





Importance of correcting genomic relationships in single-locus QTL mapping model with an advanced backcross population

Boby Mathew ^{1,*†}, Jens Léon,¹ Said Dadshani ¹, Klaus Pillen ², Mikko J. Sillanpää ^{3,‡} and Ali Ahmad Naz^{1,‡}

¹Institute of Crop Science and Resource Conservation, Department of Plant Breeding, University of Bonn, 53115 Bonn, Germany

²Department of Plant Breeding, Institute of Agricultural and Nutritional Sciences, Martin-Luther University Halle-Wittenberg, 06120 Halle (Saale), Germany

³Department of Mathematical Sciences, FIN-90014 Oulu, Finland

[†]Present address: Bayer CropScience, Monheim am Rhein, 40789, Germany.

[‡]These authors contributed equally to this work.

*Corresponding author: boby.mathew@hotmail.com

Abstract

Advanced backcross (AB) populations have been widely used to identify and utilize beneficial alleles in various crops such as rice, tomato, wheat, and barley. For the development of an AB population, a controlled crossing scheme is used and this controlled crossing along with the selection (both natural and artificial) of agronomically adapted alleles during the development of AB population may lead to unbalanced allele frequencies in the population. However, it is commonly believed that interval mapping of traits in experimental crosses such as AB populations is immune to the deviations from the expected frequencies under Mendelian segregation. Using two AB populations and simulated data sets as examples, we describe the severity of the problem caused by unbalanced allele frequencies in quantitative trait loci mapping and demonstrate how it can be corrected using the linear mixed model having a polygenic effect with the covariance structure (genomic relationship matrix) calculated from molecular markers.

Keywords: Bayesian multi-locus model; Kinship correction in experimental populations

Introduction

QTL (quantitative trait loci) mapping has been proven to be very useful in crop breeding to identify genetic regions associated with a trait of interest (Morrell *et al.* 2011). Development of the experimental populations is the first step in QTL mapping and biparental populations such as F₂, backcrosses (BC), doubled haploids (DH), or recombinant inbred lines (RIL), can be utilized for the mapping. Advanced backcross QTL analysis (AB-QTL) proposed by Tanksley and Nelson (1996) has been widely used as a method for combining QTL analysis with variety development in plant breeding programs (Bauer *et al.* 2009; Nagata *et al.* 2015; Wang *et al.* 2017b), see Wang and Chee (2010) for a more detailed review. One of the main purposes of the development of AB experimental population is to transfer the favorable QTL alleles from unadapted (*e.g.*, landraces, wild forms) to cultivated gene pool (Tanksley and Nelson 1996). During the development of these populations, selection (both natural and artificial) of agronomically adapted traits will remove the unfavorable alleles coming from the donor parent from the population and this selection will reduce the frequency of the donor genome in each of the AB lines. In addition, due to distinct crossing (wild vs cultivated), there may also be other biological phenomenon or disturbances present that can influence selectively to the process [*e.g.*, hybrid

necrosis (Bomblies and Weigel 2007); hybrid sterility (Ouyang *et al.* 2010); hybrid lethality (Garner *et al.* 2016)]. Thus, the allele frequencies in the AB families are skewed toward the alleles from the recurrent parent (Grandillo and Tanksley 2005), which may lead to substructure in the AB population.

Traditional methods for performing QTL mapping are based on standard regression techniques and they assume that the population is identically and independently distributed (*i.e.*, individuals are equally related to each other). However, this assumption is not valid when selection of favorable genes and appropriate parents is performed during the development of the experimental population. Such selection process will lead to hidden substructure in the experimental population. It is well known from association and QTL mapping studies based on multiparental population that hidden population structure, cryptic relatedness (some members of the population are more closely related to another), and polygenicity (many small genetic effects) all can yield false-positive QTL signals (Kang *et al.* 2008a; Würschum and Kraft 2015; Sul *et al.* 2018). However, it is commonly believed that mapping traits in experimental crosses of (biparental) populations are immune to cryptic relatedness problem and the only motivation for including polygenic term to the model in these crosses are because of polygenicity (see *e.g.*, Taylor and Verbyla 2011). Moreover, segregation distortion

Received: February 09, 2021. Accepted: March 18, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

is a common phenomenon associated with QTL mapping studies in experimental population. Additionally, if the markers, which cause segregation distortion is in linkage disequilibrium (LD) with genes under selection, these markers cannot be considered as neutral markers (Xu 2008). Thus, depending on the type of population, prior to the QTL mapping study, markers that show significant segregation distortion are commonly removed. However, in high LD populations such as experimental populations with high-density markers, this might not be trivial due to the availability of genome-wide marker information.

Despite the availability of genome-wide marker information, standard genome-wide association study (GWAS) analysis methods consider one marker at a time and identify the marker-trait association using a single-locus model. The single-locus model is the most commonly used model to identify marker-trait association (Balding 2006; Huang et al. 2010; Zhao et al. 2011; Li et al. 2018). Various correction methods have been proposed to correct for the hidden population structure in single-locus association analysis [e.g., principal component analysis (PCA) (Price et al. 2006; McVean 2009); mixed-model approach (Yu et al. 2006; Kang et al. 2008b); structured association (Pritchard et al. 2000)]. For a review of different methods, see Sillanpää (2011). Mixed-model approach including a random polygenic effect in the model, which describes relationships between individuals in a population, is the widely used method to correct for the population substructure and polygenic effect in plant, animal, and human GWAS studies (Kang et al. 2008a; Listgarten et al. 2010; Parks et al. 2013; Mora et al. 2016; Yano et al. 2016). However, current correction methods including mixed-model approach cannot distinguish between inflated test statistics due to population structure (or cryptic relatedness) and polygenic genetic architecture because they both lead to increased number of false positives in GWAS.

Multi-locus model (Pikkuhookana and Sillanpää 2009; Kärkkäinen and Sillanpää 2012; Wen et al. 2019) is another interesting alternative to correct for the confounding due to population structure/cryptic relatedness without having polygenic effect in the model. In contrast to the single-locus model, multi-locus model jointly fit all markers and by considering all markers simultaneously in the model, it can increase the power to detect association signals. However, Bayesian multi-locus association analysis using Markov Chain Monte Carlo (MCMC) is computationally demanding.

The main focus of this study is to find out, if inclusion of polygenic correction term (with genomic relationships) to the single-locus model improves QTL mapping power and control of false-positive QTL also in experimental crosses of AB population (which is not common practice) similarly as in mixture of multiple strains/multiparental populations (Pascual et al. 2015; Sul et al. 2018) or in population association studies. Additionally, we also want to explore the widely accepted view that mapping traits in experimental crosses without notable genotype by environment interactions are immune to cryptic relatedness problem. Finally, we want to see if multi-locus QTL model works similarly in biparental populations as in multiparental populations or in population association studies. We present the results based on two wheat AB populations along with simulated data sets. In addition, we compare the results based on single-locus model with the Bayesian multi-locus model along with the traditionally used interval mapping method.

Materials and methods

Let us consider the single-locus model with the polygenic random effect as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{W}\mathbf{v} + \boldsymbol{\epsilon}. \quad (1)$$

Here, \mathbf{Y} is a phenotypic vector, which is the mean of phenotype over different environments and replications for n unique varieties and $\boldsymbol{\beta}$ is the vector of fixed effects (in this case only the grand mean) with known incidence matrix \mathbf{X} , whereas, \mathbf{g} is an $n \times 1$ vector of polygenic effects with the incidence matrix \mathbf{Z} and $\mathbf{g} \sim N(0, \mathbf{K}\sigma_g^2)$ (assuming no interaction between genotype and environment). Here, \mathbf{K} defines the covariance structure that describes the relatedness among individuals and can be calculated either based on marker information or with the pedigree. In this study, we calculated the \mathbf{K} matrix with the function *A.mat* available in the R package *rrBLUP* (Endelman 2011) using the marker information. Moreover, \mathbf{W} is the incidence matrix for the marker being tested for the association and with single-locus model, the marker association is tested one marker at a time with the null hypothesis, that is, $v=0$ against the alternative hypothesis, that is, $v \neq 0$, here v is treated as a fixed effect. Additionally, $\boldsymbol{\epsilon}$ corresponds to the vector of residuals, following a normal distribution as $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma_\epsilon^2)$. The multi-locus association model can be defined as:

$$\mathbf{Y} = \boldsymbol{\mu} + \sum_{j=1}^m \mathbf{M}_j \mathbf{b}_j + \boldsymbol{\epsilon}. \quad (2)$$

Here, $\mathbf{Y} = \{Y_i\}_{i=1}^n$ is a phenotypic vector, which is the mean of phenotypes over different environments and replications for n unique varieties, m is the total number of markers, \mathbf{M}_{ij} is the genotypic value of line i at marker j coded as 0, 1, 2 for the genotype AA, Aa, aa, respectively. Moreover, \mathbf{M}_j is the vector of genotypic values of a line, \mathbf{b}_j is the random marker effect associated with marker j , and $\boldsymbol{\epsilon}$ corresponds to the residual, following a normal distribution as $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma_\epsilon^2)$.

In Bayesian estimation, one needs to specify the prior distribution for the unknown parameters in the Equation (2). Following Xu (2003) and Meuwissen et al. (2001), the random marker effects (\mathbf{b}_j) were assigned a normal distribution with mean zero and marker-specific variance σ_j^2 . Here, the marker-specific variances were assigned a Jeffreys' scale-invariant prior, thus, $p(\sigma_j^2) \propto 1/\sigma_j^2$ for $j = 1, \dots, m$. The prior density for the mean $\boldsymbol{\mu}$ is $p(\boldsymbol{\mu}) \propto 1$. Let $\mathbf{b} = \{\mathbf{b}_j\}$ and $\boldsymbol{\sigma}^2 = \{\sigma_j^2\}$ for $j = 1, 2, \dots, m$ be the unknown model parameters, then the likelihood of the observation vector \mathbf{Y} is:

$$p(\mathbf{Y}|\mathbf{b}, \boldsymbol{\sigma}^2) \propto (\sigma_0^2)^{-n/2} \times \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (\mathbf{Y}_i - \boldsymbol{\mu} - \sum_{j=1}^m \mathbf{M}_{ij} \mathbf{b}_j)^2\right). \quad (2)$$

Here, σ_0^2 is the residual error variance. By Bayes theorem, the joint posterior distribution of the model parameters is proportional to:

$$p(\mathbf{b}, \boldsymbol{\sigma}^2|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{b}, \boldsymbol{\sigma}^2)p(\mathbf{b}, \boldsymbol{\sigma}^2). \quad (4)$$

Following Xu (2003), we applied Gibbs sampling to draw samples from the above joint posterior distribution. R code used in this study is publicly available along with Supplementary materials.

When multi-locus association models are applied for Bayesian association analysis, one needs to perform additional

confirmatory-test for identifying the positive association signals. Permutation test is one of the commonly used methods with multi-locus model to identify the significant marker-trait association (Xu 2003). However, the phenotype permutation test with Bayesian multi-locus model is computationally challenging and additionally it also highly depends on the collinearity in the marker data. So in this study, as an alternative as proposed by Mathew et al. (2018), we used five different MCMC chains, and only picked markers that were constantly appearing in all chains as evidence of decisive association. In each of the MCMC chains, we used 50,000 iterations with a burn-in period of 10,000 iterations and retained every 50th iteration (thinning). For the estimation of the posterior mean of marker effect, we calculated the average marker effect over five MCMC chains.

The package *rrBLUP* uses linear mixed model and includes a random polygenic effect, which describes the genomic relationships between individuals, in order to correct the sample structure and relatedness in the data set.

The field data sets used in this study were collected in multi-environmental trial with replication. We used the mean of phenotype over different environments and replications as the phenotypic vector (\mathbf{Y}) in both models because the data set was collected in 2008 and we were not able to retrieve the raw data set. But it is possible to identify QTL by joint analysis across multiple environments in multi-environment trials [see Gogel et al. (2018) for more details, for package see Taylor and Verbyla (2011)].

PCA (Jolliffe 2003) is a widely used method to identify patterns of genetic substructure in populations (McVean 2009; Ma and Amos 2012). The top principal components (PCs) reflect the variations due to genetic substructure in the sample and the scatter plot of the lines based on the first two PCs can be used to visualize the patterns of genetic variation in the population. In this study, we used PCA in order to visualize the substructure among the lines in the AB populations. Here, we also used the receiver operating characteristic (ROC) curve (Fawcett 2006) to visualize the estimation accuracy of single-locus and Bayesian multi-locus model.

Field data of the AB populations

The following field data sets were used in this study along with the simulation replicates to demonstrate the importance of correcting genomic relationships in QTL mapping with AB populations.

B22 population

A mapping population designated as B22 comprising of 250 BC2F3 lines was used for this study. To develop this population, the winter wheat cultivar Batis was crossed with the synthetic wheat accession Syn022L and two backcrosses were made to Batis (as recurrent parent), using the F1 and BC1F1 plants as the maternal parents. Hereafter, we refer this population as B22 population. As described by Kunert et al. (2007), the resulting BC2F1 plants were self-pollinated, and single seed descent was used in order to obtain 250 BC2F3 plants. We used the phenotype thousand grain weight (TGW), which is the average weight of 1000 kernels for this study. This population was genotyped using 15k iSelect single nucleotide polymorphism (SNP) arrays. After excluding markers with minor allele frequency (MAF) ≤ 0.05 and missing values $\geq 20\%$, 2745 SNPs were available for the QTL analysis. After the SNP filtering, missing markers were imputed by random sampling based on the allele distribution in the population using the R package *synbreed* (Wimmer et al. 2012).

Z86 population

For the development of the Z86 population, the winter wheat cultivar Zentos was crossed with synthetic wheat Syn086L and two backcrosses were made to Zentos (as recurrent parent) resulting in an AB population of 150 BC2F3 lines. This population was genotyped using 15k iSelect SNP arrays. After excluding markers with MAF ≤ 0.05 and missing values $\geq 20\%$, 5149 SNPs were available for the analysis. In this population, we used the trait yield, which is the average yield (YLD) of a plot as the phenotype for the QTL analysis.

Phenotyping

The measurement of traits grain yield (YLD) and TGW in both AB populations was carried out in field conditions at five different locations in 2 years (10 environments) across Germany. The field stations were distributed on the following locations: (1) research station University of Bonn at Dikopshof (west Germany), (2) Limagrain–Nickerson field stations at Adenstedt (central Germany), (3) Fr. Strube Saatzeit field station at Jerxheim (central Germany), (4) Saatzeit Josef Breun field station at Morgenrot (East Germany), and (5) Lochow–Petkus field station at Wohlde (North Germany). At each test location, AB lines and their recurrent parents were sown in randomized block design comprising of 1 plot of individual AB lines and 20 plots of Batis and 10 plots of Zentos. The plot sizes were 4.5–6.3 m² where in each plot seed density was 310–360 kernels per m². Standard field management and fertilizer application were made according to local practice. From each location, grain yield was measured in one-tenth of a ton per hectare calculated from weight of grain harvested per plot and designated as desi ton per hectare (dt/ha). The net weight of 1000 kernels was taken and measured as TGW.

Simulated data sets

In order to estimate the QTL mapping accuracy of different models, we simulated two data sets conditionally on real genotype data of the wheat BC2F3 population. For the first simulation, we randomly selected five markers as QTL in such a way that not two QTL are coming from the same chromosome, where the marker effects were generated from a uniform distribution $U(8,10)$. In the second simulation to assess the effect of polygenicity, along with the five main QTL, we randomly selected another 100 markers with small effects and their effects were generated from a normal distribution with mean 0 and variance 1 (here 20% of the genetic variance was due to the polygenic variance [see Pikkuhookana and Sillanpää (2009) for more details]). Hereafter, we refer to the first simulated data set as Simulation 1 and second data set as Simulation 2. Additionally, we also analyzed another 50 simulation replicates in order to compare the estimation accuracy of single- and multi-locus association model. The simulation replicates were generated based on the Simulation 2 data set by sampling different error term. The joint heritability of the simulated traits was about 0.6.

Data availability

Genotypic and phenotypic information of the B22 population are contained in the files *B22_genotype.txt* and *B22_pheno.txt*, respectively. Whereas, the files *Z86_genotype.txt* and *B22_pheno.txt* contain the genotypic and phenotypic information for the Z86 population. Additionally, the file *Supplementary Figures* contains all the supplementary figures mentioned in the manuscript. Supplementary material is available at figshare: <https://doi.org/10.25387/g3.12579737>.

Results

In this study, we identified the marker-trait association using single-locus model as well the Bayesian multi-locus model using real and simulated data sets. The results are presented in the following sections.

Marker-trait association using field data

The AB populations used in this study involve two backcrosses to the recurrent parent and due to the selected back crossing the lines will be closely related to the recurrent parent. In order to visualize the substructure among the lines in the AB populations, we performed a PCA and plotted the scatter plot of the lines based on the first two PCs, which is shown in Figure 1, A and B. PCA shows that the lines in both populations are closely related to the recurrent parents Batis and Zentos in B22 and Z86 populations, respectively. A primary reason behind this might lie on two subsequent backcrosses in the development of this population.

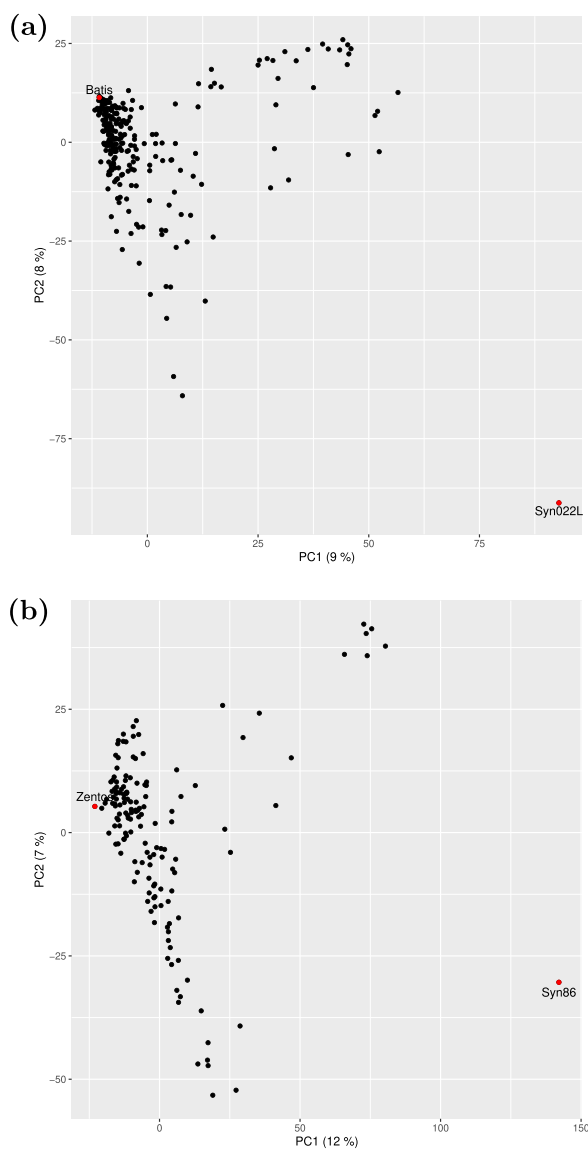


Figure 1 Population structure plot for the wheat B22 (A) and Z86 (B) populations. This scatter plot presents the first two principal components (PC1 and PC2) and the corresponding parent genotypes are shown in red color. Additionally, the proportion of variance explained the first two PCs are also provided.

In addition, heading date and grain threshability were primary criterion behind selecting the favorable genotypes during the development of the populations (Kunert et al. 2007). All these factors might have resulted in the skewed relationship toward the corresponding recurrent parent. Thus, it is important to correct for this skewed relationship, while identifying the marker-trait association in this population.

First, we used the single-locus model with the polygenic effect (estimated from the marker information) to identify the marker-trait associations. In the B22 population, we used the trait TGW and identified two significant QTL [at chosen level of false discovery rate (FDR) = 0.05]. FDR correction (Storey and Tibshirani 2003) is commonly used to control the rate of false discoveries in QTL mapping studies (Devlin et al. 2003; Nelson et al. 2017; Marees et al. 2018). The Manhattan plot along with the name of significant markers are shown in Figure 2A. Then, we also identified the

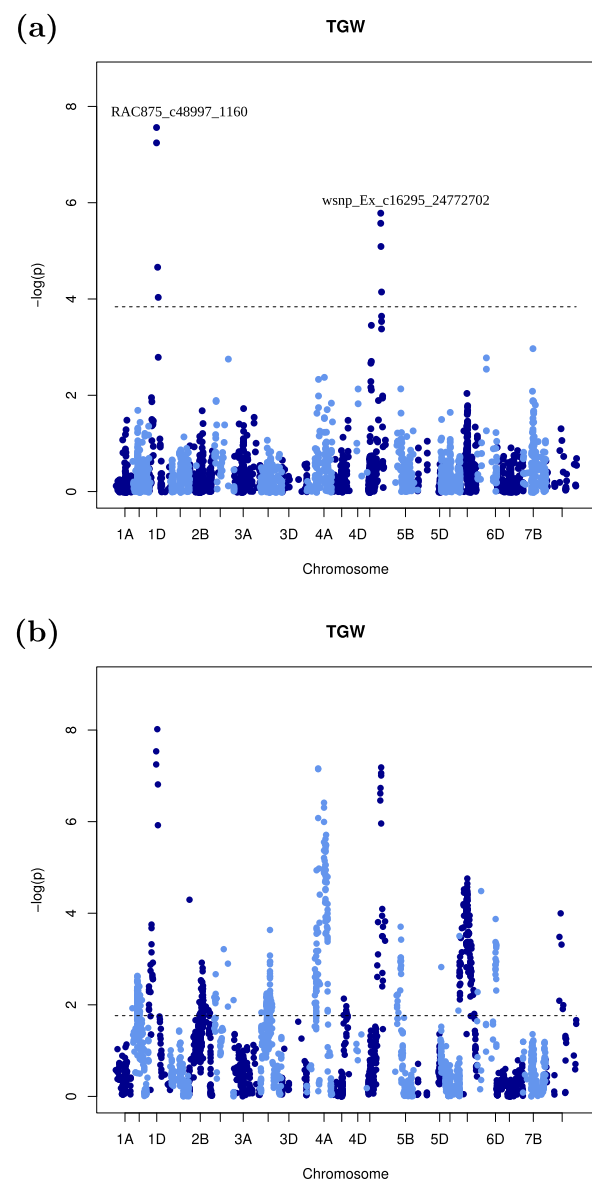


Figure 2 Manhattan plot based on the single-locus model having the polygenic effect in the model (A) and without the polygenic effect (B), using the B22 population for the trait TGW. The dashed line corresponds to an FDR rate of 0.05. Additionally, names of the significant markers (RAC875_c48997_1160, wsnp_Ex_c16295_24772702) are also shown.

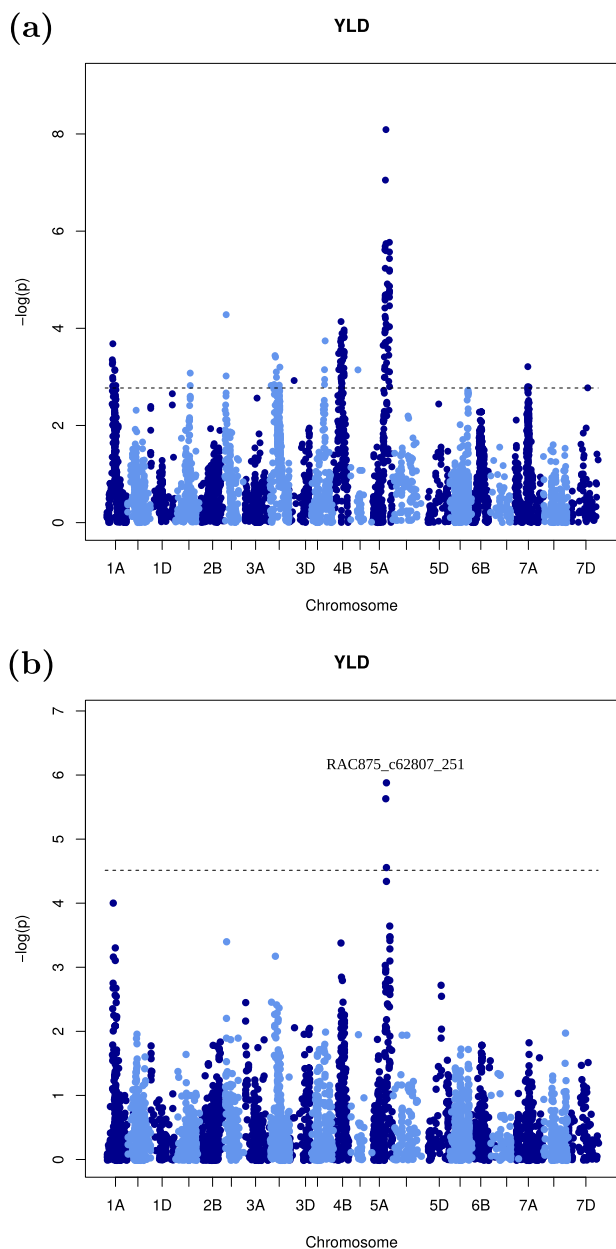


Figure 3 Manhattan plot based on the single-locus model without the polygenic effect (A) and having the polygenic effect in the model (B), using the Z86 population for the trait yield (YLD). The dashed line corresponds to an FDR rate of 0.05. Additionally, names of the significant markers (RAC875_c62807_251) are also shown.

marker-trait association for the trait TGW using the single-locus model without the polygenic effect and the Manhattan plot is shown in Figure 2B. The corresponding analysis based on the Z86 population is shown in Figure 3. From Figures 2 and 3, one can see that single-locus model with the polygenic effect can significantly reduce the number of false positives by correcting for the sample structure in the data set. The quantile-quantile (Q-Q) plot (which is the graphical representation of the proportion of significant markers compared to the expected number of significant SNPs based on P -values) is a commonly used method in GWAS studies based on SNP analyses to monitor the number of false positives (Balding 2006). Spurious Q-Q plot inflation occurs when the population structure/cryptic relatedness is not taken into account in the analysis. Thus, we also

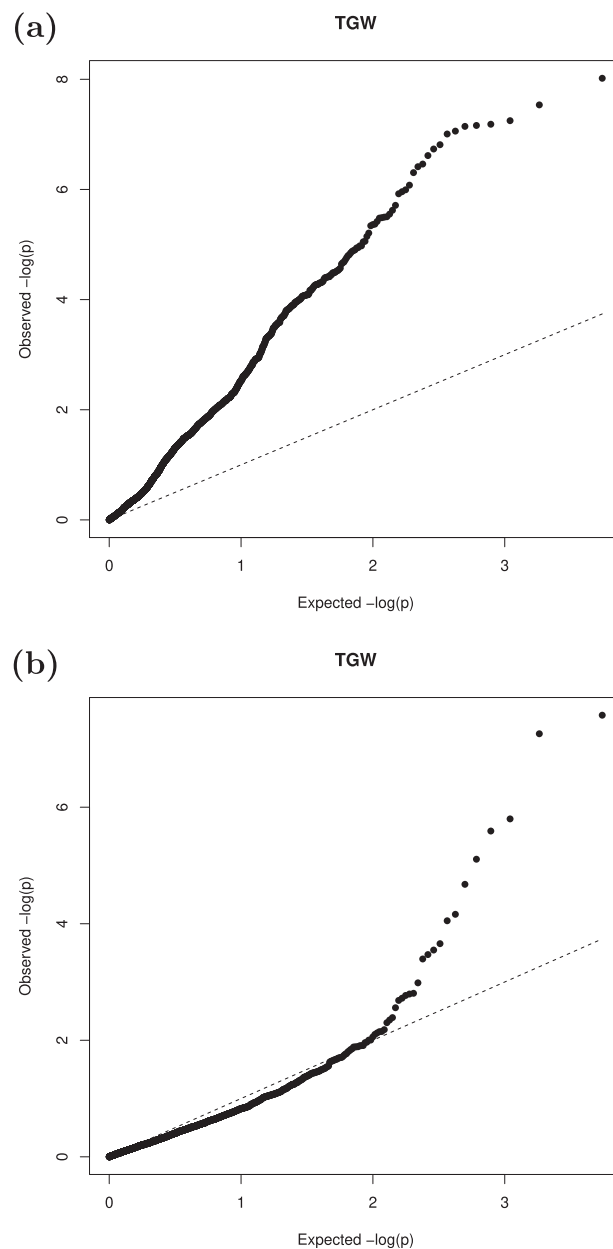


Figure 4 Q-Q plots based on the single-locus model without the polygenic effect (A) and with the polygenic effect in the model (B) using the B22 population for the trait TGW.

checked the behavior of Q-Q plot with both the real data sets to assess the number of false positives and the plots are shown in Figures 4 and 5 for the B22 and Z86 populations, respectively. From Figures 4 and 5, it is very clear that single-locus model with polygenic effect is able to control the false-positive association signals by effectively correcting the population substructure in the data set.

For TGW, the most significant SNP marker RAC875_c48997_1160 on chromosome 1D underlie a Dihydrolipoyl dehydrogenase 1 (TraesCS1D02G067500) gene carrying FAD/NAD(P)-binding domain, which involved in oxidation-reduction process in plants (Timm et al. 2015). The second SNP marker wsnp_Ex_c16295_24772702 on chromosome 5A was located at a transmembrane protein-related (TraesCS5A02G485700) gene of uncharacterized molecular function. In addition, the chromosomal 5A revealed association with trait YLD, where the most

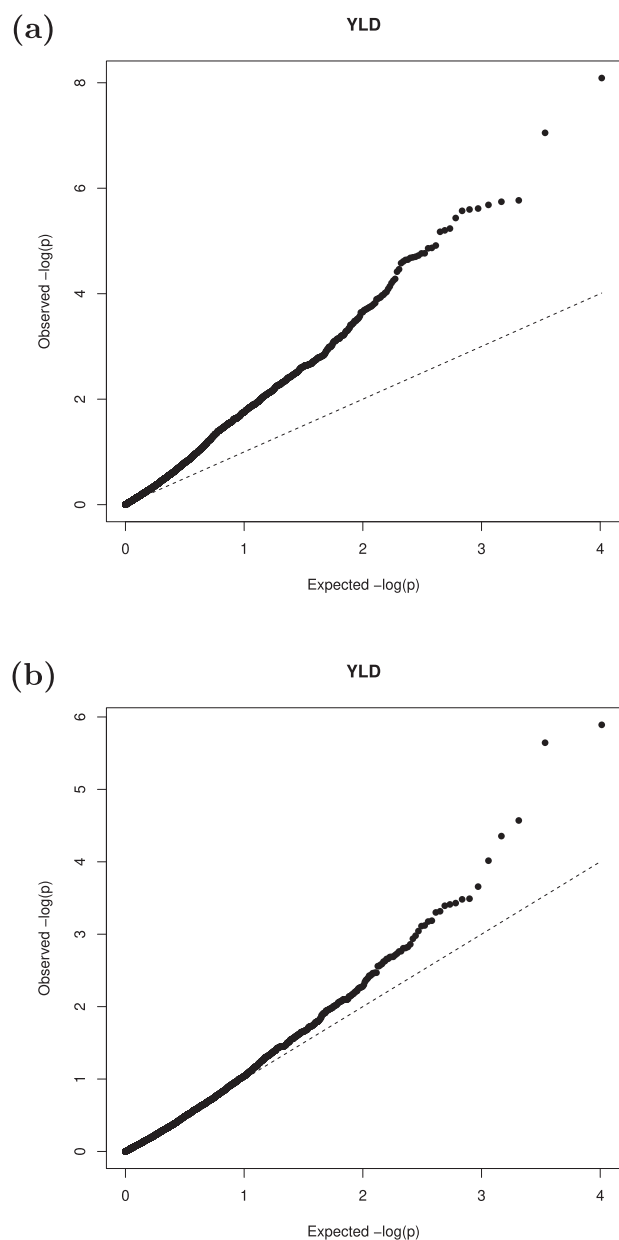


Figure 5 Q-Q plots based on the single-locus model without the polygenic effect (A) and with the polygenic effect in the model (B) using the Z86 population for the trait yield (YLD).

significant marker RAC875_c62807_251 was located at TraesCS5A02G447800 gene, which belongs to a family of hypothetical protein abundantly present in cereal species. The candidate genes predictions are based on the location of associated SNP markers and further work is needed to validate their function using high-resolution recombination analysis. It is important to mention that the designed AB population does not offer a high-resolution candidate gene analysis. Alternatively, genome-editing method like CRISPR/Cas system can be employed for the functional characterization of identified candidates.

Interval mapping is commonly used to identify QTL in experimental crosses. Thus, we also identified the QTL in both populations with interval mapping approach using the R-package R/qtl (Broman *et al.* 2003). Logarithm of the odds (LOD) curve based on

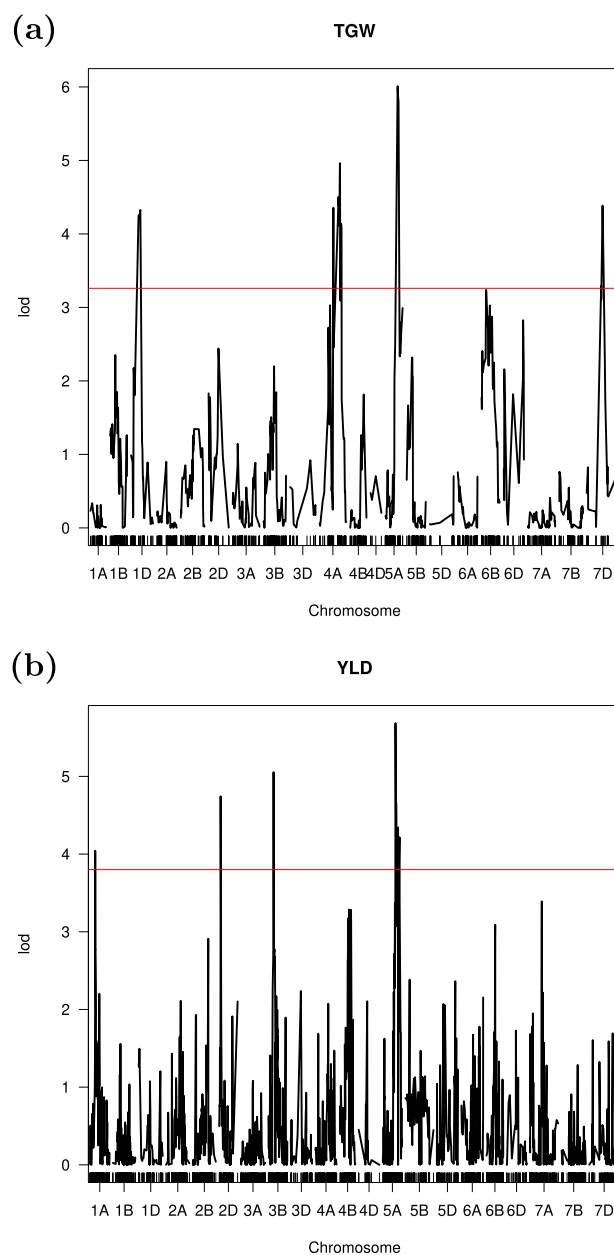


Figure 6 LOD scores based on the interval mapping approach using the population B22 for the trait TGW (a) and Z86 for the trait YLD (b). The red line shows an LOD threshold for the significant QTL determined with r/qtl using 1000 permutations at $P=0.05$.

the interval mapping approach using B22 and Z86 populations is shown in Figure 6. The LOD curve in Figure 6 confirms that the interval mapping approach is not able to correct for the population structure in AB populations.

We also identified the marker-trait association using the Bayesian multi-locus model. The significant marker effects identified using the Bayesian multi-locus model is shown in Figure 7. For the estimation of the posterior mean of marker effects, we used five different MCMC chains, each having length 50,000 iterations with a burn-in period of 10,000 iterations and averaged over five different MCMC chains. Based on the plot it can be concluded that Bayesian whole-genome regressions can estimate marker effects while accounting for polygenicity at the same time by regressing on the entire markers simultaneously.

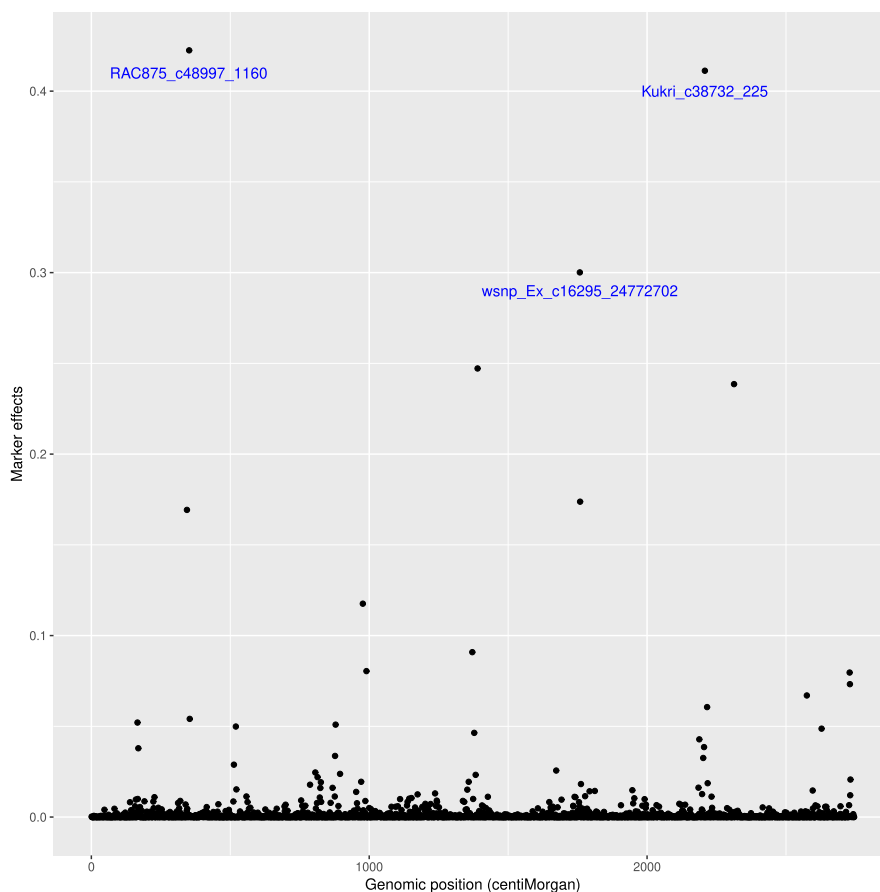


Figure 7 Marker effects estimated for the trait thousand grain yield with the Bayesian multi-locus association model plotted against the corresponding markers in the wheat BC2F3 population. Additionally, names of the significant markers (RAC875_c48997_1160, wsnp_Ex_c16295_24772702, Kukri_c38732_225) are also shown.

Marker-trait association using simulated data sets

We also identified the significant marker-trait association with the two simulated data sets using the single- and multi-locus models. The Manhattan plots based on Simulation 1 data set with single-locus model without the polygenic effect in the model is shown in [Figure 8A](#), whereas the plot based on the single-locus model having the polygenic effect is shown in [Figure 8B](#). Both models were able to identify the true simulated QTL; however, the number of false positives was effectively controlled by the single-locus model with the polygenic effect in the model ([Figure 8B](#)). The Manhattan plots based on Simulation 2 data set (here we also considered another 100 markers with small effects along with the five main effect QTL) using the single-locus model without and with the polygenic effect, is shown in [Figure 9, A and B](#), respectively. From [Figure 9B](#), it can be seen that there is an increase in the actual false positives when the marker-trait association was identified using the single-locus model without the polygenic effect in the model. The corresponding Q-Q plots for the Simulation 1 and 2 data sets are provided as the Supplementary material (S3 and S4). We also identified the QTL in simulated data set (Simulation 1) using the interval mapping approach and provided similar results like the single-locus model without the polygenic effect. The corresponding plot is provided as the Supplementary material (S5).

The significant marker effects identified using the Bayesian multi-locus model with Simulation 1 and Simulation 2 data sets

are shown in [Figure 10, A and B](#), respectively. Finally, we also plotted the (ROC) curve to visualize the estimation accuracy of single-locus and Bayesian multi-locus model based on 50 simulation replicates, shown in [Figure 11](#). Traditional ROC curve is plotted based on the true-positive and false-positive rates; however, in this study, we used the average number of true and false positives identified with the 50 simulation replicates to plot the ROC curve. Based on [Figure 11](#), one can conclude that Bayesian multi-locus model is an efficient method to correct for the population structure and polygenic effect in mapping studies.

Discussion

During the past decade, many statistical methods have been developed to correct for the population substructure and relatedness in mapping studies using association panel. More recently, many studies ([Pascual et al. 2015](#); [Wei and Xu 2016](#); [Sul et al. 2018](#)) pointed out that it is important to include polygenic term while perform mapping studies in multi-parent populations (which is a compromise between biparental population and association panel) and population based on mixture of multiple strains. In this study, we showed the importance to account for genetic substructure while identifying the marker-trait association in AB populations. We believe that it was especially important here that the covariance structure (genomic relationship matrix) of the polygenic effect was calculated from the molecular markers, because the genomic relationship matrix is able to provide more

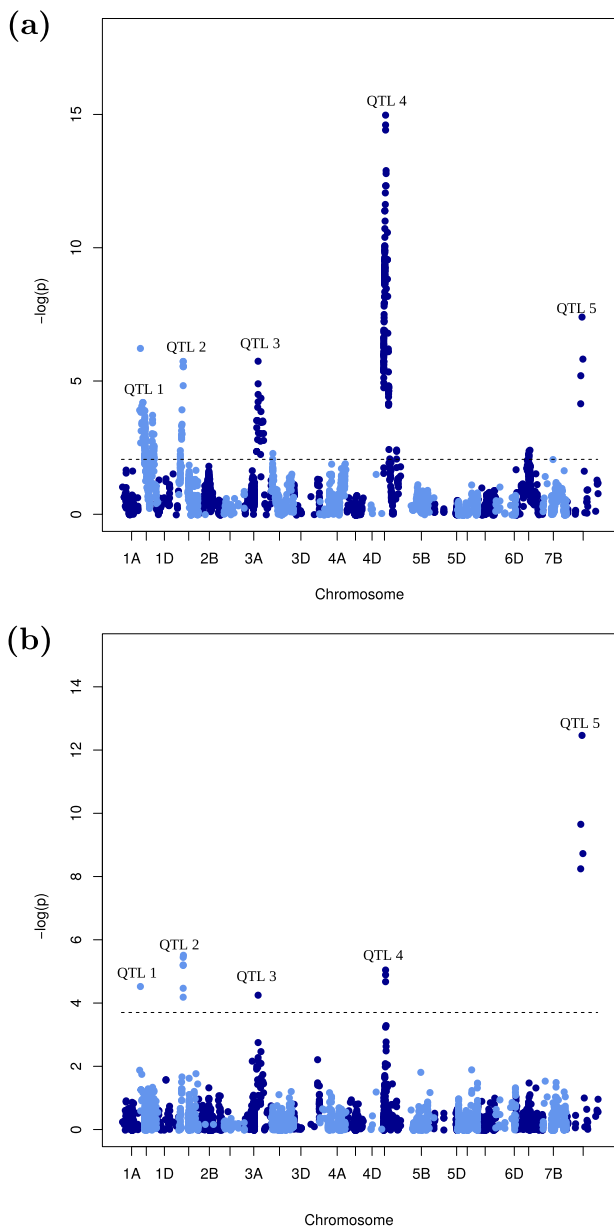


Figure 8 Manhattan plot based on Simulation 1 using single-locus model without the polygenic effect (A) and with the polygenic effect in the model (B), additionally, the simulated QTL are also shown. The dashed line corresponds to an FDR rate of 0.05

precise information on the proportion of genome shared by the relatives than the conventional pedigree-based relationship matrix. To our knowledge, this is the first study to point out the importance of accounting for genomic relationships in QTL mapping with AB populations.

The single-locus model considers one marker at a time and tests for the marker-trait association, which is computationally less intensive. But the single-locus model is known to have some limitations. One of the main drawbacks is that the effect of a single marker may be quite small in a single-locus model, but might have strong joint effects, and by estimating all marker effects simultaneously, which will increase the statistical power to detect their joint activity. Bayesian multi-locus model is an efficient solution to this problem, because Bayesian multi-locus model

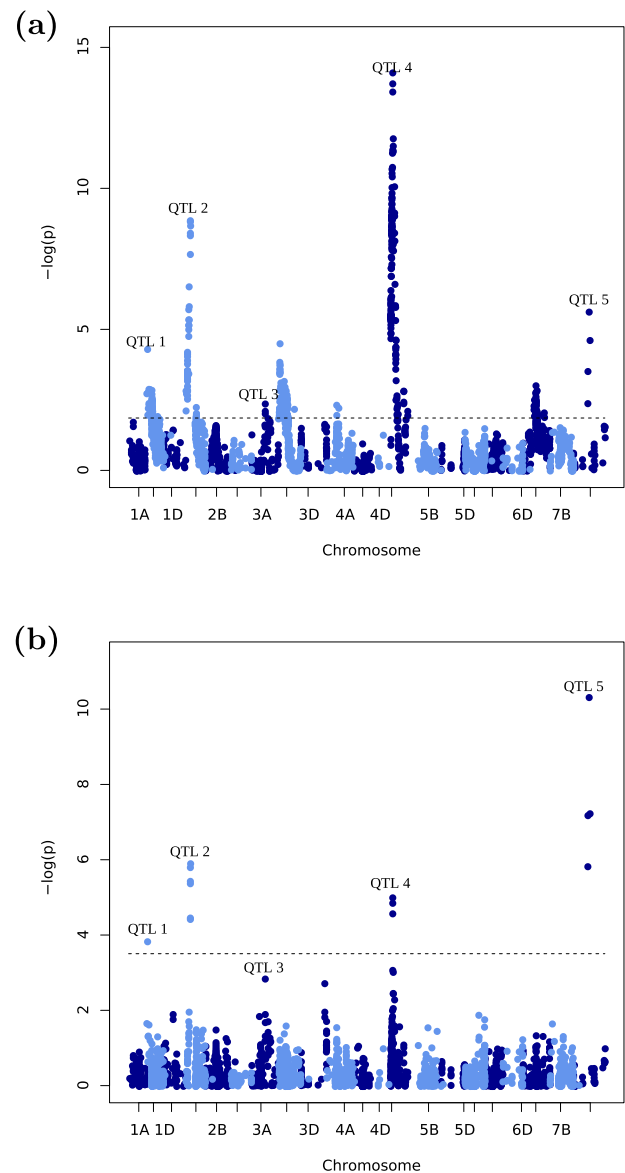


Figure 9 Manhattan plots based on Simulation 2 using single-locus model without the polygenic effect (A) and with the polygenic effect in the model (B), additionally, the simulated QTL are also shown. The dashed line corresponds to an FDR rate of 0.05

jointly estimates all marker effects. In our real data analysis, using multi-locus model, we identified one significant marker, which was not identified by the single-locus model and we believe that this is likely due to the better statistical power gained by the joint estimation of all marker effects using the Bayesian multi-locus model.

In the context of association mapping, many studies already reported the capability of Bayesian multi-locus model to automatically correct the confounding due to substructure in Bayesian (Iwata *et al.* 2007, 2009; Pikkuhookana and Sillanpää 2009; Kärkkäinen and Sillanpää 2012) and non-Bayesian (Würschum and Kraft 2015) framework. However, this is the first study to show the effectiveness of Bayesian multi-locus model to control the false positives by accounting for the substructure in an AB population.

A permutation test proposed by Churchill and Doerge (1994) and Xu (2003) for frequentist and Bayesian analyses, respectively,

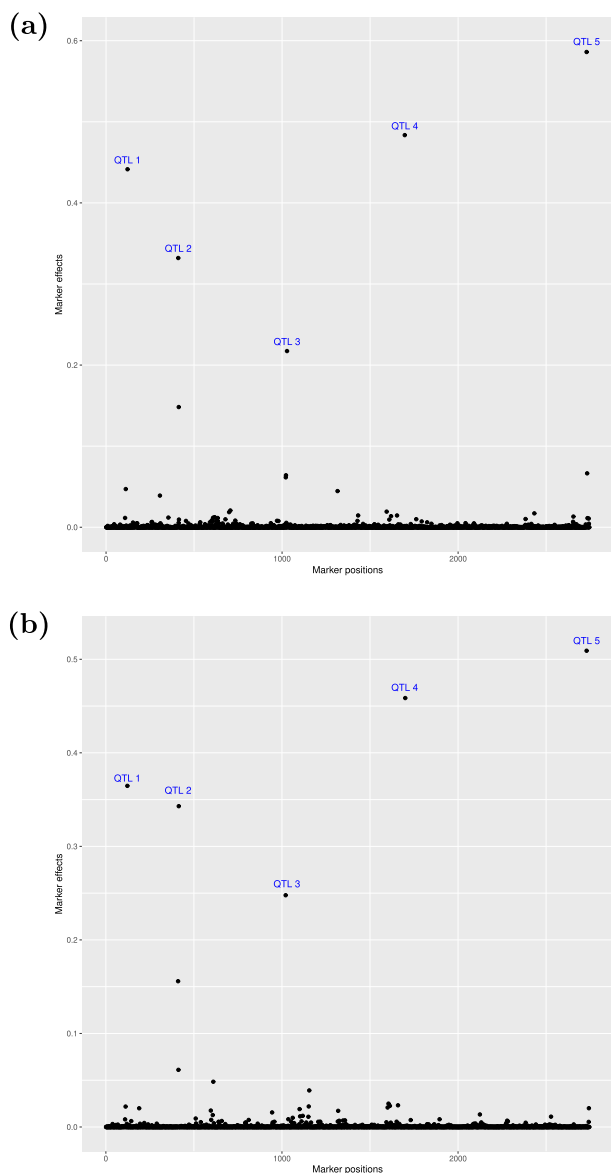


Figure 10 Marker effects estimated with the Bayesian multi-locus association model plotted against the corresponding markers in Simulation 1 (A) and Simulation 2 (B), additionally, the simulated QTL are also shown.

is commonly used in mapping studies to identify the significant threshold for marker-trait association with single- and multi-locus model. However, permutation test with Bayesian multi-locus model is sensitive to the collinearity in the marker data (Mathew et al. 2018), so we decided to consider only the SNPs, which appear repetitively in all separate analyses (in all five MCMC chains) as the significant ones. In this calculation, all SNPs within the given window were considered to be the same. For each chromosome, the window size was determined based on LD-plot (see Supplementary material). Running many MCMC chain is still computationally challenging, as a complementary alternative one can identify the significance level (at chosen level of FDR = 0.05) based on a single-locus model and use this significance level as a base line to identify the significant markers in a multi-locus association model.

In QTL mapping studies involving multi-environment data, it is essential to account for both environmental main effects and

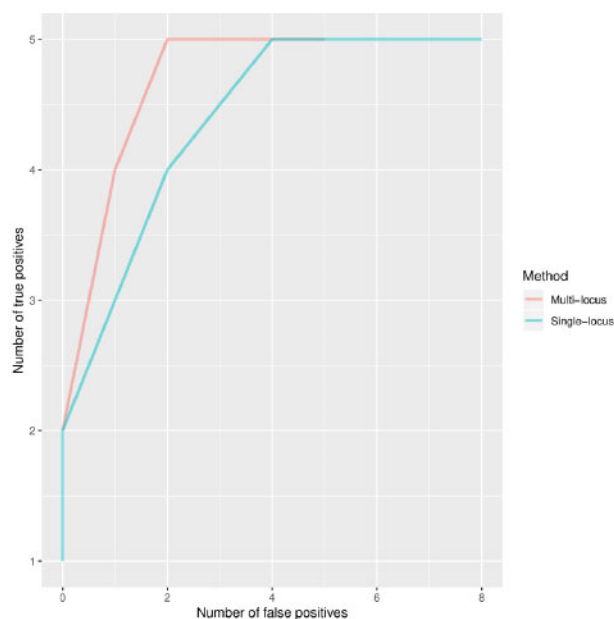


Figure 11 ROC curve based on average of 50 simulation replicates for the multi-locus and single-locus model.

environment (GxE) interactions because the phenotypes can be influenced by different environmental conditions. In the present study, we were unable to account for environmental main effects and GxE interaction effects in the QTL mapping model due to the lack of data availability of individual environments. Thus, QTL reported in this study might be biased due to the influence by environments. Nevertheless, the aim of the present study was to test the effect of genetic substructure caused by allele frequencies on the outcome of QTL analysis. Our results suggested that QTL mapping studies using experimental populations such as AB populations, it is important to account for genetic substructure in the population. We believe this outcome remain true for the QTL analysis models with or without GxE interactions.

Interval mapping is seen as historical method of the time when marker maps were very sparse. Genome-wide mapping with high-density marker information can improve the precision of QTL localization along with the detection of small- and medium-sized QTL (Stange et al. 2013). Interval mapping is believed to be robust to genetic substructure unlike genome-wide association mapping; however, our results suggest that interval mapping is suffering from allele frequency deviations in the population. Thus, our results suggest it might be a good practice to use GWAS model, which account for genetic substructure and allele frequency deviations in mapping studies with experimental populations using genome-wide marker information.

AB populations (where an exotic donor parent crossed to an adapted recurrent parent) have been successful to identify beneficial alleles in several crops, such as tomato (Fulop et al. 2016), rice (Thomson et al. 2003), wheat (Narasimhamoorthy et al. 2006), maize (Ho et al. 2002), cotton (Wang et al. 2017a), and barley (Pillen et al. 2003). These studies mainly relied on regression methods to identify the significant QTL. However, the controlled crossing scheme along with the phenotypic abnormalities caused by factors such as hybrid necrosis (Bomblies and Weigel 2007), hybrid sterility (Ouyang et al. 2010), hybrid lethality (Garner et al. 2016), and the selection of agronomically adapted traits during the development of these lines will lead to unbalanced allele

frequencies in an AB population. In this study, we showed the importance to correct for this during the QTL mapping in an AB population using the linear mixed model with a random polygenic effect. Even though, in this study, we showed the importance of correction in an AB population, it might be a good practice to use the polygenic term in a QTL mapping model to correct for the polygenicity/substructure, when the progenies in other experimental biparental populations such as F₂, DH, or RIL have experienced the deviation from the expected genotype ratios due to both natural and artificial selection. Finally, in this study, for the single-locus marker-trait association, we used the R package rrBLUP and for the Bayesian multi-locus association estimation, we used the available code used in the study by Mathew *et al.* (2018). However, for the Bayesian multi-locus association one can also use the R packages such as BGLR (Pérez and de los Campos 2014) and VIGOR (Onogi and Iwata 2016).

B.M., J.L., M.J.S., and A.A.N. were involved in the conception and design of the study. B.M. performed the statistical analyses, simulation and drafted the manuscript. K.P. and S.D. were involved in the phenotyping work for the population B22 and Z8, respectively. B.M., J.L., S.D., K.P., M.J.S., and A.A.N. participated in the interpretation of the results. All the authors critically revised the manuscript.

Acknowledgments

We are thankful to the cooperation of plant breeders, Dr. E. Kazman (Saatzucht Josef Breun), Dr. J. Schacht (Limagrain-Nickerson), Dr. E. Ebmeyer (Lochow-Petkus), and Dr. A. Spanakakis (Fr. Strube Saatzucht), and their teams for carrying out the field experiments. We are also grateful to the editor and two anonymous reviewers as well as Karin Voitl for their suggestions and comments which helped us to improve our manuscript.

Funding

The phenotyping and genotyping works were funded by the German Plant Genome Research Initiative (GABI) of the Federal Ministry of Education and Research (BMBF, project 312862) and by the Federal Ministry of Agriculture and Nutrition (Grant # IdMaRo-100203349), respectively.

Conflicts of interest: The authors declare that they have no conflict of interest.

Literature cited

- Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nat Rev Genet.* 7:781–791.
- Bauer AM, Hoti F, Korff MV, Pillen K, Léon J, *et al.* 2009. Advanced backcross-QTL analysis in spring barley (*H. vulgare* ssp. *spontaneum*) comparing a REML versus a Bayesian model in multi-environmental field trials. *Theor Appl Genet.* 119:105–123.
- Bomblies K, Weigel D. 2007. Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nat Rev Genet.* 8: 382–393.
- Broman KW, Wu H, Sen S, Churchill GA. 2003. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics.* 19:889–890.
- Churchill GA, Doerge RW. 1994. Empirical threshold values for quantitative trait mapping. *Genetics.* 138:963–971.
- Devlin B, Roeder K, Wasserman L. 2003. False discovery or missed discovery? *Heredity.* 91:537–538.
- Endelman JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome.* 4:250–255.
- Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognit Lett.* 27:861–874.
- Fulop D, Ranjan A, Ofner I, Covington MF, Chitwood DH, *et al.* 2016. A new advanced backcross tomato population enables high resolution leaf QTL mapping and gene identification. *G3 (Bethesda).* 6: 3169–3184.
- Garner AG, Kenney AM, Fishman L, Sweigart AL. 2016. Genetic loci with parent-of-origin effects cause hybrid seed lethality in crosses between *mimulus* species. *New Phytol.* 211:319–331.
- Gogel B, Smith A, Cullis B. 2018. Comparison of a one-and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. *Euphytica.* 214:44.
- Grandillo S, Tanksley SD. 2005. Advanced backcross QTL analysis: results and perspectives. The wake of the double helix: from the green revolution to the gene revolution. *edizioni Avenue Media, Italy,* p. 115–132.
- Ho J, McCouch S, Smith M. 2002. Improvement of hybrid yield by advanced backcross QTL analysis in elite maize. *Theor Appl Genet.* 105:440–448.
- Huang X, Sang T, Zhao Q, Feng Q, Zhao Y, *et al.* 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet.* 42:961–967.
- Iwata H, Ebana K, Fukuoka S, Jannink J-L, Hayashi T. 2009. Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. *Theor Appl Genet.* 118:865–880.
- Iwata H, Uga Y, Yoshioka Y, Ebana K, and, Hayashi T. 2007. Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. *Theor Appl Genet.* 114:1437–1449.
- Jolliffe I. 2003. Principal component analysis. *Technometrics.* 45:276.
- Kang HM, Ye C, Eskin E. 2008a. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics.* 180:1909–1925.
- Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, *et al.* 2008b. Efficient control of population structure in model organism association mapping. *Genetics.* 178:1709–1723.
- Kärkkäinen HP, Sillanpää MJ. 2012. Robustness of Bayesian multilocus association models to cryptic relatedness. *Ann Hum Genet.* 76:510–523.
- Kunert A, Naz A, Dedeck O, Pillen K, Léon J. 2007. AB-QTL analysis in winter wheat: I. detection of favorable exotic alleles for baking quality traits introgressed from synthetic hexaploid wheat (*T. turgidum* ssp. *Dicoccoides* 9 *T. tauschii*). *Theor Appl Genet.* 115:683–695.
- Li Y-H, Reif JC, Hong H-L, Li H-H, Liu Z-X, *et al.* 2018. Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. *Plant Sci.* 266: 95–101.
- Listgarten J, Kadie C, Schadt EE, and, Heckerman D. 2010. Correction for hidden confounders in the genetic analysis of gene expression. *Proc Natl Acad Sci U S A.* 107:16465–16470.
- Ma J, Amos CI. 2012. Principal components analysis of population admixture. *PLoS One.* 7:e40115.
- Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, *et al.* 2018. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int J Methods Psychiatr Res.* 27:e1608.
- Mathew B, Léon J, Sannemann W, Sillanpää MJ. 2018. Detection of epistasis for flowering time using Bayesian multilocus estimation in a barley MAGIC population. *Genetics.* 208:525–536.
- McVean G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet.* 5:e1000686.

- Meuwissen T, Hayes B, and Goddard M. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157:1819–1829.
- Mora F, Quiral YA, Matus I, Russell J, Waugh R, et al. 2016. SNP-based QTL mapping of 15 complex traits in barley under rain-fed and well-watered conditions by a mixed modeling approach. *Front Plant Sci*. 7:909.
- Morrell PL, Buckler ES, Ross-Ibarra J. 2011. Crop genomics: advances and applications. *Nat Rev Genet*. 13:85–96.
- Nagata K, Ando T, Nonoue Y, Mizubayashi T, Kitazawa N, et al. 2015. Advanced backcross QTL analysis reveals complicated genetic control of rice grain shape in a *japonica* × *indica* cross. *Breed Sci*. 65:308–318.
- Narasimhamoorthy B, Gill B, Fritz A, Nelson J, Brown-Guedira G. 2006. Advanced backcross QTL analysis of a hard winter wheat × synthetic wheat population. *Theor Appl Genet*. 112:787–796.
- Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, et al.; EPIC-CVD Consortium. 2017. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat Genet*. 49:1385–1391.
- Onogi A, Iwata H. 2016. VIGOR: variational Bayesian inference for genome-wide regression. *J Open Res Softw*. 4:e11.
- Ouyang Y, Liu Y-G, Zhang Q. 2010. Hybrid sterility in plant: stories from rice. *Curr Opin Plant Biol*. 13:186–192.
- Parks BW, Nam E, Org E, Kostem E, Norheim F, et al. 2013. Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell Metab*. 17:141–152.
- Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L, et al. 2015. Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnol J*. 13:565–577.
- Pérez P, and de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. 198:483–495.
- Pikkuhookana P, and Sillanpää MJ. 2009. Correcting for relatedness in Bayesian models for genomic data association analysis. *Heredity*. 103:223–237.
- Pillen K, Zacharias A, Léon J. 2003. Advanced backcross QTL analysis in barley (*Hordeum vulgare* L.). *Theor Appl Genet*. 107:340–352.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 38:904–909.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. *Am J Hum Genet*. 67:170–181.
- Sillanpää MJ. 2011. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity*. 106:511–519.
- Stange M, Utz HF, Schrag TA, Melchinger AE, Würschum T. 2013. High-density genotyping: an overkill for qtl mapping? lessons learned from a case study in maize and simulations. *Theor Appl Genet*. 126:2563–2574.
- Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 100:9440–9445.
- Sul JH, Martin LS, and Eskin E. 2018. Population structure in genetic studies: confounding factors and mixed models. *PLoS Genet*. 14:e1007309.
- Tanksley S, Nelson J. 1996. Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theor Appl Genet*. 92:191–203.
- Taylor J, Verbyla A. 2011. R package wgamim: QTL analysis in bi-parental populations using linear mixed models. *J Stat Softw*. 40:1–18.
- Thomson M, Tai T, McClung A, Lai X, Hinga M, et al. 2003. Mapping quantitative trait loci for yield, yield components and morphological traits in an advanced backcross population between *Oryza rufipogon* and the *Oryza sativa* cultivar Jefferson. *Theor Appl Genet*. 107:479–493.
- Timm S, Wittmiß M, Gamlien S, Ewald R, Florian A, et al. 2015. Mitochondrial dihydrolipoyl dehydrogenase activity shapes photosynthesis and photorespiration of *Arabidopsis thaliana*. *Plant Cell*. 27:1968–1984.
- Wang B, Chee PW. 2010. Application of advanced backcross quantitative trait locus (QTL) analysis in crop improvement. *J Plant Breed Crop Sci*. 2:221–232.
- Wang B, Draye X, Zhuang Z, Zhang Z, Liu M, et al. 2017a. QTL analysis of cotton fiber length in advanced backcross populations derived from a cross between *Gossypium hirsutum* and *G. mustelinum*. *Theor Appl Genet*. 130:1297–1308.
- Wang B, Zhuang Z, Zhang Z, Draye X, Shuang L-S, et al. 2017b. Advanced backcross QTL analysis of fiber strength and fineness in a cross between *Gossypium hirsutum* and *G. mustelinum*. *Front Plant Sci*. 8:1848.
- Wei J, Xu S. 2016. A random-model approach to QTL mapping in multiparent advanced generation intercross (MAGIC) populations. *Genetics*. 202:471–486.
- Wen Y-J, Zhang Y-W, Zhang J, Feng J-Y, Dunwell JM, et al. 2019. An efficient multi-locus mixed model framework for the detection of small and linked QTLs in F2. *Brief Bioinform*. 20:1913–1924.
- Wimmer V, Albrecht T, Auinger H-J, Schön C-C. 2012. synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*. 28:2086–2087.
- Würschum T, Kraft T. 2015. Evaluation of multi-locus models for genome-wide association studies: a case study in sugar beet. *Heredity*. 114:281–290.
- Xu S. 2003. Estimating polygenic effects using markers of the entire genome. *Genetics*. 163:789–801.
- Xu S. 2008. Quantitative trait locus mapping can benefit from segregation distortion. *Genetics*. 180:2201–2208.
- Yano K, Yamamoto E, Aya K, Takeuchi H, Lo P-C, et al. 2016. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat Genet*. 48:927–934.
- Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*. 38:203–208.
- Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, et al. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun*. 2:467.