Theor Appl Genet (2012) 124:233–246
DOI 10.1007/s00122-011-1697-2

ORIGINAL PAPER

Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars

Minghui Wang · Ning Jiang · Tianye Jia · Lindsey Leach · James Cockram · Robbie Waugh · Luke Ramsay · Bill Thomas · Zewei Luo

Received: 27 April 2011 / Accepted: 29 August 2011 / Published online: 14 September 2011

Abstract Genome-wide association study (GWAS) has become an obvious general approach for studying traits of agricultural importance in higher plants, especially crops. Here, we present a GWAS of 32 morphologic and 10 agronomic traits in a collection of 615 barley cultivars genotyped by genome-wide polymorphisms from a recently developed barley oligonucleotide pool assay. Strong population structure effect related to mixed sampling based on seasonal growth habit and ear row number is present in this barley collection. Comparison of seven statistical approaches in a genome-wide scan for significant associations with or without correction for confounding by population structure, revealed that in reducing false positive rates while maintaining statistical power, a mixed linear model solution outperforms genomic control, structured association, stepwise regression control and principal components adjustment. The present study reports significant associations for sixteen morphologic and nine agronomic traits and demonstrates the power and feasibility of applying GWAS to explore complex traits in highly structured plant samples.

Communicated by J. Yu.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1697-2) contains supplementary material, which is available to authorized users.

M. Wang · N. Jiang · T. Jia · Z. Luo (✉)
School of Biosciences, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
e-mail: z.luo@bham.ac.uk

N. Jiang · R. Waugh · L. Ramsay · B. Thomas
BioSS Unit, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK

L. Leach
Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

J. Cockram
John Bingham Laboratory, National Institute of Agricultural Botany, Cambridge CB3 0LE, UK

Introduction

With the growing availability of genome sequence data and advances in technology for rapid identification and scoring of genetic markers, linkage disequilibrium (LD) based genome-wide association study (GWAS) has gained favour in higher plants, especially crops, for the mapping of genetic factors responsible for complex trait variation (Remington et al. 2001; Gupta et al. 2005; Mackay and Powell 2007; Cockram et al. 2008; Sneller et al. 2009; Atwell et al. 2010). While conventional linkage analysis works on an experimental population derived from a cross of bi-parents divergent for a trait of interest, association mapping applies to collections of samples of a much wider germplasm base. Providing the intrinsic nature of exploiting historical recombination events, association mapping offers increased mapping resolution to polymorphisms at sequence level and should therefore enhance the efficiency of gene discovery and facilitate marker assisted selection (MAS) in plant breeding (Gupta et al. 2005; Moose and Mumm 2008). Plants offer an ease of genetic manipulation allowing production of genetically uniform cultivars through inbreeding, making it possible to conduct replicated assays for many different traits under multiple environmental conditions. Once the plant cultivars are genotyped with high-density markers, association mapping

is promising in resolving the genetic basis of complex traits of both economic and ecological importance.

In the present study, we applied GWAS to analyse a number of continuous and categorical traits in a collection of 615 elite UK barley cultivar samples, some of which were recently reported in an association mapping study (Cockram et al. 2010). Barley (*Hordeum vulgare* L.), the world's fourth most important cereal crop, is an economically important model plant for genetics research (Hayes et al. 2003; Taketa et al. 2008). Due to its narrow genetic base of breeding and also the population bottleneck during the domestication of modern barley cultivars, it was reported that barley exhibited an extensive extent of LD (Kraakman et al. 2004; Rostoks et al. 2006; Malysheva-Otto et al. 2006). Barley cultivars thus potentially provide extant genetic resources that allow successful association mapping using a relatively small density of markers, although the resolution could be limited. Plant species such as barley present exclusive features that could cause striking difference from human GWAS. For example, barley is a diploid and hermaphroditic species in which self-pollination and homozygosity are normal, thus observed heterozygosity is limited in barley cultivars (Rostoks et al. 2006). Moreover, due to the nature of inbreeding and isolation by distance, barley samples may present a much larger scale of population structure and relatedness, introducing the potential for serious confounding in the association study (Balding 2006; Atwell et al. 2010; Hamblin et al. 2010; Platt et al. 2010). In addition, human intervention plays an essential role in modern barley cultivation, hence strong selection on agronomic/economically important traits is expected (Rostoks et al. 2006). Furthermore, the whole genome sequences of many plant species including barley are currently unavailable, hindering any attempt to fine map genetic determinants to sequence level. Given the various complicating factors exhibited in plant samples, it raises concerns about the general applicability of many standard population genetics models well established for human GWAS to plants such as barley.

It is recognized that the barley germplasms are highly partitioned, predominantly due to the number of ear rows (two-row and six-row samples), and the requirement of vernalization (winter- and spring-sown samples) (von Zitzewitz et al. 2005; Yan et al. 2006; Rostoks et al. 2006; Komatsuda et al. 2007; Hamblin et al. 2010). The present barley samples comprise several combinations of these characters. As the molecular bases of ear row number and vernalization requirement in barley have been relatively well characterised (Cockram et al. 2008; Komatsuda et al. 2007), these barley samples present an ideal test bed to evaluate the applicability of various statistical methods established for association mapping in highly structured samples. In this paper, we first investigated the LD structure and its relationship with the barley population division. Methodologically, we tested the performances of one parametric model (STRUCTURE) and one dimensional reduction technique (principal component analysis, PCA) in the inference of population structure of the barley samples. Second, we undertook a genome-wide scan for significant markers associated with a number of traits using six major population structure correction methods (reviewed extensively in Astle and Balding 2009). A comparison of empirical *P* value distribution and a simulation study of statistical power were conducted to evaluate the performances of different methods. Finally, association mapping results from the best structure correction method were reported.

Materials

Germplasm and genotyping

The present study recruits the barley cultivars that have undergone at least 2 years of the UK National List (NL) trials since 1993, together with additional confidential elite lines supplied by major UK barley breeding companies. In total, there are 615 UK barley cultivars (they are also referred as to samples in followings) collected in the present study, among which 490 cultivars were reported in a GWAS of barley morphological traits in Cockram et al. (2010) (ESM, Table S1). Among this collection, 461 samples have records for ear row number, in which 433 are two-row barley and 28 are six-row barley. 472 samples are known for seasonal growth habit, in which 256 are winter barley and 216 are spring barley. The 433 two-row barley samples are almost equally partitioned into winter (220) and spring (209) groups, while 4 are unknown for seasonal growth habit. Of the 28 six-row barley samples, 27 are winter barleys and 1 is unknown for seasonal growth habit.

The barley samples were genotyped at 1,536 single nucleotide polymorphism (SNP) markers by using a recently developed genotyping platform, the Illumina GoldenGate oligonucleotide pool assay 1 (BOPA1). Details were described elsewhere for development of the markers as well as construction of the genetic consensus linkage map based on genotypes of BOPA1 and its sister production assay (BOPA2) (Close et al. 2009). Markers genotyped in less than 95% of the samples, with a minor allele frequency (MAF) below 0.1 or unmapped in the barley consensus genetic map were excluded from further analysis. Heterozygous genotypes were rare (0.8%) in the present data as expected from extensive inbreeding and backcrossing in cultivation and were removed to simplify the subsequent analysis. After quality control checking, 1,042 markers remained for further analysis.

Phenotype

Phenotypic data was recorded on 32 highly heritable morphological characters as well as on 10 agronomic characters including yield and malting traits (ESM, Table S2a, b). The 32 morphological characters, including seasonal growth habit and ear row number, were scored as binary or categorical characters according to the guidelines of International Union for the Protection of New Varieties of Plants (UPOV) protocols (http://www.upov.int/) (Cockram et al. 2010). These morphological phenotype data were loaded at NIAB (http://www.niab.com/) and named Distinctness, Uniformity and Stability (DUS) traits hereafter. For each of the agronomic characters, the phenotype data were collected from a series of performance assessment trials involving at least four site by season combinations, and thus the data so collected represent agronomic performance of the cultivars in a wide range of environments. A mixed linear model was first fitted to account for the time, trial and regional effects together with their interactions and the predicted mean values after removing the fixed effects were taken as the final agronomic trait measure for each barley sample. For convenience, we called the summarized agronomic measurements as BLUP traits which were sourced as the archived database at the Association Genetics of UK Elite Barley (AGOUEB, http://www.agoueb.org). In this primitive BLUP analysis of the agronomic traits, considerable year and site differences as well as their interactions were observed for most of the agronomic traits while the differences between trials were rarely significant and generally negligible. With regard to the difference between the genotypes, heritabilities were generally high with an average of 36% for spring and 45% for winter barley (data not shown). It should be noted that, the present barley collection has 125 extra samples compared to those used in Cockram et al. (2010). Although none of these extra samples has phenotypic records for the DUS or BLUP traits, they had been genotyped with the BOPA1 platform and hence were recruited in the present study for the purpose to investigate the population structure in a larger collection of UK barley germplasm.

Methods

Unless otherwise stated, all analyses were carried out using R, a statistical software package freely available at http://www.r-project.org.

Inference of population structure

Program STRUCTURE v2.2.3 (http://pritch.bsd.uchicago.edu/structure.html) was first applied to estimate the number of historical populations in the present barley samples using default setting of admixture model for the ancestry of individuals and correlated allele frequencies. Population sub-structure was modelled with a burnin of 2.5×10^5 cycles followed by 10^6 Markov Chain Monte Carlo (MCMC) repeats for prior assumed ancestral population number, $K = 1,\ldots, 20$. Since STRUCTURE analysis is computationally intensive, we did not load in all 1,042 markers but instead used 4 subsets of markers. The first marker set contained seven markers, each of which was selected from the middle of a chromosome while the remaining three sets consisted of 305, 96 and 54 markers selected every 2, 10 or 20 cM, respectively, along the barley chromosomes.

Principal component analysis (PCA) was also used as in Price et al. (2006) to infer the population structure in the barley samples. Let $\boldsymbol{g}$ be a matrix of genotypes with element $\boldsymbol{g}_{ij}$ being the genotype of variety i at marker j, where $i = 1$ to N (the number of samples) and $j = 1$ to M (the number of markers). Re-scale matrix $\boldsymbol{g}$ by subtracting the column mean and then dividing by the column standard deviation for each entry in column j, to give a matrix denoted by $\boldsymbol{X}$. An $N \times N$ sample covariance matrix was computed from $\boldsymbol{X}$ and then decomposed into eigenvalues and eigenvectors $\boldsymbol{V}$. In Price et al. (2006), the kth top eigenvector in $\boldsymbol{V}$ was regarded as the kth axis of variation due to ancestry difference.

Association scan

A naive single marker association (SMA) test without correction for confounding was first carried out to search for associations between trait phenotype and marker genotype. SMA refers to linear regression for continuous quantitative traits, Wilcoxon rank-sum test for ordinal categorical traits and chi-square test for other categorical traits. Six statistical methods widely used for controlling population structure were applied to analyze the data. These are structured association (SA) (Pritchard et al. 2000b), genomic control (GC) (Devlin and Roeder 1999), EIGENSTRAT (Price et al. 2006), stepwise regression (SWR) (Setakis et al. 2006), and the mixed linear model (MLM) methods with or without incorporating an inferred population structure matrix as cofactor (Yu et al. 2006).

Except for Wilcoxon rank-sum test and chi-square test, the other methods can be formulated within standard regression models that express the expected value of y_i, the phenotype of the ith sample, as a function of its genotype x_i at the test marker:

$$g(\mathrm{E}[y_i]) = \alpha + x_i\beta \quad (1)$$

where g is a link function, α is model intercept, and β is genetic effect parameter at the marker. Here, x_i is simply coded as 0 and 1 for two different homozygous genotypes

at a test marker. For categorical traits, g is a logit link function, while for quantitative traits, g is an identity link function.

The GC approach was originally developed to correct inflation in a chi-square statistic estimated by Armitage's trend test in structured case-control samples. Bacanu et al. (2002) demonstrated the GC can also be applied to the analysis of quantitative traits by taking $[\hat{\beta}/\mathrm{SE}(\hat{\beta})]^2$ (where $\hat{\beta}$ is the estimate of regression coefficient β and SE is the standard error) as a chi-square test statistic and hence the inflation factor λ can be estimated from a number of 'null' markers using similar methodology to the case–control settings (Devlin and Roeder 1999). As we do not know a priori which markers are 'null', parameter λ was estimated in the present study as the median of genome-wide chi-square scores divided by 0.455, the median of the empirical chi-square distribution under null hypothesis as suggested in Devlin and Roeder (1999). Whenever λ is larger than 1, the chi-square statistic at each test marker is divided by λ.

In the standard SA approach, association test is conditional on the assignment of a population structure inferred by STRUCTURE (the **Q** matrix). In the present study, a reliable inference of **Q** matrix was not achievable (see "Results"); instead, a design matrix indicating the cluster membership derived from k-means clustering of top three PCA axes (termed **P** matrix for simplicity) was incorporated into Eq. 1 with a form of

$$g(\mathrm{E}[y_i]) = \alpha + x_i\beta + P_i v \quad (2)$$

where $\mathbf{P}_i$ is the ith row of the **P** matrix and v is a column vector of regression coefficients. The best cluster model of the k-means clustering was determined through a model selection based on Bayesian information criterion (BIC) which is detailed in "Results".

The EIGENSTRAT method adjusts the genotype and phenotype by using eigenvectors $\boldsymbol{V}$ estimated from the PCA method as described. In detail, let $\boldsymbol{g}_{ij}$ be the genotype of individual i at marker j, the adjusted genotype is $g_{ij,adj} = g_{ij} - \gamma_j a_i$, where a_i is the ancestry of the ith individual along a given axis of variation and $\gamma_j = \sum_i a_i g_{ij} / \sum_i a_i^2$ is a regression coefficient (Price et al. 2006). Adjustments are performed using the top ten axes of variation following Price et al. (2006). Phenotype y_i is adjusted analogously. A linear regression is then carried out to test the association between adjusted phenotype and genotype.

The SWR approach uses a stepwise regression procedure to select for a 'random' set of markers as covariates to control for population structure. By including into the regression model the genotype of covariate markers, the linear equation becomes

$$g(\mathrm{E}[y_i]) = \alpha + x_i\beta + \sum_j C_{ij}\varphi_j \quad (3)$$

where C_{ij} is the genotype of the ith individual at the jth covariate marker and φ_j is the partial regression coefficient. Covariate markers are selected prior to testing a marker of interest by using a forward and backward (stepwise) selection from a random set of 305 markers chosen every 2 cM along the chromosomes. In this study, the stepwise algorithm for selecting covariate markers was implemented by an R function step, using a penalty score $\log(n)$, where n is the number of markers. It should be stressed that stepwise regression is computationally intensive in R and it may even become impractically doable when all markers are incorporated in the model fitting. Moreover, because barley exhibit extensive extent of LD along the chromosomes, it is not necessary to include all markers as adjacent markers tend to provide redundant information regarding the population structure.

The MLM model extends equation (2) by including a random polygenic effect term such that the model is expressed as

$$\mathrm{E}[y_i|\delta_i] = \alpha + x_i\beta + P_i v + \delta_i \quad (4)$$

where δ_i is interpreted as a polygenic contribution to the phenotype (Yu et al. 2006; Astle and Balding 2009) and assumed to have a distribution of $\delta \sim N(0, 2K\sigma_g^2)$, where **K** is a kinship matrix and σ_g^2 is the genetic variance attributable to genome-wide effects. In this study, **K** is estimated as a pairwise identical-by-state (IBS) allele-sharing matrix (Kang et al. 2008). As both population structure and kinship were incorporated, we called this full model the MLM (P + K) model. Meanwhile, we tested a **K** only model, called MLM (K), which omits the population structure **P** from the full model. Both MLM (P + K) and MLM (K), together with the calculation of IBS allele-sharing matrix, have been implemented in R package EMMA (Kang et al. 2008).

Simulation study

A candidate quantitative trait nucleotide (QTN) simulation similar to that of Yu et al. (2006) was conducted. The simulation starts by randomly choosing from the 1,042 markers, a marker with minor allele frequency in the range of 0.1–0.4 as a causal marker. Next, a constant genetic effect explaining a proportion of phenotype variance is added to this causal marker. Later, the phenotype value for an individual is generated by summation of population mean, its genotypic value at the causal marker and a random number. Given the size of genetic effect, a, the percentage of phenotypic variation explained by the causal marker in a pure homozygote type population is calculated

by $p(1-p)a^2/\sigma^2$, where p is the allele frequency and σ^2 is the total phenotype variance (Falconer and Mackay 1996).

For a categorical trait with n categories, we considered a liability model that a set of thresholds θ_j(s) determines the explicit categories from the (underlying) quantitative phenotype values. The ith sample falls in trait category j if its simulated quantitative trait value T_i satisfying the condition of $\theta_{j-1} < T_i \leq \theta_j$ ($j = 1, \ldots, n$), in which $\theta_n = -\theta_0 = \infty$. In the present simulation study, thresholds θ_j(s) are determined in a way that maintains the same categorical distributions (i.e. keeping the same proportion of samples in each category) as that of the observed traits.

Results

Statistical prediction of population structure

The genome-wide LD structure in the 615 barley samples was inferred with software Haploview (see Supplementary Material Fig. S2a–c for the genome-wide r^2 pattern). High LD values were observed across a wide range of the genome with 88.6% significant pairwise marker associations (Bonferroni corrected P value threshold 9.2×10^{-8}) being inter-chromosomal, suggesting strong population structure effect in the barley samples. Furthermore, significant intra-chromosomal LD was evident across the full length of chromosomes (mean = 57.4 cM). When isolating winter samples and spring samples from the whole barley collection, we found that 52.6% of the markers had allele frequencies differing more than twofold between winter and spring subpopulations. Analysis of the LD structure in the two separate seasonal growth type samples revealed that the proportion of significant inter-chromosome marker-pair associations reduced to 18.9% in winter samples and 10.8% in spring samples while the mean span of significant intra-chromosomal LD was 8.5 and 5.2 cM accordingly. This result suggests the presence of significant population structure in the 615 barley samples is primarily due to the divergence in winter and spring growth habit.

The absence of seasonal growth habit information for some barley samples necessitated a more detailed investigation of the population structure in the whole barley collection, for which we employed two statistical approaches making use of marker genotype data. The first approach was using the program STRUCTURE, a parametric method developed by Pritchard et al. (2000a), which uses a Bayesian Markov Chain Monte Carlo (MCMC) approach to infer the number of ancestral subpopulations, K, and then to assign samples probabilistically to each of the K subpopulations. Under the same program parameter setting we carried out STRUCTURE analyses using four sets of markers selected according to the criteria described above. Figure 1a shows the estimated log probabilities of the sample data on subpopulation number K from 1 to 20. As illustrated, none of the four marker sets was able to achieve a convergence for the estimation of parameter K, implying failure to obtain a reliable inference of population structure. It is also observed from Fig. 1a an upward trend in the estimated log probabilities of the data as parameter K increases when using genotypes of 305 markers which were chosen every 2 cM along barley chromosomes. However, this trend is not evident when genotypes of fewer markers were used, suggesting disagreement in STRUCTURE inferences drawn from different marker genotypes.

The second approach used PCA analysis to search for the internal patterns of population structure in the barley samples. Figure 1b and c presents the top three principal components (PCs) decomposed from the covariance matrix. The first three PCs explained, respectively, 20.1, 5.2 and 3.1% of the total variance among the barley collections. The majority of winter samples and spring samples were clearly separated from each other by the first PC, except for five winter samples which were clustered with spring samples. An isolated cluster was also evident consisting of more than 30 samples unknown for both ear row number and seasonal growth habit (Fig. 1b). Of the 27 six-row and also winter samples, 22 tend to aggregate to a separate cluster together with 1 two-row sample and some samples with unknown ear row number (Fig. 1b). The third PC reveals dispersions within either winter or spring samples, e.g. a group of 57 winter barleys are apparently separated from the majority within the winter barley cluster (Fig. 1c). These observations confirm the winter and spring growth habit as the major source that gives rise to the population structure in the barley samples, and additionally suggest the existence of subdivisions in either winter or spring barley samples. The results also manifest the separation between two-row and six-row types. However, clustering with the top three features extracted from PCA by using an x-means cluster algorithm (Pelleg and Moore 2000), which extends the commonly used k-means cluster with efficient estimation of the optimized number of clusters based on Bayesian information criterion (BIC), shows that the best cluster model has a cluster number $k = 2$ (BIC = −160.1). The resulting cluster membership with $k = 2$ was written in a matrix form and taken as the **P** matrix in the following association analyses.

Confounding due to population structure

Since variation of winter and spring growth habit contributes the major source of barley population structure, we were interested in whether the traits were distributed differentially between winter and spring barley samples.

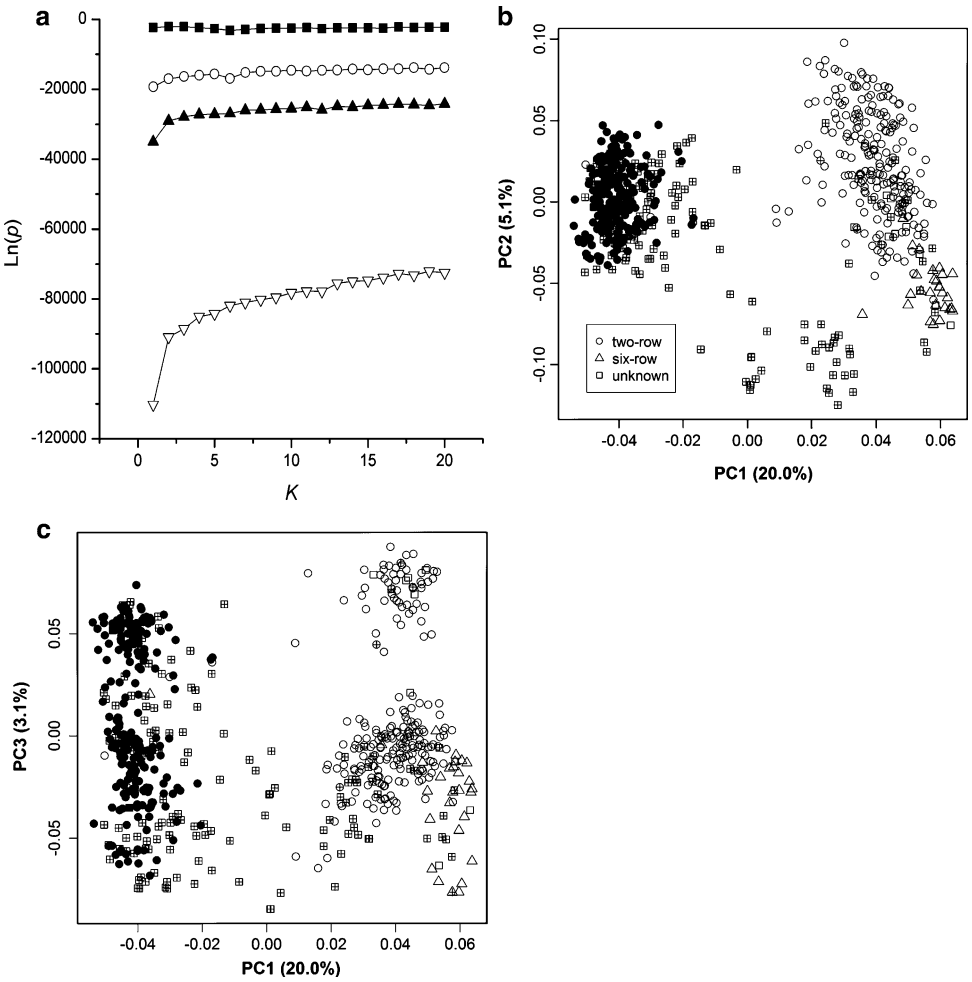

Fig. 1 **a** Estimated log probability, Ln(p), of the data on different population number K from STRUCTURE analyses. Curves from top to bottom are for analyses using genotypes from markers selected by picking one marker on every chromosome, or markers at intervals of 20 cM, 10 cM, and 2 cM, respectively. **b, c** The top three principal components (PCs) in PCA analysis of the variation of the present 615 barley samples. *Blank*, *shadowed* and *crossed* symbols indicate barley samples with winter, spring and unknown seasonal growth habit, respectively

Supplementary Material (ESM, Fig. S2a, b) presents the distributions of BLUP and DUS traits in the two seasonal growth habit samples. All BLUP and a number of DUS traits clearly presented differential distribution between winter and spring barley samples (those significantly associated with seasonal growth habit are indicated in Table S2). This observation raises a concern that any marker that differs in allele frequency between seasonal growth types will show association with these differentially distributed traits. Because more than half of the markers had allele frequencies differing more than twofold between winter and spring barley samples, an association test without accounting for the population structure will result in an increased rate of systematic false positives (Balding 2006).

A single marker association (SMA) test was first carried out to scan for markers in significant association with trait variation in the winter and spring barley samples separately. As shown in Supplementary Material (ESM, Table S3), a greater number of significant markers were predicted in winter barley varieties than in spring barley varieties. By visualizing the P value distribution in a quantile–quantile plot (Wilk and Gnanadesikan 1968), we found that all BLUP and a majority of DUS traits gave rise to a distribution of P values strongly skewed towards significance (GC $\lambda > 1.0$) (ESM, Fig. S4a, b), though we also observed distributions of P values that showed skewness against significance in four DUS traits (numbers 1, 7, 18 and 31). Because population structure may lead to excess of both false positive and false negative associations (Ziv and Burchard 2003), the present observation of skewed P value distributions suggests that the expected confounding indeed exists in winter and spring samples.

By analyzing all 615 barley samples, a large number of genome-wide distributed markers were detected to be significant from the naive SMA approach in all of the BLUP and in about half of the 32 DUS traits (ESM, Table S3). For example, with DUS trait number 28 (seasonal growth habit), there were 793 significant markers (more than 76% of the total markers). Quantile–quantile plot of the P value

distributions indicated a very strong skew towards significance for almost every trait (ESM, Fig. S4a, b). Although we expected some of the significant associations to be true, an excessive number of significant detections obfuscate the genuine association signals from the true causal genes.

Statistical approaches for reducing confounding

Six different structure correction methods were applied in an attempt to separate the genuine associations from background noise in the genome-wide mapping using all 615 barley samples. The number of significant markers surpassing Bonferroni genome-wide *P* value threshold 4.8×10^{-5} from each of these methods was listed in Supplementary Material (ESM, Table S3), from which we observed, as expected, a remarkable drop in the number of significant associations from the six correction methods compared to the naive SMA method. For most of the BLUP and DUS traits, the SWR and SA methods had a comparable number of significant predictions, followed in descending order by MLM (K), MLM (P + K), and lastly GC. EIGENSTRAT presented different conservativeness between DUS and BLUP traits; it is a bit more liberal than MLM (K) in DUS traits but as conservative as GC in BLUP traits.

Following a similar approach used by Potokina et al. (2008) to separate *cis*- from *trans*- expression quantitative trait loci (eQTL), we divided the barley genome into segments of 2 cM length (hereafter referred to as bins), and represented the significance of each segment by the *P* value of the most significant marker within it. In doing so, the barley genome was split into 303 chromosome bins. Since the barley genome displays extensive LD blocks, this effectively reduced the redundancy of significant signals from linked loci within a short chromosome region, especially for the naive SMA method (ESM, Table S3). A comparison of shared significant chromosome bins among seven analytical methods was summarized as mutual predictabilities and given in Table 1. Here, mutual predictability of method *j* to method *i* is calculated as the percentage of significant predictions from method *i* that were recovered by method *j*. For example, in DUS traits, SMA predicted 43.1% of those significant markers detected from SWR, while the mutual predictability of SWR to SMA is 4.3%. In general, a method with a larger number of predictions had higher mutual predictability than a method with fewer predictions, leading SMA (the most liberal method) to have the highest mutual predictability and GC (the most conservative method) to have the lowest mutual predictability. SWR and SA had more predictions and hence higher mutual predictabilities than EIGENSTRAT, MLM (K), MLM (P + K) and GC. However, SWR and SA showed poor mutual predictability with each other (e.g., in DUS traits, SWR recovered a proportion of 21.2% of significant chromosome bins called by SA, while SA recovered 12.6% of significant chromosome bins called by SWR), which diminished the credibility of their added predictions in comparison to methods with fewer predictions. EIGENSTRAT had higher mutual predictabilities in DUS than in BLUP traits, which is consistent with the fact that EIGENSTRAT was more liberal in DUS than in BLUP traits. For the two MLM models, the MLM (K) model had better mutual predictability in BLUP traits while the two were comparable in DUS traits.

Since the six structure correction methods presented poor agreement on significant predictions, we evaluated their control of type I error and statistical power by following a similar process to Yu et al. (2006), in which the type I error was assessed through comparing observed and expected cumulative *P* value distributions, while the power was assessed through a QTN simulation study as described above. Figure 2a, b shows the cumulative distributions of observed *P* values in genome scans from six different methods in two barley DUS traits, seasonal growth habit and ear row number. Under the null hypothesis that random markers are not in LD with the genetic loci controlling the trait, approaches that have appropriate control of type I errors are expected to show a uniform distribution of *P* values (a diagonal line in these plots). Generally, two mixed model analyses, MLM (P + K) and MLM (K), showed good approximation to the expected *P* value distributions in the two traits. EIGENSTRAT gave liberal results in both traits and GC and SA gave conservative results for seasonal growth habit, whereas results from SWR were liberal for seasonal growth habit but slightly conservative for ear row number. The statistical power simulation was conducted by adding a genetic effect to each of 100 randomly selected markers, one at a time, and then testing, in each simulation, whether the QTN marker could be detected by different models under empirical Bonferroni *P* value threshold (4.8×10^{-5}). The proportion of QTN detected across all random markers was used as the measurement of the control of type II error for each model. Figure 2c, d present the results of statistical power simulation in barley seasonal growth habit and ear row number traits. The average statistical power was consistently higher for MLM (K) than for MLM (P + K), SA, EIGENSTRAT and GC. For seasonal growth habit, the SWR had a slightly higher power than the MLM (K), but for ear row number, the opposite was true. It is noted that all six-structure correction methods had higher powers in barley seasonal growth habit than in barley ear row number at any given variance proportion, which could be attributed to the fact that seasonal growth habit had equal numbers of samples in two different trait categories. Further results of

Table 1 Comparison of shared significant predictions among seven analytical methods

Methods	#	Mutual predictability (%)						
		SMA	SWR	SA	EIGENSTRAT	MLM (P + K)	MLM (K)	GC
BLUP								
SMA	2,347	100.0	13.8	6.4	0.3	0.8	1.4	0.8
SWR	327	99.1	100.0	21.7	1.5	4.9	9.5	5.5
SA	171	87.7	41.5	100.0	2.3	10.5	17.5	8.2
EIGENSTRAT	6	100.0	83.3	66.7	100.0	50.0	50.0	0.0
MLM (P + K)	19	94.7	84.2	94.7	15.8	100.0	89.5	36.8
MLM (K)	34	97.1	91.2	88.2	8.8	50.0	100.0	38.2
GC	18	100.0	100.0	77.8	0.0	38.9	72.2	100.0
DUS								
SMA	2,522	100.0	4.3	4.4	4.8	1.7	2.0	0.3
SWR	253	43.1	100.0	12.6	11.9	8.7	9.1	1.2
SA	151	74.2	21.2	100.0	46.4	20.5	22.5	2.6
EIGENSTRAT	180	67.8	16.7	38.9	100.0	18.9	20.0	2.8
MLM (P + K)	43	97.7	51.2	72.1	79.1	100.0	97.7	14.0
MLM (K)	51	100.0	45.1	66.7	70.6	82.4	100.0	11.8
GC	7	100.0	42.9	57.1	71.4	85.7	85.7	100.0

#, the total number of significant chromosome bins of 2 cM length pooled from all of the BLUP or DUS traits. Mutual predictability is calculated as the percentage (%) of significant predictions from the method in row i (i = 1st,…,7th row) that were also predicted from the method in column j (j = 1st,…,7th column). For example, in DUS traits, SMA predicted 43.1% of those significant markers detected from SWR, while conversely, the predictability of SWR to SMA is 4.3%

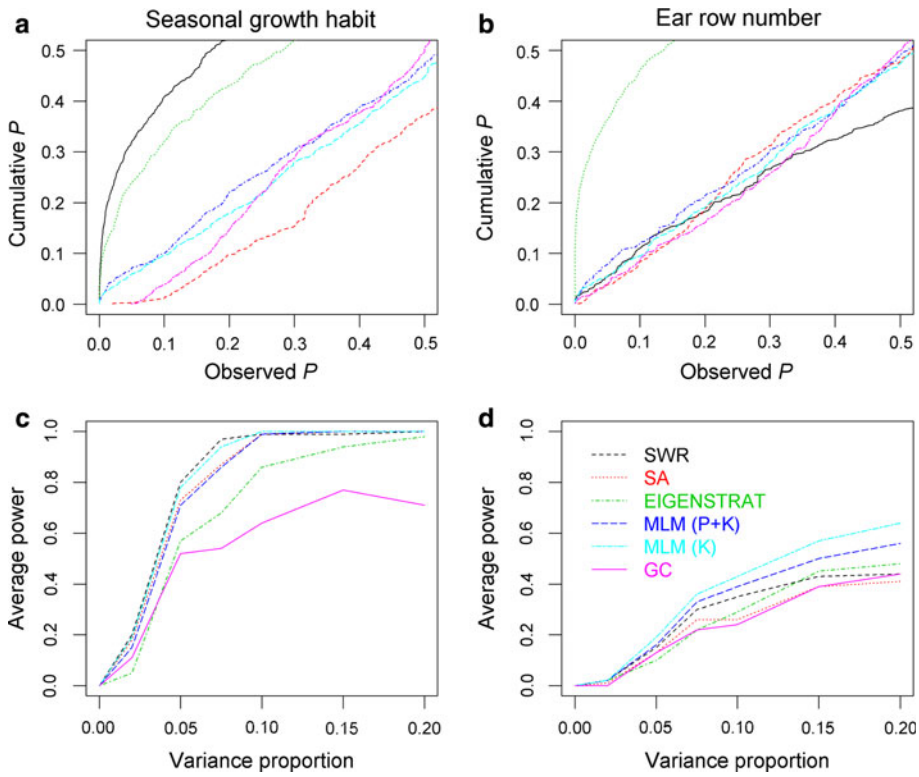

Fig. 2 Model comparison with barley seasonal growth habit and ear row number traits. **a, b** Cumulative distribution of the P values in the genome-wide association scan in 615 barley samples. **c**, **d** Average statistical power to detect a QTN based on Bonferroni correction threshold ($0.05/100 = 5 \times 10^{-4}$). The power is averaged across all 100 simulated QTN for each given variance proportion

cumulative P value distributions as well as statistical power for all 10 BLUP traits were given in Supplementary Material (ESM, Fig S4). Without exception, the MLM (K) method outperformed its rivals in terms of both controlling false positives and maintaining statistical power for all BLUP traits.

Mapping results

As the MLM (K) method consistently achieved good control of false positives while yielding the highest power among six structure correction methods, we concentrated on the significant predictions from this method in the following text. The MLM (K) method detected significant associations (exceeding Bonferroni genome-wide P value threshold 4.8×10^{-5}) for 9 out of 10 BLUP and 16 out of 32 DUS traits (Fig. 3a–d). The complete list of significant marker-trait associations, as well as peak markers, marker names, R^2_{LR} (a likelihood-ratio based R^2-like statistic), and putative rice (*Oryza sativa*) homologue loci were provided in Table 2 for BLUP traits and Supplementary Material (ESM, Table S4) for DUS traits. In the present paper, marker names were designated in the form of Xn-nn.nn (where n is a digit), in which the digit after letter X indicates chromosome number while the digits after the hyphen indicate map position. A number m of additional markers at the same chromosomal position are appended "$.i$", where $i = 1,\ldots, m$ (e.g. X2-96.82.1 is the second marker at 96.82 cM on chromosome 2). The R^2_{LR} statistic was estimated following a likelihood-ratio based formula suggested by Sun et al. (2010), $R^2_{LR} = 1 - \exp[2(L_0 - L_M)/n]$, where L_M and L_0 are respectively the maximum log-likelihood of the mixed models with or without incorporating a SNP of interest, and n is the number of individuals. While the R^2_{LR} statistic, like the coefficient of determination (R^2) in linear fixed effect model, serves as a measurement of how well different model agrees with the data, it also provides an intuitive indication of the genetic effect of the SNP of interest; the changes in R^2_{LR} values resulted from fitting with different SNPs suggest the relative importance of these SNPs in explaining the phenotypic variation. We also experimented another R^2-like statistic used in some literatures (e.g. Atwell et al 2010), $R^2_{\hat{\beta}} = \hat{\beta}^2 \sum_i (x_i - \bar{x})^2 / \sum_i (y_i - \bar{y})^2$, where $\hat{\beta}$ is the estimate of SNP fixed effect derived from mixed model analysis. Comparison showed that these two R^2-like statistics produced very similar estimates with a median correlation coefficient of 0.904 although in majority of the cases the $R^2_{\hat{\beta}}$ statistic tended to provide higher estimates than the likelihood ratio based statistic (data not shown). The putative rice homologue loci were derived through BLASTX sequence alignments with SNP marker source sequences against the version 6 rice genome sequence as in Close et al. (2009).

For DUS traits, we replicated most of the previously reported association signals, with the exception of significance on barley chromosome 3 for traits seasonal growth habit and ear row number reported in Cockram et al. (2010). This discrepancy was probably due to use of different models to account for population structure in the mixed model analyses (K only model here, Q+K model in Cockram et al. 2010 where Q is a fractional subpopulation membership matrix estimated from STRUCTURE). Among the association signals detected for barley seasonal growth habit, we noted that significant marker X1-96.92.1 (P value $< 1.0 \times 10^{-220}$) had a R^2_{LR} value of 40.9%. An analysis of colinearity with rice genome indicated that this marker mapped to the barley short-day photoperiod locus *PPD-H2*, one major QTL for seasonal growth habit (Cockram et al. 2010). Detailed examination of allele frequencies revealed that different alleles at this marker were almost fixed within winter (0.988) and spring (0.005) samples; similarly, the significant marker X4-26.19.4 was almost fixed for different alleles between the two-row (0.998) and six-row (0.071) barley samples, with R^2_{LR} value of 45.9%.

Several BLUP traits showed significant associations at markers also detected for the two barley diversity characteristics, ear row number and seasonal growth habit (Fig. 3b–d; Table 2). For example, marker X4-26.19.4, which was highly significantly associated with variation of ear row number, was detected for two grain weight traits TKW and SpecW. In addition, several markers at or near 96.92 cM on chromosome 1 were highly significantly associated with variation of seasonal growth habit, and also detected for four BLUP traits (GrainN, MaltN, SolN and GE4ml). Common association predictions among these traits may be coincidental or represent pleiotropic effects of underlying genetic factors. However, traits MaltN and GE4ml presented high levels of phenotypic correlations with seasonal growth habit (Pearson correlation coefficients 0.97–0.98), likely explaining the common predictions.

Some BLUP traits were closely connected to each other and hence phenotypic similarities and common predictions were observed as expected. For example, two-grain weight related traits, TKW and SpecW, had a Pearson correlation coefficient of 0.427 and shared a common signal at marker X4-26.19.4 (Fig. 3d; Table 2). GrainN, MaltN and SolN measured the percentage of nitrogen in dry grain, dry malted grain and malt extract, respectively. While GrainN and MaltN showed a high degree of phenotypic similarities (Pearson correlation coefficient is 0.888), SolN showed a strong negative correlation with GrainN and MaltN (Pearson correlation coefficients is −0.647 and −0.824, respectively). Association analyses revealed a common association at 93.95–96.92 cM on chromosome 1 shared among the three traits, which was detected for barley seasonal growth habit. There are also several unique association signals for each trait, such as marker X5-2.81 for GrainN, X5-63.31 for MaltN and X1-54.73.1 for SolN (Fig. 3d; Table 2).

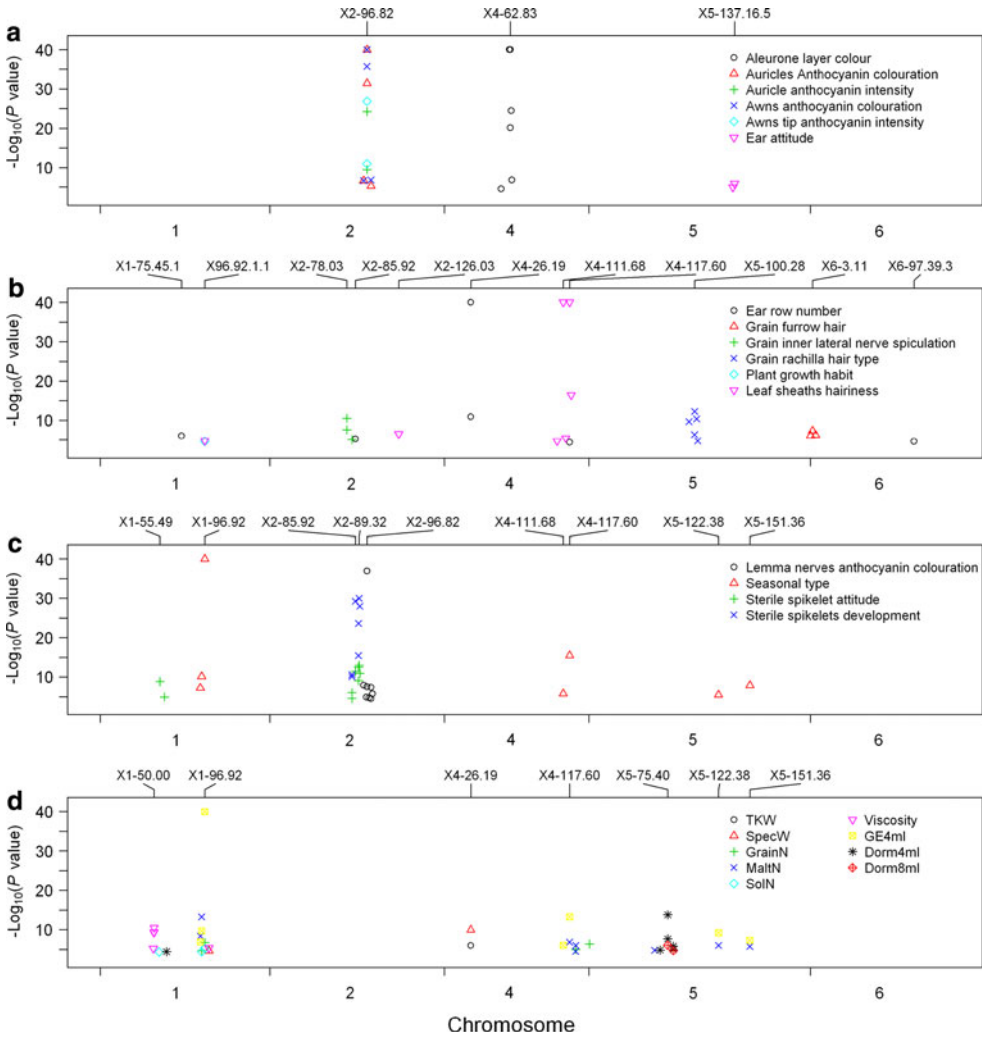

Fig. 3 Significant associations detected from the MLM (K) method for (**a**–**c**) DUS and **d** BLUP traits under Bonferroni genome-wide *P* value threshold 4.8×10^{-5}. The most significant SNPs within 2 cM chromosome bins are labelled above each figure. SNP name is designated in a form of Xn-nn.nn, in which the digit n after letter X indicates chromosome while the digits after the hyphen indicate map position (see main text for detail). Any $-\text{Log}_{10}(P \text{ value})$ above 40 is truncated to 40. Chromosomes 3 and 7 with no significant signals are omitted

Three BLUP traits, GE4ml, Dorm4ml and Dorm8ml, measured grain germinations (seedling development) one month after watering in 4 ml water (GE4ml), or 72 h after watering in 4 ml (Dorm4ml) and 8 ml (Dorm8ml) of water, respectively. GE4ml had negative correlations with the latter two germination traits (Pearson correlation coefficients −0.562 and −0.483, respectively). As stated above, GE4ml presented a very strong correlation with seasonal growth habit and hence had identical association markers. In contrast, the latter two germination traits presented positive correlation (Pearson correlation coefficient 0.667) and their association markers were different from those of GE4ml. Dorm4ml was associated with three markers at 61.53 cM on chromosome 1 and with a region at 68.35–80.61 cM on chromosome 5, while Dorm8ml gave associations at 75.4–80.61 cM on chromosome 5 (Fig. 3d; Table 2). We also performed a joint analysis of treating Dorm4ml and Dorm8ml as a single trait but incorporating water treatment as a fixed effect in the MLM (K) model. In the joint analysis, only the region at 68.35–80.61 cM on chromosome 5 was predicted significant, while markers at 61.53 cM on chromosome 1 did not exceed genome-wide threshold (*P* value 3.3×10^{-4}). One marker at 61.53 cM on chromosome 1 is mapped to barley gene UniRef90_Q02400, which encodes a late embryogenesis abundant (LEA) protein associated with desiccation tolerance of seeds (Goyal et al. 2005). For the region on chromosome 5, BLASTX sequence alignments revealed that a marker at 69.90 cM (BOPA2 marker 12_30080) mapped to Rice homologue *LOC_Os09g26620* (*E* score 2.0×10^{-51}), a putative auxin-repressed gene. This marker locus is also mapped to *Arabidopsis* homologue gene *AT1G56220.1* (*E* score 2.0×10^{-16}), which belongs to a dormancy/auxin associated gene family. The fact that the growth (germination) of seeds in dormant state commences with the uptake of water under suitable condition (Bewley 1997) suggests we have successfully identified two barley germination controlling loci.

Viscosity of wort is a cytolytic character mainly related to the breakdown of β-glucan, the main structure material

Table 2 Significant marker-trait associations exceeding Bonferroni threshold (P value 4.8×10^{-5}) in the genome scan of BLUP traits

No.	Trait	Chromosome (interval, cM)	Peak marker name		P value	R^2_{LR}(%)	Putative rice homologue
			Present[a]	BOPA[b]			
2	TKW	4H (26.19)[c]	X4-26.19.4	11_20606	1.1×10^{-6}	5.6	LOC_Os03g50040.1
3	SpecW	1H (100.69)	X1-100.69	11_10357	2.4×10^{-5}	3.6	LOC_Os07g10256.1
		4H (26.19)[c]	X4-26.19.4	11_20606	1.0×10^{-10}	7.8	LOC_Os03g50040.1
4	GrainN	1H (93.95-96.92)[d]	X1-96.92.1	11_10396	1.6×10^{-7}	5.9	LOC_Os05g44760.1
		4H (123.29)	X4-123.29	11_20013	1.1×10^{-5}	4.5	LOC_Os10g25060.1
		5H (2.81)	X5-2.81	11_20553	4.1×10^{-7}	6.6	LOC_Os12g44310.2
5	MaltN	1H (92.8-96.92)[d]	X1-96.92.1	11_10396	7.3×10^{-47}	35.9	LOC_Os05g44760.1
		4H (117.6-123.29)[c,d]	X4-117.60.2	11_21210	1.4×10^{-7}	7.8	LOC_Os03g01750.5
		5H (63.31)	X5-63.31	11_11281	1.6×10^{-5}	5.5	LOC_Os09g23350.1
		5H (122.38)[d]	X5-122.38	11_10094	9.3×10^{-7}	6.4	LOC_Os09g38030.1
		5H (151.36)[d]	X5-151.36.3	11_20100	1.7×10^{-6}	6.5	LOC_Os03g57220.2
6	SolN	1H (54.73)	X1-54.73.1	11_21217	4.4×10^{-5}	5.0	LOC_Os10g42780.1
		1H (93.95-96.92)[d]	X1-96.92.1	11_10396	5.2×10^{-6}	5.8	LOC_Os05g44760.1
7	Viscosity	1H (49.34-50.0)	X1-50.00.1	11_10438	3.0×10^{-11}	11.7	LOC_Os07g10420.1
		1H (100.69)	X1-100.69	11_10357	4.2×10^{-6}	6.3	LOC_Os07g10256.1
8	GE4 ml	1H (92.8-96.92)[d]	X1-96.92.1	11_10396	1.7×10^{-83}	40.8	LOC_Os05g44760.1
		4H (111.68)[d]	X4-111.68	11_11299	8.7×10^{-7}	6.4	LOC_Os03g02750.1
		4H (117.6)[c,d]	X4-117.60.2	11_21210	5.1×10^{-14}	15.1	LOC_Os03g01750.5
		5H (122.38)[d]	X5-122.38	11_10094	5.6×10^{-10}	10.5	LOC_Os09g38030.1
		5H (151.36)[d]	X5-151.36.3	11_20100	5.2×10^{-8}	8.5	LOC_Os03g57220.2
9	Dorm4 ml	1H (61.53)	X1-61.53	11_11049	3.6×10^{-5}	5.3	LOC_Os05g28210.1
		5H (68.35-80.61)	X5-75.40	11_21001	1.6×10^{-14}	16.0	LOC_Os09g24980.1
10	Dorm8 ml	5H (75.4-80.61)	X5-75.40	11_21001	8.5×10^{-7}	7.2	LOC_Os09g24980.1

R^2_{LR}, a likelihood ratio based R^2-like statistics. Putative rice homologue for each peak marker is derived by sequence alignment as indicated in Close et al. (2009). Trait number is the same as in Supplementary Table S2a

[a] Marker name is designated as Xn-nn.nn, in which the first digit n after letter X indicates chromosome and digits after the hyphen indicate map position. A number m of additional markers at the same chromosomal position are appended "$.i$", where $i = 1,\ldots,m$

[b] Original marker name designated in Close et al (2009)

[c] Significant association also detected for barley ear row number

[d] Significant association also detected for barley seasonal growth habit

in barley endosperm cell walls (Schmalenbach and Pillen 2009). Here, we identified two regions on chromosome 1 (49.34–50.0 and 100.69 cM) significantly associated with viscosity and consistent with strong quantitative trait locus (QTL) effects on chromosome 1 at 39–70 cM detected by Schmalenbach and Pillen (2009).

Discussion

This paper presents a GWAS of agronomic and morphologic traits in a collection of 615 barley cultivars. Compared to our previous association analysis (Cockram et al. 2010), the present study recruited additional samples, reported associations with 10 additional agronomic traits, and conducted a comprehensive survey of association test methods. The present analysis revealed that more than half of the DUS traits and all BLUP traits had shown divergent phenotypic distributions between winter and spring samples (ESM, Fig. S2). We also found shared common associations between a number of agronomic traits and two barley diversity traits, seasonal growth habit and ear row number (Table 2). All these suggest that barley diversification had a profound impact on variations in barley morphological and agronomic traits, of which the latter are undoubtedly key issues for cultivated barley. In other words, a number of traits had undergone strong selection along with diversification of seasonal growth habit in barley cultivars, complicating the association mapping in the present structured barley samples. Although we did not explicitly model the selection effect in the present study, effective correction for population structure should have simultaneously accounted for the confounding from selection which accompanied the diversification of barley samples.

Significant associations were detected for 9 out of 10 BLUP and 16 out of 32 DUS traits. Notably, we found novel associations at two chromosome regions for barley seed germination, 61.53 cM on chromosome 1 and 68.35–80.61 cM on chromosome 5. Although the full barley genome sequence is currently unavailable, information from gene annotation and homologous sequence alignment for marker sequences in the two regions suggests two possible candidate genes for barley germination, a late embryogenesis abundant (LEA) protein gene and a putative gene homologous to Rice and *Arabidopsis* dormancy/auxin associated gene. While further genetic study is required to confirm the discovery, the present finding highlights the feasibility of high resolution mapping with GWAS in barley samples. As the markers in the present barley genotyping platform (BOPA) were entirely developed from transcribed gene SNPs (Close et al. 2009), efficient interspecies homology sequence comparison, particularly through gene synteny, between barley and other grass genomes including Rice and *Arabidopsis thaliana* (Dubcovsky et al. 2001; Bennetzen and Ma 2003), provides a powerful tool for identifying and refining QTL in the un-sequenced barley genome.

Factors including geographic localization, breeding patterns and even human intervention (e.g., selective breeding based on economic/agronomically significant traits during crop domestication) may lead to complications such as strong population structure and familial relatedness within plant samples assembled in association mapping studies (Atwell et al. 2010). In the present barley samples, strong population structure is demonstrated primarily due to division by seasonal growth habit and by ear row number, both of which have close connection with breeding activities of cultivated barley. In GWAS of human complex diseases, population structure has been considered an important cause of spurious associations and an explanation of failure to replicate significant predictions, making statistical methods accounting for population structure essential to validate standard association tests (Balding 2006). This paper presents an evaluation of six association test methods which are popular in correcting for population structure, including genomic control (GC) (Devlin and Roeder 1999), stepwise regression (SWR) (Setakis et al. 2006), structured association (SA) (Pritchard et al. 2000a, b), EIGENSTRAT (Price et al. 2006), and two mixed linear model (MLM) analyses (Yu et al. 2006). While a rich literature has reported comparisons of statistical methods for population structure correction, most studies concentrated on case and control sampling designs (Astle and Balding 2009; Price et al. 2010; Wu et al. 2010). Zhao et al. (2007) compared some of the methods considered here in a sample of *Arabidopsis thaliana* inbred lines, though their study is severely under-powered due to limited sample size (95 accessions). The present study exploits a much larger sample of barley cultivars collected from a wide range of germplasm resources and hence delivers more reliable statistical inference. In addition, this paper explored a larger number of complex traits of different types, including both continuous (BLUP) and categorical (DUS) traits, to achieve a more comprehensive comparison of available methodologies.

Before correcting for population structure, it is critical to detect and infer the hidden structure in a sample. Perhaps the most commonly accepted statistical method to detect population structure is a model-based cluster approach, STRUCTURE (Pritchard et al. 2000a, b), which uses multilocus genotype data to infer the subpopulation number K and creates a subpopulation membership matrix $\mathbf{Q}$ to represent the samples. However, with the present structured barley collection, the STRUCTURE method failed to infer a reliable parameter K despite using different panels of genotypes with varying number of markers selected (Fig. 1a). This failure is expected because STRUCTURE attempts to account for population structure by allocating population groupings in such a way that Hardy–Weinberg Equilibrium (HWE) is met within subpopulations, whereas the assumption of HWE is actually invalid given the nature of extensive inbreeding in barley samples. Indeed, heterozygous genotypes were rare (0.8%) and thus removed in the present analysis, rendering this method entirely impractical.

Dimension reduction techniques such as PCA do not require a model assumption and hence can robustly predict the hidden structure by extracting principal components from a covariance matrix. Although PCA is computationally fast and visually appealing in representing broad differences across samples in a dataset, it could be difficult in practice to make biological interpretations from principal components (PCs) as population structure surrogates, and hence further statistical assessment like clustering with extracted PCs would be essential. We compared the cluster result from PCA with that from another multidimensional scaling (MDS) method, the principal coordinate (PCO) analysis of pairwise IBS kinship estimates, which has been widely utilized in GWAS to predict the hidden population structure (Purcell et al. 2007; Simon-Sanchez et al. 2009). K-means clustering using the top three axes extracted from PCA and PCO methods showed similar partition patterns when setting cluster number $k = 2$ but different partition patterns when setting $k = 3$ (ESM, Fig. S5). A Bayesian model selection with x-means cluster algorithm (Pelleg and Moore 2000) indicated that the best cluster model from PCO axes had cluster number $k = 3$ (BIC $= -70.5$), different from the cluster result from PCA axes, from which the best cluster model had cluster number $k = 2$ (BIC $= -160.1$). As the latter had a smaller BIC value, we

preferred the PCA based cluster membership to create the **P** matrix in the present association analysis.

Association analyses with the six above mentioned structure correction methods predicted highly varying numbers of significant markers in the strongly structured barley samples, suggesting caution should be taken to interpret the associations predicted from different methods. Through evaluating empirical *P* value distributions and a power simulation study, we demonstrated that two MLM approaches, especially the MLM (K) method, outperformed their rivals for controlling the rate of false positives while maintaining statistical power. The poor performance of other methods probably results from failure to correct the confounding caused by relatedness (genetic co-ancestry) presented in the barley samples, which was captured through a kinship matrix in the mixed model analyses. For example, a model selection according to BIC indicated the best cluster with top principal components to have cluster number $k = 2$, which was too simplistic given the number of combinations of seasonal growth habit and ear number characteristics. Note that the MLM (P + K) method had similar performance in controlling false positives but showed a decreased statistical power when compared to the MLM (K) method, which is probably because using the **K** matrix alone is sufficient in capturing the complicating factors in the present barley data, while a combination of **P** and **K** matrices might lead to over-correction.

It should be stressed that DUS traits in the present study were either binary or categorical characters and hence a generalized linear mixed model (GLMM) is more appropriate than the MLM based analysis. However, GLMM method with correlated variance components is computationally inefficient with the currently available algorithms and it has been suggested that a standard linear regression framework is useful in binary phenotype analysis such as with a case and control design (Kang et al. 2010). MLM has been utilized in GWAS of categorical traits, for example, Atwell et al. (2010) used MLM to correct for population structure in a GWAS of both continuous and binary or categorical traits in *Arabidopsis thaliana* inbred lines. An intensive comparison of various methods in the present study demonstrates that the MLM method is the most promising for analyzing either continuous and or categorical traits for GWAS of plant populations with extensive structure.

Acknowledgments This study was supported by research grants for the 'Association Genetics of UK Elite Barley' project, which was funded by BBSRC and RERAD as part of the Sustainable Arable LINK programme with industrial support from HGCA, KWS (UK), LS Plant Breeding, Syngenta Seeds, Groupe Limagrain, Secobra UK, Svalof Weibull, Perten Instruments AB, The Maltsters Association of Great Britain, The Scotch Whisky Research Institute and Campden BRi. ZWL is also supported by the Leverhulme Trust (RCEJ1471) of UK, NSFC (31071084) and The Basic Research Program (2012CB316505) of China.

References

Alonso-Blanco C, Aarts MGM, Bentsink L, Keurentjes JJB, Reymond M, Vreugdenhil D, Koornneef M (2009) What has natural variation taught us about plant development, physiology, and adaptation? Plant Cell 21:1877–1896

Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Stat Sci 24:451–471

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 traits in *Arabidopsis thaliana* inbred lines. Nature 465:627–631

Aulchenko YS, de Koning D-J, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177:577–585

Bacanu SA, Devlin B, Roeder K (2002) Association studies for quantitative traits in structured populations. Genet Epidemiol 22:78–93

Balding D (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781–791

Bennetzen JL, Ma J (2003) The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. Curr Opin Plant Biol 6:128–133

Bewley JD (1997) Seed germination and dormancy. Plant Cell 9:1055–1066

Close T, Bhat P, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson J, Wanamaker S, Bozdag S, Roose M, Moscou M, Chao S, Varshney R, Szucs P, Sato K, Hayes P, Matthews D, Kleinhofs A, Muehlbauer G, DeYoung J, Marshall D, Madishetty K, Fenton R, Condamine P, Graner A, Waugh R (2009) Development and implementation of high-throughput SNP genotyping in barley. BMC Genom 10:582

Cockram J, White J, Leigh FJ, Lea VJ, Chiapparino E, Laurie DA, Mackay IJ, Powell W, O'Sullivan DM (2008) Association mapping of partitioning loci in barley. BMC Genet 9:16

Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, Luo Z, Kearsey MJ, Werner P, Harrap D, Tapsell C, Liu H, Hedley PE, Stein N, Schulte D, Steuernagel B, Marshall DF, Thomas WTB, Ramsay L, Mackay I, Balding DJ, Consortium TA, Waugh R, O'Sullivan DM (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proc Natl Acad Sci USA (published ahead of print)

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan L, Shiloff BA, Bennetzen JL (2001) Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. Plant Physiol 125:1342–1353

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edition. Longman Press, New York

Goyal K, Walton LJ, Tunnacliffe A (2005) LEA proteins prevent protein aggregation due to water stress. Biochem J 388:151–157

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461–485

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

Hayes PM, Castro A, Marquez-Cedillo L, Corey A, Henson C, Jones BL, Kling J, Mather D, Matus I, Rossi C, Sato K (2003) Genetic diversity for quantitatively inherited agronomic and malting quality traits. Diversity in Barley (*Hordeum vulgare*). Elsevier, Amsterdam

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492–1504

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

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, Sabatti C, Eskin E (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–354

Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, Stein N, Graner A, Wicker T, Tagiri A, Lundqvist U, Fujimura T, Matsuoka M, Matsumoto T, Yano M (2007) Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. Proc Natl Acad Sci USA 104:1424–1429

Kraakman ATW, Niks RE, Van den Berg PMMM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435–446

Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57–63

Malysheva-Otto LV, Ganal MW, Roder 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

Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiol 147:969–977

Pelleg D, Moore A (2000) X-means: extending k-means with efficient estimation of the number of clusters. In: Proceedings of the seventeenth international conference on machine learning. Morgan Kaufmann, San Francisco, pp 727–734

Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, Ågren J, Bossdorf O, Byers D, Donohue K, Dunning M, Holub EB, Hudson A, Le Corre V, Loudet O, Roux F, Warthmann N, Weigel D, Rivero L, Scholl R, Nordborg M, Bergelson J, Borevitz JO (2010) The scale of population structure in *Arabidopsis thaliana*. PLoS Genet 6:e1000843

Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M (2008) Gene expression quantitative trait locus analysis of 16000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J 53:90–101

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67:170–181

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

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

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Grainer A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc Natl Acad Sci USA 103:18656–18661

Schmalenbach I, Pillen K (2009) Detection and verification of malting quality QTLs using wild barley introgression lines. Theor Appl Genet 118:1411–1427

Setakis E, Stirnadel H, Balding DJ (2006) Logistic regression protects against population structure in genetic association studies. Genome Res 16:290–296

Simon-Sanchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, Berg D, Paisan-Ruiz C, Lichtner P, Scholz SW, Hernandez DG, Kruger R, Federoff M, Klein C, Goate A, Perlmutter J, Bonin M, Nalls MA, Illig T, Gieger C, Houlden H, Steffens M, Okun MS, Racette BA, Cookson MR, Foote KD, Fernandez HH, Traynor BJ, Schreiber S, Arepalli S, Zonozi R, Gwinn K, van der Brug M, Lopez G, Chanock SJ, Schatzkin A, Park Y, Hollenbeck A, Gao JJ, Huang XM, Wood NW, Lorenz D, Deuschl G, Chen HL, Riess O, Hardy JA, Singleton AB, Gasser T (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat Genet 41:1308–1309

Sneller CH, Mather DE, Crepieux S (2009) Analytical approaches and population types for finding and utilizing QTL in complex plant populations. Crop Sci 49:363–380

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

Szűcs P, Skinner J, Karsai I, Cuesta-Marcos A, Haggard K, Corey A, Chen T, Hayes P (2007) Validation of the VRN-H2/VRN-H1 epistatic model in barley reveals that intron length variation in VRN-H1 may account for a continuum of vernalization sensitivity. Mol Genet Genom 277:249–261

Taketa S, Amano S, Tsujino Y, Sato T, Saisho D, Kakeda K, Nomura M, Suzuki T, Matsumoto T, Sato K, Kanamori H, Kawasaki S, Takeda K (2008) Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. Proc Natl Acad Sci USA 105:4062–4067

von Zitzewitz J, Szucs P, Dubcovsky J, Yan L, Francia E, Pecchioni N, Casas A, Chen TH, Hayes PM, Skinner JS (2005) Molecular and structural characterization of barley vernalization genes. Plant Mol Biol 59:449–467

Wilk MB, Gnanadesikan R (1968) Probability plotting methods for the analysis of data. Biometrika 55:1–17

Wu C, DeWan A, Hoh J, Wang Z (2010) A comparison of association methods correcting for population stratification in case–control studies. Ann Hum Genet 00:1–10

Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede M, Sanchez A, Valarik M, Yasuda S, Dubcovsky J (2006) The wheat and barley vernalization gene VRN3 is an orthologue of FT. Proc Natl Acad Sci USA 103:19581–19586

Yu JM, 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

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4

Ziv E, Burchard EG (2003) Human population structure and genetic association studies. Pharmacogenomics 4:431–441