Assessment of phenotypic, genomic and novel approaches for soybean breeding in Central Europe

Dissertation to obtain the doctoral degree of Agricultural Sciences
(Dr. sc. agr.)

Faculty of Agricultural Sciences
University of Hohenheim

Institute of Plant Breeding, Seed Science and Population Genetics

submitted by
Xintian Zhu

from *Henan*

2022

This thesis was accepted as a doctoral dissertation in fulfilment of the regulations for the doctoral degree “Doktor der Agrarwissenschaften” (Dr. sc. Agr. / PhD. in Agricultural Sciences) by the Faculty of Agricultural Sciences at the University of Hohenheim on 13th July, 2022.

Date of the oral examination: 14.12.2022

Examination Committee:

Chairperson of the oral examination:	Prof. Dr. Uwe Ludewig
Supervisor and Reviewer:	Prof. Dr. Tobias Würschum
Co-Reviewer:	Prof. Dr. Frank Ordon
Additional examiner:	apl. Prof. Dr. Thomas Miedaner

Contents

General Introduction

Soybean [*Glycine max* (L.) Merr.] is the economically most important leguminous crop worldwide. Its seeds are rich in both protein and oil and are used for human consumption, livestock feed, and industrial purposes. In Europe, soybean is mainly considered as a protein crop, owing to its high protein content of approximately 40%. In addition, as a legume, soybean can play an important role in crop rotation due to its ability of nitrogen fixation through the symbiosis with the rhizobacterium *Bradyrhizobium japonicum*.

Soybean originates from Asia, where it was first domesticated approximately 5,000 years ago in China, and was introduced to Japan and Korea about 2,000 years ago (Carter et al. 2004; Wilson, 2008; Zhang et al. 2021). Soybean has an attractive nutritional value and is the basis for a large range of traditional Asian types of food, such as tofu, natto (fermented soybean), edamame (vegetable soybean), soybean sprouts, soymilk, and soy sauce. Soybean was introduced to North America in 1765 by Samuel Bowen (Hymowitz and Shurtleff 2005), and since the 19th century, soybean spread rapidly in the USA and became one of the major crops.

Soybean breeding in Europe

Europe is strongly dependent on soybean imports, mainly from the Americas. The protein deficit in the European Union has recently received increased attention also from politics, and expanding the planting area of leguminous crops was suggested as a promising approach to increase protein production (Häusling 2011). In Europe, soybean is mainly used as a plant protein source for animal feed, but there is also an increasing demand for human nutrition, fueled by the shift to a more vegetarian or vegan lifestyle. Indeed, the acreage of soybean planting shows a growing trend in the past two decades and as a consequence, the production of European soybean increased substantially (Figure 1). However, the import quantity of soybean in Europe was not reduced despite the increase in production and even shows a slight increase in the past decade, which illustrates that Europe still has a strong dependency on soybean imports. Soybean production in Europe is currently focused on a few countries, including the Ukraine, Italy, Serbia, France, Romania, and Croatia, which in 2020 accounted for most of the European soybean production (FAOSTAT 2021). From this it becomes apparent, that at present Southern Europe is the major soybean production region in Europe. This is mainly due to the fact that soybean is a photoperiod sensitive crop, which hinders its cultivation at higher latitude regions.

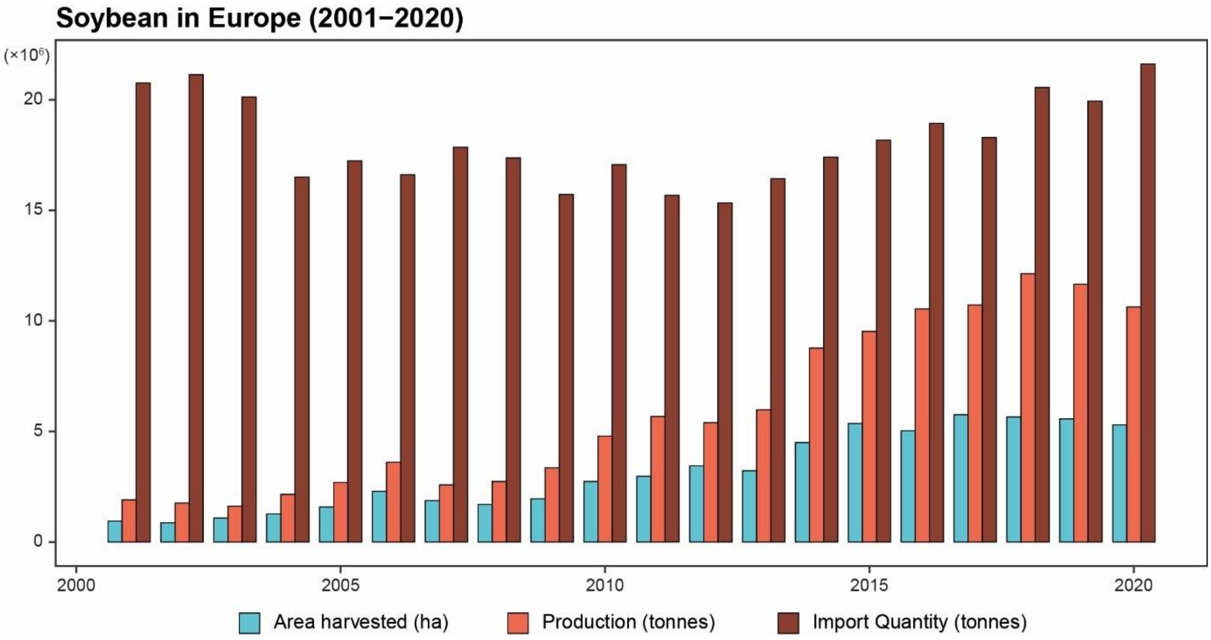


Figure 1 Annual acreage, production and import quantity of soybean in Europe from 2001 to 2020 (FAOSTAT 2021).

Similar to Europe, China is also highly dependent on soybean imports, despite it being the center of origin for soybean. Geographically, Southern and Central Europe correspond to the north of China (Figure 2), and in China soybean cultivars are grown in a broad latitudinal range from 20°N to approximately 53°N (Zhang et al. 2020b). In 2020, the main soybean producing areas were the central and northern regions, and the Heilongjiang and Inner Mongolia provinces, that are both located in the north of China, were the two top provinces with regard to soybean production. Consequently, soybean cultivars might be grown at similar latitudes in Europe and produce high seed yield, which would be in northern Germany and potentially even further north. Thus, cultivation of soybean in Central and Northern Europe appears feasible, but requires the breeding of soybean cultivars with an improved adaptation.

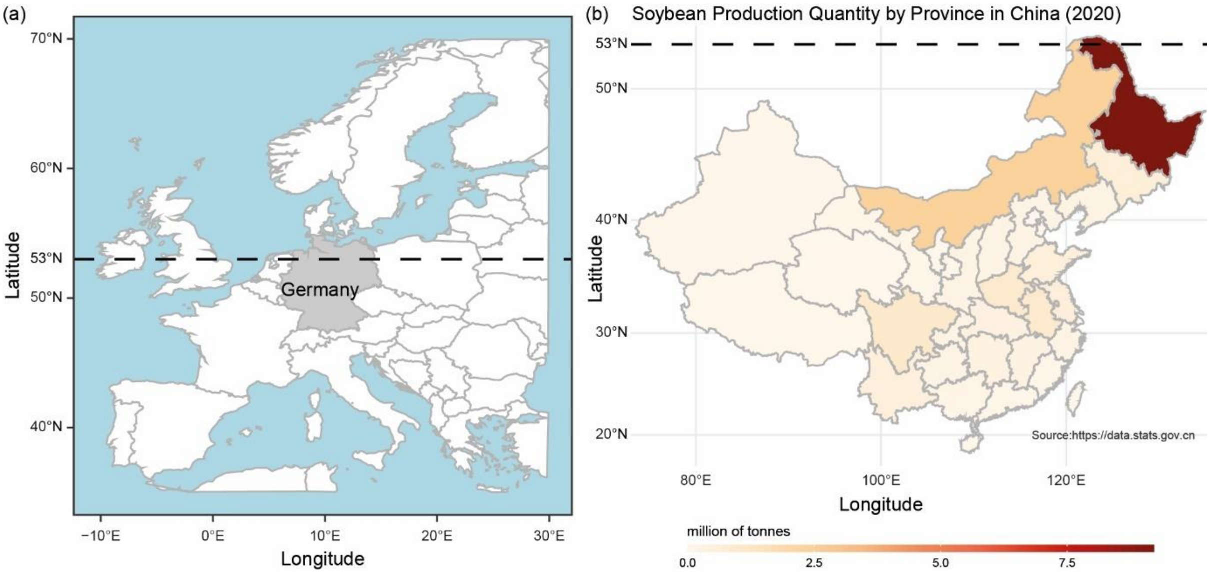


Figure 2 Comparison of the latitude of (a) Europe and (b) China, and (b) soybean production quantity by province in China in 2020. The dashed line shows 53°N, which is the northernmost region where soybean cultivars are grown in China (Zhang et al. 2020b).

Agricultural relevance and genetic basis of target traits

Protein and oil content

Protein and oil content are two important quality traits in soybean breeding, which, however, show a strong negative correlation with each other (Kurasch et al. 2017a). In Europe, soybean breeders are more focused on the improvement of protein content, owing to the need of plant protein. Until now, several studies have investigated the genetic architecture underlying protein and oil content, and hundreds of quantitative trait loci (QTL) were identified for each of the two trait and recorded in SoyBase (https://www.soybase.org/). Two large-effect QTL located on chromosomes 15 and 20 were identified in several experiments, especially the QTL on chromosome 20 that explains a large proportion of the genotypic variance (Bandillo et al. 2015; Patil et al. 2017; Lee et al. 2019).

Seed yield and thousand-seed weight

Seed yield is the most important agronomic trait and its inheritance is highly quantitative. Thus, breeders always need to take seed yield into account during the introduction and selection of soybean cultivars to the specific target regions. Thousand-seed weight is one of the most important yield component traits and generally positively correlated with seed yield. Hundreds of QTL have been reported controlling seed yield or thousand-seed weight in different populations. So far, however, there are only few genes that have been cloned and validated as the causative genes for seed weight or size, for example the *PP2C* gene from wild soybean that functions as a causal gene regulating seed weight (Lu et al. 2017b; Zhang et al. 2021).

Plant architecture

Plant architecture is a key factor affecting seed yield, and includes plant height, stem growth habit, number of nodes on the main stem, internode length, and branch number. Similar to rice and maize, the ideal architecture is also essential for the 'green revolution' of soybean (Liu et al. 2020b). This means to optimize the branching and the stem vertical growth, e.g. to increase the number of internodes, reduce internode length, and branch number. For plant height and stem growth habit, two key loci, *Dt1* and *Dt2*, with an epistatic effect have been described and the underlying genes identified (Liu et al. 2010; Tian et al. 2010; Ping et al. 2014). Besides these two well-known loci, other genes were reported to affect plant architecture. Sun et al. (2019), for example, reported that the overexpression of a micro RNA, *GmmiR156b*, resulted in an increase of the number of branches, nodes, and pods through targeting *SPL* transcripts, while there was no effect on plant height, and therefore yield increased by 46-63% per plant. Another example is the overexpression of a transcription factor, *GmMYB14*, which resulted in soybean lines that showed compact plant architecture which could allow to increase the plant density and consequently yield under field conditions (Chen et al. 2021).

Flowering time and maturity

Soybean is a short-day crop, and consequently the degree of photoperiod sensitivity of soybean cultivars is a key factor that determines their successful adaptation to regions outside its temperate center of origin. A unified standard has been defined to classify the adaptability of soybean lines to the target environments, which is the concept of maturity groups (MGs). In the USA, soybean cultivars are classified into 13 maturity groups, ranging from MG 000 for the earliest maturing soybean lines to MG X for the latest maturing ones. There are also 13 maturity groups established for Chinese soybean cultivars but these systems are flexible, and for example an even earlier maturity group MG 0000 was recently introduced (Jia et al. 2014; Li et al. 2017). In Europe, Kurasch et al. (2017b) performed a mega-environment trial at 22 locations with 75 European soybean cultivars from five maturity groups (MGs 000-II). Soybean lines from MG 000 could reach maturity at four locations above 50°N, but cultivars in later maturity groups could not. This indicates that the cultivation of soybean is limited by the cultivars' photoperiod sensitivity as well as by their response to temperature.

Previous studies have revealed several loci related to the regulation of soybean photoperiod (Cao et al. 2017; Lin et al. 2021; Zhang et al. 2021), including *E1* (Xia et al. 2012), *E2* (Watanabe et al. 2011), *E3* (Watanabe et al. 2009), *E4* (Liu et al. 2008), *E6* (Cober 2011), *E7* (Cober and Voldeng 2001), *E8* (Cober et al. 2010), *E9* (Kong et al. 2014; Zhao et al. 2016), *E10* (Zhai et al. 2014; Samanfar et al. 2017), *E11* (Wang et al. 2019), *J*

(Lu et al. 2017a; Yue et al. 2017), *Tof11*, *Tof12* (Lu et al. 2020), *Tof5* (Dong et al. 2021a), *Tof16* (Dong et al. 2021b), *Tof18* (Kou et al. 2022) and the stem growth habit genes *Dt1* (Tian et al. 2010) and *Dt2* (Ping et al. 2014). Fang et al. (2021) reported *E6* to be a novel allele of *J*.

Several studies have investigated the effect of the *E* series genes *E1-E4* and showed that the combination of alleles at these maturity loci is a key factor determining soybean's adaptation ability, allowing adaptation in a broad latitudinal range (Xu et al. 2013; Kurasch et al. 2017b; Miladinović et al. 2018). Different allelic variants of these four genes have been reported and markers have been developed to trace these alleles (Liu et al. 2008; Watanabe et al. 2009, 2011; Xia et al. 2012; Tsubokura et al. 2013, 2014). Nevertheless, to date the adaptation to higher latitudes is not well understood and is likely to rely on yet unidentified alleles at known loci as well as on novel loci.

Other important agronomic traits

Apart from the above-mentioned traits, there are additional agronomic traits that need to be considered in soybean breeding programs. Lodging, for example, is a quantitative trait that can affect seed yield and quality, and is itself also strongly affected by the environment. Similar to seed yield, lodging is also influenced by several traits, for example, plant height, stem strength, and stem diameter, which can serve as indirect traits to assist improvement of lodging resistance. Furthermore, tolerance to biotic and abiotic stresses needs to be considered in soybean breeding programs. Until now, however, breeding for resistances plays no role in Central Europe, probably due to the short history of cultivation and the still small acreage, but this can be expected to change in the future.

Genomic-assisted selection

Marker-assisted selection (MAS) has emerged in the early stages of molecular marker applications and genomic analyses as an approach to assist breeding. It refers to indirect selection of the targeted traits through the selection of molecular markers closely linked to QTL. After the introduction of molecular markers in the early 1980s, there were many studies identifying QTL and illustrating how to integrate favorable alleles in breeding programs. For simple-inherited traits controlled by one or few loci, e.g. some disease resistances, marker-assisted selection has proven its value. However, most agronomic traits are complex traits meaning that they are controlled by many loci with small effects that in addition often depend on the genetic background. In this case, marker-assisted selection is not effective for the improvement of the target traits (Holland, 2004).

Meuwissen et al. (2001) proposed genomic selection (GS) as a tool to assist the improvement of complex traits. This approach jointly uses all genome-wide markers to estimate their effects and based on these to calculate genomic estimated breeding values (GEBVs). The rapid development of genotyping and sequencing technologies has paved the way for this approach, as high-density genome-wide marker data can be routinely generated. Genomic selection requires a training set that consists of individuals with both genotypic and phenotypic data used to estimate the marker effects. Then, these model parameters are used to calculate the breeding values of untested lines with only genotypic data. For soybean, there are some studies that have evaluated the performance of genomic selection to target disease resistance, stress resistance, yield or yield-related traits (Bao et al. 2014; Ma et al. 2016; Matei et al. 2018; Wen et al. 2018; Đorđević et al. 2019; Jähne et al. 2019). The performance of genomic selection is affected by different factors, for example the genetic relatedness between the lines in the training set and in the prediction set, the size of the training set, the genetic complexity of the target trait, the linkage disequilibrium between markers and QTL, or the employed model (Wang et al. 2015; Merrick and Carter 2021). Especially the design of the training set is of utmost importance to maximize the power of genomic selection and requires further research, particularly using experimental data to complement results from simulation studies.

Phenomic selection

Genomic selection has been shown to be a promising tool for the improvement of complex traits, but a potential limitation of this approach are the costs required for the genotyping, especially when thousands or tens of thousands of selection candidates need to be predicted in early generations of the breeding cycle. Phenomic selection is an alternative approach proposed by Rincent et al. (2018). The difference is that near-infrared spectroscopy (NIRS) data are used as predictors instead of molecular markers. NIRS data is based on the absorption of electromagnetic radiation in the near-infrared region and its advantage is that it is non-destructive. It has been widely used for the determination of protein or oil content in soybean as well as in other crops, since some specific wavelengths can be used for prediction (Blanco and Villarroya 2002). The NIRS machine is therefore a common instrument used for phenotyping quality traits in breeding programs. Rincent et al. (2018) used NIRS data from wheat and poplar to predict various traits by phenomic prediction and genomic prediction with molecular marker data was used as a reference. The results showed that both approaches achieved comparable predictive abilities. So far, only few studies have explored the performance of phenomic selection based on either NIRS data or hyperspectral data, but these results also indicated the potential of phenomic selection in plant breeding (Cuevas et al. 2019; Krause et al. 2019; Parmley et al. 2019; Galán et al. 2020; Lane et al. 2020). Unlike genomic selection, phenomic selection is still in its infancy and many questions still need to be addressed. For example, it is unclear whether the factors affecting the predictive ability of genomic selection also affect that of phenomic selection, whether NIRS data from different environments has a different performance and how these can be combined, and how phenomic selection can be implemented in routine breeding programs. Especially with more spectral data from phenotyping platforms becoming available in the future, the question is how to optimally exploit this novel approach for breeding.

Current soybean breeding in Hohenheim and in Europe

The soybean breeding program in Hohenheim was built up in recent years. Previous work revealed that European soybean lines are closer to Swiss and Canadian lines than to the USA and Asian lines according to genetic diversity and population structure analysis (Hahn and Würschum 2014). However, there are rather few founder lines that have been used for the initial crosses and it is now necessary to introduce further germplasm to enrich the genetic diversity of European soybean. A potential limitation to this is the above-mentioned photoperiod sensitivity that needs to be taken into account.

Due to its long history, soybean has a large genetic diversity and, for example, over 20,000 *Glycine max* accessions are included in the USDA soybean germplasm collection that consequently presents a valuable germplasm resource for soybean breeding.

New technologies are emerging that can also assist soybean breeding. For example, the conventional approach to obtain homozygous breeding materials is recurrent selfing or the development of doubled haploid lines. In soybean no doubled haploid technology is available and there are also some disadvantages of it, e.g. not supporting early generation selection, low recombination rate and high costs. Jähne et al. (2020) developed a speed breeding protocol for short-day plants, including soybean, which allows up to five generations per year. This novel approach is currently further explored to improve the efficiency of soybean breeding in Hohenheim through a reduction of the cycle length and thus a faster generation of improved lines. In addition, NIRS data are already routinely generated to determine seed protein and oil content, but are not used further for phenomic selection, which might be promising to increase the selection gain.

Objectives of the study

The aim of this study was to evaluate, characterize, and develop approaches to improve breeding of soybean. An emphasis is given to European soybean breeding, the adaptation to Central Europe and a possible expansion of soybean cultivation to even more northern regions of Europe.

In particular, the objectives were to:

- explore the genetic architecture underlying important quality and agronomic traits towards the use of QTL in marker-assisted selection,
- evaluate the performance of genomic selection in soybean breeding and optimize the design of the training set with biparental families,
- assess parameters affecting the performance of phenomic selection and elaborate guidelines for its use in breeding programs, and
- dissect the genetic basis of soybean adaptation in early-maturing material and provide an example for the identification of genes towards a future targeted use in breeding.

Publications

Identification of seed protein and oil related QTL in 944 RILs from a diallel of early-maturing European soybean

Xintian Zhu [a], Willmar L. Leiser [a], Volker Hahn [a], Tobias Würschum [a]

[a] State Plant Breeding Institute, University of Hohenheim, Stuttgart 70593, Germany

The Crop Journal (2021) 9(1): 238-247

The original publication is available at:

https://doi.org/10.1016/j.cj.2020.06.006

Abstract

Soybean [*Glycine max* (L.) Merr.] is a global protein source and is currently expanding in Central and Northern Europe. Protein and oil content are two important quality traits that have been studied in different germplasm, however, their genetic architecture in early-maturing European soybean has not been investigated yet. In this study, we therefore performed QTL mapping for both traits using 944 recombinant inbred lines derived from eight families from a half-diallel crossing design. We identified five QTL for each trait, with the QTL on chromosomes 8, 15, and 20 being identified for both protein content and oil content. The known major QTL on chromosome 20 was detected in four families whereas the other QTL were only found in single families. Further analyses revealed the QTL to have pleiotropic but inverse effects on both traits. The effect of the major QTL was comparable between families, illustrating that it is largely independent from the genetic background. Collectively, our results illustrate the quantitative nature of protein and oil content in early European soybean. Marker-assisted selection for the QTL is possible, but the inverse effect on protein and oil content should be kept in mind.

Identification of QTL for seed yield and agronomic traits in 944 soybean (*Glycine max*) RILs from a diallel cross of early-maturing varieties

Xintian Zhu[1], Willmar L. Leiser[1], Volker Hahn[1], Tobias Würschum[2]

[1] State Plant Breeding Institute, University of Hohenheim, Stuttgart 70593, Germany

[2] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart 70593, Germany

Plant Breeding (2021) 140: 254-266

The original publication is available at:

https://doi.org/10.1111/pbr.12900

Abstract

Increasing soybean yield plays a key role in meeting the high demand for protein in Europe and other countries. The aim of this study was to dissect the genetic architecture underlying seed yield, plant height, protein yield and thousand-seed weight in early-maturing soybean. To this end, we performed QTL mapping based on 944 RILs derived from a half-diallel crossing design of five parents. We identified five to eight QTL for each of the four agronomic traits and some explained a considerable proportion of the genotypic variance. The three major QTL showed pleiotropic effects on two or more traits. Fine characterization revealed the maturity genes *E1* and *E3*, and the stem growth habit gene *Dt2* as likely candidates underlying these QTL. In general, the allele increasing seed yield also resulted in taller plants, which needs to be considered during selection due to an increased risk of lodging. Collectively, our results underline the strong effect of some loci like the *E1* gene on a range of traits including seed yield, making them attractive targets for a marker-assisted selection.

Training set design in genomic prediction with multiple biparental families

Xintian Zhu[1], Willmar L. Leiser[1], Volker Hahn[1], Tobias Würschum[2]

[1] State Plant Breeding Institute, University of Hohenheim, Stuttgart 70593, Germany

[2] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart 70593, Germany

The Plant Genome (2021) 14: e20124

The original publication is available at:

https://doi.org/10.1002/tpg2.20124

Zhu et al. 2021. Plant Genome. 14:1–15.

Abstract

Genomic selection is a powerful tool to reduce the cycle length and enhance the genetic gain of complex traits in plant breeding. However, questions remain about the optimum design and composition of the training set. In this study, we used 944 soybean [*Glycine max* (L.) Merr.] recombinant inbred lines from eight families derived through a partial–diallel mating design among five parental lines. The cross-validated prediction accuracies for the six traits seed yield, 1,000-seed weight, protein yield, plant height, protein content, and oil content were high, ranging from 0.79 to 0.87. We investigated among-family predictions, making use of the special mating design with different degrees of relatedness among families. Generally, the prediction accuracy decreased from full-sibs to half-sib families to unrelated families. However, half-sib and unrelated families also showed substantial variation in their prediction accuracy for a given family, which appeared to be caused at least in part by the shared segregation of quantitative trait loci in both the training and prediction sets. Combining several half-sib families in composite training sets generally led to an increase in the prediction accuracy compared with the best family alone. The prediction accuracy increased with the size of the training set, but for comparable prediction accuracy, substantially more half-sibs were required than full-sibs. Collectively, our results highlight the potential of genomic selection for soybean breeding and, in a broader context, illustrate the importance of the targeted design of the training set.

Phenomic selection is competitive with genomic selection for breeding of complex traits

Xintian Zhu[1,2], Willmar L. Leiser[2], Volker Hahn[2], Tobias Würschum[1]

[1] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart 70593, Germany

[2] State Plant Breeding Institute, University of Hohenheim, Stuttgart 70593, Germany

The Plant Phenome Journal (2021) 4: e20027

The original publication is available at:

https://doi.org/10.1002/ppj2.20027

Abstract

The efficiency of breeding programs depends on the ability to screen large numbers of individuals. For complex traits like yield, this can be assisted by genomic selection, which is based on estimating breeding values with genome-wide marker data. Here, we evaluate phenomic prediction, which, similar to its genomic counterpart, aims to predict the performance of untested individuals but using near-infrared spectroscopy (NIRS) data. In a large panel of 944 soybean [*Glycine max* (L.) Merr.] recombinant inbred lines phenotyped for seed yield, thousand-seed weight, and plant height at three locations, we demonstrate that the phenomic predictive abilities are high and comparable with those obtained by genomic prediction. We found that ridge regression best linear unbiased prediction performs well for phenomic prediction and that the number of wavelengths can be reduced without a decrease in predictive ability. For prediction at different locations, NIRS data from a single location can be used. However, NIRS data from different environments, like years, should be connected by common genotypes in training and prediction sets. Phenomic prediction appears to be less susceptible to relatedness between individuals in training and prediction sets than genomic prediction, as generally half-sib but also unrelated families achieved high predictive abilities. Moreover, for the same training set sizes phenomic prediction resulted in higher predictive abilities compared to genomic prediction. Phenomic prediction can be applied at different stages in a breeding program, and collectively our results highlight the potential of this approach to increase genetic gain in plant breeding.

The performance of phenomic selection depends on the genetic architecture of the target trait

Xintian Zhu[1,2], Hans Peter Maurer[2], Mario Jenz[2,3], Volker Hahn[2], Arno Ruckelshausen[3], Willmar L. Leiser[2], Tobias Würschum[1]

[1] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart 70593, Germany

[2] State Plant Breeding Institute, University of Hohenheim, Stuttgart 70593, Germany

[3] Hochschule Osnabrück, Sedanstr. 26, 49076 Osnabrück, Germany

Theoretical and Applied Genetics (2022) 135(2): 653-665

The original publication is available at:

https://link.springer.com/article/10.1007/s00122-021-03997-7

Abstract

Genomic selection is a powerful tool to assist breeding of complex traits, but a limitation is the costs required for genotyping. Recently, phenomic selection has been suggested, which uses spectral data instead of molecular markers as predictors. It was shown to be competitive with genomic prediction, as it achieved predictive abilities as high or even higher than its genomic counterpart. The objective of this study was to evaluate the performance of phenomic prediction for triticale and the dependency of the predictive ability on the genetic architecture of the target trait. We found that for traits with a complex genetic architecture, like grain yield, phenomic prediction with NIRS data as predictors achieved high predictive abilities and performed better than genomic prediction. By contrast, for mono- or oligogenic traits, for example, yellow rust, marker-based approaches achieved high predictive abilities, while those of phenomic prediction were very low. Compared with molecular markers, the predictive ability obtained using NIRS data was more robust to varying degrees of genetic relatedness between the training and prediction set. Moreover, for grain yield, smaller training sets were required to achieve a similar predictive ability for phenomic prediction than for genomic prediction. In addition, our results illustrate the potential of using field-based spectral data for phenomic prediction. Overall, our result confirmed phenomic prediction as an efficient approach to improve the selection gain for complex traits in plant breeding.

Unraveling the potential of phenomic selection within and among diverse breeding material of maize (*Zea mays* L.)

Thea Mi Weiß[1,2], Xintian Zhu[1,2], Willmar L. Leiser[1], Dongdong Li[3], Wenxin Liu[3], Wolfgang Schipprack[2], Albrecht E. Melchinger[2], Volker Hahn[1], Tobias Würschum[2]

[1] State Plant Breeding Institute, University of Hohenheim, Stuttgart 70593, Germany

[2] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart 70593, Germany

[3] Key Laboratory of Crop Heterosis and Utilization, Ministry of Education, Key Laboratory of Crop Genetic Improvement, Beijing Municipality, National Maize Improvement Center, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China

G3 Genes|Genomes|Genetics (2022) 12(3): jkab445

The original publication is available at:

https://doi.org/10.1093/g3journal/jkab445

Abstract

Genomic selection is a well-investigated approach that facilitates and supports selection decisions for complex traits and has meanwhile become a standard tool in modern plant breeding. Phenomic selection has only recently been suggested and uses the same statistical procedures to predict the targeted traits but replaces marker data with near-infrared spectroscopy data. It may represent an attractive low-cost, high-throughput alternative but has not been sufficiently studied until now. Here, we used 400 genotypes of maize (*Zea mays* L.) comprising elite lines of the Flint and Dent heterotic pools as well as 6 Flint landraces, which were phenotyped in multienvironment trials for anthesis-silking-interval, early vigor, final plant height, grain dry matter content, grain yield, and phosphorus concentration in the maize kernels, to compare the predictive abilities of genomic as well as phenomic prediction under different scenarios. We found that both approaches generally achieved comparable predictive abilities within material groups. However, phenomic prediction was less affected by population structure and performed better than its genomic counterpart for predictions among diverse groups of breeding material. We therefore conclude that phenomic prediction is a promising tool for practical breeding, for instance when working with unknown and rather diverse germplasm. Moreover, it may make the highly monopolized sector of plant breeding more accessible also for low-tech institutions by combining well established, widely available, and cost-efficient spectral phenotyping with the statistical procedures elaborated for genomic prediction - while achieving similar or even better results than with marker data.

The genetic architecture of soybean photothermal adaptation to high latitudes

Xintian Zhu[1,2], Willmar L. Leiser[2], Volker Hahn[2], Tobias Würschum[1]

[1] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, 70599 Stuttgart, Germany

[2] State Plant Breeding Institute, University of Hohenheim, 70599 Stuttgart, Germany

Highlight

Our results illustrate how combinations of loci and their interactions with the environment facilitate the expansion of soybean cultivation to regions with photoperiod and temperature conditions far beyond those of its center of origin.

Abstract

Soybean is a major plant protein source for both human food and animal feed, but to meet global demands as well as a trend toward regional production, soybean cultivation needs to be expanded to higher latitudes. In this study, we developed a large diversity panel consisting of 1,503 early-maturing soybean lines and used genome-wide association mapping to dissect the genetic architecture underlying two crucial adaptation traits, flowering time and maturity. This revealed several known maturity loci, *E1*, *E2*, *E3*, and *E4*, and the growth habit locus *Dt2* as causal candidate loci and also a novel putative causal locus, *GmFRL1*, encoding a protein homologous to the vernalization pathway gene *FRIGIDA-like 1*. In addition, the scan for QTL-by-environment interactions identified *GmAPETALA1d* as a candidate gene for a QTL with environment-dependent reversed allelic effects. The polymorphisms of these candidate genes were identified using whole-genome resequencing data of 338 soybeans, which also revealed a novel *E4* variant, *e4-par*, carried by 11 lines with nine of them originating from Central Europe. Collectively, our results illustrate how combinations of QTL and their interactions with the environment facilitate the photothermal adaptation of soybean to regions far beyond its center of origin.

Keywords

Soybean; flowering time; maturity; high latitude adaptation; low-temperature adaptation; environment interaction; genome-wide association mapping

Introduction

Soybean (*Glycine max* [L.] Merr.) is the economically most important legume crop, providing plant protein for human consumption and animal feed, and is also a source for edible oil. In Europe, the production of soybean falls far behind the actual demand and the amount of annual soybean import shows a continuously increasing trend in the last decade (FAOSTAT, 2021). Although soybean cultivation in Europe has increased in recent years (FAOSTAT, 2021), a further expansion is needed to increase soybean production. Not only in Europe, such an expansion requires the adaptation of soybean cultivars to the photothermal conditions of high latitude regions.

Soybean was domesticated in humid continental to humid subtropical regions of China between 32° and 40°N. It is a facultative short-day plant and thus photoperiod sensitivity is a crucial factor determining soybean's adaptability. During domestication and subsequent selection, soybean underwent adaptive changes that allowed a latitudinal expansion to both higher and lower latitudes. Modern soybean cultivars can be grown across a broad latitudinal range (Zhang et al., 2020), indicating that a further expansion to higher latitude regions such as Central and Northern Europe is possible. Notably, this not only requires the adaptation to the longer photoperiod but also to the cooler temperature conditions of higher latitudes. To specify the adaptability to different target regions, soybean cultivars can be classified into various maturity groups (MGs, Caldwell and Howell, 1973). The US classification system is the most commonly used one and includes 13 MGs, ranging from the extremely early MG 000 to the latest MG X and may be further expanded by earlier MGs such as MG 0000 (Jia et al., 2014). Kurasch et al. (2017) performed a mega-environment field trial at 22 locations in Europe using 75 European soybean cultivars from MGs 000 to II and the results illustrated the effect of photoperiod as well as temperature on soybean adaptation.

Several genes involved in the photoperiodic regulation of flowering time and maturity of soybean have been characterized to date, including *E1* (Bernard 1971; Molnar et al., 2003; Xia et al., 2012), *E2* (Bernard 1971; Akkaya et al., 1995, Cregan et al., 1999; Watanabe et al., 2011), *E3* (Buzzell 1971; Molnar et al., 2003; Watanabe et al., 2009), *E4* (Buzzell and Voldeng 1980; Abe et al., 2003; Molnar et al., 2003; Liu et al., 2008), *E7* (Cober and Voldeng, 2001; Molnar et al., 2003), *E8* (Cober et al., 2010), *E9* (Kong et al., 2014; Zhao et al., 2016), *E10* (Zhai et al., 2014; Samanfar et al., 2017), *J* and one of its alleles identified as *E6* (Ray et al., 1995; Bonato and Vello, 1999; Lu et al., 2017; Yue et al., 2017; Fang et al., 2021), *Tof5* (Dong et al., 2022), *Tof16* (Dong et al., 2021), and *Tof18* (Kou et al., 2022). In addition, homologs of *FLOWERING LOCUS T* (*FT*), *FT2a* (*E9*) and *FT5a*, were shown to act downstream of *E1* to control soybean flowering and maturity (Nan et al., 2014; Cai et al., 2020). By contrast, very little is known about the effect of temperature on flowering and maturity of soybean (Zhang et al. 2020). Soybean stem growth habit is another key adaptation trait as it directly affects plant height, flowering time and maturity, and two determinate stem genes *Dt1* and *Dt2* were identified (Bernard 1972; Shoemaker and Specht, 1995; Cregan et al., 1999; Liu et al., 2010; Tian et al., 2010; Ping et al., 2014). Previous studies revealed that the loss-of-function alleles at the *E1-E4* genes and their combinations are essential components of the adaptation of soybean cultivars to high latitude regions (Kurasch et al., 2017; Liu et al., 2020). Recently, mutations in *Tof11* and *Tof12* have been shown to likely have facilitated the adaptation of soybean to higher latitudes during domestication (Lu et al., 2020). Both genes promote expression of *E1*, thereby delaying flowering under long-day conditions, and consequently have no effect in an *e1* null background. In addition, Dong et al. (2022) performed whole-genome sequencing of 372 soybean accessions from the north of China, Russia and the north of Northern America and by genome-wide association mapping identified a novel locus, named *Tof5*, that has been selected in cultivated and wild soybean to facilitate adaptation to high latitude regions. These results illustrate the complexity of deciphering the genetic basis underlying soybean adaptability.

The objective of this work was to investigate the genetic mechanisms underlying adaptation to the photoperiodic and temperature conditions at higher latitudes, which have not been extensively characterized yet. To this end, we evaluated flowering time and maturity in a large diversity panel consisting of 1,503 early-maturing soybean genotypes and combined the publicly available SNP50K chip data with SNPs from whole-genome sequencing of 338 soybeans to construct the genotypic data used for genome-wide association mapping. Our results revealed the phenotypic plasticity of the two adaptation traits in response to the environment and their complex genetic architecture. Both traits are also characterized by the interaction of the genotype with the environment, which was substantiated on a genetic level by the identification of QTL with strong environment-dependent allelic effects. Further characterization identified *GmFRL1* as a candidate gene, suggesting a possible role of vernalization pathway genes in soybean adaptation, and a novel polymorphism in the maturity gene *E4* that was found in cultivars selected for adaptation to higher latitudes. The combination of major-effect QTL formed several haplotypes that allow to tailor flowering time and maturity to the target environment. Taken together, our results illustrate the plasticity and the genetic architecture of soybean photothermal adaptation as a basis for a further expansion of soybean cultivation to regions with challenging photoperiod and temperature conditions.

Materials and methods

Overview of plant materials

This study was based on a large diversity panel consisting of 1,503 soybean genotypes, a subset of 338 of them were used for whole-genome sequencing (Table S1). Within the resequencing panel, 69 are European soybean cultivars, 251 are accessions from the USDA soybean germplasm collection, and 18 are from the National Agricultural and Food Research Organization (NARO) in Japan. The remaining accessions in the diversity panel are all from the USDA soybean germplasm collection. The soybean accessions obtained from the USDA collection are all classified as maturity groups 000 to III and were in a first step evaluated for their suitability to be grown under our field conditions. Only accessions that matured and produced sufficient seeds were included in the diversity panel used for this study. The maturity groups (MGs) and origin information of the USDA accessions were obtained from GRIN-Global database (Postman et al., 2010). For 1,416 of the USDA accessions included in our panel, the maturity group classification was available and there were 43 in MG 000, 180 in MG 00, 337 in MG 0, 841 in MG I, 12 in MG II and 3 in MG III.

Sequencing, genotypic data and molecular analyses

The DNA samples of the 338 genotypes in the resequencing panel were sent to Novogene to generate the 20× coverage whole-genome sequencing data using an Illumina sequencer and seven accessions were sequenced twice. Raw paired-end resequencing reads were processed with fastp v0.20 to remove adapters and for quality check (Chen et al., 2018). The retained clean reads were mapped to the Williams 82 genome, Wm82.a2.v1 with BWA v0.7.17-r1188 using the default parameters (Li and Durbin, 2009; Schmutz et al., 2010). The output bam files were merged by individuals and sorted, and then used to generate the index files using SAMtools v1.10 (Li et al., 2009). Next, we used BCFtools v1.10.2 to perform variant calling with the default parameters (Li, 2011). SNPs or Indels with minor allele frequency (MAF) below 0.02 or a heterozygous rate above 0.8 were discarded by VCFtools v4.2 (Danecek et al., 2011) and BCFtools, respectively. The effect of the variants was predicted with SnpEff v4.3 (Cingolani et al., 2012) and the reference genome was the same to the above mapping process. The known polymorphisms of candidate genes were scored by the identified SNPs and using the Integrative Genomics Viewer (IGV, Robinson et al. 2011).

SoySNP50K chip data of the USDA soybean germplasm have been published (Song et al., 2015) and the SNPs in variant calling format that mapped to Wm82.a2.v1 were downloaded from SoyBase (Grant et al., 2010). First, SNP50K data was checked for consistency between the REF allele and the reference genome Wm82.a2.v1. Next, SNP50K chip data was merged with the identified SNPs from the resequencing panel which resulted in a total of 33,486 SNPs in the combined genotypic data set.

Principle component analysis (PCA) was performed with the R package 'SNPRelate' (Zheng et al., 2012). Linkage disequilibrium (LD) was calculated as r^2 between pairs of markers with plink v1.9, the parameters were set as --r2 --ld-window-kb 5000 --ld-window-r2 0 --ld-window 99999 (Purcell et al., 2007). Then the smoothing method was used to fit a cubic smoothing spline for the decay of LD with physical distance.

Phenotypic data

The diversity panel with a total 1,503 genotypes was evaluated at six environments. Most of the lines were grown at Eckartsweier (EWE, 48°31'17"N, 7°52'13"E, 140 m asl) and Hohenheim (HOH, 48°42'53"N, 9°11'16"E, 400 m asl) in 2018 and 2020. Notably, in 2019 only a part of the individuals from the resequencing panel were grown at the same two locations (Fig. S1). The field trial was unreplicated due to the large number of genotypes. The sowing date ranged from the end of April to early May. Experiments were conducted in plots of a single row of 1.2 m length. Five traits were investigated in one or several environments (location-year combination), including days to beginning of flowering (except HOH2019), days to maturity (all environments), flower color (except HOH2019), hilum color (EWE2019, HOH2019), and pubescence color (HOH2019). Flowering (R1) and maturity (R8) dates were recorded when 50% of the plants in a plot had reached the R1 or R8 stage, respectively, and were then used to calculate the days from sowing to the start of flowering and to maturity. Flower color was recorded as white or purple. Hilum color was scored from 1 (light color) to 8 (dark color). Pubescence color was recorded as tawny or grey.

The raw phenotypic data was filtered for outliers with the Bonferroni-Holm test (Bernal-Vasquez et al., 2016). Best linear unbiased estimates (BLUEs) of each accession across environments were estimated by the following model:

$$y_{ijk} = \mu + g_i + e_j + (ge)_{ij} + \varepsilon_{ijk}$$

where y_{ijk} indicates the observed trait value of each plot, μ the intercept, g_i the effect of the ith genotype, e_j the effect of the jth environment, $(ge)_{ij}$ the genotype-by-environment interaction between ith genotype and the jth environment, and ε_{ijk} the residual error. The residual variance was assumed as heterogenous between environments. To calculate BLUEs, g_i was modeled as a fixed term.

Broad-sense heritability was estimated by the following formula (Piepho and Möhring, 2007), $H^2 = 1 - \frac{\bar{v}_{\text{BLUP}}}{2\sigma_g^2}$, where $\bar{v}_{\text{BLUP}}$ is the mean variance of the difference of two estimated best linear unbiased predictors (BLUPs) and σ_g^2 is the genotypic variance component. The calculation of BLUPs was conducted with the above model by treating g_i as a random term. In addition, BLUEs within each environment were calculated. All mixed model calculations were performed with ASReml4 (Butler et al., 2018).

Average daily Crop Heat Units (CHU) were calculated based on the previously described approach (Brown and Bootsma, 1993) and then used to sum the daily CHU until flowering or maturity to obtain the cumulative CHU since the sowing date in each environment.

Genome-wide association mapping

For the genome-wide association mapping, we used all 1,503 genotypes. For 334 genotypes from the resequencing panel that had a SNP missing rate smaller than 0.15 in the combined genotypic data set, the marker data were derived from the resequencing data. For the accessions with both resequencing data and SNP50K array data, we used the genotypic data generated from whole-genome sequencing for the further analyses. SNPs with minor allele frequency above 0.05 and missing rate below 0.10 were filtered for genome-wide association mapping and further analyses, resulting in a total of 31,357 markers.

Genome-wide association mapping was performed with the R package 'GAPIT3' and the Blink model (Wang and Zhang, 2021). First, three principle components were used as covariates. The significant marker-trait associations were determined by a Bonferroni-corrected threshold ($P <$ 0.05). Before estimating the proportion of explained genotypic variance of each significant marker, the genotypic data was imputed with Beagle v5.1 (Browning et al., 2018). The proportion of explained genotypic variance was then estimated by fitting each marker in a linear model to calculate $p_{G\text{-}Single}$ or by fitting all significant markers in a joint model in the order of a decreasing strength of association to calculate $p_{G\text{-}Joint}$. The allele substitution effect was calculated as the regression coefficient of a linear model fitted with each significant marker separately. To detect QTL-by-environment interactions, we used the R package 'gwasQ×E' (Yamamoto and Matsunaga, 2021) with a Bonferroni-corrected significance threshold to define QTL.

Multiple comparisons of polymorphism groups of each locus were done with the LSD.test function in the R package ‘agricolae’ with the default alpha level at 0.05 and p-values adjusted by the Bonferroni method (Steel et al., 1997; de Mendiburu, 2021).

Results

Molecular analysis of the diversity panel

For this study, we established a large diversity panel with 1,503 early-maturing soybean genotypes, mainly of diverse global origin obtained from the USDA genebank but also many of the older and current European cultivars. From this panel, 338 cultivars were subjected to whole-genome sequencing. The genotypic data was obtained from the combination of the released SNP50K chip data of USDA accessions and the same SNPs extracted from the resequencing data. This resulted in a total of 33,486 SNP markers that allowed a joint analysis of all the lines and even the comparison with the entire genotyped USDA soybean collection.

Principle component analysis showed the extensive genetic diversity captured by both the diversity panel and the resequencing panel (Fig. 1a). Nevertheless, the entire molecular space is not covered equally, which is due to the sampling of early-maturing genotypes in our panel. While there is no clear separation, the accessions classified by USDA as MGs 000-0 have a certain center that consequently overlaps with that of our diversity panel (Fig. 1b). Regarding their origin, soybeans from Asia or USA showed a certain distinctness, while the European soybeans can be found over most of the molecular diversity space with similarity to both Asian and US material (Fig. 1c).

The mapping resolution of association mapping is determined by the linkage disequilibrium. In the diversity panel, linkage disequilibrium (r^2) decayed with physical distance below a threshold of 0.2 after approximately 500 Kb (Fig. S2). In addition to the analysis of linkage disequilibrium, we performed genome-wide association mapping for three traits of more simple inheritance (flower color, hilum color and pubescence color) to obtain an idea about the mapping resolution in this panel, and association signals close to the causal loci (*W1*, *I*, *T*) were identified (Fig. S3), illustrating the potential of this diversity panel for fine-mapping.

Statistical overview of the phenotypic data

The majority of the lines in the diversity panel were grown at four environments (1,279 in EWE2018, 1,492 in EWE2020, 1,450 in HOH2018, 1,419 in HOH2020) and a subset of around 220 were also grown at EWE2019 and HOH2019 (Fig. S1). Two adaptation traits, flowering time and maturity, were assessed for each genotype. Both showed a wide phenotypic variation across environments as well as within single environments, with ranges of 42.3 days for flowering time and 64 days for maturity across environments (Fig. 2). The heritabilities of the two traits within and across environments were high or very high, ranging from 0.79 to 0.96 (Fig. S1).

As temperature is expected to have a substantial effect on soybean adaptation, we also investigated the accumulated crop heat units for the time to flowering and maturity within each single environment. The average time to flowering and maturity differed between environments and was for example shorter for both locations in 2018 than in 2020 (Fig. 2). When taking the temperature into account by calculating accumulated crop heat units, these differences largely diminished and all environments had rather similar average time to flowering and maturity. These results substantiate the effect of temperature on the phenology and thus the adaptation of soybean in cooler climates. Taken together, the observed variation and the high heritabilities form an excellent basis to study the underlying genetic architecture.

Association mapping for flowering time and maturity

We then performed genome-wide association mapping for the two adaptation traits flowering time and maturity (Fig. 3, Table S2). This identified 30 putative QTL for flowering time and 27 for maturity across environments. Seven QTL were found to control both traits as the physical distance between the respective significant markers was no more than 50 Kb: qFT4/qMT3, qFT10/qMT5, qFT12/qMT7, qFT15/qMT10, qFT23/qMT15, qFT28/qMT22, and qFT29/qMT25, with qFT12/qMT7 being the by far most significant QTL for both traits. This finding is in line with the observed correlation between flowering time and maturity of $r = 0.75$. For flowering time, we identified three QTL (qFT12, qFT20, qFT28) explaining more than five percent of the genotypic variance and two such QTL (qMT2, qMT4) for maturity. Jointly, the identified QTL explained 62.45% of the genotypic variation of flowering time and 61.71% of the variation of maturity. We also performed genome-wide association mapping within each single environment (Fig. S4, S5, Table S3). This revealed that some QTL were identified in several environments, for example the QTL for flowering time and maturity on chromosome 10, whereas other QTL were specific for only one environment.

This prompted us to scan the genome for QTL-by-environment interactions. We identified such a QTL interacting with the environment for flowering time on chromosome 10, that likely corresponds to the maturity gene *E2* (Fig. 4a, Table S4). Considering the strong effect of this locus, the analysis was performed again in a subset of 1,233 genotypes of the population fixed for *E2*. Interestingly, this identified another environment-dependent QTL on chromosome 2 (Fig. 4b, Table S4). We then analyzed the effects of the alleles at these two loci in the different environments (Fig. 4c). For *E2*, the effect varied substantially among environments and for the

QTL on chromosome 2, we even observed a change in the order of the allelic groups, meaning that depending on the environment, one or the other allele promoted or delayed maturity, respectively.

Identification of known loci related to flowering time and maturity

To further characterize the identified QTL, we searched the neighboring ±250 Kb regions for candidate genes, thereby focusing on QTL identified across environments or in multiple environments (Table S5). The most significant QTL for both traits was on chromosome 10, with the associated markers being located ~15 Kb downstream of the known maturity locus *E2*, thus strongly suggesting *E2* to be the causal gene underlying this QTL (Fig. S6a). Another major maturity locus is *E1* and a QTL for flowering time was identified around 160 Kb upstream of *E1* while a QTL for maturity was about 4 Mb upstream of it (Fig. S6b). For chromosome 19, we found a QTL for both traits located between *Dt1* and *E3* but closer to *E3* with a distance less than 500 Kb (Fig. S6c). Similarly, QTL for both traits were found downstream of *E4* (Fig. S6d). Apart from the *E* loci, the stem growth habit gene *Dt2* was found around 16 Kb downstream of a QTL for maturity (Fig. S6e). Given their known effect on flowering time and/or maturity, these genes are likely candidates underlying the QTL.

Identification of allelic variants of *E4*, *GmFRL1* and *GmAP1d*

Several polymorphisms have been identified for the *E1-E4* maturity genes in soybean accessions from East Asia. We used the resequencing data to screen for novel polymorphisms that may be specific to early-maturing soybeans. For *E4*, we identified the wild-type allele *E4*, the known alleles *e4-SORE-1*, *e4-kam*, and *e4-kes*, as well as a novel variant, termed *e4-par* (Fig. 5a-c). Similar to *e4-kam* and *e4-kes*, *e4-par* is caused by a single base nucleotide deletion in exon 2 of *E4* (*Glyma.20G090000*) and results in a frameshift at amino acid 832 and consequently in a truncated protein with 852 amino acids. Eleven out of the 338 accessions in the resequencing panel carry this *e4-par* allele and all originate from the early maturity groups 000 and 00. Nine of them have their origin in Central Europe and another two are from Canada and China. These results indicate that the novel *e4-par* allele may be involved in the adaptation of soybean to higher latitude conditions.

We, therefore, also investigated the frequencies of alleles at the major QTL dependent on the country of origin (Fig. S7). For *E1*, there was a clear trend that genotypes originating from Canada or Central Europe and thus from regions of higher latitude, had a much higher frequency of the early-maturing *e1-nl* allele. A similar picture was observed for the frequency of the *e4-SORE-1* allele.

Further characterization also identified a possible candidate gene for the flowering time and maturity QTL qFT10/qMT5 on chromosome 8 (Fig. 5d-e). The significant SNP is located close to the gene *Glyma.08G325700* that encodes a homolog of *FRIGIDA-like 1* (*FRL1*) in *Arabidopsis*. There are 13 SNPs within the coding region of *GmFRL1* leading to synonymous or missense variants and these can be grouped into three haplotypes (Fig. 5e, Table S6). Several of these polymorphisms were in perfect linkage disequilibrium (r^2 = 1.00) with the peak SNP of the QTL.

For the QTL-by-environment interaction on chromosome 2, the MADS-box transcription factor *GmAPETALA1d* (*GmAP1d*, *Glyma.02G121600*) was identified as a candidate (Fig. 4, 5f). Sequence analysis revealed a SNP resulting in a missense variant (Table S6), which in the resequencing panel was in perfect linkage disequilibrium (r^2 = 1.00) with the peak SNP of the QTL.

Allelic effect of causal genes or major-effect QTL on flowering time and maturity

We further assessed the effect of alleles at the causal genes or major-effect QTL on days to flowering and maturity in our panel of early-maturing soybeans (Fig. 6a). Effects of polymorphisms in *E1*, *E2*, *E3*, *E4*, *Dt2* and *GmFRL1* (Table S1) were evaluated in the resequencing panel and two QTL explaining more than five percent of the genotypic variance for either flowering time or maturity were assessed in the entire diversity panel. For *E1*, we observed a significant effect of the null allele *e1-nl* on flowering time and on maturity, whereas the effect of the weaker *e1-as* allele was only significant for flowering time. Despite this strong effect on the target traits, the QTL likely identifying this locus, qMT4, was located ~4 Mb upstream of *E1* (Fig. S6b). We therefore examined the frequencies of the three *E1* alleles in the two allelic groups of the QTL marker (Fig. 6b). This revealed a similar frequency of the wild-type allele and reversed frequencies of the hypomorphic *e1-as* and *e1-nl* alleles, with one or the other being predominant in the two QTL allelic groups. This marker therefore mainly portrays differences between the *e1-as* and *e1-nl* alleles.

For *E2*, the dysfunctional *e2-ns* allele significantly differed from the wild-type *E2* allelic group for maturity but not for flowering time (Fig. 6a). Notably, in this panel only 25 genotypes (7.4 %) carry the wild-type allele. We detected five *E3* polymorphisms in the resequencing panel, the wild-type *E3-Mi* and *E3-Ha* alleles and the alleles *e3-tr*, *e3-fs*, and *e3-ns* reported to affect *E3* function. Multiple group comparisons showed no significant differences for both flowering time and maturity, maybe also due to the in part small group sizes. *e3-ns* genotypes matured on average around 7 days earlier than the wild-type, but only 15 individuals carry this allele. Likewise, *E4* also had five variant groups, *E4*, *e4-SORE-1*, *e4-kes*, *e4-kam* and the novel *e4-par* allele. *e4-SORE-1* and *e4-par* were earlier flowering, though this was not significant, and showed a significantly earlier maturity of around 11 and 16 days, respectively. For *Dt2*, the known missense variant showed an effect on both traits.

The long juvenile locus *J* was identified for flowering time in HOH2018 and HOH2020 (Table S5). The polymorphisms of *J* were the same as described previously and can be grouped into three haplotypes (HT1, HT15, HT23) of which H23 is only carried by four individuals (Fig. S8). Both flowering time and maturity of HT1 were earlier than HT15. For the candidate gene *GmFRL1*, there were significant differences between the three haplotypes. HT1 and HT2 showed an earlier flowering than HT3, and for maturity HT2 was 7 days earlier than HT3. The two major QTL, qFT28 on chromosome 19 and qMT2 on chromosome 4, also showed significant differences for both flowering time and maturity.

We next investigated the combined allelic effects of the two loci explaining the highest proportion of genotypic variance for flowering (*E2* and qFT28) time and maturity (*E1* and qMT4). Neglecting groups with less than five individuals, this resulted in 11 haplotypes (Fig. 6c, Table S2). In general, the results suggest at least in part additive effects of these loci, as combinations of more alleles advancing flowering or maturity showed an increasingly earlier flowering and maturity. The same picture was observed for combinations of alleles of *E1*, *E2*, *E3* and *E4* (Fig. S9). Taken together, this revealed the in part substantial effect of single loci on the two adaptation traits as well as the possibility to combine their effects to expand the photothermal range of soybean cultivation.

Phenotypic and genetic analyses in maturity groups

Maturity groups are a concept to classify soybean adaptation to different target regions. This classification is based on phenotypic assessment in the field, but also depends on genotype-by-environment interaction and thus on the test environment(s). Consequently, flowering and maturity dates of accessions within the same maturity group can differ substantially. We therefore evaluated the flowering time and maturity of the USDA accessions based on their maturity group information. Notably, only accessions that matured under our field conditions were included in the early-maturing diversity panel. As a consequence, MGs 000-III were present, but only three accessions were from MG III and 12 accessions from MG II (Fig. 7a). We observed a generally increasing and significantly different flowering time and maturity, the higher the USDA maturity group classification. Nevertheless, there was also substantial variation within each group that in MGs 00 to I amounted to approximately 20 days difference in maturity.

To investigate the effect of major loci on this classification, we exemplarily assessed the frequency of the three *E1* alleles in each MG in the resequencing panel (Fig. 7b). We found the highest frequency of the wild-type *E1* allele in MG I and the highest frequency of the dysfunctional *e1-nl* allele in MG 000. The previous analysis had shown that combinations of alleles at *E1-E4* or at major QTL allow to advance flowering time and maturity. We therefore next assessed the frequency of these alleles and of the *E1-E4* haplotypes in the different maturity groups (Fig. 7c, Table S1, S7). As described for *E1*, this revealed for many loci a trend of an increasing frequency of the early-flowering or -maturity allele in earlier maturity groups. Notably, however, even for the major-effect loci, these alleles are not fixed in the earliest MG 000 and thus, many of these lines do not carry all the flowering or maturity advancing alleles. This was substantiated by the analysis of *E1-E4*, for which there are different haplotypes for each of the maturity groups. These results corroborated the substantial effect of single loci but also the complexity of the genetic architecture underlying flowering time and maturity in soybean.

Discussion

Soybean is a short-day, photoperiod-sensitive crop, originating from humid continental to humid subtropical regions, which naturally restricted its area of cultivation. Previous studies have illustrated the complex genetic architecture of soybean flowering time and maturity (Cao et al., 2017; Lin et al., 2021; Zhang et al., 2022), but less is known about the loci and alleles governing the expansion of soybean to regions with long-day and cooler temperature conditions (Jia et al., 2014; Dong et al., 2022). The aim of this study was therefore to dissect the genetics underlying the photothermal adaptation of soybean to such conditions using a large diversity panel of early-maturing soybeans.

Effect of temperature on soybean adaptation

The growth and development of soybean are not only affected by photoperiod but also by temperature (Hesketh et al., 1973; Cregan and Hartwig, 1984). In Canada, for example, the criterion for evaluating the suitability of cultivars for a target region is to calculate the cumulative crop heat units during the growth season (Bootsma et al., 1992; Brown and Bootsma, 1993). Our results corroborate the substantial effect of temperature on soybean development, as the mean flowering time and maturity differed among the six environments, whereas the cumulative crop heat units from sowing to the begin of flowering or maturity were comparable (Fig. 2). Compared with photoperiod sensitivity, the genetic basis underlying temperature response of soybean is much less understood (Zhang et al., 2020) and also shows an interaction with photoperiod (Cober et al., 2001; Wu et al., 2015; Mao et al., 2017), which warrants further research.

In general, plants' responses to light and temperature are in part related, even on a molecular level. Phytochromes, namely *PHYB*, have been attributed a dual function as thermo- and photosensors in *Arabidopsis* (Jung et al. 2016). Also in *Arabidopsis*, the *PHYTOCHROME-INTERACTING FACTOR* (*PIF*) family, a group of helix-loop-helix transcription factors, was shown to act as a central hub in the regulation of growth in response to different environmental cues, including seasonal temperature variations (Wigge 2013; Cordeiro et al. 2022). The role of these *PIF*s in crops, by contrast, is just starting to be understood (Cordeiro et al. 2022). In this context, it is interesting that a *PIF3* homolog (*Glyma.19G222000*) was identified in our study neighboring associations on chromosome 19 for both flowering time and maturity (Fig. S10, Table S5). This gene may therefore represent a candidate for a temperature-response QTL and sequence analysis in the resequencing panel revealed two groups of linked polymorphisms with allelic differences for flowering time or maturity. In addition, as transcriptional regulation and protein stability have also been shown to be mechanisms underlying *PIF* action (Cordeiro et al. 2022), this gene might be a promising candidate for further research.

Population structure of the diversity panel in relation to the USDA germplasm

The center of origin of soybean lies in China and in many Asian countries soybean is a common crop with a longstanding history. In the 18th century soybean was introduced to North America and to Europe, but in Europe it has not become a major crop. Breeding of cultivars like the 'Fiskeby' series in Sweden in the middle of the last century showed that soybean adaptation is possible even at high latitudes of 58° N. Nevertheless, only rather recently have European cultivars of early maturity groups been bred and facilitated the expansion of the cultivation of soybean in Central Europe. We compared the genetic diversity of our panel with that of the genotyped accessions of the USDA genebank. Overall, the panel covers a broad space of the molecular diversity, despite the fact that only early-maturing soybeans were included (Fig. 1). Nevertheless, similar to a previous study, there was a certain pattern of population stratification among maturity groups (Bandillo et al., 2015), and genotypes in the diversity panel were genetically more close to the early maturity groups (Fig. 1b). Regarding the genetic background of cultivars originating from Europe, our results corroborated previous findings showing that they incorporate diversity from both America and Asia (Fig. 1c, Hahn and Würschum 2014). This is not unexpected, as the more recently released cultivars trace back to material from North America, mainly Canada, as this possesses the required level of cold tolerance. Interestingly, this material goes back in part to the European breeding efforts and founder lines like 'Fiskeby'. The earlier breeding efforts as well as some longstanding Central European breeding programs have also incorporated Asian soybeans, explaining the relatedness to that material.

The genetic architecture of flowering time and maturity in early-maturing soybeans

For flowering time and maturity across environments, we identified 30 and 27 QTL, respectively, substantiating the complex genetic architecture underlying the two adaptation traits (Fig. 3). Seven QTL were found for both traits, indicating that flowering time and maturity have an in part shared genetic basis, but also loci that are specific for only one of the two traits (Zhang et al., 2015).

In addition, we also performed association mapping with the data from the single environments, which identified additional QTL that were specific to one or some of the environments (Fig. S4, S5). One possible reason for this are the different numbers of genotypes that were evaluated at the six environments (Fig. S1). However, at four environments the full panel was evaluated and the detected QTL still showed considerable differences. This illustrates that flowering time and maturity are controlled by both environment-sensitive and environment-insensitive loci (Mao et al., 2017). The environment-dependency of some QTL was further corroborated by the identification of QTL-by-environment interactions, which identified *E2* and an additional locus on chromosome 2 (Fig. 4).

Causal genes regulating flowering time and maturity

Further characterization resulted in the identification of several genes as candidates underlying the identified QTL, including the maturity genes *E1*, *E2*, *E3* and *E4* (Fig. 3, S6). *E2*, for example, could be identified by markers just a few kb away from the gene. Interestingly, *E1* as a major QTL for maturity was identified by a marker that is located ~4 Mb upstream of *E1*. This is similar to previous studies that also found association signals for *E1* in a rather large genomic region and we therefore investigated the possible cause for this. *E1* is present in this panel with three alleles, the wild-type allele, a weak allele *e1-as* resulting in an amino acid change and the dysfunctional allele *e1-nl* with the deletion of the entire gene region. We found that the associated marker mainly captures differences between the two hypomorphic alleles *e1-as* and *e1-nl* (Fig. 6b). In general, this highlights the limitation of genome-wide association mapping when a multi-allelic locus like *E1* should be detected by bi-allelic markers.

Our analyses also identified the *J* locus as a candidate for a flowering time QTL detected in two environments. *J* is known as a key locus required for the adaptation of soybean to the lower latitudes of the tropics, as *j* mutant alleles allow to extend the vegetative phase and thereby improve yield under short-day conditions (Lu et al., 2017). The J protein associates with the *E1* promoter and acts as a direct transcriptional repressor of *E1*, which in turn relieves the *E1*-dependent repression of *FT2a* and *FT5a* to promote flowering. Lu et al. (2017) identified 40 polymorphisms in the *J* coding sequence and defined 34 haplotypes, of which only six generate frameshifts and were considered to be clear loss-of-function alleles while two non-synonymous SNPs present in low-latitude accessions were characterized as weak loss-of-function alleles. Thus, eight *j* mutant alleles were described to facilitate adaptation to lower latitudes, but these do not include the haplotypes HT1 and HT15 identified in our panel of early-maturing soybean and the function of these as well as of the remaining haplotypes remains elusive. Interestingly, Lu et al. (2017) reported no difference between HT1 and HT15 on *E1* promoter activation in transient assays. Our results suggest a role of *J* not only in the adaptation to lower but also to higher latitudes. Given its role in the transcriptional regulation of *E1*, that itself is central to the adaptation to higher latitudes, this appears plausible. Thus, further molecular work is required to investigate the effect of the different *J* haplotypes on *E1* under different natural environmental conditions as well as possible variations in the promoter region and thus in the expression of *J*.

Adaptation of soybean to higher latitudes may not only be achieved by novel loci, but also by yet unknown alleles at known loci. We therefore used the resequencing data to screen the known maturity genes for sequence polymorphisms. Indeed, this identified a novel *E4* variant, termed *e4-par*, that was found in 11 genotypes, mainly originating from Central Europe (Fig. S7). Tsubokura et al. (2013) investigated *E4* polymorphisms in accession with various origin in East Asia and reported four dysfunctional alleles (*e4-oto*, *e4-tsu*, *e4-kam*, *e4-kes*) caused by single-base deletions that result in a truncated protein sequence. Another dysfunctional allele is *e4-SORE-1* caused by a 6238 bp *Ty1/copia*-like retrotransposon insertion in exon 1 (Liu et al., 2008), and all these dysfunctional alleles originate from soybean landraces from East Asia (Kanazawa et al., 2009; Tsubokura et al., 2013; Xu et al., 2013; Langewisch et al., 2014). Likewise, *e4-par* is also a single-base deletion that produces a frameshift and a premature stop codon after 852 amino acids. We found a strong effect of this allele and genotypes carrying it were the on average earliest flowering and maturing ones (Fig. 6a). It must be noted here, that in a diversity panel this must always be interpreted with caution, as the genetic background may vary between allelic classes, which could also contribute to the observed differences. The origin of this allele requires further research, but the one Chinese line identified in our resequencing panel to carry it, was collected in the Northeast of China. As nine of the eleven genotypes carrying this allele originate from Central or Northern Europe and one from Canada, it appears likely that this allele was specifically selected for the adaptation to higher latitude growth conditions.

A possible role of vernalization pathway genes in soybean adaptation

The known maturity *E* loci illustrated that fine-mapping is possible in this diversity panel, but this also depends on the genomic region and on the locus. For both flowering time and maturity, we identified a QTL qFT10/qMT5 that was close to the *GmFRL1* locus, encoding a homolog of FRIGIDA-like 1 protein in *Arabidopsis*. *FRL1* is an essential component for the up-regulation of the flowering inhibitor *FLOWERING LOCUS C* (*FLC*) mediated by *FRIGIDA* (*FRI*) in the winter-annual growth habit type (Michaels et al., 2004). Interestingly, vernalization pathway genes have been retained in the soybean genome, even though soybean does not require exposure to low temperatures to initiate flowering (Jung et al., 2012), and another locus on chromosome 5 encoding FRIGIDA-like 3 protein was recently identified as a candidate gene for soybean flowering time (Li et al., 2019). Previous studies also suggested that the vernalization pathway-related genes *GmVRN1-like* and *GmFLC-like* play crucial roles in the low temperature-induced regulation of soybean flowering time (Lü et al., 2015; Lyu et al., 2020). *GmFRL1* identified here therefore appears as an interesting candidate involved in the regulation of flowering and maturity in early-maturing soybeans, and its validation and molecular characterization warrant further research.

GmAP1d as a candidate for a QTL-by-environment interaction

A homolog of the *Arabidopsis* floral meristem and organ identity gene *APETALA1*, *GmAP1d*, was identified as a candidate gene for the environment-dependent QTL on chromosome 2. *GmAP1* was reported to be induced by *GmFT2a* and *GmFT5a* during flowering induction (Nan et al., 2014). The soybean genome contains four *AP1* homologs, and the quadruple mutant of *GmAP1* showed a delayed flowering under short-day conditions, whereas overexpression of *GmAP1a* resulted in an earlier flowering (Chen et al., 2020). In addition, Xu et al. (2021) recently proposed a cotyledon-based model for soybean adaptation to high latitude regions, in which *GmFT2a* is upregulated in a photoperiod-dependent pathway to activate *GmAP1* expression in the stem apex to induce flowering. Moreover, Li et al. (2021) confirmed the *GmAP1* homologs as primary targets of *FT2a*/*FT5a* in the regulation of flowering time as well as their role in conferring a long-juvenile trait required for adaptation to the low latitude regions in the tropics. Collectively this makes *GmAP1d* a likely candidate for the environment-interaction QTL. We identified a single amino acid change in *GmAP1d*, but given the potential effect of differential expression of *GmAP1* on flowering induction, also polymorphisms in regulatory regions appear as possible candidates. Interestingly, this QTL showed reversed allelic effects at both locations in 2019 compared to the other four environments, illustrating that its effect depends on specific environmental factors like precipitation or temperature. In general, this example illustrates that QTL effects are often not static but dependent on environmental cues, which highlights the importance of understanding the genotype-by-environment interaction on a molecular level towards a tailored adaptation of crops to different target environments and their prevailing climatic conditions.

Transferability and genetic classification of maturity groups

Soybean cultivars are classified into maturity groups and also a large part of the USDA germplasm has been evaluated for maturity and based on that assigned to maturity groups. This is very valuable as it allows to choose accessions that are more likely to fit to a certain target region. However, maturity is also dependent on the environment and in our case this classification is based on the assessment at a much lower latitude than the target region in Central Europe. Kurasch et al. (2017) performed a mega-environment field trial using lines from MGs 000-II and found that some accessions from MGs I and II could not reach maturity in the most northern mega-environments, which indicated that lines from MGs 000-0 may be best suited for Central and Northern Europe. We found that the available maturity group classification on average matched the flowering time and maturity of the accessions (Fig. 7). Nevertheless, we also found substantial overlap between the groups and the flowering time and maturity of some accessions in MG I were as early as those of the earlier maturity groups. This is on the one hand due to noise in the maturity group designation of the accessions, but also in line with the observed environment-specific QTL and illustrates the substantial genotype-by-environment interaction for both traits. Thus, prediction of the suitability of a genotype for a target region requires a better understanding of the effects of QTL or specific QTL alleles under different environmental conditions.

Another important question related to this is how to design genotypes suited for a certain early maturity group. Our results revealed a trend towards the early maturity-conferring alleles in earlier maturity groups, but also showed that these alleles are not fixed, not even in the earliest maturity group MG 000. While there is a certain inaccuracy in the classification of the genotypes to the maturity groups, this illustrates that none of these alleles is absolutely required for an early-maturing genotype. Not having one allele can be compensated by the alleles at other loci. What is observed for the major loci applies even more so to the many loci with small effect and thus, for them the trend of an increasing frequency of the early allele in earlier maturity groups is less pronounced. In addition to the effect of a locus on maturity, the allele frequencies will also depend on the breeding history of the lines, the availability of certain alleles in breeding programs, and the possible pleiotropic effects on other target traits. In summary, this illustrates that many allele combinations can lead to a certain maturity group and even to early maturity. This suggests that in breeding programs for higher latitude conditions with their long photoperiods and cooler temperatures, also crosses of lines of early with lines from later maturity groups can be envisaged. Nevertheless, an increasing number of early maturity-conferring alleles is required to achieve earlier maturity and the number of suitable haplotypes of the relevant loci can be expected to be further reduced when even earlier soybeans are bred.

Conclusions

In this study, we used a large panel of early-maturing soybean lines with various origin to explore the genetics underlying adaptation to the photothermal conditions of high latitude regions. Our result not only revealed the substantial effect of known loci but also novel alleles as well as candidate genes that affect flowering time and maturity under these environmental conditions. Overall, this study illustrates the availability of genetic variation and thus the plasticity of soybean adaptation, which forms the basis for a further expansion of the soybean cultivation area.

Acknowledgments

We thank the technical team of the State Plant Breeding Institute for excellent work and the USDA and NARO genebanks for providing seeds.

Conflict of Interest

The authors declare no conflict of interest.

Funding

This work was funded by the Federal Ministry of Education and Research of Germany (BMBF, grant numbers 031B0339A, 031B0339B).

Data Availability

All study data are included in the article and/or supporting information.

References

Abe J, Xu D, Miyano A, Komatsu K, Kanazawa A, Shimamoto Y (2003) Photoperiod-insensitive Japanese soybean landraces differ at two maturity loci. Crop Sci **43**(4): 1300–1304

Akkaya MS, Shoemaker RC, Specht JE, Bhagwat AA, Cregan PB (1995) Integration of simple sequence repeat DNA markers into a soybean linkage map. Crop Sci **35**(5): 1439–1445

Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, Lorenz A (2015) A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. Plant Genome. doi: 10.3835/plantgenome2015.04.0024

Bernal-Vasquez A-M, Utz H-F, Piepho H-P (2016) Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. Theor Appl Genet **129**: 787–804

Bernard RL (1971) Two major genes for time offlowering and maturity in soybeans. Crop Sci **11**:242–244

Bernard RL (1972) Two genes affecting stem termination in soybeans. Crop Sci **12**:235–239

Bonato ER, Vello NA (1999) *E6*, a dominant gene conditioning early flowering and maturity in soybeans. Genetics and Molecular Biology **22**: 229-232

Bootsma A, Gordon R, Read G, Richards WG (1992) Heat units for corn in the Maritime Provinces. Atl Comm Agrometeorol Publ **92**: 8

Brown D, Bootsma A (1993) Crop Heat Units for Corn and Other Warm Season Crops in Ontario. Ontario Ministry of Agriculture and Food, Toronto. Factsheet No. 93-119, Agdex 111/31

Browning BL, Zhou Y, Browning SR (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet **103**: 338–348

Butler DG, Cullis BR, Gilmour AR, Gogel BJ, Thompson R (2018) ASReml-R Reference Manual Version 4.

Buzzell RI (1971) Inheritance of a soybean flowering response to fluorescent-daylength conditions. Can J Genet Cytol **13**:703–707

Buzzell RI, Voldeng HD (1980) Inheritance of insensitivity to long day length. Soybean Genet Newsl **7**: 26–29

Caldwell BE, Howell RW (1973) Soybeans: Improvement, production, and uses. American Society of Agronomy Madison

Cai Y, Wang L, Chen L, Wu T, Liu L, Sun S, Wu C, Yao W, Jiang B, Yuan S, Han T (2020) Mutagenesis of *GmFT2a* and *GmFT5a* mediated by CRISPR/Cas9 contributes for expanding the regional adaptability of soybean. Plant Biotechnol J **18**(1): 298-309.

Cao D, Takeshima R, Zhao C, Liu B, Jun A, Kong F (2017) Molecular mechanisms of fowering under long days and stem growth habit in soybean. J Exp Bot **68**: 1873–1884

Chen L, Nan H, Kong L, Yue L, Yang H, Zhao Q, Fang C, Li H, Cheng Q, Lu S, Kong F (2020) Soybean *AP1* homologs control flowering time and plant height. Journal of Integrative Plant Biology **62**(12):1868-1879

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**: i884–i890

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) **6**: 80–92

Clough SJ, Tuteja JH, Li M, Marek LF, Shoemaker RC, Vodkin LO (2004) Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of CHS genes comprising the *I* locus. Genome **47**: 819–831

Cober, ER, Molnar SJ, Charette M, Voldeng HD (2010) A new locus for early maturity in soybean. Crop Sci **50**(2):524-527

Cober ER, Stewart DW, Voldeng HD (2001) Photoperiod and Temperature Responses in Early-Maturing, Near-Isogenic Soybean Lines. Crop Sci **41**: 721–727

Cober ER, Voldeng HD (2001) A new soybean maturity and photoperiod-sensitivity locus linked to *E1* and *T*. Crop Sci **41**: 698-701

Cordeiro AM, Andrade L, Monteiro CC, Leitão G, Wigge PA, Saibo NJM (2022) PHYTOCHROME-INTERACTING FACTORS: a promising tool to improve crop productivity. J Exp Bot **73**: 3881–3897

Cregan PB, Hartwig EE (1984) Characterization of Flowering Response to Photoperiod in Diverse Soybean Genotypes. Crop Sci **24**: 659–662

Cregan PB, Jarvik TY, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, VanToai TT, Lohnes DG, Chung J, Specht JE (1999) An integrated genetic linkage map of the soybean genome. Crop Sci **39**: 1464–1490

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al (2011) The variant call format and VCFtools. Bioinformatics **27**: 2156–2158

Dong L, Cheng Q, Fang C, Kong L, Yang H, Hou Z, Li Y, Nan H, Zhang Y, Chen Q, et al (2022) Parallel selection of distinct *Tof5* alleles drove the adaptation of cultivated and wild soybean to high latitudes. Mol Plant **15**(2): 308-321

Dong L, Fang C, Cheng Q, Su T, Kou K, Kong L, Zhang C, Li H, Hou Z, Zhang Y, et al (2021) Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. Nat Commun **12**: 5445

Fang C, Liu J, Zhang T, Su T, Li S, Cheng Q, Kong L, Li X, Bu T, Li H, et al (2021) A recent retrotransposon insertion of *J* caused *E6* locus facilitating soybean adaptation into low latitude. J Integr Plant Biol **63**: 995–1003

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**: 1–14

FAOSTAT (2021) FAOSTAT statistical database. Food and Agriculture Organization of the United Nations, https://www.fao.org/faostat/en/#home

Grant D, Nelson RT, Cannon SB, Shoemaker RC (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res **38**: D843–D846

Hahn V, Würschum T (2014) Molecular genetic characterization of Central European soybean breeding germplasm. Plant Breed **133**: 748–755

Hesketh JD, Myhre DL, Willey CR (1973) Temperature Control of Time Intervals Between Vegetative and Reproductive Events in Soybeans. Crop Sci **13**: 250–254

Jia H, Jiang B, Wu C, Lu W, Hou W, Sun S, Yan H, Han T (2014) Maturity group classification and maturity locus genotyping of early-maturing soybean varieties from high-latitude cold regions. PLoS One **9**: 1–9

Jung CH, Wong CE, Singh MB, Bhalla PL (2012) Comparative genomic analysis of soybean flowering genes. PLoS One, **7**(6): e38250

Jung JH, Domijan M, Klose C, Biswas S, Ezer D, Gao M, Khattak AK, Box MS, Charoensawan V, Cortijo S, et al (2016) Phytochromes function as thermosensors in Arabidopsis. Science **354**: 886–889

Kanazawa A, Liu B, Kong F, Arase S, Abe J (2009) Adaptive Evolution Involving Gene Duplication and Insertion of a Novel Ty1/copia-Like Retrotransposon in Soybean. J Mol Evol **69**: 164–175

Kasai A, Kasai K, Yumoto S, Senda M (2007) Structural features of *GmIRCHS*, candidate of the *I* gene inhibiting seed coat pigmentation in soybean: implications for inducing endogenous RNA silencing of chalcone synthase genes. Plant Mol Biol **64**: 467–479

Kong F, Nan H, Cao D, Li Y, Wu F, Wang J, Lu S, Yuan X, Cober ER, Abe J, et al (2014) A new dominant gene *E9* conditions early flowering and maturity in soybean. Crop Sci **54**: 2529–2535

Kou K, Yang H, Li H, Fang C, Chen L, Yue L, Nan H, Kong L, Li X, Wang F, et al (2022) A functionally divergent SOC1 homolog improves soybean yield and latitudinal adaptation, Current Biology **32**: 1728–1742

Kurasch AK, Hahn V, Leiser WL, Vollmann J, Schori A, Bétrix CA, Mayr B, Winkler J, Mechtler K, Aper J, et al (2017) Identification of mega-environments in Europe and effect of allelic variation at maturity *E* loci on adaptation of European soybean. Plant Cell Environ **40**: 765–778

Langewisch T, Zhang H, 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**(4): e94150

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**: 2987–2993

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1000 Genome Project Data Processing (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079

Li M, Liu Y, Tao Y, Xu C, Li X, Zhang X, Han Y, Yang X, Sun J, Li W, et al (2019) Identification of genetic loci and candidate genes related to soybean flowering through genome wide association study. BMC Genomics **20**: 1–13

Li X, Fang C, Yang Y, Lv T, Su T, Chen L, Nan H, Li S, Zhao X, Lu S, Dong L (2021) Overcoming the genetic compensation response of soybean florigens to improve adaptation and yield at low latitudes. Current Biology **31**(17): 3755-3767

Lin X, Liu B, Weller JL, Abe J, Kong F (2021) Molecular mechanisms for the photoperiodic regulation of flowering in soybean. J Integr Plant Biol **63**: 981–994

Liu B, Kanazawa A, Matsumura H, Takahashi R, Harada K, Abe J (2008) Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. Genetics **180**: 995–1007

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

Liu L, Song W, Wang L, Sun X, Qi Y, Wu T, Sun S, Jiang B, Wu C, Hou W, et al (2020) Allele combinations of maturity genes *E1-E4* affect adaptation of soybean to diverse geographic regions and farming systems in China. PLoS One **15**: 1–15

Lü J, Suo H, Yi R, Ma Q, Nian H (2015) Glyma11g13220, a homolog of the vernalization pathway gene *VERNALIZATION 1* from soybean [*Glycine max* (L.) Merr.], promotes flowering in Arabidopsis thaliana. BMC Plant Biol **15**: 1–12

Lu S, Dong L, Fang C, Liu S, Kong L, Cheng Q, Chen L, Su T, Nan H, Zhang D, et al (2020) Stepwise selection on homeologous *PRR* genes controlling flowering and maturity during soybean domestication. Nat Genet **52**: 428–436

Lu S, Zhao X, Hu Y, Liu S, Nan H, Li X, Fang C, Cao D, Shi X, Kong L, et al (2017) Natural variation at the soybean *J* locus improves adaptation to the tropics and enhances yield. Nat Genet **49**: 773–779

Lyu J, Cai Z, Li Y, Suo H, Yi R, Zhang S, Nian H (2020) The Floral Repressor *GmFLC-like* Is Involved in Regulating Flowering Time Mediated by Low Temperature in Soybean. Int J Mol Sci **21**: 1322

Mao T, Li J, Wen Z, Wu T, Wu C, Sun S, Jiang B, Hou W, Li W, Song Q, et al (2017) Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. BMC Genomics **18**: 1–17

de Mendiburu F (2021) R - Package agricolae.

Michaels SD, Bezerra IC, Amasino RM (2004) FRIGIDA-related genes are required for the winter-annual habit in Arabidopsis. Proc Natl Acad Sci USA **101**: 3281–3285

Molnar SJ, Rai S, Charette M, Cober ER (2003) Simple sequence repeat (SSR) markers linked to *E1*, *E3*, *E4*, and *E7* maturity genes in soybean. Genome **46**(6):1024–1036

Nan H, Cao D, Zhang D, Li Y, Lu S, Tang L, Yuan X, Liu B, Kong F (2014) GmFT2a and GmFT5a redundantly and differentially regulate flowering through interaction with and upregulation of the bZIP transcription factor GmFDL19 in soybean. PoS One, 9(5), p.e97669

Piepho HP, Möhring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. Genetics **177**: 1881–1888

Ping J, Liu Y, Sun L, Zhao M, Li Y, She M, Sui Y, Lin F, Liu X, Tang Z, et al (2014) *Dt2* is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. Plant Cell **26**: 2831–2842

Postman J, Hummer K, Ayala-Silva T, Bretting P, Franko T, Kinard G, Bohning M, Emberland G, Sinnott Q, Mackay M, et al (2010) GRIN-GLOBAL: AN INTERNATIONAL PROJECT TO DEVELOP A GLOBAL PLANT GENEBANK INFORMATION MANAGEMENT SYSTEM. Acta Hortic. International Society for Horticultural Science (ISHS), Leuven, Belgium, pp 49–55

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet **81**: 559–575

Ray JD, Hinson K, Mankono EB, Malo FM (1995) Genetic control of a long-juvenile trait in soybean. Crop Sci **35**:1001–1006

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol **29**: 24–26

Samanfar B, Molnar SJ, Charette M, Schoenrock A, Dehne F, Golshani A, Belzile F, Cober ER (2017) Mapping and identification of a potential candidate gene for a novel maturity locus, *E10*, in soybean. Theor Appl Genet **130**: 377–390

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature **463**: 178–183

Shoemaker RG, Specht JE (1995) Integration of the soybean molecular and classical linkage groups. Crop Sci **35**(2):436–446

Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F (2015) Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol J **13**: 211–221

Song Q, Hyten DL, Jia G, Quigley C V, Fickus EW, Nelson RL, Cregan PB (2015) Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3 Genes|Genomes|Genetics **5**: 1999–2006

Steel RGD, Torrie JH, Dickey DA (1997) Principles and Procedures of Statistics: A Biometrical Approach. McGraw-Hill

Tian Z, Wang X, Lee R, Li Y, Specht JE, Nelson RL, McClean PE, Qiu L, Ma J (2010) Artificial selection for determinate growth habit in soybean. Proc Natl Acad Sci USA **107**: 8563–8568

Toda K, Yang D, Yamanaka N, Watanabe S, Harada K, Takahashi R (2002) A single-base deletion in soybean flavonoid 3′-hydroxylase gene is associated with gray pubescence color. Plant Mol Biol **50**: 187–196

Tsubokura Y, Matsumura H, Xu M, Liu B, Nakashima H, Anai T, Kong F, Yuan X, Kanamori H, Katayose Y, et al (2013) Genetic variation in soybean at the maturity locus *E4* is involved in adaptation to long days at high latitudes. Agronomy **3**: 117–134

Wang J, Zhang Z (2021) GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. Genomics Proteomics Bioinformatics. doi: https://doi.org/10.1016/j.gpb.2021.08.005

Watanabe S, Hideshima R, Zhengjun X, Tsubokura Y, Sato S, Nakamoto Y, Yamanaka N, Takahashi R, Ishimoto M, Anai T, et al (2009) Map-based cloning of the gene associated with the soybean maturity locus *E3*. Genetics **182**: 1251–1262

Watanabe S, Xia Z, Hideshima R, Tsubokura Y, Sato S, Yamanaka N, Takahashi R, Anai T, Tabata S, Kitamura K, et al (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

Wigge PA (2013) Ambient temperature signalling in plants. Curr Opin Plant Biol **16**: 661–666

Wu T, Li J, Wu C, Sun S, Mao T, Jiang B, Hou W, Han T (2015) Analysis of the independent- and interactive-photo-thermal effects on soybean flowering. J Integr Agric **14**: 622–632

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

Xu M, Xu Z, Liu B, Kong F, Tsubokura Y, Watanabe S, Xia Z, Harada K, Kanazawa A, Yamada T, et al (2013) Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. BMC Plant Biol **13**: 91

Xu X, Zhang L, Cao X, Liu L, Jiang B, Zhang C, Jia H, Lyu X, Su Y, Cai Y, Liu L (2021) Cotyledons facilitate the adaptation of early‐maturing soybean varieties to high‐latitude long‐day environments. Plant Cell Env **44**(8): 2551–2564

Yamamoto E, Matsunaga H (2021). Exploring efficient linear mixed models to detect squantitative trait locus-by-environment interactions. G3, *11*(8), jkab119.

Yue Y, Liu N, Jiang B, Li M, Wang H, Jiang Z, Pan H, Xia Q, Ma Q, Han T, et al (2017) A Single Nucleotide Deletion in *J* Encoding GmELF3 Confers Long Juvenility and Is

Associated with Adaption of Tropic Soybean. Mol Plant **10**: 656–658

Zabala G, Vodkin LO (2007) A Rearrangement Resulting in Small Tandem Repeats in the F3′5′H Gene of White Flower Genotypes Is Associated with the Soybean *W1* Locus. Crop Sci **47**: S-113-S-124

Zhai H, Lü S, Liang S, Wu H, Zhang X, Liu B, Kong F, Yuan X, Li J, Xia Z (2014) *GmFT4*, a homolog of *FLOWERING LOCUS T*, is positively regulated by *E1* and functions as a flowering repressor in soybean. PLoS One **9**(2): e89030

Zhang J, Song Q, Cregan PB, Nelson RL, Wang X, Wu J, Jiang G-L (2015) Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. BMC Genomics **16**: 217

Zhang L xin, Liu W, Tsegaw M, Xu X, Qi Y ping, Sapey E, Liu L ping, Wu T ting, Sun S, Han T fu (2020) Principles and practices of the photo-thermal adaptability improvement in soybean. J Integr Agric **19**: 295–310

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**(2): 256–282

Zhao C, Takeshima R, Zhu J, Xu M, Sato M, Watanabe S, Kanazawa A, Liu B, Kong F, Yamada T, et al (2016) A recessive allele for delayed flowering at the soybean maturity locus *E9* is a leaky allele of *FT2a*, a *FLOWERING LOCUS T* ortholog. BMC Plant Biol **16**: 1–15

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics **28**: 3326–3328

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

Figure legends

Figure 1 Soybean genetic diversity. (a) Principle component analysis (PCA) based on genome-wide molecular markers for the diversity panel used in this study (with the subset of resequenced lines shown in red) and ~17,000 USDA soybean accessions (except those included in the diversity panel). (b) USDA accessions classified according to maturity groups (MGs). (c) All genotypes classified based on their origin.

Figure 2 Environmental plasticity of flowering time and maturity. Histograms showing days to flowering (red) and maturity (blue) across environments or at each single environment. The dashed vertical black lines indicate the trait means. The red line shows the accumulated crop heat units (CHU) in each environment since the sowing date.

Figure 3 Results of genome-wide association mapping in the diversity panel for days to flowering (top) and maturity (below) across environments. The red and black points indicate the significantly associated markers with an explained proportion of genotypic variance ($p_{G\text{-}joint}$) larger or smaller than 5 %, respectively. The horizontal solid gray line shows the Bonferroni-corrected significance threshold ($P < 0.05$). The positions of the maturity loci *E1-E4* and *J*, the two stem growth habit genes *Dt1* and *Dt2*, and the homolog of *FRIGIDA-like 1* (*GmFRL1*, *Glyma.08G325700*) are indicated by vertical lines.

Figure 4 Results of genome-wide association mapping for QTL-by-environment interactions (a) for days to flowering in the whole diversity panel (1,503 lines) and (b) for maturity in a subset of the panel (1,233 lines) fixed for *E2*. The horizontal gray line shows the Bonferroni-corrected significance threshold ($P < 0.05$). (c) Average flowering time and maturity of the two allelic groups of the most significant SNP of the two QTL shown for each environment.

Figure 5 Identification of novel alleles and candidate genes. (a-c) *E4* variants identified in the resequencing panel. (a) Five detected *E4* alleles. (b) Maturity group and origin information of the 11 genotypes carrying the novel *e4-par* variant. $^\$$ European cultivar classification is shown in brackets. (c) A single base deletion of *e4-par* produces a truncated protein with 852 amino acids. (d-e) Identification of *GmFRL1* as a candidate gene. (d) Fine-mapping of *GmFRL1* on chromosome 8 based on the association mapping result for days to flowering (top) and maturity

(below) across environments. (e) Alleles of *GmFRL1* and three haplotypes identified in the resequencing panel. (f) Fine-mapping of the QTL-by-environment interaction on chromosome 2 and identification of *GmAP1d* as a candidate gene.

Figure 6 Effect of identified loci on adaptation. (a) Boxplots of days to flowering and maturity across environments based on the allelic groups of candidate genes (*E1*, *E2*, *E3*, *E4*, *Dt2*, *GmFRL1*) in the resequencing panel and two major-effect QTL (qFT28, qMT2) in the diversity panel. Letters above the boxplots show the result of multiple group comparisons and different letters indicate significant difference. (b) Frequency of *E1* variants in the two allelic groups (C/T) of the major QTL for maturity, qMT4 (marker ss715593464), and boxplots for days to flowering and maturity within each allelic group in the resequencing panel. (c) Stacking of QTL for days to flowering and maturity based on 11 haplotype groups determined by *E1*, *E2* and two major-effect QTL (qFT28, qMT2) in the resequencing panel. Alleles which delay flowering or maturity are colored in orange.

Figure 7 Maturity group classification. (a) Boxplots of days to flowering and maturity in the diversity panel and (b) the resequencing panel based on the USDA maturity group classification. In (b) the allelic state at *E1* is shown for the individuals in each group. Letters show the result of multiple group comparisons and different letters indicate significant difference. MG, maturity group. (c) Allele frequencies of different loci in the resequencing panel shown for four maturity groups.

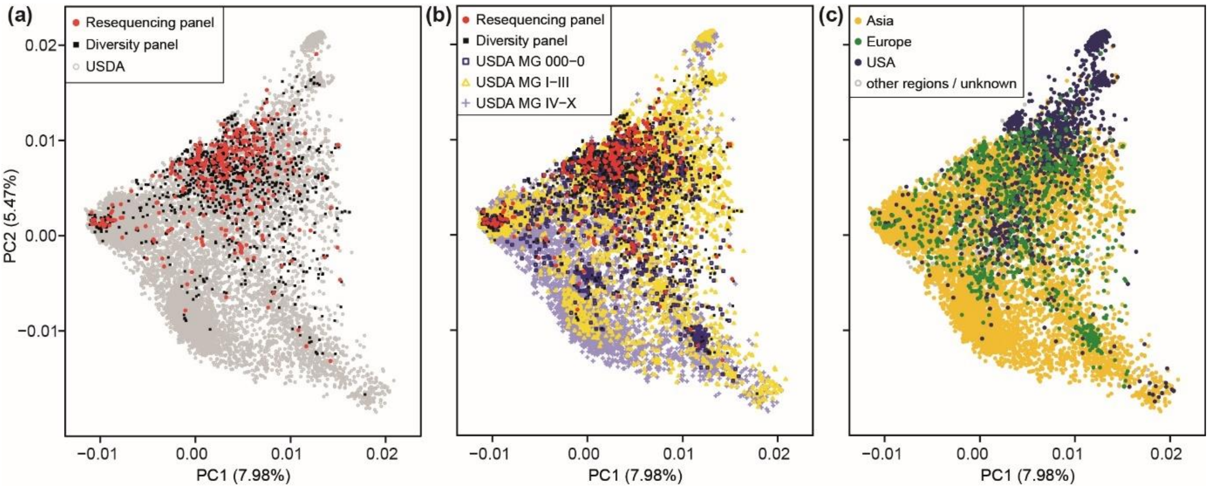


Figure 1 Soybean genetic diversity. (a) Principle component analysis (PCA) based on genome-wide molecular markers for the diversity panel used in this study (with the subset of resequenced lines shown in red) and ~17,000 USDA soybean accessions (except those included in the diversity panel). (b) USDA accessions classified according to maturity groups (MGs). (c) All genotypes classified based on their origin.

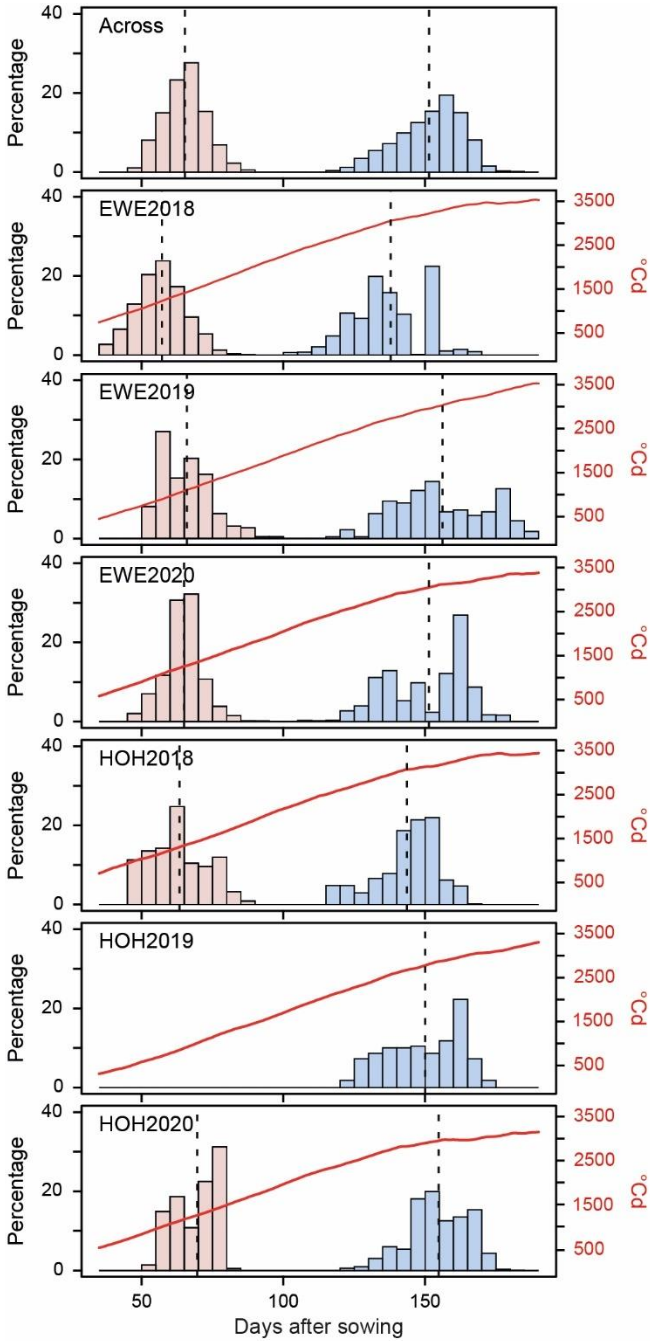


Figure 2 Environmental plasticity of flowering time and maturity. Histograms showing days to flowering (red) and maturity (blue) across environments or at each single environment. The dashed vertical black lines indicate the trait means. The red line shows the accumulated crop heat units (CHU) in each environment since the sowing date.

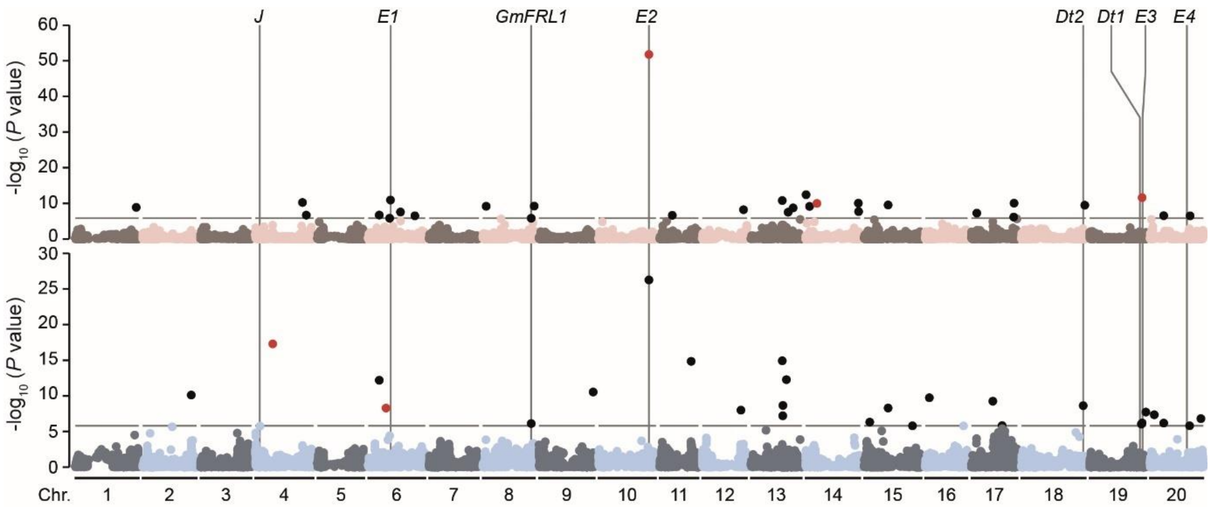


Figure 3 Results of genome-wide association mapping in the diversity panel for days to flowering (top) and maturity (below) across environments. The red and black points indicate the significantly associated markers with an explained proportion of genotypic variance ($p_{G\text{-}joint}$) larger or smaller than 5 %, respectively. The horizontal solid gray line shows the Bonferroni-corrected significance threshold ($P < 0.05$). The positions of the maturity loci *E1*-*E4* and *J*, the two stem growth habit genes *Dt1* and *Dt2*, and the homolog of *FRIGIDA-like 1* (*GmFRL1*, *Glyma.08G325700*) are indicated by vertical lines.

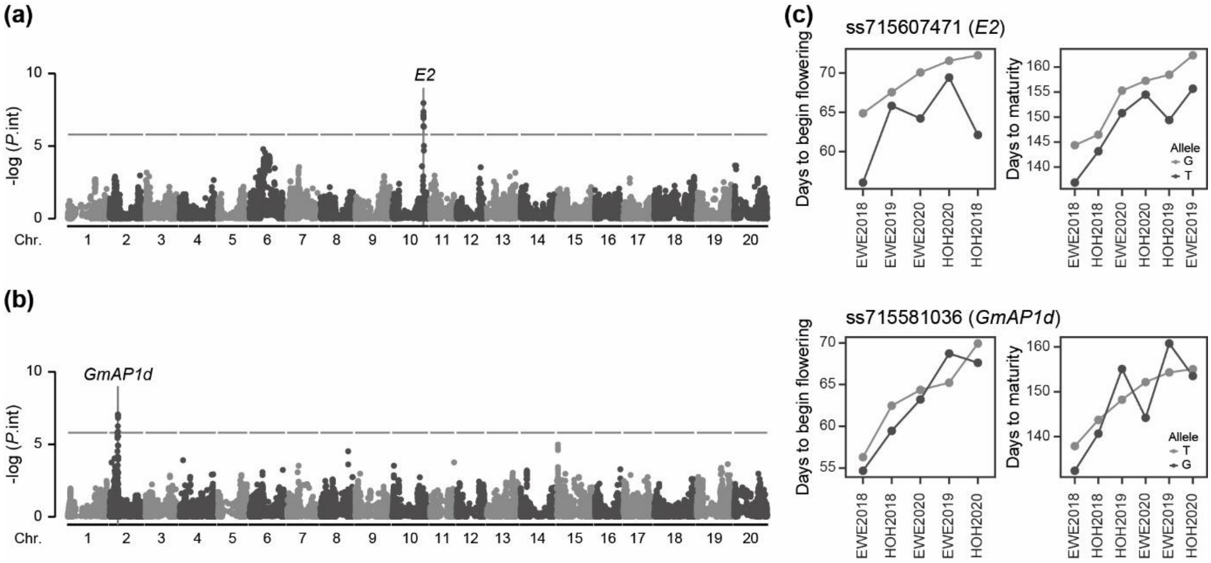
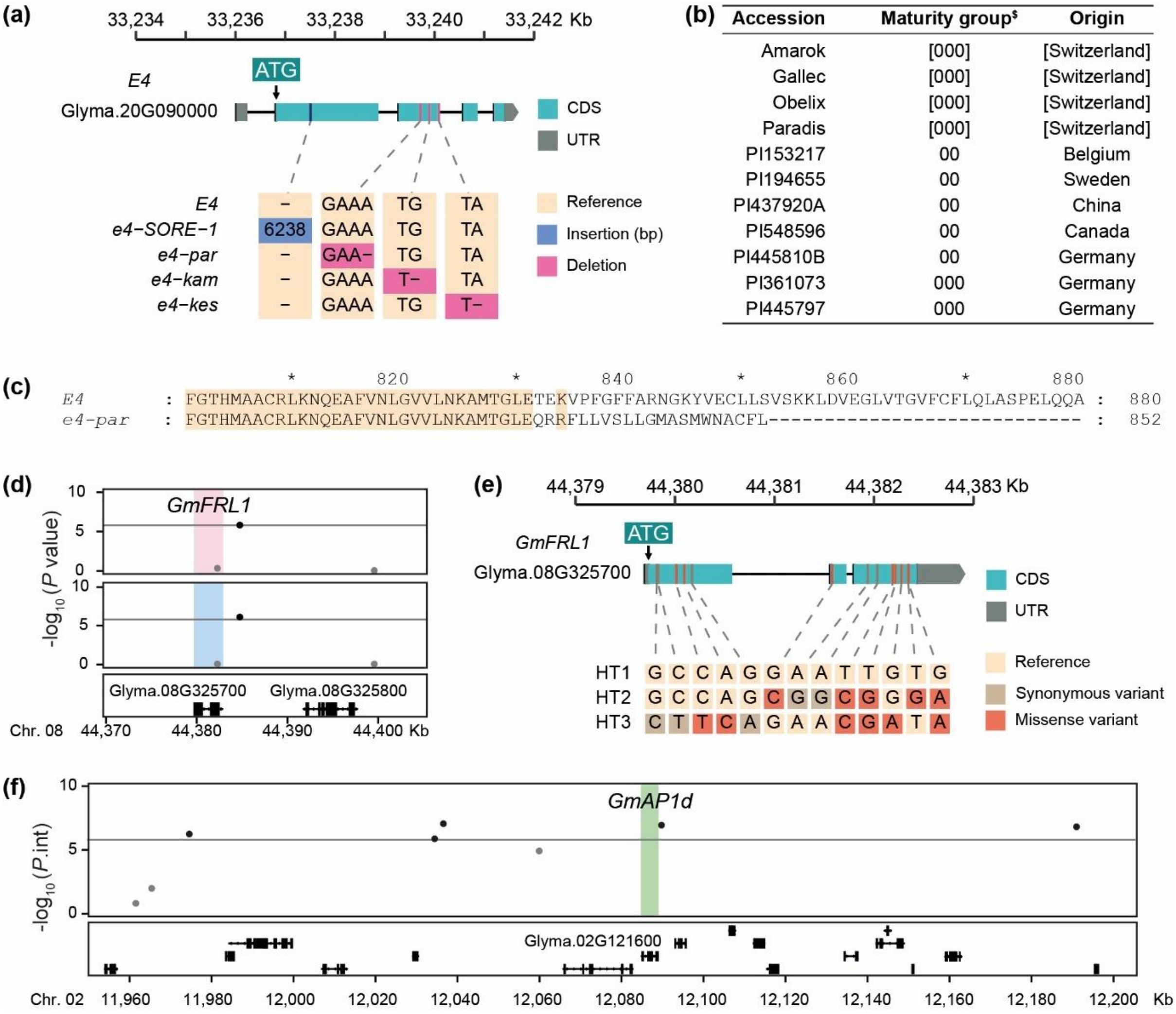


Figure 4 Results of genome-wide association mapping for QTL-by-environment interactions (a) for days to flowering in the whole diversity panel (1,503 lines) and (b) for maturity in a subset of the panel (1,233 lines) fixed for *E2*. The horizontal gray line shows the Bonferroni-corrected significance threshold ($P < 0.05$). (c) Average flowering time and maturity of the two allelic groups of the most significant SNP of the two QTL shown for each environment.



Accession	Maturity group$	Origin
Amarok	[000]	[Switzerland]
Gallec	[000]	[Switzerland]
Obelix	[000]	[Switzerland]
Paradis	[000]	[Switzerland]
PI153217	00	Belgium
PI194655	00	Sweden
PI437920A	00	China
PI548596	00	Canada
PI445810B	00	Germany
PI361073	000	Germany
PI445797	000	Germany

Figure 5 Identification of novel alleles and candidate genes. (a-c) *E4* variants identified in the resequencing panel. (a) Five detected *E4* alleles. (b) Maturity group and origin information of the 11 genotypes carrying the novel *e4-par* variant. $ European cultivar classification is shown in brackets. (c) A single base deletion of *e4-par* produces a truncated protein with 852 amino acids. (d-e) Identification of *GmFRL1* as a candidate gene. (d) Fine-mapping of *GmFRL1* on chromosome 8 based on the association mapping result for days to flowering (top) and maturity (below) across environments. (e) Alleles of *GmFRL1* and three haplotypes identified in the resequencing panel. (f) Fine-mapping of the QTL-by-environment interaction on chromosome 2 and identification of *GmAP1d* as a candidate gene.

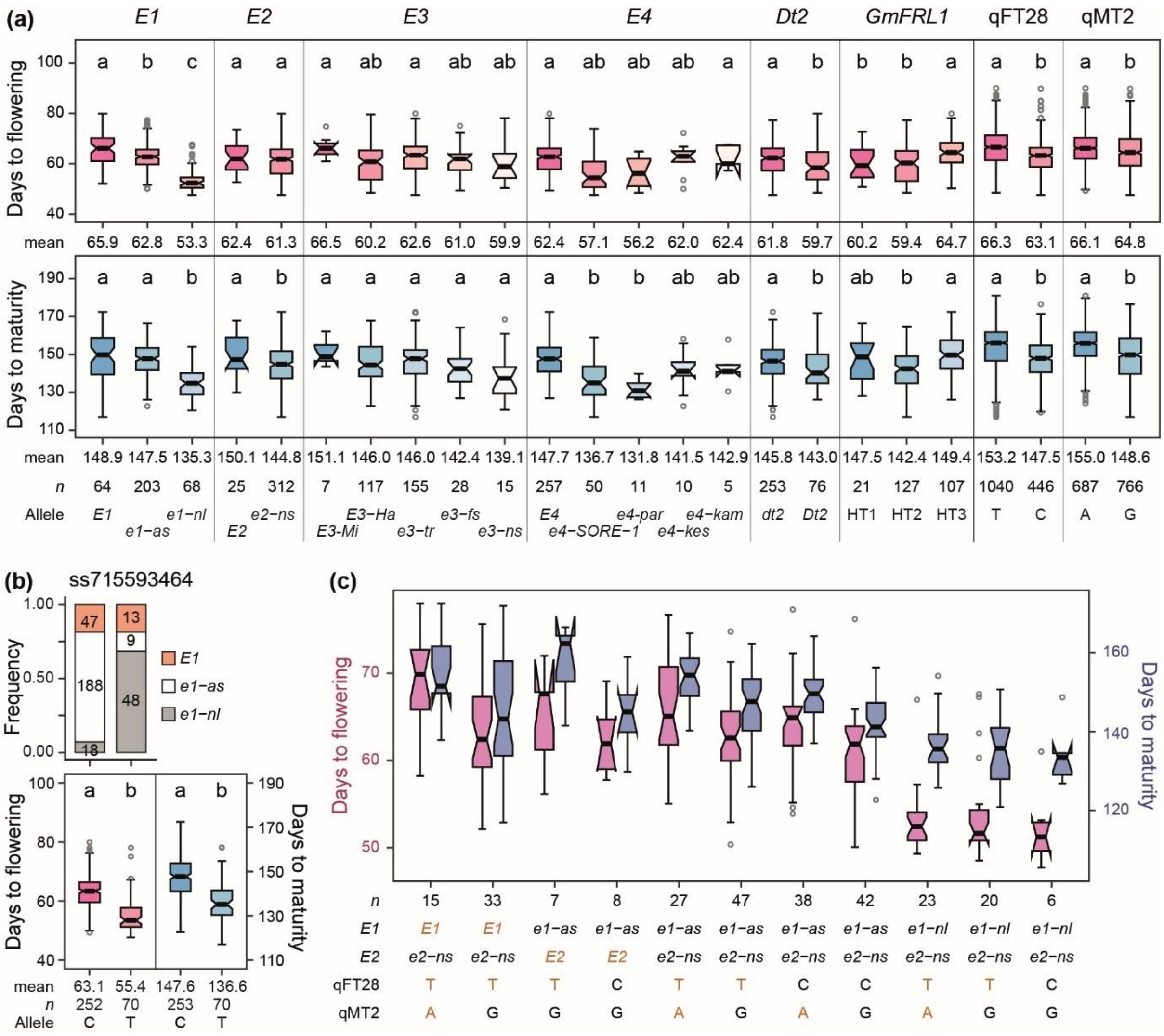


Figure 6 Effect of identified loci on adaptation. (a) Boxplots of days to flowering and maturity across environments based on the allelic groups of candidate genes (*E1*, *E2*, *E3*, *E4*, *Dt2*, *GmFRL1*) in the resequencing panel and two major-effect QTL (qFT28, qMT2) in the diversity panel. Letters above the boxplots show the result of multiple group comparisons and different letters indicate significant difference. (b) Frequency of *E1* variants in the two allelic groups (C/T) of the major QTL for maturity, qMT4 (marker ss715593464), and boxplots for days to flowering and maturity within each allelic group in the resequencing panel. (c) Stacking of QTL for days to flowering and maturity based on 11 haplotype groups determined by *E1*, *E2* and two major-effect QTL (qFT28, qMT2) in the resequencing panel. Alleles which delay flowering or maturity are colored in orange.

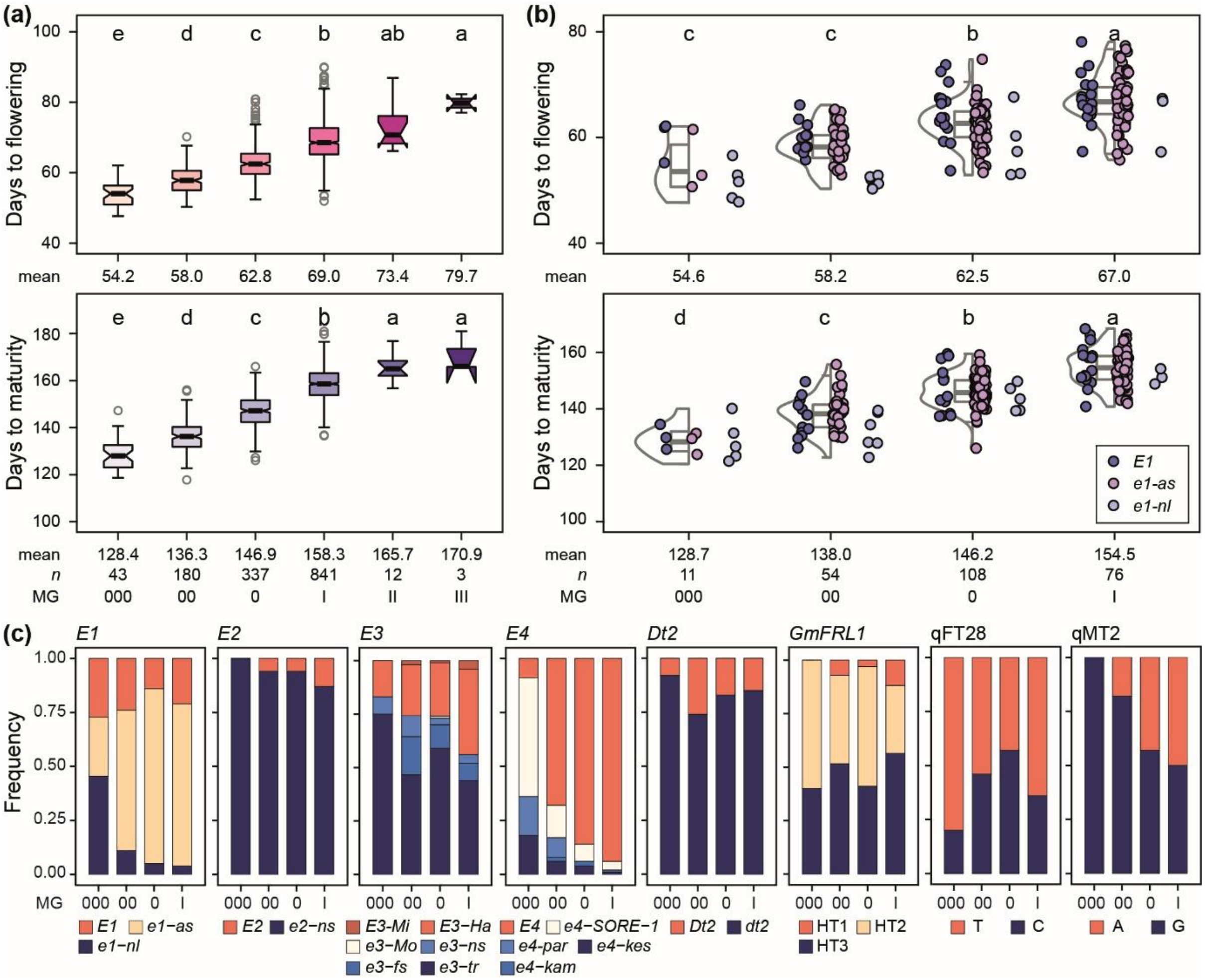


Figure 7 Maturity group classification. (a) Boxplots of days to flowering and maturity in the diversity panel and (b) the resequencing panel based on the USDA maturity group classification. In (b) the allelic state at *E1* is shown for the individuals in each group. Letters show the result of multiple group comparisons and different letters indicate significant difference. MG, maturity group. (c) Allele frequencies of different loci in the resequencing panel shown for four maturity groups.

The genetic architecture of soybean photothermal adaptation to high latitudes

Xintian Zhu[1,2], Willmar L. Leiser[2], Volker Hahn[2], Tobias Würschum[1]

[1] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, 70599 Stuttgart, Germany

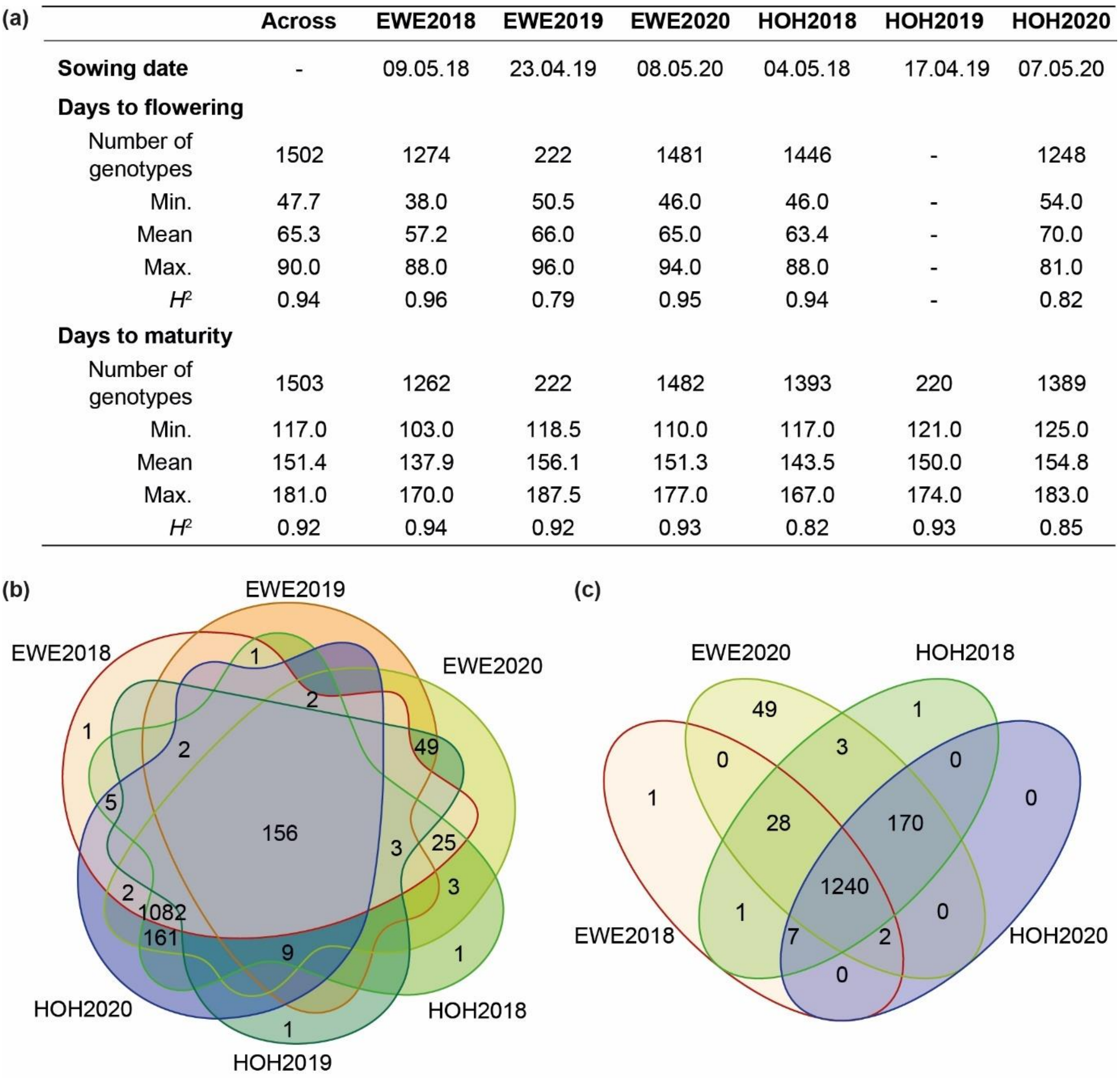
[2] State Plant Breeding Institute, University of Hohenheim, 70599 Stuttgart, Germany

Supporting information

Table S1 Information on the resequencing panel and the diversity panel.

Table S3 Results from genome-wide association mapping for days to begin of flowering and maturity at each single environment.

(not published data)



(a)

	Across	EWE2018	EWE2019	EWE2020	HOH2018	HOH2019	HOH2020
Sowing date	-	09.05.18	23.04.19	08.05.20	04.05.18	17.04.19	07.05.20
Days to flowering							
Number of genotypes	1502	1274	222	1481	1446	-	1248
Min.	47.7	38.0	50.5	46.0	46.0	-	54.0
Mean	65.3	57.2	66.0	65.0	63.4	-	70.0
Max.	90.0	88.0	96.0	94.0	88.0	-	81.0
H^2	0.94	0.96	0.79	0.95	0.94	-	0.82
Days to maturity							
Number of genotypes	1503	1262	222	1482	1393	220	1389
Min.	117.0	103.0	118.5	110.0	117.0	121.0	125.0
Mean	151.4	137.9	156.1	151.3	143.5	150.0	154.8
Max.	181.0	170.0	187.5	177.0	167.0	174.0	183.0
H^2	0.92	0.94	0.92	0.93	0.82	0.93	0.85

Figure S1 (a) Sowing date and statistical summary for days to flowering and maturity across environments and for each single environment. (b) Venn diagrams for accessions grown in all six environments. (c) Venn diagrams for accessions grown in four environments (EWE2018, EWE2020, HOH2018, and HOH2020).

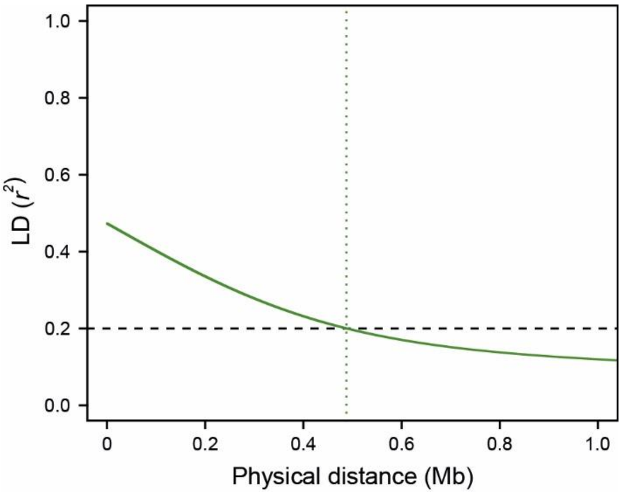


Figure S2 Decay of linkage disequilibrium (LD) with physical distance estimated in the diversity panel used in this study for genome-wide association mapping. The dashed horizonal line shows the threshold of 0.2 and the vertical dotted line indicates the distance after which LD decays below this threshold.

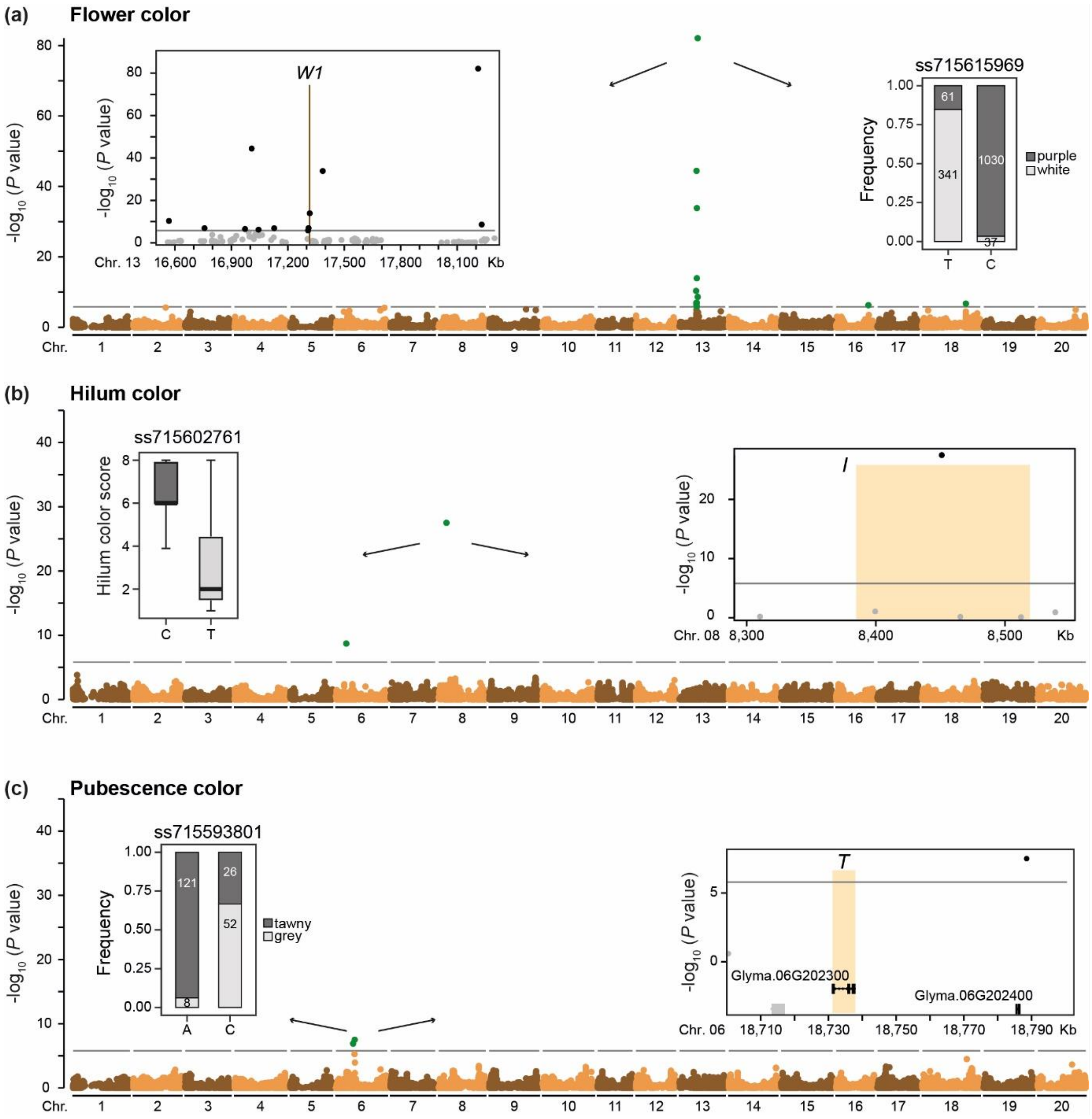


Figure S3 (a) Genome-wide association mapping for flower color across environments. Fine-mapping of the candidate gene *W1* on chromosome 13 (left inset) and frequency of flower color within allelic groups of the most significant SNP (right inset). For flower color, the 11 significant markers on chromosome 13 spanned an interval of around 1.6 Mb which incorporated the known causal gene *W1* (Glyma.13G072100) encoding flavonoid 3′5′-hydroxylase (F3′5′H) (Zabala and Vodkin, 2007; Sonah et al., 2015; Zhou et al., 2015; Fang et al., 2017). (b) Result for hilum color across environments. Fine-mapping of the candidate gene *I* on chromosome 8 (right inset) and boxplots showing hilum color score for the two allelic groups of the most significant SNP (left inset). For hilum color, the most significant marker was found on chromosome 8 within the causal *chalcone synthase* (CHS) gene cluster region known as *I* locus (Clough et al., 2004; Kasai et al., 2007). (c) Result for pubescence color in HOH2019. Fine-mapping of the candidate gene *T* on chromosome 6 (right inset) and frequency of tawny and grey pubescence within the two allelic groups of the most significant SNP (left inset). Pubescence color is controlled by the *T* locus (Glyma.06G202300) that encodes flavonoid 3'-hydroxylase (F3'H) (Toda et al., 2002) and the most significant marker was about 50 Kb downstream of the *T* locus.

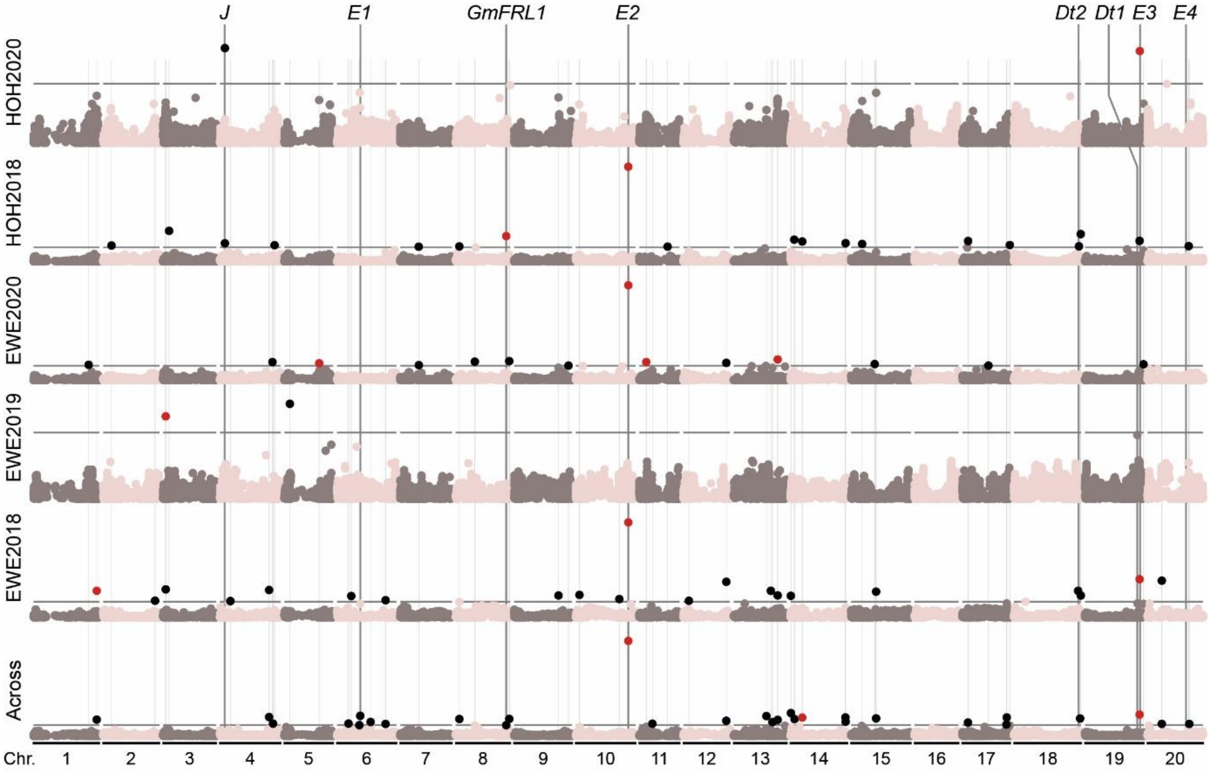


Figure S4 Genome-wide association mapping for days to flowering in each single environment and across environments. The red and black points indicate significant markers with an explained proportion of genotypic variance ($p_{G\text{-}joint}$) > 5 % and < 5 %, respectively. The horizontal solid gray line shows the Bonferroni-corrected significance threshold ($P < 0.05$).

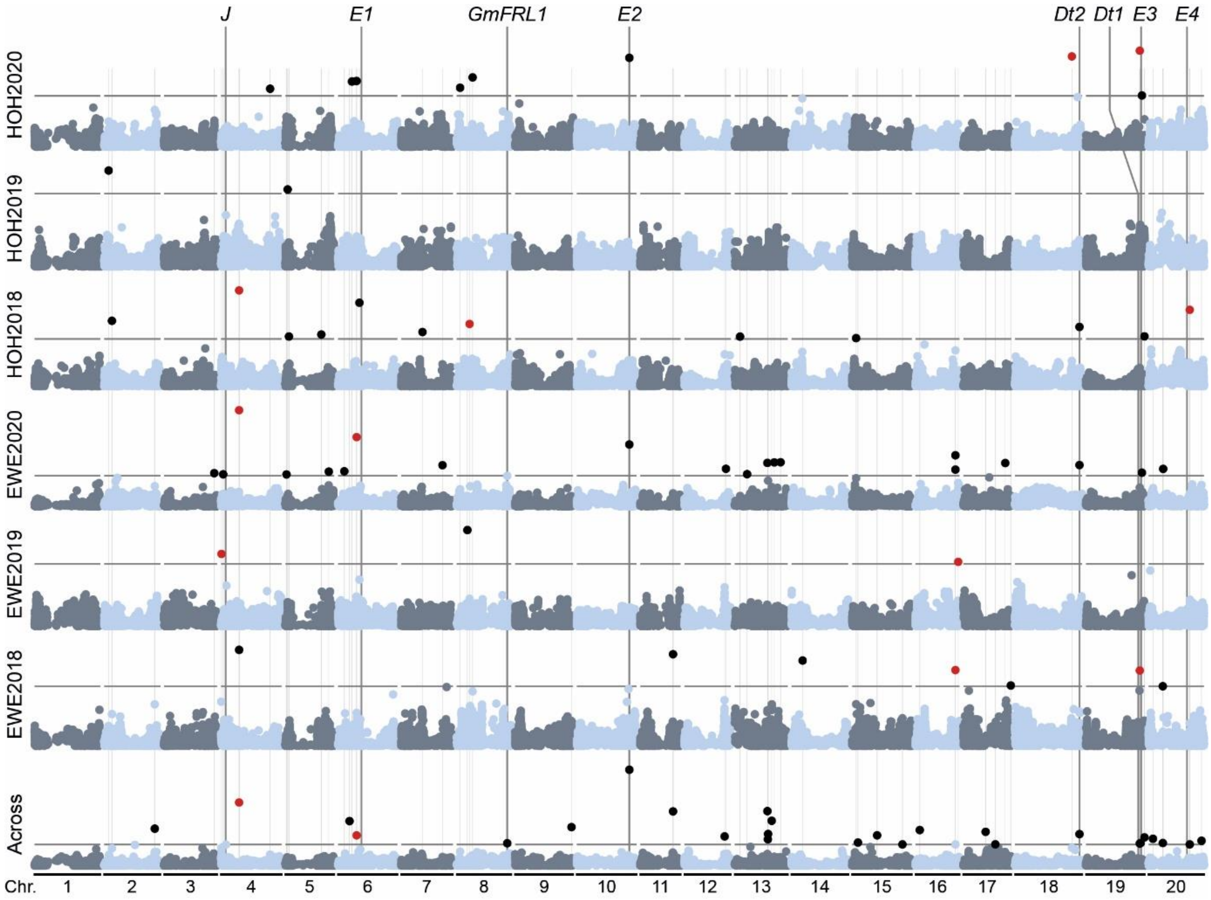


Figure S5 Genome-wide association mapping for days to maturity in each single environment and across environments. The red and black points indicate significant markers with an explained proportion of genotypic variance ($p_{G\text{-}joint}$) > 5 % and < 5 %, respectively. The horizontal solid gray line shows the Bonferroni-corrected significance threshold ($P < 0.05$).

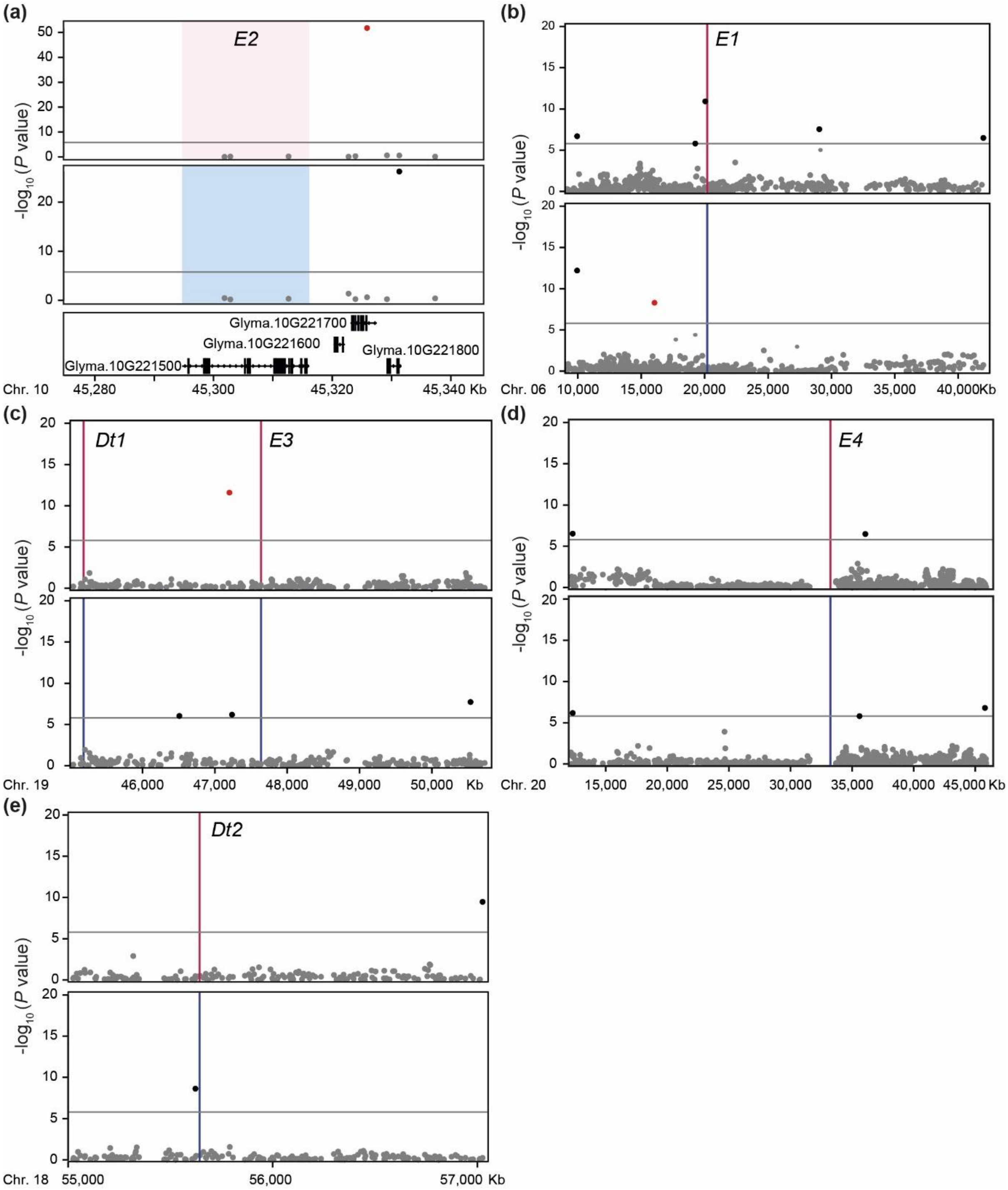


Figure S6 (a-e) Identification of known candidate genes for days to flowering (top) and maturity (below) based on the genome-wide association mapping across environments. The red and black points indicate significant markers with an explained proportion of genotypic variance ($p_{G\text{-}joint}$) > 5 % and < 5 %, respectively. The horizontal solid gray line shows the Bonferroni-corrected significance threshold (P < 0.05).

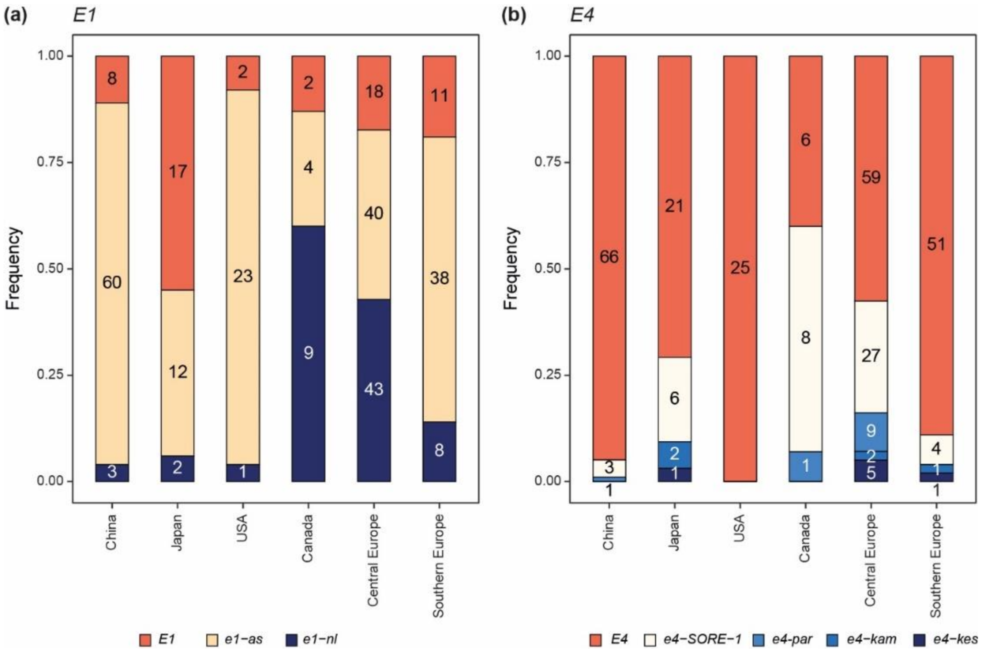


Figure S7 (a) Frequency of polymorphisms of *E1* and (b) *E4* in the resequencing panel separated by the origin of the soybean genotypes. Few heterozygous genotypes are not shown.

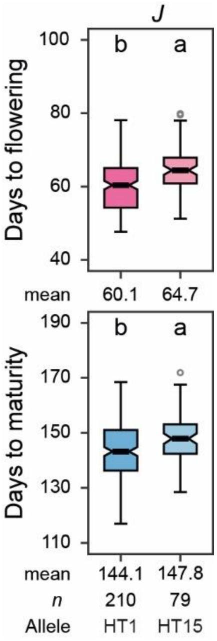
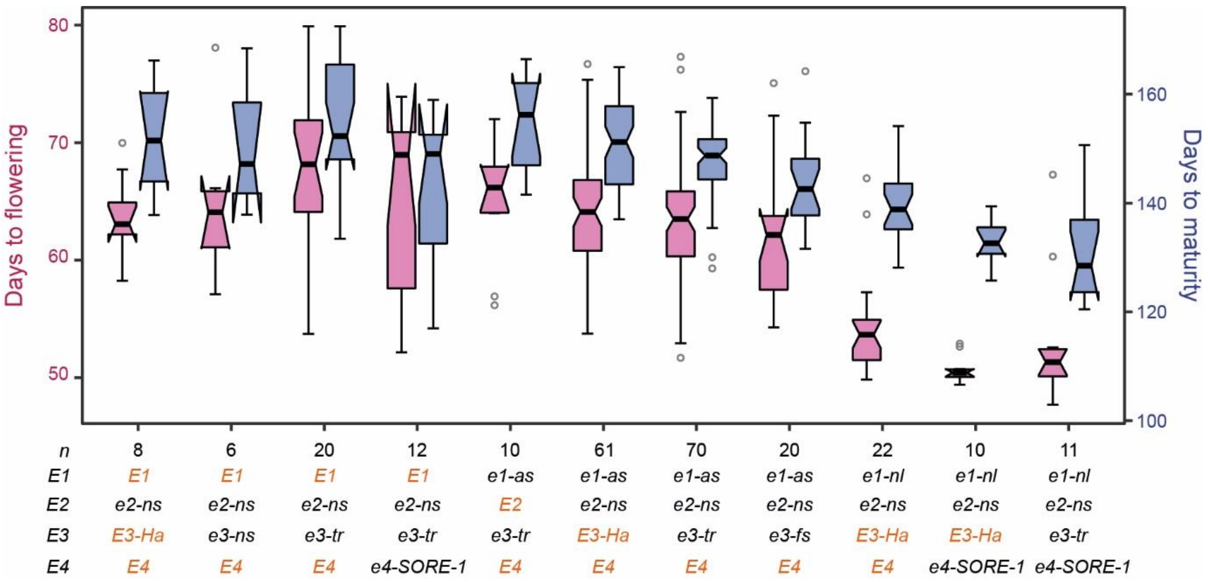


Figure S8 Boxplots of days to flowering and maturity across environments based on two haplotypes, HT1 and HT15, of the candidate gene *J* in the resequencing panel. Letters above boxplots show the result of multiple group comparisons and different letters indicate significant difference.



n	8	6	20	12	10	61	70	20	22	10	11
E1	E1	E1	E1	E1	e1-as	e1-as	e1-as	e1-as	e1-nl	e1-nl	e1-nl
E2	e2-ns	e2-ns	e2-ns	e2-ns	E2	e2-ns	e2-ns	e2-ns	e2-ns	e2-ns	e2-ns
E3	E3-Ha	e3-ns	e3-tr	e3-tr	e3-tr	E3-Ha	e3-tr	e3-fs	E3-Ha	E3-Ha	e3-tr
E4	E4	E4	E4	e4-SORE-1	E4	E4	E4	E4	E4	e4-SORE-1	e4-SORE-1

Figure S9 Stacking of QTL for days to flowering and maturity based on 11 haplotype groups of the maturity loci *E1-E4* in the resequencing panel. Alleles which delay flowering and maturity are colored in orange.

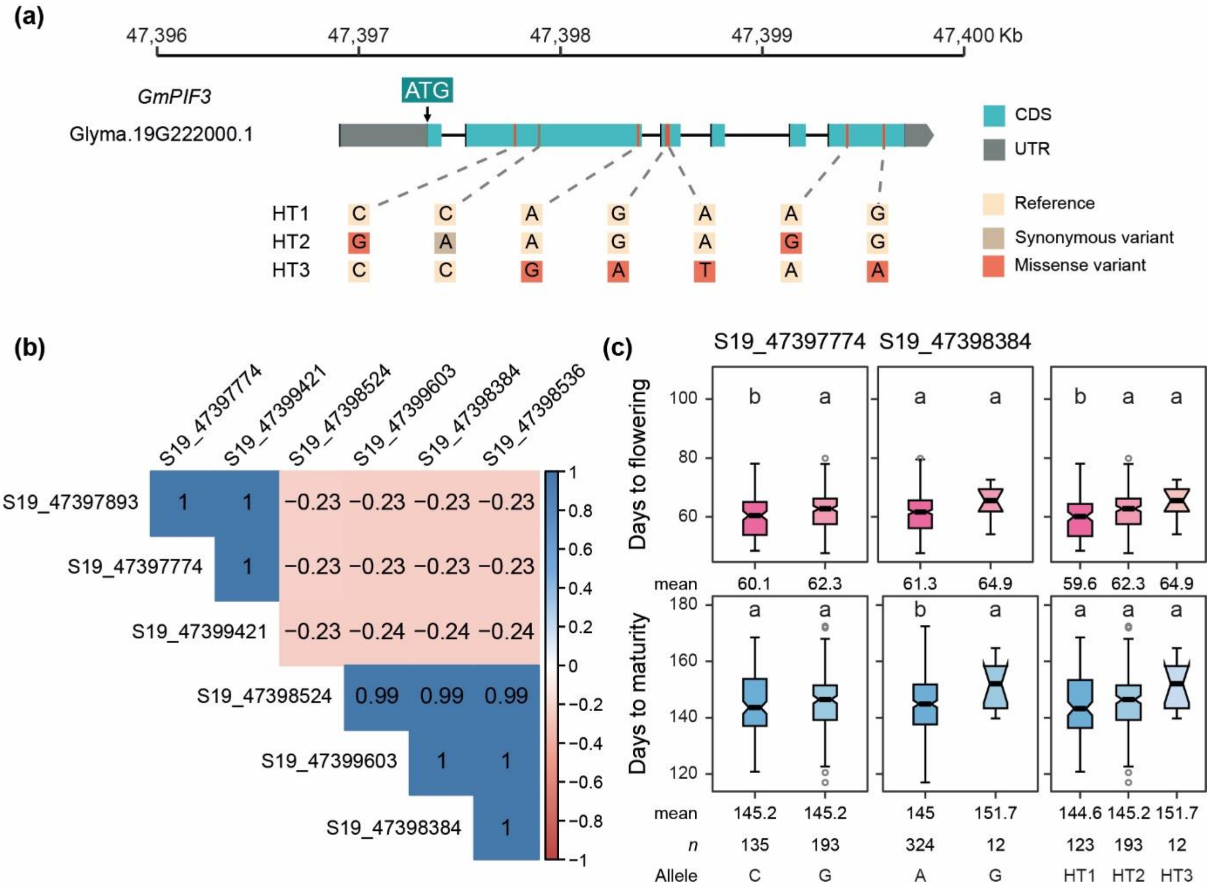


Figure S10 Identification of allelic variation of *GmPIF3* and effect on flowering time and maturity. (a) Alleles of *GmPIF3* and three haplotypes identified in the resequencing panel. (b) Heatmap showing the correlation among alleles of *GmPIF3* in the resequencing panel. Names refer to the position in the reference genome. (c) Boxplots of days to flowering and maturity across environments based on one polymorphism for each of the two groups of linked polymorphisms and the three haplotypes (HT1, HT2 and HT3) of *GmPIF3* in the resequencing panel. Letters above the boxplots show the result of multiple group comparisons and different letters indicate significant difference.

Table S2 Results from genome-wide association mapping for days to begin of flowering and maturity across environments.

QTL	SNP	Chr.	Pos. (bp)	REF /ALT	REF FREQ	-log10 (*P* value)	$p_{G\text{-}Single}$	$p_{G\text{-}Joint}$	Effect
Days to begin flowering									
qFT1	ss715580461	1	54,637,046	G/T	0.55	8.86	12.99	4.46	2.62
qFT2	ss715587931	4	42,521,312	G/A	0.68	10.23	1.20	3.49	-0.85
qFT3	ss715588243	4	45,898,325	C/A	0.85	6.67	1.25	1.39	1.12
qFT4	ss715595707	6	9,946,974	C/T	0.90	6.70	0.38	2.39	0.74
qFT5	ss715593814	6	19,261,720	G/A	0.58	5.81	4.37	3.15	1.53
qFT6	ss715593837	6	20,048,520	A/G	0.93	10.92	0.14	0.35	-0.51
qFT7	ss715594064	6	29,039,651	A/G	0.59	7.55	0.46	1.14	-0.50
qFT8	ss715594403	6	41,974,977	G/T	0.81	6.49	2.05	1.56	-1.30
qFT9	ss715601623	8	3,879,313	T/C	0.92	9.17	3.09	0.67	-2.27
qFT10	ss715602218	8	44,384,746	A/C	0.69	5.82	12.05	0.88	-2.72
qFT11	ss715602474	8	47,038,818	C/T	0.48	9.22	0.27	0.13	0.38
qFT12	ss715607486	10	45,325,872	C/T	0.12	51.79	10.61	10.61	-3.61
qFT13	ss715609447	11	11,214,137	C/A	0.51	6.63	0.55	0.15	0.54
qFT14	ss715612722	12	37,031,186	C/T	0.79	8.21	0.70	0.47	-0.74
qFT15	ss715614589	13	28,249,992	T/C	0.72	10.77	0.04	0.07	-0.16
qFT16	ss715615426	13	33,504,867	A/G	0.34	7.50	0.09	0.84	-0.23
qFT17	ss715616006	13	38,026,151	G/A	0.11	8.70	8.55	1.96	-3.33
qFT18	ss715618372	14	338,515	C/T	0.22	12.41	2.73	1.86	1.44
qFT19	ss715618428	14	3,444,258	C/A	0.50	9.13	7.50	2.27	1.98
qFT20	ss715617372	14	10,094,156	T/C	0.93	9.96	9.03	5.27	4.37
qFT21	ss715619390	14	47,398,552	C/T	0.85	10.03	5.61	3.78	2.37
qFT22	ss715619417	14	47,590,507	A/C	0.87	7.68	5.51	1.91	2.49
qFT23	ss715621278	15	21,715,509	G/A	0.10	9.52	3.03	2.66	-2.11
qFT24	ss715628011	17	5,101,603	T/G	0.94	7.25	0.43	1.23	0.97
qFT25	ss715627445	17	38,309,629	G/A	0.51	6.14	0.58	0.10	0.55
qFT26	ss715627503	17	38,595,915	T/C	0.83	10.03	4.48	3.05	-2.05
qFT27	ss715632413	18	57,025,570	A/C	0.12	9.47	0.31	0.27	0.62
qFT28	ss715635642	19	47,200,665	T/C	0.70	11.59	4.27	5.47	-1.63
qFT29	ss715636763	20	12,321,598	T/C	0.69	6.53	3.82	1.38	1.53
qFT30	ss715637655	20	36,074,213	C/T	0.90	6.49	2.27	0.37	-1.80
Days to maturity									
qMT1	ss715583027	2	43,784,710	A/G	0.74	10.12	0.58	0.69	-0.97
qMT2	ss715587166	4	15,716,751	A/G	0.47	17.30	7.32	9.80	-3.01
qMT3	ss715595707	6	9,946,974	C/T	0.90	12.20	2.49	4.36	2.90
qMT4	ss715593464	6	16,046,818	C/T	0.90	8.29	15.10	8.51	-7.31
qMT5	ss715602218	8	44,384,746	A/C	0.69	6.12	11.38	3.37	-4.05
qMT6	ss715604927	9	48,858,116	T/C	0.42	10.54	5.04	3.32	-2.53
qMT7	ss715607488	10	45,331,299	T/C	0.12	26.27	2.18	2.18	-2.49
qMT8	ss715609985	11	28,206,933	T/C	0.94	14.85	6.44	3.95	-5.85
qMT9	ss715612471	12	34,718,187	A/G	0.82	8.00	2.90	2.90	2.46
qMT10	ss715614589	13	28,249,992	T/C	0.72	14.92	0.59	0.57	0.95

qMT11	ss715614687	13	28,679,294	T/C	0.16	7.21	0.03	0.32	0.25
qMT12	ss715614709	13	28,854,797	C/T	0.81	8.66	5.46	3.81	3.31
qMT13	ss715615231	13	31,977,980	T/C	0.14	12.28	0.00	0.05	0.06
qMT14	ss715622823	15	5,184,552	G/A	0.52	6.31	4.72	0.99	2.41
qMT15	ss715621278	15	21,715,509	G/A	0.10	8.30	0.53	1.82	-1.35
qMT16	ss715621917	15	43,631,120	C/T	0.25	5.80	0.09	0.12	-0.38
qMT17	ss715624797	16	3,559,844	A/C	0.11	9.74	0.52	0.55	1.27
qMT18	ss715626365	17	19,503,588	G/A	0.81	9.26	0.75	0.58	1.21
qMT19	ss715626617	17	27,871,580	C/T	0.81	5.82	0.60	0.40	1.08
qMT20	ss715632223	18	55,622,046	T/G	0.90	8.63	0.43	2.78	-1.21
qMT21	ss715635577	19	46,509,345	A/G	0.72	6.04	6.47	1.13	-3.13
qMT22	ss715635651	19	47,236,627	T/C	0.41	6.19	1.72	1.27	-1.48
qMT23	ss715636052	19	50,534,870	G/A	0.90	7.73	6.43	3.53	-4.67
qMT24	ss715638025	20	3,814,870	G/A	0.67	7.35	8.14	1.67	-3.35
qMT25	ss715636763	20	12,321,598	T/C	0.69	6.18	1.96	0.51	1.67
qMT26	ss715637603	20	35,601,857	A/G	0.90	5.80	17.11	1.13	-7.54
qMT27	ss715638659	20	45,783,064	T/C	0.48	6.80	8.61	2.24	-3.25

Table S4 Results from genome-wide association mapping for QTL-by-environment interactions for days to begin flowering in the whole diversity panel and for maturity in a subset of the panel fixed for *E2*. *P*.int is the P value of the QTL-by-environment interaction effect, *P*.ame of the additive main effect and *P*.all of the total effect.

	SNP	Chr.	Pos. (bp)	*P*.int	*P*.ame	*P*.all
Whole diversity panel						
	ss715607470	10	45,226,484	7.11	11.88	16.98
	ss715607471	10	45,250,482	7.96	14.71	20.54
	ss715607475	10	45,269,968	7.14	16.59	21.50
	ss715607477	10	45,301,855	6.36	14.57	18.77
	ss715607480	10	45,302,838	6.36	13.61	17.86
	ss715607481	10	45,312,644	6.36	13.61	17.86
	ss715607483	10	45,322,752	6.97	16.43	21.18
	ss715607485	10	45,323,915	6.93	16.15	20.88
	ss715607486	10	45,325,872	6.93	16.15	20.88
	ss715607487	10	45,329,231	6.93	16.15	20.88
	ss715607488	10	45,331,299	7.36	13.95	19.20
	ss715607489	10	45,337,346	6.36	13.61	17.86
	ss715607517	10	45,826,997	6.33	9.32	13.76
***E2* subset of diversity panel**						
	ss715581033	2	11,974,580	6.24	0.45	5.94
	ss715581035	2	12,034,409	5.87	0.35	5.53
	ss715581036	2	12,036,555	7.06	0.54	6.79
	ss715581043	2	12,089,749	6.94	0.56	6.69
	ss715581049	2	12,190,975	6.81	0.61	6.60

Table S5 Candidate genes for QTL inferred by screening the neighboring 250 Kb regions of the significant markers.

SNP	Chr.	Pos. (bp)	Environment	Gene	Loci	Distance (~Kb)	Reference
Days to begin flowering							
ss715588228	4	4,307,731	HOH2018	Glyma.04G050200	*J*	+226	Lu et al., 2017
ss715588228	4	4,307,731	HOH2020	Glyma.04G050200	*J*	+226	Lu et al., 2017
ss715588448	4	47,242,528	HOH2018	Glyma.04G202000	*GmLFY1*	-238	Meng et al., 2007
ss715593837	6	20,048,520	Across	Glyma.06G207800	*E1*	-159	Xia et al., 2010
ss715601623	8	3,879,313	Across	Glyma.08G046500	*GmFKF1-2*	+218	Li et al., 2013
ss715601617	8	3,866,379	HOH2018	Glyma.08G046500	*GmFKF1-2*	+205	Li et al., 2013
ss715602218	8	44,384,746	Across	Glyma.08G325700	*GmFRL1*	+2	-
ss715602217	8	44,382,274	HOH2018	Glyma.08G325700	*GmFRL1*	0	-
ss715606507	10	37,596,706	EWE2018	Glyma.10G141400	*GmphyA1*	+101	Liu et al., 2008
ss715607486	10	45,325,872	Across	Glyma.10G221500	*E2*	+10	Watanabe et al., 2011
ss715607488	10	45,331,299	EWE2018	Glyma.10G221500	*E2*	+15	Watanabe et al., 2011
ss715607486	10	45,325,872	EWE2020	Glyma.10G221500	*E2*	+10	Watanabe et al., 2011
ss715607475	10	45,269,968	HOH2018	Glyma.10G221500	*E2*	-25	Watanabe et al., 2011
ss715627503	17	38,595,915	Across	Glyma.17G231600	*ELF3*	-49	-
ss715632278	18	56,027,056	HOH2018	Glyma.18G278100	*GmCOL2*	+60	Cao et al., 2015
ss715632491	18	57,673,097	EWE2018	Glyma.18G298900	*GmFT1a*	+14	Guo et al., 2015
ss715632484	18	57,587,071	HOH2018	Glyma.18G298900	*GmFT1a*	-67	Guo et al., 2015
ss715632491	18	57,673,097	EWE2018	Glyma.18G299000	*GmFT1b*	+(<1)	Guo et al., 2015
ss715632484	18	57,587,071	HOH2018	Glyma.18G299000	*GmFT1b*	-83	Guo et al., 2015
ss715635642	19	47,200,665	Across	Glyma.19G222000	*GmPIF3*	-196	-
ss715635642	19	47,200,665	EWE2018	Glyma.19G222000	*GmPIF3*	-196	-
ss715635642	19	47,200,665	HOH2018	Glyma.19G222000	*GmPIF3*	-196	-
ss715635674	19	47,378,001	HOH2020	Glyma.19G222000	*GmPIF3*	-19	-
ss715636060	19	50,566,017	EWE2020	Glyma.19G260900	*LHY2a*	+156	Lu et al., 2020
Days to maturity							
ss715587301	4	1,952,235	EWE2020	Glyma.04G022100	*FD*	+221	-
ss715601559	8	3,636,582	HOH2020	Glyma.08G046500	*GmFKF1-2*	-22	Li et al., 2013
ss715602218	8	44,384,746	Across	Glyma.08G325700	*GmFRL1*	+2	-
ss715607488	10	45,331,299	Across	Glyma.10G221500	*E2*	+15	Watanabe et al., 2011
ss715607488	10	45,331,299	EWE2020	Glyma.10G221500	*E2*	+15	Watanabe et al., 2011
ss715607488	10	45,331,299	HOH2020	Glyma.10G221500	*E2*	+15	Watanabe et al., 2011
ss715621693	15	3,573,258	HOH2018	Glyma.15G044400	*GmTOE4a*	+41	Zhao et al., 2015
ss715632223	18	55,622,046	Across	Glyma.18G273600	*Dt2*	-16	Ping et al., 2014
ss715632223	18	55,622,046	EWE2020	Glyma.18G273600	*Dt2*	-16	Ping et al., 2014
ss715632223	18	55,622,046	HOH2018	Glyma.18G273600	*Dt2*	-16	Ping et al., 2014
ss715635651	19	47,236,627	Across	Glyma.19G222000	*GmPIF3*	-160	-
ss715636052	19	50,534,870	Across	Glyma.19G260900	*LHY2a*	+125	Lu et al., 2020
ss715636052	19	50,534,870	HOH2018	Glyma.19G260900	*LHY2a*	+125	Lu et al., 2020

Table S6 Polymorphisms in the CDS region of *GmFRL1* and *GmAP1d*.

	Chr.	Pos. (bp)	REF	ALT	ANN
GmFRL1					
	8	44,379,816	G	C	synonymous variant, p.Leu28Leu
	8	44,379,831	C	T	synonymous variant, p.Ser33Ser
	8	44,380,012	C	T	missense variant, p.His94Tyr
	8	44,380,095	A	C	missense variant, p.Leu121Phe
	8	44,380,170	G	A	synonymous variant, p.Ala146Ala
	8	44,381,579	G	C	missense variant, p.Asp288His
	8	44,381,936	A	G	synonymous variant, p.Lys382Lys
	8	44,382,032	A	G	synonymous variant, p.Gln414Gln
	8	44,382,187	T	C	missense variant, p.Leu466Ser
	8	44,382,213	T	G	missense variant, p.Ser475Ala
	8	44,382,274	G	A	missense variant, p.Gly495Asp
	8	44,382,343	T	G	missense variant, p.Val518Gly
	8	44,382,348	G	A	missense variant, p.Ala520Thr
GmAP1d					
	2	12,087,053	T	A	missense variant, p.Val133Asp

Table S7 *E1-E4* haplotypes in four maturity groups in the resequencing panel.

	MG 000	**MG 00**	**MG 0**	**MG I**
E1/E2/E3/E4				1
E1/E2/e3/e4			1	
E1/e2/E3/E4		2	3	4
E1/e2/e3/E4		5	7	8
E1/e2/e3/e4	3	5	4	3
e1-as/E2/e3/E4		2	1	8
e1-as/E2/e3/e4			1	
e1-as/e2/E3/E4		9	21	26
e1-as/e2/E3/e4	2			
e1-as/e2/e3/E4		15	51	20
e1-as/e2/e3/e4	1	6	6	1
e1-nl/E2/E3/E4				1
e1-nl/E2/E3/e4			1	
e1-nl/e2/E3/E4			1	1
e1-nl/e2/E3/e4		2		
e1-nl/e2/e3/E4		1	1	
e1-nl/e2/e3/e4	4	3	2	1

e2: *e2-ns*; *E3*: *E3-Mi*, *E3-Ha*; *e3*: *e3-tr*, *e3-fs*, *e3-ns*, *e3-Mo*; *e4*: *e4-SORE-1*, *e4-par*, *e4-kes*, *e4-kam*

General Discussion

One promising step towards EU protein self-sufficiency is the expansion of soybean cultivation in Europe and the improvement of its productivity. This means that new soybean cultivars are required that combine adaptation to European climatic conditions, including higher latitude regions in Central and Northern Europe, with agronomic performance, especially protein yield. Plant breeding is based on the generation of new genetic variation followed by selection of genotypes with the desired characteristics as candidates for cultivar registration. The efficiency of breeding can be seen from the annual selection gain, which is determined by different parameters as expressed in the breeder's equation:

$$\Delta G = \frac{ih\sigma_G}{L}$$

where ΔG is the annual selection gain, i the selection intensity, h the square root of the heritability of the target trait or more general, the selection accuracy, σ_G the square root of the additive genetic variation within the population, and L the number of years per breeding cycle. In this thesis, different approaches were characterized and devised that allow to increase the annual selection gain through one or the other parameter. Marker-assisted selection, genomic selection and phenomic selection are selection tools that can increase the selection intensity but could also improve the selection accuracy and potentially the speed in a breeding program and thus the cycle length. In addition, the genetic variation can be increased by introgression of new germplasm, which requires the prior screening and assessment. In the following, the different approaches and their potential for soybean breeding are discussed in more detail.

The genetic basis of agronomic traits and the potential for marker-assisted selection

The breeding goals vary with the target region and in Europe the priorities of soybean breeding are different from the USA or Asia. Besides seed yield, which is always a main target of a breeding program, the focus has so far been on selecting soybean cultivars with high protein content and the improvement of the adaptation in order to expand the soybean cultivation in Central Europe. Thus, a better understanding of the genetic basis of the target traits could assist future selection if QTL suitable for marker-assisted selection can be identified in European material.

Protein and oil content are the two most important quality traits in soybean breeding. In our study, 944 recombinant inbred lines (RILs) derived from eight biparental families were used for QTL mapping. For protein content and oil content, we found two major-effect QTL on chromosomes 15 and 20 that were identified for both traits but with inverse effects (Zhu et al. 2020). Especially the QTL on chromosome 20 had the largest effect, which is consistent with previous research (Vaughn et al. 2014; Bandillo et al. 2015; Lee et al. 2019). Identification of the causal genes underlying these two QTL has drawn the attention of many researchers. *GmSWEET10a* (Glyma.15G049200) and *GmSWEET10b* (Glyma.08G183500), a pair of *SWEET* homologs, underwent a stepwise selection and affect protein content, oil content, and seed size simultaneously (Miao et al. 2020; Wang et al. 2020; Zhang et al. 2020a). *GmSWEET10a* is located in the region of the QTL on chromosome 15 and is therefore a possible candidate for the underlying gene. Fliege et al. (2022) reported fine-mapping to narrow down the known QTL cqSeed protein-003 on chromosome 20 in *G. soja* and further compared the sequence to the reference genome of William 82 to detect polymorphisms. This identified a candidate gene Glyma.20G85100 encoding a CCT-domain protein and a 321 bp transposon insertion that contributes to low seed protein content. Interestingly, Marsh et al. (2022) used a large panel of wild soybean, landraces and soybean cultivars and combined with association and linkage analysis identified the same candidate gene underlying cqSeed protein-003. Further analyses revealed structural variation of Glyma.20G85100, with a 304 bp insertion/deletion in the fourth exon and trinucleotide tandem repeats of various length in the second exon of this gene that were highly correlated with protein content, suggesting two possible causal variants. The validation of this gene as the causal gene underlying the QTL on chromosome 20 came shortly after from Goettel et al. (2022), who performed association studies and functional characterization analyses including the construction of transgenic lines, subcellular localization and gene expression. This demonstrated the candidate gene, named *POWR1*, as being causal and showed it to affect protein, oil content, seed weight and yield simultaneously. It is hypothesized to do so by regulating the expression of genes involved in lipid metabolism and seed nutrient transport. Results of allele distributions and genomic scans revealed that *POWR1* is a domestication gene and showed a nearly complete fixation of a transposable element insertion allele, that increases oil content and seed weight and decreases protein content, in *G. max* cultivars compared to *G. soja* due to artificial selection during soybean domestication. Thus, identification of novel, protein content-favorable allele(s) for use in protein-targeted soybean breeding programs might be expanded to *G. soja*, which warrants further research.

Seed yield is a major target in soybean breeding but a highly complex inherited trait. So far, few genes are known that are related to its component traits such as thousand-seed weight (Lu et al. 2017b). QTL mapping was also used in this study to investigate the genetic architecture of seed yield and other important agronomic traits within the RIL population. We found QTL with pleiotropic effects located on chromosomes 6, 19, and 18, controlling both seed yield and plant height. Further fine-mapping showed that they most likely correspond to the known loci *E1*, *E3*, and *Dt2* (Zhu et al. 2021). Phenotypic analyses also showed a significant positive correlation between seed yield and plant height (Kurasch et al. 2017a). *E1* and *E3* regulate maturity and *Dt2* the stem growth habit and their alleles can also be used to adjust seed yield, but this depends on the target environment. In Europe, lines with a too high level of photoperiod sensitivity bear the risk of not maturing in time, while early-maturing lines reduce the potential to achieve a higher yield. *Dt1* is another growth habit gene and has an epistatic interaction with *Dt2*. The determinate phenotype of soybean is caused by *dt1*.

In the background of *Dt1*, the plants are indeterminate in growth and *Dt2* can lead to a semi-determinate growth habit. The determinate and semi-determinate growth habits both have their own advantages for soybean grown in Europe. These examples illustrate the inter-relatedness between seed yield and maturity as well as agronomic traits and breeders therefore need to consider all traits when selecting promising material for future cultivars.

Soybean is a short-day photoperiod-sensitive crop and thus, soybean adaptation to the target region is always of high importance for soybean breeding in Europe. Mega-environment field trials revealed the effect of the known maturity loci *E1-E4* on adaptation under the Central and Northern European conditions (Kurasch et al. 2017b). In recent years, aided by the release of the soybean genome sequence and the development of low-cost whole-genome sequencing technologies, several loci responsible for the adaptation of soybean to higher as well as lower latitudes have been reported (Zhang et al. 2021). This also includes the identification of novel genes responsible for an early-maturing habit under high latitude conditions, such as the recently identified *Tof5* and *Tof18* (Dong et al. 2021a; Kou et al. 2022). This illustrates the complexity of the regulatory networks of soybean flowering time and maturity and suggests the presence of additional, yet unknown genes. In our study, we developed a large panel of 1,503 early-maturing soybean lines, including USDA and NARO (Japan) genebank accessions and European soybean cultivars. The panel was evaluated at six environments and the data was used for genome-wide association mapping. In addition, 338 of the lines were subjected to whole-genome sequencing, which allowed to zoom into QTL regions and identify candidate genes and alleles. This approach identified some of the known loci, like *E1*, *E2*, *E3*, *E4* and *Dt2*, which can serve as a validation of the underlying data. In addition, a novel candidate gene, *GmFRL1* (Glyma.08G325700), encoding a homolog of FRIGIDA-like 1 protein in *Arabidopsis*, was found for both days to flowering and maturity. *FRIGIDA LIKE 1* (*FRL1*) is necessary for the up-regulation of the floral inhibitor *FLOWERING LOCUS*

C (*FLC*) by *FRIGIDA* (*FRI*) in winter-annual habit *Arabidopsis* (Michaels et al. 2004). Interestingly, even though soybean does not require a vernalization process, as for example the winter cereals like winter wheat, it has retained the vernalization genes in its genome. Previous studies already indicated that genes related to the vernalization pathway may also have an effect on soybean flowering time (Lü et al. 2015; Lyu et al. 2020). Thus, *GmFRL1* is a promising candidate gene underlying this QTL and adds further support for the role of vernalization pathway genes in soybean adaptation.

We also performed an analysis to identify QTL-by-environment interactions, which was facilitated with the data from six environments. One significant QTL was found and identified as *E2*, and another significant QTL on chromosome 2 was detected in a subpopulation fixed for *E2*, for which *GmAP1d* (Glyma.02G121600) was identified as candidate gene. There are four *AP1* homologs in soybean and the quadruple mutant showed late-flowering under short-day conditions (Chen et al. 2020). *FT2a* and *FT5a* interact with *FDL19*, which binds to the promoter of *GmAP1a* to induce its transcription and promote flowering (Li et al. 2021). This makes *GmAP1d* a strong candidate gene underlying this QTL. Interestingly, we found not only changes in effect size as for *E2*, but inversed allelic effects of this QTL in the two environments in 2019 compared to the other environments. QTL and their effects are generally regarded as static, so a QTL is identified and assumed to have a certain effect on the target trait. However, these results highlight the plasticity and the environmental dependency of QTL effects and emphasize the importance of a molecular understanding of genotype-by-environment interactions. Breeders always breed for certain target environments and the interaction of the genotype with the environment further complicates matters. Especially with the consequences of climate change, weather conditions will become more variable and less predictable. Thus, the aim can either be to select alleles that show no or only little interaction with the environment, or if this is not possible, to at least understand the behavior of certain alleles in order to select the most suitable ones for a given target environment and its prevailing climatic conditions.

Collectively, the obtained results improve our understanding of the genetic architecture of the target traits in early-maturing soybean under European conditions. This lays the foundation for a utilization in European soybean breeding programs and determines the optimal choice of the breeding strategy.

Evaluation of marker-assisted selection for soybean breeding

Marker-assisted selection is a breeding tool that aims to pyramid favorable alleles of QTL for target traits. For simple-inherited traits, this approach is very promising and has proven its value. In soybean, it has been used to select for resistance against soybean cyst nematode which resulted in disease-resistant accessions (Arelli et al. 2006, 2017). Based on the results of our studies, marker-assisted selection could be applied in our or other European soybean breeding programs for some QTL. For example, the QTL on chromosome 20 for protein content and oil content, which explained over 20% or even 50% of the genotypic variance within the biparental families, or the QTL on chromosome 6 for seed yield and plant height. Phenotypic analyses showed that protein content is negatively correlated with oil content in the RIL population (Kurasch et al. 2017a), and consistent with this, the QTL on chromosome 20 has an inverse effect on protein content and oil content. In the diversity panel of early-maturing soybeans evaluated at six environments, we also observed this significant negative correlation between protein content and oil content (Figure 3a). In addition, the same marker identified for this major QTL in the 944 RILs was identified in the SNP dataset of the resequencing panel and then used to investigate the allelic effect of this SNP on protein and oil content in this diversity panel. This revealed a similar picture to the previous result from the biparental RIL populations, with inverse effects of the two alleles on protein and oil content (Figure 3b). Notably, only 12 soybean lines in the resequencing panel carry the protein content favorable allele. This allele was found to be associated with lower yield (Sebolt et al. 2000) and thus selection for higher seed yield may have favored the other allele. On the other hand, protein content is only the focus of selection in Europe, whereas in many other countries the selection is targeted at improving oil content. Analysis of the origin of the 12 lines carrying the protein favorable allele revealed that most of them originated from Europe, while none is from the USA, China, or South America, and thus, the different breeding goals in different regions may underlie the different allele frequencies (Figure 3c). The allele increasing protein content also had a negative effect on seed yield in our RIL population (Zhu et al. 2021a). This illustrates that even when QTL suitable for marker-assisted selection have been identified, the decision on how to use them may not be straightforward due to pleiotropic effects of the QTL alleles on more than one target trait. In general, the selection for one allele will then result in a disadvantage for another trait which must be compensated by other non-pleiotropic loci. In the case of this protein content QTL, selection depends on the primary aims of the breeding program but selection of the protein content-increasing allele may still be worthwhile, as illustrated by the fact that several cultivars carry it.

As QTL mapping in the RIL population identified *E1*, *E3* and *Dt2* as candidate genes underlying seed yield and plant height, we assessed their allelic effects on plant height in the resequencing panel (Figure 3d). This confirmed that plant height is associated with the polymorphisms in these genes. We further investigated the effect of these three candidate genes on protein and oil content. Accessions carrying wild-type *E1* showed high protein content and low oil content and lines carrying the *e1-nl* allele had high protein and oil content. There was no significant difference within *E3* and *Dt2* allelic groups. This illustrates that also the selection for maturity can affect quality traits, probably as a pleiotropic effect through the reduced or extended time available for seed filling and maturation.

Previous studies and our own results substantiated the important roles of the *E* series maturity genes in soybean adaptation and consequently their potential value for marker-assisted selection. Phenotypic selection done in breeding so far has resulted in different combinations of *E1-E4* alleles that can result in a similar maturity as illustrated by cultivars from the same maturity group that carry different *E* haplotypes (Kurasch et al. 2017b). Combined with two major QTL identified in our diversity panel (qFT28 and qMT2), we simulated two scenarios of marker-assisted selection using either the major QTL *E1*/*E2*/qFT28/qMT2 or the four *E* loci *E1*/*E2*/*E3*/*E4*. qFT28 was also reported for flowering time in a previous study with the physical distance between the two significant markers being smaller than 50 Kb (Mao et al. 2017) and qMT2 is a novel major QTL for maturity explaining 9.8% of the genotypic variance. Our result showed that the combination of the alleles promoting flowering or maturing substantially advanced both traits and thus allows tailoring adaptation to the target region. Kurasch et al. (2017b) reported three photoperiod-insensitive *E1-E4* haplotypes suited for the northernmost of the evaluated locations in Central Europe, and none of them carried photoperiod-sensitive alleles of the *E1* and *E2* genes. Liu et al. (2020a) observed that both *E1* and *E2*, particularly *E2*, significantly affected flowering time and maturity. In the present study, *E1* was found to have a large effect and consequently, the haplotype groups that carry *e1-nl* were the earliest to flower and mature. Notably, only 25 genotypes in the resequencing panel carried the wild-type *E2* allele substantiating the important role of *E2* for soybean adaptation to high latitude regions. In addition, we found that haplotype groups that only differ in *E3* alleles had only slight variations in flowering and maturity. In contrast, haplotype groups with the *E4* or *e4-SORE-1* alleles showed more pronounced differences for both traits. Thus, the available *e4-SORE-1* alleles might be more important than the *E3* alleles in the adaptation of soybean to high latitudes (Tsubokura et al. 2013; Kurasch et al. 2017b), but these comparisons must be treated with some caution due to the unbalancedness of the different *E* alleles in the genetic background of such a diversity panel.

However, yield is another important agronomic trait and Miladinović et al. (2018) reported *e1-as/e2/E3/E4* as the optimal combination of *E1-E4* for Central-Eastern Europe. Thus, from a breeding perspective, not only adaptation and quality but also yield needs to be considered when selecting certain alleles at maturity loci.

Taken together, our results showed that marker-assisted selection can help to select candidate lines with high protein content and a desired morphology in soybean breeding programs. However, the effect of this selection on other target traits may need to be counter-acted and taken into account when deciding on the selection intensity and the required population size of a segregating family. For traits with a more complex genetic architecture, no large-effect QTL were identified that could be used for marker-assisted selection.

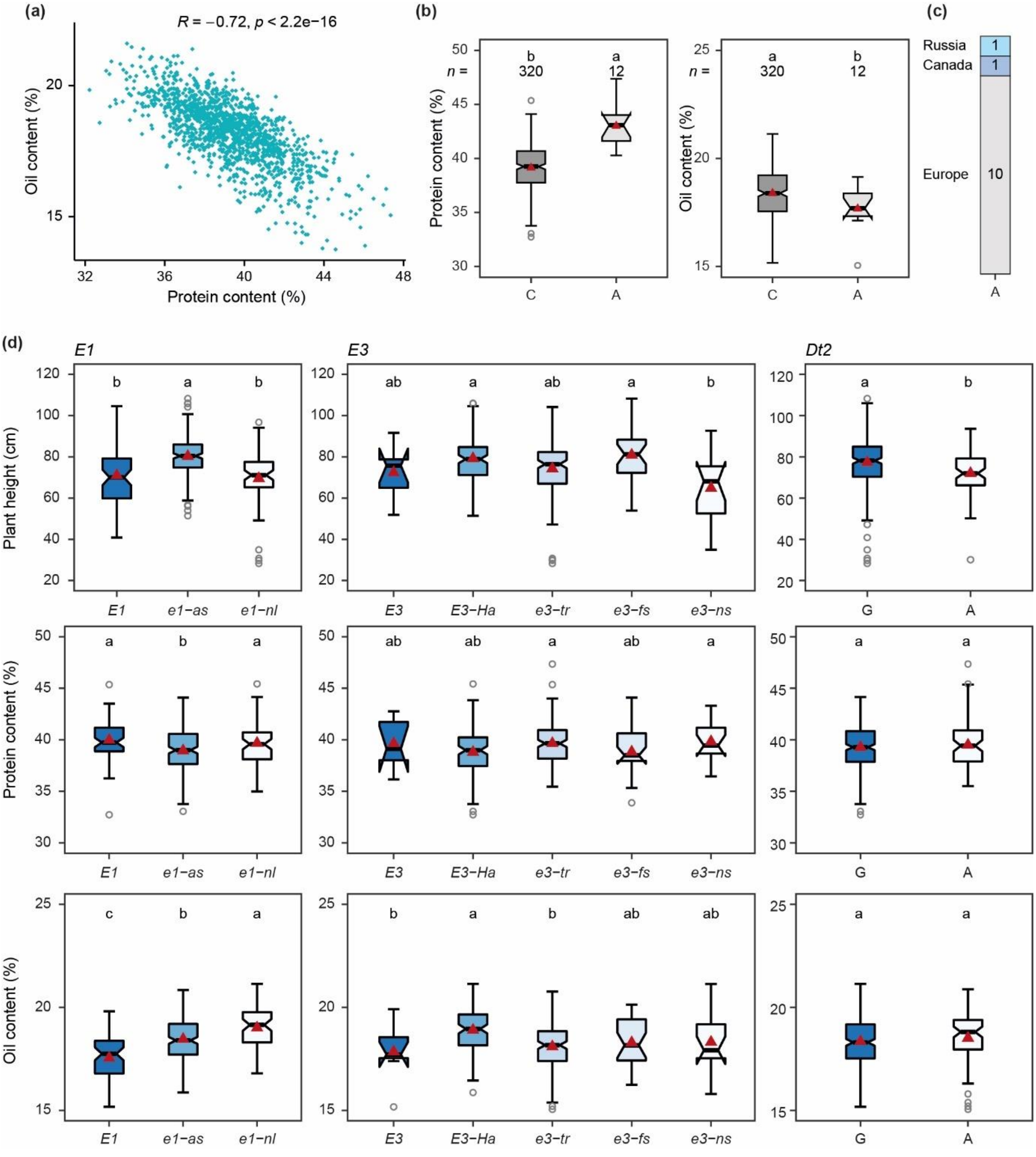


Figure 3 Characterization of the effect of identified QTL or candidate genes from the RIL population in the resequencing panel. (a) Correlation between BLUEs of protein content and oil content across environments in the diversity panel. (b) Allelic effect of the protein and oil content QTL on chromosome 20 in the resequencing panel. (c) Origin of twelve accessions carrying the protein content favorable allele. (d) Allelic effect of *E1*, *E3*, and *Dt2* on plant height, protein and oil content in the resequencing panel.

Utilization of genomic selection in soybean breeding

As mentioned above, marker-assisted selection has its limitations when selecting for complex traits. Within the last two decades, the rapid improvement of genomic sequencing technologies that allow to generate genome-wide marker profiles, paved the way for genomic selection (GS) that was suggested by Meuwissen et al. in 2001. Since then, genomic selection has been further refined and applied in both animal and plant breeding programs. Here, we used the RIL population to evaluate the potential of genomic selection in soybean breeding (Zhu et al. 2021b). This population is particularly suited for this purpose, as it mirrors the situation in breeding programs, with several biparental families and selection done within and among the individuals of such families. We found that for six traits, including seed yield, the prediction accuracies were high with a range from 0.79 to 0.87. This illustrates the potential of this approach to assist breeding of complex traits in soybean, but it is known that different aspects of genomic selection can be optimized in order to maximize the prediction accuracy.

Previous studies have revealed the potential effect of the biometric model used for the analysis. An approach that was already tested by Meuwissen et al. (2001) and that can be regarded as a reference for genomic prediction is rrBLUP. Considering that the rrBLUP approach might not fully capture the effects of large-effect QTL, these were here included as fixed effects in the rrBLUP model, but the result showed only a slight or no improvement of the prediction accuracy. This is likely due to the large linkage blocks present in biparental families, so that even the effects of strong QTL can be captured in rrBLUP, given that a sufficiently high number of markers is available. Other algorithms for model training were also evaluated, including Bayesian models, random forest (RF), or gradient boosting (GB), but we observed that they achieved similar prediction accuracies as rrBLUP. This is in line with previous studies that reported no significant difference in prediction accuracy using various models (Wang et al. 2015). Thus, owing to its predictive performance and its ease of implementation, rrBLUP can also be recommended for genomic selection in soybean breeding.

Another important aspect of genomic selection is the design of the training set. The half-diallel design of the RIL population allowed to investigate different degrees of relatedness between training and prediction set by using full-sibs, half-sibs or material that at least by pedigree is unrelated. Our results substantiated previous work, showing that if full-sibs can be used, they allow to achieve a high prediction accuracy with a minimum training set size (Marulanda et al. 2015). Full-sibs are available if only a part of a family is phenotyped that can then be used to predict the remaining individuals from the same family, which, however, is often not the case in breeding programs. Thus, if no lines from the family to be predicted are available to serve as training set, composite training sets should be used. Our results showed that these should include half-sib families or genetically related material. Importantly, our results also substantiated previous findings from a simulation study, showing that if unrelated families are used in the training and prediction sets, this can lead to very low or even negative prediction accuracies (Brauner et al. 2020). This means that in such cases the selection is pretty much random and the expenses required for genotyping are wasted. We also found that if QTL with large effect are segregating in both the training and the prediction set, such as the QTL for protein content on chromosome 20, higher prediction accuracies can be obtained even for unrelated families. Our results further showed that if half-sibs and unrelated families are combined in a training set, the prediction accuracy was still high. Thus, large composite training sets comprising individuals from many families should be used for genomic selection. If enough families are available to form the training set, this will include half-sibs or related material and should thereby ensure the success of genomic selection in a breeding program. If, however, crosses with exotic material from outside the breeding program are made, caution is required as the prediction accuracy for such a family may be low. In general, the training set will require a constant optimization and model update with each selection cycle. In conclusion, genomic selection can achieve high prediction accuracies also for complex traits like seed yield and thus complements marker-assisted selection as a tool for soybean breeding. A potential limitation, however, are the costs required for genotyping, which especially for smaller breeding programs or minor crops – as soybean still is in Germany and other European countries – cannot be afforded.

Potential of phenomic selection for soybean breeding

Rincent et al. (2018) introduced a new approach for selection in breeding programs, called phenomic selection (PS). The idea is to use the same biometric approach as for genomic selection, but using near-infrared spectroscopy (NIRS) data instead of molecular marker data for prediction. The reflectance of seed samples can be captured with a NIRS machine, which in soybean is already routinely done to assess the protein and oil content, so NIR spectra are available anyhow (Blanco and Villarroya 2002). Inspired by the prospect of having a cheap and high-throughput approach available for selection of complex traits, we investigated phenomic selection for soybean breeding as well as aspects of this approach for plant breeding in general. The results of our analyses showed that phenomic selection is competitive with genomic selection regarding its predictive ability and may even outperform it with regard to some aspects (Zhu et al. 2021c, 2022; Weiß et al. 2022).

First, a rather small number of variables was sufficient for phenomic selection. Variable selection based on correlation properties of NIRS data reduced the number of NIRS wavelengths to 23, which resulted in no or only a slight decrease of the predictive ability compared to all wavelengths. In contrast, the same number of markers used for genomic selection caused a substantial decrease of the predictive ability. For genomic selection, if the selected markers are closely linked to causal loci and the target traits are simply inherited, predictive abilities could still reach to a high value, similar to marker-assisted selection, but not for traits with a complex genetic basis. This result opens the opportunity for customized NIRS machines with fewer wavelengths, which could decrease costs and increase throughput. Thus, even if no NIRS machine is available yet in a breeding program, the required investments into infrastructure are manageable.

Second, genetic relatedness is a crucial factor affecting the prediction accuracy of genomic selection. Interestingly, our result showed that phenomic selection is less sensitive to genetic relatedness between the training set and the prediction set. In the RIL population, the prediction for seed yield of a single family using either full-sibs, half-sib families or unrelated families showed that predictive abilities were similar, which is in stark contrast to genomic prediction (Zhu et al. 2021b, c). In order to validate these findings and to shed light on other aspects of phenomic selection, we extended the analyses to a triticale and a maize dataset. In the study based on a large triticale dataset consisting of two DH populations and a diversity panel, we again observed a strong difference when performing genomic and phenomic prediction among populations (Zhu et al. 2022). Using each of the two DH populations as training set to predict grain yield in the diversity panel achieved a high predictive ability and combining them could further improve the phenomic predictive ability, whereas the genomic predictive ability was low. This finding of a robust among-population prediction of phenomic selection was further substantiated in maize. Here, the phenomic predictive ability was less affected by population structure than its genomic counterpart when predicting between the two heterotic groups Flint and Dent (Weiß et al. 2022). These findings not only indicate less restrictions of phenomic selection compared to genomic selection regarding the training set design, but also open the opportunity of extending the selection to more exotic germplasm without the need of special training sets, which warrants further research.

Third, compared to genomic selection, a smaller size of the training set is sufficient for phenomic selection to achieve the same predictive ability. In the RIL population, we observed a rapid increase of the phenomic predictive ability with training set size when using half-sibs, which reached a plateau with approximately 50 individuals at a predictive ability comparable to that using full-sibs. From a breeding perspective, every individual in the training set is associated with resources that have to be spend, making genomic selection more labor- and money-consuming when there are thousands or tens of thousands of selection candidates. Thus, the lower costs for a similar selection gain makes phenomic selection attractive for breeders.

However, our analyses also revealed that phenomic selection might only be suitable for complex inherited traits but not for mono- or oligo-genic traits. In the triticale dataset, we dissected the genetic basis underlying the studied traits grain yield, plant height, thousand-kernel weight, and the disease resistances powdery mildew and yellow rust and compared the performance of genomic prediction and phenomic prediction for these five traits. We found large-effect QTL for the disease resistance traits, especially for yellow rust in one DH population, which explained over 90% of the genotypic variance. In this case, the genomic predictive ability or that of marker-assisted selection was much higher than the phenomic predictive ability. For grain yield, by contrast, no large-effect QTL were found and in this case, phenomic prediction performed better than genomic prediction. Thus, phenomic selection can be applied to genetically complex traits, whereas more simple inherited traits are better tackled by single markers that allow to trace the major QTL or by classical phenotypic selection.

With regard to the characterization and implementation of phenomic selection in breeding, open questions remain. At present the idea is that the NIRS reflectance signals associate with endophenotypes which allow to predict complex traits like grain yield, but do not necessarily have to be directly related to these traits. Notably, phenomic selection may also use non-additive genetic effects to achieve its predictive ability. Thus, an important question for future research is how much the breeding value is improved when utilizing phenomic selection. Another question concerns the transferability of prediction models across environments and the optimal environment to generate the NIRS data.

The predictive ability of NIRS data from different environments can vary. Rincent et al. (2018) hypothesized that stress environments may be more powerful for prediction than normal conditions, which is a very interesting point that could not be addressed in this study but requires further research. In the RIL population, we used NIRS data from different environments for the training set and prediction set to predict phenotype BLUEs across environments and for seed yield, the predictive ability was reduced to a low level in this setting. If NIRS data from two environments were combined and adjusted NIRS data used for prediction, the high predictive abilities were restored. In order to optimize this approach for breeding, we tested the number of genotypes required to connect environments and observed that as few as 20~30 connecting genotypes are sufficient. While this approach should be validated with independent datasets, such connecting genotypes and the use of NIRS data adjusted across environments are valuable for the use of phenomic predictive in breeding programs. Here, NIRS data may often come from different environments, as even with a central location for the breeding activities, data from different years should be combined.

Our diversity panel with available NIRS data was used as an additional dataset to examine phenomic prediction and its predictive abilities for soybean target traits. Variance component analysis showed a similar picture to the previous study (Zhu et al. 2021c), as genotypic variance exists along the whole spectrum but the proportion varies for the different wavelengths (Figure 4a). Likewise, also the genotype-by-environment interaction variance is variable along the spectrum. The phenomic predictive abilities for yield- and adaptation-related traits obtained by cross-validation ranged from 0.67 to 0.86 (Figure 4b). This substantiates the promising phenomic predictive abilities obtained in the RIL population.

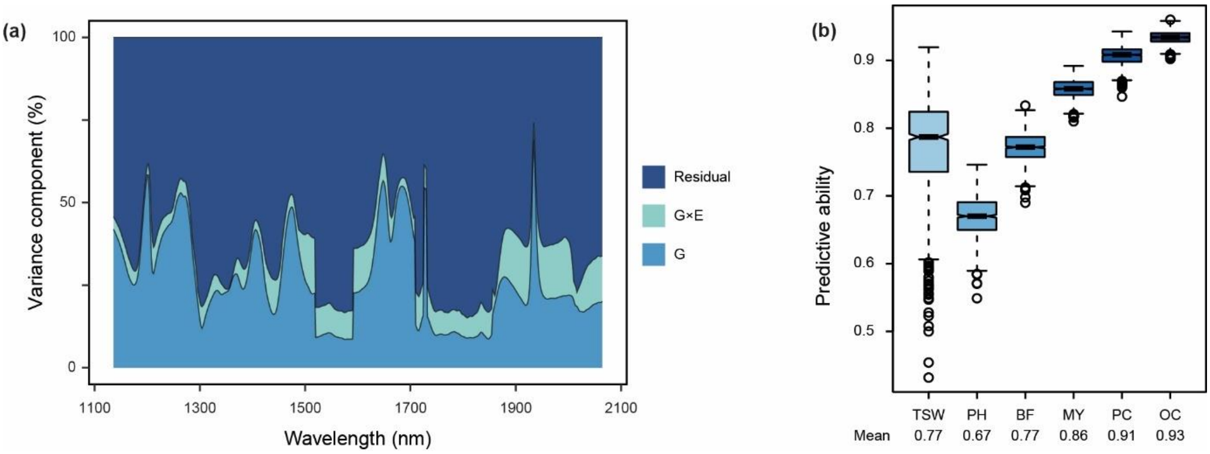


Figure 4 Phenomic prediction in the soybean diversity panel. (a) Variance component analysis for each wavelength along the NIRS profile. (b) Phenomic predictive abilities for trait value across environments obtained by 1,000 runs of five-fold cross-validation. TSW, thousand-seed weight; PH, plant height; BF, days to begin flowering; MY, days to maturity; PC, protein content; OC, oil content.

Phenomic selection can be applied using NIRS data obtained from seed samples (Figure 5), but also using other spectral data obtained from seeds or in the field. Field-based phenotyping systems are a field of active research and allow to collect spectral imaging data from the plots in the field, even at different developmental stages of the crops. We investigated the potential of this approach using triticale data from the Breedvision platform, which is a prototype phenotyping system that incorporates multiple sensors that are mounted on a carrier vehicle (Busemeyer et al. 2013). In our study, we used hyperspectral data in the spectral range from 976.1 to 1689.4 nm in 2014 (from 930 to 1700 nm in 2015) collected at several locations to predict the grain yield of triticale. Similar to a previous study (Krause et al. 2019), the predictive abilities using single-environment hyperspectral data were promising, for example reaching 0.82 for grain yield. Our results therefore indicate that field-based (hyper)spectral imaging data can be used for phenomic prediction of complex traits. Data collected at different developmental stages resulted in different predictive abilities, in this dataset the latest stage was better than the earlier ones. Thus, the choice of the optimal developmental stage or environmental conditions warrants further research. Field-based high-throughput phenotyping is not restricted to the ground, but can also be based on unmanned aerial vehicles, that allow to increase the throughput and collect the data of an entire field in a much shorter time. The predictive abilities of such an approach were evaluated recently for biomass yield in rye, which also showed promising results that were comparable to the genomic counterpart (Galán et al. 2020).

Collectively, these results provided strong support for the potential of phenomic selection in plant breeding. This novel tool is comparably cheap and amenable to high-throughput, which allows to screen the thousands of candidates generated in each cycle in breeding programs and thereby can become an important pillar to improve the selection gain.

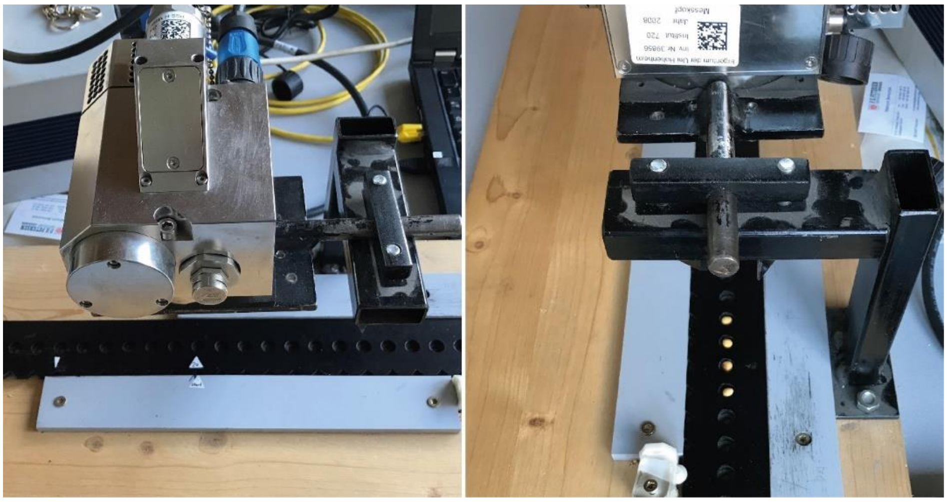


Figure 5 Polytec 2121 diode array spectrometer (Polytec GmbH, Waldenbronn, Germany) in Hohenheim and a modification that allows the measurement of single seed samples.

Utilization of diverse soybean germplasm for breeding

Genetic diversity is essential for the success of breeding programs. As illustrated by the breeder's equation, genetic variation is required to ensure the desired selection gain. In addition, genes or specific alleles may need to be introgressed, for example to facilitate adaptation or for resistance to abiotic and biotic stress. Soybean breeding for Central Europe is comparably young and goes back to rather few cultivars and lines. Thus, a broadening of the genetic diversity of the breeding pool is required. Moreover, marker-assisted selection and genome editing profit from or even require the identification of the causal genes and their polymorphisms underlying the target traits.

Here, we used a large diversity panel of early-maturing soybean to exemplarily exploit this diversity source for valuable variation and for the potential to dissect the genetic architecture of two adaptation traits (Zhu et al. submitted). The panel was shown to incorporate a large genetic variation for adaptation that can now be utilized in breeding to expand soybean's latitudinal adaptation and thus its cultivation in Central and Northern Europe (Figure 2). As mentioned before, two candidate genes *GmFRL1* and *GmAP1d* were identified, which after further validation can serve as targets for marker-assisted selection as well as for allele mining approaches or even genome editing to generate designed alleles. In addition, we found a novel allele of *E4*, called *e4-par* (after the European cultivar 'Paradis' carrying it), which is caused by a single-base deletion and leads to a frameshift and premature stop codon after 852 amino acids. Phenotypic analysis showed that lines carrying this novel *e4-par* allele showed the on average earliest flowering and maturity in this panel, being similar or maybe even earlier than the *e4-SORE-1* allele. Nine of the eleven lines carrying *e4-par* originate from Europe and one from Canada, which indicates that it may have been selected to facilitate adaptation to these high-latitude regions. This allele therefore warrants further characterization and holds promise for soybean breeding in Central and Northern Europe. These examples illustrate that the objectives mentioned above can be achieved with such diversity panels: utilization of soybean germplasm from other regions or genebanks provides new variation for breeding and can be used to dissect the genetic basis of target traits down to causal genes and alleles towards their targeted deployment. Once the genes and alleles underlying adaptation are known and characterized, it will also be possible and faster to introgress material from other maturity groups not adapted here, in order to make use of the breeding progress achieved elsewhere.

Integrated strategies for soybean breeding

To meet the protein demand in Europe as well as globally, soybean production should be increased and its cultivation expanded, which requires integrated breeding strategies that combine various classical approaches and new technologies. Here, we illustrate how the different approaches might be combined for the improvement of a classic soybean breeding program (Figure 6). Notably, the innovations devised in this thesis not only serve soybean breeding but can also assist breeding of other crops.

Tools that assist selection can be utilized to increase the selection intensity. This can be marker-assisted selection or genomic selection or a combination of the two, depending on whether major QTL are present for the trait of interest. Even if such genomics-assisted approaches are used, the costs can be restricted. One could of course apply both approaches to the early generation lines, which however, would require the most expenses. Alternatively, a hierarchical approach can be taken, where the early generation lines are first assessed in the field and selected for simple inherited traits. Then, marker-assisted selection for major QTL is done and only the remaining candidates are subjected to genomic selection, which requires genome-wide genotyping. The acreage of soybean in Germany and Central Europe is still rather small and soybean breeding programs, just as the breeding program of the University of Hohenheim, may not be able to afford these expenses required for the genomics-assisted selection tools. Phenomic selection based on NIRS data from seeds or other spectral data obtained in the field is an attractive alternative in this situation, as our results revealed that its predictive ability is competitive with that of genomic selection in different crops, i.e. in soybean, triticale, and maize investigated here. NIRS data are already routinely generated in soybean breeding and can be used for the prediction of complex traits. If molecular marker data are also available, these can be combined with the NIRS or hyperspectral data to achieve higher predictive abilities, especially for the traits with major QTL. As elaborated in detail in Zhu et al. (2021c) and as shown in Figure 6, there are different approaches how to incorporate and utilize phenomic selection in breeding programs. A promising approach that warrants further research is the single-seed prediction, as this would allow to select among but also within the segregating families of the early generation material. This requires some adjustments in the NIRS infrastructure as shown in Figure 5, which could be further refined to allow for a higher throughput and a real-time prediction and sorting of the seeds.

Time is money, also in plant breeding. The annual selection gain depends on the length of the breeding cycle and thus on the time required for each generation. Watson et al. (2018) proposed speed breeding as a strategy for long-day crops, which is mainly based on the extension of the photoperiod to 22 hours per day. Obviously, this will not work for short-day crops such as soybean. Nevertheless, a speed breeding protocol was also developed for short-day crops by Jähne et al. (2020), which is based on light-emitting diodes (LEDs) with adjustable light quality. Under a blue-light enriched, far-red-deprived light spectrum with 10 hours of light per day, soybean is able to reach up to five generations per year. This speed breeding system can be another key element for an efficient soybean breeding and is already beginning to be routinely applied in our soybean breeding program.

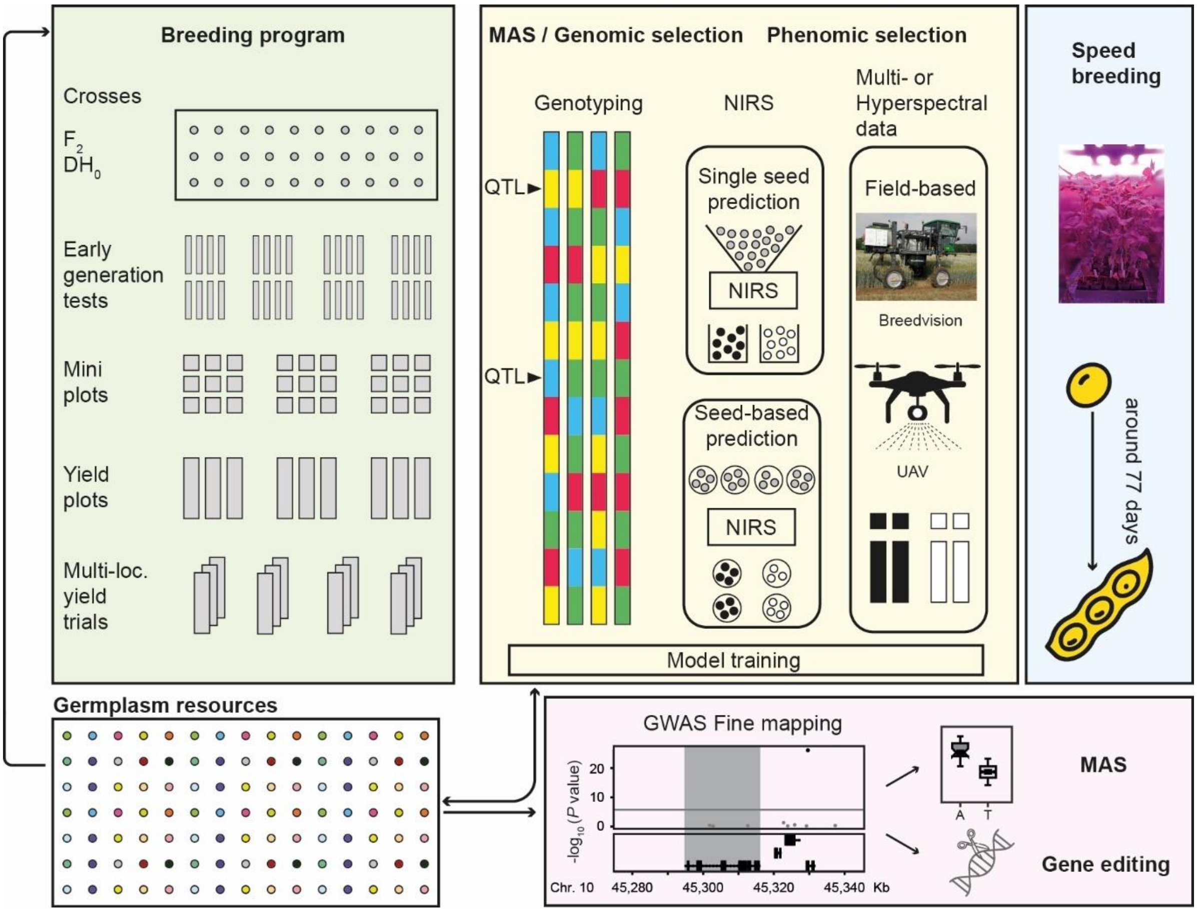


Figure 6 Integration of different breeding strategies for the improvement of a soybean breeding program. Picture of BreedVision is photographed by Tobias Würschum, the picture of soybean plants growing in a speed breeding chamber is from Jähne et al. (2020).

In addition to the selection strategy and the shortening of the generation time, the genetic diversity of the breeding pool is another key factor determining the success of soybean breeding programs. Breeders also need to select promising lines or lines carrying favorable alleles for target traits from other germplasm resources, e.g. from genebanks. On the one hand, these lines can be used in the breeding program as parents for crosses, for example to improve adaptation as illustrated here. On the other hand, forward genetic approaches using QTL or association mapping are needed to identify QTL controlling the target traits. Once identified, these can be used in marker-assisted selection or to search for novel alleles. Furthermore, genome editing can be applied in the future to generate designed alleles with the desired properties.

Taken together, a range of phenotypic, genomic and novel approaches are available for soybean breeding, which can be integrated and jointly hold great potential to assist the expansion of soybean cultivation in Central Europe through breeding of adapted and agronomically improved cultivars.

Summary

Soybean is the economically most important leguminous crop worldwide and serves as a main source of plant protein for human nutrition and animal feed. Europe is dependent on plant protein imports and the EU protein self-sufficiency, which is an issue that has been on the political agenda for several decades, has recently received renewed interest. The protein imports are mainly in the form of soybean meal, and soybean therefore appears well-suited to mitigate the protein deficit in Europe. This, however, requires an improvement of soybean production as well as an expansion of soybean cultivation and thus breeding of new cultivars that combine agronomic performance with adaptation to the climatic conditions in Central Europe. The objective of this thesis was to characterize, evaluate and devise approaches that can improve the efficiency of soybean breeding.

Breeding is essentially the generation of new genetic variation and the subsequent selection of superior genotypes as candidates for new cultivars. The process of selection can be supported by marker-assisted or genomic selection, which are both based on molecular markers. A first step towards the utilization of these approaches in breeding is the characterization of the genetic architecture underlying the target traits. While such quantitative trait locus (QTL) mapping studies are available for soybean, little was known about European soybean breeding material under our environmental conditions. In this study, we therefore performed QTL mapping for six target traits in a large population of 944 recombinant inbred lines from eight biparental families. We identified a few major-effect QTL, for example for protein and oil content on chromosomes 15 and 20, or the loci *E1*, *E3* and *Dt2* as likely candidates for three pleiotropic QTL affecting seed yield and plant height. These results showed that some major-effect QTL are present that could be utilized in marker-assisted selection, but in general the target traits are quantitatively inherited.

For such traits controlled by numerous small-effect QTL, genomic selection has proven as a powerful tool to assist selection in breeding programs. We therefore also evaluated the genomic prediction accuracy and found this to be high and promising for the six traits of interest. Furthermore, this dataset with the eight families is representative for the situation in breeding and we used it to analyze the performance of genomic prediction dependent on the genetic relatedness between the training set and the prediction set. The results showed that prediction works best if related material like full-sibs or half-sibs are available, and generally suggested that composite training sets should be used when performing genomic selection. In conclusion, these results illustrated the potential of genomic selection for soybean breeding programs, but a potential limitation of this approach are the costs required for genotyping with molecular markers.

Phenomic selection is an alternative approach that uses near-infrared or other spectral data for prediction instead of the marker data used for its genomic counterpart. Here, we evaluated the phenomic predictive ability in soybean as well as in triticale and maize. Phenomic prediction based on near-infrared spectroscopy (NIRS) of seeds showed a comparable or even slightly higher predictive ability than genomic prediction. We further characterized this novel approach and devised strategies for its implementation in breeding programs. Interestingly, phenomic prediction was much less sensitive than genomic prediction to the genetic relatedness between the training and prediction sets, and in addition, a smaller training set is sufficient for a similar predictive ability. NIRS data from a single environment can be used to predict phenotypic values across environments, and if NIRS data of the training set and the prediction set are from different environments, connecting genotypes can be added to adjust the NIRS data. Collectively, our results illustrate the potential of phenomic selection for breeding of complex traits in soybean and other crops. The advantage of this approach is that NIRS data are often available anyhow and can be generated with much lower costs than the molecular marker data, also in high-throughput required to screen the large numbers of selection candidates in breeding programs.

Soybean is a short-day plant originating from temperate China, and thus adaptation to the climatic conditions of Central Europe is a major breeding goal. In this study, we established a large diversity panel of 1,503 early-maturing soybeans, comprising of European breeding material and accessions from genebanks. This panel was evaluated in six environments, which revealed valuable genetic variation that can be introgressed into our breeding programs. In addition, we deciphered the genetic architecture of the adaptation traits flowering time and maturity, which revealed known loci (*E1*, *E2*, *E3*, *E4*, *Dt2*) but also novel candidate genes like *GmFRL1*. Moreover, *GmAP1d* was identified as a candidate gene for a QTL-by-environment interaction whose favorable allele varies between environments. A subset of 338 lines of the diversity panel was resequenced, which revealed a novel allele of the *E4* maturity locus, *e4-par*, that was found in eleven soybean lines and may have been selected to improve soybean adaptation to high latitude regions. Different combinations of alleles of the identified loci allow to tailor flowering time and maturity, which lays the basis for a further expansion of soybean in Central Europe. These examples also illustrate how genes underlying the target traits can be identified towards a future targeted utilization of natural or engineered alleles in breeding.

Taken together, the findings of this study show the potential of several phenotypic, genomic and novel approaches that can be integrated to improve the efficiency of soybean breeding and thus hold great promise to assist the expansion of soybean cultivation in Central Europe through breeding of adapted and agronomically improved cultivars.

Zusammenfassung

Die Sojabohne ist die wirtschaftlich wichtigste Leguminose weltweit und dient als eine Hauptquelle für pflanzliches Eiweiß in der menschlichen Ernährung und im Tierfutter. Europa ist von pflanzlichen Eiweißimporten abhängig und die Selbstversorgung der EU mit Eiweiß, ein Thema, das seit mehreren Jahrzehnten auf der politischen Agenda steht, hat in letzter Zeit wieder an Interesse gewonnen. Die Eiweißimporte erfolgen hauptsächlich in Form von Sojaschrot und Soja scheint daher gut geeignet das Eiweißdefizit in Europa abzumildern. Dies erfordert jedoch eine Steigerung der Sojaproduktion sowie eine Ausweitung des Sojaanbaus und damit die Züchtung neuer Sorten, die agronomische Leistung mit Anpassung an die klimatischen Bedingungen in Mitteleuropa verbinden. Ziel dieser Arbeit war es, Ansätze zu charakterisieren, zu bewerten und zu entwickeln, die die Effizienz der Sojazüchtung verbessern können.

Züchtung basiert im Wesentlichen auf der Erzeugung neuer genetischer Variation und der anschließenden Selektion überlegener Genotypen als Kandidaten für neue Sorten. Dieser Selektionsprozess kann durch markergestützte oder genomische Selektion unterstützt werden, die beide auf molekularen Markern beruhen. Ein erster Schritt zur Nutzung dieser Ansätze in der Züchtung ist die Charakterisierung der den Zielmerkmalen zugrundeliegenden genetischen Architektur. Während für Soja solche Quantitative Trait Locus (QTL)-Kartierungsstudien bereits durchgeführt waren, war über europäisches Sojazuchtmaterial unter unseren Umweltbedingungen wenig bekannt. In dieser Studie wurde daher eine QTL-Kartierung für sechs Zielmerkmale in einer großen Population von 944 rekombinanten Inzuchtlinien aus acht biparentalen Familien durchgeführt. Dadurch wurden einige QTL mit großem Effekt identifiziert, z. B. für Protein- und Ölgehalt auf den Chromosomen 15 und 20, oder die Loci *E1*, *E3* und *Dt2* als wahrscheinliche Kandidaten für drei pleiotrope QTL, die den Kornertrag und die Pflanzenhöhe beeinflussen. Diese Ergebnisse zeigten, dass es einige QTL mit großen Effekten gibt, die für die markergestützte Selektion genutzt werden könnten, aber im Allgemeinen werden die Zielmerkmale quantitativ vererbt.

Bei solchen Merkmalen, die von zahlreichen QTL mit kleinem Effekt kontrolliert werden, hat sich die genomische Selektion als leistungsfähiges Instrument zur Unterstützung der Selektion in Zuchtprogrammen erwiesen. Deshalb wurde auch die genomische Vorhersagegenauigkeit untersucht und festgestellt, dass diese für die sechs Zielmerkmale hoch und damit vielversprechend ist. Außerdem ist dieser Datensatz mit acht Familien repräsentativ für die Situation in Zuchtprogrammen und wurde deshalb genutzt um die genomischen Vorhersagegenauigkeit in Abhängigkeit von der genetischen Verwandtschaft zwischen den für die Kalibration der Modelle genutzten und den vorherzusagenden Linien

zu analysieren. Die Ergebnisse zeigten, dass die Vorhersage am besten funktioniert, wenn verwandtes Material wie Vollgeschwister oder Halbgeschwister verfügbar sind und legten allgemein nahe, dass bei der genomischen Selektion gemischte Kalibrationssets verwendet werden sollten. Zusammenfassend lässt sich sagen, dass die Ergebnisse das Potenzial der genomischen Selektion für Sojazuchtprogramme verdeutlichten, wobei jedoch die Kosten für die Genotypisierung mit molekularen Markern ein möglicher Nachteil dieses Ansatzes sind.

Die phänomische Selektion ist ein alternativer Ansatz, bei dem Nahinfrarot- oder andere Spektraldaten für die Vorhersage verwendet werden, anstatt der Markerdaten wie bei der genomische Selektion. Hier wurde die phänomische Vorhersagefähigkeit bei Soja sowie bei Triticale und Mais untersucht. Die phänomische Vorhersage mittels Nahinfrarotspektroskopie (NIRS) von Bohnen/Körnern zeigte eine vergleichbare oder sogar leicht höhere Vorhersagegenauigkeit als die genomische Vorhersage. Dieser neue Ansatz wurde weiter charakterisiert und Strategien für den Einsatz in Zuchtprogrammen entwickelt. Interessanterweise war die phänomische Vorhersage viel weniger empfindlich als die genomische Vorhersage im Hinblick auf die genetische Verwandtschaft zwischen den Kalibrations- und Vorhersagesets, und außerdem ist ein kleineres Kalibrationsset ausreichend für eine vergleichbare Vorhersagegenauigkeit. NIRS-Daten von einer einzigen Umwelt können zur Vorhersage der Merkmale in verschiedenen Umwelten verwendet werden, und wenn die NIRS-Daten des Kalibrationssets und des Vorhersagesets aus verschiedenen Umwelten stammen, können verknüpfende Genotypen verwendet werden. Insgesamt veranschaulichen unsere Ergebnisse das Potenzial der phänomischen Selektion für die Züchtung komplexer Merkmale bei Soja und anderen Nutzpflanzen. Der Vorteil dieses Ansatzes besteht darin, dass NIRS-Daten häufig ohnehin erhoben werden und mit wesentlich geringeren Kosten als molekulare Markerdaten generiert werden können, auch im Hochdurchsatz, der für das Screening der großen Anzahl von Selektionskandidaten in Züchtungsprogrammen erforderlich ist.

Die Sojabohne ist eine Kurztagspflanze, die ursprünglich aus den gemäßigten Zonen in China stammt, weshalb die Anpassung an die klimatischen Bedingungen in Mitteleuropa ein wichtiges Zuchtziel ist. In dieser Studie wurde ein großes Diversitätspanel mit 1.503 frühreifenden Sojalinien erstellt, das aus europäischem Zuchtmaterial und Akzessionen aus Genbanken besteht. Dieses Panel wurde in sechs Umwelten evaluiert, was wertvolle genetische Variationen aufzeigte, die in Zuchtprogramme eingebracht werden kann. Darüber hinaus wurde die genetische Architektur der Adaptationsmerkmale Blühzeitpunkt und Reife entschlüsselt und dabei bekannte Loci (*E1*, *E2*, *E3*, *E4*, *Dt2*), aber auch neue Kandidatengene wie *GmFRL1* entdeckt. Darüber hinaus wurde *GmAP1d* als ein Kandidatengen für eine QTL-by-Environment-Interaktion identifiziert, dessen vorteilhaftes Allel je nach Umgebung

variiert. Ein Teil von 338 Linien des Diversitätspanels wurde resequenziert, wodurch ein neues Allel des *E4*-Reifelocus, *e4-par*, identifiziert wurde, das in elf Sojalinien gefunden wurde und möglicherweise selektiert wurde, um die Anpassung von Soja an Regionen in hohen Breitengraden zu verbessern. Unterschiedliche Kombinationen von Allelen der identifizierten Loci ermöglichen es, den Blühzeitpunkt und die Reife anzupassen, was die Grundlage für eine weitere Expansion der Sojabohne in Mitteleuropa bildet. Diese Beispiele veranschaulichen auch, wie Gene, die den Zielmerkmalen zugrunde liegen, identifiziert werden können, um in Zukunft gezielt natürliche oder designte Allele in der Züchtung zu nutzen.

Insgesamt zeigen die Ergebnisse dieser Studie das Potenzial verschiedener phänotypischer, genomischer und neuer Ansätze, die zur Verbesserung der Effizienz der Sojazüchtung integriert werden können und vielversprechend sind, um die Ausweitung des Sojaanbaus in Mitteleuropa durch die Züchtung angepasster und agronomisch verbesserter Sorten zu unterstützen.

References

Arelli PR, Shannon JG, Mengistu A, et al (2017) Registration of Conventional Soybean Germplasm JTN-4307 with Resistance to Nematodes and Fungal Diseases. J Plant Regist 11:192–199. https://doi.org/https://doi.org/10.3198/jpr2016.10.0058crg

Arelli PR, Young LD, Mengistu A (2006) Registration of High Yielding and Multiple Disease-Resistant Soybean Germplasm JTN-5503. Crop Sci 46:2723–2724. https://doi.org/https://doi.org/10.2135/cropsci2005.12.0471CRG

Bandillo N, Jarquin D, Song Q, et al (2015) A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome 8:. https://doi.org/10.3835/plantgenome2015.04.0024

Bao Y, Vuong T, Meinhardt C, et al (2014) Potential of Association Mapping and Genomic Selection to Explore PI 88788 Derived Soybean Cyst Nematode Resistance. Plant Genome 7:1–13. https://doi.org/10.3835/plantgenome2013.11.0039

Blanco M, Villarroya I (2002) NIR spectroscopy: a rapid-response analytical tool. TrAC Trends Anal Chem 21:240–250. https://doi.org/https://doi.org/10.1016/S0165-9936(02)00404-1

Brauner PC, Müller D, Molenaar WS, Melchinger AE (2020) Genomic prediction with multiple biparental families. Theor Appl Genet 133:133–147. https://doi.org/10.1007/s00122-019-03445-7

Busemeyer L, Mentrup D, Möller K, et al (2013) BreedVision — A Multi-Sensor Platform for Non-Destructive Field-Based Phenotyping in Plant Breeding. Sensors 13

Cao D, Takeshima R, Zhao C, et al (2017) Molecular mechanisms of fowering under long days and stem growth habit in soybean. J Exp Bot 68:1873–1884. https://doi.org/10.1093/jxb/erw394

Carter TE, Nelson R, Sneller CH, Cui Z (2004) Soybeans: improvement, production and uses. Agronomy 16:303-416.

Chen L, Nan H, Kong L, et al (2020) Soybean *AP1* homologs control flowering time and plant height. J Integr Plant Biol 62:1868–1879. https://doi.org/10.1111/jipb.12988

Chen L, Yang H, Fang Y, et al (2021) Overexpression of *GmMYB14* improves high-density yield and drought tolerance of soybean through regulating plant architecture mediated by the brassinosteroid pathway. Plant Biotechnol J 19:702–716. https://doi.org/https://doi.org/10.1111/pbi.13496

Cober ER (2011) Long Juvenile Soybean Flowering Responses under Very Short Photoperiods. Crop Sci 51:140–145. https://doi.org/https://doi.org/10.2135/cropsci2010.05.0262

Cober ER, Molnar SJ, Charette M, Voldeng HD (2010) A New Locus for Early Maturity in Soybean. Crop Sci 50:524–527. https://doi.org/https://doi.org/10.2135/cropsci2009.04.0174

Cober ER, Voldeng HD (2001) Low R:FR light quality delays flowering of *E7E7* soybean lines. Crop Sci 41:1823–1826. https://doi.org/10.2135/cropsci2001.1823

Cuevas J, Montesinos-López O, Juliana P, et al (2019) Deep Kernel for genomic and near infrared predictions in multi-environment breeding trials. G3 Genes, Genomes, Genet 9:2913–2924. https://doi.org/10.1534/g3.119.400493

Dong L, Cheng Q, Fang C, et al (2021a) Parallel selection of distinct *Tof5* alleles drove the adaptation of cultivated and wild soybean to high latitudes. Mol Plant 1–14. https://doi.org/10.1016/j.molp.2021.10.004

Dong L, Fang C, Cheng Q, et al (2021b) Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. Nat Commun 12:1–11. https://doi.org/10.1038/s41467-021-25800-3

Đorđević V, Ćeran M, Miladinović J, et al (2019) Exploring the performance of genomic prediction models for soybean yield using different validation approaches. Mol Breed 39:74. https://doi.org/10.1007/s11032-019-0983-6

Fang C, Liu J, Zhang T, et al (2021) A recent retrotransposon insertion of *J* caused *E6* locus facilitating soybean adaptation into low latitude. J Integr Plant Biol 63:995–1003. https://doi.org/https://doi.org/10.1111/jipb.13034

FAOSTAT (2021) FAOSTAT statistical database. In: Nations, Food Agric. Organ. United. https://search.library.wisc.edu/catalog/999890171702121

Fliege CE, Ward RA, Vogel P, et al (2022) Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20. Plant J 1–15. https://doi.org/10.1111/tpj.15658

Galán RJ, Bernal-Vasquez AM, Jebsen C, et al (2020) Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye. Theor Appl Genet 133:3001–3015. https://doi.org/10.1007/s00122-020-03651-8

Goettel W, Zhang H, Li Y, et al (2022) *POWR1* is a domestication gene pleiotropically regulating seed quality and yield in soybean. Nat Commun 13, 3051. https://doi.org/10.1038/s41467-022-30314-7

Hahn V, Würschum T (2014) Molecular genetic characterization of Central European soybean breeding germplasm. Plant Breed 133:748–755. https://doi.org/10.1111/pbr.12212

Häusling M (2011) The EU protein deficit: what solution for a long-standing problem? (2010/2111(INI)). Committee on Agriculture and Rural Development

Holland JB (2004) Implementation of molecular markers for quantitative traits in breeding programs—challenges and opportunities. In New Directions for a Diverse Planet: Proceedings for the 4th International Crop Science Congress. Regional Institute, Gosford, Australia, www. cropscience. org. au/icsc2004.

Hymowitz T and Shurtleff WR (2005) Debunking Soybean Myths and Legends in the Historical and Popular Literature. Crop Science, 45: 473-476. https://doi.org/10.2135/cropsci2005.0473

Jähne F, Balko C, Hahn V, et al (2019) Cold stress tolerance of soybeans during flowering: QTL mapping and efficient selection strategies under controlled conditions. Plant Breed 138:708–720. https://doi.org/10.1111/pbr.12734

Jähne F, Hahn V, Würschum T, Leiser WL (2020) Speed breeding short-day crops by LED-controlled light schemes. Theor Appl Genet 133:2335–2342. https://doi.org/10.1007/s00122-020-03601-4

Jia H, Jiang B, Wu C, et al (2014) Maturity group classification and maturity locus genotyping of early-maturing soybean varieties from high-latitude cold regions. PLoS One 9:1–9. https://doi.org/10.1371/journal.pone.0094139

Kong F, Nan H, Cao D, et al (2014) A new dominant gene *E9* conditions early flowering and maturity in soybean. Crop Sci 54:2529–2535. https://doi.org/10.2135/cropsci2014.03.0228

Kou K, Yang H, Li H, et al (2022) A functionally divergent *SOC1* homolog improves soybean yield and latitudinal adaptation. Curr Biol. https://doi.org/https://doi.org/10.1016/j.cub.2022.02.046

Krause MR, González-Pérez L, Crossa J, et al (2019) Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat. G3 Genes, Genomes, Genet 9:1231–1247. https://doi.org/10.1534/g3.118.200856

Kurasch AK, Hahn V, Leiser WL, et al (2017a) Phenotypic analysis of major agronomic traits in 1008 RILs from a diallel of early European soybean varieties.

Crop Sci 57:726–738. https://doi.org/10.2135/cropsci2016.05.0318

Kurasch AK, Hahn V, Leiser WL, et al (2017b) Identification of mega-environments in Europe and effect of allelic variation at maturity *E* loci on adaptation of European soybean. Plant Cell Environ 40:765–778. https://doi.org/10.1111/pce.12896

Lane HM, Murray SC, Montesinos-López OA, et al (2020) Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels. Plant Phenome J 3:1–19. https://doi.org/10.1002/ppj2.20002

Lee S, Van K, Sung M, et al (2019) Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. Theor Appl Genet 132:1639–1659. https://doi.org/10.1007/s00122-019-03304-5

Li J, Wang X, Song W, et al (2017) Genetic variation of maturity groups and four *E* genes in the Chinese soybean mini core collection. PLoS One 12:e0172106

Li X, Fang C, Yang Y, et al (2021) Overcoming the genetic compensation response of soybean florigens to improve adaptation and yield at low latitudes. Curr Biol 31:3755-3767.e4. https://doi.org/10.1016/j.cub.2021.06.037

Lin X, Liu B, Weller JL, et al (2021) Molecular mechanisms for the photoperiodic regulation of flowering in soybean. J Integr Plant Biol 63:981–994. https://doi.org/10.1111/jipb.13021

Liu B, Kanazawa A, Matsumura H, et al (2008) Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. Genetics 180:995–1007. https://doi.org/10.1534/genetics.108.092742

Liu B, Watanabe S, Uchiyama T, et al (2010) The soybean stem growth habit gene *Dt1* is an ortholog of arabidopsis *TERMINAL FLOWER1*. Plant Physiol 153:198–210. https://doi.org/10.1104/pp.109.150607

Liu L, Song W, Wang L, et al (2020a) Allele combinations of maturity genes *E1-E4* affect adaptation of soybean to diverse geographic regions and farming systems in China. PLoS One 15:1–15. https://doi.org/10.1371/journal.pone.0235397

Liu S, Zhang M, Feng F, Tian Z (2020b) Toward a 'Green Revolution' for Soybean. Mol Plant 13:688–697. https://doi.org/https://doi.org/10.1016/j.molp.2020.03.002

Lü J, Suo H, Yi R, et al (2015) Glyma11g13220, a homolog of the vernalization pathway gene *VERNALIZATION 1* from soybean [*Glycine max* (L.) Merr.], promotes flowering in Arabidopsis thaliana. BMC Plant Biol 15:1–12. https://doi.org/10.1186/s12870-015-0602-6

Lu S, Dong L, Fang C, et al (2020) Stepwise selection on homeologous *PRR* genes controlling flowering and maturity during soybean domestication. Nat Genet 52:428–436. https://doi.org/10.1038/s41588-020-0604-7

Lu S, Zhao X, Hu Y, et al (2017a) Natural variation at the soybean *J* locus improves adaptation to the tropics and enhances yield. Nat Genet 49:773–779. https://doi.org/10.1038/ng.3819

Lu X, Xiong Q, Cheng T, et al (2017b) A *PP2C-1* Allele Underlying a Quantitative Trait Locus Enhances Soybean 100-Seed Weight. Mol Plant 10:670–684. https://doi.org/10.1016/j.molp.2017.03.006

Lyu J, Cai Z, Li Y, et al (2020) The Floral Repressor *GmFLC-like* Is Involved in Regulating Flowering Time Mediated by Low Temperature in Soybean. Int J Mol Sci 21:1322. https://doi.org/10.3390/ijms21041322

Ma Y, Reif JC, Jiang Y, et al (2016) Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). Mol Breed 36:1–10. https://doi.org/10.1007/s11032-016-0504-9

Mao T, Li J, Wen Z, et al (2017) Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. BMC Genomics 18:1–17. https://doi.org/10.1186/s12864-017-

3778-3

Marsh JI, Hu H, Petereit J, et al (2022) Haplotype mapping uncovers unexplored variation in wild and domesticated soybean at the major protein locus cqProt-003. Theor Appl Genet. https://doi.org/10.1007/s00122-022-04045-8

Marulanda JJ, Melchinger AE, Würschum T (2015) Genomic selection in biparental populations: Assessment of parameters for optimum estimation set design. Plant Breed 134:623–630. https://doi.org/10.1111/pbr.12317

Matei G, Woyann LG, Milioli AS, et al (2018) Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. Mol Breed 38:117. https://doi.org/10.1007/s11032-018-0872-4

Merrick LF, Carter AH (2021) Comparison of genomic selection models for exploring predictive ability of complex traits in breeding programs. Plant Genome 14:1–19. https://doi.org/10.1002/tpg2.20158

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

Miao L, Yang S, Zhang K, et al (2020) Natural variation and selection in *GmSWEET39* affect soybean seed oil content. New Phytol 225:1651–1666. https://doi.org/10.1111/nph.16250

Michaels SD, Bezerra IC, Amasino RM (2004) *FRIGIDA*-related genes are required for the winter-annual habit in *Arabidopsis*. Proc Natl Acad Sci U S A 101:3281–3285. https://doi.org/10.1073/pnas.0306778101

Miladinović J, Ćeran M, Đorđević V, et al (2018) Allelic variation and distribution of the major maturity genes in different soybean collections. Front Plant Sci 9:1–8. https://doi.org/10.3389/fpls.2018.01286

Parmley K, Nagasubramanian K, Sarkar S, et al (2019) Development of Optimized Phenomic Predictors for Efficient Plant Breeding Decisions Using Phenomic-

Assisted Selection in Soybean. Plant Phenomics 2019:1–15. https://doi.org/10.34133/2019/5809404

Patil G, Mian R, Vuong T, et al (2017) Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. Theor Appl Genet 130:1975–1991. https://doi.org/10.1007/s00122-017-2955-8

Ping J, Liu Y, Sun L, et al (2014) *Dt2* is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. Plant Cell 26:2831–2842. https://doi.org/10.1105/tpc.114.126938

Rincent R, Charpentier JP, Faivre-Rampant P, et al (2018) Phenomic selection is a low-cost and high-throughput method based on indirect predictions: Proof of concept on wheat and poplar. G3 Genes, Genomes, Genet 8:3961–3972. https://doi.org/10.1534/g3.118.200760

Samanfar B, Molnar SJ, Charette M, et al (2017) Mapping and identification of a potential candidate gene for a novel maturity locus, *E10*, in soybean. Theor Appl Genet 130:377–390. https://doi.org/10.1007/s00122-016-2819-7

Sebolt AM, Shoemaker RC, Diers BW (2000) Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci 40:1438–1444. https://doi.org/10.2135/cropsci2000.4051438x

Sun Z, Su C, Yun J, et al (2019) Genetic improvement of the shoot architecture and yield in soya bean plants via the manipulation of *GmmiR156b*. Plant Biotechnol J 17:50–62. https://doi.org/https://doi.org/10.1111/pbi.12946

Tian Z, Wang X, Lee R, et al (2010) Artificial selection for determinate growth habit in soybean. Proc Natl Acad Sci U S A 107:8563–8568. https://doi.org/10.1073/pnas.1000088107

Tsubokura Y, Matsumura H, Xu M, et al (2013) Genetic variation in soybean at the maturity locus *e4* is involved in adaptation to long days at high latitudes.

Agronomy 3:117–134. https://doi.org/10.3390/agronomy3010117

Tsubokura Y, Watanabe S, Xia Z, et al (2014) Natural variation in the genes responsible for maturity loci *E1*, *E2*, *E3* and *E4* in soybean. Ann Bot 113:429–441. https://doi.org/10.1093/aob/mct269

Vaughn JN, Nelson RL, Song Q, et al (2014) The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3 Genes, Genomes, Genet 4:2283–2294. https://doi.org/10.1534/g3.114.013433

Wang F, Nan H, Chen L, et al (2019) A new dominant locus, *E11*, controls early flowering time and maturity in soybean. Mol Breed 39:70. https://doi.org/10.1007/s11032-019-0978-3

Wang S, Liu S, Wang J, et al (2020) Simultaneous changes in seed size, oil content and protein content driven by selection of *SWEET* homologues during soybean domestication. Natl Sci Rev 7:1776–1786. https://doi.org/10.1093/nsr/nwaa110

Wang X, Yang Z, Xu C (2015) A comparison of genomic selection methods for breeding value prediction. Sci Bull 60:925–935. https://doi.org/https://doi.org/10.1007/s11434-015-0791-2

Watanabe S, Hideshima R, Zhengjun X, et al (2009) Map-based cloning of the gene associated with the soybean maturity locus *E3*. Genetics 182:1251–1262. https://doi.org/10.1534/genetics.108.098772

Watanabe S, Xia Z, Hideshima R, et al (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. https://doi.org/10.1534/genetics.110.125062

Watson A, Ghosh S, Williams MJ, et al (2018) Speed breeding is a powerful tool to accelerate crop research and breeding. Nat Plants 4:23–29.

https://doi.org/10.1038/s41477-017-0083-8

Weiß TM, Zhu X, Leiser WL, et al (2022) Unraveling the potential of phenomic selection within and among diverse breeding material of maize (*Zea mays* L.). G3 Genes, Genomes, Genet 12(3): jkab445. https://doi.org/10.1093/g3journal/jkab445

Wen Z, Tan R, Zhang S, et al (2018) Integrating GWAS and gene expression data for functional characterization of resistance to white mould in soya bean. Plant Biotechnol J 16:1825–1835. https://doi.org/10.1111/pbi.12918

Wilson RF (2008). Soybean: market driven research needs. In Genetics and genomics of soybean. Springer, New York, NY 3-15.

Xia Z, Watanabe S, Yamada T, et al (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:. https://doi.org/10.1073/pnas.1117982109

Xu M, Xu Z, Liu B, et al (2013) Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. BMC Plant Biol 13:1–14. https://doi.org/10.1186/1471-2229-13-91

Yue Y, Liu N, Jiang B, et al (2017) A Single Nucleotide Deletion in *J* Encoding GmELF3 Confers Long Juvenility and Is Associated with Adaption of Tropic Soybean. Mol Plant 10:656–658. https://doi.org/10.1016/j.molp.2016.12.004

Zhai H, Lü S, Liang S, et al (2014) *GmFT4*, a homolog of *FLOWERING LOCUS T*, is positively regulated by *E1* and functions as a flowering repressor in soybean. PLoS One 9:. https://doi.org/10.1371/journal.pone.0089030

Zhang H, Goettel W, Song Q, et al (2020a) Dual use and selection of GmSWEET39 for oil and protein improvement in soybean. PLoS Genet 16:. https://doi.org/10.1371/JOURNAL.PGEN.1009114

Zhang L xin, Liu W, Tsegaw M, et al (2020b) Principles and practices of the photothermal adaptability improvement in soybean. J Integr Agric 19:295–310. https://doi.org/10.1016/S2095-3119(19)62850-9

Zhang M, Liu S, Wang Z, et al (2021) Progress in soybean functional genomics over the past decade. Plant Biotechnol J 1–27. https://doi.org/10.1111/pbi.13682

Zhao C, Takeshima R, Zhu J, et al (2016) A recessive allele for delayed flowering at the soybean maturity locus *E9* is a leaky allele of *FT2a*, a *FLOWERING LOCUS T* ortholog. BMC Plant Biol 16:1–15. https://doi.org/10.1186/s12870-016-0704-9

Zhu X, Leiser WL, Hahn V, Würschum T (2020) Identification of seed protein and oil related QTL in 944 RILs from a diallel of early-maturing European soybean. Crop J 9:238–247. https://doi.org/10.1016/j.cj.2020.06.006

Zhu X, Leiser WL, Hahn V, Würschum T (2021a) Identification of QTL for seed yield and agronomic traits in 944 soybean (*Glycine max*) RILs from a diallel cross of early-maturing varieties. Plant Breed 140:254–266. https://doi.org/10.1111/pbr.12900

Zhu X, Leiser WL, Hahn V, Würschum T (2021b) Training set design in genomic prediction with multiple biparental families. Plant Genome 14:1–15. https://doi.org/10.1002/tpg2.20124

Zhu X, Leiser WL, Hahn V, Würschum T (2021c) Phenomic selection is competitive with genomic selection for breeding of complex traits. Plant Phenome J 4:1–21. https://doi.org/10.1002/ppj2.20027

Zhu X, Maurer HP, Jenz M, et al (2022) The performance of phenomic selection depends on the genetic architecture of the target trait. Theor Appl Genet 135:653–665. https://doi.org/10.1007/s00122-021-03997-7

Zhu X, Leiser WL, Hahn V, Würschum T (submitted) The genetic architecture of soybean photothermal adaptation to high latitudes.

List of all publications:

Zhu X, Leiser WL, Hahn V, Würschum T (2020) Identification of seed protein and oil related QTL in 944 RILs from a diallel of early-maturing European soybean. The Crop Journal 9:238–247. https://doi.org/10.1016/j.cj.2020.06.006

Zhu X, Leiser WL, Hahn V, Würschum T (2021a) Identification of QTL for seed yield and agronomic traits in 944 soybean (*Glycine max*) RILs from a diallel cross of early-maturing varieties. Plant Breeding 140:254–266. https://doi.org/10.1111/pbr.12900

Zhu X, Leiser WL, Hahn V, Würschum T (2021b) Training set design in genomic prediction with multiple biparental families. Plant Genome 14:1–15. https://doi.org/10.1002/tpg2.20124

Zhu X, Leiser WL, Hahn V, Würschum T (2021c) Phenomic selection is competitive with genomic selection for breeding of complex traits. Plant Phenome Journal 4:1–21. https://doi.org/10.1002/ppj2.20027

Zhu X, Maurer HP, Jenz M, et al (2022) The performance of phenomic selection depends on the genetic architecture of the target trait. Theoretical and Applied Genetics 135:653–665. https://doi.org/10.1007/s00122-021-03997-7

Weiß TM, **Zhu X**, Leiser WL, et al (2022) Unraveling the potential of phenomic selection within and among diverse breeding material of maize (*Zea mays* L.). G3 Genes, Genomes, Genetics 12(3): jkab445. https://doi.org/10.1093/g3journal/jkab445

Zhu X, Leiser WL, Hahn V, Würschum T (submitted) The genetic architecture of soybean photothermal adaptation to high latitudes.

Shen E, Chen T**, Zhu X,** et al (2020), Expansion of *MIR482/2118* by a class-II transposable element in cotton. Plant Journal, 103: 2084-2099. https://doi.org/10.1111/tpj.14885

List of publications included in doctoral thesis:

Zhu X, Leiser WL, Hahn V, Würschum T (2020) Identification of seed protein and oil related QTL in 944 RILs from a diallel of early-maturing European soybean. The Crop Journal 9:238–247. https://doi.org/10.1016/j.cj.2020.06.006

Zhu X, Leiser WL, Hahn V, Würschum T (2021a) Identification of QTL for seed yield and agronomic traits in 944 soybean (*Glycine max*) RILs from a diallel cross of early-maturing varieties. Plant Breeding 140:254–266. https://doi.org/10.1111/pbr.12900

Zhu X, Leiser WL, Hahn V, Würschum T (2021b) Training set design in genomic prediction with multiple biparental families. Plant Genome 14:1–15. https://doi.org/10.1002/tpg2.20124

Zhu X, Leiser WL, Hahn V, Würschum T (2021c) Phenomic selection is competitive with genomic selection for breeding of complex traits. Plant Phenome Journal 4:1–21. https://doi.org/10.1002/ppj2.20027

Zhu X, Maurer HP, Jenz M, et al (2022) The performance of phenomic selection depends on the genetic architecture of the target trait. Theoretical and Applied Genetics 135:653–665. https://doi.org/10.1007/s00122-021-03997-7

Weiß TM, **Zhu X**, Leiser WL, et al (2022) Unraveling the potential of phenomic selection within and among diverse breeding material of maize (*Zea mays* L.). G3 Genes, Genomes, Genetics 12(3): jkab445. https://doi.org/10.1093/g3journal/jkab445

Zhu X, Leiser WL, Hahn V, Würschum T (submitted) The genetic architecture of soybean photothermal adaptation to high latitudes.

Acknowledgments

Time always goes by so quickly. I feel like I have just arrived in Germany and now I am writing these Acknowledgements of my PhD thesis. The past three years in Hohenheim are a fantastic part of my life and a lot of wonderful memories will be kept in my mind forever. When I look back now, I feel like I had a very long, but a fabulous dream being in Germany. There were so many first experiences and so many people I have never thought about to have and to meet.

The first person I want to thank is my supervisor Prof. Dr. Tobias Würschum. Thanks for accepting me as PhD student in Hohenheim and let me join your lab. Your enthusiasm and attitude to science inspired me a lot. You were always there to help me when I had some questions and explained everything patiently and clearly to me and provided me new ideas. Your optimism also motives me and makes me believe that my research can contribute a little progress to solving specific scientific question.

Then I want to thank my second supervisor, Dr. Willmar L. Leiser. Thanks for your suggestions and ideas, often thinking from the perspective of a breeder which taught me a lot. Also thank you for your help with some analyses. Also, I want to thank Dr. Volker Hahn. Thanks for your excellent field work that produced the fantastic dataset and for your fast replies whenever I needed your help. Thank you, Dr. Hans Peter Maurer, you are very kind and helped me solve any technical problem and guaranteed my efficient work on the servers. Thank you also for to the triticale dataset and your creative input. In addition, I want to thank associate Prof. Wenxin Liu from China Agriculture University, it was my pleasure to meet you in Hohenheim. Thank you for caring about my life abroad and for all the support.

Also, thanks to all my colleagues from the State Plant Breeding Institute and the Institute of Plant Breeding, thanks to all of you! Dr. Volker Hahn, Dr. Willmar L. Leiser, apl. Prof. Dr. C. Friedrich H. Longin, apl. Prof. Dr. Thomas Miedaner, Dr. Kim Steige, Prof. Tobias Würschum. Thanks to Mrs. Kurka and Mrs. Kösling for help with administrative and life matters. Thanks to all PhD students and friends, Muhammad Afzal, Dr. Khaoula EL Hassouni, Lea Schwarzwälder, Dr. David Sewordor Gaikpa, Dr. Ana L. Galiano-Carneiro, Dr. Rodrigo José Galán, Dr. Ana Kodisch, Paul Gruner, María Belén Kistner, Félicien Akohoue, Johannes Trini, Jan Neuweiler, Dr. Felix Jähne, Dr. Alena K. Kurasch, Dr. Patrick Thorwarth, Thea Mi Weiß, Cleo Döttinger, Sandra Roller and all unmentioned colleagues. The time with you was full of fun and is unforgettable.

Last but not least, I want to thank my family, who supported and encouraged me a lot. Especially my brother, who always listens to me and kindly gives his suggestion and supports me all the time.

Curriculum vitae

Name:	Xintian Zhu
Date and place of birth:	22th April 1995 in Henan, China

Education

07/2019-12/2022	PhD candidate of Plant breeding Institute of Plant Breeding, Seed Science and Population Genetics (350a), University Hohenheim, Stuttgart, Germany
09/2016-03/2019	Master of Bioinformatics College of Agricultural and Biotechnology, Zhejiang University, Hangzhou, China
09/2012-06/2016	Bachelor of Plant Science and Technology College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China

8th June 2022, Hohenheim

Xintian Zhu

Xintian Zhu

Annex 3

Declaration in lieu of an oath on independent work

according to Sec. 18(3) sentence 5 of the University of Hohenheim's Doctoral Regulations for the Faculties of Agricultural Sciences, Natural Sciences, and Business, Economics and Social Sciences

1. The dissertation submitted on the topic

Assessment of phenotypic, genomic and novel approaches ....

for soybean breeding in Central Europe ....

is work done independently by me.

2. I only used the sources and aids listed and did not make use of any impermissible assistance from third parties. In particular, I marked all content taken word-for-word or paraphrased from other works.

3. I did not use the assistance of a commercial doctoral placement or advising agency.

4. I am aware of the importance of the declaration in lieu of oath and the criminal consequences of false or incomplete declarations in lieu of oath.

I confirm that the declaration above is correct. I declare in lieu of oath that I have declared only the truth to the best of my knowledge and have not omitted anything.

Hohenheim , 08. 06. 2022

Place, Date

Xintian Zhu

Signature