Genome-wide association study and genomic selection for plant height, maturity, seed weight, and yield in soybean

Waltram Ravelombola
University of Arkansas
Jun Qin
Hubei Academy of Agricultural Sciences
Ainong Shi (ashi@uark.edu)
University of Arkansas Fayetteville https://orcid.org/0000-0002-1066-7920
Fengmin Wang
Hubei Academy of Agricultural Sciences
Yan Feng
Hubei Academy of Agricultural Sciences
Yaning Meng
Hubei Academy of Agricultural Sciences
Chunyan Yang
Hubei Academy of Agricultural Sciences
Mengchen Zhang
Hubei Academy of Agricultural Sciences

Research article

Keywords: Glycine max, Genome-wide association study, Genomic selection, Genotyping by sequencing, Maturity, Seed weight, Yield, Plant height, Single nucleotide polymorphism

Posted Date: January 3rd, 2020

DOI: https://doi.org/10.21203/rs.2.20026/v1

Abstract

Background

Soybean [Glycine max (L.) Merr.] is a legume of great interest worldwide. Enhancing genetic gain for agronomic traits via molecular approaches has been long considered as the main task for soybean breeders and geneticists. The objectives of this study were to conduct a genome-wide association study (GWAS) for these traits and identify SNP markers associated with the four traits, and to assess genomic selection (GS) accuracy.

Results

A total of 250 soybean accessions were evaluated for maturity, plant height, seed weight, and yield over three years. This panel was genotyped with a total of 10,259 high quality SNPs postulated from genotyping by sequencing (GBS). Population structure was inferred using STRUCTURE 2.3.4, GWAS was performed using a Bayesian Information and Linkage Disequilibrium Iteratively Nested Keyway (BLINK) model, and GS was evaluated using a ridge regression best linear unbiased predictor (rrBLUP) model. The results revealed that: a total of 20, 31, 37, 31, and 23 SNPs were significantly associated with the average 3-year data for maturity, plant height, seed weight, and yield, respectively; some significant SNPs were mapped into previously described loci (E2, E4, and Dt1) affecting maturity and plant height in soybean and a new locus mapped on chromosome 20 was significantly associated with plant height; Glyma.10g228900, Glyma.19g200800, Glyma.09g196700, and Glyma.09g038300 were candidate genes found in the vicinity of the top or the second best SNP (if no annotated genes found close the top one) for maturity, plant height, seed weight, and yield, respectively; a 11.5-Mb region of chromosome 10 was associated with both seed weight and yield; and GS accuracy was trait-, year-, and population structure-dependent.

Conclusions

The SNP markers identified from this study for plant height, maturity, seed weight and yield can be used to improve the four agronomic traits in soybean through marker-assisted selection (MAS) and GS in breeding programs. After validation, the candidate genes can be transferred to new cultivars using the linked SNP markers through MAS. The high GS accuracy has confirmed that the four agronomic traits can be selected in molecular breeding through GS.

Background

Soybean [*Glycine max* (L.) Merr.] is one of the most important legumes grown worldwide. It provides about 60% of the vegetable-derived proteins and more than 57% of oilseed (http://www.fas.usda.gov/). Soybean biofuel has recently become attractive with a value estimated to be more than $35 billion in the United States (www.soystats.com). The success of a newly developed soybean variety that meets end-users' needs relies on a large number of characteristics. Of which, maturity, plant height, seed weight, and

yield are critical [1]. Soybean breeders have been focusing on improving these traits in their cultivar development pipeline. Highly heritable traits such as maturity are selected at an earlier stage, whereas complex traits such as yield are tested at a later stage in the breeding pipeline. With the rapid development of genomic-related tools and DNA sequencing technology, the improvement of agricultural traits of interest could be performed faster.

Genome-wide association study (GWAS) and genomic selection (GS) are powerful tools to assist with understanding the genetic architecture governing complex traits of importance in agriculture. GWAS has been conducted to identify markers associated with agronomic traits in soybean [2,3]. To date, more than 60 loci found through GWAS have been reported to be associated with maturity in soybean [4–9]. These loci are distributed across the soybean genome with chromosome 16 having the highest number of significant loci (https://www.soybase.org/). Candidate genes such as *Glyma11g14150*, *Glyma16g02840*, and *Glyma16g03580* have been suggested to control maturity in soybean [8]. The most described loci affecting maturity in soybean are *E1-E10* and *J* [10].

Efforts towards identifying significant loci controlling plant height in soybean via GWAS have also been undertaken. Recently, 68 loci affecting plant height have been identified using GWAS [5,7,8,11,12]. Out of the 68 loci identified for plant height, 19 were mapped on chromosome 19 (https://www.soybase.org/). Candidate genes such as *Glyma09g34601*, *Glyma10g08930*, *Glyma12g00320*, *Glyma16g06333*, *Glyma16g25570*, *Glyma17g36350*, *Glyma18g49080*, *Glyma19g37180*, *Glyma.13g362300*, *Glyma.13g365500*, *Glyma.02g155600*, *Glyma.19g196000*, *Glyma19g37890*, *Dt1*, *LOC100789162*, *LOC100787543*, *LOC100786140*, *LOC100804065*, *LOC100777767*, *LOC100810047*, *LOC100804944*, and *LOC100788304* have been reported to affect plant height in soybean [11,12].

Molecular markers associated with seed weight have been previously investigated in soybean. GWAS identified a total of more than 95 loci controlling seed weight with chromosome 4 harboring 12 loci [1,4,6,7,11–16]. *Glyma11g03360*, *Glyma11g03430*, *Glyma11g05760*, *Glyma18g05240*, *Glyma18g43500*, *Glyma18g43500*, *Glyma.03g166700*, *Glyma.11g164700*, *Glyma.11g164700*, *LOC100784416*, *Glyma05g09390*, *Glyma.07g076800*, *Glyma10g28250*, *Glyma14g08050*, *Glyma14g08280*, *Glyma19g33421*, *Glyma19g33550*, and *Glyma19g35180* have been reported as potential candidate genes for seed weight in soybean [1,11,12,14].

GWAS has been proven to be efficient in identifying molecular markers associated with yield in soybean. A total of 219 soybean accessions were evaluated for yield and four SNPs were found to be associated with yield across multiple environments via GWAS [12]. The candidate genes *Glyma.13g073900*, *Glyma.06g050300*, *Glyma.03g169700*, and *Glyma.03g171900* were found in the vicinity of these SNPs. In addition, a total of 139 soybean accessions were genotyped using the BARCSoySNP6K in order to conduct GWAS for yield [11], reporting a total of six significant SNPs associated with yield, of which, four were located on chromosome 12. An additional study also suggested six significant loci associated with yield in soybean [6]. Copley et al. [4] reported three significant SNPs associated with yield through GWAS.

GS has allowed for the estimation of the effects of all markers across the genome. These markers effects have been coined genomic estimated breeding values (GEBVs) and can be used for predicting the performance of a line [17]. GS has been shown to outperform the traditional marker-assisted selection (MAS) when complex traits such as yield are involved [18]. The establishment of GS for complex traits can permit the achievement of a rapid genetic gain per unit of time [18]. Matei et al. [19] showed that selection cycle for yield and seed weight can be significantly shortened using GS. The efficiency of GS heavily relies on its accuracy. However, previous reports showed discrepancy regarding the accuracy of GS for complex trait in soybean. Duhnen et al. [20] reported an accuracy of 0.39 for yield in soybean. Jarquín et al. [21] suggested an accuracy of up to 0.80 for soybean yield in an elite population consisting of 301 individuals.

Conducting GWAS and GS for maturity, plant height, seed weight, and yield in soybean can contribute towards discovering new loci of interest and confirming the previously reported ones, which will significantly enhance breeding for agronomic traits in soybean. In addition, GS-related studies in soybean still remain limited. Therefore, the objectives of this study were to investigate the population structure within a soybean panel, to conduct GWAS for maturity, plant height, seed weight, and yield, to identify SNP markers associated with these traits, and to assess the accuracy of GS for these traits.

Results

Phenotypic variation in maturity, plant height, seed weight and yield in soybean germplasm

Data on maturity were collected over 3 years on a total of 250 soybean genotypes. Average maturity was slightly stable over years. Average maturity values were 118.1 days, 106.4 days, and 103.8 days in 2008, 2009, and 2010, respectively (Table 1). Average maturity was in the range of 77.3 days to 139.0 days with an average of 109.5 days for the 3-year study (Table 1) (Fig. 1). Maturity was significantly different between years (F value = 600, p-value < 0.0001) and between genotypes (F value = 5.67, p-value < 0.0001) (Table 2). The top 5 soybean genotypes with the longest average maturity were Hualvdou (123.9 days), Miyangxiaozihuang (122.8 days), Miyangbaidou (121.9 days), Huanggandou (120.4 days), and Zhongnongshi-1 (119.8 days), whereas the ones with the shortest average maturity were Huangdou 2 (92.7 days), Xiaohuangdou (92.3 days), Zhongdou 36 (91.6 days), Shuyangchunheidoubing (88.7 days), and PI159764 (79.8 days) (Table 3).

Plant height was 90.5 cm, 101.8 cm, and 76.1 cm in 2008, 2009, and 2010, respectively (Table 1). Plant height ranged between and 18.5 cm and 162.9 cm over 3 years (Table 1) (Fig. 1). Significant differences in plant height were found between years (F value = 496.08, p-value < 0.0001) and between genotypes (F value = 18.51, p-value < 0.001) (Table 2). Based on the 3-year pooled data, the tallest genotypes were Lvpihuangdou (152.7 cm), Miyunlaoyelian (143.3 cm), Jindou 21 (139.5 cm), Xiaheidou (138.4 cm), and

Sijiaoqi (133.7 cm), whereas the shortest ones consisted of Elf (38.3 cm), Gnome 85 (35.9 cm), Kaohsiung 1 (31.6 cm), Zhonghuang 18 (30.0 cm), and PI159764 (20.2 cm) (Table 3).

Average data on seed weight were, in g per 100 seeds, 17.3, 18.8, and 16.8 in 2008, 2009, and 2010, respectively (Table 1) (Fig. 2). Seed weight varied from 18.5 (g/100 seeds) to 162.9 (g/100 seeds) over 3 years. Seed weight was significantly different between years (F value = 135.67, p-value < 0.0001) and between genotypes (F value = 28.49, p-value < 0.0001) (Table 2). Overall, the genotypes with the highest seed weight were Zhongpindaheidou (34.0 g/100 seeds), KD01 (33.7 g/100 seeds), Lv 75 (32.6 g/100 seeds), Lvrouheipidou (31.6 g/100 seeds), and Verde (29.8 g/100 seeds), whereas those with the lowest seed weight were Liushiribaidou (9.5 g/100 seeds), Qingdou (9.2 g/100 seeds), Xiaoheidou (8.3 g/100 seeds), Shuyangchunheidoubing (6.9 g/100 seeds), and Banyesheng (4.3 g/100 seeds) (Table 3).

Average soybean yield was 2309.3 kg/hm^2, 2207.9 kg/hm^2, and 2212.1 kg/hm^2 in 2008, 2009, and 2010, respectively (Table 1). Over the 3 years, the highest recorded yield was 4286.6 kg/hm^2, whereas the lowest one was 649.4 kg/hm^2 (Table 1). Significant differences in yield were identified between years (F value = 5.32, p-value = 0.0052) and between genotypes (F value = 4.2, p-value < 0.0001) (Table 2). The high-yielding genotypes were Delsoy 4710 (3406.1 kg/hm^2), Franklin (3204.9 kg/hm^2), Lu 99 − 1 (3195.8 kg/hm^2), Newton (3134.7 kg/hm^2), and Lu 99 − 7 (3127.9 kg/hm^2), whereas the accessions with the lowest yield for the 3-year average were Xiaoheidou (1251.8 kg/hm^2), Banyesheng (1173.5 kg/hm^2), Verde (1114.1 kg/hm^2), PI159764 (784.4 kg/hm^2), and Kim (655.1 kg/hm^2) (Table 2).

Correlation between maturity, plant height, seed weight, and yield

Pearson's correlation coefficients (r) between maturity, plant height, seed weight, and yield were computed. Maturity was almost not correlated with seed weight (r = 0.054) and yield (r = 0.019), respectively (Table 4). However, a medium correlation was found between maturity and plant height (r = 0.439). A low correlation was identified between seed weight and yield (r = 0.080). Seed weight was negatively correlated with plant height (r=-0.257). Plant height was not correlated with yield (r=-0.004) (Table 4).

For each trait, Pearson's correlation coefficients were also calculated between the three years. The correlation between years was high for maturity, seed weight, and plant height, and relatively moderate for yield. For maturity, the highest correlation was between data from 2008 and 2010 (r = 0.763) (Table 4), whereas the lowest one was between 2009 and 2010 (r = 0.657) (Table 4). For seed weight, the data were highly consistent over years since the lowest correlation was 0.876, corresponding to seed weight in 2008 and 2010 (Table 4). Yield data varied from year to year where the highest correlation was found between 2009 and 2009 (r = 0.604), and the lowest one between 2008 and 2010 (r = 0.484) (Table 4). Similarly to seed weight, data were also consistent over years for plant height where the lowest correlation was 0.848 for the data obtained in 2009 and 2010 (Table 4).

Population structure and genetic diversity analysis

SNP filtering yielded a total of 10,259 high-quality SNPs that were used for GWAS and GS. SNP number per chromosome varied from 292 (chromosome 12) to 785 (chromosome 18), with an average SNP number of 513 per chromosome (Table 5). There was also a large variation in the SNP density per chromosome with chromosome 16 having the lowest SNP density (73.29 kb between two adjacent SNPs) and chromosome 12 having the highest one (136.50 kb) (Table 5). The average distance between two adjacent SNPs was 94.86 kb. The average minor allele frequency (%) per chromosome ranged between 16.24% and 24.73%, with an average of 21.02% (Table 5). The average percentage of heterozygous SNPs per chromosome varied from 0.42–0.74%, with an average of 0.59%. There was not significant discrepancy in the average percentage of missing SNPs per chromosome (Table 5).

STRUCTURE Harvester indicated a delta K peak at K equal to 2 (Fig. 3A), indicating that the panel involving the 250 soybean genotypes consisted of two subpopulations. The bar plot from STRUCTURE 2.3.4 showed the two-well differentiated subpopulations where the subpopulation 1 was shown in red and the subpopulation shown in 2 in green (Fig. 3B). The first cluster consisted of 146 genotypes, accounting for 57.2% of the total population. The second subpopulation involved a total of 103 genotypes, which corresponded to 40.4% of the whole panel. In addition to the two distinct subpopulations, an admixture consisting of 6 genotypes was also identified and accounted for only 2.4% of the soybean panel used in this study. The mean inbreeding coefficients of the subpopulation relative to the total population were 0.565 and 0.043 for the subpopulation 1 and subpopulation 2, respectively (Table S2). The average distances between individuals in the same cluster were 0.332 (subpopulation 1) and 0.154 (subpopulation 2) (Table S2). The overall proportions of membership of a genotype within each cluster were 0.567 and 0.433 for subpopulation 1 and subpopulation 2, respectively (Table S2). The average allele frequency divergence among populations was 0.095 (Table S2).

A genetic diversity analysis was also carried out along with population structure as shown in Fig. 3C. In the phylogenetic tree, the subpopulation 1 was displayed using solid red circles, the subpopulation 2 was shown using solid green circles, and the admixture using solid blue circles (Fig. 3C). A good correlation was found between the genetic diversity analysis and the structure analysis. A large number of genotypes belonging to the subpopulation 1 was placed on the left-hand side of the phylogenetic tree (Fig. 3C), whereas most of the genotypes grouped into subpopulation 2 were clustered on the right part of the phylogenetic tree (Fig. 3C). The admixture genotypes were randomly scattered in the phylogenetic tree.

Genome-wide association study (GWAS) in maturity, plant height, seed weight and yield

Maturity

The significant SNPs associated with maturity varied between years (Fig. 4). When the maturity data were combined over three years, a total of 20 SNPs associated with maturity were found (Table 6). These

SNPs were distributed across the soybean genome (Fig. 4D) with chromosome 20 having the most significant SNPs associated with maturity (Table 6). The top five SNPs with the highest LOD values were Chr10_45903960 (LOD = 10.47, MAF = 38.31%), Chr13_33362588 (LOD = 6.83, MAF = 6.61%), and Chr09_30962080 (LOD = 6.53, MAF = 16.60%), Chr08_3672982 (LOD = 6.32, MAF = 22.53%), and Chr01_10725106 (LOD = 6.22, MAF = 14.04%), which were located on chromosomes 10, 13, 9, 8, and 1, respectively (Table 6). The t-test analysis conducted to compare the genotypic class from the aforementioned SNPs was significant expect for Chr13_33362588 (Figs. 8A-E). The SNP Chr10_45903960 was the only significant SNP which was consistent across three years with LOD values equal to 9.95, 3.22, 3.86, and 10.47 in 2008, 2009, 2010, and the combined data, respectively (Tables 6 and S3).

A total of 21, 16, and 33 SNPs associated with maturity were found in 2008, 2009, and 2010, respectively (Figs. 4A-4C) (Table S3). The top five SNPs found in 2008 were Chr20_43647960 (LOD = 11.87, MAF = 8.27%), Chr10_45903960 (LOD = 9.95, MAF = 38.31%), Chr01_9100366 (LOD = 7.92, MAF = 15.26%), Chr02_33386127 (LOD = 6.31, MAF = 27.60%), and Chr13_23406473 (LOD = 5.68, MAF = 5.60%), and located on chromosomes 20, 10, 1, 2, and 13, respectively (Fig. 4A) (Table S3). In 2009, the most significant SNPs associated with maturity were Chr04_46043518 (LOD = 7.77, MAF = 6.75%), Chr17_7918542 (LOD = 6.97, MAF = 13.01%), Chr03_35339536 (6.92, MAF = 34.44%), Chr16_5522373 (LOD = 5.93, MAF = 12.16%), and Chr08_19444425 (LOD = 5.52, MAF = 6.07%), which were found on chromosomes 4, 17, 3, 16, and 8, respectively (Fig. 4B) (Table S3). The data from 2010 provided with the largest number of SNPs among the three years. The top SNPs found from the 2010 maturity were Chr17_11801370 (LOD = 4.10, MAF = 20.78%), Chr03_2051343 (LOD = 3.90, MAF = 33.73%), Chr04_42265893 (LOD = 3.90, MAF = 15.97%), Chr10_45903960 (LOD = 3.87, MAF = 38.31%), and Chr12_38319250 (LOD = 3.87, MAF = 11.65%), located on chromosomes 17, 3, 4, 10, and 12, respectively (Fig. 4C) (Table S3). Despite the data from 2010 having the largest number of significant SNPS, the LOD values were slightly lower than the 2 other years. The SNP Chr17_7918542 was found in both 2008 and 2010. Of the 21 SNPs found in 2008, 6 were located on chromosomes 11 and 17 with 3 SNPs each (Table S3). In 2009, chromosome 4 harbored the most SNP (Table S3), and 6 SNPs were located on chromosome 6 for the data from 2010 (Table S3).

Plant height

The number and location of SNPs associated with plant height varied between years (Fig. 5). The data from 2009 suggested a total of 43 SNP markers associated with plant height, whereas the one from 2008 had a total of 18 significant SNPs (Table S3). On average, the significant SNPs associated with the 2010 plant height data had higher LOD values than the two other years (Table S3). The top five significant SNPs associated with plant height in 2008 were Chr05_4382157 (LOD = 11.80, MAF = 12.65%), Chr06_20321899 (LOD = 10.12, MAF = 46.64%), Chr19_44603046 (LOD = 7.56, MAF = 10.36%), Chr03_35230252 (LOD = 7.13, MAF = 22.13%), and Chr19_45769759 (LOD = 5.71, MAF = 23.53%), which were found on chromosomes 5, 6, 19, 3, and 19, respectively (Fig. 5A) (Table S3). Out of the 18 significant SNPs found for the 2008 data, 4 were mapped on chromosome 10 within a 600-Kb distance. In 2009, the

top five SNPs with the highest LOD values were Chr05_4341777 (LOD = 12.56, MAF = 18.11%), Chr19_45769759 (LOD = 8.70, MAF = 23.53%), Chr02_33386127 (LOD = 6.72, MAF = 27.60%), Chr05_27107594 (LOD = 6.32, MAF = 6.53%), and, Chr19_44603046 (LOD = 6.04, MAF = 10.36%), located on chromosomes, 5, 19, 2, 5, and 19, respectively (Fig. 5B). A total of 8 SNPs out of the 43 found in 2009 were located on chromosome 5. The SNPs with the highest LOD values associated with plant height in 2010 were Chr19_45359939 (LOD = 22.63, MAF = 20.24%), Chr05_4341777 (LOD = 9.72, MAF = 18.11%), Chr19_45270675 (LOD = 7.02, MAF = 14.06%), Chr08_3672982 (LOD = 6.85, MAF = 22.53%), and Chr02_11936018 (LOD = 6.78, MAF = 14.52%), which were located on chromosomes 19, 5, 19, 8, and 2, respectively (Fig. 5C) (Table S3). The two SNPs on chromosome 19 were within a distance of 80 Kb. The SNP marker Chr20_34864638 was the only one which was consistent over three years (LOD_2008 = 3.41, LOD_2009 = 4.59, LOD_2010 = 3.40, MAF = 21.37%). The SNPs Chr19_45769759, Chr19_44603046, and Chr05_27107594 were significant in both 2008 and 2009, and the SNP Chr05_4341777 was significant in both 2009 and 2010 (Table S3).

When the plant height data were combined over three years, a total of 31 SNPs were identified (Table 6). The top five SNPs associated with the combined data were Chr19_45769759 (LOD = 11.78, MAF = 23.53%), Chr05_4341777 (LOD = 7.97, MAF = 18.11%), Chr20_43000412 (LOD = 6.76, MAF = 43.83%), Chr19_45270675 (LOD = 6.60, MAF = 14.06%), and Chr19_44603046 (LOD = 6.21, MAF = 10.36%), which were found on chromosomes 19, 5, 20, 19, and 19, respectively (Fig. 5) (Table S3). The t-test analysis conducted to compare the data from the two genotypic classes defined by the aforementioned SNPs was significant (Figs. 8F-J). Despite the lack of consistency between the SNPs associated with plant height over three years, the region 44 Mb- 45 Mb on chromosome 19 displayed significant SNPs regardless of the year (Fig. 5), indicating that this region could contain important QTL(s) for plant height in soybean.

Seed weight

Similarly to the pattern of SNPs found for maturity and plant height, the SNP number and location varied between years for seed weight (Fig. 6). For the combined data (2008, 2009, and 2010), a total of 37 SNPs were found to be associated with seed weight (Table 7). The SNP Chr04_36949349 was the only one which was consistent across three years (LOD_2008 = 13.55, LOD_2009 = 9.01, LOD_2010 = 26.76, MAF = 16.47%) (Table S3). The top five significant SNPs were Chr04_36949349 (LOD = 22.58, MAF = 16.47%), Chr09_42124679 (LOD = 12.97, MAF = 35.90%), Chr01_42908212 (LOD = 10.47, MAF = 5.14%), Chr17_15088391 (LOD = 9.68, MAF = 23.08%), and Chr08_47483065 (LOD = 7.85, MAF = 25.83%), which were located on chromosomes 4, 9, 1, 17, and 8, respectively (Fig. 6) (Table 7). T-test analysis was significant for genotypic class comparison for these SNPs (Figs. 8K-O). Out of the 37 SNPs associated with the average seed weight over three years, 10 were located on chromosome 10 (Table 7).

A total of 29 SNPs were significantly associated with seed weight in 2008 (Table S3). Of the 29 SNPs, 7 were located on chromosome 10 with a genomic region of about 7 Mb. The top significant SNPs were Chr09_42124679 (LOD = 14.27, MAF = 35.90%), Chr04_36949349 (LOD = 13.55, MAF = 16.47%), Chr11_18308497 (LOD = 7.82, MAF = 16.61%), Chr18_54032147 (LOD = 6.66, MAF = 6.69%), and Chr08_47483065 (LOD = 6.44, MAF = 6.69%), located on chromosomes 9, 4, 11, 18, and 8, respectively

(Fig. 6A) (Table S3). The data from 2009 suggested a total of 30 SNPs associated with seed weight. The SNPs with the highest LOD values were Chr09_3931828 (LOD = 26.87, MAF = 6.91%), Chr02_23396978 (LOD = 17.05, MAF = 5.67%), Chr18_53835731 (LOD = 14.09, MAF = 14.96%), Chr13_34849223 (LOD = 13.62, MAF = 11.62%), and Chr11_32073461 (LOD = 13.27, MAF = 33.60%), found on chromosomes 9, 2, 18, 13, and 11, respectively (Fig. 6B) (Table S3). Interestingly, chromosome 9 harbored the largest number of SNPs for 2008, whereas chromosome 10 had the largest number of SNPs for the data from 2008 (Figs. 6A and 6B). The results from 2010 revealed the highest number of SNPs among the three years, where a total of 42 SNPs were found to be associated with seed weight (Fig. 6C). The top significant SNPs were Chr04_36949349 (LOD = 26.76, MAF = 16.47%), Chr02_19239630 (LOD = 15.27, MAF = 5.98%), Chr07_38110956 (LOD = 14.65, MAF = 33.05%), Chr09_42124679 (LOD = 12.65, MAF = 35.90%), and Chr08_47483065 (LOD = 10.16, MAF = 25.83%), which were found on chromosomes 4, 2, 7, 9, and 8, respectively (Table S3). The SNPs Chr09_42124679 and Chr08_47483065 were consistent in both 2008 and 2010, and the SNP Chr17_15088391 overlapped between the data from 2009 and 2010 (Table S3).

Yield

The number of significant SNPs associated with yield was consistent over years. A total of 17, 17, and 16 significant SNPs were found in 2008, 2009, and 2010, respectively (Table S3) (Fig. 7). However, the SNP locations were different between years. Only two significant SNPs were overlapping between years for the yield phenotype. The SNP Chr14_191942 was found in both 2009 and 2010, and the SNP Chr16_30202844 was identified in both 2008 and 2009. For the 2008 yield data, the top significant SNPs were Chr01_33020051 (LOD = 14.74, MAF = 9.02%), Chr18_1543178 (LOD = 13.59, MAF = 5.69%), Chr19_44507961 (LOD = 10.95, MAF = 24.80%), Chr13_7416534 (LOD = 7.25, MAF = 30.57%), and Chr07_7610107 (LOD = 7.13, MAF = 49.17%), located on chromosomes 1, 18, 19, 13, and 7, respectively (Fig. 7A).

For the 2009 data, the top SNPs associated with yield were Chr10_44639359 (LOD = 8.78, MAF = 29.41%), Chr11_10192448 (LOD = 5.27, MAF = 22.75%), Chr05_4106987 (LOD = 5.16, MAF = 14.22%), Chr18_51842219 (LOD = 4.98, MAF = 10.55%), and Chr16_30202844 (LOD = 4.93, MAF = 17.14%), which were found on chromosomes 10, 11, 5, 18, and 16, respectively (Table 7) (Fig. 7B). The average LOD values of the significant SNPs found in 2009 was lower than that of found in 2008 (Table S3).

The average LOD values of the significant SNPs in 2010 was also lower than that of found in 2008 but was almost similar to the 2009 results (Table S3). The SNP markers for the 2010 yield data were scattered across the soybean genome (Fig. 7C). The top significant SNPs associated with yield in 2010 were Chr08_43763537 (LOD = 9.83, MAF = 35.18%), Chr05_5758793 (LOD = 8.64, MAF = 6.69%), Chr18_7828843 (LOD = 7.90, MAF = 34.71%), Chr03_20676666 (LOD = 7.29, MAF = 14.16%), and Chr14_191942 (LOD = 6.38, MAF = 13.65%) (Table S3), which were located on chromosomes 8, 5, 18, 3, and 14, respectively (Fig. 7C).

When the yield data over three years were combined, a total of 23 significant SNPs were identified (Table 7) (Fig. 7D). The top significant SNPs for the combined data were Chr09_3204462 (LOD = 8.92,

MAF = 25.98%), Chr19_9586720 (LOD = 8.52, MAF = 27.97%), Chr08_47747059 (LOD = 8.43, MAF = 41.43%), Chr10_2089552 (LOD = 7.02, MAF = 24.31%), and Chr07_7610107 (LOD = 6.69, MAF = 49.17%), which were located on chromosomes 9, 19, 8, 10, and 7, respectively (Table 7). The variation between each genotypic class defined by the aforementioned SNPs was visualized in Figs. 8P-T. Of the 23 significant, 6 were located on chromosome 10 (Fig. 7D). Of the 6 significant SNPs found on chromosome 10, 3 were mapped within a 3-Mb genomic region (Table 7), indicating a strong likelihood of QTL(s) affecting soybean yield in this region.

Candidate genes selection

Candidate genes found within a 10-kb genomic region spanning a significant SNP associated with the combined data over three years for maturity, plant height, seed weight, and yield were investigated. For maturity, a total of 20 significant SNPs were identified (Table 6). Of which, 14 were mapped to genomic regions harboring annotated genes in Soybase (www.soybase.org). The annotated genes had a wide variety of functions. Leucine-rich repeat (LRR) domain was prevalent as shown in Table 6. The candidate genes associated with the most significant SNPs were Glyma.10g228900, Glyma.13g220200, Glyma.08g046800, Glyma.17g100600, and Glyma.08g176000, which encoded for LRR receptor-like protein kinase, F-box/leucine rich repeat protein, sensor histidine kinase, malate and lactate dehydrogenase, and replication protein A-related, respectively (Table 6). A candidate gene involved in epigenetics such as O-methyltransferase was also identified.

Of the 31 significant SNPs associated with the combined data for plant height, 25 were found in the vicinity of annotated genes (Table 6). Functional annotations of the candidate genes were diverse and mainly included transcription factors, kinases, and biomolecule transporters. The candidate genes found near the most significant SNPs were Glyma.19g200800, Glyma.05g048800, Glyma.19g195500, Glyma.19g187900, and Glyma.06g134200, encoding for transcription factor NF-Y alpha-related, cleavage and polyadenylation specificity factor, ubiquitin, PHD-finger, and serine/threonine-protein kinase, respectively (Table 6).

A total of 24 candidate genes were found for seed weight. Of which, 19 had functional annotations and one has an uncharacterized function (Table 7). GWAS for seed weight revealed a high probability of QTL(s) on chromosome 10; however, most of the SNPs found on the chromosome 10 were not mapped in the vicinity of an annotated gene (Table 7). Functional annotations related to the candidate genes had diverse functions. The candidate genes with defined functional annotations and associated with the most significant SNPs were Glyma.09g196700, LOC100801248, Glyma.20g181100, Glyma.02g161100, and Glyma.18g251800, which encoded for ring finger domain-containing, protein TIC 56 chloroplastic-like, LRR receptor-like protein kinase, camp-response element binding protein-related, and putative serine/threonine protein kinase, respectively (Table 7).

A total of 15 annotated genes were found near the significant SNPs associated with the combined data for yield over three years. Of which, 14 had functional annotations. Similarly to what was found for seed weight, only one candidate gene was identified on chromosome 10, which harbored 10 SNPs associated

with yield (Table 7). Most of the candidate genes encoded for transpiration factors and transferases. The candidate genes with functional annotations and found close to the most significant SNPs for yield were Glyma.09g038300, Glyma.08g366900, Glyma.07g082800, Glyma.18g021100, and Glyma.01g093300, encoding for calmodulin-binding transcription activator (CAMTA), zinc finger protein with krab and scan domains, homo-oligomeric flavin containing cys decarboxylase family, gamma-glutamyltransferase, and RNA recognition motif, respectively (Table 7).

Genomic selection (GS) in maturity, plant height, seed weight and yield

GS accuracy for maturity, plant height, seed weight, and yield were estimated. Using a 5-fold cross-validation within the 250-soybean accession panel, GS accuracy was found to be trait-dependent (Fig. 9) (Table S4). On average, the accuracy of GS was the highest for seed weight (0.84) and was the lowest for maturity (0.47) (Table S4). Variations in GS accuracy were found across years (Fig. 9). The between-year variation was important for yield where the highest accuracy was 0.72 in 2008 and the lowest one was 0.50 in 2010 (Fig. 9). The estimation of GS accuracy was also conducted using the two subpopulations derived from structure analysis. Cross-validation using all 250 soybean accessions and samples from the Q1 group provided similar trend as shown in Fig. 8. However, GS accuracy significantly dropped down for maturity when samples from the Q1 cluster were used (Fig. 9) (Table S4). Interestingly, GS averaged 0.64 with a less variation between traits and across years when samples from subpopulation 2 were used to estimate GS accuracy (Fig. 9). In addition, accuracy for predicting maturity was the best using Q2 samples (Fig. 9) (Table S4). The ANOVA results indicated statistically significant interaction effect between year and population type on the accuracy of GS for maturity (F value = 5.77, p-value = 0.0001), plant height (F value = 11.47, p-value < 0.0001), seed weight (F value = 76.17, p-value < 0.0001), and yield (F value = 12.82, p-value < 0.0001) (Table 8). The results indicated that year and population structure were important factors to take into account when incorporating GS in a breeding program.

GS accuracy using datasets from different years had similar trends to that of obtained from a within-year cross validation (Figs. 9 and 10). Overall, there is lack of consistency between GS accuracy and the training year across traits and population types. This could be explained by the significant interaction effect of population structure and years on GS of plant height (F value = 2.86, p-value = 0.0091), seed weight (F value = 4.56, p-value = 0.0001), and yield (F value = 17.37, p-value < 0.0001) (Table 9). Yield in 2010 was better predicted using dataset from 2009 than using yield data obtained in 2008 when all samples within the panel and individuals from Q2 were used for cross-validation, respectively (Fig. 10) (Table S5). However, plant height in 2010 was better predicted by the dataset recorded in 2008 for all sample- and Q1 sample-cross validation. Different results were found for plant height when cross-validation was carried out using individuals from the Q2 group (Fig. 10). Genomic prediction appeared to be year-independent but subpopulation-dependent for maturity (Fig. 10) (Table S5). ANOVA results showed that there was no significant interaction effect between population structure and year in predicting maturity (F value = 1.68, p-value = 0.1229) (Table 9). In addition, year effect on GS of maturity was not significant (F value = 0.88, p-value = 0.4519) (Table 9). However, GS for maturity was heavily

influenced by population structure (F value = 1232.94, p-value < 0.0001) (Table 9). Maturity could be better predicted than the other traits when samples from the group Q2 were used for cross-validation (Fig. 10). Despite the stability of maturity prediction between training and testing year as shown in Fig. 9, GS performed poorly for this trait when all samples and samples from the subpopulation 1 were used, respectively (Fig. 10) (Table S5).

GS was also conducted using samples from subpopulation 1 as training set and samples from subpopulation 2 as testing set and vice versa. For cross-validation using data from the same year (Table S6), discrepancy in GS accuracy was found when samples from one subpopulation was used to predict the ones from the other group (Fig. 11). Overall, selection accuracy was slightly higher for most traits when the GS model was trained using samples from subpopulation 1 (Fig. 11). This difference was substantial for maturity. The average GS accuracy for maturity was 0.49 and 0.20 for Q1-based training set and Q2-based population, respectively (Table S6). Unlike the two previous approaches where cross-validation was carried out using within-subgroup samples, variation of selection accuracy between years was more pronounced when prediction was done across subpopulations (Fig. 11). In addition, the interaction effect of population structure and year on GS was significant for maturity (F value = 89.30, p-value < 0.0001), plant height (F value = 14.02, p-value < 0.0001), seed weight (F value = 56.77, p-value < 0.0001), and yield (F value = 184.92, p-value < 0.0001) (Table 10). When the training set was taken from subgroup 1 in order to predict yield of subpopulation 2, selection accuracy was 0.72 and 0.41 for 2009 and 2008, respectively (Table S6). A similar trend was found for yield when the training samples were derived from subpopulation 2 (Fig. 11).

In addition to performing a within-year GS using two subpopulations, trait prediction for other years of a genetically- distant population subset was also conducted. Overall, results indicated a similar trend to what was found using a within-year GS approach (Figs. 11 and 12). The interaction effect of population structure and year on GS was significant for maturity (F value = 13.18, p-value < 0.0001), plant height (F value = 14.70, p-value < 0.0001), seed weight (F value = 264.27, p-value < 0.0001), and yield (F value = 25.23, p-value < 0.0001) (Table 11). Prediction accuracy for most traits was slightly higher when the samples from Q1 were used to train the GS model (Table S7). Training the model on Q1 resulted in prediction accuracy of yield being 0.62, 0.40, 0.36, and 0.41 when the training/testing year pair was 2008/2009, 2008/2010, 2009/2010, average 2008_2009/2010, respectively (Table S7). When the GS model was trained using samples from Q2, the prediction accuracy for yield was 0.28, 0.13, 0.31, and 0.21 corresponding to the training/testing year pair s2008/2009, 2008/2010, 2009/2010, average 2008_2009/2010, respectively (Table S7). Updating data on plant height, maturity date, and seed weight each year helped increase GS when the model was trained under Q2. The update was achieved by averaging the training set data from 2008 and 2009 to predict the testing set in 2010. GS accuracy was 0.47, 0.07, and 0.60 for plant height, maturity date, and seed weight, respectively, using the 2008 data from the Q2 samples to predict the 2010 data from the Q1 samples (Fig. 11) (Table S7). By taking the average data from 2008 and 2009 to establish the training set, plant height, maturity date, and seed weight of the Q2 samples were predicted with an accuracy of 0.52, 0.09, and 0.68, respectively (Table S7).

These results indicated that GS accuracy was impacted by multiple factors such as population structure and the variable year from which the training set was established.

Discussion

The soybean accessions used in this study were evaluated over 3 years for plant height, maturity, seed weight, and yield. These agronomic traits are key characters to improve in a soybean breeding program. Soybean varieties with good agronomic performance could be highly competitive in the soybean seed market. Efforts towards providing soybean growers with high performing varieties have been of upmost importance in a soybean breeding program. In this report, significant differences in plant height, maturity, seed weight, and yield were identified among the genotypes. In addition, the year effect was also significant. The design used in this research did not allow for an estimation of a possible genotype by year interaction effect. This will be conducted in a further study. However, the findings were in agreement with previously reported studies where the year effect can substantially impact plant height, maturity, seed weight, and yield in soybean [22–26]. In this report, seed weight was negatively correlated with plant height. This result was different from that of reported by Kato et al [25] who found that seed weight was positively correlated with plant height. Since the germplasm genotypes used by Kato et al [25] was different from the ones used in this investigation, the discrepancy found between the two studies suggested that agronomic trait should be investigated on a per population-type basis in a breeding program, which provides additional rationale to our study.

GWAS was conducted using a BLINK model. BLINK is one the latest and most improved statistical models to conduct GWAS [27]. Spurious associations could be reduced by incorporating population structure and Kinship effects into the GWAS model [28]. This has been established into the BLINK algorithm [29]. Therefore, we did not run the MLM (Q + K) of Tassel 5 [30] for GWAS since previous investigations had successfully demonstrated that BLINK had more statistical power in identifying true associations and reducing false positives for a large number of traits [27]. The only purpose of running population structure (Q) in this study was to assess GS accuracy between two unrelated subpopulations, which will be further discussed in this report.

A total of 20, 37, 31, and 23 SNPs were found to be significantly associated with maturity, seed weight, plant height, and yield, respectively, in soybean using the average data obtained over 3 years (Tables 6 and 7). Diers et al [31] also reported a similar range of SNP number associated with these traits using a nested association mapping (NAM) soybean population. A total of 19, 29, 15, and 23 SNPs were reported to be associated with maturity, seed weight, plant height, and yield, respectively [31]. Assefa et al. [32] found a total of 14, 10, and 9 SNPs associated with seed weight, plant height, and yield, respectively, upon a GWAS study involving a total of 419 soybean accessions. After SNP validation is performed, the information from this report could be used in trait introgression efforts in soybean breeding programs. This has been successfully demonstrated by Hegstad et al. [33] who introgressed large-effect QTL regions from commercial soybean cultivars with high yield into the Corteva Agriscience soybean accessions.

Previous investigations reported a total of more than 60 loci controlling maturity in Soybase (https://www.soybase.org/). Chromosome 16 has been shown to harbor the most significant loci affecting soybean maturity. Of the 20 SNPs found to be associated with the average maturity over 3 years in this study, one was mapped at 31 Mb on chromosome 16. Diers et al. [31] found a total of 19 regions for maturity on chromosome 16. However, Xia et al. [34] reported a significant discrepancy in SNPs associated with maturity across multiple environments, which was also consistent with that of reported in this study where different SNPs were reported in different years with different environmental conditions. A cluster of significant SNPs were identified in a 232-kb region of chromosome 20. This region spanned a significant SNP associated with maturity that was reported by Zatybekov et al. [35]. A high-LOD SNP (LOD > 10), Chr10_45903960, associated with maturity was also found on chromosome 10. A total of 10 loci on chromosome 10 were reported to be associated with maturity in Soybase (https://www.soybase.org/). One of those loci harbored the SNP Chr10_45903960. Two of the SNPs associated with maturity, Chr10_45903960 and Chr20_41339091, were located into the E2 and E4 loci that have been reported to control maturity in soybean [10]. For the plant height-related SNPs, previous investigations showed discrepancy in terms of SNP location. Zatybekov et al. [35] reported two SNPs significantly associated with plant height on chromosome 9 and 20, which were mapped at 42 Mb and 8 Mb, respectively. Our results did not indicate any plant-height associated SNPs on chromosome 9, whereas the one mapped on chromosome 20 is about 7.5 Mb away from that of reported by Zatybekov et al. [35]. Of the 68 loci associated with plant height in Soybase (https://www.soybase.org/), 19 loci were found on chromosome 19. The plant height-associated SNP with the highest LOD value found in this study was mapped on chromosome 19. We found that all significant SNPs associated with plant height and mapped on chromosome 19 overlapped with the Dt1 locus, which has been well-described for controlling plant height in soybean [5]. These findings suggest that the present investigation contributes towards enriching SNP markers associated with plant height, which is essential in efficiently establishing a breeding pipeline for agronomic trait improvement in soybean.

Zhang et al. [1] reported that seed weight was a complex trait controlled by a large number of loci. This statement appeared to be sound when taking into account the number of SNPs associated with maturity reported in this study. To date, more than 100 QTLs affecting seed weight have been reported (https://www.soybase.org/). A large number of these QTLs were mapped on chromosomes 2, 4, 5, 7, 17, 18, and 20 [31, 35]. Our results revealed a SNP with an LOD of 22.58 at 37 Mb location on chromosome 4, indicating the likelihood of a strong QTL affecting seed weight in this region. In addition, a large cluster of SNPs were found on chromosome 10. Some of which overlapped with previously reported seed weight-related QTLs (https://www.soybase.org/). For yield, previous reports showed that SNP markers associated with yield were scattered across the soybean genome. To date, more than 170 loci have been associated with yield in soybean (https://www.soybase.org/). Zatybekov et al. [35] mapped SNP markers associated with soybean yield on chromosomes 14, 17, and 20. Diers et al. [31] reported 23 loci affecting soybean yield on chromosome 16 alone. Two of the SNPs on chromosome 19 reported in this study were in the vicinity of a significant yield SNP marker identified by Assefa et al. [32]. Despite the lack of overlapping SNPs between traits, overlapping significant loci were identified to control two or more traits.

A 7.2-Mb region of chromosome 2 and defined by the SNP markers Chr02_12086588 and Chr02_19239630 were associated with both seed weight and plant height. A 9-Mb region of chromosome 4 harboring the SNPs Chr04_46043483, Chr04_46043518, and Chr04_36949349 was significantly associated with both maturity date and seed weight. A genomic DNA sequence spanning a 450-Kb region of chromosome 7 harbored the SNPs Chr07_33588669 and Chr07_7610107 and was associated with seed weight and yield, respectively. In addition, a 270-Kb region of chromosome 8 defined by the significant SNP markers Chr08_47483065 and Chr08_47747059 were associated with both seed weight and yield. One of the most important loci reported in this investigation is defined by an 11.5-Mb region of chromosome 10 containing the significant SNPs Chr10_18370776, Chr10_19170955, Chr10_19620114, Chr10_20805615, Chr10_24454215, Chr10_24773660, Chr10_29894008, Chr10_19477000, Chr10_24773517, Chr10_26343503, Chr10_27017034, and Chr10_29778879. These significant SNPs were associated with seed weight and yield. A 4-Mb region of chromosome of chromosome 19 was found to be associated with maturity, plant height, and yield in soybean, which was in agreement with a study investigated by Assefa et al. [32]. A 5.4-MB region of chromosome 20 harbored loci that were associated with maturity, plant height, and seed weight. These findings were consistent with previously reported studies [31, 35].

Our results suggested that maturity-associated candidate genes encoding for Leucine-rich repeat proteins were prevalent. Osakabe et al. [36] showed that LRR proteins acted as a key regulator for maturity in plant. In addition, Jinn et al. [37] reported that these proteins were also involved in floral organ abscission. These previous investigations supported that the LRR domains found in this study could be good candidate genes for maturity in soybean and could be further investigated towards validation. The findings indicated O-methyltransferase being a candidate gene for maturity in soybean. Held et al. [38] found that O-methyltransferase-related genes were highly expressed during cell maturation in maize. The candidate genes associated with plant height consisted of transcription factors, kinases, and biomolecule transporters. One of these transcription factors is NF-Y alpha-related. Zhao et al. [39] reported that this transcription factor regulated plant growth, indicating that NF-Y alpha-related could be a good candidate gene for plant height in soybean. The SNP markers associated with plant height and mapped in the Dt1 locus on chromosome 19 were in the vicinity of a gene that encodes for a tetratricopeptide repeat protein (TPR). However, the role of TPR domains in affecting plant height has been poorly investigated. The candidate genes associated with seed weight that were reported in this study had diverse functional annotations. A large number of candidate genes playing significant roles in seed development were hormone-signaling [40]. However, no candidate genes involved in hormone signaling pathways were identified. A large number of candidate genes found for yield were transcription factors and transferases. These results were consistent with that of reported by Diers et al. [31] who reported candidate transcription factors that could affect yield in soybean. Candidate genes associated with maturity, plant height, seed weight, and yield were reported in this study. Additional investigations would be required to validate these candidate genes.

GS accuracy for maturity, plant height, seed weight, and yield was assessed. Howard and Jarquin [41] reported that GS performed better than phenotypic selection in soybean when dealing with complex trait.

In this study, we found that the trend of GS accuracy was similar for all traits when all samples and Q1 samples, respectively, were used for cross validation (Figs. 8 and 9). Interestingly, GS for plant height was lower than yield when cross-validation was conducted using data from the same year from all samples and Q1 samples, respectively. These results were different from that of reported by Ma et al. [42] who also used a one-year data to estimate GS accuracy for plant height and yield in soybean. They reported an accuracy of 0.86 and 0.47 for plant height and yield, respectively. However, when cross-validation was conducted using data from different years regardless of the subpopulation, GS was higher for plant height than yield. This could be explained by the fact that plant height has higher heritability than yield [42], thus resulting in a higher genomic prediction accuracy. In addition, inconsistency in results has been found in previous studies investigating GS for yield in soybean. Jarquín et al. [21] suggested an accuracy of 0.64 for yield, whereas Stewart-Brown et al. [43] and Duhnen et al. [44] reported an accuracy of 0.26 and 0.39, respectively. We have also highlighted the effect of population structure on the accuracy of GS. The results indicated that GS accuracy for maturity was heavily affected by population structure. This could be explained by the fact maturity can cause a structure within a population, thus using maturity data from one subpopulation to predict the maturity data from another unrelated subpopulation would not work. We have also found that year can significantly affect GS, implying that updating the training model each year will be necessary for efficiently establishing a GS pipeline within a breeding program for agronomic trait improvement in soybean.

Conclusions

GWAS and GS were conducted for four agronomic traits: maturity, plant height, seed weight, and yield in an association panel consisted of 250 soybean accessions. A total of 20, 37, 31, and 23 SNPs were found to be significantly associated with maturity, seed weight, plant height, and yield, respectively, in this soybean panel using the average data obtained over 3 years. Glyma.10g228900, Glyma.19g200800, Glyma.09g196700, and Glyma.09g038300 were identified as candidate genes for maturity, plant height, seed weight, and yield, respectively. A 11.5-Mb region of chromosome 10 was associated with both seed weight and yield. The GS accuracy was trait-, year-, and population structure-dependent. The SNP markers for plant height, maturity, seed weight and yield can be used in molecular breeding to improve the four agronomic traits in soybean through MAS and GS. After validation, the candidate genes can be transferred to new soybean cultivars using the linked SNP markers through MAS. The high GS accuracy has confirmed that the four agronomic traits can be selected in soybean molecular breeding through GS.

Methods

Plant materials and phenotyping

A total of 250 soybean accessions were used in this study and they were originated collected from China, the United States, South Korea and Japan with 168 (67.2% out of 250), 76 (30.4%), 3 (1.2%), and 2 (0.8%) accessions, respectively plus with one with unknown origin (S1 Table). Soybean accessions were planted in Shijiazhuang (114°83'E, 38°03'N) in Hebei province, China with a randomized complete block design

(RCBD) with three replicates during the growing season of 2008, 2009, and 2010. A total of 36 soybean seeds were sown in two rows with a row length of 2 meters and space between rows was 0.5 meter. Phenotyping was conducted for maturity, plant height, seed weight, and yield.

Data distribution was visualized using the 'MASS' package of R v. 3.4.2 [45]. Distribution of each trait was displayed separately for each year and then combined. ANOVA for each trait was conducted using PROC MIXED of SAS® v. 9.4. The statistical model for the analysis was the following

$Y_{ij} = \mu + Gi + T_j + \varepsilon_{ij}$ with i = 1,2, … ,250 and j = 1,2,3

Y_{ij} was the response from the i^{th} genotype in the j^{th} year, μ was the overall mean, G_i represented the effect of the i^{th} genotype (fixed effect), T_j was the effect of the j^{th} year (fixed effect), and ε_{ij} was the experimental error associated with the ij^{th} observation.

Genotyping

DNA was extracted from young soybean leaves using the CTAB (hexadecyltrimethyl ammonium bromide) method [46]. DNA library was prepared using the restriction enzyme ApeKI following the GBS protocol described by Elshire et al. [47] and DNA sequencing was performed using GBS method [47, 48]. The 90-bp, double-end sequencing was performed on each soybean genotype using the GBS protocol by an Illumina HiSeq at the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. The GBS dataset contained 3.26 M short-reads or 283.74 Mbp of sequence for each accession. The short reads were aligned to the soybean whole genome sequence (Wm82.a1.v1)

(https://www.soybase.org/GlycineBlastPages/archives/Gma1.01.20140304.fasta.zip; https://www.soybase.org/GlycineBlastPages/index.php?db_select=Gma1.01) using SOAPaligner/soap2 (http://soap.genomics.org.cn/) and SOAPsnp v. 1.05 was used for SNP calling [49, 50]. Approximately a half million SNPs were identified upon SNP calling. Minor allele frequency (MAF) threshold was set to 5%, SNP with heterozygosity more than 10% was discarded, and SNP with missing data more than 15% were also removed. Upon SNP filtering, a total of 10,259 high quality SNPs were retained and used for further analysis.

Population structure analysis

Population structure was inferred using STRUCTURE 2.3.4 through a Bayesian resampling technique [51]. Out of the 10,259 filtered SNPs, a total of randomly chosen 5,129 SNPs were used for inferring population structure (K). Using a randomly selected SNP for conducting a population structure analysis is a valid approach when the use of the complete set of SNPs is computationally heavy as described by Huang et al. [52]. Analysis was run using an admixture model along with a correlated allele frequency model, which was independent of each run [53].

A total of ten runs were performed for each estimated K. The Markov chain Monte Carlo (MCMC) length of the burn-in period was 30,000 and the number of MCMC iterations was 50,000. The identification of

optimal K was done using STRUCTURE Harvester [54, http://taylor0.biology.ucla.edu/structureHarvester/] and based upon the equation established by Evanno et al [55]. The optimal K was that of corresponding the highest delta K value.

The inferred population structure K was used to generate the structure Q-matrix consisting of K vectors. Each soybean accession was assigned to a Q cluster using a cut-off probability of 0.55. Any soybean accessions that could not be grouped in any of the clusters would be considered as admixture. Population structure was visualized using STRUCTURE PLOT with the option “sort by Q” [56].

Genetic diversity analysis

The genetic diversity analysis and the establishment of the phylogenetic tree were done using MEGA 7 [57]. Statistical inference was achieved using a maximum likelihood procedure and the parameters for drawing the phylogenetic tree was the following: Phylogeny Reconstruction; Statistical method: Maximum Likelihood; Test of phylogeny: None; Substitutions type: Nucleotide; Model/Method: Tamura-Nei Model; Rates among sites: Gamma distributed with Invariant sites (G + I); No of Discrete Gamma Categories: 5; Gaps/Missing Data treatment: Use all sites; ML Heuristic Method: Nearest-Neighbor-Interchange (NNI); Initial Tree for ML: Make initial tree automatically (Default - NJ/BioNJ); Branch Swap Filter: Moderate; Number of threads: 1.

The Q-matrix component was imported into MEGA 7 and was used along with the genetic diversity analysis when drawing the phylogenetic tree. Consistency or discrepancy between the population structure analysis and the genetic diversity analysis could be easily captured by doing so. The shape of “node/subtree marker” and the “branch line” had the same color as the bar plots obtained from STRCUTURE PLOT for the sub-trees of each cluster (Q).

Genome-wide study analysis (GWAS)

GWAS was conducted using a Bayesian Information and Linkage Disequilibrium Iteratively Nested Keyway (BLINK) and run in R using the package ‘BLINK’ [27]. Significant SNPS were those with an LOD value greater than 3 [58].

BLINK was an improved model version of Fixed and Random Model Circulating Probability Unification (FarmCPU) and has been shown to be statistically powerful and efficient in identifying significant SNPs associated with trait of importance [27]. FarmCPU involved a fixed effect model (FEM) and a random effect model (REM), which were run iteratively. FarmCPU assumed an even distribution of markers within the genome. However, this assumption was relaxed in BLINK where a Linkage Disequilibrium information was used instead [27]. The REM part was replaced by a second FEM in BLINK. The two FEM models used in BLINK were the following.

FEM (1): $y_i = M_{i1}b_1 + M_{i2}b_2 + \ldots + M_{ik}b_k + M_{ij}d_j + e_i$

FEM (2): $y_i = M_{i1}b_1 + M_{i2}b_2 + \ldots + M_{ij}b_j + e_i$

where y_i was the phenotypic data from the i^{th} sample; M_{i1},$M_{i2}b_2$, …, M_{ik} were the genotypes of k pseudo QTNs, which were initially empty and with effects b_1, b_2, …, b_k, respectively; M_{ij} represented the j^{th} genetic marker of the i^{th} sample; and e_i was the residual having a distribution with mean zero and a variance σ^2_e.

Candidate gene discovery

A 10 kb-genomic region spanning a significant SNP was used for candidate gene search using the G. max Williams 82.a2 reference in Soybase (https://www.soybase.org/). The SNPs associated with the combined data over three years for maturity date, plant height, seed weight, and yield were used for candidate gene discovery. Functional annotations pertaining to each postulated candidate gene were also investigated using Soybase (https://www.soybase.org/).

Genomic-selection and cross-validation

GS was conducted using a ridge regression best linear unbiased predictor (rrBLUP) model. rrBLUP has been shown to be effective in estimating the effects of loci controlling complex traits [59]. The rrBLUP model was $y = W\beta + \varepsilon$ [17]. In this equation, y represented the vector phenotypic data; β denoted the marker effect with $\beta \sim N(0,I\sigma^2_\beta)$; W represented the incidence matrix relating the genotype to the vector phenotype; G was the matrix displaying the genetic marker; and ε referred to the random error. The solution for rrBLUP was $\beta\_hat = (Z^TZ + I\lambda)^{-1}Z^Ty$ with $Z = WG$. The ridge parameter was defined as $\lambda = \sigma^2_e/\sigma^2_\beta$, where σ^2_e was the residual variance and σ^2_β the marker effect variance. rrBLUP was carried out in R using the 'rrBLUP' package [60].

Cross-validation was conducted using 4 different approaches. The first approach consisted of sampling accessions from and cross-validating within the 255 soybean genotypes (whole panel), the genotypes belonging to subpopulation 1 (Q1) from structure analysis (Q1 panel), and the genotypes from subpopulation 2 (Q2) from structure analysis (Q2 panel), respectively. For each defined subgroup, the training dataset was taken from a year and the validation dataset was also extracted from the same year. The second approach differed from the first one in a way that the training dataset from a year was used to predict the genotype's performance in the succeeding year(s). The third strategy used samples from the Q1 panel to predict the Q2 samples' performance within the same year and vice versa. The fourth methodology consisted of training the GS model using dataset from the Q1 panel from a particular year that was used to predict the traits of the Q2 panel from another year and vice versa. Due to the relatively small sample size of each panel, a 5-fold cross-validation was carried out for GS involving the whole panel and the Q1 panel, and a 3-fold cross validation was performed for GS using the Q2 panel. Doing so allowed for an accurate estimation of the Person's correlation coefficient (GS accuracy) that was established based on a sample size (> 30) that could be statistically valid under such constraint.

ANOVA was conducted to assess the interaction effect of year and population type (whole panel, Q1 panel, and Q2 panel) on GS accuracy using PROC MIXED of SAS® v. 9.4. The statistical model for this analysis was the following.

$Y_{ijk} = \mu + T_i + P_j + + YP_{ij} + \varepsilon_{ijk}$ with i = 1,2,3, j = 1,2,3, and k = 1,2,3,…,100

Y_{ijk} was the GS accuracy from the i^{th} year using the j^{th} population type at the k^{th} replication, T_i represented the effect of the i^{th} year (fixed effect), P_j was the effect of the j^{th} population type (fixed effect), and ε_{ijk} was the experimental error associated with the ijk^{th} observation.

Declarations

Ethics approval and consent to participate

All data and materials are not related to human and animals. All research materials of soybean germplasm accessions are obtained from Chinese Academy Of Agricultural Sciences. This research is not related to any plant specimens to be deposited as vouchers or any other association for this section.

Consent to publish

Not applicable

Availability of data and materials

SNP data can be found in the ENA using accession number PRJEB34546 (https://www.ebi.ac.uk/ena/data/view/PRJEB34546) and other data supporting this article have been listed in the supplementary materials

Competing interests

The authors declare that they have no competing interests.

Funding

(1) National key R&D program of China (2017YFD0101403); (2) The construction specialty program of industrial technology system of national modern agriculture (CARS-04-PS06); and (3) National Natural Science Foundation of China (31471522). (4) Key R & D Program Project in Hebei Province (16227516d).

Author contributions

JQ carried out genotyping. CY, FW, YF, and YM carried out phenotyping. RW, AS, and JQ analyzed the data. RW composed the draft of the manuscript. MZ and CY directed and managed this research. AS and JQ reviewed and edited the manuscript. All authors read, corrected and approved the final manuscript.

Acknowledgements

The authors would like to thank Prof. Lijuan Qiu (Chinese Academy Of Agricultural Sciences) for providing 250 soybean germplasm accessions seeds.

References

1. Zhang J, Song Q, Cregan PB, Jiang GL. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). Theor Appl Genet. 2016;129(1):117–130.
2. Cao Y, Li S, He X, Chang F, Kong J, Gai J, Zhao T. Mapping QTLs for plant height and flowering time in a Chinese summer planting soybean RIL population. Euphytica. 2017;213(2):39.
3. Yao D, Liu ZZ, Zhang J, Liu SY, Qu J, Guan SY, Pan LD, Wang D, Liu JW, Wang PW. Analysis of quantitative trait loci for main plant traits in soybean. Genet Mol Res. 2015;14(2):6101–6109.
4. Copley TR, Duceppe MO, O'Donoughue LS. Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines. BMC Genomics. 2018;19(1):167.
5. Fang C, Ma Y, Wu S, Liu Z, Wang Z, Yang R, Hu G, Zhou Z, Yu H, Zhang M, Pan Y. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol. 2017;18(1):161.
6. Hu Z, Zhang D, Zhang G, Kan G, Hong D, Yu D. Association mapping of yield-related traits and SSR markers in wild soybean (*Glycine soja* Sieb. and Zucc.). Breed Sci. 2014;63(5):441–449.
7. Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol J. 2015;13(2):211–221.
8. Zhang H, Hao D, Sitoe HM, Yin Z, Hu Z, Zhang G, Yu D. Genetic dissection of the relationship between plant architecture and yield component traits in soybean (*Glycine max*) by association analysis across multiple environments. Plant Breed. 2015;134(5):564–572.
9. Zuo Q, Hou J, Zhou B, Wen Z, Zhang S, Gai J, Xing H. Identification of QTLs for growth period traits in soybean using association analysis and linkage mapping. Plant Breed. 2013;132(3):317–323.
10. Miladinovic J, Ćeran M, Đorđević V, Balešević-Tubić S, Petrović K, Đukić V, Miladinović D. Allelic variation and distribution of the major maturity genes in different soybean collections. Front Plant Sci. 2018;9:1286.
11. Contreras-Soto RI, Mora F, de Oliveira MAR, Higashi W, Scapim CA, Schuster I. A Genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. PLoS one. 2017;12(2):e0171105.
12. Zhang J, Song Q, Cregan PB, Nelson RL, Wang X, Wu J, Jiang GL. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. BMC Genomics. 2015;16(1):217.
13. Hao D, Cheng H, Yin Z, Cui S, Zhang D, Wang H, Yu D. Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. Theor Appl Genet. 2012;124(3):447–458.
14. Wang J, Chu S, Zhang H, Zhu Y, Cheng H, Yu D. Development and application of a novel genome-wide SNP array reveals domestication history in soybean. Sci Rep. 2016;6(1):20728.

15. Yan L, Hofmann N, Li S, Ferreira ME, Song B, Jiang G, Ren S, Quigley C, Fickus E, Cregan P, Song Q. Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. BMC Genomics. 2017;18(1):529.
16. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, Fang C. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33(4):408–414.
17. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;257(4):1819–1829.
18. Heffner EL, Jannink JL, Sorrells ME. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome. 2011;4(1):65–75.
19. Matei G, Woyann LG, Milioli AS, de Bem Oliveira I, Zdziarski AD, Zanella R, Coelho AS, Finatto T, Benin G. Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. Mol Breed. 2018;38(9):117.
20. Duhnen A, Gras A, Teyssèdre S, Romestant M, Claustres B, Daydé J, Mangin B. Genomic Selection for Yield and Seed Protein Content in Soybean: A study of breeding program data and assessment of prediction accuracy. Crop Sci. 2017;57(3):1325.
21. Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics. 2014;15(1):740.
22. Baig D, Khurshid H, Arshad M, Jan SA, Khan MA, Nawaz N. Evaluation of soybean genotypes for yield and other economically important traits under rainfed condition . Pure Appl Biol. 2018;7:1-7.
23. Dalló SC, Zdziarski AD, Woyann LG, Milioli AS, Zanella R, Conte J, Benin G. Across year and year-by-year GGE biplot analysis to evaluate soybean performance and stability in multi-environment trials. Euphytica. 2019;215(6):113.
24. Jiang GL, Rutto LK RS. Evaluation of soybean lines for edamame yield traits and trait genetic correlation. HortScience. 2018;53(12):1732–1736.
25. Kato S, Sayama T, Ishimoto M, Yumoto S, Kikuchi A, Nishio T. The effect of stem growth habit on single seed weight and seed uniformity in soybean (*Glycine max* (L.) Merrill). Breed Sci. 2018;68(3):352–359.
26. Wiggins B, Wiggins S, Cunicelli M, Smallwood C, Allen F, West D, Pantalone V. Genetic gain for soybean seed protein, oil, and yield in a recombinant inbred line population. J Am Oil Chem Soc. 2019;96(1):43–50.
27. Huang M, Liu X, Zhou Y, Summers RM, Zhang Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience. 2019;8(2):giy154.
28. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38(2):203.
29. Liu X, Huang M, Fan B, Buckler ES, Zhang Z, Bradbury P. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet.

2016;12(2):e1005767.

30. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633–2635.
31. Diers BW, Specht J, Rainey KM, Cregan P, Song Q, Ramasubramanian V, Graef G, Nelson R, Schapaugh W, Wang D, Shannon G. Genetic architecture of soybean yield and agronomic traits. G3 Genes, Genomes, Genet. 2018;8(10):3367–3375.
32. Assefa T, Otyama PI, Brown AV, Kalberer SR, Kulkarni RS, Cannon SB. Genome-wide associations and epistatic interactions for internode number, plant height, seed weight and seed yield in soybean. BMC Genomics. 2019;20(1):527.
33. Hegstad JM, Nelson RL, Renny-Byfield S, Feng L, Chaky JM. Introgression of novel genetic diversity to improve soybean yield. Theor Appl Genet. 2019;132(9):2541–2552.
34. Xia Z, Wang Y, Li Y, Wu H, Hu B, Zheng J, Zhai H, Lv S, Liu X, Chen X, Qiu H. Genotyping of soybean cultivars with medium-density array reveals the population structure and QTNs underlying maturity and seed traits. Front Plant Sci. 2018;9:610.
35. Zatybekov A, Abugalieva S, Didorenko S, Gerasimova Y, Sidorik I, Anuarbek S, Turuspekov Y. GWAS of agronomic traits in soybean collection included in breeding pool in Kazakhstan. BMC Plant Biol. 2017;17(S1):179.
36. Osakabe Y, Maruyama K, Seki M, Satou M, Shinozaki KYSK. Leucine-rich repeat receptor-like kinase1 is a key membrane-bound regulator of abscisic acid early signaling in *Arabidopsis*. Plant Cell. 2005;17(4):1105–1119.
37. Jinn TL, Stone JM, Walker JC. HAESA, an *Arabidopsis* leucine-rich repeat receptor kinase, controls floral organ abscission. Genes Dev. 2000;14(1):108–117.
38. Held BM, Wang H, John I, Wurtele ES, Colbert JT. An mRNA putatively coding for an O-methyltransferase accumulates preferentially in maize roots and is located predominantly in the region of the endodermis. Plant Physiol. 1993;102(3):1001–1008.
39. Zhao H, Wu D, Kong F, Lin K, Zhang H, Li G. The *Arabidopsis thaliana* nuclear factor Y transcription factors. Front Plant Sci. 2017;7:2045.
40. Jing Y, Zhao X, Wang J, Teng W, Qiu L, Han Y, Li W. Identification of the genomic region underlying seed weight per plant in soybean (*Glycine max* L. Merr.) via high-throughput single-nucleotide polymorphisms and a genome-wide association study. Front Plant Sci. 201811;9:1392.
41. Howard R, Jarquin D. Genomic prediction using canopy coverage image and genotypic information in soybean via a hybrid model. Evol Bioinformatics. 2019;15:117693431984002.
42. Ma Y, Reif JC, Jiang Y, Wen Z, Wang D, Liu Z, Guo Y, Wei S, Wang S, Yang C, Wang H. Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). Mol Breed. 2016;36(8):113.
43. Stewart-Brown BB, Song Q, Vaughn JN, Li Z. Genomic selection for yield and seed composition traits within an applied soybean breeding program. G3 Genes, Genomes, Genet. 2019;9(7):2253–2265.

44. Duhnen A, Gras A, Teyssèdre S, Romestant M, Claustres B, Daydé J, Mangin B. Genomic selection for yield and seed protein content in soybean: a study of breeding program data and assessment of prediction accuracy. Crop Sci. 2017;57(3):1325–1337.
45. Team RDC. R: a language and environment for statistical computing. r foundation for statistical computing, Vienna. 2011
46. Kisha TJ, Sneller CH, Diers BW. Relationship between genetic distance among parents and genetic variance in populations of soybean. Crop Sci. 1997;37(4):1317–1325.
47. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS one. 2011;6(5):e19379.
48. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, Normandeau É, Laroche J, Larose S, Jean M, Belzile F. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS one. 2013;8(1):e54603.
49. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009;19(6):1124-1132.
50. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-2993.
51. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–959.
52. Huang L, Zeng A, Chen P, Wu C, Wang D, Wen Z. Genomewide association analysis of salt tolerance in soybean [*Glycine max* (L.) Merr.]. Plant Breed. 2018;137(5):714–720.
53. Shi A, Buckley B, Mou B, Motes D, Morris JB, Ma J, Xiong H, Qin J, Yang W, Chitwood J, Weng Y. Association analysis of cowpea bacterial blight resistance in USDA cowpea germplasm. Euphytica. 2016;208(1):143–155.
54. Earl DA, VonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 2011;4(2):359–361.
55. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005;14(8):2611–2620.
56. Ramasamy RK, Ramasamy S, Bindroo BB, Naik VG. STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. Springerplus. 2014;3(1):431.
57. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis Version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870–1874.
58. Kaler AS, Dhanapal AP, Ray JD, King CA, Fritschi FB, Purcell LC. Genome-wide association mapping of carbon isotope and oxygen isotope ratios in diverse soybean genotypes. Crop Sci. 2017;57(6):3085–3100.
59. Haile JK, N'Diaye A, Clarke F, Clarke J, Knox R, Rutkoski J, Bassi FM, Pozniak CJ. Genomic selection for grain yield and quality traits in durum wheat. Mol Breed. 2018;38(6):75.

60. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome J. 2011;4(3):250–255.

Tables

Due to technical limitations, the tables are only available as a download in the supplemental files section.

Supplementary File

Table S1. List of genotypes evaluated for maturity, plant height, seed weight, and yield over three years. SD represents the standard deviation for each trait over three years

Table S2. Mean Fst values, average distance between samples within the same subpopulation, average probability value of from each individual within each cluster, and allele frequency divergence among populations.

Table S3. Significant SNPs associated with the average maturity date, plant height, seed weight, and yield in 2008, 2009, and 2010, chromosome and physical position of the significant SNPs, LOD(-log10(p-value)) values, minor allele frequency at the SNP locus, and gene ID and functional annotation.

Table S4. Genomic selection accuracy for maturity, plant height, seed weight, and yield using 100 replications and where cross-validation was performed within all samples, samples from subpopulation Q1, and samples from subpopulation Q2, respectively.

Table S5. Genomic selection accuracy where cross-validation was performed within all samples, samples from subpopulation Q1, and samples from subpopulation Q2, respectively. Cross-validation within each group was conducted as following. The data from 2008 were used to predict the data from 2009 and 2010, respectively, the data from 2009 were used to predict the data from 2010, and the average data from 2008 and 2009 were used to predict the data from 2010.

Table S6. Genomic selection accuracy for maturity, plant height, seed weight, and yield using samples from subpopulation 1 (Q1) as training set and samples from subpopulation 2 (Q2) as testing set and vice versa. Estimation of genomic selection accuracy was done using 100 replications.

Table S7. Genomic selection for maturity, plant height, seed weight, and yield using samples from subpopulation 1 (Q1) as training set and samples from subpopulation 2 (Q2) as testing set, and vice versa. Data from 2008 in the training set were used to predict that of 2009 and 2010 in the testing set, respectively. Data from 2009 in the training set were used to predict that of 2010 in the testing set. The average data from 2008 and 2009 in the training set were used to predict that of 2010 in the testing set.

Figures

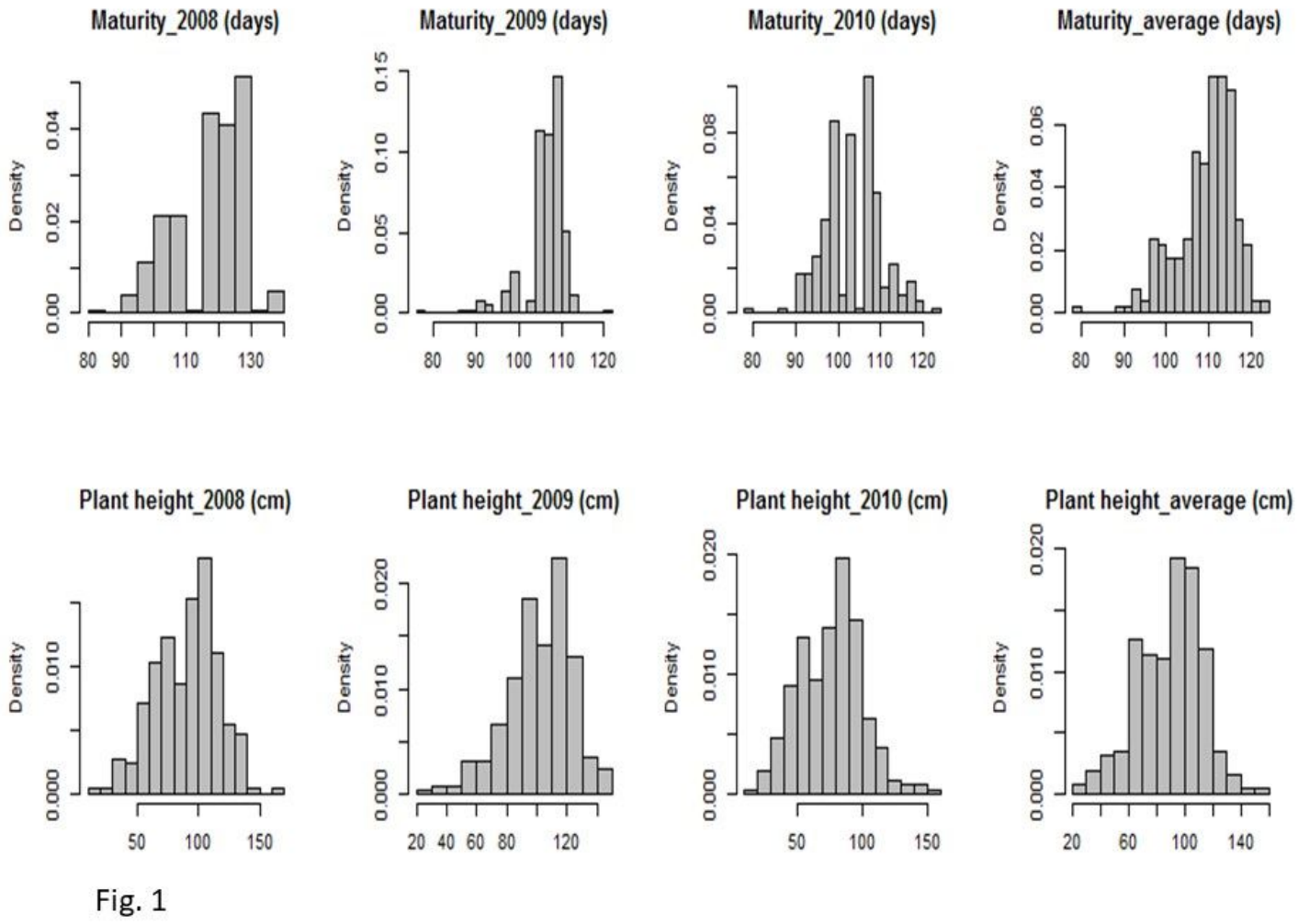

Fig. 1

Figure 1

Distribution of maturity and plant height in 2008, 2009, and 2009, and the average over the 3-year data.

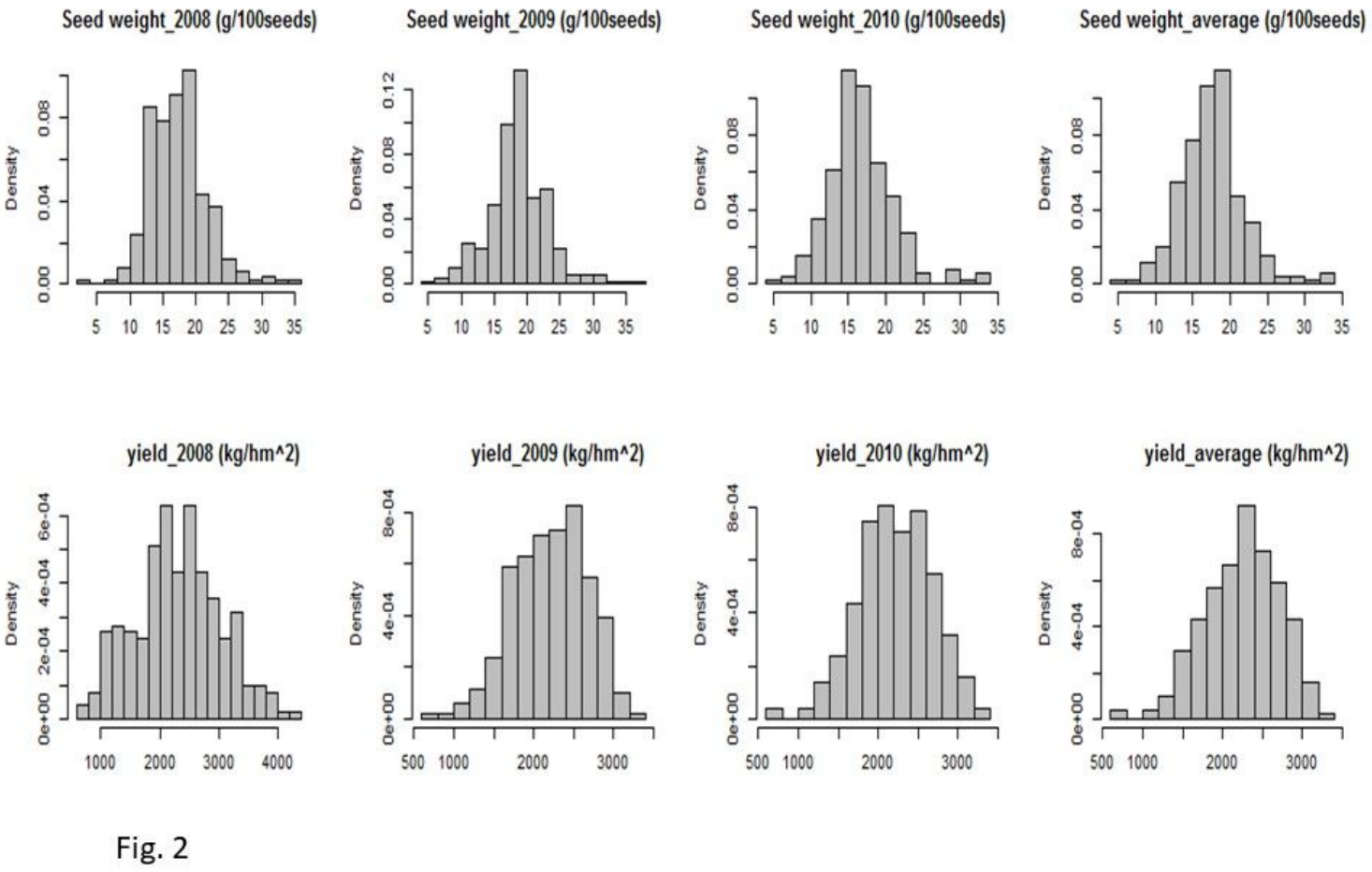

Fig. 2

Figure 2

Distribution of seed weight and yield in 2008, 2009, and 2009, and the average over the 3-year data.

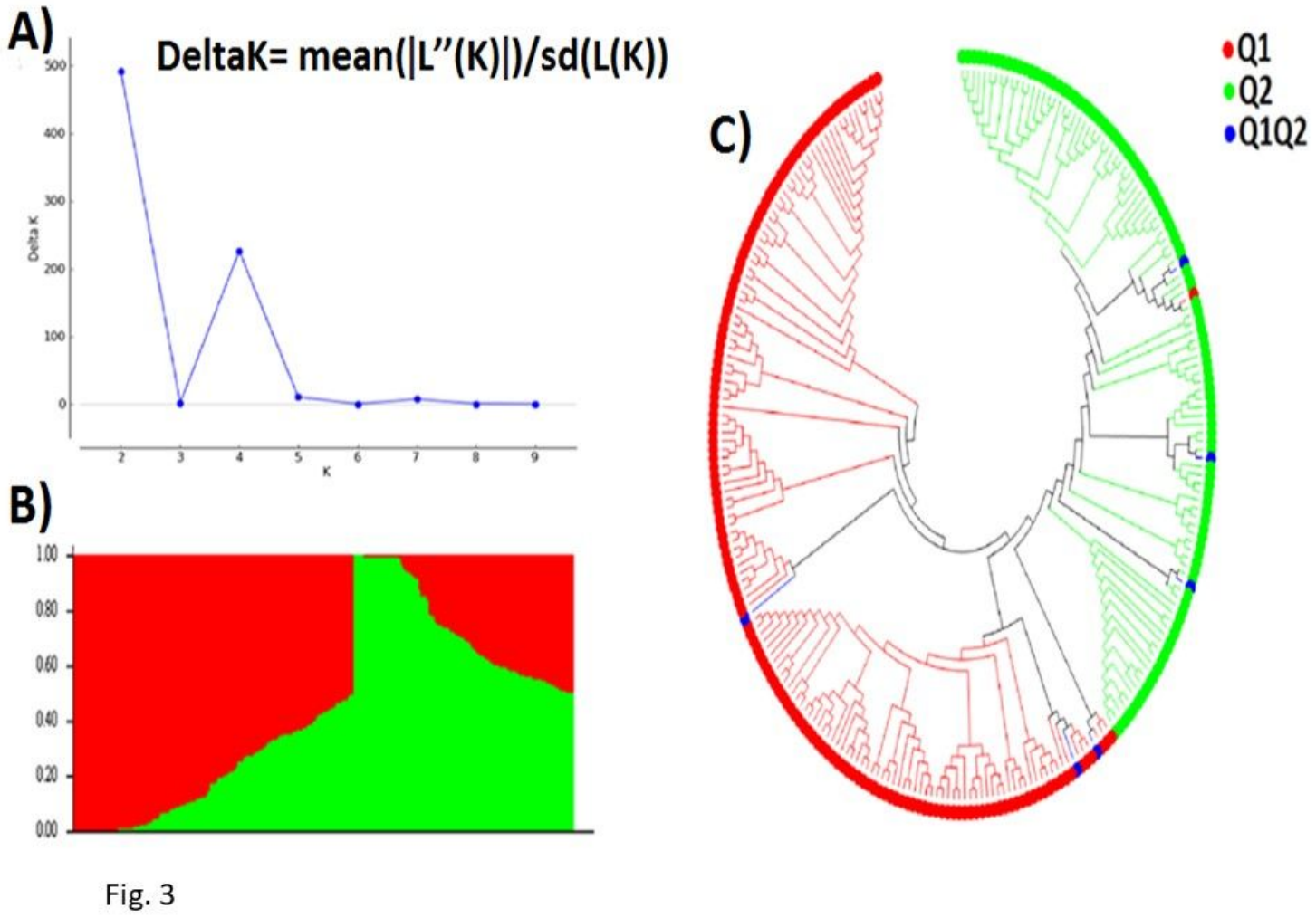

Fig. 3

Figure 3

Population structure and genetic diversity analysis. (A) Plot showing the delta K values on the y-axis and the corresponding K values on the x-axis. The plot was obtained from STRUCTURE Harvester (Earl and VonHoldt, 2011; http://taylor0.biology.ucla.edu/structureHarvester/). The delta K peak corresponds to K=2. (B) Bar plot showing the population structure using STRUCTURE 2.3.4 (Pritchard et al., 2000) where the red color corresponds to cluster 1 and the green one to cluster 2. The y-axis of the bar plot indicates the proportion of membership of a genotype to each cluster. (C) Phylogenetic tree involving a combined analysis between population structure and genetic diversity. The subpopulation 1 is represented by the solid red circles. The subpopulation population 2 is indicated by the solid green circles. Admixture is represented by the solid blue circles.

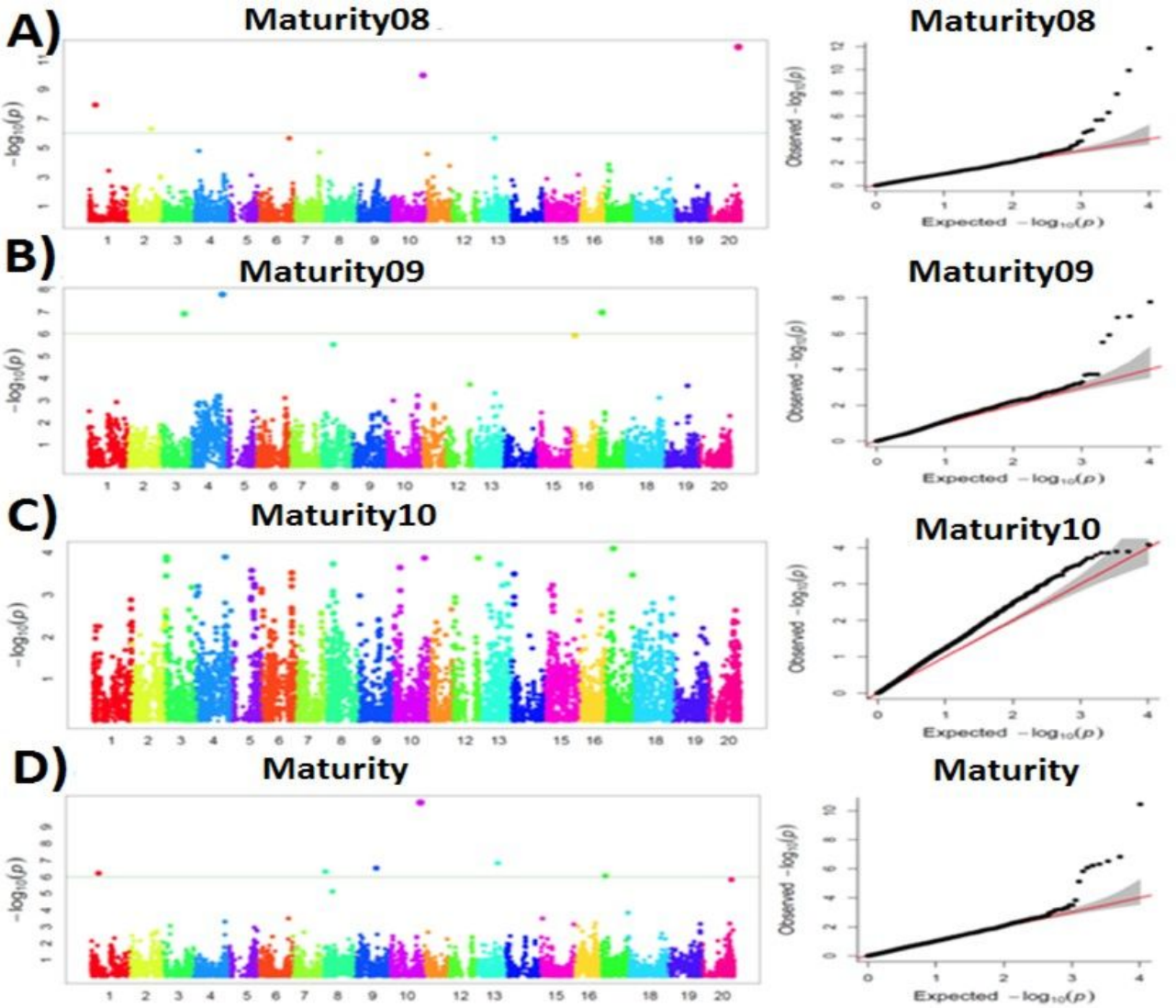

Fig. 4

Figure 4

Manhattan plots and QQ-plots for maturity in 2008 (A), 2009 (B), 2010 (C), and the average data over 3 years (D).

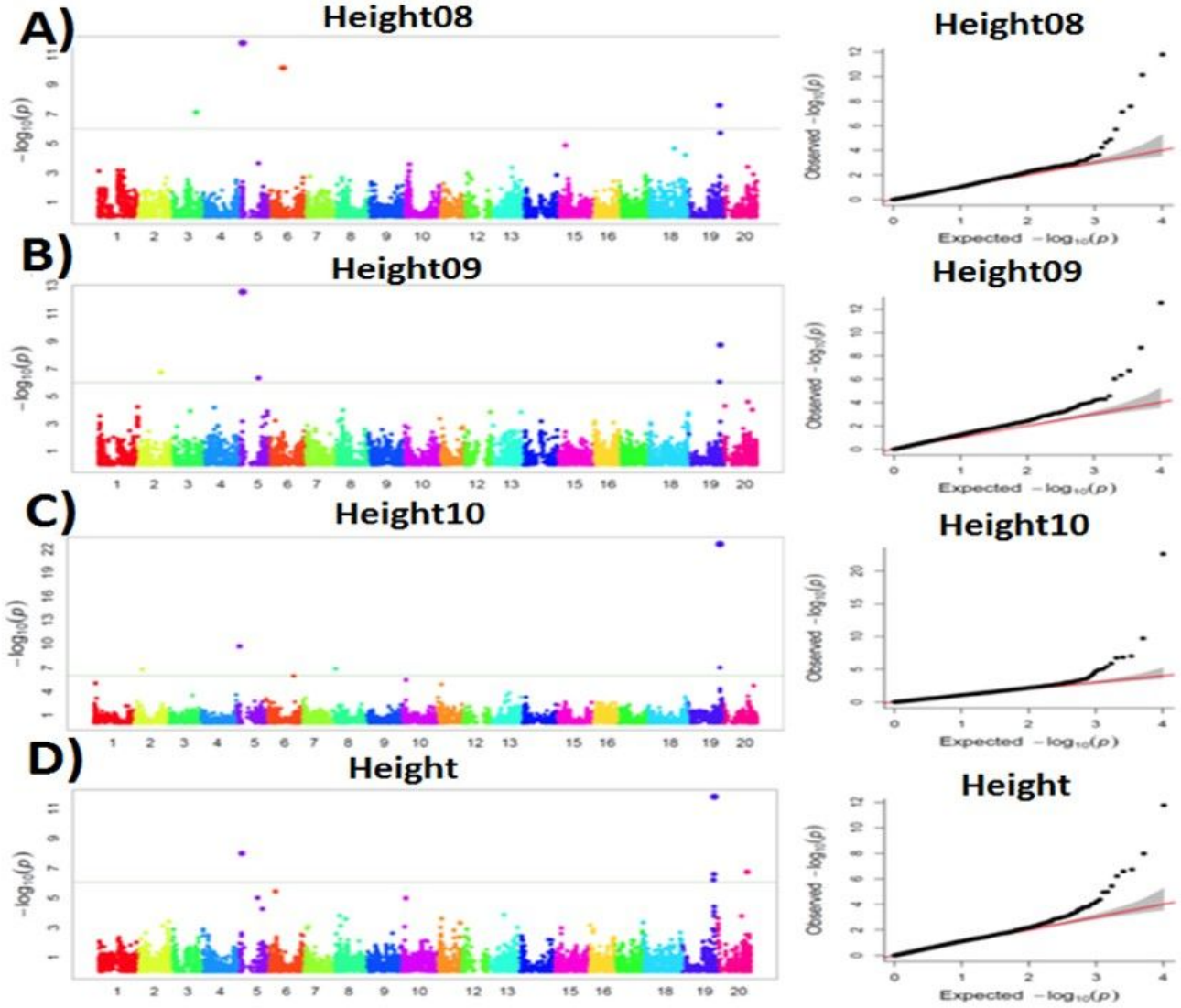

Fig. 5

Figure 5

Manhattan plots and QQ-plots for plant height in 2008 (A), 2009 (B), 2010 (C), and the average data over 3 years (D).

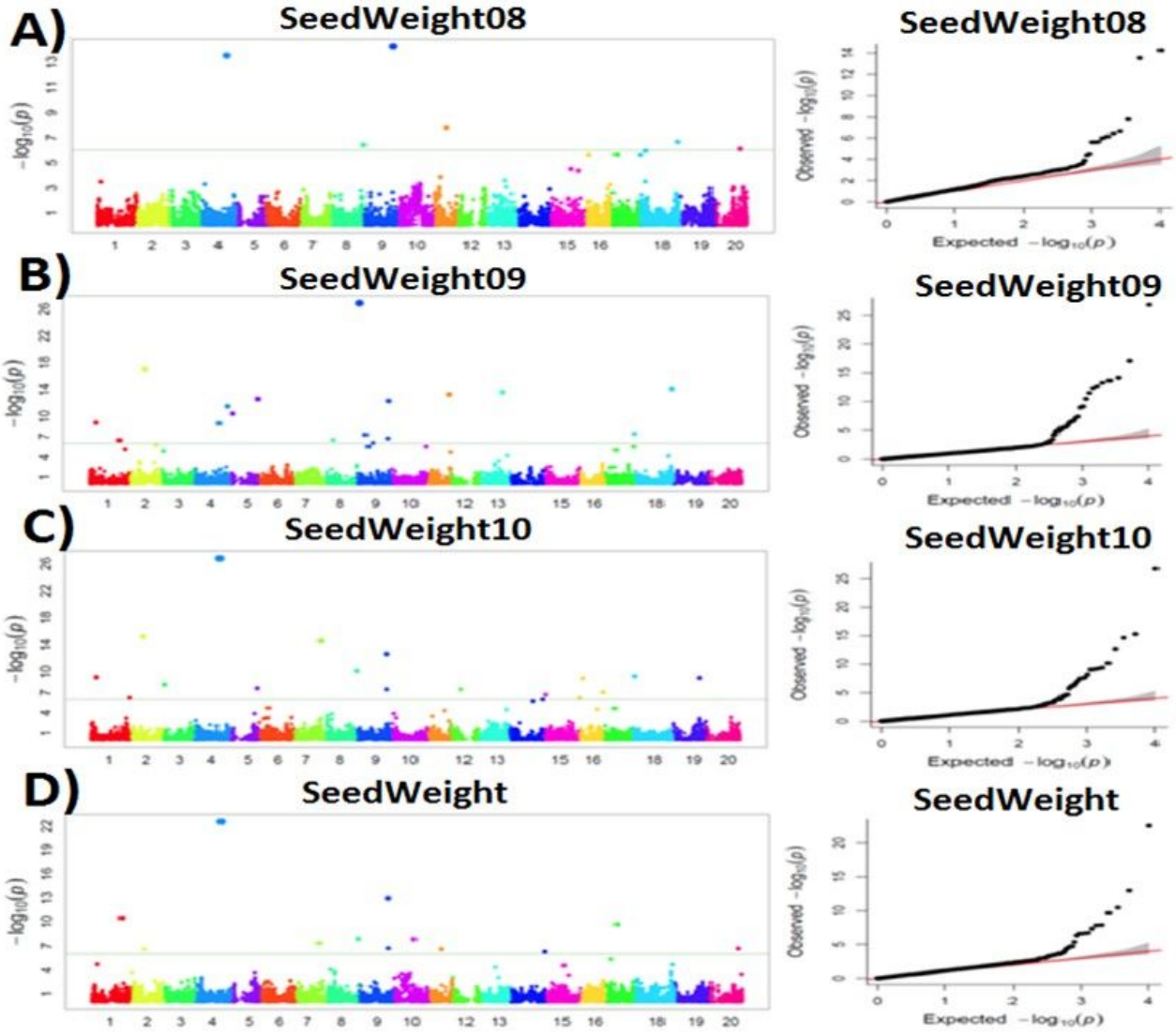

Fig. 6

Figure 6

Manhattan plots and QQ-plots for seed weight in 2008 (A), 2009 (B), 2010 (C), and the average data over 3 years (D).

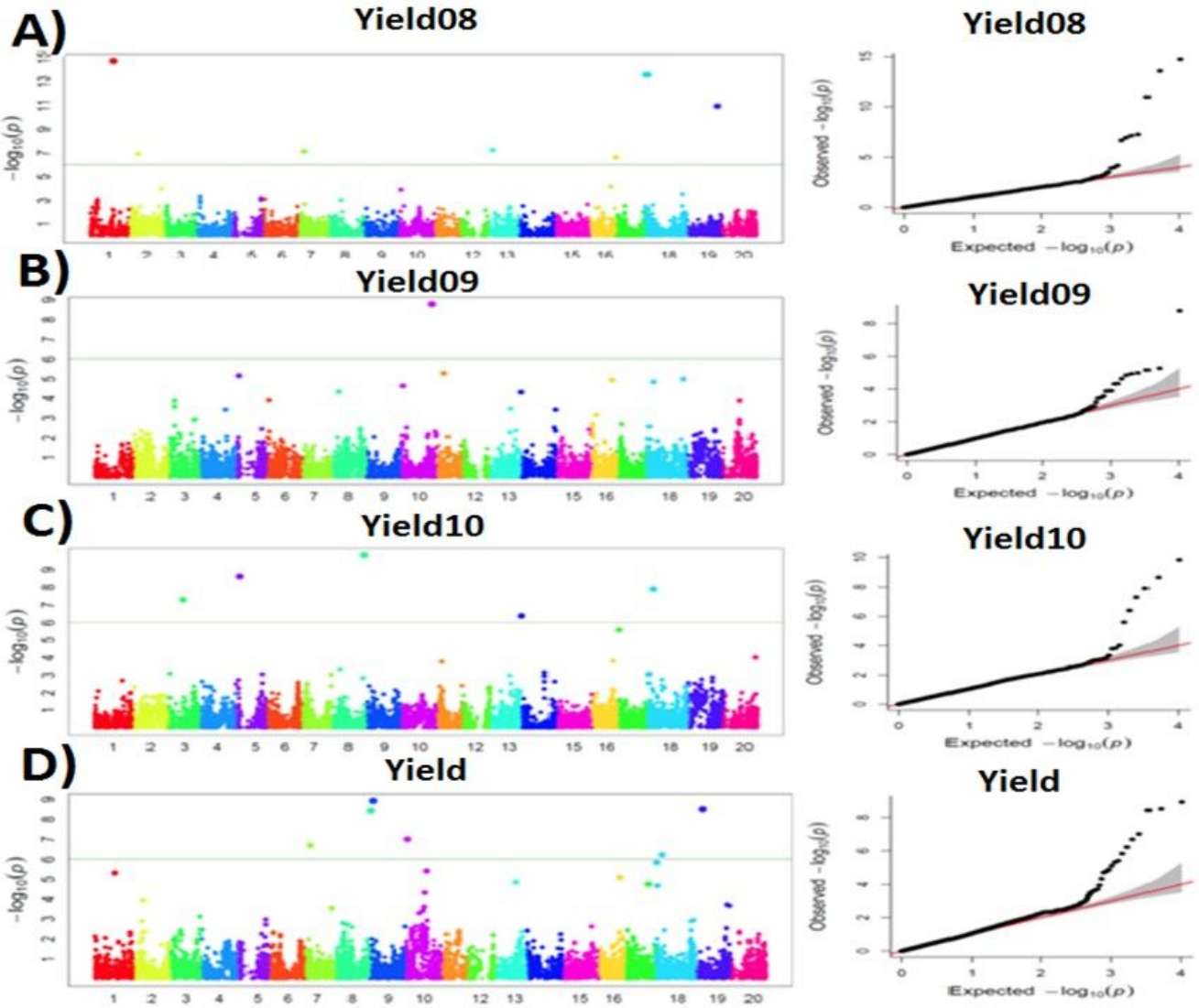

Fig. 7

Figure 7

Manhattan plots and QQ-plots for yield in 2008 (A), 2009 (B), 2010 (C), and the average data over 3 years (D).

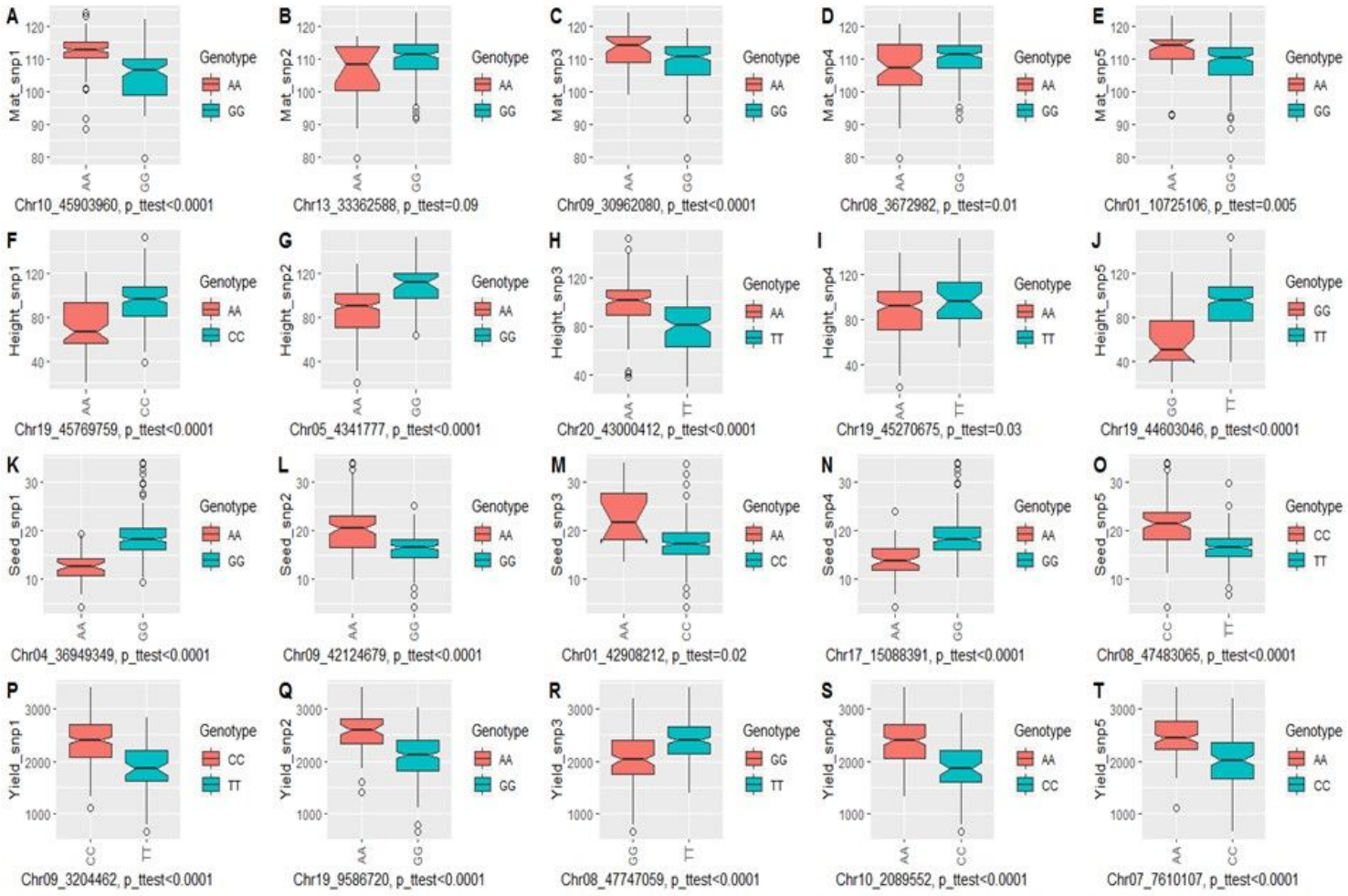

Fig. 8

Figure 8

Boxplots showing the variation in plant maturity (A, B, C, D, and E), plant height (F, G, H, I, and J), 100-seed weight (K, L, M, N, and O), and grain yield (P, Q, R, S and T) within each genotypic class defined by the top 5 significant SNPs for each trait. The x-axis showed the genotypic class from each SNP, whereas the y-axis represented the phenotypic value for each trait of interest. On the y-axis, mat_snp denotes the maturity date, height_snp represents the plant height, seed_snp is the 100-seed weight, and yield_snp refers to the grain yield data. The SNP name was shown below of each x-axis. A two-tailed t-test was conducted to statistically compare the trait value from each genotypic class defined by the SNP. The t-test was conducted under unequal variance assumption. p_ttest refers to the p-value obtained from the two-tailed t-test analysis. The sample size corresponding to each genotypic class defined by each significant SNP was the following: Chr10_45903960 (LOD= 10.47){nAA=149, nGG=95}, Chr13_33362588 (LOD= 6.83) {nAA=16, nGG=226}, Chr09_30962080 (LOD= 6.76) {nAA=40, nGG=198}, Chr08_3672982 (LOD= 6.32) {nAA=57, nGG=193}, Chr01_10725106 (LOD= 6.22) {nAA=33, nGG=192}, Chr19_45769759 (LOD= 11.78) {nAA=60, nCC=190}, Chr05_4341777 (LOD= 7.97) {nAA=204, nGG=46}, Chr20_43000412 (LOD= 6.76) {nAA=132, nTT=103}, Chr19_45270675 (LOD= 6.60) {nAA=214, nTT=35}, Chr19_44603046 (LOD=

6.21) {nGG=25, nTT=225}, Chr04_36949349 (LOD= 22.58) {nAA=41, nGG=207}, Chr09_42124679 (LOD= 12.97) {nAA=84, nGG=148}, Chr01_42908212 (LOD= 10.47 {nAA=12, nCC=238}, Chr17_15088391 (LOD= 9.38) {nAA=57, nGG=187}, Chr08_47483065 (LOD= 7.85) {nCC=62, nTT=178}, Chr09_3204462 (LOD= 8.92) {nCC=181, nTT=66}, Chr19_9586720 (LOD= 8.52) {nAA=66, nGG=163}, Chr08_47747059 (LOD= 8.43) {nGG=104, nTT=145}, Chr10_2089552 (LOD= 7.02) {nAA=185, nCC=64}, and Chr07_7610107 (LOD= 6.69) {nAA=119, nCC=123}. Outliers are represented by empty dots.

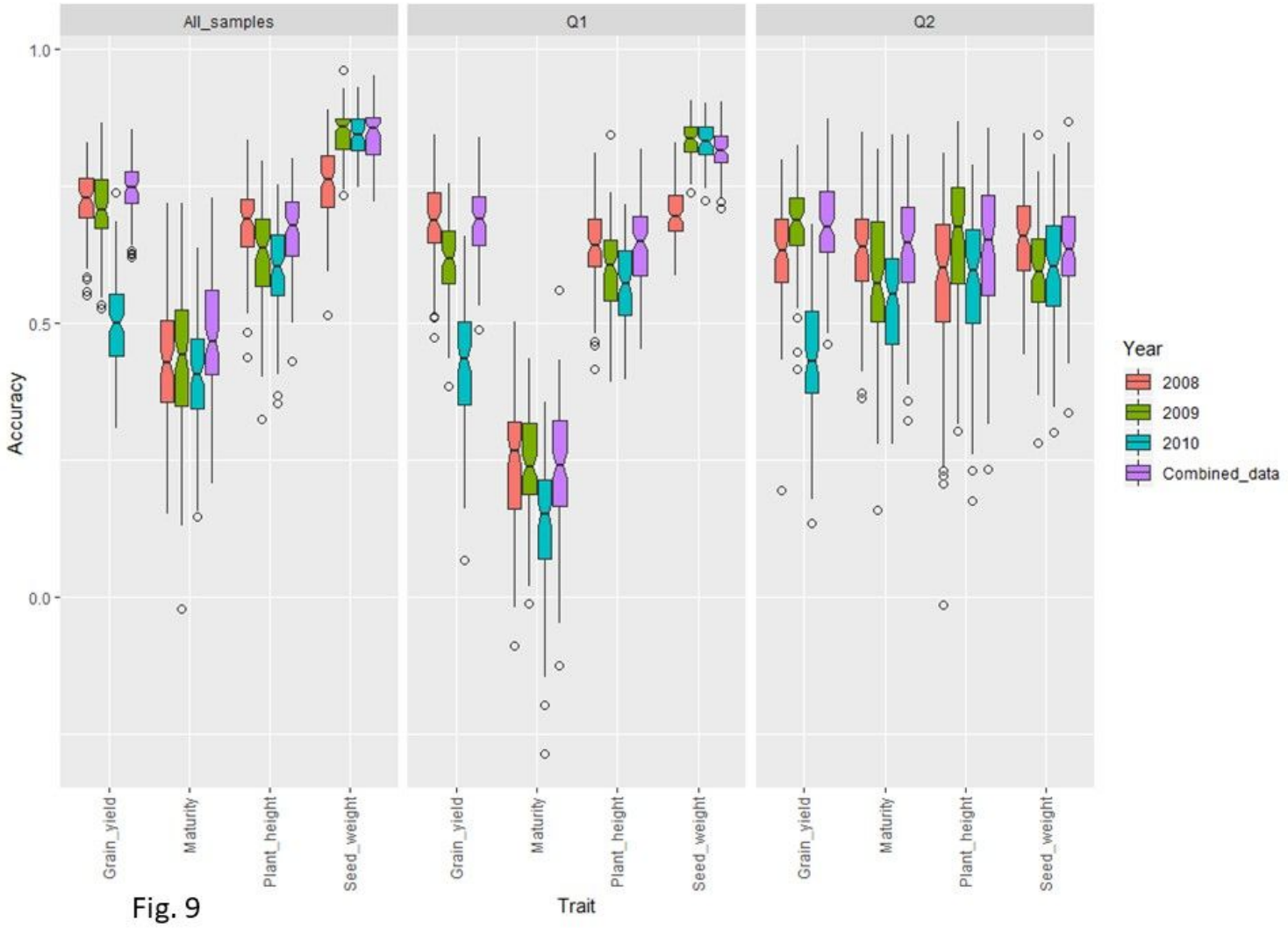

Figure 9

Genomic selection accuracy for yield, maturity, plant height, and seed weight using training/testing sets from all 225 soybean accessions (all samples), samples derived from Q1, and samples from the Q2 subpopulation, respectively. Cross-validation was conducted using the data from the same year.

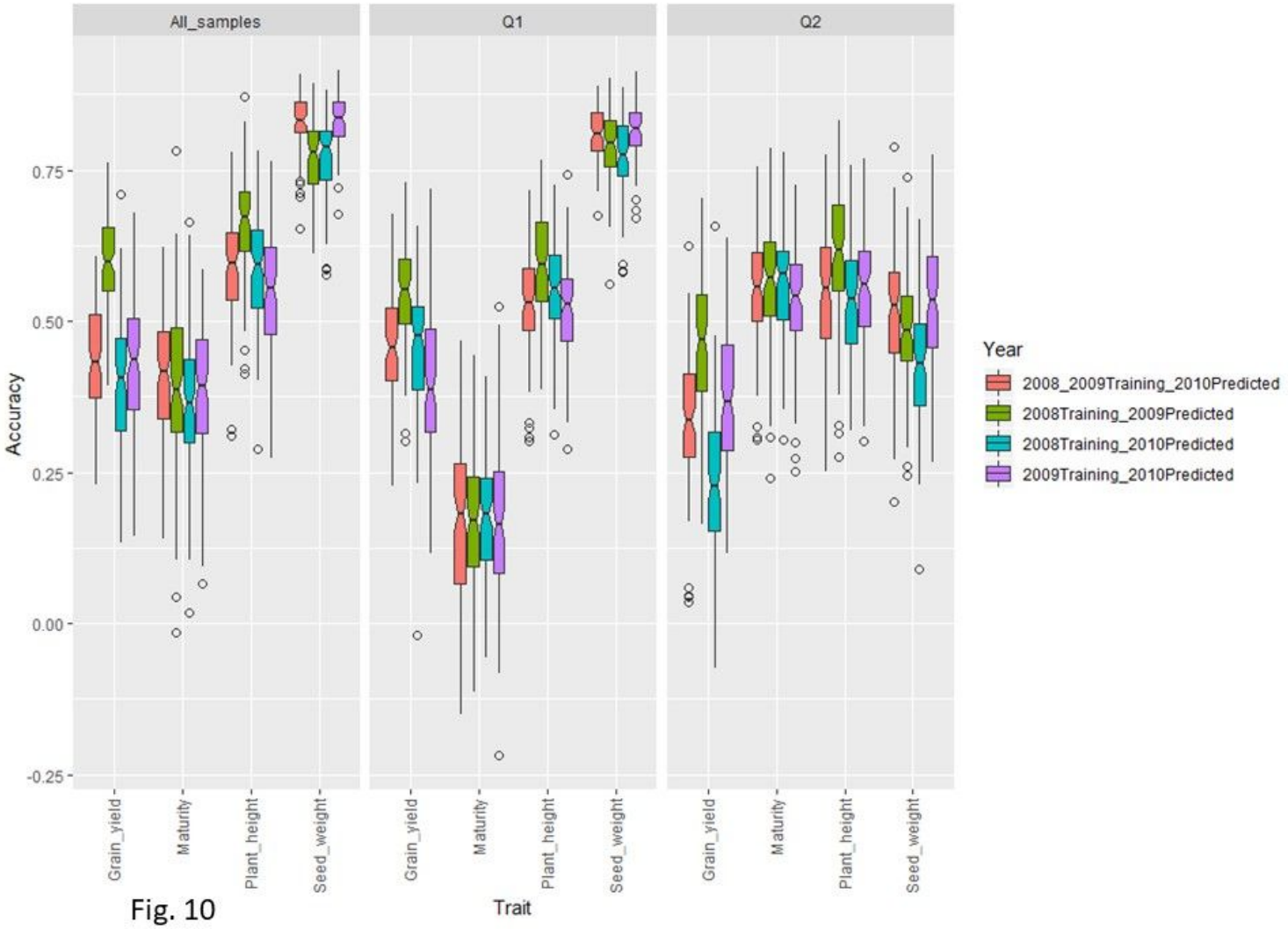

Fig. 10

Figure 10

Genomic selection accuracy for yield, maturity, plant height, and seed weight using training/testing sets from all 225 soybean accessions (all samples), samples derived from Q1, and samples from the Q2 subpopulation, respectively. Cross-validation was conducted in a way that the data from a year was used to predict that of from the succeeding year(s).

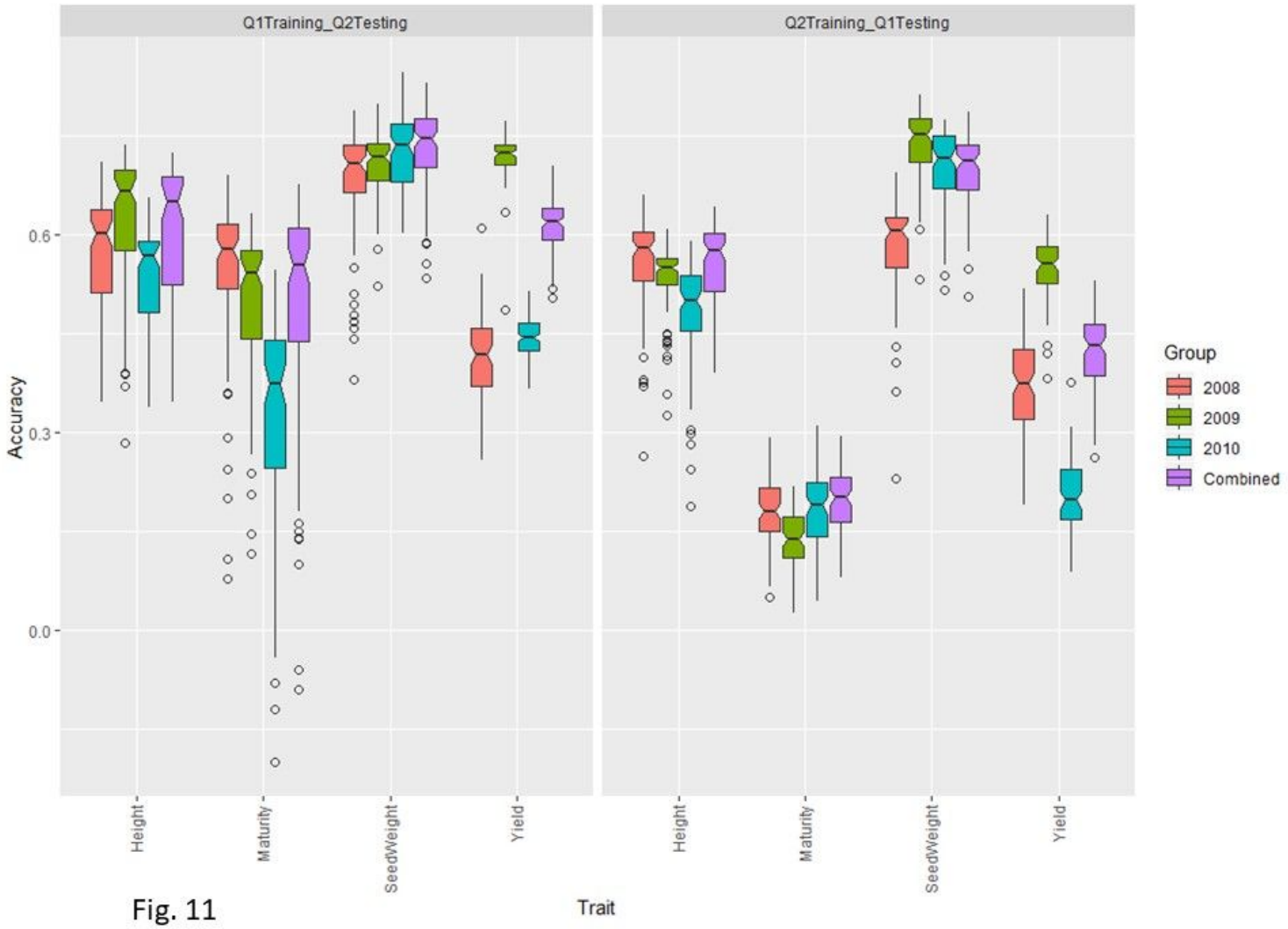

Figure 11

Genomic selection accuracy for plant height, maturity, 100-yield, and yield using samples from Q1 as a training set and individuals from Q2 as a testing set, and vice versa. Cross-validation was done using data from the same year.

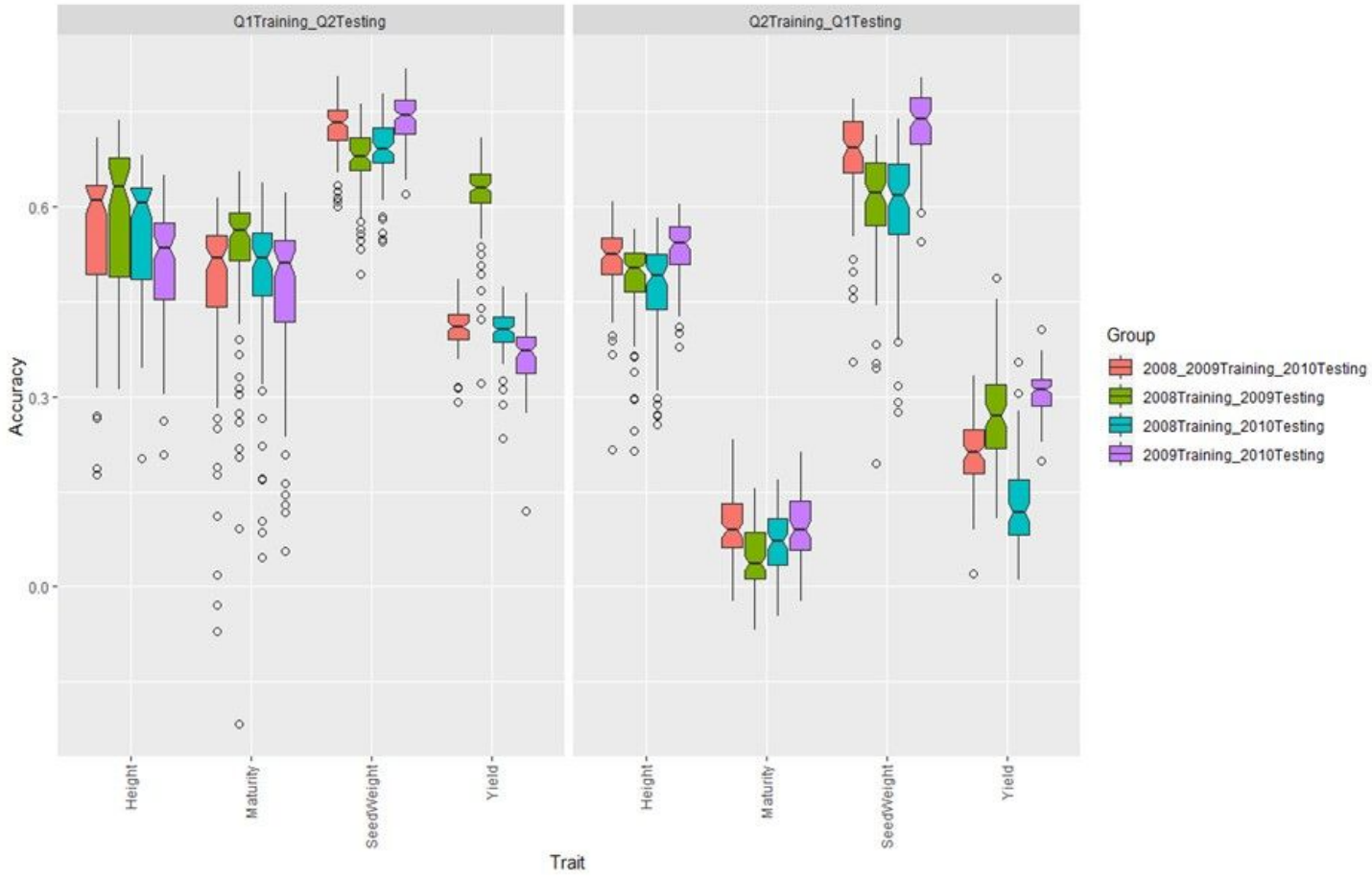

Fig. 12

Figure 12

Genomic selection accuracy for plant height, maturity, 100-yield, and yield using samples from Q1 as a training set and individuals from Q2 as a testing set, and vice versa. Cross-validation was performed using the data from a year to predict that of from the succeeding year(s).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableAgronomicTraits.xlsx
- Table9.xlsx
- Table10.xlsx
- Table11.xlsx
- Table4.xlsx

- Table5.xlsx
- Table6.xlsx
- Table7.xlsx
- Table1.xlsx
- Table2.xlsx
- Table3.xlsx
- Table8.xlsx