Molecular Biology and Evolution
"In review"

Article : Discoveries

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

Artem Pankin[1,2,3,*] 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 Cologne, 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 the 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 were homologs of the genes, in other plant species, implicated in the regulation of light signaling, the circadian clock, hormone, and carbohydrate metabolism. Phylogeographic analyses revealed a mosaic ancestry of the domestication-related loci, which originated from wild barley populations from both the eastern and western parts of the Fertile Crescent. This indicates that assembly of the mosaic of the cultivated barley genomes was part of the domestication process, which supports a protracted domestication model.

Introduction

Domesticated barley (*Hordeum vulgare* ssp. *vulgare*) is one of the Neolithic founder crops, which facilitated the establishment of the early agricultural societies (Lev-Yadun et al. 2000). Due to its striking environmental plasticity, barley is of utmost importance as a staple crop in a wide range of agricultural environments (Dawson et al. 2015). The first traces of barley cultivation were found at archaeological sites in the Fertile Crescent, which dated back to ~10,000 B.C. (Zohary et al. 2012). 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 (Harlan and Zohary 1966). 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 (Gottlieb 1984). 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 (Purugganan and Fuller 2011; Abbo et al. 2014; Pourkheirandish et al. 2015). 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 (Hammer 1984). 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 (Ross-Ibarra et al. 2007; Shi and Lai 2015). 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 (Huang et al. 2012; Hufford et al. 2012; Lin et al. 2014; Schmutz et al. 2014; Zhou et al. 2015).

The circumstances of barley domestication are debatable and its genome-wide effects on the domesticated barley genomes remain poorly understood. The early models, based on diversity analyses 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 (Negassa 1985; Molina-Cano et al. 1999; Badr et al. 2000; Morrell and Clegg 2007; Dai et al. 2012;

Pourkheirandish et al. 2015). Archaeological and recent molecular evidence suggested that barley domestication was a protracted process, involving the polyphyletic origin of the non-brittle spikes and the heterogeneous (mosaic) ancestry of cultivated barley genomes (Fuller et al. 2012; Allaby 2015; Poets et al. 2015; Pourkheirandish et al. 2015). To further unravel the increasingly complex model of barley domestication, a detailed understanding of the genes and demographic 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. Using population genomics, we disentangled the effects of demography and domestication-related selection on barley genomic diversity following a common two-step approach, which assumes that the genome-wide patterns of variation capture the demographic effects, whereas the extreme outliers represent the instances of selection. The genome-wide diversity patterns indicated a recent domestication bottleneck, which was stronger than previous estimates. The selection scans, besides the spike brittleness locus, identified multiple loci and novel candidate genes affected by the selection during domestication. Finally, we revealed a heterogeneous ancestry of the candidate domestication loci, which suggested the origin of domestication genes in different parts of the Fertile Crescent. This updates the current model of the mosaic ancestry of cultivated barley genomes (Poets et al. 2015) by linking the formation of the genome mosaic to the assembly of the domestication syndrome during the apparently protracted process of domestication.

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 (**Supplementary Table 1**). 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 comprised the whole variety of domesticated barley lifeforms, namely two- and six-row genotypes with winter and spring growth habits. Additionally, domesticated barley is classified into the *btr1*

(*btr1Btr2*) and *btr2* (*Btr1btr2*) types based on the allelic status of the spike brittleness genes *Btr1* and *2;* independent mutations in either of these genes convert the wild-type brittle spikes into the non-brittle spikes of the domesticated forms (Pourkheirandish et al. 2015). To further verify representativeness of the selected genotypes, we screened for the *Btr* mutations using allele-specific markers. In our genotype set, the *btr1* and *btr2* types were represented by 71% and 29% of the domesticated accessions, respectively (**Supplementary Table 1**).

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 2**). Cumulatively, the captured regions comprised approximately 13.8 Mbp (**Supplementary Table 3**) 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 3**). 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 on par with the genome-wide Arabidopsis estimates (2.4) (**Supplementary note**) (Ossowski et al. 2010) but higher than the previous barley estimates (1.15 - 1.70) (Duran et al. 2009; Kono et al. 2016). 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 expected distributions - a large proportion of rare polymorphisms and a 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**). fastSTRUCTURE analysis revealed patterns of admixture in 36% and 12% of the domesticated and wild genotypes, respectively, which corresponded to the intermediate PCA genotypes (**Fig. 1a**, **Supplementary Fig. 4a**). Both fastSTRUCTURE and INSTRUCT models produced matching admixture patterns ($r^2 > 0.99$) (**Supplementary Fig. 4b**). In domesticates, the landraces constituted 95% of the admixed individuals and they did not originate from any specific locality (**Supplementary table 1**). 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 admixed 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 (Zeven 1998). 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 fastSTRUCTURE and PCA (**Supplementary Fig. 5**) and, therefore, were treated as a single group of domesticated barley in the diversity analyses and the selection scans.

The extent of linkage disequilibrium (LD) characterizes a recombination landscape and a 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 (Morrell et al. 2005; Caldwell et al. 2006). 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**).

Genetic diversity in wild and domesticated barley

Domestication results in loss of genetic diversity via the so-called domestication bottleneck, which has been observed in all studied crops (Doebley et al. 2006). However, the impact of the domestication bottleneck on the genetic diversity greatly varied among crop species. Here, we compared genetic diversity in the wild and domesticated subgroups using various population genetic parameters to estimate the intensity of the domestication bottleneck in barley. Wild barley comprised ~7x more segregating sites than domesticated genotypes (**Fig. 1c**) and 88% of the sites resided in the non-coding regions. An unbiased estimator of the population mutation rate Watterson's θ_w (not to be confused with the neutral mutation rate μ), which provides correction for uneven sample size, was 5x higher in wild barley ($\theta_w = 7.36 \times 10^{-3}$) than in the domesticates ($\theta_w = 1.47 \times 10^{-3}$). Nei's nucleotide diversity π_n or θ_π, an estimator of an average number of pairwise differences between two sequences randomly drawn from a population per nucleotide, suggested that the domesticates ($\pi_n=1.53 \times 10^{-3}$) retained 52% of the wild barley nucleotide diversity ($\pi_n=2.97 \times 10^{-3}$). This diversity ratio (π_{dom} / π_{wild}), which indicates the intensity of the domestication bottleneck, was similar to the genome-wide estimates reported in tomato and soybean but higher than those in maize (83%), rice (80%) and common bean (83%) (Huang et al. 2012; Hufford et al. 2012; Lin et al. 2014; Schmutz et al. 2014; Zhou et al. 2015). In barley, previous diversity ratio estimates varied dramatically in the range from 36% to 117% based on averaging diversity estimates in a few isolated genes or on a SNP genotyping assay surveying variation at ~1000 SNPs (Buckler et al. 2001; Caldwell et al. 2006; Kilian et al. 2006; Saisho and Purugganan 2007; Russell et al. 2011; Fu 2012; Morrell et al. 2014). In a recent exome re-sequencing study, which compared the genome-wide diversity of the landrace and wild barley genotypes, the diversity ratio was approximately 73%, which is higher than our estimate (Russell et al. 2016). Therefore, we next investigated whether the diversity ratio calculations are robust against variations in the genotype sampling strategies and SNP calling procedures that could explain the observed discrepancy between the studies. Surprisingly, varying sample sizes in both domesticated and wild subgroups did not noticeably affect the diversity ratio values even in the small subsets comprising 20 – 30 genotypes (**Supplementary Fig. 7a,b**). Similarly, the diversity ratio remained the same after removing the singleton SNPs, which are the most common artifacts of SNP calling or sequencing procedures (**Table 1**). However, including the cultivated samples admixed with the wild genotypes into the calculations increased the diversity ratio to 65%. The estimates of

nucleotide diversity separately in the landraces and advance cultivars suggested another weaker bottleneck associated with crop improvement ($\pi_{cultivar} / \pi_{landrace} = 0.86$).

The diversity estimates suggested that the distributions of allele frequencies greatly differed between wild and domesticated barley. Indeed, the rare alleles were enriched in wild barley (Tajima's D=-1.91 / -1.40 without singletons), whereas the folded site frequency spectra (SFS) in the domesticates were skewed toward the common alleles (**Table 1**, **Fig. 1c**). The difference between the D values in the wild and domesticated subgroups was consistent along the individual chromosomes, but the variation of D was much greater in the domesticates (**Supplementary Fig. 8**). Coalescent simulations performed with the assumption of no selective pressure and the characteristics of an idealized Wright-Fisher population estimated the range of neutral variation of D between -1.74 and 2.46 for the wild population (p-value < 0.01). Thus, the basal levels of D in wild barley, which crossed the simulated thresholds, rejected the assumptions of the neutral model. This questions the utility of the threshold values of D that are based on the simulations assuming the neutral model to infer selection in barley.

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. 1d**). 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 genome-wide fixation index (Fst=0.29), which characterizes divergence between wild and domesticated barley, 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) (Russell et al. 2011; Huang et al. 2012; Hufford et al. 2012; Zhou et al. 2015). 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 during barley domestication. The chromosomes 4 and 7 were significantly more differentiated than the other chromosomes ($p < 0.01$) (**Supplementary Fig. 9**).

Phylogeography of barley domestication

We identified nine population of wild barley using both fastSTRUCTURE and phylogenetic analyses (**Fig. 2abc**; **Supplementary Fig. 10**). 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 genotypes of mixed ancestry occurring mostly in the border overlapping areas (**Supplementary Figs. 11, 12**). Only 23 wild genotypes had a highly admixed ancestry and could not be attributed to any of the nine populations (**Fig. 2c**). Rooting the phylogeny to the outgroup species *H. bulbosum* and *H. pubiflorum* enabled tracing the population differentiation in time. The most ancestral wild 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**). The 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 northern Israel, indicating the isolation-by-distance model.

On the genome-wide rooted phylogenetic tree, the cluster of domesticated barley appeared as a monophyletic sister group branching off at the root before the divergence of the wild barley populations (**Fig. 2b**). This apparently inaccurate position of the domesticated barley cluster on the phylogram presumably reflected the inability of the phylogenetic models to reconstruct population histories that cannot be described by a simple bifurcating tree, as in the case of the gene flow (Felsenstein 1982; Allaby et al. 2008; but see Pickrell and Pritchard 2012), and, therefore, suggested a highly reticulate ancestry of the modern domesticated genotypes. To discover the presence of the historical gene flow between wild and domesticated barley, which could not be identified by the STRUCTURE model, we reconstructed the local ancestry patterns in the domesticated barley genomes. We assumed that if an allele from a domesticated accession was evolutionary closest to a wild allele specific to a single wild population, as determined by maximum-likelihood distance, this population was ancestral for this domesticated allele. The cases where the ancestral wild allele was found in several wild populations were excluded from the analysis to minimize the negative effects of the incomplete lineage sorting on the ancestry estimates. Following this approach, the wild ancestor populations and the corresponding geographic ancestral locations were assigned for 1,232 reference contigs separately for each of the domesticated genotypes (**Fig. 3abc**). This analysis revealed a heterogeneous ancestry of the domesticated barley genomes, i.e. wild barley

populations from different parts of the Fertile Crescent contributed to the cumulative domesticated genome, and the proportion of their individual contributions did not noticeably differ between the domesticated genotypes (**Fig. 3abc, Supplemental Fig. 13**).

Footprints of domestication-related selection

To discover candidate regions and genes that likely experienced selection under domestication, we used a combination of tests - mean r^2 (LD), Fay & Wu's H_{norm} and diversity ratio (π_w/π_d). Using a combination of the tests that explore different aspects of variation helps catalog signatures of different selection scenarios and thus obtain a more complete list of candidate domestication loci (Innan and Kim 2004).

Both selective sweeps and background selection may result in elevated local LD manifested by positively correlated recombination and nucleotide variation rates (Charlesworth et al. 1993). 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 Fig. 14**; **Supplementary table 4**). In wild barley, two of the outliers were co-located with the major flowering loci *PpdH1* and *VRN-H1*, which modulate 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 (Cockram et al. 2011). The *VRN-H1* gene is tightly linked to several flowering-related genes, including *HvPHYC*, a homolog of Arabidopsis *PHYTOCHROME C* gene (Nishida et al. 2013; Pankin et al. 2014). 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.

To reveal signatures of positive selection under domestication, we computed H_{norm} values using a panel of 64,977 SNPs with assigned ancestral status for individual reference contigs and sliding windows (**Table 1**). The amplitude of interspecies variation of H_{norm} was

low compared to D, indicating lesser influence of demography on the H_{norm} estimator. Simulation analysis has demonstrated that the H_{norm} statistics is much less sensitive to the bottleneck than the D test and the compromising effect of the bottleneck diminishes rapidly with time (Zeng et al. 2006). The π_w/π_d scan, while not being a formal test of selection, builds on the premise that the severe depletion of nucleotide diversity in the domesticated population at certain genomic regions detected as statistical outliers is likely a signature of a domestication selective sweep.

In the domesticates, the H_{norm} and π_w/π_d scans identified 13 regions (10-31 cM) and 178 gene-bearing contigs, including 41 target genes, on all barley chromosomes (**Fig. 4ab**, **Supplementary tables 4**, **5**). Only 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 (Taketa et al. 2004; Pourkheirandish et al. 2015). However, the *NUD* gene itself did not carry a selection signature (π_w/π_d=0.7) and thus was not the target of selection in this region as suggested in the previous study (Russell et al. 2016). Indeed, both hulless and hulled genotypes are ubiquitously present in the domesticated barley genepool and apparently represent an improvement but not domestication trait (Saisho and Purugganan 2007).

Among the candidate domestication loci were homologs of genes of light signaling, photoperiod, circadian clock, abscisic acid (ABA) and carbohydrate metabolism pathways (**Fig. 4d**). 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 (Hufford et al. 2012; Schmutz et al. 2014). 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 (Schmutz et al. 2014). Intriguingly, HvCUL4 (seq442; AK371672) and an ortholog of the Arabidopsis SUPRESSOR OF PHYA 2 (SPA2, seq108; MLOC_52815), 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. 4c**) (Zhu et al. 2008). 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) (Hufford et al. 2012). A barley ortholog of

OsMADS56 (seq411; AK369282) 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; FJ951828; 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 (Papaefthimiou et al. 2010).

Phylogeographic analyses demonstrated that, similarly to the genome-wide data, the candidate domestication loci had a mosaic ancestry, descending from nine wild barley populations (**Figs. 3c, 4a**). This illustrates that the mosaic ancestry patterns in cultivated barley genomes were formed in the process of domestication.

Discussion

Understanding domestication history involves comparing the genetic diversity in both domesticated and wild progenitor species to predict demographic events associated with domestication; screening for selected candidate domestication loci to identify domestication alleles and define elusive domestication syndrome traits; and, finally, tracing the ancestry of the cultivated genomes and domestication genes back to specific wild populations to predict where and how many times a crop has been domesticated.

Here, we estimated that domesticated barley retained only half of the genetic diversity found in its wild progenitor species as measured by Nei's π, which was less than the previous genome-wide estimates. Variation of the diversity ratios that are based on a few Sanger-sequenced loci is expected given the large variance of π along the barley chromosomes, whereas the variation of the genome-wide estimates likely reflects differences in the genotype sampling strategies. Interestingly, we found that the diversity ratio calculations were robust to varying numbers of samples in the compared groups – the subsets of tens and hundreds of genotypes returned comparable estimates. However, the recent gene flow between the compared subspecies strongly affected the diversity ratios. In contrast to previous studies (Russell et al. 2011; Russell et al. 2016), we excluded the recently admixed genotypes identified by STRUCTURE from the diversity analyses, which possibly explains the differences in the diversity estimates. This raises the question of whether excluding admixed genotypes from the diversity analyses is justified. It has been suggested that the outcrossing

between the modern wild and domesticated barley genotypes may occur either naturally in sympatric stands within the Fertile Crescent (Russell et al. 2011) or during the *ex situ* reproduction in the germplasm banks (Jakob et al. 2014). Our analyses hinted that both scenarios occurred in our selection of genotypes (see **Supplementary note**); however, distinguishing between these alternatives remains a challenge for future studies. Whereas the *in situ* outcrossing may represent an evolutionary mechanism alleviating detrimental effects of the bottlenecks on genetic diversity in cultivated populations (Verhoeven et al. 2011), including the samples that unintentionally hybridized during the genebank reproduction in the diversity screens and selection scans apparently may lead to erroneous interpretations of the revealed patterns.

We found that, in wild barley, the rare alleles were overrepresented above the thresholds expected by the neutral model as suggested by the strongly negative genome-wide D values as a function of the site frequency spectrum (SFS). In the studies based on the sequencing of several isolated genes the D values for most of the loci were also negative (Lin et al. 2001; Morrell et al. 2005; Bedada et al. 2014). The excess of rare polymorphisms could be a signature of either directional selection or various demographic histories such as an exponential population growth and multiple mergers (Wright and Gaut 2005; Eldon et al. 2015). These alternatives are difficult to distinguish due to the similarity of their effects on SFS. In Arabidopsis, the genome-wide enrichment of rare alleles was mainly attributed to the demography of the species and only partly to selection, but statistical testing did not support any specific demographic model (Nordborg et al. 2005; Schmid et al. 2005). A demographic history of wild barley remains largely unexplored. It has been suggested that wild barley may have experienced large-scale postglacial range expansion using spatial analysis of genetic diversity (Jakob et al. 2014; Russell et al. 2014). Such demographic scenario involving exponential population growth (perhaps conjointly with bottleneck) could explain the observed genome-wide excess of rare alleles.; however, its fit to the empirical data awaits statistical testing.

In domesticated barley, rare polymorphisms were severely depleted compared with the wild progenitor SFS. This together with an increased LD, reduced diversity, a small number of private alleles, and large variance of the diversity parameters across the domesticated barley chromosomes are the signatures of a strong recent bottleneck in the demographic history of cultivated genotypes apparently associated with domestication (Nei et al. 1975; McVean 2002).

Accumulating molecular and archaeological evidence has been shifting our views on

the demography of barley domestication from an early model of a fast monophyletic event toward a more complex protracted model of domestication (reviewed in Pankin and von Korff 2017). The protracted model postulates a slower non-centric process of domestication (Purugganan and Fuller 2011). Accordingly, domestication traits may have reached fixation in a cultigen on a longer timescale and the archaeological data corroborate this assumption (Tanno and Willcox 2012). Distinction between the two models and their conceptualization have been a subject of recent debates (Fuller et al. 2012; Heun et al. 2012).

Previously, the patterns of barley population structure suggested independent domestication lineages in the east and west of the Fertile Crescent, which descended from the respective eastern and western wild barley subpopulations roughly bisected by the Zagros range in modern Iran (Morrell and Clegg 2007; Morrell et al. 2014). The notion of the partitioning of wild barley genetic diversity into eastern and western clusters gained support primarily from the earlier studies of population structure, which suggested the ancestral population split in the area of the Zagros Mountains (Morrell and Clegg 2007; Fang et al. 2014). However, our analysis of the wild barley population structure at a finer scale in terms of the number of genotypes and markers than the previous studies identified that the oldest *H. spontaneum* populations were established in the north of modern Israel based on the hierarchy of population splits in both the rooted phylogenetic analysis and the series of K in the fastSTRUCTURE clustering. Although this finding did not directly refute the independent domestications in the east and west of the Fertile Crescent, it challenged the position of the Zagros range as a primary axis separating the wild barley diversity. Apparently, the eastern Fertile Crescent wild barley populations were among the youngest and arrived at the Zagros range after differentiation of the older populations along the northward migration route.

The patterns of the *Btr1/2* locus diversity implicated at least two independent domestication events; however, in both loci, the domestication mutations seemed to originate in the western horn of the Fertile Crescent (Pourkheirandish et al. 2015). According to a recent more complex model of barley domestication based on the genome-wide data, five wild barley populations from both the eastern and western Fertile Crescent contributed to the genomes of all modern barley landraces (Allaby 2015; Poets et al. 2015). However, the extent to which the mosaic composition of the domesticated barley genomes was related to the neutral processes, domestication, or later environmental adaptations remained unclear. We found that the domesticated barley genomes comprised ancestral fragments that originated from nine wild barley populations residing in different parts of the Fertile Crescent thus reinforcing the mosaic

model. The mosaic ancestry of the candidate domestication loci let us further expand on this model by suggesting the direct relationship between the heterogeneous ancestry and the process of domestication itself.

Interestingly, the ancestry diagrams of the individual domesticated genotypes were remarkably similar even across the *btr1* and *btr2* types, which apparently represent the independent events that led to the origin of non-brittle spikes. Poets et al. (2006) revealed the ancestral patterns in the landrace barley that were similar between the subpopulations from both the east and west of the Fertile Crescent. It hints at the possibility that a single highly admixed lineage – wild or (proto)domesticated - has been at the root of all modern domesticated barley genotypes. In this case, the protracted process of recurrent introgressions into the ancestral lineage could gradually lead to the assembly of the domestication syndrome (DS) of modern cultivated barley. This scenario assumes a monophyletic lineage of the modern barley domesticates but polyphyletic origin of the specific domestication alleles (see Pankin and Korff, 2016). Using simulations, Allaby (2005) suggested that, in a scenario of truly independent domestication lineages, a gene flow between the domesticates may mask the patterns of a polyphyletic origin so that all the domesticated genotypes will form a monophyletic cluster on a phylogenetic tree. However, the question whether the gene flow may homogenize the independent lineages to the extent that they become nearly identical at the level of the ancestry diagrams requires further investigation. The gene flow between the wild and (proto)domesticated populations as a part of the domestication process has been documented in several crops (Huang et al. 2012; Civáň et al. 2013; Hufford et al. 2013; Mascher et al. 2016). A highly reticulate history of the crop genomes exemplified by barley in this study complicates distinction between the mono- and polyphyletic models of crop origin (Olsen and Gross 2008; Huang et al. 2012; Civáň et al. 2015; Huang and Han 2015). Among the problems that we face in this endeavor are that the alternative demographies may have left (nearly) identical signatures in the domesticated genomes or that the truly independent domesticated lineages existing in the past may not have left their descendants in the modern domesticated barley genepool. Finally, defining from what exact point a cultivated population becomes domesticated is crucial but represents a conceptual difficulty in the domestication studies (Larson et al. 2014).We therefore argue that discovering and tracing the origin of the individual domestication traits and selected mutations provides a complementary and more targeted approach in untangling reticulate domestication histories (Pourkheirandish et al. 2015; Win et al. 2016).

Understanding of the traits and genes that constitute the barley DS is extremely limited. Spike brittleness is the only studied example of a crucial DS trait (sensu (Abbo et al. 2014). Besides, seed dormancy, seed size, synchrony of flowering, and number and angle of side shoots, also known in cereals as tillers, have been suggested to constitute the DS (Doebley et al. 2006; Meyer and Purugganan 2013). Here, we identified genomic regions that likely experienced selection under domestication and proposed novel candidate domestication genes.

The top domestication candidates were the homologs of genes, in other plant species, 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 (Zhu et al. 2015). 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 their wild ancestor species, which grow in isolated patches. Following this hypothesis may be the key to understanding the involvement of the modulators of light signaling, circadian clock and shade avoidance pathways in domestication (Müller et al. 2016; Shor and Green 2016).

It is noteworthy that the domestication loci detected in this study may be only a representative sample of the truly selected loci. Some loci might have experienced selection regimes leaving signatures that frequently escape detection or that are confounded with the effects of demography (Teshima et al. 2006)**.** Other domestication loci could have been missed because of the gaps in certain regions of the chromosomes in the Popseq genetic map (Mascher, Muehlbauer, et al. 2013). A resolved physical map of the barley genome will be required in future studies to obtain sharper selection signals and to unlock the genomic regions that remained unexplored using the genetic map.

Given the complex ancestry of the modern cultivated barley, constructing a realistic model of its domestication becomes a major challenge. To this end, it is critical to partition the nucleotide variation in barley populations into the neutral demographic and non-neutral selected components. In this study, we comprehensively surveyed nucleotide diversity in both wild and domesticated barley populations to predict their demographic history and to identify

genomic regions and novel candidate genes that carried signatures of a selective sweep associated with domestication. We updated the mosaic model of barley domestication by linking the heterogeneous origin of cultivated barley with the process of domestication itself. Our findings highlighted the need to statistically resolve the alternative demographies of both wild and cultivated barley subspecies in future studies and provided a solid basis for identification, experimental validation, and tracing the ancestry of the novel adaptive alleles and traits that suited barley to the agricultural environments. Taken together, this will facilitate development of a more coherent narrative of barley domestication history.

Materials and methods

Plant material and Btr genotyping assay

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 1**). 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 area of the Fertile Crescent. 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.

The DNA samples of domesticated barley were genotyped using PCR markers distinguishing loss-of-function alleles of the brittleness genes *Btr1* and *2*. The markers were amplified using allele-specific primer pairs Btr1f 5'-CCGCAATGGAAGGCGATG-3' / Btr1r 5'-CTATGAAACCGGAGAGGC-3' (~200 bp fragment, presence - *Btr1* / absence - *btr1*) and

Btr2f 5'-AATACGACTCACTATAGGGTTCGTCGAGCTCGCTATC-3' / Btr2r 5'-GTGGAGTTGCCACCTGTG-3' (~ 160 bp fragment, 15 bp deletion in the *btr2* allele). PCR reactions (1 x PCR buffer, 0.1 M primers, 1 U Taq polymerase, 100 ng DNA) were incubated in the PTC DNA Engine thermocycler (Bio-Rad, Hercules, CA, USA) under the following conditions: 95°C for 3 min; 30 cycles of 95°C for 20 s, 60°C for 30 s, 72°C for 1 min; 72°C for 5 min.

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 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 (Higgins et al. 2010). 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 6**). 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) (IBGSC, 2012) 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 (Wheelan et al. 2001). 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. 15**.

A set of 1000 additional HarvEST genes was randomly selected such that they had no homology to target genes as determined by BLASTN and were evenly spread over all barley linkage groups according to the GenomeZipper map (Mayer et al. 2011). 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 2**). 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 a 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 (Guo et al. 2012). To identify such regions, the Illumina reads from 10 randomly selected barley genotypes were mapped to the complete Morex genome reference set (IBGSC, 2012). 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 (International Barley Genome Sequencing Consortium (IBGSC). 2012; Mascher, Muehlbauer, et al. 2013)

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. 16**). 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) (Fu et al. 2012), 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 (Song et al. 2014). 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) (Li and Durbin 2009; McKenna et al. 2010). 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 (McKenna et al. 2010). 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 (Danecek et al. 2011; Quinlan 2014). Functional effects of the SNPs were predicted using SnpEff 3.6b software using the custom CDS coordinates as a reference genome (Cingolani et al. 2012). 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

Site frequency spectra (SFS) for various genomic regions and bootstrapping of the rare SNPs (1000 random draws) were calculated using two different approaches: based on the SNP allele counts from the results of the SNP calling and based on the raw 'bam' files using ANGSD 0.913 (Korneliussen et al. 2014). 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 (Chang et al. 2015). For estimations of population parameters, only a subset of SNPs with minor allele frequency (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 (Patterson et al. 2006).

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 (Hill and Weir 1988). 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 the fastSTRUCTURE software, which implements the Bayesian clustering algorithm STRUCTURE, assuming Hardy-Weinberg equilibrium between alleles, in a fast and resource-efficient manner (Raj et al. 2014). This algorithm efficiently detects recent gene flow events but not the historical admixture. The runs were executed with 20 iterations for a predefined number of populations (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 output matrices from the iteration runs were

summarized using CLUMPAK (Kopelman et al. 2015), reordered and plotted using an in-house R script. Additionally, the population structure in wild barley was determined using INSTRUCT software, which extends the STRUCTURE model to include selfing (Gao et al. 2007). Due to very hight computational intensity of INSTRUCT, we ran the analyses on 10 randomly drawn subsamples of 1000 SNP markers for five independent chains and summarized the runs using descriptive statistics.

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 diversity parameters, such as number of segregating sites (S), Watterson's estimator (θ_w) per genotyped site (Watterson 1975), Nei's (sometimes referred as Tajima's) nucleotide diversity (π) per genotyped site (Nei and Li 1979), fixation index (Fst) (Hudson et al. 1992), as well as the frequency-based selection tests, such as Tajima's D (Tajima 1989) and normalized Fay and Wu's H_{norm} (Zeng et al. 2006) 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).
Hudson's Fst values were calculated for the entire SNP dataset, as a ratio of average numerator and denominator values (“ratio of averages”), and for each individual SNP separately to compare the Fst patterns along the chromosomes. The “ratio of averages” and the average of the individual SNP Fst values produce highly discrepant genome-wide estimates – approximately two-fold difference (Bhatia et al. 2013). As recommended by Bhatia et al. (2013), we report the genome-wide Fst estimated as a “ratio of averages”.

To determine the ancestral status of 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 (Mascher, Richmond, et al. 2013).

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 (Hudson 2002). 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 in the 10-cM windows with a sliding step of 1 cM and separately for the individual loci (cut-off > 5 SNPs / locus). The putative 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 . Since the simulated thresholds of neutral variation for the H_{norm} test were based on a model that likely does not accurately reflect a demographic history of barley, we used a very conservative threshold (p-value < 0.001) to identify outliers. The overlapping outlier windows were merged into putatively selected regions.

The genome-wide 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 (Stamatakis 2014). Wild barley *H. bulbosum* and *H. pubiflorum* were used as outgroup species. The trees were visualized and collapsed using Dendroscope 3.5.7 (Huson and Scornavacca 2012).

To estimate ancestry of the domesticated barley loci, we calculated pairwise ML distances between each wild and domesticated genotypes separately for each locus, i.e. individual contig in the mapping reference, using GTRGAMMA model in RAxML 8.2.8 (Stamatakis 2014). For each domesticated genotype, if a locus had the smallest ML distance only with a single wild population, this population was deemed ancestral. If a locus from domesticated barley was equally distant from the corresponding locus in several wild barley populations, such loci were not taken into the ancestry analysis.

Availability of data and materials

The scripts and the accompanying files used for analysis are available in an online repository at https://github.com/artempankin/korffgroup. Raw Illumina sequence reads have been deposited at NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA329198.

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.

Acknowledgements

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. This work was supported by the Max Planck Society and by the Deutsche Forschungsgemeinschaft grants (DFG 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.

References

Abbo S, van-Oss RP, Gopher A, Saranga Y, Ofner I, Peleg Z. 2014. Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. Trends Plant Sci. 19:351–360.

Allaby RG. 2015. Barley domestication: the end of a central dogma? Genome Biol. 16:1.

Allaby RG, Fuller DQ, Brown TA. 2008. The genetic expectations of a protracted model for the origins of domesticated crops. Proc. Natl. Acad. Sci. 105:13982–13986.

Badr A, M K, Sch R, Rabey HE, Effgen S, Ibrahim HH, Pozzi C, Rohde W, Salamini F. 2000. On the Origin and Domestication History of Barley (Hordeum vulgare). Mol. Biol. Evol. 17:499–510.

Bedada G, Westerbergh A, Nevo E, Korol A, Schmid KJ. 2014. DNA sequence variation of wild barley Hordeum spontaneum (L.) across environmental gradients in Israel. Heredity 112:646–655.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting FST: The impact of rare variants. Genome Res. 23:1514–1521.

Buckler ES, Thornsberry JM, Kresovich S. 2001. Molecular diversity, structure and domestication of grasses. Genet. Res. 77:213–218.

Caldwell KS, Russell J, Langridge P, Powell W. 2006. Extreme Population-Dependent Linkage Disequilibrium Detected in an Inbreeding Plant Species, Hordeum vulgare. Genetics 172:557–567.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation

PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:1.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. 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.

Civáň P, Craig H, Cox CJ, Brown TA. 2015. Three geographically separate domestications of Asian rice. Nat. Plants 1:15164.

Civáň P, Ivaničová Z, Brown TA. 2013. Reticulated Origin of Domesticated Emmer Wheat Supports a Dynamic Model for the Emergence of Agriculture in the Fertile Crescent. PLOS ONE 8:e81955.

Cockram J, Hones H, O'Sullivan DM. 2011. Genetic variation at flowering time loci in wild and cultivated barley. Plant Genet. Resour. 9:264–267.

Dai F, Nevo E, Wu D, Comadran J, Zhou M, Qiu L, Chen Z, Beiles A, Chen G, Zhang G. 2012. Tibet is one of the centers of domestication of cultivated barley. Proc. Natl. Acad. Sci. 109:16969–16973.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Dawson IK, Russell J, Powell W, Steffenson B, Thomas WTB, Waugh R. 2015. Barley: a translational model for adaptation to climate change. New Phytol. 206:913–931.

Doebley JF, Gaut BS, Smith BD. 2006. The Molecular Genetics of Crop Domestication. Cell 127:1309–1321.

Duran C, Appleby N, Vardy M, Imelfort M, Edwards D, Batley J. 2009. Single nucleotide polymorphism discovery in barley using autoSNPdb. Plant Biotechnol. J. 7:326–333.

Eldon B, Birkner M, Blath J, Freund F. 2015. Can the Site-Frequency Spectrum Distinguish Exponential Population Growth from Multiple-Merger Coalescents? Genetics 199:841–856.

Fang Z, Gonzales AM, Clegg MT, Smith KP, Muehlbauer GJ, Steffenson BJ, Morrell PL. 2014. Two Genomic Regions Contribute Disproportionately to Geographic Differentiation in Wild Barley. G3 GenesGenomesGenetics 4:1193–1203.

Felsenstein J. 1982. How can we infer geography and history from gene frequencies? J. Theor. Biol. 96:9–20.

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

Fu Y-B. 2012. Population-based resequencing analysis of wild and cultivated barley revealed

weak domestication signal of selection and bottleneck in the Rrs2 scald resistance gene region. Genome 55:93–104.

Fuller DQ, Asouti E, Purugganan MD. 2012. Cultivation as slow evolutionary entanglement: comparative data on rate and sequence of domestication. Veg. Hist. Archaeobotany 21:131–145.

Gao H, Williamson S, Bustamante CD. 2007. A Markov Chain Monte Carlo Approach for Joint Inference of Population Structure and Inbreeding Rates From Multilocus Genotype Data. Genetics 176:1635–1651.

Gottlieb LD. 1984. Genetics and morphological evolution in plants. Am. Nat.:681–709.

Guo Y, Long J, He J, Li C-I, Cai Q, Shu X-O, Zheng W, Li C. 2012. Exome sequencing generates high quality data in non-target regions. BMC Genomics 13:1.

Hammer K. 1984. Das domestikationssyndrom. Kult. 32:11–34.

Harlan JR, Zohary D. 1966. Distribution of Wild Wheats and Barley. Science 153:1074–1080.

Heun M, Abbo S, Lev-Yadun S, Gopher A. 2012. 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.

Higgins JA, Bailey PC, Laurie DA. 2010. Comparative genomics of flowering time pathways using Brachypodium distachyon as a model for the temperate grasses. PLoS One 5:e10065.

Hill WG, Weir BS. 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. 33:54–78.

Huang X, Han B. 2015. Rice domestication occurred through single origin and multiple introgressions. Nat. Plants 2:15207.

Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, et al. 2012. A map of rice genome variation reveals the origin of cultivated rice. Nature 490:497–501.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583–589.

Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. 2013. The Genomic Signature of Crop-Wild Introgression in Maize. PLOS Genet 9:e1003477.

Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, et al. 2012. Comparative population genomics of maize domestication and improvement. Nat. Genet. 44:808–811.

Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61:1061–1067.

Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. U. S. A. 101:10667–10672.

International Barley Genome Sequencing Consortium (IBGSC). 2012. A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716.

Jakob SS, Rödder D, Engler JO, Shaaf S, Özkan H, Blattner FR, Kilian B. 2014. Evolutionary History of Wild Barley (Hordeum vulgare subsp. spontaneum) Analyzed Using Multilocus Sequence Data and Paleodistribution Modeling. Genome Biol. Evol. 6:685–702.

Kilian B, Özkan H, Kohl J, Haeseler A von, Barale F, Deusch O, Brandolini A, Yucel C, Martin W, Salamini F. 2006. Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. Mol. Genet. Genomics 276:230–241.

Kono TJY, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, Smith KP, Tiffin P, Fay JC, Morrell PL. 2016. The Role of Deleterious Substitutions in Crop Genomes. Mol. Biol. Evol. 33:2307–2317.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. 15:1179–1191.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15:356.

Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, Barton L, Vigueira CC, Denham T, Dobney K, et al. 2014. Current perspectives and the future of domestication studies. Proc. Natl. Acad. Sci. 111:6139–6146.

Lev-Yadun S, Gopher A, Abbo S. 2000. The Cradle of Agriculture. Science 288:1602–1603.

Lin J-Z, Brown AHD, Clegg MT. 2001. Heterogeneous geographic patterns of nucleotide sequence diversity between two alcohol dehydrogenase genes in wild barley (Hordeum vulgare subspecies spontaneum). Proc. Natl. Acad. Sci. 98:531–536.

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, et al. 2014. Genomic analyses provide insights into the history of tomato breeding. Nat. Genet. 46:1220–1226.

Mascher M, Muehlbauer GJ, Rokhsar DS, Chapman J, Schmutz J, Barry K, Muñoz-Amatriaín M, Close TJ, Wise RP, Schulman AH. 2013. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J. 76:718–727.

Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, Sampath D, Ayling S, Steuernagel B, Pfeifer M, D'Ascenzo M. 2013. Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. Plant J. 76:494–505.

Mascher M, Schuenemann VJ, Davidovich U, Marom N, Himmelbach A, Hübner S, Korol A, David M, Reiter E, Riehl S, et al. 2016. Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. Nat. Genet. 48:1089–1093.

Mayer KF, Martis M, Hedley PE, Šimková H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H. 2011. Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23:1249–1263.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

McVean GA. 2002. A genealogical interpretation of linkage disequilibrium. Genetics 162:987–991.

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. Nat. Rev. Genet. 14:840–852.

Molina-Cano JL, Moralejo M, Igartua E, Romagosa I. 1999. Further evidence supporting Morocco as a centre of origin of barley. Theor. Appl. Genet. 98:913–918.

Morrell PL, Clegg MT. 2007. Genetic evidence for a second domestication of barley (Hordeum vulgare) east of the Fertile Crescent. Proc. Natl. Acad. Sci. 104:3289–3294.

Morrell PL, Gonzales AM, Meyer KKT, Clegg MT. 2014. Resequencing Data Indicate a Modest Effect of Domestication on Diversity in Barley: A Cultigen With Multiple Origins. J. Hered. 105:253–264.

Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2005. 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.

Müller NA, Wijnen CL, Srinivasan A, Ryngajllo M, Ofner I, Lin T, Ranjan A, West D, Maloof JN, Sinha NR. 2016. Domestication selected for deceleration of the circadian clock in cultivated tomato. Nat. Genet. 48:89–93.

Negassa M. 1985. 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.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. 76:5269–5273.

Nei M, Maruyama T, Chakraborty R. 1975. The bottleneck effect and genetic variability in populations. Evolution:1–10.

Nishida H, Ishihara D, Ishii M, Kaneko T, Kawahigashi H, Akashi Y, Saisho D, Tanaka K, Handa H, Takeda K, et al. 2013. Phytochrome C Is A Key Factor Controlling Long-Day Flowering in Barley. Plant Physiol. 163:804–814.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, et al. 2005. The Pattern of Polymorphism in Arabidopsis

thaliana. PLoS Biol 3:e196.

Olsen KM, Gross BL. 2008. Detecting multiple origins of domesticated crops. Proc. Natl. Acad. Sci. 105:13701–13702.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327:92–94.

Pankin A, Campoli C, Dong X, Kilian B, Sharma R, Himmelbach A, Saini R, Davis SJ, Stein N, Schneeberger K. 2014. 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.

Pankin A, von Korff M. 2017. Co-evolution of methods and thoughts in cereal domestication studies: a tale of barley (Hordeum vulgare). Curr. Opin. Plant Biol. 36:15–21.

Papaefthimiou D, Likotrafiti E, Kapazoglou A, Bladenopoulos K, Tsaftaris A. 2010. 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.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet 2:e190.

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genet 8:e1002967.

Poets AM, Fang Z, Clegg MT, Morrell PL. 2015. Barley landraces are characterized by geographically heterogeneous genomic origins. Genome Biol. 16:1.

Pourkheirandish M, Hensel G, Kilian B, Senthil N, Chen G, Sameri M, Azhaguvel P, Sakuma S, Dhanagond S, Sharma R, et al. 2015. Evolution of the Grain Dispersal System in Barley. Cell 162:527–539.

Purugganan MD, Fuller DQ. 2011. Archaeological data reveal slow rates of evolution during plant domestication. Evolution 65:171–183.

Quinlan AR. 2014. BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinforma.:11.12. 1-11.12. 34.

Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573–589.

Ross-Ibarra J, Morrell PL, Gaut BS. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc. Natl. Acad. Sci. 104:8641–8648.

Russell J, Dawson IK, Flavell AJ, Steffenson B, Weltzien E, Booth A, Ceccarelli S, Grando S, Waugh R. 2011. 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.

Russell J, Mascher M, Dawson IK, Kyriakidis S, Calixto C, Freund F, Bayer M, Milne I, Marshall-Griffiths T, Heinen S, et al. 2016. Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. Nat. Genet. 48:1024–1030.

Russell J, van Zonneveld M, Dawson IK, Booth A, Waugh R, Steffenson B. 2014. 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.

Saisho D, Purugganan MD. 2007. Molecular phylogeography of domesticated barley traces expansion of agriculture in the Old World. Genetics 177:1765–1776.

Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T. 2005. A Multilocus Sequence Survey in Arabidopsis thaliana Reveals a Genome-Wide Departure From a Neutral Model of DNA Sequence Polymorphism. Genetics 169:1601–1615.

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, et al. 2014. A reference genome for common bean and genome-wide analysis of dual domestications. Nat. Genet. 46:707–713.

Shi J, Lai J. 2015. Patterns of genomic changes with crop domestication and breeding. Curr. Opin. Plant Biol. 24:47–53.

Shor E, Green RM. 2016. The Impact of Domestication on the Circadian Clock. Trends Plant Sci. 21:281–283.

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

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

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Taketa S, Kikuchi S, Awayama T, Yamamoto S, Ichii M, Kawasaki S. 2004. 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.

Tanno K, Willcox G. 2012. Distinguishing wild and domestic wheat and barley spikelets from early Holocene sites in the Near East. Veg. Hist. Archaeobotany 21:107–115.

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16:702–712.

Verhoeven KJF, Macel M, Wolfe LM, Biere A. 2011. Population admixture, biological invasions and the balance between local adaptation and inbreeding depression. Proc. R. Soc. Lond. B Biol. Sci. 278:2–8.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276.

Wheelan SJ, Church DM, Ostell JM. 2001. Spidey: a tool for mRNA-to-genomic alignments. Genome Res. 11:1952–1957.

Win KT, Yamagata Y, Doi K, Uyama K, Nagai Y, Toda Y, Kani T, Ashikari M, Yasui H, Yoshimura A. 2016. A single base change explains the independent origin of and selection for the nonshattering gene in African rice domestication. New Phytol.:in press.

Wright SI, Gaut BS. 2005. Molecular Population Genetics and the Search for Adaptive Evolution in Plants. Mol. Biol. Evol. 22:506–519.

Zeng K, Fu Y-X, Shi S, Wu C-I. 2006. Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. Genetics 174:1431–1439.

Zeven AC. 1998. Landraces: a review of definitions and classifications. Euphytica 104:127–139.

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. 33:408–414.

Zhu D, Maier A, Lee J-H, Laubinger S, Saijo Y, Wang H, Qu L-J, Hoecker U, Deng XW. 2008. Biochemical characterization of Arabidopsis complexes containing CONSTITUTIVELY PHOTOMORPHOGENIC1 and SUPPRESSOR OF PHYA proteins in light control of plant development. Plant Cell 20:2307–2323.

Zhu L, Bu Q, Xu X, Paik I, Huang X, Hoecker U, Deng XW, Huq E. 2015. CUL4 forms an E3 ligase with COP1 and SPA to promote light-induced degradation of PIF1. Nat. Commun. 6.

Zohary D, Hopf M, Weiss E. 2012. Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin. OUP Oxford

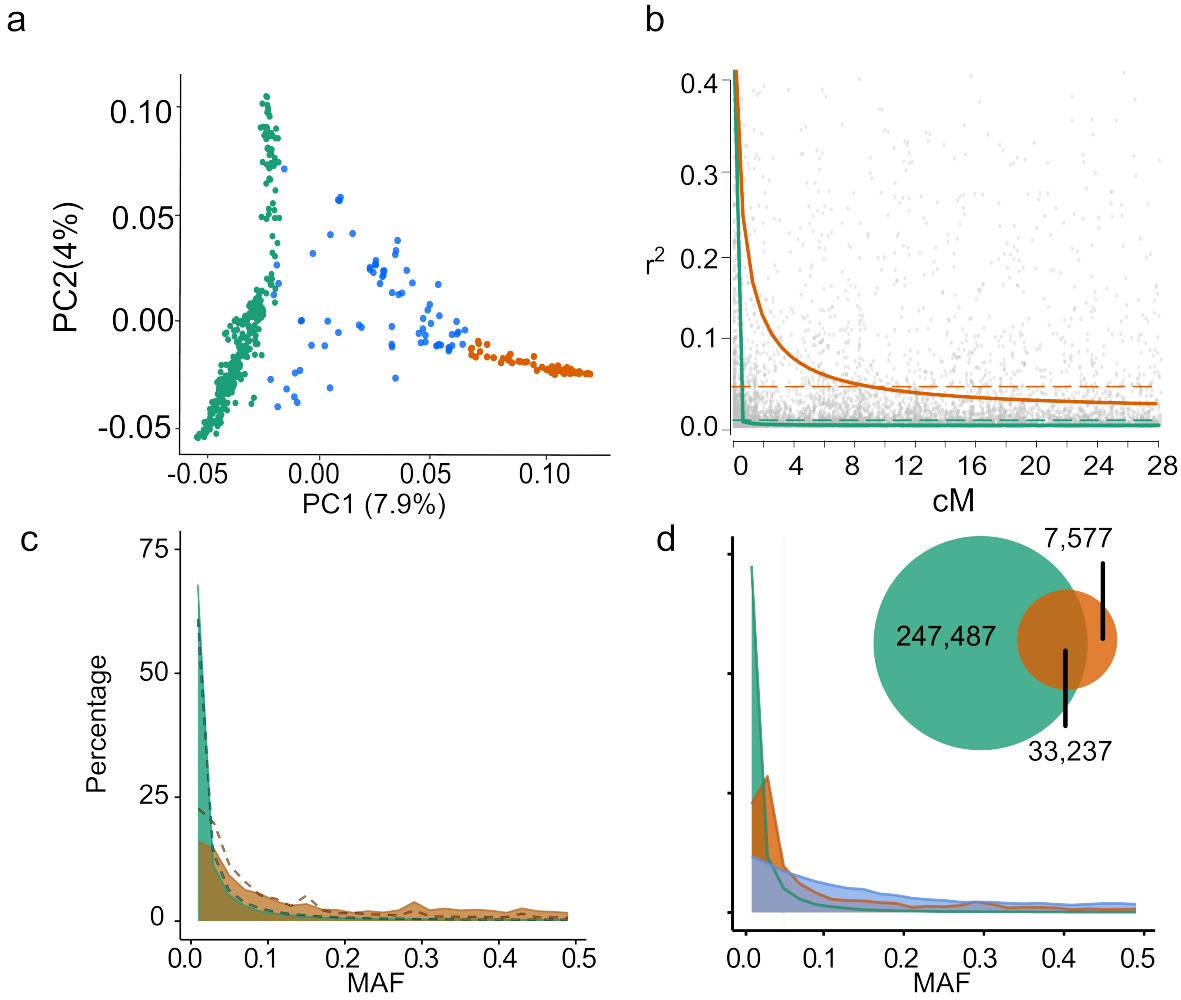

Figure 1 Genome-wide analysis of nucleotide diversity.

(**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**) Folded site frequency spectra (SFS) in wild (green) and domesticated (orange) barley based on the SNP calling (solid line) and ANGSD algorithms (dashed line).

(**d**) Proportion and SFS of shared (blue) and private alleles in wild (green) and domesticated (orange) barley.

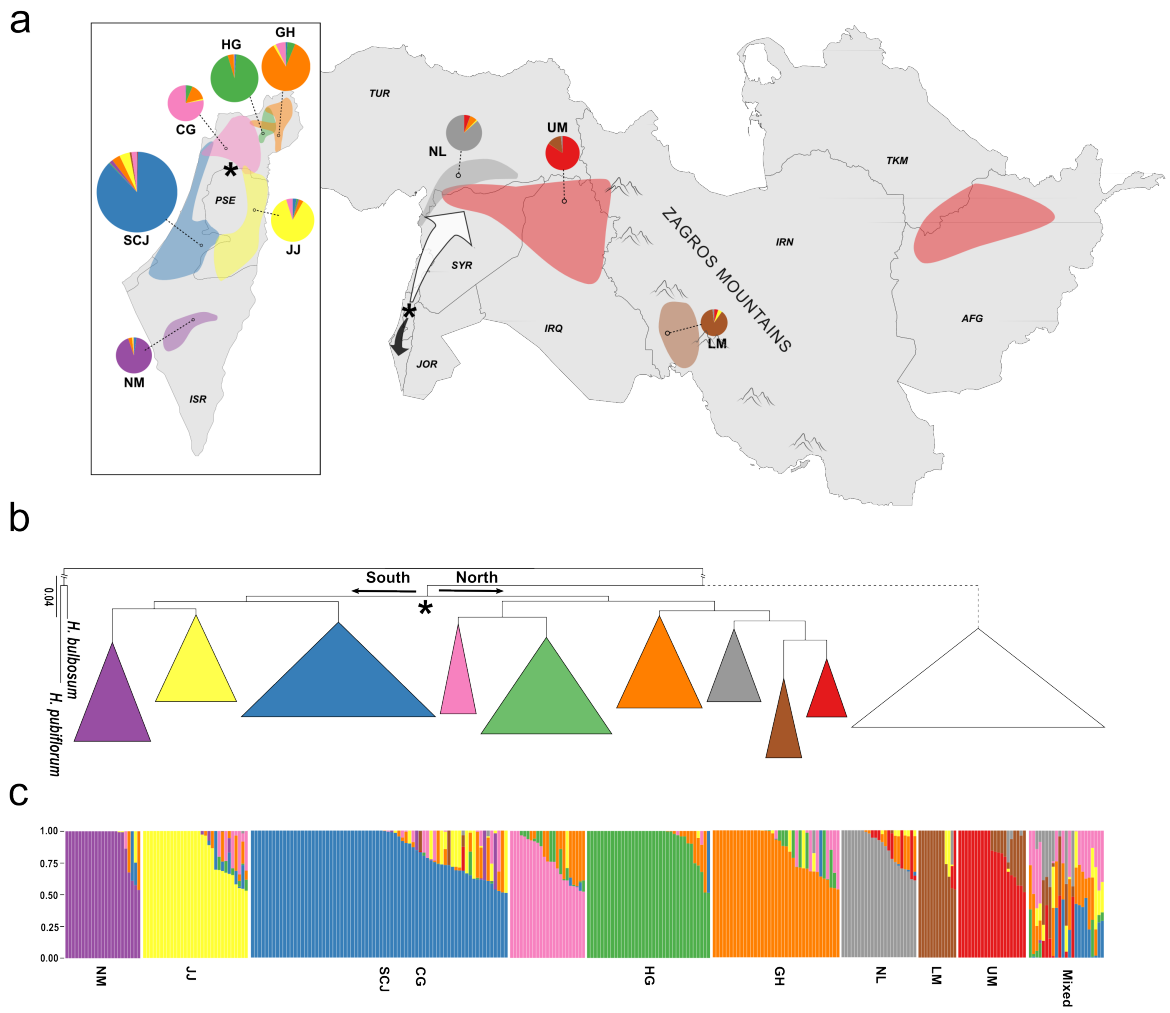

Figure 2 Geographic distribution, structure and phylogeny of nine wild barley populations.

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; gray); 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 the ancestral composition of the populations as determined by fastSTRUCTURE and are connected to the geographic centers of population distributions by dashed lines. The 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**) Population structure of wild barley as determined by fastSTRUCTURE for K=9. Vertical bars correspond to individual genotypes and colors indicate their membership in the nine subpopulations.

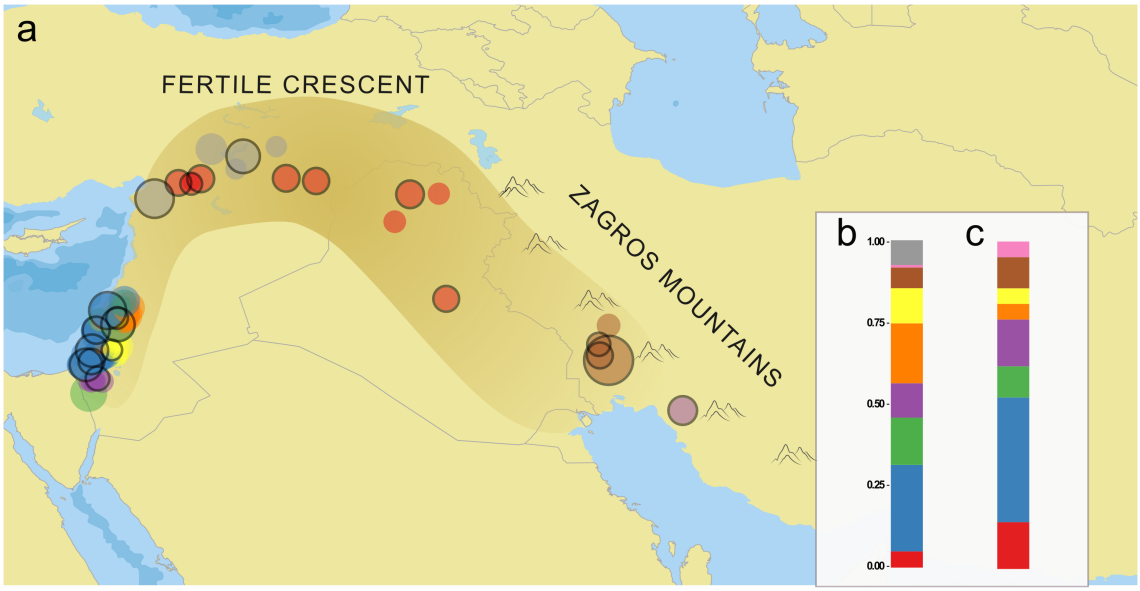

Figure 3 Geographic distribution of the Fertile Crescent wild barley genotypes carrying ancestral haplotypes of domesticated barley genotypes **(a)**. Colors correspond to nine wild barley population as in Fig. 2. Locations of the haplotypes that are ancestral to putative domestication loci carrying footprints of selection are shown as dark gray circles. Ancestry diagrams illustrating the proportional contribution of the nine wild barley populations to the domesticated barley genomes inferred using the complete dataset (**b**) and only the genes carrying footprints of selection under domestication (**c**).

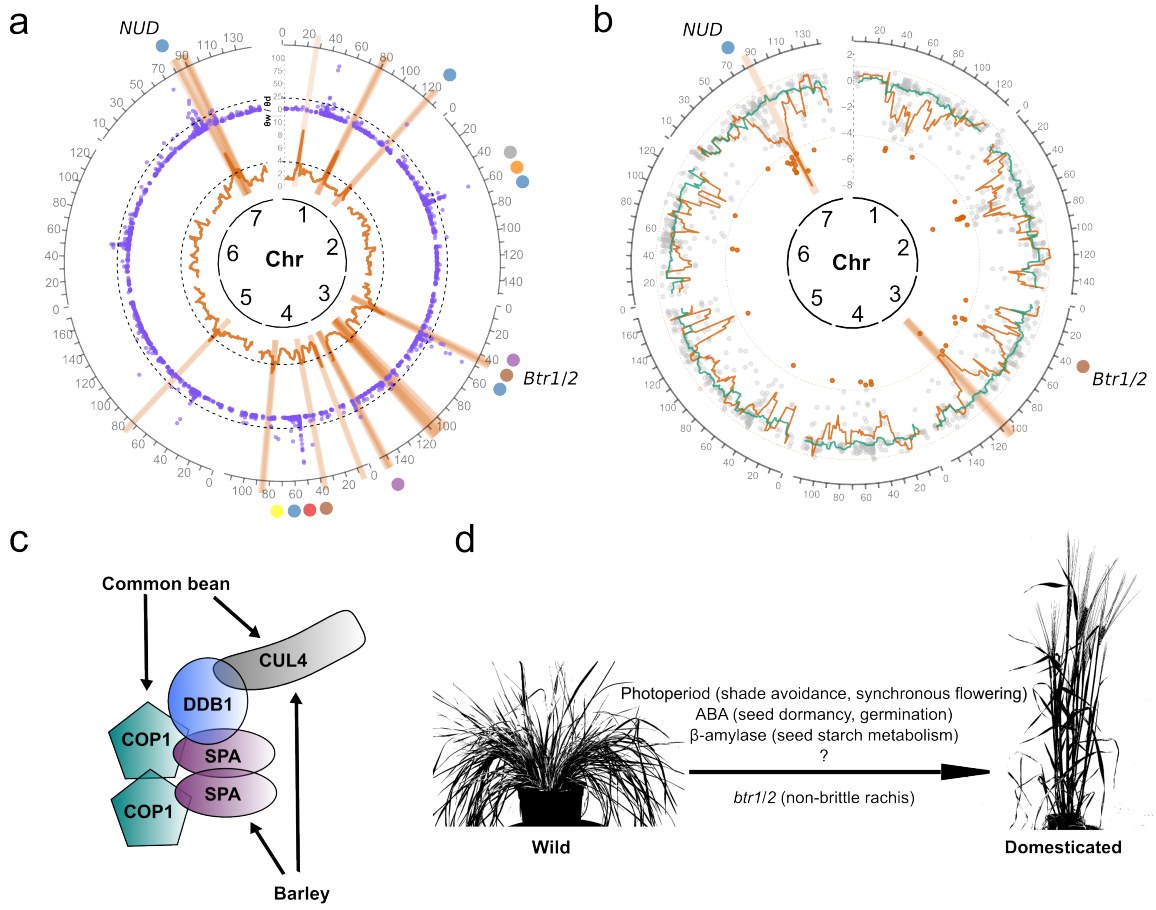

Figure 4 Genomic signatures of selection under domestication, candidate genes and pathways. (**a**) Diversity reduction (θ_w / θ_d) values for sliding windows and 2 369 individual targets. The innermost circle represents barley linkage groups followed by the sliding-window (orange) and individual target (violet) scans. The sliding window-based candidate selected regions in domesticated barley are shown by orange segments. The statistical thresholds ($p < 0.01$) are shown by dashed lines. *Btr1/2* – brittleness genes implicated in domestication (Pourkheirandish et al. 2015). The ancestry of the outlier loci is shown as dots color-coded as the corresponding wild populations in Fig. 2.

(**b**) Selection scan using the normalized Fay & Wu's H (H_{norm}) test. The innermost circle represents barley linkage groups and genetic distances in cM are shown on the outermost gray scale. The H_{norm} values along the chromosomes (sliding window 10 cM, 1-cM step) are shown for wild and domesticated subgroups by the green and orange lines, respectively. The sliding window-based candidate selected regions in domesticated barley are shown by orange

segments. The H_{norm} values of the individual loci above and below the significance thresholds are shown by orange and gray points, respectively. The orange dashed lines are the simulated thresholds of H_{norm} neutral variation (p-value < 0.001) in domesticated barley. Genetic locations of the *Btr1/2* and *NUD* genes are shown as brown and blue round symbols, respectively.

(**c**) 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 (Schmutz et al. 2014).

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

Table 1. Nucleotide diversity parameters in wild and domesticated barley.

	S, x1000 [a]	θw ($x10^{-3}$) [b]	π ($x10^{-3}$) [c]	D [d]	H_{norm} [e]
Wild	298 / 206 [f]	7.36 / 7.59	2.97 / 4.27	-1.91 / -1.40	-0.05 / -0.1
Domesticated	43 / 41	1.47 / 2.09	1.53 / 2.26	0.15 / 0.25	-1.02 / -1.07

a - number of segregating sites; b - Watterson's θ per genotyped nucleotide;

c - Nei's π_n; d - Tajima's D; e - Fay&Wu's H_{norm}; f - with / without singleton SNPs