Population Structure and Linkage Disequilibrium in U.S. Barley Germplasm: Implications for Association Mapping

Martha T. Hamblin, Timothy J. Close, Prasanna R. Bhat, Shiaoman Chao, Jennifer G. Kling, K. Joseph Abraham, Tom Blake, Wynse S. Brooks, Blake Cooper, Carl A. Griffl ey, Patrick M. Hayes, David J. Hole, Richard D. Horsley, Donald E. Obert, Kevin P. Smith, Steven E. Ullrich, Gary J. Muehlbauer, and Jean-Luc Jannink*

ABSTRACT
Previous studies have shown that there is considerable population structure in cultivated barley (Hordeum vulgare L.), with the strongest structure corresponding to differences in row number and growth habit. U.S. barley breeding programs include six-row and two-row types and winter and spring types in all combinations. To facilitate mapping of complex traits in breeding germplasm, 1816 barley lines from 10 U.S. breeding programs were scored with 1536 single nucleotide polymorphism (SNP) genotyping assays. The number of SNPs segregating within breeding programs varied from 854 to 1398. Model-based analysis of population structure showed the expected clustering by row type and growth habit; however, there was additional structure, some of which corresponded to the breeding programs. The model that fit the data best had seven populations: three two-row spring, two six-row spring, and two six-row winter. Average linkage disequilibrium (LD) within populations decayed over a distance of 20 to 30 cM, but some populations showed long-range LD suggestive of admixture. Genetic distance (allele-sharing) between populations varied from 0.11 (six-row spring vs. six-row spring) to 0.45 (two-row spring vs. six-row spring). Analyses of pairwise LD revealed that the phase of allelic associations was not well correlated between populations, particularly when their allele-sharing distance was >0.2. These results suggest that pooling divergent barley populations for purposes of association mapping may be inadvisable.

Genome-wide association studies (GWAS) in populations of unrelated individuals provide an efficient way to map the genome.
locations of quantitative trait loci (QTL), particularly when high-throughput genotyping methods are used to score genetic markers. However, association mapping faces challenges that arise from the complex history of the populations employed, in contrast to linkage mapping populations whose allele frequencies and recombination history are determined by experimental design. In the case of domesticated plants, each crop’s ancestral population characteristics, domestication history, breeding history, and mating system will have generated distinct patterns of population structure, linkage disequilibrium (LD), and allele frequency distribution. Knowledge of these attributes is critical to achieve experimental designs of maximal power and resolution.

Cultivars of barley, whose wild form (Hordeum vulgare L. ssp. spontaneum K. Koch) is distributed throughout the Middle East, usually fall into distinct groups based on morphology, phenology, and end use: six row vs. two row, spring varieties vs. winter varieties, malting barley vs. feed barley. Several large studies of diversity and LD in cultivated barley, focusing on worldwide (Malysheva-Otto et al., 2006; Saisho and Purugganan, 2007) or European germplasm (Rostoks et al., 2006), have been published recently. These studies show that there is strong genetic structure within diverse collections of barley germplasm, with major divisions corresponding roughly to row number and growth habit (i.e., spring vs. winter varieties). The population genetic characteristics of these barley groups are expected to reflect their differences in breeding history, and possibly also differences in domestication history, since barley is thought to have been domesticated twice (Kilian et al., 2006; Morrell and Clegg, 2007).

Recently, association mapping in barley, using a candidate gene approach, has been successful at identifying variation at previously identified major photoperiod loci (Cockram et al., 2008; Jones et al., 2008; Stracke et al., 2009), in spite of the fact that variation at these loci is confounded with population structure (Zhao et al., 2007). However, the goal of GWAS, namely to identify all alleles underlying complex traits, is more challenging. Methodologies accounting for population structure in GWAS have been the subject of many papers in human and plant genetics (e.g., Marchini et al., 2004; Yu et al., 2006; Zhao et al., 2007), but the population genetic properties of barley are very different from those of natural populations such as humans and Arabidopsis, or an outbred species like maize (Zea mays L.). For example, in other analyses based on bi-allelic DNA variation, the highest average \( F_{ST} \) observed between human populations is about 0.15 (Hinds et al., 2005), while \( F_{ST} \) between Ethiopian and other barley populations ranges from 0.37 to 0.86 (Saisho and Purugganan, 2007).

The Barley Coordinated Agricultural Project (CAP; www.BarleyCAP.org [verified 22 Nov. 2009]) is a community effort to develop genomic and statistical tools for GWAS and integrate them into U.S. barley breeding programs (Waugh et al., 2009). To further this goal, a large collection of germplasm (3840 lines) from 10 U.S. breeding programs (the CAP population) is being genotyped for 3072 single nucleotide polymorphism (SNP) assays developed from alignments of expressed sequence tags (ESTs) sequenced in a small number of diverse lines (Kota et al., 2008; Rostoks et al., 2006). These 3072 SNP assays were selected from 4608 “pilot” SNPs that were validated first in a much smaller but more diverse panel, referred to as the Core panel. The patterns of population structure, LD, and allele frequency distribution in the Core are likely to differ from those of the breeding populations, which are also likely to differ from each other. In this study, based on the data available at the end of 2007, we identify population structure within 1816 lines of the CAP population and characterize those populations with respect to their allelic diversity and their patterns of LD. We also look at correlations of LD across populations, to assess whether QTL in different populations are likely to be associated with the same allele at linked markers. Finally, we compare the results of our analyses with similar analyses of other barley collections (Close et al., 2009).

MATERIALS AND METHODS

Germplasm and SNP Data

A table of germplasm in the Core panel is provided as Supplementary Table S1. The 1816 lines of the barley CAP germplasm were contributed by the 10 CAP participating programs. These programs constitute a large portion of the barley improvement research community in the United States and service the entire spectrum of barley growing regions and market end-uses. The University of Minnesota (K. Smith; MN), North Dakota State University (R. Horsley; N2, N6) breeding programs service the Upper Midwest production area and emphasize spring malting barley and to a lesser extent feed barley. Efforts are directed to both six- and two-rowed cultivars with the greater emphasis on six-rowed. The primary traits that are important for this region are yield, lodging, disease resistance (Fusarium head blight, net blotch, Septoria speckled leaf blotch, stem rust, and spot blotch), and malting quality. The expansion of barley production into central and western North Dakota has placed increased emphasis on grain protein concentration, plant height, and adaptation to more arid environments. The Oregon State University (P. Hayes; OR) program concentrates on winter barley cultivars for malting. The USDA-Aberdeen, Idaho (D. Obert; AB) and the Utah State University (D. Hole; UT) programs breed both spring and winter two- and six-rowed malting, feed, and food barleys. The Idaho program also develops types specifically for aquaculture. The Washington State University (S. Ullrich; WA) program develops spring two- and to a lesser extent six-rowed barley varieties for malt, feed, and food uses, including hulless, waxy, and proanthocyanidin-free types. In addition to quality, yield and resistance to lodging, stripe rust, and root rots are emphasized. The Montana State University (T. Blake;
MT) program develops spring two-rowed barley cultivars for both malt and feed uses. The important traits for this region are yield (including water-limiting conditions), lodging resistance, disease resistance (stripe rust, barley yellow dwarf virus, and net blotch), and malting and feed quality. Busch Agricultural Resources, Inc. (D.B. Cooper; BA) targets six-row cultivars to the same areas as the Minnesota and North Dakota programs and two-row cultivars to the Intermountain Region also serviced by the USDA and Montana programs. The Virginia Tech (C. Grifﬁey; VT) program breeds for the southeastern United States and focuses on winter barley for feed, food, and ethanol production. The important traits for this region are yield, test weight, seed quality, lodging resistance, disease resistance (leaf rust, net blotch, and powdery mildew), and hull adherence.

Single nucleotide polymorphism genotyping methods are described in Rostoks et al. (2006). A table of SNP assays is provided in Supplementary Table S2. The SNP data sets are available from The Hordeum Toolbox (http://www.hordeum-toolbox.org/ [verified 22 Nov. 2009]). Diversity arrays technology (DArT) genotyping is described in Jaccoud et al. (2001) and Wenzl et al. (2004).

Population Structure
Principal components analysis (PCA) was performed in R (R Development Core Team, 2009) using the function prcomp and a matrix of Manhattan distances among lines. Structure (Pritchard et al., 2000) analyses with the full data set (1816 individuals × 1415 markers) using the linkage model (because many of the markers were closely linked) performed poorly, failing to converge after very lengthy runs. We therefore devised the following strategy. Separate analyses were performed to determine the number of subpopulations (k) vs. the genome fraction of lines originating from each subpopulation. We wanted all markers to contribute to determining k but needed to run analyses with marker subsets so that the analyses would converge. We ran 40 analyses on markers randomly selected every 10 cM but making sure that no two markers were closer than 5 cM, leading to subsets of 116 markers. We used the no-linkage/admixture model with correlated allele frequencies across subpopulations. We judged models to have converged if they ranged in log posterior density by <200 from Markov chain Monte Carlo iterations 50,001 to 150,000. About 90% of models converged by this criterion. For each k value, we averaged the log posterior density of the converged models and performed the analysis suggested by Evanno et al. (2005) to determine the best value for k. Having determined the appropriate k, we wanted as many markers as possible to contribute to the estimation of line genome fractions originating from each subpopulation. We wanted to identify the largest marker subset for which all markers had a minor allele frequency of at least 0.01, and a linkage disequilibrium (r²) with any other marker that did not exceed 0.25. We developed a heuristic algorithm (to be presented in a later publication) that, in smaller data sets, reliably identiﬁed subsets close to the largest exhaustive-search result. A subset of 486 markers was identiﬁed that we used to run Structure (Pritchard et al., 2000) under the no-linkage/admixture model with 100,000 burn-in and 200,000 analysis iterations. Eight replicate runs of this model produced very similar likelihood estimates.

Summary Statistics
Allele frequencies, expected heterozygosity, and genetic distances were calculated in PowerMarker (Liu and Muse, 2005). Measures of LD (r and r²) were calculated in PLINK (Purcell et al., 2007) (http://pngu.mgh.harvard.edu/~purcell/plink/ [veriﬁed 22 Nov. 2009]). The vast majority of SNP genotypes were homozygous; heterozygous sites were treated as missing data.

Simulations
Simulated data sets for the ascertainment analysis were made by sampling 200 lines, with replacement, from the actual population samples, implemented in the programming language R.

RESULTS

Core Panel SNP Data Set
Three pilot 1536–SNP assay arrays were developed from SNPs that had been discovered as polymorphic in ESTs from a small number of diverse barley lines (Kota et al., 2008; Rostoks et al., 2006). These 4608 SNP assays were initially tested in a panel of 102 diverse barley lines (the Core panel; Supplementary Table S1) that emphasizes malting varieties (60%) and lines of North American origin (91%). Most of the DNAs from the Core panel were also hybridized with a DArT array (Jaccoud et al., 2001; Wenzl et al., 2004). The final data set had 3233 polymorphic markers (2517 SNPs and 716 DArTs), 2940 of which have minor allele frequency (MAF) ≥ 0.05. At the time of this analysis, 3131 of the markers had been mapped, for an average density of about 2.5 markers cM⁻¹. However, the distribution of these markers is far from uniform: presumably because of low rates of recombination in large genomic regions (Kunzel et al., 2000), 60% of adjacent marker pairs map to the same genetic location.

CAP Germplasm SNP Data Set
Based on the results in the Core panel, two final sets of 1536 SNPs each (Barley Oligonucleotide Pool Assay [BOPA] 1 and BOPA2) were developed for genotyping in a total of 3840 barley lines from 10 U.S. breeding programs (the CAP germplasm; www.BarleyCAP.org [veriﬁed 22 Nov. 2009]), representing the major types of barley important in U.S. agriculture. Although the BOPAs were chosen to maximize genome coverage, close to 50% of the adjacent SNP pairs with MAF ≥ 0.05 have the same genetic map position. Genotyping of 1816 lines (Years 1 and 2) of the CAP germplasm with BOPA1 was complete at the time of this analysis. After removal of markers that were monomorphic or had a failure rate >20%, the data set comprised 1416 SNPs, which are identiﬁed in Supplementary Table S2. The origins and characteristics of the breeding lines, as well as the number of SNPs segregating in lines from each program, are shown in Table 1.
Population Structure

Because barley is known to have strong population structure, we performed both distance-based and model-based analyses to assign the individual lines to population clusters (see Methods). When the Core germplasm was analyzed by PCA, Components 1 and 2 explained 75.9% and 11.7% of the variation, respectively. Most individuals fell into one of three clusters that corresponded to two-row, six-row spring, and six-row winter lines (data not shown). For the CAP germplasm, PCA Components 1 and 2 explained 76.5 and 15.7% of the variation, and about 85% of individuals fell into one of four fairly tight clusters that corresponded to six-row spring, six-row winter, and two groups of two-row spring lines. For both sets of germplasm, the first principal component separates the lines based on row number, while the second component separates them based on growth habit. One of the two-row clusters in the CAP germplasm is comprised entirely of lines from North Dakota; most of the lines from Oregon, and all the lines from Utah, fell outside the well-defined clusters identified in this analysis.

We also analyzed the CAP germplasm using the model-based method implemented in Structure (see Methods), which assigns each individual a membership coefficient for each cluster. The best fit to the data was a model of 10 populations without admixture, or seven populations with admixture. Interestingly, the 10 populations in the no-admixture model roughly corresponded to the 10 breeding programs (see Fig. 1A), and almost all individuals had very high probability of membership in one of the populations.

Under the admixture model, which is more reasonable for this sample, 763 of the 1816 lines had <80% membership in any of the seven clusters (Fig. 1B). In Table 2, we show the composition of these clusters, using either 50 or 80% membership as a criterion for assignment. Each population is largely composed of individuals sharing the same row number and growth habit. Using 50% membership as a criterion, most populations contain a small number (<9%) of “other” individuals; only seven of those individuals remain in the populations under the strict assignment criterion. For subsequent analyses, the term “population” refers to the seven groups identified by Structure, with individuals assigned by the less stringent criterion of >50% membership.

Properties of the Seven Populations

While 1416 SNPs are segregating in the set of 1816 lines, Table 2 shows that between 80 and 97% of the SNPs in the full sample are segregating in each population. There is also considerable variation in the allele frequency spectrum, as shown by differences in expected heterozygosity \((H_e)\). For example, in the six-row spring populations, almost half the SNPs are rare, producing much lower estimates of \(H_e\). These differences in SNP number and frequency reflect substantial genetic differentiation produced by differences in the size and breeding history of these populations.

The great majority of lines clustered according to row number and growth habit, but 77 lines (about 5%) were assigned to populations dominated by individuals of a different row type or growth habit. Most of these individuals, which presumably descended from wide crosses between the major barley groups, had <80% membership in the population to which they were assigned.

Ascertainment Bias

Comparisons of SNP number and frequency based on genotyping assays may not be an entirely accurate reflection of the differences in diversity between populations. This is because of using a small number of lines to discover polymorphic SNPs introduces ascertainment bias (Nielsen, 2000); that is, the SNPs do not necessarily capture the same fraction of diversity present in different sets of germplasm. Because of SNP ascertainment, there is the possibility that the low variation observed in the six-row spring CAP populations is due, at least in part, to this bias.

To try to evaluate the extent of any ascertainment bias, we used the Core panel to compare population-specific patterns of SNP and DArT markers. Although DArT markers are biased toward common alleles, since they were chosen for high polymorphic information content (PIC) values, the number of lines used to screen for polymorphism was much larger than the seven lines used to discover the BOPA SNPs. A DArT survey of 33 barley cultivars showed a median PIC value of 0.42, and under-representation of alleles only for two Southeast Asian cultivars (Wenzl et al., 2004). Table 3 shows, separately, the percentages of SNPs and DArTs that are fixed, at low MAF, or at MAF \(\geq 0.05\), in two-row and six-row spring populations of the Core panel. Because the sample sizes in these two populations are quite different, we also validated these results by simulating for each population four data sets of 200 individuals each (see Methods); the patterns in the simulated data were virtually identical to those reported in Table 3 (data not shown). In both the two-row and six-row populations, the percentage of fixed and low MAF alleles is smaller for the DArTs than for the SNPs, as expected because of the DArT bias toward common alleles. Comparing across populations, the two-row population has far fewer monomorphic markers of both types, but the patterns for DArTs and SNPs are similar. This suggests that, while the SNP data cannot be used to draw rigorously quantitative conclusions about the relative levels of variation in the CAP populations, population-specific bias is not extreme in the BOPA markers.

Relationships among Populations

Table 4 shows, for each pair of populations, the number of segregating SNPs in common and the genetic distance
as measured by shared alleles. As expected, the genetic distances are smaller between populations that share the same row number (2 or 6) and growth habit (winter [wi] or spring [sp]) [e.g., populations 2sp(BA) and 2sp(N2)] and are much larger in comparisons across those groups. But it also appears that at least some of the differences in genetic distance have been produced by breeding programs. For example, the two-row North Dakota sample, which formed a separate cluster in the PCA analysis, constitutes the vast majority of population 2sp(N2), and the six-row North Dakota sample forms the vast majority of population 6sp(N6). Population 2sp(N2) is both the most divergent within the three two-row populations, and the least divergent from the six-row populations, especially population 6sp(N6), the other population dominated by North Dakota germplasm.

**Linkage Disequilibrium**

Differences in effective population size and breeding history are expected to generate different patterns of LD,
which would affect the design of association studies using this material. In Fig. 2, we show the decay of average LD ($r^2$) with recombination distance across the genome in the total CAP sample and each of the seven populations. At shorter distances, there are large differences among the populations, with populations 2sp(WA) and 6sp(MN) having the highest LD and population 6wi(VT) and the total sample having the lowest. Linkage disequilibrium in all populations decays to a basal level over a distance of 20 to 30 cM, but the basal level differs among populations. The two populations with lowest LD at 0 cM have higher basal LD, while populations 6wi(OR) and 2sp(WA) have higher LD levels at all distances.

What Fig. 2 does not reveal is the large variance in $r^2$ in each of these bins. While average $r^2$ varies from ~0.4 to 0.8 for pairs of markers with the same genetic map position, Table 5 shows that even in populations with the highest average LD, almost 10% of those pairs have $r^2 \leq 0.1$. Conversely, some unlinked SNP pairs (50 cM apart) can have high LD, especially in the winter populations.

Correlation of Linkage Disequilibrium between Populations

Since the success of association studies depends on LD between marker alleles and alleles at unobserved QTL, it is important to know whether those associations are likely to be conserved across populations. One way to assess this is to find the correlation of $r$, the signed correlation coefficient, for the same pairs of SNPs in different populations (de Roos et al., 2008). One would expect correlations to be higher for closely linked SNP pairs in all populations, and to be maintained over greater recombination distance in more closely related populations. To test these hypotheses, we identified, for each of the 21 pairs of populations, the subset of SNPs that was segregating at MAF $\geq 0.05$ in both populations; the numbers of shared SNPs are shown in Table 4. In each population, we calculated $r$ for...
all SNP pairs on each chromosome and binned them by the recombination distance between the SNPs. We then found the correlation of \( r \) values between populations for each bin. Figure 3 shows the correlation of \( r \) for SNPs at 0 and 5 cM apart, for all 21 pairs of populations, as a function of the allele-sharing genetic distance between the populations (Table 4). There is a significantly negative relationship (\( t \) test, \( P < 0.0001 \)) between allele-sharing distance between populations and the correlation of \( r \) for tightly linked SNPs (0 cM). For SNP pairs 5 cM apart, the relationship is not significant (\( t \) test, \( P = 0.12 \)), and several of the correlations are negative. A Wilcoxon signed-rank test of the difference in correlation at 0 and 5 cM has a \( P \) value of <0.0001. Figure 3 also shows the correlation of \( r \) for a pair of populations that are very closely related; these two populations were created by randomly splitting population 2sp(BA) into halves. As expected, the correlation for this pair was very high, and remained high even at 20 cM (data not shown).

Table 5. Distribution of \( r^2 \).

<table>
<thead>
<tr>
<th>Summary statistic</th>
<th>Population*</th>
<th>Mean ( r^2 ) at 0 cM</th>
<th>% ( r^2 \leq 0.1 ) at 0 cM</th>
<th>Mean ( r^2 ) at 50 cM</th>
<th>% ( r^2 &gt; 0.1 ) at 50 cM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2sp(BA)</td>
<td>2sp(N2)</td>
<td>2sp(WA)</td>
<td>6sp(MN)</td>
<td>6sp(N6)</td>
</tr>
<tr>
<td>Mean ( r^2 ) at 0 cM</td>
<td>0.556</td>
<td>0.57</td>
<td>0.679</td>
<td>0.672</td>
<td>0.616</td>
</tr>
<tr>
<td>% ( r^2 \leq 0.1 ) at 0 cM</td>
<td>17.9</td>
<td>19.4</td>
<td>14.2</td>
<td>13.8</td>
<td>16.3</td>
</tr>
<tr>
<td>Mean ( r^2 ) at 50 cM</td>
<td>0.019</td>
<td>0.015</td>
<td>0.042</td>
<td>0.014</td>
<td>0.022</td>
</tr>
<tr>
<td>% ( r^2 &gt; 0.1 ) at 50 cM</td>
<td>1.7</td>
<td>1.9</td>
<td>8.9</td>
<td>0.5</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>6wi(VT)</td>
<td>6wi(OR)</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ( r^2 ) at 0 cM</td>
<td>0.379</td>
<td>0.628</td>
<td>0.387</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% ( r^2 \leq 0.1 ) at 0 cM</td>
<td>31.2</td>
<td>17</td>
<td>21.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ( r^2 ) at 50 cM</td>
<td>0.052</td>
<td>0.082</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% ( r^2 &gt; 0.1 ) at 50 cM</td>
<td>26</td>
<td>26</td>
<td>11.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The population name indicates the row number (2 or 6), the growth habit (winter [wi] or spring [sp]) of the majority of the barley lines that make up the population, and the breeding program that contributes the largest number of individuals to the population. Breeding program: BA, Busch Agric. Resources, Inc.; MN, Univ. of Minnesota; N2 and N6, North Dakota State Univ.; OR, Oregon State Univ.; VT, Virginia Tech; WA, Washington State Univ.

Figure 2. Mean \( r^2 \) by recombination distance in seven barley populations and the total sample. AB, USDA-Aberdeen, Idaho; BA, Busch Agric. Resources, Inc.; MN, Univ. of Minnesota; N2 and N6, North Dakota State Univ.; OR, Oregon State Univ.; VT, Virginia Tech; WA, Washington State Univ.; 2, two-row; 6, six-row; sp, spring; wi, winter; SNPs, single nucleotide polymorphisms.
DISCUSSION

Most association studies in crop plants have used populations representing global or regional diversity (see summary in Zhu et al., 2008), including accessions from germplasm banks; the barley CAP collection, in contrast, focuses almost entirely on breeding germplasm from U.S. programs. While this approach misses some of the diversity present in worldwide collections (>260,000 accessions; Matus and Hayes, 2002), the advantage is that any favorable alleles identified in the project will already be present in adapted lines, increasing the efficiency of subsequent breeding strategies. The breeding process has the potential to introduce new genetic structure, which can cause spurious associations (Stich et al., 2007), and we have found evidence of this in the CAP germplasm. It is important to understand the patterns of population structure, relatedness, and LD in this collection so that the methods used are adequate to account for them.

Population Structure

We found significant population structure associated with differences in row number and growth habit, as has been reported by others (e.g., Malyshova-Otto et al., 2006; Rostoks et al., 2006), as well as structure that appeared to have been generated by recent breeding history. This was particularly apparent when a no-admixture model was applied, results of which supported a model with 10 populations corresponding roughly to material from the 10 breeding programs (Fig. 1A). Even the UT lines, most of which were classified as “mixed” under the admixture model, formed their own population. Under the admixture model with broad assignment (Table 2), the N2, N6, VT, and OR samples contributed 78, 88, 76, and 98% of populations 2sp(N2), 6sp(N6), 6wi(VT), and 6wi(OR), respectively. When population assignment was strictly defined, only population 2sp(BA) was not dominated by lines from a particular program, consistent with a high level of exchange of two-row spring malting germplasm among the midwestern breeding programs.

Limited exchange among most of the other breeding programs could have produced the structure observed in Fig. 1. The VT breeding program, for example, has had very little exchange with other programs because of differences in growth habit, adaptation, and end use. It is also possible that local adaptation has contributed to this structure by generating differences in allele frequencies in these gene pools. If this were true, it might also explain why the six-row and two-row North Dakota gene pools are somewhat more similar (genetic distance = 0.33) than other six-row and two-row populations (genetic distance = 0.36–0.45).

The Effect of Breeding History on Diversity

The populations inferred by Structure analysis differed in the number and frequency distribution of SNPs segregating within them. According to a simplistic model of diversity in cultivated barley, we would expect to find the most diversity in the two-row winter types because these are the closest to the wild ancestor, *H. vulgare* ssp. *spontaneum*, which has two rows and requires vernalization (Pourkheirandish and Komatsuda, 2007). While this is not
the case, the converse is true: the least diverse populations are the six-row spring types. Since both the spring habit and six-row trait are derived, there would have been two selective bottlenecks associated with the establishment of this gene pool. However, the six-row trait appeared very early in barley domestication (Pourkheirandish and Komatsuda, 2007), and six-row spring barley predominated in Europe for several millennia. Backes et al. (2003), discussing European barley, suggest that six-row spring barley was once much more diverse, but that its replacement by two-row spring barley during the 18th and 19th centuries led to a loss of diversity.

While domestication history is important, several studies suggest that recent breeding history, dominated by the requirements of the malting and brewing industries, has had a more dramatic effect on the diversity of barley germplasm, including reduction of the diversity of the six-row spring types. According to Peel and Rasmussen (2000), “Breeding malting barley in the upper midwestern USA began with a relatively narrow genetic base and the core germplasm has become narrower with time.” Mikel and Kolb (2008) used pedigree data to estimate the coefficient of parentage (CP) for contemporary North American barley cultivars released between 1970 and 2006. They found the lowest diversity in six-row malting types (CP = 0.246), but the greatest diversity was in six-row feed types (CP = 0.026), which are not well represented in the CAP germplasm. Most of the six-row feed types in the CAP panel are winter lines from the VT program and spring lines from the UT program, which did not group with the six-row spring populations. Figure 1B shows that most of the UT lines have high membership (>0.8) in any of the seven populations of the admixture model. The UT lines trace back to North African lines, while the midwestern six-row lines trace back to Central Asia (Molina-Cano et al., 2002). Furthermore, no UT lines were included in the Core, so there was no optimization for SNPs segregating in this set of germplasm.

Another study (Condon et al., 2008) documents changes in simple sequence repeat allelic diversity in elite lines produced from 1958 to 1998 in the MN program, which breeds six-row spring malting barley. Comparing pedigree and marker data, they found that pedigree data always overestimated diversity, since parents assumed to be unrelated were actually related, and allelic diversity was reduced over the four decades. They also found evidence for fixation of alleles associated with selection on traits such as disease resistance and malting quality.

The two-row spring and six-row winter populations are more diverse than the six-row spring germplasm, but probably for very different reasons. The six-row winter lines come from only two breeding programs, OR and VT. The VT lines are primarily feed and food types, which means they have not been subject to the very strict selection criteria applied to malting types. The OR program focuses on malting types, but has used more diverse germplasm, including feed types. In addition, the OR samples have included some research materials such as genetic stocks.

The two-row spring lines are mostly malting types, but they come from many different programs, and constitute the largest group in the CAP panel (41%, by the broad definition in Table 2). According to Matus and Hayes (2002), the North American two-row malting varieties were selected from European introductions, but “…are distinct from the two-row malting accessions from Japan, Australia, and Europe. The requirements of the North American malting and brewing industries are quite different from those employed throughout the rest of the world.” Perhaps the greater diversity in two-row malting barley is due to the greater diversity in the European introductions from which it was selected, since European six-row barley had already lost most of its diversity.

**Ascertainment Bias and Limits to Population Genetic Inference**

It is well known that SNP genotyping data sets do not accurately reflect the allele frequency distributions observed in full resequencing data, and are skewed toward intermediate frequency alleles (Clark et al., 2005; Nielsen, 2000). There may also be population-specific bias due to the composition of the SNP discovery panel, compromising comparisons of diversity between populations. We observed differences in heterozygosity across the CAP populations, and at least some of that variation is consistent with patterns observed in other studies using other types of data (see previous section).

However, without a full resequencing data set from the same material for comparison, we cannot know how accurately those differences reflect the distribution of all SNPs segregating in those populations. Furthermore, the “bottom-up” population genetic approaches used to study the evolution of plant domestication (e.g., Ross-Ibarra et al., 2007; Wright and Gaut, 2005) depend heavily on information derived from the full allele frequency spectrum. As a consequence, this SNP data set, while very well suited for association mapping of QTL, could not easily be used for those sorts of analyses.

**Patterns of Linkage Disequilibrium**

Wild barley has remarkably low levels of LD for a self-pollinating plant (Morrell et al., 2005); the extent of LD in cultivated barley is much greater. Unlike populations at equilibrium, in which the decay of LD with genetic distance depends only on the relative rates of mutation and recombination, gene pools undergoing domestication experience dramatic changes in allele frequencies due
to genetic drift or selection, and some allelic combinations may be lost, generating new, more extensive LD. In a different process, when divergent materials are crossed, extensive “admixture LD” can also be generated, that is, LD that does not decay as a function of genetic distance. The effects of these events are seen in the patterns of average LD in the various populations making up the barley CAP germplasm (Fig. 2). All populations show extensive LD due to domestication and breeding, with \( r^2 \) decaying to basal levels only at distances of 20 to 30 cM. The higher basal levels in some populations, even at distances of 50 cM, represent admixture LD, as is seen most clearly for the total sample.

Some portion of the processes of selection and drift described above have occurred separately in the major groups of barley, resulting in different allelic combinations in those populations. In the most extreme case, when all four gametic types are not observed at a pair of linked SNP loci, alleles may be in coupling phase in one population and in repulsion phase in another. We found that, for populations that are not closely related, this has happened often enough that the correlation of \( r \) is very low or even negative (Fig. 3). Some of the negative correlations may reflect divergent selection on key traits such as row number and vernalization requirement. If the correlation between alleles at different marker loci is representative of the correlation between alleles at QTL and marker loci, these results suggest that combining divergent populations for purposes of QTL mapping may obscure associations between markers and QTL.

CONCLUSIONS

This set of 1816 individuals scored at 1416 SNPs does not represent the final CAP genotyping data set, which will consist of 3840 lines scored at about 3000 SNPs: each of the breeding programs represented in this study will contribute an additional 192 lines for genotyping, and all lines will be scored with a second BOPA array. However, it is unlikely that the fundamental observations reported here will change dramatically in the larger data set. Given the strong population structure, and the uncorrelated allelic associations between some populations, it will be necessary to test whether association mapping will be more successful in the full data set, or whether the more closely related populations should be pooled and tested separately. We are undertaking studies to address this question. Regardless of the outcome of those studies, because of the large number of lines in the CAP study, even subsets of the data should have sample sizes providing reasonable power to detect QTL of modest effect.

Supplemental Information Available

Supplemental information is available free of charge at http://crop.scijournals.org.

ACKNOWLEDGMENTS

Hiroyoshi Iwata and Peter Bradbury provided valuable comments during the course of this study. This research was supported by USDA-CSREES-NRI Grant no. 2006-55606-16722 “Barley Coordinated Agricultural Project: Leveraging Genomics, Genetics, and Breeding for Gene Discovery and Barley Improvement.” MTH and JLI designed the study, and analyzed and interpreted the data. MTH wrote the paper. PRB, SC, and TJ designed, implemented, and scored the SNP genotyping assays. JGK curated Barley CAP core pedigree and Barley CAP germplasm data. KJA wrote the algorithm used to select unassociated markers. TB, WSB, BC, CG, PH, DH, RH, DO, KS, and SEU selected the germplasm and contributed knowledge of it; they are listed in alphabetical order. GJM is Project Director of the Barley CAP project, which generated the data, and of which this work is a part.

REFERENCES


