ORIGINAL ARTICLE

An innovative procedure of genome-wide association analysis fits studies on germplasm population and plant breeding

Jianbo He[1] · Shan Meng[1] · Tuanjie Zhao[2,4] · Guangnan Xing[2,4] · Shouping Yang[3,4] · Yan Li[3,4] · Rongzhan Guan[4,5] · Jiangjie Lu[1] · Yufeng Wang[1] · Qiuju Xia[6] · Bing Yang[6] · Junyi Gai[1,2,3,4,5]

Received: 27 January 2017 / Accepted: 1 August 2017

Abstract

Key message **The innovative RTM-GWAS procedure provides a relatively thorough detection of QTL and their multiple alleles for germplasm population characterization, gene network identification, and genomic selection strategy innovation in plant breeding.**

Abstract The previous genome-wide association studies (GWAS) have been concentrated on finding a handful of major quantitative trait loci (QTL), but plant breeders are interested in revealing the whole-genome QTL-allele constitution in breeding materials/germplasm (in which tremendous historical allelic variation has been accumulated) for genome-wide improvement. To match this requirement, two innovations were suggested for GWAS: first grouping tightly linked sequential SNPs into linkage disequilibrium blocks (SNPLDBs) to form markers with multi-allelic haplotypes, and second utilizing two-stage association analysis for QTL identification, where the markers were preselected by single-locus model followed by multi-locus multi-allele model stepwise regression. Our proposed GWAS procedure is characterized as a novel restricted two-stage multi-locus multi-allele GWAS (RTM-GWAS, https://github.com/njau-sri/rtm-gwas). The Chinese soybean germplasm population (CSGP) composed of 1024 accessions with 36,952 SNPLDBs (generated from 145,558 SNPs, with reduced linkage disequilibrium decay distance) was used to demonstrate the power and efficiency of RTM-GWAS. Using the CSGP marker information, simulation studies demonstrated that RTM-GWAS achieved the highest QTL detection power and efficiency compared with the previous procedures, especially under large sample size and high trait heritability conditions. A relatively thorough detection of QTL with their multiple alleles was achieved by RTM-GWAS compared with the linear mixed model method on 100-seed weight in CSGP. A QTL-allele matrix (402 alleles of 139 QTL × 1024 accessions) was established as a compact form of the population genetic constitution. The 100-seed weight QTL-allele matrix was used for genetic characterization, candidate gene prediction, and genomic selection for optimal crosses in the germplasm population.

Communicated by Dr. Mikko J. Sillanpaa.

Jianbo He and Shan Meng have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-017-2962-9) contains supplementary material, which is available to authorized users.

✉ Junyi Gai
sri@njau.edu.cn

[1] Soybean Research Institute, Nanjing Agricultural University, Nanjing 210095, China

[2] National Center for Soybean Improvement, Ministry of Agriculture, Nanjing 210095, China

[3] Key Laboratory of Biology and Genetic Improvement of Soybean (General), Ministry of Agriculture, Nanjing 210095, China

[4] State Key Laboratory for Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China

[5] Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing 210095, China

[6] State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China

Published online: 21 August 2017

Introduction

Green revolution is a great success of conventional plant breeding in which the basic breeding procedure includes two major steps: the first is to choose parental materials from germplasm or breeding materials to design optimal crosses, and the second is to select the best progenies for further testing in the segregating generations. As the crosses determine the potential of progeny selection, optimal cross design is a key of the conventional breeding. The concept of "Breeding by Design" through designing parental crosses based on quantitative trait loci (QTL) was proposed for direct genotypic selection for potential recombination in plant breeding (Peleman and van der Voort 2003). Meanwhile, genomic selection (GS) focusing mainly on progeny selection was proposed as a marker-assisted selection procedure for animal breeding with no need of QTL information but using a training population to establish the relationship between whole-genome markers and the phenotype values termed genomic estimated breeding value (GEBV) as selection criterion (Meuwissen et al. 2001). Although plant scientists also consider using GS for both cross selection and progeny selection in plant breeding (Heffner et al. 2009; Mohammadi et al. 2015), the GS approach for animal breeding cannot be readily applied to complex plant breeding (Jonas and de Koning 2013). The difficulties of using the classic GS approaches in plants lie in the uncertain accuracy of an indirect black box relationship, inadequate estimation of recombination potential, and high genotyping costs in large amount of progeny materials (Desta and Ortiz 2014). Although some efforts have been made to improve the accuracy of genomic prediction (De Coninck et al. 2016; Jiang and Reif 2015; Vazquez et al. 2016), in plant conventional breeding, following the "Breeding by Design" concept, GS based on whole-genome QTL-allele constitution seems to be a potential approach to both optimal crosses and superior progenies. This GS approach can be termed as QTL-allele-based GS, which in fact is a direct genotype selection method.

To utilize QTL-allele-based GS efficiently for both the optimal design of crosses and the selection of superior progenies in plant breeding, a relatively thorough detection of the whole-genome QTL-allele constitution in a germplasm/breeding population is an essential requirement. Plant germplasm is a genetic reservoir from which improved varieties were developed. The QTL-allele constitution of all the derived breeding populations can be inferred from that of the germplasm population. And if the genetic constitution of all breeding target traits in the germplasm has been explored, the GS can be utilized in both plant breeding stages, that means at the first stage with GS for optimal cross design based on the genetic structure information of the parental materials/germplasm and at the second stage with GS for elite progenies using a set of markers associated with target traits.

The genome-wide association study (GWAS) was found to be a potential approach for detecting whole-genome QTL with multiple alleles in a large natural population, especially in plant germplasm population, based on linkage disequilibrium (LD) (Nordborg and Weigel 2008; Huang and Han 2014). However, GWAS has been suffered from a high false-positive rate caused by unknown population structure, which could interfere in the QTL detection. The population structure bias may be caused by the admixture of heterogeneous populations or the inbreeding-caused relatedness among the materials, or mainly a mixture of them (Devlin and Roeder 1999; Campbell et al. 2005; Voight and Pritchard 2005). The structured association analysis (SA) using model-based clustering and principal components analysis (PCA) of the marker covariance matrix are two widely used methods to correct admixture bias (Pritchard et al. 2000; Price et al. 2006). The approaches based on linear mixed models (LMM) that incorporate kinship matrix were proposed to correct population structure bias from both admixture and relatedness (Yu et al. 2006; Kang et al. 2008, 2010; Zhang et al. 2010), and the LMM method had been the preferred statistical method for GWAS in plants (Atwell et al. 2010; Huang et al. 2010; Jia et al. 2013; Li et al. 2013; Morris et al. 2013; Dhanapal et al. 2015). The widely used SA, PCA, and LMM methods in many studies performed association tests individually on a single marker basis (single-locus model), where the accumulated contribution of the detected QTL may be inflated obviously due to the correlations among neighboring loci, and, therefore, leading to the overflowing heritability problem under single-locus model. Therefore, it is more reasonable to explicitly include multiple loci in the statistical model (Zeng 1994). Recently, statistical methods based on multi-locus model under the framework of LMM have been also proposed for GWAS (Segura et al. 2012; Rakitsch et al. 2013; Wang et al. 2016), but they have not been widely used due to lack of user-friendly computer software and the time-consuming computations for large-scale GWAS.

However, there are still some difficulties in using the previously reported GWAS procedures for characterization of germplasm (or breeding populations) in plants. The first is that the bi-allelic SNP marker could not match the multi-allele property of traits in the germplasm materials. The second is high false-negative rate or the missing heritability problem which causes the large gap between the detected QTL contribution and the total genetic contribution (the overall heritability) due to the stringent experiment-wise significance level, such as the Bonferroni correction. For example, only an average of five loci was identified in GWAS of 41 traits in rice, accounting for about 22% of the total phenotypic variation (Huang et al. 2010; Zhao et al. 2011). Therefore, a relatively thorough detection of genome-wide QTL is required to provide the full information of the genetic constitutions of the germplasm population. The third is high false-positive rate or the overflowing heritability problem which causes the inflation (even much more than 100%) of the total phenotypic variation explained by detected QTL under single-locus model. In addition, the difficulties also lie in how to choose the optimal crosses based on the genetic structure of the germplasm/breeding materials and how to reduce the genotyping cost for the large number of progenies at each selection generation in plant breeding.

In the present study, two innovations were suggested for a relatively thorough QTL-allele detection required in germplasm/breeding population study and genome-wide selection in breeding programs: first grouping tightly linked sequential SNPs into LD blocks (SNPLDB) to generate multi-allelic haplotypes, and second performing two-stage association analysis for QTL identification, with a single-locus model pre-selection of markers followed by multi-locus multi-allele model stepwise regression for QTL identification. Accordingly, a novel restricted two-stage multi-locus multi-allele GWAS procedure (RTM-GWAS) was assembled and then its effectiveness was demonstrated using simulation experiments in comparison with a set of previous GWAS procedures. The usability of RTM-GWAS was further demonstrated with a case study on 100-seed weight of Chinese soybean germplasm population (CSGP), from which the established QTL-allele matrix was used in characterizing the germplasm population, annotating candidate genes, and applying genomic selection for optimal crosses.

Materials and methods

Plant materials and field experiments

The Chinese soybean germplasm population (CSGP) consists of 1024 soybean accessions, including wild soybean (*Glycine soja* Sieb. & Zucc.) and cultivated soybean [*Glycine max* (L.) Merr.], which was sampled from the germplasm storage at the National Center for Soybean Improvement, Nanjing, China. The wild soybean part consists of 203 annual wild accessions (WA), and the cultivated soybean part consists of 375 farmers' landraces (LR) and 446 released cultivars (RC). The genetic relationship among the CSGP accessions was presented in Fig. S1. Theoretically, this population contains a wide range of genetic variation accumulated during the soybean domestication and breeding history in China, and therefore, the QTL-allele information obtained from the CSGP may be passed onto its derived breeding materials. In fact, the use of the germplasm from multiple sources covering WA, LR, and RC of the soybean is a challenge to the GWAS procedure to be established.

The CSGP accessions were planted in a randomized complete block design experiment with three replications, at Jiangpu Experimental Station of Nanjing Agricultural University, Nanjing, China, (32°07°N, 118°62°E), in 2010, 2011, and 2012. A modification was made for the experiment design that the wild accessions and cultivated accessions were planted in two separate sub-blocks in each replication/block because of the different plot sizes between wild and cultivated soybeans, i.e., 1.0×1.0 m^2 hill plots for WA and 0.8×0.8 m^2 hill plot for LR and RC due to the large and vining plant of WA.

After the thinning at V3 (third node) stage, eight and five plants were kept in each hill plot for cultivated soybean and wild soybean, respectively. To hold the twining stem of the wild soybeans upward, a bamboo stick was used in each hill plot starting at V6 (sixth node stage). The field management including weed control and fertilizer application was conducted normally. The matured seeds were harvested and dried under 35–40 °C till a constant weight. The 100-seed weight (g) was measured five times per plot for the whole experiment.

Statistical analysis of the phenotypic data

The analysis of variance (ANOVA) of randomized block design under multiple years/environments was used for 100-seed weight of the 1024 accessions, and here, we assume that the effects of the sub-blocks and plot sizes for WA and LR/RC on 100-seed weight are small and may be compensated under multiple environments. The mixed-effects model was used for ANOVA, and the phenotypic value for *i*th environment (year), *j*th block nested in *i*th environment, and *k*th genotype is expressed as $y_{ijk} = \mu + e_i + r_{j(i)} + g_k + (ge)_{ik} + \varepsilon_{ijk}$, where μ is the overall mean, and the environment and genotype-by-environment interaction effects were considered as random effects. The heritability of 100-seed weight was estimated as $\hat{h}^2 = \hat{\sigma}_g^2/[\hat{\sigma}_g^2 + \hat{\sigma}_{ge}^2/s + \hat{\sigma}^2/(s \cdot r)]$, where $\hat{\sigma}_g^2$, $\hat{\sigma}_{ge}^2$, and $\hat{\sigma}^2$ are estimated variances of genotype, genotype-by-environment, and random error, respectively, and s is the number of environments and r is the number of replications (Hanson et al. 1956). The variance components were estimated using the REML method of PROC VARCOMP in SAS/STAT software (SAS Institute Inc., Cary, NC, USA).

SNP genotyping

The RAD-Seq (restriction site-associated DNA sequencing) was used for SNP genotyping for all the 1,024 soybean accessions in the present study, which was done at BGI Tech, Shenzhen, China. The genomic DNA was extracted from 6–8 to 8–10 bulked leaves (V4–V5 vegetative growth stage) of *G. max* and *G. soja* per accession according to the hexadecyltrimethylammonium bromide (CTAB) method (Murray and Thompson 1980). Restriction site-associated DNA (RAD) libraries were created as previously described (Baird et al. 2008). The barcoded adapters were ligated to each sample. The barcodes were 6 nucleotides (nt) long and differed from each other by at least 2 nt to facilitate unambiguous identification of each specimen following sequencing. The specimens were pooled and sheared, and the 400–700 base pair (bp) size fractions were purified from an agarose gel after electrophoresis. Following ligation of a second adaptor, the fragments containing both adapters were PCR-amplified (12 cycles) and the 400–700 bp size fraction was isolated as described above. The RAD libraries were sequenced on an Illumina HiSeq 2000 instrument through multiplexed shotgun genotyping method (Andolfatto et al. 2011).

All sequence reads were aligned against the genome of Williams 82 using SOAP2 (Li et al. 2009; Schmutz et al. 2010). The RealSFS was used for SNP calling at population level based on the Bayesian estimation of site frequency. For imputation quantity control, the resulted SNPs were pruned with a maximum missing and heterozygous allele call rate of 20% and a minimum minor allele frequency (MAF) of 1% (if a third allelic phenomenon appeared only in one individual, it was treated as missing allele) in the population. Then, missing genotype data were imputed using software fastPHASE (Scheet and Stephens 2006) and SNPs with MAF < 1% were excluded.

Simulation studies for key issues in GWAS

Simulation study on LD decay for different inbreeding coefficients

A total of 13 populations with different inbreeding coefficient were simulated using forward-time population simulation package simuPOP (Peng and Kimmel 2005). The proportion of individuals under self-mating schemes (equivalent to inbreeding coefficient) was set as 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, and 1.0, respectively. The population size was set to 10,000. Each population was initialized from an F_1 hybrid and then evolved for 100 generations with ten replications. The LD between loci with varying recombination fractions was recorded. For evaluating LD decay distance, the recombination fraction was approximately transformed to genetic distance and then to physical distance by assuming that 1 cM is approximately equivalent to 400 kb according to the soybean genetic and sequence map at SoyBase (http://soybase.org).

Simulation study on marker–trait association power for different QTL heritabilities

The theoretical power of marker-trait association test for different LD and heritability levels was calculated using the PROC POWER in SAS/STAT software (SAS Institute Inc., Cary, NC, USA) based on tests for simple correlation coefficient between marker and trait with a sample size of 1000. The correlation between marker and trait under additive genetic model can be measured as $\sqrt{r^2h^2}$, where r^2 is the LD between the typed marker and causal locus, and h^2 is the heritability of the causal locus (Weir 2008).

Design and establishment of the RTM-GWAS procedure

SNPLDB marker construction

Since the SNPs usually distribute unevenly on the genome, the tight and loose linkage between neighboring SNPs leads to a block-like structure of genomic sequences which may transmit together to the offspring. The different combinations of linked SNPs in a block could be classified as multiple haplotypes/alleles. This should overcome the low polymorphism and redundant QTL detection of the bi-allelic SNP markers and theoretically more effective and powerful than SNP markers in GWAS. Accordingly, the multi-allelic marker-type termed SNP LD block (SNPLDB) was suggested as genomic markers for plant germplasm population.

Genomic blocks were defined first using the block partitioning approach with confidence intervals based on genome-wide LD (D') pattern (Gabriel et al. 2002). Then, the SNPs within an LD block were grouped into an SNPLDB marker with multiple haplotypes as its alleles. The genotype for each individual was determined by its corresponding SNP haplotype at each locus. To control the minor allele frequency for further statistical analysis, the haplotypes with extremely low frequency (such as <0.01) were approximately replaced with the most similar haplotype in a same SNPLDB. The similarity between two haplotypes was defined as the proportion of SNP sites in identity-by-state. Any individual SNP that could not be grouped into blocks was treated as an individual SNPLDB marker. Therefore, there are two types of SNPLDB marker, the SNPLDB with multiple SNPs and the SNPLDB with a single SNP. The LD between the multi-allelic SNPLDBs was calculated as the weighted average of LD values for all allele combinations (Farnir et al. 2000).

Construction of the genetic similarity coefficient matrix using SNPLDB markers

Since the marker-based genetic relationship matrix used for correction of the population structure bias in the previous methods (Patterson et al. 2006; Price et al. 2006; VanRaden 2008) is suitable for the bi-allelic SNPs but not for the proposed multi-allelic SNPLDBs, here, we suggest a more flexible genetic similarity coefficient (GSC) matrix based on SNPLDBs to estimate the comprehensive population structure. The GSC between two individuals is defined as the proportion of loci that are in identity-by-state, $s_{ij} = \sum_{k=1}^{m} c_{ijk}/(2m)$, where c_{ijk} is the number of common alleles (taking a value of 0, 1, or 2) at kth SNPLDB between ith and jth individual, and m is the total number of SNPLDBs. In spite of the population structure varies uncertainly due to variable admixture components, inbreeding schemes, and even both, the GSC matrix may be used as a general approach to estimate the varied population structure biases without any requirement of pre-set assumptions. In practice, the top J eigenvectors with largest eigenvalues of the GSC matrix calculated from genome-wide SNPLDBs were suggested to correct the population structure bias in association analysis based on linear model. The population structure effect here is considered as fixed effect rather than random effect, since the population is pre-determined rather than randomly formed, and therefore, the GSC correction for population structure should be directly on the population.

Two-stage multi-locus multi-allele association analysis

In general, hundreds of thousands or even millions of molecular markers are used in GWAS, but in fact, most of them may be irrelevant to the target trait. To effectively reduce the huge model space for an efficient multiple QTL modeling, we suggest using a two-stage strategy to perform GWAS. Here, the individuals are all assumed homozygous at each locus for simplicity; otherwise, a genotypic genetic model can be implemented. At the first stage, a single-locus association test is performed to eliminate redundant SNPLDBs, and the linear model is expressed as:

$$y_i = \mu + \sum_{j=1}^{J} w_{ij}\alpha_j + \sum_{l=1}^{L} x_{il}\beta_l + \varepsilon_i, \tag{1}$$

where y_i is the observed phenotypic value of the ith individual and μ is the overall mean; w_{ij} is the coefficient of the jth eigenvector of the SNPLDB GSC matrix of the ith individual, and a_j is the effect of the jth eigenvector of the SNPLDB GSC matrix, while J is the number of eigenvectors chosen for population structure correction; x_{il} is a value of 0 or 1, indicating the allele state of the lth allele (L is the total number of alleles) at the SNPLDB under testing for the ith individual and β_l is the effect of the lth allele; ε_i is the residual effect assumed to be normally distributed.

At the second stage, a multi-locus model extended from Eq. (1) is built based on the candidate SNPLDBs selected at the first stage,

$$y_i = \mu + \sum_{j=1}^{J} w_{ij}\alpha_j + \sum_{k=1}^{K}\sum_{l=1}^{L_k} x_{ikl}\beta_{kl} + \varepsilon_i, \tag{2}$$

where x_{ikl} is a value of 0 or 1, indicating the allele state of the lth allele at the kth locus for the ith individual, and β_{kl} is the effect of the lth allele at the kth locus, while L_k is the total number of alleles at the kth locus. Other terms and parameters are the same as in Eq. (1).

Equation (1) can be solved using regression analysis, and Eq. (2) can be solved efficiently using stepwise regression featured with forward selection and backward elimination. The normal significance level 0.05 or 0.01 is recommended for the F test under single-locus model to pre-select candidate markers, and also for multi-locus model of the stepwise regression approach as the built-in control for the experiment-wise error rate. Since the QTL detection is carried out at the second stage under multi-locus model, the total genetic variation explained by detected QTL will be less than the total genetic variation among accessions or the heritability of the trait.

RTM-GWAS

The two-stage multi-locus multi-allele association analysis combined with using SNPLDB markers is defined as restricted two-stage multi-locus multi-allele genome-wide association study procedure, coded as RTM-GWAS. The RTM-GWAS is characterized with the following key points: (1) SNPLDB genome markers with multiple alleles; (2) two-stage association analyses, i.e., association analysis under single-locus model for pre-selection of markers and stepwise regression analysis under multi-locus model for genome-wide QTL-allele detection, with total genetic contribution limited to the overall heritability value; (3) population structure correction using GSC matrix calculated from multi-allelic markers; and (4) normal built-in experiment-wise significance criterion.

Simulation comparisons among different GWAS methods

The SNPLDB markers of the real CSGP and a simulated ideal population were used for simulation comparisons. The ideal population was simulated from the real SNPLDB markers of the CSGP, in which the genotype data were randomly shuffled across whole genome. The population structure is assumed to be eliminated in the ideal population, and therefore, the ideal population provided a reference for other populations to be compared. To simulate the QTL detection, a total of 100 SNPLDBs were randomly sampled as the causal loci, and their associated effects were drawn from an exponential distribution with a rate of 1. The random error was drawn from the normal distribution with a scaled variance to fix the trait heritability to 0.9 (about the similar in the present study for 100-seed weight under replicated experiments). The phenotype value for each individual was obtained as the sum of overall genotypic value plus a random error.

In addition, to examine the performance of different GWAS methods under six different architectures, i.e., two QTL number levels (10 and 100) by three trait heritability levels (0.2, 0.5, and 0.9), were also simulated based on the real genotype data of CSGP using the same simulation method as above. To assess the influence of sample size on RTM-GWAS, four populations with different sample size, i.e., 200, 400, 600, and 800, were randomly sampled from the ideal population and used for simulations under the 100-locus model with a trait heritability of 0.9.

The RTM-GWAS method was compared to the Naïve, PCA, LMM, and the haplotype-based association test (HAT) methods. The Naïve method performs the association test based on the simple linear model without control of population structure. The PCA method performs the association test based on a general linear model, in which the top 10 eigenvectors of SNPLDB GSC matrix are incorporated as covariates to correct for population structure. The LMM performs the association test based on a linear mixed model, in which the SNPLDB GSC matrix is incorporated as covariance structure of the random effect for population structure correction. In the present LMM analysis, the EMMAX algorithm (Kang et al. 2010) was used. For performing the HAT method, the --hap-assoc command in the PLINK software was used (Purcell et al. 2007). As the --hap-assoc command can include only one covariate, the phenotypic values were adjusted through regression analysis to leave out the effects of eigenvectors, and then, the adjusted phenotypic values were used in HAT. In this way, the HAT procedure can keep consistent with other GWAS procedures which use the top 10 eigenvectors of SNPLDB GSC matrix for population structure correction. A uniform Bonferroni-adjusted significance level of 0.05 was used for Naïve, PCA, LMM, HAT, and the second stage of RTM-GWAS, while a threshold of 0.05 was used for the first stage of RTM-GWAS. To evaluate the influence of significance level on GWAS, nine fixed significance levels without Bonferroni correction were also used in comparisons among the five GWAS methods, including the significance levels of 1e−10, 1e−9, 1e−8, 1e−7, 1e−6, 1e−5, 1e−4, 1e−3, and 1e−2, respectively.

In computing the detection efficiency or false discovery rate (FDR) and detection power, an associated marker was considered as a false positive if none of the causal loci were found within a 100-kb window centered at the causal locus; otherwise, it was considered as a true positive. A causal locus was counted as a detected locus if at least one marker was found significantly associated within a 100-kb window centered at the causal locus. A total of 100 independent replications were performed to get an average of the detection power and FDR. For comparisons among the different GWAS methods, the following indicators were calculated: Power, overall detection power of the 100 simulated causal loci; PVE, phenotypic variation explained by detected causal loci; FDR, false discovery rate; RPO, relative detection power, i.e., Power × (1 − FDR); RPV, relative PVE, i.e., PVE × (1 − FDR).

GWAS of 100-seed weight in the CSGP

The genetic structure (neighbor-joining tree) of the population was estimated based on the GSC matrix calculated from genome-wide SNPLDBs. The pairwise distance matrix was calculated as one minus the SNPLDB GSC matrix built from all the SNPLDBs. Then, the PHYLIP 3.6 software (Felsenstein 1989) was used to construct the neighbor-joining tree. Based on it, the GWAS of 100-seed weight of the CSGP was performed using the RTM-GWAS procedure established above.

Gene annotation and candidate gene selection

The gene system conferring soybean 100-seed weight from the detected QTL system was established as follows: at first, the annotated genes were searched within the interval (with a 30-kb flanking expansion) of the associated SNPLDBs. Then, to identify the candidate genes for the trait, the Chi-square tests were performed to test the linkages between the detected SNPLDB and the SNPs within each annotated gene. The tests were conducted for all the SNPs in an annotated gene. The significance level was set to 0.05. The gene calls and annotations were retrieved from *G. max* (version Wm82.a1.v1.1) SoyBase (http://soybase.org).

Genetic differentiation analysis of the population

The analysis of molecular variance (AMOVA) for WA–LR–RC subpopulation variation was conducted using the Arlequin software (Excoffier and Lischer 2010) for the complete SNPLDB data set and for the detected 100-seed weight QTL/marker data set, respectively. Chi-square test was used to test the independence of allele frequency distribution among subpopulation for each QTL. F_{ST} was estimated for all SNPLDBs using a sliding window of 500-kb interval centered on each SNPLDB.

Genomic selection for optimal crosses

For optimal cross selection in plant conventional breeding, all possible single crosses (523,776) were generated in silico, each with 2000 homozygous progenies derived randomly from an F_1 hybrid through continuously selfing. For linkage model, the number of crossovers on a chromosome is simulated according to the Poisson process with a rate parameter of λ under no interference, where λ is the length of chromosome in Morgan. For independent assortment model, all loci were assumed independent from each other and the alleles of different loci assorted independently. The genotypic value of a progeny was predicted as the sum of QTL genotypic values according to QTL-allele matrix. The predicted phenotypic value of a progeny derived from a cross between parental line i and j was calculated as $y_{ij} = g_{ij} + (y_i - g_i + y_j - g_j)/2$, where g_{ij} is the predicted genotypic value of the progeny, y_i and y_j are the observed phenotypic values of the two parental lines, respectively, g_i and g_j are the predicted genotypic values of the two parental lines, respectively, and the predicted genotypic value was calculated as the sum of QTL-allele effects. Accordingly, all the predicted values of the possible crosses were obtained for optimal cross selection based on genomic QTL-allele information.

Results

Key issues in GWAS of plant germplasm population

The theoretical base of using GWAS in detecting QTL in a natural population is the tight linkage disequilibrium (LD) between a marker and the involved QTL. In self-pollinated plants and the inbred lines of cross-pollinated plants, the inbreeding coefficient (IBC) is greater than 0.95, and it is about 0.50–0.95 in often cross-pollinated plants. The LD decay distance in populations with different IBCs was simulated (Fig. S2A). In general, the population with higher inbreeding exhibits slower LD decay, especially for IBC = 0.9–1.0. When IBC is less than 0.8, the LD starts to decay rapidly, but may still extend for a long distance. The estimated LD decay distances (Table S1) for IBC = 0.9–1.0 (self-pollinated plants) increase rapidly in comparison with those for IBC less than 0.8 (540–3822 kb vs. 306 kb for $r^2 = 0.5$, and 1260–8919 kb vs. 714 kb for $r^2 = 0.3$). Therefore, a key issue of GWAS in self-pollinated plants, such as soybean, is to cope with the inbreeding-caused population structure bias for a shortened LD decay distance and an effective QTL detection (here admixture being considered as the cause secondary to inbreeding for the increased LD decay distance).

The statistical power of GWAS in detecting QTL with varying heritability (h^2) at an experiment-wise significance level of 5×10^{-7} (Bonferroni adjustment for 100,000 markers) was simulated (Fig. S2B). The QTL with higher heritability have higher statistical power to be detected in GWAS. For example, at $r^2 = 0.5$, a QTL with $h^2 = 7\%$ or larger can have a statistical power more than 0.8 to be detected, while for a small effect or a small h^2 QTL, the power drops down quickly ($h^2 = 1\%$ with power less than 0.1). The GWAS can keep a relatively high power for QTL with high heritability even when the LD is not high; this means that in GWAS of populations with high inbreeding coefficient, the markers far away from the QTL can be significantly associated, but in fact as a noise rather than a useful locus. For QTL with low heritability, high power of GWAS requires very high LD and the detection power of a small genetic contribution QTL is much less than that of a large genetic contribution QTL. Therefore, the fluctuation of QTL detection will more likely happen for small contribution QTL than for large contribution QTL, especially in populations with increased inbreeding. Therefore, another key issue of GWAS in plants is to ensure that both large and small contribution QTL can be detected for understanding the entire genetic architecture of the trait.

The previous GWAS strategy in plants was mainly based on individual SNP association test. However, in fact, due to the historical mutation, recombination, and introgression, multiple alleles widely accumulated on each locus in natural populations. The bi-allelic SNP marker in the previous GWAS could not represent the multi-allelic nature of a QTL. The geneticists are interested in detecting QTL/genes, while the breeders are more likely interested in finding the best alleles on multi-allelic loci. Therefore, the third key issue of GWAS in plants is to detect the entire QTL-alleles with their effects estimated.

Properties of the SNPLDB marker

The SNPLDB markers in the Chinese soybean germplasm population (CSGP) were studied based on the genotyping by sequencing through RAD-seq (restriction site-associated DNA sequencing). A total of 145,558 high-quality SNPs with minor allele frequency (MAF) >0.01 were identified (Table S2). Genome-wide LD patterns from SNPs indicated that the LD was high with the maximum of half decay distance about 500 kb for r^2 (Fig. 1a, b). Accordingly, a total of 36,952 SNPLDBs were constructed with 2–14 haplotypes/alleles per SNPLDB marker (Fig. 1c; Table S2). Among the SNPLDBs, 70.3% of them are composed of a single SNP termed as S.SNPLDB, while 29.7% are built from multiple SNPs termed as M.SNPLDB (Fig. 1c). A great part of the

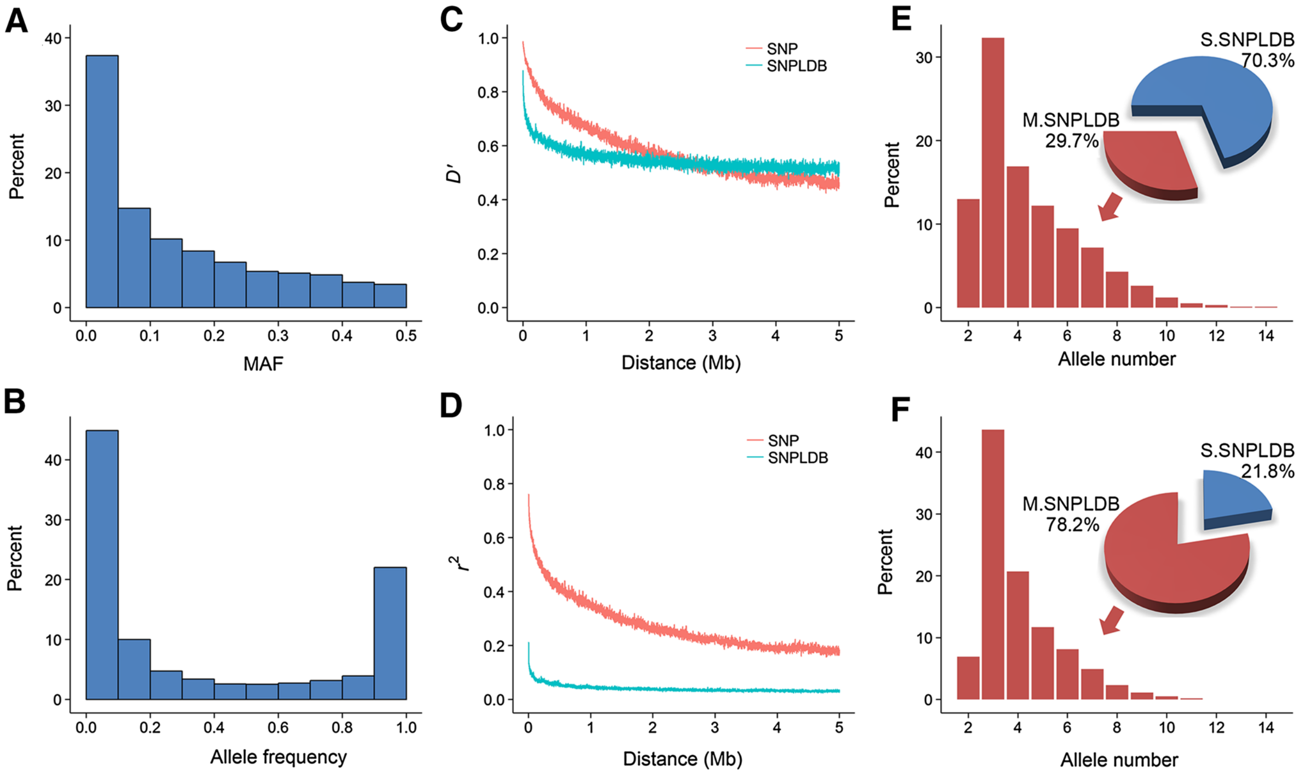

Fig. 1 Characterization of genome-wide SNP and SNPLDB. **a** Minor allele frequency of SNP in CSGP. **b** Allele frequency of SNPLDB in CSGP. **c** Decay of D' in CSGP. **d** Decay of r^2 in CSGP. **e** Allele number of SNPLDB in CSGP. **f** Allele number of SNPLDB in 302 resequenced soybean accessions population

SNPLDBs (74.2%) in CSGP have only two alleles. However, along with the increased coverage of the sequencing, the S.SNPLDBs would be very likely merged into LD blocks with multiple SNPs, as shown in Fig. 1d, where up to 78.2% of the SNPLDBs were M.SNPLDBs from the soybean resequencing genotype data by Zhou et al. (2015). The LD decay distance calculated from the SNPLDB marker data was shorter than those from the original SNP data. The distance at which D' decays to 0.6 was about 2000–3000 kb for SNPs but 200–500 kb for SNPLDBs (Table S3). Thus, the multi-allelic SNPLDB marker with shortened LD decay distance could improve the power of GWAS.

Simulation demonstration of RTM-GWAS in terms of detection efficiency and power

Simulations were performed to assess the detection efficiency and power of RTM-GWAS based on the SNPLDB data of CSGP. An ideal population without structure bias was simulated using the real CSGP SNPLDB data in which the alleles on each locus were randomly shuffled among the 1024 accessions across the whole genome. This ideal population was used as a reference compared with the CSGP. The detection efficiency (in terms of FDR) and power (in terms of detected QTL and PVE) were evaluated, and then, their combined indicators, relative detection power (RPO), and relative phenotypic variation explained by detected causal loci (RPV) were used as the comprehensive merit of the GWAS procedure. The RTM-GWAS method was compared to the PCA, LMM, and HAT methods, but SNPLDB GSC matrix rather than SNP covariance matrix was used for RTM-GWAS, PCA, LMM, and HAT. A uniform Bonferroni-adjusted significance level of 0.05 was used as the threshold to detect significant QTL for all methods, even it is not necessary for RTM-GWAS (because Bonferroni correction was proposed for an experiment-wise test criterion in a single-locus model analysis, while the built-in experiment-wise criterion is already in the multi-locus model analysis). The simulation results of the 100-locus model with a trait heritability of 0.9 (Table 1) indicated that RTM-GWAS outperformed PCA, LMM, and HAT in both CSGP and the simulated ideal population. However, all methods encountered a high false discovery rate (FDR, generally >0.3) in CSGP, but the RTM-GWAS method exhibited the lowest FDR, about 17% less than that of LMM with twofold detection power increase. In the simulated ideal population, all the single-locus model methods, i.e., Naive, PCA, LMM, and HAT, performed similarly as expected, while RTM-GWAS showed

Table 1 Performance comparison of the association analysis methods in 100-locus model simulations under a trait heritability of 0.9

Simulation	Method	Power	PVE	FDR	RPO	RPV	Time[a]
CSGP	Naïve	0.685	0.779	0.979	0.014	0.016	3.5 min
	PCA	0.164	0.587	0.915	0.014	0.050	12.7 min
	LMM	0.080	0.456	0.524	0.038	0.217	58.2 min
	HAT	0.095	0.455	0.878	0.012	0.056	2.7 h
	RTM	0.175	0.615	0.357	0.112	0.395	13.0 h
Ideal	Naïve	0.101	0.556	0.007	0.101	0.552	3.8 min
	PCA	0.099	0.551	0.008	0.099	0.547	12.4 min
	LMM	0.098	0.549	0.004	0.098	0.547	58.6 min
	HAT	0.098	0.543	0.012	0.096	0.537	2.7 h
	RTM	0.303	0.804	0.010	0.300	0.796	3.1 h

Power, overall detection power of the 100 simulated causal loci; PVE, phenotypic variation explained by detected causal loci; FDR, false discovery rate; RPO, relative detection power [Power × (1 − FDR)]; RPV, relative PVE [PVE × (1 − FDR)]. Naïve, association test based on simple linear model without control for population structure; PCA, association test based on linear model, and the top 10 eigenvectors with largest eigenvalues of SNPLDB genetic similarity coefficient (GSC) matrix were incorporated as covariates to correct for population structure; LMM, association test based on linear mixed model, and SNPLDB GSC matrix was incorporated as covariance structure of random effect to correct for population structure; HAT, the haplotype-based association test implemented in PLINK software (--hap-assoc); RTM, association test using the proposed restricted two-stage multi-locus multi-allele GWAS method (RTM-GWAS). A uniform Bonferroni-adjusted significance level of 0.05 was used for Naïve, PCA, LMM, HAT, and the second stage of RTM-GWAS, while a threshold of 0.05 was used for the first stage of RTM-GWAS. The CSGP simulation was based on Chinese soybean germplasm population, comprising 1024 individuals, genotyped at 36,952 SNPLDBs; the ideal simulation was based on an ideal population simulated from real SNPLDB genotype data of CSGP, in which the genotype data were randomly shuffled across whole genome

[a] All computing was performed on a single core of an Intel Xeon E5-2670 2.60 GHz CPU, and the time is the total computing time elapsed for the 100 replications

threefold increase in detection power. Furthermore, although the RTM-GWAS method showed a notable improvement in detection power, there was still a big gap between the variation explained by the detected QTL and total phenotypic variation. This indicated that the Bonferroni correction used in RTM-GWAS is too stringent, since the extra Bonferroni correction was added to the built-in experiment-wise error control, which caused a redundant error control. The power and FDR under different significance levels without Bonferroni correction were evaluated (Fig. S3), and the results indicated that the RTM-GWAS method performed the best at each significance level for both CSGP and simulated ideal population. However, when using a very stringent significance level such as 1e−10, the RTM-GWAS method, and LMM had similar performance for CSGP.

The performance of association analysis methods was also evaluated under six different genetic architectures based on CSGP. The results of relative detection power (Table 2) showed that RTM-GWAS performed the best under 10-locus model at all heritability levels, and all methods failed (RPO < 1%) to detect association under 100-locus model with an extremely low trait heritability ($h^2 = 0.2$), but RTM-GWAS performed the best under 100-locus model with higher heritability ($h^2 = 0.5, 0.9$). With the decrease of trait heritability, the detection power dropped rapidly as expected for all methods in the 100-locus model simulation (Table S4). For simple traits, i.e., the 10-locus model, the detection power decreased from 0.687 to 0.225 along with the decrease of trait heritability from 0.9 to 0.2 (Table S5). These results indicated that the heritability of complex traits or the precision of the experiment is very important for the detection power of GWAS. The influence of sample size on the RTM-GWAS method was evaluated under the 100-locus model with a trait heritability of 0.9 based on the ideal population. The simulation result (Table S6) showed that the detection power decreased and the FDR increased along with the decrease of sample size. For population with a small sample size, e.g., 200, the detection power of RTM-GWAS is as low as 0.022 with a high FDR of 0.326, but the detection power increases rapidly at a sample size of 400. The results indicated that both high heritability of the traits (or experiment precision) and large sample size are required for GWAS of complex traits.

The RTM-GWAS of 100-seed weight in Chinese soybean germplasm population

At first, the genetic structure of the population was estimated based on the GSC matrix calculated from genome-wide SNPLDBs. The CSGP is highly structured and clearly divided into three clusters corresponding to WA, LR and RC (Fig. 2a). The frequency distribution, descriptive statistics,

Table 2 Relative detection power of the association analysis methods under six genetic architectures in simulations based on real soybean genotype data

No. QTL	Method	$h^2 = 0.2$	$h^2 = 0.5$	$h^2 = 0.9$
10	Naïve	0.033	0.007	0.003
	PCA	0.069	0.04	0.021
	LMM	0.089	0.083	0.087
	HAT	0.072	0.052	0.029
	RTM	0.161	0.377	0.631
100	Naïve	0.022	0.013	0.014
	PCA	0.006	0.012	0.014
	LMM	0.004	0.018	0.038
	HAT	0.002	0.011	0.012
	RTM	0.004	0.029	0.112

No. QTL, the number of causal loci simulated to generate a quantitative trait; h^2, trait heritability; Naïve, association test based on simple linear model without control for population structure; PCA, association test based on linear model, and the top 10 eigenvectors with largest eigenvalues of SNPLDB genetic similarity coefficient (GSC) matrix were incorporated as covariates to correct for population structure; LMM, association test based on linear mixed model, and SNPLDB GSC matrix was incorporated as covariance structure of random effect to correct for population structure; HAT, the haplotype-based association test implemented in PLINK software (--hap-assoc); RTM, association test using the proposed restricted two-stage multi-locus multi-allele GWAS method (RTM-GWAS). The simulation was based on the Chinese soybean germplasm population, comprising 1024 individuals, genotyped at 36,952 SNPLDBs; the relative detection power is calculated as [detection power × (1 − false discovery rate)]

and analysis of variance are listed in Table S7 and Table S8. The distribution of 100-seed weight in CSGP exhibited a wide variation ranged 0.85–35.98 g, with a variation coefficient of 54.4% and heritability of 98.9% (Fig. 2b; Table S7 and Table S8). Second, both RTM-GWAS and LMM methods were used to identify QTL for 100-seed weight. Only three loci were identified by LMM and 16 loci by RTM-GWAS using the Bonferroni criterion (Fig. 2c; Table S9). The high false-negative rate indicated that the extra Bonferroni criterion is too stringent for detecting the genome-wide QTL. Therefore, the significance level of 0.01 was then used for RTM-GWAS and 139 loci were identified covering 148 out of the 203 reported QTL regions for 100-seed weight, and 35 loci were new ones (Fig. 2d, Table S9 and Table S10) in comparison with the data in SoyBase (http://soybase.org). The Q–Q plot of association test p values (Fig. S4) showed that both LMM and RTM-GWAS exhibited an efficient correction for population structure.

The 139 loci for 100-seed weight distributed on all of the 20 soybean chromosomes, and there were 1–14 loci per chromosome with the chromosome 18 having the most. The phenotypic variation explained by each locus ranged from 0.07 to 9.84%. Among the 139 loci, there were 22 large contribution loci ($R^2 > 1\%$) explaining a total of 61.8% phenotypic variation, and 117 small contribution loci ($R^2 < 1\%$) explaining a total of 36.4% phenotypic variation; in a total, the 139 loci contributed 98.2% of the phenotypic variation. Since the undetected QTL (mainly small contribution QTL) explained only 0.7% of the phenotypic variation, a relatively thorough or full detection of the 100-seed weight was achieved using the RTM-GWAS under the experiment precision conditions. Furthermore, the most loci detected by Bonferroni-adjusted significance level were also detected by non-adjusted significance level, but the detected loci by the latter usually cover those by the former. In the present study, the 139 loci include 15 out of the 16 loci identified using additional Bonferroni criterion and also include the two loci detected by LMM.

QTL-allele matrix and candidate genes of 100-seed weight in the CSGP

There were 47 loci with multiple SNPs (M.SNPLDBs) among the 139 detected loci. The allele number per locus ranged from 2 to 10, with an average of 2.9 and a total of 402 (detailed data in Spreadsheet S1). The effect of each allele was estimated by RTM-GWAS, including 201 positive and 201 negative effect alleles. The allele effects ranged from 0.0189 to 9.5207 g for positive effect alleles and from −5.2758 to −0.0002 g for negative effect alleles, and approximately 89.3% of the allele effects were between −1.0 and 1.0 (Fig. 3a, Spreadsheet S1). The detected 100-seed weight loci with their 402 allele effects can be further organized into a 139 × 1024 (locus × accession) or 402 × 1024 (allele × accession) matrix (Fig. 3b), which in fact acts as the genetic structure of the whole population (CSGP). The matrix characterizes the population with the genetic richness and diversity on each locus as well as the population, from which all of the allele frequencies can be obtained for further study.

From the detected QTL system of soybean 100-seed weight, there are a total of 766 annotated genes within or neighboring to 126 out of the 139 loci according to SoyBase (http://soybase.org), among which 136 candidate genes (with 281 SNPs) were tightly associated with 74 loci (Table S11). The Gene Ontology analysis showed that these 136 genes are involved in various biological processes that could be grouped into nine categories, including primary metabolism, secondary metabolism, seed development, cell growth, photosynthesis, flower development, response to stress, signal transduction, and unknown process (Fig. S5A). In addition, the haplotypes of a candidate gene can be found from the genomic sequence data. For example, as shown in Fig. S5B, the candidate gene Glyma14g10915 has nine haplotypes based on the six SNPs, which correspond to five haplotypes/alleles in the locus Gm14_BLOCK112_9120910_9131797

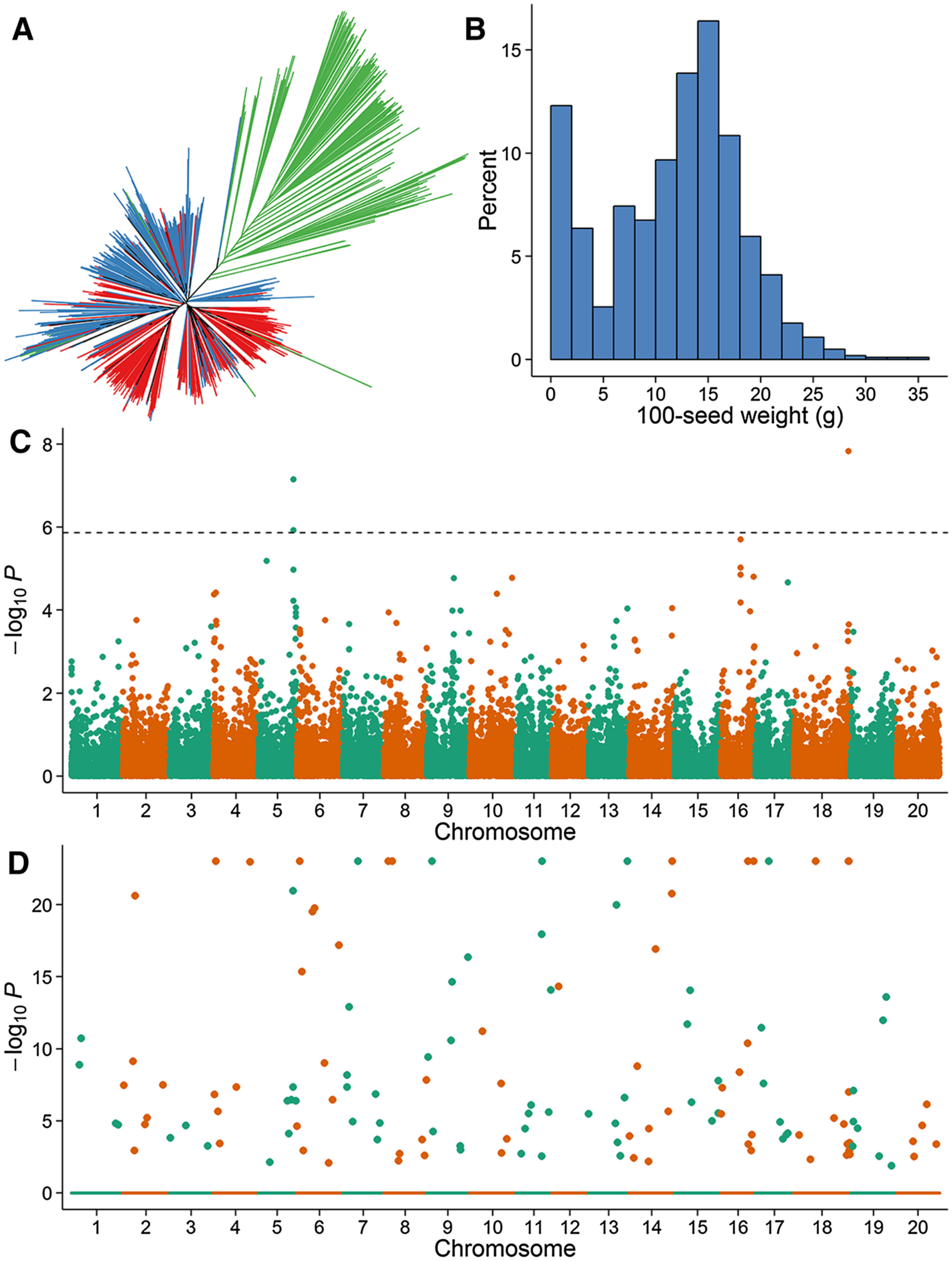

Fig. 2 Genetic dissection of phenotypic variation for 100-seed weight in CSGP. **a** Neighbor-joining tree of CSGP based on SNPLDB, where *green* for wild soybean, *blue* for soybean landrace, and *red* for released cultivar. **b** Histogram of 100-seed weight in CSGP. **c** Manhattan plot for LMM method. Association test was based on linear mixed model with EMMAX algorithm, and SNPLDB GSC matrix was incorporated as covariance structure of random effect. **d** Manhattan plot for RTM-GWAS method with a threshold of 0.05 for the first stage and a significance level of 0.01 for the second stage, and the −log10 *p* values greater than 22.4 (the maximum is 129.8) were shown as 22.4 (color figure online)

(The number of gene haplotype is greater than that of QTL-allele, because the expanding region of ±30 kb was used to search candidate genes, and therefore, the first two SNPs of Glyma14g10915 were not in the SNPLDB interval). In this way, the RTM-GWAS can provide a way to predict the allelic information of gene-allele system in a germplasm population, rather than only a small part of them.

Genetic characterization of the CSGP

From the established QTL-allele matrix, the genetic differentiation can be detected and characterized at the entire population level or subpopulation level, involving whole-genome QTL or a group of QTL. The differentiation in allele frequency between WA and LR subpopulations (WA–LR), between LR and RC subpopulations (LR–RC), and among the three subpopulations (WA–LR–RC) was examined based on both genome-wide SNPLDBs and QTL. The AMOVA based on genome-wide 36,952 SNPLDBs showed that 26.36, 4.09, and 19.31% of genomic variation were explained by the subpopulation for WA–LR, LR–RC, and WA–LR–RC, respectively (Table S12). The AMOVA based on the 139 loci for 100-seed weight exhibited a similar pattern with the results from the total markers, but with larger between(among)-subpopulation variation for LR–RC (4.82% vs. 4.09 in comparison with 20.7 1 vs. 26.36% for

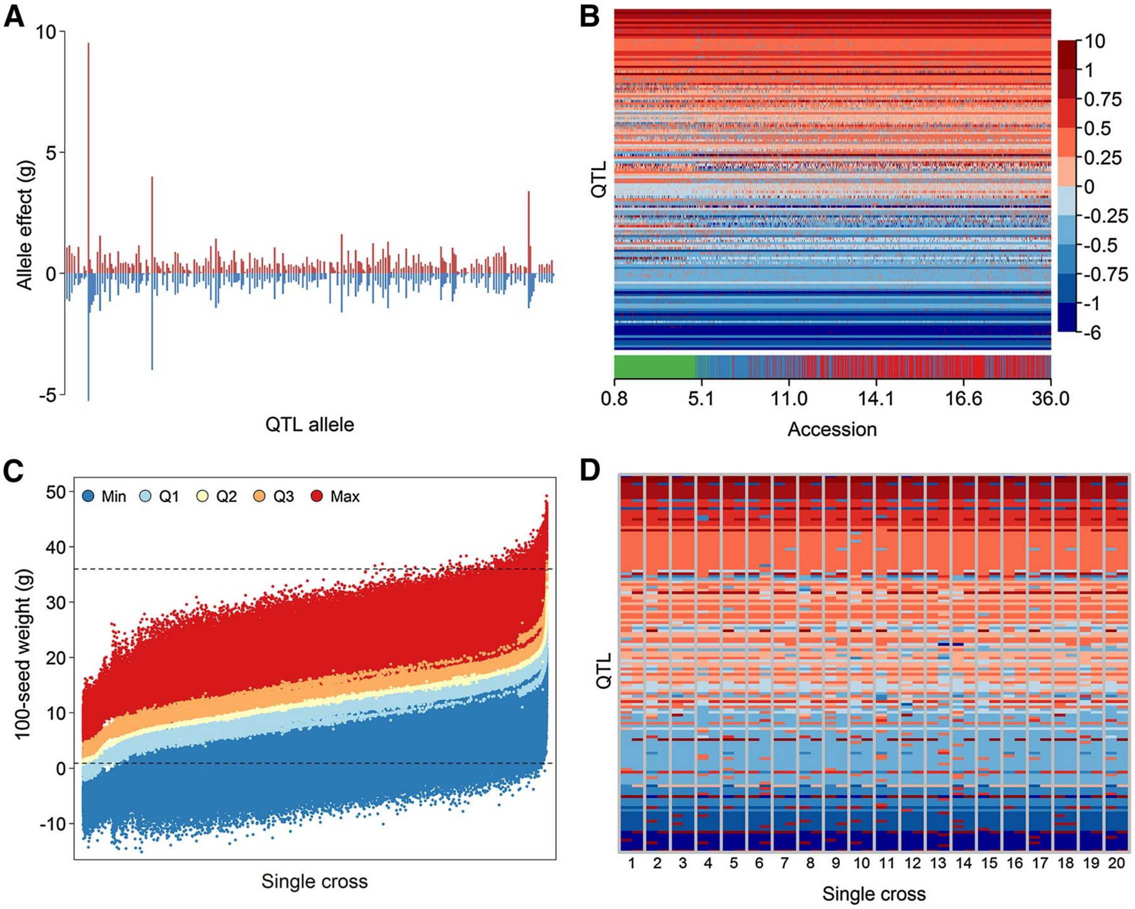

Fig. 3 QTL-allele matrix and optimal recombination design for soybean 100-seed weight. **a** Distribution of genetic effect of 402 alleles of the 139 QTL for soybean 100-seed weight. **b** Heat map representation of the QTL-allele matrix (locus × accession) of soybean 100-seed weight, which the *right color bar* represents *colors* for the ten intervals of allele effects and the *bottom color bar* represents accessions with 100-seed weight arranged in ascending order where *green* for wild soybean, *blue* for landrace soybean, and *red* for released cultivar. **c** Distribution of predicted 100-seed weight of the simulated progenies of all possible single crosses, with the maximum and minimum (*top* and *bottom horizontal line*) values of parental lines. **d** *Top* 20 superior single crosses under linkage model. The optimal crosses were identified according to the 99th percentile of the sample as its breeding potential. Each cross was presented with QTL genotype of the two parental lines in one column, while different *colors* represent the size of QTL-allele effect as in **b** (color figure online)

LR–RC and 15.90 vs. 19.31% for WA–LR–RC), which might coincide with the genetic improvement from LR to RC.

Each of the 100-seed weight loci was also statistically tested for differentiation among subpopulations (Spreadsheet S1). Among the 139 loci, 83 (59.7%), 84 (60.4%), and 107 (77.0%) loci showed significant differentiation in allele frequency for WA–LR, LR–RC, and WA–LR–RC, respectively, at a significance level of 0.01; and the proportion of loci (60.4%) with a significant differentiation among LR–RC is greater than that on the genome-wide level (39.7%). The allele frequency of the 139 loci showed that 339 (84.3%) alleles existed in all three subpopulations, and 42 (10.4%) alleles existed in only two subpopulations, especially 21 (5.2%) alleles existed in only one subpopulation. The frequency of positive and negative effect alleles was investigated in each subpopulation and whole CSGP (Fig. S6). There are 46.1, 46.9, 48.3, and 47.4% positive effect alleles in WA, LR, RC, and CSGP, respectively. The subpopulation exhibited different pattern of distribution of positive and negative effect alleles. In general, the number of positive alleles was less than that of negative alleles for accessions with 100-seed weight <10 g, but was greater for accessions with 100-seed weight >10 g. Since the space in the present paper is limited, we will have another paper to present the differentiation and evolutionary relationship among the subpopulations, including the newly emerged or extinct old alleles.

Genomic selection for optimal crosses of 100-seed weight in the CSGP

Using the QTL-allele matrix in GS for optimal crosses, in the present study, all possible crosses (523,776) among the 1024 accessions each with 2000 homozygous progenies were simulated based on 100-seed weight QTL-allele matrix built from 139 QTL, with the 99th percentile representing the breeding potential for each single cross. The prediction was performed in two ways, with-linkage and without-linkage; for the with-linkage model, the natural linkage among the detected QTL in the population was maintained, while for the without-linkage model, independent assortment among the detected QTL was considered. Among the 523,776 single crosses, the top 20 crosses with largest predicted 100-seed weight were identified with 12.4–19.9% improvement over the accession with largest 100-seed weight in the CSGP based on linkage model (Fig. 3c, d; Table S13). The linkage model is reasonable for optimal cross prediction in this case and the transgressive potential at both positive and negative directions is very large. However, the without-linkage or independent assortment model might be appropriate if breaking the negative linkages can reveal more potential (up to 30.7% improvement, Fig. S7 and Table S13).

Discussion

The efficiency and power of the RTM-GWAS procedure

Our purpose was to explore the full QTL system in a plant germplasm/breeding population rather than only a handful of major QTL through the improvement of GWAS procedure, or in other words, to increase the QTL detection efficiency and power of the GWAS procedure. The above results indicated that the RTM-GWAS provides a much better QTL detection efficiency and power (about 5–8 or more times of) than the other four methods. The RTM-GWAS is characterized by two innovations: the multi-allelic SNPLDB genomic markers derived from bi-allelic SNPs and the two-stage multi-locus multi-allele GWAS model.

First innovation: the SNPLDB marker with varied number of alleles

Compared with SNP-based GWAS methods, the multi-allelic SNPLDB marker can match a QTL with multiple alleles and multiple SNPLDB markers can match a series of QTL with varied number of alleles. As the multi-allelic genetic variation is a natural property of the plant germplasm population, the SNPLDB marker theoretically matches the genetic loci with varied number of alleles and detecting QTL by SNPLDB marker should be more appropriate than the bi-allelic SNP marker. However, different haplotype block partitioning algorithms can be employed for SNPLDB marker construction, and may give very different results (Ding et al. 2005). Therefore, the block definition may have strong effect on the construction of SNPLDB marker. Although several methods based on different concepts have been proposed for block partitioning, including the LD-based (Gabriel et al. 2002), the recombination-based (Wang et al. 2002), and the diversity-based (Zhang et al. 2002) methods, the approach based on LD confidence intervals is suggested for SNPLDB construction in the present study. Because the genomic block is determined by the recombination history of the population, the LD pattern is the appropriate measurement of the historic recombination (Pattaro et al. 2008).

Second innovation: the two-stage multi-locus multi-allele GWAS model

The two-stage association analysis, where the markers were preselected by single-locus model followed by multi-locus multi-allele model stepwise regression, is benefited from the distant noisy markers reduced at the first stage to keep a relatively less-noisy mapping environment, the total phenotypic contribution controlled within the trait heritability value, the built-in experiment-wise threshold without additional Bonferroni correction, and population structure adjustment with GSC. The simulation studies indicated that RTM-GWAS is much more efficient and powerful than the PCA, LMM, and HAT methods, especially under the conditions of high heritability, large sample size, and large number of involved QTL. Although the overall relative detection efficiency and power in Tables 1 and 2 are not high enough outwardly, there are two reasons involved. One is that the Bonferroni correction was used, which in fact is not appropriate for a procedure with a built-in experiment-wise significance criterion in RTM-GWAS. The results of the simulation studies were well supported by the case study on QTL detection of 100-seed weight in CSGP, where 16 QTL vs. 139 QTL were identified by the thresholds with vs. without Bonferroni correction, which showed very large difference in the detection power even for the same set of data. The large number of detected QTL (139) in the CSGP (with 1024 accessions) may be convinced by the fact that these QTL covered 148 of the 203 QTL regions reported at SoyBase (from about 59 parental materials). Another reason for a not high enough efficiency and power (but higher than the other procedures) in the simulation studies for RTM-GWAS is that the method to generate our simulation studies is different from the previous studies. In the present study, 100 causal loci were simulated for a quantitative trait, and an overall detection power was calculated on all causal loci. However, in the previous studies in the literature, usually only a handful of causal loci were simulated or the detection power was only calculated on a part of the causal loci (Segura et al. 2012; Wang et al. 2016). As a quantitative trait is more likely controlled by a large number of genes, a 100-locus model for quantitative trait is more reasonable. Logically, the evaluation of detection power should be based on all causal loci rather than on a small part of them.

Benefit of GSC based on SNPLDB

In addition to the two innovations in the RTM-GWAS procedure, the use of GSC obtained from a large number of genome-wide SNPLDB for population structure correction also contributes to the improvement of detection efficiency and power. The germplasm/breeding population structure varies greatly, which can be caused from both admixture (such as geographic isolation, migration, and artificial selection) and inbreeding (such as differential mating scheme), as well as their mixture with the proportions even not estimable (Yu et al. 2006). Usually, the accessions in germplasm were developed by farmers and collected from different geographic regions, which had inbred for many generations, even no pedigree/kinship can be traced for this kind of population. Therefore, separation of the whole population into subpopulations through model-based population structure correction does not necessarily match the real situation. Thus, GSC is the most possible estimation of the genetic relationship among accessions that can be estimated from the SNPLDB markers for population structure correction. The estimated relationship may contain all the genetic information of the population, including the comprehensive results from admixture and differential inbreeding schemes, as well as both of them. The effectiveness of population structure correction with SNPLDB GSC matrix is about the same as EIGENSTRAT (Price et al. 2006), VanRaden1, and VanRaden2 (VanRaden 2008) based on SNP markers (Table S14). However, the SNPLDB GSC matrix takes into account the varied number of alleles, and should fit better the genetic relationship among the materials in the population. Therefore, using the SNPLDB GSC matrix for population structure correction (even based on sampling a representative working population if needed) is suggested as a universal approach in RTM-GWAS.

Further consideration on model selection

The stepwise regression algorithm used in the second stage of RTM-GWAS may not be the best solution for model selection. Actually, there are several sophisticated variable selection algorithms that have been widely used for QTL mapping, such as the Lasso methods (Li and Sillanpaa 2012) and the Bayesian methods (Karkkainen and Sillanpaa 2012). However, these algorithms are typically designed for continuous variable with one degree of freedom, and cannot be readily applied to the multi-allelic SNPLDB maker which can have two or more degrees of freedom. The group Lasso algorithm (Yuan and Lin 2006) may be a potential solution on this, but more investigations are still needed to adapt these algorithms to RTM-GWAS. Furthermore, as indicated by the simulation results, it is also suggested that the sample size in RTM-GWAS should be large enough (such as >400), and the trait heritability should be controlled at a high level (such as >0.8) which can be achieved through experimental design and careful operation of the experiments.

Potential utilization of the RTM-GWAS procedure

Germplasm population characterization

As indicated above, the detected full-size QTL-allele matrix can be used directly to characterize the population, to study the differentiation of the population, and to reveal the evolutionary relationship among populations based on allele frequencies. Moreover, in plants, the collected germplasm is a genetic reservoir from which all varieties were developed. If the genetic constitution in terms of gene/QTL-allele composition of all breeding target traits in the germplasm reservoir has been explored, the genetic constitution of its derived breeding populations can be inferred from their genome-wide markers using the established relationship between QTL-alleles and molecular markers in the reservoir. Since this kind of QTL/allele-marker relationship was obtained from the reservoir, the genetic information can be used for its derived population. For example, the present QTL-allele matrix obtained from the CSGP may be used to infer the QTL-allele structure of its derived populations, especially in the tested ecoregion (lower valleys of Changjiang and Huai-river and their neighboring regions in China, but not necessarily in regions far away from this region due to the environmental differences of ecoregions which might cause different phenotypes and mapping results). Since the CSGP is composed of a large size (1024) of representative germplasm accessions, including the three sources of WA, LR, and RC from all regions in China, the QTL-allele structure of all materials derived from the population may be inferred accordingly, and the breeding population can be organized differently with its corresponding QTL-allele matrix meeting the requirements of different breeding programs. The geneticists and breeders can use the QTL-allele structure of the secondary or re-organized breeding populations to do further genetic study and breeding operations. Thus, genotyping the germplasm resources thoroughly in a state or a country is an efficacious forever program and should be arranged as a state- or country-wide public program.

Genetic system and gene finding for quantitative traits

A series of parallel studies on the utilization of RTM-GWAS have been also carried out in our group (Zhang et al. 2015a; Meng et al. 2016), a recent work was on drought tolerance of soybean. Using the RTM-GWAS, a total of 268 QTL were detected for four drought-tolerance traits in a soybean germplasm population. From which, 684 genes were annotated, and the expression patterns of 320 genes in response to drought were verified with qRT-PCR. The annotated genes are involved in the biological processes of ABA responser, stress responser, transport, development factor, protein metabolism factor, transcription factor, protein kinase, and unknown others or involved in a gene network (Wang et al. 2017, personal communication). The verified expression pattern and annotation demonstrated the relevance of the detected QTL system of the traits, and, therefore, further demonstrated the usefulness of RTM-GWAS procedure in detection of the QTL system in plant germplasm population. Another example of potential utilization of the RTM-GWS procedure was on mapping QTL conferring flowering date of soybean in a nested association mapping (NAM) population composed of four recombinant inbred line (RIL) populations with a joint parent, in which 139 QTL with 496 alleles were detected from RTM-GWAS, while only 7–16 QTL with 14–72 alleles were identified by other mapping procedures, including composite interval mapping (CIM), mixed model-based composite interval mapping (MCIM), joint inclusive composite interval mapping (JICIM), and mixed linear model GWAS (MLM-GWAS) (Li et al. 2017). The reason for a high power of RTM-GWAS in RIL and NAM populations is that these bi-parental populations fit well in the hypothetical conditions for using LD to detect the tightly associated QTL since there is no population structure problem in RIL populations.

Genomic selection based on QTL-allele matrix in plant breeding

Following the "Breeding by Design" concept by Peleman and van der Voort (2003), we proposed the genomic selection based on QTL-allele matrix (QTL-allele-based GS). This strategy is different from the classic GS based on GEBV (GEBV-based GS) by Meuwissen et al. (2001). The latter was started mainly for animal breeding but has been considered to fit the requirement on GS for crosses and progenies in plant breeding, while the former was concentrated on cross and genotype design and selection. In fact, the QTL-allele-based GS is a direct genotype selection for trait alleles/genes, and the GEBV-based GS is a selection for a linear combination of SNPs for comprehensive traits or indirect selection for alleles/genes. Both strategies are characterized with their advantages and short comings, and are to be further improved and evaluated in breeding practices.

Theoretically, the QTL-allele-based GS is characterized with: direct alleles/genes selection but limited for additive effect selection; selection accuracy depends on QTL mapping accuracy; multiple usefulness of the obtained QTL-allele matrix of the germplasm/breeding population; and easiness of optimal cross prediction with its recombination potential and reduced genotyping cost for progeny selection due to reduced marker number. While the GEBV-based GS is characterized with: indirect alleles/genes selection using GEBV with additive and epistasis effects mixed in GEBV; selection accuracy depends on the performance of prediction models; the usefulness of GEBV is limited to the breeding program; high genotyping cost for keeping all GEBV-related markers in progeny selection. According to the "Breeding by Design" concept, we tend to use a direct selection for alleles/genes if the QTL/gene mapping can be accurate and complete enough. It is especially due to the large number of materials (more than 100 crosses and 5000–10,000 progenies each year) to be tested and genotyped in plant breeding programs. We compared our detected QTL with those in the literature (SoyBase, http://soybase.org); our results can cover basically those in the literature, even with some more new ones, which convinces us of the accuracy of our QTL-allele matrix and GS prediction. In addition, in a QTL pyramiding study using marker-assisted selection for seed protein content, a transgressive progeny with 54.15% was obtained from a cross between two lines obtained from two respective RIL populations with protein content 35.35–44.83% in their four parents (Zhang et al. 2015b). This example also convinces us to use "Breeding by Design" strategy as our GS strategy.

However, the difficulty in genotyping a large amount of progenies is still a problem to be resolved, even the needed markers have been limited to the involved QTL in QTL-allele-based GS. Furthermore, in a real breeding program, multiple traits might be considered simultaneously where multiple QTL-allele matrices are involved. In this case, the linear combination of the matrices should be considered in the two GS stages. In addition, GS for different crossing patterns (two-way, three-way, even multiple-way crosses) and GS for traits with pleiotropy and epistasis should also be considered in the future.

Author contribution statement JG conceived and designed the study. JH performed the simulations and developed the computer software. JH and SM analyzed and interpreted the results. SM, TZ, GX, SY, and YL performed the field experiments. RG, JL, YW, QX, and BY performed the genome sequencing. JH, SM, and JG drafted the manuscript.

Acknowledgements This work was supported by the China National Key R & D Program for Crop Breeding (2016YFD0100304), the China National Key Basic Research Program (2011CB1093), the China National Hightech R&D Program (2012AA101106), the Natural Science Foundation of China (31571695), the MOE 111 Project (B08025), Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT13073), the MOA Public Profit Program (201203026-4), the MOA CARS-04 program, the Jiangsu Higher Education PAPD Program, and the Jiangsu JCIC-MCP Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res 21:610–617

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 JD, 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 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nat Genet 37:868–872

De Coninck A, De Baets B, Kourounis D, Verbosio F, Schenk O, Maenhout S, Fostier J (2016) Needles: toward large-scale genomic prediction with marker-by-environment interaction. Genetics 203:543–555

Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592–601

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

Dhanapal AP, Ray JD, Singh SK, Hoyos-Villegas V, Smith JR, Purcell LC, Andy King C, Cregan PB, Song Q, Fritschi FB (2015) Genome-wide association study (GWAS) of carbon isotope ratio ($\delta^{13}C$) in diverse soybean [*Glycine max* (L.) Merr.] genotypes. Theor Appl Genet 128:73–91

Ding K, Zhou K, Zhang J, Knight J, Zhang X, Shen Y (2005) The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. Mol Biol Evol 22:148–159

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567

Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, Georges M (2000) Extensive genome-wide linkage disequilibrium in cattle. Genome Res 10:220–227

Felsenstein J (1989) PHYLIP—phylogeny inference package (version 3.2). Cladistics 5:164–166

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Hanson CH, Robinson HF, Comstock RE (1956) Biometrical studies of yield in segregating populations of Korean *Lespedeza*. Agron J 48:268

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol 65:531–551

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

Jia G, Huang X, Zhi H, Zhao Y, Zhao Q, Li W, Chai Y, Yang L, Liu K, Lu H, Zhu C, Lu Y, Zhou C, Fan D, Weng Q, Guo Y, Huang T, Zhang L, Lu T, Feng Q, Hao H, Liu H, Lu P, Zhang N, Li Y, Guo E, Wang S, Wang S, Liu J, Zhang W, Chen G, Zhang B, Li W, Wang Y, Li H, Zhao B, Li J, Diao X, Han B (2013) A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). Nat Genet 45:957–961

Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. Genetics 201:759–768

Jonas E, de Koning DJ (2013) Does genomic selection have a future in plant breeding? Trends Biotechnol 31:497–504

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 SY, 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

Karkkainen HP, Sillanpaa MJ (2012) Back to basics for Bayesian model building in genomic selection. Genetics 191:969–987

Li Z, Sillanpaa MJ (2012) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet 125:419–435

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N, Liu J, Warburton ML, Cheng Y, Hao X, Zhang P, Zhao J, Liu Y, Wang G, Li J, Yan J (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45:43–50

Li S, Cao Y, He J, Zhao T, Gai J (2017) Detecting the QTL-allele system conferring flowering date in a nested association mapping population of soybean using a novel procedure. Theor Appl Genet. doi:10.1007/s00122-017-2960-y

Meng S, He J, Zhao T, Xing G, Li Y, Yang S, Lu J, Wang Y, Gai J (2016) Detecting the QTL-allele system of seed isoflavone content in Chinese soybean landrace population for optimal cross design and gene system exploration. Theor Appl Genet 129:1557–1576

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Mohammadi M, Tiede T, Smith KP (2015) PopVar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. Crop Sci 55:2068

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES, Kresovich S (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci USA 110:453–458

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8:4321–4325

Nordborg M, Weigel D (2008) Next-generation genetics in plants. Nature 456:720–723

Pattaro C, Ruczinski I, Fallin DM, Parmigiani AG (2008) Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies. BMC Genomics 9:405

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190

Peleman JD, van der Voort JR (2003) Breeding by design. Trends Plant Sci 8:330–334

Peng B, Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. Bioinformatics 21:3686–3687

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

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

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, 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

Rakitsch B, Lippert C, Stegle O, Borgwardt K (2013) A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics 29:206–214

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q, Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44:825–830

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Vazquez AI, Veturi Y, Behring M, Shrestha S, Kirst M, Resende MF Jr, de Los Campos G (2016) Increased proportion of variance explained and prediction accuracy of survival of breast cancer patients with use of whole-genome multiomic profiles. Genetics 203:1425–1438

Voight BF, Pritchard JK (2005) Confounding from cryptic relatedness in case–control association studies. PLoS Genet 1:e32

Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet 71:1227–1234

Wang S-B, Feng J-Y, Ren W-L, Huang B, Zhou L, Wen Y-J, Zhang J, Dunwell JM, Xu S, Zhang Y-M (2016) Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep 6:19444

Weir BS (2008) Linkage disequilibrium and association mapping. Annu Rev Genom Hum Genet 9:129–142

Yu J, Pressoir G, Briggs WH, Vroh Bi I, 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

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc B 68:49–67

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468

Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360

Zhang Y, He J, Wang Y, Xing G, Zhao J, Li Y, Yang S, Palmer RG, Zhao T, Gai J (2015a) Establishment of a 100-seed weight quantitative trait locus-allele matrix of the germplasm population for optimal recombination design in soybean breeding programmes. J Exp Bot 66:6311–6325

Zhang Y, Liu M, He J, Wang Y, Xing G, Li Y, Yang S, Zhao T, Gai J (2015b) Marker-assisted breeding for transgressive seed protein content in soybean [*Glycine max* (L.) Merr]. Theor Appl Genet 128:1061–1072

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun 2:467

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, Fang C, Shen Y, Liu T, Li C, Li Q, Wu M, Wang M, Wu Y, Dong Y, Wan W, Wang X, Ding Z, Gao Y, Xiang H, Zhu B, Lee SH, Wang W, Tian Z (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33:408–414