barley paths of distribution. The study by Muñoz-Amatriain et al. (2014) detected high genetic differentiation among barley germplasm groups, representing the whole world, at the 5H region mentioned in the previous paragraph. In this region of 5H, there is co-location of population differentiation and association of markers to agroclimatic variables. We cannot conclude that these two facts are causally related, due to the aforementioned lack of
protection against false positives. The commonalities found, however, are striking, and confirm the interest of this genomic region for further research.

The number of SNPs associated to agroclimatic variables was low, only 36, i.e., 0.5% of all SNPs with position. This low number is probably caused by the stringent statistical thresholds used, and is in line with the expectations that only a small portion of the genome will be associated with adaptation to climate (Meirmans, 2015). An examination of the genes present in the reference genome near the markers significant under the population covariance model found a high number of gene models, from which it is difficult to pinpoint potential candidates. A few of the genes harbouring SNPs shown in Table 4 had also prior information linking them to effects related to responses to the agroclimatic variables associated. These genes will be discussed in more detail.

Markers, BOPA1_946-2500 and BOPA1_1977-1385, associated to water availability variables, occur within genes that could be linked to stress responses, such as sucrose synthase 4 and Protein dehydration-induced 19 homolog. Marker BOPA2_12_10979 corresponds to a gene coding for an isoprenyl transferase, and was associated to variables describing vernalization potential. This kind of genes has been connected to water stress responses (Pei et al., 1998) and oxidative stress in plants (Grassman et al., 2002). Marker BOPA1_4025-300 occurs within a gene coding for a cathepsin B-like cysteine proteinase. These enzymes have been reported to increase their expression in barley in response to cold treatment (Martínez et al., 2003), and to provide frost protection in wheat (Talanova et al., 2012), although the association in our study occurred with a variable related to water availability.
Two marker-agroclimatic associations deserve further comment. A distal region on the long arm of 7H, around 127 cM, appeared related to frost variables under both the null and the covariance models. The marker with the largest BF for number of days of frost in January-February was actually within a gene coding for a polyamine oxidase. Polyamines have been frequently reported as part of the response of *Arabidopsis thaliana* to abiotic stresses and, in particular, to cold tolerance (Cuevas et al., 2008). The diverse roles of several polyamines in response to cold stress in crop species, particularly in winter cereals, was revealed in a number of studies (reviewed by Pecchioni et al., 2014). Wang et al. (2015) also found a role for polyamines in wheat, not only in cold stress responses, but also in heat stress responses, as already pointed out previously by Goyal and Asthir (2010). Further experimental evidence is needed to confirm whether or not they have any role in barley adaptation.

Finally, BOPA1_2208-279 and BK_23 also fall within a potential candidate gene. These markers are placed within *CBF4*, the last of the well-known cluster of *CBF* genes on chromosome 5H, identified as the responsible for the most important frost tolerance QTL in barley, *FrH2* (Francia et al., 2007; Francia et al., 2016). Additionally, markers 3258527 and SCRI_RS_224251 were the only ones associated with altitude, which could be a proxy for frost variables. They are located just 600 bp and 7Mb away from *CBF4*, so they probably are part of the same signal.

*CBFs* and their regulon are major determinants of low-temperature tolerance. *CBFs* are transcription factors that regulate suites of genes during drought and low temperature stresses. Their evolution was involved in Pooideae adaptation to cold climates (Sandve and Fjellheim, 2010). Specifically, members of the CBF3/4-subfamilies are thought to play roles in Pooideae adaptation to freezing stress (Li et al., 2012). Their role in frost tolerance has been related to different temperatures of threshold induction, different expression levels (Galiba
et al., 2009, Knox et al., 2010) and to copy number variation (Tondelli et al., 2011; Francia et al., 2016). CNV occurrence is highly dynamic at this locus and, therefore, it is not surprising that it appears linked to altitude and frost occurrence in a collection of landrace germplasm. The identification of associations of markers in CBF4 as associated to vernalization and frost variables is a case in point that validates the GEA strategy followed.

These findings reveal new information about the environmental drivers of genetic diversification of barley. Winter temperatures, affecting frost occurrence and vernalization potential, are behind the genetic differences between the groups of Spanish barley landraces. The relevance of this conclusion exceeds the local interest, as the Iberian Peninsula is the endpoint of the routes of Neolithic expansion that encompassed the entire Mediterranean basin. In some cases, we were able to pinpoint new candidate genes associated to specific environmental conditions that open further avenues for research on cereal crops adaptation.

Given the genetic closeness of wheat and barley, and the parallel history of expansion of these crops, the occurrence of similar processes in wheat deserves investigation. This knowledge, after further experimental validation and allele mining, will be applicable to pre-breeding and breeding of barley.
ACKNOWLEDGEMENTS

We thank Torsten Günther for advice on the use Bayenv2. This work was supported by the FACCE ERA-NET Plus project ClimBar (618105) and by Spanish Agencia Estatal de Investigación projects CGL2014-52135-C3-3-R, AGL2013-48756-R, AGL2016-80967-R and RFP2015-00006-00-00. BCM was funded by Fundación ARAID.

REFERENCES


Casas, A. M., Contreras-Moreira, B., Cantalapiedra, C. P., Sakuma, S., Gracia, M. P., Moralejo, M., ... Igartua, E. (2018). Resequencing the *Vrs1* gene in Spanish barley landraces revealed...


37
DATA AND CODE ACCESSIBILITY

Climate data, genotype and marker information are provided in the supplementary excel file. All the data files and scripts used for the selection of climate variables and the association analyses are described in R markdown documents, available at https://eead-csic-combio.github.io/barley-agroclimatic-association, doi:10.5281/zenodo.1886991

AUTHOR CONTRIBUTIONS

EI, AMC, SB, BCM, designed the study. RSN, SB, collected the climatic data, prepared the climate database, and derived the agroclimatic variables. AMC, NEM, carried out the genetic diversity analyses. BCM, CPC, performed the Bayenv, LFMM and BayPass analyses. All authors performed statistical analyses linking geographic and genetic data. AMC, CPC, BCM, curated the genotypic data and prepared genotypic databases. AMC secured funding. AMC, EI, RSN, SB, BCM, wrote the manuscript. All authors edited and approved the manuscript.
Table 1. List of agroclimatic and geographic variables used on this study. One hundred and forty-seven variables were initially used, from which twenty were selected as most representative and least redundant via cluster analysis.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Variable description and unit</th>
<th>Scale of aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>pcp</td>
<td>average cumulative precipitation (mm)</td>
<td>season, month</td>
</tr>
<tr>
<td>tmed</td>
<td>average daily mean temperature (°C)</td>
<td>season, month</td>
</tr>
<tr>
<td>tmax</td>
<td>average daily max temperature (°C)</td>
<td>season, month</td>
</tr>
<tr>
<td>tmin</td>
<td>average daily min temperature (°C)</td>
<td>season, month</td>
</tr>
<tr>
<td>tamp</td>
<td>average daily thermal amplitude (°C)</td>
<td>season, month</td>
</tr>
<tr>
<td>frost</td>
<td>average number of frost days (-)</td>
<td>season, month</td>
</tr>
<tr>
<td>pffrost</td>
<td>average first day in the year where P(tmin&lt;0) ≤ 0.10</td>
<td>annual</td>
</tr>
<tr>
<td>verna</td>
<td>average potential vernalization (days) (-)</td>
<td>month</td>
</tr>
<tr>
<td>verna_nd</td>
<td>average number of days since 15th November to reach n = 10, 20, 30 and 40 vernalization days (-)</td>
<td>annual</td>
</tr>
<tr>
<td>ET₀</td>
<td>average cumulative reference evapotranspiration (mm)</td>
<td>annual, season, month</td>
</tr>
<tr>
<td>bal</td>
<td>climatic water balance (pcp - eto) (mm)</td>
<td>annual, season, month</td>
</tr>
<tr>
<td>lon</td>
<td>longitude, in UTM zone 30N projection (km)</td>
<td>—</td>
</tr>
<tr>
<td>lat</td>
<td>latitude, in UTM zone 30N projection (km)</td>
<td>—</td>
</tr>
<tr>
<td>alt</td>
<td>elevation (m above mean sea level)</td>
<td>—</td>
</tr>
<tr>
<td>dummy</td>
<td>random data with spatial coherence (-)</td>
<td>—</td>
</tr>
</tbody>
</table>
Table 2. Measure of population differentiation (Fst) between the four barley germplasm groups identified (for codes, see text). Diagonal, average diversity values (heterozygosities).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.217</td>
<td>0.242</td>
<td>0.246</td>
<td>0.227</td>
</tr>
<tr>
<td>2</td>
<td>0.331</td>
<td>0.378</td>
<td>0.352</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.196</td>
<td>0.186</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>0.224</td>
</tr>
</tbody>
</table>
Table 3. Relationship of agroclimatic variables with molecular markers. Number of SNPs associated to agroclimatic variables above the combined BF and rho threshold (see full explanation in the Materials and Methods section) for the Bayenv2 null model. Also shown, Pearson correlation coefficients between BF and genetic differentiation between the 4 germplasm groups (XtX, calculated with BayPass). The standard deviation for the correlation coefficients with the 12 dummy variables is also shown.

<table>
<thead>
<tr>
<th>Variables</th>
<th>#associations</th>
<th>r BF-XtX</th>
</tr>
</thead>
<tbody>
<tr>
<td>dummies (12)</td>
<td>5</td>
<td>0.027 ± 0.08</td>
</tr>
<tr>
<td>Alt</td>
<td>74</td>
<td>0.256</td>
</tr>
<tr>
<td>bal_aut</td>
<td>0</td>
<td>-0.066</td>
</tr>
<tr>
<td>bal_jun</td>
<td>34</td>
<td>0.114</td>
</tr>
<tr>
<td>bal_mar_apr_may</td>
<td>0</td>
<td>0.180</td>
</tr>
<tr>
<td>bal_win</td>
<td>3</td>
<td>-0.051</td>
</tr>
<tr>
<td>ET0_spr</td>
<td>12</td>
<td>0.264</td>
</tr>
<tr>
<td>frost_apr_may</td>
<td>11</td>
<td>0.238</td>
</tr>
<tr>
<td>frost_jan_feb</td>
<td>193</td>
<td>0.192</td>
</tr>
<tr>
<td>P_frost</td>
<td>233</td>
<td>0.295</td>
</tr>
<tr>
<td>Lat</td>
<td>70</td>
<td>0.103</td>
</tr>
<tr>
<td>Lon</td>
<td>78</td>
<td>0.015</td>
</tr>
<tr>
<td>pcp_aut</td>
<td>9</td>
<td>-0.018</td>
</tr>
<tr>
<td>pcp_mar_apr</td>
<td>0</td>
<td>-0.061</td>
</tr>
<tr>
<td>pcp_may_jun</td>
<td>17</td>
<td>0.145</td>
</tr>
<tr>
<td>pcp_win</td>
<td>11</td>
<td>-0.023</td>
</tr>
<tr>
<td>tamp_spr</td>
<td>0</td>
<td>0.071</td>
</tr>
<tr>
<td>tamp_win</td>
<td>0</td>
<td>0.033</td>
</tr>
<tr>
<td>verna_30d</td>
<td>40</td>
<td>0.268</td>
</tr>
<tr>
<td>verna_jan_feb</td>
<td>85</td>
<td>0.299</td>
</tr>
<tr>
<td>verna_mar_apr</td>
<td>30</td>
<td>0.263</td>
</tr>
</tbody>
</table>
Table 4. Reference genome search for the markers with highest BF factors (Bayenv) and lowest P-values (LFMM) within each genomic region associated with an agroclimatic variable (other than longitude and latitude). Highlighted, significant values with the stringent thresholds for each method. Markers reported had significant values for at least one of the analyses, and were beyond percentile 99 for the distribution of the other. The search encompassed a window of 1Mb at each side of each marker. The number of neighbor gene models found in these windows is broken down into number of high confidence, low confidence and unclassified genes. Gene hits are provided when the SNP falls within a gene. Accessions of gene models and descriptions are provided in the two rightmost columns.

<table>
<thead>
<tr>
<th>Marker</th>
<th>chr</th>
<th>cM</th>
<th>Position (bp)</th>
<th>Agroclimatic variable</th>
<th>BF Bayenv</th>
<th>-log10(P) LFMM</th>
<th>neighbor genes (HORVU)</th>
<th>gene_hit</th>
<th>gene_description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3255550</td>
<td>F</td>
<td>0</td>
<td>1H 114.08 533885730</td>
<td>pcp_may_jun</td>
<td>7.5</td>
<td>5.62</td>
<td>54 60 0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3255919</td>
<td>F</td>
<td>0</td>
<td>2H 0.6 1839535</td>
<td>frost_y_feb pfpfrost</td>
<td>13.2</td>
<td>4.28</td>
<td>54 49 0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3256699</td>
<td>F</td>
<td>0</td>
<td>2H 31.37 46833247</td>
<td>frost_y_feb pcp_may_jun</td>
<td>10.9</td>
<td>3.77</td>
<td>31 16 2</td>
<td>2Hr1G013830</td>
<td>Protein WEAK CHLOROPLAST MOVEMENT UNDER BLUE LIGHT 1</td>
</tr>
<tr>
<td>3261758</td>
<td>F</td>
<td>0</td>
<td>2H 31.37 46833484</td>
<td>pcp_may_jun</td>
<td>14.3</td>
<td>5.23</td>
<td>31 16 2</td>
<td>2Hr1G013830</td>
<td>Protein WEAK CHLOROPLAST MOVEMENT UNDER BLUE LIGHT 1</td>
</tr>
<tr>
<td>SCR1_RS_144592</td>
<td>2H</td>
<td>49.43 114309420</td>
<td>bal_jun</td>
<td>14.1</td>
<td>6.50</td>
<td>10 8 1</td>
<td>2Hr1G030810</td>
<td>F-box only protein 13</td>
<td></td>
</tr>
<tr>
<td>3262943</td>
<td>F</td>
<td>0</td>
<td>2H 49.43 114710786</td>
<td>bal_jun</td>
<td>10.0</td>
<td>7.28</td>
<td>12 11 1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BOPA1_946-2500</td>
<td>2H</td>
<td>49.43 115322116</td>
<td>bal_jun</td>
<td>13.0</td>
<td>7.43</td>
<td>11 15 1</td>
<td>2Hr1G030870</td>
<td>sucrose synthase 4</td>
<td></td>
</tr>
<tr>
<td>BOPA1_2601-171</td>
<td>2H</td>
<td>49.43 118999843</td>
<td>bal_jun</td>
<td>15.7</td>
<td>4.73</td>
<td>14 13 1</td>
<td>2Hr1G031310</td>
<td>Protein TIFY 3B</td>
<td></td>
</tr>
<tr>
<td>BOPA1_Consensus</td>
<td>2H</td>
<td>63.77 636292482</td>
<td>bal_jun</td>
<td>11.2</td>
<td>5.74</td>
<td>11 15 0</td>
<td>2Hr1G089020</td>
<td>Disease resistance protein</td>
<td></td>
</tr>
<tr>
<td>BOPA1_1613-291</td>
<td>2H</td>
<td>104.48 723245814</td>
<td>bal_mar_apr</td>
<td>25.1</td>
<td>5.36</td>
<td>36 28 0</td>
<td>2Hr1G111600</td>
<td>Adenine nucleotide alpha hydrolases-like superfamily protein</td>
<td></td>
</tr>
<tr>
<td>3256997</td>
<td>F</td>
<td>0</td>
<td>2H 108.11 727985335</td>
<td>pcp_may_jun</td>
<td>9.5</td>
<td>2.40</td>
<td>37 26 0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3261860</td>
<td>F</td>
<td>0</td>
<td>2H 121.46 751637636</td>
<td>tamp_spr</td>
<td>15.4</td>
<td>7.29</td>
<td>79 26 6</td>
<td>2Hr1G121480</td>
<td>unknown function</td>
</tr>
<tr>
<td>BOPA1_1283-332</td>
<td>2H</td>
<td>121.46 751886865</td>
<td>tamp_spr</td>
<td>35.6</td>
<td>7.06</td>
<td>77 29 6</td>
<td>2Hr1G121990</td>
<td>calreticulin 1b</td>
<td></td>
</tr>
<tr>
<td>Gene ID</td>
<td>Dataset</td>
<td>Fold Change</td>
<td>p-value</td>
<td>q-value</td>
<td>E-value</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCRI_RS_123364</td>
<td>2H</td>
<td>127.47</td>
<td>763961511</td>
<td>10.5</td>
<td>3.68</td>
<td>72</td>
<td>6</td>
<td>2Hr1G126610</td>
<td>transportin 1</td>
</tr>
<tr>
<td>SCRI_RS_186444</td>
<td>3H</td>
<td>22.58</td>
<td>27751234</td>
<td>13.9</td>
<td>3.72</td>
<td>33</td>
<td>27</td>
<td>1</td>
<td>3Hr1G012800</td>
</tr>
<tr>
<td>3260736</td>
<td>3H</td>
<td>32.86</td>
<td>32635613</td>
<td>10.1</td>
<td>4.25</td>
<td>15</td>
<td>9</td>
<td>0</td>
<td>3Hr1G061550</td>
</tr>
<tr>
<td>BOPA1_1977-1385</td>
<td>3H</td>
<td>47.2</td>
<td>46826574</td>
<td>12.4</td>
<td>4.40</td>
<td>12</td>
<td>10</td>
<td>0</td>
<td>3Hr1G061690</td>
</tr>
<tr>
<td>SCRI_RS_220192</td>
<td>3H</td>
<td>58.51</td>
<td>535469596</td>
<td>9.4</td>
<td>4.10</td>
<td>13</td>
<td>10</td>
<td>0</td>
<td>3Hr1G070850</td>
</tr>
<tr>
<td>SCRI_RS_225540</td>
<td>3H</td>
<td>58.51</td>
<td>536074314</td>
<td>3.6</td>
<td>4.61</td>
<td>21</td>
<td>36</td>
<td>1</td>
<td>3Hr1G070860</td>
</tr>
<tr>
<td>BOPA2_12_30399</td>
<td>3H</td>
<td>59.11</td>
<td>544458340</td>
<td>7.2</td>
<td>5.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOPA1_4025-300</td>
<td>3H</td>
<td>97.75</td>
<td>624282646</td>
<td>21.1</td>
<td>4.33</td>
<td>37</td>
<td>18</td>
<td>2</td>
<td>3Hr1G088270</td>
</tr>
<tr>
<td>BOPA1_6841-637</td>
<td>4H</td>
<td>89.39</td>
<td>599525496</td>
<td>23.8</td>
<td>2.58</td>
<td>24</td>
<td>19</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3256603</td>
<td>5H</td>
<td>59.53</td>
<td>529391379</td>
<td>3.6</td>
<td>5.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOPA1_2208-279</td>
<td>5H</td>
<td>87.29</td>
<td>560732601</td>
<td>7.2</td>
<td>5.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BK_23</td>
<td>5H</td>
<td>87.29</td>
<td>560732648</td>
<td>7.7</td>
<td>5.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOPA1_1628-410</td>
<td>6H</td>
<td>23.49</td>
<td>16749790</td>
<td>2.8</td>
<td>4.44</td>
<td>75</td>
<td>56</td>
<td>1</td>
<td>6Hr1G009400</td>
</tr>
<tr>
<td>3254663</td>
<td>6H</td>
<td>54.48</td>
<td>396127093</td>
<td>30.2</td>
<td>4.23</td>
<td>24</td>
<td>13</td>
<td>0</td>
<td>6Hr1G059780</td>
</tr>
<tr>
<td>SCRI_RS_226361</td>
<td>7H</td>
<td>2.41</td>
<td>5168993</td>
<td>13.3</td>
<td>4.66</td>
<td>64</td>
<td>57</td>
<td>1</td>
<td>7Hr1G002810</td>
</tr>
<tr>
<td>Marker</td>
<td>Start</td>
<td>End</td>
<td>PI</td>
<td>PI2</td>
<td>PI3</td>
<td>7HR1G027120</td>
<td>Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>----------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>----------------</td>
<td>----------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pcp_aut</td>
<td>12.3</td>
<td>3.92</td>
<td></td>
<td></td>
<td></td>
<td>7HR1G027120</td>
<td>Isoprenyl transferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC2</td>
<td>36.3</td>
<td>4.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOPA2_12_10979^a</td>
<td>7H 43.38</td>
<td>48001348</td>
<td>12.8</td>
<td>2.71</td>
<td>28</td>
<td>7HR1G113760</td>
<td>Transcription initiation factor TFIID subunit 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOPA2_12_31241</td>
<td>7H 116.35</td>
<td>638089288</td>
<td>83.1</td>
<td>5.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOPA1_8758-564</td>
<td>7H 120.53</td>
<td>642239530</td>
<td>16.4</td>
<td>5.68</td>
<td>61</td>
<td>7HR1G115780</td>
<td>30S ribosomal protein S5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOPA1_5595-297</td>
<td>7H 127.15</td>
<td>647000345</td>
<td>21.2</td>
<td>4.93</td>
<td>45</td>
<td>7HR1G118230</td>
<td>Protein MEMO1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3257624</td>
<td>F</td>
<td>0</td>
<td>7H 127.15</td>
<td>647041453</td>
<td>16.8</td>
<td>5.11</td>
<td>7HR1G118240</td>
<td>Polyamine oxidase 1</td>
<td></td>
</tr>
</tbody>
</table>

^a The geographic distributions of these markers are presented in Fig. 6.
Figure 1. Diagram illustrating the agroclimatic variables selected for this study, matched to the milestones and phases of growth of the barley plant. Latitude, longitude and altitude are not represented. Dates correspond to an average autumn sowing in the Iberian Peninsula. Phases and milestones adapted from Slafer and Rawson (1994), and Sreenivasulu and Schnurbusch (2012).
Figure 2. Top: group membership probabilities resulting from the Structure analysis run with a 100K burn-in and 100K MCMC iterations, for K=4 subpopulations. Cluster 1, blue, 6-rowed barleys closer to European cultivars; cluster 2, yellow, 2-rowed barleys; cluster 3, red, 6-rowed barleys; cluster 4, green, six-rowed barleys. Bottom: geographic distribution of the four subpopulations over elevation in mainland Spain (colour coded as in the top graph), represented in a UTM-30N projection, axes in meters.
Figure 3. Genome wide diversity and genetic differentiation among germplasm groups of Spanish barleys, presented both as 4cM sliding windows (lines), and as single SNPs (grey dots) with values beyond significance thresholds (XtX) or beyond the 1 (H) or 99 percentiles (LD). a) Four-cM sliding windows for XtX values calculated for the four germplasm groups with BayPass (solid black line), the black dashed line indicates the 95 percentile value for the 4cM window scores; single SNP values above the BayPass threshold, 9.56, are marked with grey dots; b) heterozygosity (H) values for all accessions in 4 cM sliding windows (solid blue line), with a reference line drawn at percentile 5 (dashed blue line); single SNP values marked with grey dots indicate values below percentile 1; c) linkage disequilibrium corrected for population structure
(r^2_s) averaged for groups of 9 contiguous SNPs, in 4cM windows (solid red line), with a reference line (dashed red) drawn at the 95 percentile calculated separately at the three chromosomal zones described by Mascher et al. (2017), in which vertical spikes indicate the positions of centromeres, and single SNP values above percentile 99 are indicated with grey dots; d) conceptual summary of graphs a-c, vertical scale offset, displaying positions of XtX and r^2_s sliding window values above percentile 95 and heterozygosity values below percentile 5. In graph d), color-coded dots (enlarged for better visualization) correspond to positions in which 4 cM sliding windows values exceed those thresholds. Crosses at the bottom of this graph indicate the positions of well-known flowering time and domestication genes, according to the updated POPSEQ map (Beier et al 2017), and listed in Supplementary File 3.
Figure 4. Triplot of the first two axes of a redundancy analysis with 103 agroclimatic variables (geographic variables longitude, latitude and altitude removed), showing genotypes, variables and germplasm groups. Variables are colour-coded according to each category. The two variables which explained most variance in a multiple regression analysis are indicated with diamond icons.
Figure 5. Association between XtX scores (population differentiation) and BF for selected agroclimatic variables. Red circles indicate XtX scores among germplasm groups (only values above 8 are shown for better visualization); the dashed red line establishes the XtX significance threshold, at 9.56. Black circles indicate BF (only values above 25, for better visualization) for 3 frost agroclimatic variables, and 3 vernalization variables (blue times signs).
Figure 6. Maps of four agroclimatic variables related to vernalization (verna_mar_apr), late frost hazard (pstorm), hydric status (pcp_win) and drought (bal_jun). Four SNPs are shown as squares, circles and triangles placed in the original locations where landraces were collected.