Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation

Russell, Joanne; Mascher, Martin; Dawson, Ian K.; Kyriakidis, Stylianos; Calixto, Cristiane; Freund, Fabian; Bayer, Micha; Milne, Iain; Marshall-Griffiths, Tony; Heinen, Shane; Hofstad, Anna; Sharma, Rajiv; Himmelbach, Axel; Knauft, Manuela; van Zonneveld, Maarten; Brown, John; Schmid, Karl; Kilian, Benjamin; Muehlbauer, Gary J.; Stein, Nils; Waugh, Robert

Published in:
Nature Genetics

DOI:
10.1038/ng.3612

Publication date:
2016

Document Version
Peer reviewed version

Citation for published version (APA):
Russell, J., Mascher, M., Dawson, I. K., Kyriakidis, S., Calixto, C., Freund, F., ... Waugh, R. (2016). Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. Nature Genetics, 48, 1024-1030. DOI: 10.1038/ng.3612

Title: Adaptation of barley to different environments revealed in the exomes of landraces and wild relatives.

Running Title: Adaptation in barley

Joanne Russell[1,10], Martin Mascher[2,3,10], Ian K Dawson[1], Stylianos Kyriakidis[1], Cristiane Calixto[4], Fabian Freund[5], Micha Bayer[1], Iain Milne[1], Tony Marshall-Griffiths[1], Shane Heinen[6], Anna Hofstad[6], Rajiv Sharma[2,4], Axel Himmelbach[2], Manuela Knauft[2], Maarten van Zonneveld[7], John WS Brown[1,4], Karl Schmid[5], Benjamin Kilian[2,8], Gary J. Muehlbauer[6,9,11], Nils Stein[2,11] and Robbie Waugh[1,4,11]

[1]Cell and Molecular Sciences, The James Hutton Institute, Invergowrie, Dundee, DD2 5DA

[2]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, OT Gatersleben, 06466 Stadt Seeland, Germany

[3]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, D-04103 Leipzig, Germany

[4]Division of Plant Sciences, School of Life Sciences, The University of Dundee, The James Hutton Institute, Invergowrie, Dundee, DD2 5DA

[5]Department of Crop Biodiversity and Breeding Informatics, University of Hohenheim, Stuttgart, Germany

[6]Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108

[7]Bioversity International, Costa Rica Office, Turrialba, Costa Rica

[8]Bayer CropScience NV, Innovation Centre, BCS Breeding & Trait Development, Technologiepark 38, 9052 Zwijnaarde (Gent), Belgium

[9]Department of Plant Biology, University of Minnesota, St. Paul, MN 55108

[10] contributed equally to this work

[11] joint corresponding authors

Author for Correspondence: Robbie Waugh
The James Hutton Institute
Invergowrie
Dundee DD2 5DA
Scotland, UK
Tel: (44) 1382 568734 E-mail: robbie.waugh@hutton.ac.uk

After domestication, during a process of widespread range extension, barley adapted to a broad spectrum of agri-environments. To explore how the barley genome responded to the environmental challenges it encountered, we sequenced the exomes of a collection of 267 geo-referenced landraces and wild accessions. A combination of genome-wide analyses showed that patterns of variation have been strongly shaped by geography and that individual gene variant-by-environment associations are prominent in our dataset. We observed significant correlations between days to heading (flowering) and height, with seasonal temperature and dryness variables in common garden experiments, suggesting these were major drivers of environmental adaptation in sampled germplasm. A detailed analysis of known flowering-associated genes revealed that many contain extensive sequence variation and that patterns of single- and multi-gene haplotypes exhibit strong geographical structuring. While these appear to have contributed significantly to range-wide eco-geographical adaptation, many unidentified factors remain key to regional success. (148 words)

Producing sufficient food for a growing human population during a period of restricted natural resources, climate change and competition for land and water is a key global challenge[1]. Primary production needs to increase, and crops that use resources more efficiently and display increased resilience to unpredictable climatic events urgently need to be developed[2]. As the use of natural genetic variation in plant breeding continues to underpin the improvement of all our major crops, identifying and understanding the mechanisms by which selection of derived or pre-existing genetic variants have permitted crop plants to flourish in new and challenging environments would simultaneously provide the knowledge and the germplasm to develop resilient varieties equipped to cope with predicted environments across a wide geographical range[3].

Barley is an ancient crop[4]. It is the world's 4th most important cereal and is especially important as a staple in regions most vulnerable to climate change[5]. It is also highly adapted. Since its domestication in the near east Fertile Crescent over 10,000 years ago, it has undergone extensive and well-documented range expansion[6]. Regionally known as 'the last crop before the desert', it is now cultivated from the Arctic Circle to the equatorial highlands and to southerly latitudes. As such its ability to yield in marginal environments could increase in importance in the coming decades. Here we explore the variation recorded in its genome by sequencing the exomes[7] of a representative geo-referenced collection of locally-adapted barley landraces from across the geographic range of the species and its progenitor wild genotypes, and begin to elucidate genomic signatures that underlie barleys' adaptive responses.

RESULTS

Characterisation of molecular diversity

We generated and analyzed exome sequence data from 267 barley accessions, primarily wild material and landraces that we had purified by two rounds of single seed descent (SSD) to eliminate accession heterogeneity and reduce residual heterozygosity (**Supplementary Table 1**). Sequencing all 267 samples generated 30.7 billion paired-end reads with an average of 108 million reads per accession. These were mapped against the barley draft genome assembly[8,9], yielding a total of 59.5 Mbp of genomic sequence covered by at least ten reads in ≥ 95% of the accessions. Analysis of this dataset with the GATK[10] pipeline and customized scripts (**Supplementary Note**) revealed a total of 1,688,807 single nucleotide

polymorphisms (SNPs) and 143,872 short insertions or deletions (indels). We assigned eighty-two percent of variants to approximate chromosomal locations using an ultra-dense POPSEQ linkage map of the barley genome[11]. On average, only 1.7% heterozygous genotype calls were observed per variant position, as expected for an inbreeding species with low outcrossing rates subject also to SSD[12]. We observed good concordance between our variant calls and SNP array[13] whole genome datasets (> 98%). The majority (64.5%) of SNPs were located in annotated exons and distributed across 20,729 'high-confidence' and 20,453 'low-confidence' genes. The prediction of functional effects of SNP variants in near-complete, high-confidence genes revealed 148,560 synonymous and 157,363 non-synonymous variants, a deficiency of non-synonymous SNPs compared to neutral expectations that could indicate purifying selection[14]. 2,325 SNPs gave rise to premature stop codons, changed start codon positions or affected splice sites. Data are summarized in **Supplementary Table 2**.

Overall, genome-wide diversity was highly structured in our collection. Principal components analysis (PCA) partitioned the 267 accessions into two clearly defined groups (**Fig. 1**). In general terms, the first principal component (PC) separated wild and landrace germplasm. A number of accessions that were classified as wild were genetically close to cultivated material and may represent ancestral forms, hinting at multiple independent domestication events. However, as they occur in sympatry[15] we cannot exclude the possibility that they represent either feral forms or unintended mix-ups. We therefore removed accessions whose domestication state and provenance were ambiguous, along with those with ≥ 5% missing and/or ≥ 5% heterozygous data. This left 228 accessions (91 classed as *H. vulgare* ssp. *spontaneum* [wild or 'spontaneum'] and 137 as *H. vulgare* ssp. *vulgare* ['landrace'] for which we had very high quality SNP data, and accurate longitude and latitude collection site information (**Online Methods**).

Within our collection the majority of SNPs were rare (minor allele frequency below 5%, **Supplementary Fig. 1a-c**). We detected a higher proportion of rare SNPs at non-synonymous than synonymous sites, a situation also observed in wheat landraces[16] and attributed in humans to purifying selection[17], which is consistent with our observations on the overall numbers of SNPs in non-synonymous and synonymous categories (see above). The excess of rare variants was more pronounced in the wild individuals that carried on average 5,191 private alleles, representing a large pool of standing genetic variation that could be exploited in future crop improvement. In contrast, the domesticated landrace barleys exhibited an average of 429. This is consistent with the bottleneck of domestication in landraces resulting in a loss of diversity relative to wild progenitors[18].

A conspicuous feature of the genomes of barley and other Triticeae is the severely reduced recombination in the pericentromeric and centromeric regions that on a physical level occupy approximately two-thirds of each chromosome[19,20]. These regions were characterized by reduced diversity, which was strongest on chromosomes 1H, 4H and 7H, confirming previous studies[8,21]. Across the genome we observed an average reduction of nucleotide diversity (RoD) of 27% in landraces relative to wild barleys, which is similar to previous estimates[15,22] (**Fig. 2a**). Some regions that exhibited a strong RoD (measured for 5 cM intervals) corresponded to known domestication loci. The RoD maximum observed on the long arm of 7H (75 to 80 cM) contains the *NUDUM (HvNUD)* locus that is responsible for the free-threshing phenotype selected by early farmers who used barley grain for human

consumption[23]. A second region on chromosome 3H contained *BRITTLE RACHIS 1* and *2* (*HvBTR1* and *HvBTR2)*, two domestication genes that abolish seed shattering in cultivated barley[24]. A third on chromosome 4H contained, amongst others, the flowering-associated genes *HvFKF1,* and *HvPRR59*, while three further regions on 1H and 7H did not correspond with known domestication loci. Reduced nucleotide diversity is commonly associated with low recombination rates (*ρ)* (**Fig. 2b**), the formation of long haplotype blocks, and selection constraints at the level of sequence polymorphism[25]. These are each reflected in the higher levels of linkage disequilibrium (LD) observed in the landrace group (**Supplementary Fig. 2**). Differentiation between groups (fixation index F_{ST}) (**Fig. 2c**) is greatest on chromosome 2H and 3H near *HvBTR1/HvBTR2,* while chromosome 6H shows little differentiation along its entire length. As expected, the exome capture space was concentrated within the low-recombining peri-centromeric regions that are estimated to contain 20-30% of all barley genes (**Fig. 2d**).

Partitioning of diversity according to geography

We explored the major clusters defined by PCA. Diversity within the spontaneum group indicated that overall genetic structure is largely related to geography and that our collection partitioned into an estimated $K = 9$ ancestral populations according to individual admixture coefficients calculated using sNMF[26] (**Fig. 3a** and **Supplementary Fig. 3**). The first two PCs ordered samples along the Fertile Crescent, starting from Southern Israel, traversing the Levant from South to North following a path through Southern Turkey, and arriving in Iran. Accessions originating in Central Asia clustered together with accessions from the eastern Fertile Crescent. A more formal assessment of the relationship between inter-individual genetic identities of pairs of accessions and geographic distances obtained by spatial autocorrelation analysis (**Online Methods**, **Supplementary Fig. 4**) indicated significant ($P < 0.01$) overall reductions in genetic identity with geographic distance. However, the 'hockey stick' profile indicates that a simple isolation-by-distance model does not entirely explain overall genomic divergence in this case. It is nevertheless consistent with the known demographic history of wild barley and the presence of refugia within the tested sample range in both the Eastern Mediterranean and Central Asia at the Last Glacial Maximum (~ 20,000 years BP)[27].

Within the landrace group, PCA, admixture coefficients (based on an optimal $K = 14$) and spatial autocorrelation indicated higher geographic-based structuring than in wild germplasm (**Fig. 3b, Supplementary Figs. 3, 4**). This group can be divided into two categories according to whether the inflorescence contains two or six rows of grain, a feature that until relatively recently was used to define domesticated barley as two different species, *H. distichum* L. and *H. hexastichum* L. We now know that both types were derived from two-rowed *H. vulgare* ssp. *spontaneum*, with archaeobotanical evidence indicating that either during or soon after domestication mutations in a single gene, *SIX-ROWED SPIKE 1 (HvVRS1)*[28], were necessary and sufficient to convert the inflorescence from having two to six rows of grain. Our collection contained fifty-five two-row and eighty-two six-row landrace accessions. These two genepools were most differentiated (F_{ST}) close to the location of *HvVRS1* (~ 80 cM) on chromosome 2H (**Supplementary Fig. 5**). A second F_{ST} peak on chromosome 4H corresponds to the *INTERMEDIUM-C* (*INT-C*) gene, a modifier of lateral spikelet fertility that is epistatic to *HvVRS1* and whose allelic state is differentiated between two- and six-rowed germplasm[29]. Strong F_{ST} peaks in regions without known row-type

genes were also detected on other chromosomes, especially 1H and 3H. These may represent two- vs. six-row genepool-specific loci that affect plant performance.

Signatures of selection

For both spontaneum and landrace groups we searched for within-group signatures of selective sweeps. We found few regions with clear patterns. A SweeD[30] analysis (**Online Methods**) did not detect genetic diversity patterns differing strongly from what is expected in a pure-drift model in either group (all CLR values smaller than 2). However, LD patterns analysed with OmegaPlus[31] did show sweep signals (**Supplementary Fig. 6**). Within landrace material, sweep candidates were identified on chromosome 1H (4.1 cM) and 3H (43.9 cM), while in the spontaneum group a candidate region was identified on chromosome 1H (104.9 cM). The sweep position for the landrace group on 3H corresponds to a SNP in a pentatricopeptide repeat (PPR) protein (one of a large family of RNA-binding proteins involved in regulating post transcriptional control in chloroplasts and mitochondria) that is located between *HvBTR1/HvBTR2* (40.7 cM) and the flowering-associated genes *HvFT2* (45.6 cM) and *HvGI* (45.8 cM). The corresponding region exhibited a high RoD in landrace compared to spontaneum material (see above, **Fig. 2a**).

Interactions between diversity and environment

We explored whether associations between genetic diversity and environment across the sample range were apparent in our dataset. We assembled information on 36 bioclimatic variables across the range and reduced dimensionality by clustering those that were highly correlated to obtain 7 synthetic and 8 individual environmental variables (**Online Methods**, **Supplementary Table 3**). Collectively these exhibit complex relationships, indicative of their independence from geography (**Supplementary Fig. 7a**). We analyzed SNP-environment associations independently for geo-referenced spontaneum and landrace accessions based on the Bayenv 2 approach[32], which searches for associations adjusted to correct for genetic drift. We treated each accession as a sample of a different sub-population. For each environmental variable, we observed that SNPs with a high level of association tended to cluster in different genomic regions, suggesting that each of these regions harbor genes relevant for adaptation to an environmental parameter (**Supplementary Fig. 7b**). The regions generally differed between spontaneum and landrace accessions, indicating that adaptation involved different genes and occurred independently in separate groups. Given the differences in evolutionary timescales and geographic ranges between groups this may not be entirely unexpected. Overall our analysis provided a strong indication of an effect of environmental variables on allele frequencies for the most highly associated SNPs (median Bayes factor at least of the order 10^6). However, the level of correlation for any SNP was generally weak (absolute Spearman correlation 0.07 or less).

We then estimated the X^TX statistic with Bayenv 2[32] across optimal *K*-group sub-populations with no admixture **(Online Methods)**. X^TX gives a measure comparable to F_{ST} with high values possibly indicative of selection processes. We first observed that SNPs ranked within the top 0.1% of values for spontaneum accessions and landraces occupied a wide range of chromosome positions (**Supplementary Fig. 7c**). In landrace material, positions at 51.2 cM and 51.4 cM on 4H flanked the flowering-associated gene *HvPRR59*, while a position at 48.4 cM on 5H was coincident with another flowering-associated gene, *HvELF4-like*, and a position at 67.7cM on 7H included a non-synonymous SNP within a further flowering-associated gene, *HvCO1*. Several of the top X^TX SNPs showed strong latitudinal

differentiation that is especially obvious when allelic states were plotted on geographic maps (see subsequent discussion). We interpret this as indicative of preferential selection, identifying candidate genes or genomic regions involved in domesticated barley's latitudinal range expansion.

Growth in Common Gardens

Across the sampling range two major environmental variables are seasonal photoperiod (day-length) and temperature. Within the landrace group in particular, barley has adapted to these variables by changing its lifestyle from a facultative winter-planted species in its center of origin, to a short season spring-planted crop on the North-Western fringes of Europe and highland plateaus of Central Asia. We explored whether the main developmental pathways subject to adaptive selection were those related to two key life history traits. We conducted a series of common garden experiments (6 site by season combinations in Dundee [56.5^{o}N, 3.0^{o}W], Gatersleben [51.8^{o}N, 11.3^{o}E] and St. Paul, MN [45.0^{o}N, 93.2^{o}W]) using 127 of the landrace accessions and recorded days to heading and plant height. Across sites and seasons, correlations for days to heading ranged from R^2 = 0.37 (Dundee, 2013 vs. 2014) – 0.827 (Gatersleben, 2013 vs. 2014) and for height from R^2 = 0.522 (St Paul, 2014 vs. Dundee, 2014) – 0.834 (Gatersleben, 2014 vs. St. Paul, 2015). All correlations for individual traits were significant. As several factors likely influence both of these traits, we used a multivariate analysis to derive principal components that allowed us to graphically explore the patterns observed for all 6 site/season combinations and provide a preliminary view of this complex dataset. PC1 accounted for 67.7% and 76.9% of the variation for days to heading and height respectively **(Fig. 4a, b)**. Plotting the PC1 values for each accession against the fifteen environmental variables **(Fig. 4c, d)** revealed that the strongest overall correlation across sites for both traits was with the same synthetic variable, PET+BIO5+BIO9+BIO10 (annual global potential evapotranspiration + maximum temperature of warmest month + mean temperature of driest quarter + mean temperature of warmest quarter, respectively; R^2 = 0.182, P < 0.001 for days to heading and R^2 = 0.292, P < 0.001 for height). Dissecting this synthetic variable into its individual components revealed that PET showed the strongest relationship with days to heading (R^2 = 0.196, P < 0.001)(**Supplementary Fig. 8a)** and BIO9 with height (R^2 = 0.281, P < 0.001) **Supplementary Fig. 8b)**. Investigating the positions of the SNPs with the highest level of association with this particular combination of variables in our earlier test of SNP-environment associations revealed that six flowering-associated genes were located within 1 cM windows of these high value SNPs in the landrace group, with two (*HvELF4-like* and *HvGI*) at coincident chromosome positions.

Flowering-associated genes

Given these observations, and prior indications from genome-wide studies of diversity[13], we undertook a more detailed investigation of the network of genes that encode known components of flowering-associated pathways (**Supplementary Fig. 9**). We chose 19 genes that were recently identified as homologues of well-characterized genes in *Arabidopsis,* including core circadian clock and clock output genes[33] (**Supplementary Table 4**). The genes are distributed across all chromosomes on the barley genome (**Supplementary Fig. 10**). Allelic states were scored at 773 nucleotide positions (total of 729 polymorphic positions in geo-referenced accessions, **Supplementary Table 5**). Spatial autocorrelation analysis **(Supplementary Fig. 11)** once again showed that variation in most genes is consistent with an isolation-by-distance model. For *HvFKF1*, high structuring in the landrace group appears

due to comparatively rare alleles in high linkage disequilibrium in geographically proximate accessions. For each gene we superimposed the observed nucleotide variation on hand-curated barley gene models and assembled SNP profiles and haplotypes for each accession (**Supplementary Figs. 12–30; Fig. 5a-j)**.

All genes except *HvCO2* exhibited 10 or more haplotypes. *HvCO2*, *HvELF4-like*, *HvGRP7a, HvGRP7b* and *HvZTLb* contained no non-synonymous SNPs in the accessions surveyed. Of the remainder, *HvELF3, HvPPD-H1, HvPRR95* and *HvTOC1* showed exceptionally high levels of non-synonymous variants, with 25, 35, 17 and 20, respectively. *HvCEN, HvCO1, HvFT2, HvPPD-H1, HvTOC1* and *HvZTLa* each contained non-synonymous substitutions in highly conserved protein domains. We used SIFT[34] to predict the potential impact of observed variation. All fourteen genes containing non-synonymous substitutions had variants predicted to affect protein function, and in all cases these occurred at low frequency in the population (with several tagged as low confidence). In agreement with the spatial autocorrelation analysis, visual inspection of the individual gene haplotypes plotted in geographical space **(Fig. 5c-f** for *HvCEN* and *HvPPD-H1***,** and **Supplementary Figs. 12-30)** showed some geographic structuring for most flowering-associated genes. Comparisons revealed that in no case were spontaneum and landrace haplotypes mutually exclusive, and that haplotype diversity was significantly richer ($P < 0.01$) in the spontaneum genepool. For each gene, median joining networks revealed that the major haplotypes contained spontaneum, two-rowed and six-rowed landrace accessions, with *HvGI* and *HvPRR95* the exceptions, where landraces were exclusively associated with the major haplotype.

As combinations of allelic variants in the coding sequences of different flowering-associated genes have been previously identified as modulating the phenotypic response[35], we investigated whether specific multi-gene allelic complexes were enriched in different eco-geographical zones. We first explored relationships among pairs of genes using the Mantel statistic based on genic SNP genetic distance matrices (**Online Methods**). As would be expected from stochastic sorting alone, many gene combinations revealed highly significant positive values ($P < 0.01$). However, this analysis also revealed highly significant negative values in five instances in the landrace group (*HvFKF1* vs. *HvFT2*, *HvFKF1* vs. *HvZTLb*, *HvFT1* vs. *HvGI*, *HvTOC1* vs. *HvZTLa* and *HvFT1* vs. *HvFT2*), with the last combination representing two related phosphatidylethanolamine-binding protein (PEBP) genes being very highly significant ($P < 0.001$). None of these genic combinations showed highly significant negative or positive values in the spontaneum group. Based on its very high significance, it appears that the *HvFT1-HvFT2* interaction may be of particular adaptive importance in landraces.

To investigate whether higher-order allelic complexes were assembled we pooled the polymorphic SNP information from all 19 flowering-associated genes and derived fractional sNMF assignments for spontaneum and landrace groups **(Supplementary Fig. 31a, b**). We found that even this relatively small sample of genes distributed across all chromosomes revealed a strong tendency for groups of alleles to co-locate geographically, suggesting that co-adapted gene complexes, at least in part, underlie the observed distribution of the accessions. A comparison of flowering-associated-gene SNP median X^TX values with random SNP samples for landrace and spontaneum groups (**Online Methods**) indicated that *HvPPD-H1* was more differentiated in both cases ($P < 0.05$ for two Bayenv runs for each barley group), suggestive of selection. In landrace barley we also found one gene, *HvTOC1*, significantly less differentiated than a random SNP set ($P < 0.01$ for both Bayenv runs).

HvCEN, *HvPPD-H1* and *HvFT1* have been considered elsewhere as causative for variation in days to heading in cultivated barley[13,36,37,38]. Our assessment indicated that the causal SNP 274284:918 in *HvCEN* was most associated with latitude of all tested flowering-associated gene SNPs in the landrace collection (see **Fig. 5c** for haplotype distributions). We set this SNP as a benchmark to determine additional SNPs within the exome dataset that were equally or more associated with days to heading in common garden trials to identify candidates that may contribute to adaptation during latitudinal range expansion. We adopted three approaches to detect candidate SNPs for testing. First, a naïve approach was applied in which the entire latitudinal distribution of landrace barleys was split into North and South based on median latitude. We then used a X^2 test to detect 400 SNPs with a MAF of > 0.2 that had allelic distributions that were at least as unlikely as SNP 274284:918 to be evenly distributed over latitude. Second, 130 SNPs identified by Bayes factor ranks as highly associated with latitude were also considered. Third, 207 SNPs identified by X^TX as deviating strongly from assigned *K*-group sub-populations were also included. These data were pooled for analysis against days to heading measurements taken in the common garden trials (*N* = 127) (correlation coefficient for heading data across sites = 0.61). Of the total set of candidate SNPs, 27 (3.7% - from a total of 15 sequence contigs)(**Supplementary Table 6**) were equally or more associated with days to heading based on $-log_{10}P$ values than SNP 274284:918 at two of the trial sites (**Online Methods**) and are therefore putatively involved in providing adaptive capacity for geographical expansion. Strong values of association were observed for three SNPs within sequence contig 2548323 on 7H (69.3 cM) located between the flowering-associated genes *HvCO1* (67.9 cM) and *HvLHY* (70.8 cM). Contig 2548323 contains a homologue of a nucleolar gar2-related protein. As its role in latitudinal differentiation is difficult to predict, we favor an explanation that it may simply be associated through close genetic linkage with a more appropriate candidate. Of the three approaches to detect candidate SNPs, the highest proportion equaling or exceeding *HvCEN* SNP 274284:918 for association with days to heading at both trial sites was observed for the naïve X^2 method (6.5% of tested SNPs).

DISCUSSION

Our analysis revealed a complex pattern of allelic variation in barley associated with geographical site of origin and inherent environmental variables that appears typical of the species[39,40]. Common garden experiments highlighted the relationship between days to heading and height with seasonal temperature and dryness variables, and the importance of flowering-associated genes hinted that combinations of alleles at a subset of the loci studied contribute to the overall matching of environment and life history traits. Despite this connection, our data indicate that many other factors have been involved in shaping the allelic complement of the barley genome as it spread across its current range and that combinations of genomic responses have contributed to the partitioning of diversity apparent in genome-wide genetic data.

(3637 words)

METHODS

Materials and Methods and associated references and URLs are available in the online version of the manuscript.

Accession Codes

All Accession codes are given in the Online Methods section. Raw read files can be retrieved from the European Nucleotide Archive (ENA) under project ID PRJEB8044 (ERP009079)[a]. ENA accessions for each sample are included in **Supplementary Table 1**. Genotype matrices for SNPs and indels are available under DOI[b,c]. DOIs were registered with e!DAL[41].

URL's

a. http://www.ebi.ac.uk/ena/data/view/ERP009079
b. http://dx.doi.org/10.5447/IPK/2016/4
c. http://dx.doi.org/10.5447/IPK/2016/5
d. http://www.illumina.com
e. www.cgiar-csi.org/data/global-aridity-and-pet-database
f. https://cran.r-project.org/web/packages/cluster/
g. http://biodiversityinformatics.amnh.org/open_source/gdmg/index.php
h. www.vsni.co.uk/downloads/genstat/17th-edition/
i. http://popart.otago.ac.nz
j. https://www.arcgis.com/home/
k. http://www.geomidpoint.com/
l. http://biology-assets.anu.edu.au/GenAlEx/Welcome.html

ACKNOWLEDGEMENTS

We would like to acknowledge funding from the Scottish Government Research Program to RW, JR, IKD, MB and IM and from the EU FP7 WHEALBI project to JR, RW, NS, IKD, SK and BK. MvZ has been supported by the CGIAR Climate Change, Agriculture and Food Security (CCAFS) program. The work would not have been possible without funding from BBSRC Grant No. BB/I00663X/1 to RW, German Science Foundation (DFG) SPP1530 Grant No. KI1465/6-1 to BK and BMBF TRITEX 0315954 A to NS. Funding to GJM was provided by the United States Department of Agriculture - National Institute of Food and Agriculture (USDA-NIFA) as part of the Triticeae Coordinated Agricultural Project (TCAP), grant no. 2011-68002-30029. We specifically thank Bill Thomas and Allan Booth (JHI) for helpful discussions around the common garden experiments. Analysis done by FF and KS was mainly performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and the Arts Baden-Württemberg and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

AUTHOR CONTRIBUTIONS

RW, GJM, NS, and BK conceived the study. BK selected the germplasm for inclusion and purified lines by 2 rounds of single seed descent. BK, RS, GJM, SH, AHo and JR conducted the common garden experiments and analysed the data. MK, AHi and NS generated the exome sequence data. JR, IKD, KS, FF, MM, SK, CC, MB, JWSB, MvZ, TM-G and IM analysed the data and provided information included in the supplementary files. RW, MM, IKD, JR, FF, NS and GJM wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

References

1. Beddington, J. *et al*. What next for agriculture after Durban? *Science* **335**, 289-290 (2012).
2. Challinor, A.J. *et al.* A meta-analysis of crop yield under climate change and adaptation. *Nat. Clim. Change* **4,** 287-291 (2014).
3. McCouch, S. *et al.* Agriculture: feeding the future. *Nature* **499**, 23-24 (2013).
4. Zohary, D, Hopf, M, and Weiss, E. Domestication of plants in the old world (Fourth edition). Oxford University Press, ISBN 978-0-19-954906-1 (2013).
5. Dawson, I.K. *et al.* Barley: a translational model for adaptation to climate change. *New Phytol*. **206**, 913-931 (2015).
6. Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop domestication. *Cell* **127**, 1309-1321 (2006).
7. Mascher, M. *et al*. Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J*. **76**, 494-505 (2013).
8. International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711-716 (2012).
9. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
10. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297-1303 (2010).
11. Mascher, M. *et al*. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J*. **76**, 718-727 (2013).
12. Kilian, B. *et al*. Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. *Mol. Gen. Genomics* **276**, 230-241 (2006).
13. Comadran, J. *et al*. Natural variation in a homolog of *Antirrhinum CENTRORADIALIS* contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat. Genet.* **44**, 1388-1392 (2012).
14. Hurst, L.D. Genetics and the understanding of selection. *Nat. Rev. Genet.* **10**, 83-93 (2009).
15. Russell, J. *et al*. Analysis of >1000 single nucleotide polymorphisms in geographically matched samples of landrace and wild barley indicates secondary contact and chromosome-level differences in diversity around domestication genes. *New Phytol*. **191**, 564-578 (2011).
16. Cavanagh, C.R. *et al.* Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci. USA* **110**, 8057-8062 (2013).
17. Li, Y. *et al*. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969-972 (2010).
18. Meyer, R.S., & Purugganan, M.D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet*. **14**, 840-852 (2013).
19. Kunzel, G., Korzun, L. & Meister, A. Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* **154**, 397-412 (2000).

20. Kunzel G., & Waugh R. Integration of microsatellite markers into the translocation-based physical RFLP map of barley chromosome 3H. *Theor. Appl. Genet.* **105,** 660-665 (2002).
21. Baker, K. *et al*. The low-recombining pericentromeric region of barley restricts gene diversity and evolution but not gene expression. *Plant J.* **79,** 981-992 (2014).
22. Morrell, P.L., Gonzales, A.M., Meyer, K.K. & Clegg, M.T. Resequencing data indicate a modest effect of domestication on diversity in barley: a cultigen with multiple origins. *J. Hered*. **105**, 253-264 (2014).
23. Taketa, S. *et al*. Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. *Proc. Natl. Acad. Sci. USA* **105**, 4062-4067 (2008).
24. Pourkheirandish, M. *et al*. Evolution of the grain dispersal system in barley. *Cell* **162**, 527-539 (2015).
25. Begun, D.J. & Aquadro, C.F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519-520 (1992).
26. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & Francois, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**, 973-983 (2014).
27. Russell, J. *et al*. Genetic diversity and ecological niche modelling of wild barley: refugia, large-scale post-LGM range expansion and limited mid-future climate threats? *PLoS One* **9**, e86021 (2014).
28. Komatsuda, T. *et al*. Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc. Natl. Acad. Sci. USA* **104**, 1424-1429 (2007).
29. Ramsay, L. *et al*. *INTERMEDIUM-C*, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene *TEOSINTE BRANCHED 1*. *Nat. Genet.* **43**, 169-172 (2011).
30. Pavlidis, P., Živković, D., Stamatakis, A., & Alachiotis, N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol. Biol. Evol.* **30**, 2224-2234 (2013).
31. Alachiotis, N., Stamatakis, A., & Pavlidis, P. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics* **28**, 2274-2275 (2012).
32. Günther, T., & Coop, G. Robust identification of local adaptation from allele frequencies. *Genetics* **195**, 205-220 (2013).
33. Calixto, C.P.G., Waugh, R. & Brown, J.W.S. Evolutionary relationships among barley and *Arabidopsis* core circadian clock and clock-associated genes. *J. Mol. Evol.* **80**, 108-119 (2015).
34. Kumar, P., Henikoff, S., & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc*. **4**, 1073-1081 (2009).
35. Faure, S. *et al.* Mutation at the circadian clock gene *EARLY MATURITY 8* adapts domesticated barley (*Hordeum vulgare*) to short growing seasons. *Proc. Natl. Acad. Sci. USA* **109**, 8328-8333 (2012).
36. Jones, H. *et al.* Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the Fertile Crescent. *Mol. Biol. Evol.* **25**, 2211–2219 (2008).

37. Turner, A., Beales, J., Faure, S., Dunford, R.P. & Laurie, D.A. The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. *Science* **310**, 1031-1034 (2005).
38. Casas, A. M *et al. HvFT1* (*VrnH3*) drives latitudinal adaptation in Spanish barleys. *Theor. Appl. Genet*. **122**, 1293-1304 (2011).
39. Morrell, P.L., Lundy, K.E. & Clegg, M.T. Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite migration. *Proc. Natl. Acad. Sci. USA* **100**, 10812-10817 (2003).
40. Poets, A.M., Fang, Z., Clegg, M.T. & Morrell, P.L. Barley landraces are characterized by geographically heterogeneous genomic origins. *Genome Biol.* **16**, 173. (2015).
41. Arend, D. *et al.* e!DAL – a framework to store, share and publish research data. *BMC Bioinformatics* **15**, 214 (2014).

Figure Legends

Figure 1. Principal components analysis of diversity in the barley collection. Red = Spontaneum group (*N* = 91), blue = Landrace group (*N* = 137), grey = Other (*N* = 39).

Figure 2. Chromosome-level profiles of different barley group accessions. Red = Spontaneum (*N* = 91), blue = Landrace (*N* = 137). (**a**) Nucleotide diversity (π). The grey line indicates the proportional reduction in diversity in the landrace group compared to the wild group ($[\pi_{spont.}-\pi_{land.}]/\pi_{spont.}$; scaled as in parentheses). (**b**) Recombination rate (ρ). (**c**) Differentiation (F_{ST}) between spontaneum and landrace groups. (**d**) Cumulative number of exome capture targets. Given the absence of a physically ordered genome sequence, values are for 5 cM bins. Some features discussed in the text are indicated with arrows and positions of centromeres with dashed lines.

Figure 3. Geographical partitioning of diversity. The locations of 228 geo-referenced accessions, with fractional sNMF assignments, are represented for (**a**) Spontaneum (*N* = 91, *K* = 9) and (**b**) Landrace (*N* = 137, *K* = 14) groups, respectively, indicating the geographic structuring of exome capture sequence variation. To visualize the majority of individual points, some positions are marginally offset. The placing of a single accession in central China, above the state of Arunachal Pradesh in northeast India, is shown appended to the main spontaneum group map.

Figure 4. Common garden experiments. (**a**, **b**) Plots of first and second principal components from multivariate analysis of 6 sites/seasons for (**a**) days to heading and (**b**) height accounting for 67.7% (PC1) and 14.1% (PC2), and 76.9% (PC1) and 10.1% (PC2), of variation, respectively. (**c**, **d**) Regression plots of PC1 of the most significant (synthetic) environmental variable (PET+BIO5+BIO9+BIO10) on the *x*-axis with PC1 of (**c**) days to heading and (**d**) height from plots (**a**) and (**b**) on the *y*-axis. R^2 and significance values are shown. Individual accessions are color-coded according to their individual *K*-groups without admixture.

Figure 5. Molecular and spatial variation in *HvCEN* and *HvPPD-H1*. Genomic structure of (**a**) *HvCEN* and (**b**) *HvPPD-H1* with relative positions and effect of the 14 and 129 SNPs identified, respectively. Synonymous SNPs and SNPs in UTRs and introns are generally indicated by small grey arrows. Non-synonymous SNPs are represented by large black or colored arrows, the latter representing SNPs in conserved protein domains. Three additional SNPs in *HvPPD-H1* (9, 57 and 82) with functions outlined in **Supplementary Fig. 25** are also indicated by large black arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded red or green. SNP 9 in *HvCEN* has been previously predicted to be causal for days to heading[13]. SNPs 70 and 73 in *HvPPD-H1* are associated with *PPD-H1* or *ppd-H1* alleles according to Jones *et al*[36] and Turner *et al*[37], respectively. (**c**–**f**) show haplotype distributions according to geography in the landrace and spontaneum groups. Haplotypes are color coded from the most (red, A) to the least (brown, G) frequent, with unique and rare (< 5 individuals) haplotypes combined into a single class (orange,I). (**g**, **h**) Venn diagrams of haplotype number and sharing between different barley groups. (**i**, **j**) Median joining networks for *HvCEN* and *HvPPD-H1* haplotypes, respectively. Red = Two-row landrace, blue = Six-row landrace, yellow = Spontaneum.

ONLINE METHODS

Plant material, library preparation and exome sequencing

A list of all accessions included in the present study is available as **Supplementary Table 1**. The collection was designed to track the diffusion of early domesticates out of the Fertile Crescent along major dispersal routes. In the choice of landrace accessions, we covered a large geographical area and gave some priority to regions where proximate wild materials would allow a comparison of wild and landrace accessions growing in related environments. For both wild and landrace material we considered known population structure and distribution ranges[27,41,42] and critically required trusted geo-reference collection site data associated with each accession. We omitted early cultivar selections originating from North-Central and North-Western Europe (as used previously[36,43]) due to a lack of accurate collection site information that we considered a requirement for our current purposes.

DNA was extracted from fresh young plant material at the three leaf stage using a cetyl-trimethylammonium bromide-based (CTAB) protocol as described previously[44]. Exome capture and Illumina sequencing were performed as described elsewhere[7]. Apart from previously published data[7], four bar-coded samples were captured together with the same capture kit and sequenced on one lane of an Illumina HiSeq2000 (2 x 100 bp) at IPK Gatersleben. Demultiplexing was performed with CASAVA 1.8.2[d]. Read counts per sample are included in **Supplementary Table 1**.

Read mapping, read depth analysis, SNP calling and prediction of function

Details of the bioinformatics pipelines used for read mapping, variant calling and validation are given in **Supplementary Note 1**.

Accessions for barley group assessment

From the original 267 barley accessions, 228 that had geographic location data, ≤ 5% missing data and ≤ 5% heterozygous states, were assembled into two defined barley groups to allow for subsequent geospatial analyses: 'spontaneum' (*N* = 91), and 'landrace' (*N* = 137). Of the landraces, 55 had two-row inflorescences and 82 were six-row. Assignment of each accession was based on the available passport data from collection expeditions and supplementary data sources. The latter included: observations from field trials, previous molecular marker studies, preliminary PCA of SNP data and information on geographic distributions from archaeological sources[45]. The assignment process revealed that six of the 228 accessions had likely been previously misclassified as spontaneum and should be considered as within the landrace group for current analysis (see **Supplementary Table 1**). In addition, we 'forced' six putative *Hordeum agriocrithon* accessions, based on the closest neighbouring genotypes, into either our spontaneum or landrace categories so that they could be incorporated into group analyses.

Environmental data

Environmental data for 34 variables for all 228 spontaneum and landrace accessions were assembled from online databases from geo-referenced sample points: 19 bioclimatic variables (all 30 second resolution layers[46]) and altitude/elevation (m) from the WorldClim database; monthly solar radiation data (12 months, 1.5 minute resolution layers[e]); and

Global potential evapotranspiration (PET) and Global aridity index (GAI) data (annual averages, 1950 to 2000; both 30 second resolution layers[e]). PET is a measure of the ability of the atmosphere to remove water through evapotranspiration processes, while GAI is defined as mean annual precipitation divided by mean annual potential evapotranspiration. Latitude and longitude were also considered as environmental inputs, totaling 36 variables for analysis (**Supplementary Table 3**). For further use we summarized environmental information from all locations into seven synthetic and eight individual variables. This was undertaken through clustering of variables based on pairwise correlations between them with the "pam" function in the R package cluster[f]. Fourteen groups were assigned, which was the minimum number > 2 with no negative silhouette width for each trait. Latitude, with its particular importance in barley range expansion, was also retained as a separate variable. The assignment of traits to groups is shown in **Supplementary Table 3**. For all groups consisting of more than a single variable, PCA was performed in R using the ade4 package[47]. The first PC always explained at least 85% of the variation within a group and we therefore represented each group by this measure during further analysis.

Overall genetic structure

Principal components analysis was performed with the R package SNPRelate[48]. Partitioning of accessions into *K* ancestral groups and estimation of individual ancestry coefficient was performed for spontaneum and landrace groups and all scored SNPs with sNMF[26]. We chose sNMF for this analysis because previous reports had shown its good performance in species with high levels of inbreeding. The command "sNMF" was called with the parameter "-m 1" (assuming haploid data for the predominantly inbreeding barley) for *K* between 1 and 15 for our two groups. For each *K*, 15 replications were performed. The optimal *K* was determined using the cross-entropy criterion within sNMF[26]. *Q* proportions were averaged across all 15 replications with CLUMPP (version 1.1.2[49]). Due to the difference in calculation times in CLUMPP, which depends on the value of *K*, the method for combining sNMF replications varied to allow for approximately equal run length. When *K* = 9 (for spontaneum), the 'Greedy' algorithm with 500 permutations was applied; when *K* = 14 (for landrace), the 'LargeKGreedy' algorithm with 100,000 permutations was used. The fractional assignments of individuals to groups is given in **Supplementary Table 1** (see also **Fig. 3**). The sNMF analysis with the same number of *K* groups (to aid comparison with earlier results) and the same level of replication was also applied to pooled genic SNPs from 19 flowering-associated genes (see below), again using CLUMPP with the same parameters to combine results.

Chromosome-level diversity, recombination and differentiation statistics

The diversity statistic π was calculated with the program compute, which is part of the libsequence library[50]. Values were calculated for each genetically anchored whole-genome shotgun contig for both spontaneum and landrace groups. Per base pair estimates were calculated by dividing the number of polymorphic sites by the size of target regions on each contig. Reduction of diversity (RoD) scores were calculated on a per-contig level using the formula $(\pi_{spontaneum} - \pi_{landrace})/\pi_{spontaneum}$. Recombination rate ($\rho$) in wild barleys was estimated with Maxhap[51] on a per-contig level as described previously[52]. We considered only SNPs with minor allele counts ≥ 3 located in contigs that contained at least 20 SNPs. Values of ρ per base pair were estimated across a grid of values from 10^{-4} to 1, assuming no

homologous gene conversion. For visualization along the genetically ordered barley draft genome in **Fig. 2**, rolling medians were calculated across 101 WGS contigs that are at adjacent positions in the POPSEQ map with the R package zoo[53]. Differentiation (F_{ST}) between groups was calculated using Hudson's estimator[51] with the explicit formula given as equation 10 in Bhatia *et al*[54]. Single SNP estimates were averaged across all polymorphic SNPs in 5 cM windows using the ratio-of-averages method[55]. A neighbor-joining tree (not shown) was drawn with the R package ape[55] from a distance matrix calculated with the function "snpgdsDiss" of the SNPRelate package[48].

Signatures of selection

Using near-complete high-confidence gene-sequence SNPs, we employed OmegaPlus (version 2.3.0[31]) to scan for LD signatures of positive selection in contigs within spontaneum and landrace groups, based on the ω-statistic. For each window ranging from 200 bp to maximally the length of the contig, centred at equidistant (number of SNP) positions across the contig, we computed ω without using imputation. For each contig, we computed approximate *P* values for the observed ω under a neutral model. We used the standard coalescent with recombination as simulated in ms[51] as a neutral model to produce haploid segments with sample sizes as in the observed data for the barley groups. For each contig, we set the scaled mutation rate Θ to the Watterson estimate and the scaled recombination rate ρ to the results of the Maxhap analysis. For contigs where we did not compute a ρ estimate, we used the mean ρ across contigs where it was computed. Note that we implicitly use the coalescent with (complete) selfing as described elsewhere[56]. However, it is not necessary to rescale the parameters Θ and ρ, since they are already scaled properly using the standard coalescent to estimate them. We assume that our simulated individuals are completely homozygous, which causes a negligible bias. Approximate *P* values were produced via the Monte Carlo approach from Besag and Clifford[57] with h=100 and a maximum of 10^6 simulations per contig. Simulations were performed with *ms*. The ω-statistic for the simulations was computed in OmegaPlus with the same parameters as for the observed contig. Correction for multiple testing was done by controlling for a family-wise error rate of 0.05 with Holm's method[58].

Complementary with the above, SweeD (version 3.3.2[30]) was used to scan for typical signatures of positive selection in the folded site-frequency spectrum (fSFS) for spontaneum and landrace groups with the same SNP data. Individual accessions were treated as haploid, and heterozygous base calls were identified with equal probability as either reference or alternative allele. Only SNPs with no missing data were included in all steps of the analysis. The SweeD analysis was performed within each contig, with the genome-wide fSFS as input for each. The composite likelihood ratio (CLR) as in the parametric approach described elsewhere[59] was computed at each SNP position. We reported the maximum CLR for each contig. Since we found no strong selection signals (CLR values were < 2 for all contigs and all groups), we refrained from comparing CLR scores against simulations from a neutral model.

Finally, for both spontaneum and landrace groups we identified SNPs from the near-complete high-confidence gene sequence pool with the strongest 'non-drift' differentiation between inferred sub-populations as an approach to detect positions under selection pressure. We employed a method based on deviation from *K*-group designations of

individual accessions similar to those identified through sNMF analysis (see above under section 'Overall genetic structure'), but instead of fractional *Q* assignments we used no admixture assignments generated with STRUCTURE (version 2.3.4[60]) using the optimum number for *K* derived from the entire SNP dataset according to the cross-entropy criterion applied in sNMF. Input data which consisted of a representative set of 4,000 SNPs sampled from across the genome from the near-complete high-confidence gene sequence pool were treated as haploid and 10,000 burnin and 50,000 further steps were employed in STRUCTURE, conditions which were sufficient in trial runs to ensure the convergence of key parameters. The analysis was undertaken on 15 separate occasions and the results synthesized with CLUMPP with parameters as earlier for sNMF. Final membership of each accession to its *K*-group was then based on the highest proportional representation to a group. Using Bayenv 2.0[32] we then estimated the variance-covariance matrix Ω between *K*-groups twice based on two sets of 50,000 randomly chosen intronic SNPs with 150,000 MCMC steps; an additional replicate was checked for strong deviation from these estimates, but not used further. In our use of Bayenv (see also below) all accessions were treated as haploid. For both Ω estimates, we then estimated the X^TX statistic for each SNP using 200,000 MCMC steps. The two runs showed high correspondence (Spearman's correlation for spontaneum and landrace = 0.86 and 0.93, respectively). We reported the consensus set of SNPs being in the top 0.1% of X^TX values for both runs.

Spatial autocorrelation analysis

Spatial autocorrelation analysis of SNP data was undertaken with SPAGeDi (version 1.4c[61]) to assess the relationship between inter-individual genetic identities of pairs of accessions and geographic distances for spontaneum and landrace groups. Pairwise geographic distances between accessions were calculated with the Geographic Distance Matrix Generator (version 1.2.3[g]), to account for the Earth's curvature, and input as a separate matrix into SPAGeDi. To accommodate the input limits of SPAGeDi, the approach was applied to representative samples of 5,000 SNPs from each of four SNP categories (intron, nonsynonymous, synonymous and UTR) sampled systematically across the genome from near-complete high-confidence gene sequences with chromosome physical map assignments. The analysis was also applied to genic SNPs from 19 flowering-associated genes (see below). All analyses were based on Ritland's kinship coefficient (other coefficients gave similar results) as a measure of identity, on 10 geographic distance classes of equal sample size (distance classes expressed as *ln* transformations), and on 1,000 permutations for testing if the distribution of identities over geographic distances deviates from the null hypothesis of no relationship between identity and geographic distance. Values for identity are averages for all pairwise comparisons within a particular distance class, expressed as the difference from the average value of randomly sampled pairs of individuals.

Environmental association analyses

We used Bayenv 2.0[32] to identify the SNPs from the near-complete high-confidence gene sequence SNP pool with the strongest associations with seven synthetic and eight individual environmental variables (see above, 'Environmental data'). We treated each accession as being sampled from a different sub-population within its spontaneum (N = 91) or landrace (N = 137) group. We estimated the variance-covariance matrix Ω between these sub-

populations twice, as described in the section 'Signatures of selection', but with 100,000 MCMC iterations. For both estimates, we ran Bayenv 2.0 to estimate Bayes factors (drift-only model vs. environment-associated model) and Spearman's correlation coefficient between each environmental variable and the standardised allele frequencies for each SNP, using 200,000 MCMC iterations. We reported the consensus set of the SNPs being in the top 1% of Bayes factors for both runs that were also in the top 5% of absolute correlations in each run. Bayenv runs for different SNPs were conducted in parallel using GNU parallel software[62]. Reasonable correlation was observed between Bayenv runs (Spearman's correlation coefficient ranged from 0.55 to 0.70 and from 0.50 to 0.67 for spontaneum and landrace groups, respectively, for the 15 environmental variables).

Common garden experiments

Seeds of 127 landrace accessions were grown under field conditions in Dundee (56.5^{o}N, 3^{o}W), Gatersleben (51.8^{o}N, 11.3^{o}E) and St Paul, MN (45.0^{o}N, 93.2^{o}W) in 2013 (Dundee, Gatersleben), 2014 (Dundee, Gatersleben, St. Paul) and 2015 (St. Paul). Plants were scored for days to heading (growth stage 55 on Zadoks scale; half of inflorescence emerged in half of the plot or row), and height to ear tip. Local practices for fertilizer and disease control were adopted for each trial site. Data was collated and analyzed using GenStat 17th Edition[h] for correlations between site/seasons and PCA (for 99 accessions common to all six trials). We conducted a genome-wide analysis using a combination of SNPs identified from the X^TX, Bayenv and naïve X^2 test (742 SNPs in total) and days to heading scores for two sites (Dundee and Minnesota, 2014) with all 127 accessions (GenStat, QTL commands for single trait association with no correction for population structure).

Generation of specific gene-sequence haplotypes and diversity analysis for flowering-associated genes

Literature on crop domestication[63] and our initial findings indicate the importance of understanding the role of variation in genes controlling time to flowering in cultivated crop range expansion. Genic SNPs were extracted from the relevant contigs of 19 individually curated flowering-associated genes. These included circadian clock genes that have wide regulatory functions in the genome including in controlling flowering and that were recently identified in barley as homologues of well-characterized genes in *Arabidopsis*[33] (**Supplementary Table 4**). These SNPs were analysed with sNMF and SPAGeDi (as described above under 'Overall genetic structure' and 'Spatial autocorrelation analysis', respectively) and haplotypes constructed by summing SNP profiles within each gene, excluding accessions with missing data points. In addition, median-joining haplotype networks[64] were constructed with the software tool PopART[i]. Haplotypes were plotted on geographic maps with ArcGIS[j]. In addition, the distribution of individual-gene SNPs were plotted on geographic maps for visual inspection of variation and geographic midpoints for allelic placements determined for key SNPs with the Geographic Midpoint Calculator using the 'center of gravity' method[k]. Venn diagrams of haplotypes of all 19 sequences were generated from haplotype overlaps between barley groups with eulerAPE (Version 3.0.0[65]).

We further tested the genetic differentiation of flowering-associated gene SNPs within the spontaneum and landrace groups. We extracted X^TX values calculated for each gene SNP (see under 'Signatures of selection') and compared the median of SNPs for each gene with

the median of a group of randomly chosen SNPs of the same number from the overall distribution. 100,000 iterations of random sampling provided a test for whether the median value for a flowering-associated gene ranks significantly lower or higher than expected based on a null hypothesis of no difference. Correction for multiple testing across flowering-associated genes involved adjusting *P* values with Holm's method using the R function p.adjust. *P* values were reported for both Bayenv runs.

Mantel Tests of flowering-associated gene pairs

SNP panels for each flowering-associated gene were used to generate genic genetic distance matrices for landrace and spontaneum groups. Of the 19 flowering-associated genes, two (*HvCO1* and *HvLHY*) with > 25% missing SNP data for any barley accession were first excluded from analysis. Any accessions with > 10% missing data for any gene were then also excluded from the (entire) analysis, leaving 134 and 81 accessions in landrace and spontaneum groups, respectively, and 17 flowering-associated genes for comparison. Genetic distance matrices based on simple matching at each SNP locus and scaled to 0,1 across loci for each gene were generated in GenAlEx[l](version 6.502). The Mantel statistic was then calculated for each pair of genic distance matrices[66] using the same software package, with a test for significance based on 999 random permutations.

(2851 words)

Online Methods References

41. Jakob, S.S. *et al.* Evolutionary history of wild barley (*Hordeum vulgare* subsp. *spontaneum*) analyzed using multilocus sequence data and paleodistribution modeling. *Genome Biol. Evol.* **6**, 685-702 (2014).
42. Pasam, R.K. *et al.* Genetic diversity and population structure in a legacy collection of spring barley landraces adapted to a wide range of climates. *PLoS One* **9**, e116164 (2014).
43. Jones, H. *et al.* Evolutionary history of barley cultivation in Europe revealed by genetic analysis of extant landraces. *BMC Evol. Biol.* **11**, 320 (2011).
44. Doyle, J.J. & Doyle, J.L. Isolation of plant DNA from fresh tissue. *Focus* **12**, 13-15 (1990).
45. Harlan, J.R. Crops and Man. The American Society of Agronomy and the Crop Science Society of America, Madison, Wisconsin (1975).
46. Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Clim.* **25**,1965-1978 (2005).
47. Dray, S. & Dufour, A.B. The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Soft*. **22**, 1-20 (2007).
48. Zheng, X. *et al*. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-3328 (2012).
49. Jakobsson, M. & Rosenberg, N.A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806 (2007).
50. Thornton, K. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325-2327 (2003).
51. Hudson, R.R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805-1817 (2001).
52. Hufford, M.B. *et al*. Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808-811 (2012).

53. Zeileis, A. & Grothendieck, G. zoo: S3 infrastructure for regular and irregular time series. *J. Stat. Soft.* **14**, 6 (2005).
54. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A.L. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Res.* **23**, 1514-1521 (2013).
55. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289-290 (2004).
56. Nordborg, M. & Donnelly, P. The coalescent process with selfing. *Genetics* **146**, 1185-1195 (1997).
57. Besag, J. & Clifford, P. Sequential Monte Carlo *p*-values. *Biometrika* **78,** 301-304 (1991).
58. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65-70 (1979).
59. Nielsen, R. *et al*. Genomic scans for selective sweeps using SNP data. *Genome Res*. **15**, 1566-1575 (2005).
60. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587 (2003).
61. Hardy, O.J. & Vekemans X. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* **2**, 618-620 (2002).
62. Tange, O. GNU parallel: the command-line power tool. *USENIX Magazine* **36**, 42-47 (2011).
63. Nakamichi, N. Adaptation to the local environment by modifications of the photoperiod response in crops. *Plant & Cell Physiol*. **56**, 594-604 (2015).
64. Bandelt, H. J., Forster, P. & Rohl, A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16,** 37-48.
65. Micallef, L. & Rodgers**,** P. eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses. *PLoS ONE* **9**, e101717 (2014).
66. Smouse, P.E., Long, J.C. & Sokal, R.R. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool*. **35**, 627-632 (1986).