Assessment of Power and False Discovery Rate in Genome-Wide Association Studies using the BarleyCAP Germplasm

Peter Bradbury, Thomas Parker, Martha T. Hamblin, and Jean-Luc Jannink*

ABSTRACT
Success in genome-wide association studies (GWAS) is dependent on the power to detect quantitative trait loci (QTL) with a minimal rate of false discovery. The objective of this study was to determine the potential for GWAS within barley (Hordeum vulgare L.) by evaluating several linear models that varied in the way they accounted for population structure (model-based STRUCTURE or principle component analysis [PCA]) and familial relatedness. Using genotype data from the Barley Coordinated Agricultural Project (BarleyCAP), phenotypic effects were simulated for different numbers of QTL with three heritability levels. Under each scenario, power and false discovery rate were calculated for sample sizes of 100 or 300 individuals. A mixed model that accounted for familial relatedness but not population structure performed as well as or better than all other models across all heritability levels, QTL numbers, and sample sizes tested. Simulations with 100 lines performed poorly for QTL detection but simulations with 300 lines performed adequately, suggesting that the BarleyCAP data can be used successfully for GWAS if sample sizes are adequate.

Genome-wide association studies (GWAS), that is, the identification of statistical associations between molecular markers and quantitative traits, offer a promising approach for revealing the genetic basis of phenotypic variation. In this method, association mapping (AM) is performed on populations of individuals that lack a well-defined family structure and share common ancestors in the relatively distant past. Because linkage disequilibrium (LD; association of alleles) decays over generations, statistical associations in such populations should only be found for those markers closely linked to the trait of interest. Using coalescent simulations to generate both genotypic and phenotypic data, Long and Langley (1999) showed that power to detect such associations is dependent on the extent of LD, marker density, sample size, quantitative trait loci (QTL) number, and QTL effect size.

Long and Langley (1999) simulated genotypes from a random-mating, panmictic population and did not consider the effects of population structure, kinship, or mating system. While most LD is...
a consequence of physical linkage, population structure and polygenic background effects can cause strong correlations between unlinked loci, leading to spurious associations between a marker and phenotype (Kennedy et al., 1992), a phenomenon that has been observed frequently in empirical data (e.g., Marchini et al., 2004; Aranzana et al., 2005; Yu et al., 2006). The need to control these Type I errors has motivated the development of new statistical methods for AM. Candidate gene simulations based on single nucleotide polymorphism (SNP) data support use of a mixed-model approach (Yu et al., 2006) that adjusts for population structure and polygenic background for maize and polygenic background for human data. Other researchers have compared specific methods used in mixed-model kinship estimation (Zhao et al., 2007) and developed more efficient algorithms (Kang et al., 2008). Recently, Iwata et al. (2007; 2009) showed that Bayesian AM methods can perform well in detection of multiple QTL in rice, a species with substantial population structure.

In empirical data sets, population structure, kinship, and mating system may vary dramatically among AM populations such that questions of power and methodology need to be examined on a case-by-case basis. Self-pollinating plant species such as Arabidopsis and Hordeum vulgare L. show more extensive LD, which facilitates GWAS with fewer markers (Aranzana et al., 2005; Rostoks et al., 2009). Use of restricted germplasm can facilitate GWAS (e.g., Breseghello and Sorrells, 2006; Kraakman et al., 2006), though this reduces resolution. The effect of population structure will also vary depending on the trait of interest: traits that are highly correlated with population structure will present problems for GWAS, regardless of methodology (Zhao et al., 2007; Yu et al., 2006).

The Barley Coordinated Agricultural Project (BarleyCAP; http://www.barleycap.org [verified 21 Sept. 2010]) was initiated to integrate the resources of barley genomics and genetics for the development of superior barley cultivars. A main objective of BarleyCAP is the development of a moderately high-density LD genetic map for use in marker-assisted selection and QTL identification through GWAS. A collection of 3840 diverse barley genotypes contributed from 10 barley breeding programs (http://www.barleycap.org) are being genotyped at ~3000 SNPs to meet this goal (Close et al., 2009; Waugh et al., 2009). While the number of SNPs is not high, it may be sufficient given the extent of LD in barley.

Population structure is substantial in the BarleyCAP germplasm (Hamblin et al., 2010), and some of the lines are closely related. The object of our study was to assess the power of GWAS given the SNP density and the characteristics of this germplasm. More specifically, we evaluated the power to detect QTL in relation to false discovery rate (FDR) as a function of QTL size, number, and heritability; the nominal p-value to be used in declaring significance; the best statistical model to be used in data analysis; and the effect of sample size on the power to detect QTL. To accomplish this, we simulated QTL across a range of heritabilities and QTL effect sizes in subsets of the BarleyCAP germplasm and obtained estimates of power and FDR using a variety of statistical models that account for population structure and kinship. Results of these analyses will provide guidelines for BarleyCAP researchers while, more generally, providing information on GWAS in self-pollinated crop species.

**MATERIALS AND METHODS**

**Plant Material and Genotype Data**

The germplasm was submitted in years 2006 and 2007 from 10 U.S. barley breeding programs (described in Hamblin et al., 2010). DNA from these lines were genotyped with the 3072 SNPs of the barley oligonucleotide pool assay (BOPA) 1 and 2 developed as part of the BarleyCAP (Rostoks et al., 2009; Close et al., 2009). Unmapped SNPs and those with minor allele frequencies (MAF) <0.028 were removed from the data set (i.e., at least 50 lines carried the minor allele). After removing duplicate lines and lines with large amounts of missing data, the final data set consisted of 2198 mapped SNPs scored in 1803 lines.

Previous model-based structure analysis (Pritchard et al., 2000) of this germplasm, using genotype data from BOPA1 only, identified seven subpopulations (Hamblin et al., 2010). We did not redo the structure analysis with the additional data from BOPA2 because the Structure program does not perform well with large numbers of linked markers; for this reason, Hamblin et al. (2010) only used a subset of the BOPA1 markers. Additionally, principle component analysis (PCA) was performed on the SNP data set using the function prcomp in the stats package of R (R Development Core Team, 2008). For all analyses, the number of PCA axes was equal to the number of Q-vectors minus one (i.e., six) so that the two methods could be compared directly.

Germplasm subsamples of various sizes were created with the following approach, designed to capture as much genotypic diversity as possible: K-means cluster analysis with the Hartigan-Wong algorithm (R Development Core Team, 2008) was used on the 1803 line × 2198 marker data to group lines into n clusters based on their genotypes, and one individual was randomly chosen from each cluster to form a sample of n lines. Six such subsamples were created where n = 100, 200, 300, 400, 800, and 1600. An additional data set, called population MN (for Minnesota), included 304 lines identified in the population structure analysis as having at least half their alleles contributed by the 6-row spring population dominated by Minnesota germplasm (Hamblin et al., 2010). For each of those data sets, markers with a MAF less than 0.05 were removed. In addition, if any pairs of markers had a correlation greater than 0.99, one of the pair was removed. Removal of low-MAF and correlated markers resulted in 100 × 1891, 300 × 1891, and 304 × 776 genotypic data sets used in the model comparison simulations and 200 × 1759, 300 × 1812, 400 × 1841, 800 × 1866, and 1600 × 1889 genotypic data sets used in the sample size comparisons. (The different numbers of markers in the two 300-line data sets is due to chance sampling.)
Phenotype Simulation
The 1-, 2-, and 10-QTL phenotypic data sets were created using the genotype data sets. To generate a genotype score for one QTL, a single marker was chosen. Values of 0 and 1 were assigned to the alleles at random. The genotype score of an individual was taken to be the sum of the scores of the two alleles for that marker. For 2- and 10-QTL phenotypic data, the appropriate number of markers was chosen at random with the restrictions, in the 2-QTL case, that the markers be on different chromosomes and, in the 10-QTL case, that no markers be closer than 10 cM. Allele values were assigned at random, and allele scores summed for each marker. Subsequently, the scores for each marker were divided by their standard deviation to standardize the genotypic variance of the QTL. These standardized scores were then summed across markers to create a phenotype. For each of the scenarios, 1000 phenotypes were generated and a normally distributed random error was added to the summed marker genotype scores to standardize heritability at 0.25, 0.75, and 1.0. Note that, because a subset of the loci is associated with population structure, the phenotype will be correlated with that structure when loci from that subset are used as QTL. Thus, though we did not simulate a population effect for the phenotype, such an effect occurs inevitably in the simulation process.

Quantitative trait loci in the MN population were simulated in two different ways. Method A: the simulation procedure was applied independently to the MN population so that each phenotype was the result of 10 segregating QTL. Method B: the QTL and error term used for the 300 line simulation were applied to the MN lines, which segregated for only 3.5 QTL, on average, rather than 10. Also, since the genetic variance was reduced but the error variance remained the same, Method B resulted in lower heritabilities for the MN population than for the 300 lines.

Statistical Analysis
A fixed-effects model was fit using lm in the stats package of R (Kang et al., 2008) or by using software written by the authors in Java, while the mixed model analysis was performed using Efficient Mixed-Model Analysis (EMMA) in R (Kang et al., 2008). The following six models were used:

1) Simple: \[ y = \mu + A\alpha + \varepsilon; \]
2) P-only: \[ y = \mu + A\alpha + P\nu + \varepsilon; \]
3) Q-only: \[ y = \mu + A\alpha + Q\nu + \varepsilon; \]
4) K-only: \[ y = \mu + A\alpha + Z\nu + \varepsilon; \]
5) P + K: \[ y = \mu + A\alpha + P\nu + Z\nu + \varepsilon; \] and
6) Q + K: \[ y = \mu + A\alpha + Q\nu + Z\nu + \varepsilon. \]

In these models, \( y \) was the phenotypic response vector; \( \mu \) was the population mean; \( \alpha \) was the vector of allelic effects for individual SNPs; \( \nu \) was the vector of population effects; \( \nu \) was the vector of polygenic background effects; \( \varepsilon \) was the vector of residual effects; \( Q \) was a matrix of vectors calculated by STRUCTURE relating \( y \) to \( \nu \); \( P \) was a matrix composed of eigenvectors from PCA relating \( y \) to \( \nu \), and \( A \) and \( Z \) were the incidence matrices of ones and zeroes relating model parameters \( \alpha \) and \( \nu \) to \( y \). The parameters \( \nu \) and \( \varepsilon \) were fit as random effects with \( \text{Var}(\nu) = \sigma^2_K \) and \( \text{Var}(\varepsilon) = \sigma^2_I \) resulting in the phenotypic variance \( \text{Var}(y) = \sigma^2_K ZK' + \sigma^2_I \), in which \( K \) is the kinship matrix and \( I \) is an identity matrix. Kinship was calculated as the proportion of shared alleles, as implemented in EMMA (Kang et al., 2008). All other parameters were fit as fixed effects.

Power and False Discovery Rate
The \( p \)-values from the AM analysis were used to calculate power and FDR as follows: For each scenario (i.e., QTL number and size), a nominal \( p \)-value was obtained for each of the phenotypes. The range of nominal \( p \)-values was divided into 100 log intervals, yielding a set of 100 threshold maximum values (\( p_{\text{max}} \)). For each of these 100 \( p_{\text{max}} \) values, markers with \( p \)-value below \( p_{\text{max}} \) were declared “discoveries,” and their chromosome and position noted. A discovery was declared true if it was within 10 cM of a QTL. Conversely, a QTL was declared identified if it was within 10 cM of a discovery. Power was the number of identified QTL divided by the number of simulated QTL. The number of true discoveries was counted. False discovery rate for each phenotype was calculated as 1–(true discoveries/total discoveries). If no discoveries were obtained for a phenotype, that phenotype was not included. The overall FDR for each \( p_{\text{max}} \) (i.e., empirical FDR in Tables 1 and 2) was the average of the FDRs for each phenotype. Similarly, overall power was the average of power calculated for individual phenotypes. In Tables 1 and 2, FDR was also calculated according to Benjamini and Hochberg (1995); empirical \( p \)-value or type I error rate was calculated as (the number of false discoveries)/(the number of markers not within 10 cM of a QTL). Calling this value the empirical \( p \)-value emphasizes that the nominal \( p \)-value is estimating it based on the assumption that the \( F \)-statistic has an \( F \)-distribution. Likewise, the Benjamini-Hochberg false discovery rate (BHFDK) is an estimate of the empirical FDR. In all the figures, we present power as a function of FDR, since the critical \( p \)-value used to achieve a given FDR varied among the simulations.

RESULTS
Statistical Model
We compared the power to detect a single simulated QTL using six models that account for population structure and relatedness among lines in different ways (see Methods). Figure 1 shows the results for sample sizes of 100 and 300 lines with two QTL at three heritabilities. While the simple model has almost no power to detect QTL at reasonable FDR, the other models show a dramatic improvement. The K-only model (kinship matrix only) performed as well as or slightly better than all other models across all heritability levels and QTL sizes. Models that included K performed better than those that did not. Power increased substantially when heritability increased from 0.25 to 0.75, with smaller increases in power when heritability increased from 0.75 to 1.0. Using 300 lines provided substantially more power than the simulation with 100 lines.

Phenotypic Model
Using the same six models, the power to detect QTL was calculated over a range of QTL numbers and sizes. The power decreased as the number of QTL controlling the trait increased (Fig. 1C and 1D). At a heritability of 0.25 with 1
QTL, or at heritabilities of 0.75 and 1.0 with 1 or 2 QTL, power was high even at low FDR (0.1). When the number of QTL controlling the trait increased to 10, larger QTL (7.5–10%) could still be detected at low FDR, but power to detect a small QTL (2.5%) at a FDR of 0.1 was near zero. Table 1 shows, for various combinations of QTL size and heritability, the nominal \( p \)-values associated with specific levels of false discovery that were used to declare that QTL were significant and the corresponding power values.

### The Effect of Sample Diversity

To evaluate the effect of population diversity and structure, we compared results for 300 diverse lines to those for a sample of 304 lines all belonging to the same sub-population [MN (Minnesota)] as defined by model-based structure analysis (see Methods). These results are shown for simulations of 10 QTL only (Fig. 2). Quantitative trait loci in the MN population were simulated in two different ways: Method A addresses the impact of reduced diversity without the complication of changing QTL number, while Method B addresses the scenario in which the same trait is analyzed in two different populations (see Methods). For method A, power in MN was slightly higher both at moderate (0.75) and low (0.25) heritability: FDR of 0.1 resulted in 0.71 power to detect a 7.5% QTL compared with 0.48 for the 300-line data set. When the heritability was 1.0, the models including K had slightly more power in the MN population but the P- and Q-only models had less power. For method B, power was higher for the MN population at heritabilities of 1 and 0.75 but worse at 0.25. The K-only models were as good as or better than the other models except at 0.75 heritability when K + Q had slightly more power over a range of FDR.

### The Effect of Sample Size

Using a single phenotypic and genetic model (P-only, 10 QTL, and heritability of 0.75), we determined power as a function of sample size (Fig. 3). This model was used to

---

### Table 1. Nominal \( p \)-value, power, and false discovery rate associated with quantitative trait loci (QTL) detection. Simulations with 300 lines were analyzed using the K-only model. \( p \)-values and power at specific false discovery rate (FDR) were interpolated based on original power and FDR output.

<table>
<thead>
<tr>
<th>Empirical FDR</th>
<th>Heritability</th>
<th>QTL No.</th>
<th>Nominal ( p )-value</th>
<th>Power</th>
<th>Empirical ( p )-value</th>
<th>BH FDR†</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1</td>
<td>1</td>
<td>( 5.61 \times 10^{-5} )</td>
<td>0.980</td>
<td>( 1.83 \times 10^{-4} )</td>
<td>0.019</td>
</tr>
<tr>
<td>0.05</td>
<td>0.75</td>
<td>1</td>
<td>( 2.20 \times 10^{-5} )</td>
<td>0.969</td>
<td>( 1.43 \times 10^{-4} )</td>
<td>0.012</td>
</tr>
<tr>
<td>0.05</td>
<td>0.25</td>
<td>1</td>
<td>( 1.27 \times 10^{-6} )</td>
<td>0.743</td>
<td>( 5.12 \times 10^{-6} )</td>
<td>0.004</td>
</tr>
<tr>
<td>0.05</td>
<td>1</td>
<td>2</td>
<td>( 5.95 \times 10^{-4} )</td>
<td>0.976</td>
<td>( 3.03 \times 10^{-4} )</td>
<td>0.118</td>
</tr>
<tr>
<td>0.05</td>
<td>0.75</td>
<td>2</td>
<td>( 1.18 \times 10^{-4} )</td>
<td>0.934</td>
<td>( 1.98 \times 10^{-4} )</td>
<td>0.081</td>
</tr>
<tr>
<td>0.05</td>
<td>0.25</td>
<td>2</td>
<td>( 2.93 \times 10^{-6} )</td>
<td>0.359</td>
<td>( 3.35 \times 10^{-6} )</td>
<td>0.011</td>
</tr>
<tr>
<td>0.05</td>
<td>1</td>
<td>10</td>
<td>( 4.25 \times 10^{-3} )</td>
<td>0.753</td>
<td>( 4.68 \times 10^{-4} )</td>
<td>0.469</td>
</tr>
<tr>
<td>0.05</td>
<td>0.75</td>
<td>10</td>
<td>( 7.41 \times 10^{-4} )</td>
<td>0.357</td>
<td>( 1.58 \times 10^{-4} )</td>
<td>0.310</td>
</tr>
<tr>
<td>0.05</td>
<td>0.25</td>
<td>10</td>
<td>( 4.62 \times 10^{-8} )</td>
<td>0.001</td>
<td>( 8.46 \times 10^{-7} )</td>
<td>0.001</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>( 1.34 \times 10^{-3} )</td>
<td>0.990</td>
<td>( 7.19 \times 10^{-4} )</td>
<td>0.173</td>
</tr>
<tr>
<td>0.1</td>
<td>0.75</td>
<td>1</td>
<td>( 2.24 \times 10^{-4} )</td>
<td>0.983</td>
<td>( 3.64 \times 10^{-4} )</td>
<td>0.079</td>
</tr>
<tr>
<td>0.1</td>
<td>0.25</td>
<td>1</td>
<td>( 6.22 \times 10^{-6} )</td>
<td>0.792</td>
<td>( 1.25 \times 10^{-4} )</td>
<td>0.016</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>2</td>
<td>( 2.69 \times 10^{-3} )</td>
<td>0.986</td>
<td>( 9.20 \times 10^{-4} )</td>
<td>0.326</td>
</tr>
<tr>
<td>0.1</td>
<td>0.75</td>
<td>2</td>
<td>( 5.27 \times 10^{-4} )</td>
<td>0.963</td>
<td>( 5.56 \times 10^{-4} )</td>
<td>0.219</td>
</tr>
<tr>
<td>0.1</td>
<td>0.25</td>
<td>2</td>
<td>( 1.44 \times 10^{-5} )</td>
<td>0.467</td>
<td>( 1.03 \times 10^{-4} )</td>
<td>0.041</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>( 9.76 \times 10^{-3} )</td>
<td>0.834</td>
<td>( 1.33 \times 10^{-3} )</td>
<td>0.741</td>
</tr>
<tr>
<td>0.1</td>
<td>0.75</td>
<td>10</td>
<td>( 1.93 \times 10^{-3} )</td>
<td>0.470</td>
<td>( 4.93 \times 10^{-4} )</td>
<td>0.456</td>
</tr>
<tr>
<td>0.1</td>
<td>0.25</td>
<td>10</td>
<td>( 1.94 \times 10^{-8} )</td>
<td>0.001</td>
<td>( 1.06 \times 10^{-6} )</td>
<td>0.000</td>
</tr>
</tbody>
</table>

†The Benjamini-Hochberg false discovery rate (BH FDR) is the FDR value calculated according to Benjamini and Hochberg (1995).

---

### Table 2. Empirical false discovery rates (FDRs) and \( p \)-values associated with a BH FDR† of 0.1 under different scenarios of quantitative trait loci (QTL) number and size. Simulations with 300 lines were analyzed using the K-only model.

<table>
<thead>
<tr>
<th>BH FDR</th>
<th>Heritability</th>
<th>QTL No.</th>
<th>Nominal ( p )-value</th>
<th>Power</th>
<th>Empirical ( p )-value</th>
<th>BH FDR†</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>( 3.62 \times 10^{-4} )</td>
<td>0.993</td>
<td>( 4.95 \times 10^{-4} )</td>
<td>0.100</td>
</tr>
<tr>
<td>0.1</td>
<td>0.75</td>
<td>1</td>
<td>( 3.09 \times 10^{-4} )</td>
<td>0.985</td>
<td>( 4.28 \times 10^{-4} )</td>
<td>0.110</td>
</tr>
<tr>
<td>0.1</td>
<td>0.25</td>
<td>1</td>
<td>( 1.69 \times 10^{-4} )</td>
<td>0.875</td>
<td>( 5.20 \times 10^{-4} )</td>
<td>0.237</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>2</td>
<td>( 4.61 \times 10^{-4} )</td>
<td>0.978</td>
<td>( 2.87 \times 10^{-4} )</td>
<td>0.048</td>
</tr>
<tr>
<td>0.1</td>
<td>0.75</td>
<td>2</td>
<td>( 3.52 \times 10^{-4} )</td>
<td>0.941</td>
<td>( 2.60 \times 10^{-4} )</td>
<td>0.061</td>
</tr>
<tr>
<td>0.1</td>
<td>0.25</td>
<td>2</td>
<td>( 1.23 \times 10^{-4} )</td>
<td>0.556</td>
<td>( 2.87 \times 10^{-4} )</td>
<td>0.170</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>( 3.59 \times 10^{-4} )</td>
<td>0.517</td>
<td>( 3.82 \times 10^{-5} )</td>
<td>0.007</td>
</tr>
<tr>
<td>0.1</td>
<td>0.75</td>
<td>10</td>
<td>( 1.14 \times 10^{-4} )</td>
<td>0.196</td>
<td>( 2.79 \times 10^{-5} )</td>
<td>0.015</td>
</tr>
<tr>
<td>0.1</td>
<td>0.25</td>
<td>10</td>
<td>( 6.90 \times 10^{-6} )</td>
<td>0.030</td>
<td>( 1.43 \times 10^{-4} )</td>
<td>0.348</td>
</tr>
</tbody>
</table>

†The Benjamini-Hochberg false discovery rate (BH FDR) is the FDR value calculated according to Benjamini and Hochberg (1995).
Figure 1. Power and false discovery rate (FDR) for identifying 2 quantitative trait loci (QTL) using six models at three heritability ($h^2$) levels (0.25, 0.75, and 1.0). Simple: no control for population structure. K: control using a kinship matrix. P: control using principal component eigenvectors; Q: control using admixture percentages from STRUCTURE.

Figure 2. Comparison of power in a diverse sample of 300 lines and a less diverse subpopulation of 304 lines (MN Population) with 10 simulated quantitative trait loci (QTL) at three heritability levels. FDR, false discovery rate.
avoid the large amount of computational time required for the K-only model with 1600 lines. For 1600 lines at 0.10 FDR, power was 0.74, only 4% higher than for 800 lines. For sample sizes smaller than 800, power at FDR of 0.1 was 0.48 or lower.

Empirical vs. Nominal False Discovery Rate
It is standard to employ a procedure such as that suggested by Benjamini and Hochberg (1995) to estimate the false discovery rate when a large number of tests are performed. We were interested in comparing such an estimated FDR with the empirical FDR calculated from our simulation results. For the K-only analysis, we used the Benjamini and Hochberg procedure to calculate a BH- FDR of 0.1 and found the empirical FDR as well as the nominal and empirical p-values (Table 2). For traits with a few QTL and high heritability, the BH- FDR is a reasonable approximation of the empirical FDR, but for larger numbers of QTL, BH- FDR overestimates the FDR. For traits with low heritability and few QTL, BH- FDR underestimates the FDR. When the QTL number is high and heritability is low, power to detect QTL is poor and FDR cannot be estimated accurately.

DISCUSSION
Because populations vary in the extent of LD and population structure, it is necessary to empirically determine which AM method is most appropriate for the particular population under investigation and the necessary marker density. Our simulations examined the effects of several phenotypic (QTL number and effect size) and statistical models on our ability to detect simulated QTL in samples of barley germplasm drawn from the BarleyCAP data set.

Our results indicate that the SNP density and the lines available in the BarleyCAP germplasm are sufficient for detection of QTL using GWAS, supporting the conclusion of Rostoks et al. (2009) that a recent history of artificial outcrossing, coupled with extensive LD, facilitates GWAS in barley. Simulations with 300 diverse lines resulted in sufficient power to detect QTL when the trait was controlled by a few QTL with large to moderate effect sizes, although power to detect small effect size QTL was poor. Due to computational limitations, our sample sizes were fairly small; the coordinated agricultural project germplasm will eventually include 3840 lines, which will provide increased power to detect smaller QTL.

Because our analyses were based on simulated QTL, we were able to calculate an exact FDR. However, in studies of real phenotypic data, the choice of an exact cutoff value for significance is problematic. Table 1 shows that the critical p-values required to achieve desirable FDRs vary with both heritability and QTL number. This difficulty arises, in part, from the fact that the p-values from the tests of association between markers and phenotypes, on which BH- FDR is based, indicate statistical significance only. On the other hand, the empirical FDR tests proximity of markers to QTL. While association implies proximity, the relationship is not exact. One result of this study is to demonstrate that the relationship between tests of association and physical linkage varies with the number and size of QTL. In spite of this difficulty, because some type of correction for multiple testing is necessary, a reasonable approach would be to use BH- FDR to decide which associations to accept. This test is likely to be conservative but not as conservative as a Bonferroni correction.

One of the goals of this study was to determine what form of the mixed model would best account for population structure. The curves in Fig. 1 clearly show that accounting for population structure dramatically increases power; this is expected because the empirical genotype data come from a strongly structured set of germplasm. For this barley data set, the K-only model performed as well as or better than all other models. Prior studies on AM (Yu et al., 2006; Zhao et al., 2007; Stich et al., 2008; Zhu and Yu, 2009) have found that Q + K and P + K mixed models outperform the K-only model; these studies, however, have used data from maize, Arabidopsis spp., and wheat. There are two factors that might explain why we found the K-only model to be superior in our simulations. First, it may be that including structure in the models would improve performance for some of the simulated phenotypes even though on average that was not the case. To examine this possibility we divided phenotypes from the single QTL case into quartiles based on their association with population structure as assessed by PCA. We then calculated power and FDR for each set, hypothesizing that markers in the top quartile would perform better under a model that included population structure. This was not
the case: the K-only model gave the best results in all quartiles, although the P-only model approached the K-only model for those markers in the top quartile. To explain these results, we hypothesize that structure control in the statistical model is a double-edged sword: while it reduces false positives from spurious associations with structure, it also reduces true positives from QTL that are correlated with structure. Presumably, the K-only model strikes the best balance and is least wounded by these two edges.

Factors affecting the performance of the various mixed models may include the method of calculating the kinship matrix (Zhao et al., 2007), the number of markers used to calculate kinship (Yu et al., 2009), or the specific phenotype being analyzed. In our analyses, we calculated kinship as a proportion of shared alleles as implemented in EMMA (Kang et al., 2008). None of the values in our kinship matrix were set to zero as they were by Yu et al. (2006). Furthermore, we used a large number of markers, so possibly the K-matrix in this study more accurately modeled true kinship than in some earlier studies. In addition, our study involved data simulated using a strictly additive model. Nonadditive genetic effects might behave differently.

Simulation with sample sizes from 100 to 1600 showed, not surprisingly, that larger samples resulted in higher power. Although the P-only model was used in this simulation, we believe, based on other results not shown here, that the K-only model would produce a similar outcome: Fig. 1 shows that the K-only model outperformed the P-only model and showed a similar increase in performance with increased sample size under all six scenarios tested. We therefore conclude that the K-only model, at 1600 lines, would also outperform the P-only model. Assuming a 10% increase in power of the K-only model over the P-only model at 0.10 FDR, that would put power of detection well above 0.80 in this scenario of 10 QTL and heritability of 0.75. Given that the final BarleyCAP data set will include 3840 lines, it should be possible to achieve increases in power and the ability to detect QTL of even smaller effect sizes.

To make best use of the BarleyCAP data, it is unclear whether testing should be done on all 3840 lines together or within subpopulations. We saw an increase in power in population MN (304 lines) compared to the 300-line full data set, even though the marker density of MN was lower (776 markers as compared to 1812 in the 300-line full data set). We have shown (Hamblin et al., 2010) that there are higher levels of LD within the smaller population, in which individuals are more closely related to each other than in the larger, more diverse population. These higher levels of LD result in greater power to detect QTL in all size classes and heritability levels, although resolution is decreased. Hamblin et al. (2010) presented this issue from a different perspective, showing that allelic associations may be negatively correlated in comparisons between more highly diverged populations of the BarleyCAP germplasm. Such negative correlations would have the effect of canceling out associations when divergent populations are pooled, causing LD to decay more rapidly. In addition, for most complex traits, a less diverse population will be segregating fewer QTL. That provides greater power to detect those QTL present but of course no power to detect QTL that are not segregating. These factors taken together suggest that a reasonable approach would be to test as diverse a population as possible given the goals of the experiment and the resources available. If the population under study has distinct and highly differentiated subpopulations, then, in addition to evaluating the population as a whole, the subpopulations should be analyzed separately.

The differences in population structure between diverse and narrow germplasm may also contribute to the observed differences in power. Although we did not explicitly incorporate population structure into our simulated phenotypes, the nature of the genotype data, which is strongly structured, ensured that many simulated QTL were correlated with that structure. In diverse populations, both simulated and real QTL would more often be correlated with population structure and thus harder to detect. These results suggest that, while larger sample size generally results in greater power, it may sometimes be appropriate to conduct analyses with subsets of the BarleyCAP germplasm that are more similar to the germplasm found in a typical breeding program. Separate analysis of 2-row and 6-row germplasm, for example, would usually be a good practice.

CONCLUSIONS

The results of our GWAS simulations support the use of GWAS in the BarleyCAP data. While larger sample sizes generally provide higher power, analysis of less diverse subsets will be appropriate in some cases. The results described strongly support the use of mixed models with a kinship matrix to correct for population structure. Although our simulation results suggest that the K-only model performs best under a wide variety of conditions, the assumptions of the model may not always be true in practice; for some phenotypes, it may be appropriate to include population structure covariates by using either a P + K or Q + K model.

References


