Dissection The Practical Soybean Breeding Pipeline By Developing High Throughput Functional Array ZDX1

Rujian Sun
College of Agriculture, Northeast Agricultural University
Bincheng Sun
Sam Higginbottom Institute of Agriculture Technology and Sciences Faculty of Animal Husbandry and Dairying: Sam Higginbottom University of Agriculture Technology and Sciences
Yu Tian
Chinese Academy of Agricultural Sciences
Shanshan Su
Beijing Compass Biotechnology Co, Ltd.
Yong Zhang
Keshan Branch of Heilongjiang Acdemy of Agricultural Sciences
Wanhai Zhang
Hulunbuir Institute of Agriculture and Animal Husbandry
Bingfu Guo
Chinese Academy of Agricultural Sciences
Huihui Li
Chinese Academy of Agricultural Sciences
Yanfei Li
Chinese Academy of Agricultural Sciences
Huawei Gao
Chinese Academy of Agricultural Sciences
Yongzhe Gu
Chinese Academy of Agricultural Sciences
Lili Yu
Chinese Academy of Agricultural Sciences
Yansong Ma
Chinese Academy of Agricultural Sciences
Jingshun Wang
Hulunbuir Institute of Agriculture and Animal Husbandry
Ping Yu
Hulunbuir Insititute of Agriculture and Animal Husbandry
Erhu Su
Inner Mongolia Academy of Agricultural and Animal Husbandry Sciences
Qiang Li
Inner Mongolia Academy of Agricultural and Animal Husbandry Sciences
Xingguo Hu

Hulunbuir Institute of Agriculture and Animal Husbandry
Qi Zhang
Hublunbuir Institute of Agriculture and Animal Husbandry
Rongqi Guo
Hulunbuir Institute of Agriculture and Animal Husbandry
Shen Chai
Hulunbuir Institute of Agriculture and Animal Husbandry
Lei Feng
Hulunbuir Institute of Agriculture and Animal Husbandry
Jun Wang
Chinese Academy of Agricultural Sciences
Huilong Hong
Chinese Academy of Agricultural Sciences
Jiangyuan Xu
Chinese Academy of Agricultural Sciences
Jing Wen
Chinese Academy of Agricultural Sciences
Jiqiang Liu
Beijing Compass Biotechnology Co, Ltd.
Yinghui Li
College of Agriculture, Northeast Agricultural University
Lijuan Qiu (qiulijuan@caas.cn)
College of Agriculture, Northeast Agricultural University https://orcid.org/0000-0003-0112-5713

Research Article

Keywords: soybean, SNP array, genetic diversity, functional loci, parental selection, genomic selection

Posted Date: September 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-837237/v1

Version of Record: A version of this preprint was published at Theoretical and Applied Genetics on February 21st, 2022. See the published version at https://doi.org/10.1007/s00122-022-04043-w.

Abstract

Microarray technology facilitates rapid, accurate, and economical genotyping. Here, using resequencing data from 2,214 representative soybean accessions, we developed the ZDX1 high-throughput functional soybean array, containing 158,959 SNPs, covering 90.92% of soybean genes and sites related to agronomically important traits. We genotyped 817 soybean accessions using ZDX1, including parental lines, non-parental lines, and progeny from a practical breeding pipeline. It was clarified that non-parental lines had highest genetic diversity, and 235 SNPs were identified to be fixed in the progeny. The unknown soybean cyst nematode-resistant and early maturity accessions were identified by using allele combinations. Notably, we found that breeding index was a good indicator for progeny selection, in which the superior progeny were derived from the crossing more distantly related parents with at least one parent having a higher breeding index. Based on this rule, two varieties were directionally developed. Meanwhile, redundant parents were screened out and potential combinations were formulated. GBLUP analysis displayed that the markers in genic regions had priority to be higher accuracy on predicting four agronomic traits compared with either whole genome or intergenic markers. Then we used progeny to expand the training population to increase the prediction accuracy of breeding selection by 32.1%. Collectively, our work provided a versatile array for high accuracy selecting and predicting both parents and progeny that can greatly accelerate soybean breeding.

Key Message

We provided the ZDX1 high-throughput functional soybean array for high accuracy evaluating and selecting both parents and progeny that can greatly accelerate soybean breeding.

Introduction

The goal of crop breeding is to develop plant varieties with ideal traits, such as higher yield, improved quality, and enhanced environmental adaptability (Liu et al. 2020a). Although commercially produced soybean [*Glycine max* (L). Merr.] yield has increased, due in part to the breeding of new varieties (Rincker et al. 2014), the yield increase per unit area has not changed significantly for the past few decades (Liu et al. 2020a). This shows that reliance on traditional phenotyping methods to develop new varieties has limitations (Barabaschi et al. 2016). Innovative genotyping platforms can accelerate the process of identification, evaluation, and use of elite germplasm. In particular, SNP arrays provide a high allele detection rate (Rasheed et al. 2017) and enable rapid, low-cost, high-throughput genotyping that can accelerate breeding research (Bailey-Serres et al. 2019; Viquez-Zamora et al. 2013; Yu et al. 2014).

The publication of the soybean genome has facilitated the discovery of SNPs (Schmutz et al. 2010), and SNP arrays have become a key technology in soybean genetics research. Despite low SNP density, previously developed soybean arrays have been used in diversity analysis, genetic mapping, and association analysis, etc. (Shen et al. 2005; Song 2014; Song et al. 2020; Wang et al. 2018b). More recently, high density soybean arrays have been successively developed. For example, the 50K soybean array was used to genotype 96 elite, landrace, and wild accessions and subsequently identify candidate genomic regions shaped by domestication or recent selection (Song et al. 2013). Similarly, this array was used to correlate protein- and oil-related loci via GWAS analysis of 298 strains (Hwang et al. 2014). The 180K soybean array was used to show that morphologically intermediate soybeans are natural hybrids between cultivated and wild soybean (Lee et al. 2015), while GWAS

based on the 335K array identified a candidate interval on chromosome 20 that affected grain weight (Wang et al. 2016). These related studies have increased our understanding of soybean genetics and have laid the foundation for the application of SNP arrays in breeding programs.

One of the key challenges facing plant breeders is the selection of suitable parents for generating sufficiently rich genetic variation to allow a maximal selection response during the breeding cycle in self-pollinating crops (Ji et al. 2018). To meet this challenge, new and more effective breeding strategies that combine phenotypic data with high-throughput genotyping have been developed to better identify prospective germplasm and to evaluate progeny (Varshney et al. 2014). Soybeans of different types (Pandey et al. 2017) and from different sources (Marrano et al. 2019) can be distinguished using microarrays in order to analyze the genetic relationships between different materials, and consequently provide a basis for determining the most suitable parents. Linking genotypic information with agronomically valuable traits that are not easily scored (Rasheed et al. 2017) can also help in the early evaluation of parents and the identification of desirable progeny. With the development of microarrays, genome-wide selection (GS) based on a large number of markers can be more informative and robust in selecting for complex traits controlled by multiple genes, such as yield, seed quality, and disease resistance, accelerating the translation of genotypic data to phenotypic selection in the field(Xu et al. 2020). However, there are relatively few reports describing the integration of high-throughput sequencing with the main breeding process.

Currently, there is an urgent need for a functional SNP array that covers the entire soybean genome and also contains representative and agronomically important sites to facilitate genetic research and molecular breeding. Here, we screened representative SNPs from a wide range of germplasm resources to develop the 'Zhongdouxin No. 1' (ZDX1) functional array. Using a breeding population comprised of 817 accessions including parental lines, non-parental lines and progeny, we demonstrate the use of this array for improving steps in breeding including screening for new genetic resources, population diversity analysis, optimizing hybrid combinations, and progeny selection. The ZDX1 array described in this work, with associated breeding selection strategies can accelerate all steps in the breeding process.

Materials And Methods

SNP detection, filtering, and selection for array development

Using resequencing data from 2,214 soybean accessions as the basic information and based on the Illumina platform, we obtained the VCF file by comparison with the reference genome Wm82.a2.v1 (Gmax_275_v2.0) and obtained 11,048,862 initial polymorphic SNP sites, including commercialized array sites, important gene sites, QTL and GWAS sites, and important trait functional sites. After removal of sites with a deletion rate of >0.1 and a degree of heterozygosity >15%, 9,092,282 sites were retained. We then screened 2,379,054 sites according to the rules of ‘no interference SNP sites in 35 bp around each site’ and ‘retaining sites with MAF≥0.01’. We deleted the non-polymorphic sites and the sites with interference within 50 bp on either side, keeping the tiling order=1 sites, and 2,039,377 sites remained. We deleted the sites with errors, tested 41 sliding window gradients for site screening, and selected the 4,800 bp window. The principle for site selection was "priority + Illumina score ≥0.4 + non-AT/GC selection site (if there is no non-AT/GC, then select the sites with higher priority)", among which the priority definition principles are: I. Excellent QTL sites, GWAS sites, important genes (VIP), selective genes, common genes, terminator/alternative splicing/non-synonymous mutation sites. II. Interspecies and intraspecific

subgroup unique sites. III. Selection interval (domestication) sites. IV. Whole genome coverage sites. V. Gap-filling sites. After genotyping and clustering with GenomeStudio software (GenomeStudio 2008), and testing and adjusting the typing signal which was >3, we finally obtained 158,959 SNP sites for ZDX1.

Plant materials and phenotypic data collection

The plants used in this study are 817 accessions from the actual breeding population, including 77 parental lines and 169 non-parental lines and 571 progeny. Progeny are stable lines obtained by the pedigree method after crossing. Among them, there are 298 progeny which both parents are included in parental lines, and other 273 progeny in which only single parents (male or female) are included in parental lines. Additionally, 283 of the 571 progeny bred in 2015, 288 progeny bred in 2016.

The field identification of 817 experimental plants was performed with three replicates (designated as environments E1 and E2) in Zalantun City, Inner Mongolia (47°40' N, 122°36' E) in 2017 and 2018, and one replicate in Keshan County, Heilongjiang Province in 2018 (48°33' N, 126°8' E) (designated as environment E3). The experiment used a randomized block design and set one control line every 20 lines ('Neidou4hao' or 'Keshan1hao'), with a row spacing of 0.65 meters and a row length of 3 meters. Six quantitative traits were investigated: VE, defined as the date of emergence of the cotyledons. beginning maturity (R7), defined as the days from emergence to when one pod on the main stem has reached mature pod color (Fehr et al. 1971), and for each row, the date was defined as when 50% of the plants meet the above condition (Qiu 2006). In each plot, 20 plants were continuously harvested where there was no shortage of seedlings, the seed yield (SY), 100-seed weight (SW), protein content, and oil content were measured, and the survey was conducted as previously described. One qualitative trait, leaf shape, was recorded as either narrow or broad leaflet (Qiu 2006).

Genotypic data collection

A commercial kit (Tiangen Plant Genomic DNA Kit, DP305) was used to extract genomic DNA from young soybean leaves. We used the ZDX1 SNP array developed based on the Illumina® platform as a typing tool (Zhao et al. 2018), and used GenomeStudio software to obtain the SNP genotypes. The ZDX1 array contains 14 reported functional loci, including six for growth period which are *e1-fs*, *e1-as* (Tsubokura et al. 2014; Xia et al. 2012), *e3-fs* (Tardivel et al. 2014; Xu et al. 2013), *e4-keshuang* (Langewisch et al. 2014; Tsubokura et al. 2013), e4-oto (Langewisch et al. 2014; Tsubokura et al. 2013), and *GmGPRR3b/Tof12* (Li et al. 2020); three sites in genes for cyst nematode resistance which are *rhg1-a/GmSNAP18* (Cook et al. 2012; Shi et al. 2015), *Rhg4/GmSHMT08* (Liu et al. 2012; Shi et al. 2015), and *GmSNAP11* (Tian et al. 2019; Tian et al. 2018); leaf shape *Ln/ ln* (Jeong et al. 2012); stem termination *Dt1/Gmtfl1-ta* and *Dt1/Gmtfl1-ab* (Langewisch et al. 2014; Tian et al. 2010); seed coat color *Gm850* (Wang et al. 2018a); and seed coat gloss *Bloom1* (Zhang et al. 2018).

Population genetic analysis

PLINK v2.1.1 (Purcell et al. 2007) was used to control the genotypes: we screened out 7,099 sites with a genotyping success rate of <90%, 63,311 sites with MAF<0.01, and removed 123 sites which are on unassembled chromosomal fragments and/or are not on autosomes, which left 88,426 effective SNP sites. Linkage disequilibrium (LD) analysis was performed with Ldheatmap software, in which the maximum distance (Kb) between two SNPs was set to 1,000, and the correlation coefficient (r^2) of alleles was calculated to measure LD in each group level. The LD decay rate was defined as the chromosomal distance at which the average r^2 dropped to

half its maximum value. The Kinship matrix was calculated using the VanRaden method in Gapit software to obtain the genetic relationships between lines in the population.

After removing LD from parental lines and non-parental lines with a rule of typing success rate >0.9 and indep-pairwise 50 5 0.5 (a. consider a window of 50 SNPs, b. calculate LD between each pair of SNPs in the window, c. remove one of a pair of SNPs if the LD is greater than 0.5, d. shift the window 5 SNPs forward and repeat the procedure), we obtained 8,940 loci. We used PLINK v2.1.1 for PCA analysis and R software to draw PCA diagrams.

Best linear unbiased estimates and breeding index

Based on the phenotypic data obtained by the multi-point field identification method over several years, the R language asreml data package was used to calculate the best linear unbiased estimates (BLUE) from the phenotypic data for genomic selection (He et al. 2016) and breeding index.

We propose a selection index as a metric for breeding, we named this index value the breeding index (BI). The index is a linear combination of predicted values of comprehensive traits, each having a unique weight, as shown below:

$$I_j = \sum_{k=1}^{5} w_k \hat{y}_{jk}^*$$

where I_j is the selection index score for individual j, w_k is the economic weight for the kth trait for k = 1,2,...,5, and is the standardized predicted value for trait k from the jth individual accession that is calculated by standardizing the values for each trait by subtracting the mean value and dividing by the SD. We included all five traits in the selection index, corresponding to the following order; beginning maturity, 100-seed weight, protein, oil, and seed yield (Cui et al. 2020; Zhao et al. 2015), the weight of five taits is showed below:

w = [-0.2, 0.1, 0.2, 0.1, 0.4]

Among the 246 parents, the BI of the top 1/3, middle 1/3, and bottom 1/3 from high to low were designated as high parents, medium parents, and low parents, respectively, with 82 accessions in each group. In addition, ‘Rate over best-parent’ means the proportion of progeny with better performance than that of the ‘best’ parent.

Heritability and genomic selection

PLINK v2.1.1 was used to control the genotype according to the following criteria: we removed 7,099 sites with a genotyping success rate of <90%, eight Insertion/Deletion (Indel) sites, 745 sites on scaffolds, and 82,085 sites with MAF<0.05. This left 69,022 valid SNP sites remaining.

ABLUP (Song et al. 2019), GBLUP (Zhe et al. 2015), and HBLUP (Li et al. 2014a; Lourenco et al. 2020; Song et al. 2017) were performed by BGLR (Pérez and de Los Campos 2014), asreml (Gilmour et al. 2015), and R software, respectively. We computed the heritabilities using the following formula in QTL ICIMapping (Meng et al. 2015):

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GE}^2}{e} + \frac{\sigma_\varepsilon^2}{er}}$$

where σ_G^2 is the variance among soybean lines, σ_{GE}^2 is the genotype-by-environment interaction variance, σ_ε^2 is the residual variation, and e and r are the number of environments and replications within environments, respectively.

The Pearson correlation coefficient between the predicted and observed phenotype (rMP) was estimated, and the prediction accuracy (rGS) was calculated for the standardized rMP by the square root of the broad-sense heritability (Lehermeier et al. 2013). When comparing the prediction effects of gene regions, intergenic regions, and whole genome markers, the following strategies were adopted for marker sampling. Among the 69,022 loci retained after filtering, the number of gene regions was 33,756 and the number of intergenic regions was 35,266. In order to eliminate the influence of the number of loci on the prediction accuracy, all 33,756 loci were reserved in the gene regions, 33,733 loci were uniformly selected in the intergenic regions, and 33,761 loci were uniformly selected from the 69,022 loci over the whole genome (of which 16,457 are in genes and 17,304 in intergenic regions). When comparing different traits, different models, and different marker sampling strategies, a 5-fold cross-validation method was used to evaluate the prediction accuracy of the genome selection model. In order to reduce the sampling error, each sampling method was repeated 100 times, and the pairwise.t.test function in *R* was used to analyze the significance of the differences.

Results

SNP selection and characteristics of the ZDX1 array

In order to develop a high-throughput SNP typing array customized for the Illumina sequencing platform, SNPs were selected in re-sequence data from 2,214 soybean germplasm accessions, including 862 improved cultivars (*Glycine max* (L.) Merr.), 1,131 landraces, 218 annual wild soybean accessions (*Glycine soja* Sieb. & Zucc.), and three perennial wild soybean accessions (*Glycine Subgenus Glycine*) (Fig. S1). The SNP information was compared with the soybean reference genome Wm82.a2.v1 (Phytozome, https://phytozome.jgi.doe.gov/pz/portal.html), and 11,048,862 initial SNP sites were obtained after screening and quality control. Following removal of SNPs with no polymorphism, 50 bp interference, or one-sided probe length of 0, a total of 2,039,377 loci were retained. Using GenomeStudio software for genotyping and adjustment (Fig. S2), we then screened these loci for previously described QTL, GWAS, and developed array loci, as well as other published loci associated with valuable traits, ultimately resulting in a set of 158,959 high-quality sites (Fig. 1a, Supplemental Table 1).

We next examined the genomic distribution of the 158,959 high-quality SNPs and found that they were evenly distributed across the 20 soybean chromosomes. The number of SNP sites on each chromosome ranged from 6,085 to 9,314, of which, 90.23% fell within 10 kb (6.0 kb average distance) (Supplemental Table 2). In addition, SNP number showed a highly significant positive correlation with chromosome length, with a Pearson's coefficient of 0.98 (p = 8.61E-14) (Fig. 1b). We mapped 64,435 of the candidate SNPs to 50,592 annotated genes, accounting for 90.92% of the total number of predicted genes in the soybean reference genome (Fig. 1c). In addition, another 4.29% of the large-effect SNPs could potentially affect gene function, including 5,684 non-synonymous SNPs, 119

Stoploss SNPs (four of which are both nonsynonymous or Stoploss), 604 Stopgain SNPs, six frameshift SNPs, and 414 alternative splicing SNPs.The SNPs selected for inclusion in the ZDX1 array also included 14,685 synonymous sites, 6,120 unknown sites, 14,845 sites located in intronic regions, 12,158 sites located within the 1,000 bp upstream or downstream of a gene, 9,804 sites located in UTR regions, and 94,524 sites located in intergenic regions (Supplemental Table 1). A/G and T/C (transitions) represented the main nucleotide variants on the chip, with 54,219 and 54,262, respectively, accounting for 68.25% of the total SNPs. The site frequency spectrum (SFS) for the 2,114 re-sequenced accessions showed that the sites with minor allele frequency (MAF) > 0.1 accounted for 81.3% of the total. SNPs with MAFs between 0.10–0.20, 0.20–0.30, 0.30–0.40, and 0.40–0.50 accounted for 31.48%, 19.20%, 15.79%, and 14.85%, respectively (Fig. 1d). Collectively, these data showed that the SNPs selected for ZDX1 were evenly distributed across the soybean chromosomes and that the array had high gene coverage and utilization, thus enabling population structure analysis, whole-genome-based selection, and other related studies.

In addition, the ZDX1 array was also designed to retain high-priority loci, including 2,402 SNPs for genes related to important traits and 627 SNPs for genes that underwent domestication or improvement (Supplemental Table 3). In addition, Analysis using Soybase showed that the candidate SNPs for ZDX1 also included 953 SNPs in QTL intervals, 547 GWAS-identified SNPs (https://soybase.org), and 110,811 SNPs that differed between ecological groups (Supplemental Table 4). Moreover, 3,869 SNPs from two low-density arrays, the 1.5K BeadChip (Shen et al. 2005) and the BARCSoySNP6K (Song 2014) array were also included (Supplemental Table 5). Compared with the three high-density arrays SoySNP50K, 180K AXIOM®, and NJAU 355K SoySNP, the ZDX1 array contains 134,737 unique sites (Fig. 1e), with a specificity rate as high as 84.8%. In addition, 14 important functional sites (causal SNPs) related to traits such as growth period, resistance to cyst nematodes, leaf type, pod setting habit, seed coat color, seed dormancy, and phosphorus efficiency (Supplemental Table 6) were selected for the array, thus facilitating identification of economically and agronomically valuable traits and screening for elite germplasm.

As a final step in marker selection, we evaluated the accuracy of the marker information. To this end, we first determined the site detection rates in 817 well-established breeding materials (Supplemental Table 7) and found that the detection rate for each sample was between 84.40% and 95.98%, with an average of 95.19%. At the same time, three DNA samples were randomly selected for two repetitions, and the genotype similarity between the repetitions was > 99.9% (Supplemental Table 8). The above data shows that this array has a high degree of accuracy and repeatability. Taken together, these results confirmed that the high-density ZDX1 array was both reliable and accurate.

Analysis of genetic diversity of breeding population and screening of fixed sites in breeding improvement

Subsequently, we applied the ZDX1 array for genotyping in a test population of 817 breeding lines in soybean breeding program, including 77 parental lines, 169 non-parental lines, and 571 stable progeny lines developed using the pedigree method after crossing. To analyze the genetic diversity of the three subpopulations, we next conducted linkage disequilibrium analysis (LD; indicated by r^2). The results showed that the attenuation rate of the non-parental lines r^2 was higher than that of the progeny and parental lines, and the distance at which r^2 decayed by half was 244 kb, 276 kb, and 303 kb, respectively (Fig. 2a). These results indicated that the genetic diversity among the non-parental lines was higher than that of the parental lines and progeny, which thus helped to broaden the genetic diversity of the parental lines. Similar to the results of LD analysis, PCA analysis (Fig. 2b)

confirmed that the distribution of the non-parental line subgroup was more scattered, further indicating higher genetic diversity.

The results of MAF showed that 46,376 sites in the test population were completely fixed, and that 38,625 sites (83.29%) contained differences between groups from geographically separated ecological regions. Furthermore, the percentages of fixed sites in the non-parental lines, parental lines, and progeny were 34.72%, 41.79%, and 34.63%, respectively (Supplemental Table 9, Fig. S3). To further clarify which sites were selected and fixed during the breeding process, 6,579 sites were selected based on their polymorphisms in the 817 accessions, as well as in the progeny subgroup (MAF = 0). It is worth noting that the minor allele types corresponding to these sites were the same across all three subgroups. Statistical analysis showed that the MAF values of the parental lines ranged between 0 and 0.0390, while the MAF values of the non-parental lines ranged between 0 and 0.1317 (Fig. 2c). Among them, 235 sites were identified where the MAF values of the parental and the non-parental lines were > 0.01, and 109 sites were located in genic regions spanning 95 genes (Supplemental Table 10). Taken together, these results suggested that these apparently informative SNP sites were fixed during the breeding process, which can be selected in future breeding.

Germplasm screening for breeding target traits using functional sites in ZDX1 array

In order to then develop elite germplasm using the functionally informative SNP sites, we selected fourteen SNP sites from the array to identify the test population, among which five were found to be non-polymorphic. These five marker sites included stem termination (*Dt1/Gmtfl1-ta*, *Dt1/ Gmtfl1-ab*), and the seed coat color (*Gm850*). Among the six sites related to maturity, *e4-oto* and *GmGPRR3b/Tof12* were completely fixed, and the MAF values for *e1-fs*, *e1-as*, *e3-fs*, and *e4-keshuang* were calculated to be between 0.001 and 0.192. In addition, the MAF values for three sites associated with cyst nematode resistance, *rhg1-a/GmSNAP18*, *Rhg4/GmSHMT08*, and *GmSNAP11*, were between 0.012 and 0.017, while the MAF value for leaf-shape *Ln/ ln* site was 0.203. These findings suggested that these unfixed sites could be related to the genetic diversity of the phenotype, and potentially controlled traits that are desirable for breeders (Fig. S4).

We next analyzed six alleles from *E1, E3, E4*, and *GmPRR3b* related to fertility, and MAF analysis showed that these six alleles were found in 11 distinct allelic combinations in the population, among which the *e1-fs*, *e1-as/ e3-fs/ e4-kes*, *e1-as/ e3-fs*, and *e1-as/ e4-kes* genotypes were associated with precocity (Supplemental Table 11). Notably, only one accession, 'Dongnong36' (80.73d), carried the *e1-fs* genotype. Among the materials with two or more genotypes that included *e1-as*, *e3-fs*, and *e4-kes*, nine parental lines/nonparental lines exhibited an earlier fertile period (87.14-97.98d), while three progeny (HJ15-1231, HJ15-896, and HJ15-897) had relatively late growth periods (109.17-114.32d). These results showed that functional SNPs associated with determinate growth period in the ZDX1 array could thus be used to identify germplasm with early maturity phenotypes.

Three nematode resistance-associated SNPs, including Gm18_1643660, Gm08_8361148, and Gm11_32970174 (located in the *rhg1*, *Rhg4*, and *SCN3-11* genes, respectively) were also covered by the array (Table 1). We found that the frequency of alleles for enhanced disease resistance in the tested materials was relatively low: 1.22%, 1.71%, and 1.47%, respectively. These three sites could be found in eight allelic combinations among the 817 accessions of the diversity panel, while seven accessions were identified that carried all of the resistance loci, including three known resistant varieties, 'Kangxian1hao', 'Kangxian5hao', and 'Kangxian8hao'. In addition to

these accessions, three new varieties not previously known to carry nematode resistance were also identified, including 'Shundou5hao', 'Qinong1hao', 'Fengdou 23', as well as the progeny 'HJ15-863'. The proportion of resistant progeny was extremely low, potentially due to the difficulty of large-scale phenotypic identification and the lack of directional selection against SCN in the progeny.

Table 1
Allelic combinations at the *rhg1-a*, *Rhg4*, and *GmSNAP11* loci

Combination	*rhg1*-a/GmSNAP18 Gm18_1643660	*Rhg4*/GmSHMT08 Gm08_8361148	*GmSNAP11* Gm11_32970174	Number of parental lines	Number of non-parental lines	Number of progeny
Com1	GG	GG	TT	0	6	1
Com2	CC	CC	CC	76	162	557
Com3	CC	CC	TT	0	0	3
Com4	GG	CC	TT	0	0	2
Com5	CC	GG	CC	1	0	6
Com6	GG	CC	CC	0	0	1
Com7	CG	CC	CC	0	0	1
Com8	CG	GC	TC	0	1	0

SNPs mapped to the *Ln/ln* loci were also used to genotype 817 test materials. This screen revealed that 649 narrow leaflet soybean accessions all carried the CC genotype, 166 broad leaflet accessions harbored the GG genotype, and two heterozygous CG accessions showed both broad and narrow leaflet. This result reflected the prevalence of narrow leaflet among the soybean breeding lines used in Northeast China, and demonstrates the high accuracy of leaf shape detection using SNPs associated with functional loci in the ZDX1 array. Indeed, a greater proportion of round-leaf accessions were present in the non-parental lines (32.0%), while round-leaf accessions in the parental lines and progeny accounted for 10.4% and 18.4% of these populations, respectively. These proportions again reflected the preference by breeders for narrow leaflet (Fig. S5). We also found that breeders in the high latitudes of Northeast China have a preference for narrow-leaf breeding materials. This is related to the fact that the narrow leaflet lines have > 4 pods (Fang et al. 2013) and greater light transmission through the canopy.

Using breeding index and genetic distance to explore the method of parental selection

In order to illustrate how the ZDX1 array can improve the parental selection process, we next used genotype data to generate a kinship matrix for the full accessions, which revealed pairwise genetic distance that ranged between − 0.54 and 2.56 (with larger values indicating closer kinship; see Fig. S6). Analysis of R7 (beginning maturity), SW (100-seed weight), protein content, oil content, and SY (seed yield) in 298 progeny derived from the parental lines for both parents showed that the rate over best-parent of each trait was non-significantly negatively correlated with the genetic relationship between the parents ($p = 0.30–0.97$), and the correlation coefficients (r_{hd}) were − 0.02 to -0.42, suggesting that the more distant the parental relationship, the greater the possibility that a higher

proportion of progeny would outperform the parental lines. In addition, the mean value of each trait among progeny was positively correlated with the average parental value, with correlation coefficients (r_{po}) were between 0.33 and 0.73. Among them, mean oil and seed weight of the progeny showed an extremely significant ($p < 0.01$) correlation with the mean of these trait values among the parental lines (Fig. 3), which indicated that the use of elite parents in hybrid combinations allows the selection of elite progeny.

While high yield is the most important goal in soybean breeding, traits such as maturation time and seed quality (i.e., protein and oil contents) should also be comprehensively considered during selection. To this end, we included all of the five traits in the selection index, defined as the breeding index (BI), with which we scored the parental lines and 298 progeny which both parents are included in parental lines. Based on BI values, the parents could be categorized as high, medium, or low phenotypes (Supplemental Table 7). The 30 (top 10%) high-performance progeny could then be divided into five types based on the parental BI index. Using this system, we identified two high×high types, 11 high×medium types, nine high×low types, three medium×medium types, and five medium×low types. Of these types, 73.3% involved contributions from at least one high type parent (Fig. 4). When evaluating new lines with BI, the commonly used control varieties ‘Keshan1hao’ (BI = 0.53) and ‘Neidou4hao’ (BI = 0.25) were rated as a “High” type. This standard was also used to screen out two new varieties ‘Mengdou1137’ and ‘Mengdou640’ that passed the national certification, the "High×Low" parental combination was used to generate these two varieties, the average genetic relationship was relatively distant (-0.15). These results indicated that the selection of more distantly-related parents, among which at least one parent has strong multiple trait indexes, will more likely produce progeny with highest composite agronomic performance for these traits. It provides a reference for us to select suitable parents in complex self-bred crop breeding.

Following identification of candidate parental lines with suitable genetic distance and high-performance phenotypes, we also needed to enable efficient breeding decisions by eliminating redundant germplasm accessions from the diversity panel, which otherwise results in considerable genetic redundancy in the selected parental subgroup. We have counted the parental lines of the bottom 30 progeny (bottom 10%), among these parents, 12 parents including ‘Dengke4hao’ and ‘Hujiao1120’ did not derive excellent progeny (top 10%) (Supplemental Table 12), they can no longer be used in future breeding. Meanwhile, based on the genetic relationships indicated by different metrics for genetic distance, including 0.5 ~ 1.0, we identified the non-parental lines with higher similarity to the parental lines used in crosses (Fig. S7A), and finally eliminated 21 redundant non-parental lines including ‘Mei1’ and ‘Nenao08-1092’ based on kinship scores of > 1.0 (Supplemental Table 13). After screening out redundant parents, the improved combinations were proposed for future breeding. Meanwhile, the high-performing progeny lines should also be included in the parent nursery. We therefore selected the accessions with the top 10% of BI values, and calculated the number of potential combinations that could be formulated using different metrics for genetic distance, including − 0.5 ~ 0.0 (Fig. S7B). Using a kinship score of <-0.3 as the standard, we selected 46 high-potential combinations for use in future breeding experiments (Supplemental Table 14). By eliminating redundant parents and formulating potential combinations, the parent population structure is optimized and breeding efficiency is improved.

Different strategies based on ZDX1 array improve the accuracy of genomic selection in theoretical and actual breeding

We next explored the efficiency of different strategies to improve the accuracy of genomic selection using the ZDX1 array. The results of GBLUP (genomic best linear unbiased prediction) analysis to test the accuracy of selection based on the ZDX1 array revealed that the prediction accuracy for the five traits of beginning maturity, seed weight, protein content, oil content, and seed yield were 0.79, 0.73, 0.78, 0.77, and 0.69 respectively; these scores were all significantly higher than those of ABLUP (pedigree-based best linear unbiased prediction), based on pedigree relationships, and HBLUP (combined best linear unbiased prediction) based on both pedigree relationship and genotype data ($p < 0.01$) (Fig. 5a, Supplemental Table 15). These results indicated that the genomic information provided by the array can better reflect the population structure than pedigree relationships.

We subsequently identified 33,756, 33,733, and 33,761 sites that were respectively selected as marker subsets from gene regions, intergenic regions, or the whole genome. GBLUP analysis confirmed these three marker sampling methods showed no significant differences in their accuracy for predicting yield. For the other traits, the accuracy of prediction using markers for genic regions was 2.33% higher than that of SNP markers for intergenic regions, with highly significant differences among methods for each of the four traits ($p < 0.01$). Also, markers associated with genic regions were more accurate by an average of 0.57% than those sampled from across the whole genome, and were significantly more accurate for predicting 100-seed weight, protein content, and oil content ($p < 0.01$). Furthermore, use of only the 33,756 SNPs in genic regions also significantly improved the predictive accuracy ($p < 0.01$) for selecting 100-seed weight, and protein and oil contents compared with accuracy provided by using all 69,022 SNPs (Fig. 5b, Supplemental Table 15). These results showed that the sites on the genic-region in the array include more useful genetic information. For most traits, the strategy of sampling SNP markers for gene-encoding regions can reduce the number of requisite markers while also improving the accuracy of genomic selection.

In order to evaluate the efficiency of ZDX1 in predicting progeny in actual breeding, we also selected 246 parents as training population I, and 283 of the 571 progeny bred in 2015 as Predicted Population I. We used ZDX1 array to predict the top 50% of the 283 progeny. The prediction accuracy for the five traits in these high-value lines ranged from 0.30 to 0.45 (Circle1). We then used these selected 141 high-value progeny to expand training population I to generate training population II to further predict the 288 progeny bred in 2016 (Predicted Population □) (Fig.5c). The results showed that with the exception of yield, the predictive accuracy was improved for these traits, ranging from 0.48 to 0.67 (Circle2), while the average accuracy was significantly increased by 32.1% (p=0.024) (Fig.5d, Supplemental Table15). Collectively, these results demonstrate that the predictive accuracy of breeding decisions based on the ZDX1 array can be improved by establishing a model using the parental lines and continuously expanding the model with high-performing progeny.

Discussion

Pan-genomic studies in soybean have shown that the use of a large number of accessions is more conducive to identifying genetic variation (Li et al. 2014b; Liu et al. 2020b). However, the initial site data in previously developed soybean arrays were derived from only a few to dozens of cultivated or wild species (Lee et al. 2015; Song et al. 2013; Wang et al. 2016). The initial locus information in ZDX1 is derived from 2,214 soybean germplasm accessions from a wide range of sources and types, which reduces the possibility that rare alleles originating in a small number of samples will affect the marker identification results and that its high representativeness. A similar strategy has been successfully applied in the development of arrays for other species such as *Eucalyptus* (Silva-Junior et al. 2015). Previous studies have shown that SNPs with higher MAF should be the first choice for

array design, which is another advantage of our array, together with the robustness of the Illumina bead chip system (Gunderson 2009). Although arrays with similar or higher density have been used for soybean genotyping (Lee et al. 2015; Wang et al. 2016), we paid more attention to the distribution of the SNP sites throughout the genome. Due to this even genomic distribution, the ZDX1 array provides unprecedented practicability and functionality. In particular, its extremely high coverage of annotated genes and many important sites makes it useful for correlation analysis and genetic mapping, while the moderate density reduces costs. These characteristics of the ZDX1 array contribute to its versatility and reliability for soybean breeding and genetic research.

When screening germplasm for potential use as parents, phenotypic identification is time-consuming and laborious, and the results are greatly affected by the environment. Therefore, molecular marker-assisted selection represents an efficient and effective method for screening target traits (Barabaschi et al. 2016). Previous studies have shown that some breeding materials which perform well locally are unique in terms of genotype and may therefore contribute useful genetic variation to breeding programs (Iquira et al. 2010).Other Studies have shown that a lack of genetic diversity can hinder efforts to increase yield potential in new varieties (Hegstad et al. 2019). We hypothesized that non-parental lines or progeny with different genotypes could be used to expand the pool of parental lines, while the frequency of genetic resources with a higher degree of similarity to the parental lines could be reduced. If the genetic variation and distance between least-related accessions are sufficiently large in the parent population, then a progeny population with greater genetic variation can be obtained (Mikel et al. 2010). Using our SNP array data, we confirmed that the greater the genetic distance among parents, the higher the rate over best-parent of progeny. Intuitively, to obtain progeny with higher absolute trait values, the parental lines should also show high performance for those traits. In summary, our results suggest a strategy for assembling parents with a greater chance of obtaining excellent progeny, while avoiding blindly formulating a large number of suboptimal combinations.

In soybean genomic selection research, scientists have used germplasm resources (Shu et al. 2013; Zhang et al. 2016) and breeding varieties (Jarquín et al. 2014; Ma et al. 2016; Xavier et al. 2016) or unrelated germplasm and recombinant inbred lines (Matei et al. 2018; Stewart-Brown et al. 2019) in their studies. In this study, predictive accuracy provided by GBLUPs (i.e., based on genomic information) was higher than that of ABLUP and HBLUP models, reaching an average of 0.75, and the accuracy of predicting performance in beginning maturity, 100-seed weight, protein, oil, and seed yield traits was similar or higher to that in previously reported results (Supplemental Table 16). These findings further indicate that the SNPs used in the ZDX1 are broadly representative of soybean varieties used in various breeding studies. Moreover, genomic information can better reflect the genetic structure of the breeding population than pedigree relationships.

In terms of strategies to improve the accuracy of prediction, previous studies have shown that selecting a subset of high effect markers can effectively improve accuracy (e Sousa et al. 2019; Liu et al. 2019; Ma et al. 2016), although the sampling strategy is complex and varies from population to population. Here, we found that sampling SNPs located within genic regions was more informative than sampling SNPs from intergenic or random regions, and therefore marker effects do not need to be considered for each different population. Indeed, sampling SNPs from genic regions can ensure or even significantly improve the accuracy of prediction and reduce sequencing costs (Song et al. 2020). In addition, expanding the training group can also improve the accuracy of predictions (Hao et al. 2019). In the current study, we expanded the training population of parental lines to include progeny with higher predicted values for traits of interest, which dramatically increased the accuracy of predicting

beginning maturity, 100-seed weight, protein content and oil content values among progeny by 32.1% on average. Marker selection, BLUPs modeling, and expanding the training set based on marker data from the ZDX1 SNP array can thus improve the accuracy of predicting progeny phenotypes.

In traditional plant breeding, breeders mainly rely on phenotype and experience, which may be confounded by a range of factors(Barabaschi et al. 2016). Molecular breeding is therefore considered the best option for improving breeding efficiency (Chen et al. 2014). However, molecular techniques have thus far failed to effectively integrate high-throughput genotyping with the whole breeding process. In this study, we propose an optimization strategy to comprehensively improve the breeding processes of parental evaluation, selection for crosses, and progeny selection using the ZDX1 array (Fig. 6). The low availability of phenotypically ideal germplasm accessions that can be used as parental lines represents a major bottleneck in the breeding progress, so it is necessary to introduce new, high quality germplasm that affects any given trait (Liu et al. 2020a). However, when breeders introduce new genetic resources, genotypic information should be examined first to eliminate germplasm that is redundant with lines in the original pool of parents. When selecting parents, breeders often rely on phenotype or experience to make parent combinations in subsequent crosses. To address this issue, Qi et al (Qi et al. 2021) once proposed the “Potalaization” concept of selecting highly diverse genetic materials from elite germplasm collections as parents. In our study, accessions with high breeding values were used as the candidate parents for crosses to ensure that progeny in the segregating population had higher breeding index values (Vianna et al. 2003). We then further screened the progeny of crosses between High BI parents with distant relatives to expand the distribution range of the progeny population and effectively increase the probability of obtaining elite progeny. If the main agronomic traits of the parents are similar, it is difficult to identify the F_1 based on phenotype alone, whereas heterozygosity in the microarray data can help genotype the progeny. Segregation data in the F_2 and subsequent generations should be combined with phenotypic selection and genome-wide selection to rapidly increase the frequency of molecular marker alleles associated with favorable traits in the breeding population (Liu et al. 2018). Functional markers can be used in foreground selection of target traits and combined with array data for whole-genome background scanning and selection of progeny with traits similar to that of recurrent parents in the breeding population (Jordan et al. 2011), ultimately improving the efficiency of backcrosses. In this study, we combined an affordable and high-throughput functional SNP array ZDX1 to improve conventional breeding procedures, thus successfully applying molecular breeding principles to both theoretical breeding models and selection in the field.

Declarations

Acknowledgements

This research was supported by the Natural Science Foundation of Iner Mongolia (2020MS03004), the Agricultural Science and Technology Innovation Program (ASTIP) of Chinese Academy of Agricultural Sciences (CAAS-ZDRW202003-1), National Key R&D Program of China (2017YFD0102002), National Key R&D Program (2019YFE0105900), Project of Sino-Uruguayan Joint Laboratory (2018YFE0116900), Youth Innovation Fund Project of Inner Mongolia Academy of Agriculture & Animal Husbandry Sciences (2020QNJJNO5).

Author contribution statement

LQ, YL and RS designed the study; BS, JW, PY, WZ, ES, QL, SC and LF provided soybean accessions; RS performed the experiments, with the assistance of LQ, YL, YZ, XH, QZ and RG; LQ, YL, SS, and JL desiged the array; RS, YT,

BG, HL, YL, HG, YG, LY, YM, HH, JX, and JW analyzed the data; RS, LQ and YL interpreted the results and wrote the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. Bailey-Serres J, Parker JE, Ainsworth EA, Oldroyd GED, Schroeder JI (2019) Genetic strategies for improving crop yields. Nature 575:109–118
2. Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, Vale G, Cattivelli L (2016) Next generation breeding. Plant Sci 242:3–13
3. Chen HD, Xie WB, He H, Yu HH, Chen W, Li J, Yu RB, Yao Y, Zhang WH, He YQ, Tang XY, Zhou FS, Deng XW, Zhang QF (2014) A high-density SNP genotyping array for rice biology and molecular breeding. Mol Plant 7:541–553
4. Cook DE, Lee TG, Guo XL, Melito S, Wang K, Bayless AM, Wang JP, Hughes TJ, Willis DK, Clemente TE, Diers BW, Jiang J, Hudson ME, Bent AF (2012) Copy number variation of multiple genes at *rhg1* mediates nematode resistance in soybean. Science 338:1206–1209
5. Cui YR, Li RD, Li GW, Zhang F, Zhu TT, Zhang QF, Ali J, Li ZK, Xu SZ (2020) Hybrid breeding of rice via genomic selection. Plant Biotechnol J 18:57–67
6. e Sousa MB, Galli G, Lyra DH, Granato ÍSC, Matias FI, Alves FC, Fritsche-Neto R (2019) Increasing accuracy and reducing costs of genomic prediction by marker selection. Euphytica 215:18
7. Fang C, Li WY, Li GQ, Wang Z, Zhou ZK, Ma YM, Shen YT, Li CC, Wu YS, Zhu BG, Yang WC, Tian ZX (2013) Cloning of *Ln* gene through combined approach of map-based cloning and association study in soybean. J Genet Genomics 40:93–96
8. Fehr WR, Caviness CE, Burmood DT, Pennington JS (1971) Stage of development descriptions for soybeans, *Glycine Max* (L.) Merrill. Crop Sci 11:929–931
9. GenomeStudio (2008) GenomeStudioTM genotyping module v1.0 user guide
10. Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R (2015) ASReml user guide release 4.1 functional specification. VSN International Ltd, Hemel Hempstead
11. Gunderson KL (2009) Whole-Genome Genotyping on Bead Arrays. Methods mol biol 529:197–213
12. Hao YF, Wang HW, Yang XH, Zhang HW, He C, Li DD, Li HH, Wang GY, Wang JK, Fu JJ (2019) Genomic prediction using existing historical data contributing to selection in biparental populations: a study of kernel oil in maize. Plant Genome 12:1–9
13. He S, Schulthess AW, Mirdita V, Zhao YS, Korzun V, Bothe R, Ebmeyer E, Reif JC, Jiang Y (2016) Genomic selection in a commercial winter wheat population. Theor Appl Genet 129:641–651
14. Hegstad JM, Nelson RL, Renny-Byfield S, Feng L, Chaky JM (2019) Introgression of novel genetic diversity to improve soybean yield. Theor Appl Genet 132:2541–2552

15. Hwang EY, Song QJ, Jia GF, Specht JE, Hyten DL, Costa J, Cregan PB (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genom 15:1–12
16. Iquira E, Gagnon E, Belzile F (2010) Comparison of genetic diversity between Canadian adapted genotypes and exotic germplasm of soybean. Genome 53:337–345
17. Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genom 15:1–10
18. Jeong N, Suh SJ, Kim MH, Lee S, Moon JK, Kim HS, Jeong SC (2012) *Ln* is a key regulator of leaflet shape and number of seeds per pod in soybean. Plant Cell 24:4807–4818
19. Ji Y, Zhao D, Chen X, Yong Z, Wang J (2018) Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). The Crop J 6:353–365
20. Jordan DR, Mace ES, Cruickshank AW, Hunt CH, Henzell RG (2011) Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. Crop Sci 51:1444–1457
21. Langewisch T, Zhang HX, Vincent R, Joshi T, Xu D, Bilyeu K (2014) Major soybean maturity gene haplotypes revealed by SNPViz analysis of 72 sequenced soybean genomes. PLoS One 9:94150
22. Lee YG, Jeong N, Kim JH, Lee K, Kim KH, Pirani A, Ha BK, Kang ST, Park BS, Moon JK, Kim N, Jeong SC (2015) Development, validation and genetic analysis of a large soybean SNP genotyping array. Plant J 81:625–636
23. Lehermeier C, Wimmer V, Albrecht T, Auinger HJ, Gianola D, Schmid VJ, Schön CC (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. Stat Appl Genet Mol Biol 12:375–391
24. Li C, Li YH, Li YF, Lu H, Hong HL, Tian Y, Li HY, Zhao T, Zhou XW, Liu J, Zhou X, Jackson SA, Liu B, Qiu LJ (2020) A domestication-associated gene *GmPRR3b* regulates the circadian clock and flowering time in soybean. Mol Plant 13:745–759
25. Li XJ, Wang S, Huang J, Li LY, Zhang Q, Ding XD (2014a) Improving the accuracy of genomic prediction in Chinese Holstein cattle by using one-step blending. Genet Sel Evol 46:66
26. Li YH, Zhou GY, Ma JX, Jiang WK, Jin LG, Zhang ZH, Guo Y, Zhang JB, Sui Y, Zheng LT, Zhang SS, Zuo QY, Shi XH, Li YF, Zhang WK, Hu YY, Kong GY, Hong HL, Tan B, Song J, Liu ZX, Wang YS, Ruan H, Yeung CK, Liu J, Wang HL, Zhang LJ, Guan RX, Wang KJ, Li WB, Chen SY, Chang RZ, Jiang Z, Jackson SA, Li RQ, Qiu LJ (2014b) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32:1045–1052
27. Liu SL, Zhang M, Feng F, Tian ZX (2020a) Toward a "Green Revolution" for soybean. Mol Plant 13:688–697
28. Liu SM, Kandoth PK, Warren SD, Yeckel G, Heinz R, Alden J, Yang CL, Jamai A, El-Mellouki T, Juvale PS, Hill J, Baum TJ, Cianzio S, Whitham SA, Korkin D, Mitchum MG, Meksem K (2012) A soybean cyst nematode resistance gene points to a new mechanism of plant resistance to pathogens. Nature 492:256–260
29. Liu XG, Wang HW, Hu XJ, Li K, Liu ZF, Wu YJ, Huang CL (2019) Improving genomic selection with quantitative trait loci and nonadditive effects revealed by empirical evidence in maize. Front Plant Sci 10:1129
30. Liu XG, Wang HW, Wang H, Guo ZF, Xu XJ, Liu JC, Wang SH, Li WX, Zou C, Prasanna BM, Olsen MS, Huang CL, Xu YB (2018) Factors affecting genomic selection revealed by empirical evidence in maize. The Crop J 6:341–352
31. Liu YC, Du HL, Li PC, Shen YT, Peng H, Liu SL, Zhou GA, Zhang HK, Liu Z, Shi M, Huang XH, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang CZ, Tian ZX (2020b) Pan-genome of wild and cultivated soybeans. Cell

182:162–176 e113

32. Lourenco D, Legarra A, Tsuruta S, Masuda Y, Aguilar I, Misztal I (2020) Single-step genomic evaluations from theory to practice: using SNP chips and sequence data in BLUPF90. Genes 11:790
33. Ma YS, Reif JC, Jiang Y, Wen ZX, Wang DC, Liu ZX, Guo Y, Wei SH, Wang SM, Yang CM, Wang HC, Yang CY, Lu WG, Xu R, Zhou R, Wang RZ, Sun ZD, Chen HZ, Zhang WH, Wu JA, Hu GH, Liu CY, Luan XY, Fu YS, Guo T, Han TF, Zhang MC, Sun BC, Zhang L, Chen WY, Wu CX, Sun S, Yuan BJ, Zhou XA, Han DZ, Yan HR, Li WB, Qiu LJ (2016) Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). Mol Breed 36:113
34. Marrano A, Martinez-Garcia PJ, Bianco L, Sideli GM, Di Pierro EA, Leslie CA, Stevens KA, Crepeau MW, Troggio M, Langley CH, Neale DB (2019) A new genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density Axiom *J. regia* 700K SNP genotyping array. Plant Biotechnol J 17:1027–1036
35. Matei G, Woyann LG, Milioli AS, de Bem Oliveira I, Zdziarski AD, Zanella R, Coelho ASG, Finatto T, Benin G (2018) Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. Mol Breed 38:1–13
36. Meng L, Li HH, Zhang LY, Wang JK (2015) QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. The Crop J 3:269–283
37. Mikel MA, Diers BW, Nelson RL, Smith HH (2010) Genetic diversity and agronomic improvement of north american soybean germplasm. Crop Sci 50:1219–1229
38. Pandey MK, Agarwal G, Kale SM, Clevenger J, Nayak SN, Sriswathi M, Chitikineni A, Chavarro C, Chen X, Upadhyaya HD, Vishwakarma MK, Leal-Bertioli S, Liang X, Bertioli DJ, Guo B, Jackson SA, Ozias-Akins P, Varshney RK (2017) Development and evaluation of a high density genotyping 'Axiom_Arachis' array with 58K SNPs for accelerating genetics and breeding in groundnut. Sci Rep 7:1–10
39. Pérez P, de Los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495
40. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, Bakker P, Daly MJ (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81:559–575
41. Qi XP, Jiang BJ, Wu TT, Sun S, Wang CJ, Song WW, Wu CX, Hou WS, Song QJ, Lam HM (2021) Genomic dissection of widely planted soybean cultivars leads to a new breeding strategy of crops in the post-genomic era. The Crop J. https://doi.org/10.1016/j.cj.2021.01.001
42. Qiu LJ, Chang RZ, Liu ZX, Guan RX, Li YH (2006) Descriptors and data standard for soybean (*Glycine* spp.). China Agriculture Press, Beijing
43. Rasheed A, Hao YF, Xia XC, Khan A, Xu YB, Varshney RK, He ZH (2017) Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol Plant 10:1047–1064
44. Rincker K, Nelson R, Specht J, Sleper D, Cary T, Cianzio SR, Casteel S, Conley S, Chen P, Davis V, Fox C, Graef G, Godsey C, Holshouser D, Jiang GL, Kantartzi SK, Kenworthy W, Lee C, Mian R, McHale L, Naeve S, Orf J, Poysa V, Schapaugh W, Shannon G, Uniatowski R, Wang D, Diers B (2014) Genetic improvement of U.S. soybean in maturity groups II, III, and IV. Crop Sci 54:1419–1432
45. Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, Hyten DL, Song QJ, Thelen JJ, Cheng JL, 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 SQ, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du JC, Tian ZX, Zhu LC, 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

46. Shen R, Fan JB, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Garcia EW, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A (2005) High-throughput SNP genotyping on universal bead arrays. Mutat Res 573:70–82
47. Shi Z, Liu SM, Noe J, Arelli P, Meksem K, Li ZL (2015) SNP identification and marker assay development for high-throughput selection of soybean cyst nematode resistance. BMC Genom 16:1–12
48. Shu YJ, Yu DS, Wang D, Bai X, Zhu YM, Guo CH (2013) Genomic selection of seed weight based on low-density SCAR markers in soybean. Genet Mol Res 12:2178–2188
49. Silva-Junior OB, Faria DA, Grattapaglia D (2015) A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. New Phytol 206:1527–1540
50. Song H, Zhang J, Jiang Y, Gao H, Tang S, Mi S, Yu F, Meng Q, Xiao W, Zhang Q (2017) Genomic prediction for growth and reproduction traits in pig using an admixed reference population. J Anim Sci 95:3415–3424
51. Song HL, Zhang JX, Zhang Q, Ding XD (2019) Using different single-step strategies to improve the efficiency of genomic prediction on body measurement traits in pig. Front Genet 9:1–10
52. Song QJ (2014) Soybean BARCSoySNP6K beadchip - a Tool for soybean genetics research. Plant & Animal Genome:10–15
53. Song QJ, Hyten DL, Jia GF, Quigley CV, Fickus EW, Nelson RL, Cregan PB (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8:e54985
54. Song QJ, Yan L, Quigley C, Fickus E, Wei H, Chen LF, Dong FM, Araya S, Liu JL, Hyten D, Pantalone V, Nelson RL (2020) Soybean BARCSoySNP6K: An assay for soybean genetics and breeding research. Plant J 104:800–811
55. Stewart-Brown BB, Song QJ, Vaughn JN, Li ZL (2019) Genomic selection for yield and seed composition traits within an applied soybean breeding program. G3-Genes Genom Genet 9:2253–2265
56. Tardivel A, Sonah H, Belzile F, Donoughue LSO (2014) Rapid identification of alleles at the soybean maturity gene *E3* using genotyping by sequencing and a haplotype-based approach. Plant Genome 7:1–9
57. Tian Y, Liu B, Shi XH, Reif JC, Guan R, Li YH, Qiu LJ (2019) Deep genotyping of the gene *GmSNAP* facilitates pyramiding resistance to cyst nematode in soybean. The Crop J 7:677–684
58. Tian Y, Yang L, Li YH, Qiu LJ (2018) Development and utilization of KASP marker for *SCN3-11* locus resistant to soybean cyst nematode. Acta Agronomica Sinica 44:26–37
59. Tian ZX, Wang XB, Lee R, Li YH, Specht JE, Nelson RL, McClean PE, Qiu LJ, Ma JX (2010) Artificial selection for determinate growth habit in soybean. Proc Natl Acad Sci (PNAS) 107:8563–8568
60. Tsubokura Y, Matsumura H, Xu ML, Liu BH, Nakashima H, Anai T, Kong FJ, Yuan XH, Kanamori H, Katayose Y, Takahashi R, Harada K, Abe J (2013) Genetic variation in soybean at the maturity Locus *E4* is involved in adaptation to long days at high latitudes. Agronomy 3:117–134
61. Tsubokura Y, Watanabe S, Xia ZJ, Kanamori H, Yamagata H, Kaga A, Katayose Y, Abe J, Ishimoto M, Harada K (2014) Natural variation in the genes responsible for maturity loci *E1*, *E2*, *E3* and *E4* in soybean. Ann Bot 113:429–441

62. Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLoS Biol 12:1001883
63. Vianna BPA, Olívio GI, Carneiro VML, Eduardo PC, Santos DCTD (2003) Predicting performance of soybean populations using genetic distances estimated with RAPD markers. Genet Mol Biol 26:343–348
64. Viquez-Zamora M, Vosman B, van de Geest H, Bovy A, Visser RGF, Finkers R, van Heusden AW (2013) Tomato breeding in the genomics era: insights from a SNP array. BMC Genom 14:1–13
65. Wang J, Chu SS, Zhang HR, Zhu Y, Cheng H, Yu DY (2016) Development and application of a novel genome-wide SNP array reveals domestication history in soybean. Sci Rep 6:1–10
66. Wang M, Li WZ, Fang C, Xu F, Liu YC, Wang Z, Yang R, Zhang M, Liu SL, Lu SJ, Lin T, Tang JY, Wang YQ, Wang HR, Lin H, Zhu BG, Chen MS, Kong FJ, Liu BH, Zeng DL, Jackson SA, Chu CC, Tian ZX (2018a) Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat Genet 50:1435–1441
67. Wang YY, Li YQ, Wu HY, Hu B, Zheng JJ, Zhai H, Lv SX, Liu XL, Chen X, Qiu HM, Yang JY, Zong CM, Han DZ, Wen ZX, Wang DC, Xia ZJ (2018b) Genotyping of soybean cultivars with medium-density array reveals the population structure and QTNs underlying maturity and seed traits. Front Plant Sci 9:610
68. Xavier A, Muir WM, Rainey KM (2016) Assessing predictive properties of genome-wide selection in soybeans. G3-Genes. Genom Genet 6:2611–2616
69. Xia ZJ, Watanabe S, Yamada T, Tsubokura Y, Nakashima H, Zhai H, Anai T, Sato S, Yamazaki T, Lu SX, Wu HY, Tabata S, Harada K (2012) Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering. Proc Natl Acad Sci (PNAS) 109:2155–2164
70. Xu ML, Xu ZH, Liu BH, Kong FJ, Tsubokura Y, Watanabe S, Xia ZJ, Harada K, Kanazawa A, Yamada T, Abe J (2013) Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. BMC Plant Biol 13:91
71. Xu YB, Liu XG, Fu JJ, Wang HW, Wang JK, Huang CL, Prasanna BM, Olsen MS, Wang GY, Zhang AM (2020) Enhancing genetic gain through genomic selection: from livestock to plants. Plant Commun 1:100005
72. Yu HH, Xie WB, Li J, Zhou FS, Zhang QF (2014) A whole-genome SNP array (RICE6K) for genomic breeding in rice. Plant Biotechnol J 12:28–37
73. Zhang DJ, Sun LJ, Li S, Wang WD, Ding YH, Swarm SA, Li LH, Wang XT, Tang XM, Zhang ZF, Tian ZX, Brown PJ, Cai C, Nelson RL, Ma JX (2018) Elevation of soybean seed oil content through selection for seed coat shininess. Nat Plants 4:30–35
74. Zhang JP, Song QJ, Cregan PB, Jiang GL (2016) Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). Theor Appl Genet 129:117–130
75. Zhao SL, Jing W, Samuels DC, Sheng QH, Shyr Y, Guo Y (2018) Strategies for processing and quality control of Illumina genotyping arrays. Brief Bioinform 19:765–775
76. Zhao YS, Li Z, Liu GZ, Jiang Y, Maurer HP, Wurschum T, Mock HP, Matros A, Ebmeyer E, Schachschneider R, Kazman E, Schacht J, Gowda M, Longin CF, Reif JC (2015) Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc Natl Acad Sci (PNAS) 112:15624–15629
77. Zhe Z, Erbe M, He JL, Ober U, Li JQ (2015) Accuracy of whole genome prediction using a genetic architecture enhanced variance-covariance matrix. G3-Genes Genom Genet 5:615–627

Figures

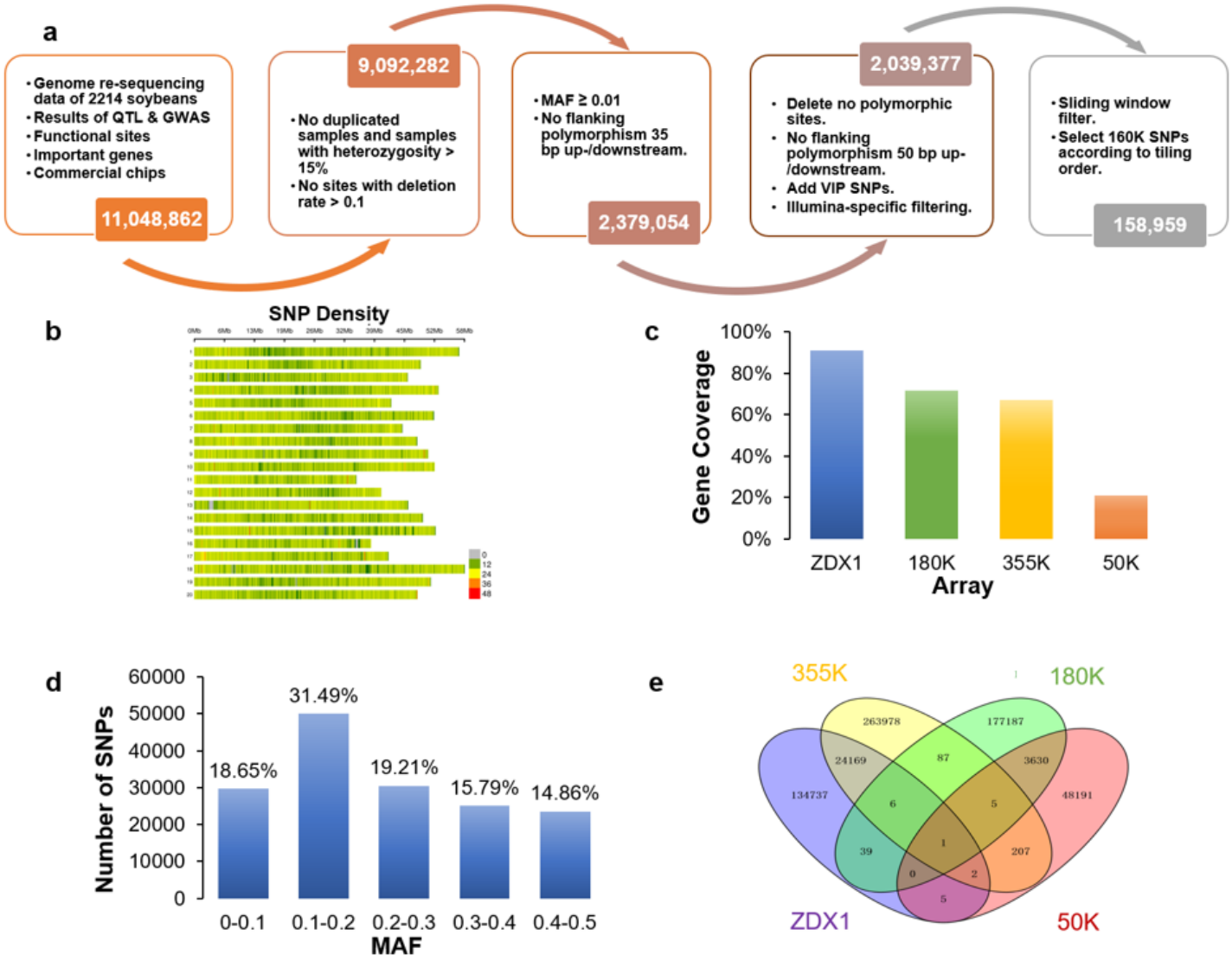

Figure 1

Summary information content of ZDX1 array. (a) Pipeline of SNP identification and selection for the ZDX1 array. (b) The distribution of SNP loci on the soybean chromosomes. (c) The percentage of gene coverage in the ZDX1 array, the SoySNP50K array, the 180K AXIOM® array, and the NJAU 355K SoySNP array. (d) The number of SNPs belonging to different minor allele frequency (MAF) classes based on 2,214 soybean accessions. (e) Venn diagram showing the overlap of SNP positions between the ZDX1, SoySNP50K, 180K AXIOM®, and NJAU 355K SoySNP arrays

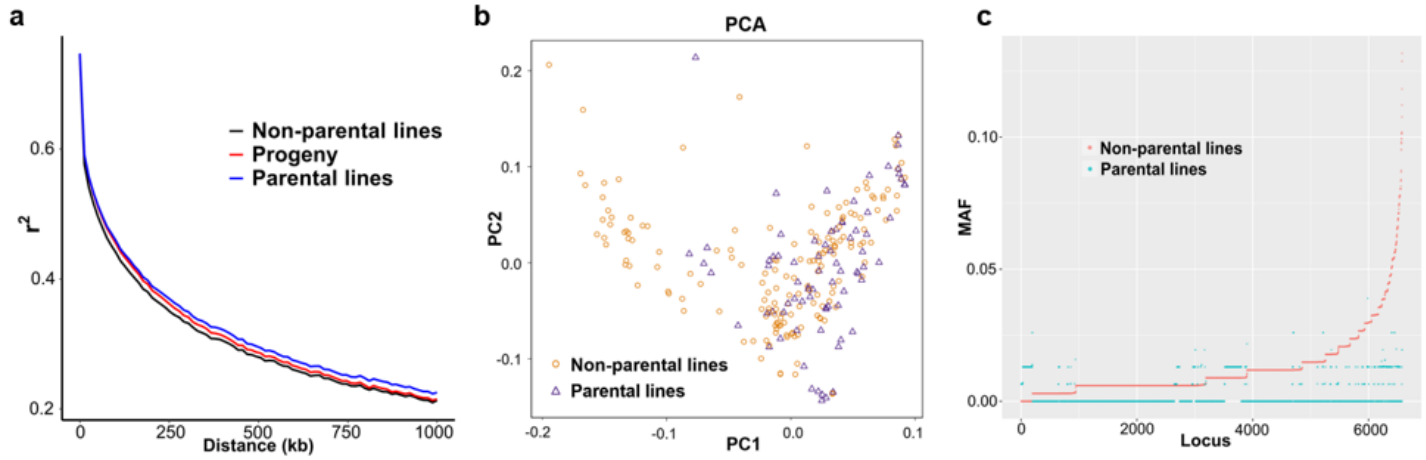

Figure 2

Analysis of genetic diversity of breeding population and screening of fixed sites in breeding improvement. (a) LD decay of r2 and physical distance between SNP markers in parental lines, non-parental lines, and progeny. (b) Principal component analysis (PCA) of 77 parental lines and 169 non-parental lines based on kinship. Individuals from the same species are shown in the same color. (c) A scatter plot showing the minor allele frequencies (MAFs) for the parental lines and non-parental lines at 6,579 sites with MAF of progeny =0

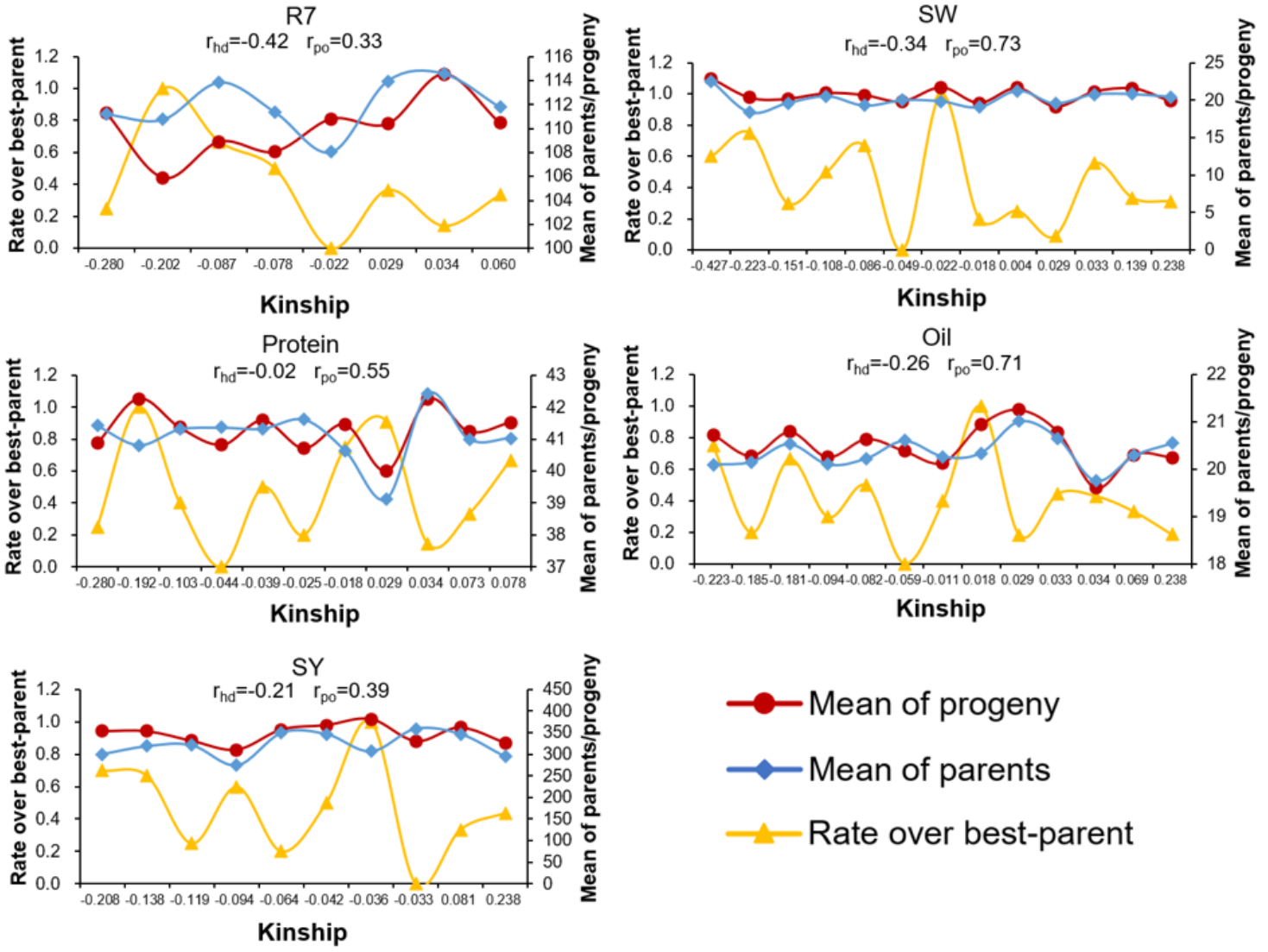

Figure 3

Mean value of parents and progeny, and the rate over best-parent of progeny for five traits at plotted against genetic distance. The blue diamonds represent the average parental value, the red circles represent the average progeny, and the yellow triangles represents the rate over best-parent of progeny. The genetic distance is the mean value under different rate over best-parent; rhd represents the correlation coefficient between the rate over best-parent of progeny and the genetic relationship between parents; and rpo represents the correlation coefficient between the mean value of progeny and the mean value of parents

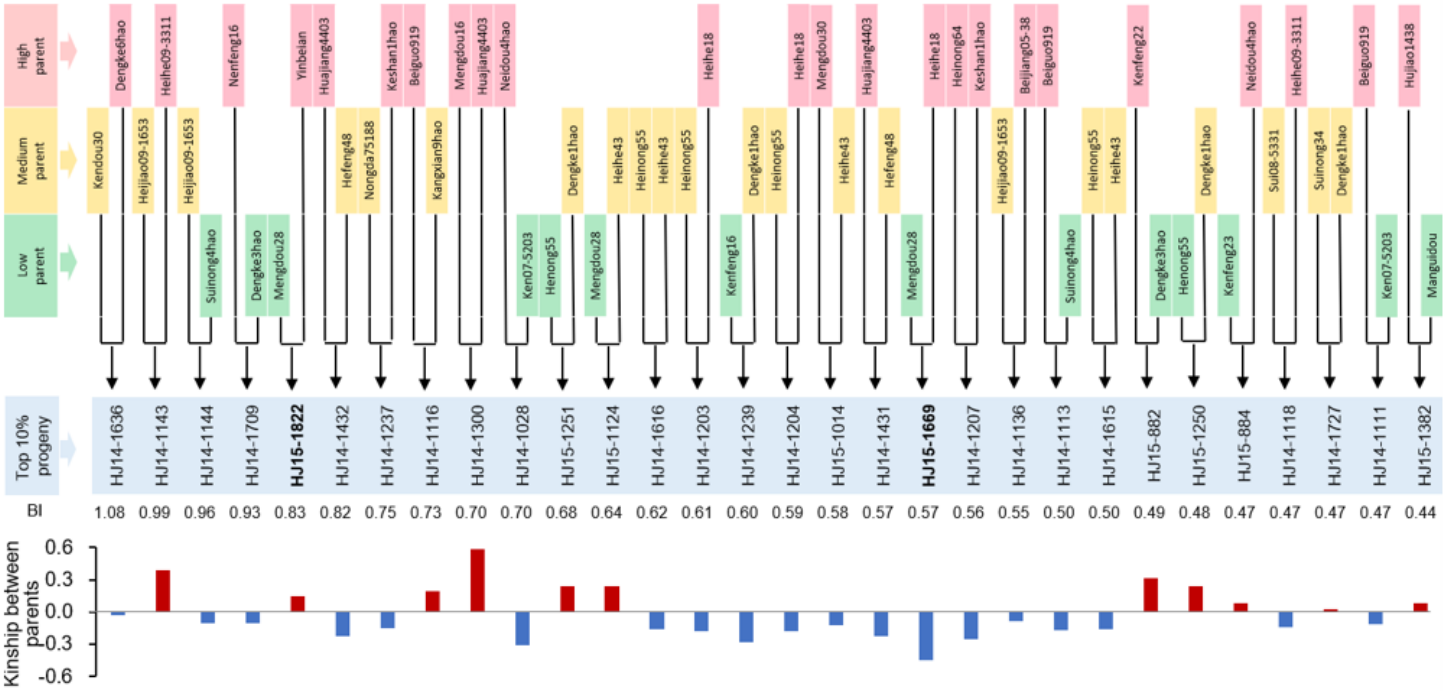

Figure 4

The relationship between the top 10% of progeny in multiple traits and their parental lines. The blue box in the center is the top 10% of progeny with BI. They are arranged in order from high to low from left to right. The BI values are given below the box. The parents of these lines are classified by BI value; the top 1/3 of lines with the highest BI values are the high parents, the middle 1/3 are the medium parents, and the bottom 1/3 are the Low parents. The bar graph at the bottom shows the kinship between the parental lines

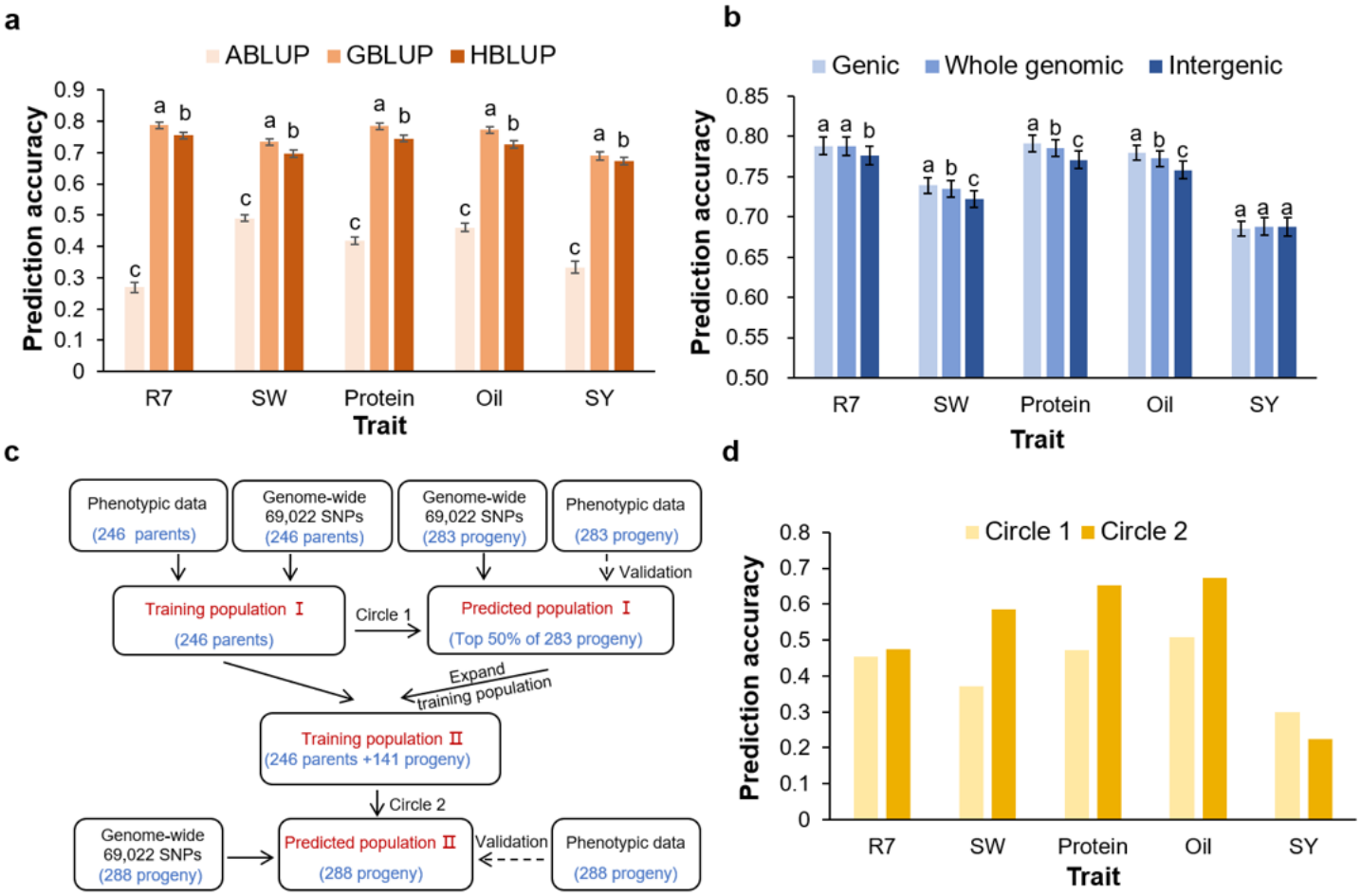

Figure 5

Different strategies based on ZDX1 array in genomic selection. (a) The prediction accuracy (rGS) of three models for five traits with 100 repetitions using 5-fold cross-validation. The prediction accuracy is shown as the mean value ± standard deviation. (b) Prediction accuracy of selected sites for gene region, whole genome, and intergenic region markers. The prediction accuracy is shown as the mean value ± standard deviation. (c) Simulating the process of predicting progeny performance by parental resources in actual breeding and the prediction process after using progeny to expand the training population. (d) Prediction accuracy for five traits for the 246 parents (Training Population I) and 246 parents + 141 progeny (Training Population II) used as training populations for prediction

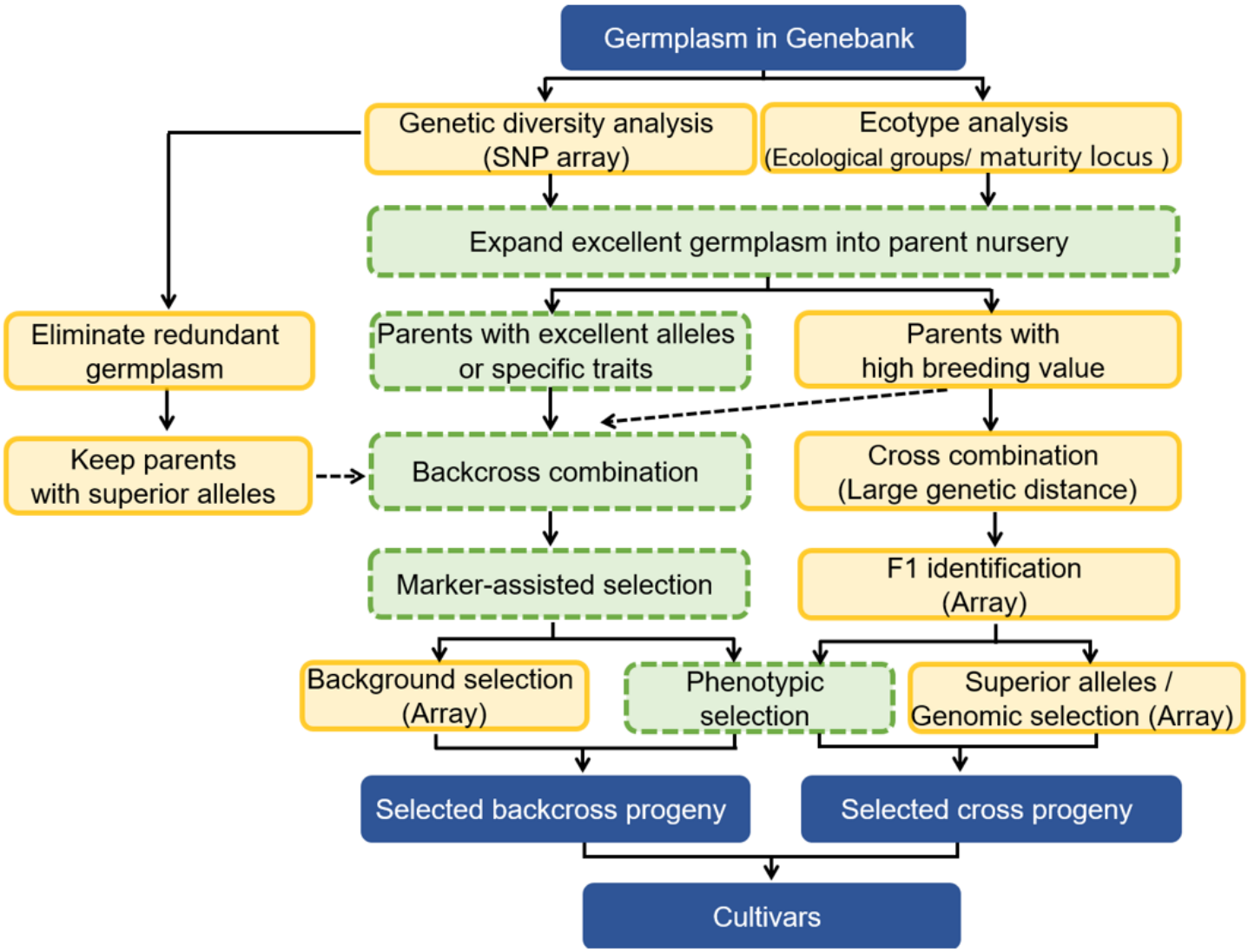

Figure 6

Optimized scheme for using genome-wide molecular marker breeding combined with array screening. Germplasm resources are introduced from a resource bank, redundant accessions are eliminated through genetic diversity analysis, and accessions with excellent alleles are retained. Germplasm accessions with higher breeding index (BI) are used as one of the candidate parents in cross breeding, and the superior resources are further screened for those with highly distant genetic relationships for cross breeding. A microarray is then used for F1 identification, hybrid segregation combined with phenotypic selection, and whole-genome selection. Germplasm with high breeding values with excellent multiple traits can also be used as recurrent parents, when germplasm with specific traits is used for backcross improvement, functional markers can be used for foreground selection, and microarrays can be used for genome-wide background scanning, combined with phenotypes for selection, and excellent stable lines can be selected. The green dashed boxes indicate the commonly used breeding method, and the boxes enclosed by solid yellow lines represent the improved scheme proposed in this study

Supplementary Files

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

- SupplementaryFigures.docx
- SupplementalTablesZDX1soybeanarraySunRujian.xlsx