Supplementary Notes

Components of the targeted enrichment assay

The targeted re-sequencing assay captured approximately 13.8 Mbp of the barley genome that mapped to 23,408 contigs. The captured regions comprised the genes implicated in the candidate domestication pathways, random loci and the so-called off-target captured regions.

A set of the candidate domestication loci consisted of 666 genes, including exon, intron and promoter sequences putatively implicated in modulation of domestication-related traits, such as reproductive development, photoperiod sensitivity, plant architecture, tillering, carbohydrate metabolism, seed dormancy and abiotic stress tolerance (**Supplementary Table 3**). Of all the targets, 88% were selected in a form of cDNA and 85% comprised putative promoter regulatory regions > 100 bp. The target sequences were mined from various barley genomic and transcript databases and the predicted open reading frames (ORF) of 126 genes were longer than those of the MLOC genes currently used as a barley reference gene set (IBGSC, 2011). For 52 % of the genes the complete ORFs could be mapped to the IBGSC Morex contigs, whereas the rest of the ORFs were partially or completely absent from the IBGSC reference genome. These apparently represent either the genic regions not yet incorporated in the Morex reference genome or the unique allelic variants.

To attenuate effects of the biased selection of genes on the estimates of genetic diversity, we selected fragments of 1000 random genes spread over the barley chromosomes. The enrichment design baits tiled in total 2.42 Mbp of barley gene space (**Supplementary Table 2**).

It has been shown that the hybridization-based enrichment assays, particularly NimbleGen SeqCap, are prone to generate off-target reads in the human exome capture assays (Bodi et al., 2013). In the human exome sequencing, large high-quality SNP datasets that originate from the off-target enrichment regions have been documented (Guo et al., 2012). Likewise, in this study, the size of the off-target captured regions was approximately six times larger than the size of the target capture design and yielded ~ 400.000 SNPs.

Heterozygous SNP calls

To investigate the origin of heterozygous SNP calls, we compared SNP calling results from the pooled and individually sequenced DNA samples. To this end, we mapped the exome libraries of three barley genotypes (acc. nos. ERR271694, ERR271717 and ERR271720; Mascher et al., 2013b), which were individually enriched and sequenced, and extracted the SNPs using the pipeline

describe in this study. The amount of heterozygous genotypes derived from the individually-enriched samples was comparable to the number of heterozygotes observed in the pooled-enriched dataset described in this study (Supplementary Fig. 19). This result suggested that the heterozygous calls apparently originated from the stacking of conserved paralogs, missing from the barley reference genome, rather than from the artefactual "jumping PCR", commonly contaminating samples that were enriched in pools (Kircher et al., 2011). These findings support the hypothesis that the barley genome contains multiple probably pseudogenized copies of the developmental genes, which, according to the earlier proposed model, presumably associate with the mobile elements (Wicker et al., 2011).

Transition to transversion bias (Ti/Tv)

The Ti/Tv ratio is a characteristic of mutational dynamics of genomes. It varies between different species and genomic regions. It also serves as a SNP calling quality parameter; it should not strongly deviate from the species average values (Guo et al., 2013). The estimation of the Ti/Tv bias is important not only to understand species-specific patterns of variation, but also to tailoring nucleotide substitution likelihood models used in the selection tests such as the synonymous to non-synonymous substitution ratio (K_a/K_s, also known as d_n/d_s or ω) (Yang and Bielawski, 2000). In most species, transitions are more common than transversions, whereas they constitute only half the number of transversions in the unbiased scenario. In this study, the Ti/Tv genome-wide ratio (2.48) was in par with the genome-wide Arabidopsis estimates (2.4) but higher than previously reported values in other crops (0.92 – 1.83) (Bus et al., 2012; Morton et al., 2006; Ossowski et al., 2010; Simko et al., 2006; Zhu et al., 2003) and in barley ESTs (1.15) (Duran et al., 2009). For different genomic regions, the Ti/Tv ratios varied in the range from 1.66 to 3.2 with the genome-wide value of 2.48, which is five times higher than expected if all substitutions would happen at equal rate (**Supplementary Fig. 2b**). The synonymous and non-synonymous SNPs had the highest and the lowest Ti/Tv values, respectively, whereas the ratio did not vary dramatically between the coding and non-coding regions with only slightly higher bias towards transitions in the latter group. The number of different transitions was balanced (191,635 A:G and 196,835 C:T), whereas the number of possible transversions varied in the range from 30,601 to 46,346.

Linkage disequilibrium decay and minor allele frequency

Averaging LD via regression curves obscures important aspects of the LD structure, since LD is often discontinuous and intermittent rather than smoothly declining with the distance. Exploration of the refined LD patterns may offer insights into the causes of LD and help understand

its application to diversity analysis. It has been suggested that the measures of LD show some dependency on the allele frequency (Hedrick, 1987) and that this phenomena may be related to the age of alleles and selection (Chakravarti, 1999). With that in mind, we reconstructed the LD decay curve for the marker pairs with matching MAF in 0.1 bins and compared it to the unmatched LD calculations in wild barley. LD decay curves reconstructed for the individual MAF-matched bins revealed that the dynamics of LD decay was frequency dependent. In both subspecies, LD of the low frequency alleles decayed faster than of the medium frequency alleles (**Supplementary Fig. 20**). Intriguingly, in both Arabidopsis and barley, the rate of LD decay negatively correlated with the allele frequency (Gan et al., 2011), whereas, in human data, no difference between the different allele frequency bins was found (Eberle et al., 2006).

Interpretation of the negative basal values of Tajima's D in barley

Coalescent simulations performed with the assumption of no selective pressure and the characteristics of idealized Wright-Fisher population estimated the range of neutral variation in D between -1.74 and 2.46 for the wild population (p-value < 0.01). Intriguingly, crossing the simulated thresholds, the basal levels of D in wild barley rejected assumptions of the neutral model. Significantly negative D values are traditionally interpreted as the evidence of either the background (negative) selection (BS) or of a departure of population characteristics, chiefly in terms of demographic history and recombination properties, from the Wright-Fisher's assumptions. The latter seemed to play a bigger role in shifting D to the negative values, which were observed in this study (see Discussion).

Influence of missing data and unbalanced number of samples on the diversity estimates

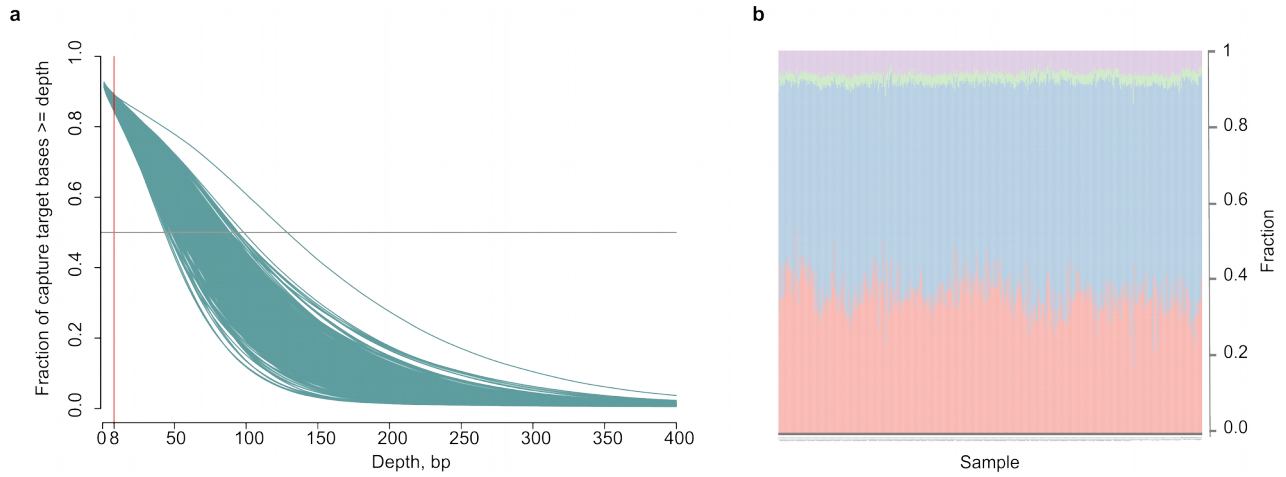
Linkage disequilibrium. Since the wild subset was ~ 5-fold larger than the domesticates, we examined whether the remarkable difference in the rate of LD decay can be attributed to the unbalanced number of individuals in the two groups. The bootstrapping of the wild subset resulted only in slightly higher estimates of LD decay (median ~ 0.6 cM, p-value < 0.001), demonstrating robustness of the LD comparisons in the subgroups unbalanced in the number of samples (**Supplementary Fig. 21**). The number of SNPs in the MAF bins greatly varied, e.g. 19,772 and 3,510 SNPs in the 0.1 and 0.2 MAF bins, respectively. The bootstrapping revealed that this difference in the number of SNP did not notably affect the LD calculations (**Supplementary Fig. 22**). To further test robustness of the LD estimations, I compared the LD decay curves obtained using the standard dataset, which contains less than 50% of missing data, and the dataset without any missing data (**Supplementary Fig. 20a**). This experiment revealed that the varying amount of

missing data did not significantly alter the LD decay estimates.

Tajima's D To evaluate the influence of considerable amount of missing data present in the dataset and of the singleton alleles, which are the fraction most prone to false positive SNP calls, we estimated D for the wild barley SNP subsets without missing data and separately without singletons. The complete removal of missing data resulted in even lower D value (-2.49), whereas cutting off singletons expectedly elevated D (-1.18), which nevertheless remained strongly negative.

Private and common alleles We estimated the influence of the unbalanced numbers of genotypes in the wild and domesticated subgroups on the amount of private alleles. This experiment revealed that, in the bootstraps using equalized wild barley subsets, the private allele ratios (57% and 65%) remained markedly higher than in the domesticates (**Supplementary Fig. 23**).

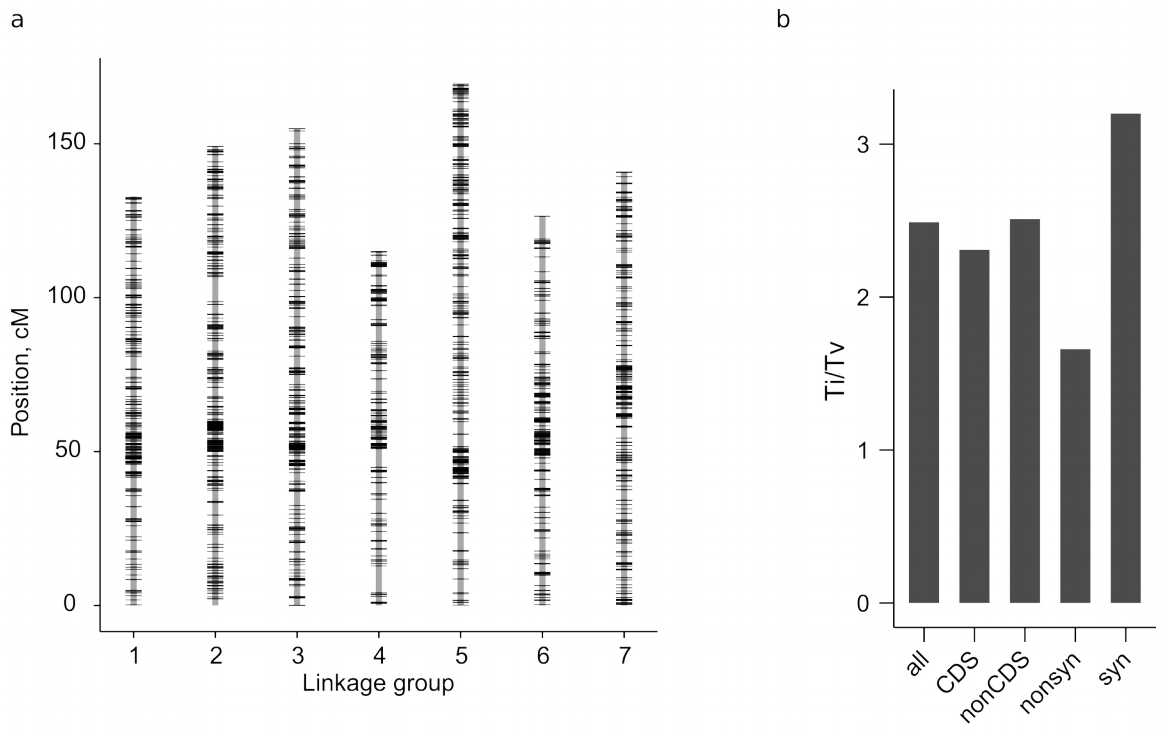
Supplementary Figures



Supplementary Figure 1 Characteristics of coverage and polymorphisms.

(**a**) A fraction of target nucleotides covered at a certain depth in the individual samples shown as cyan curves. A cut-off coverage threshold for the SNP calling and the median coverage are shown as vertical red and horizontal gray lines, respectively.

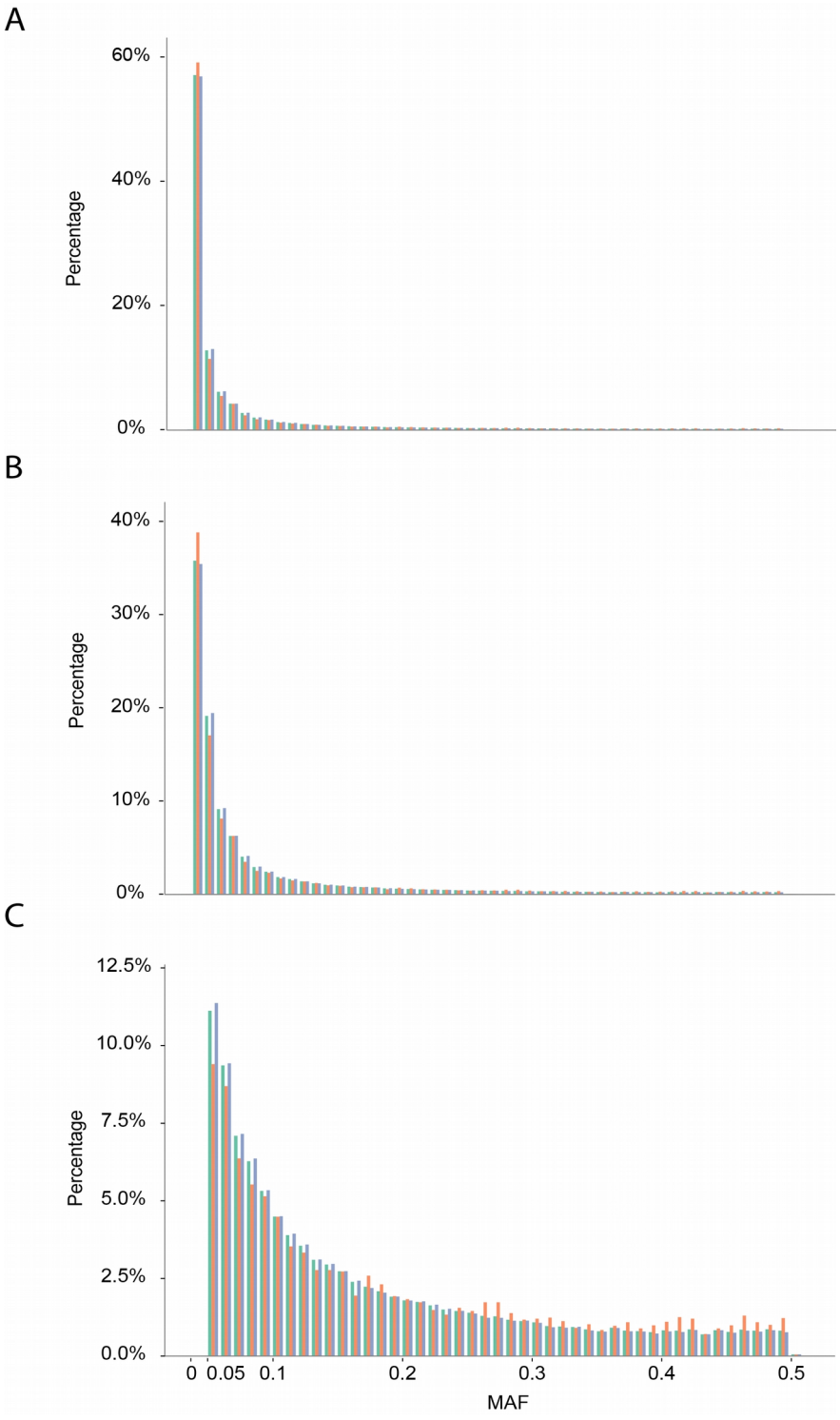
(**b**) Proportion of different call categories in the individual samples. Fractions of missing data, reference, hetero- and homozygous alleles are shown as red, blue, green and magenta bar charts, respectively.



Supplementary Figure 2 Distribution of SNP markers over the barley chromosomes and transition / transversion (Ti / Tv) ratio.

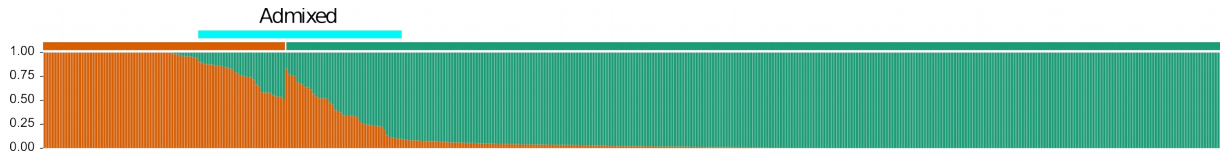
(**a**) Mapping location of the SNP markers on barley linkage group based on the PopSeq map (Mascher et al., 2013a). The linkage groups and marker positions are shown as vertical gray and horizontal black bars, respectively.

(**b**) Ti/Tv ratio of the SNP markers in different structural subgroups – coding (CDS), non-coding (nonCDS), coding synonymous (syn) and coding non-synonymous (nonsyn).



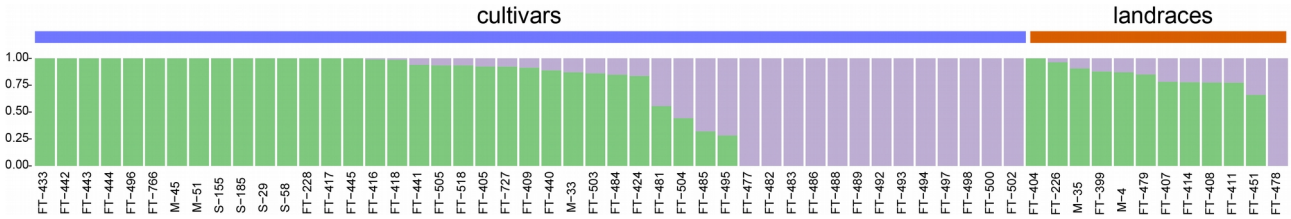
Supplementary Figure 3 Distribution of minor allele frequency (MAF) of the SNPs.

The MAF spectra are shown for the complete set with singleton SNPs (A), without singleton SNPs (B) and for the dataset with MAF >= 0.05 (C). All the sub-sets contain the SNPs with < 0.5 missing data frequency. Cyan, red and blue bars represent the MAF distributions in all, coding and non-coding SNP subsets, respectively.



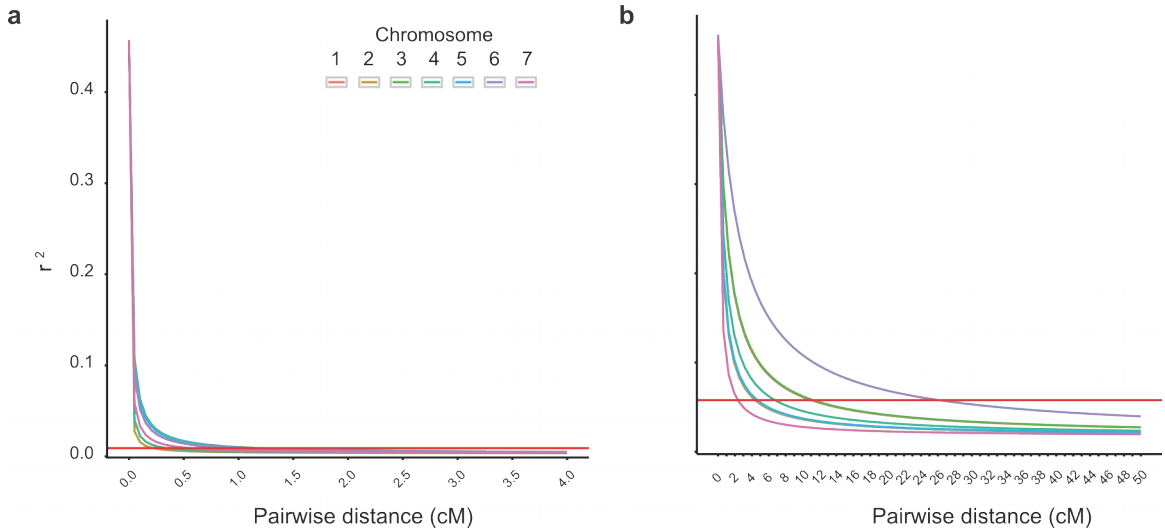
Supplementary Figure 4 Global genetic ancestry of the wild and domesticated barley genotypes as determined by the population structure analysis.

Wild and domesticated genotypes are shown as green and orange vertical bars, respectively. Admixed genotypes, which were defined as carrying more that 10 % of mixed ancestry, are shown under the cyan bar.



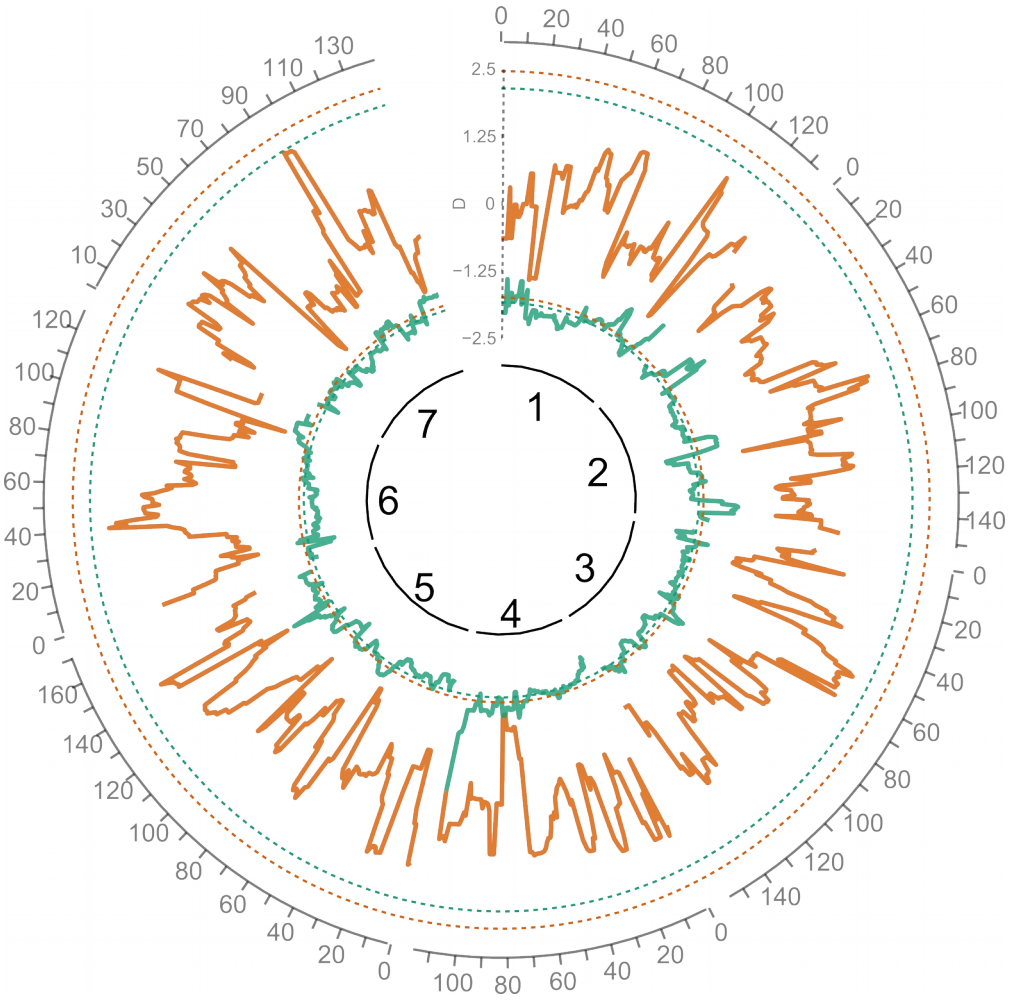
Supplementary Figure 5 Population structure analysis of the cultivar and landrace genotypes (K=2).

Vertical bars correspond to individual genotypes and color indicates their membership in the subpopulations. The cultivar and landrace subgroups are labeled by blue and orange horizontal bars, respectively.



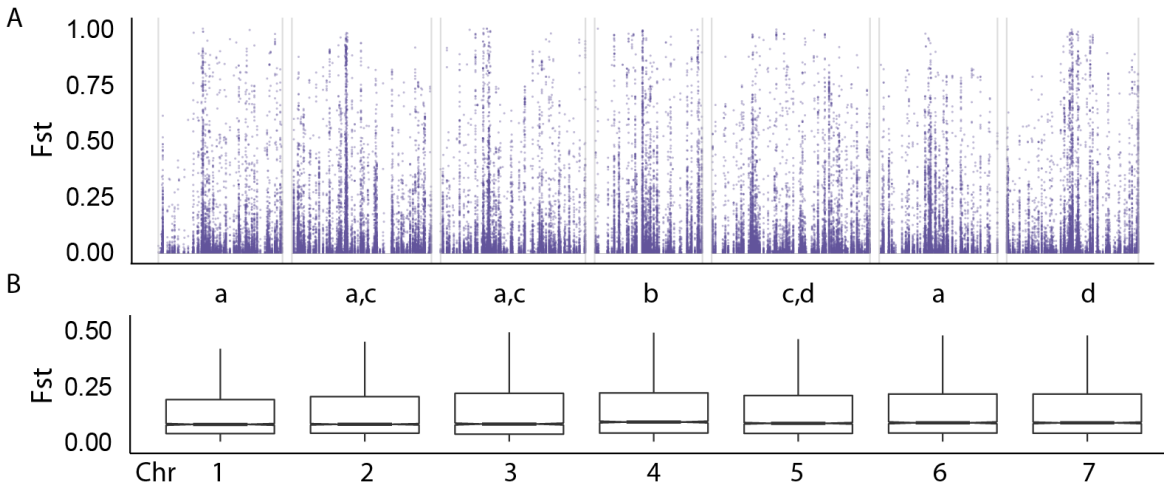
Supplementary Figure 6 LD decay of individual chromosomes in wild (a) and domesticated (b) barley.

Red horizontal lines show the background LD. Colors of the regression line correspond to individual chromosomes according to the legend.



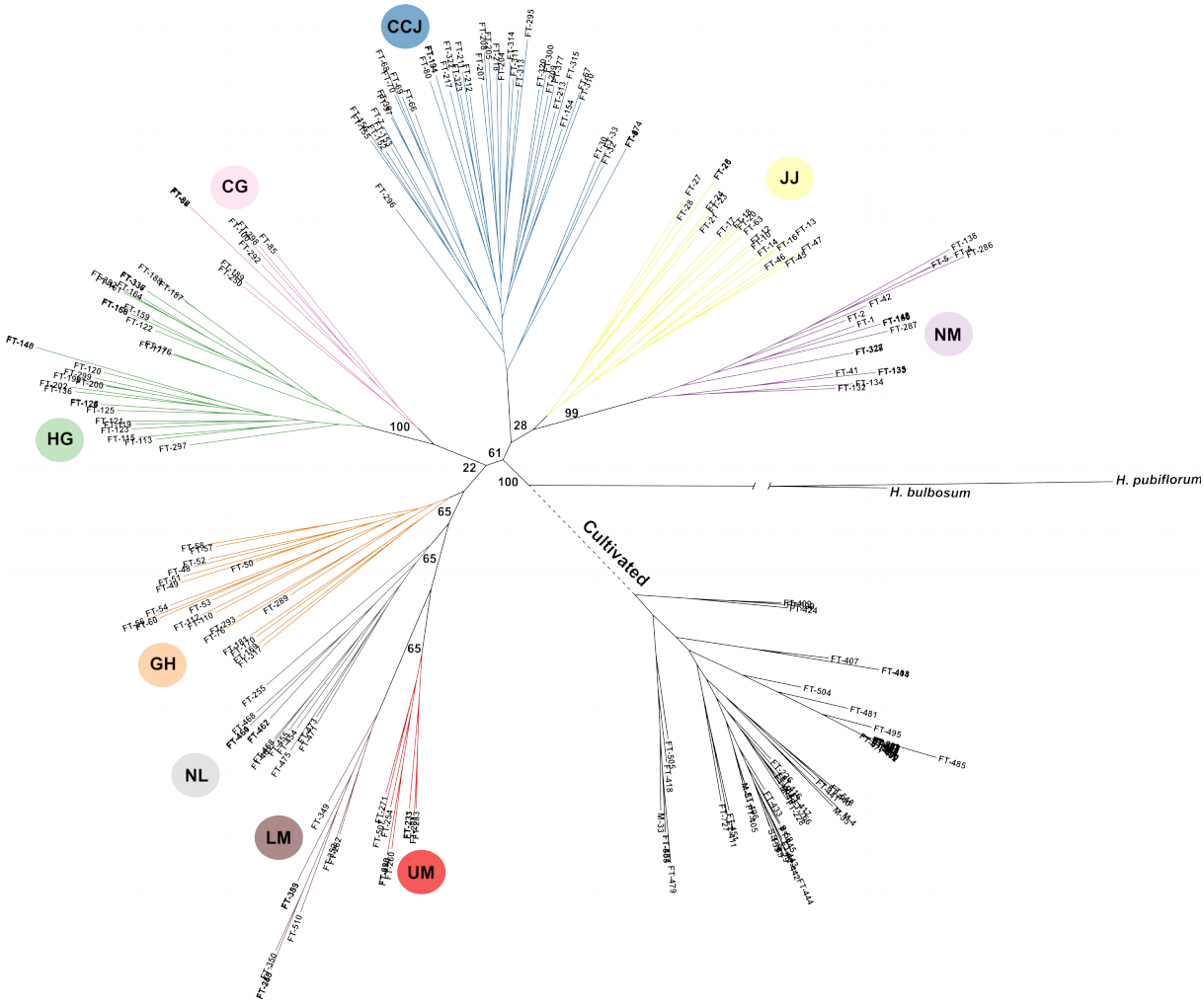
Supplementary Figure 7 Distribution of Tajima's D values along the chromosomes.

The numbers inside the innermost circle indicate barley linkage groups and genetic distances in cM are shown on the outermost gray scale. Tajima's D values (sliding window 10 cM, 1-cM step) are shown for the wild (green) and domesticated (orange) barley subgroups. Thresholds of neutral D variation obtained by coalescent simulations are shown for wild and domesticated barley as green and orange dashed lines, respectively. The scale of the D values is shown at the '0' position of the linkage group 1.

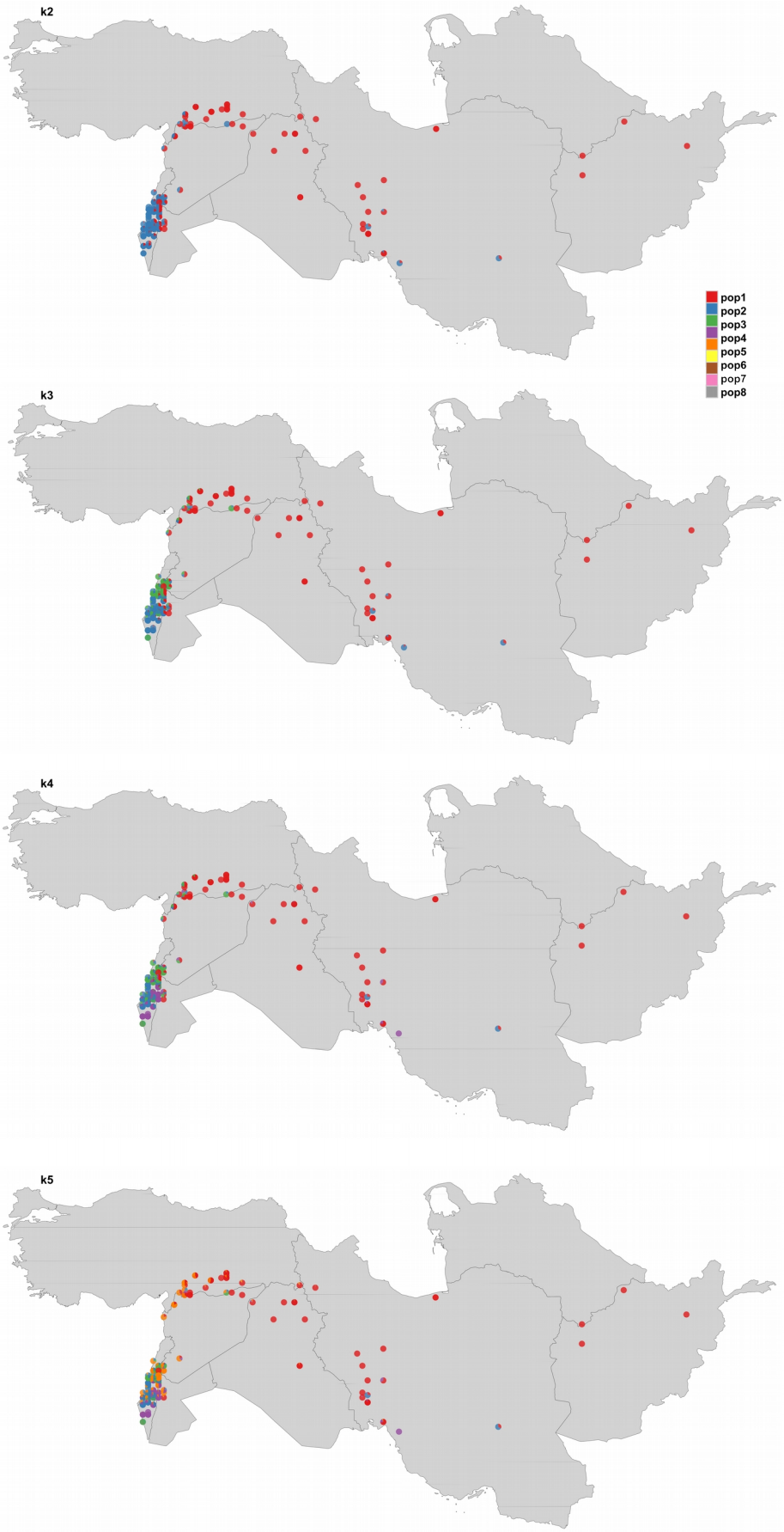


Supplementary Figure 8 Fixation index (Fst) of all pairwise SNP comparisons between wild and domesticated barley.

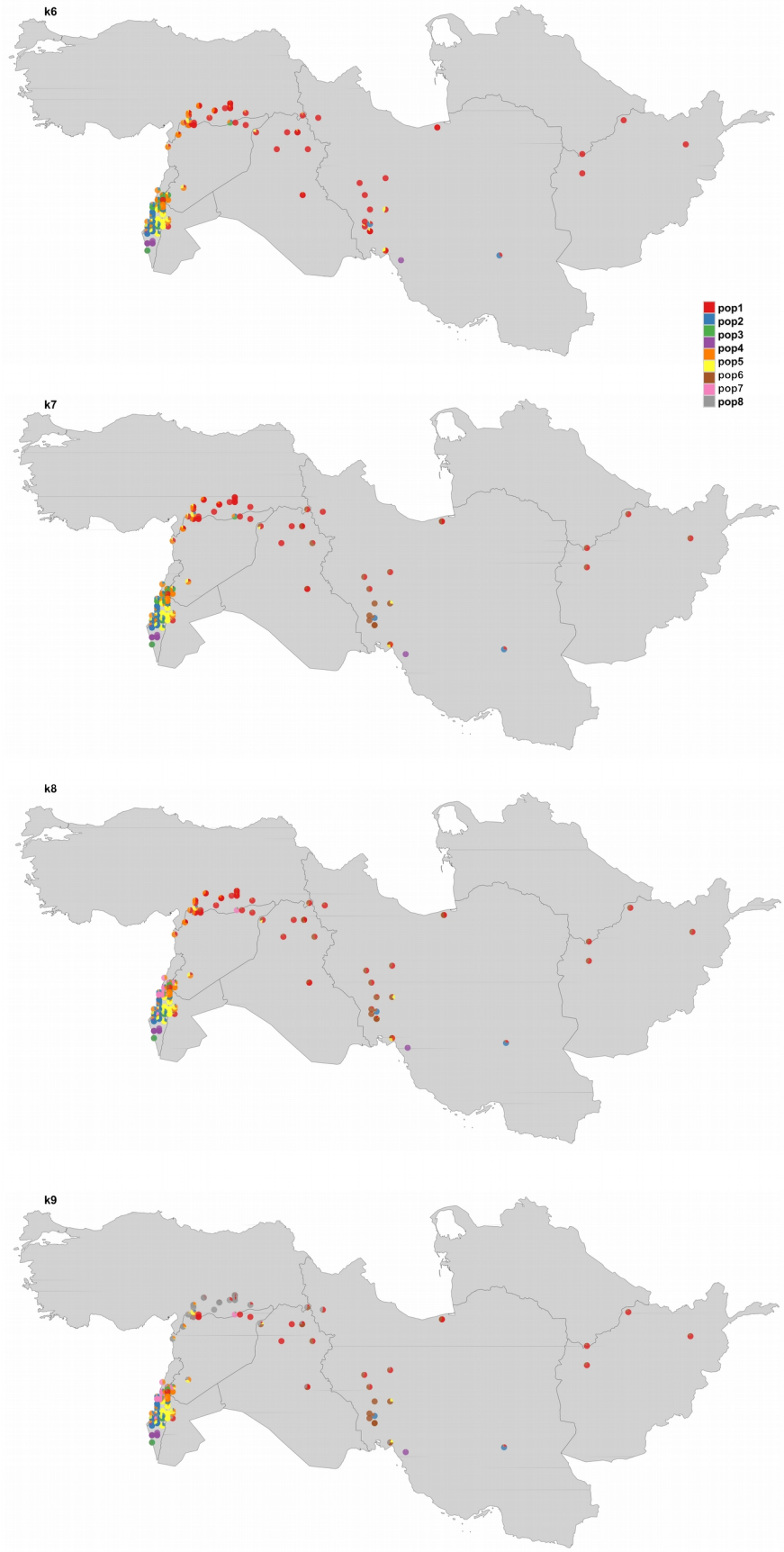
(A) The individual Fst values plotted along the barley chromosomes.
(B) Summary statistics for distributions of the Fst values by chromosome. The letters indicate significantly different distributions ($p < 0.01$).



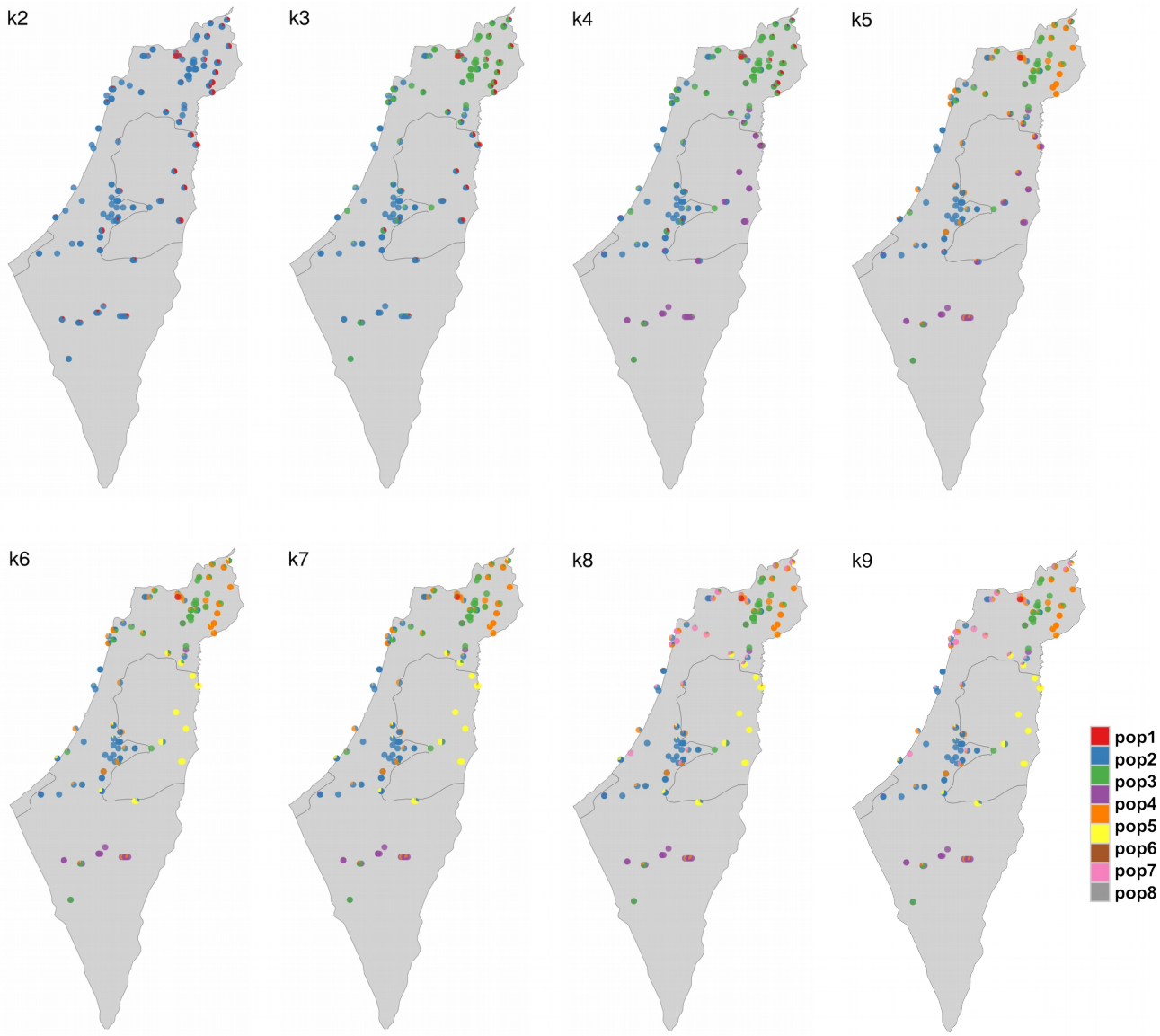
Supplementary Figure 9 The Maximum Likelihood (ML) unrooted phylogeny of 230 barley accessions. Coloured clusters correspond to the nine wild barley (H. vulgare ssp. spontaneum) populations. Carmel & Galilee (CG; pink); Golan Heights (GH; orange); Hula Valley & Galilee (HG; green); Judean Desert & Jordan 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). Cultivated barley (H. vulgare ssp. vulgare) is shown as a black cluster. The dashed line indicates that the phylogenetic placement of the cultivated barley cluster may be uncertain due to its complex hybrid origin. Wild barley *H. bulbosum* and *H. pubiflorum* are used as distant outgroup species and the length of the outgroup branch is artificially shortened. The bootstrap values are shown at the corresponding nodes.



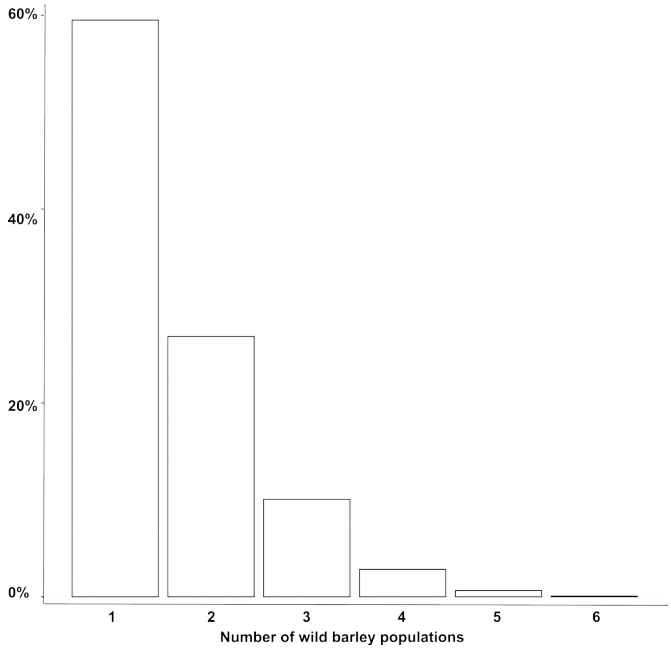
Supplementary figure 10 Distribution of the wild barley populations within the Fertile Crescent. The pie charts, reflecting ancestral composition of the individual genotypes as determined by fastSTRUCTURE for K from 2 to 9, are shown at geographic location of the genotypes.



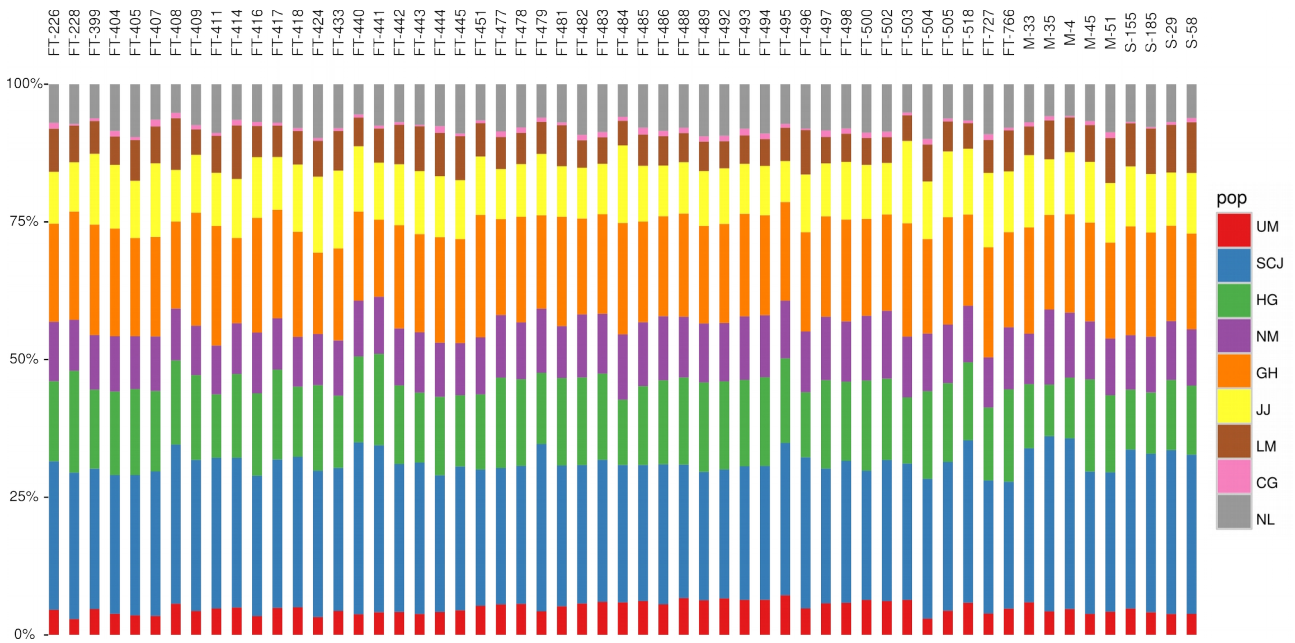
Supplementary figure 10 *(continued)* Distribution of the wild barley populations within the Fertile Crescent. The pie charts, reflecting ancestral composition of the individual genotypes as determined by fastSTRUCTURE for K from 2 to 9, are shown at geographic location of the genotypes.



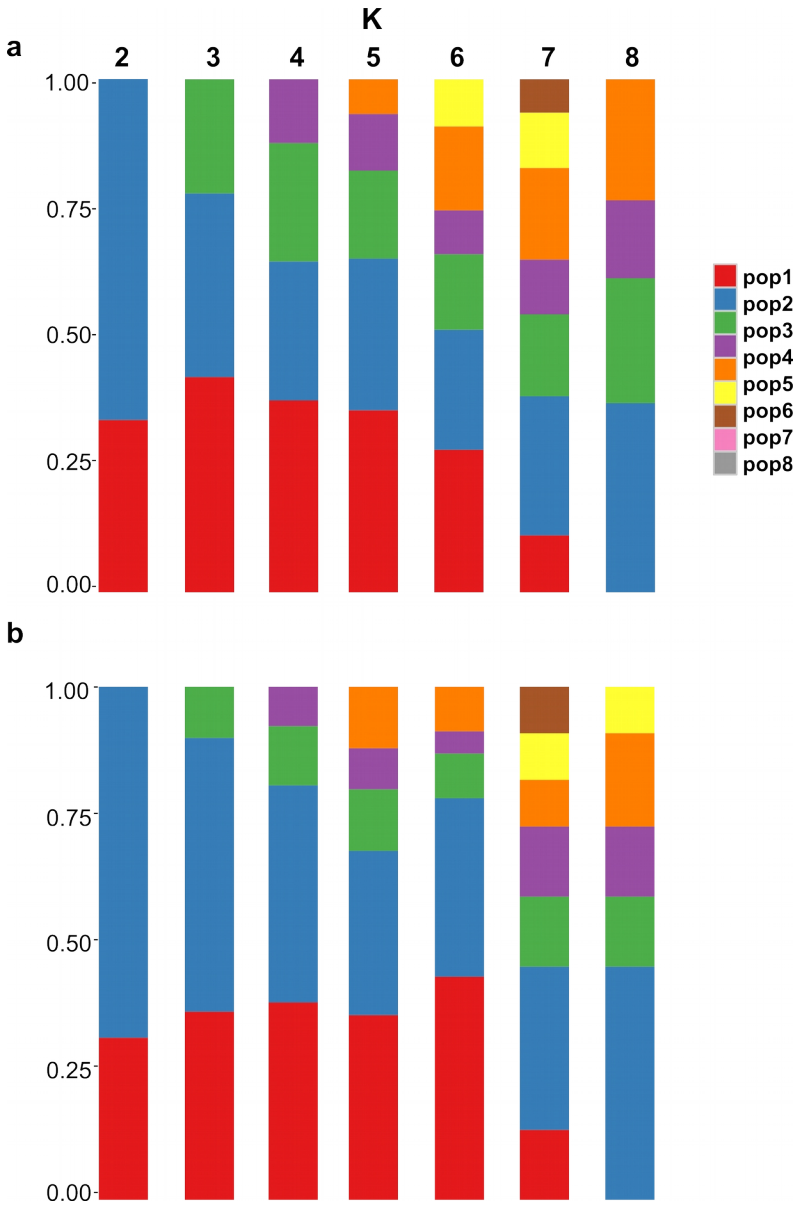
Supplementary Figure 11 Distribution of the wild barley populations within Israel and Palestine. The pie charts, reflecting ancestral composition of the individual genotypes as determined by fastSTRUCTURE for K from 2 to 9, are connected to their geographic location.



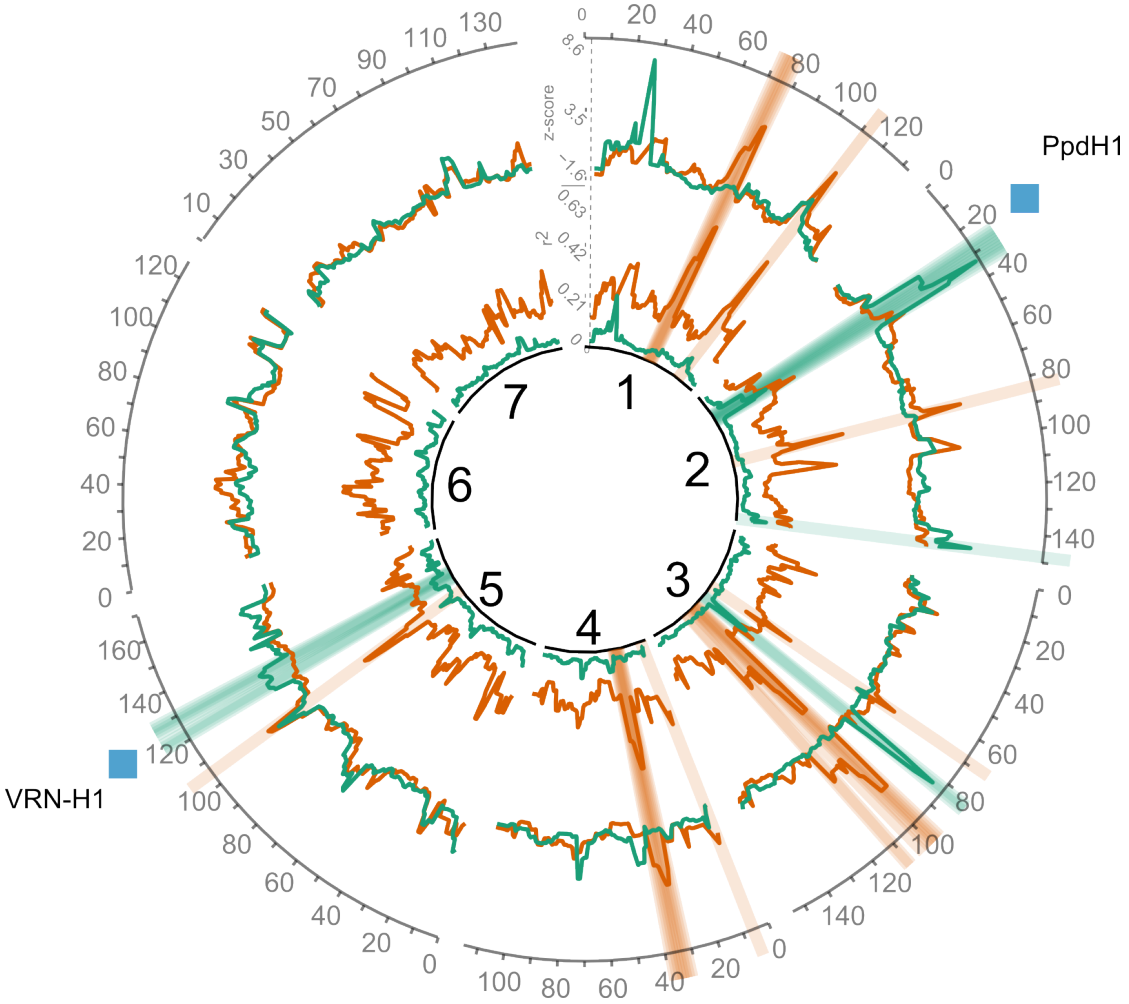
Supplemental Figure 12 Proportion of domesticated barley genes and contigs (out of 1232 contigs) that originated from a single (monophyletic) or multiple (polyphyletic) wild barley populations.



Supplementary Figure 13 Proportional contribution of nine wild barley populations to genomes of the individual barley genotypes. Carmel & Galilee, CG; Golan Heights, GH; Hula Valley & Galilee, HG; Judean Desert & Jordan Valley, JJ; Lower Mesopotamia, LM; Negev Mountains, NM; North Levant, NL; Sharon, Coastal Plain & Judean Lowlands, SCJ; Upper Mesopotamia, UM.

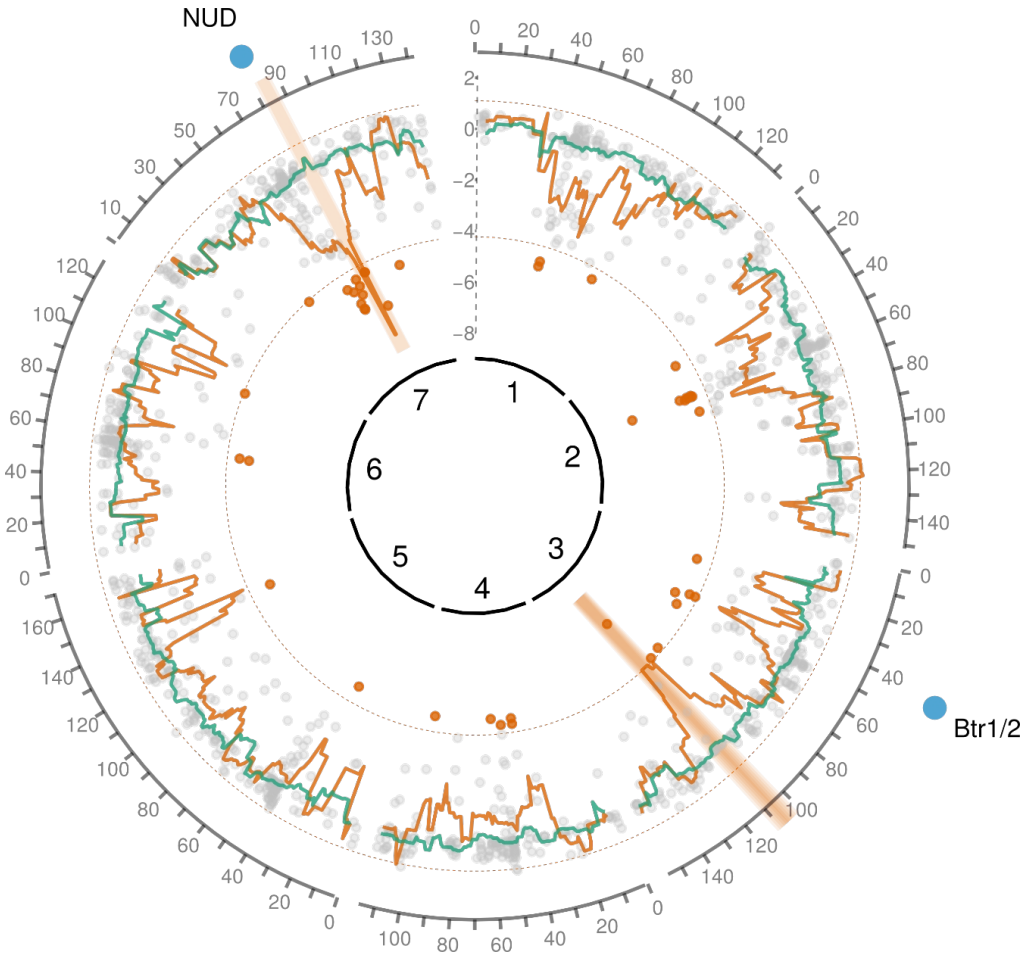


Supplementary Figure 14 Proportional contribution of wild barley populations, number of populations, K from 2 to 8, to composition of the domesticated barley genomes inferred using the complete dataset (**a**) and only the genes carrying footprints of selection under domestication (**b**)



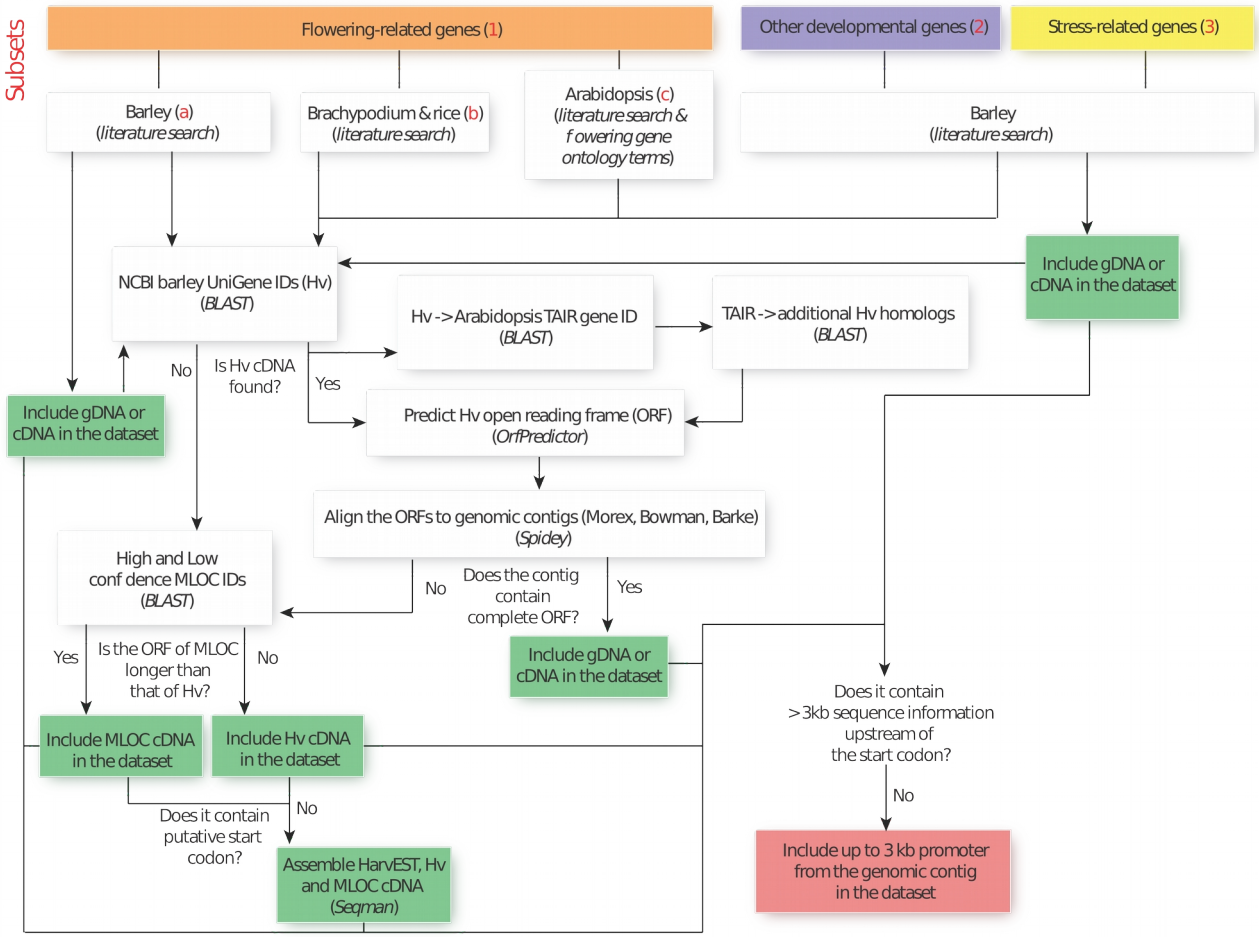
Supplementary Figure 15. Domestication-related selection signatures based on patterns of linkage disequilibrium (LD).

The numbers inside the innermost circle indicate barley linkage groups and genetic distances in cM are shown on the outermost gray scale. LD patterns (sliding window 10 cM, 1-cM step) are shown for the wild (green) and domesticated (orange) barley subgroups. The inner circle displays the original data, whereas the outer circle shows the data normalized by the z-scores. The selection sweeps defined as significant deviations ($p < 0.05$) in LD between wild and domesticated subgroups are shown by green and orange segments, respectively. Genetic locations of barley genes that have been implicated in domestication and adaptation are shown as blue pictograms.



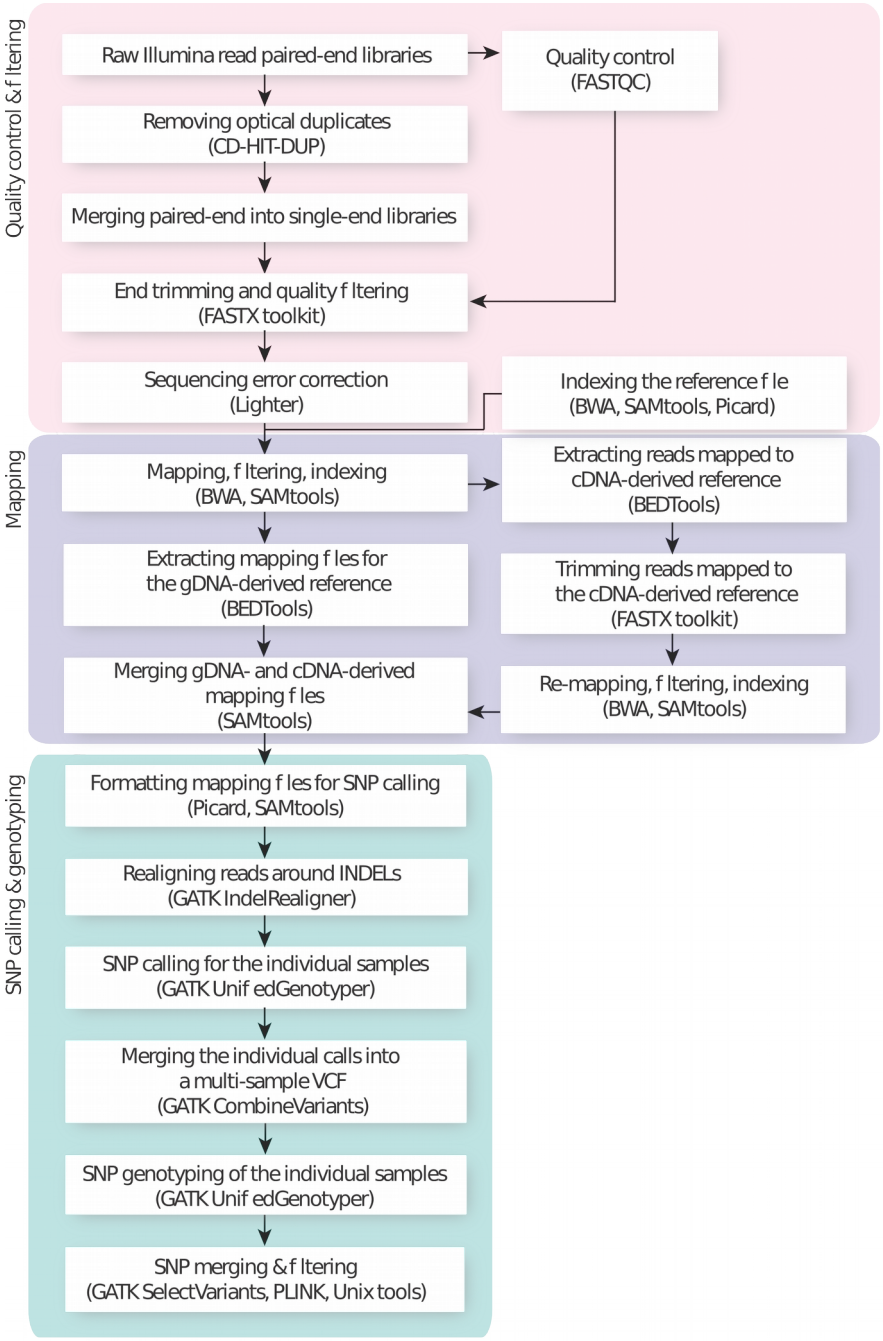
Supplementary Figure 16 Domestication-related selection signatures based on patterns of normalized Fay & Wu's H (Zeng et al., 2006) values (H_{norm}).

The numbers inside the innermost circle indicate barley linkage groups and genetic distances in cM are shown on the outermost gray scale. Distribution of H_{norm} along the chromosomes (sliding window 10 cM, 1-cM step) is shown for wild and domesticated subgroups by the green and orange lines, respectively. The window-based selection sweeps in domesticated barley are shown by the orange segments. The H_{norm} values of the individual loci that exceed the significance thresholds are shown by orange points for the domesticated subgroup. Whereas, the H_{norm} values below the thresholds are shown in gray. The orange dashed lines are the thresholds of H_{norm} neutral variation (p-value < 0.001) in domesticated barley, respectively. Genetic location of the Btr1/2 genes is shown as a blue round symbol.

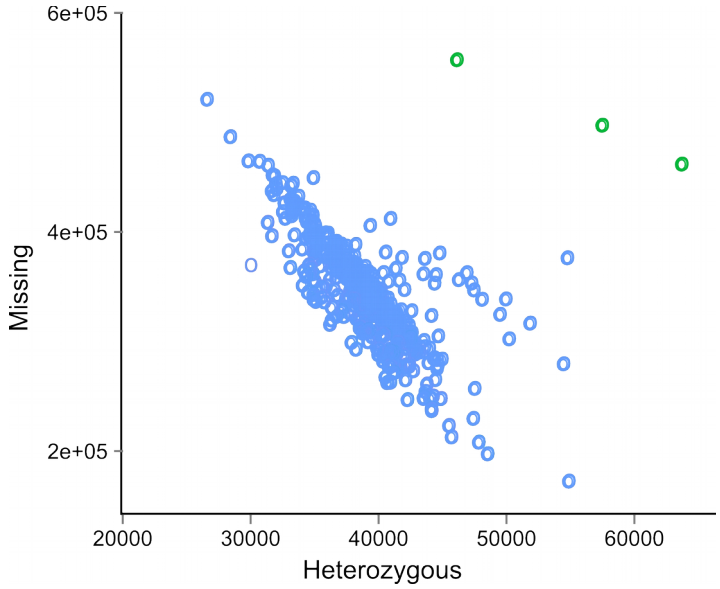


Supplementary Figure 17 Selection of target genes.

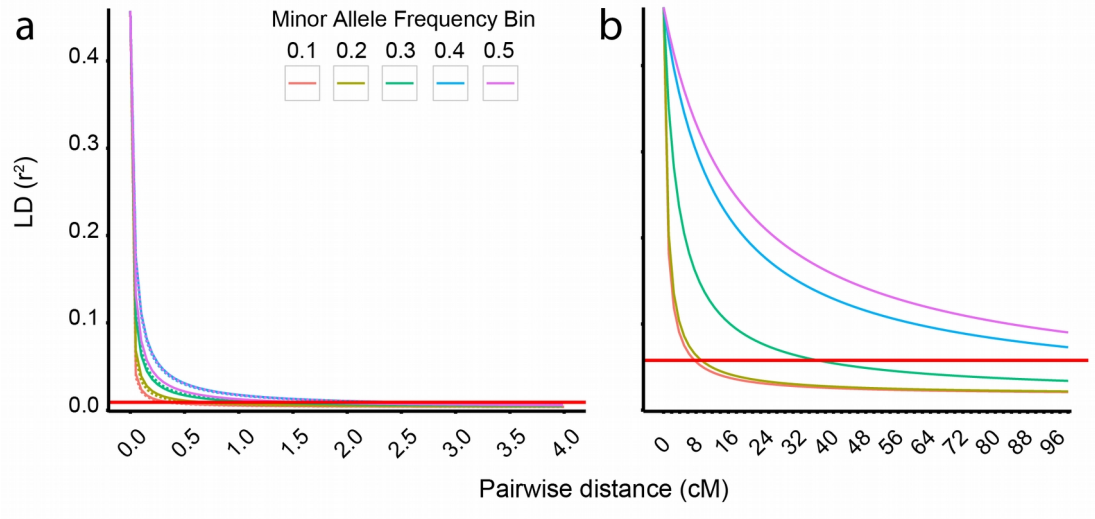
Subsets of three different functional categories of genes are highlighted in orange, violet and yellow. The output steps of the decision-making processes of selecting gene body sequences and promoter regions are highlighted in green and red, respectively.



Supplementary Figure 18 The data analysis pipeline - read filtering, mapping, SNP calling and genotyping.

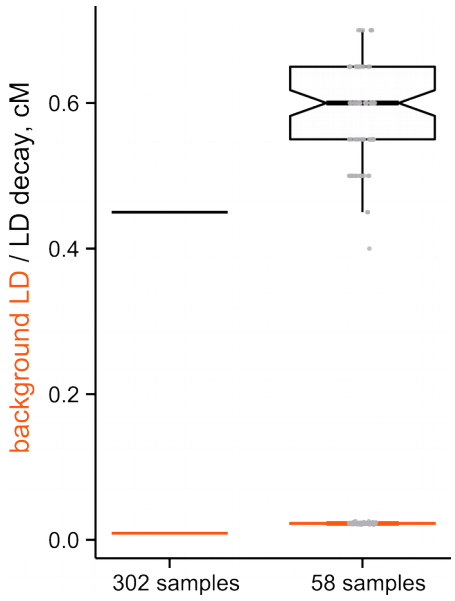


Supplementary Figure 19 Distribution of heterozygous and missing genotypes in the SNP dataset. The pooled and individually-sequenced samples are depicted as blue and green pictograms, respectively.



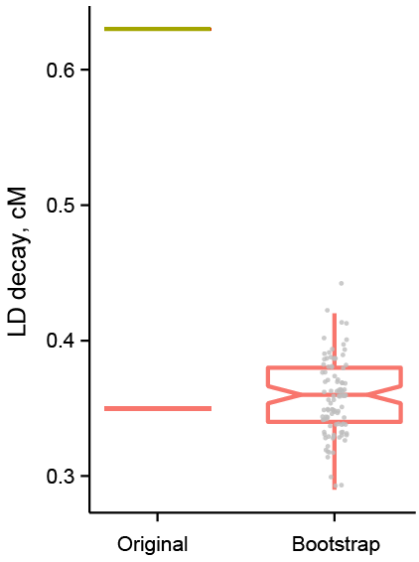
Supplementary Figure 20 LD dependency on the allele frequency.

LD decay curves obtained using the SNP datasets matched by the minor allele frequency (MAF) in bins in wild (a) and domesticated (b) barley (missing data < 50%). Color of the curves correspond to the MAF bins as shown in the legend. Red horizontal lines indicate the background LD levels. The dashed curves illustrate the LD decay in the wild barley dataset (a) without missing data.



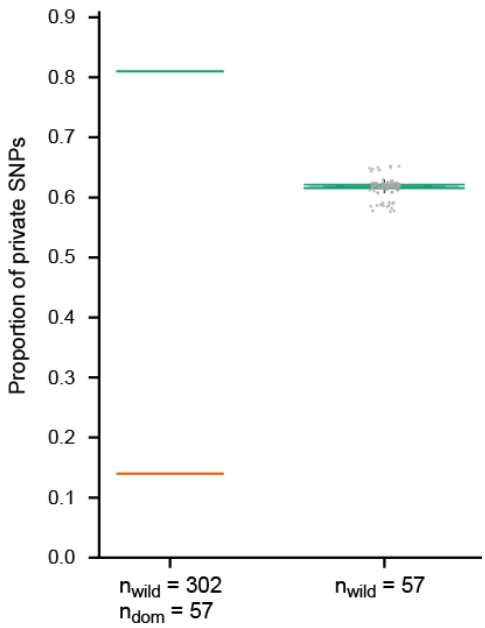
Supplementary Figure 21 Variation of linkage disequilibrium (LD) parameters in the randomly sub-sampled wild genotypes.

58 genotypes were 1000 times randomly drawn from a total of 302 wild genotypes. The red horizontal line and box plot illustrate variation in the background LD, the black horizontal line and box plot in the LD decay. Gray dots correspond to the individual LD estimates for each random subsample.



Supplementary Figure 22 Effect of the varying numbers of frequency-matched SNP on the LD decay estimates in wild barley.

The original dataset contains 19,772 and 3,510 SNPs in the 0.1 (red) and 0.2 (green) MAF bins. The bootstrapped dataset contains 1000x random draws of the 3,510 SNPs from the 0.1 MAF bin.



Supplementary Figure 23 Effect of the unequal number of samples on estimates of private allele ratios in wild and domesticated barley.

The horizontal green and red lines show the ratios calculated from the original wild and domesticated barley datasets, respectively. The green boxplot and the gray dots illustrate variation of the ratios in the equalized wild barley dataset.

Supplementary Tables

Supplementary Table 2 Characteristics of the enrichment assay and SNP calling.

	Selected size, Mbp	**Captured, Mbp**	**Captured CDS, Mbp**	**Homozygous SNPs**		
				Total, with singletons	Total, w/o singletons	Filtered set*
Target	2.42	2.24	0.85	121,294	83,752	20,954
Off-target	-	11.56	0.48	423,024	270,858	34,682
Total	9.91	13.80	1.33	544,318	354,610	55,636

* - minor allele frequency < 0.05; missing data frequency < 0.5

Supplementary Table 5 Selection signatures based on the sliding window genome scans.

ID*	**Chromosome**	**Start, cM**	**End, cM**	**Size, cM**	**Total number of genes****	
					HighConfidence	**LowConfidence**
FWH_feat1_3	3	95	116	21	209	260
FWH_feat1_7	7	61	92	31	1718	3094
NPR_feat1_1	1	22	32	10	44	154
NPR_feat2_1	1	73	86	13	104	90
NPR_feat3_1	1	118	130	12	160	214
NPR_feat1_3	3	37	51	14	521	820
NPR_feat2_3	3	95	118	23	258	306
NPR_feat3_3	3	145	159	14	137	297
NPR_feat1_4	4	2	13	11	26	47
NPR_feat2_4	4	32	43	11	64	66
NPR_feat3_4	4	79	91	12	149	209
NPR_feat1_5	5	68	79	11	111	144
NPR_feat1_7	7	75	97	22	534	806
dLD_feat1_1	1	67	82	15	126	126
dLD_feat2_1	1	107	117	10	46	109
dLD_feat1_2	2	78	88	10	125	147
dLD_feat1_3	3	64	74	10	193	215
dLD_feat2_3	3	91	116	25	228	289
dLD_feat1_4	4	1	11	10	66	92
dLD_feat2_4	4	27	43	16	86	111
dLD_feat1_5	5	100	110	10	62	133
wLD_feat1_2	2	20	39	19	124	181
wLD_feat2_2	2	142	152	10	135	266
wLD_feat1_3	3	77	89	12	131	149
wLD_feat1_5	5	116	135	19	334	598

* - feat(ure) – putatively selected region; FWH – Fay&Wu's H_{norm}, NPR – π_w/π_d, dLD and wLD – LD outlier regions in domesticated and wild barley, respectively; ** - high and low confidence genes as defined in IBGSC, 2011

Supplementary References

Bodi, K., A. G. Perera, P. S. Adams, D. Bintzler, K. Dewar, D. S. Grove, J. Kieleczawa et al. “Comparison of commercially available target enrichment methods for next-generation sequencing.” *Journal of biomolecular techniques: JBT* 24, no. 2 (2013): 73.

Bus, Anja, Jochen Hecht, Bruno Huettel, Richard Reinhardt, and Benjamin Stich. “High-throughput polymorphism detection and genotyping in Brassica napus using next-generation RAD sequencing.” *BMC genomics* 13, no. 1 (2012): 281.

Chakravarti, A. “Population genetics - making sense out of sequence.” *Nature genetics*, *21*, (1999):56-60.

Duran, Chris, Nikki Appleby, Megan Vardy, Michael Imelfort, David Edwards, and Jacqueline Batley. “Single nucleotide polymorphism discovery in barley using autoSNPdb.” *Plant biotechnology journal* 7, no. 4 (2009): 326-333.

Eberle, Michael A., Mark J. Rieder, Leonid Kruglyak, and Deborah A. Nickerson. "Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome." *PLoS Genet* 2, no. 9 (2006): e142.

Gan, Xiangchao, Oliver Stegle, Jonas Behr, Joshua G. Steffen, Philipp Drewe, Katie L. Hildebrand, Rune Lyngsoe et al. "Multiple reference genomes and transcriptomes for Arabidopsis thaliana." *Nature* 477, no. 7365 (2011): 419-423.

Guo, Yan, Jirong Long, Jing He, Chung-I. Li, Qiuyin Cai, Xiao-Ou Shu, Wei Zheng, and Chun Li. "Exome sequencing generates high quality data in non-target regions." BMC genomics 13, no. 1 (2012): 1.

Guo, Yan, Fei Ye, Quanghu Sheng, Travis Clark, and David C. Samuels. “Three-stage quality control strategies for DNA re-sequencing data.” *Briefings in bioinformatics* (2013): bbt069.

Hedrick, Philip W. "Gametic disequilibrium measures: proceed with caution."*Genetics* 117.2 (1987): 331-341.

International Barley Genome Sequencing Consortium (IBGSC) “A Physical, Genetic and Functional Sequence Assembly of the Barley Genome.” Nature 491 (2012): 711–16.

Kircher, Martin, Susanna Sawyer, and Matthias Meyer. “Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform.“ Nucleic acids research (2011): gkr771.

Mascher, Martin, Gary J. Muehlbauer, Daniel S. Rokhsar, Jarrod Chapman, Jeremy Schmutz, Kerrie Barry, María Muñoz-Amatriaín et al. “Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ).” The Plant Journal 76, no. 4 (2013a): 718-727.

Mascher, Martin, Todd A. Richmond, Daniel J. Gerhardt, Axel Himmelbach, Leah Clissold,

Dharanya Sampath, Sarah Ayling et al. “Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond.” *The Plant Journal* 76, no. 3 (2013b): 494-505.

Morton, Brian R., Irie V. Bi, Michael D. McMullen, and Brandon S. Gaut. “Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition.” *Genetics* 172, no. 1 (2006): 569-577.

Ossowski, Stephan, Korbinian Schneeberger, José Ignacio Lucas-Lledó, Norman Warthmann, Richard M. Clark, Ruth G. Shaw, Detlef Weigel, and Michael Lynch. “The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana.” S*cience* 327, no. 5961 (2010): 92-94.

Simko, Ivan, Kathleen G. Haynes, and Richard W. Jones. “Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers.” *Genetics* 173, no. 4 (2006): 2237-2245.

Wicker, Thomas, Klaus FX Mayer, Heidrun Gundlach, Mihaela Martis, Burkhard Steuernagel, Uwe Scholz, Hana Šimková et al. „Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives.” *The Plant Cell* 23, no. 5 (2011): 1706-1718.

Yang, Ziheng, and Joseph P. Bielawski. “Statistical methods for detecting molecular adaptation.” *Trends in ecology & evolution* 15, no. 12 (2000): 496-503.

Zeng, K., Fu, Y. X., Shi, S., & Wu, C. I. “Statistical tests for detecting positive selection by utilizing high-frequency variants.” *Genetics*, *174*(3), (2006): 1431-1439.

Zhu, Y. L., Q. J. Song, D. L. Hyten, C. P. Van Tassell, L. K. Matukumalli, D. R. Grimm, S. M. Hyatt, E. W. Fickus, N. D. Young, and P. B. Cregan. “Single-nucleotide polymorphisms in soybean.” *Genetics* 163, no. 3 (2003): 1123-1134.