GENOMIC PROVENANCE AND GENETIC PROVIDENCE IN DOMESTICATED BARLEY

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Ana Maria Poets

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Peter L. Morrell

August 2015

ACKNOWLEDGEMENTS

I am thankful for the support I received from my friends and family during my time in graduate school. I specially thank my mother, Maria Teresa and my sisters, Lucia and Carmiña, for their encouragement and support that overcame geographic barriers. I am immensely thankful to my husband, Jacob Poets, and his limitless love and support. I thank God, who gave me this opportunity and the strength to culminate it successfully, to Him be the glory!

I am grateful to my advisor Dr. Peter L. Morrell, for his guidance not only in the development of this thesis, but also his enlightening advice on my future career. Dr. Morrell gave me all the support I needed to achieve this goal. Thank you Dr. Morrell for trusting in me and helping me to discover the scientist within me. Dr. Morrell is a remarkable mentor and friend that have helped me to foster my curiosity for population genetics. I have learned from Dr. Morrell's example how to be a good researcher and colleague.

I would also like to thank my thesis committee members: Dr. James Anderson, Dr. Robert Stupar, Dr. David Moeller and Dr. Peter Tiffin for your involvement in my degree completion.

I thank everyone in the Morrell Lab, especially to Thomas Kono, Kiran Seth and Zhou Fang for their friendship and support, including commenting and editing my manuscripts; also for helping me to prepare for oral presentations. Thanks to all the other

members in the Morrell Lab Ashley Rozmarin (my first mentee), Chaochih Liu, Kevin Volz, Beau Miller, Paul Hoffman and Amber Eule-Nashoba for your help in my projects. Thanks to Michael Kantar and Justin Anderson for helpful discussions about my projects and science in general.

I would like to thank Dr. Clegg, Mary Durbin and Kapua Meyer (Mele) at the University of California Irvine, the co-authors of my very first author publication. It was an incredible experience working with you, coordinating a job 100% remotely was very easy having you at the other end of the line. Thanks for being so supportive of me as new scientist and for extending your friendship to me. Dr. Clegg, you set a bar very high for what it means to be an excellent contributor, I am thankful I got to learn this from you.

I am thankful for the funding from the US Department of Agriculture National Institute for Food and Agriculture through the Triticeae Coordinate Agricultural Project. It has been a great experience to be part of such a remarkable group of scientist (faculty and students) working together towards the advancement of agriculture. I am grateful for the Doctoral Dissertation Fellowship from the University of Minnesota Graduate School that supported me through my last year of training. The work presented here was carried out using computing resources at the University of Minnesota Supercomputing Institute.

Chapter 1: I am the first author on this project published in Genome Biology. Zhou Fang, Michael T. Clegg and Peter L. Morrell were co-authors on this paper. AMP and PLM designed the study, executed the analyses, and wrote the manuscript. MTC provided input on analysis and both MTC and ZF contributed to writing the manuscript

Chapter 2: I am the first author on this project. Mohsen Mohammadi, Kiran Seth, Hongyun Wang, Thomas J. Y. Kono, Zhou Fang, Gary J. Muehlbauer, Kevin P. Smith, and Peter L. Morrell were co-authors. AMP, HW, GJM, KPS, PLM designing the project. AMP and PLM wrote the manuscript. AMP, MM, ZF and TJYK, executed the analyses and contributed to writing the manuscript. I would like to thank Dr. David Hole, Dr. Patrick Hayes, and Dr. Jerome Franckowiak for sharing information about their breeding programs.

ABSTRACT

In the context of plant improvement, it is important to identify the sources of genetic diversity that are available for use, and to understand how current genetic diversity is being utilized. Comparative population genetic analyses provide a means of identifying new sources of genetic variation as well as changes in allele frequencies governed by selection (natural or artificial) or demographic processes. These analyses have the potential to identify variants contributing to local adaptation, and determine possible limits to selection.

Considering the broad geographic distribution of both wild and cultivated barley, it is likely that many traits have multiple origins. Thus, the identification of the source population that contributed to a specific chromosomal region in landraces (primitive barley domesticates) can help identify genetic variants that contribute to agronomic traits. In Chapter 1, I use genotyping data for a set of 803 landrace and 277 wild barley accessions to address two primary questions (*i*) Do specific wild populations contribute disproportionately to barley landraces?, (*ii*) does the genetic contribution of wild populations to landraces vary across the genome or across the broad geographical range of landrace cultivation? I find that multiple wild populations contributed to the genetic composition of the landraces. Their contribution differs across the genome and across the geographic range. We rule out recent introgression, suggesting that these contributions are ancient. The over-representation in landraces of genomic segments from local wild

populations suggests that wild populations contributed locally adaptive variation to landraces. This chapter has been published in Genome Biology (see Poets *et. al.,* 2015).

Barley (*Hordeum vulgare* ssp. *vulgare*) was introduced to North America by early European colonist as early as 1602 on the East Coast of what is now the United States. The initial introduction was for beer production, since that time barley has been adapted to diverse environments and produced across the nation. This adaptation has involved many generations of selection towards ideal phenotypes constrained by the end-market requirements. In Chapter 2, I analyze a data set comprised of 3,613 barley accessions representative of North American barley breeding programs. I aim to determine: (*i*) the patterns of recent and long-term selection in these programs, (*ii*) assess the effects of drift and linked selection in breeding populations, and (*iii*) identify the extent of gene flow among barley breeding programs. Applying population genetics approaches I identify loci known to be controlling major traits and a series of loci putatively involved in recent and older bouts of selection. There is clear evidence of genetic drift and linked selection acting in these populations. There is evidence of possible gene flow among populations with similar growth habit and inflorescence type. Finally, I emphasize the effects of ascertainment bias towards spring six-row types in the Barley Oligo Pooled Assay (BOPA) platform.

Supplemental Material

Large supporting information files (Tables) are presented as electronic attachments:

Chapter 1:

Table S1.1 803 landrace accessions used in this study with latitude and longitude information.

Table S1.2 1,896 SNPs shared between wild barley and landrace populations.

Table S1.5 Genome-wide ancestry as a function of distance from wild populations. Proportions of ancestry for individual landraces, and the great circle distance between each individual and the closest accession from each wild population. Attached as Supplemental Material.

Table S1.7 Frequency of alleles private to the wild present in each of the landrace populations. Private SNPs in wild barley that are present in the landraces, including linkage group and their frequency in each landrace population. Attached as Supplemental Material

Table S1.9 Chromosome painting. Individual landrace ancestry inferred at each genomic region. The cells are colored according to their inferred ancestry from the wild populations. Two haplotypes (rows) per landrace accession are depicted. Attached as Supplemental Material.

Chapter 2:

Table S2.1 Sample information for the 3,613 barley accessions representing North American breeding programs.

Table S2.2 Single Nucleotide Polymorphism (SNP) markers used in this study with linkage group, genetic position, and annotation information.

Table S2.4 Annotations for SNPs having outlier F_{ST} values in the breeding programs comparison

Table S2.5 Annotations for SNPs having outlier F_{ST} values in the growth habit comparison

Table S2.6 Annotations for SNPs having outlier F_{ST} values in the row-type comparison

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 CHAPTER 1

BARLEY LANDRACES ARE CHARACTERIZED BY GEOGRAPHICALLY HETEROGENEOUS GENOMIC ORIGINS

The genetic provenance of domesticated plants and the routes along which they were disseminated in prehistory have been a long-standing source of debate. Much of this debate has focused on identifying centers of origins for individual crops. However, many important crops show clear genetic signatures of multiple domestications, inconsistent with geographically circumscribed centers of origin. To better understand the genetic contributions of wild populations to domesticated barley, we compare single nucleotide polymorphism frequencies from 803 barley landraces to 277 accessions from wild populations.

We find that the genetic contribution of individual wild populations differs across the genome. Despite extensive human movement and admixture of barley landraces since domestication, individual landrace genomes indicate a pattern of shared ancestry with geographically proximate wild barley populations. This results in landraces with a mosaic of ancestry from multiple source populations rather than discrete centers of origin. We rule out recent introgression, suggesting that these contributions are ancient. The over-representation in landraces of genomic segments from local wild populations suggests that wild populations contributed locally adaptive variation to primitive varieties.

This study increases our understanding of the evolutionary process associated with the transition from wild to domesticated barley. Our findings indicate that cultivated barley is comprised of multiple source populations with unequal contributions traceable across the genome. We detect putative adaptive variants and identify the wild progenitor conferring those variants.

1.1 INTRODUCTION

The domestication of plants and animals around 10,500 YBP initiated the development of complex human societies and provided the raw material on which modern agriculture still depends (Diamond, 2002; Harlan & Zohary, 1966; Pourkheirandish & Komatsuda, 2007). Barley and early forms of wheat, and later pea, lentil, chickpea, and a number of other species were the primary plants in the Neolithic agropastoral package that originated in the Fertile Crescent and later spread across North Africa and most of Eurasia (Harris & Gosden, 1996; Zohary, Hopf, & Weiss, 2012). A growing body of archeological evidence suggests that Fertile Crescent agriculture involved a gradual transition from plant collection into management and cultivation (Harlan & Zohary, 1966; Fuller, Willcox, & Allaby, 2011; Weiss, Kislev, & Hartmann, 2006). Having started with the collection of seed from fully wild barley populations that began as much as 50,000 YBP (Zohary et al., 2012; Lev, Kislev, & Bar-Yosef, 2005) agricultural practices were ultimately widely disseminated through a mix of cultural and demic diffusion (Harris & Gosden, 1996; Fuller et al., 2011; Ammerman & Cavalli-Sforza, 1984; Willcox, 2013).

Extensive archeological remains at human Neolithic sites capture the timing and phenotypic transition from wild to cultivated barley across the Near East (Harlan & Zohary, 1966; Zohary et al., 2012; Willcox, 1991; Willcox, 2005; Zohary, 1969) making barley a particularly desirable system to study the evolution of domestication. The

biology of the species also facilitates genetic studies because it is a diploid, self-fertilizing species with a genetically diverse wild progenitor that has a broad geographic distribution (Harlan & Zohary, 1966) marked by substantial genetic differentiation among wild populations (Fang et al., 2014). Recent genetic studies of wild and landrace (primitive domesticate) barley collections (Morrell & Clegg, 2007; Saisho & Purugganan, 2007) and evidence of independent origins of important domestication-related traits (Komatsuda, Maxim, Senthil, & Mano, 2004; Takahashi & Hayashi, 1964; Tanno & Willcox, 2012) support the hypothesis of at least two independent domestication events followed by some degree of admixture among domesticates from distinct portions of the geographic range of the wild barley distribution. This scenario is also consistent with minimal loss of diversity in cultivated barley relative to its wild ancestor (Morrell, Gonzales, Meyer, & Clegg, 2014). Here we address the following questions: (1) Do specific wild populations contribute disproportionately to barley landraces? and (2) does the genetic contribution of wild populations to landraces vary across the genome or across the broad geographical range of landrace cultivation?

Multiple lines of evidence, presented here, indicate that barley landraces have mosaic ancestry, reflecting the contribution of all major geographic portions of the range of the wild progenitor species. A broad contribution of wild progenitor populations to the landraces is consistent with archeological evidence for a gradual transition to cultivation (Purugganan & Fuller, 2011). This is demonstrated by phenotypic change, particularly non-shattering of the inflorescence, which is essential for barley domestication, identified

at many Neolithic sites. Identification of putatively adaptive contributions from wild progenitor populations provides a potential means of detection of loci contributing to locally adaptive variation (for example, for climatic adaptation).

1.2 MATERIALS AND METHODS

We used 803 barley landrace accessions from the 2,446 landrace and cultivated lines in the National Small Grains Collection (NSGC) Core Collection from the USDA. These 803 individuals include all landraces collected in Europe, Asia, and North Africa constituting the range of dissemination of cultivated barley in human pre-history (Figure 1.1, Table S1.1).

We also make use of the 284 wild barley accessions from the Wild Barley Diversity Collection (WBDC) (Steffenson et al., 2007) analyzed in (Fang et al., 2014). Accessions represent the entire geographic range of wild barley including the Fertile Crescent, Central Asian, and adjacent North African regions.

1.2.1 Genotypic data

A collection of 2,446 landraces and cultivated accessions from the NSGC were genotyped with 7,864 SNPs using the Illumina Infinium SNP genotyping platform (hereafter referred to as the 9K). The 9K chip includes 5,010 SNPs discovered in a panel of 10 barley varieties, composed primarily of European two row cultivars. In addition, a set of 2,832 SNPs used for the existing BOPA (Barley Oligo Pooled Assay 1 and 2) on the Illumina Golden Gate genotyping platform (Close et al., 2009) was included. Additionally, 22 SNPs from resequencing studies were added, giving a total of 7,864 SNP assays on the chip (Comadran et al., 2011). The BOPA SNPs derived principally from one wild barley accession and eight malting barley cultivars, from Europe, the United States, and Japan (Briscoe et al., 1994).

We used automated genotype calling implemented in the software ALCHEMY (Wright et al., 2010). ALCHEMY uses a Bayesian model of the raw intensity data files. This approach does not assume Hardy-Weinberg Equilibrium; and each single nucleotide polymorphism (SNP) call is independent of other genotype calls at the SNP. SNP calls with posterior probability >0.95 were recorded; calls below this threshold were marked as missing data. The accuracy of calls was verified following the method explained previously (Morrell & Clegg, 2007).

SNP quality control procedures consisted of the removal of SNPs that were monomorphic, had more than 10 % missingness, or had more than 10 % heterozygosity (see reference Fang et al., 2014). We retained 6,152 SNP for all 2,446 landraces and cultivated lines after quality control. The curated SNP dataset was used to identify potential duplicate individuals in the NSGC barley core. The details of the procedure used to identify duplicate accessions are explained in Muñoz-Amatriaín et al. (2014). We retain 803 landrace accessions after quality control.

The 284 accessions from the WBDC were genotyped with 3,072 SNPs (Close et al., 2009), a subset of the 9K platform. After quality control this dataset consisted of 2,624 SNPs for each of the 284 accessions (see reference Fang et al., 2014) for specific information about these populations and SNP quality control steps.

We used the consensus genetic map described in (Muñoz-Amatriaín et al., 2014) which is the result of merging the 11 genetic maps of the 2011 consensus map developed by Muñoz-Amatriaín et al. (2011) with the iSelect SNP platform map based on the Morex

x Barke mapping population (Comadran et al., 2011). This map, referred to here as the 'iSelect map', identifies genetic position for 4,527 of the SNPs used to genotype the NSGC accessions.

We infer the phase of heterozygous sites (approximately 0.1 % of sites) using PHASE v.2.1.1 (Stephens & Donnelly, 2003; Stephens, Smith, & Donnelly, 2001) for all 1,896 SNPs which were shared between landraces and wild barley, and had genetic map positions (Table S1.2). The runs are set to the default values for number of iterations = 100, thinning intervals = 1, and burn-in = 100. We consider only phased calls with probabilities of at least 90 %. All imputed sites for missing data are re-set to missing values using a customized R script (R Project for Statistical Computing, http://www.r-project.org/). Experimentally phased haplotypes are used in two analyses where they are critical to inference, that is, the estimation of admixture proportions and assessment of identity by state between wild and landrace accessions.

Linkage disequilibrium (LD) as measured by r^2 (Hill & Robertson, 1968) is calculated for all possible pairwise comparisons on each linkage group based on the 4,527 SNPs included in the iSelect genetic map. We considered SNPs with minor allele frequency (MAF) >5 %. The LDheatmap package in R (Shin, Blay, McNeney, & Graham, 2006) was used to generate plots of LD relative to genetic distance (Figure S1.1).

1.2.2 Genetic assignment

To determine the geographic population structure among the 803 landraces in our dataset, we used a Bayesian clustering algorithm implemented in STRUCTURE (Pritchard et al., 2000; Falush, Stephens, & Pritchard, 2003). We explored the numbers of clusters (referred to as *K*) ranging from 1 to 7 (Figure S1.2). For each value of *K* we used 10 replicated runs, with a burn-in length and run length of 100,000 iterations. We used an admixture model because archeological and genetic evidence suggest extensive movement of barley and thus likely admixture (Harris & Gosden, 1996; Zohary et al., 2012; Morrell & Clegg, 2007; Morrell et al., 2014; Ordon, Schiemann, & Friedt, 1997). We used the uncorrelated allele frequency model, which is more conservative. STRUCTURE analysis was run based on the 6,152 SNPs for the 803 landraces. Considering the high selfing rate of barley >98.2 % (Abdel-Ghani et al., 2004) we used a haploid model (option PLOIDY=1). To summarize the assignment results for all replications we used CLUMPP (Jakobsson & Rosenberg, 2007). CLUMPP deals with label switching (that is, when cluster names change between replicates); and multimodality problems (that is, when individual samples change clusters in each replication).

We used two *ad hoc* approaches, ΔK (Evanno, Regnaut, & Goudet, 2005) and Clusterdness (Rosenberg et al., 2005), to determine the number of clusters that best explain the population structure among the landraces. ΔK is based on the second order rate of change of the log probability of data between successive *K* values (Evanno et al.,

2005), and Clusterdness (Rosenberg et al., 2005) is the extent to which individuals are estimated to belong to a single cluster rather than to a combination of clusters (Figure S1.3).

The primary population structure identified here ($K = 2$, Figure S1.2) agrees with previous observations of population differentiation of landrace and wild barley accessions east and west of the Zagros Mountains (Morrell & Clegg, 2007; Morrell et al., 2014). The large sample considered here permits greater resolution of the geographic differentiation among barley landraces (Figure S1.2).

We estimated the degree of differentiation among individuals by PCA. For this analysis we use all the 4,527 SNPs with known genetic position for the 803 landrace accessions. The PCA was performed in the SmartPCA program from the EIGENSOFT package (Patterson, Price, & Reich, 2006). SmartPCA permits PCA analysis with SNP loci that include missing data, thus our analysis is based on the full SNP genotyping dataset. Procrustes analysis (Wang et al., 2010) implemented in the vegan package in R (Jari Oksanen, F. Guillaume Blanchet, & Roeland Kindt, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Helene Wagner, 2015) was used to identify the optimal rotation that maximizes the similarity between genetic variation on PCA plot and geographic maps of sample locations (Figure S1.4).

We used SharedPoly and compute from the libsequence library (Thornton, 2003) to calculate summary statistics, including number of segregating sites, number of private

alleles in each cluster, and the percent pairwise diversity scaled by number of segregating sites (Table 1.1).

To further analyze the degree of differentiation between these populations we calculated *F*-statistics (Weir & Cockerham, 1984; Wright, 1951) for individual SNPs (6,152 SNPs) genome-wide implemented in the R package HierFstat (de Meeus & Goudet, 2007). To detect genetic differentiation in individual groups of landraces we used focal comparisons of each population to the overall dataset (Table S1.3).

1.2.3 Admixture inference

Using a Maximum Likelihood approach implemented in TreeMix (Pickrell & Pritchard, 2012), we infer the patterns of population split and mixtures between the six wild barley populations identified in Fang et al. (2014). Populations were identified as Caspian Sea (seven accessions), Central Asian (53 accessions), Northern Levant (42 accessions), Northern Mesopotamia (41 accessions), Southern Levant (107 accessions), and Syrian Desert (34 accessions) (see Table S1 in Fang et al. (2014), for geographic location of WBDC accessions). The TreeMix analysis included all wild populations and the four landrace populations identified here. We ran 25 replications of the tree without bootstrapping, and 25 replications with bootstrapping including five SNPs and 25 SNPs at a time. From this we determine that the wild population from the Caspian Sea is more closely related to landrace populations than to other wild populations (Figure S1.5) thus suggesting the possibility of a more recent introgression with the landraces (Hufford et al., 2013). Including the Caspian Sea wild population in the ancestry analysis of the

landraces results in greater contribution from the Caspian Sea wild population than expected based on historical human migration information (data not shown). Although, the Caspian Sea wild individuals resemble wild barley morphologically (with a shattering inflorescence and extensive branching) other traits such as seed size and erect tillers could suggest either convergent evolution of phenotypes or that the Caspian Sea wild population has been in more recent contact with domesticated material, a result which could potentially bias our inferences of ancestry. Based on this observation the seven individuals from the Caspian Sea wild population are excluded from analysis of population ancestry, retaining 277 wild barley accessions. We note here that the original WBDC encompasses 318 accessions. Fang et al. (2014) identified 30 accessions that appear to be duplicated within the sample or have genotypic composition suggestive of recent introgression. These along with four other samples were removed from the study due to missing latitude and longitude information, resulting in 284 wild barley accessions in our sample.

We utilized a machine learning approach implemented in the software SupportMix (Omberg et al., 2012) to identify the contribution of each of the wild barley populations to individual genomic segments in the landraces. SupportMix can perform admixture analysis by simultaneously analyzing a large number of possible source populations, regardless of their relationship to the focal population and without making assumptions regarding population demographic history or specific population genetic parameters (Omberg et al., 2012). SupportMix is a two-level method. First, it uses a

support vector machine for the classification of the ancestral populations at each genomic segment independent of each other. Once the model is trained to distinguish the source populations it takes one sample at a time (for a specific genomic segment) and assigns it to a putative source population. Second, after all genomic segments are classified for each accession it continues with a smoothing step using a Hidden Markov Model to detect transitions between the different ancestral groups, this approach considers correlations between genetic blocks to limit the effect of regions with poor information content. The wild population with the highest genetic similarity is assigned as the source for that genomic segment and given a probability of assignment to that source population.

The five wild barley populations identified as clearly distinct groups from the landraces are used as potential source populations for the landraces. For this analysis we used 1,896 SNPs found on the iSelect genetic map that were common between the wild barley (SNPs are polymorphic in all 277 wild accessions) and the collection of landraces (803 individuals) (Table S1.2). We ran SupportMix on genomic segments comprised of 50, 75, and 100 SNPs. The wild population with highest similarity is assigned as the source population for that segment. Individual assignment probabilities below 95 % are treated as missing. Inference of admixture using 50 SNP windows results in a large proportion (45 %) of genomic segments with probabilities of assignment below our threshold of 0.95. Thus, SNP windows shorter than 50 SNPs are not used. Increasing the window size to 75 and 100 SNPs results in a higher confidence of ancestry assigned for each genomic region. In these two analyses there were 17.5 % and 19.5 % of the genomic

segments across the seven linkage groups in our sample of landraces with probability of assignment below 0.95, respectively. These segments coincide primarily with the boundaries of linkage groups and are treated as missing data. Therefore we use windows of 75 SNPs. The proportion of ancestry genome-wide is estimated as the percentage of contribution of each wild population to the complete landraces dataset.

The predictive accuracy of SupportMix for genetic assignment of individual genomic segments was evaluated by cross-validation using a subset of wild barley individuals as testing samples, maintaining the remaining wild accessions as the validation sample. The test was run 50 times, sampling four accessions (eight haplotypes) from each wild population per iteration, without replacement. As in the landrace assignment, we used windows of 75 SNPs. An average of 16.4 % of genomic segments per individual could not be assigned with confidence (probability of assignment <0.95) to any population of origin. Window sizes smaller than 75 SNPs resulted in > 80 % of the genome being unassigned (data not shown). In summary, among genomic segments that are assigned with high confidence, 69 % are correctly assigned to the population of origin. A notable exception to assignment of genomic segments of wild individuals back to population of origin occurred in the Northern Levant population, where proportional assignment to the Northern Levant wild population averaged 43 %, with 31 % of segments assigning to the geographically proximate Southern Levant (Figure S1.6 and Table S1.4).

1.2.4 Genetic contribution from proximate wild populations into landraces

We determined the genetic contribution of wild populations in landraces for those growing in the same geographic range as the natural range of wild barley (Table S1.5 and Figure S1.7). East African landraces are outside this range; therefore they were not considered in this analysis. We calculated the great circle distance between each landrace and the nearest wild individual from each wild population using the R package pracma (Borchers, 2015). We then calculated the correlation between distance and the proportion of ancestry assigned in SupportMix.

1.2.5 Private/Shared alleles analysis

Using the 1,896 SNPs in common between landraces and wild barley, we identified alleles private to each of the five wild barley populations using the software SharedPoly from the libsequence library (Thornton, 2003). We found 115, 20, 20, 17, and nine private alleles corresponding to Southern Levant, Northern Levant, Central Asian, Northern Mesopotamia, and Syrian Desert wild barley populations, respectively (Table S1.6). We search for the presence of these SNPs that are private to individual wild populations in each of the landrace populations; this class of variants is referred to as shared alleles and their observed frequency in each landrace population is shown in Figure S1.8 (see also Table S1.7). The estimation of frequency is based on diploid sample size, thus at a given SNP, heterozygous individuals contribute one allele private to the wild population analyzed, and homozygous sites are counted either as zero or two.

1.2.6 Identity by State

An Identity by State (IBS) analysis between the wild and landrace barley lines is used to test for shared genomic segments between populations, consistent with recent introgression. The IBS analysis used PLINK v.1.90 (Chang et al., 2015) with window sizes of 30 SNPs. Larger window sizes resulted in no shared segments between these two datasets. Therefore, we report the results for windows of 30 SNPs. Only segments with 100 % match for the 30 SNPs were considered as significant. There are 37 non-overlapping IBS segments between landraces and wild, with 18 % of wild individuals sharing segments with 36 % of the landraces within each landrace population (Figure S1.9). On average the IBS segments composed of 30 SNPs represent 10.48 cM genomic regions (Figure S1.9 and Table S1.8).

1.3 RESULTS AND DISCUSSION

1.3.1 Population structure and genetic differentiation among barley landraces

To investigate the contribution of wild to domesticated barley we first examined the extent of population structure among landraces using genotyping data from 6,152 SNPs in 803 landrace accessions collected in Europe, Asia, and North Africa (Figure 1.1A, Table S1.1). Population structure was estimated using a Bayesian clustering algorithm implemented in STRUCTURE (Pritchard, Stephens, & Donnelly, 2000). Four major groups of landraces were identified: Coastal Mediterranean, Central European, East African, and Asian (Figure 1.1B, see Table 1 for summary statistics of these populations). The first three groups are nested within a Western primary population (when $K = 2$) while Asian landraces correspond to the Eastern partition (Figure S1.2), similar to the structure reported in previous studies (Morrell & Clegg, 2007; Saisho & Purugganan, 2007; Morrell et al., 2014). The genetic assignment results agree with estimates of the degree of differentiation among landrace individuals by Principal Component Analysis (Figure S1.4), and with the genetic differentiation identified by *F*-statistics (Weir & Cockerham, 1984) (see Table S1.3 for a summary of pairwise F_{ST} comparisons). In summary, western wild barley populations appear to contribute most directly to the genetic constitution of African and European landraces, while eastern wild barley populations made a greater contribution to Asian landraces.

1.3.2 Inference of the genetic contribution of wild populations at specific genomic segments

Beyond evidence for the primary genetic composition and origins of landraces, there are more subtle patterns of genetic exchange. Each of the populations identified in wild barley (Fang et al., 2014) (Figure S1.10) contributes to the genetic composition of the four landrace populations, but this contribution is heterogeneous across genomic segments (Figure 1.2B, Figure S1.11, Table S1.9). This is demonstrated by an analysis of admixture, based on genetic assignment using five of the six wild barley populations as learning samples. These are used to identify the contribution of each wild barley population to individual genomic segments in the landrace populations (see Materials and Methods for the rationale for removing the Caspian Sea wild population from the learning sample). This analysis is based on SupportMix, a tool designed to examine admixture proportions across the genome (Omberg et al., 2012). The analysis is focused on 75 SNP windows because this window size maximized assignment probabilities while permitting the comparison of a large number of genomic segments. Only 17.6 % of genomic segments have a probability of assignment below 0.95, and are marked as missing data (unassigned, Figure S1.12). The genome-wide proportion of ancestry is estimated as the proportional contribution of each wild population to all landraces (Figure 1.2A, Table S1.10). We then estimated the excess or deficit of ancestry (referred to here as Δ ancestry) for each genomic segment in each landrace population. Δ ancestry is the difference between the contributions from each wild population for a particular genomic

segment to the average genome-wide proportion of ancestry derived from that wild population (Figure 1.2B, Figure S1.11). The predictive accuracy of this approach was evaluated by using accessions from the wild barley populations to assign individual genomic segments relative to their known population of origin (cross-validation). This analysis indicates that the power of SupportMix to infer ancestries at any given genomic segment (that is, the potential to accurately assign an individual back to a known population of origin) in our dataset averages 69 % (among genomic segments with probability of assignment 0.95) (Figure S1.6 and Table S1.4). This value, although slightly lower than previously reported values of robustness for estimators of ancestry of genomic segments (approximately 80 %) (Hellenthal et al., 2014), is consistent with the challenge of resolving the contribution of five possible source populations for each genomic segment across all 803 landrace accessions in our sample.

Across all landrace populations, for the fraction that had 0.95 probability of assignment (82.5 %), the largest genome-wide proportion of ancestry derives from the Southern Levant wild population (57 %) (Table S1.10 and Figure S1.13). These results agree with previous archeological and genetic data that identified the Southern Levant (present-day Israel) as the primary contributor to domesticated barley (Lev et al., 2005). Higher assignment to wild barley from the western portion of the range (particularly the Southern Levant) is also expected due to greater representation of SNPs discovered in this region on the genotyping platform (Fang et al., 2014; Russell et al., 2011). Along with the Southern Levant contribution, the genetic composition of landrace populations

reflects an average contribution of 12 % from Northern Levant, 11 % Central Asian, 10 % Northern Mesopotamia, and 9 % Syrian Desert wild populations (Figure 1.2A). Although, the average genome-wide ancestry among landrace populations is similar (Figure 1.2A), the within population variation indicates that the contribution from wild populations differs among individuals in a population (Figure S1.13). Moreover, the genetic composition of landrace populations varies across genomic regions (Figure 1.2B, Figure S1.11). The indication that multiple wild populations contributed to current genetic composition is similar to the patterns observed for domesticated emmer wheat (Civan, Ivanicova, & Brown, 2013). The East African landrace population is inferred to have highly admixed ancestry from multiple wild barley populations (Figure S1.11D). This is consistent with earlier conjecture that barley was imported to Ethiopia from domesticated sources (Zohary, 1970).

1.3.3 There is a higher genetic contribution from proximate wild populations into landraces

There is abundant archeological evidence of human mediated movement and dissemination of cultivated barley beyond the initial range of domestication, beginning approximately 8,000 YBP (Harris & Gosden, 1996; Zohary et al., 2012). In addition, our study shows that landraces frequently carry genomic segments with inferred ancestry that most closely resembles proximate wild populations (Figure 1.2B and Figure S1.11). For example, a higher contribution of proximate wild populations is evident at 13 % (4/29) of

the genomic segments in Asian landraces (Figure 1.2B), with an excess of ancestry derived from the Central Asian wild population compared to the average landrace ancestry genome-wide. The proportional contribution of wild populations to proximate landraces is reflected in greater genome-wide similarity relative to great circle distance from the neighboring wild population (Figure 1.3, Table S1.5 and Figure S1.7). This is evident in a negative correlation (r) between geographic distance and genetic contribution of the Central Asian wild population to the Asian landraces (Figure 1.3, Figure S1.7) (r = -0.47). A similar pattern is observed in all other comparisons between the proportion of ancestry and distance from each wild population. This correlation is consistent with isolation by distance, with r equal to -0.28 and -0.27 for comparisons to Northern Levant and Syrian Desert wild populations, respectively. There is very limited correlation with distance (0.04) for Northern Mesopotamia and Southern Levant populations. Within each landrace population, individual samples have distinct genetic compositions, with some accessions carrying higher proximate wild ancestries than the average in their population. For example, the Northern Mesopotamia wild population contributed 12 % of the genomic segments in Asian landraces (Figure S1.12 and Table S1.10), but variance among individuals results in > 20 % contribution to some individual landrace accessions in this population (Table S1.10).

1.3.4 Private alleles provide direct evidence of contributions of wild populations to landraces

The frequency of SNPs in the landraces that are unique (private) to any of the wild populations (181 SNPs total, Table S1.7) is examined to further delineate the contribution of individual wild populations to the genetic composition of barley. We find 127/161 (79 %) of alleles private to Western (as opposed to Eastern) wild populations present in Asian landraces at an average frequency of 19.5 %. There are 18/20 (90 %) of alleles private to the Eastern wild population present in the Coastal Mediterranean and Central European landrace populations with an average frequency of 24 % (Table S1.6 and Figure S1.8). The larger number and frequency of private alleles from Western wild populations present in Asian landraces are consistent with the genetic assignment analysis reported above. This indicates a greater contribution of Western wild barley to Asian landraces than Eastern wild barley to Coastal Mediterranean and Central European landraces, consistent with previous results based on resequencing (Morrell & Clegg, 2007; Morrell et al., 2014). The private/shared allele comparison also identifies a greater contribution of Southern Levant private alleles to all landrace populations (Table S1.6 and S1.7). The higher contribution from the Southern Levant wild population should be treated as preliminary, as ascertainment bias could influence our observations. Using coalescent simulations, Fang et al. (2013) found that the discovery panel for this set of SNPs is best modeled as derived from eight inbred lines, retaining variants with a minimum of three occurrences in the discovery panel. This accords well with the

discovery scheme reported by Close et al. (2009), which includes an eastern wild barley and Japanese cultivar, but is generally weighted toward European and North American barley cultivars where the genetic composition is contributed largely by western wild populations (Morrell & Clegg, 2007; Morrell et al., 2014). When using private alleles to estimate the contribution from other wild populations, this will have the (conservative) effect of underestimating the contribution of other wild populations to landraces.

There is an uneven genetic representation of wild populations across various landrace populations at specific genomic regions (Figure S1.8), perhaps suggesting that particular adaptations have been combined in landraces from geographically diverse wild populations. This is evident, for example, in the higher frequency (76.6 %) of SNP variant 11_21184 (linkage group 2H) private to Northern Mesopotamia wild populations found in all landrace populations except in the Coastal Mediterranean (Figure S1.8 and Table S1.7). A similar pattern is observed for SNP variant 11_10480 (linkage group 4H) that is private to the Syrian Desert wild population, but is found in high frequency (81.3 %) in all landrace populations except for the Asian population (Figure S1.8 and Table S1.7).

The increased genetic resemblance between landraces and proximate wild populations indicates the potential adaptive nature of alleles found in genomic segments with higher positive Δ ancestry, or high frequency private alleles. For example, Δ ancestry values indicate regions on linkage groups 1H, 2H, and 5H in the Asian landraces (Figure 1.3B, Figure S1.11B) that have an elevated contribution from the Central Asian

wild population. Although, this excess cannot be explained solely by the presence of Central Asian private alleles, there is one Central Asian wild private SNP variant 11_21286 (linkage group 2H) at 63 % frequency in Asian landraces and virtually absent or in low frequency in the Coastal Mediterranean, Central European, and East African populations (Figure S1.8 and Table S1.7). Likewise, Coastal Mediterranean, Central European, and East African populations have a higher proportion of Northern Mesopotamia ancestry (Δ ancestry) at two genomic segments at linkage group 4H (Figure S1.11). We identify two SNP variants private to Northern Mesopotamia (11_10756 and 12_30136) at high frequency (77 %) in the middle segment on linkage group 4H (Table S1.7). The private/shared alleles analysis also confirms the admixed nature of the East African population, yet with a larger contribution from Western wild populations. The East African population includes 12 private alleles derived from Eastern wild populations (33 % frequency) and 103 private alleles from Western wild populations (44 % frequency) (Table S1.6).

1.3.5 Similarity between wild and landrace populations cannot be explained by recent introgression

An alternative hypothesis for the mosaic ancestry of landraces involves recent or ongoing introgression from proximate wild populations (Morrell & Clegg, 2007). Population genetic effects of recent introgression include large chromosomal regions in (admixture) linkage disequilibrium (LD) (Briscoe, Stephens, & O'Brien, 1994; Chakraborty & Smouse, 1988; Chakraborty & Weiss, 1988) or extended genomic tracts

of shared ancestry (Gusev et al., 2009; Gusev et al., 2012). Admixture LD breaks down quickly in outcrossing species, but should be more readily detectable in self-fertilizing species such as barley. The estimated rate of outcrossing 1.8 %, averaged across samples of wild and cultivated barley (Abdel-Ghani, Parzies, Omary, & Geiger, 2004) should greatly reduce the rate of effective recombination (see reference Nordborg, 2000), dramatically increasing the number of generations for the decay of admixture LD. An analysis of identity by state (IBS) among the landraces and wild barley populations conditioning on a complete match over 30 SNPs (which constitutes approximately 1/15 of the SNPs per linkage group) identifies 37 non-overlapping IBS segments (Figure S1.9). Only 18 % of wild and 36 % of landrace individuals contribute to this perfect-match IBS (Table S1.8) whereas differential ancestry for individual genomic segments can involve >80 % of landraces (for example, from Northern Mesopotamia wild population in the first genomic segment in linkage group 4H in Asian landraces; Figure S1.12 B). Some degree of IBS is expected among distantly related individuals from distinct populations, owing to the expectations of deep patterns of shared descent within a species (Ralph & Coop, 2013). IBS comparisons fail to identify large shared segments (constituting half or one-quarter of linkage groups), as expected after introgression (Chakraborty & Weiss, 1988). The low levels of genome-wide LD (Figure S1.1) and small blocks of IBS (average 10.5 cM) suggest that contributions from wild populations into the cultigen are not recent and in some cases may date back to early in the history of

widespread barley cultivation which started around 8,500 YBP (Diamond, 2002; Harlan & Zohary, 1966; Zohary et al., 2012; Willcox, 2005).

1.4 CONCLUSIONS

In summary, the genetic composition of barley landraces indicates a genetic contribution from multiple wild progenitor populations that in turn must reflect the pattern of initial domestication and later patterns of trade and migration of early agriculturalists along the axes of Europe, Africa, and Asia. Although multiple populations contribute to the genetic composition of the cultigen, the contribution from the broad geographic range of wild barley populations also varies across the genome as well as across landrace populations. The clear contribution from proximate wild populations, at specific genomic regions, raises the intriguing possibility of adaptive contributions based on regional and local environments.

Table 1.1 Summary statistics.
Results for four landrace populations, based on 6,152 SNPs. Values for sample size, number of segregating sites, number of private alleles, and percent pairwise diversity scaled by number of segregating sites are reported

Landraces	Sample Size	Segregating sites	Private alleles	Pairwise diversity
Central European	210	6004	70	0.337
Asian	279	5541	26	0.268
Coastal Mediterranean	228	5950	40	0.309
East African	86	4298	3	0.210

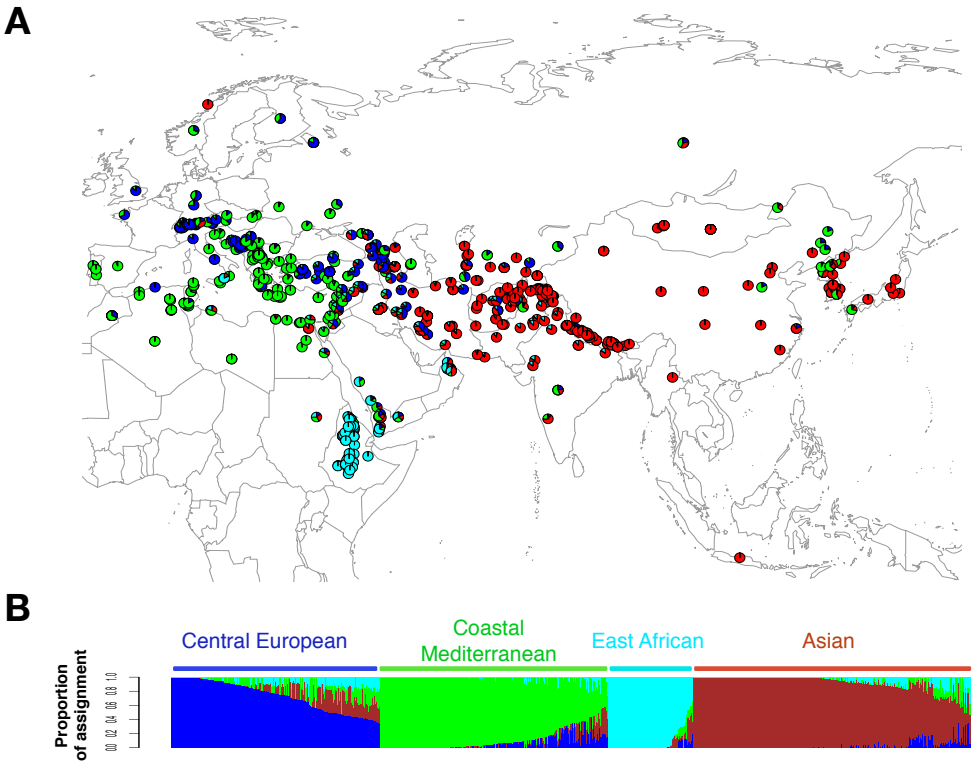


Figure 1.1 Distribution and populations structure of the barley landraces

Samples from the Old World used in this study. **A**. Colors correspond to the proportion of assignment to each of the four populations identified among landraces. **B.** Secondary population subdivision for the optimal K = 4: Central European, Mediterranean, East African, and Asian. Each color represents the majority of assignment for four landrace populations. The Y-axis is percent composition with samples sorted along the X-axis geographically by longitude from west to east.

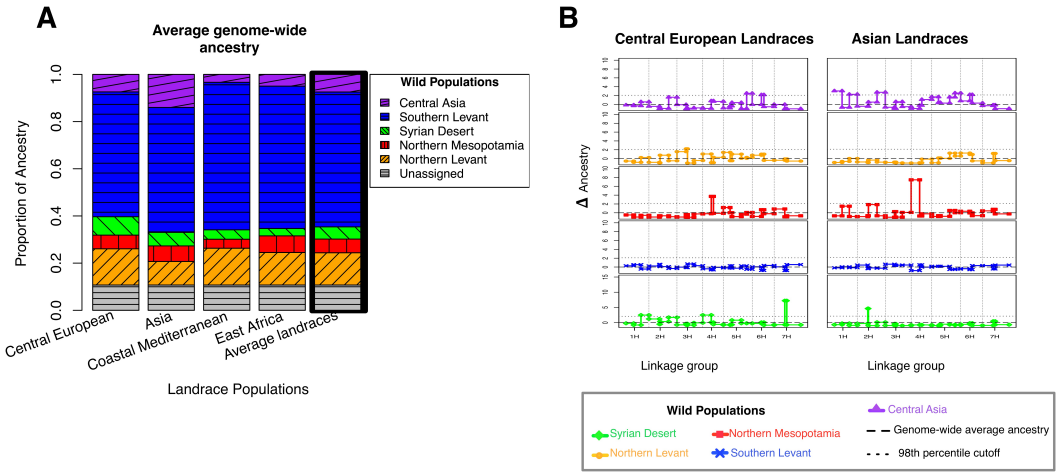


Figure 1.2 Proportions of genetic ancestry

A. Genome-wide proportion of ancestry for each landrace population. Colors represent the population of origin from the wild (Unassigned sites are not considered). The barplot on the far right (inside a black box) represents the average genome-wide ancestry among all four landrace populations, used also as a base line for panel **B** (average contribution). **B** Excess or deficit of ancestry (Δ ancestry) for the Central European and Asian landrace populations. Δ ancestry is measured as the deviation from the average contribution of each wild population at the genome-wide level (that is, how many times more/less of that ancestry is observed at each genomic segment with respect to genome-wide ancestry proportion) (black dashed line). Each line corresponds to one of the five populations identified in wild barley. Positive values indicate X times of excess ancestry and negative values a deficit of ancestry with respect to the genome-wide average ancestry of a particular wild population. The dotted horizontal line indicates the 98th percentile cutoff from the distribution of excess or deficit of each wild population across all genomic segments at each landrace population

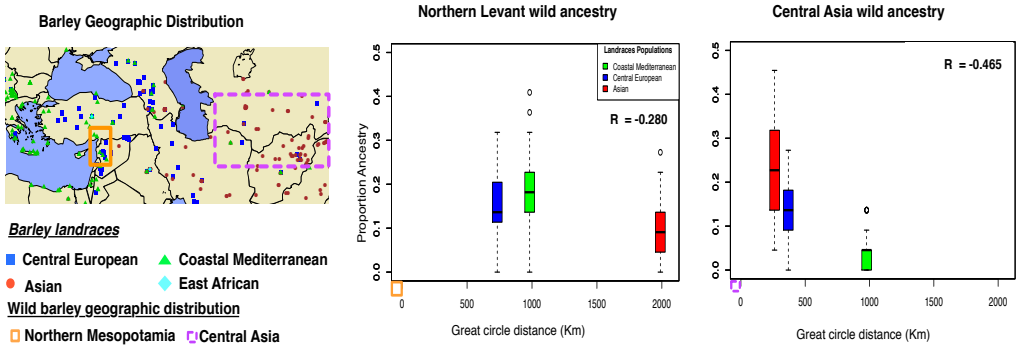


Figure 1.3 Genome-wide ancestry as a function of distance from wild populations

The map on the left indicates the distribution of landraces sympatric with populations of wild barley. The orange and purple boxes represent the geographic distribution of Northern Levant and Central Asian wild populations. The panels on the right indicate the distribution of proportion of ancestry (Y-axis) in each of the landrace populations as a function of distance (X-axis) from the source wild population with the closest proximity. The boxplot for each landrace population was placed at the population's median value of the great circle distance between each landrace individual to the closest wild barley individual in the wild barley population analyzed (depicted at coordinates 0,0). The correlation (*r*) between distance and proportion of ancestry is indicated in each comparison. East African landraces are not included in the depiction due to small sample size (two individuals) in the geographic range analyzed

1.5 SUPPORTING INFORMATION

1.5.1 Supplementary Figures

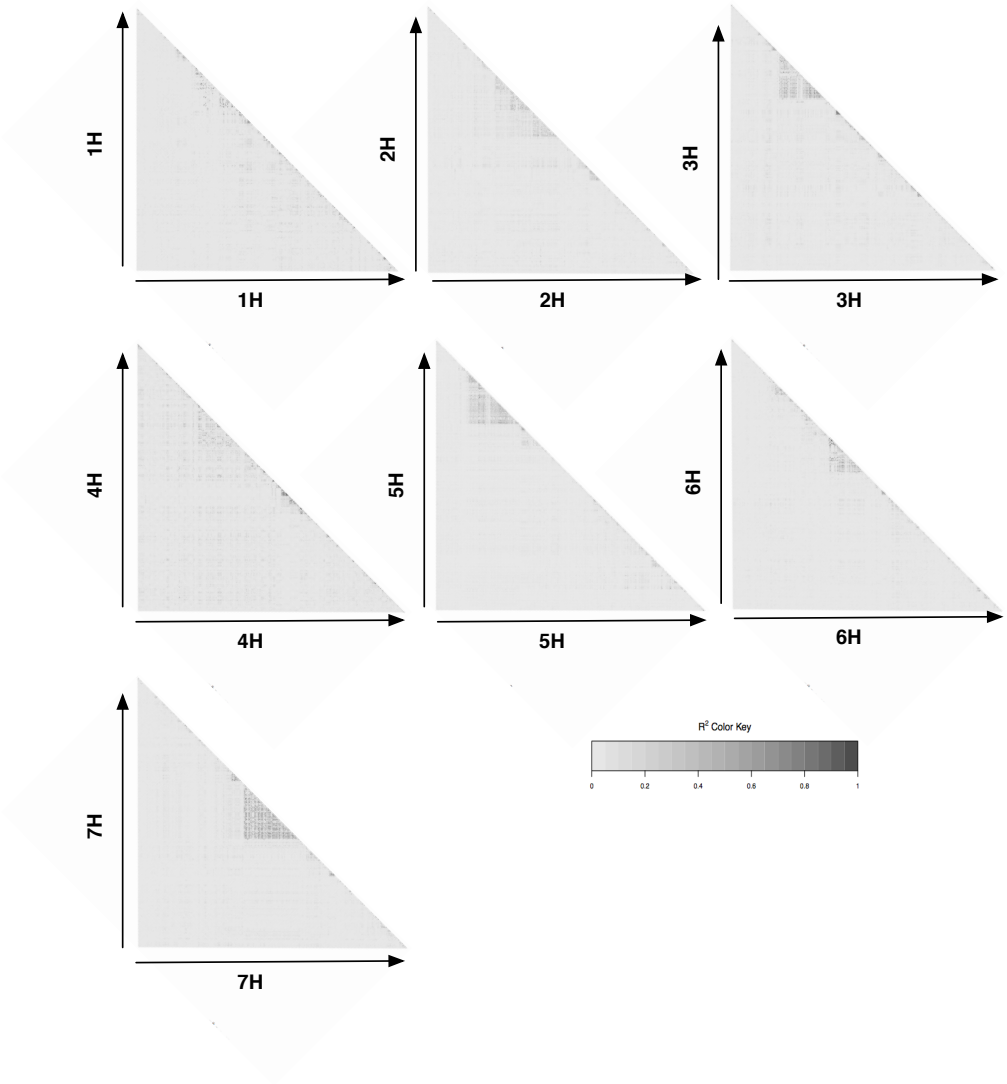


Figure S 1.1 Linkage disequilibrium (r^2).

Linkage disequilibrium determined by a pairwise comparison of the SNPs in each linkage group in the landraces.

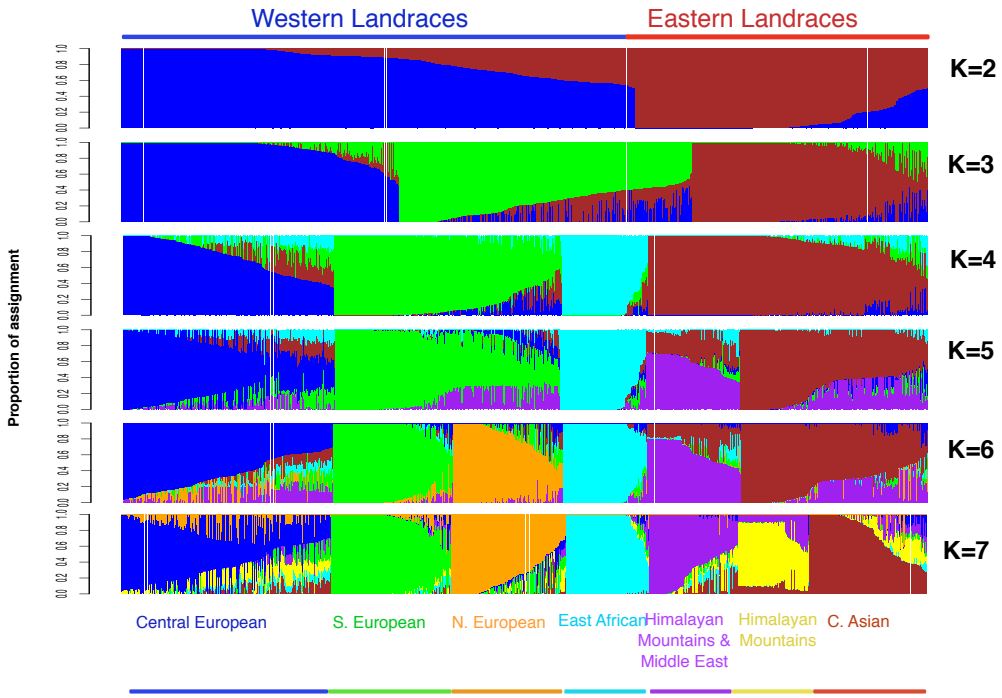
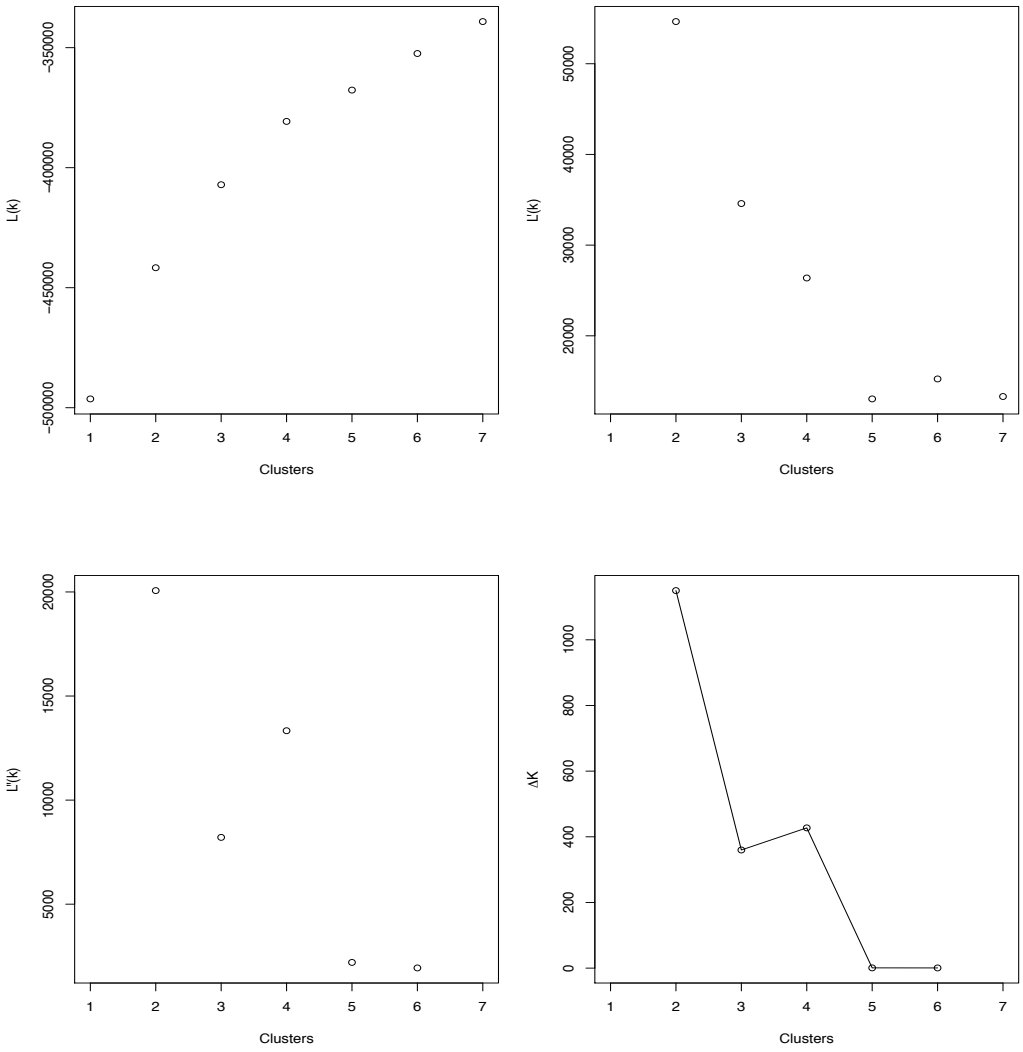


Figure S 1.2 Population structure of barley landraces.

All clusters from *K* = 2 to 7: Central European, Southern European, Northern European, East African, the Himalayan Mountains, Himalayan Mountains and Middle Eastern, and Central Asian. The Y-axis is percent composition and the X-axis displays accessions sorted geographically from west to east.

A


L(k)
−350000
−400000
−450000
−500000
Clusters
L'(k)
50000
40000
30000
20000
Clusters
L''(k)
20000
15000
10000
5000
Clusters
ΔK
1000
800
600
400
200
0
Clusters

B

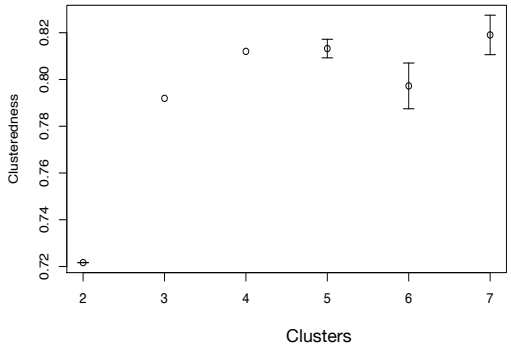

Clusteredness
0.82
0.80
0.78
0.76
0.74
0.72
Clusters

Figure S 1.3 Identification of the optimal number of groups *K*.

(A) ΔK, description of the four steps to determine the number of clusters that best explain the population structure among the landraces; **(B)** Clusterdness, the extent to which individuals were estimated to belong to a single cluster rather than to a combination of clusters.

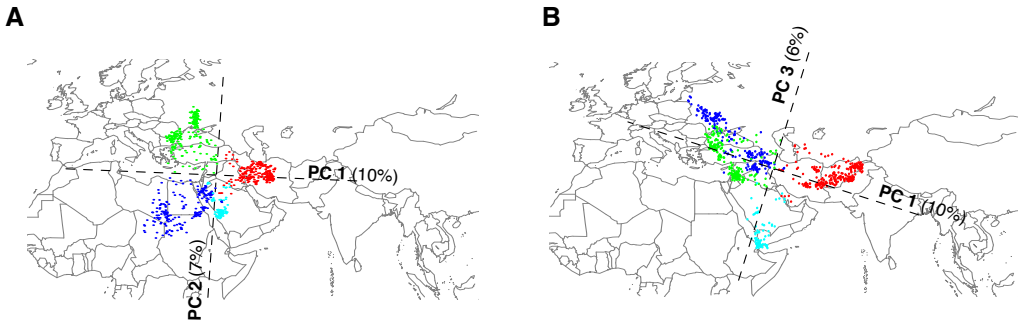


Figure S 1.4 Relationship of barley landrace accessions based on principal components.

(A) Principal Component Analysis transformation of the genetic variation in barley landraces. Compares projected locations to sample localities as depicted in Figure 1, by rotating PC1 versus PC2 93° clockwise. **(B)** Principal Component Analysis transformation of the genetic variation in barley landraces. Compares projected locations to sample localities as depicted in Figure 1, by rotating PC1 versus PC3 70° clockwise. This comparison results in a greater separation of the East African population from other landrace populations.

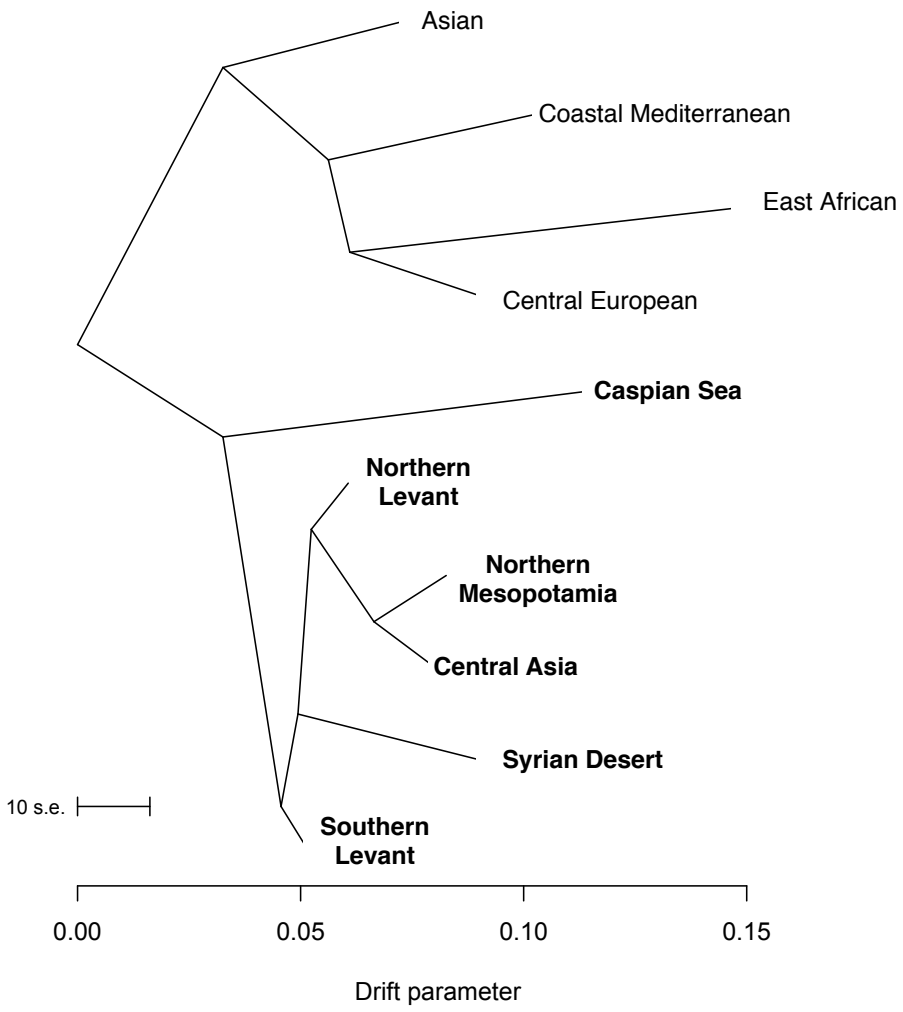


Figure S 1.5 Maximum Likelihood tree.

Tree among wild (bold font) and barley landraces as inferred by TreeMix.

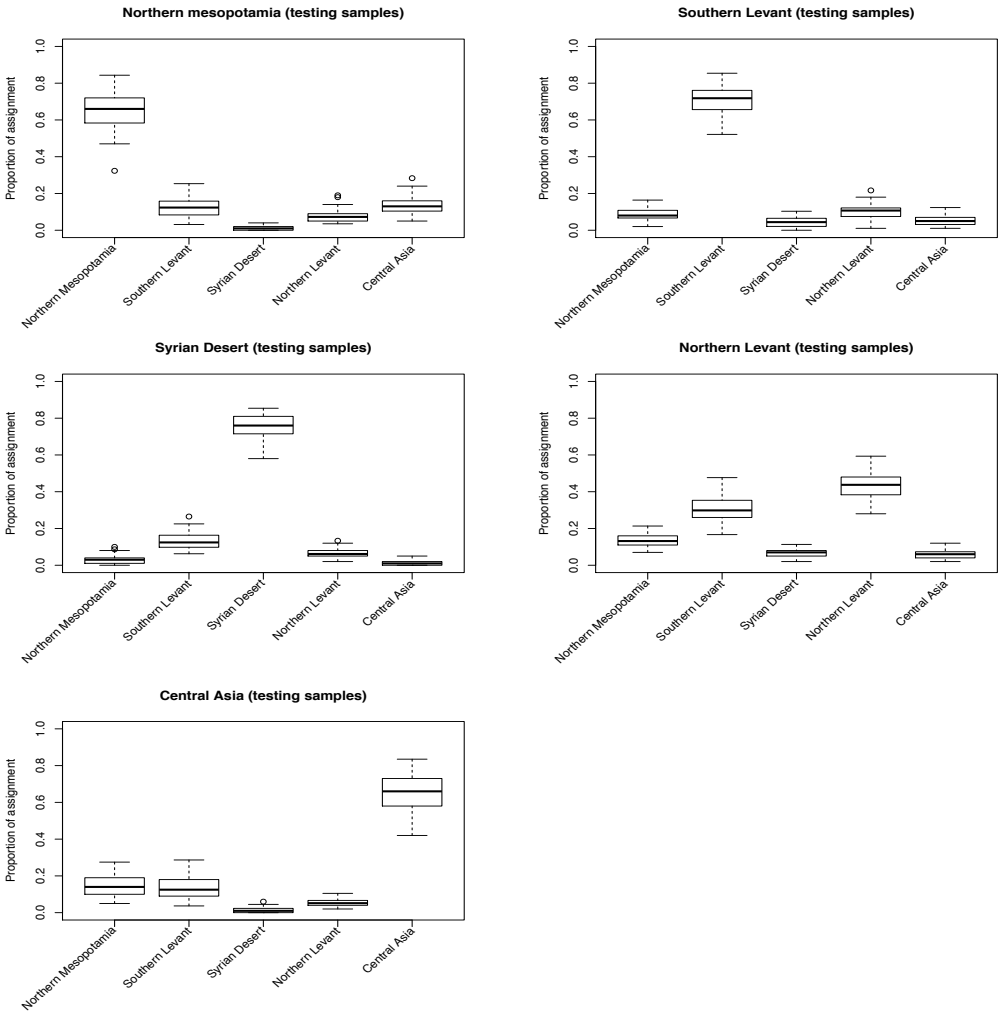


Figure S 1.6 Predictive accuracy of SupportMix by cross-validation.

Each panel represents the average proportion of ancestry assigned to individuals from a wild population used as a test dataset compared to a training dataset composed of all remaining wild barley individuals. The analysis was run 50 times for four individuals from each wild population (proportions represent only sites with assigned ancestry).

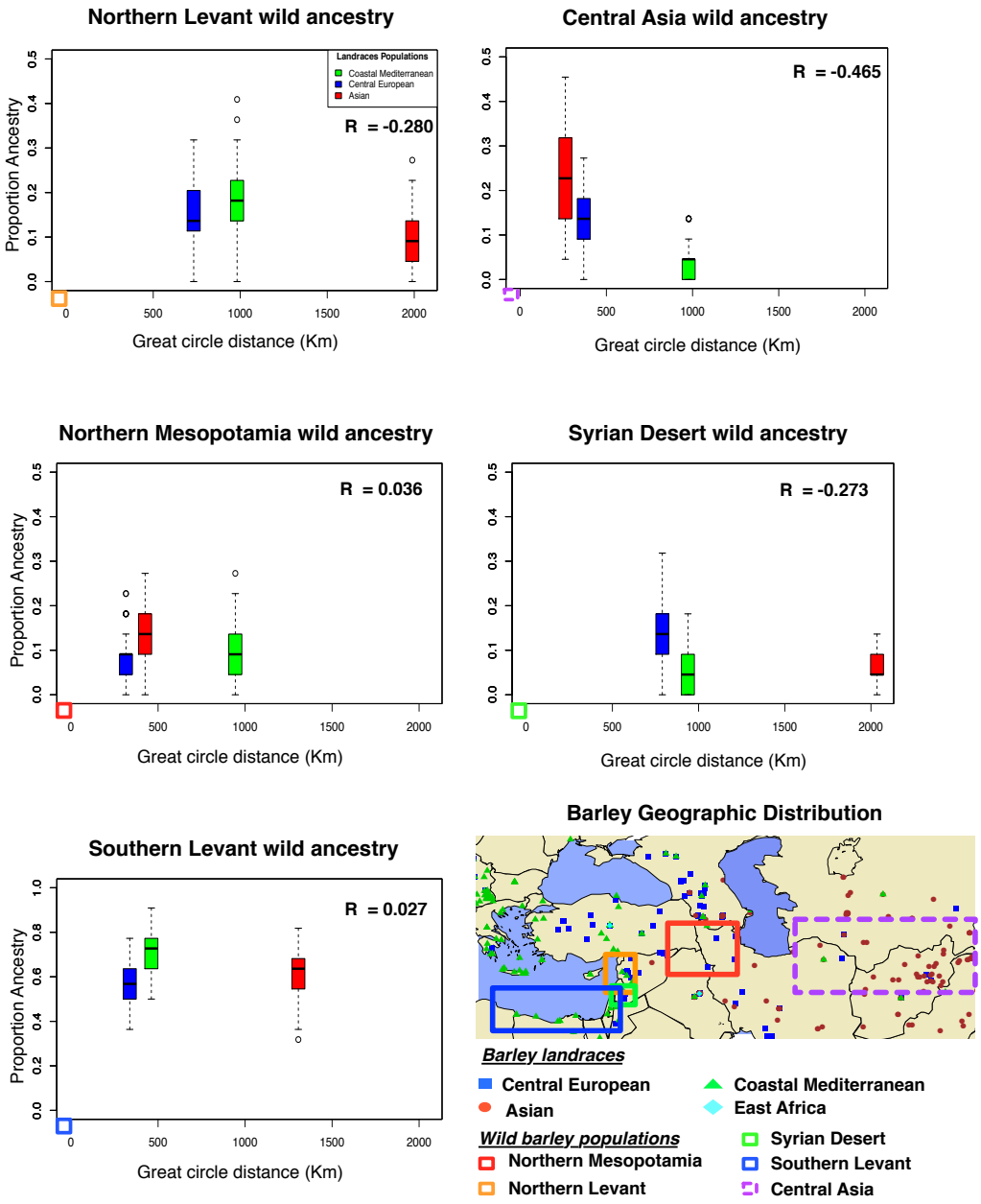


Figure S 1.7 Genome-wide ancestry as a function of distance from wild populations.

The map on the bottom right shows the distribution of landraces sampled from within the natural range of wild barley. The boxes represent the geographic distribution of Southern Levant, Northern Mesopotamia, Syrian Desert, Northern Levant, and Central Asian wild populations. The other panels indicate the distribution of proportion of ancestry (Y-axis) in each of the landrace populations as a function of distance (X-axis) from the ancestral wild population. The boxplots for each landrace population are at the median of the distribution

of distances calculated for each landrace and the closest wild accession (depicted at coordinates 0,0). The correlation (*r*) between distance and proportion of ancestry is indicated in each comparison. East African landraces are not included in the depiction due to small sample size (two individuals) in the geographic range analyzed.

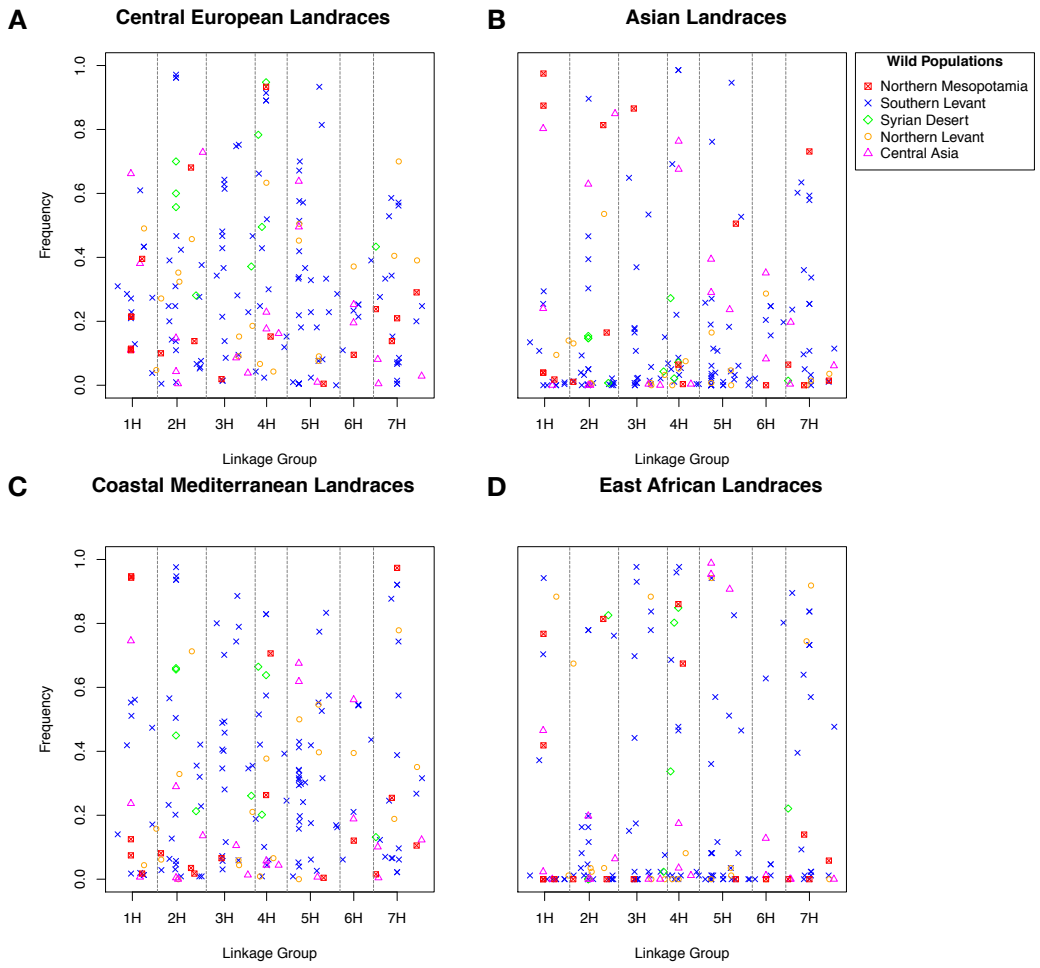


Figure S 1.8 Frequency of alleles private to the wild populations present in each of the landrace populations.

Linkage groups are separated by gray dashes.

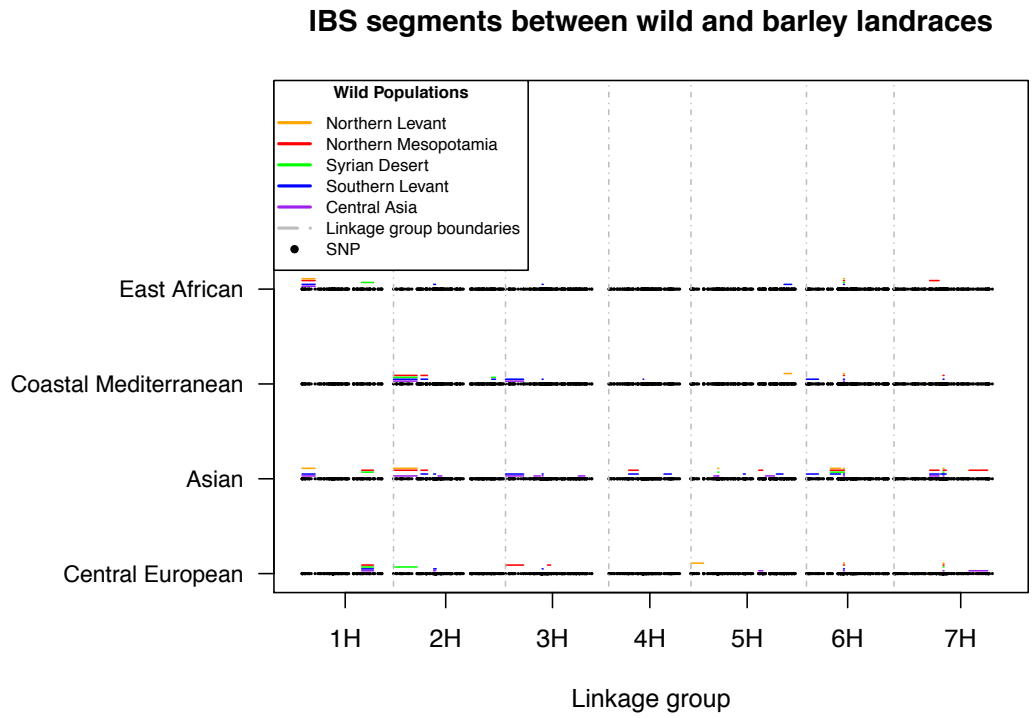


Figure S 1.9 Identical by State segments between wild and cultivated barley.

Black dots represent SNPs in each landrace population. The x-axis is the genomic location of each SNP. The vertical gray dashed lines define the limits between linkage groups. The lines represent the location and extend of IBS between each wild and landrace population. Each segment is 30 SNPs long.

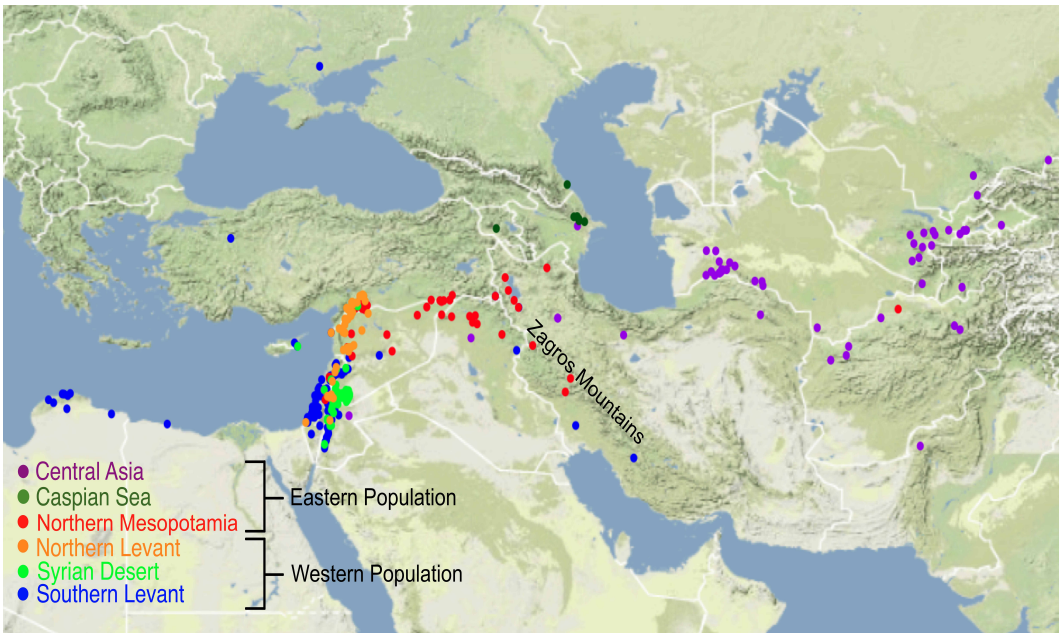


Figure S 1.10 Population structure in wild barley.

Each of the six colors represents one of the six subpopulations. Three different subpopulations are nested in the Eastern and Western populations, respectively. This figure has been reproduced from Fang et al. (2014).

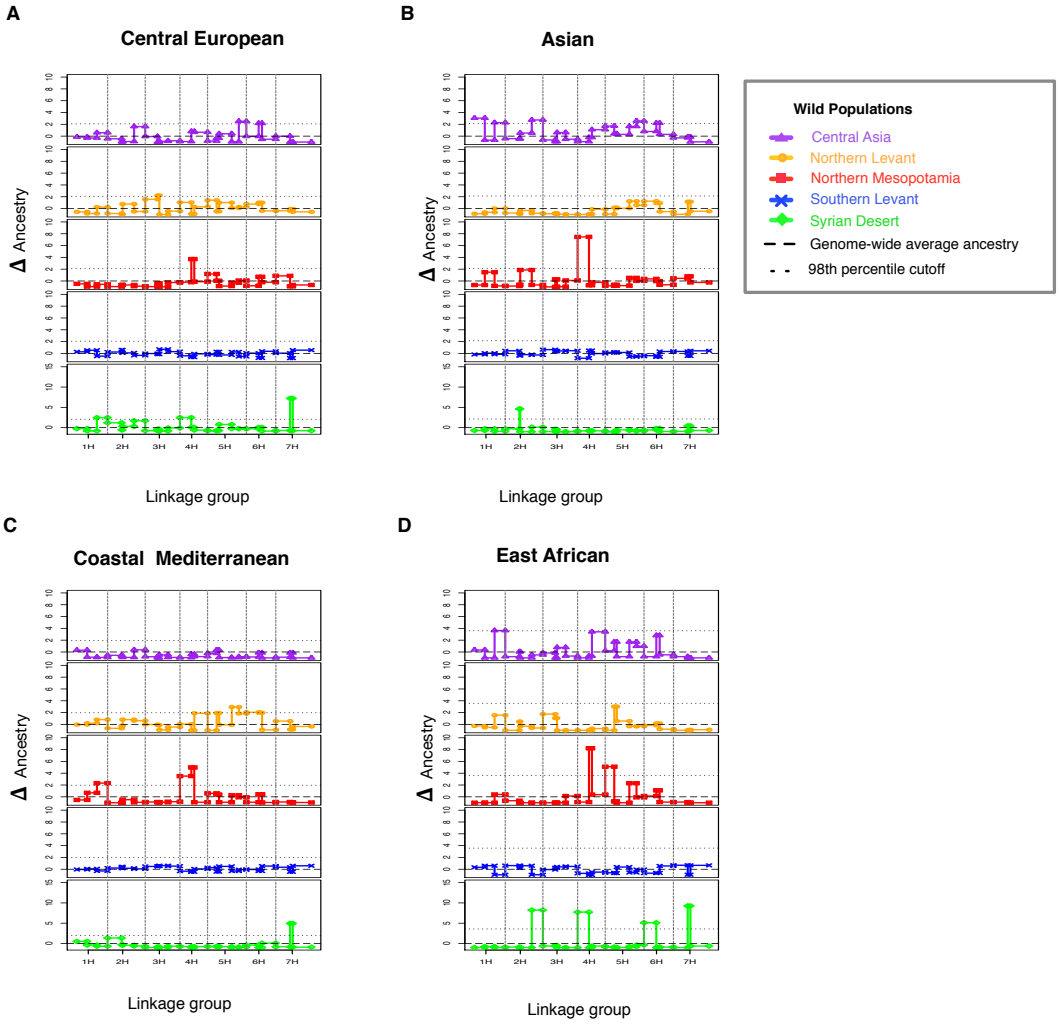


Figure S 1.11 Excess or deficit of ancestry for barley landrace populations.

Excess or deficit (Δ ancestry) measured as the deviation from average contribution of each wild population from average genome-wide contributions (black dashed line). Each panel corresponds to the five populations identified in wild barley (Figure S1.10). Positive values indicate an excess and negative values a deficit of ancestry from a particular wild population. The dotted horizontal line indicates the 98th percentile cutoff from the distribution of excess or deficit of each wild population across all genomic segments for each landrace population.

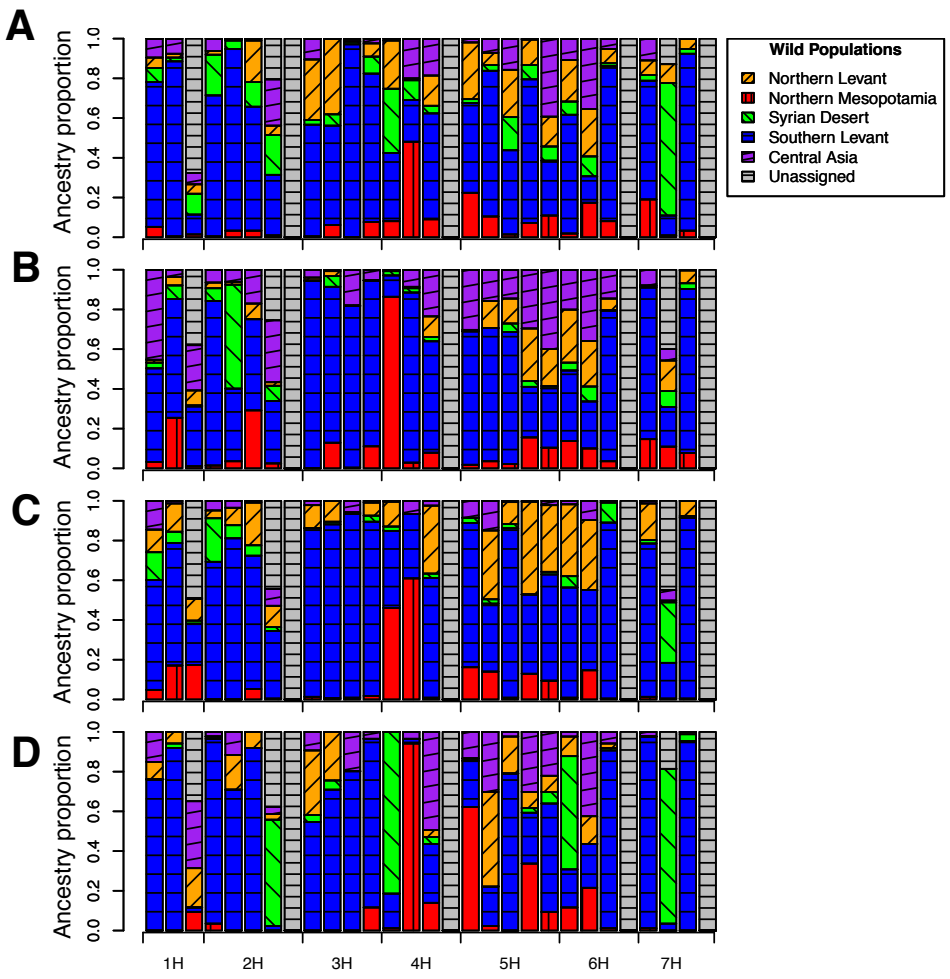


Figure S 1.12 Proportion of ancestry in barley landrace populations at each genomic segment.

Ancestry proportions include unassigned sites. **(A)** Central European landrace population, **(B)** Asian landrace population, **(C)** Coastal Mediterranean landrace population, **(D)** East African landrace population. The tick marks on the x-axis in panel **D** indicate the linkage group boundaries.

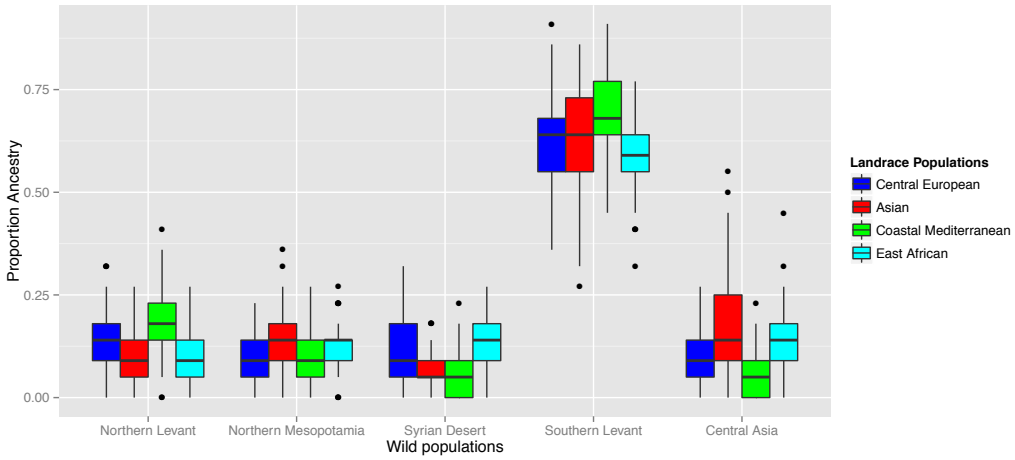


Figure S 1.13 Distribution of the genome-wide proportion of ancestry from wild to landrace barley populations.

Unassigned genomic regions are not considered. The boxplots are order (from left to right) according to the legend.

1.5.2 Supplementary Tables

Table S 1.1 803 landrace accessions used in this study with latitude and longitude information.
Attached as Supplemental Material.

Table S 1.2 1,896 SNPs shared between wild barley and landrace populations.
Attached as Supplemental Material.

Table S 1.3 Median and maximum Focal F_{ST} values from comparisons of each landrace population to all the other landraces.

	Median F_{ST}	Maximum F_{ST}
Central European	0.064	0.807
Asian	0.102	0.879
Coastal Mediterranean	0.08	0.838
East African	0.121	0.963

Table S 1.4 Predictive accuracy of SupportMix by cross-validation.
Average across 50 runs of the genome-wide proportion of ancestry in subsets of wild barley analyzed as testing samples, using the remaining wild individuals as the validation data set.

	Genetic Assignment: Genome-wide ancestry				
Testing Samples	**Northern Mesopotamia**	**Southern Levant**	**Syrian Desert**	**Northern Levant**	**Central Asia**
Northern Mesopotamia	0.64	0.13	0.01	0.08	0.14
Southern Levant	0.09	0.71	0.05	0.1	0.05
Syrian Desert	0.03	0.13	0.76	0.07	0.02
Northern Levant	0.14	0.31	0.07	0.43	0.06
Central Asia	0.15	0.14	0.01	0.05	0.65

Table S 1.5 Genome-wide ancestry as a function of distance from wild populations.
Proportions of ancestry for individual landraces, and the great circle distance between each individual and the closest accession from each wild population. Attached as Supplemental Material.

Table S 1.6 Summary of number of private alleles from wild barley populations present in the landraces.
Average allele frequency of alleles private to wild populations in the landraces (in parenthesis). The number of private alleles at each wild population is shown in brackets.

	Wild barley populations				
Landrace Populations	**Northern Mesopotamia [17]**	**Southern Levant [115]**	**Syrian Desert [9]**	**Northern Levant [20]**	**Central Asia [20]**
Central European	16 (0.25)	106 (0.34)	9 (0.57)	20 (0.30)	18 (0.25)
Asian	14 (0.37)	89 (0.20)	9 (0.10)	15 (0.11)	13 (0.46)
Coastal Mediterranean	16 (0.30)	115 (0.34)	9 (0.43)	18 (0.29)	17 (0.23)
East African	7 (0.53)	77 (0.34)	6 (0.51)	13 (0.40)	11 (0.33)

Table S 1.7 Frequency of alleles private to the wild present in each of the landrace populations.
Private SNPs in wild barley that are present in the landraces, including linkage group and their frequency in each landrace population. Attached as Supplemental Material

Table S 1.8 Proportion of individuals involved in IBS.
IBS segments (30 SNP each) between each landrace population and the 277 wild barley lines

	No. Landraces	**Wild (%)**	**Landraces (%)**
Central European	210	0.14	0.30
Asian	279	0.36	0.62
Coastal Mediterranean	228	0.12	0.24
East African	86	0.09	0.28
Average	201	0.18	0.36

Table S 1.9 Chromosome painting. Individual landrace ancestry inferred at each genomic region.
The cells are colored according to their inferred ancestry from the wild populations. Two haplotypes (rows) per landrace accession are depicted. Attached as Supplemental Material.

Table S 1.10 Genome-wide proportion of ancestry among each landrace population.
Genome-wide average proportions of genetic ancestry in barley landrace populations for all regions that have a probability of assignment >95%.

	Genetic Assignment: Genome-wide Ancestry				
Landraces	**Northern Mesopotamia**	**Southern Levant**	**Syrian Desert**	**Northern Levant**	**Central Asia**
Central European	0.08	0.57	0.12	0.13	0.1
Asian	0.12	0.57	0.05	0.08	0.17
Coastal Mediterranean	0.1	0.64	0.06	0.16	0.04
East African	0.11	0.52	0.14	0.1	0.13
Average	0.1	0.57	0.09	0.12	0.11

2 CHAPTER 2

THE EFFECTS OF BOTH RECENT AND LONG-TERM SELECTION AND GENETIC DRIFT ARE READILY EVIDENT IN NORTH AMERICAN BARLEY BREEDING POPULATIONS

Barley was introduced to North America ~400 years ago but adaptation to modern production environments is more recent. Comparisons of allele frequencies among different growth habits and inflorescence types in North America indicate significant genetic differentiation has accumulated in a relatively short evolutionary time span. Allele frequency differentiation is greatest among barley with two-row versus six-row inflorescences, and then by spring versus winter growth habit. Large changes in allele frequency among breeding programs suggest a major contribution of genetic drift and linked selection on genetic variation. Despite this, comparisons of 3,613 modern North American cultivated breeding lines that differ for row type and growth habit permit the discovery of 183 SNP outliers putatively linked to targets of selection. For example, SNPs within the *Cbf4, Ppd-H1*, and *Vrn-H1* loci which have previously been associated with agronomically-adaptive phenotypes, are identified as outliers. Analysis of extended haplotype-sharing identifies genomic regions shared within and among breeding programs, suggestive of a number of genomic regions subject to recent selection. Finally, we are able to identify recent bouts of gene flow between breeding programs that could point to the sharing of agronomically-adaptive variation. These results are supported by pedigrees and breeders understanding of germplasm sharing.

2.1 INTRODUCTION

The evolution of breeding populations encompasses processes that reasonably mimic the evolution of natural populations, but in accelerated time (Ross-Ibarra, Morrell, & Gaut, 2007). Comparative population genetic approaches can be used for the identification of genes underlying adaptive variation and to understand the effects of demographic patterns on diversity without specific phenotypic information (Ross-Ibarra et al., 2007; Nielsen, Hellmann, Hubisz, Bustamante, & Clark, 2007).

Initiation and evolution of breeding populations may involve episodes such as founder events (Martin, Blake, & Hockett, 1991), bottlenecks, and gene flow from other breeding programs or exotic sources. These demographic effects can contribute to differences in allele frequency among breeding programs due to genetic drift, local adaptation, and selection (or linked selection). Understanding the selection and demographic history, including the effects of genetic drift and migration in breeding populations can accelerate crop improvement (Ross-Ibarra et al., 2007), for example by the identification of loci involved in domestication and improvement (e.g., Cavanagh et al., 2013); identification of introgression between domesticates and wild relatives (e.g., Hufford et al., 2013); and the determination of specific donor individuals contributing, for example, disease resistance variants (e.g., Fang et al., 2013).

History of selection can be investigated by the identification of large allele frequency differences among populations (at individual loci) (Cavalli-Sforza, 1966). Allele frequency differences between subdivided populations can be measured by

fixation indices or *F*-statistics such as $F_{ST,}$ a measure of differentiation in allele frequencies in sub-populations relative to the total population (Lewontin & Krakauer, 1973). The identification of loci with large differences in allele frequency among populations (as identified from F_{ST} outliers) works especially well for older or longer-term selective events (Nielsen et al., 2007).

Patterns of more recent selection (Pritchard, Pickrell, & Coop, 2010; Horton et al., 2012) can be identified through approaches that detect a high degree of haplotype sharing between individuals. These approaches can identify specific haplotypes (and their underlying sequence variants) subject to selection based on their relative frequencies in a population (Innan, Zhang, Marjoram, Tavaré, & Rosenberg, 2005; Hudson, Bailey, Skarecky, Kwiatowski, & Ayala, 1994).

Migration between populations can contribute to adaptive variation (Slatkin, 1987). Identity by state (IBS) analysis is a sensitive approach for the identification of specific genomic regions involved in migration between distinct populations. IBS analyses have been used in studies of barley and maize to identify specific genomic regions involved in adaptation to new environments and disease resistance genes derived from exotic germplasm (Fang et al., 2013; Hufford et al., 2013).

In the present study we use single nucleotide polymorphism (SNP) genotype data from the Barley Coordinated Agricultural Project to investigate the breeding history of North American barley breeding populations. Barley (*Hordeum vulgare* ssp. *vulgare*) was introduced to North America by European colonists as early as 1602 as a crop essential

for beer production (Weaver, 1950). Early successes in barley production in North America involved introduction of varieties adapted to similar environmental conditions. In the eastern growing region of the United States barley was introduced from northern Europe and England malting varieties; and in the western growing region from Mediterranean feed varieties (Weaver, 1943). Notable exceptions from outside northern Europe included "Manchuria" from northeastern China which was well adapted to the Upper Midwest growing environment, Stavropol from Russia was the most important variety in the Lower Midwest (especially Kansas), and "Trebi" from the southern shores of the Black Sea was found to be particularly productive in western regions under irrigated conditions (Wiebe & Reid, 1961). Despite the use of varieties from multiple Old World sources, the set of founder varieties was relatively narrow, which likely contributed to reduced diversity in modern North American cultivars (Martin et al., 1991).

Barley production is divided into spring and winter growth habit. Spring barley production dominates in North America, while winter barley is grown in more southerly latitudes and moderate coastal climates (Weaver, 1950). Spring and winter barley have been bred separately by breeding programs located in different geographic regions. Further trait requirements by end use markets (i.e., malting, feed, and food) and the establishment of breeding programs for two-row and six-row inflorescence type, contribute to highly structured barley populations (Cuesta-Marcos et al., 2010; Hamblin et al., 2010; Wang et al., 2012; Zhou, Muehlbauer, & Steffenson, 2012).

Our analysis focuses on four major questions. First, which of the factors, including breeding programs, growth habit, and row-type contribute most directly to genetic differentiation among samples? Second, to what extent are loci identified as major contributors to phenotypic variance in Old World barley contributing to allele frequency differences in North American breeding populations? Third, can we identify evidence of recent or long-term selection acting on breeding populations and what loci are involved? Fourth, how have patterns of shared ancestry and migration contributed to diversity and relatedness in current breeding populations?

2.2 MATERIALS AND METHODS

2.2.1 Plant Materials

Genotypic data for a total of 3,971 barley accessions including varieties, advanced lines, and genetic stocks from barley breeding programs that participated in the Barley Coordinated Agricultural Project (CAP) were downloaded from The Triticeae Toolbox (T3) (http://triticeaetoolbox.org/). Barley samples are representatives of four years of germplasm enhancement (2006-2009) from 10 breeding programs. These breeding programs include most barley growing regions and market end-uses in the United States. For more information about the programs (see Hamblin et al., 2010; Wang et al., 2012).

Breeding two-row and six-row barley within a single program often involves different objectives, thus we consider these types as independent populations resulting in 16 breeding populations (see Table 2.1). Hereafter, the following notation to refer to each

breeding population was used: “breeding program’s name abbreviation” followed by 2 or 6 for two-row and six-row, respectively as described in Table 2.1. The Bush Agricultural lines from the international program, referred here as BAI2, were separated from the North American two-row lines, referred here as BA2. Analyses included a total of 16 populations (Table 2.1). There were 10 six-row accessions from Oregon mislabeled in T3 as two-row (Dr. Patrick Hayes, personal communication), we used the corrected row-type (see Supplemental Information for more details). The current sample is divided hierarchically at the highest level into spring and winter growth habits and then within each growth habit by breeding programs (Figure 2.1).

2.2.2 Genotyping

All 3,971 accessions were genotyped with 2,882 Barley Oligo Pooled Assay Single Nucleotide Polymorphism, referred to as BOPA SNPs (Close et al., 2009) using Illumina GoldenGate Technology (Illumina, San Diego, CA); genotypic data was downloaded from http://triticeaetoolbox.org. BOPA SNPs were identified primarily from resequencing of expressed sequenced tags (ESTs) (Close et al., 2009). SNP order along each linkage group is based on the consensus genetic map (Muñoz-Amatriaín et al., 2014). Sampled accessions were self-fertilized to at least the F_4 generation before genotyping (Wang et al., 2012; Hamblin et al., 2010), resulting in average expected heterozygosity of 6.25%. Genotyping information was downloaded from (http://triticeaetoolbox.org/) with all filters set to zero.

SNP annotations and metadata information, including gene names and whether SNPs occur in genic or non-genic regions, were obtained using SNPMeta (Kono, Seth, & Poland, 2013).

2.2.3 Quality Control

The dataset was filtered for monomorphic SNPs, and SNPs or accessions with more than 25% missing data. Additionally, we removed accessions with incomplete sample information (i.e., row type or growth habit) and accessions where single lines represented a population. We removed accessions that presented > 6.25% heterozygosity. Finally, we removed genetic stocks or near-isogenic lines such that one accession from each near-isogenic line set was retained in the dataset.

2.2.4 Summary Statistics

Basic descriptive statistics were calculated for each of the 16 breeding populations (Table 2.1). The degree of inbreeding was estimated by the inbreeding coefficient F_{IS} ($1 - H_o/H_e$) using a custom R script. The similarity between samples was estimated by the percent pairwise diversity calculated using the *compute* program from the libsequence library (Thornton, 2003). Monomorphic SNPs in each population were excluded and heterozygous and ambiguous calls were treated as missing data. The SharedPoly program from libsequence (Thornton, 2003) was used to count the number of private SNPs in each breeding population.

2.2.5 Derived Site Frequency Spectrum (SFS)

To infer ancestral state, SNP states from *Hordeum bulbosum* as reported for BOPA SNPs were used (Fang et al., 2014). This involved alignment of RNAseq data from one accession of *H. bulbosum* (Cb2920/4) to the Morex draft assembly (Mayer et al., 2012), and calling the *H. bulbosum* nucleotide at the BOPA SNP positions. Ambiguous nucleotide calls, trans-specific polymorphisms, and sites at which *H. bulbosum* segregates for different nucleotides than barley (*H. vulgare ssp. vulgare)* were treated as missing data. Derived site frequency spectra were calculated and plotted for each of the 16 populations.

2.2.6 Joint derived Site Frequency Spectrum

Following the same procedure as for the derived SFS for each breeding population we calculated the derived SFS within winter or spring accessions and within two-row or six-row accessions using a custom R script. The derived SFS was compared between growth habit and between row types. The joint derived SFSs were plotted using the R package grDevices (Team, 2012).

2.2.7 Population Structure

The degree of differentiation among individuals from all breeding programs was estimated by Principal Component Analysis (PCA). The analysis was performed using the SmartPCA program from the EIGENSOFT package (Patterson,

Price, & Reich, 2006). SmartPCA permits PCA analysis with SNP loci that include missing data.

2.2.8 Maximum likelihood tree of relatedness and migration

The population relatedness and patterns of gene flow between breeding populations was inferred using a maximum likelihood approach implemented in TreeMix (Pickrell & Pritchard, 2012). To polarize the divergence among populations we used genotyping data for 438 landrace accessions from Europe (Poets, Fang, Clegg, & Morrell, 2015) (see https://github.com/AnaPoets/BarleyLandraces). The majority of founders of the North American barley population derive from Europe (Weaver, 1944). The 2,021 SNPs shared between data sets were used to build a phylogeny tree using landraces as an outgroup. We ran 25 replicates of the tree, bootstrapping with 75 SNP windows. We used the replicate with the lowest standard error for the residuals as the base tree topology and inferred the likelihood of having between one and five migration events among breeding programs. The plot of the residuals was used to evaluate which tree best fit the data. Candidates for admixture can be identified by those population pairs with residuals above zero standard error, which represent populations that are more closely related to each other in the data than in the best-fit tree (Pickrell & Pritchard, 2012).

2.2.9 Changes in allele frequency

To identify putative targets of long-term selection involved in the row-type and growth habit differentiation among breeding programs we used the Weir and Cockerham

(Weir & Cockerham, 1984) measure of allele frequency differentiation, F_{ST}, as implemented in the R package hierfstat (Goudet, 2005). F_{ST} was calculated for the following partitions of the data: 1) spring versus winter, 2) two-row versus six-row, and 3) among breeding programs. Owing to the high level of inbreeding in the data set, a haploid model for F_{ST} estimation was used. Heterozygous SNPs were treated as missing data. An empirical genome-wide threshold for the top 2.5% of F_{ST} values was used to identify SNPs with large differences in frequency relative to the genome-wide average. To identify the degree of differentiation that each breeding program has with respect to other programs we report F_{ST} for all pairwise comparisons.

To characterize average allele frequency divergence for breeding populations an analogous to F_{ST} was calculated according to Nicholson *et al*.,(Nicholson et al., 2002) and reported as *c*. The *c* is an estimate of the degree of divergence of a population from ancestral allele frequencies. For this analysis, the data set was divided into two groups: spring six-row, and spring two-row. We do not report *c* for winter barleys owing to limited sampling. Monomorphic SNPs were removed from each group. The parameter *c* was calculated using the popdiv program from the popgen package in R (Marchini, 2013). We used a burn-in period of 1,000 iterations followed by a run length of 10,000 iterations with the scale parameter of Dirichlet distribution used to update global allele frequencies $m = 10$ (see Supplemental Information for more details).

2.2.10 Analysis of resequencing data for known genes contributing to phenotypic differentiation

Resequencing data for 10 accessions for vernalization sensitivity loci (*Vrn-H3*) (von Zitzewitz et al., 2005) and 96 accessions for *Vrs1* gene controlling row-type differentiation (Komatsuda et al., 2007) were obtained from NCBI Popsets (UID #157652625 and 219664771). Contextual sequences for SNPs known to occur in these genes were downloaded from T3. To determine the position of SNPs within these genes and their correlation with growth habit and row-type differentiation SNP contextual sequence for individual SNPs were aligned to resequencing data set in Geneious v.7.1.9 (Kearse et al., 2012).

2.2.11 Identity by State

We used an identity by state (IBS) analysis to identify shared genomic segments between breeding programs potentially indicative of recent introgression. The analysis used PLINK v.1.90 (Chang et al., 2015) with windows sized of 50 and 100 SNPs allowing for up to 10% mismatch. The frequency of shared segments between two populations was estimated for each SNP window. This analysis made use of phased genotyping data, with phase inferred using fastPHASE v1.2 (Scheet & Stephens, 2006). Missing genotypic state was treated as missing (*i.e.,* genotypic state inferred during phasing were ignored). The phased data were only used for IBS and pairwise haplotype sharing analyses.

2.2.12 Pairwise Haplotype Sharing

To explore recent events of selection within populations, we used the pairwise haplotype sharing (PHS) approach (Toomajian et al., 2006). A shared haplotype is defined as a genomic segment that extends out from a focal SNP, and is shared among individuals in a population. PHS is a form of IBS analysis that compares the extent of shared haplotypes among individuals normalized by genome-wide sharing. A PHS score depends on the length of the shared haplotype and its frequency in the population. Extended shared haplotypes are potentially suggestive of recent or ongoing selection, owing to limited potential for recombination to break down genomic regions subject to recent selection (Horton et al., 2012). PHS was calculated within each breeding population using a customized Perl script (Cavanagh et al., 2013). An empirical threshold of 2.5% and a minimum allele frequency of 10% within each population were used to identify outliers in the distribution of PHS.

2.2.13 Four-population test for gene flow detection

The robustness of patterns of migration inferred using TreeMix (see Results for details) was assessed using the four-population test *(f_4-test)* (Keinan, Mullikin, Patterson, & Reich, 2007; Reich, Thangaraj, Patterson, Price, & Singh, 2009). The *f_4-test* is designed to distinguish introgression from incomplete lineage sorting. The test evaluates trees of relatedness among populations and measures genetic drift along lineages quantitatively based on the variance in allele frequencies. Significant deviations from zero in three possible tree topologies (see Supplemental Information for details) indicate that the tree

evaluated does not fit the data, suggesting the presence of gene flow.

We used the topology inferred in TreeMix as our hypothesized relationship between populations. We inferred that a tree of relatedness with three migration events fit the data better than having more or less migrations (see the Results section for more details). Following the ((A, B), (C,D)) notation for a tree topology with four populations (S2.1), we assessed migration among the following sets of populations: ((N2, X),(Y, Y)), ((UT6, UT2), (Y, Y)) and ((OR2, X), (OR6,VT6)), where X and Y were replaced iteratively for any two-row or six-row barley populations, respectively. The populations UT6, OR2 and N2 were chosen because they had more specific signal of gene flow according to the TreeMix results. These populations were paired with populations from the same branch of the tree inferred by TreeMix, assuming that these populations are more similar to each other due to shared ancestral polymorphisms. The other pair of populations to be compared to was selected from the branch containing the population putatively involved in the migration event (connected by the arrow, Figure 2.6). Since the population putatively involved in migration was not clearly defined, we ran the analysis iteratively between pairs of populations taken from the same branch. We used the fourpop option in the TreeMix software (Pickrell & Pritchard, 2012) to estimate the f_4*-value* for each configuration. Significance of f_4*-values* was determined at $p < 0.05$. We infer that migration has occurred when the three possible trees had a significant non-zero f_4*-value.* All code used for analysis and figures is available at https://github.com/MorrellLAB/NorthAmerica_Fst

2.3 RESULTS

The original dataset included 3,971 accessions genotyped with 2,882 SNPs representative of 10 breeding programs across the United States of America. We removed 340 SNPs monomorphic in the full panel, 241 accessions with ≥25% missing data, 22 accessions with heterozygosity > 6.25% across SNPs, 13 accessions with missing growth habit (spring versus winter) or row type (two versus six rows) information, 79 near isogenic lines, and three accessions with single accessions representing a population. After quality control our data set consisted of 3,613 barley accessions (Table 2.1) and 2,542 SNPs (Table 2.2).

2.3.1 Characterization of North American breeding programs

Analysis of the structure of the North American populations using principal component analysis (Figure S2.2) revealed that the primary population structure (PC1 = 19.6% variance) is explained by differences in row-type, which corresponds to an average F_{ST} of 0.23 (Figure S2.3). PC2 indicates that 9.1% of the variance among lines is explained by differentiation in growth habit, with average F_{ST} of 0.17. These results are congruent with earlier analyses on a subset of these populations (Wang et al., 2012; Cuesta-Marcos et al., 2010; Hamblin et al., 2010) and on a comparable set of samples analyzed by Zhou *et al.* (Zhou et al., 2012).

The 16 breeding populations (*i.e.,* breeding programs separated by row-type) were represented by an average sample size of 225 lines with a minimum of 30 lines (from UT2), and a maximum sample size of 386 for MN6 (Table 2.1). On average, only two

SNPs were private to each of the breeding programs, with a maximum of 11 private SNPs in OR6. Winter programs had 16 private SNPs respect to spring programs which in turn had 72 private SNPs (Table 2.3). The average inbreeding coefficient (F_{IS}) across populations was 0.98 as expected after four generations of self-fertilization. Percent pairwise diversity ranged from 0.15 to 0.36 and averaged 0.25 across breeding programs (Table 2.1, Figure 2.2A).

The joint unfolded site frequency spectrum showed that there are slightly more rare variants in spring than winter programs and slightly more rare segregating variants in six-row than in two-row programs (Figure S2.4), reflecting minimal impact of ascertainment bias in these partitions. In individual breeding populations there is an excess of rare and high frequency variants relative to neutral expectations on a model of a population in equilibrium (Kimura, 1983) (Figure S2.5). OR2, OR6, UT6 and UT2 have a higher proportion of mid-frequency variants than other populations. When the breeding populations are considered jointly, the derived SFS (Figure S2.6) displays an elevated number of mid-frequency variants, consistent with retention of variants segregating at an average minor allele frequency of 24% in the discovery panel (Close et al., 2009).

2.3.2 Genome-wide scan for evidence of selection in the North America breeding programs

The distribution of F_{ST} statistics for comparisons of growth habit, row-type, and breeding programs showed that among the three classifications, the greatest differentiation in allele frequencies is found among breeding programs with an average

F_{ST} of 0.37, while 0.23 for inflorescence type and 0.17 for growth habit (Figure 2.3, Figure S2.3). For each of the three partitions, we identified 61 SNPs in the upper ≥0.975 of F_{ST} values, for a total of 183 SNP outliers (Tables 2.4, S2.5, S2.6), 34 SNPs were common outliers between the row-type and breeding program comparisons, and seven were common between breeding program and growth habit comparisons.

2.3.3 Known genes contributing to phenotypic differentiation

We identified a SNP outlier (12_30883, linkage group 5H) in the F_{ST} comparison for growth habit located in the vernalization sensitivity locus (*Vrn-H1*) (von Zitzewitz et al., 2005) known to be involved in the growth habit differentiation. SNPs within two additional well characterized genes photoperiod response-H1 (*Ppd-H1*) (Jones et al., 2008; Turner, Beales, Faure, Dunford, & Laurie, 2005) and c-repeat binding factor 4 (*Cbf-4*) (Haake et al., 2002) were also found to be F_{ST} outliers (Figure 2.3B). Both of these genes have reported functions related to growth habit differentiation. *Ppd-H1* alters flowering time, thus making it possible to avoid extreme unfavorable seasonal conditions (Lister et al., 2009) whereas *Cbf-4* contributes to cold acclimation (Skinner et al., 2005). While vernalization sensitivity and photoperiod response are important determinants that contribute to growth habit through an environmental response, there are also loci that contribute to flowering time independent of environmental cues. This class of genes in barley is referred to as Earliness *per se* (EPS) loci; one characterized example is the early maturity 6 locus (*Eam6)* (Boyd et al., 2003). A linkage mapping study placed early maturity 6 locus (*Eam6*) (Laurie, Pratchett, Snape, & Bezant, 1995) near the centromere

of linkage group 2H (cM 67.8), a region where we identify six SNPs with elevated F_{ST} (average F_{ST}= 0.87) (see Table 2.5, Figure 2.3). The two SNPs (11_20438 and 11_20366) most strongly associated with flowering time in a recent GWAS study (Comadran et al., 2011b) that helped to the identification of the barley *Centroradialis* gene (HvCEN) responsible for flowering time variation (Comadran et al., 2012) were not identified as outliers in our F_{ST} comparison of growth habit. However, both SNPs, 11_20438 and 11_20366, have an above average F_{ST} 0.52 and 0.25, respectively.

In an association mapping (AM) study Cuesta-Marcos *et al.* (2010) identified two significant SNPs in linkage groups 1H and 2H (12_31319 and 11_10213, respectively) associated with row-type differentiation. In our analysis, 12_31319 has an outlier F_{ST} value of 0.86, while 11_10213 has an above average F_{ST} = 0.47 but is not defined as an outlier (Figure 2.3C). Additionally, there are two SNPs (11_20422 and 11_20606) in linkage group 4H that have been identified near the *Intermedium-C* gene (*Int-c*) (Ramsay et al., 2011), which is a modifier of lateral spikelets in barley. After quality control, only 11_20422 remained in our dataset. This SNP has a F_{ST} value of 0.81 and is thus considered an outlier in the comparisons between row types (Figure 2.3C).

Other SNPs occurring in well-characterized genes did not appear as outliers despite the previously reported contribution to function. There were three SNPs in our data set (12_30893, 12_30894, and 12_30895 on linkage group 7H) occurring within the vernalization sensitive locus 3 (*Vrn-H3)* (Yan et al., 2006) that were not outliers in the F_{ST} comparison between spring and winter types (Figure 2.3). This gene is an ortholog of the

Arabidopsis thaliana flowering locus T (*FT*) that promotes flowering time under long days (Turck, Fornara, & Coupland, 2008). In barley, nine linked polymorphisms in the first intron have been predicted to be responsible for the variation in flowering time at this locus (Yan et al., 2006). Alignment of the three outlier SNPs to resequencing data of this gene from 10 barley accessions (Karsai et al., 2008) identified three of the SNPs segregating in the first intron of *Vrn-H3* (Figure S2.7A). In Yan *et al.* (2006) nucleotide states "A" and "G" in 12_30894 and 12_30895, respectively, were associated with spring barley types, while, "T" and "C" with winter types. The resequencing data in part support this association having two out of four spring barleys with the inferred haplotype while five out of six winter barleys carried the correct inferred haplotype. However, in our larger data set of spring and winter accessions these SNPs are segregating at an average allele frequency of 50% in both spring and winter growth habits (Figure S2.7B), showing no association between these SNPs and spring versus winter growth habit (maximum F_{ST} = 0.04).

The *Vrs1* gene, a well characterized contributor to row-type (Komatsuda et al., 2007) was not identified in the F_{ST} comparison between two-row and six-row accessions. Our SNP panel included five SNPs (12_30896, 12_30897, 12_30899, 12_30900, and 12_30901, linkage group 2H) in *Vrs1*. However, none of the SNPs reach the empirical cutoff for F_{ST} in this comparison, with a maximum observed F_{ST} of 0.45. The "G" nucleotide state at SNP 12_30900, results in an amino acid substitution associated with the six-row phenotype (*vrs1.a3* allele), however, the alternative nucleotide state "C" can

be found in either six-row or two-row accessions (Komatsuda et al., 2007; Youssef, Koppolu, & Schnurbusch, 2012). In a panel of 96 European accessions of cultivated barley (Popset ID 219664771) (Figure S2.7 B) the "G" state always resulted in a six-row phenotype. In our sample of North American breeding programs 2% of individuals that carry this variant state were reported as two-row barleys, which is similar to previous results that found this SNP significant for row-type differentiation segregating in two-row accessions with an allele frequency of 1% (Cuesta-Marcos et al., 2010). The "C" state segregated in 50% frequency in each of the row-type partitions.

2.3.4 Haplotype sharing and evidence for recent selection

The pairwise haplotype sharing (PHS) analysis permits the identification of genomic regions that are putatively involved in more recent selection. PHS analysis within individual breeding populations identified a total of 775 SNPs in the upper ≥0.975 of the PHS distribution (Table 2.7, Figure S2.8). In a small number of cases, focal SNPs in the PHS analysis were identified as outliers in more than one breeding population. Sharing of PHS outliers is greatest for OR6 and UT6 (22), BA6 and WA6 (17), and AB2 and OR2 (14) (Table 2.8). Average values were considerably lower, with three SNP shared within two-row populations, four within six-row populations, and two SNPs between two-row and six-row populations. Out of the 775 SNPs, 77 were in genes with known function. The haplotypes for these significant PHS values varied in length from 19.7 cM to 139.6 cM with mean 55.9 cM across breeding programs (Figure S2.9A, Table 2.9). The frequency of the SNP state with significant PHS ranged from 10% (the

minimum value we consider) to 61% with mean 24% (Figure S2.9B).

Within BAI2, BA6, MN6, UT2, UT6, WA6, and OR2 we observed long runs of haplotype sharing (average length 112.45 cM) at an average frequency of 24%, significantly exceeding genome-wide similarities. These regions were putatively subject to recent selection. Among outliers for PHS, BA2 showed significant PHS surrounding three SNPs (12_30893, 12_30894, 12_30895, in linkage group 7H) with all three SNPs occurring within *Vrn-H3*. AB2, MT2, OR2 and OR6 had an outlier PHS value for SNPs 12_30901 (linkage group 2H) in the *Vrs1* gene. The SNPs in *Vrn-H3* and *Vrs1* were at an average frequency of 22.25 % in each of these populations, with an average length of shared haplotype of 66.98 cM (Tables 2.7 and S2.9). There are three SNP (12_20368, 12_20593, and 12_21049 in linkage group 2H) with significant PHS value in OR2 (frequency 11 % and haplotype length 138.9 cM). SNPMeta annotations identified SNP 12_20593 within the nicotianamine synthase 2 (*nashor2*) gene in barley (Herbik et al., 1999) and SNP 12_20368 within a gene with sequence similarity to galactinol synthase 2 gene in wheat (*TaGolS* 2), which in turn is orthologous to the characterized gene *TaGolS* 2 in *Arabidopsis thaliana* (Taji et al., 2002). In *A. thaliana, TaGolS* 2 has been identified as playing an important role in drought-stress tolerance (Taji et al., 2002).

2.3.5 Drift from an ancestral allele frequency

Quantification of allele frequency divergence based on *c* (Nicholson et al., 2002) showed that six-row breeding programs have experienced more divergence than two-row programs, with mean *c* of 23.1% and 17.1%, respectively (Figure 2.4, Table 2.10).

Among the two-row programs Utah two-row has diverged the most while Idaho two-row resembles ancestral allele frequencies, suggesting reduced effects of drift or linked selection in the latter population. Among the six-row programs Utah six-row was the closest to ancestral allele frequencies while Minnesota six-row has experienced the most divergence.

2.3.6 Gene flow between breeding programs

To identify genomic segments subject to recent introgression, we tested for regions with a high degree of identity by state between breeding populations. Window sizes of 50 (0.63 - 44.96 cM) and 100 (13 – 62.25 cM) SNPs were used, and we permitted up to 10% mismatch among haplotypes. As expected, populations within growth habit and row-type had a higher degree of haplotype sharing than between these partitions. The degree of haplotype sharing was lower within row-type than within growth habit (Table 2.11 and S2.12, and Figures S2.10 and S2.11). There was a high frequency of shared haplotypes among two-row populations at 50 SNPs windows, but this was reduced when 100 SNPs windows were considered. At 100 SNPs windows BAI2 and N2 presented the lowest degree of shared haplotypes with other two-row populations, while BA2 shared the most haplotypes (Figure S2.11C,E,H). For both 50 and 100 SNPs windows, high frequency shared haplotypes were common for all six-row populations see for example Figure 2.5A, except for the Utah populations (Figures S2.10 and S2.11, Tables 2.13 and S2.14). This contrasted with VT6 which showed very low degree of haplotype sharing, particularly with the various spring breeding populations (Figure 2.5B). Haplotype

sharing for VT6 occurred primarily with the OR2 and OR6 populations; *i.e.*, with the only other programs with similar growth habit. However, OR2 and OR6 showed higher levels of allele frequency similarity with other breeding programs than VT6 (Figure S2.10J,K,N). These results are consistent with differentiation in allele frequency observed in F_{ST} comparisons between breeding populations (Table 2.15).

For 50 SNPs windows, there were a high number of IBS segments shared between row types. However, when windows of 100 SNPs were considered, the majority of the IBS haplotypes were at low frequency (<20% frequency) in one of the two populations compared (Tables S2.13 and S2.14). An exception to this was the comparison between OR2 and OR6, where shared haplotypes were quite common. Sharing occurred at every 100 SNPs window, with shared haplotypes at frequencies as high as 90% in the two populations (Figure S2.11K) with an average frequency within populations of 0.46 and 0.58 (for OR2 and OR6 respectively) (Table 2.14). IBS with 50 SNPs windows for N2 and N6 span 94% of all windows, with shared haplotypes occurring at average frequencies of 22% and 50%, respectively (Figure S2.10, Table 2.13). Increasing the window size to 100 SNPs identified many fewer shared haplotypes between N2 and N6, leaving 70% of the genome shared at an average frequency of 8% in N2 and 35% in N6 (Figure S2.11, Table 2.14).

2.3.7 Maximum likelihood tree of relatedness and migration

We determined that the tree topology that best describes the 16 populations separates the programs first by row-type followed by growth habit (Figure 2.6),

consistent with the F_{ST} and PCA results, with the exception of UT2 and UT6 that are not separated by row-type. In the TreeMix analysis, the two Utah and winter populations were more similar to ancestral allele frequencies, while MN6 was the most diverged.

Adding three migration events resulted in the lowest residuals and standard errors relative to trees with no migration, or one, two, four, or five migrations (Figure S2.12). With three migrations, we infer exchange between spring six-row programs and N2; VT6/OR6 and OR2, and spring six-row programs and UT6 (Figure 2.6).

Based on the *f_4-test*, six different topologies yielded significant results suggesting gene flow between UT6 and spring six-row populations (Table 2.16), with z-scores significantly different from zero (significance at $p < 0.05$). The *f_4-test* for introgression between N2 and any spring six-row program, resulted in values significantly different from zero, consistent with gene flow. The f_4-test did not support introgression between OR2 and VT6 and/or OR6 since none of the possible trees topologies involving these populations were significant.

2.4 DISCUSSION

Comparative analyses of allele frequency differentiation and extend haplotype sharing in a sample of 3,613 barley accessions representing 16 barley breeding populations results in five primary conclusions: i) barley breeding populations in North America have strongly differentiated allele frequencies among breeding populations. Across programs, inflorescence type followed by growth habit account for the greatest proportion of variance in allele frequency, ii) a number of loci previously identified as

major contributors to growth habit adaptation in barley are readily identifiable as outliers in allele frequency in our F_{ST} analyses, iii) we identify putative signals of recent and long-term selection similar in magnitude to previously isolated genes of known function, iv) the average per SNP allele frequency divergence at the population level appears consistent with the dominant action of genetic drift and linked selection, and v) identification of populations are recently subject to exchange of genetic material. There are low levels of genetic exchange across inflorescence row-type boundaries.

2.4.1 Identification of loci putatively subject to recent and long-term selection

Allele frequency differentiation sufficient to be detected in an F_{ST} outlier analysis is generally the result of long-term directional selection (Beaumont & Balding, 2004). The SNPs identified as outliers for spring versus winter growth habit include previously isolated genes of large effect, for example *Vrn-H1* and *Ppd-H1*, congruent with a previous F_{ST} analysis in European barleys (Comadran et al., 2012). Allele frequency differentiation between row-type, which explains a larger portion of the allele frequency divergence in our sample, identified two SNPs in the genomic region of genes contributing to row-type, however, SNPs within the well-characterized *Vrs1* locus were not identified as F_{ST} outliers (see below). Many additional F_{ST} outliers are newly identified as putative targets of selection.

Despite a relatively recent introduction to North America, multiple genes found to contribute to agronomic traits in Eurasia appear to recapitulate allele frequencies in current, highly structured, breeding populations. It bodes well for the potential to

translate genetic results including efforts to identify particular causative genes and mutations across breeding populations and regions with distinct demographic and breeding histories when the phenotype is conferred by a single allele (as oppose to an allelic series, as for example, *Vrs1*).

PHS analysis in the North American breeding programs detected a series of loci that are putatively targets of long-term selection within barley breeding programs, comparable to findings in wheat (Cavanagh et al., 2013). The length of the haplotypes shared (average 81.38 cM) and their frequencies (average 25%) suggest that these events have taken place in recent generations so that recombination has not had time to break down the haplotypes. This analysis indicates that selection and linked selection are altering patterns of haplotype diversity across the genome. However, there are some chromosomes that have little or no evidence of selection in most recent generations (*e.g.*, linkage group 2H in BA2 and WA2. See Figure S2.8).

2.4.2 Limitations of allele frequency comparisons

Despite having five SNPs positioned in the *Vrs1* gene (controlling the fertility of lateral spikelet) (Komatsuda & Mano, 2002), including one SNP (12_30900) variant co-segregating with the six-row phenotype, the comparison of allele frequencies differentiation between row-type fails to identify outliers in *Vrs1*. One important contributing factor may be that the six-row phenotype can result from multiple disruptions of the *Vrs1* gene (Komatsuda et al., 2007). The phenomenon of multiple mutational paths to the same phenotype has been identified as "functionally equivalent

mutations" (Ralph & Coop, 2010) and can result in multiple variants at modest frequency controlling a phenotype. The maintenance of alleles of large effect at modest frequency in the population reduces the power to detect phenotypic associations, in part because any given variant explains a small portion of phenotypic variation (Thornton, Foran, & Long, 2013).

The BOPA SNP platform used here is largely derived from variants identified from cDNA libraries, from cultivated accessions (Close et al., 2009), and thus is highly enriched for SNPs in genic regions (Kono et al., 2013). However, SNPs genotyped within a locus can have limited correlation with a phenotype, depending on the haplotype on which they occur (see Nordborg & Tavaré, 2002for discussion of this issue). This may be the case for the previously cloned *Vrn-H3* where we see limited correlation between SNPs in this gene and allele frequency difference in the growth habit comparison. It should also be noted that the resequencing panel in which *Vrn-H3* alleles were found to be associated with the phenotype included only five spring and eight winter individuals (Yan et al., 2006). It is perhaps not surprising that sampling error (*i.e.,* small sample sizes) accounts for inconsistencies in the phenotype-genotype associations reported in previous studies and our study using a panel of >3,000 accessions.

2.4.3 The effects of linked selection and drift in breeding programs

The derived site frequency spectrum in individual breeding population identifies an excess of rare and high frequency derived variants (Figure S2.5) with respect to neutral model of a population in equilibrium. During selection, alleles linked to the target

variant can be carried to high frequencies. Therefore, an excess of high and rare frequency of derived variants can be associated with the effects of linked selection (Fay & Wu, 2000) as has been observed in a resequencing study of *Oryza sativa* (Caicedo et al., 2007). The PHS results presented here suggest considerable potential to detect the effect of linked selection, particularly when the selection was recent, impacting large genomic regions. One of the largest genomic region detected as an outlier in the PHS analysis is found in OR2, this haplotype involves 138.9 cM on linkage group 2H, observed at 11% frequency. Individuals carrying this haplotype derived from a two-row by six-row cross (Merlin x Strider) and were selected to recover the two-row phenotype (Dr. Hayes, personal communication). This resulted in a long shared haplotype centered on the *Vrs1* gene contributing to row-type (identified by SNPs 12_30901). Two of the SNPs that permitted the identification of the outlier haplotype occur within an iron homeostasis (*nashor 2*) and drought-stress tolerance genes (*TaGolS 2*).

Our comparisons of allele frequency differences among individual barley breeding populations (see Figure 2.4 and Table 2.15) suggests that genetic drift (and likely linked selection) play a major role in differentiation among populations. The cumulative effects of selection and genetic drift over several generations of breeding may result in reduced response to selection (Hanrahan, Eisen, & Lagates, 1973), with larger effects of drift when the population is small and/or highly inbred (Robertson, 1960). The magnitude of drift suggests that the majority of barley breeding programs have small effective population sizes. Gerke *et al*. (2014) noted that genetic drift played a major role in

changes in allele frequency over the history of maize breeding in North America. Although, in the larger partitions of barley populations, signals of directional selection are clearly evident at loci of large effect, drift may be extremely important to the loss of variation for variants of small effect (or for phenotypes determined by multiple genes), especially when the selective pressure is sporadic (*e.g.*, drought tolerance).

2.4.4 Gene flow among North American breeding programs

Gene flow between populations can be detected by the presence of large chromosomal regions in (admixture) linkage disequilibrium (Briscoe, Stephens, & O'Brien, 1994; Chakraborty & Weiss, 1988; Chakraborty & Smouse, 1988) or extended genomic regions of shared ancestry (Gusev et al., 2009; Gusev et al., 2012). Shorter shared haplotypes indicate shared ancestry a larger number of generation before present (more generations of recombination). Considering the high frequency of IBS segments (and associated low F_{ST}) across the genome between populations of similar inflorescence type, we speculate that these segments could reflect a history of shared ancestry among these programs, possibly pre-dating the separation of breeding programs (Martin et al., 1991).

Gene flow that involves adaptive variation can result in differential retention of genomic segments in the recipient population, and can also be evident as shared genomic regions with reduced diversity (Hufford et al., 2013). For example, MN6 demonstrates a high degree of IBS with other six-row populations (Figure 2.6A), and these genomic regions have lower average pairwise diversity than other regions when the six-row

populations are analyzed together (Figure S2.13).

With regard to IBS, VT6 is the most clearly differentiated from other North American populations, showing similarity only to the other winter programs, OR2 and OR6 (for 50 SNPs windows, Figure S2.10). Pairwise F_{ST} with other breeding programs and genetic differentiation as measured by PCA (Figure S2.2B) suggest that this isolation has been maintained over many generations. Large IBS segments (100 SNPs windows) shared between VT6 and OR2 and OR6 (Figure 2.5B) indicate that winter programs have recently exchanged genetic material. The demographic history inferred in the TreeMix analysis supports this relationship (Figure 2.6), particularly when invoking three migration events, the topology best supported by the data. However, the formal examination of migration using the *f_4-test* failed to support gene flow among two-row and six-row winter programs, thus suggesting that haplotype similarities across row-types are due to ancestral history rather than recent introgression. In addition to the postulated shared ancestry, genetic similarity is expected between these programs under the premise that both Virginia and Oregon breeding programs have been subject to relative recent introgression from leaf rust resistance lines from the International Maize and Wheat Improvement Center (CIMMYT) (Dr. Griffey and Dr. Hayes, personal communication). Isolated populations, like VT6, can carry adaptive variants absent in other populations (Kristiansson, Naukkarinen, & Peltonen, 2008). Therefore, VT6 could be used for the identification of new variants for yield or disease resistance, as well as a source of genetic variation to increase diversity in other breeding programs.

The high degree IBS segments shared between the Utah populations and most of the six-row populations yielded a significant signal of introgression based on the *f_4-test*. It is possible that the signal detected here is due to the ongoing genetic exchange between UT2 and UT6 (as documented in pedigrees) since 2001 (Dr. David Hole, personal communication) and the genetic history shared between six-row barleys. This may also account for the elevated pairwise diversity in the Utah populations. Additionally, there is evidence of introgression between N2 and two six-row populations (WA6 and AB6), but not with N6. The introgression between the six-row populations and N2 are uncommon (Dr. Jerome Franckowiak, personal communication), thus this result needs more exploration.

Using the tree of relatedness as our hypothetical relationship, a necessary component of the *f_4-test*, forces populations of the same row-type to be assumed as related by shared history. Therefore, gene flow within row-types is not investigated. Although, crosses across row-type are not common in barley breeding programs due to concerns of recombining desirable alleles with less favorable ones (Martin et al., 1991), we detect two instances of introgression across row-type that result in higher levels of genetic diversity as it was speculated by Martin *et al*. (Martin et al., 1991).

2.4.5 Implications

Comparative population genetic approaches have the potential to uncover breeding history at a level of detail not previously possible. Our applications of allele frequency differentiation analyses in barley breeding programs are able to identify

candidate genes or linked markers controlling major traits. Though the nature of the target of selection identified by individual SNPs is not always readily apparent, the identification of F_{ST} outliers reported here provides a reasonable first step toward the discovery of genes underlying agronomic adaptation or potential markers linked to these genes.

F_{ST} analyses also reveal the effect that linked selection and drift have in breeding programs. Taking into account that with any degree of linkage a proportion of the genetic variation may not be immediately available for selection (Robertson, 1970; Hill & Robertson, 1966); furthermore, beneficial variants could be in linkage with deleterious mutations (Felsenstein, 1974). Gains from selection in breeding programs will depend on the disassociation of these linkage blocks (Morrell, Buckler, & Ross-Ibarra, 2012; Rodgers-Melnick et al., 2015). This can be achieved by increasing the effective amount of recombination. Hill and Robertson (Hill & Robertson, 1966) suggested that a relaxed generation with a large number of contributing parents between each generation of selection could increase the amount of recombination needed to disassociate these variants.

2.4.6 Caveats of the analysis

It is important to note that the results presented here are dependent on the samples submitted for genotyping (https://triticeaetoolbox.org) by each breeding program. Given that one of the goals of the Barley Coordinated Agricultural Project

(http://www.barleycap.org) was to evaluate the diversity in the North American barley breeding program each program was encourage to submit representative lines. We assume that the data represent the diversity within each breeding programs and that the same sampling scheme across breeding programs is comparable. Deviations from these assumption could influence the results in four major ways: *i*) under estimation of the diversity within breeding programs; *ii*) over estimation of the role of drift in breeding populations due to reduced representation of the parental lines used in the breeding programs; *iii*) over or under estimation of allele frequency differentiation between partitions of the data; and *iv*) excess of shared haplotypes due to accessions highly related by pedigree (*e.g.,* sibs and half-sibs). With these caveats in mind, we made the best use of the dataset to uncover the underlying genetic response to breeder's efforts. We encourage readers of this paper and barley breeders to evaluate the results with an understanding of the limitations of the sampling.

Table 2.1 Descriptive statistics by breeding program and row type.

Breeding Program	Abbreviation	Row Type	No. Markers	Sample Size	Average Pairwise Diversity	Mean F_{IS}	Private SNP
Spring barley			**2392**	**2952**	**0.331**	**0.980**	**72**
University of Idaho in Aberdeen	AB2	2	2103	239	0.266	0.983	3
	AB6	6	1791	142	0.257	0.984	1
Busch Agricultural Resources, Inc.	BA2	2	2029	172	0.203	0.997	0
	BA6	6	1770	147	0.166	0.997	0

Busch Agricultural Resources, Inc.(International)	BAI2	2	1508	60	0.295	0.997	0
University of Minnesota	MN6	6	1650	386	0.148	0.989	0
Montana State University	MT2	2	2041	317	0.264	0.989	1
North Dakota State University Two-row	N2	2	2177	353	0.241	0.959	5
North Dakota State University Six-row	N6	6	1722	380	0.197	0.980	2
Washington State University	WA2	2	2032	351	0.248	0.971	1
	WA6	6	1665	32	0.256	0.976	0
Winter barley			**2336**	**661**	**0.291**	**0.972**	**16**
Oregon State University	OR2	2	2052	73	0.293	0.960	0
	OR6	6	2272	268	0.274	0.968	11
Utah State University	UT2	2	1650	30	0.360	0.975	1
	UT6	6	2283	343	0.303	0.970	6
Virginia Polytechnic Institute and State University	VT6	6	2025	320	0.246	0.985	2
Average			1923.125	225.813	0.251	0.980	6.722

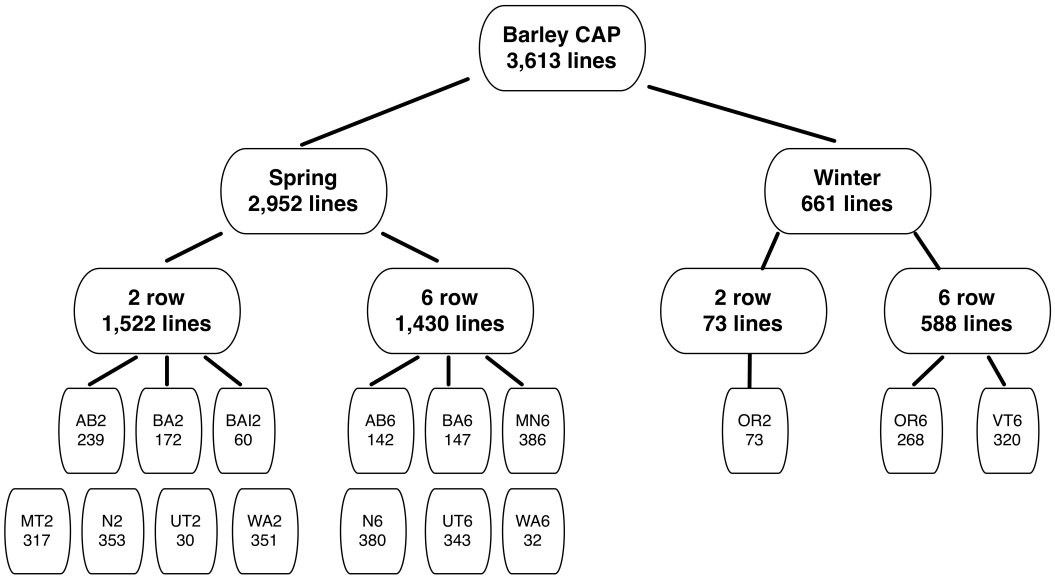


Figure 2.1 Breeding programs.

University of Idaho in Aberdeen (AB), Busch Agricultural Resources, Inc. (BA), Busch Agricultural Resources, Inc. International lines (BAI), Oregon State University (OR), Utah State University (UT), Washington State University (WA), Montana State University (MT), Virginia Polytechnic Institute and State University (VT), North Dakota State University two-rows (N2), North Dakota State University six-row (N6), and University of Minnesota (MN).

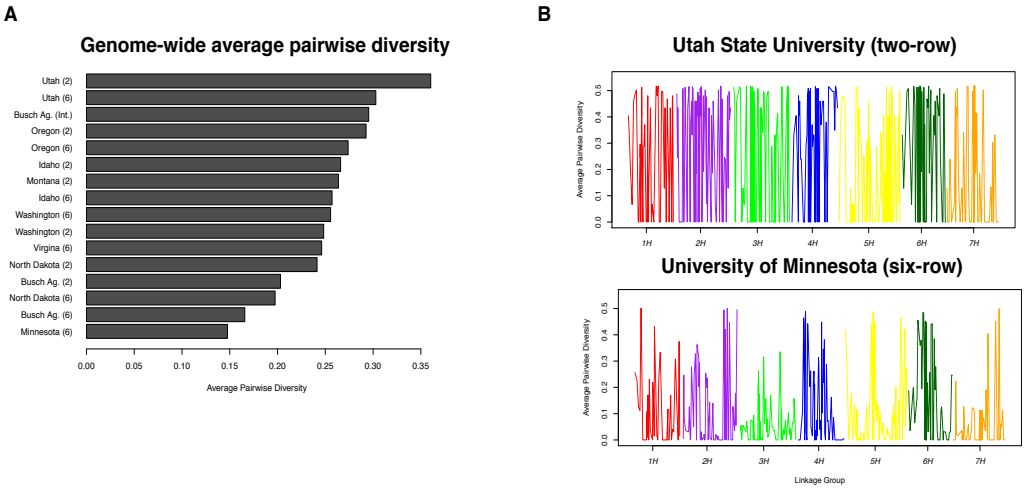


Figure 2.2 Genetic diversity in breeding programs.
(**A)** Average pairwise diversity in each breeding program. (**B)** Genome-wide pairwise diversity for the most and least diverse populations according to average diversity, Utah State University (two-row) and University of Minnesota (six-row), respectively. Diversity values were averaged in 10-SNP sliding windows with a step of five SNPs.

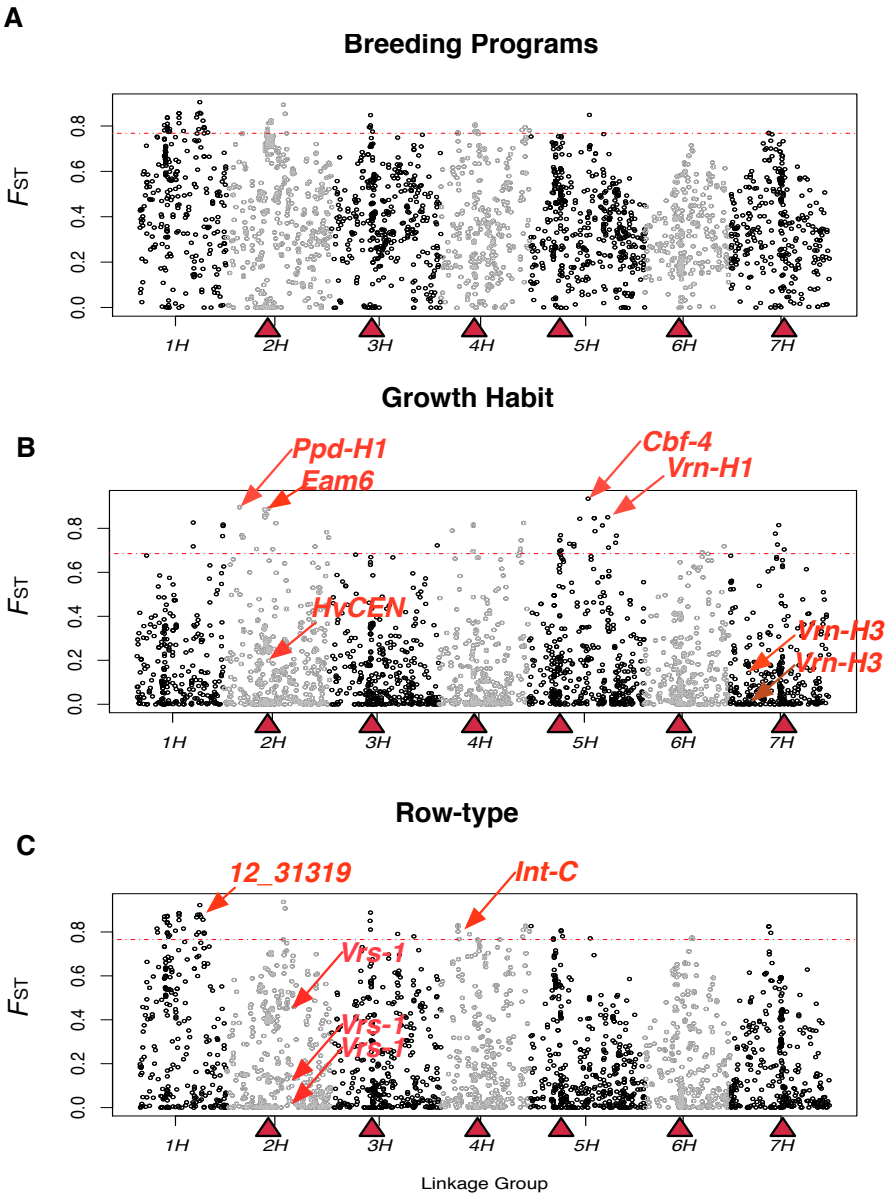


Figure 2.3 Distribution of F_{ST} values from comparisons between different partitions of the data.

(A) Among breeding programs; **(B)** between spring and winter types; and **(C)** between six-row and two-row types. The mean of the distribution is indicated by the dotted line. Characterized genes contributing to the trait are indicated.

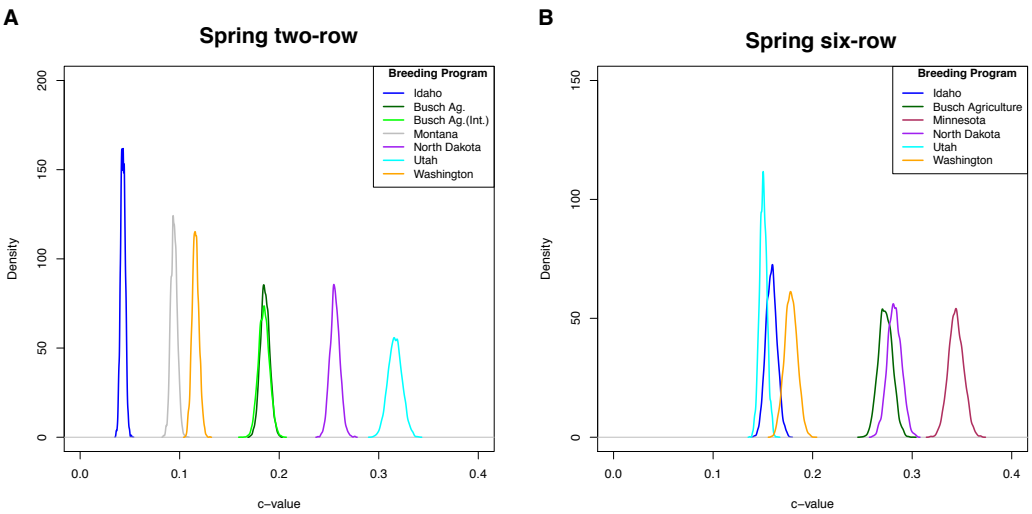


Figure 2.4 Marginal posterior density plots obtained for *c* for spring barley populations.

Density distribution after 10, 000 iterations. Plot shows the amount of drift estimated from ancestral allele frequencies (0,0 coordinates). **(A)** Two-row barley populations; **(B)** Six-row barley populations.

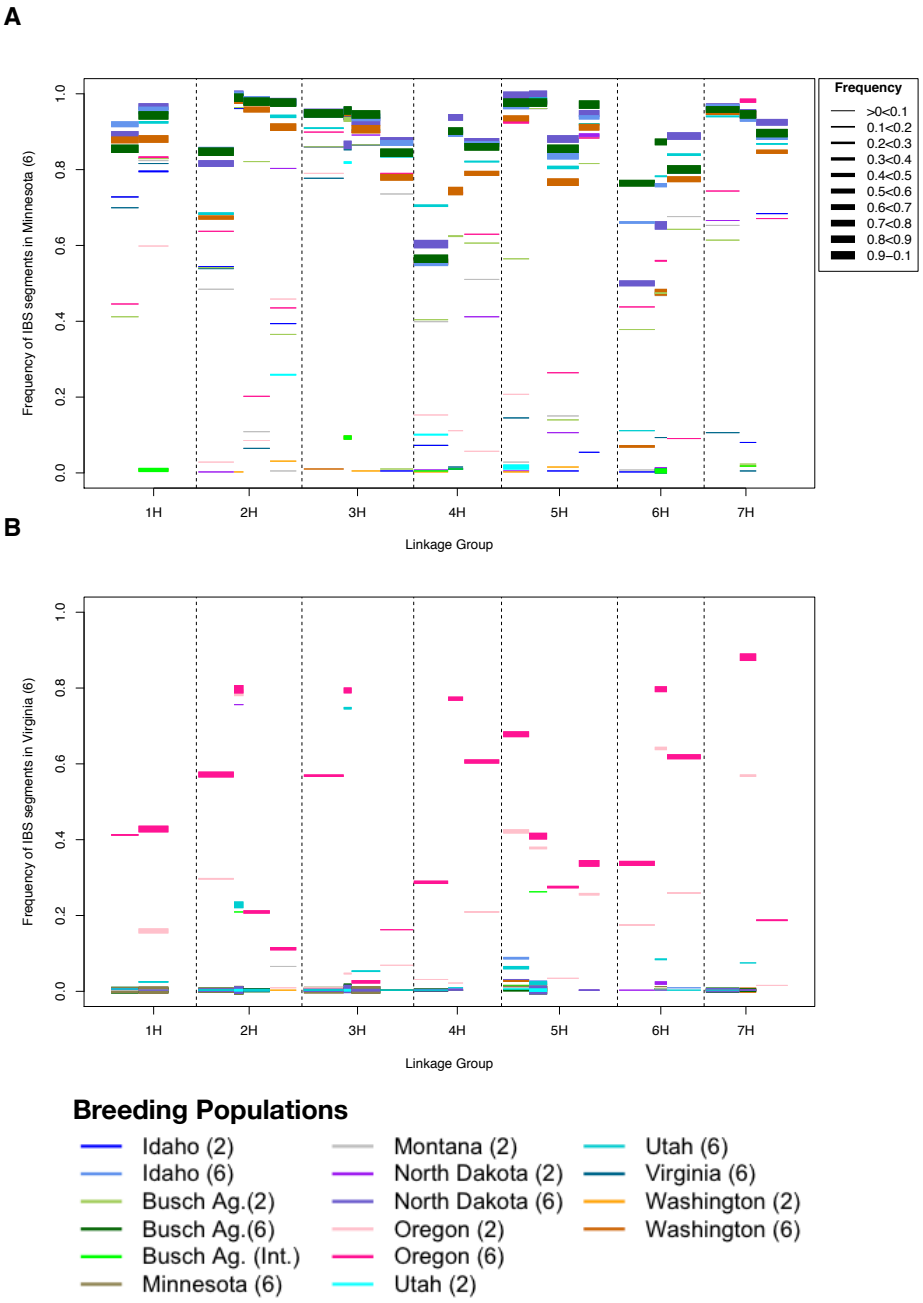


Figure 2.5 Frequency of identity by state haplotypes.

(A) University of Minnesota breeding programs and **(B)** Virginia Polytechnic Institute and State University. Using 100 SNPs windows. Linkage groups are in the X-axis. The Y-axis is the frequency of the haplotype in the population indicated in the Y-axis label ("base population"). The color of the bars in the plot represent the distinct barley breeding programs that share the same segment with the "base population". The width of each bar corresponds to the frequency of that haplotype in the compared populations.

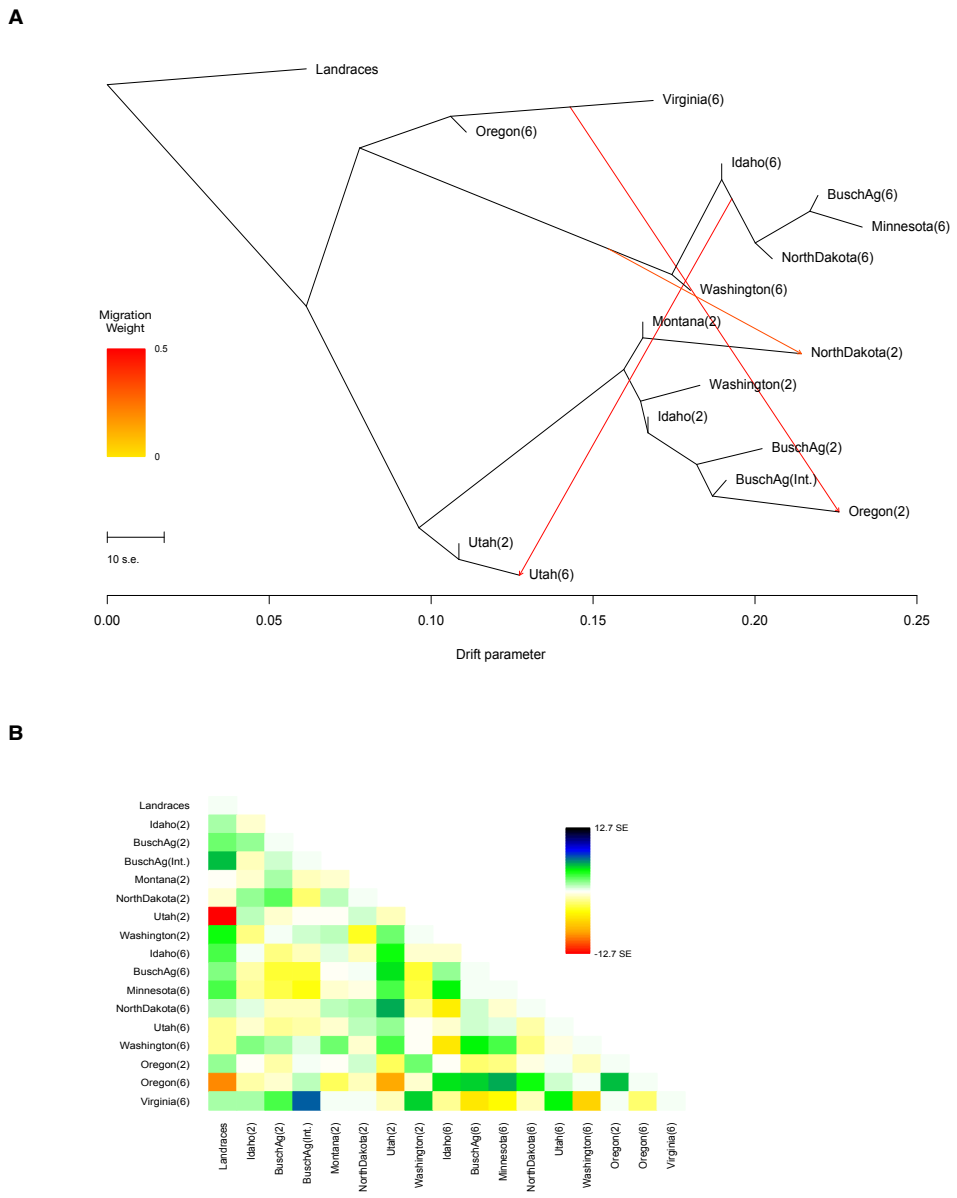


Figure 2.6 Tree of relatedness among North American breeding programs.

(A) Plotted is the structure of the graph inferred by TreeMix for the 16 North American breeding populations, allowing three migration events. The arrows (migration events) are colored according to their weight. Horizontal branch length is proportional to the amount of genetic drift in the branch. The scale bar in the left shows ten times the average standard error of the entries in the sample covariance matrix. (**B)** Residual fit of inferred tree of relatedness. Residuals above zero represent populations that are more closely related to each other in the data than in the best-fit tree, thus are considered candidates for admixture events. Row-type for each population is presented in parenthesis.

2.5 SUPPORTING INFORMATION

2.5.1 Supplementary Figures

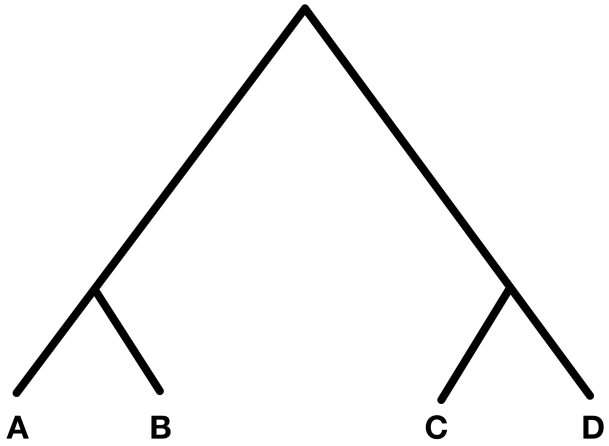


Figure S 2.1 Unrooted tree for the relationship among four populations.

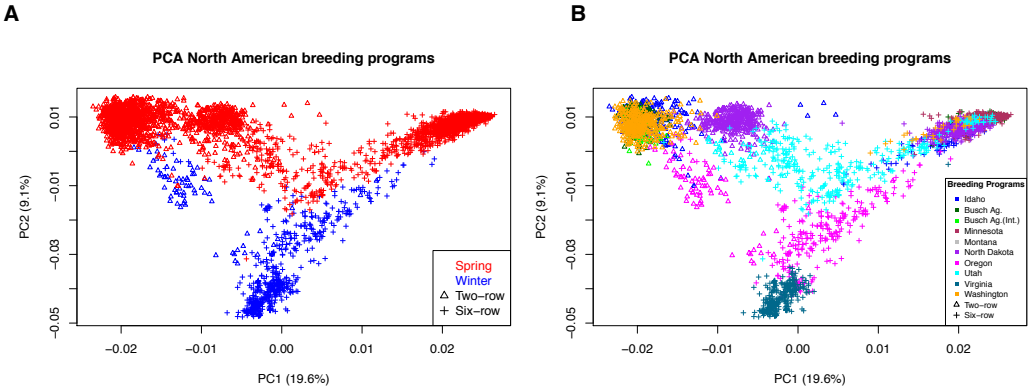


Figure S 2.2 Relationship of barley lines from the North American breeding programs by principal components.

(A) Colors separate growth habit and shapes separate row type; **(B)** Colors separate breeding programs and shapes distinguish two from six row types.

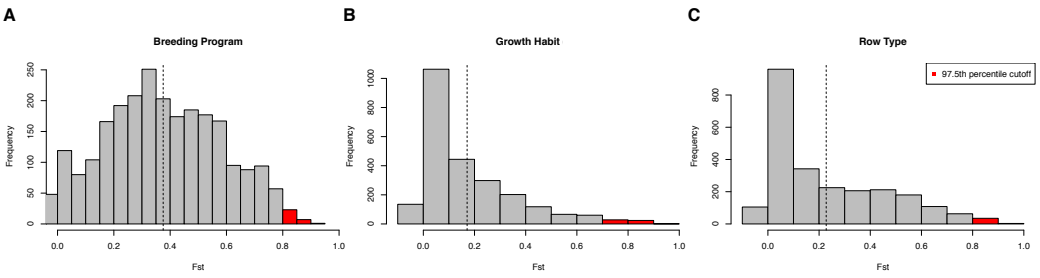


Figure S 2.3 Distribution of F_{ST} values from comparisons between different partitions of the data.

(A) Among breeding programs; **(B)** between spring and winter types; and **(C)** between six-row and two-row types. SNPs falling in the 97.5 percentile cutoff are depicted in red. The mean of the distribution is indicated by the dotted line.

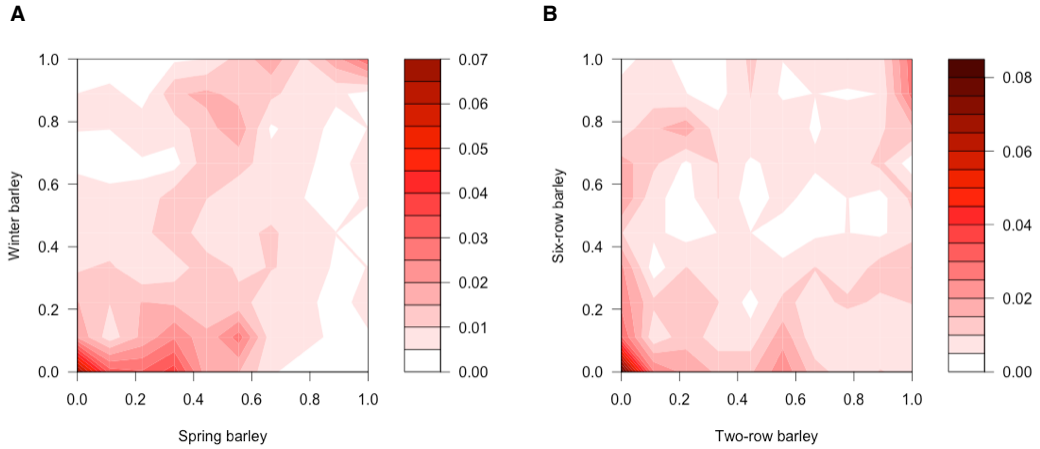


Figure S 2.4 The joint unfolded (derived) site frequency spectrum.

(**A**) Winter versus spring cultivars; (**B**) six-row versus two-row cultivars. The scale in the right indicates the number of times for each observation given in percentage.

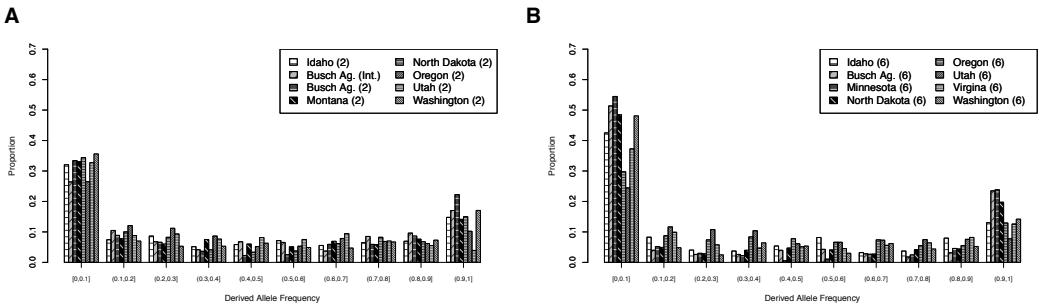


Figure S 2.5 Derived site frequency spectrum in breeding programs separated by row type.

(A) Two-row barley programs; **(B)** six-row barley programs.

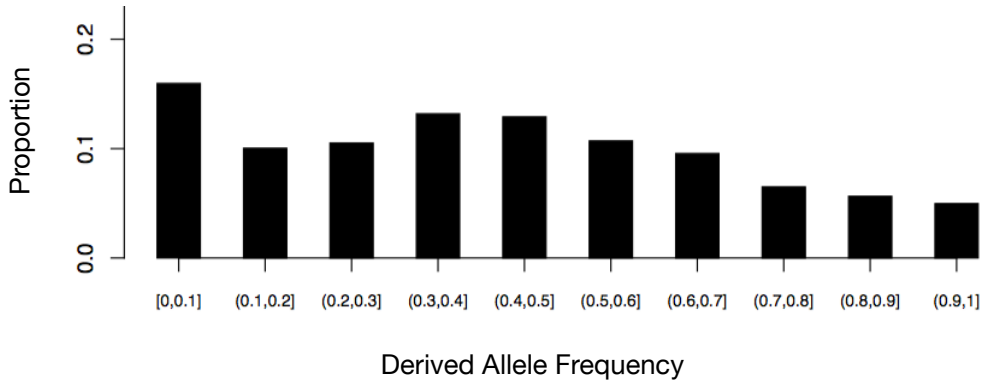


Figure S 2.6 Derived site frequency spectrum for all 3,613 barley lines combined.

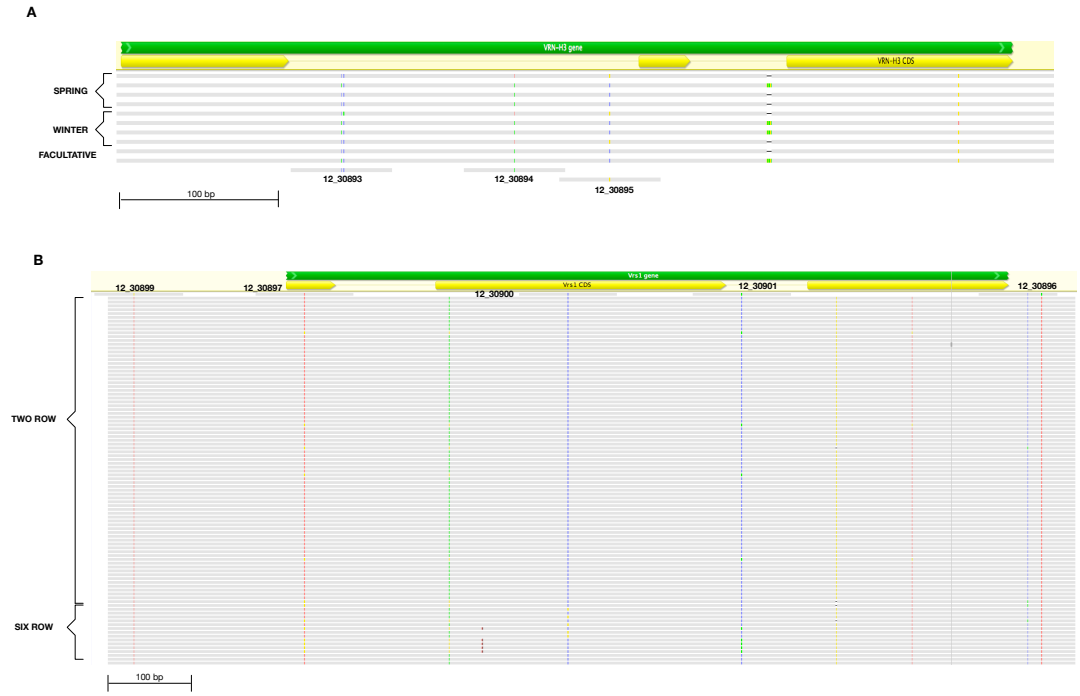


Figure S 2.7Alignments of SNPs contextual sequences to resequencing data.

(**A)** Alignment of three SNP contextual sequences to resequencing data for *Vrn-H3* (B) Alignment of five SNP contextual sequences to resequencing data for *Vrs1*.

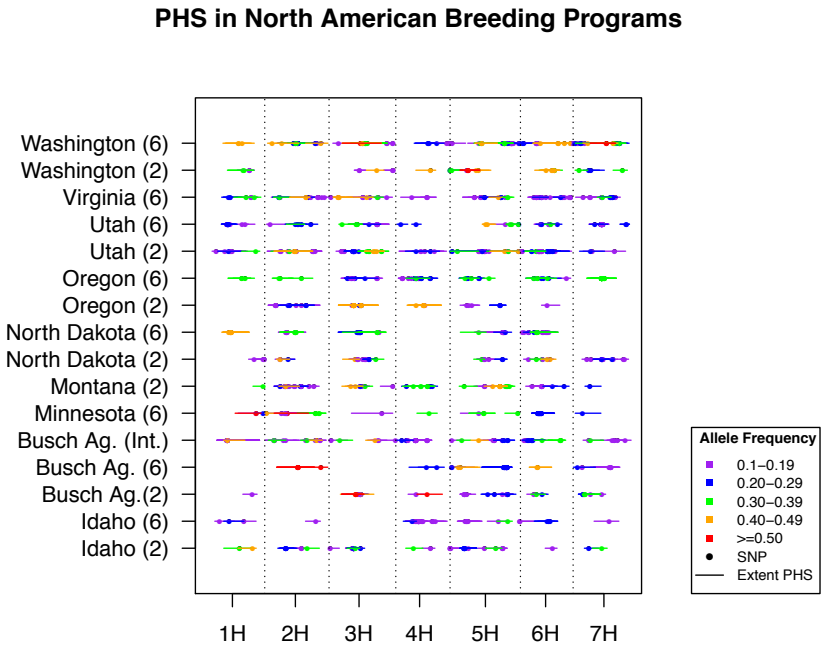


Figure S 2.8 Pairwise haplotype sharing.

Extent of shared haplotypes among individuals in each breeding population, normalized by how much is shared genome-wide. Average length of significant shared haplotypes ranges from 19.5 cM to 139.6 cM. Colors correspond to the frequency of the shared haplotypes within a program.

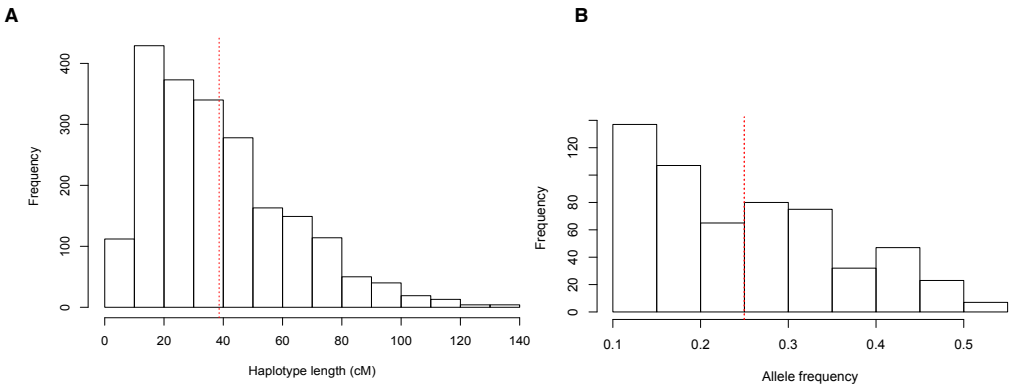
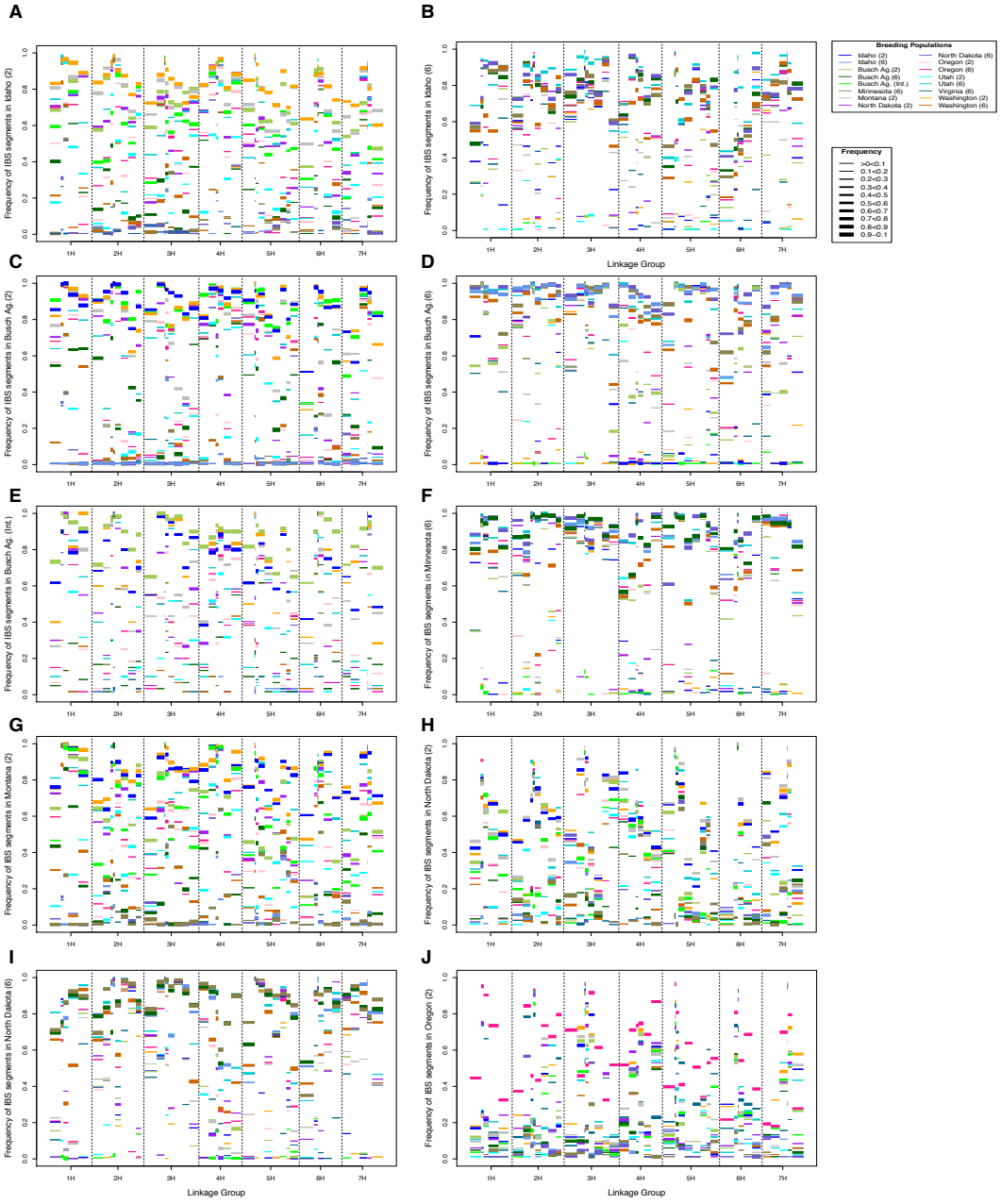


Figure S 2.9 Frequency distribution.

(A) Haplotype length and **(B)** haplotype frequency for SNPs with outlier PHS values. Dashed line indicates the distribution's mean.


A
B
C
D
E
F
G
H
I
J
Frequency of IBS segments in Idaho (2)
Frequency of IBS segments in Idaho (6)
Frequency of IBS segments in Busch Ag (2)
Frequency of IBS segments in Busch Ag (6)
Frequency of IBS segments in Busch Ag. (Int.)
Frequency of IBS segments in Minnesota (6)
Frequency of IBS segments in Montana (2)
Frequency of IBS segments in North Dakota (2)
Frequency of IBS segments in North Dakota (6)
Frequency of IBS segments in Oregon (2)
1H
2H
3H
4H
5H
6H
7H
Linkage Group
Breeding Populations
Idaho (2)
Idaho (6)
Busch Ag.(2)
Busch Ag.(6)
Busch Ag. (Int.)
Minnesota (6)
Montana (2)
North Dakota (2)
North Dakota (6)
Oregon (2)
Oregon (6)
Utah (2)
Utah (6)
Virginia (6)
Washington (2)
Washington (6)
Frequency
>0<0.1
0.1<0.2
0.2<0.3
0.3<0.4
0.4<0.5
0.5<0.6
0.6<0.7
0.7<0.8
0.8<0.9
0.9<0.1

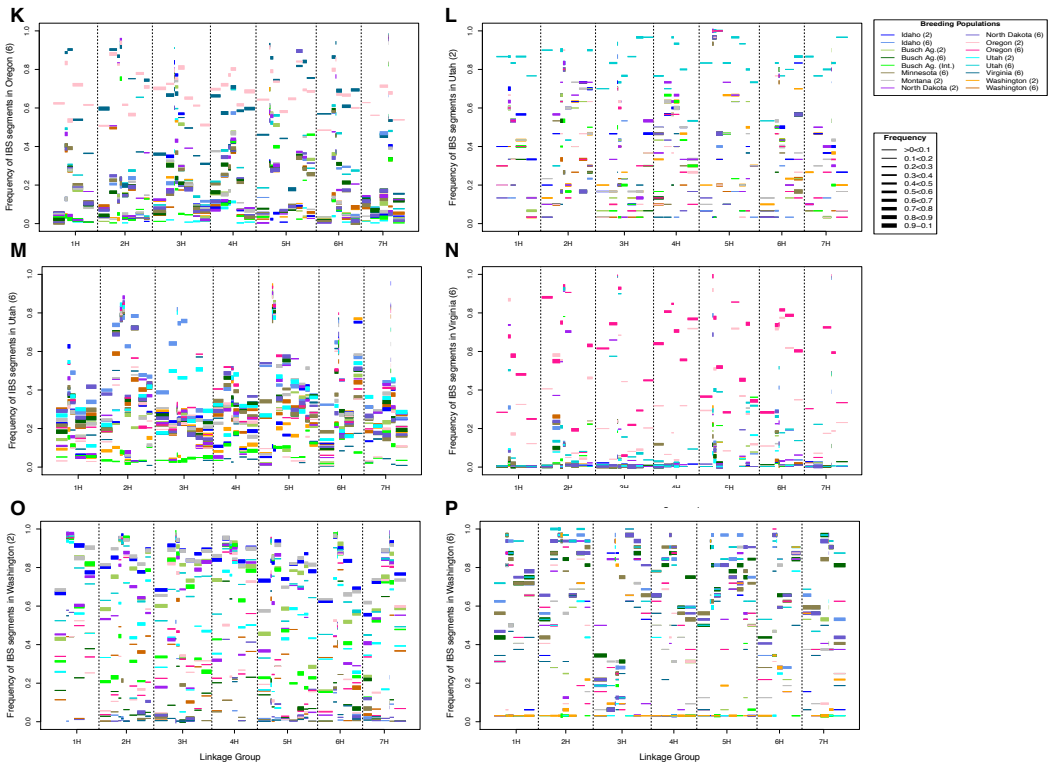


Figure S 2.10 Frequency of identity by state haplotypes. Using 50 SNPs windows.

Linkage groups are in the X-axis. The Y-axis is the frequency of the haplotype in the population indicated in the Y-axis label ("based population"). The colors of the bars in the plot represent the distinct barley breeding programs that share the same segment with the "base population". The width of each bar corresponds to the frequency of that haplotype in the compared population. Each panel represents the frequency of haplotypes shared between populations A) Idaho two-row, B) Idaho six-row, C) Busch Ag. two-row, D) Busch Ag. six-row, E) Busch Ag. International, F) Minnesota six-row, G) Montana two-row, H) North Dakota two row, I) North Dakota six-row, J) Oregon two-row, K) Oregon six-row, L) Utah two-row, M) Utah six-row, N) Virginia six-row, O) Washington two-row, and P) Washington six-row.


A
B
C
D
E
F
G
H
I
J
Frequency of IBS segments in Idaho (2)
Frequency of IBS segments in Idaho (6)
Frequency of IBS segments in Busch Ag (2)
Frequency of IBS segments in Busch Ag (6)
Frequency of IBS segments in Busch Ag. (Int.)
Frequency of IBS segments in Minnesota (6)
Frequency of IBS segments in Montana (2)
Frequency of IBS segments in North Dakota (2)
Frequency of IBS segments in North Dakota (6)
Frequency of IBS segments in Oregon (2)
1H
2H
3H
4H
5H
6H
7H
Linkage Group
Breeding Populations
Idaho (2)
Idaho (6)
Busch Ag (2)
Busch Ag (6)
Busch Ag. (Int.)
Minnesota (6)
Montana (2)
North Dakota (2)
North Dakota (6)
Oregon (2)
Oregon (6)
Utah (2)
Utah (6)
Virginia (6)
Washington (2)
Washington (6)
Frequency
>0<0.1
0.1<0.2
0.2<0.3
0.3<0.4
0.4<0.5
0.5<0.6
0.6<0.7
0.7<0.8
0.8<0.9
0.9<0.1

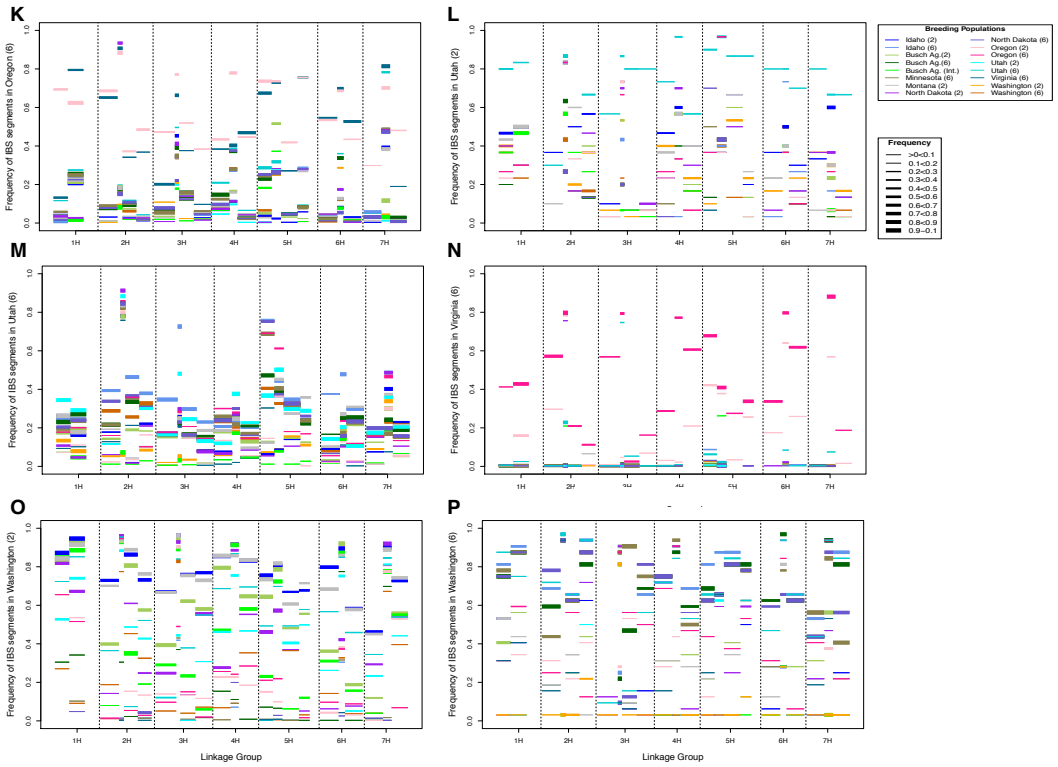
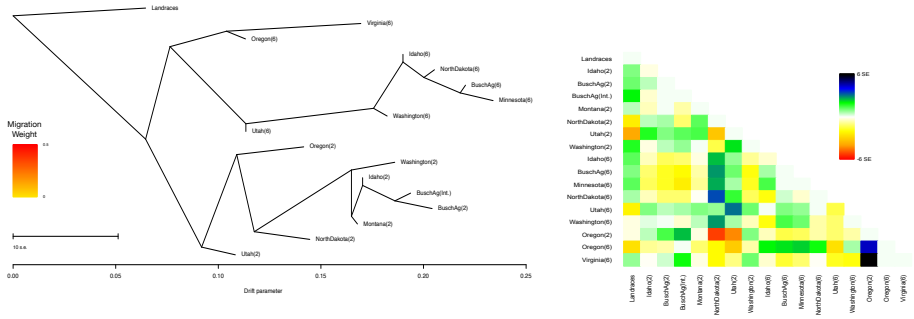


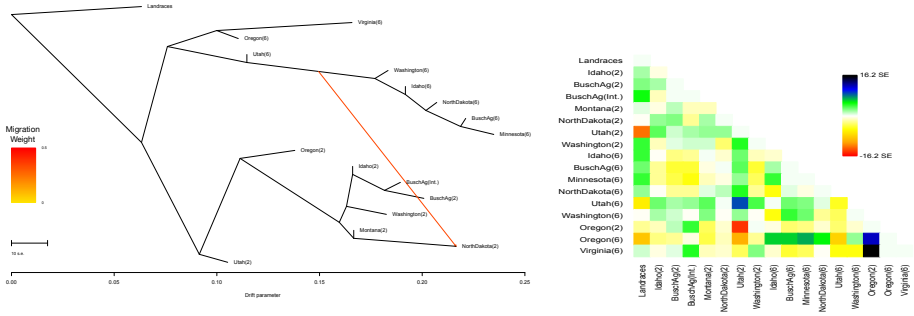
Figure S 2.11 Frequency of identity by state haplotypes. Using 100 SNPs windows.

Using 100 SNPs windows. Linkage groups are in the X-axis. The Y-axis is the frequency of the haplotype in the population indicated in the Y-axis label ("based population"). The color of the bars in the plot represent the distinct barley breeding programs that share the same segment with the "base population". The width of each bar corresponds to the frequency of that haplotype in the compared population. Each panel represents the frequency of haplotypes shared between populations A) Idaho two-row, B) Idaho six-row, C) Busch Ag. two-row, D) Busch Ag. six-row, E) Busch Ag. International, F) Minnesota six-row, G) Montana two-row, H) North Dakota two row, I) North Dakota six-row, J) Oregon two-row, K) Oregon six-row, L) Utah two-row, M) Utah six-row, N) Virginia six-row, O) Washington two-row, and P) Washington six-row.

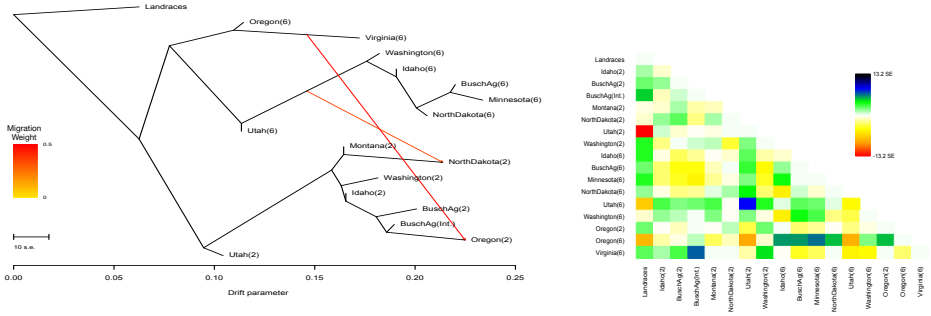
A. No-Migration



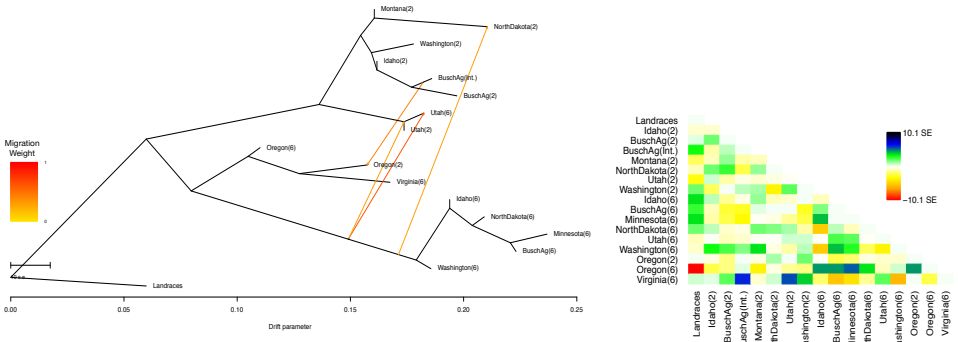
B. One-migration



C. Two-migrations



D. Four-migrations



E. Five-migrations

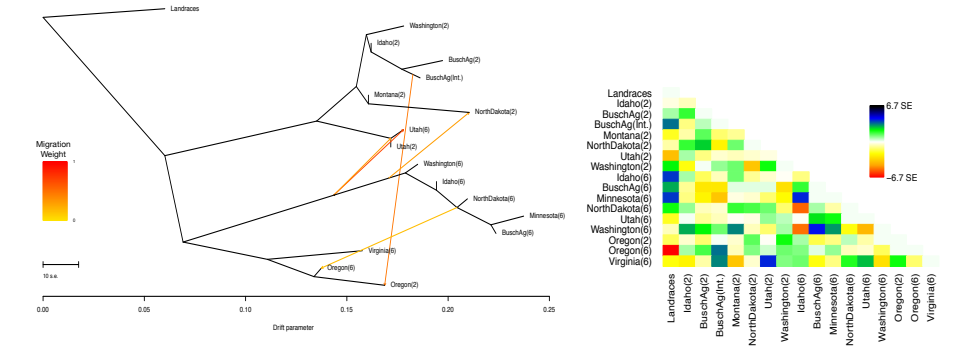


Figure S 2.12 Tree topology for the 16 North American breeding programs as inferred by TreeMix, allowing for different levels of migration.

(A) No-migration; **(B)** One-migration; **(C)** Two-migrations; **(D)** Four migrations; **(E)** Five-migrations. The arrows (migration events) are colored according to their weight. Horizontal branch length is proportional to the amount of genetic drift in the branch. The scale bar shows ten times the average standard error of the entries in the sample covariance matrix. Next to each tree is the residual fit of inferred tree of relatedness. Residuals above zero represent populations that are more closely related to each other in the data than in the best-fit tree, thus are considered candidates for admixture events.

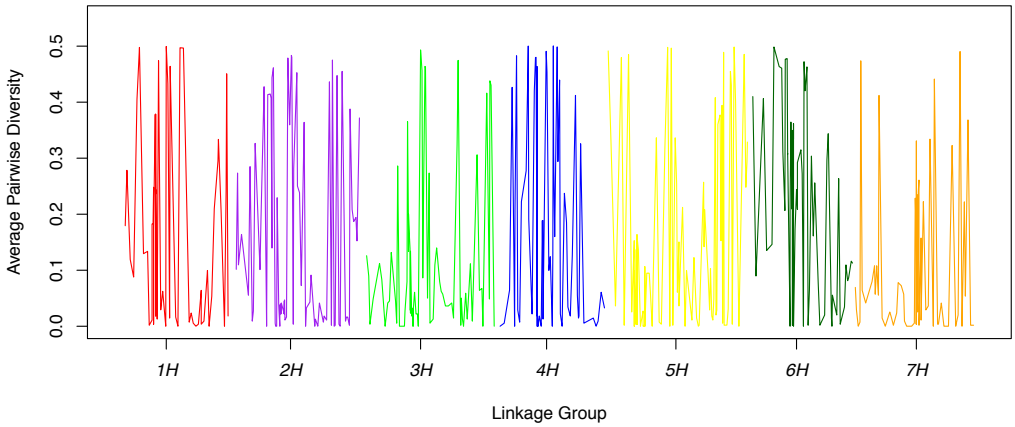


Figure S 2.13 Genome-wide diversity among six-row spring breeding programs. Diversity values were averaged in 10-SNP sliding windows with a step of five SNPs.

2.5.2 Supplementary Tables

Table S 2.1 Samples information for the 3,613 barley accessions representing North American breeding programs.

Table S 2.2 Single Nucleotide Polymorphism (SNP) markers used in this study with linkage group, genetic position, and annotation information.

Table S 2.3 Annotations for SNPs private to winter and spring breeding programs

SNP name	Linkage Group	Position (cM)	Gene short name	Position in gene	Silent
Winter					
12_11409	1H	117.24	-	-	-
12_10718	2H	3.1	-	non-coding	yes
12_21415	2H	3.84	-	-	-
12_30775	2H	10.86	LOC10083704 2	2	no
12_10154	2H	69.05	-	non-coding	yes
12_11369	2H	69.05	-	non-coding	yes
12_11102	3H	59.83	-	non-coding	yes
12_21345	3H	163.49	-	non-coding	yes
12_11060	4H	26.2	-	1	no
12_30394	4H	26.2	-	non-coding	yes
12_11134	5H	86.08	-	1	no
12_30238	5H	154.08	-	non-coding	yes
12_11325	5H	157.61	-	-	-
12_11104	6H	56.06	-	-	-
12_10278	6H	60.65	-	-	-
12_11309	7H	151.1	-	-	-
Spring					
12_11311	1H	3.21	-	-	-
12_10506	1H	42.42	-	non-coding	yes
12_30059	1H	45.2	-	non-coding	yes
11_10176	1H	58.59	-	non-coding	yes
12_11173	1H	103.99	-	3	yes
12_21172	1H	121.56	-	non-coding	yes
11_10903	1H	133.47	-	non-coding	yes

12_30379	2H	48.8	-	-	-
12_30491	2H	49.5	-	-	-
12_20688	2H	55.46	-	2	no
11_21096	2H	56.64	-	non-coding	yes
11_21388	2H	57.29	-	non-coding	yes
12_30338	2H	60.89	-	-	-
12_11316	2H	73.89	-	1	no
12_21476	2H	74.55	OPR2	1	no
12_11347	2H	76.61	-	3	yes
11_11072	2H	77.76	-	non-coding	yes
12_11388	2H	89.68	-	-	-
12_21396	2H	129.3	-	3	no
12_31406	2H	136.33	SSIIIb	non-coding	yes
12_30248	2H	154.03	-	non-coding	yes
12_20027	2H	155.68	-	-	-
12_20090	3H	3.97	-	3	yes
12_10103	3H	3.97	-	-	-
12_30915	3H	14.91	-	-	-
12_31009	3H	50.85	-	non-coding	yes
12_11069	3H	53.2	-	1	no
12_30721	3H	55.22	-	-	-
12_30126	3H	67.86	-	-	-
12_31010	3H	67.86	-	-	-
12_31014	3H	67.86	-	non-coding	yes
12_11150	3H	75.27	-	non-coding	yes
12_30088	3H	75.27	-	3	no
12_30005	3H	77.37	-	non-coding	yes
12_11338	3H	134.71	-	non-coding	yes
12_30963	3H	137.74	-	non-coding	yes
12_31496	3H	150.11	-	non-coding	yes
12_21500	3H	178.25	-	non-coding	yes
12_21117	4H	0	-	1	no
12_11485	4H	10.77	-	-	-
12_10626	4H	26.2	-	non-coding	yes
12_30907	4H	30.37	-	non-coding	yes
11_10261	4H	53.94	-	-	-
12_31156	4H	56.22	-	-	-
12_21137	4H	57.54	-	3	yes
12_30718	4H	102.26	-	-	-

12_30046	4H	105.83	-	-	-
12_11183	4H	111.81	-	3	yes
12_20059	5H	45.64	-	1	no
12_31062	5H	51.51	-	3	no
12_31064	5H	51.51	Pepc	1	no
11_20018	5H	93.66	-	2	no
11_21061	5H	99.39	-	1	no
12_11245	5H	113.51	-	-	-
12_11298	5H	115.45	-	-	-
12_20045	5H	119.65	-	non-coding	yes
12_11472	5H	123.98	-	non-coding	yes
12_10273	5H	157.61	NAC	non-coding	yes
11_20826	5H	162.03	-	-	-
12_31509	6H	59.25	-	3	yes
12_30681	6H	60.65	-	non-coding	yes
12_30581	7H	79.08	-	3	yes
12_11536	7H	81.78	-	1	no
12_10268	7H	81.78	LOC100825330	non-coding	yes
12_31000	7H	81.78	-	non-coding	yes
12_11477	7H	81.78	-	non-coding	yes
12_11091	7H	85.28	-	3	no
12_11055	7H	85.28	-	1	no
12_11529	7H	85.28	-	-	-
11_20349	7H	86.84	-	-	-
12_30419	7H	91.67	LOC100824781	non-coding	yes
11_11521	7H	127.74	-	1	no

Table S 2.4 Annotations for SNPs having outlier F_{ST} values in the breeding programs comparison

Table S 2.5 Annotations for SNPs having outlier F_{ST} values in the growth habit comparison

Table S 2.6 Annotations for SNPs having outlier F_{ST} values in the row-type comparison

Table S 2.7 Annotations for SNPs outliers on the pairwise haplotype sharing analysis.

Table S 2.8 Number of SNPs with significant PHS values shared between breeding populations

	Busch Ag.(2)	Busch Ag.(Int.)	Montana (2)	North Dakota(2)	Oregon (2)	Utah (2)	Washington (2)	Idaho (6)	Busch Ag.(6)	Minnesota (6)	North Dakota(6)	Oregon (6)	Utah (6)	Virginia (6)	Washington (6)
Idaho(2)	9	3	10	0	14	1	6	0	0	1	1	2	1	0	1
Busch Ag.(2)	-	5	4	0	2	1	1	0	2	0	3	1	1	3	1
Busch Ag.(Int.)	-	-	3	1	1	3	2	3	1	3	4	3	1	4	5
Montana(2)	-	-	-	1	3	5	3	1	6	3	1	4	1	8	0
North Dakota(2)	-	-	-	-	1	5	1	2	0	0	3	3	2	5	5
Oregon(2)	-	-	-	-	-	0	0	1	0	0	1	3	0	2	0
Utah(2)	-	-	-	-	-	-	2	2	1	4	7	8	5	6	5
Washington(2)	-	-	-	-	-	-	-	0	2	0	0	1	0	0	2
Idaho(6)	-	-	-	-	-	-	-	-	1	0	1	10	1	0	1
Busch Ag.(6)	-	-	-	-	-	-	-	-	-	0	3	2	4	5	17
Minnesota(6)	-	-	-	-	-	-	-	-	-	-	2	0	0	1	3
North Dakota(6)	-	-	-	-	-	-	-	-	-	-	-	4	7	4	9
Oregon(6)	-	-	-	-	-	-	-	-	-	-	-	-	22	2	3
Utah(6)	-	-	-	-	-	-	-	-	-	-	-	-	-	2	5
Virginia(6)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5

Table S 2.9 SNPs outliers in the PHS analysis at each population. Including the PHS value, and haplotype frequency and length.

Table S 2.10 Mean and standard deviation for *c* in spring two-row and six-row barley breeding programs.

	Mean *c*	Standard Deviation
Two-row		
University of Idaho	0.043	0.002
Busch Agricultural Resources Inc.	0.186	0.005
Busch Agricultural Resources (International)	0.185	0.006
Montana State University	0.094	0.003
North Dakota State University	0.256	0.005
Utah State University	0.316	0.007
Washington State University	0.116	0.004
Average	**0.171**	**0.005**
Six-row		
University of Idaho	0.159	0.006
Busch Agricultural Resources Inc.	0.273	0.007
University of Minnesota	0.344	0.008
North Dakota State University	0.282	0.007
Utah State University	0.150	0.004
Washington State University	0.178	0.006
Average	**0.231**	**0.006**

Table S 2.11 Frequency of identity by state segments for 50 SNPs windows between pair of populations at each chromosomal segment.

Table S 2.12 Frequency of identity by state segments for 100 SNPs windows between pair of populations at each chromosomal segment.

Table S 2.13 Identity by state segments in 50 SNPs windows.

Pairwise genome-wide proportions of IBS segments shared between populations and average frequency of IBS segments in each population.

	Population 2														
Population 1	**Idaho (6)**	**Busch Ag.(2)**	**Busch Ag.(6)**	**Busch Ag. (Int.)**	**Minnesota (6)**	**Montana (2)**	**North Dakota (2)**	**North Dakota (6)**	**Oregon (2)**	**Oregon (6)**	**Utah (2)**	**Utah (6)**	**Virginia (6)**	**Washington (2)**	**Washington (6)**
Idaho (2)	0.79 (0.08/0.44)	1 (0.78/0.9)	1 (0.33/0.43)	1 (0.67/0.79)	0.88 (0.16/0.52)	1 (0.79/0.85)	1 (0.64/0.62)	0.88 (0.16/0.41)	1 (0.49/0.32)	1 (0.54/0.17)	1 (0.39/0.44)	1 (0.68/0.28)	0.6 (0.15/0.04)	1 (0.85/0.85)	0.98 (0.3/0.36)
Idaho (6)	--	0.98 (0.47/0.07)	1 (0.74/0.93)	0.38 (0.17/0.14)	1 (0.72/0.9)	1 (0.53/0.09)	0.81 (0.54/0.19)	1 (0.76/0.85)	0.92 (0.42/0.1)	1 (0.63/0.23)	0.69 (0.27/0.28)	1 (0.83/0.48)	0.85 (0.48/0.08)	0.6 (0.35/0.1)	1 (0.7/0.75)
Busch Ag.(2)	--	--	1 (0.46/0.63)	1 (0.88/0.88)	0.98 (0.12/0.66)	1 (0.81/0.73)	1 (0.62/0.44)	0.98 (0.14/0.46)	0.96 (0.53/0.31)	0.98 (0.52/0.21)	0.9 (0.36/0.42)	1 (0.7/0.31)	0.75 (0.11/0.06)	1 (0.87/0.76)	0.96 (0.28/0.48)
Busch Ag.(6)	--	--	--	0.9 (0.08/0.3)	1 (0.92/0.91)	1 (0.66/0.32)	1 (0.53/0.29)	1 (0.95/0.86)	0.94 (0.48/0.17)	1 (0.78/0.22)	0.58 (0.4/0.37)	1 (0.91/0.37)	0.88 (0.54/0.03)	1 (0.25/0.25)	1 (0.87/0.74)
Busch Ag. (Int.)	--	--	--	--	0.44 (0.27/0.19)	1 (0.7/0.62)	1 (0.53/0.35)	0.4 (0.33/0.18)	0.85 (0.51/0.34)	0.9 (0.47/0.15)	0.77 (0.39/0.33)	1 (0.57/0.12)	0.46 (0.26/0.13)	1 (0.79/0.59)	0.69 (0.34/0.16)
Minnesota (6)	--	--	--	--	--	1 (0.7/0.16)	0.85 (0.64/0.25)	1 (0.91/0.88)	0.85 (0.52/0.12)	1 (0.78/0.2)	0.46 (0.46/0.38)	1 (0.9/0.34)	0.81 (0.53/0.04)	0.79 (0.33/0.14)	1 (0.81/0.72)
Montana (2)	--	--	--	--	--	--	1 (0.68/0.59)	1 (0.15/0.55)	0.98 (0.51/0.33)	1 (0.52/0.22)	1 (0.44/0.45)	1 (0.74/0.35)	0.67 (0.14/0.07)	1 (0.85/0.85)	1 (0.33/0.51)
North Dakota (2)	--	--	--	--	--	--	--	0.94 (0.22/0.5)	0.96 (0.38/0.35)	1 (0.42/0.26)	0.94 (0.31/0.45)	1 (0.52/0.32)	0.71 (0.17/0.19)	1 (0.56/0.63)	0.94 (0.22/0.48)
North Dakota (6)	--	--	--	--	--	--	--	--	0.85 (0.45/0.13)	1 (0.6/0.21)	0.58 (0.33/0.32)	1 (0.79/0.37)	0.83 (0.45/0.05)	0.79 (0.3/0.15)	1 (0.74/0.7)
Oregon (2)	--	--	--	--	--	--	--	--	--	1 (0.62/0.7)	0.81 (0.18/0.36)	1 (0.33/0.27)	0.98 (0.35/0.36)	0.98 (0.37/0.41)	0.85 (0.19/0.41)
Oregon (6)	--	--	--	--	--	--	--	--	--	--	0.75 (0.11/0.41)	1 (0.34/0.38)	1 (0.59/0.59)	0.98 (0.16/0.43)	1 (0.18/0.62)
Utah (2)	--	--	--	--	--	--	--	--	--	--	--	0.98 (0.82/0.4)	0.44 (0.32/0.09)	0.98 (0.43/0.5)	0.71 (0.31/0.29)
Utah (6)	--	--	--	--	--	--	--	--	--	--	--	--	0.96 (0.26/0.2)	1 (0.24/0.73)	1 (0.34/0.77)
Virginia (6)	--	--	--	--	--	--	--	--	--	--	--	--	--	0.69 (0.07/0.14)	0.85 (0.05/0.5)
Washington (2)	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1 (0.33/0.24)

Notation: Proportion of genome sharing IBS segments (average frequency of IBS segment in Population 1/average Frequency of IBS segment in Population 2)

Table S 2.14 Identity by state segments in 100 SNPs windows.

Pairwise genome-wide proportions of IBS segments shared between populations and average frequency of IBS segments in each population.

	Population 2														
Population 1	**Idaho (6)**	**Busch Ag.(2)**	**Busch Ag.(6)**	**Busch Ag. (Int.)**	**Minnesota (6)**	**Montana (2)**	**North Dakota (2)**	**North Dakota (6)**	**Oregon (2)**	**Oregon (6)**	**Utah (2)**	**Utah (6)**	**Virginia (6)**	**Washington (2)**	**Washington (6)**
Idaho (2)	0.65 (0.08/0.35)	1 (0.66/0.84)	1 (0.24/0.32)	1 (0.57/0.68)	0.74 (0.09/0.41)	1 (0.69/0.77)	1 (0.48/0.49)	0.74 (0.05/0.27)	0.91 (0.41/0.23)	1 (0.39/0.09)	0.91 (0.33/0.42)	1 (0.52/0.19)	0.35 (0.01/0)	1 (0.76/0.79)	1 (0.18/0.23)
Idaho (6)	--	1 (0.38/0.04)	1 (0.69/0.93)	0.13 (0.06/0.33)	1 (0.68/0.89)	1 (0.42/0.07)	0.57 (0.42/0.08)	1 (0.69/0.79)	0.83 (0.31/0.05)	1 (0.5/0.15)	0.43 (0.26/0.28)	1 (0.73/0.37)	0.78 (0.32/0.01)	0.35 (0.03/0.1)	1 (0.61/0.72)
Busch Ag.(2)	--	--	1 (0.32/0.55)	1 (0.82/0.82)	0.96 (0.05/0.58)	1 (0.71/0.57)	0.96 (0.48/0.3)	1 (0.03/0.3)	0.83 (0.45/0.2)	0.96 (0.35/0.11)	0.74 (0.3/0.34)	1 (0.53/0.2)	0.52 (0.01/0)	1 (0.79/0.64)	0.96 (0.16/0.37)
Busch Ag.(6)	--	--	--	0.78 (0.01/0.21)	1 (0.92/0.9)	1 (0.57/0.22)	0.78 (0.38/0.17)	1 (0.94/0.85)	0.78 (0.43/0.12)	1 (0.69/0.15)	0.48 (0.32/0.27)	1 (0.86/0.27)	0.7 (0.51/0)	0.87 (0.02/0.17)	1 (0.83/0.68)
Busch Ag. (Int.)	--	--	--	--	0.26 (0.24/0.02)	1 (0.55/0.48)	0.87 (0.41/0.26)	0.3 (0.09/0.01)	0.83 (0.36/0.21)	0.83 (0.31/0.07)	0.65 (0.31/0.25)	1 (0.36/0.05)	0.17 (0.06/0.12)	1 (0.63/0.46)	0.57 (0.27/0.11)
Minnesota (6)	--	--	--	--	--	1 (0.59/0.08)	0.7 (0.48/0.08)	1 (0.89/0.84)	0.65 (0.48/0.06)	1 (0.68/0.13)	0.3 (0.45/0.29)	1 (0.85/0.27)	0.65 (0.46/0)	0.48 (0.02/0.09)	1 (0.77/0.68)
Montana (2)	--	--	--	--	--	--	1 (0.53/0.48)	1 (0.05/0.41)	0.96 (0.39/0.22)	1 (0.34/0.12)	0.91 (0.38/0.37)	1 (0.57/0.24)	0.52 (0.01/0.01)	1 (0.77/0.77)	1 (0.24/0.4)
North Dakota (2)	--	--	--	--	--	--	--	0.7 (0.08/0.35)	0.87 (0.25/0.25)	0.91 (0.25/0.15)	0.78 (0.19/0.36)	0.96 (0.35/0.18)	0.43 (0.05/0.08)	1 (0.43/0.47)	0.83 (0.09/0.29)
North Dakota (6)	--	--	--	--	--	--	--	--	0.65 (0.32/0.05)	1 (0.46/0.13)	0.3 (0.38/0.3)	1 (0.69/0.28)	0.7 (0.31/0.01)	0.43 (0.07/0.04)	1 (0.67/0.64)
Oregon (2)	--	--	--	--	--	--	--	--	--	1 (0.46/0.58)	0.74 (0.13/0.27)	0.96 (0.2/0.15)	0.87 (0.21/0.22)	0.96 (0.27/0.28)	0.7 (0.12/0.33)
Oregon (6)	--	--	--	--	--	--	--	--	--	--	0.52 (0.08/0.4)	1 (0.19/0.27)	1 (0.48/0.45)	0.91 (0.07/0.27)	0.96 (0.12/0.53)
Utah (2)	--	--	--	--	--	--	--	--	--	--	--	0.96 (0.76/0.29)	0.17 (0.23/0)	0.91 (0.31/0.43)	0.52 (0.21/0.22)
Utah (6)	--	--	--	--	--	--	--	--	--	--	--	--	0.74 (0.16/0.08)	1 (0.11/0.59)	1 (0.25/0.68)
Virginia (6)	--	--	--	--	--	--	--	--	--	--	--	--	--	0.33 (0.01/0.01)	0.87 (0.01/0.31)
Washington (2)	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1 (0.25/0.09)

Notation: Proportion of genome sharing IBS segments (average frequency of IBS segment in Population 1/average Frequency of IBS segment in Population 2)

Table S 2.15 Pairwise F_{ST} between breeding programs. Row-type is shown in parenthesis.

	Idaho (2)	Idaho (6)	BuschAg (2)	BuschAg (6)	BuschAg (Int.)	Minnesota (6)	Montana (2)	North Dakota (2)	North Dakota (6)	Oregon (2)	Oregon (6)	Utah (2)	Utah (6)	Virginia (6)	Washington (2)	Washington (6)
Idaho (2)	-0.004	0.518	0.121	0.587	0.102	0.652	0.054	0.247	0.586	0.269	0.393	0.249	0.339	0.487	0.111	0.489
Idaho (6)	--	-0.007	0.607	0.121	0.572	0.179	0.516	0.457	0.169	0.497	0.361	0.447	0.207	0.516	0.554	0.162
BuschAg (2)	--	--	-0.006	0.679	0.129	0.726	0.173	0.348	0.660	0.370	0.463	0.398	0.415	0.549	0.225	0.598
BuschAg (6)	--	--	--	-0.007	0.673	0.150	0.578	0.518	0.138	0.603	0.437	0.574	0.274	0.589	0.616	0.210
BuschAg (Int.)	--	--	--	--	-0.017	0.733	0.152	0.344	0.651	0.282	0.411	0.328	0.371	0.506	0.166	0.556
Minnesota (6)	--	--	--	--	--	-0.003	0.639	0.583	0.253	0.676	0.512	0.654	0.359	0.642	0.667	0.334
Montana (2)	--	--	--	--	--	--	-0.003	0.228	0.577	0.289	0.404	0.261	0.342	0.494	0.101	0.491
North Dakota (2)	--	--	--	--	--	--	--	-0.003	0.510	0.356	0.389	0.318	0.308	0.490	0.343	0.438
North Dakota (6)	--	--	--	--	--	--	--	--	-0.003	0.589	0.444	0.538	0.287	0.575	0.613	0.233
Oregon (2)	--	--	--	--	--	--	--	--	--	-0.014	0.239	0.312	0.319	0.363	0.317	0.470
Oregon (6)	--	--	--	--	--	--	--	--	--	--	-0.004	0.327	0.259	0.254	0.435	0.354
Utah (2)	--	--	--	--	--	--	--	--	--	--	--	-0.035	0.156	0.428	0.297	0.424
Utah (6)	--	--	--	--	--	--	--	--	--	--	--	--	-0.003	0.365	0.379	0.203
Virginia (6)	--	--	--	--	--	--	--	--	--	--	--	--	--	-0.003	0.508	0.519
Washington (2)	--	--	--	--	--	--	--	--	--	--	--	--	--	--	-0.003	0.539
Washington (6)	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	-0.033

Table S 2.16 Population relatedness best explained by introgression.

Populations involved in trees with significant ($p<0.05$) f_4-values.

P1	P2	P3	P4
((Utha(2),Utha(6));(six-row,six-row))			
UT2	UT6	AB6	BA6
UT2	UT6	AB6	MN6
UT2	UT6	BA6	N6
UT2	UT6	BA6	WA6
UT2	UT6	MN6	N6
UT2	UT6	MN6	WA6
((North Dakota (2-row), two-row); (six-row, six-row))			
N2	AB2	BA6	AB6
N2	AB2	BA6	WA6
N2	AB2	AB6	MN6
N2	AB2	AB6	UT6
N2	AB2	MN6	WA6
N2	AB2	MN6	UT6
N2	AB2	WA6	N6
N2	AB2	UT6	N6
N2	AB2	UT6	WA6
N2	BAI2	BA6	AB6
N2	BAI2	MN6	AB6
N2	BAI2	WA6	BA6
N2	BAI2	UT6	BA6
N2	BAI2	UT6	AB6
N2	BAI2	MN6	WA6
N2	BAI2	MN6	UT6
N2	BAI2	N6	WA6
N2	BAI2	N6	UT6
N2	BAI2	WA6	UT6
N2	BA2	N6	BA6
N2	BA2	WA6	BA6
N2	BA2	UT6	BA6
N2	BA2	AB6	UT6
N2	BA2	MN6	UT6

N2	BA2	N6	WA6
N2	BA2	N6	UT6
N2	BA2	WA6	UT6
N2	MT2	BA6	AB6
N2	MT2	AB6	MN6
N2	MT2	BA6	WA6
N2	MT2	BA6	UT6
N2	MT2	MN6	WA6
N2	MT2	MN6	UT6
N2	MT2	N6	WA6
N2	MT2	N6	UT6
N2	MT2	WA6	UT6
N2	VT2	BA6	UT6
N2	VT2	AB6	UT6
N2	VT2	MN6	UT6
N2	VT2	N6	UT6
N2	WA2	AB6	BA6
N2	WA2	AB6	MN6
N2	WA2	AB6	N6
N2	WA2	BA6	WA6
N2	WA2	BA6	UT6
N2	WA2	AB6	UT6
N2	WA2	N6	UT6
N2	WA2	WA6	UT6
N2	WA2	MN6	UT6
N2	WA2	N6	WA6

2.5.3 Supplementary Materials and Methods

2.5.3.1 Plant materials

We note that 10 samples from Oregon (STAB7-079, OR76-071, F6-1, F6-2, F6-3, F6-4,F6-6,F6-7,F6-8 and G388) were mislabeled in the T3 data base as two-row (Dr. Patrick Hayes, personal communication). In this manuscript we classified them correctly as six-row type. There were three accessions removed from the data set because they

where the only ones representing a population: one accession from Montana and two accessions from Oregon two-row Oregon)

2.5.3.2 Nicholson's c

This approach uses a Markov Chain Monte Carlo model (MCMC) to estimate allele frequency changes at each population from a comparison to a common ancestral allele frequency (Nicholson et al., 2002). The model explicitly includes the SNP ascertainment process and corrects allele frequency differentiation influenced by it. The differences in allele frequencies using *c* are assumed to be the results of demographic events rather than selection, however the fitting of the model could result in the discovery of loci targets of selection (Nicholson et al., 2002). The model assumes a star phylogeny in which populations split and evolve in subsequent isolation under drift alone.

2.5.3.3 Four-population test

The idea of the *f_4-test* is that for each set of four populations A, B, C, and D there are three possible unrooted trees that best describe the relationship in the absence of gene flow: ((A,B),(C,D)), ((A,C),(B,D)), and ((A,D),(B,C)). If the assumed topology is ((A,B),(C,D)) the f_4 statistic is calculated as the product of the difference of allele frequencies between A and B, and between C and D. Thus for one SNP $f_4 = (p_A\text{-}p_B) \times (p_C\text{-}p_D)$. Genome-wide the f_4 is just the mean of the *f_4-values* at every SNP. If the tree topology was correct, then the allele frequency between pairs of populations should be uncorrelated resulting in $f_4 = 0$. Significant deviations from zero indicate that the tree evaluated does not fit the data. The significance of deviations from zero was assessed by

calculating a normally distributed Z-score, using a Block Jackknife to obtain a standard error correcting for linkage disequilibrium among SNPs. Having three trees being significantly different from zero implies that the relationship between the populations compared is more complex, suggesting the presence of gene flow between the populations involved in the trees.

3 BIBLIOGRAPHY

Abdel-Ghani, A. H., Parzies, H. K., Omary, A., & Geiger, H. H. (2004). Estimating the outcrossing rate of barley landraces and wild barley populations collected from ecologically different regions of Jordan. *Theor Appl Genet*, *109*(3), 588-595.

Ammerman, A. J., & Cavalli-Sforza, L. L. (1984). *The Neolithic transition and the genetics of populations in Europe.* Prinston, New Jersey: Prinston University Press.

Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol*, *13*(4), 969-980.

Boyd, W. J. R., Li, C. D., Grime, C. R., Cakir, M., Potipibool, S., Kaveeta, L. et al. (2003). Conventional and molecular genetic analysis of factors contributing to variation in the timing of heading among spring barley (*Hordeum vulgare L.*) genotypes grown over a mild winter growing season. *Crop Pasture Sci*, *54*(12), 1277-1301.

Briscoe, D., Stephens, J. C., & O'Brien, S. J. (1994). Linkage disequilibrium in admixed populations: applications in gene mapping. *J Hered*, *85*(1), 59-63.

Caicedo, A. L., Williamson, S. H., Hernandez, R. D., Boyko, A., Fledel-Alon, A., York, T. L. et al. (2007). Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet*, *3*(9), 1745-1756.

Cavalli-Sforza, L. L. (1966). Population structure and human evolution. *Proc. R. Soc. Lond. B*, 362-379.

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S. et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *P Natl Acad Sci USA*, *110*(20), 8057-8062.

Chakraborty, R., & Smouse, P. E. (1988). Recombination of haplotypes leads to biased estimates of admixture proportions in human populations. *Proc Natl Acad Sci USA*, *85*(9), 3071-3074.

Chakraborty, R., & Weiss, K. M. (1988). Admixture as a tool for finding linked genes

and detecting that difference from allelic association between loci. *Proc Natl Acad Sci USA*, *85*(23), 9119-9123.
Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*, 7.
Civan, P., Ivanicova, Z., & Brown, T. A. (2013). Reticulated origin of domesticated emmer wheat supports a dynamic model for the emergence of agriculture in the Fertile Crescent. *PloS ONE*, *8*(11), e81955.
Close, T. J., Bhat, P. R., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L. et al. (2009). Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*, *10*, 582.
Comadran, J., Ramsay, L., MacKenzie, K., Hayes, P., Close, T. J., Muehlbauer, G. et al. (2011a). Patterns of polymorphism and linkage disequilibrium in cultivated barley. *Theor Appl Genet*, *122*(3), 523-531.
Comadran, J., Russell, J. R., Booth, A., Pswarayi, A., Ceccarelli, S., Grando, S. et al. (2011b). Mixed model association scans of multi-environmental trial data reveal major loci controlling yield and yield related traits in *Hordeum vulgare* in Mediterranean environments. *Theor Appl Genet*, *122*(7), 1363-1373.
Comadran, J., Kilian, B., Russell, J., Ramsay, L., Stein, N., Ganal, M. et al. (2012). Natural variation in a homolog of *Antirrhinum* CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat Genet*, *44*(12), 1388-1392.
Cuesta-Marcos, A., Szucs, P., Close, T. J., Filichkin, T., Muehlbauer, G. J., Smith, K. P. et al. (2010). Genome-wide SNPs and re-sequencing of growth habit and inflorescence genes in barley: implications for association mapping in germplasm arrays varying in size and structure. *BMC Genomics*, *11*, ARTN 707.
de Meeus, T., & Goudet, J. (2007). A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infect Genet Evol*, *7*(6), 731-735.
Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature*, *418*(6898), 700-707.
Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*, *14*(8), 2611-2620.
Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567-1587.
Fang, Z., Gonzales, A. M., Clegg, M. T., Smith, K. P., Muehlbauer, G. J., Steffenson, B. J. et al. (2014). Two genomic regions contribute disproportionately to geographic differentiation in Wild Barley. *G3*, *4*(7), 1193-1203.
Fang, Z., Eule-Nashoba, A., Powers, C., Kono, T. Y., Takuno, S., Morrell, P. L. et al. (2013). Comparative analyses identify the contributions of exotic donors to disease

resistance in a barley experimental population. *G3*, *3*(11), 1945-1953.
Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405-1413.
Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, *78*(2), 737-756.
Fuller, D. Q., Willcox, G., & Allaby, R. G. (2011). Early agricultural pathways: moving outside the 'core area' hypothesis in Southwest Asia. *J Exp Bot*, *63*(2), 617-633.
Gerke, J., Edwards, J., Ke, G., Ross-Ibarra, J., & McMullen, M. D. (2014). The genomic impacts of drift and selection for hybrid performance in maize. *arXiv:1307.7313*.
Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F statistics. *Mol Ecol Notes*, *5*(1), 184-186.
Gusev, A., Palamara, P. F., Aponte, G., Zhuang, Z., Darvasi, A., Gregersen, P. et al. (2012). The architecture of long-range haplotypes shared within and across populations. *Mol Biol Evol*, *29*(2), 473-486.
Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L. et al. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Res*, *19*(2), 318-326.
Haake, V., Cook, D., Riechmann, J. L., Pineda, O., Thomashow, M. F., & Zhang, J. Z. (2002). Transcription factor CBF4 is a regulator of drought adaptation in *Arabidopsis*. *Plant Physiol*, *130*(2), 639-648.
Hamblin, M. T., Close, T. J., Bhat, P. R., Chao, S., Kling, J. G., Abraham, K. J. et al. (2010). Population structure and linkage disequilibrium in US barley germplasm: implications for association mapping. *Crop Sci*, *50*(2), 556-566.
Hanrahan, J. P., Eisen, E. J., & Lagates, J. E. (1973). Effects of population size and selection intensity of short-term response to selection for postweaning gain in mice. *Genetics*, *73*(3), 513-530.
Borchers, H. W. (2015). pracma: Practical Numerical Math Functions.
Harlan, J. R., & Zohary, D. (1966). Distribution of wild wheats and barley. *Science*, *153*(3740), 1074-1080.
Harris, D. R., & Gosden, C. (1996). The beginnings of agriculture in western Central Asia. In D. R. Harris (Ed.), *The origins and spread of agriculture and pastoralism in Eurasia* (pp. 370-389). London: University College of London.
Hellenthal, G., Busby, G. B., Band, G., Wilson, J. F., Capelli, C., Falush, D. et al. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747-751.
Herbik, A., Koch, G., Mock, H. P., Dushkov, D., Czihal, A., Thielmann, J. et al. (1999). Isolation, characterization and cDNA cloning of nicotianamine synthase from barley. A key enzyme for iron homeostasis in plants. *Eur J Biochem*, *265*(1), 231-239.
Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet*, *38*(6), 226-231.
Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet Res*, *8*(03), 269-294.
Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A. et al.

(2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat Genet*, *44*(2), 212-216.
Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J., & Ayala, F. J. (1994). Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics*, *136*(4), 1329-1340.
Hufford, M. B., Lubinksy, P., Pyhajarvi, T., Devengenzo, M. T., Ellstrand, N. C., & Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLoS Genet*, *9*(5), e1003477.
Innan, H., Zhang, K., Marjoram, P., Tavaré, S., & Rosenberg, N. A. (2005). Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics*, *169*(3), 1763-1777.
Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801-1806.
Jari Oksanen, F. Guillaume Blanchet, & Roeland Kindt, P. L., Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Helene Wagner. (2015). vegan: Community ecology package.
Jones, H., Leigh, F. J., Mackay, I., Bower, M. A., Smith, L. M. J., Charles, M. P. et al. (2008). Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the fertile crescent. *Mol Biol Evol*, *25*(10), 2211-2219.
Karsai, I., Szucs, P., Koszegi, B., Hayes, P. M., Casas, A., Bedo, Z. et al. (2008). Effects of photo and thermo cycles on flowering time in barley: a genetical phenomics approach. *J Exp Bot*, *59*(10), 2707-2715.
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S. et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647-1649.
Keinan, A., Mullikin, J. C., Patterson, N., & Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet*, *39*(10), 1251-1255.
Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: University Press, Cambridge, United Kingdom.
Komatsuda, T., & Mano, Y. (2002). Molecular mapping of the intermedium spike-c (int-c) and non-brittle rachis 1 (btr1) loci in barley (*Hordeum vulgare L.*). *Theor Appl Genet*, *105*(1), 85-90.
Komatsuda, T., Maxim, P., Senthil, N., & Mano, Y. (2004). High-density AFLP map of nonbrittle rachis 1 (btr1) and 2 (btr2) genes in barley (*Hordeum vulgare L.*). *Theor Appl Genet*, *109*(5), 986-995.
Komatsuda, T., Pourkheirandish, M., He, C., Azhaguvel, P., Kanamori, H., Perovic, D. et al. (2007). Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *P Natl Acad Sci USA*, *104*(4), 1424-1429.
Kono, T. J. Y., Seth, K., & Poland, J. A. (2013). SNPMeta: SNP annotation and SNP

Metadata collection without a reference genome. *Mol Ecol*.
Kristiansson, K., Naukkarinen, J., & Peltonen, L. (2008). Isolated populations and complex disease gene identification. *Genome Biol*, *9*(8), 109.
Laurie, D. A., Pratchett, N., Snape, J. W., & Bezant, J. H. (1995). RFLP mapping of five major genes and eight quantitative trait loci controlling flowering time in a winter× spring barley (Hordeum vulgare L.) cross. *Genome*, *38*(3), 575-585.
Lev, E., Kislev, M. E., & Bar-Yosef, O. (2005). Mousterian vegetal food in Kebara cave, Mt. Carmel. *J Archaeol Sci*, *32*(3), 475-484.
Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175-195.
Lister, D. L., Thaw, S., Bower, M. A., Jones, H., Charles, M. P., Jones, G. et al. (2009). Latitudinal variation in a photoperiod response gene in European barley: insight into the dynamics of agricultural spread from 'historic' specimens. *J Archaeol Sci*, *36*(4), 1092-1098.
Marchini, J. (2013). Statistical and Population Genetics.
Martin, J. M., Blake, T. K., & Hockett, E. A. (1991). Diversity among North American spring barley cultivars based on coefficients of parentage. *Crop Sci*, *31*(5), 1131-1137.
Mayer, K. F., Waugh, R., Langridge, P., Close, T. J., Wise, R. P., Graner, A. et al. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, *491*, 711-716.
Morrell, P. L., Buckler, E. S., & Ross-Ibarra, J. (2012). Crop genomics: advances and applications. *Nat Rev Genet*, *13*(2), 85-96.
Morrell, P. L., & Clegg, M. T. (2007). Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent. *P Natl Acad Sci USA*, *104*(9), 3289-3294.
Morrell, P. L., Gonzales, A. M., Meyer, K. K., & Clegg, M. T. (2014). Resequencing data indicate a modest effect of domestication on diversity in barley: A cultigen with multiple origins. *J Hered*, *105*(2), 253-264.
Muñoz-Amatriaín, M., Cuesta-Marcos, A., Endelman, J. B., Comadran, J., Bonman, J. M., Bockelman, H. E. et al. (2014). The USDA barley core collection: Genetic diversity, population structure, and potential for genome-wide association studies. *PloS ONE*, *9*(4), e94688.
Muñoz-Amatriaín, M., Moscou, M. J., Bhat, P. R., Svensson, J. T., Bartoš, J., Suchánková, P. et al. (2011). An improved consensus linkage map of barley based on flow-sorted chromosomes and single nucleotide polymorphism markers. *Plant Genome*, *4*(3), 238-249.
Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, Ó., Stefánsson, K., & Donnelly, P. (2002). Assessing population differentiation and isolation from single‐nucleotide polymorphism data. *J R Stat Soc Series B Stat Methodol*, *64*(4), 695-715.
Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nat Rev Genet*, *8*(11), 857-868.

Nordborg, M. (2000). Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics*, *154*(2), 923-929.
Nordborg, M., & Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends Genet*, *18*(2), 83-90.
Omberg, L., Salit, J., Hackett, N., Fuller, J., Matthew, R., Chouchane, L. et al. (2012). Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC Genet*, *13*, 49.
Ordon, F., Schiemann, A., & Friedt, W. (1997). Assessment of the genetic relatedness of barley accessions (*Hordeum vulgare* sl) resistant to soil-borne mosaic-inducing viruses (BaMMV, BaYMV, BaYMV-2) using RAPDs. *Theor Appl Genet*, *94*(3-4), 325-330.
Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, *2*(12), e190.
Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*, *8*(11), e1002967.
Poets, A. M., Fang, Z., Clegg, M. T., & Morrell, P. (2015). Barley landraces are characterized by geographically heterogeneous genomic origins. *Genome Biol*,*16(1),173*.
Pourkheirandish, M., & Komatsuda, T. (2007). The importance of barley genetics and domestication in a global perspective. *Annals of Botany*, *100*(5), 999-1008.
Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*, *20*(4), R208-15.
Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945-959.
Purugganan, M. D., & Fuller, D. Q. (2011). Archaeological data reveal slow rates of evolution during plant domestication. *Evolution*, *65*(1), 171-183.
Team, R. D. C. (2012). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.
Ralph, P., & Coop, G. (2010). Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics*, *186*(2), 647-668.
Ralph, P., & Coop, G. (2013). The geography of recent genetic ancestry across Europe. *PLoS Biol*, *11*(5), e1001555.
Ramsay, L., Comadran, J., Druka, A., Marshall, D. F., Thomas, W. T., Macaulay, M. et al. (2011). INTERMEDIUM-C, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene TEOSINTE BRANCHED 1. *Nat Genet*, *43*(2), 169-172.
Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*(7263), 489-494.
Robertson, A. (1970). A theory of limits in artificial selection with many linked loci. In *Mathematical topics in population genetics* (pp. 246-288). Springer.
Robertson, A. (1960). A theory of limits in artificial selection. *Proc R Soc B*, *153*(951), 234-249.

Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., Mitchell, S. E. et al. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc Natl Acad Sci USA*, *112*(12), 3823-3828.
Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet*, *1*(6), e70.
Ross-Ibarra, J., Morrell, P. L., & Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *P Natl Acad Sci USA*, *104*, 8641-8648.
Russell, J., Dawson, I. K., Flavell, A. J., Steffenson, B., Weltzien, E., Booth, A. et al. (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*(2), 564-578.
Saisho, D., & Purugganan, M. D. (2007). Molecular phylogeography of domesticated barley traces expansion of agriculture in the Old World. *Genetics*, *177*(3), 1765-1776.
Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hu Genet*, *78*(4), 629-644.
Shin, J.-H., Blay, S., McNeney, B., & Graham, J. (2006). LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J Stat Softw*, *16*, 1-9.
Skinner, J. S., von Zitzewitz, J., Szucs, P., Marquez-Cedillo, L., Filichkin, T., Amundsen, K. et al. (2005). Structural, functional, and phylogenetic characterization of a large CBF gene family in barley. *Plant Mol Biol*, *59*(4), 533-551.
Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, *236*(4803), 787-792.
Steffenson, B. J., Olivera, P., Roy, J. K., Jin, Y., Smith, K. P., & Muehlbauer, G. J. (2007). A walk on the wild side: mining wild wheat and barley collections for rust resistance genes. *Crop Pasture Sci*, *58*(6), 532-544.
Stephens, M., & Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, *73*(5), 1162-1169.
Stephens, M., Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, *68*(4), 978-989.
Taji, T., Ohsumi, C., Iuchi, S., Seki, M., Kasuga, M., Kobayashi, M. et al. (2002). Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *Plant J*, *29*(4), 417-426.
Takahashi, R., & Hayashi, J. (1964). Linkage study of two complementary genes for brittle rachis in barley. *Bericht des Ohara Instituts für Landwirtschaftliche Biologie*, *12*(2), 99-105.
Tanno, K.-, & Willcox, G. (2012). Distinguishing wild and domestic wheat and barley

spikelets from early Holocene sites in the Near East. *Veg Hist Archaeobot*, *21*(2), 107-115.
Thornton, K. (2003). libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, *19*(17), 2325-2327.
Thornton, K. R., Foran, A. J., & Long, A. D. (2013). Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. *PLoS Genet*, *9*(2), e1003258.
Toomajian, C., Hu, T. T., Aranzana, M. J., Lister, C., Tang, C., Zheng, H. et al. (2006). A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Biol*, *4*(5), e137.
Turck, F., Fornara, F., & Coupland, G. (2008). Regulation and identity of florigen: FLOWERING LOCUS T moves center stage. *Annu. Rev. Plant Biol.*, *59*, 573-594.
Turner, A., Beales, J., Faure, S., Dunford, R. P., & Laurie, D. A. (2005). The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Science*, *310*(5750), 1031-1034.
von Zitzewitz, J., Szucs, P., Dubcovsky, J., Yan, L., Francia, E., Pecchioni, N. et al. (2005). Molecular and structural characterization of barley vernalization genes. *Plant Mol Biol*, *59*(3), 449-467.
Wang, C., Szpiech, Z. A., Degnan, J. H., Jakobsson, M., Pemberton, T. J., Hardy, J. A. et al. (2010). Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat Appl Genet Mol Biol*, *9*, Article 13.
Wang, H., Smith, K. P., Combs, E., Blake, T., Horsley, R. D., & Muehlbauer, G. J. (2012). Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theor Appl Genet*, *124*(1), 111-124.
Weaver, J. C. (1943). Barley in the United States: A Historical Sketch. *Geographical Review*, *33*(1), 56-73.
Weaver, J. C. (1944). *United States Malting Barley Production* (34, No. 2). Taylor & Francis, Ltd. on behalf of the Association of American Geographers.
Weaver, J. C. (1950). *American barley production: A study in agricultural Geography*. Minneapolis, MN: Burgess Pub. Co.
Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358-1370.
Weiss, E., Kislev, M. E., & Hartmann, A. (2006). Autonomous cultivation before domestication. *Science*, *5780*, 1608.
Wiebe, G. A., & Reid, D. A. (1961). Classification of barley varieties growing in the United States and Canada in 1958. *U.S. Department of Agriculture*, 210.
Willcox, G. (2005). The distribution, natural habitats and availability of wild cereals in relation to their domestication in the Near East: multiple events, multiple centres. *Veg Hist Archaeobot*, *14*(4), 534-541.
Willcox, G. (2013). Anthropology. The roots of cultivation in southwestern Asia. *Science*, *341*(6141), 39-40.

Willcox, G. (1991). La culture inventée, la domestication inconsciente: le début de l'agriculture au Proche-Orient. *Travaux de la Maison de l'Orient*, *20*(1), 9-29.

Wright, M. H., Tung, C. W., Zhao, K., Reynolds, A., McCouch, S. R., & Bustamante, C. D. (2010). ALCHEMY: a reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations. *Bioinformatics*, *26*(23), 2952-2960.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*(1), 323-354.

Yan, L., Fu, D., Li, C., Blechl, A., Tranquilli, G., Bonafede, M. et al. (2006). The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc Natl Acad Sci U S A*, *103*(51), 19581-19586.

Youssef, H. M., Koppolu, R., & Schnurbusch, T. (2012). Re-sequencing of vrs1 and int-c loci shows that labile barleys (Hordeum vulgare convar. labile) have a six-rowed genetic background. *Genet Resour Crop Ev*, *59*(7), 1319-1328.

Zhou, H., Muehlbauer, G., & Steffenson, B. N. (2012). Population structure and linkage disequilibrium in elite barley breeding germplasm from the United States. *J Zhejiang Univ Sci B*, *13*(6), 438-451.

Zohary, D. (1969). The progenitors of wheat and barley in relation to domestication and agricultural dispersal in the Old World. In P. J. Ucko & G. Dimbleby (Eds.), *The domestication and exploitation of plants and animals* (pp. 47-66). Duckworth, London.

Zohary, D. (1970). Center of diversity and center of origin. In O. H. Frankel & E. Bennett (Eds.), *Genetic resources in plants - their exploration and conservation* (pp. 33-42). Oxford: Blackwell Scientific Publications.

Zohary, D., Hopf, M., & Weiss, E. (2012). *Domestication of plants in the Old World: The origin and spread of domesticated plants in south-west Asia, Europe, and the Mediterranean Basin* (4 ed.). Oxford: Oxford University Press.

4 APPENDIX: ADDITIONAL PUBLISHED WORK

In addition to the work described in this dissertation, I participated in five published scientific research in the course of completing this degree. The following paragraphs describe my role in each of the publications and include the abstract from each publication. On all of these publications my name appears as AM Gonzales.

1. Gonzales, A. M., Fang, Z., Durbin, M. L., Meyer, K. K., Clegg, M. T., & Morrell, P. L. (2012). Nucleotide sequence diversity of floral pigment genes in Mexican populations of *Ipomoea purpurea* (morning glory) accord with a neutral model of evolution. J Hered, 103(6), 863-872.

As the primary author of this publication I participated in the design of the project. I coordinated the collection of Sanger resequencing data with our partners at the University of California, Irvine. I carried out all the analysis, from DNA sequence assembling and alignment to the statistical analyses and interpretation of results. I wrote the majority of the manuscript.

Abstract

The common morning glory (*Ipomoea purpurea*) is an annual vine native to Central and Southern Mexico. The genetics of flower color polymorphisms and interactions with the biotic environment have been extensively studied in *I. purpurea* and in its sister species *I. nil*. In this study, we examine nucleotide sequence polymorphism in 11 loci, 9 of which are known to participate in a pathway that produces floral pigments. A sample of 30 *I. purpurea* accessions from the native range of Central and Southern Mexico comprise the data, along with one accession from each of the two sister species *I. alba* and *I. nil*. We observe moderate levels of nucleotide sequence polymorphism of ~1%. The ratio of recombination to mutation parameter estimates (ρ/θ) of ~2.5 appears consistent with a mixed-mating system. *Ipomoea* resequencing data from these genic regions are noteworthy in providing a good fit to the standard neutral model of molecular evolution. The derived silent site frequency spectrum is very close to that predicted by coalescent simulations of a drift-mutation process, and Tajima's D values are not significantly different from expectations under neutrality.

2. Fang, Z., Gonzales, A. M., Durbin, M. L., Meyer, K. K., Miller, B. H., Volz, K. M. et al. (2013). Tracing the geographic origins of weedy *Ipomoea purpurea* in the southeastern United States. J Hered, 104(5), 666-677

I participated in the design of the project. I coordinated the collection of Sanger resequencing data with our partners at the University of California, Irvine. I worked on the DNA sequence assembly and alignment for all 11 loci. I carried out the basic summary statistics within *I. nil* and estimations of diversity between *I. nil* and *I. purpurea.* I collaborated on the writing of the paper.

Abstract

Ipomoea purpurea (common morning glory) is an annual vine native to Mexico that is well known for its large, showy flowers. Humans have spread morning glories worldwide, owing to the horticultural appeal of morning glory flowers. *Ipomoea purpurea* is an opportunistic colonizer of disturbed habitats including roadside and agricultural settings, and it is now regarded as a noxious weed in the Southeastern US. Naturalized populations in the Southeastern United States are highly polymorphic for a number of flower color morphs, unlike native Mexican populations that are typically monomorphic for the purple color morph. Although *I. purpurea* was introduced into the United States from Mexico, little is known about the specific geographic origins of US populations relative to the Mexican source. We use resequencing data from 11 loci and 30 *I. purpurea* accessions collected from the native range of the species in Central and Southern Mexico and 8 accessions from the Southeastern United States to infer likely geographic origins in Mexico. Based on genetic assignment analysis, haplotype composition, and the degree of shared polymorphism, *I. purpurea* samples from the Southeastern United States are genetically most similar to samples from the Valley of Mexico and Veracruz State. This supports earlier speculation that *I. purpurea* in the Southeastern United States was likely to have been introduced by European colonists from sources in Central Mexico.

3. Mascher, M., Richmond, T. A., Gerhardt, D. J., Himmelbach, A. , Clissold, L., Sampath, D. , Ayling, S. , Steuernagel, B. , Pfeifer, M. , D'Ascenzo, M., Akhunov, E. D. , Hedley, P. , Gonzales, A. M., Morrell, P. L., Kilian, B., Blattner, F. R., Scholz, U., Mayer, K. F. X., Flavell, A. J., Muehlbauer, G. J., Waugh, R., Jeddeloh, J. A., and Stein, N. (2013). Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. The Plant Journal 76: 494-505.

In this project I retrieved from public repositories (*i.e.,* NCBI) complete gene sequences for all characterized loci in barley, as they could be used later to evaluate the performance of the exome-capture. This resulted in a set of ~400 genes. The sequences were compared with the contigs sequences that were used for the design of the exome-capture platform. Any characterized barley gene that was not present in the original exome capture design was added.

Abstract

Advanced resources for genome-assisted research in barley (*Hordeum vulgare*) including a whole-genome shotgun assembly and an integrated physical map have recently become available. These have made possible studies that aim to assess genetic diversity or to isolate single genes by whole-genome resequencing and in silico variant

detection. However such an approach remains expensive given the 5 Gb size of the barley genome. Targeted sequencing of the mRNA-coding exome reduces barley genomic complexity more than 50-fold, thus dramatically reducing this heavy sequencing and analysis load. We have developed and employed an in-solution hybridization-based sequence capture platform to selectively enrich for a 61.6 megabase coding sequence target that includes predicted genes from the genome assembly of the cultivar Morex as well as publicly available full-length cDNAs and de novo assembled RNA-Seq consensus sequence contigs. The platform provides a highly specific capture with substantial and reproducible enrichment of targeted exons, both for cultivated barley and related species. We show that this exome capture platform provides a clear path towards a broader and deeper understanding of the natural variation residing in the mRNA-coding part of the barley genome and will thus constitute a valuable resource for applications such as mapping-by-sequencing and genetic diversity analyzes.

4. Morrell, P. L., Gonzales, A. M., Meyer, K. K., & Clegg, M. T. (2014). Resequencing data indicate a modest effect of domestication on diversity in barley: A cultigen with multiple origins. J Hered, 105(2), 253-264.

My contribution to this project was primarily in data analysis. I calculated the descriptive statistics and estimators of nucleotide sequence diversity at silent sites for seven loci for partitions of the data. I was involved with the estimation of the recombination and mutation ratios. I assembled DNA sequences from *Hordeum bulbosum* that represented 22 distinct haplotypes from resequenced loci. This data was used to determine the ancestral state for mutations at each segregating site.

Abstract

The levels of diversity and extent of linkage disequilibrium in cultivated species are largely determined by diversity in their wild progenitors. We report a comparison of nucleotide sequence diversity in wild and cultivated barley (*Hordeum vulgare ssp. spontaneum and ssp. vulgare*) at 7 nuclear loci totaling 9296bp, using sequence from *Hordeum bulbosum* to infer the ancestral state of mutations. The sample includes 36 accessions of cultivated barley, including 23 landraces (cultivated forms not subject to modern breeding) and 13 cultivated lines and genetic stocks compared to either 25 or 45 accessions of wild barley for the same loci. Estimates of nucleotide sequence diversity indicate that landraces retain >80% of the diversity in wild barley. The primary population structure in wild barley, which divides the species into eastern and western populations, is reflected in significant differentiation at all loci in wild accessions and at 3 of 7 loci in landraces. "Oriental" landraces have slightly higher diversity than "Occidental" landraces. Genetic assignment suggests more admixture from Occidental

landraces into Oriental landraces than the converse, which may explain this difference. Based on θ_π for silent sites, modern western cultivars have ~73% of the diversity found in landraces and ~71% of the diversity in wild barley.

5. Fang, Z., Gonzales, A. M., Clegg, M. T., Smith, K. P., Muehlbauer, G. J., Steffenson, B. J. et al. (2014). Two genomic regions contribute disproportionately to geographic differentiation in wild barley. G3, 4(7), 1193-1203.

In this paper I was involved in the design, calculation of basic summary statistics, and contributed substantially to the writing.

Abstract

Genetic differentiation in natural populations is driven by geographic distance and by ecological or physical features within and between natural habitats that reduce migration. The primary population structure in wild barley differentiates populations east and west of the Zagros Mountains. Genetic differentiation between eastern and western populations is uneven across the genome and is greatest on linkage groups 2H and 5H. Genetic markers in these two regions demonstrate the largest difference in frequency between the primary populations and have the highest informativeness for assignment to each population. Previous cytological and genetic studies suggest there are chromosomal structural rearrangements (inversions or translocations) in these genomic regions. Environmental association analyses identified an association with both temperature and precipitation variables on 2H and with precipitation variables on 5H.