12-2012

Genome-wide association study for oat (Avena sativa L.) beta-glucan concentration using germplasm of worldwide origin

Mark A. Newell
The Samuel Roberts Noble Foundation

Franco G. Asoro
Iowa State University

M. Paul Scott
United States Department of Agriculture

Pamela J. White
Iowa State University, pjwhite@iastate.edu

William D. Beavis
Iowa State University, wdbeavis@iastate.edu

See next page for additional authors

Genome-wide association study for oat (Avena sativa L.) beta-glucan concentration using germplasm of worldwide origin

Abstract
Detection of quantitative trait loci (QTL) controlling complex traits followed by selection has become a common approach for selection in crop plants. The QTL are most often identified by linkage mapping using experimental F2, backcross, advanced inbred, or doubled haploid families. An alternative approach for QTL detection are genome-wide association studies (GWAS) that use pre-existing lines such as those found in breeding programs. We explored the implementation of GWAS in oat (*Avena sativa* L.) to identify QTL affecting β-glucan concentration, a soluble dietary fiber with several human health benefits when consumed as a whole grain. A total of 431 lines of worldwide origin were tested over 2 years and genotyped using Diversity Array Technology (DArT) markers. A mixed model approach was used where both population structure fixed effects and pair-wise kinship random effects were included. Various mixed models that differed with respect to population structure and kinship were tested for their ability to control for false positives. As expected, given the level of population structure previously described in oat, population structure did not play a large role in controlling for false positives. Three independent markers were significantly associated with β-glucan concentration. Significant marker sequences were compared with rice and one of the three showed sequence homology to genes localized on rice chromosome seven adjacent to the *CslF*gene family, known to have β-glucan synthase function. Results indicate that GWAS in oat can be a successful option for QTL detection, more so with future development of higher-density markers.

Keywords
Agronomy

Disciplines
Agricultural Science | Agriculture | Agronomy and Crop Sciences | Biochemistry | Biotechnology | Food Chemistry | Genetics | Plant Breeding and Genetics

Authors
Mark A. Newell, Franco G. Asoro, M. Paul Scott, Pamela J. White, William D. Beavis, and Jean-Luc Jannink

Theor Appl Genet (2012) 125:1687–1696
DOI 10.1007/s00122-012-1945-0

ORIGINAL PAPER

Genome-wide association study for oat (*Avena sativa* L.) beta-glucan concentration using germplasm of worldwide origin

Mark A. Newell · Franco G. Asoro · M. Paul Scott · Pamela J. White · William D. Beavis · Jean-Luc Jannink

Received: 16 January 2012 / Accepted: 15 July 2012 / Published online: 3 August 2012

Abstract Detection of quantitative trait loci (QTL) controlling complex traits followed by selection has become a common approach for selection in crop plants. The QTL are most often identified by linkage mapping using experimental F_2, backcross, advanced inbred, or doubled haploid families. An alternative approach for QTL detection are genome-wide association studies (GWAS) that use pre-existing lines such as those found in breeding programs. We explored the implementation of GWAS in oat (*Avena sativa* L.) to identify QTL affecting β-glucan concentration, a soluble dietary fiber with several human health benefits when consumed as a whole grain. A total of 431 lines of worldwide origin were tested over 2 years and genotyped using Diversity Array Technology (DArT) markers. A mixed model approach was used where both population structure fixed effects and pair-wise kinship random effects were included. Various mixed models that differed with respect to population structure and kinship were tested for their ability to control for false positives. As expected, given the level of population structure previously described in oat, population structure did not play a large role in controlling for false positives. Three independent markers were significantly associated with β-glucan concentration. Significant marker sequences were compared with rice and one of the three showed sequence homology to genes localized on rice chromosome seven adjacent to the *CslF* gene family, known to have β-glucan synthase function. Results indicate that GWAS in oat can be a successful option for QTL detection, more so with future development of higher-density markers.

Communicated by I. Mackay.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1945-0) contains supplementary material, which is available to authorized users.

M. A. Newell
The Samuel Roberts Noble Foundation, Ardmore, OK 73401, USA
e-mail: manewell@noble.org

F. G. Asoro · W. D. Beavis
Department of Agronomy, Iowa State University, Ames, IA 50011, USA

M. P. Scott
Corn Insects and Crop Genetics Research Unit, USDA-ARS, Ames, IA 50011, USA

P. J. White
Department of Food Science and Human Nutrition, Iowa State University, Ames, IA 50011, USA

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

The objective of quantitative trait locus (QTL) mapping is to identify genomic regions that are associated with a phenotype of interest. The identified regions, linked to a causal genetic variant, can be selected in a breeding program with the goal of improving genetic gain per unit time (Lande and Thompson 1990). Furthermore, identification of causal variants increases our understanding of the mechanisms that affect a trait, which may in turn lead to improved selection methods. In linkage studies, experimental and random F_2, backcross, advanced inbred, or doubled haploid families are developed. Although this approach is powerful in QTL detection, the shortcomings

of the approach are numerous (Jannink et al. 2001). The high power to detect QTL linked to marker loci is due to the extensive linkage disequilibrium (LD), spanning large chromosomal regions, generated from the mating of two inbred lines. A positive impact of such LD is the low marker density required to adequately cover the genome. Conversely, QTL positioning has low resolution such that the marker could be as much as 10–30 cM (centi-Morgans) from the causal allele (Kearsey and Farquhar 1998).

An alternative approach to QTL mapping is genome-wide association studies (GWAS), also known as LD mapping. In contrast to QTL mapping based on bi-parental populations, GWAS uses a sample of lines from the broader breeding population, unrelated by any specific crossing design (Zhu et al. 2008). In such studies, associations between genotype and phenotype depend on historical LD broken down by many generations of recombination. Hence, for GWAS a larger number of markers are required to assure LD between markers and causative alleles throughout the genome, thus enabling fine-scale mapping. The reward, however, is that the short LD blocks that exist in such groups of lines can result in high-resolution mapping of QTL. This situation is advantageous because the LD utilized in selection for such QTL will not easily be broken down by recombination. GWAS has been widely used in human genetic studies where the development of experimental populations is impossible. In contrast to the experimental populations developed for linkage mapping, a major issue facing GWAS is the unknown relationship among individuals, also known as population structure that can lead to spurious associations (Kennedy et al. 1992). To statistically control for structure and the covariance among individuals, a mixed model analysis (Yu et al. 2006) that fits population structure, marker, and polygenic effects has been widely implemented.

Oat (*Avena sativa* L.), a grass species grown as a grain or forage crop predominantly in temperate short-season regions, poses another issue for linkage mapping. Oat lacks a consensus map, making comparisons with other QTL studies difficult. Together, the adequate levels of LD and a marker system that has the ability to saturate the genome make GWAS a superior approach to identification of QTL in oat. Newell et al. (2011) explored genome-wide LD in oat and showed that to attain values of $r^2 = 0.2$ between markers, one marker per centi-Morgan (cM) was needed. The most comprehensive oat map available, 'Kanota' × 'Ogle,' is 1,890 cM (Wight et al. 2003); thus, approximately 2,000 markers would be required to reach an average LD between markers and causal alleles of 0.2. Recent advances in Diversity Array Technology (DArT) markers in oat and current single nucleotide polymorphism (SNP) development can provide such density requirements for oat.

Although oat production worldwide has been decreasing, it is still highly prized for its positive health benefits. The health benefits associated with consuming oat as a whole grain is attributed to (1-3, 1-4)-β-D-glucan (hereafter referred to as β-glucan), a hemicellulose found in cereal endosperm cell walls (Fincher 2009). Research on the role of oat β-glucan in the human diet has shown that it improves health with respect to blood pressure (Keenen et al. 2002), diabetes (Jenkins et al. 2002), cholesterol (Braaten et al. 1994), and immune response (Estrada et al. 1997). β-glucan viscosity is a primary factor affecting the aforementioned health benefits, although the mechanisms involved are not well understood (Colleoni-Sirghie et al. 2003). Independent studies in oat and barley have demonstrated a positive relationship between viscosity and β-glucan concentration (Chernyshova et al. 2007; Izydorczyk et al. 1998). Thus, β-glucan concentration is a good target for selection in oat breeding programs.

An unintended consequence of the breeding process is the loss of genetic variants that control valuable traits (Robertson, 1960; Hill and Robertson, 1968). This is often the case for elite material where intense selection, possibly for other traits, has occurred and the useful genetic variants are lost due to fixation of the undesired allele at a locus. Thus, the identification of QTL in germplasm from worldwide origin that includes breeding lines and landraces may enable the use of genetic variants not currently found in elite varieties. The objectives of this study were to (1) conduct a GWAS to identify QTL associated with increased β-glucan concentration in oat germplasm of worldwide origin and, (2) determine the effects of population structure in mixed model association analyses for oat.

Materials and methods

Genetic material

Genetic material was requested from the National Small Grains Collection within the National Plant Germplasm System. Selection of accessions was based on three criteria, the standardized β-glucan values from the Germplasm Resources Information Network (GRIN), the accession origin, and a pre-screening of the materials to confirm their ability to flower. Three data sets in the GRIN database included β-glucan information, these included oat.beta-glucan.madison.07, 91, and 95. Together these data sets included over 6,000 varieties, breeding lines, and landraces of worldwide origin. Because the three data sets were measured in different years, and each set contained different lines, the values were standardized within each data set. Standardized β-glucan values were calculated within each data set as $Z = \frac{x-\mu}{\sigma}$ where x is the raw GRIN β-glucan

value, μ is the mean for the particular data set, and σ is the standard deviation for the particular data set. In order to increase power for the analysis, lines were chosen that spanned the tails of the standardized β-glucan distribution. The second criterion for selection was based on the origin of accessions: lines were selected to maximize the diversity of the germplasm set. This was done to sample the array of alleles present in available oat germplasm. Approximately, half of the lines selected were from the upper tail and half were from the lower tail of the distribution while taking into account the origin of the materials. At this stage, 607 lines were requested from GRIN and pre-screened in a single environment to confirm their ability to flower in the target environment. Information about the 466 lines included in the study is provided in Online Resource 1.

Genotypic and phenotypic analysis

Plants were grown under greenhouse conditions and tissue was collected from a single plant of each accession. Extraction of DNA was done with methods prescribed by Diversity Arrays P/L, Canberra, Australia and described by Tinker et al. (2009). Accessions, derived from a single DNA parent plant, were grown as hill plots in Ames, Iowa in 2009 and 2010 in an incomplete block design. Years, replicates, and incomplete blocks were considered as fixed effects and accessions as random effects. Two replicates were grown in both 2009 and 2010 where incomplete blocks consisted of 5×5 hill plots. For the 2009 and 2010 season, hill plots were grown at 40 and 12 in. apart, respectively. Field checks for β-glucan included nine varieties and breeding lines representing a range of β-glucan concentration. Plots were harvested, threshed, cleaned, and 0.5–3 g of seed per hill, depending on availability, were dehulled using a compressed-air oat laboratory dehuller manufactured by Codema Inc. (Eden Prairie, MN). The field design was conserved for laboratory analysis of β-glucan. An enzymatic approach for evaluation of β-glucan concentration was implemented using the streamlined mixed linkage β-glucan kit (Megazyme Int., Wicklow, Ireland) with minor modifications. The laboratory protocol was modified to increase the throughput capability by reducing reagent amounts by 90 %, thus enabling use of a 96-well plate for evaluation (Newell et al. 2012). Although 0.5–3 g of seed per hill was initially ground, 8–12 mg of the flour was used for the β-glucan assay. Statistical analysis for β-glucan was implemented in SAS version 9.2 (SAS Institute 2010) using PROC MIXED for mixed-effects models.

Data cleaning

To remove possible errors and redundancies in markers and lines that may cause false associations, a data-cleaning step was performed using the R statistical software (R Development Core Team 2009). This included a four-step process, all of which have been previously described as necessary steps in preparation of GWAS (Miyagawa et al. 2008). Initially, the data set consisted of 466 accessions and 1001 DArT markers. First, markers with call rates of less than 0.8 were removed; this step was implemented to remove markers that likely contained errors. This step removed only one marker, resulting in 466 accessions and 1000 markers. Second, markers with minor allele frequency (MAF) of less than 0.01 were removed, as they do not contribute substantially to the variation in the data. This step reduced the number of markers from 1,000 to 982 markers. We realize that alleles at a frequency of 0.01 will only be present in five individuals, giving them little power to be associated with a QTL. At the same time, if a QTL allele segregates at this low frequency, it will only be possible to tag it with a low-frequency marker. We therefore opted to leave in such low-frequency markers. Third, markers were merged that diverged by less than 1 % across the genotyped lines, thus combining markers that were in near perfect LD. This step resulted in a matrix of 466 accessions and 796 markers. Lastly, accessions that differed by less than 1 % on the markers were merged, thus removing accession redundancies. After implementation of this step, the final data set was reduced to 431 accessions and 796 markers (Online Resource 2). After data cleaning, the mean value for each marker across lines was used to impute missing values for each marker.

Association analysis

Association analysis to identify QTL controlling oat β-glucan was implemented in R using the GWA function with modification in the rrBLUP package (Endelman 2011). The GWA function applies a mixed-linear model that can account for both population structure and marker-based kinship, denoted by K, originally described by Yu et al. (2006). The model used for association analysis was $Y = \text{mean} + Ma + Pv + Zu + \text{e}$ where Y is a vector of β-glucan BLUPs from the analysis of phenotype (above), a is a vector of marker fixed effects, v is a vector of population structure fixed effects, and u is a vector of random polygenic effects. M is a matrix of marker scores for the markers included in the model. P is described in detail below. Z is the incidence of the polygenic effects in the observations and was taken to be an identity matrix. The variance of u is assumed to be $\text{var}(u) = 2KVg$ where K is the marker-based kinship matrix and Vg is the polygenic effect variance. Marker-based kinship was calculated using the emma.kinship function in the emma package (Kang et al. 2008). Models that did not include K in the mixed model used an identity matrix indicating no relationship between individuals.

Models accounting for differing levels of population structure fixed effects with and without K were assessed. The first model, denoted by P1, included the n_p principal components that were significantly correlated with the response variable at $p \leq 0.01$. Hence, n_p was chosen based purely on the number of significant axes. The second model, denoted by P2, included the first n_p principal components, a common approach used when principal component analysis (PCA) is used to account for population structure. For both models, the number of dimensions was equal to n_p, thus comparisons could be made across models. In all, six models were assessed including P1, P1K, P2, P2K, K, and a simple model where neither P nor K was included in the model. The six models were assessed for their ability to control for type I errors by plotting the distribution of p values for the markers, where uniformly distributed p values indicate proper control for type I errors. The Benjamini and Hochberg (1995) false discovery rate (FDR) at 0.25 was used to control for multiple testing. Two R^2 measures were used to assess the amount of variability explained by each marker. In addition to the standard measure, R^2, the likelihood ratio-based R^2 denoted by R^2_{LR} was also calculated, as it has been shown to be a better estimate of R^2 in GWAS (Sun et al. 2010). In order to determine how the significant markers affect percent β-glucan, the preferred model from the above analysis was used for analysis of the subset of significant markers.

Rice sequence homology

It is unlikely that the DArT markers identified as significant are functional; instead, they likely rely on LD with the causal locus. Therefore, the sequences of the significant DArT markers were compared for their sequence homology with the rice (*Oryza sativa* L.) genome in a three-step approach. First, a set of candidate genes in rice were identified that included all of the *Csl* and the *CesA* gene families. These genes were chosen because the *Csl* gene family has been shown to be involved in β-glucan biosynthesis (reviewed by Burton and Fincher 2009) and the *CesA* family in cellulose synthase. The *CesA* gene family was also included because it has been shown to encode the catalytic subunit of cellulose synthase and is likely co-regulated with genes in the *Csl* gene family (Burton et al. 2004). Second, significant DArT sequences (available in Tinker et al. 2009) were compared based on their sequence homology with the entire rice genome with an E-value threshold of 1×10^{-15} and a hit score of greater than 500 (Ouyang et al. 2007). This level of stringency was used because it is expected that there may be differences in sequence given the interspecies nature of the sequences. Lastly, because it is likely that the significant DArT sequences are not functional, but depend on LD with the causal locus, we tested if the DArT sequences were adjacent to the rice candidate genes. A threshold distance for declaring a DArT sequence to be adjacent to a rice candidate gene was determined by picking a point at random in the rice genome and determining its distance in kb with the nearest rice candidate gene. This process was repeated 1,000,000 times to construct the distribution of distances under the null hypothesis that DArT sequence positions were random relative to rice candidate genes. The distance at the 5 % quantile of this null distribution was taken as the threshold to declare adjacency to a candidate gene. Thus, a DArT sequence within the 5 % quantile, 247 kb, of a rice candidate gene is said to be adjacent to that gene.

Results

Germplasm selection

In total, 466 accessions were selected ranging in standardized β-glucan from −3.27 to 4.74 % with a bimodal distribution of β-glucan concentration reported in GRIN. The number of lines that were classified as either high or low based on the standardized β-glucan values were 238 and 228, respectively. The lines in the distribution with lower β-glucan values ranged from −3.27 to −1.33 %, whereas the higher distribution ranged from 0.57 to 4.74 %. Thus, selection based on this criterion was apparent in the distribution. The second criterion was selection of lines in order to increase the diversity of the germplasm set. In total, the selected accessions were from 49 countries from around the world (Fig. 1). The majority of lines were from the USA, Turkey, Germany, and the Russian Federation with 171, 32, 28, and 27 lines, respectively. For the top 14 countries that accounted for most of the lines in the set, most were evenly split between the high and low β-glucan classifications according to GRIN.

Phenotypic analysis

Raw β-glucan values were lower than expected and ranged from 1.44 to 6.20 % with an average of 3.90 % β-glucan. Best linear unbiased predictions (BLUPs) for β-glucan ranged from −1.38 to 2.40. Model assumptions were diagnosed by graphical representation of the residuals and the correlation between the residuals and fitted values. Residuals were normally distributed and there was no evidence of correlation between the residuals and fitted values, and thus model assumptions were met. There was a significant (p value < 0.0001) and non-significant (p value $= 0.4081$) genotype × year and genotype × rep within year interaction, respectively, for β-glucan content. Although there was a significant genotype × year interaction, data were analyzed jointly because predictions were highly

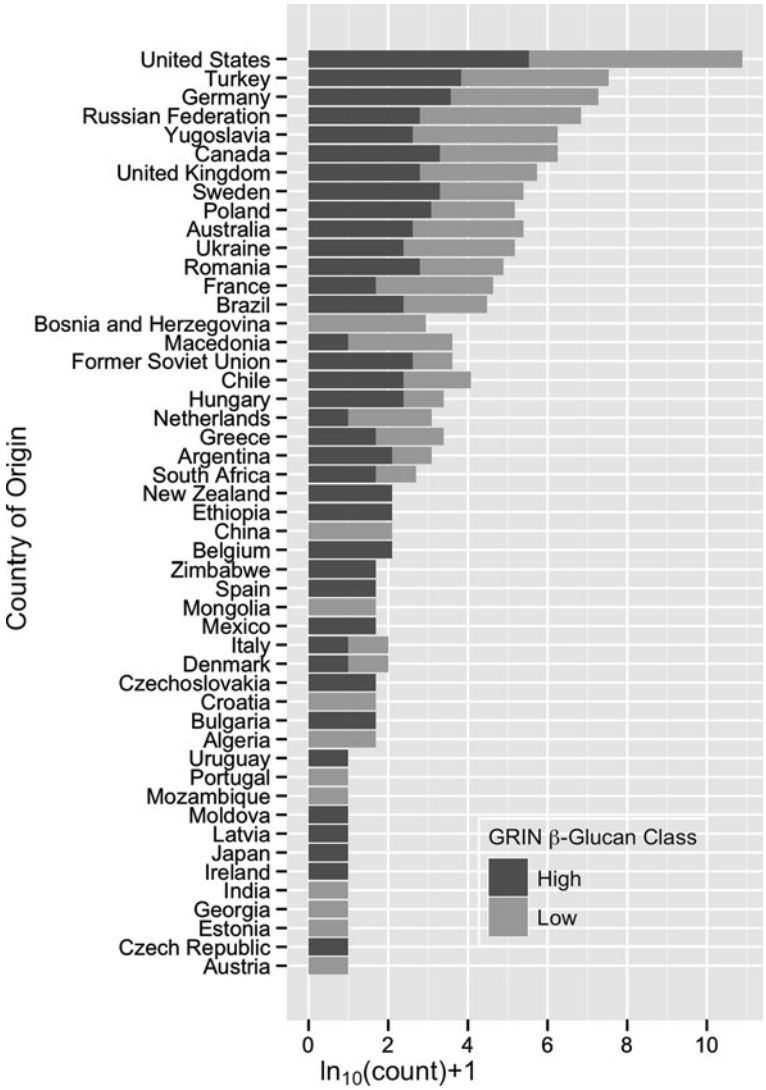

Fig. 1 *Bar graph* showing the countries of origin for the lines included in the study colored by the GRIN β-glucan classification. Counts are represented by the ln(count) + 1 with values for *High* and *Low* GRIN classification stacked within each country

correlated across years (0.87) and this enables identification of stable alleles across environments. As expected, due to the procedure in which the lines were selected, the distribution of β-glucan values was bimodal (Fig. 2). Field checks ranged from −0.64 for Buff, a naked oat bred for high protein content to 2.40 for N979-5-1-22, an Iowa State University line bred for high β-glucan concentration. Three of the checks (HiFi, ND030288, and N979-5-1-22) had β-glucan values greater than any of the lines included in the study. The average β-glucan BLUPs for the two selection groups according to the GRIN classifications were −0.34 and 0.34 for the low and high class, respectively. There was a highly significant correlation ($r = 0.68$) between the GRIN and β-glucan BLUPs (Fig. 3).

Population structure and kinship

The level of population structure in the data set was explored to gain insight into its possible effect on the association analysis. Principal component analysis on the marker data (with missing scores for a marker imputed as the mean value for that marker) showed that the first three axes accounted for only 14.5, 6.1, and 3.7 % of the total variation in the data. These low levels of variation explained, along with visualization of the principal components, indicated that population structure among the lines was weak as compared to other grass species such as barley (Hamblin et al. 2010). There are three small groups of lines that do tend to deviate from the average set of lines included in the study; these groups can easily be seen in principal components two and three (Fig. 4). The first group, designated as group A, contains 17 lines in which the majority is from Turkey and all but three are landraces. The second group, designated group B, contains eight lines, all being from MD, USA except one that is from South Africa. The third group, designated group C, contains 16 lines, all breeding lines from Maryland, USA. Besides these three groups of lines (41 in total), the remaining lines

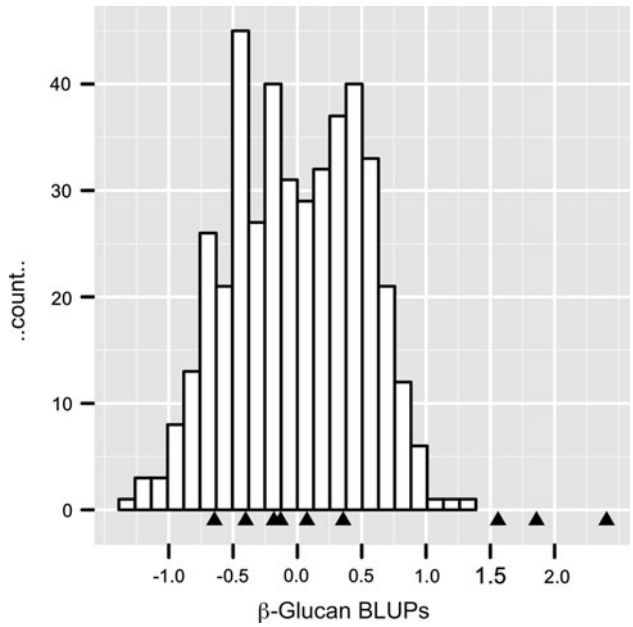

Fig. 2 Distribution of β-glucan BLUPs for the lines showing the bimodal distribution as a result of the selection process. *Triangles* beneath the distribution represent the β-glucan BLUPs for field checks including Buff, Excel, Winona, Cherokee, IA02130-2-2, Baker, HiFi, ND030288, and N979-5-1-22 from *left* to *right*, respectively

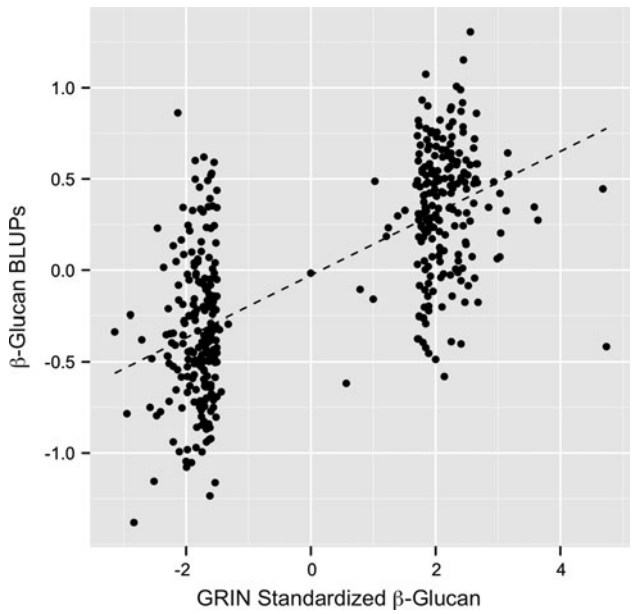

Fig. 3 *Scatter plot* showing the relationship between the GRIN standardized β-glucan values and the β-glucan BLUPs based on 2 years with two replicates per year. The *dashed line* represents the regression between the two measures (correlation = 0.68)

do not form distinct clusters with respect to the first three principal components. Mean β-glucan values for the groups were 3.57, 4.17, 3.62, and 4.06 for groups A, B, and C, and

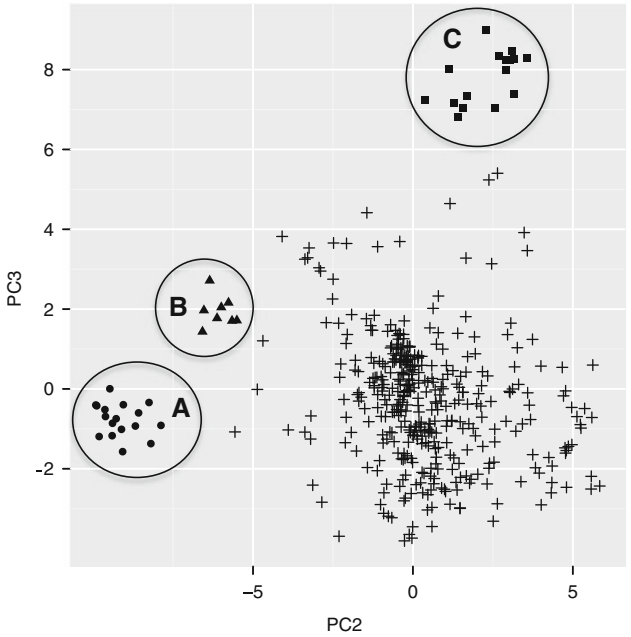

Fig. 4 Principal component 2 (PC2) versus PC3 computed on the marker data for the 466 accessions used in this study. PC2 and PC3 accounted for 6.1 and 3.7 % of the variation in the data, respectively. Three groups, referred to as A, B, and C, are clearly separated from the remaining *lines*

the remaining lines, respectively. These results of low levels of population structure are in agreement with previous results for oat that included a wide variety of germplasm of worldwide origin (Newell et al. 2011).

Evaluation of P in the mixed model

The effect of population structure in the mixed model approach was tested by observation of each model's ability to reduce the number of false positives. In order to assess a model's ability to account for this, the distribution of observed p values for the six models was plotted in the negative $\log_{10}$ scale (Fig. 5). The null hypothesis, or expectation, follows a uniform distribution represented by a diagonal line. When there is an over-abundance of low p values, the distribution of p values does not follow this line, but rises above it on the negative log scale. In contrast, a model that sufficiently accounts for the number of false positives follows the expectation except for the few significant markers. Five principal components (5, 14, 25, 30, and 31) were significantly correlated to β-glucan, and thus the population structure fixed effects included five dimensions. For P1, 5 % of the total variation in the data was explained by the significant PCs. This is far less than for P2 in which the first five principal components accounted for 30 % of the total variation. Among the six models tested, P1, P2, and the simple model did not sufficiently reduce the number of false positives. The only model that showed

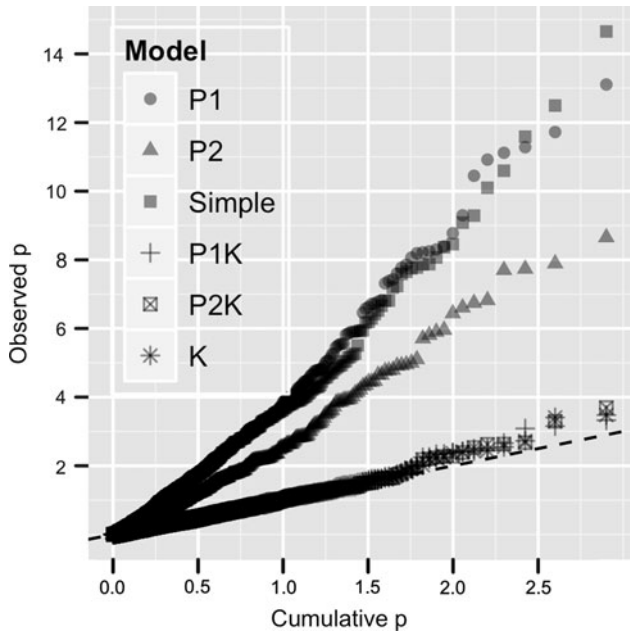

Fig. 5 Distribution of *p* values for the six models included for evaluation of P in the mixed model. *Axes* represent the cumulative *p* versus the observed *p* in the negative $\log_{10}$ scale where the *dashed line* represents the null expectation. Models that do not include K (P1, P2, and simple) do not adequately account for false positives in contrast to P1K, P2K, and K that effectively reduce the number of false positives

improvement over the simple model in decreasing the number of false positives for these three models was P2; this result is most likely due to the model fixed effects accounting for a large amount of the variation in the marker data (30 %). The P1 model did not show an improvement over the simple model most likely for the same reason. In contrast, when K was included in each of those three models, the distribution of *p* values followed the expected uniform distribution. This indicated that the addition of K in the model sufficiently accounted for relationships between individuals and effectively reduced the number of false positives. In addition, it also demonstrated the small effect that principal components can have on the number of false positives regardless of whether they comprise a small (P1) or large (P2) proportion of the variation in the data.

Association analysis

As expected from the evaluation of P in the mixed model, models that did not include K identified a large number of significant, false positive markers. The P1, P2, and simple models had 398, 286, and 441 significant (FDR < 0.25) markers, respectively. Given such large numbers of significant markers that are likely false positives, these models were excluded from further analyses. The numbers of significant markers were greatly reduced with the addition of K in the mixed model, where the P1K, P2K, and K models had only three, two, and two markers significantly associated with β-glucan (FDR < 0.25; Fig. 6). Two of the significant markers were in common to all models. These were oPt.0133 and oPt.17174/oPt.8715, where the forward slash refers to makers that were merged during data cleaning. A third marker, oPt.6825/oPt.0112, was significantly associated with β-glucan according to the P1K model. Thus, three independent markers were identified as significant in the P1K model and were not in LD with one another (between marker r^2 values ranged from 0.004 to

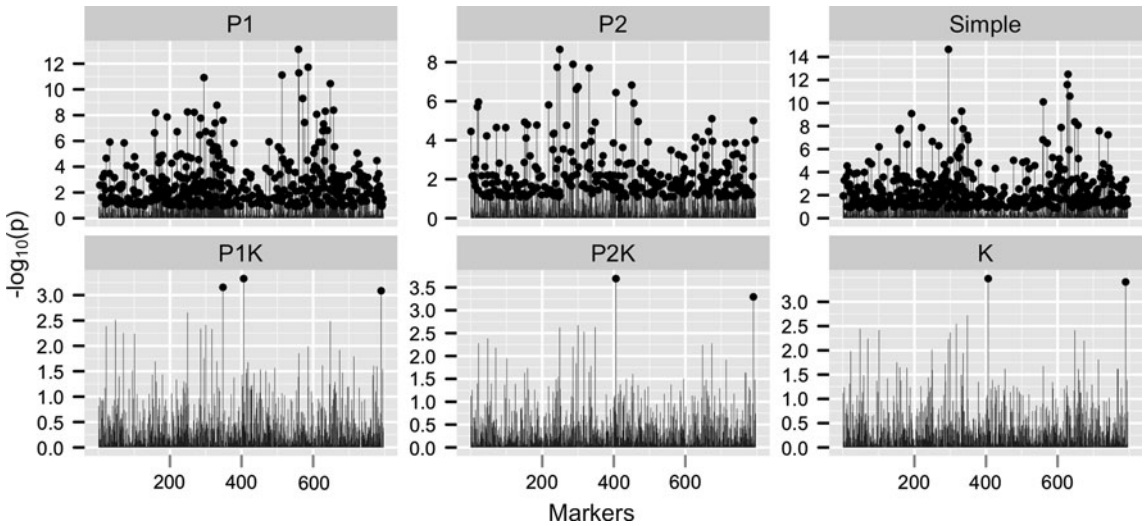

Fig. 6 *Manhattan plots* for the eight models used for association analysis showing the scores for each marker in no particular order calculated as $-\log_{10}(p)$. Significant scores using an FDR of 0.25 are represented by *bold points*

Table 1 Results of the significant markers for the P1K, P2K, and K models including the score, R^2, R^2_{LR}, the FDR q-value obtained using the Benjamini and Hochberg method for multiple testing, and the marker effects

	Marker name		
	oPt.0133	oPt.6825/ oPt.0112	oPt.17174/ oPt.8715
Model	P1K, P2K, K	P1K	P1K, P2K, K
Score [$-\log_{10}$(p)]	3.33, 3.69, 3.48	3.15	3.09, 3.29, 3.41
R^2 (%)	2.7, 3.2, 3.0	2.6	2.5, 2.8, 2.9
R^2_{LR} (%)	2.8, 3.1, 2.9	2.6	2.5, 2.7, 2.8
q-value	0.22, 0.16, 0.16	0.22	0.22, 0.20, 0.16
Marker effect	−0.366	−0.264	−0.248

The FDR cutoff for significance was 0.25. Marker names separated by a forward slash represent merged markers as a result of the data-cleaning steps

0.031). The R^2 and R^2_{LR} values varied around about 3 % for all of the markers across models. The three markers that were identified, oPt.0133, oPt.6825/oPt.0112, and oPt.17174/oPt.8715, affected beta-glucan concentration by 0.37, 0.26, and 0.25 %, respectively (Table 1).

Rice sequence homology

As part of the three-step process for the rice sequence comparison, 47 candidate rice genes were identified that were within the *Csl* and *CesA* gene families. The 47 rice candidates spanned all of the rice chromosomes with most occurring on rice chromosome seven. The fewest rice candidates occurred on rice chromosome 11 with only one DArT sequence that had sequence homology to the rice candidates. In total, five oat DArT sequences were compared with rice for sequence homology because two of the significant markers had been merged. The DArT marker sequences included were oPt.0133, oPt.6825, oPt.0112, oPt.17174, and oPt.8715. Two of the markers, oPt.17174 and oPt.8715, which were merged, resulted in no sequence homology with the rice genome and were thus excluded from further evaluation. The remaining three markers, oPt.0112, oPt.6825, and oPt.0133, had significant sequence homologies with 1, 3, and 33 sequences, respectively, to the rice genome. By our definition of adjacent, one of the DArT sequences, opt.0133, with homology to rice was adjacent to the *CslF* gene family, including *CslF1*, *2*, *3*, *4*, *8*, and *9*. The sequence with homology to oPt.0133 in rice on chromosome seven was within 137 kb of all of the genes in the *CslF* gene family. The *CslF* gene family is known to have β-glucan synthase function (reviewed by Fincher 2009). Information concerning the rice sequence homology including rice candidates and DArT marker homology are included in Online Resource 3.

Discussion

Numerous research studies have been implemented for GWAS using the mixed model approach that accounts for population structure and pair-wise kinship, initially described by Yu et al. (2006). For this study, we evaluated the inclusion of population structure fixed effects in the model. We found that including P, as principal components, in the model did not substantially decrease the number of false positives. Also, the number of false positives decreased as the amount of marker variation that the principal components accounted for increased. Similar results were found for simulated data where the K model performed as well or better than models including population structure (Bradbury et al. 2011). The Bradbury et al. (2011) study found this result to be consistent across varying numbers of QTL and heritability estimates. Similarly, the initial publication by Yu et al. (2006) found that including population structure showed an improvement over not including it only for traits highly correlated to population structure. This could partially explain the results in our study where the small influence of P in the mixed model was indicative of the low levels of population structure that exist in oat. One concern for oat is the effect of population structure due to spring and winter types and two, but inter-breeding species, *A. sativa* and the red oats (*A. sativa* ssp. *byzantia* K. Koch); however, visualization of the principal components did not exhibit these types in the panel evaluated. Newell et al. (2011) suggests that although these four groups do exist in oat, the majority are spring *sativa* types and population structure due to these groups is small because of admixture. In addition, the third criterion for selection of lines included a pre-evaluation for vernalization requirement, thus excluding winter types from the panel. In contrast to these results, Stich et al. (2008) implemented GWAS in wheat and found that including P in the mixed model improved control of false positives relative to just including K. However, as pointed out in Stich et al. (2008), inclusion of P in the mixed model had a large effect most likely because of the high levels of population structure that exists in wheat.

Three independent markers were identified to be associated with increased β-glucan concentration, two of which were in common for the three models that included K and one that was only significant in the P1K model. Previous studies have identified QTL associated with increased β-glucan concentration in oat (Kianian et al. 2000; De Koeyer et al. 2004). These were linkage mapping studies conducted on recombinant inbred line populations derived by crossing two inbred lines. Unfortunately, there was no agreement between the results presented here and previous studies. It is difficult to make a good comparison between this and previous studies because the map position of only

one of the three significant markers in our study is known (oPt.6825/oPt.0112 on linkage group 8; Tinker et al. 2009). Lastly, given the diversity of alleles represented in this study compared to previous studies, one might not expect a high level of agreement between them.

Despite the wide range of β-glucan in our panel, we identified only three independent markers significantly associated with β-glucan concentration. There are three possible explanations for this low number of associations. First, the marker density we had available may have been insufficient given the decay of LD to $r^2 = 0.2$ at 1 cM, (Newell et al. 2011). Hence, polymorphisms causing variation in β-glucan may have been in linkage equilibrium with our markers, and higher marker densities could have uncovered more QTL. A concurrent GWAS in elite oat using a similar marker set identified 15 markers significantly associated with β-glucan concentration (FG Asoro, personal communication). Though a less stringent FDR was used in that study (0.33), the greater number of associations identified may have come from the less rapid decay of LD in that panel of North American elite lines as compared to the global panel used in this study. A second reason for the low number of associations identified here may be that our global collection had more rare alleles causing variation in β-glucan than did the elite panel. Rare alleles cause less variation in the data and therefore may not be detected. Rare alleles are a leading hypothesis for the observation of "missing heritability" in human association studies (Yang et al. 2010). A third possibility that we cannot exclude is that the DArT markers were developed from a genetically narrow set of germplasm in relation to the lines used in this study. However, the DArT markers were developed from a panel of 182 accessions of global representation; thus, we do not believe that there would be much ascertainment bias in the marker development given this broad panel. It is also important to point out that there is the possibility that loci controlling β-glucan can appear co-located with loci controlling yield due to dilution effects. In order to test this, we also analyzed the phenotypic data with yield as a fixed effect. However, BLUPs from the analyses with and without yield as a fixed effect had a correlation of greater than 0.99, giving a strong indication that this was not the case.

The five oat DArT sequences identified as significant were also compared for sequence homology with rice in a three-step process to enable the significant markers to be matched by location to potential rice candidates. The three-step approach was implemented because it is unlikely that the DArT sequences are functional themselves but instead rely on LD with the causal locus. Synteny between oat and rice, however, would show the DArT marker to fall in a region with rice candidate genes. DArT sequences that fall within regions of likely candidates may support the notion that the DArT marker is truly associated with β-glucan concentration. Three of the markers had sequence homology to rice; one of these DArT sequences, opt.0133, was located on rice chromosome seven and was, by our definition, adjacent to the *CslF* gene family. However, it must be taken into consideration that the opt.0133 DArT marker has sequence homology with 33 sequences in the rice genome. Though our test does not pass multiple testing as it was designed only for single sequence tests, it does satisfy greater thresholds of 519 and 3.1×10^{-17} for hit score and *E* value, respectively, which we believe is of high value. The *CslF* gene family was previously shown in rice (Burton et al. 2006) and barley (Burton et al. 2008) to have β-glucan synthase function. To date, β-glucan QTL have not been identified in oat within proximity to the *CslF* gene family based on comparative genomics methods using information from the other grass species. The results presented indicate that the lines in the NPGS could possess valuable alleles for β-glucan concentration not found in elite oat.

Acknowledgments 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".

References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57:289–300

Braaten TJ, Wood PJ, Scott FW, Wolynetz MS, Lowe MK, Bradley-Whyte P (1994) Oat β-glucan reduces blood cholesterol concentration in hypercholesterolemic subjects. Eur J Clin Nutr 48:465–474

Bradbury P, Parker T, Hamblin MT, Jannink J-L (2011) Assessment of power and false discovery rate in genome-wide association studies using the barley CAP germplasm. Crop Sci 51:52–59

Burton RA, Fincher GB (2009) (1,3;1,4)-beta-D-glucans in cell walls of the Poaceae, lower plants, and fungi: a tale of two linkages. Mol Plant 2:873–882

Burton RA, Shirley NJ, King BJ, Harvey AG, Fincher GB (2004) The *CesA* gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. Plant Physiol 134:224–236

Burton RA, Wilson SM, Hrmova M, Harvey AJ, Shirley NJ, Medhurst A, Stone BA, Newbigin EJ, Bacic A, Fincher GB (2006) Cellulose synthase-like *CslF* genes mediate the synthesis of cell wall (1,3;1,4)-β-D-glucans. Science 311:1940–1942

Burton RA, Jobling SA, Harvey AJ, Shirley NJ, Mather DE, Bacic A, Fincher GB (2008) The genetics and transcriptional profiles of the cellulose synthase-like *HvCslF* gene family in barley. Plant Physiol 146:1821–1833

Chernyshova AA, White PJ, Scott MP, Jannink J-L (2007) Selection for nutritional function and agronomic performance in oat. Crop Sci 47:2330–2339

Colleoni-Sirghie M, Fulton B, White PJ (2003) Structural features of water soluble (1–3), (1–4)-β-D-glucan from high-β-glucan and traditional oat lines. Carbohydr Polym 54:237–249

De Koeyer DL, Tinker NA, Wight CP, Deyl J, Burrows VD, O'Donoughue LS, Lybaert A, Molnar SJ, Armstrong KC, Fedak G, Wesenberg DM, Rossnagel BG, McElroy AR (2004) A molecular linkage map with associated QTLs from a hulless × covered spring oat population. Theor Appl Genet 108:1285–1298

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255

Estrada A, Yun CH, Van Kessel A, Li B, Hauta S, Laarveld B (1997) Immunomodulatory activities of oat beta-glucan in vitro and in vivo. Microbiol Immunol 41:991–998

Fincher GB (2009) Exploring the evolution of (1,3;1,4)-β-D-glucans in plant cell walls: comparative genomics can help! Curr Opin Plant Biol 12:140–147

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

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Izydorczyk MS, Macri LJ, Mac Gregor AW (1998) Structure and physiochemical properties of barley non-starch polysaccharides-I. water-extractable β-glucans and arabinoxylans. Carbohydr Polym 35:249–258

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

Jenkins AL, Jenkins DJA, Zdravkovic U, Würsch P, Vuksan V (2002) Depression of the glycemic index by high levels of beta-glucan fiber in two functional foods tested in type 2 diabetes. Eur J Clin Nutr 56:622–628

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723

Kearsey MJ, Farquhar AGL (1998) QTL analysis in plants; where are they now? Heredity 80:137–142

Keenen JM, Pins JJ, Frazel C, Moran A, Turnquist L (2002) Oat ingestion reduces systolic and diastolic blood pressure in patients with mild or borderline hypertension: a pilot study. J Family Pract 51:369

Kennedy BW, Quinton M, van Arendonk JAM (1992) Estimation of effects of single genes on quantitative traits. J Anim Sci 70:2000–2012

Kianian SF, Phillips RL, Rines HW, Fulcher RG, Webster FH, Stuthman DD (2000) Quantitative trait loci influencing β-glucan content in oat (*Avena sativa*, 2n = 6x = 42). Theor Appl Genet 101:1039–1048

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Miyagawa T, Nishida N, Ohashi J, Kimura R, Fugimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotch J, Okaxaki Y, Tsuji S, Tokunaga K (2008) Appropriate data cleaning methods for genome-wide association study. J Hum Genet 53:886–893

Newell MA, Cook D, Tinker NA, Jannink J-L (2011) Population structure and linkage disequilibrium in oat (*Avena sativa* L.): implications for genome-wide association studies. Theor Appl Genet 122:623–632

Newell MA, Kim HJ, Asoro FG, Lauter A, White PJ, Scott MP, Jannink J-L (2012) Micro-enzymatic evaluation of oat (*Avena sativa* L.) beta-glucan for high-throughput phenotyping. Cereal Chem. (in review)

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35:D883–D887

Robertson A (1960) A theory of limits in artificial selection. Proc Royal Soc London, Series B, Biological Sciences 153:234–249

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

SAS Institute (2010) SAS/STAT® 9.2 User's Guide, SAS Campus Drive, Cary, North Carolina 27513

Stich B, Möhring J, Piepho H-P, Heckenberger M, Buckler ES, Melchinger AE (2008) Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754

Sun G, Zhu C, Kramer MH, Yang S-S, Song W, Piepho H-P, Yu J (2010) Variation explained in mixed-model association mapping. Heredity 105:333–340

Tinker NA, Kilian A, Wight CP, Heller-Uszynska K, Wenzl P, Rines HW, Bjørnstad Å, Howarth CJ, Jannink J-L, Anderson JM, Rossnagel BG, Stuthman DD, Sorrells ME, Jackson EW, Tuvesson S, Kolb RL, 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

Wight CP, Tinker NA, Kianian SF, Sorrells ME, O'Donoughur LS, Hoffman DL, Groh S, Scoles GJ, Li CD, Webster FH, Phillips RL, Rines HW, Livingston SM, Armstrong KC, Fedak G, Molnar SJ (2003) A molecular marker map in 'Kanota' × 'Ogle' hexaploid oat (*Avena* spp.) enhanced by additional markers and a robust framework. Genome 46:28–47

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20