Theor Appl Genet (2011) 122:623–632
DOI 10.1007/s00122-010-1474-7

ORIGINAL PAPER

Population structure and linkage disequilibrium in oat (*Avena sativa* L.): implications for genome-wide association studies

M. A. Newell · D. Cook · N. A. Tinker · J.-L. Jannink

Received: 17 June 2010 / Accepted: 11 October 2010 / Published online: 2 November 2010

Abstract The level of population structure and the extent of linkage disequilibrium (LD) can have large impacts on the power, resolution, and design of genome-wide association studies (GWAS) in plants. Until recently, the topics of LD and population structure have not been explored in oat due to the lack of a high-throughput, high-density marker system. The objectives of this research were to survey the level of population structure and the extent of LD in oat germplasm and determine their implications for GWAS. In total, 1,205 lines and 402 diversity array technology (DArT) markers were used to explore population structure. Principal component analysis and model-based cluster analysis of these data indicated that, for the lines used in this study, relatively weak population structure exists. To explore LD decay, map distances of 2,225 linked DArT marker pairs were compared with LD (estimated as r^2). Results showed that LD between linked markers decayed rapidly to $r^2 = 0.2$ for marker pairs with a map distance of 1.0 centi-Morgan (cM). For GWAS, we suggest a minimum of one marker every cM, but higher densities of markers should increase marker-QTL association and therefore detection power. Additionally, it was found that LD was relatively consistent across the majority of germplasm clusters. These findings suggest that GWAS in oat can include germplasm with diverse origins and backgrounds. The results from this research demonstrate the feasibility of GWAS and related analyses in oat.

Communicated by B. Keller.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1474-7) contains supplementary material, which is available to authorized users.

M. A. Newell
Department of Agronomy,
Iowa State University, Ames, IA 50011, USA

D. Cook
Department of Statistics,
Iowa State University, Ames, IA 50011, USA

N. A. Tinker (✉)
Agriculture and Agri-Food Canada, ECORC,
960 Carling Ave, Ottawa, ON K1A OC6, Canada
e-mail: nick.tinker@agr.gc.ca

J.-L. Jannink (✉)
USDA-ARS, Robert W. Holley Center for Agriculture and Health,
Cornell University Department of Plant Breeding and Genetics,
407 Bradfield Hall, Ithaca, NY 14853, USA
e-mail: jeanluc.jannink@ars.usda.gov

Introduction

Oat (*Avena sativa* L.) is a grass species grown as a grain or forage crop predominantly in temperate short-season regions. Oat has received significant attention in recent years due to the human health benefits of consuming it as a whole-grain food. There are many active oat breeding programs around the world where improved oat varieties are developed through phenotypic selection for complex traits such as disease resistance, yield, lodging, and stress tolerance. For example, there are at least 12 publicly funded oat breeding programs in the USA and Canada. Recently, methods to improve the response to selection have focused on the identification of individual loci, known as quantitative trait loci (QTL), controlling these complex traits. There are numerous QTL mapping studies that have utilized linkage-based analysis of bi-parental populations in oat (e.g. reviewed by Rines et al. 2006; Holland 2007). Although this has proven to be a powerful approach for QTL detection, it delivers low-resolution, population-specific QTL,

and samples only a small portion of the allelic diversity present in the germplasm available (Zhu et al. 2008). Genome-wide association studies (GWAS) attempt to overcome the pitfalls associated with linkage mapping in bi-parental populations. Genome-wide association studies have the ability to identify useful allelic diversity and to map this diversity with high resolution within complex plant pedigrees that are typical of breeding programs (Jannink et al. 2001). From a practical perspective, GWAS have been applied in many grain crops, including rice, maize, barley, and wheat (Agrama et al. 2007; Beló et al. 2008; Kraakman et al. 2006; Zheng et al. 2009). Implementation of GWAS in oat for QTL detection could be valuable to the oat community.

The ability of GWAS to deliver high-power, high-resolution results is largely dependent on the extent of linkage disequilibrium (LD) within the working population. Also known as gametic phase disequilibrium, LD is defined as the non-random association of alleles at two loci (Falconer and Mackay 1996) and is affected by mutation, admixture, selection, drift, population structure associated with breeding history, and reproductive biology (reviewed by Flint-Garcia et al. 2003). Additionally, since the mechanisms mentioned may differentially affect different genomc regions, this can introduce LD heterogeneity across the genome. This makes the power and resolution achieved in GWAS highly dependent on the species and the population being evaluated.

The extent of LD among the common grass species varies with respect to the crop and the population chosen for evaluation. For example, in maize, an allogamous species, LD decays over relatively shorter distances compared with autogamous crops. Remington et al. (2001) reported that LD (measured as r^2) declined to 0.1 within 1,500 basepairs (bp) in a set of 102 maize inbred lines representing breeding germplasm from temperate and tropical regions. Tenaillon et al. (2001) found similar results in a group of 25 maize lines consisting of 16 landraces and nine elite inbreds: LD decayed to 0.15 at 500 bp for the combined dataset. Unlike maize, barley is a self-pollinated crop with strong population structure due to variation in growth habit and kernel row number. Zhang et al. (2009) reported that LD extends to 2.6 cM at r^2 equal to 0.2 for a group of elite Canadian lines. This is in agreement with Hamblin et al. (2010) in a study of North American elite germplasm. Similar to barley, oat is a self-pollinated species; thus it is expected that LD decay will occur over relatively long map distances.

In addition to LD, the power and resolution of GWAS is also dependent on marker density. Until recently, the lack of genetic markers and a method to deliver high-throughput genotyping have limited the options for identifying QTL in oat. Diversity Array Technology (DArT) markers developed recently for oat have greatly increased the density of available markers (Tinker et al. 2009). These markers were developed based on random clones isolated from 60 elite varieties of diverse global origin, making them useful in diversity analysis as well as in linkage mapping. Since they can be applied in parallel using a cost-effective assay, they show good potential for use in QTL detection, comparative mapping, marker-assisted selection (MAS), and genomic selection.

Due to the fact that the power and resolution of GWAS depends greatly on the extent of LD across the genome, it is important to survey this extent. In this study, we determine and discuss the population structure and the extent of LD among DArT markers in an extensive worldwide collection of oat consisting of varieties, breeding lines, and landraces.

Materials and methods

Plant material

Datasets from four independently assembled germplasm collections were combined in this study to increase the diversity and representation of the results. The four component datasets consisted of 462, 466, 198, and 279 lines, each set having been assembled for a variety of other purposes that will be published elsewhere. The set of 462 lines consisted of current North American elite varieties. The set of 466 lines was a world collection of oat germplasm from the Germplasm Resources Information Network (USDA ARS 2010) consisting of varieties, breeding lines, and landraces. The set of 198 lines consisted of varieties of global origin that were used by Tinker et al. (2009) in the initial DArT development work, and the set of 279 lines was an extension of this set intended for use in association mapping. The combined dataset represented a total of 1,405 lines from 53 countries.

DArT genotyping

Plants were grown under greenhouse conditions and tissue was collected from single plants or from multiple bulked seedlings originating from seed of a single oat panicle. Isolation of DNA was performed by a variety of methods in use by different collaborating laboratories. It was already evident in the work of Tinker et al. (2009) that these different extraction methods would not affect the DArT assays. DArT marker analysis was performed by Diversity Arrays P/L, Canberra, Australia using methods described by Tinker et al. (2009). Due to the fact that DArT markers were under development during the initial stages of this work, the four datasets submitted for DArT genotyping provided datasets with varying numbers of markers.

Data curation

Because four independently assembled component datasets were merged, some duplicate lines existed. Previous research has suggested that this could contribute to biased estimates of LD (Breseghello and Sorrells 2006). In addition to duplicate lines, the datasets ranged in numbers of markers from 1,001 to 1,958; thus a core marker set was required. Data curation was implemented to accomplish the following: (1) identify and merge lines that were submitted more than once across the datasets and (2) identify and merge redundant markers, and (3) eliminate lines and markers with insufficient data points.

The DArT marker assay produces dominant marker scores, which were represented as a matrix of 1s and 0s. Genetic distances (measured as Manhattan distance) were calculated across all genotype pairs and expressed as the proportion of the maximum. Pairs of lines with genetic distances <5% that had similar names were merged, retaining the line with the most complete data. Pairs with genetic distances <2% that shared at least 200 markers in common were also merged regardless of nomenclature. The rationale for this was that even if the two lines are in fact distinct, they must be strongly related, and that representing them as a single entry would be more meaningful in the determination of LD.

After removal of redundant lines, redundant markers were also removed using a similar fashion. Markers based on DArT clones having DNA sequence data represented in the same contiguous DNA assembly (Tinker et al. 2009), and with scores that differed by <2% were merged. When two markers belonged to the same sequence assembly, but differed by more than 2%, the marker with most missing data was removed. This approach assumes that all markers belonging to the same sequence assembly are identical. Markers with identical scores were also merged if the scores were non-ambiguous across 100 or more lines. This merging process was performed for compatibility with the mapping data set and the resulting linkage maps. This process resulted in the merging of only four pairs of markers (1% of the total marker number) and is unlikely to have affected the results because such markers would be at distance zero and in perfect disequilibrium. Although these assumptions of marker identity may occasionally be incorrect, these procedures were selected as a conservative approach to remove redundant markers that would otherwise cause an upward bias in the estimate of LD for the short artificial linkage intervals caused by slight variations in the scoring of identical markers.

Because many of the markers were not scored on all of the datasets, markers that were scored in fewer than 80% of the lines were removed, followed by removal of all monomorphic markers. These final steps ensured that only markers and lines with a sufficient amount of data were retained. Genetic distances among all lines were re-computed after these final data curation steps. The resulting dataset after data curation consisted of 1,205 lines and 402 markers.

Model-based cluster analysis

The Mclust package (Fraley and Raftery 2006) in the statistical software R was used to identify clusters among lines. A model-based approach was used because it determines the number of clusters and cluster membership simultaneously and it does not have underlying genetic assumptions that are rarely met. The package identifies the optimal model according to the Bayesian Information Criterion (BIC) for expectation–maximization (EM) initialized by hierarchical clustering for parameterized Gaussian mixture models (Fraley and Raftery 2007). Due to the large number of dimensions (402 markers), cluster analysis was implemented on the principal components. By using the principal components instead of the marker data, it was possible to fit models of varying shape, size, and orientation. Models with between 2 and 30 clusters were compared.

Accounting for population structure

Principal components analysis (PCA) was implemented to account for population structure (Price et al. 2007). First, missing marker values were replaced by the mean for the marker. PCA was applied to the lines using the prcomp function in the statistical software R, which adequately handles computational issues of high-dimensional data. The choice of the number of principal components used was based on the scree plot of eigenvalues (Cattell 1966). Singular value decomposition was used to account for population structure using the appropriate number of principal components from above. A matrix representing marker scores expected on the basis of population structure was calculated as $R = UDV'$, where U is a matrix of left singular vectors, D is a diagonal matrix of singular values, and V is a matrix of right singular vectors. The population structure matrix (R) was subtracted from the marker data, and LD was calculated as described below.

Population structure was accounted for in all LD calculations using the aforementioned approach. For the six clusters (percent of the variation shown in parentheses), four (23.4), three (18.1), three (16.2), four (24.8), four (22.6), and seven (62.2) principal components were used, respectively. For the entire sample, five principal components (22.8) were used to account for population structure.

Linkage disequilibrium

Three common methods for calculating LD exist in plants, denoted by D, D', and r^2. For this research, the correlation

squared (r^2) was used because (1) it is not as highly influenced by small sample sizes and low allele frequencies (Flint-Garcia et al. 2003), and (2) it is relevant for QTL mapping because it relates the amount of variance explained by the marker to the amount of variance generated by the associated QTL (Zhu et al. 2008). The calculation used is as follows: $r^2 = [\sigma_{XY}/(\sigma_X\sigma_Y)]^2$, where σ_{XY} is the covariance between marker X and marker Y, and σ_X and $\sigma_Y$are the standard deviations for marker X and marker Y, respectively. This calculation was applied using the R statistical package (R Development Core Team 2009) to each marker pair using all available data points. The functional relationship between LD and map distance was determined by fitting the nonlinear model (Sved 1971) $r^2 = 1/(1 + 4ad)$, where d is the map distance in cM and a is an estimated regression coefficient. The parameter a can also be interpreted as the effective population size of the population to which the analysis was applied.

Map distances

The current lack of a consensus map in oat presents the issue of deciding on a map distance measure that will adequately describe the LD decay. Most commonly, the LD for a pair of markers is compared to a map distance that is taken directly from a consensus map. However, the only map on which a large number of markers have been resolved is the updated Kanota × Ogle map (Tinker et al. 2009), where there are approximately twice as many linkage groups as there are chromosomes in oat. In order to avoid bias and artifacts introduced by the map, it was decided to use direct counts of recombination events between each available pair of markers in the published Kanota × Ogle mapping data as the primary measure of map distance. These recombination estimates were expressed as centi-Morgan (cM) distances using the Kosambi mapping function and are identified hereafter as "PairD". For example, if two markers "A" and "C" showed a direct pair-wise map distance of 20 cM, this distance would be used in estimating LD decay regardless of what their distance was on the resolved map, even if they were not resolved to the same linkage group on the published map. Markers at distances of greater than 40 cM were considered unlinked. In order to compare this approach to the more common approach based on a resolved linkage map, we reconstructed the same analysis using resolved cM distances (identified as "MapD") from the published map, which included 665 linked pair-wise LD measurements. Furthermore, since other unpublished map data were available, and since this would allow estimation of a greater number of pair-wise marker distances, we tested a third additional approach: estimates of recombination between each pair of markers were computed from all available mapping populations, and these were averaged (represented hereafter as "AveD"). Pairs of markers were excluded under the following situations: (1) when the minimum distance was less than or equal to 5 cM but individual estimates varied by more than 10 cM and (2) when the minimum distance was greater than 5 cM and less than or equal to 20 cM and the distances varied by more than 200% of minimum. Pairs with AveD greater than 40 cM were considered unlinked.

Results

Data curation

The final dataset contained 1205 lines and 402 markers (additional data for the final dataset are given in Online Resource 1). Lines from the United States, Canada, and Germany were highly represented, accounting for 44, 16.5, and 4.6% of the lines in the study, respectively. Other countries represented by greater than 2% of the lines were Sweden, Turkey, the United Kingdom, and the Russian Federation. The remaining 46 countries accounted for 19.3% of the lines included in the study. Some lines had multiple origins upon merging those with similar genotype that varied in origin, and these accounted for 2.2% of the lines. A further 2.8% of the lines had unknown origin. In total, there were 2,225 linked and 15,541 unlinked (>40 cM) pair-wise LD estimates used for the primary analysis of this study (PairD).

Population structure

Principal component analysis and model-based cluster analysis were used to explore population structure. Cluster analysis was implemented on the first five principal components explaining 22.8% of the variation. The BIC for the different models were similar beyond six clusters but a defined peak was not present; therefore, the decision was made to use only six clusters (C1–C6). All of the clusters can be separated in the first three principal components that account for 8.8, 4.91, and 3.6% of the variation, respectively (Fig. 1). Although the clusters can be separated by the first three PCs, the data represent a cloud in space where distinct clusters are not readily seen.

The number of lines per cluster ranged from 20 to 334 (Table 1). Geographic origins of lines making up >5% of a cluster were evaluated to assess the relationship between the clusters and their origins. Because US and Canadian lines account for 60% of the 1,205 lines in the study, it was expected that these countries would make up a substantial amount of all clusters. In general, clusters C1 and C2 contain mainly Canadian lines, and clusters C3, C4, and C5 contain mainly US lines. Although it is difficult to distin-

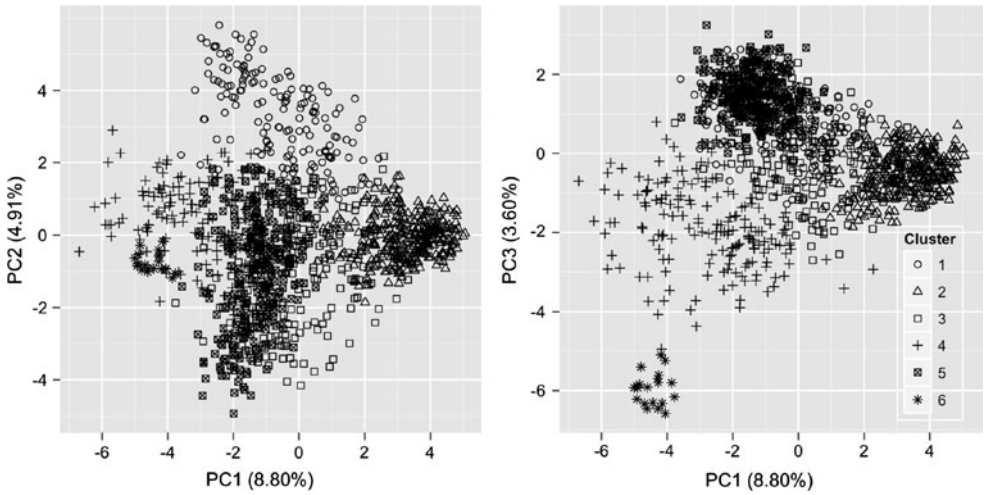

Fig. 1 Scatter plots of principal component 1 (PC1) versus PC2 (**a**) and PC1 versus PC3 (**b**) showing the 1,205 lines making up the six clusters. Percent of the total variation accounted for by each PC is denoted in the axes titles in *parentheses*

Table 1 Descriptions of the six oat clusters designated C1 to C6 identified using model-based cluster analysis

Cluster	Number of lines	Origins[a]	Representative line[b]
C1	166	CA, US	Assiniboia
C2	334	DE, US, SE, CA, RU, YU	Triple crown
C3	209	US, CA, Unk	Jay
C4	184	US, TR, UK, AU	Kanota
C5	292	US, CA	Ogle
C6	20	US, CA, UK, Mul	Red oats

AU Australia, *CA* Canada, *DE* Germany, *RU* Russian Federation, *SE* Sweden, *TR* Turkey, *US* The United States, *YU* Yugoslavia, *Mul* multiple, *Unk* unknown

[a] Origins are identified as countries from which >5% of the lines in a cluster originated

[b] An arbitrary but widely recognized line was selected to represent each cluster

Table 2 Average pair-wise genetic distances between the six germplasm clusters

	C2	C3	C4	C5	C6
C1	28	70	80	45	157
C2	–	43	82	75	165
C3		–	36	24	138
C4			–	98	73
C5				–	155

guish the clusters by origin due to the high frequency of US and Canadian lines, clusters could be associated with important released oat lines. These include Assiniboia, Triple Crown, Jay, Kanota, and Ogle for clusters C1 to C5, respectively. Cluster C6 was a small cluster consisting of red oats (*A. sativa* ssp. *byzantina* K. Koch), typically grown as winter oats in the southern US.

Cluster relationships

Differences exist in the pair-wise relationships between clusters as seen in the PC scatter plots. Quantitative results for genetic distances (measured as Manhattan distance) between clusters are shown in Table 2; a graphical representation is shown in Fig. 3. C6, a small group of winter red oats, is most distant from all other clusters, with an average distance of 138, whereas C3 is most closely related to all other clusters with an average distance of 62. All other clusters have similar average distances to other clusters in the range of 74–91. Interestingly, the genetic distances are in agreement with the origin and/or adaptation of the representative lines (Table 1) falling within each cluster. That is, C3 and C5 are most closely related (24) and are defined by Jay and Ogle, Indiana and Illinois lines, respectively, from the USA. The next most closely related are C1 and C2 (28), which contain Assiniboia and Triple Crown. Assiniboia was bred in Manitoba, Canada, while Triple Crown originated in Sweden but is released in, and highly adapted to, Western Canada. Last, C4 and C6 have a pair-wise genetic distance of 73; these clusters contain Kanota (C4), a winter red oat from Kansas and another small group of mostly red oat lines (C6). These results suggest that clustering was efficient in separating major lines and oat types for the germplasm used in this study.

Linkage disequilibrium

The extent of LD in a species determines the power and resolution of GWAS. For oat, it was expected that decay of LD would be over relatively long map distances because of its breeding history and reproductive biology. One way of summarizing breeding history is by estimating the effective population size of the sample analyzed. For this study, the effective population size for the entire sample estimated from non-linear regression of r^2 on map distance was 92

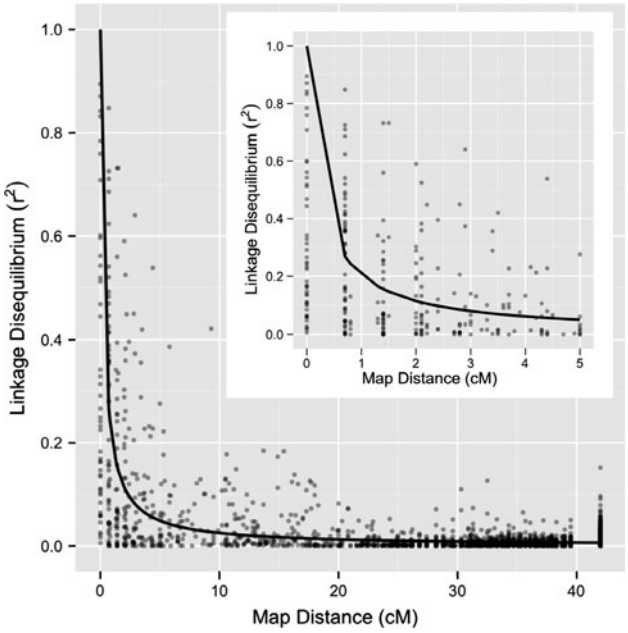

Fig. 2 Scatter plot of LD (r^2) decay for all 2,225 linked marker pairs as a function of map distances (cM) for the 1,205 oat lines. The LD for 14,122 unlinked marker pairs is shown at an arbitrary distance of 42 cM. The *inset graph* shows the LD decay from 0 to 5 cM

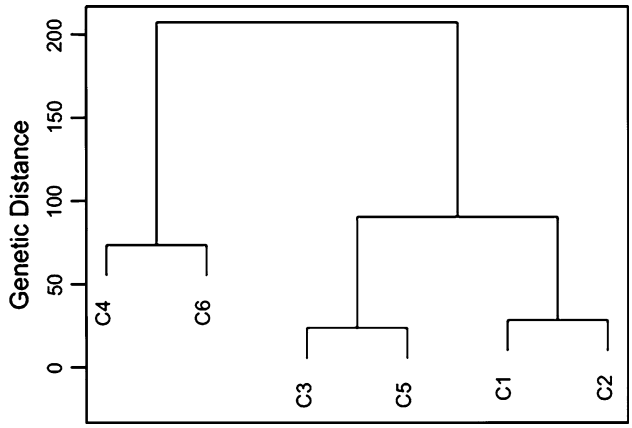

Fig. 3 Cluster dendrogram showing the genetic distances (measured as Manhattan distance) between clusters using wards linkage

with a standard error of six. In other words, an "ideal" randomly mating population of 92 diploid individuals (Falconer and Mackay 1996) would be expected to have a similar rate of LD decay as the global oat population. Based on decay of LD in barley (Hamblin et al. 2010) and wheat (Chao et al. 2007), this effective population size appears typical of elite cultivated small grains. For the 2,225 linked marker pairs used in this study, LD decays such that r^2 is equal to 0.1 at 2.5 cM (Fig. 2). Within clusters, C1 to C5 show similar trends in LD decay, but quantitative differences can be seen (Table 3). In C6, the small cluster of mainly red oat lines, LD decays sporadically most likely due to the small sample size (Table 3). Cluster C5 had a relatively slower LD decay compared with the entire sample. A possible reason for this result could be the frequent use of the variety Ogle in this group as a parent in crosses for variety development (Figs. 3, 4).

Two alternate distance measures were used to determine if decay was dependent on the choice of distance. The first, MapD, was based on map distances in the Kanota × Ogle genetic map, and the second, AveD, was based on average recombination distances from six populations of oat, each composed of 75–150 recombinant inbred lines at generations 4–6 (unpublished data). MapD and AveD were composed of 665 and 3,228 linked pair-wise LD measurements, respectively. Both alternate measures of map distances showed similar trends compared with PairD, the primary distance used for analysis in this study, with some minor differences (Fig. 5). MapD resulted in a slower decay that was more sporadic than both PairD and AveD. However, AveD resulted in a curve very similar to PairD, suggesting that decay is not highly dependent on the distance measurement.

Identification of disequilibrium between unlinked marker pairs can be useful since these markers can affect GWAS. Unlinked marker pairs with high disequilibrium could indicate that unknown linkages or pseudo-linkages are present in oat. Quantitative results of LD between unlinked marker pairs are shown in Table 3 and are expectedly low. The average LD between unlinked markers is lower for the entire sample (0.004) than it is within each germplasm cluster (0.006–0.075). The estimate of LD

Table 3 Distribution of r^2 for the six germplasm clusters and the entire sample that included 1,205 lines

Map distance (cM)	Linkage disequilibrium (r^2)						
	Cluster						Entire sample
	C1	C2	C3	C4	C5	C6	
0	0.280	0.221	0.308	0.260	0.363	0.239	0.320
>0 to 5	0.164	0.093	0.151	0.117	0.217	0.080	0.156
>5 to 10	0.087	0.043	0.057	0.058	0.095	0.109	0.057
>10 to 40	0.015	0.009	0.010	0.012	0.012	0.079	0.008
Unlinked (>40)	0.009	0.006	0.007	0.010	0.007	0.075	0.004

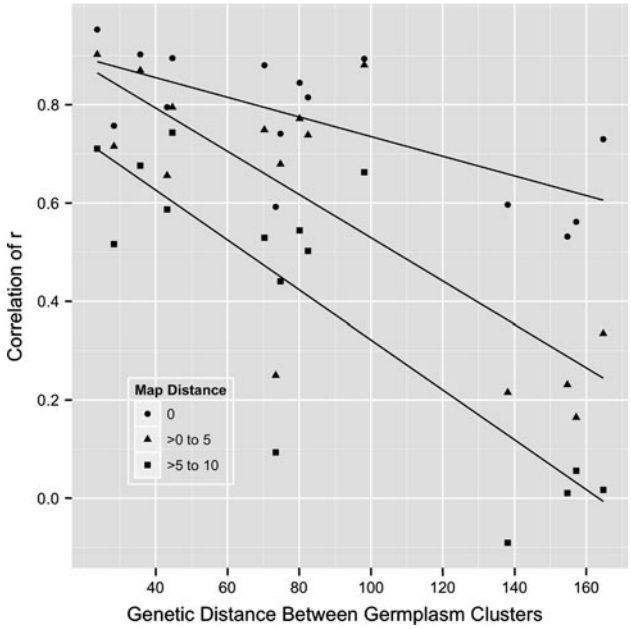

Fig. 4 Correlation of *r* as a function of genetic distance between germplasm clusters at map distances of 0, >0 to 5, and >5 to 10 cM. Regression *lines* refer to the three map distances from *top* to *bottom*, respectively

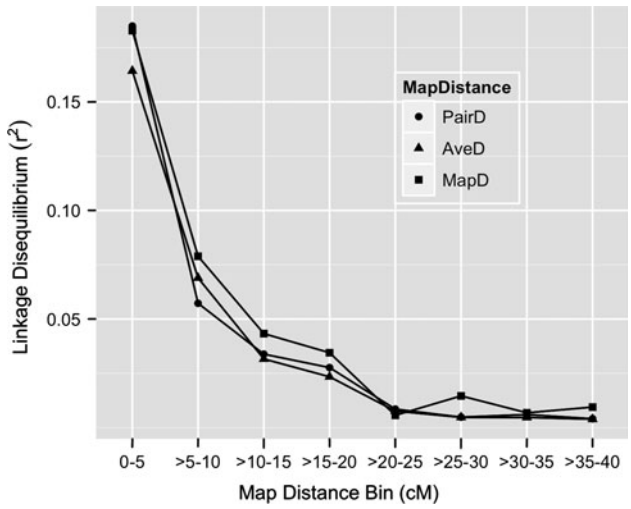

Fig. 5 Average linkage disequilibrium (measured as r^2) at binned map distances for the three distance measures used in this study, *PairD*, *AveD*, and *MapD*

between unlinked marker pairs is shown in Fig. 2 at an arbitrary distance of 42 cM. The LD for these marker pairs are below an r^2 value of 0.1 except for one point that has a value of 0.15.

Correlation of *r* between clusters

The design of GWAS depends on the consistency of gametic or LD phase across germplasm clusters. To address this, the correlation of LD (measured as the correlation of *r*) between cluster pairs was explored at map distances of zero, greater than zero to five, and greater than five to 10 cM (Fig. 4). The correlation of *r*, rather than r^2, was used because *r* has a signed value, making it more relevant to discuss consistency. Ten thousand permutations of the analysis of variance (ANOVA) were used to test for a significant relationship between correlation of *r* and the genetic distance for varying map distances. There was a significantly negative relationship between the correlation of *r* and the genetic distance for all map distance intervals ($P = 0.01$). These results match our expectations as well as previous results found in barley (Hamblin et al. 2010). However, when cluster C6 is removed from the analysis, there was no longer a significant relationship at any of the three map distance intervals. Quantitative differences in consistency of LD between the six germplasm clusters from map distances 0 to 5 and >5 to 40 cM are shown in Table 4. At the interval of 0–5 cM, the average correlation of *r* between clusters is 0.81, ranging from 0.71 to 0.92 when cluster C6 was excluded. Pair-wise correlations of *r* for cluster pairs that included C6 were lower, ranging from 0.32 to 0.49, which would be expected given the small sample size of this cluster ($n = 20$) and its more distant relationship with all other clusters (Table 2). For the interval of >5 to 40 cM, the average correlation of *r* for cluster pairs was relatively lower with a value of 0.20. Given the weak population structure in oat that was described earlier, one would expect LD phase to be consistent across germplasm, as was shown here. These results indicate that LD is consistent across most oat germplasm, most notably at short-range map distances up to 5 cM where marker-QTL associations are most likely to occur.

Table 4 Average correlation of LD (measured as *r*) between clusters at map distances of 0–5 (above diagonal) and >5 to 40 cM (below diagonal)

	C1	C2	C3	C4	C5	C6
C1	–	0.74***	0.79***	0.80***	0.83***	0.32***
C2	0.24***	–	0.71***	0.77***	0.71***	0.49***
C3	0.24***	0.33***	–	0.89***	0.92***	0.37***
C4	0.21***	0.25***	0.30***	–	0.89***	0.39***
C5	0.39***	0.19***	0.39***	0.29***	–	0.36***
C6	0.09**	0.06*	−0.02	0.04	0.03	–

*** Significant at 0.0001

** Significant at 0.01

* Significant at 0.05

Discussion

Population structure

The level of population structure in a species has implications on the design and analysis of GWAS. For this study,

PCA and cluster-analysis were used to explore the amount of population structure. Major population structure in an organism can be observed by plotting the first few PCs (Menozzi et al. 1978; Price et al. 2007). For oat, examination of the first three principal components indicates that there is weak population structure within the germplasm evaluated. In contrast, barley is known to have strong population structure due to 2-row, 6-row, spring, and winter types (Hamblin et al. 2010). Plotting of the first two PCs for barley results in non-overlapping, distinct clusters similar to what would be expected for a species with strong population structure (Malysheva-Otto et al. 2006; Zhang et al. 2009). Oat also has four recognizable types, including naked, hulled, spring, and winter. Interestingly, although it was not tested because the majority of lines included in the study were spring, hulled types or unknown, the observed population structure is most likely independent of these types, other than for the small but distinct group of red winter varieties in cluster C6. Possible reasons for this could include different practices in exchange of germplasm among breeding programs, or breeding methods that more frequently utilize crosses among different oat types. Fortuitously, the relatively weak population structure of oat reported here suggests that GWAS can successfully span a wide diversity of oat types.

Linkage disequilibrium

For this study, LD decay was explored for a set of germplasm consisting of 1205 lines, and for six derived germplasm clusters identified using model-based cluster analysis. The amount of LD in the combined population decayed at a rate very similar to that within the derived germplasm clusters. Theory developed to predict LD in the presence of population structure suggests that overall LD should be similar to sub-population LD when extensive migration occurs between sub-populations (Sved 2009). When migration is high, LD within the overall population should behave as in an unstructured, large population, as is the case for oat. While we have no records of the frequency of crosses made by breeding programs between the clusters that we found, we expect that they do occur, given the low level of differentiation between clusters. In contrast, barley is known to have strong population structure and shows large differences between germplasm clusters with respect to LD decay (Hamblin et al. 2010; Zhang et al. 2009). Most importantly, the design of GWAS depends on the consistency of gametic or LD phase across germplasm clusters. If the phase of LD differs among clusters, independent GWAS need to be conducted within each cluster. It is expected that LD for closely linked markers will be most similar between closely related clusters and that this similarity will decrease more slowly at greater distances in closely related clusters than it will in more distant clusters. Our results demonstrated that this relationship was significant only if cluster C6 was included. A possible reason for this result is that the remaining clusters are not genetically distinct enough for a large change in the correlation of r to occur. The consistency of LD phase for oat across most germplasm clusters identified here indicates that GWAS can include germplasm with diverse origins and backgrounds.

Using alternate estimates of map distance taken directly from the map (MapD), a similar but slightly slower and less consistent rate of LD decay was observed. The Kanota × Ogle genetic map is incomplete and contains more linkage groups than the 21 chromosomes in oat. Therefore, it is possible that some of the inconsistency in decay is due to the much smaller number of marker pairs for which an estimate of map distance was available. However, the slower rate of decay observed using MapD may also result from ubiquitous errors in map construction. For example, if markers “A” and “C” are separated by several other markers on the map, and some mis-scoring has occurred in these intervening markers, then the map distance between markers “A” and “C” will be artificially stretched, resulting in an apparent slower decay in LD at longer map distances. For this reason, we think that the estimates based on PairD provide a more accurate representation of LD decay.

One concern for this study and for future research is the sub-optimal application of dominant bi-allelic markers in LD and GWAS studies, since a single dominant or recessive allele class can include genotypes with multiple alleles at a target locus. When this occurs, the ability to detect associations with a specific target allele is weakened. This problem will also apply to the use of bi-allelic SNP markers, except that the co-dominant nature of SNPs can eliminate confounding effects of residual heterozygosity. DArT and SNP markers may also differ in that a variety of mutation mechanisms can lead to a change of DArT allele (e.g., both point mutation in the DArT restriction site or a long insertion into the DArT fragment could lead to loss of fragment amplification), whereas generally only point mutations will cause a change of SNP allele. DArT marker mutation rates may therefore be higher than SNP mutation rates (which are known to be very low). The implications in the current study are that LD values are probably underestimated relative to what they would be if multi-allelic markers were available. The use of dominant or bi-allelic markers may become more powerful if multi-locus haplotypes can be used for GWAS, especially if higher-density SNP resources are developed at a later date.

Another source of error may arise due to the presence of markers that segregate as multiple loci but are identified by a single marker name. When this occurs, the dominant alleles of two or more loci will be confounded in the set of diversity data, even though a single locus may segregate

normally within a given mapping population. This scenario is known to exist for DArT markers in other polyploid species such as wheat and likely exists to some degree in oat. Based on indirect estimates, the frequency of markers that segregate as multiple loci in oat is potentially about 5% (Tinker et al. 2009). Thus, two markers identified as being unlinked in Kanota × Ogle could be linked in a portion of the diversity panel, or vice versa. To address this, we have tested a third set of estimates for map distance that were derived from averaging the recombination fractions across multiple mapping populations. Although this analysis provided a greater number of data points, the results were highly similar to the primary analysis using PairD from Kanota × Ogle, demonstrating that the estimates of LD decay are quite robust and probably not influenced by segregation of duplicate markers.

The power and resolution of GWAS is dependent on the extent of LD in a given population, assuming a suitable marker system is in place. In practice, r^2 between a marker and a QTL is equal to the percent of phenotypic variation of a QTL that can be explained by a marker. For oat, LD was on average 0.2 for DArT markers separated by 1.0 cM. Thus, the results from this study indicate that a marker every cM (2,000 markers total) would explain, on average, 20% of QTL variance. Since a marker and a QTL must also have similar minor allele frequencies to be in LD, we suggest that the number of markers should be on the order of 10,000 to increase the probability of identifying a marker that is in high LD with a QTL. At the current rate of development for oat DArT markers, this marker density is approachable in the near future. The authors are also engaged with collaborators in the development of new SNP marker resources for oat, a process that has been greatly assisted by the use of DArT markers to select diverse germplasm. However, these results do not imply that GWAS cannot succeed at much lower densities. The markers employed in this study are somewhat clustered, and since DArT markers are designed to target gene-rich regions, it is possible that many of these will be the same clusters that contain QTL. Many linkage blocks of favorable QTL alleles may also have been deliberately or inadvertently selected in breeding programs, and these same linkage blocks will contain markers in high LD. Therefore, while we encourage the development of high-density maps for GWAS, we do not discourage the exploration and utilization of QTL associations using existing molecular tools. Most importantly, this work demonstrates the distances at which LD can be expected, and the non-dependence of LD on population structure.

Acknowledgments We thank Adrienne Moran-Lauter, Charlene Wight, and Andrezej Kilian for excellent technical assistance. We thank the following for contributing germplasm: Franco Asoro, Jennifer Mitchel Fetch, Weikai Yan, Stephen Harrison, Heidi Kaeppler, Ron Barnet, Mike McMullen, Herbert Ohm, Deon Stuthman, Brian Rossnagel, Tom Fetch, Luiz Carlos Federizzi, Axel Diedericsen, Hermann Bürstmayr. We thank the following for allowing us to access unpublished marker data from additional segregating oat populations: Luiz Carlos Federizzi, Itamar Nava, Steve Molnar, Kyle Gardner, Brian Rossnagel, Aaron Beattie, Eric Jackson, and Rebekah Oliver. We greatly appreciate the contributions of many additional collaborators, too numerous to list, to the germplasm collections and to the additional mapping work. Funding for this work was provided by USDA-NIFA grant number 2008-55301-18746 "Association genetics of beta-glucan metabolism to enhance oat germplasm for food and nutritional function," Agriculture and Agri-Food Canada, and the North American Milling Association.

References

Agrama HA, Eizenga GC, Yan W (2007) Association mapping of yield and its components in rice cultivars. Mol Breed 19:341–356

Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li B, Tingey S (2008) Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. Mol Genet Genomics 279:1–10

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.). Genetics 172:1165–1177

Cattell RB (1966) The scree test for the number of factors. Multivar Behav Res 1:245–276

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 47:1018–1030

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman Group Limited, London

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Fraley C, Raftery AE (2006) MCLUST version 3 for R: normal mixture modeling and model-based clustering. Technical report no. 504, Department of Statistics, University of Washington

Fraley C, Raftery AE (2007) Bayesian regularization for normal mixture estimation and model-based clustering. J Classif 24:155–181

Hamblin MT, Close TJ, Bhat PR, Chao S, Kling JG, Abraham KJ, Blake T, Brooks WS, Cooper B, Griffey CA, Hayes PM, Hole DJ, Horsley RD, Obert DE, Smith KP, Ullrich SE, Muehlbauer GJ, Jannink J-L (2010) Population structure and linkage disequilibrium in U.S. barley germplasm: implications for association mapping. Crop Sci 50:556–566

Holland JB (2007) Genetic architecture of complex traits in plants. Curr Opin Plant Biol 10:156–161

Jannink J-L, Bink CAM, Jansen RC (2001) Using complex plant pedigrees to map valuable genes. Trends Plant Sci 6:337–342

Kraakman ATW, Martinez F, Mussiraliev B, van Eeuwijk FA, Niks RE (2006) Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. Mol Breed 17:41–58

Malysheva-Otto LV, Ganal MW, Röder MS (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). BMC Genet 7:6

Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. Science 201:786–792

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2007) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. PNAS 98:11479–11484

Rines HW, Molnar SJ, Tinker NA, Phillips RL (2006) Oat. In: Berlin KC (ed) Genome mapping and molecular breeding in plants, cereals and millets volume 1. Springer, Berlin, pp 211–242

Sved JA (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor Popul Biol 2:125–141

Sved JA (2009) Correlation measures for linkage disequilibrium within and between populations. Genet Res Camb 91:183–192

R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphisms along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). PNAS 98:9161–9166

Tinker NA, Kilian A, Wight CP, Uszynska KH, Wenzl P, Rines HW, Bjornstad A, Howarth CJ, Jannink J-L, Anderson JM, Rossnagel BG, Stuthman DD, Sorrells ME, Jackson EW, Tuvesson S, Kolb FL, Olsson O, Federizzi LC, Carson ML, Ohm HW, Molnar SJ, Scoles GJ, Eckstein PE, Bonman JM, Ceplitis A, Langdon T (2009) New DArT markers for oat provide enhanced map coverage and global germplasm characterization. BMC Genomics 10:39

USDA, ARS, National Genetic Resources Program (2010) *Germplasm Resources Information Network*—(*GRIN*) [Online Database]. National Germplasm Resources Laboratory, Beltsville, Maryland. Available: http://www.ars-grin.gov/cgi-bin/npgs/acc/query.pl. Accessed 7 April 2010

Zhang LY, Marchand S, Tinker NA, Belzile F (2009) Population structure and linkage disequilibrium in barley assessed by DArT markers. Theor Appl Genet 119:43–52

Zheng S, Byrne PF, Bai G, Shan X, Reid SD, Haley SD, Seabourn BW (2009) Association analysis reveals effects of wheat glutenin alleles and rye translocations on dough-mixing properties. J Cereal Sci 50:283–290

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:2–5