SCIENCE CHINA
Life Sciences

•LETTER TO THE EDITOR•

September 2022 Vol.65 No.9: 1898–1901
https://doi.org/10.1007/s11427-022-2130-8

GenoBaits Soy40K: a highly flexible and low-cost SNP array for soybean studies

Yucheng Liu[1†], Shulin Liu[1†], Zhifang Zhang[1], Lingbin Ni[1,3], Xingming Chen[2], Yunxia Ge[2], Guoan Zhou[1] & Zhixi Tian[1,3*]

[1]*State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China;*
[2]*MolBreeding Biotechnology Co., Ltd., Shijiazhuang 050035, China;*
[3]*University of Chinese Academy of Sciences, Beijing 100039, China*

Received April 2, 2022; accepted May 17, 2022; published online May 26, 2022

Citation: Liu, Y., Liu, S., Zhang, Z., Ni, L., Chen, X., Ge, Y., Zhou, G., and Tian, Z. (2022). GenoBaits Soy40K: a highly flexible and low-cost SNP array for soybean studies. Sci China Life Sci 65, 1898–1901. https://doi.org/10.1007/s11427-022-2130-8

Dear Editor,

Soybean (*Glycine max* [L.] Merr.) provides more than half of the oilseeds and more than a quarter of protein worldwide. It is estimated that the production of soybean has to be doubled by 2050 to meet the needs of the rapidly increasing consumption of soybean seeds along with a continuously increasing population (Ray et al., 2013). As such, development of a genotyping platform with high throughput, high efficiency and high precision but low-cost is urgently needed to accelerate soybean functional study and molecular design breeding.

SNP array is a robust high-throughput genotyping technology that is less expensive than next-generation sequencing and genotyping by sequencing. To date, SNP arrays have been widely used in various studies, including those involving population structure investigations, genetic dissection, and molecular breeding (Rasheed et al., 2017). However, routine chip SNP arrays have a substantial deficiency: once the designed SNP probes are fixed in the chip array, the targeting SNPs are determined and cannot be adjusted. Liquid chip technology, referred to as genotyping by target sequencing (GBTS), has shown promise for organizing and updating SNP markers in a timely manner; moreover, the cost of this technology is much lower than that of routine chip SNP arrays (Xu et al., 2020).

To develop a GBTS liquid chip for soybean, polymorphisms from the pan-genome of 27 soybean accessions and whole-genome resequencing data of 2,898 accessions (Liu et al., 2020b) were used to select the genetic background, and polymorphisms from the regions conferring agronomic traits (Zhang et al., 2022) were used to select the genetic foreground. After a selection pipeline (Figure 1A; Figure S1 in Supporting Information) and two rounds of genotyping quality evaluation with 562 soybean accessions, polymorphisms from 40,334 regions were ultimately selected as targets. Of these 40,334 regions, 40,019 contained SNPs only, and 315 contained insertions-deletions (INDELs). Each region of this array contained 1 to 7 detectable polymorphisms. The GBTS liquid chip was named GenoBaits Soy40K array.

Of the GenoBaits Soy40K array, more variations occurred on the chromosome arms than within the centromere (Figure 1B). The average interval length between adjacent regions was 24.7 kb, with 44.4% primarily distributed from 10 to 20 kb, and 7.2% had an interval larger than 50 kb (Figure 1C). The Pearson correlation coefficient (R) between polymorphism number and chromosome length was 0.81 (P

†Contributed equally to this work
*Corresponding author (email: zxtian@genetics.ac.cn)

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

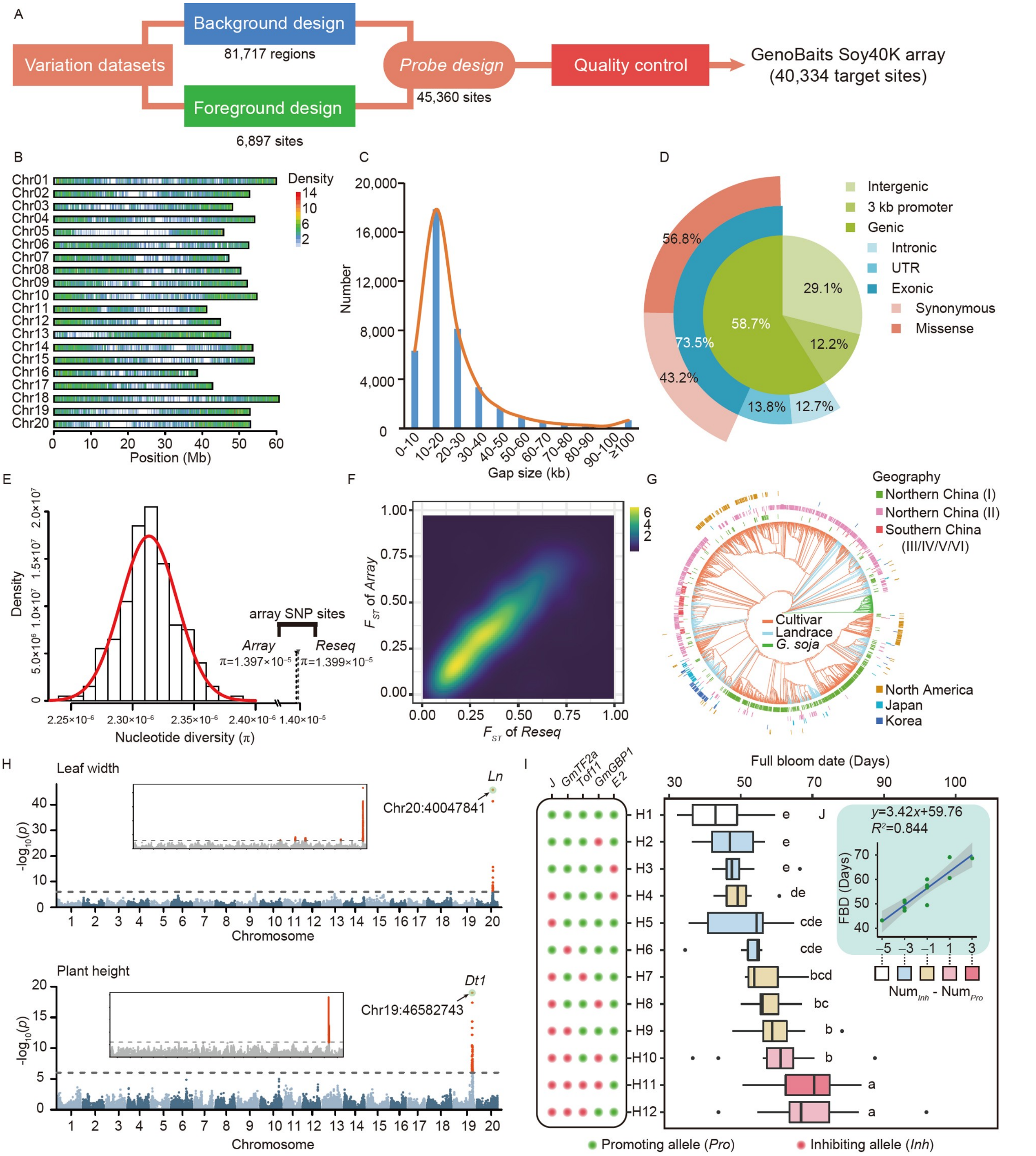

Figure 1 Character of GenoBaits Soy40K array and its application scenario. A, Design pipeline of the soybean GenoBaits Soy40K array. B, SNP marker density within 500-kb windows throughout the whole genome. C, Number of variations in different gap-size sections between adjacent markers. D, SNP locations within the genome. E, Distribution of average π values calculated by 200 random selections and the normal distribution estimation (histogram and red line). The average π is calculated by the GenoBaits Soy40K array (*Array*) and previous resequencing work (*Reseq*), respectively, and are labeled on the graph. F, F_{ST} correlations between values by the GenoBaits Soy40K arrays (*Array*) and previous resequencing work (*Reseq*). G, ML tree comprising 2,078 soybean accessions. The colored lines represent groups of cultivar, landrace and *G. soja*; the colored blocks represent geographic regions, such as North America, Japan, Korea and the I–VI ecoregions of soybean in China. H, GWAS results of leaflet width (top) and plant height (bottom); the Manhattan plots in the small frames represent GWAS results from previous resequencing work for the same sample set; Chr20:40047841 and Chr19:46582743 are reported functional SNPs related to trait. I, Full bloom date of haplotypes formed by the genotypes of 5 flowering-related genes. Haplotypes formed with different number of inhibiting (*Inh*, red) and promoting (*Pro*, green) loci are colored white, blue, yellow, pink and red. The number below the haplotype shows the number of samples with phenotype/genotype data. Multiple comparisons were performed by the SNK test, with a significance level equal to 0.05. J, Relationship between Num_{Inh}−Num_{Pro} and full bloom date.

value=1.46×10^{-5}). Of the polymorphisms, 58.7% were located in the genic region, 12.2% were in the 3 kb region upstream from the start codon, and 29.1% were in the intergenic region (Figure 1D). According to the annotation of the ZH13 reference genome (version 2) (Shen et al., 2019), the target polymorphisms covered 39.9% (22,164) of the protein-coding genes. Of the polymorphisms located in exonic regions, 56.9% were synonymous variations, and 43.1% were missense variations (Figure 1D).

To evaluate the application of the GenoBaits Soy40K array in genetic diversity analysis, we genotyped 2,078 soybean accessions (2,012 *G. max* and 66 *G. soja*) collected from major countries and regions in which soybean is cultivated. The genotyping results showed that 97.9% of the samples had a calling rate greater than 95%, implying a good performance of this array in genotyping. Subsequently, we compared the genetic diversity features derived from the GenoBaits Soy40K array with those of our previous 2,898 resequencing dataset. Of these two datasets, 1,940 accessions overlapped. First, we compared the nucleotide diversity (π) of these 1,940 accessions using genotyping data from the two datasets. We observed that the π value distribution of randomly picked windows from GenoBaits Soy40K array showed normal distribution pattern as expect (Figure 1E). Consistently, similar mean π values were observed (1.397×10^{-5} for the GenoBaits Soy40K array and 1.399×10^{-5} for the resequencing data) (Figure 1E), suggesting a good genetic diversity representativeness of the GenoBaits Soy40K array. In addition, F_{ST} values between *G. soja* and *G. max* from these two datasets showed a high correlation (R^2=0.914) (Figure 1F). Pericentromeric regions have higher linkage disequilibrium (LD) values than chromosome arms (Zhou et al., 2015). Because more polymorphisms were selected from the chromosome arm than from the pericentromeric regions, the LD at the whole-genome level from the GenoBaits Soy40K array was higher than that from the resequencing dataset (Figure S2A in Supporting Information). However, when the LD from the chromosome arms was compared, the gap between the array and resequencing datasets was reduced (Figure S2B in Supporting Information).

We also assessed the performance of the GenoBaits Soy40K array in population structure analyses. We found that *G. soja* and *G. max* could be clearly classified into two monophyletic groups, and the cultivated accessions exhibited significant geographic specificity (Figure 1G). PCA showed that the top 5 principal components (PCs) explained approximately 60% of the genetic variation, and the genetic diversity of *G. soja* had a higher value than that of landraces and cultivated soybean (Figure S3 in Supporting Information). These results are consistent with the findings of previous reports (Liu et al., 2020b; Zhou et al., 2015). To assess the performance of the GenoBaits Soy40K array in terms of its ability to mine genetic loci using genome wide association study (GWAS), we performed GWAS for 26 traits using the genotypic data from the GenoBaits Soy40K array and phenotypic data from a previous study, then we compared the results from the GenoBaits Soy40K array with those of previous report (Fang et al., 2017). We found that the GWAS from the GenoBaits Soy40K array showed a high consistence with that from resequencing, particular for the major effective loci. Moreover, some of the polymorphisms with the most genetic variation were located in the sequences of functional genes. For example, the GWAS of leaflet width revealed a significant signal past the threshold ($P=1\times10^{-6}$) surrounding gene *Ln* (*SoyZH13_20G103500*) on chromosome 20; and GWAS of plant height detected a significant signal surrounding gene *Dt1* (*SoyZH13_19G179500*) (Figure 1H). The variation Chr20:40047841 with the lowest $P=1.98\times10^{-46}$ and Chr19:46582743 with the lowest $P=1.36\times10^{-19}$ are the causal allelic variations that control leaflet shape and plant height (Fang et al., 2017).

Molecular breeding aims to disrupt linkage drag and to pyramid desirable genetic alleles affording important traits into a single variety. In the GenoBaits Soy40K array, we preloaded variations of 49 reported functional genes that govern important agronomic traits (Figure S4 in Supporting Information). To explore the GenoBaits Soy40K array in function-aware haplotype map-assisted breeding, we tested the allele combination of flowering time related genes. In this study, we preloaded the functional alleles of 12 flowering-related genes in the GenoBaits Soy40K array. In the studied population, 128 homozygous haplotypes of all these flowering genes were generated based on genotyping from the GenoBaits Soy40K array. Of the total homozygous haplotypes, 12 haplotypes from 5 genes were determined to be major haplotypes in the studied population, as they were present in more than 20 individuals. According to the functional characterization of these genes in controlling flowering in soybean, we coded the different alleles of the five genes into promoting allele (*Pro*) or inhibiting (*Inh*) allele. Together with previously flowering phenotyping records (Fang et al., 2017), we investigated the relationship between the difference in inhibiting/promoting allele number ($\mathrm{Num}_{Inb}-\mathrm{Num}_{Pro}$) and date of full blooming of samples with major haplotypes (Figure 1I). We found that the date of full blooming was linearly related to $\mathrm{Num}_{Inb}-\mathrm{Num}_{Pro}$ (R^2=0.844): the lines with low $\mathrm{Num}_{Inb}-\mathrm{Num}_{Pro}$ flowered earlier; in contrast, those with a greater $\mathrm{Num}_{Inb}-\mathrm{Num}_{Pro}$ flowered later (Figure 1J). These results indicated that the flowering-related loci had a roughly additive effect. Therefore, a combination of inhibiting/promoting alleles from different genes leads to a relatively continuous flowering time of varieties, such that full bloom date lasts from 43.2 days (H1) to 69.0 days (H12).

In summary, we developed a soybean SNP array, the

GenoBaits Soy40K, based on our previous genotyping of a large number of representative soybean accessions by deep resequencing. By testing with 2,078 soybean accessions, we suggested that the GenoBaits Soy40K array had a good performance in genotyping, population structure analysis, GWASs and molecular breeding. The GenoBaits Soy40K array has more advantages than other genotyping approaches in low-cost and flexibility. Moreover, along with the development of high-throughput resequencing technology, the price will continuously decrease accordingly. Although substantial progress has been achieved in soybean functional genomics, far more genes controlling important traits need to be identified (Liu et al., 2020a; Liu and Tian, 2020; Zhang et al., 2022). As the knowledge produced by the efforts of the whole soybean research society increases, the SNP array will be conveniently improved with the adding or deleting target polymorphisms, which will better serve soybean functional studies and molecular breeding.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA24030501), the National Natural Science Foundation of China (31788103), and the Modern Seed Industry Science and Technology Innovation Team of Hebei Province (21326313D-4).*

References

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G., Zhou, Z., Yu, H., Zhang, M., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol 18, 161.

Liu, S., Zhang, M., Feng, F., and Tian, Z. (2020a). Toward a "Green Revolution" for soybean. Mol Plant 13, 688–697.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M., et al. (2020b). Pan-genome of wild and cultivated soybeans. Cell 182, 162–176.e13.

Liu, Y., and Tian, Z. (2020). From one linear genome to a graph-based pan-genome: a new era for genomics. Sci China Life Sci 63, 1938–1941.

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R.K., and He, Z. (2017). Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol Plant 10, 1047–1064.

Ray, D.K., Mueller, N.D., West, P.C., and Foley, J.A. (2013). Yield trends are insufficient to double global crop production by 2050. PLoS ONE 8, e66428.

Shen, Y., Du, H., Liu, Y., Ni, L., Wang, Z., Liang, C., and Tian, Z. (2019). Update soybean Zhonghuang 13 genome to a golden reference. Sci China Life Sci 62, 1257–1260.

Xu, Y.B., Yang, Q.N., Zheng, H.J., Xu, Y.F., Sang, Z.Q., Guo, Z.F., Peng, H., Zhang, C., Lan, H.F., Wang, Y.B., et al. (2020). Genotyping by target sequencing (GBTS) and its applications (in Chinese). Sci Agric Sin 53, 2983–3004.

Zhang, M., Liu, S., Wang, Z., Yuan, Y., Zhang, Z., Liang, Q., Yang, X., Duan, Z., Liu, Y., Kong, F., et al. (2022). Progress in soybean functional genomics over the past decade. Plant Biotechnol J 20, 256–282.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33, 408–414.

SUPPORTING INFORMATION

The supporting information is available online at https://doi.org/10.1007/s11427-022-2130-8. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.