Targeted re-sequencing reveals the genetic signatures and the reticulate history of barley domestication

Artem Pankin[1,3,*] Janine Altmüller[4], Christian Becker[4] and Maria von Korff[1,2,3*]

[1] Institute of Plant Genetics, Heinrich-Heine-University, 40225 Düsseldorf, Germany

[2] Cluster of Excellence on Plant Sciences "From Complex Traits towards Synthetic Modules" 40225 Düsseldorf, Germany

[3] Max Planck Institute for Plant Breeding Research, 50829, Cologne, Germany

[4] Cologne Center for Genomics (CCG), University of Cologne, 50931, Germany

***Corresponding Authors:** Artem Pankin (pankin@mpipz.mpg.de), Maria von Korff (korff@mpipz.mpg.de)

Abstract

Barley (*Hordeum vulgare* L.) is one of the Neolithic founder crops of the early agricultural societies. The circumstances of its domestication and genomic signatures that underlie barley transition from a weed to a crop remain obscure. We explored genomic variation in a diversity set of 433 wild and domesticated barley accessions using targeted re-sequencing that generated a genome-wide panel of 544,318 high-quality SNPs. We observed a ~50% reduction of genetic diversity in domesticated compared to wild barley and diversity patterns indicative of a strong domestication bottleneck. Selection scans discovered multiple selective sweep regions associated with domestication. The top candidate domestication genes have been implicated in the regulation of light signaling, circadian clock, hormone and carbohydrate metabolism. Phylogeographic analyses revealed a mosaic ancestry of the domestication-related loci, which originated from nine wild barley populations. This indicates that recurrent introgression and selection of wild alleles apparently shaped the domesticated gene pool, which supports a protracted domestication model.

Introduction

Domesticated barley (*Hordeum vulgare* ssp. *vulgare*) is one of the Neolithic founder crops, which facilitated establishment of the early agricultural societies.[1] Due to its striking environmental plasticity, barley is of utmost importance as a staple crop in a wide range of agricultural environments.[2] The first traces of barley cultivation were found at archaeological sites in the Fertile Crescent, which dated back to ~10,000 B.C.[3] The Fertile Crescent is the primary habitat of the crop progenitor wild barley (*H. vulgare* ssp. *spontaneum*). However, its isolated populations have spread as far as North African and European shores of the Mediterranean and East Asia. Wild barley is a rich yet underutilized reservoir of novel alleles for breeding of barley cultivars better adapted to predicted future climatic perturbations.

In contrast to some other crops, the visible phenotype of domesticated barley did not dramatically diverge from its wild form.[4] So far, the spike rachis brittleness has remained the only well characterized domestication trait that exhibits a clear dimorphism between the wild and domesticated subgroups, which are characterized by the brittle and non-brittle spikes, respectively[5–7]. Other such traits and underlying genes that define the barley domestication syndrome (DS), as a complex of all characters that characterize the domesticated phenotype, are yet undiscovered.[8] When adaptive phenotypes are not clearly defined, the so-called bottom-up approach, which starts with the identification of genome-wide signatures of selection, has proven instrumental in reconstructing the genetic architecture of the DS.[9,10] In other crops, the selection scans detected multiple selective sweep regions associated with domestication, which comprised hundreds of candidate domestication genes apparently modulating yet unstudied aspects of the domestication phenotypes.[11–15]

The circumstances of barley domestication are debatable and its effects on the domesticated barley genomes remain poorly understood. The early models, based on the diversity analysis of isolated genes and neutral DNA markers, proposed the Israel-Jordan area as a primary center of cultivated barley origin and hinted at the East Fertile Crescent, the Horn of Africa, Morocco and Tibet as the alternative centers of domestication.[5,16–20] Archaeological and recent molecular evidence suggested that barley domestication was a protracted process, involving the polyphyletic origin of the non-brittle spikes and the recurrent gene flow between the wild and domesticated germplasm.[5,21–23] To further unravel the increasingly complex model of barley domestication, a detailed understanding of the genes and processes involved in transition of barley from the wild to domesticated form is crucial.

Here we developed a genome-wide targeted re-sequencing assay to interrogate ~ 544,000 SNPs in a diversity panel comprising 344 wild and 89 domesticated barley genotypes. The population genetic analyses shed new light on the structure of genetic diversity in the wild populations and on how domestication affected diversity of the cultivated barley genotypes. Besides the spike brittleness locus, the selection scans identified multiple genomic regions and candidate genes affected by the selection under domestication. Finally, we untangled a complex ancestry of the domesticated barley genomes, which consisted of at least nine wild ancestral populations, and demonstrated heterogeneous origin of the candidate domestication loci.

Results

>500,000 SNPs discovered by the targeted re-sequencing assay

A total of 433 barley genotypes, including 344 wild and 89 domesticated barley genotypes were analyzed in this study. To maximize diversity, the wild barley genotypes were selected to cover the entire range of its habitats in the Fertile Crescent. The domesticated barley included landraces from the Fertile Crescent, North and East Africa and advanced cultivars from Europe, Australia, USA and the Far East. This set represented the whole variety of domesticated barley lifeforms, namely two- and six-row genotypes with winter and spring growth habits.

Illumina enrichment re-sequencing of 23,408 contigs in 433 barley genotypes yielded ~ 8 billion reads (0.56 Tb of data; **Supplementary note** and **Supplementary Table 1**). Cumulatively, the captured regions comprised approximately 13.8 Mbp (**Supplementary Table 2**) and 1.33 Mbp of which resided in the coding regions (CDS). Per sample analysis of the coverage revealed that approximately 87% of the captured regions were covered above the SNP calling threshold and that the between-sample variation was relatively low with the median depth of coverage varying from 45 to 130 (**Supplementary Fig. 1a**). The SNP calling pipeline identified 544,318 high quality SNPs including approximately 190,000 of singletons (**Supplementary Table 2**). On average, each sample carried 6% and 3% of the homozygous and heterozygous SNPs per variant position, respectively (**Supplementary Fig. 1b**, **Supplemental note**). Of all the SNPs, 37,870 resided in CDS and approximately 43% of them fell into the non-neutral category based on the predictions of the snpEff software. The CDS were more conserved than the non-coding regions with the average SNP density of 29 and 41

SNPs per Kbp, respectively. 45% of the SNPs were located on the barley genetic map, whereas for 37% of the SNPs, only the chromosome could be assigned (**Supplementary Fig. 2a**). The transition to transversion bias (Ti/Tv) genome-wide ratio (2.48) was in par with the genome-wide Arabidopsis estimates (2.4) (**Supplementary note**).[24] The minor allele frequency (MAF) spectra did not reveal any systematic bias, e.g. lack of rare variants often attributed to the ascertainment bias, and resembled the MAF distributions simulated based on the standard neutral coalescent model - large proportion of rare polymorphisms and rapid exponential decrease in the number of SNPs with the higher MAFs (**Supplementary Fig. 3**).

Admixture and linkage disequilibrium

In domestication studies, where patterns of genetic variation are contrasted between wild and domesticated genotypes, it is critical to distinguish these subgroups and exclude genotypes of unverified provenance. The PCA revealed two distinct clusters corresponding to the domesticated and wild subspecies with the multiple genotypes scattered between these clusters (**Fig. 1a**). STRUCTURE analysis revealed patterns of admixture in 36% and 12% of the domesticated and wild genotypes, respectively, corresponding to the intermediate PCA genotypes (**Supplementary Fig. 4**). In domesticates, the landraces constituted 95% of the admixed individuals and they did not correspond to any specific locality (**Supplementary table 4**). Similarly, in wild subspecies, the admixed genotypes were spread all over the Fertile Crescent, indicating that the admixture was not restricted to any particular geographical area. These genotypes of ambiguous provenance were removed from the further analyses.

Landraces and cultivars are two recognized groups of domesticated barley. The former are tentatively defined as locally adapted varieties traditionally cultivated and selected by farmers in the field, whereas the latter are the products of the breeding programs.[25] Despite these generally accepted differences in definitions, sorting extant domesticated genotypes into these two groups is not without controversy, partly owing to the use of landrace material in modern breeding. In this study, the landraces did not differentiate from the cultivars based on the result of both STRUCTURE and PCA (**Supplementary Fig. 5**) and thus were treated as a single group of domesticated barley.

The extent of linkage disequilibrium (LD) characterizes the recombination landscape and haplotype diversity of a group. The LD is mostly maintained by the physical properties of

a chromosome, as a function of physical distance between markers. Additionally, other processes, such as selection and demographic history, may create peculiar LD patterns. In wild and domesticated barley, the LD decayed to the background levels at the distances of 0.45 cM and 8.55 cM, respectively, and showed some dependency on MAF (**Fig. 1b**; **Supplementary note**). Such 20-fold difference in the extent of LD between the groups apparently resulted from the limited amount of historical recombination in domesticated barley and was consistent with previous reports.[26,27] The rate of LD decay varied between the individual chromosomes in a range from 0.2 to 0.8 cM in the wild barley and in a much bigger range from 2 cM to 26 cM in the domesticated subspecies (**Supplementary Fig. 6**).

Effect of domestication on genetic diversity

Domestication results in loss of genetic diversity via the so-called domestication bottleneck.[28] In this study, wild barley comprised ~7x more segregating sites than domesticated genotypes (**Fig. 1c**) and 88% of the sites resided in the non-coding regions. As measured by Watterson's θ_w, an unbiased estimator, which provides correction for the sample size, the mutation rate in wild barley (θ_w=7.36 x 10^{-3}) was 5x higher than in the domesticates (θ_w=1.47 x 10^{-3}). Nei's nucleotide diversity π_n or θ_π, an estimator of average number of pairwise differences between two randomly drawn sequences per nucleotide, suggested that the domesticates (π_n=1.53 x 10^{-3}) retained 52% of the wild nucleotide diversity (π_n=2.97 x 10^{-3}). Reduction of diversity in domesticated barley compared with the wild subspecies was similar to that in tomato and soybean but higher than in maize (17%), rice (20%) and common bean (17%).[11–15] Rare alleles were enriched in wild barley (Tajima's D=-1.908; **Supplementary note**), whereas, MAF spectra in the domesticates were severely skewed toward common alleles (**Fig. 1d**). The former indicated that, similarly to Arabidopsis, genome-wide nucleotide variation in wild barley did not follow assumptions of a Wright-Fisher neutral model, hinting either at a rapid population growth after a bottleneck or at a large-scale range expansion in wild barley demographic history.[29,30] The difference between the D values in wild and domesticated barley was consistent along the individual chromosomes (**Supplementary Fig. 7**).

To detect the episodes of positive selection or selective sweeps, we computed genome-wide and rolling Fay and Wu's H_{norm} values using a panel of 64,977 SNPs with assigned ancestral status (**Fig. 1c**). The amplitude of interspecies variation of H_{norm} was low compared to

D, indicating lower influence of demography on the H_{norm} estimator. The H_{norm} values were lower in coding than in non-coding SNPs, suggesting a link between functionality and the abundance of the high frequency derived alleles.

In wild barley, the private alleles constituted 81% of the total number of SNPs, whereas their share in domesticates was much lower (14%). Wild barley contained ~ 27x more private polymorphisms than the domesticates (**Fig. 1e**). The MAF distribution of the shared alleles was severely skewed toward the more common alleles. This strongly suggests that most shared polymorphisms originated from the common ancestor (identity-by-descent) rather than occurred independently in two subspecies (identity-by-state). The small share of private alleles in domesticates along with a reduction of diversity, an increase of LD and a depletion of rare alleles suggested a strong domestication bottleneck in cultivated barley.[31,32]

The genome-wide divergence between wild and domesticated barley (Fst=0.29) was similar to the previously reported values in barley (Fst=0.26), soybean (Fst=0.29) and rice (Fst=0.27), but higher than in maize (Fst=0.11).[11,12,15,33] Differentiation between the non-synonymous polymorphisms (Fst=0.27) was higher than that of the synonymous SNPs (Fst=0.25), suggesting the action of adaptive selection under barley domestication. Divergence of chromosomes 4 and 7 was significantly higher than that of the other chromosomes ($p < 0.01$) (**Supplementary Fig. 8**).

Phylogeography of barley domestication

We identified nine population of wild barley using STRUCTURE and phylogenetic analyses (**Fig. 2abc**; **Supplementary Fig. 9**). Six populations, Carmel and Galilee (CG); Golan Heights (GH); Hula Valley and Galilee (HG); Judean Desert and Jordan Valley (JJ); Negev Mountains (NM); Sharon, Coastal Plain and Judean Lowlands (SCJ), were concentrated in the South Levant and the other three, Lower Mesopotamia (LM), North Levant (NL) and Upper Mesopotamia (UM), occupied large areas of the Northern and Eastern Fertile Crescent. Habitats of the wild populations were distinct with very few immigrants and admixed genotypes occurring mostly in the border overlapping areas (**Supplementary Figs. 10, 11**). Only 23 wild genotypes had highly admixed ancestry and could not be attributed to any of the nine populations (**Fig. 2c**). Rooting the phylogeny to the outgroup *Hordeum* species enabled tracing the population differentiation in time. The most ancestral population split was located

in the north of modern Israel followed by migration of the populations along the two routes, the short one to the south until the Negev Desert and the longer route to the eastern part of the Fertile Crescent (**Fig. 2ab**). Hierarchy of the populations splits on the phylogram, as a function of genetic distance, followed geographical patterns of differentiation and spread of the wild populations away from the Northern Israel, indicating the isolation-by-distance model.

In both the genome-wide STRUCTURE and phylogenetic analyses the cluster of domesticated barley appeared as a sister group relative to all wild barley populations, suggesting reticulated history of the domesticated genotypes. To investigate this, using the maximum likelihood distance approach, we identified wild population ancestors for 1,232 contigs in all domesticates. 60% of the contigs in the domesticated barley were monophyletic (**Supplemental Fig. 12**). All nine wild barley populations contributed to the cumulative domesticated genome and their individual contributions did not drastically differ between the domesticated genotypes (**Fig. 2d, Supplemental Figs. 13, 14**).

Footprints of domestication-related selection

We used a combination of mean r^2, H_{norm} and π_w/π_d tests, which reveal various signatures of selection, to obtain a catalog of candidate regions and genes that likely experienced selection in the domestication context. Both selective sweeps and background selection may result in elevated local LD manifested by positively correlated recombination and nucleotide variation rates.[34] In this study, LD strongly and negatively correlated with nucleotide diversity (Pearson's $r = -0.68$; $p < 0.001$). The patterns of rolling r^2 values were heterogeneous along the chromosomes, and, in the domesticates, the amplitude of variation was high compared with wild barley. The LD scan identified twelve regions on chromosomes 1H, 2H, 3H, 4H and 5H, significantly deviating from the mean values in wild and domesticated barley (**Supplementary figure 15**; **Supplementary table 5**). In wild barley, two of the outliers were co-located with the major flowering loci *PpdH1* and *VRN-H1*, for which mutations lead to reduced photoperiod and vernalization sensitivity, respectively. The location of the *VRN-H1* gene, a key barley regulator of vernalization response, within the region of extended LD on the chromosome 5H in wild barley is noteworthy. It has been shown that 98% of wild barley possess the wild-type winter *VRN-H1* allele, which delays flowering until the vernalization requirement is fulfilled.[35] The *VRN-H1* gene is tightly linked to several flowering-related genes, including *HvPHYC*, a

homolog of Arabidopsis *PHYTOCHROME C* gene.[36] The variation in *HvPHYC* modulates photoperiodic flowering and the early flowering mutant allele is private to domesticates. It is tempting to speculate that, in wild barley, extended LD at *VRN-H1* is a signature of background selection, purging novel mutations and maintaining integrity of this gene cluster, which is apparently critical for flowering. In the domesticates, the H_{norm} and π_w/π_d scans for selective sweeps identified 13 regions (10-31 cM) and 178 gene-bearing contigs, including 41 target genes, on all barley chromosomes (**Fig. 3a**, **Supplementary Fig. 16**, **Supplementary tables 5**, **6**). 94 of the outlier contigs were located on the genetic map.

In both tests, candidate regions on chromosomes 3 and 7 overlapped with the spike-brittleness locus *Brt1/2* and the *NUD* locus, controlling the naked (hulless) grain phenotype.[5,37] However, the *NUD* gene itself did not carry a selection signatures (π_w/π_d=0.7) and thus was not the target of selection. Indeed, both hulless and hulled genotypes are ubiquitously present in the domesticated barley genepool and apparently represent an improvement but not domestication trait.[38]

Among the candidate domestication loci were homologs of genes of light signaling, photoperiod, circadian clock, abscisic acid (ABA) and carbohydrate metabolism genetic pathways (**Fig. 3c**). None of the candidate domestication genes identified in this study have been functionally characterized in barley, however, putative function can often be inferred from homology. In other crops, several flowering loci have been reported among the candidate domestication genes.[12,14] In common bean, the orthologs of light signaling genes, encoding two different members of the same protein complex CONSTITUTIVELY PHOTOMORPHOGENIC 1 (COP1) and CULLIN4 (CUL4) have been independently targeted by selection in two separate domestication events Mesoamerican and Andean, respectively.[14] Intriguingly, HvCUL4 (seq442) and an ortholog encoding Arabidopsis SUPRESSOR OF PHYA 2 (SPA2, seq108), encoding another member of the COP1-CUL4-SPA protein complex, carried very strong signatures of selection (CUL4, H_{norm}=-5.1; SPA2, π_w/π_d=64)(Fig. 2b).[39] In maize, a homolog of Arabidopsis *AGAMOUS-LIKE20* gene (AGL), encoding SUPRESSOR OF OVEREXPRESSION OF CONSTANS1 (SOC1) protein, was a domestication candidate (mentioned as an ortholog of rice OsMADS56).[12] A barley ortholog of OsMADS56 (seq411) resided within a sharp selection π_w/π_d signal on the chromosome 1H but did not carry a signature of selection itself. However, three other AGL candidate genes were in the 75th percentile of the π_w/π_d scan. Another notable domestication candidate, *HvGCN5* (seq612; π_w/π_d = 104), encoding a homolog of the GNAT/MYST histone acetyltransferases, has been

implicated in regulation of seed maturation, dormancy and germination based on the expression analysis and its regulation by the phytohormone ABA.[40]

Phylogeographic analyses demonstrated that, similarly to the genome-wide data, the candidate domestication loci had mosaic ancestry, consisting of nine wild barley populations (**Fig. 2e**). This illustrates that the recurrent historic admixture of heterogeneous wild germplasm into the domesticates was relevant to the process of domestication.

Discussion

Accumulating molecular and archaeological evidence has been shifting our views on the process of barley domestication from an early model of a fast monophyletic event toward a more complex protracted model of domestication.[41] The protracted model postulates a slower gradual process of domestication and a polyphyletic origin of the barley crop.[6] Accordingly, domestication traits may reach fixation in a cultigen on a longer timescale and the archaeological data corroborate this assumption.[42] Distinction between the two models and their conceptualization has been a subject of recent debates.[22,43] At least two independent domestication events occurred in the Middle East based on the evolution patterns of the *Btr1*/*2* locus.[5] According to a recent more complex model of barley domestication, five wild barley populations contributed to the genomes of modern landraces.[23] However, whether the mosaic composition of the domesticated germplasm was related to domestication remained unclear. We significantly expanded on this model by demonstrating that nine wild barley populations gave rise to the domesticated genomes. Moreover, the mosaic ancestry of the domestication candidate genes suggested a direct relationship between the recurrent introgressions of wild germplasm and the process of domestication, thus supporting the protracted model.

The concept of a crucial domestication syndrome (DS) trait, which is present in a derived form in all domesticates and either segregates or is fixed in the wild form, has been coined in order to clarify the domestication glossary.[7] Understanding of the barley crucial DS traits is extremely limited. Spike brittleness is the only studied example of a crucial DS trait in barley. Besides, seed dormancy, synchrony of flowering, and number and angle of tillers have been suggested to constitute the DS.[28,44] Here, we cataloged novel candidate genes residing in the selective sweep regions associated with domestication. The top candidates were enriched for homologs of genes implicated in regulation of light signaling, photoperiod response, circadian clock, ABA and carbohydrate metabolism - processes closely linked to the putative

DS traits. Intriguingly, domestication-related selection seems to converge on homologous developmental pathways and protein complexes in different species. The signatures of selection in the components of the E3 ubiquitin-ligase COP1-CUL4-SPA in barley and bean species is a particularly vivid example. The COP1-CUL4-SPA complex is a critical component of the far-red light signaling, photoperiod and circadian clock pathways.[45] We hypothesize that, in this case, parallelism in the putative targets of selection may stem from commonality of the crop adaptation to agricultural practices. Dramatic changes in the light environment resulted from cultivating barley and bean plants in dense stands compared with the wild ancestor species. It might be the key to understanding the involvement of the modulators of light signaling, circadian clock and shade avoidance pathways in domestication.[46]

In conclusion, this study provides a valuable resource to identify and characterize novel barley domestication genes and thereby unravel the increasingly complex model of barley domestication. Informed exploitation of wild germplasm may help introduce favorable alleles into cultivated barley particularly at genomic regions of low diversity. We demonstrate that population genomics facilitates identification of such loci. Therefore, we expect that the results of this study will boost future breeding efforts aimed at alleviating detrimental effects of genetic bottlenecks and selfing on genetic diversity.

Materials and methods

Plant material

A panel consisting of 344 wild and 89 domesticated lines and a single genotype of *H. vulgare* ssp. a*griocrithon* were selected to maximize genetic diversity and to cover the entire range of the wild and landrace barley habitats in the Fertile Crescent (**Supplementary table 4**). The advanced barley cultivars were sampled to represent Northern European, East Asian, North American and Australian breeding programs. The largest part of the germplasm set, 98% of wild and 40% of domesticated barley genotypes originated from the countries of the Fertile Crescent area. The selection of domesticated barley originated from various breeding programs and represented the whole variety of cultivated barley lifeforms, namely two- (71%) and six-row (29%) genotypes with winter (45%) and spring (55%) growth habits based on the passport data. All material was purified by single-seed descent to eliminate accession heterogeneity.

Leaf samples for DNA extraction were collected from a single 3-week old plant of every genotype. The DNA extraction was performed using the DNeasy Plant Mini kit (QIAGEN, Hilden, Germany) following manufacturer's recommendations. The DNA samples were quantified using the NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA) and the DNA integrity was assessed using electrophoresis in the 0.8% agarose gel.

Design of the targeted enrichment assay

To interrogate the genetic diversity of barley populations in the domestication context, we designed a custom target enrichment sequencing assay that included the loci implicated in the candidate domestication pathways in barley and other species and the neutral loci to attenuate effects of the biased selection.

A set of genic sequences comprised a comprehensive subset of loci related to flowering time and development of meristem and inflorescences. Additionally, it contained a selection of genes related to agronomic traits putatively affected by domestication, e.g. tillering, seed dormancy, carbohydrate metabolism. First, scientific literature was mined for the genes

implicated in the aforementioned processes and the corresponding nucleotide sequences were extracted from NCBI GenBank. Second, flowering genes from the other grass species, such as Brachypodium and rice, were selected.[47] Third, a set of 259 Arabidopsis genes characterized by the flowering-related gene ontology (GO) terms that have been confirmed experimentally was assembled (**Supplementary table 7**). The barley homologs of all these genes were extracted from the NCBI barley UniGene set (Hv cDNA, cv. Haruna Nijo, build 59) either by the BLASTN search (e-value < 1e-7) or, in the case of Arabidopsis genes, by searching the annotation table downloaded from the NCBI UniGene server (ftp://ftp.ncbi.nih.gov/repository/UniGene/Hordeum_vulgare). This table was further used to reciprocally extract additional Hv homologs based on the Arabidopsis gene identifiers. If the BLAST search failed to identify a reliable Hv homolog, the homologs were searched in the barley High and Low confidence genes (MLOC cDNA)[48] and in the HarvEST unigene assembly 35 (http://harvest.ucr.edu).

Open reading frames (ORF) of Hv cDNA were predicted using OrfPredictor guided by the BLASTX search against Arabidopsis TAIR 10 database.[49] The predicted ORFs were aligned to the genomic contigs of barley cultivars Morex, Bowman and Barke using the Spidey algorithm implemented in the NCBI toolkit. The ORFs of the selected sequences were categorized as complete or partial based on the presence or absence of putative start and stop codons. The complete complementary DNA (cDNA) were selected and, if the complete cDNA was absent, partial gDNA and cDNA were included in the dataset. For several genes with previously characterized intronic regions, e.g. predicted to contain regulatory elements, complete genomic DNA (gDNA) were selected. In case when only partial cDNA was available, chimeric sequences were assembled from the Hv, MLOC and HarvEST cDNA using SeqMan software (DNASTAR Lasergene®8 Core Suite, Madison, WI, USA). The selected sequences were cross-annotated with NCBI UniGene Hv and IBGSC MLOC identifiers using reciprocal BLASTN (e-value < 1e-05). In addition to the coding regions and introns, the selection contained sequences up to 3 kilobase pairs (Kbp) upstream of the predicted start codon, which presumably corresponded to regulatory promoter regions. The target selection workflow is schematically outlined on **Supplementary Fig. 17**.

A set of 1000 additional HarvEST genes was randomly selected such that they had no homology to target genes as determined by BLASTN and evenly spread over all barley linkage groups according to the GenomeZipper map.[50] The 100-bp stretches of each of these genes were included in the enrichment library.

The target sequences were filtered and tiled with 100-bp selection baits using Nimblegen proprietary algorithm and the library of baits was synthesized as a part of the SeqCap EZ enrichment kit (design name 130830_BARLEY_MVK_EZ_HX3; Roche NimbleGene, Madison, WI). Barcoded Illumina libraries were individually prepared, then enriched and sequenced in 24-sample pools at the Cologne Center for Genomics facilities following the standard protocols.

The genic sequences from a variety of barley genotypes were used to design the enrichment library to ensure that the longest ORF and promoter regions were selected. However, most advanced physical and genetics maps have been developed for the barley cultivar Morex. Since mapping information is essential for the downstream analyses, the so-called Morex genomic contigs were used as a mapping reference provided that they comprised the entire regions tiled by the baits (**Supplementary Table 1**). If such contigs were not available, the genomic contigs of the barley genotypes Bowman and Barke or the templates that were used for the bait design were included in the mapping reference.

Targeted enrichment assays are known to capture large amount of sequences, which are homologous to the selected targets but not included in the original enrichment design. Such off-target enrichment is known to generate high quality SNP datasets.[51] To identify such regions, the Illumina reads from 10 randomly selected barley genotypes were mapped to the complete Morex genome reference set.[48] All genomic contigs that had at least one read mapped to them were included in the mapping reference. This thinning of the complete Morex genome dataset helped avoid excessive computational load in the downstream steps of the SNP calling pipeline. The Morex contigs were masked with “N”s at the regions of longer than 100 bp that exhibited more than 97% homology with the original capture targets. The PopSeq and IBGSC ‘Morex” genetic maps were used to extract mapping positions of the reference sequences.[48,52]

Read mapping and SNP calling

The SNP calling pipeline consisted of three modules: quality control and filtering of Illumina read libraries; mapping the reads to the reference; and SNP calling, genotyping and filtering (**Supplementary Fig. 18**). The quality parameters of the paired-end Illumina libraries were assessed using FastQC tool (v. 0.11.2; http://www.bioinformatics.babraham.ac.uk/projects/fastqc). After filtering out optical

duplicates, resulting from a PCR amplification, using the CD-HIT-DUP software (v. 0.5)[53], the paired-end read files were merged and henceforth treated as a single-end dataset. Next, based on the FastQC results, the reads were trimmed from both ends to remove low quality sequencing data, filtered to remove the remaining adaptor sequences and low-complexity artifacts using the FASTX toolkit (v. 0.0.14; http://hannonlab.cshl.edu/fastx_toolkit). The sequencing errors in the dataset were corrected using the Bloom-filter tool Lighter with the conservative set of parameters: k-mer size 23, alpha 0.2, and maximum corrections per read 2.[54] The reference file was indexed for the downstream processing using Burrows-Wheeler Aligner 0.5.9-r16 (BWA), SAMtools and Picard tools (http://broadinstitute.github.io/picard).[55,56] The groomed read datasets were mapped onto the reference genome using BWA (modules 'aln' and 'samse') with the following stringency parameters: missing probability (-n) 0.05, maximum number of gaps (-o) 2, and gap extensions (-e) 12. Some of the reference loci were present in the form of cDNA and the gDNA-derived reads mapped onto such targets may generate false positive SNP calls at the intron-exon junctions. To alleviate this problem, the reads that mapped to cDNA-derived targets were extracted, additionally trimmed by 14 bp from each end and remapped following the described procedure. Reads that mapped to several locations were filtered out.

The regions containing INDELs are prone to alignment errors and thus may generate false positive polymorphism calls. To tackle this issue, the reads were locally realigned around INDELs using RealignerTargetCreator and IndelRealigner walkers of the GATK suite.[57] Raw SNP calling was performed for each sample library separately using the GATK UnifiedGenotyper walker with the default parameters. Afterwards, the output lists of polymorphisms, the so-called VCF files, were merged into a multi-sample VCF file using the GATK CombineVariants walker. The *de novo* SNP discovery using GATK emits a call only if there was a nucleotide substitution compared with the reference genome without distinction between a reference allele and zero coverage (missing data). To obtain a dataset containing both reference and non-reference calls, the genotyping mode of the GATK UnifiedGenotyper was applied to the individual bam files using the raw calls as the reference set of alleles. The output VCF files were merged into a multi-sample VCF file, which contained only the biallelic homozygous SNPs passing the following filters: depth of coverage (DP) > 8, mapping quality (MQ) > 20, Fisher strand (FS) < 60. For the downstream analyses, all heterozygous SNPs were a treated as missing data. This pipeline was implemented in a series of bash scripts adapted for high-performance parallelized computation.

Characterization of the assay

To describe the capture quality parameters, two different sets of reference regions were defined as following: target capture regions tiled by the baits and the regions covered by the reads outside of the target and predicted capture regions. *De facto* captured regions were defined as those with the depth of coverage ≥ 8, set as the SNP calling threshold, in at least one of the samples. The depth of coverage was analyzed using bedtools v.2.16.2, vcftools v.0.1.11 and R.[58,59] Functional effects of the SNPs were predicted using SnpEff 3.6b software using the custom CDS coordinates as a reference genome.[60] The CDS coordinates were mapped on the target genomic contigs based on the Spidey predictions and extracted from the IBGSC annotation file for the additional genomic contigs. Transition / transversion ratios (Ti/Tv) were calculated using VariantEval walker of the GATK package.

Population genetics analyses

Minor allele frequency (MAF) spectra for various genomic regions and bootstrapping of the rare SNPs (1000 random draws) were calculated using R. The SNPs were tentatively divided into neutral and non-neutral subsets defined by the SnpEff flags, which, for the neutral subgroup, carried the UTR, DOWNSTREAM, UPSTREAM, INTERGENIC, INTRON and SILENT SnpEff flags. The vcf files were converted into the ped format using tabix utility of Samtools, PLINK 1.9.[61] For estimations of population parameters, only a subset of SNPs with MAF > 0.05, missing data frequency (MDF) < 0.5 was selected. For the structure and principal component analyses (PCA), the SNPs in very high LD (r^2 >0.99) were pruned using PLINK 1.9. The PCA was performed using smartpca utility of the EIGENSOFT software version 5.0.2.[62]

The linkage disequilibrium (LD) estimator r^2 was calculated for each SNP pair separately in the wild and domesticated barley subsets using PLINK 1.9. The background LD was defined as an average of the interchromosomal r^2 values (95^{th} percentile). Rate of LD decay was estimated using a nonlinear least-square (nls) regression fit to the intrachromosomal or intergenic r^2 values using Hill and Weir's formula, providing adjustment for sample size.[63]

The nls regression analysis was implemented in R. The LD decay value was defined at the intersection point of the regression curve with the background LD. To estimate the robustness of LD estimated in unbalanced samples, i.e. varying number of individuals or markers, the balanced sub-samples were 1000x randomly drawn from the larger sub-group. Variation of the LD estimates in these bootstrap experiments was assessed using standard summary statistics.

The structure of barley populations was inferred using fastSTRUCTURE software, which implements Pritchard's STRUCTURE algorithm in a fast and resource-efficient manner.[64] This algorithm very efficiently detects recent gene flow events but not the historical admixture. The runs were executed with 20 iterations for a predefined number of population (K). To identify admixture between wild and domesticated barley K was set at 2. The optimal K for wild barley was chosen to represent the model with maximum marginal likelihood tested for K from 2 to 25 as implemented in fastSTRUCTURE. The geographic centers of the populations were calculated as a median of the latitude and longitude of the genotypes comprising the populations. The vector geographic map dataset was downloaded from Natural Earth repository and manipulated in R (http:/www.naturalearthdata.com).

The Maximum Likelihood (ML) phylogeny was constructed from the genome-wide SNP dataset using GTRCAT model with Lewis' ascertainment bias correction to account for the absence of invariant sites in the alignment and the majority-rule tree-based criteria for bootstopping (autoMRE_IGN) implemented in RAxML 8.2.8.[65] Wild barley *H. bulbosum* and *H. pubiflorum* were used as outgroup species. The trees were visualized and collapsed using Dendroscope 3.5.7.[66]

To estimate ancestry of the domesticated barley loci, we calculated pairwise ML distances between each wild and domesticated genotypes for each locus using GTRGAMMA model in RAxML 8.2.8.[65] For each domesticated genotype, if a locus had the smallest ML distance only with a single wild population, this population was deemed ancestral.

The diversity parameters, such as number of segregating sites (S), Watterson's estimator (θ_w) per genotyped site[67], Nei's (sometimes referred as Tajima's) nucleotide diversity (π) per genotyped site[68], fixation index (Fst)[69], as well as the frequency-based selection tests, such as Tajima's D[70] and normalized Fay and Wu's H_{norm}[71] were calculated separately for the wild and domesticated barley using mstatspop software with 1000 permutations (release 0.1b 20150803; http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html). To determine ancestral status of the SNPs, which is a prerequisite for the H test, the SNPs were

genotyped in two wild barley species, *H. bulbosum* and *H. pubiflorum*, and alleles that were identical in both species were tentatively assigned as ancestral. The genotyping was performed following the mapping and SNP calling pipeline described above using the *Hordeum* exome Illumina datasets.[72]

The D and H_{norm} values vary greatly at different genomic regions due to the neutral random processes, e.g. genetic drift, and the range of this variation depends on the properties of the examined populations, such as the population size and demographic history. To estimate confidence intervals for the distribution of the D and H_{norm} under a Wright-Fisher neutral model in the wild and domesticated barley, coalescent simulations of 1000 datasets were performed using the ms software with the number of samples (n) and θ_w used as the variable parameters describing the populations.[73] Variation of the D and H_{norm} in the simulated neutral datasets was assessed using the msstats and statsPs software (https://github.com/molpopgen/msstats).

The selective sweeps and selection signatures in individual loci were discovered using the diversity reduction index ($\pi_{w(ild)}/\pi_{d(omesticated)}$) and the H_{norm} test. The scans were performed in the wild and domesticated barley sub-sets genome-wide in the 10-cM windows with a sliding step of 1 cM and separately for the individual loci (cut-off > 5 SNPs). The sweeps and targets of selection were statistically defined based on the z-score test for outliers (p-value < 0.05) for the π_w/π_d and based on the simulated thresholds of neutral variation for the H_{norm} test (p-value < 0.001). The overlapping outlier windows were merged into putatively selected regions.

Code availability

The scripts and the accompanying files used for analysis are available in an online repository at https://github.com/artempankin/korffgroup.

Availability of data and materials

Raw Illumina sequence reads have been deposited at NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA329198.

Competing Interests

The authors declare no competing financial interests.

Funding

This work was supported by the Max Planck Society and by DFG grants SPP1530 ("Flowering time control: from natural variation to crop improvement") and the Excellence Cluster EXC1028. A.P. was supported by an IMPRS fellowship from the Max Planck Society.

Author Contributions

A.P. and M.K. conceived and designed the experiments. J.A. and D.B. conducted the enrichment sequencing experiments. A.P. analyzed the data. A.P. and M.K. wrote the manuscript.

Acknowledgement

We cordially thank Kerstin Luxa, Teresa Bisdorf, Caren Dawidson, Elisabeth Kirst and Andrea Lossow for excellent technical assistance. We thank Eyal Fridman, Hakan Özkan and Benjamin Kilian for barley seeds.

References

1. Lev-Yadun, S., Gopher, A. & Abbo, S. The Cradle of Agriculture. *Science* **288,** 1602–1603 (2000).
2. Dawson, I. K. *et al.* Barley: a translational model for adaptation to climate change. *New Phytol.* **206,** 913–931 (2015).
3. Zohary, D., Hopf, M. & Weiss, E. *Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin*. (OUP Oxford, 2012).
4. Gottlieb, L. D. Genetics and morphological evolution in plants. *Am. Nat.* 681–709 (1984).
5. Pourkheirandish, M. *et al.* Evolution of the Grain Dispersal System in Barley. *Cell* **162,** 527–539 (2015).
6. Purugganan, M. D. & Fuller, D. Q. Archaeological data reveal slow rates of evolution during plant domestication. *Evolution* **65,** 171–183 (2011).
7. Abbo, S. *et al.* Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. *Trends Plant Sci.* **19,** 351–360 (2014).
8. Hammer, K. Das domestikationssyndrom. *Die Kulturpflanze.* **32,** 11–34 (1984).
9. Shi, J. & Lai, J. Patterns of genomic changes with crop domestication and breeding. *Curr. Opin. Plant Biol.* **24,** 47–53 (2015).
10. Ross-Ibarra, J., Morrell, P. L. & Gaut, B. S. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl. Acad. Sci.* **104,** 8641–8648 (2007).
11. Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490,** 497–501 (2012).
12. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44,** 808–811 (2012).
13. Lin, T. *et al.* Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46,** 1220–1226 (2014).
14. Schmutz, J. *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46,** 707–713 (2014).
15. Zhou, Z. *et al.* Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33,** 408–414 (2015).
16. Badr, A. *et al.* On the Origin and Domestication History of Barley (Hordeum vulgare). *Mol. Biol. Evol.* **17,** 499–510 (2000).

17. Molina-Cano, J. L., Moralejo, M., Igartua, E. & Romagosa, I. Further evidence supporting Morocco as a centre of origin of barley. *Theor. Appl. Genet.* **98,** 913–918 (1999).

18. Negassa, M. Patterns of phenotypic diversity in an Ethiopian barley collection, and the Arussi-Bale Highland as a center of origin of barley. *Hereditas* **102,** 139–150 (1985).

19. Dai, F. *et al.* Tibet is one of the centers of domestication of cultivated barley. *Proc. Natl. Acad. Sci.* **109,** 16969–16973 (2012).

20. Morrell, P. L. & Clegg, M. T. Genetic evidence for a second domestication of barley (Hordeum vulgare) east of the Fertile Crescent. *Proc. Natl. Acad. Sci.* **104,** 3289–3294 (2007).

21. Allaby, R. G. Barley domestication: the end of a central dogma? *Genome Biol.* **16,** 1 (2015).

22. Fuller, D. Q., Asouti, E. & Purugganan, M. D. Cultivation as slow evolutionary entanglement: comparative data on rate and sequence of domestication. *Veg. Hist. Archaeobotany* **21,** 131–145 (2012).

23. Poets, A. M., Fang, Z., Clegg, M. T. & Morrell, P. L. Barley landraces are characterized by geographically heterogeneous genomic origins. *Genome Biol.* **16,** 1 (2015).

24. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *science* **327,** 92–94 (2010).

25. Zeven, A. C. Landraces: a review of definitions and classifications. *Euphytica* **104,** 127–139 (1998).

26. Morrell, P. L., Toleno, D. M., Lundy, K. E. & Clegg, M. T. Low levels of linkage disequilibrium in wild barley (Hordeum vulgare ssp. spontaneum) despite high rates of self-fertilization. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 2442–2447 (2005).

27. Caldwell, K. S., Russell, J., Langridge, P. & Powell, W. Extreme Population-Dependent Linkage Disequilibrium Detected in an Inbreeding Plant Species, Hordeum vulgare. *Genetics* **172,** 557–567 (2006).

28. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127,** 1309–1321 (2006).

29. Schmid, K. J., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B. & Mitchell-Olds, T. A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169,** 1601–1615 (2005).

30. 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).

31. McVean, G. A. A genealogical interpretation of linkage disequilibrium. *Genetics* **162,** 987–991 (2002).
32. Nei, M., Maruyama, T. & Chakraborty, R. The bottleneck effect and genetic variability in populations. *Evolution* 1–10 (1975).
33. 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).
34. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134,** 1289–1303 (1993).
35. Cockram, J., Hones, H. & O'Sullivan, D. M. Genetic variation at flowering time loci in wild and cultivated barley. *Plant Genet. Resour.* **9,** 264–267 (2011).
36. Pankin, A. *et al.* Mapping-by-sequencing identifies HvPHYTOCHROME C as a candidate gene for the early maturity 5 locus modulating the circadian clock and photoperiodic flowering in barley. *Genetics* **198,** 383–396 (2014).
37. Taketa, S. *et al.* Monophyletic origin of naked barley inferred from molecular analyses of a marker closely linked to the naked caryopsis gene (nud). *Theor. Appl. Genet.* **108,** 1236–1242 (2004).
38. Saisho, D. & Purugganan, M. D. Molecular phylogeography of domesticated barley traces expansion of agriculture in the Old World. *Genetics* **177,** 1765–1776 (2007).
39. Zhu, D. *et al.* Biochemical characterization of Arabidopsis complexes containing CONSTITUTIVELY PHOTOMORPHOGENIC1 and SUPPRESSOR OF PHYA proteins in light control of plant development. *Plant Cell* **20,** 2307–2323 (2008).
40. Papaefthimiou, D., Likotrafiti, E., Kapazoglou, A., Bladenopoulos, K. & Tsaftaris, A. Epigenetic chromatin modifiers in barley: III. Isolation and characterization of the barley GNAT-MYST family of histone acetyltransferases and responses to exogenous ABA. *Plant Physiol. Biochem.* **48,** 98–107 (2010).
41. Allaby, R. G., Fuller, D. Q. & Brown, T. A. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc. Natl. Acad. Sci.* **105,** 13982–13986 (2008).
42. Tanno, K. & Willcox, G. Distinguishing wild and domestic wheat and barley spikelets from early Holocene sites in the Near East. *Veg. Hist. Archaeobotany* **21,** 107–115 (2012).

43. Heun, M., Abbo, S., Lev-Yadun, S. & Gopher, A. A critical review of the protracted domestication model for Near-Eastern founder crops: linear regression, long-distance gene flow, archaeological, and archaeobotanical evidence. *J. Exp. Bot.* **63,** 4333–4341 (2012).

44. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14,** 840–852 (2013).

45. Zhu, L. *et al.* CUL4 forms an E3 ligase with COP1 and SPA to promote light-induced degradation of PIF1. *Nat. Commun.* **6,** (2015).

46. Müller, N. A. *et al.* Domestication selected for deceleration of the circadian clock in cultivated tomato. *Nat. Genet.* **48,** 89–93 (2016).

47. Higgins, J. A., Bailey, P. C. & Laurie, D. A. Comparative genomics of flowering time pathways using Brachypodium distachyon as a model for the temperate grasses. *PLoS One* **5,** e10065 (2010).

48. International Barley Genome Sequencing Consortium (IBGSC)., I. B. G. S. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491,** 711–716 (2012).

49. Wheelan, S. J., Church, D. M. & Ostell, J. M. Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.* **11,** 1952–1957 (2001).

50. Mayer, K. F. *et al.* Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* **23,** 1249–1263 (2011).

51. Guo, Y. *et al.* Exome sequencing generates high quality data in non-target regions. *BMC Genomics* **13,** 1 (2012).

52. Mascher, M. *et al.* Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.* **76,** 718–727 (2013).

53. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).

54. Song, L., Florea, L. & Langmead, B. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol.* **15,** 1 (2014).

55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

56. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

57. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

58. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

59. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinforma.* 11.12. 1–11.12. 34 (2014).

60. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).

61. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 1 (2015).

62. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2,** e190 (2006).

63. Hill, W. G. & Weir, B. S. Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* **33,** 54–78 (1988).

64. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197,** 573–589 (2014).

65. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

66. Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61,** 1061–1067 (2012).

67. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7,** 256–276 (1975).

68. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76,** 5269–5273 (1979).

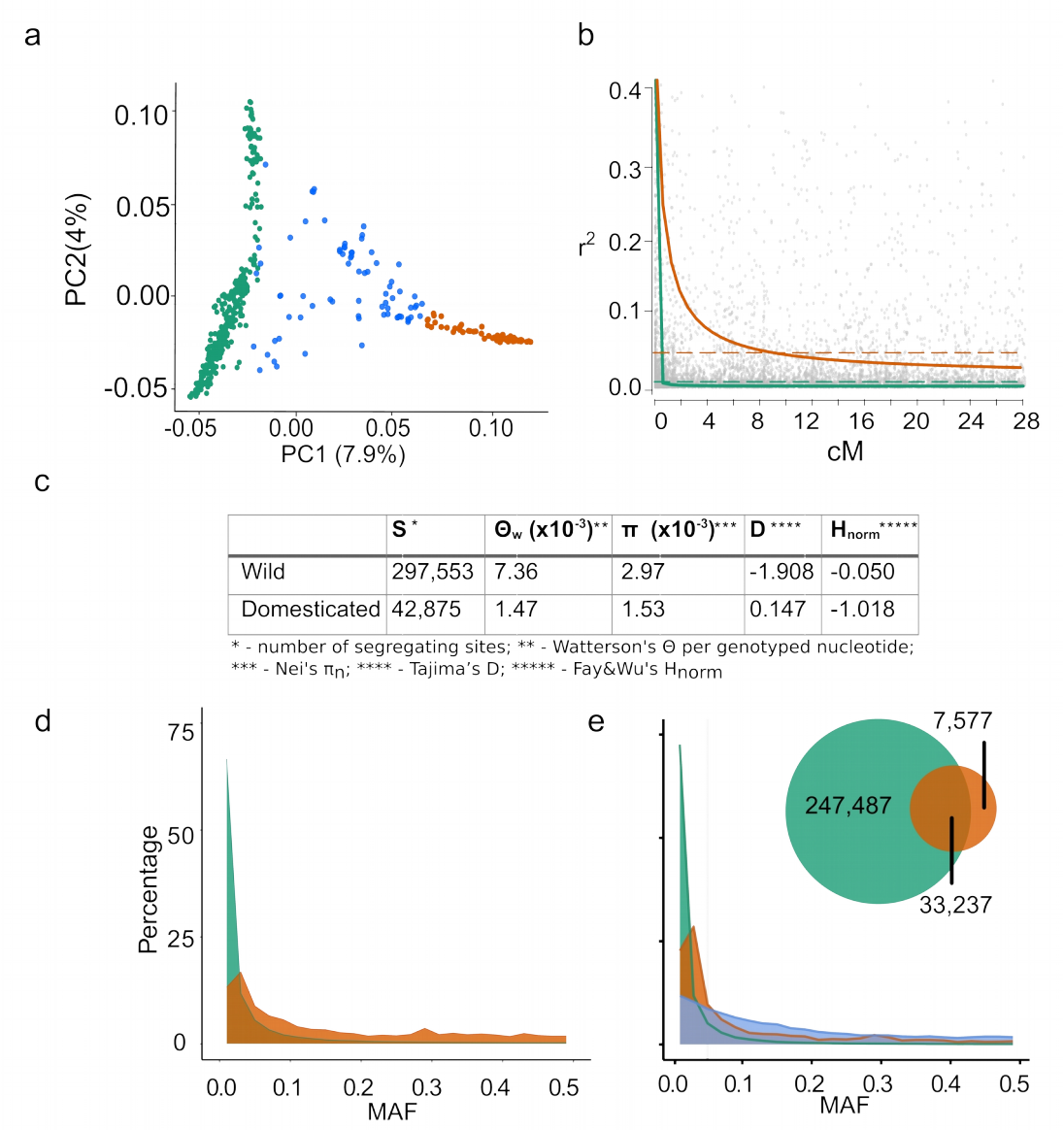
69. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* 1358–1370 (1984).

70. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).

71. Zeng, K., Fu, Y.-X., Shi, S. & Wu, C.-I. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174,** 1431–1439 (2006).

72. 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).

73. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).



	S *	Θ_w (x10^{-3})**	π (x10^{-3})***	D ****	H_{norm}*****
Wild	297,553	7.36	2.97	-1.908	-0.050
Domesticated	42,875	1.47	1.53	0.147	-1.018

* - number of segregating sites; ** - Watterson's Θ per genotyped nucleotide; *** - Nei's π_n; **** - Tajima's D; ***** - Fay&Wu's H_{norm}

Figure 1 Genome-wide analysis of nucleotide diversity.

(**a**) Principal component (PC) analysis of 433 barley genotypes. The first two PC discern subgroups of wild (green), domesticated (orange) barley and admixed (blue) genotypes. A percentage of the total variation explained by the PC is shown in parentheses.

(**b**) Linkage disequilibrium (LD) decay as a function of genetic distance. The non-linear regression curves for pairwise r^2 values are shown for wild (green) and domesticated (orange) barley. The background levels of LD are shown as horizontal dashed lines.

(**c**) Nucleotide diversity parameters in wild and domesticated barley.

(**d**) Distribution of minor allele frequencies (MAF) in wild (green) and domesticated (orange) barley.

(**e**) Distribution and MAF spectra of shared (blue) SNPs and private SNPs in wild (green) and domesticated (orange) barley.

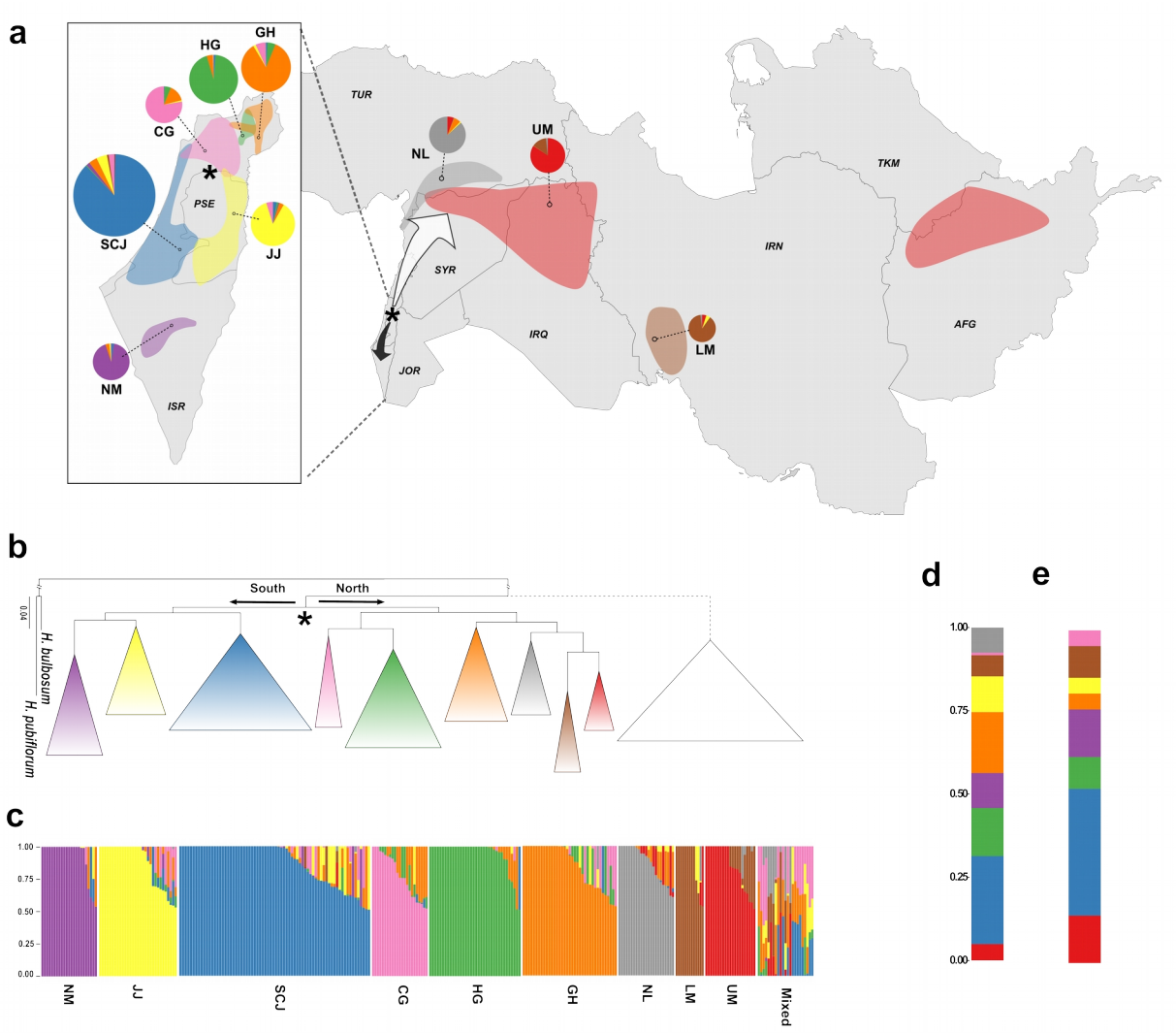


Figure 2 Geographic distribution, structure and phylogeny of nine wild barley populations and their contribution to the domesticated barley genomes.

The colors correspond to the nine wild barley (*H. vulgare* ssp. *spontaneum*) populations. Carmel & Galilee (CG; pink); Golan Heights (GH; orange); Hula Valley & Galilee (HG; green); Judean Desert & Jordan Valley (JJ; yellow); Lower Mesopotamia (LM; brown); Negev Mountains (NM; magenta); North Levant (NL; grey); Sharon, Coastal Plain & Judean Lowlands (SCJ; blue); Upper Mesopotamia (UM; red).

(**a**) Distribution of the wild barley populations within the Fertile Crescent. The pie charts represent ancestral composition of the populations as determined by fastSTRUCTURE and are connected to the geographic centers of population distributions by dashed lines. Size of the pie charts reflects the number of genotypes in the populations. An approximate location of the ancestral wild barley population is shown by an asterisk, and the northward and southward migration routes are indicated by the white and black arrows, respectively. The country codes (ISO 3166) are shown in italics.

(**b**) The Maximum Likelihood (ML) phylogeny of 359 barley accessions. Wild barley clusters collapsed based on the population assignment. Cultivated barley (*H. vulgare* ssp. *vulgare*) is shown as an unfilled cluster. The dashed line indicates uncertainty of the phylogenetic placement of the cultivated barley due to its complex hybrid origin. The ancestral population split is indicated by an asterisk. H. *bulbosum* and *H. pubiflorum* were used as distant outgroup species and the length of the outgroup branch was artificially shortened. (**c**) Ancestral proportions of wild barley as determined by fastSTRUCTURE. (**d, e**) Proportional contribution of nine wild barley populations to composition of the domesticated barley genomes inferred using the complete dataset (**d**) and only the genes carrying footprints of selection under domestication (**e**).

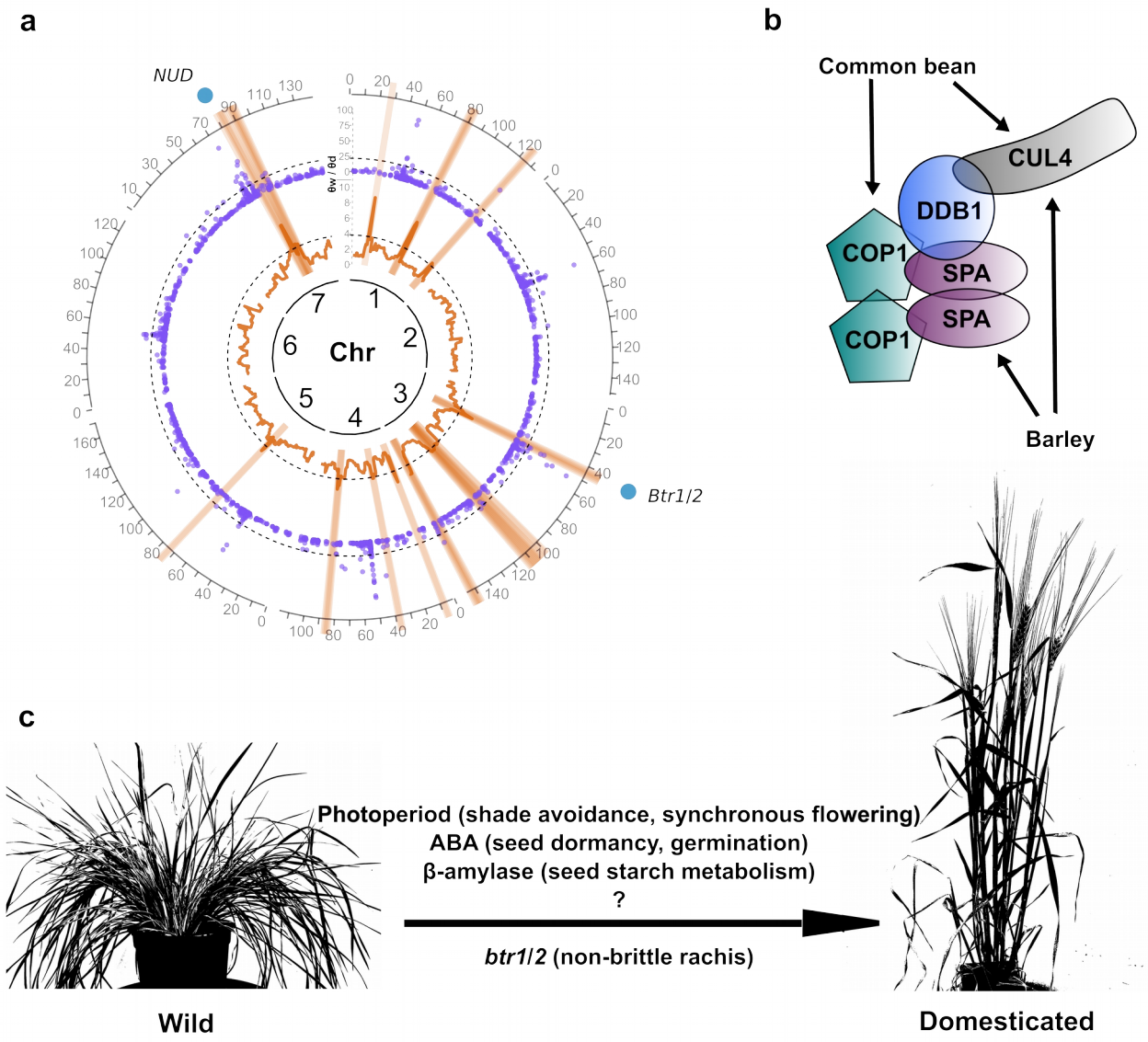


Figure 3 Genomic signatures of selection under domestication, candidate genes and pathways. (**a**) Diversity reduction (θ_w / θ_d) values for sliding windows and 2 369 individual targets. The innermost circle represents barley chromosomes followed by the sliding-window (orange) and individual target (violet) scans. The candidate selection sweep regions are shown by orange blocks. The statistical thresholds ($p < 0.01$) are shown by dashed lines. *Btr1/2* – brittleness loci implicated in domestication.[5]

(**b**) Members of the CUL4-COP1-SPA protein complex affected by selection under domestication in barley and common bean as an example of parallelism in domestication.[14]

(**c**) Candidate domestication pathways and traits. ABA, abscisic acid.