Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Population structure and linkage disequilibrium in elite barley breeding germplasm from the United States*

Hao ZHOU[†1], Gary MUEHLBAUER[2], Brian STEFFENSON[1]

([1]*Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108, USA*)

([2]*Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA*)

[†]E-mail: zhoux222@umn.edu

Received Dec. 31, 2011; Revision accepted May 16, 2012; Crosschecked May 9, 2012

Abstract: Cultivated barley is known to have a complex population structure and extensive linkage disequilibrium (LD). To conduct robust association mapping (AM) studies of economically important traits in US barley breeding germplasm, population structure and LD decay were examined in a complete panel of US barley breeding germplasm (3840 lines) genotyped with 3072 single nucleotide polymorphisms (SNPs). Nine subpopulations (sp1–sp9) were identified by the program STRUCTURE and subsequently confirmed by principle component analysis (PCA). Out of the nine subpopulations, seven were very similar to the respective subpopulations identified by Hamblin *et al*. (2010) which were based on half of the germplasm and half of the SNP markers, but two subpopulations were found to be new. One subpopulation was dominated by six-rowed spring lines from Utah State University (UT) and the other was composed of six-rowed spring lines from multiple breeding programs (USDA-ARS Aberdeen (AB), Busch Agricultural Resources Inc. (BA), UT, and Washington State University (WA)). LD was found to decay across a range from 4.0 to 19.8 cM. This result indicates that the germplasm genotyped with 3072 SNPs would be robust for mapping and possibly identifying the causal polymorphisms contributing to disease resistance and perhaps other traits.

Key words: Association mapping (AM), Structure, Linkage disequilibrium (LD)
doi:10.1631/jzus.B1200003 **Document code:** A **CLC number:** Q344+.4

1 Introduction

The primary aim of association mapping (AM) is to identify the causal genetic variants that are responsible for phenotypic variation in a collection of individuals. However, since the individuals collected are usually not independent and belong to different subpopulations, spurious associations may be found between phenotypic traits and markers (Ewens and Spielman, 1995). For example, if disease resistance is common in one subpopulation and rare in other subpopulations, all markers that are fixed or in high allele frequency are likely to be associated with that resistance trait (Pritchard and Rosenberg, 1999; Pritchard *et al.*, 2000). Thus, population structure analysis, which can identify and account for the relationships among individuals in a mapping panel, is usually the first and one of the most important steps in AM.

The plant population structure can be complex (Yu *et al.*, 2005). In barley (*Hordeum vulgare* L.), populations can be divided into groups: based on spike morphology (two-rowed vs. six-rowed), intended use (malting vs. feed or food), and growth habit (winter/facultative vs. spring types). In individual breeding programs, lines are usually under independent selection for different target traits such as yield, malting quality, and disease resistance, leading to further population subdivisions. In addition, crosses between elite breeding lines and exotic germplasm (e.g., ranging from breeding lines of other programs to wild barley accessions) are often made to

* Project supported by the Barley Coordinated Agricultural Project (No. USDA-CSREES-2006-55606-16722) of the USDA National Institute of Food and Agriculture and the Lieberman-Okinow Endowment at the University of Minnesota, USA

improve various barley traits and also can contribute to population structure. Recent studies have confirmed the presence of strong genetic structure in cultivated barley (Malysheva-Otto *et al.*, 2006; Rostoks *et al.*, 2006; Saisho and Purugganan, 2007; Comadran *et al.*, 2009; Hamblin *et al.*, 2010).

In addition to population structure, another critical factor in AM is linkage disequilibrium (LD) which is non-random association of alleles between two loci. Due to the large genome size of most plant species, the number of available molecular markers will not sufficiently capture all of the existing genetic variation (Mangin *et al.*, 2012). Even if markers could capture all of this variation, the current prohibitive costs of such genotyping would necessitate that only a subset of markers spanning the genome be used in genome-wide mapping studies. For example, there are about 10 million genetic variants (single nucleotide polymorphisms or SNPs) in the human genome, but the HapMap phase II project only used 3.1 million SNP markers (31%) to identify the common disease-associated loci (Frazer *et al.*, 2007; Bhangale *et al.*, 2008). The rationale for using a portion of available genetic variants to represent all genetic variation in a species is based on LD. If the causal gene is in strong LD with one or more markers, these markers will be identified as being significantly associated with the trait through a whole genome scan. However, in extreme cases where all genetic variants are independent, one would have to sequence the entire genome and examine every variant until the causal one is found. LD is therefore a critical factor in AM because its rate of decay in a species determines the density of molecular markers needed for such analyses (Rafalski, 2002). If LD extends over a long distance, e.g., in many self-pollinated species such as soybean (Hyten *et al.*, 2007), wheat (Chao *et al.*, 2007), barley (Malysheva-Otto *et al.*, 2006; Comadran *et al.*, 2009; Hamblin *et al.*, 2010), and Arabidopsis (Nordborg *et al.*, 2002), fewer markers are needed to cover the entire genome. On the other hand, if LD extends over a very short distance, a much higher marker density is required to cover the genome. Such is the case in humans where LD extends to only one kilobase and researchers have to select millions of SNP markers identified from whole genome sequencing for conducting genome-wide association studies (Frazer *et al.*, 2007; Ku *et al.*, 2010). A similar situation exists in maize, an out-crossing plant species (Remington *et al.*, 2001). If marker density is not sufficiently high, the extent of LD across the entire genome cannot be rigorously assessed and thus portions of the genome will remain poorly described (Rafalski, 2002). Thus, characterizing the extent of LD in a germplasm panel is necessary for the interpretation of AM results.

The developments of numerous molecular markers for various plant species, high throughput sequencing techniques (Metzker, 2009), and modern analytical tools (Yu *et al.*, 2005; Kang *et al.*, 2008; Kang and Sul, 2010; Zhang *et al.*, 2010) have led to new initiatives for utilizing AM to identify causal variants for diverse traits (Zhu *et al.*, 2008) such as flowering time, kernel composition, and kernel color in maize (Thornsberry *et al.*, 2001; Palaisa *et al.*, 2004; Wilson *et al.*, 2004); developmental and flowering-related traits in Arabidopsis (Atwell *et al.*, 2010); multiple agronomic traits in sugar beet (Würschum *et al.*, 2011); quality traits in potato (D′Hoop *et al.*, 2008); disease resistance in sugarcane (Wei *et al.*, 2006); wood property traits in *Pinus taeda* (González-Martínez *et al.*, 2007); disease resistance in barley (Massman *et al.*, 2011); and flowering time in ryegrass (Skøt *et al.*, 2007).

To develop the tools for the genome-wide AM of economically important traits in barley, the Barley Coordinated Agricultural Project (BCAP) (http://www.barleycap.org) was established in the United States. The complete BCAP germplasm for AM consists of 3840 elite breeding lines from ten different improvement programs (384 from each). This panel was genotyped with 3072 SNP markers that were identified from barley expressed sequence tags (ESTs) and sequenced amplicons by Close *et al*. (2009). Elite breeding lines were selected for BCAP in order to fully exploit agronomically pertinent germplasm and utilize beneficial alleles directly in subsequent breeding generations without the negative impact of linkage drag—the association of deleterious alleles with selected alleles. To robustly perform the AM of economically important traits, population structure patterns and LD in BCAP germplasm must be characterized. A study on this topic was recently completed on just a subset of the complete BCAP germplasm (1816 of 3840 total lines) and SNP marker set (1536 of 3072 total markers) by Hamblin *et al*. (2010). However, many of the current and likely future, AM

studies for BCAP will include the complete panel. Moreover, to obtain higher power and better resolution in AM studies, it is essential to use the complete BCAP germplasm as the mapping panel (Risch and Merikangas, 1996; Jannink and Walsh, 2002); thus it is meaningful to have population structure and LD analyses to be conducted for the complete germplasm and SNP marker set. Therefore, the objectives of this study were to: (1) characterize population structure within BCAP germplasm using the Bayesian Markov Chain Monte Carlo (MCMC) and principle component analysis (PCA) approaches, and (2) determine the LD decay within this same germplasm.

2 Materials and methods

2.1 Plant materials

The germplasm panel used in this study was developed by BCAP and consisted of elite barley breeding lines from ten programs in the United States (Table 1). Eight programs utilize spring type barley, and two utilize winter or winter/facultative type barleys. Aside from yield, quality, and agronomic traits, these breeding programs focus on different end uses for barley such as malting, feed, and food. All lines were inbred to at least the F_4 generation and were selected to be representative of each program (Hamblin *et al.*, 2010). Ninety-six lines were submitted from each of the ten breeding programs in each year of the project from 2006 to 2009. Thus, the total number of lines evaluated per year was 960 for a project total of 3840. The complete BCAP germplasm panel was designated as CAP.

2.2 SNP genotyping

Lyophilized leaf tissue from a single plant selection of each barley line was sent to the US Department of Agriculture (USDA)-Agricultural Research Service (ARS) Biosciences Research Laboratory in Fargo, ND for DNA extraction and genotyping. DNA was isolated according to standard procedures (Pallotta *et al.*, 2003). Following the protocols of Illumina's GoldenGate Bead Array Technology (Illumina, San Diego, CA, USA) (Fan *et al.*, 2003), two barley oligonucleotide pool assays (BOPA1 and BOPA2) (Close *et al.*, 2009) containing allele specific oligos for a set of 3072 SNPs were used to genotype the barley lines. All lines were genotyped by Illumina SNP technology (Gunderson *et al.*, 2004) on the Illumina® BeadStation 500G.

2.3 Population structure

The program STRUCTURE (Version 2.1) (Pritchard *et al.*, 2000; Falush *et al.*, 2003), which implements a Bayesian MCMC approach, was used to estimate the membership probability of each barley line to a number of hypothetical subpopulations (*K*).

Table 1 Details on the germplasm contributed by the ten US barley breeding programs comprising the BCAP

Breeding program	Location	Breeder	n_t	n_e	n_{SNP}	Growth habit	Row type	Primary use
University of Minnesota (MN)	St. Paul, MN	Kevin SMITH	384	384	1865	Spring	Six	Malting
North Dakota State University (N6)	Fargo, ND	Richard HORSLEY	384	380	2203	Spring	Six	Malting
USDA-ARS Aberdeen (AB)	Aberdeen, ID	Don OBERT	384	381	2573	Spring	Six/two	Malting/feed
Utah State University (UT)	Logan, UT	David HOLE	384	367	2620	Spring	Six/two	Feed
Busch Agricultural Resources Inc. (BA)	Ft. Collins, CO	Blake COOPER	384	381	2458	Spring	Six/two	Malting
North Dakota State University (N2)	Fargo, ND	Richard HORSLEY	384	383	2527	Spring	Two	Malting
Washington State University (WA)	Pullman, WA	Steve ULLRICH	384	383	2613	Spring	Six/two	Malting/feed/food
Montana State University (MT)	Bozeman, MT	Tom BLAKE	384	384	2315	Spring	Two	Malting/feed/food
Oregon State University (OR)	Corvallis, OR	Patrick HAYES	384	334	2713	Winter/ facultative	Six/two	Malting/feed/food
Virginia Polytechnic Inst. & State Univ. (VT)	Blacksburg, VA	Carl GRIFFEY	384	356	2440	Winter	Six	Feed

n_t: total number of lines; n_e: number of lines examined; n_{SNP}: number of polymorphic SNP markers

To avoid overestimation of subpopulation divergence caused by tightly linked SNP markers (Falush *et al.*, 2003), a subset of markers (designated as snp1) with approximately 10 cM spacing was selected for the analysis. Additionally, a second independently selected subset (snp2) of markers with similar spacing also was analyzed to provide some validation. These two subsets were then combined to form a new larger subset (snp1&2) of 205 SNP markers. The snp1&2 subset was used to confirm that a sufficient number of markers were present in the two smaller subsets for inferring subpopulations and also to calculate the final subpopulation membership matrix (***Q***), which is the fractional subpopulation membership for each barley line. For each marker subset, subpopulation numbers from 2–15 were modeled with a burn-in of 10000 cycles, followed by 50000 iterations with ten independent runs using an admixture model. Posterior probability Pr($X|K$) of each K was generated by the program and lnPr($X|K$) was plotted against each value of K. The optimal subpopulation number (K) was then estimated based on the value and variation of Pr($X|K$) as well as the rate of lnPr($X|K$) change from $K-1$ to K. The determined optimal K was then used in an additional run of STRUCTURE with the same settings, except a burn-in of 50000 cycles and run of 100000 iterations to calculate the final matrix ***Q*** and relative distance between each pair of subpopulations. A Kullback-Leibler distance (Dragalin *et al.*, 2003) (a measure of dissimilarity between two subpopulations) table also was generated.

PCA, a classical nonparametric linear dimensionality reduction technique (Jolliffe, 2002), also was conducted as an alternative method for determining population structure. This analysis was performed using the software program TASSEL (Version 3.0) (Bradbury *et al.*, 2007). SNP markers with minor allele frequency (MAF)>0.05 were used in the PCA. A matrix ***P***, which consists of principle component vectors accounting for population structure, was generated.

Structure in CAP was visualized by creating a scatter plot of the first three principle component vectors. To compare the results of population structure division, each barley line was assigned to a subpopulation based on the membership proportion from matrix ***Q*** (from the output of STRUCTURE software) and then color-coded in the scatter plot.

2.4 Linkage disequilibrium

Pair-wise measures of LD (r^2) were calculated with Haploview (Version 4.2) (Barrett *et al.*, 2005) using SNP markers with MAF>0.05 in CAP and in each subpopulation defined by STRUCTURE. Significance of each pair-wise LD comparison was determined (P-value≤0.05) using the formula $\chi^2=2nr^2$ where n is the number of individuals in the panel and $df=1$. Due to the intensive computational requirements, only intra-chromosomal LD was calculated. For each chromosome, pair-wise LDs (r^2) with P-values<0.05 were plotted against genetic distance (cM), and a second-degree smoothed loess curve (Cleveland, 1979) was fit using the program R (www.R-project.org) to summarize the relationship of LD extent and genetic distance. Background LD, which is the random association between two loci, was determined by the 95th percentile of unlinked r^2 values (i.e., LD between any pair of markers greater than 50 cM apart (Haldane, 1919) in program R). The intersection of the loess curve and background LD was considered as an estimate of LD decay (Breseghello and Sorrells, 2006).

3 Results

3.1 Marker statistics

Based on the consensus map developed by Close *et al.* (2009), 2943 out of 3072 total SNP markers (from BOPA1 and BOPA2) were mapped to the seven barley chromosomes and spanned a distance of 1099 cM. The number of polymorphic markers (MAF>0.0) was calculated for each breeding program and ranged from 1865 for MN to 2713 for OR (Table 1). SNP markers with >20% missing data or <5% MAF were removed, resulting in 2099 mapped markers. Breeding lines in the panel with >10% missing data also were removed, resulting in 3733 total lines. Thus, the final data matrix of 3733 lines×2099 marker loci was used to investigate population structure and LD in CAP.

SNP marker coverage of the genome was good for the CAP and averaged 1.9/cM. Gaps between SNP markers ranged from 0 to 10 cM with a mean of 0.5 cM. Only ten gaps larger than 5 cM were identified: one each on chromosomes 1H, 3H, and 6H, four on chromosome 5H, and three on chromosome 7H (Table 2).

Table 2 A summary of SNP marker gaps across the seven barley chromosomes for the complete panel (CAP) of US barley breeding germplasm

Gap size (cM)	Number of SNP marker gaps of the specified interval						
	Chromosome 1H	Chromosome 2H	Chromosome 3H	Chromosome 4H	Chromosome 5H	Chromosome 6H	Chromosome 7H
0–1	191	275	274	189	321	232	235
1–2	29	32	37	32	31	23	37
2–3	11	13	12	12	13	13	12
3–4	5	6	8	3	8	3	7
4–5	2	5	2	1	3	4	1
5–10	1	0	1	0	4	1	3

MAF for the complete set of SNP markers (3072) in each breeding program was calculated and plotted (Fig. 1). The median MAF ranged from 0.039 for N6 to 0.276 for BA. The MN and N6 programs had by far the lowest median MAF as more than half of the SNP markers had values less than 0.05. In contrast, the AB and BA programs had the highest median MAF. The MN program also had two special characteristics that set it apart from the other programs: (1) it had the smallest inter-quartile range (IQR: distance between the upper quartile and lower quartile—a measure of variability in the data) and (2) it had 340 outliers (MAF distributed out of the range of 1.5×IQR from the median), while the other programs had none. The heterozygosity rate was also calculated for each breeding program (Fig. 1) and ranged from 0.07% for BA to 1.22% for OR. The distribution of rare SNP markers (MAF<0.05) among the ten breeding programs is shown in Table 3 and ranged from 386 for AB to 1160 for N6. The number of shared rare SNP markers ranged from 89 (between OR and MN) to 506 (between MN and N6).

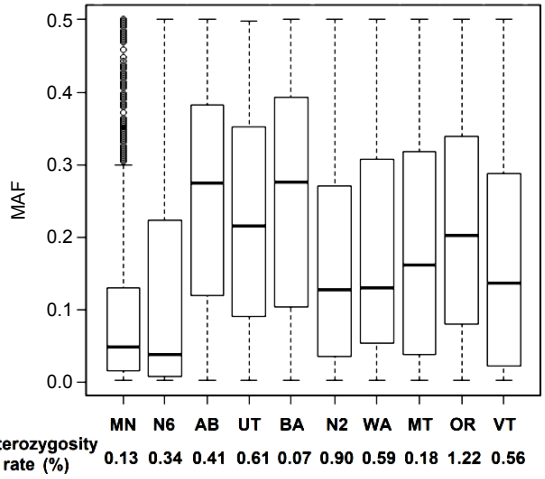


Fig. 1 Boxplots of minor allele frequency (MAF) in US barley breeding germplasm

Five statistics are represented as bars in each boxplot from bottom to top: the smallest observation, lower quartile, median, upper quartile, and largest observation, respectively. Data points which lie outside this range are extreme values. Heterozygosity rate is given at the bottom under each breeding program. Breeding programs: University of Minnesota (MN), North Dakota State University (N6), USDA-ARS Aberdeen (AB), Utah State University (UT), Busch Agricultural Resources Inc. (BA), North Dakota State University (N2), Washington State University (WA), Montana State University (MT), Oregon State University (OR), Virginia Polytechnic Institute and State University (VT)

3.2 Population structure

The STRUCTURE program was used to analyze population structure in CAP. Three subsets of SNP markers, snp1, snp2, and snp1&2, were analyzed independently to determine the optimal subpopulation number (K) within the final panel number of 3733 lines. The snp1 and snp2 subsets included 103 and 102 markers, respectively, spanned all seven chromosomes, and had gap sizes ranging from 7–12 cM. The snp1&2 subset included 205 markers across the genome with gap sizes ranging from 1.8–9.0 cM. Posterior probability increased gradually from K=3 and reached a near-stationary stage at K=9. Additionally, the first decrease in posterior probability was observed from K=9 to K=10 in all three SNP marker subsets (Table 4). Large variations within the ten independent runs were observed at K=10, 12, and 15 using the snp1&2 subset. STRUCTURE results from all three marker subsets indicated that 9 is the optimal number of subpopulations (K).

Table 3 Distribution of rare SNP markers (0<MAF<0.05) among ten US breeding programs

Program	Number of rare SNP markers									
	MN[a]	N6	AB	UT	BA	N2	WA	MT	OR	VT
MN	948									
N6	506	1160								
AB	95	195	386							
UT	188	188	149	449						
BA	142	176	186	133	404					
N2	197	278	196	192	180	752				
WA	163	257	241	183	235	253	602			
MT	220	231	134	122	155	253	242	654		
OR	89	194	105	117	94	130	177	90	408	
VT	218	337	117	146	105	230	211	213	192	792

Values on the diagonal are the number of rare SNP markers within each program. The other values are the number of rare SNP markers shared between pairs of breeding programs. [a] Breeding programs: University of Minnesota (MN), North Dakota State University (N6), USDA-ARS Aberdeen (AB), Utah State University (UT), Busch Agricultural Resources Inc. (BA), North Dakota State University (N2), Washington State University (WA), Montana State University (MT), Oregon State University (OR), Virginia Polytechnic Institute and State University (VT)

Table 4 Natural logarithm probability of data as a function of *K* and the natural logarithm probability difference between *K* and *K*−1 for three subsets of SNP markers based on ten independent runs in the STRUCTURE program

K	snp1		snp2		snp1&2	
	Mean lnPr(*X*|*K*)	Δ[a]	Mean lnPr(*X*|*K*)	Δ	Mean lnPr(*X*|*K*)	Δ
2	−201035		−191442		−389012	
3	−174864	26171	−166954	24488	−339195	49817
4	−169596	5268	−162184	4770	−328650	10545
5	−165281	4316	−158594	3590	−320312	8338
6	−161033	4248	−154671	3923	−311827	8485
7	−157518	3514	−151416	3254	−304936	6891
8	−155964	1555	−147800	3616	−302295	2641
9	−150342	5621	−146452	1348	−293325	8970
10	−151161	**−819**	−149815	**−3363**	−300727	**−7402**
11	−148509	2652	−143706	6109	−287285	13442
12	−146737	1772	−144412	−706	−316283	−28998
13	−146348	389	−143094	1318	−282061	34222
14	−145326	1022	−141075	2019	−280425	1636
15	−146296	−969	−140642	433	−291228	−10804

The first decreases in the mean lnPr(*X*|*K*) were found for *K*=9 to *K*=10 and are given in bold. [a] The value of lnPr(*X*|*K*)−lnPr(*X*|*K*−1) for *K* from 2 to 15

Most subpopulations were strongly represented by lines from specific breeding programs (Table 5). For example, 84% of lines (366) of subpopulation 2 (sp2) were from MN, 85% of lines (355) of sp3 were from VT, 76% of lines (380) of sp4 were from N2, 99% of lines (247) of sp5 were from UT, 88% of lines (369) of sp6 were from N6, and 100% of lines (207) of sp7 were from OR. In contrast, sp1, sp8, and sp9 were not as strongly dominated by lines from a single breeding program and included a wider mixture of lines.

Assignments of lines to sp1, sp2, sp3, sp4, sp6, sp7 and sp8 were, in general, very similar to the respective subpopulations of 2sp(BA) (90.4% identical assignment), 6sp(MN) (68.5%), 6wi(VT) (91.7%), 2sp(N2) (97.7%), 6sp(N6) (83.6%), 6wi(OR) (97.6%), and 2sp(WA) (91.7%) identified by Hamblin *et al*. (2010) which was based only on CAPI and CAPII and half (1536) of the total SNP markers. The lower percentage of identical line assignments for subpopulations 6sp(MN) and 6sp(N6) was due to 91 and 34 lines, respectively, being assigned to the new subpopulation sp9 in this study.

In addition to subpopulation membership, STRUCTURE also generated a Kullback-Leibler divergence table, which shows the relative distance between each pair of subpopulations (Table 6). The smallest distance found between pairs of subpopulations was 0.14 for sp6 (dominated by N6) and sp9 (dominated by AB and BA), followed by 0.22 for sp2 (dominated by MN) and sp9, and 0.33 for sp2 and sp6. The largest distance (2.23) found between subpopulations was for sp2 and sp3 (dominated by VT).

Table 5 Subpopulations defined by the STRUCTURE program in the complete panel (CAP) of US barley breeding germplasm

sp[a]	Number (proportion) of lines[c]											Assignment based on CAPI and CAPII[d]
	MN[b]	N6	AB	UT	BA	N2	WA	MT	OR	VT	Total	
sp1	0 (0.00)	0 (0.00)	171 (0.23)	21 (0.03)	232 (0.31)	1 (0.00)	86 (0.12)	170 (0.23)	62 (0.08)	0 (0.00)	743	2sp(BA)
sp2	366 (0.84)	6 (0.01)	21 (0.05)	9 (0.02)	32 (0.07)	0 (0.00)	0 (0.00)	1 (0.00)	2 (0.00)	0 (0.00)	437	6sp(MN)
sp3	0 (0.00)	0 (0.00)	0 (0.00)	2 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	59 (0.14)	355 (0.85)	416	6wi(VT)
sp4	0 (0.00)	1 (0.00)	21 (0.04)	3 (0.01)	0 (0.00)	380 (0.76)	2 (0.00)	89 (0.18)	1 (0.00)	0 (0.00)	497	2sp(N2)
sp5	0 (0.00)	0 (0.00)	2 (0.01)	247 (0.99)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	249	
sp6	9 (0.02)	369 (0.88)	13 (0.03)	0 (0.00)	20 (0.05)	0 (0.00)	6 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)	417	6sp(N6)
sp7	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	207 (1.00)	0 (0.00)	207	6wi(OR)
sp8	0 (0.00)	0 (0.00)	47 (0.10)	18 (0.04)	1 (0.00)	2 (0.00)	264 (0.58)	124 (0.27)	0 (0.00)	0 (0.00)	456	2sp(WA)
sp9	9 (0.03)	4 (0.01)	106 (0.34)	67 (0.22)	96 (0.31)	0 (0.00)	25 (0.08)	0 (0.00)	3 (0.01)	1 (0.00)	311	

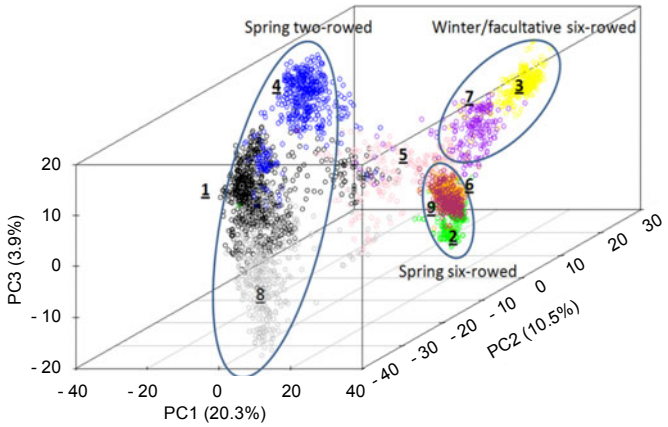
[a] sp: subpopulations defined in this study. [b] Breeding programs: University of Minnesota (MN), North Dakota State University (N6), USDA-ARS Aberdeen (AB), Utah State University (UT), Busch Agricultural Resources Inc. (BA), North Dakota State University (N2), Washington State University (WA), Montana State University (MT), Oregon State University (OR), Virginia Polytechnic Institute and State University (VT). [c] Number (proportion) of lines from individual breeding programs belonging to a subpopulation defined by the STRUCTURE program. [d] Subpopulations defined by Hamblin *et al*. (2010) based on CAPI and CAPII and half (1 536) of the total SNP markers

Table 6 Kullback-Leibler distance between pairs of subpopulations (sp1–sp9) in the complete panel (CAP) of US barley breeding germplasm

sp	Kullback-Leibler distance								
	sp1	sp2	sp3	sp4	sp5	sp6	sp7	sp8	sp9
sp1									
sp2	1.29								
sp3	1.39	*2.23*							
sp4	0.41	1.48	1.28						
sp5	0.86	1.72	0.78	0.66					
sp6	1.28	**0.33**	1.17	0.65	0.69				
sp7	1.30	1.94	0.63	0.88	0.84	1.44			
sp8	0.40	1.95	1.27	0.53	0.63	1.62	1.31		
sp9	1.26	**0.22**	1.24	0.67	0.68	**0.14**	0.89	1.25	

The three smallest distances between subpopulations are given in bold and the largest distance is in italics

PCA also was used to analyze population structure in CAP (Fig. 2). Three major clusters consisting of spring two-rowed, winter/facultative six-rowed, and spring six-rowed lines were identified. The first PC was due predominantly (20.3%) to different row types (i.e., six-rowed vs. two-rowed), and the second PC (10.5%) was due to winter/facultative vs. spring types. PC3 explained only 3.9% of the total variation, but it separated the two sub-groups within the two-rowed spring lines. Additionally, when the lines were color-coded with their subpopulation membership (Fig. 2) as defined by STRUCTURE, PC3 also separated some of these subpopulations. For example, sp1 (in black, dominated by two-rowed lines from AB, BA, MT, and WA) and sp8 (in grey, dominated by MT and two-rowed lines from WA) were clearly separated by PC3. Most of the CAP lines were captured by the three major clusters, the exception being lines from sp5 (in pink, dominated by six-rowed lines from UT), which lie in the center of three clusters.



sp1: in black, dominated by BA (31%), AB (23%), MT (23%), and WA (12%)
sp2: in green, dominated by MN (84%)
sp3: in yellow, dominated by VT (85%)
sp4: in blue, dominated by N2 (76%) and MT (18%)
sp5: in pink, dominated by UT (99%)
sp6: in orange, dominated by N6 (88%)
sp7: in purple, dominated by OR (100%)
sp8: in grey, dominated by WA (59%) and MT (27%)
sp9: in red, dominated by AB (34%) and BA (31%)

Fig. 2 Three-dimensional plot of the complete panel (CAP) of US barley breeding germplasm based on PCA
The first PC was due predominantly (20.3%) to different row types (i.e., six-rowed vs. two-rowed), and the second PC (10.5%) was due to winter/facultative vs. spring types. PC3 only explained 3.9% of the total variation, but it separated the two subgroups within the two-rowed spring lines. The nine subpopulations (sp1–sp9) defined by STRUCTURE were color-coded and numbered. The three major clusters defined by PCA also were labeled. Breeding programs: University of Minnesota (MN), North Dakota State University (N6), USDA-ARS Aberdeen (AB), Utah State University (UT), Busch Agricultural Resources Inc. (BA), North Dakota State University (N2), Washington State University (WA), Montana State University (MT), Oregon State University (OR), Virginia Polytechnic Institute and State University (VT)

STRUCTURE and PCA were both used in this study to analyze the population structure within CAP. The results of these analyses are compared in Fig. 2. In this figure, sp1 (in black), sp3 (in yellow), sp4 (in blue), sp5 (in pink), sp7 (in purple), and sp8 (grey) were, in general, separated into definite clusters by the three PCs, although some lines from sp4 (in blue, dominated by N2) were mixed into sp1 (in black, dominated by two-rowed lines from AB, BA, MT, and WA) and a few lines from sp1 were mixed into sp5 (in pink, dominated by six-rowed lines from UT). Lines in sp2 (in green, dominated by MN), sp6 (in orange, dominated by N6), and sp9 (in red, dominated by AB and BA) exhibited a greater degree of overlap with each other. However, in general, all nine subpopulations could be distinguished from the plot. This indicates that PCA and STRUCTURE subdivided the population in a very similar way and that the PCA results confirmed that K=9 was optimal. In addition, the pair-wise distances between subpopulations (defined by STRUCTURE) on the plot also generally matched the Kullback-Leibler distances between each pair of subpopulations (Table 6). For example, sp2 (in green), sp6 (in orange), and sp9 (in red) exhibited some degree of overlap in the plot (Fig. 2) and also were the ones showing the closest pair-wise distances (Table 6). In contrast, the largest distance found from the Kullback-Leibler table was between sp2 (in green, dominated by MN) and sp3 (in yellow, dominated by VT); these two subpopulations also showed a large separation in the PCA plot from the lower left corner for sp2 to the upper right corner for sp3 (Fig. 2).

3.3 Linkage disequilibrium

Classical pair-wise intra-chromosomal LD (r^2) was first calculated for the CAP and then for the nine subpopulations (sp1–sp9) defined by STRUCTURE. In CAP, 94% of the total pair-wise LD was significant, whereas in the nine subpopulations the values ranged from 43% in sp2 to 66% in both sp1 and sp5 (Table 7). In all cases, LD was extensive. Long range LD (>50 cM) was observed in CAP and all of its subpopulations. The average significant LD extent was 46.6 cM in CAP and ranged from 35.5 cM in sp2 to 44.0 cM in sp3 for the subpopulations (Table 7).

To determine LD decay, background LD was first estimated in order to establish the LD threshold in CAP and each subpopulation (Table 7). The highest extent of background LD was 0.3 in CAP and ranged from 0.04 (sp2) to 0.12 (sp5) in the subpopulations. With respect to CAP, the background LD and second degree loess curve did not intersect; thus, it was not possible to estimate LD decay (Table 7). Estimates of LD decay in the nine subpopulations ranged from 4.0 cM in both sp3 and sp5 to 19.8 cM in sp2 (Table 7).

Table 7 Intra-chromosomal linkage disequilibrium (LD) in the complete panel (CAP) of US barley breeding germplasm and each subpopulation (sp1–sp9) as defined by the STRUCTURE program

Panel	Number of marker pairs in significant LD	Average LD extent (cM)	Background LD[b]	Estimated LD decay[c] (cM)
CAP	301 329 (94%[a])	46.6	0.30	
sp1	207 711 (66%)	43.5	0.05	11.9
sp2	108 736 (43%)	35.5	0.04	19.8
sp3	165 160 (59%)	44.0	0.09	4.0
sp4	187 423 (60%)	43.2	0.06	12.0
sp5	192 339 (66%)	42.4	0.12	4.0
sp6	115 190 (45%)	37.7	0.05	15.9
sp7	164 372 (60%)	41.7	0.10	7.9
sp8	171 433 (56%)	41.4	0.05	15.9
sp9	139 095 (52%)	40.6	0.07	12.0

[a] Proportion of significant pair-wise intra-chromosomal LD among all pair-wise intra-chromosomal LD ($P<0.05$). [b] The 95th percentile of unlinked r^2 values (LD between any pair of SNP markers >50 cM apart). [c] Significant pair-wise LD (r^2) was plotted against cM and a second-degree smoothed loess curve was fitted; the intersection of the loess curve and background LD was considered as an estimate of LD decay

4 Discussion

One of the primary goals of the BCAP is to identify markers linked with beneficial traits using an AM approach. Achievement of this objective will make the breeding process more efficient and ultimately lead to the more rapid development of barley cultivars with superior yield, quality, and agronomic traits. However, it does have some disadvantages, notably the requirement for controlling population structure within the mapping panel (Yu *et al.*, 2005) and the need for high marker density. A study based on just a subset of the complete BCAP germplasm (1 816 of 3 840 total lines) and SNP marker set (1 536 of 3 072 total markers) (Hamblin *et al.*, 2010) was recently completed. However, most traits, such as yield, heading date, malting quality, disease resistance, and morphological traits, were collected from the entire CAP germplasm. Moreover, to obtain higher power and better resolution in AM studies as well as a more accurate analysis of population structure and LD extent, it is essential that all available lines and SNP markers be included in the mapping panel (Risch and Merikangas, 1996; Jannink and Walsh, 2002).

4.1 Population structure

Different end uses, growth habits, and row types of breeding lines can all affect the population structure in barley panels (Malysheva-Otto *et al.*, 2006; Rostoks *et al.*, 2006). Based on the SNP markers used in this study, clear evidence was found for population structure, i.e., large variation in polymorphic marker numbers, MAF distribution, heterozygous rate, and rare SNP numbers (Tables 1 and 3; Fig. 1). A low frequency of polymorphic markers indicates less diversity within a specific breeding program due likely to the fixation of loci within the program or possibly to an unknown contribution of ascertainment bias (Hamblin *et al.*, 2010). The Upper Midwest six-rowed malting barley programs of MN and N6 had the lowest number of polymorphic markers, indicating that they are the least diverse among all BCAP programs, although the existence of structure within each individual program could be a confounding factor. For example, the AB, WA, UT, BA and OR programs are comprised of both two-rowed and six-rowed lines (Table 1). Thus, it is possible that subgroups within a breeding program may have an even lower number of polymorphic markers. However, under the broader scope of individual breeding programs, the MN and N6 programs were genetically the least diverse. This may largely be due to the strict requirements for malting quality mandated by industry (Wych and Rasmusson, 1983). Other evidences for this low diversity include the lowest median MAF and the second and third lowest heterozygous rates (Fig. 1), both of which are important genetic diversity

indicators in population genetics (Cornuet and Luikart, 1996). The narrow gene pool of Midwest six-rowed spring barley (MN and N6) also was reported by other researchers (Wych and Rasmusson, 1983; Horsley *et al.*, 1995; Condón *et al.*, 2008; Mikel and Kolb, 2008; Hamblin *et al.*, 2010). Another important statistic in population genetics is the number of rare polymorphisms (variants). Rare variants are more likely to be recently derived than other common variants and are, therefore, more likely to be population specific (Watterson and Guess, 1977). Hence, rare variants are very sensitive indicators of the relationships among different populations. The MN and N6 programs share the highest number of rare SNP markers, suggesting a close relationship between them (Table 3). This pattern of sharing is consistent with an ancestry study showing that MN and N6 share four common ancestors out of only six and seven total ancestors for the two programs, respectively (Martin, 1991). Low values of the Kullback-Leibler distance statistics among Midwest six-rowed spring subpopulations (sp2, sp6, and sp9) from STRUCTURE analysis (Table 5) and overlapping scatter plots of Midwest six-rowed spring lines (Fig. 2) from PCA analysis are further evidences for the close relationship between these two programs.

Two widely used methods that account for population structure in AM studies are the Bayesian MCMC method implemented in the STRUCTURE program (Pritchard *et al.*, 2000) and PCA (Patterson *et al.*, 2006). These two methods were implemented for structure analysis in this study and their results compared. The results derived from PCA were very similar to those obtained from STRUCTURE for CAP (Fig. 2). Similar conclusions were also reported by others (Patterson *et al.*, 2006; Song *et al.*, 2009; Mezmouk *et al.*, 2011). Thus, it appears that there are no serious shortcomings of using PCA for population structure analysis. Moreover, PCA has the added advantages of being much faster in generating the complete analysis and also more parsimonious in its use of computational resources.

STRUCTURE is a program that classifies individuals into discrete populations (Pritchard *et al.*, 2000). In this study, the optimal subpopulation number was determined by running STRUCTURE with three sets of SNP markers. Selection of nine as the optimal subpopulation number was based on a rigorous analysis of two independent marker subsets (snp1 and snp2) as well as the combined subset (snp1&2). Moreover, it agrees with the analysis of Hamblin *et al*. (2010) on the CAPI and CAPII panels (Table 5). In the analysis of Hamblin *et al*. (2010), the germplasm was divided into seven subpopulations, 2sp(BA), 6sp(MN), 6wi(VT), 2sp(N2), 6sp(N6), 6wi(OR), and 2sp(WA), which correspond to sp1, sp2, sp3, sp4, sp6, sp7, and sp8, respectively, in our analysis. Compared to their study, the subpopulations of sp5 and sp9 defined in this study were "new". This is likely due to the fact that in their study, only 96 UT lines (from CAPII) were included, while in this study all UT lines (384 from CAPI to CAPIV) were considered. The population size of 96 for UT in Hamblin *et al*. (2010) was too small to assign a new subpopulation; thus, these 96 lines were assigned to 2sp(BA), 6sp(N6), 6wi(VT), 2sp(WA), and an unassigned admixture group. In this study, the larger set of UT lines was placed in an independent subpopulation (sp5), which consisted of 99% UT (247 lines) and 1% AB (2 lines) (Table 5). The other new subpopulation (sp9) was mainly composed of six-rowed spring lines from AB, BA, UT, and WA. Again, because of the small population size, these lines were assigned into 6sp(MN) and 6sp(N6) in the study of Hamblin *et al*. (2010).

Unlike the program STRUCTURE, which assigns individuals into discrete populations, PCA plots the coordinates of each individual along axes of variation without resorting to a model (Patterson *et al.*, 2006; McVean, 2009). Therefore, subpopulations can be visualized by plotting the first few eigenvectors (see CAP example from this study in Fig. 2). However, one disadvantage of PCA is that it does not perform well in an admixed population (Patterson *et al.*, 2006; Boyko *et al.*, 2009). In such populations, the expected allele frequency of an individual is a linear combination of the frequencies in its ancestors. PCA will fail in this case because all individuals in the admixed population will have the same ancestry proportion (Patterson *et al.*, 2006). For example, in this study, the Midwest six-rowed spring barley programs, i.e., sp2 (in green, dominated by MN), sp6 (in orange, dominated by N6), and sp9 (in red, dominated by AB and BA), may represent such an admixed population (Fig. 2). These programs have a very narrow gene pool and also share a large proportion of

founders (Horsley *et al.*, 1995; Condón *et al.*, 2008; Hamblin *et al.*, 2010). Because of gene recombination, each Midwest line has a different size and location of a founder's chromosomal segments, but the proportion of one founder's genetic material is likely to be constant. As mentioned above, if all lines have the same amount of genetic proportion contributed by its ancestors, PCA is less likely to work well for describing structure. The program STRUCTURE, with its model-based clustering method, does not depend on the ancestry proportion. Thus, with this program, Midwest lines were clearly assigned into three groups, sp2, sp6, and sp9 (Table 5). Because of the limitations of PCA in addressing structure in populations with heavy admixture history, one must be very cautious in accounting for structure in AM.

4.2 Linkage disequilibrium

LD is the nonrandom association of alleles at two or more loci in a population. However, since AM panels now tend to be quite large, the structure within them becomes a critical factor in considering the extent of LD. In this study, LD was first calculated without considering structure in CAP. Then, LD was estimated in each subpopulation as defined by STRUCTURE, which removed some degree of structure effect. Without considering the effect of structure, 94% of pair-wise intra-chromosomal LD was significant and the background LD was as high as r^2=0.3 (Table 7). The background LD was determined as the LD between two loci that are more than 50 cM apart because, in such cases, the two loci should be unlinked (Haldane, 1919). Background LD is mostly due to other factors, such as selection, genetic drift, and structure (Flint-Garcia and Thornsberry, 2003). However, because CAP is a highly structured population, the main factor contributing to high background LD may well be structure. To prove this hypothesis, LD was estimated in subpopulations sp1–sp9, which should have limited structure. The results revealed a sharp decrease in background LD (Table 7). Additionally, subpopulations with more diversity (more unexplained structure within each subpopulation) tended to have higher background LD. For example, sp5, which was composed almost entirely (99%) of six-rowed spring lines from UT, has a very diverse genetic background (Prof. David HOLE, Utah State University, May 2012, personal communication), which can be inferred from the dispersed plotting of sp5 in Fig. 2 and also the highest background LD (r^2=0.12) among all subpopulations. In contrast, Midwest six-rowed spring subpopulations (sp2, sp6, and sp9) with a very narrow genetic pool (Horsley *et al.*, 1995; Condón *et al.*, 2008; Hamblin *et al.*, 2010) had a much lower background LD from 0.04 to 0.07. These results indicate that structure is an important confounding factor for LD estimation in CAP.

LD decay is the key factor for determining the minimum marker density needed for AM. In this study, the intersection of the loess curve and background LD was used as the estimated value for LD decay. The Midwest six-rowed spring subpopulations of sp2, sp6, and sp9 have a long LD decay (Table 7) of 19.8, 15.9, and 12.0 cM, respectively. This long LD decay again confirms the narrow gene pool of these breeding programs. In contrast, six-rowed spring lines from sp5 (99% UT) and six-rowed winter lines from sp3 (dominated by VT) have the shortest LD decay, which again suggests the diverse genetic background of these two subpopulations. Due to the confounding effect of population structure, LD decay could not be estimated in the CAP using this method because there was no intersection between the loess curve and background LD (Table 7). However, overall LD decay in CAP should be between the LD in the most genetically diverse subpopulation (sp3) and the LD in the most genetically narrow subpopulation (sp2), which ranged from 4.0 to 19.8 cM (Table 7). This extensive overall LD decay indicates that genetic variations within BCAP germplasm may be well represented by BOPA1 and BOPA2 SNP markers and thus this data matrix of 3 733 lines×2 099 marker loci may be very robust for mapping causal polymorphisms contributing to disease resistance and perhaps other traits (Zhou, 2011).

References

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**(7298):627-631. [doi:10.1038/nature08800]

Barrett, J.C., Fry, B., Maller, J., Daly, M.J., 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**(2):263-265. [doi:10.1093/bioinformatics/bth457]

Bhangale, T.R., Rieder, M.J., Nickerson, D.A., 2008.

Estimating coverage and power for genetic association studies using near-complete variation data. *Nat. Genet.*, **40**(7):841-843. [doi:10.1038/ng.180]

Boyko, A.R., Boyko, R.H., Boyko, C.M., Parker, H.G., Castelhano, M., Corey, L., Degenhardt, J.D., Auton, A., Hedimbi, M., Kityo, R., 2009. Complex population structure in African village dogs and its implications for inferring dog domestication history. *PNAS*, **106**(33):13903-13908. [doi:10.1073/pnas.0902129106]

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S., 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**(19):2633-2635. [doi:10.1093/bioinformatics/btm308]

Breseghello, F., Sorrells, M.E., 2006. Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics*, **172**(2):1165-1177. [doi:10.1534/genetics.105.044586]

Chao, S., Zhang, W., Dubcovsky, J., Sorrells, M., 2007. Evaluation of genetic diversity and genome-wide linkage disequilibrium among US wheat (*Triticum aestivum* L.) germplasm representing different market classes. *Crop Sci.*, **47**(3):1018-1030. [doi:10.2135/cropsci2006.06.0434]

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**(368):829-836. [doi:10.1080/01621459.1979.10481038]

Close, T.J., Bhat, P.R., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., Druka, A., Stein, N., Svensson, J.T., Wanamaker, S., *et al*., 2009. Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*, **10**(1):582. [doi:10.1186/1471-2164-10-582]

Comadran, J., Thomas, W.T.B., van Eeuwijk, F.A., Ceccarelli, S., Grando, S., Stanca, A.M., Pecchioni, N., Akar, T., Al-Yassin, A., Benbelkacem, A., 2009. Patterns of genetic diversity and linkage disequilibrium in a highly structured *Hordeum vulgare* association-mapping population for the Mediterranean basin. *Theor. Appl. Genet.*, **119**(1):175-187. [doi:10.1007/s00122-009-1027-0]

Condón, F., Gustus, C., Rasmusson, D.C., Smith, K.P., 2008. Effect of advanced cycle breeding on genetic diversity in barley breeding germplasm. *Crop Sci.*, **48**(3):1027-1036. [doi:10.2135/cropsci2007.07.0415]

Cornuet, J.M., Luikart, G., 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**(4):2001-2014.

D'Hoop, B.B., Paulo, M.J., Mank, R.A., van Eck, H.J., van Eeuwijk, F.A., 2008. Association mapping of quality traits in potato (*Solanum tuberosum* L.). *Euphytica*, **161**(1-2):47-60. [doi:10.1007/s10681-007-9565-5]

Dragalin, V., Fedorov, V., Patterson, S., Jones, B., 2003. Kullback-Leibler divergence for evaluating bioequivalence. *Stat. Med.*, **22**(6):913-930. [doi:10.1002/sim.1451]

Ewens, W.J., Spielman, R.S., 1995. The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.*, **57**(2):455-464.

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.

Fan, J.B., Oliphant, A., Shen, R., Kermani, B.G., Garcia, F., Gunderson, K.L., Hansen, M., Steemers, F., Butler, S.L., Deloukas, P., *et al*., 2003. Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol.*, **68**:69-78. [doi:10.1101/sqb.2003.68.69]

Flint-Garcia, S.A., Thornsberry, J.M., 2003. Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.*, **54**(1):357-374. [doi:10.1146/annurev.arplant.54.031902.134907]

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164):851-861. [doi:10.1038/nature06258]

González-Martínez, S.C., Wheeler, N.C., Ersoz, E., Nelson, C.D., Neale, D.B., 2007. Association genetics in *Pinus taeda* L. I. Wood property traits. *Genetics*, **175**(1):399-409. [doi:10.1534/genetics.106.061127]

Gunderson, K.L., Kruglyak, S., Graige, M.S., Garcia, F., Kermani, B.G., Zhao, C., Che, D., Dickinson, T., Wickham, E., Bierle, J., 2004. Decoding randomly ordered DNA arrays. *Genome Res.*, **14**(5):870-877. [doi:10.1101/gr.2255804]

Haldane, J.B.S., 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.*, **8**(29):309-320.

Hamblin, M.T., Close, T.J., Bhat, P.R., Chao, S., Kling, J.G., Abraham, K.J., Blake, T., Brooks, W.S., Cooper, B., Griffey, C.A., *et al*., 2010. Population structure and linkage disequilibrium in US barley germplasm: implications for association mapping. *Crop Sci.*, **50**(2):556-566. [doi:10.2135/cropsci2009.04.0198]

Horsley, R.D., Schwarz, P.B., Hammond, J.J., 1995. Genetic diversity in malt quality of North American six-rowed spring barley. *Crop Sci.*, **35**(1):113-118. [doi:10.2135/cropsci1995.0011183X003500010021x]

Hyten, D.L., Choi, I.Y., Song, Q., Shoemaker, R.C., Nelson, R.L., Costa, J.M., Specht, J.E., Cregan, P.B., 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics*, **175**(4):1937-1944. [doi:10.1534/genetics.106.069740]

Jannink, J.L., Walsh, B., 2002. Association Mapping in Plant Populations. *In*: Kang, M.S. (Ed.), Quantitative Genetics Genomics and Plant Breeding. CAB International, Wallingford, UK, p.59-68.

Jolliffe, I., 2002. Principal Component Analysis. Springer-Verlag, New York, NY.

Kang, H.M., Sul, J.H., 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**(4):348-354. [doi:10.1038/ng.548]

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E., 2008. Efficient control of population structure in model organism association mapping. *Genetics*, **178**(3):1709-1723. [doi:10.1534/genetics.107.080101]

Ku, C.S., Loy, E.Y., Pawitan, Y., Chia, K.S., 2010. The pursuit of genome-wide association studies: where are we now? *J. Hum. Genet.*, **55**(4):195-206. [doi:10.1038/jhg.2010.19]

Malysheva-Otto, L.V., Ganal, M.W., Roder, M.S., 2006. Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). *BMC Genet.*, **7**(1):6-18. [doi:10.1186/1471-2156-7-6]

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., Cierco-Ayrolles, C., 2012. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, **108**(3):285-291. [doi:10.1038/hdy.2011.73]

Martin, J.M., 1991. Diversity among North American spring barley cultivars based on coefficients of parentage. *Crop Sci.*, **31**(5):1131-1137. [doi:10.2135/cropsci1991.00111183X003100050009x]

Massman, J., Cooper, B., Horsley, R., Neate, S., Dill-Macky, R., Chao, S., Dong, Y., Schwarz, P., Muehlbauer, G.J., Smith, K.P., 2011. Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm. *Mol. Breed.*, **27**(4):439-454. [doi:10.1007/s11032-010-9442-0]

McVean, G., 2009. A genealogical interpretation of principal components analysis. *PLoS Genet.*, **5**(10):e1000686. [doi:10.1371/journal.pgen.1000686]

Metzker, M.L., 2009. Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**(1):31-46. [doi:10.1038/nrg2626]

Mezmouk, S., Dubreuil, P., Bosio, M., Décousset, L., Charcosset, A., Praud, S., Mangin, B., 2011. Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. *Theor. Appl. Genet.*, **122**(6):1149-1160. [doi:10.1007/s00122-010-1519-y]

Mikel, M.A., Kolb, F.L., 2008. Genetic diversity of contemporary North American barley. *Crop Sci.*, **48**(4):1399-1407. [doi:10.2135/cropsci2008.01.0029]

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J., 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.*, **30**(2):190-193. [doi:10.1038/ng813]

Palaisa, K., Morgante, M., Tingey, S., Rafalski, A., 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. *PNAS*, **101**(26):9885-9890. [doi:10.1073/pnas.0307839101]

Pallotta, M.A., Asayama, S., Reinheimer, J.M., Davies, P.A., Barr, A.R., Jefferies, S.P., Chalmers, K.J., Lewis, J., Collins, H.M., Roumeliotis, S., *et al*., 2003. Mapping and QTL analysis of the barley population Amagi Nijo× WI2585. *Aust. J. Agric. Res.*, **54**(12):1141-1144. [doi:10.1071/AR02218]

Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. *PLoS Genet.*, **2**(12):e190. [doi:10.1371/journal.pgen.0020190]

Pritchard, J.K., Rosenberg, N.A., 1999. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, **65**(1):220-228. [doi:10.1086/302449]

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. *Genetics*, **155**(2):945.

Rafalski, A., 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.*, **5**(2):94-100. [doi:10.1016/S1369-5266(02)00240-6]

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M., Buckler IV, E.S., 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *PNAS*, **98**(20):11479-11484. [doi:10.1073/pnas.201394398]

Risch, N., Merikangas, K., 1996. The future of genetic studies of complex human diseases. *Science*, **273**(5281):1516-1517. [doi:10.1126/science.273.5281.1516]

Rostoks, N., Ramsay, L., MacKenzie, K., Cardle, L., Bhat, P.R., Roose, M.L., Svensson, J.T., Stein, N., Varshney, R.K., Marshall, D.F., 2006. Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *PNAS*, **103**(49):18656-18661. [doi:10.1073/pnas.0606133103]

Saisho, D., Purugganan, M.D., 2007. Molecular phylogeography of domesticated barley traces expansion of agriculture in the old world. *Genetics*, **177**(3):1765-1776. [doi:10.1534/genetics.107.079491]

Skøt, L., Humphreys, J., Humphreys, M.O., Thorogood, D., Gallagher, J., Sanderson, R., Armstead, I.P., Thomas, I.D., 2007. Association of candidate genes with flowering time and water-soluble carbohydrate content in *Lolium perenne* (L.). *Genetics*, **177**(1):535-547. [doi:10.1534/genetics.107.071522]

Song, B.H., Windsor, A.J., Schmid, K.J., Ramos-Onsins, S., Schranz, M.E., Heidel, A.J., Mitchell-Olds, T., 2009. Multilocus patterns of nucleotide diversity, population structure and linkage disequilibrium in *Boechera stricta*, a wild relative of *Arabidopsis*. *Genetics*, **181**(3):1021-1033. [doi:10.1534/genetics.108.095364]

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., Buckler, E.S., 2001. *Dwarf8* polymorphisms associate with variation in flowering time. *Nat. Genet.*, **28**(3):286-289. [doi:10.1038/90135]

Watterson, G.A., Guess, H.A., 1977. Is the most frequent allele the oldest? *Theor. Popul. Biol.*, **11**(2):141-160. [doi:10.1016/0040-5809(77)90023-5]

Wei, X., Jackson, P.A., McIntyre, C.L., Aitken, K.S., Croft, B., 2006. Associations between DNA markers and resistance

to diseases in sugarcane and effects of population substructure. *Theor. Appl. Genet.*, **114**(1):155-164. [doi:10.1007/s00122-006-0418-8]

Wilson, L.M., Whitt, S.R., Ibáñez, A.M., Rocheford, T.R., Goodman, M.M., Buckler, E.S., 2004. Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell*, **16**(10):2719-2733. [doi:10.1105/tpc.104.025700]

Würschum, T., Maurer, H.P., Kraft, T., Janssen, G., Nilsson, C., Reif, J.C., 2011. Genome-wide association mapping of agronomic traits in sugar beet. *Theor. Appl. Genet.*, **2**(2):1-11.

Wych, R.D., Rasmusson, D.C., 1983. Genetic improvement in malting barley cultivars since 1920. *Crop Sci.*, **23**(6):1037-1040. [doi:10.2135/cropsci1983.0011183X002300060004x]

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**(2):203-208. [doi:10.1038/ng1702]

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.*, **42**(4):355-360. [doi:10.1038/ng.546]

Zhou, H., 2011. Association Mapping of Multiple Disease Resistance in US Barley Breeding Germplasm. PhD Thesis, University of Minnesota, St. Paul, USA.

Zhu, C.G., Buckler, M., Yu, E.S., 2008. Status and prospects of association mapping in plants. *Plant Genome*, **1**(1):5-20. [doi:10.3835/plantgenome2008.02.0089]