Rothamsted Repository Download

A - Papers appearing in refereed journals

Li, D., Bai, D., Tian, Y., Li, Y.□H., Zhao, C., Wang, Q., Guo, S., Luan, X., Wang, R., Yang, J., Hawkesford, M. J., Schnable, J. C., Jin, X. and Qiu, L. 2022. Time series canopy phenotyping enables the identification of genetic variants controlling dynamic phenotypes in soybean. *Journal of Integrative Plant Biology.* https://doi.org/10.1111/jipb.13380

The publisher's version can be accessed at:

- https://doi.org/10.1111/jipb.13380

The output can be accessed at: https://repository.rothamsted.ac.uk/item/98v25/time-series-canopy-phenotyping-enables-the-identification-of-genetic-variants-controlling-dynamic-phenotypes-in-soybean.

© 11 October 2022, Please contact library@rothamsted.ac.uk for copyright queries.

Research Article
https://doi.org/10.1111/jipb.13380

Time series canopy phenotyping enables the identification of genetic variants controlling dynamic phenotypes in soybean[∞]

Delin Li[1†] , Dong Bai[1†] , Yu Tian[1†] , Ying-Hui Li[1*] , Chaosen Zhao[2] , Qi Wang[1,3] , Shiyu Guo[1,3] , Yongzhe Gu[1], Xiaoyan Luan[4], Ruizhen Wang[2], Jinliang Yang[5], Malcolm J. Hawkesford[6] , James C. Schnable[5*] , Xiuliang Jin[1*] and Li-Juan Qiu[1*]

1. The National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI)/Key Laboratory of Crop Gene Resource and Germplasm Enhancement (MOA)/Key Laboratory of Soybean Biology (Beijing) (MOA), Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China
2. Crops Research Institute of Jiangxi Academy of Agricultural Sciences, Nanchang 330200, China
3. College of Agriculture, Northeast Agricultural University, Harbin 150030, China
4. Soybean Research Institute, Heilongjiang Academy of Agricultural Sciences, Harbin 150086, China
5. Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, Nebraska 68583, USA
6. Plant Sciences Department, Rothamsted Research, West Common, Harpenden, Hertfordshire AL5 2JQ, UK
[†]These authors contributed equally to this work.
*Correspondences: Ying-Hui Li (liyinghui@caas.cn); James C. Schnable (schnable@unl.edu); Xiuliang Jin (jinxiuliang@caas.cn); Li-Juan Qiu (qiulijuan@caas.cn, Dr. Qiu is fully responsible for the distribution of the materials associated with this article)
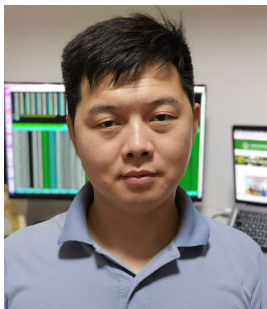
Delin Li

Li-Juan Qiu

ABSTRACT

Advances in plant phenotyping technologies are dramatically reducing the marginal costs of collecting multiple phenotypic measurements across several time points. Yet, most current approaches and best statistical practices implemented to link genetic and phenotypic variation in plants have been developed in an era of single-time-point data. Here, we used time-series phenotypic data collected with an unmanned aircraft system for a large panel of soybean (*Glycine max* (L.) Merr.) varieties to identify previously uncharacterized loci. Specifically, we focused on the dissection of canopy coverage (CC) variation from this rich data set. We also inferred the speed of canopy closure, an additional dimension of CC, from the time-series data, as it may represent an important trait for weed control. Genome-wide association studies (GWASs) identified 35 loci exhibiting dynamic associations with CC across developmental stages. The time-series data enabled the identification of 10 known flowering time and plant height quantitative trait loci (QTLs) detected in previous studies of adult plants and the identification of novel QTLs influencing CC. These novel QTLs were disproportionately likely to act earlier in development, which may explain why they were missed in previous single-time-point studies. Moreover, this time-series data set contributed to the high accuracy of the GWASs, which we evaluated by permutation tests, as evidenced by the repeated identification of loci across multiple time points. Two novel loci showed evidence of adaptive selection during domestication, with different genotypes/haplotypes favored in different geographic regions. In summary, the time-series data, with soybean CC as an example, improved the accuracy and statistical power to dissect the genetic basis of traits and offered a promising opportunity for crop breeding with quantitative growth curves.

Keywords: canopy coverage, dynamic regulation, GWAS, soybean, time series, unmanned aircraft system

INTRODUCTION

Plants undergo dynamic changes in their phenotypes throughout their life cycle as a result of both developmental changes and responses to environmental cues and stressors. A quantification of both the genetic and phenotypic differences among individuals in a given population is a necessary precondition for connecting genotype to phenotype via either linkage mapping (Ahn and Tanksley, 1993) or association mapping (Flint-Garcia et al., 2003). The development of sequencing technologies over the past decade has allowed genotyping to reach high-throughput scales that are highly accurate and cost-effective (Goodwin et al., 2016). Until recently, collecting quantitative plant traits had been comparatively slow and expensive. However, advances in plant phenotyping have brought a marked reduction in the overall cost of collecting plant trait data and in the marginal costs of collecting trait measurements during the same experiment at additional time points beyond the first one. Early reports suggested that time-series data can provide both greater detection power to identify quantitative trait loci (QTLs) and a deeper understanding of the phenotypic consequences of specific alleles at specific loci (Li and Sillanpaa, 2015). To date, many studies that use time-series trait data to map genes controlling genetic variation in plants have been performed at relatively small scale, with populations of fewer than 500 genotypes, while thousands are typically needed to reveal the genetic basis of complex traits, like yield and stress tolerance (Huang and Han, 2014), and millions are used in human studies (Lee et al., 2018).

Advances in sensors, image-processing algorithms, and deep learning algorithms are driving the emergence of diverse high-throughput phenotyping platforms (Li and Sillanpaa, 2015; Wang et al., 2020; Watt et al., 2020; Yang et al., 2020; Jin et al., 2021). One of the most widely used technology platforms for high-throughput phenotyping is the Unmanned Aircraft System (UAS). UAS refers to the combination of an unmanned aerial vehicle, one or more sensors, remote control systems, and downstream software for processing images or other sensor data to extract quantitative or qualitative measurements. UASs are robust and broadly deployable in field applications and have been successfully used to monitor plant growth and performance in response to changes in the environment, including biotic and abiotic stresses (Jin et al., 2021). For example, UASs have been used to quantify plant height (PH) in maize (*Zea mays*) (Anthony et al., 2014), flowering time (FT) in rice (*Oryza sativa*) (Guo et al., 2015), senescence rates in bread wheat (*Triticum aestivum*) (Hassan et al., 2018), nitrogen content of manila grass (*Zoysia matrella*) (Caturegli et al., 2016), multiple canopy traits in sorghum (*Sorghum bicolor*) (Shi et al., 2016), lodging in buckwheat (*Fagopyrum esculentum*) (Murakami et al., 2012), leaf chlorophyll contents in potato (*Solanum tuberosum*) (Roosjen et al., 2018), plant density in tobacco (*Nicotiana tabacum*) (Fan et al., 2018), maturity date in soybean

(*Glycine max* (L.) Merr.) (Zhou et al., 2019), and breeding programs in sorghum and maize (Pugh et al., 2018). Canopy coverage (CC) has been identified as one of the traits with the greatest potential to influence plant breeding, based on both the ease of estimation from RGB color space data and its correlation with desirable properties in crops (Jin et al., 2021).

Genome-wide association studies (GWASs) are widely used in plants to identify loci associated with variation in a diverse array of plant traits (Liu and Yan, 2019). Time-series trait data, in combination with GWASs, offer the potential to dissect the dynamic regulation of plant phenotypes. However, very few studies have combined GWASs and time-series traits to dissect the dynamic regulation of crop phenotypes, compared to the many studies that have developed methods to collect time-series data. Time-series GWASs were reported to display increased power for the identification of causal loci for PH in sorghum (Miao et al., 2020) and drought tolerance in maize (Wu et al., 2021), which were based on different optical images captured from plants grown in the greenhouse. Field-based phenotyping systems including UASs have been used to decipher several causal loci in maize (Wang et al., 2019; Anderson et al., 2020; Adak et al., 2021) and wheat (Lyra et al., 2020) behind PH at different developmental stages. Notably, these studies used fewer than four hundred genotypes for each bi-parental or natural populations and thus did not fully use the high-throughput potential of the phenotyping platform or the breadth of genetic diversity.

Soybean is a leguminous crop that was originally domesticated in China (Sedivy et al., 2017) but is now grown around the globe, where it is a critical source of both vegetable protein and oil. Canopy coverage data have been successfully collected via UASs in multiple legumes (Xavier et al., 2017; Cazenave et al., 2019; Sarkar et al., 2020). Xavier et al. (2017) phenotyped time-series CC among approximately 5,600 recombinant inbred lines (RILs) of a soybean nested association mapping (SoyNAM) population and identified seven QTLs from GWASs, based on only 4,077 single nucleotide polymorphisms (SNPs). In this study, we combined the high throughput of UASs, the diversity of soybean natural populations, and a high-density of 4 million SNPs to dissect the dynamic regulation of CC by collecting time-series CC data spanning the vegetative and reproductive stages. The population surveyed here represents the broad genetic diversity of the 23,587 cultivated soybeans from the Chinese National Soybean Gene Bank (Li et al., 2022). The 4 million SNPs with a minimum minor allele frequency of 5% were generated from whole-genome resequencing (Li et al., 2022) of the 1.1-Gb genome (Schmutz et al., 2010). Time-series CC determination of a large population improved the accuracy and statistical power of GWASs by increasing phenotype and genotype diversity and highlights the usefulness of high-throughput UAS for genetic studies. In addition, the diverse growth curves and rates of canopy closure revealed by time-series CC in this diversity panel have

potential applications in breeding programs to develop improved soybean cultivars.

RESULTS

Acquisition of time-series soybean CC

We phenotyped 1,303 soybean accessions (Figure 1A) both manually and with the help of a UAS at Nanchang (28°31′56″

N, 116°1′34″E) in 2020. These accessions were collected from around the world (Table S1) and consisted of 903 landraces and 400 improved cultivars, which comprised the four soybean sub-populations described in a previous study (Li et al., 2022). We employed a UAS (Figure 1B) to collect data reflecting the proportion of ground covered by the vertical projection of the plant canopy, here referred to as canopy coverage or CC (Figure 1C–E). The UAS deployed in the field permitted high-throughput phenotyping, as the UAS
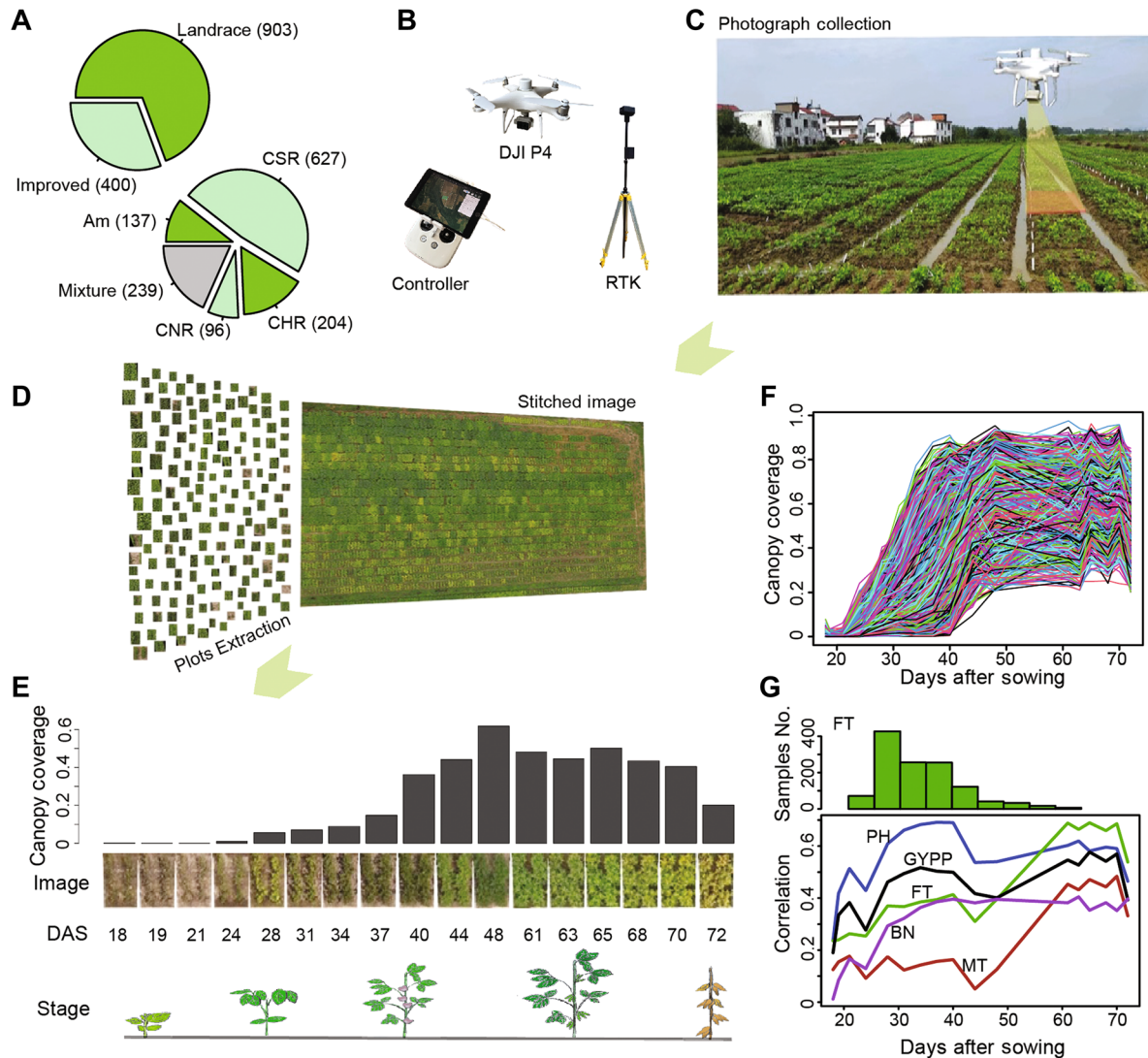

Figure 1. The soybean materials and the unmanned aircraft system used in Nanchang

(A) Breakdown of the soybean accessions used in this study between landraces and improved varieties and among the different soybean subpopulations identified (Li et al., 2022): China Northeastern region (CNR), China Huang-Huai region (CHR), China Southern region (CSR), America (Am), and non-categorized accessions (Mixture). (B) The Unmanned Aircraft System (UAS) used in Nanchang, consisting of a controller, a DJI phantom 4 multispectral drone (DJI-P4), and a D-RTK 2 mobile station (RTK) providing centimeter-level positioning. (C–E) Visual summary of the procedure used for field-based UAS phenotyping. (C) At each time point, images were collected by a UAS at a height of 12 m above the field. (D) All images are first assembled into a single field image, which was then clipped into individual images for each soybean accession at each time point. (E) An example of clipped images taken for the soybean accession "DongNong4Hao" and the corresponding canopy coverage (CC) values estimated from each clipped image. (F) Changes in CC plotted for all soybean accessions across all 17 time points. Each line indicates one accession. (G) Distribution of manually scored flowering times (FTs) for each soybean accession included in this study (top) and changes in the correlation between CC and classical developmental traits (BN, branch number; GYPP, grain yield per plant; MT, maturity time; PH, plant height) (bottom). All phenotypic data were collected in a soybean field trial conducted in Nanchang (China South) in 2020.

image collection time only took about 20 min (Table S2) as a flight altitude of 12 m above the field. This instrument facilitated image collection for 17 time points, from 18 to 72 d after sowing (DAS), which spanned the vegetative and reproductive stages. We extracted the CC values for each soybean accession from the field images for each time point, providing representative phenotypic data of most of soybean's developmental stages.

Canopy coverage is a developmental trait

The CC data derived from the UAS deployed in Nanchang showed variation across accessions starting at the fourth time point (24 DAS) (Figures 1F, S1). We classified CC across the subsequent 13 time points into two internally correlated clusters based on the corresponding Spearman's correlation coefficient matrix: a first cluster from 24 to 48 DAS and a second cluster from 61 to 72 DAS (Figure S2A). Notably, 48 DAS was the day when 94% of all soybean accessions had flowered. The CC distribution itself also differed between time points, switching from a long-tailed distribution (28–34 DAS) to a more uniform distribution (37–40 DAS), and finally to a left-skewed distribution (48–72 DAS) (Figure S1).

In parallel to UAS-acquired data, we also manually phenotyped five developmental traits: PH, FT, maturity time (MT), branch number (BN), and grain yield per plant (GYPP). Importantly, the CC data derived from the last 13 time points showed dynamic and significant (Spearman's one-tailed test, Bonferroni adjustment with $\alpha = 0.01$) correlations with these five manually collected classical traits, with the exception of CC and MT at 44 DAS (Figure 1G). Canopy coverage displayed the highest correlation with PH in the first CC cluster (24–48 DAS) and exhibited the highest correlation with FT in the second cluster (61–72 DAS). There was no significant difference in the correlations between CC with each of five classical traits and pairwise correlations between these five classical traits (Welch's two-sample t-test, P-value 0.76; Figure S2). This result supported the notion that time-series CC is a novel dynamic developmental trait that reflects the diversity of soybean accessions. The accessions with the top 15% GYPP values had a significantly higher maximum CC across all 17 time points compared to the accessions with the lowest 15% GYPP values (one-tailed Welch's two-sample t-test, P-value $< 2.2E{-}16$), indicating that CC is a major yield determinant factor in cereals.

Genetic basis of time-series CC, as revealed by GWAS

To reveal the dynamic regulation of CC, we performed a GWAS with 1,303 soybean accessions using the CC values for each of the later 13 time points (Figure S1), using a high-density set of 4,383,780 SNPs with a minimum minor allele frequency (MAF) of 5%. In total, we identified 35 CC-associated loci (Table S3) with an association cutoff benefiting from the multiple time points, as we required each region be detected in at least three time points (see Methods). Of the 35 loci, 10 (or 28.6%) overlapped with previously reported QTLs for FT or PH (covering 28.6 Mb of

the 1.1-Gb soybean genome (Schmutz et al., 2010)) or with previous GWAS results (comprising 367 leading SNPs) (Table S4). Only one (or 11.1%) out of nine loci associated with one or two time points had been previously reported. The median number of associated loci was 3 when shuffling the genotype but following the same cutoffs for 50 permutations (Figure S3), indicating that the false-positive rate of the 35 loci is below 10%. The proportion of variation explained (PVE) (Zhou and Stephens, 2014) by SNPs for each CC time point ($N = 13$) had a median value of 0.40, while the median PVE of permutated CC traits ($N = 13 \times 50 = 650$) was close to zero (6×10^{-6}).

The 35 loci showed a dynamic association with CC across the later 13 time points from 28 to 72 DAS, which represented different developmental stages (Figure 2A, B). These 35 loci formed three groups as a function of the time points at which they were identified: "all stages" ($N = 9$), "earlier stages" (28–40 DAS, $N = 11$), and "later stages" (61–72 DAS, $N = 15$) (Figure 2B). The 44 and 48 DAS time points marked a transition between earlier and later stages.

The 10 previously reported loci (Table S4) overlapped with CC-associated loci at least three times from 44 DAS onward, when 88% of studied soybean accessions flowered. Moreover, eight of these loci belonged to either the "all stages" or "later stages" CC-associated loci. The leading SNPs for these 10 loci had a more significant P-value than the other 25 (one-tailed t-test P-value 0.02; Figure S4). Among these 35 associated loci, we noticed the two known FT genes $E1$ (Xia et al., 2012) and $E2$ (Watanabe et al., 2011) (Figure 2A, B; Table S3). $E2$ was associated with CC in 10 out of 13 time points, and the presumed causal SNP, a common allele causing a premature stop codon (Watanabe et al., 2011), also had the most significant P-value (Figure 2B; Table S3). A comparison of phenotypes of the two $E2$ alleles at this presumed causal SNP indicated that the Williams 82 (W82) reference allele (AA) exhibits a higher CC (one-tailed t-test P-value $<10^{-15}$ for 28–72 DAS) than the alternative allele, which introduces the premature stop codon in $E2$ (Figure 2C).

Population size increases statistical power of GWAS

Increasing population size can improve the statistical power of GWASs (Visscher et al., 2012). Handling populations of hundreds or thousands of individuals is ideally suited to exploit the advantages of high-throughput phenotyping platforms. Here, we investigated the effect of sample size on our UAS-based time-series CC. To this end, we randomly selected subsets ranging from 10% to 90% of all 1,303 soybean accessions to repeat GWASs on a smaller sample size. The fraction of the 35 loci identified in these subsampled analyses increased linearly with population size (Figure S5A). By contrast, while the proportion of the 10 previously reported loci for FT and PH also increased with population size, they reached their maximum detection with a population comprising 60% of all accessions (Figure S5B). Importantly, we identified $E2$, which was the most significant candidate in time-series CC GWASs, in most (8 of 10) subsets consisting
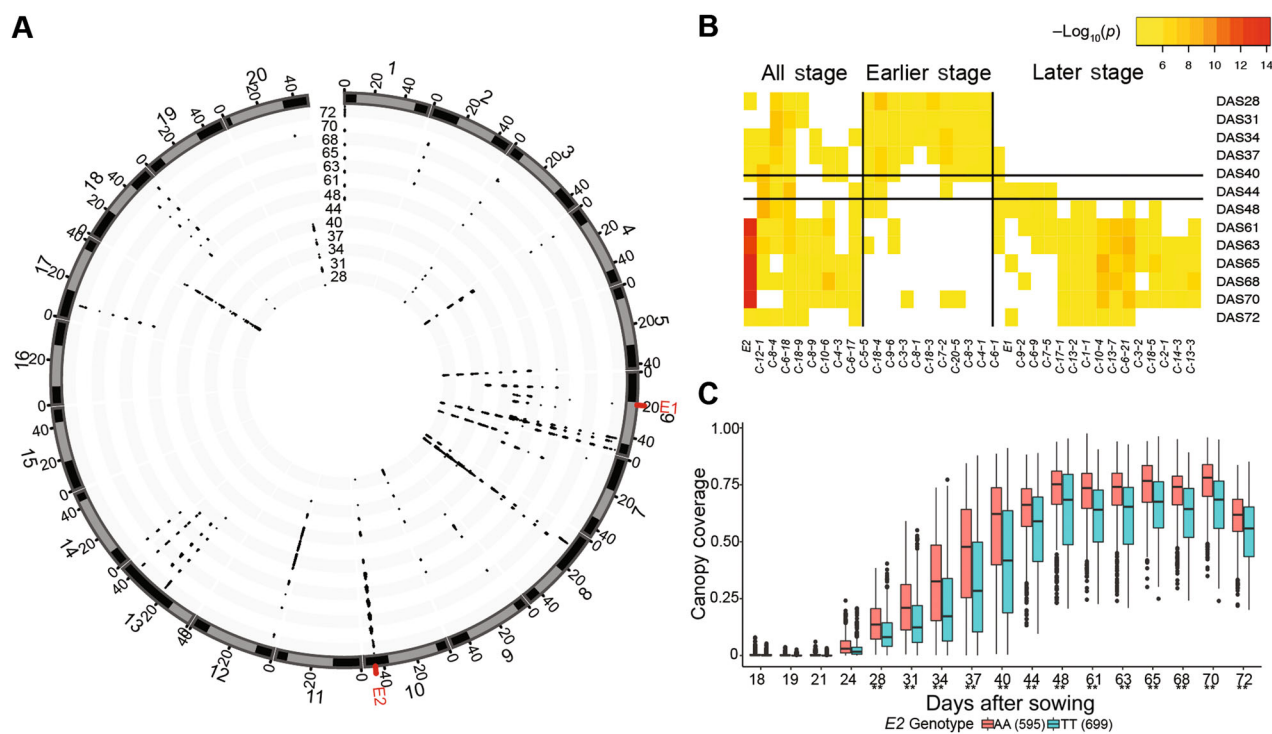
Figure 2. Dynamic association of canopy coverage (CC)
(A) Repeated identification of single nucleotide polymorphisms (SNPs) significantly associated with CC across 13 time points from 28 to 72 d after sowing (DAS). Only SNPs with a $-\log_{10}(p) >4$ and present in the 35 loci identified in at least three time points are plotted. The two known genes identified in this analysis (*E1* and *E2*) are shown in red. **(B)** Change in the $-\log_{10}(p)$ across the 13 time points for loci associated with CC in at least three time points. Trait-associated loci were categorized as: "all stages," "earlier stages," and "later stages" based on whether they were identified in earlier time points (28–40 DAS) or later time points (61–72 DAS); 44–48 DAS were transition time points. **(C)** Distributions of CC values for soybean accessions carrying different genotypes of the presumed causal SNP at the *E2* gene. $**P \leq 0.001$, as determined by one-tailed *t*-test.

of 20% of all accessions and in all cases when subpopulation size was larger than 20% of the total. These results indicated that UAS-based phenotyping of thousands of accessions can improve the statistical power of GWASs, allowing the identification of novel loci associated with a given trait.

Canopy coverage PCA and GWASs of the first two principal components

The time-series traits paved the way for new opportunities to reveal more trait-associated loci but also presented challenges for interpreting this complex trait data set. To reduce the dimensionality of the time-series data, we thus conducted a principal component analysis (PCA) using the CC trait data for all 17 time points and all 1,303 accessions. The first two principal components (PCs) explained 91.5% and 6.1% of the standing variation of the data (Figure 3A) and showed large variance between accessions (Figure S6). PC1 illustrated average CC values, while PC2 reflected the speed of canopy closure (Figures 3B, C, S7). We then performed GWASs using the first two PCs as traits, and 19 and 12 of the 35 loci were determined to be associated with PC1 or PC2, respectively. In total, GWASs of the first two PCs identified 28 (or 80%) of the 35 CC-associated loci (Table S3). Furthermore, we identified three loci, which were part of the "all stages" group, as being associated with both two PCs. In

addition, loci only associated with PC1 belonged to either the "all stages" ($N = 6$) or "later stages" ($N = 11$) group, while PC2-specific loci were part of the "earlier stages" ($N = 8$) group, during which the canopy closed (Figure S7). For example, the known FT gene *E2*, which was among the "later stage" CC-associated loci, showed an association with PC1 but not PC2 (Figure 3D, E).

The PCA not only captured the majority of polymorphisms and the genetic basis for the time-series CC, but importantly, it also quantified the growth curve for each accession with PC2. Moreover, the soybean accessions with the lowest 15% PC2 values belonged to all genetically categorized (Li et al., 2022) sub-populations: China Northeastern region (CNR; $N = 14$), China Huang-huai region (CHR; $N = 21$), China Southern region (CSR; $N = 64$), and America region (Am; $N = 48$). This observation indicated no strong bias caused by population structure, even though soybean is sensitive to photoperiod (Watanabe et al., 2012). Notably, we detected no accession from the CNR subpopulation among accessions with the top 15% PC2 values, likely reflecting their low BN (Figure S8).

Confirmation with time-series CC from an independent environment

We collected the time-series CC phenotypic data as a single replicate in Nanchang (28°31′56″N, 116°1′34″E) over 1 year,
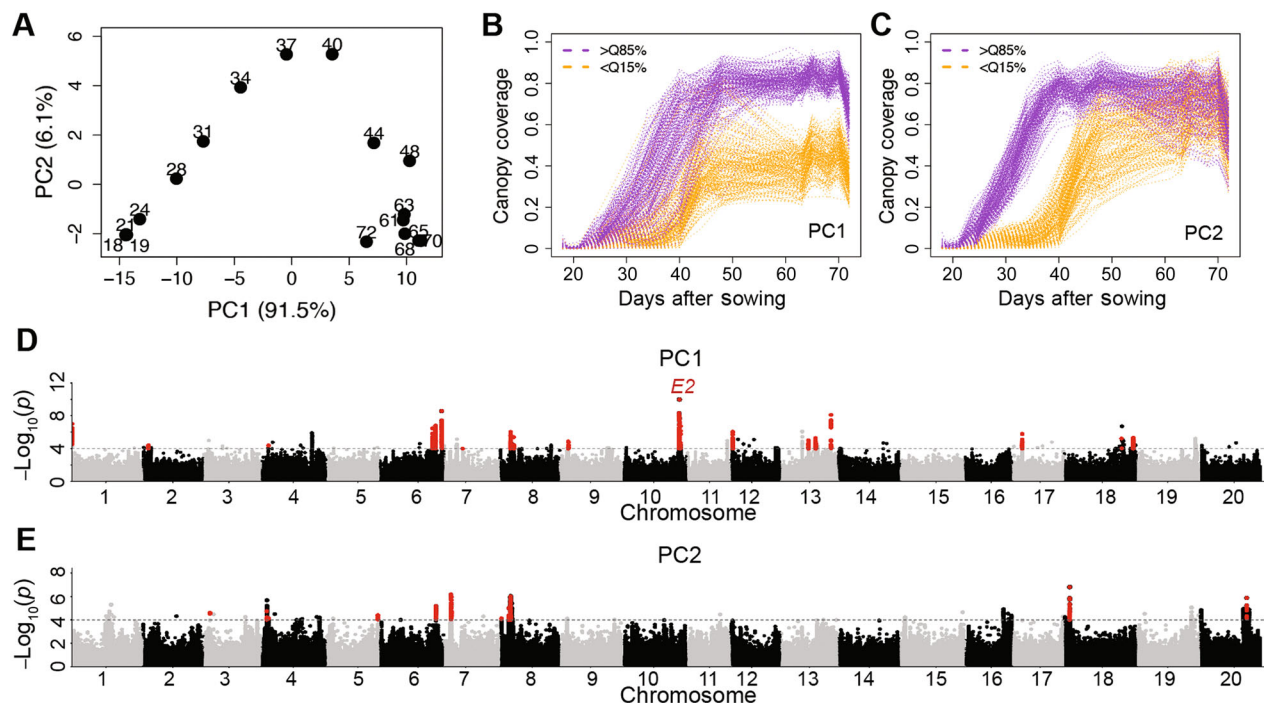
Figure 3. The first two principal components capture most of the phenotypic variation and associated loci

(A) Time-series canopy coverage (CC) dot plots of the first two principal components (PCs). Each time point is labeled next to the corresponding circle (in d after sowing). **(B, C)** Different patterns of CC values over time for soybean accessions in the top 15% (purple) or bottom 15% (orange) values for PC1 **(B)** and PC2 **(C)**. Q, quantile. **(D, E)** Genome-wide association studies (GWAS) results using PC1 **(D)** or PC2 **(E)** obtained from the principal component analysis (PCA) of CC. single nucleotide polymorphisms (SNPs) with $-\log_{10}(p) > 4$ that overlap with the 35 CC-associated loci identified at each time point are highlighted in red. The flowering time-related gene *E2* is also labeled in red.

while data from at least 2 years are typically needed for genetic analyses of crop traits. To confirm the phenotypes and association results obtained above, we planted 397 genotyped soybean cultivars (Li et al., 2022) as two to three replicates in another location, Sanya (18°23′44″N, 109°9′42″E) (Table S5). A sum of 350 of the 397 cultivars belonged to the Nanchang population (Figure S9). We analyzed time-series CC data from 21 time points, from 21 to 87 DAS (Figure S10A, B). The first two PCs also explained most of the standing variation (90.9%) and represented the average CC (PC1) and speed of canopy closure (PC2) (Figure S10C–E). We observed a significant correlation between the two sets of PC1 values from the overlapped 350 cultivars collected in Nanchang and Sanya (Spearman's correlation coefficient 0.33, *P*-value 8.16×10^{-8}). Moreover, soybean accessions with the top 15% PC1 values based on Sanya time-series CC data had a significantly higher average (one-tailed *t*-test *P*-value 2.64×10^{-5}) than the accessions with the bottom 15% PC1 values in Nanchang (Figure 4A). The soybean accessions with the top 15% values for PC2 from Sanya time-series CC data achieved maximum CC earlier (one-tailed *t*-test *P*-value 3.43×10^{-5}) than the accessions with the bottom 15% values for PC2 from Nanchang (Figure 4B).

We performed GWASs with 4,189,867 SNPs (MAF >5%) using the time-series CC values for the 397 genotyped soybean accessions for each of the 21 time points. The PVE (Zhou and Stephens, 2014) by SNPs for each time point had

a median value of 0.45 in Sanya, which was higher (one-tailed *t*-test *P*-value 0.01) than that in Nanchang. We identified 16 of 35 (45.7%) Nanchang CC-associated loci in at least one time point of Sanya, with 11 loci (31.4%) identified in at least three time points (Figure 4C). For these 11 loci, five belonged to the "all stages" (55.5%) group, another two to the "earlier stages" (18.2%) group, and the final four to the "later stages" (26.7%) group, as defined based on the Nanchang time-series CC GWAS. Notably, we identified no associated locus before 40 DAS in Sanya, while "earlier stages" were defined as 28–40 DAS in Nanchang.

One novel locus was favored during adaptation in the north of China and in the United States

Of the 35 identified CC-associated loci, we focused on a "later stage" novel locus mapping to chromosome 1 that we named *C-1-1*, as it had the smallest physical interval (280,730–296,725 bp) of all loci. *C-1-1* was significantly associated with CC at seven time points, from 48 to 72 DAS, and with PC1 (Figure 5A). The *C-1-1* locus comprised three annotated genes: *Glyma.01G001900*, *Glyma.01G002000*, and *Glyma.01G002100*. None of these genes contained SNPs with a large effect (Table S6). However, *Glyma.01G002100* had InDels (insertion/deletions) in its exons (Figure 5B). Moreover, *Glyma.01G002100* has a homolog in Arabidopsis (*Arabidopsis thaliana*), *AUXIN RESPONSE FACTOR7* (*ARF7*), encoding an auxin-regulated
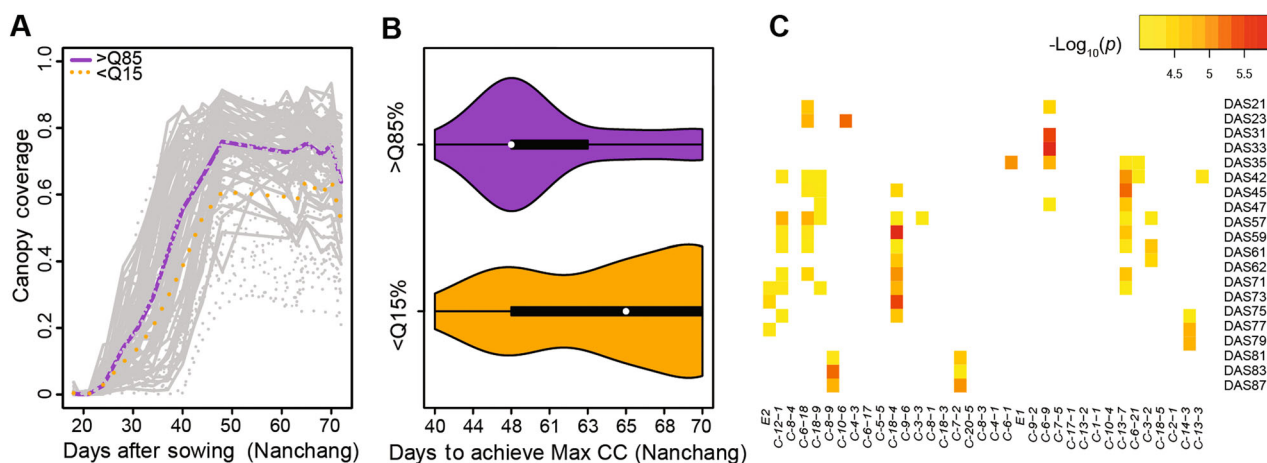
Figure 4. Time series for canopy coverage (CC) in Sanya confirms the principal component analysis (PCA) and genome-wide association studies (GWAS) results of Nanchang

(A) Nanchang time-series CC. The top 15% (>quantile 85%, solid lines) and bottom 15% (<quantile 15%, dashed lines) values are those extracted from PC1 of the CC time series in Sanya. The purple and orange lines represent the average of the top 15% ($N = 59$) and bottom 15% ($N = 39$) soybean lines. **(B)** Distribution of days after sowing to achieve maximum CC in Nanchang. The top 15% (purple, $N = 58$) and bottom 15% (orange, $N = 59$) values are those extracted from PC2 of the CC time series in Sanya. The white dots represent the average. **(C)** The $-\log_{10}(p)$ of the leading single nucleotide polymorphism (SNP) for each locus identified by GWASs for each time point. The loci are in the same order as Figure 2B.

transcriptional activator. The *Arabidopsis arf7 arf19* double mutant produces fewer inflorescence stems, suggesting enhanced apical dominance (Okushima et al., 2005), which could affect CC. We thus hypothesized that *Glyma.01G002100* is a candidate gene for *C-1-1*.

In total, we detected 12 variants within this candidate gene and five haplotypes (H1–H5) with at least five soybean accessions (Figure 5B). Soybean accessions carrying H1 showed significantly lower PC1 values than the other four haplotypes (H2–H5) (one-tailed *t*-test *P*-value 1.2×10^{-10}–3.5×10^{-35}), while five of six pairwise comparisons between H2–H5 showed no difference (one-tailed *t*-test *P*-value 0.06–0.66) (Figure 5C). Hence, H1 was presumed as the function-loss haplotype. H1 differed from the other haplotypes at two SNPs (1-292611 and 1-296596) and one InDel (1-293836), which were all linked (pairwise linkage disequilibrium (LD) $R^2 \geq 0.92$). Each of those three polymorphisms distinguished H1 from the other haplotypes (H2–H5), and genotype AA at the leading SNP (1-292611) could represent H1.

H1 showed a different distribution in three soybean sub-populations located in China (Figure 5D): 55.4% in the CNR subpopulation, none in CHR, and 0.4% in CSR. By contrast, the frequency of genotype AA (equal with H1) at the leading SNP (1-292611) in wild soybean (*Glycine soja*) ($N = 218$) reached 10% and showed no such geographical distribution bias among *G. soja* accessions with geographic information (Table S7). Furthermore, the percentage of H1 in CNR improved cultivars ($N = 58$) was 43.1% ($N = 25$). The frequency of the H1 haplotype was 46.3% ($N = 50$) among improved cultivars of the Am subpopulation and increased in the north of the United States, especially around the Great Lakes region (Figure 5D). Together, these results suggest a geographical adaptation for the *C-1-1* locus during domestication from *G. soja* to *G. max*.

The frequency of the H1 haplotype was 84.0% among CNR landraces ($N = 25$) but only 43.1% in the CNR improved cultivars ($N = 58$), however, both sample sizes here were small. We thus considered the genotypes at the leading SNP in another independent study (Liu et al., 2020). The leading SNP appeared to have the same geographical distribution among 1,934 *G. max* cultivars, in that genotype AA (equal with H1) was favored in the north of China. We also identified 654 genotyped soybean cultivars derived from two northeast provinces, Heilongjiang and Jilin, with the top two frequencies for AA (Figure S11A). The frequency of genotype AA (equal with H1) was 42.4% ($N = 53$) in this other set of landraces and 43.5% ($N = 221$) in the other improved cultivars (Figure S11B), suggesting no difference in the frequency of haplotype H1 in the two groups.

To avoid artifacts from population structure, we compared the time-series CC data between H1 ($N = 25$) and the other major haplotype H4 ($N = 32$, 55%) in CNR improved cultivars. We determined that H1 is associated with a significantly lower CC (one-tailed *t*-test *P*-value ≤ 0.05) at 65–72 DAS (Figure 5E). There were no significant differences (*t*-test *P*-value >0.05) between H1 and H4 improved cultivars for the four classical traits (FT, GYYP, MT, and PH) measured in Nanchang (Figure S12; Table S8), Heilongjiang, or Jilin ($\geq$2-year data sets) (Figure S13; Table S9). We observed that 128 improved cultivars from the independent study (Liu et al., 2020), which were developed in Heilongjiang or Jilin and with at least 2 years of regional yield trials from 1993 to 2004 (Table S10), also showed no significant difference between genotype AA at leading SNP (equal with H1, $N = 48$) and genotype GG (H2–H5, $N = 80$) for PH, MT, or yield. However, we detected a significant difference for BN (*t*-test *P*-value 1.35×10^{-3}) with genotype AA accessions having a lower BN than genotype GG accessions (Figure S14).
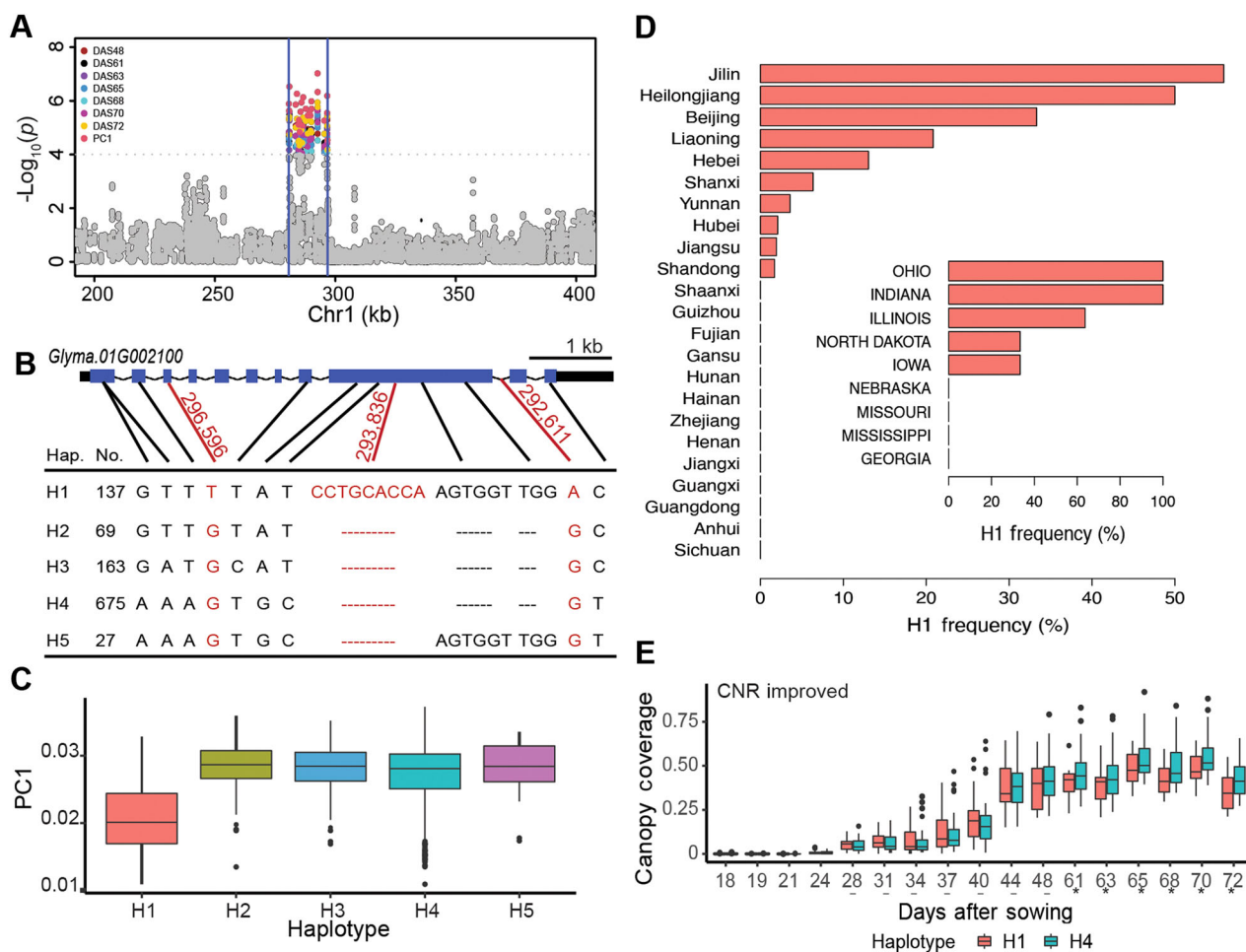
Figure 5. Different alleles in a novel canopy coverage (CC) associated locus are favored in certain geographic regions
(A) Local Manhattan plot of genome-wide association studies (GWAS) results for the *C-1-1* locus (indicated between the two blue vertical lines). All results from GWASs using CC from each time point and PC1 are plotted. (B) Schematic diagram of the candidate gene *Glyma.01G002100* and the positions of single nucleotide polymorphisms (SNPs) detected in this gene. Haplotypes observed in at least five soybean accessions are listed. The three variants (chromosome position) highlighted in red distinguish the H1 haplotype from the others. (C) Distribution of PC1 values for each of the five haplotypes. H1 is significantly different from the other haplotypes (one-tailed *t*-test *P*-value 1.2×10^{-10} to 3.4×10^{-35}), but H2–H5 are not significantly different from each other. (D) Proportion of accessions from each province (China) or state (United States) carrying the H1 haplotype at *C-1-1*. Chinese provinces with fewer than five soybean accessions and states from the United States with fewer than three soybean accessions are not shown. (E) Distribution of CC values in lines carrying the H1 or H4 haplotype among improved cultivars from the subpopulation CNR. –, not significant; *$P \leq 0.05$, as determined by one-tailed *t*-test.

C-8-9 is associated with CC and FT and was favored by low-latitude regions during domestication

In parallel to GWASs of the time-course CC data set, we also conducted GWASs for the five classical manually collected traits (Table S8), which were the average values from 2017 and 2018 in Nanchang. GWASs identified 13 loci for at least one trait, including the two known FT genes *E1* and *E2*, out of the 35 CC-associated loci identified above (Figures 6A, S15). In addition to *E1* and *E2*, we also identified another two known genes, the FT gene *PSEUDORESPONSE REGULATOR 3b* (*PRR3b*) (Li et al., 2020a) (leading SNP 12-5520945, *P*-value 1.5×10^{-16}) and the PH gene *Determinate stem 1* (*Dt1*) (Liu et al., 2010) (leading SNP 19-45143999, *P*-value 2.4×10^{-19}) (Figure S15).

We identified a locus designated here as *C-8-9* as being FT-associated (Figure 6A), CC-associated in "all stages" for seven out of the 13 time points (Figure 2B), and PC1-associated. The *C-8-9* locus encompasses 77.6 kb on chromosome 8 (1,125,492–1,203,080 bp); we further narrowed the locus to a 22.8-kb LD block (11,601,216–11,623,966 bp), which overlapped with association results for FT and CC (Figure 6B, C). Even though only the time points 34, 37, and 63 DAS had SNPs with *P*-value $<1 \times 10^{-4}$ (Figure 6D), the SNPs of this narrowed interval showed significantly lower *P*-values than the genome-wide average at all 13 time points (one-tailed *t*-test *P*-value $6.6E \times 10^{-14}$ to 3.2×10^{-64}) (Figure S16). This smaller interval contained four annotated genes, of which *Glyma.08G150600* with the leading SNP (8-11602352)
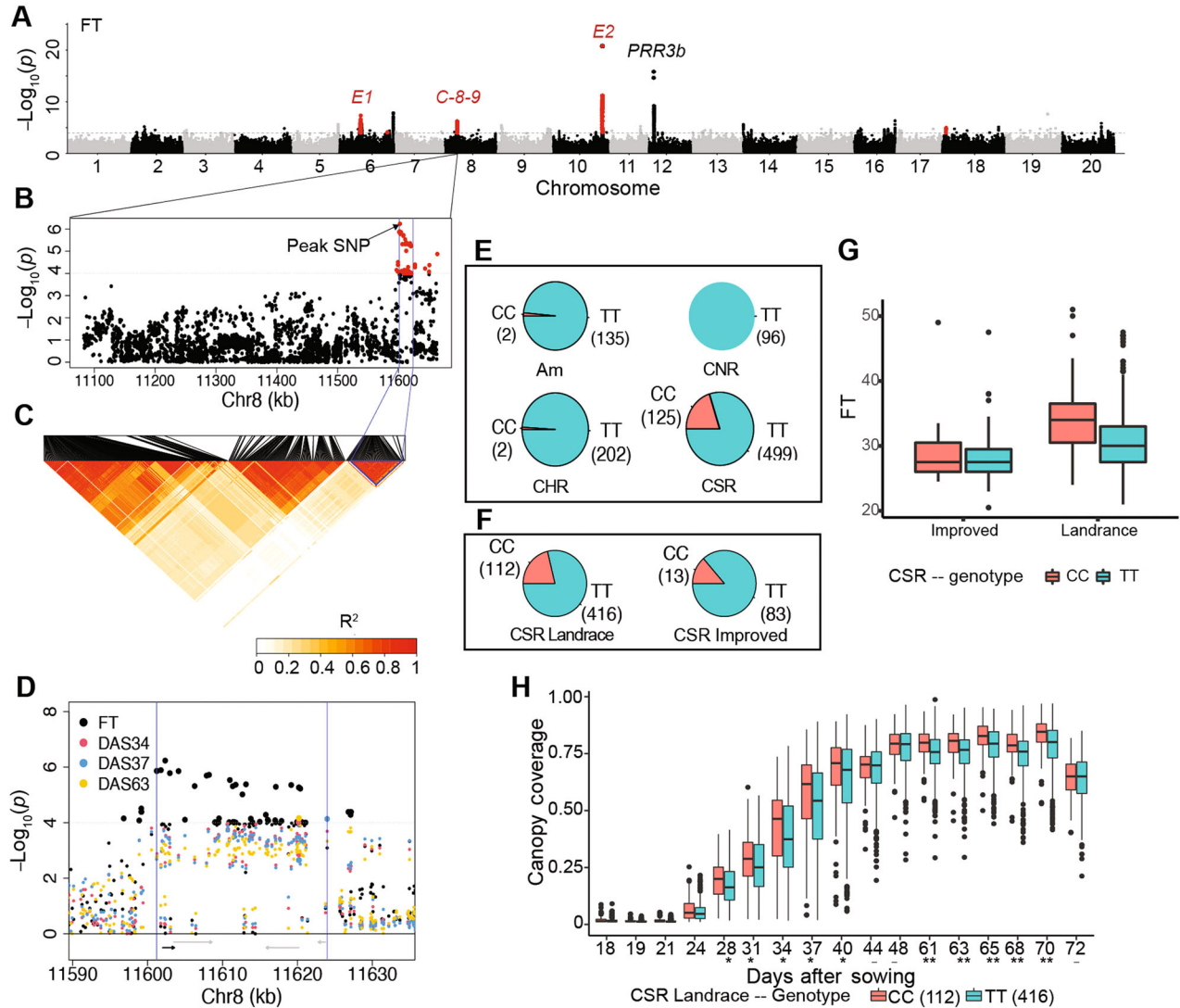
Figure 6. A pleiotropic locus associated with both flowering time and canopy coverage (CC)

(A) Manhattan plot of genome-wide association studies (GWAS) for flowering time (FT). Single nucleotide polymorphisms (SNPs) with a $-\log_{10}(p) > 4$ and also associated with CC in at least three time points are highlighted in red. The known *PRR3b* locus is labeled in black. **(B)** Local Manhattan plot near the *C-8-9* locus. The strongest FT-associated SNP is "8-11602352" (labeled as peak SNP). Most of the FT-associated SNPs (red), located between the two blue vertical lines, are present within an linkage disequilibrium (LD) block **(C)** (highlighted with the blue triangle). **(C)** LD heatmap of the *C-8-9* locus. **(D)** SNPs within the LD block between the two blue vertical lines (Chr8: 11,601,216–11,623,966 bp) are also associated with CC at three time points and with PC1. The four black or gray arrows below the local Manhattan plot mark the positions of annotated genes, with the arrow indicating the strand direction of transcription. The candidate gene harboring the peak SNP is shown in black. **(E)** Distribution of genotypes for the peak SNP (8-11602352) in each of the four soybean sub-populations. Three heterozygous accessions at the peak SNP were excluded from subpopulation China Southern region (CSR) in **(E, F)**. **(F)** Distribution of genotypes for the peak SNP (8-11602352) in landrace and improved lines in subpopulation CSR. **(G)** Distribution of FT values for the two possible genotypes in improved and soybean landraces. **(H)** Distribution of CC values across 17 time points for accessions carrying the CC or TT genotype. –, not significant; *$P \leq 0.05$; **$P \leq 0.001$, as determined by one-tailed *t*-test.

is homologous to Arabidopsis *GLYCINE-RICH PROTEIN 2B* (*GRP2B*), which was hypothesized to be involved in the transition to flowering (Nakaminami et al., 2009). Therefore, we speculate that *Glyma.08G150600* is the mostly likely candidate for this locus.

Examining the genotype of the leading SNP (T/C), we established that the frequency of the alternative allele genotype CC is 20.2% in the CSR subpopulation and <1.5% in other soybean sub-populations (Figure 6E). There was no

difference in allele frequency between CSR landraces and improved cultivars at this SNP (Figure 6F). However, the frequency of genotype CC was 56.6% in *G. soja* (*N* = 218) and was present across all of China, with a minimum allele frequency of 37.5% at the province level (Table S11). This observation was consistent with strong selection during soybean domestication, although whether the observed differences in allele frequency might have arisen via genetic drift cannot be ruled out. The frequency of the genotype CC in

landraces and improved cultivars was 21.4% and 13.5%, respectively, in the CSR subpopulation (Figure 6F). There was no difference between FT or other traits for CSR improved cultivars with different genotypes, in either trait data collected in 2020 or for the average of 2017 and 2018, but we detected a significant difference for FT, MT, and GYPP among CSR landraces (Figures 6G, S17). We also observed differences in CC between CSR landraces with different genotypes in 10 out of 13 time points (Figure 6H). This observation suggested that the C allele is a landrace-beneficial allele in the south of China, which can extend FT and MT with a large CC to obtain high yields (Figure S17).

DISCUSSION

In this study, we deployed UASs for high-throughput phenotyping to collect CC data from a collection of 1,303 soybean accessions across 17 time points and dissected the dynamic regulation of CC. We identified 35 associated loci, which only partially overlapped with known developmental genes and previously reported QTLs. In addition, we confirmed a large fraction (45.7%) of these loci from a time-series CC analysis collected in another environment. Compared to a previous soybean CC study (Xavier et al., 2017), the time-series CC data obtained here were more diverse and revealed another dimension of CC—speed of canopy closure —since we used a population representing the 23,587 cultivated soybeans from the Chinese National Soybean Gene Bank, rather than RILs of the SoyNAM population. We also used 100 times more SNPs for GWASs, which increased our statistical power to detect loci associated with soybean CC.

High-throughput phenotyping, together with high-density genotyping, improves accuracy, statistical power, and resolution of genetic studies

The high-throughput nature of UAS-powered data collection makes it feasible to generate time-series data from large and diverse populations. However, the best approaches with which to leverage time-series data and improve our understanding of underlying genetic mechanisms remain undetermined. One straightforward approach, used here, is to perform GWASs separately for each time point before looking for recurring associations across time points. When we used independent time-point analyses, we consistently identified 35 CC-associated loci in at least three of the 13 time points. Despite the high P-value threshold used, permutation-based testing suggested that fewer than 10% of these 35 CC-associated loci represented false positives. Several (10/35) of these loci were also previously reported to be associated with either PH and/or FT. The multiple-time-points approach used here by time-series GWASs provided a true positive association result of high significance.

High-throughput phenotyping can also improve statistical power by making larger field studies more easily achievable and collecting phenotypic data from larger populations.

These larger populations provide greater power to identify genetic variants linked to phenotypic variation (Park et al., 2010; Visscher et al., 2012; Korte and Farlow, 2013), particularly in cases when the minor allele at a particular locus is rare within the population (Anderson et al., 2018). Indeed, we observed that more loci could be identified as population size increased for the time-series CC data set and that the additional loci detected as the population expanded were particularly likely to be novel and not previously linked to PH or FT (Figures S4, S5). In addition, a large population size and a high density of markers increase mapping resolution (Hamblin et al., 2011), a feature of particular importance in soybean, as LD decays slowly in this predominantly self-pollinated species (Zhou et al., 2015; Li et al., 2020b). In this study, we identified four loci overlapping with known causal genes. In two such cases, E2 (Watanabe et al., 2011) and PRR3b (Li et al., 2020a), the single most significant SNP was present within the known causal gene itself.

Time-series data reveal loci uniquely involved in phenotype-missing stages

In plants, most genetic studies tend to focus on phenotyping data collected from either seedlings grown in the greenhouse or adult plants grown in the field. Comparatively fewer data sets are collected between the seedling stage and maturity, in part because of the difficulty of scoring entire populations at equivalent stages of development using manual approaches. We classified the 35 loci identified here into those observed in early-stage time points, late-stage time points, and across the entire growth period. Of the 10 cases where we identified a locus that overlapped with a known gene or QTL, eight were shared across either all stages or in later stages, suggesting that the genes involved in controlling variation in early-stage canopy closure have been comparatively under-investigated relative to those observables in adult plants.

Geographically favored selection during soybean adaptation

Soybean cultivars are sensitive to photoperiod, with flowering being induced by short days (Watanabe et al., 2012). The soybean varieties used for this study were collected from a wide range of geographical locales, with the hope that some of the genetic and phenotypic variation present in the population would reflect adaptation to local climates. C-1-1 and C-8-9 both showed distributions of genotypes/haplotypes consistent with adaptation to different latitudes during domestication and/or crop improvement. For the C-1-1 locus, H1 was the major haplotype in high-latitude soybean cultivars, while it was the minor haplotype in G. soja (wild soybean) and was almost absent in the low-latitude soybean sub-populations CHR and CSR. These improved cultivars present at high latitudes with H1 showed no difference for PH, FT, MT, or yield-related traits compared to other haplotypes from low latitudes and high latitudes, but showed a lower BN at high latitudes where they originated. The candidate gene identified for C-1-1, Glyma.01 G002100, is

homologous to an *Arabidopsis* gene involved in establishing apical dominance (Okushima et al., 2005). We hypothesize that soybean accessions harboring H1 with lower CC values are suited for higher density planting, which is important for the agronomic systems used at high latitudes where the growing season is shorter, but not well suited for the lower planting densities used at lower latitudes with longer growing seasons. The H1 haplotype had a similar frequency (42.9%–46.3%) in the improved cultivars for the United States and the North of China and a much higher frequency (>95%) in the Great Lakes region of the United States. This observation may indicate that H1 was either a target of selection in breeding programs in the north of China and in the United States or that there was a genetic bottleneck leading to a similar drift in these breeding programs.

Another locus associated with CC and FT, *C-8-9*, appeared to have experienced a dramatic shift in allele frequency between *G. max* and *G. soja*. Indeed, 56.6% of *G. soja* accessions carried the C allele. However, in domesticated soybean cultivars, the frequency of this allele was less than 1.5% in the CNR, CHR, and Am sub-populations. In the CSR subpopulation, the frequency of the C allele was 21.4% in landraces and 13.5% in improved cultivars. This distribution of allele frequencies is consistent with the C allele being strongly selected against in the CNR, CHR, and Am sub-populations, either during or after the initial domestication of soybean. In our study, the C allele was associated with increases in FT and MT, as well as with increased yield in the landraces from the CSR subpopulation, and this allele appears to be beneficial in low-latitude regions. Alleles of these two novel loci (*C-1-1* and *C-8-9*) could be used as genetic markers in breeding programs.

Dynamic CC selection could benefit breeding programs

In this study, we showed that CC was significantly correlated with five manually scored traits collected from the same field study of 1,303 soybean accessions (Figure 1G). Canopy coverage may therefore be used as an indicator to select or exclude extreme accessions. For example, the soybean accessions with lowest GYPP values always had extremely low CC and can be easily excluded much earlier than at the seed setting stage. However, a more important use of time-series CC data may be to enable selection and breeding for growth velocity, especially in early growth stages, for example, canopy closure. This trait is currently not commonly scored by breeding programs due to logistical constraints. However, early canopy closure can reduce weed pressure throughout the growing season, thus enabling farmers to reduce labor and herbicide costs. Our population of 1,303 soybean accessions showed diverse time-series CC values that were reflected in a PCA, whose first two principal components represented the average CC and speed of canopy closure, respectively. Soybean accessions that reach high canopy closure in the early stages (Figure 3C) would provide targets associated with both plant vigor and profitability and

sustainability outcomes of reduced weed pressure desirable to farmers.

Unmanned Aircraft System-based phenotyping will likely not be adopted by breeding programs unless the time required and the data acquisition and processing pipelines are sufficiently streamlined. In our study, it was possible to phenotype the entire set of 1,303 soybean accessions in a field experiment in a single flight at an altitude of 12 m in only approximately 20 min of UAS time (Table S2). The flight program may be extended to obtain PH and maturity at the same time. By contrast, CC is typically expensive to measure manually; phenotyping a large set of plants for either height or maturity could take at least one workday for manual collection. The development of software and algorithms also continues to shorten the labor time for image analysis, which is the most time-consuming step for UAS-based phenotyping. We observed a significantly improved PVE (Zhou and Stephens, 2014) by performing GWASs on the time-series CC data collected in Sanya compared to Nanchang, likely resulting from a higher resolution RGB sensor used for this second set of phenotypes and the addition of replicates for each soybean accession. We anticipate that further advances in sensors, image-processing algorithms, and experimental design will continue to increase the quality and utility of UAS for both plant genetics and plant breeding investigations. Improved UAS platforms are rapidly coming online that enable the acquisition of multispectral data and thermal infrared imaging, in addition to conventional RGB photography (Liu et al., 2021). These advances will further accelerate the availability of time-series and multi-location time-series data sets and increase the importance of identifying and using the most effective approaches for incorporating time-series data into quantitative genetics studies and plant breeding programs.

MATERIALS AND METHODS

Plants and growth conditions

Li et al. (2022) genotyped 2,214 soybean (*Glycine max* (L.) Merr.) accessions, which represent the broad genetic diversity of 23,587 cultivated soybeans from the Chinese National Soybean Gene Bank, via genome resequencing. Of these, 1,303 soybean cultivars were planted and well-grown at Nanchang (28°31′56″N, 116°1′34″E) and consisted of 903 landrace soybeans and 400 improved cultivars. Most of these landraces ($N = 897$) originated from China, while the improved cultivars originated from 11 countries (Table S1). The soybean accessions have been categorized into four geographic regions: CNR ($N = 96$), CHR ($N = 204$), CSR ($N = 627$), and Am ($N = 137$), while the remaining 239 could not be categorized (Li et al., 2022). All 1,303 accessions were sown in Nanchang on July 15, 2020. Each soybean accession was hand-planted as a 1.8-m × 0.8-m plot with two rows, with 10 cm between seedlings for each row, that is, 19 individuals per row. Another 397 soybean cultivars out of the genotyped 2,214 soybean accessions (Li et al., 2022) were sown on

December 3, 2020 in Sanya (18°23′44″N, 109°9′42″E), of which 350 were included in the 1,303-accession set planted in Nanchang. Each soybean accession was hand-planted with three independent replicates (plots), with each replicate being a 1-m × 0.65-m plot with two rows, and 10 cm between seedlings for each row, that is, 11 individuals for each row.

Phenotyping data sets

Two types of soybean traits were collected: (i) manually collected traits and (ii) UAS-derived CC. The classical traits in this study comprised PH, FT, MT, GYPP, and BN. All classical traits were determined manually in 2020 (Table S1), as well as in 2017 and 2018 (Table S8) in Nanchang, in which all accessions were planted in the same field as described for the 1,303-accession set planted in Nanchang. In addition, some improved cultivars from subpopulation CNR were also phenotyped for PH, FT, MT, and GYPP at two locations of Northeast China: Harbin (45°45′ N, 126°41′ E) of Heilongjiang province and Changchun (43°88′ N, 125°35′ E) of Jilin province, in 2018 and 2019 (Table S9). The planting protocol was the same as in Nanchang.

Moreover, there were 128 improved cultivars previously genotyped in an independent study (Liu et al., 2020) and developed from 1993–2004 in the Heilongjiang and Jilin provinces in China. These additional improved cultivars were phenotyped for BN, PH, MT, and yield, which were extracted from 2-year regional trials records (Table S10).

CC, as measured by an UAS, was collected at 17 time points from August 1 to September 24, 2020 in Nanchang for the 1,303 genotyped soybean accessions. Another 397 genotyped soybean cultivars were phenotyped for CC from December 23, 2020 to February 27, 2021 in Sanya. In both field trials, the accessions lacking genotyping data or showing poor germination or survival were not analyzed or described in this study. The acquisition of CC for each plot was described as below.

UAS photograph collection

Two different unmanned aerial vehicles were used in this study: Phantom 4 multispectral drone integrated with an RGB sensor (2.1 million effective pixels) in Nanchang and a DJI Matrice 600 Hexacopter Platform integrated with a Sony ILCE-7M2 RGB sensor (24.3 million effective pixels) in Sanya. Photographs were collected at an altitude of 12 m above the ground in Nanchang and 17 m in Sanya. The flight plans (Table S2) were automated by DJI Ground Station Pro and took place between approximately 10:00 and approximately 14:00 under a clear sky with a trajectory overlap of 75% and side overlap of 60%/75% (Fawcett et al., 2019; Stöcker et al., 2020).

Image preprocessing

Image preprocessing consisted of two stages: (i) generation of orthomosaic maps and (ii) geographical alignment and tailoring. Generating an orthomosaic map comprised three steps: (i) alignment of the original images and mosaicking; (ii) generation of point clouds, mesh, and texture; and (iii) generation of the orthographic map and export. All three steps were executed using Agisoft Photo-Scan Professional software (Version 1.2.2; Agisoft LLC., Russia), which is based on structural motion algorithms (Verhoeven, 2011).

To obtain a unified geographical reference system for all orthomosaic images, georeferencing was carried out in Esri ArcGIS (10.7; ESRI, USA). All other orthomosaic images were georeferenced by the extracted ground control points. Each plot for each soybean line was then obtained by clipping orthomosaic images. Clipping used an in-house ENVI (5.3; Exelis Visual Information Solutions, USA) IDL script. The images for each line were re-exported in tiff format using ArcGIS.

Extraction of CC

MATLAB (R2021a) was used to calculate CC in two steps. The first step was two-value processing. To identify crop targets using a computer-aided vision system, the green plant parts were separated from the soil background. Binary images were obtained by using the excess green index (EGI, (2G-R-B)/G) presented by Meyer and Neto (2008). The EGI threshold was set to 0.05 for canopy and background segmentation. EGI values greater than the green threshold correspond to canopy (binary value = 1), and values below the green threshold correspond to soil (binary value = 0). The percentage of canopy pixels in an image was defined as CC. Canopy coverage for all plots was then calculated. Plots with an average CC below 0.2 at post-FT points were excluded from subsequent analysis, as they may reflect plots with lower germination or survival rates.

Genotyping data sets

A set of 8,785,134 SNPs from resequencing data of 2,214 soybean accessions (Li et al., 2022) based on the Williams 82 (W82) v2 reference genome (Schmutz et al., 2010) were imputed with Beagle (Browning and Browning, 2007; Browning and Browning, 2016). SNPs were then filtered for a minimum minor allele frequency (MAF) of at least 5% among input samples for GWASs. Insertion/deletion (InDel) data from the resequencing of the 2,214 soybean accessions were used for filtering candidate genes without imputation. In addition, 218 wild soybean (G. soja) accessions that are part of the 2,214 soybean accessions were also analyzed in this study.

The genotypes of 2,898 soybean accessions from another study (Liu et al., 2020) based on the Zhonghuang 13 (ZH13) v2 reference genome were used here to supplement the population size. The positions of all investigated W82 SNPs were converted to ZH13 positions using their flanking sequences ±100 bp for BLASTN (Camacho et al., 2009) against the ZH13 v2 reference genome (Shen et al., 2018, 2019).

Principal component analysis

Principal component analysis (Hotelling, 1933) is used for dimensionality reduction by projecting data points into principal components (PCs) while preserving data variation as much as possible (Hotelling, 1933). Here, the PCA was performed using the built-in R (V4.0.2) function "prcomp" to investigate differences between the time-series CC values. Variation associated with each PC was calculated. The

coordinates of all soybean individuals (observations) along the first two PCs, which explained 97.6% of the standing variation, were used as traits for subsequent GWASs.

Genome-wide association studies

Genome-wide association studies were performed with the univariate linear mixed model, which uses the Wald test and centered relatedness matrix implemented in the GEMMA algorithm (Zhou and Stephens, 2012). Canopy coverage-associated loci were identified with multiple steps, as described below, to capitalize on the multiple time points, which we refer to as time-series GWASs. (i) SNPs with a P-value $\leq 1E{-}4$ from CC GWAS for each of the 13 time points (28–72 DAS) in Nanchang were used to identify candidate CC-associated loci. (ii) Significant SNPs were merged into separate SNP clusters; consecutive SNP clusters were merged into a single cluster when linkage disequilibrium (LD) R^2 was ≥ 0.5 or when they were less than 200 kb apart. Each separate associated SNP clustered with at least five unique SNPs were defined as CC candidate loci. (iii) The most significant SNP of each candidate locus was required to have a P-value $\leq 3.16E{-}6$, which was 1/pruned SNPs (as determined by PLINK v1.90 with the settings –indep-pairwise 100 kb 50 0.8) (Gaunt et al., 2007). (iv) The associated loci were required to be detected as associated for at least three independent time points. The retained loci were defined as final CC-associated loci.

Permutation tests were performed to evaluate the accuracy of the time-series CC GWASs. The SNP set for each accession was randomized across all accessions for permutations, but maintaining the same CC values across different time points for each soybean accession to maintain the CC correlation. Fifty independent permutations were conducted, followed by a GWAS to identify associated loci based on the criteria described above. The numbers of associated loci were used as the false positives to evaluate the false discovery rate of the time-series GWAS.

Published FT and PH QTLs and associated SNPs were downloaded from SoyBase (https://www.soybase.org). QTLs with a physical interval below 1 Mb were retained. If a QTL overlapped with or its associated peak SNP was included in one of the CC-associated loci above, the QTL or SNP was considered as being part of the CC-associated locus.

Phylogenetic tree

IQ-Tree (V2) (Minh et al., 2020) was used to infer the phylogenetic tree, for which a general time reversible model with unequal rates and unequal base frequency (Tavaré, 1986) was used. The pruned SNPs ($N = 316,112$) were inputted for construction of the tree. The tree was visualized by iTOL (Letunic and Bork, 2007).

Data availability statement

The resequencing data for the soybean accessions in this project were deposited in the Sequence Read Archive database of NCBI (www.ncbi.nlm.nih.gov) under accession number PRJNA681974 (Li et al., 2022) and in the Genome Sequence Archive (GSA) database of BIG Data Center

(https://ngdc.cncb.ac.cn) under bioProject number PRJCA002030 (Liu et al., 2020). Genotype data resulting from those two projects were directly used.

CONFLICTS OF INTEREST

The authors declare that there is not conflict of interest.

AUTHOR CONTRIBUTIONS

Y.-H.L., X.J., J.C.S., L.-J.Q., and D.L. conceived the research, C.Z., R.W., Y.G., S.G., Q.W., and X.L. manually collected phenotyping data, D.L. and D.B. collected unmanned aerial vehicle photos, X.J. and D.B. performed image analysis, D.L. and Y.T. performed data analysis. J.Y. and M.J.H. contributed ideas and discussions on the analysis. Y.-H.L. and L.-J.Q. supervised the research. D. L., X.J., J.C.S., and M.J. H. wrote the manuscript. All authors read and approved of the final manuscript.

REFERENCES

Adak, A., Murray, S.C., Anderson, S.L., Popescu, S.C., Malambo, L., Romay, M.C., de Leon, N., and de Leon, N. (2021). Unoccupied aerial systems discovered overlooked loci capturing the variation of entire growing period in maize. Plant Genome **14:** e20102.

Ahn, S.N., and Tanksley, S.D. (1993). Comparative linkage maps of the rice and maize genomes. Proc. Natl. Acad. Sci. U.S.A. **90:** 7980–7984.

Anderson, S.L., 2nd, Mahan, A.L., Murray, S.C., and Klein, P.E. (2018). Four parent maize (FPM) population: Effects of mating designs on linkage disequilibrium and mapping quantitative traits. Plant Genome **11:** 170102.

Anderson, S.L., 2nd, Murray, S.C., Chen, Y., Malambo, L., Chang, A., Popescu, S., Cope, D., and Jung, J. (2020). Unoccupied aerial system enabled functional modeling of maize height reveals dynamic expression of loci. Plant Direct **4:** e00223.

Anthony, D., Elbaum, S., Lorenz, A., and Detweiler, C. (2014). On crop height estimation with UAVs. 2014 *IEEE/RSJ International Conference on Intelligent Robots and Systems: IEEE*, 4805–4812. https://doi.org/10.1109/IROS.2014.6943245

Browning, B.L., and Browning, S.R. (2016). Genotype imputation with millions of reference samples. Am. J. Hum. Genet. **98:** 116–126.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. **81:** 1084–1097.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics **10:** 421.

Caturegli, L., Corniglia, M., Gaetani, M., Grossi, N., Magni, S., Migliazzi, M., Angelini, L., Mazzoncini, M., Silvestri, N., Fontanelli, M., Raffaelli, M., Peruzzi, A., and Volterrani, M. (2016). Unmanned aerial vehicle to estimate nitrogen status of turfgrasses. PLoS ONE **11:** e0158268.

Cazenave, A.-B., Shah, K., Trammell, T., Komp, M., Hoffman, J., Motes, C.M., and Monteros, M.J. (2019). High-throughput approaches for phenotyping alfalfa germplasm under abiotic stress in the field. Plant Phenome J. **2:** 190005.

Fan, Z., Lu, J., Gong, M., Xie, H., and Goodman, E.D. (2018). Automatic tobacco plant detection in UAV images via deep neural networks. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **11:** 876–887.

Fawcett, D., Azlan, B., Hill, T.C., Kho, L.K., Bennie, J., and Anderson, K. (2019). Unmanned aerial vehicle (UAV) derived structure-from-motion photogrammetry point clouds for oil palm (*Elaeis guineensis*) canopy segmentation and height estimation. Int. J. Remote Sens. **40:** 7538–7560.

Flint-Garcia, S.A., Thornsberry, J.M., and Buckler, E.S., IV. (2003). Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. **54:** 357–374.

Gaunt, T.R., Rodríguez, S., and Day, I.N.M. (2007). Cubic exact solutions for the estimation of pairwise haplotype frequencies: Implications for linkage disequilibrium analyses and a web tool "CubeX". BMC Bioinformatics **8:** 428.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: Ten years of next-generation sequencing technologies. Nat. Rev. Genet. **17:** 333–351.

Guo, W., Fukatsu, T., and Ninomiya, S. (2015). Automated characterization of flowering dynamics in rice using field-acquired time-series RGB images. Plant Methods **11:** 7.

Hamblin, M.T., Buckler, E.S., and Jannink, J.L. (2011). Population genetics of genomics-based crop improvement methods. Trends Genet. **27:** 98–106.

Hassan, M., Yang, M., Rasheed, A., Jin, X., Xia, X., Xiao, Y., and He, Z. (2018). Time-series multispectral indices from unmanned aerial vehicle imagery reveal senescence rate in bread wheat. Remote Sens. **10:** 809.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24:** 417–441.

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

Jin, X., Zarco-Tejada, P., Schmidhalter, U., Reynolds, M.P., Hawkesford, M.J., Varshney, R.K., Yang, T., Nie, C., Li, Z., Ming, B., Xiao, Y., Xie, Y., and Li, S. (2021). High-throughput estimation of crop traits: A review of ground and aerial phenotyping platforms. IEEE Geosci. Remote Sens. Mag. **9:** 200–231.

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. Plant Methods **9:** 29.

Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., and Linnér, R.K. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet. **50:** 1112–1121.

Letunic, I., and Bork, P. (2007). Interactive tree of life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics **23:** 127–128.

Li, C., Li, Y.H., Li, Y., Lu, H., Hong, H., Tian, Y., Li, H., Zhao, T., Zhou, X., Liu, J., Zhou, X., Jackson, S.A., Liu, B., and Qiu, L.J. (2020a). A domestication-associated gene *GmPRR3b* regulates the circadian clock and flowering time in soybean. Mol. Plant **13:** 745–759.

Li, Y.H, Li, D., Jiao, Y.Q, Schnable, J.C., Li, Y.F, Li, H.H, Chen, H.Z, Hong, H.L, Zhang, T., and Liu, B. (2020b). Identification of loci controlling adaptation in Chinese soya bean landraces via a combination of conventional and bioclimatic GWAS. Plant Biotechnol. J. **18:** 389–401.

Li, Y.-H., Qin, C., Wang, L., Jiao, C., Hong, H., Tian, Y., Li, Y., Xing, G., Wang, J., Gu, Y., Gao, X., Li, D., Li, H., Liu, Z., Jing, X., Feng, B., Zhao, T., Guan, R., Guo, Y., Liu, J., Yan, Z., Zhang, L., Ge, T., Li, X., Wang, X., Qiu, H., Zhang, W., Luan, X., Han, Y., Han, D., Chang, R., Guo, Y., Reif, J.C., Jackson, S.A., Liu, B., Tian, S., and Qiu, L.-j. (2022). Genome-wide signatures of the geographic expansion and breeding of soybean. Sci. China Life Sci. https://doi.org/10.1007/s11427-022-2158-7

Li, Z., and Sillanpaa, M.J. (2015). Dynamic quantitative trait locus analysis of plant phenomic data. Trends Plant Sci. **20:** 822–833.

Liu, B., Watanabe, S., Uchiyama, T., Kong, F., Kanazawa, A., Xia, Z., Nagamatsu, A., Arai, M., Yamada, T., Kitamura, K., Masuta, C., Harada, K., and Abe, J. (2010). The soybean stem growth habit gene *Dt1* Is an ortholog of Arabidopsis *TERMINAL FLOWER1*. Plant Physiol. **153:** 198–210.

Liu, H.J., and Yan, J. (2019). Crop genome-wide association study: A harvest of biological relevance. Plant J. **97:** 8–18.

Liu, S., Jin, X., Nie, C., Wang, S., Yu, X., Cheng, M., Shao, M., Wang, Z., Tuohuti, N., Bai, Y., and Liu, Y. (2021). Estimating leaf area index using unmanned aerial vehicle data: Shallow vs. deep machine learning algorithms. Plant Physiol. **187:** 1551–1576.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., and Tian, Z. (2020). Pan-genome of wild and cultivated soybeans. Cell **182:** 162–176.

Lyra, D.H., Virlet, N., Sadeghi-Tehran, P., Hassall, K.L., Wingen, L.U., Orford, S., Griffiths, S., Hawkesford, M.J., and Slavov, G.T. (2020). Functional QTL mapping and genomic prediction of canopy height in wheat measured using a robotic field phenotyping platform. J. Exp. Bot. **71:** 1885–1898.

Meyer, G.E., and Neto, J.C. (2008). Verification of color vegetation indices for automated crop imaging applications. Comput. Electron. Agric. **63:** 282–293.

Miao, C., Xu, Y., Liu, S., Schnable, P.S., and Schnable, J.C. (2020). Increased power and accuracy of causal locus identification in time series genome-wide association in sorghum. Plant Physiol. **183:** 1898–1909.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. **37:** 1530–1534.

Murakami, T., Yui, M., and Amaha, K. (2012). Canopy height measurement by photogrammetric analysis of aerial images: Application to buckwheat (*Fagopyrum esculentum Moench*) lodging evaluation. Comput. Electron. Agric. **89:** 70–75.

Nakaminami, K., Hill, K., Perry, S.E., Sentoku, N., Long, J.A., and Karlson, D.T. (2009). Arabidopsis cold shock domain proteins: Relationships to floral and silique development. J. Exp. Bot. **60:** 1047–1062.

Okushima, Y., Overvoorde, P.J., Arima, K., Alonso, J.M., Chan, A., Chang, C., Ecker, J.R., Hughes, B., Lui, A., Nguyen, D., Onodera, C., Quach, H., Smith, A., Yu, G., and Theologis, A. (2005). Functional genomic analysis of the *AUXIN RESPONSE FACTOR* gene family

members in *Arabidopsis thaliana*: Unique and overlapping functions of *ARF7* and *ARF19*. Plant Cell **17**: 444–463.

Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet. **42**: 570–575.

Pugh, N.A., Horne, D.W., Murray, S.C., Carvalho, G., Jr., Malambo, L., Jung, J., Chang, A., Maeda, M., Popescu, S., Chu, T., Starek, M.J., Brewer, M.J., Richardson, G., and Rooney, W.L. (2018). Temporal estimates of crop growth in sorghum and maize breeding enabled by unmanned aerial systems. Plant Phenome J. **1**: 170006.

Roosjen, P.P.J., Brede, B., Suomalainen, J.M., Bartholomeus, H.M., Kooistra, L., and Clevers, J.G.P.W. (2018). Improved estimation of leaf area index and leaf chlorophyll content of a potato crop using multi-angle spectral data-potential of unmanned aerial vehicle imagery. Int. J. Appl. Earth Obs. Geoinf. **66**: 14–26.

Sarkar, S., Cazenave, A.-B., Oakes, J., McCall, D., Thomason, W., Abbot, L., and Balota, M. (2020). High-throughput measurement of peanut canopy height using digital surface models. Plant Phenome J. **3**: e20003.

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

Sedivy, E.J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: The origin, genetic architecture and molecular bases. New Phytol. **214**: 539–553.

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.

Shen, Y., Liu, J., Geng, H., Zhang, J., Liu, Y., Zhang, H., Xing, S., Du, J., Ma, S., and Tian, Z. (2018). *De novo* assembly of a Chinese soybean genome. Sci. China Life Sci. **61**: 871–884.

Shi, Y., Thomasson, J.A., Murray, S.C., Pugh, N.A., Rooney, W.L., Shafian, S., Rajan, N., Rouze, G., Morgan, C.L., and Neely, H.L. (2016). Unmanned aerial vehicles for high-throughput phenotyping and agronomic research. PLoS ONE **11**: e0159781.

Stöcker, C., Nex, F., Koeva, M., and Gerke, M. (2020). High-quality UAV-based orthophotos for cadastral mapping: Guidance for optimal flight configurations. Remote Sens. **12**: 3625.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. **17**: 57–86.

Verhoeven, G. (2011). Taking computer vision aloft-archaeological three-dimensional reconstructions from aerial photographs with photoscan. Archaeol. Prospect. **18**: 67–73.

Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. **90**: 7–24.

Wang, H., Cimen, E., Singh, N., and Buckler, E. (2020). Deep learning for plant genomics and crop improvement. Curr. Opin. Plant Biol. **54**: 34–41.

Wang, X., Zhang, R., Song, W., Han, L., Liu, X., Sun, X., Luo, M., Chen, K., Zhang, Y., Yang, H., Yang, G., Zhao, Y., and Zhao, J. (2019). Dynamic plant height QTL revealed in maize through remote sensing phenotyping using a high-throughput unmanned aerial vehicle (UAV). Sci. Rep. **9**: 3458.

Watanabe, S., Harada, K., and Abe, J. (2012). Genetic and molecular bases of photoperiod responses of flowering in soybean. Breed. Sci. **61**: 531–543.

Watanabe, S., Xia, Z., Hideshima, R., Tsubokura, Y., Sato, S., Yamanaka, N., Takahashi, R., Anai, T., Tabata, S., Kitamura, K., and Harada, K. (2011). A map-based cloning strategy employing a residual heterozygous line reveals that the *GIGANTEA* gene is involved in soybean maturity and flowering. Genetics **188**: 395–407.

Watt, M., Fiorani, F., Usadel, B., Rascher, U., Muller, O., and Schurr, U. (2020). Phenotyping: New windows into the plant for breeders. Annu. Rev. Plant Biol. **71**: 689–712.

Wu, X., Feng, H., Wu, D., Yan, S., Zhang, P., Wang, W., Zhang, J., Ye, J., Dai, G., Fan, Y., Li, W., Song, B., Geng, Z., Yang, W., Chen, G., Qin, F., Terzaghi, W., Stitzer, M., Li, L., Xiong, L., Yan, J., Buckler, E., Yang, W., and Dai, M. (2021). Using high-throughput multiple optical phenotyping to decipher the genetic architecture of maize drought tolerance. Genome Biol. **22**: 185.

Xavier, A., Hall, B., Hearst, A.A., Cherkauer, K.A., and Rainey, K.M. (2017). Genetic architecture of phenomic-enabled canopy coverage in *Glycine max*. Genetics **206**: 1081–1089.

Xia, Z., Watanabe, S., Yamada, T., Tsubokura, Y., Nakashima, H., Zhai, H., Anai, T., Sato, S., Yamazaki, T., Lu, S., Wu, H., Tabata, S., and Harada, K. (2012). Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering. Proc. Natl. Acad. Sci. U.S.A. **109**: E2155–E2164.

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J.H., Batchelor, W.D., Xiong, L., and Yan, J. (2020). Crop phenomics and high-throughput phenotyping: Past decades, current challenges, and future perspectives. Mol. Plant **13**: 187–214.

Zhou, J., Yungbluth, D., Vong, C.N., Scaboo, A., and Zhou, J. (2019). Estimation of the maturity date of soybean breeding lines using UAV-based multispectral imagery. Remote Sens. **11**: 2075.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. **44**: 821–824.

Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods **11**: 407–409.

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

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article: http://onlinelibrary.wiley.com/doi/10.1111/jipb.13380/suppinfo

Scan using WeChat with your smartphone to view JIPB online

Scan with iPhone or iPad to view JIPB online