Combining Focused Identification of Germplasm and Core Collection Strategies to Identify Genebank Accessions for Central European Soybean Breeding

Max Haupt and Karl Schmid[1]

Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany

Funding source: BMEL Grant number FKZ 2814EPS012

ABSTRACT

Environmental adaptation of crops is essential for reliable agricultural production and an important breeding objectives. Genbanks provide genetic variation for the improvement of modern varieties, but the selection of suitable germplasm is frequently impeded by incomplete phenotypic data. We address this bottleneck by combining a *Focused Identification of Germplasm Strategy* (FIGS) with core collection methodology to select soybean (*Glycine max*) germplasm for Central European breeding from a collection of >17,000 accessions. By focussing on environmental adaptation to high-latitude cold regions, we selected an 'environmental precore' of 3,663 accessions using environmental data and compared the Donor Population of Environments (DPE) in Asia and the Target Population of Environments (TPE) in Central Europe in the present and in 2070. Using SNP genotypes we reduced the precore into two diverse core collections of 183 and 366 of accessions as diversity panels for evaluation in high-latitude cold regions. Tests of genetic differentiation between precore and core collections revealed differentiation signatures in genomic regions that control maturity, and novel candidate loci for environmental adaptation demonstrating the potential of diversity panels for studying environmental adaptation. Objective-driven core collections increase germplasm utilization for abiotic adaptation by breeding for a rapidly changing climate, or *de novo* adaptation of crop species to expand cultivation ranges.

Keywords: Soybean, adaptation, core collection, plant breeding, genetic diversity, genebank

[1] Correspondence: karl.schmid@uni-hohenheim.de

INTRODUCTION

Many crop species are cultivated outside their center of domestication (Diamond 2002). An expansion of the cultivation range requires adaptation to local biotic and abiotic environmental conditions, which can be limited by reduced genetic variation as a result of variation ’left behind’ in the center of origin due to founder effects (Tanksley and McCouch 1997; Hyten et al. 2006). Nevertheless, numerous crops were successfully adapted to novel environments by natural and human selection that contributed to the rapid expansion of farming beyond centers of domestication (Harlan 1992). In current plant breeding programs, reduced levels of genetic variation in elite breeding populations limit breeding progress and the adaptation to rapidly changing environmental conditions. Plant genetic resources stored in large *ex situ* germplasm collections are an important source of novel, potentially useful genetic variation to achieve these goals (Wang et al. 2017a).

The efficient utilization of plant genetic resources by breeders and researchers is frequently limited by incomplete phenotypic data for traits of interest, which makes the targeted selection of putative useful accessions from large germplasm collections very challenging. The *Focused Identification of Germplasm Strategy* (FIGS) promises a solution to this phenotyping bottleneck for traits associated with environmental adaptation. FIGS is based on the premise that native landraces evolved during centuries of cultivation in response to local eco-geographic conditions which led to local adaptation to biotic and abiotic conditions. Genomic footprints of this adaptation should be detectable if environmental parameters describing plant germplasm collection sites are used as selection criteria to ’focus in’ on a subset of accessions that potentially harbor phenotypic variation in a target trait. The comprehensive phenotypic and genomic evaluation of focused subsets is more resource efficient than of large collections (Street et al. 2008; Sanders et al. 2013). FIGS has been successfully used in a number of crops including *Hordeum vulgare* for agro-morphological characteristics and time of heading and maturity (Endresen 2010), *Triticum aestivum* for resistances to powdery mildew (Bhullar et al. 2009) and stem rust (Bari et al. 2012), and *Vicia faba* for drought tolerance (Khazaei et al. 2013). A second tool in the characterization of germplasm collections are core collections. They are small subsets of larger collections that maximize genetic and phenotypic diversity (Frankel 1984). Similar to trait-focused subsets, core collections provide a resource efficient means for further evaluation and utilization and have been established for many crops based on passport information, agronomic and morphological data as well as genetic diversity to serve varying objectives (Odong et al. 2013).

The soybean (*Glycine max*) is a typical example of a crop whose cultivation expanded outside its center of origin. It was originally domesticated in China and is still the most important legume crop throughout East Asia until today. It was first introduced to Europe and the US in the 18th century. During the 20th century it became a major crop in the Americas, and is currently the most widely cultivated legume crop of the world. In Central Europe (CEU), the area of soybean cultivation has been constantly increasing over the last decade and has revived the interest in breeding improved varieties for this production environment. This development is driven by the motivation to reduce dependencies from soy imports, to re-establish legumes in crop rotations for a more sustainable agriculture, and to mitigate the consequences of climate change by cultivating thermotolerant crops (BMEL 2016; European Soya Declaration 2017).

The adaptation of soybean to Central European production environments is an ongoing effort. So far, breeding research mainly focused on characterizing adaptive phenotypic variation of elite breeding germplasm by investigating the genetic basis of time to flowering and maturity (Kurasch et al. 2017). A portfolio of CEU cultivars that are classified into maturity groups (MGs) exists. In analogy to the system used in Northern America, MGs express the suitability of a variety for cultivation within a certain latitudinal range to ensure a full utilization of the vegetation period and timely maturation. Most cultivars suitable for production in Central Europe belong to the earliest MGs *000-0* and are mainly derived from North American material (Hahn and Würschum 2014). This narrow genetic basis (Gizlice et al. 1994) can limit long-term breeding progress and the utilization of genetic resources may alleviate this situation.

In addition to photoperiod, temperature and daily irradiation influence soybean development. Cold tolerance ensures that the vegetation period in high-latitude cold regions is fully exploited (Kurasch et al. 2017; Balko et al. 2014), but the trait has received little attention in soybean research so far. Flowering is the most susceptible developmental stage for low temperature, which causes flower and pod abscissions (Gass et al. 1996). Poor growth in early development and insufficient grain filling during pod development also reduce grain yield (Littlejohns and Tanner 1976). Phenotyping for cold tolerance, especially at flowering stage is laborious and a reliable exposure to cold stress requires controlled conditions, which limited the diversity of germplasm that has been evaluated for this trait (Funatsuki et al. 2004; Cober et al. 2013; Balko et al. 2014). Previous studies indicated a polygenic basis of cold tolerance and a potential overlap with the *E* series loci (Funatsuki et al. 2005) that are responsible for adaptation to photoperiod and soybean diffusion across latitudes (Jiang et al. 2014). Central European varieties exhibit variation in cold tolerance (Balko et al. 2014) and are used for improving cold tolerance in breeding of varieties for Northern Japan (Yamaguchi et al. 2015). Additional useful variation for cold tolerance and adaptation to high-latitude regions may be present in Asian landraces that are conserved in *ex situ* germplasm collections.

In this study, we present the development of two core collections of the USDA Soybean Germplasm Collection (Nelson 2011) with a focus on environmental adaptation to high-latitude cold regions to support the breeding of varieties for cultivation in Central Europe. Following the FIGS approach, we used environmental data characterizing the geographic origin of landraces and other germplasm in combination with current and future environmental conditions of the Central European *Target Population of Environments*. First, a precore collection consisting of accessions potentially adapted to CEU was formed based on environmental data. A scan for genetic differentiation between precore accessions and non-selected germplasm identified selection signals in genomic regions known to be involved in environmental adaptation and thus confirmed that FIGS enriched the precore for adaptive genetic variation. From the environmentally defined precore, we constructed genetically diverse, but substantially smaller core collections based on marker information. These may be used by researchers and breeders for comprehensive phenotypic characterization to facilitate trait discovery for Central European production environments. We show that a combination of FIGS and core collection methodology to define core subsets driven by targeted objectives will likely contribute to a more rapid and efficient utilization of soybean genetic resources.

MATERIALS & METHODS

Plant Material and the Donor Population of Environments

The *G. max* accessions originated from the USDA Soybean Germplasm Collection and only the *Introduced G. max* sub-collection ($N > 17,000$) was considered, because it includes landraces (Nelson 2011; Song et al. 2015). Out of 6,180 accessions with georeferences we retained 5,886 Asian accessions by excluding material collected west of 60°E to remove accessions outside the original soybean cultivation range. The collection sites of the remaining accessions constitute the *Donor Population of Environments* (DPE) of Asian landraces, which may include potentially useful germplasm for CEU soybean breeding.

Target Population of Environments

The *Target Population of Environments* (TPE) is the set of environments to which germplasm needs to be adapted for a successful cultivation. Our TPE consists of the soybean production environments in Central Europe (CEU) and includes Germany, Austria, Switzerland, Poland, Czech Republic and Slovakia. To characterize the TPE, we georeferenced locations that previously hosted soybean variety trials (Recknagel 2015; Kurasch et al. 2017). These served as proxies for the geographic extent of soybean cultivation in CEU since no detailed records regarding soybean cultivation exist for this region.

Environmental Data

Based on the georeferences of the curated collection, environmental data was retrieved from the `WorldClim` (version 2) database (Fick and Hijmans 2017) at a resolution of 30 arc-seconds ($\approx 1\ \text{km}^2$). We used variables that are informative for the soybean cropping season in the northern hemisphere which lasts from May to September. Additionally, we modified the bioclimatic variables to render them informative for the soybean cropping season: Instead of the *annual mean temperature* (`BIO1`) we computed the *seasonal mean temperature* as the average mean temperature from May to September. Likewise we proceeded with `BIO2` to `BIO4`, `BIO6`, `BIO7` and `BIO12` to `BIO15`. `BIO8` to `BIO11` (*mean temperature of wettest / driest / warmest / coldest quarter*) were replaced by *mean temperatures of wettest / driest / warmest / coldest month* between May and September. `BIO16` to `BIO19` (*precipitation of wettest / driest / warmest / coldest quarter*) were replaced by monthly substitutes. An overview of the data used is provided in Table S1.

We also calculated monthly sums of *Crop Heat Units* (CHU) to quantify the accumulation of temperature over the growing season. CHUs are commonly used in Canada to rate the regional suitability of warm-season crop varieties with differing thermal requirements (Bootsma et al. 2007) and were also adopted by Central European soybean growers. Average daily CHU were computed from monthly averages of daily maximum and minimum temperatures according to Brown and Bootsma (1993):

$$Y_{max} = 3.33 \times (T_{max} - 10) - 0.084 \times (T_{max} - 10.0)^2 \tag{1}$$

$$Y_{min} = 1.8 \times (T_{min} - 4.4) \tag{2}$$

$$\text{CHU}_{\text{daily}} = \frac{Y_{max} + Y_{min}}{2} \tag{3}$$

where T_{max} and T_{min} refer to the monthly averages of daily maximum and minimum temperatures, respectively. Negative Y_{max} and Y_{min} values were set to zero. A similar dataset was assembled for the TPE considering climate data for both current and future conditions. Climate projections for the year 2070 were retrieved from `WorldClim` version 1.4 (Hijmans et al. 2005) for the greenhouse gas scenario RCP8.5 as modeled by the *Community Climate System Model* (CCSM4.0). Available projections included average monthly climate data for minimum and maximum temperature, precipitation and the bioclimatic variables, and were used to compute derived variables as outlined above. Environmental qualities without projections for the year 2070 were substituted by their current equivalents to preserve the dimensional consistency of the datasets. Latitude information for the collection sites was included as additional environmental variable to account for the rather narrow adaptation of soybean germplasm to photoperiod.

Environmental Characterization and Selection of Precore Accessions

Principal component analysis (PCA) was performed to summarize the high dimensional datasets of the DPE and the TPE separately using the `dudi.pca()` function from `R ade4` (Dray and Dufour 2007). Variables were mean centered and normalized. Correlations (i.e. the loadings) and squared loadings between the original variables and the principal components were used to quantify the amount of shared information and the contribution of single variables to the components. Environmental data for the TPE (current and projected for 2070 climate) was used to project Central European soybean growing environments as illustrative observations into the DPE's multi-environmental space. The position of a collection site in the multivariate space was then used as 'environmental profile' associated with accessions of corresponding descent. Accessions with an environmental origin similar to environments of the TPE along the first two principal components were subsequently included in the precore to form a group of promising accessions with potential abiotic adaptation to the TPE.

Genotypic Data

The USDA Soybean Germplasm Collection has been previously genotyped with the SoySNP50K array (Song et al. 2013; Song et al. 2015) and genotypes were downloaded from *SoyBase* (`https://soybase.org/snps/`; May 16, 2017) in the version mapped to the second *G. max* genome assembly "Glyma.Wm82.a2" (Schmutz et al. 2010). Non-chromosomal SNPs were removed from the genotype dataset and missing data was imputed using `BEAGLE` version 4.1 with default settings (Browning and Browning 2016).

Phenotypic Data

Phenotypic data for the *Introduced G. max* sub-collection of the USDA Soybean Germplasm collection (Oliveira et al. 2010; Nelson 2011) was retrieved from `https://npgsweb.ars-grin.gov/gringlobal/search.aspx`. Data included qualitative and quantitative trait data like maturity group, 100 seed weight, seed yield, protein and oil contents. We used the quantitative trait data to assess changes in phenotypic performance and variance in the selection of the core collection. Homogeneity of variances in groups was tested using the modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median as implemented in `R lawstat` version 3.2 (Hui et al. 2008). Significance of differences of the means between groups was assessed with Welch's unequal variances t-test implemented in `R` version 3.4.3 with Bonferroni correction for multiple testing.

Core Sampling

Core sampling is based on reducing the number of available accessions by means of redundancy reduction through penalizing core subset solutions with too many too similar accessions. The large-scale genotyping of germplasm collections revealed the magnitude of identical and highly similar accessions in the USDA Soybean Germplasm Collection (Song et al. 2015) and other crops.

Core collections were assembled by analysing genotype data with `R Core Hunter` version 3.2 (`http://www.corehunter.org/`). All accessions that qualified after the environmental characterization and/or had a maturity group rating from 000 to I were considered for core sampling. We used `Core Hunter` in default execution mode, i.e. using the advanced parallel tempering search algorithm, and fixed the maximum number of steps without improvement to 100 (Beukelaer et al. 2017a; Beukelaer et al. 2017b) to efficiently select core subsets from the population of all possible cores (Thachuk et al. 2009). We explored different optimization strategies to identify the approach most suitable to our data and objectives: Sampling objectives regarding (1) optimization (i.e. maximization) of allelic diversity included the parameters expected heterozygosity and Shannon's diversity index; (2) maximization of genetic distance based on the average entry-to-nearest-entry distance (Beukelaer et al. 2012) using the Modified Roger's distance (MR) between accessions; (3) optimization of the auxiliary parameter allele coverage was performed the proportion of alleles retainable in core collections. For the definition of these measures the reader is referred to Thachuk et al. (2009). Core sizes of 5%, 10%, 15% and 20% of the input number of accessions were cross-evaluated for all three optimization objectives. Additionally, the simultaneous optimization of expected heterozygosity and MR was examined with varying weights to estimate the trade-off between both approaches. The measures expected heterozygosity, MR and allele coverage were also assessed for all sets used in this study to evaluate the success of the sampling process.

We also explored the effect of different sampling strategies on the subsequent utilization of the core collection: (1) Cores were sampled without any stratification; (2) they were sampled by means of classic stratification according to maturity, i.e. cores were sampled separately within three groups of accessions corresponding to maturity group ratings 000 to 0, I and II to X. A final core was then obtained by merging the result from these three groups; (3) a more refined 2-fold pseudostratification sampling strategy was devised and consisted in first sampling within accessions of maturity groups 000 to 0 and adopting the result of this first run as fixed in a second and final sampling run from all accessions, and (4) a 3-fold pseudostratification strategy first sampled within maturity groups 000 to 0, too. The second run then was limited to maturity groups 000 to I while fixing the result of the first run. In a third and final run, sampling was performed from all accessions while fixing the result from the second run. The general rationale behind all stratification strategies was primarily to retain accessions in the cores relative to their input group sizes by mitigating the effect of varying levels of genetic similarities within groups on the core sampling process. Table S4 provides a more detailed overview of the four core sampling strategies. For all reported cores the random seed was fixed to one prior to sampling for the sake of reproducibility of the core subset selection. To validate the stability of the results, re-sampling was performed for five additional independent runs with random seeds to rule out convergence effects.

Screening for Signals of Adaptation

To identify genomic regions that reflect adaptive genetic differentiation between precore and non-selected accessions we used the basic model of `BAYPASS` (Gautier 2015) for allele count data. Only marker data for loci with minor allele frequencies $\geq$0.01 in the *Introduced G. max* subcollection was considered. `BAYPASS` calculates the XtX statistic that may be seen as a SNP-specific F_{ST} and accounts for the variance–covariance structure of the groups for population structure correction (Günther and Coop 2013). Significance of the observed XtX signals was assessed from allele count data for 1,000 SNPs sampled randomly from a pseudo-observed dataset (POD). The POD was constructed with the R function `simulate.baypass()` based on the covariance matrix of population allele frequencies (Ω) and other properties of the original data (Gautier 2015). `BAYPASS` was then run on the simulated data to produce an empirical distribution of XtX estimates and the 99% quantile value was used as threshold to distinguish selection from neutrality.

RESULTS

Curation of Collection Sites

Of 6,180 USDA soybean accessions with georeferenced collection sites and `WorldClim` data, 5,886 fell within our definition of the DPE. These originated from 699 unique collection sites, constituting an average contribution of $\approx$8 accessions per site. The distribution of accession numbers per collection site (Fig. S1) revealed several locations that contributed a large number (up to 619) of accessions to the collection, suggesting that multiple accessions from a region were aggregated into a single georeference. To purge the data set of accessions with unreliable geographic information, we checked locations with $\geq$20 accessions and found that georeferences were partly assigned based on provenance from cities, districts or even provinces (`https://npgsweb.ars-grin.gov/gringlobal/search.aspx?`). We removed 1,704 accessions (29%) from ten collection sites that were considered too large for a reliable inference of environmental adaptation (Fig. S1).

Identification of Candidate Accessions for the TPE

By analyzing climate data from the geographic origin of accessions, we aimed at selecting accessions pre-adapted to cool and high latitude environments of Central Europe, and to identify the selective environmental factors acting during the soybean cropping season from May to September that characterize the DPE and TPE. Most environmental variables were strongly correlated with each other (Fig. S2 - S4). A PCA summarized 71.8% of the total environmental variation among collection sites in the first two principal components (PCs) (Fig. 1). Latitude and temperature based variables were strongly correlated with the first PC, while precipitation based variables related to the first and the second PC (Fig. S5 and S6). The first PC separated cooler from warmer regions of origin. To identify locations with accessions suitable for cultivation in Europe, we projected the European soybean cultivation environments onto the multivariate DPE to relate the TPE to the DPE (Polygons in Fig. 1). The projection shows that current CEU environments cluster in a narrow environmental range of less than 3,000 CHUs that includes a small proportion of provenances. Climate projections for 2070 indicate an average increase of more than 500 CHUs during the soybean cropping season in the TPE (Fig. S7). Even these conditions in the TPE exclude the majority of accessions because they originate from warmer regions with >3,500 CHUs. But the overlap of DPE and TPE with respect to temperature and other environmental variables increased and included North-Eastern China, Northern Japan, the Himalayas and Russia. The latter regions also contributed accessions from areas that appear to be highly unfavorable for soybean cultivation. To select potentially suitable accessions for cultivation in CEU, we included 123 environments with a position of ≥ 5 on the first PC (grey dashed line in Fig. 1) that provided 523 accessions with potential beneficial variation for abiotic adaptation to CEU (Fig. 2).

Only ≈20% of accessions from the USDA *Introduced G. max* collection were georeferenced and included in the above analysis of environmental variation in the DPE. However, the complete collection was categorized for maturity group, which is a proxy for the photo-thermal requirements of accessions and georeferenced accessions show a strong correlation between CHUs and their maturity group classification (Fig. S11). Accessions from early maturity groups originate from en-

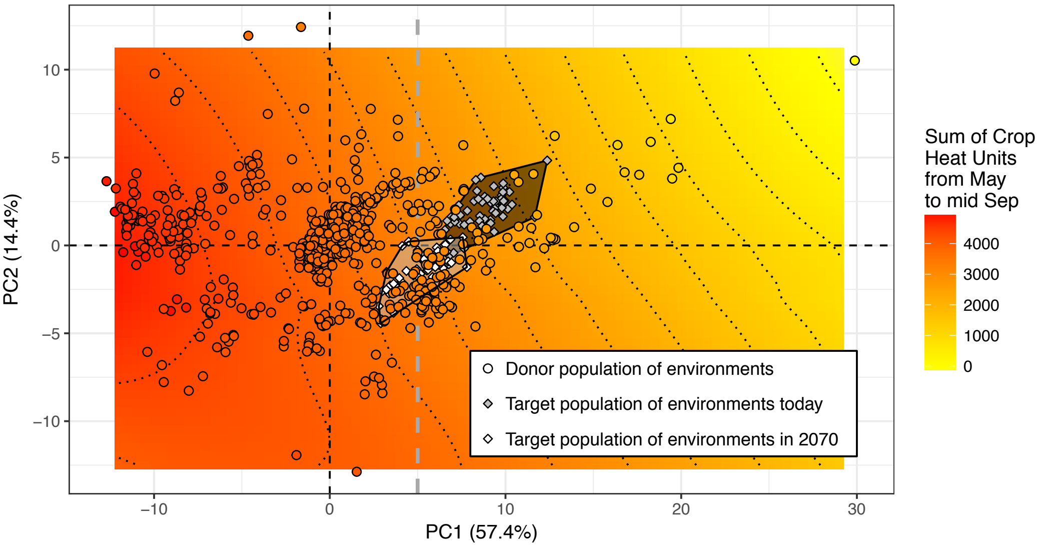

Fig. 1. Principal component analysis of environmental variables obtained from *G. max* collection sites to characterize the donor population of environments (DPE). The color of collection sites corresponds to the sum of Crop Heat Units (CHUs) of the soybean cultivation season from May to mid September as a site-specific estimate of available temperatures. Contour lines represent bins of 400 CHUs. Current (grey) and future (white) climatic conditions in Central European soybean growing environments are included as illustrative observations. The environmental scopes of Central European scenarios (today and in 2070) are indicated with polygons.

vironments that resemble CEU e.g. in terms of temperature while late maturing accessions originate from warmer regions. We also included accessions without georeferences and no environmental data, but were classified as early maturing (maturity groups 000-I) and potentially suitable for cultivation in CEU. By combining accessions selected based on environmental data or maturity group classifications, we constructed an ’environmental precore collection’ of 3,663 accessions (Tables 3 and 4).

Genomic footprints of environmental adaptation

To test whether soybean accessions are locally adapted to their original cultivation environment, we searched for genomic regions with allele frequency differences between the 3,663 environmental precore accessions and the remaining 13,354 USDA *Introduced G. max* accessions. The BAYPASS program identified a total of 52 genomic regions with a stronger genetic differentiation (i.e. higher

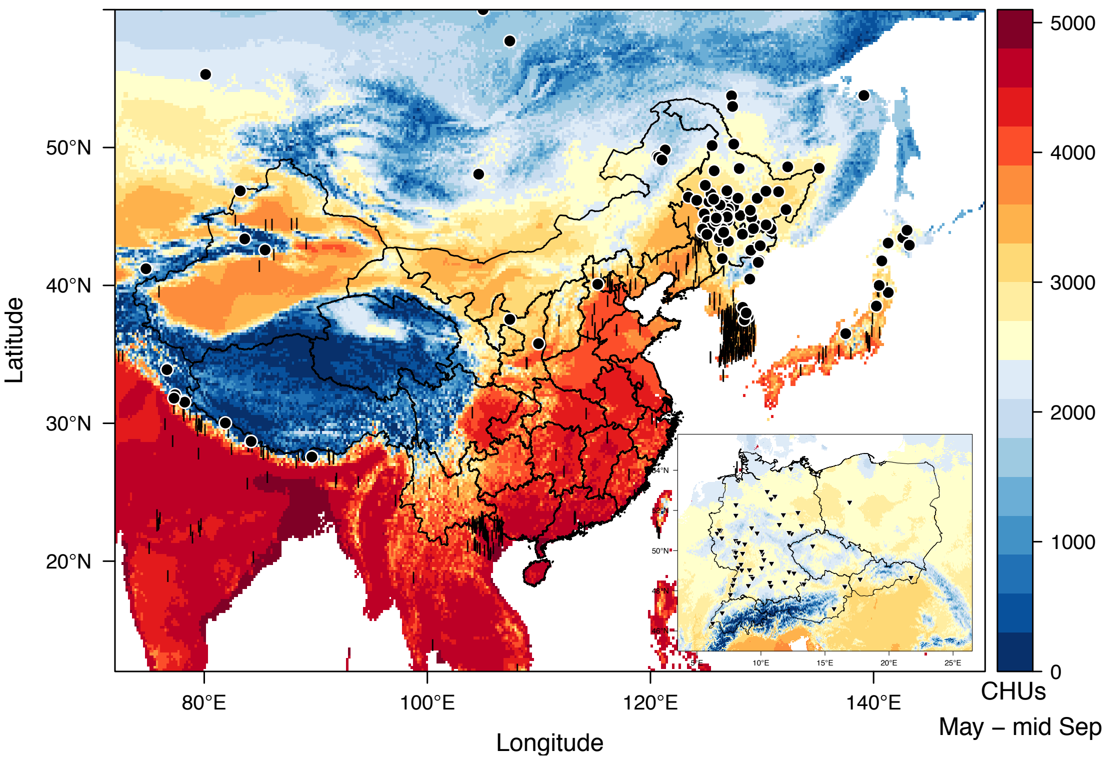

Fig. 2. Collection sites of *G. max* germplasm in the *Donor Population of Environments*: Filled dots represent sites that were identified as donors of candidate accessions for CEU. Vertical lines represent sites which were not selected to contribute germplasm. Colouring represents estimates of available termperatures during the soybean season in Crop Heat Units. The inset shows the *Target Population of Environments* (current situation) with triangles indicating the extent of soybean cultivation in CEU based on soybean variety testing locations.

XtX values) than expected given the population structure (Fig. 3). These regions likely reflect genetic differentiation due to local adaption because they harbor several well characterized maturity genes of soybean, including *E1* - *E3* on chromosomes 6, 10 and 19, which are three out of four cloned genes known to be major determinants of soybean adaptation to latitudinal zones (Jiang et al. 2014). On chromosome 19, a strongly differentiated region harbors *Dt1*, which is an ortholog of the *Arabidopsis thaliana TFL1* gene. In soybean, this gene controls the agronomically important trait indeterminacy and affects the length of flowering and reproductive period, which impacts plant height and maturity (Liu et al. 2010). The strongest differentiation occurs in a genomic region that contain the *Pdh1* and *GmFT2a* genes. The *pdh1* allele conveys shattering resistance and has been hypothesized crucial for soybean adaptation to Northern and semi-arid environments as opposed to humid South Asian environments where selection on pod dehiscence was more relaxed (Funatsuki et al. 2014). *GmFT2a* is a homolog of *Arabidopsis FLOWERING LOCUS T* (Sun et al. 2011) that has been identified as soybean maturity locus *E9* (Zhao et al. 2016). Additional significantly differentiated genomic regions harbor genes for which a role in abiotic adaptation of soybean has been demonstrated or postulated (Tab. 1 and S2). Phenotypic traits controlled by these genes include adaptation to photoperiod, heat, drought and cold stress. The limited marker density of the SoySNP50K array did not allow a fine-scaled mapping of differentiation signals, but in summary provides evidence that the environmental precore collection reflects such an adaptation. It also illustrates the possibility for mining the precore accessions for further traits that might not yet have been introduced into CEU breeding but could benefit soybean cultivation in cold and high-latitude environments. We used loci known to control environmental adaptation (i.e. *Dt1*, *E1-E3*, *E9* and *Pdh1*) to test whether genetic differentiation caused by local adaptation is better identified by environmental parameters or maturity group ratings. In general, genome scans comparing early maturing accessions (MG 000 to I) to late material (MG II to X) and comparing accessions of CEU-like origin to accessions of non CEU-like origin produced very similar results (Fig. S16). One exception was the *E1* locus region that exhibited a weaker differentiation signal in the BAYPASS run with accessions grouped solely based on environmental data as accessions of CEU-like origin were not necessarily of early maturity. This may indicate imprecise or incorrect georeferencing, which puts late maturing accessions in the wrong environments. On the other hand, these accessions might have acquired unique adaptation strategies that would not be identified by solely focusing on early maturing accessions. Therefore we decided to retain the respective accessions in the precore.

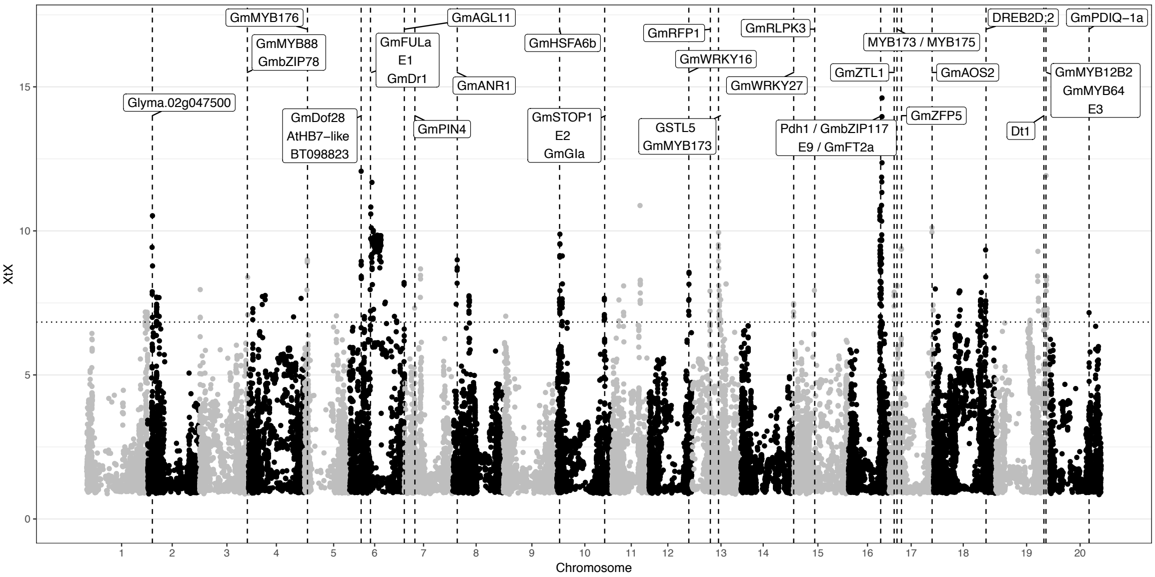

Fig. 3. Genome-wide differentiation between precore and non-candidate accessions estimated with BAYPASS. Significantly differentiated regions are located above the dotted horizontal line which represents the POD 99% quantile of *XtX* values. Labels indicate genes that are located within significantly differentiated regions and known or hypothesized to be involved in abiotic adaptation and which are located within significantly differentiated regions. Chromosome-level plots are provided in Fig. S12 - S15. Candidate genes and their function are listed in Tab. 1 and Tab. S2.

TABLE 1. Overview of genomic regions significantly differentiated between the environmental precore and remaining accessions of the USDA *Introduced G. max* collection with proximity to known or hypothetical genes for abiotic adaptation. Gene functions are provided in Tab. S2.

Marker	XtX	Chr	Pos	*N* genes ±1Mb	Candidate gene*	Gene pos
ss715580458	7.18	1	54,584,027	23	GmANS2	54,530,596 : 54,532,594
					GmANS3	54,530,547 : 54,532,598
					GmSGR2	54,554,346 : 54,556,366
ss715582735	10.53	2	4,482,526	20	Glyma.02g047500	4,373,609 : 4,374,901
ss715586578	8.39	3	44,888,926	22	GmMYB88	44,634,162 : 44,635,400
					GmbZIP78	45,015,664 : 45,021,621
ss715587948	7.30	4	4,043,556	9	GmHSFA2	4,216,891 : 4,218,094
ss715592648	9.99	5	2,624,122	23	GmMYB176	2,802,638 : 2,804,766
ss715590444	7.05	5	29,852,798	4	GmPM29	29,765,494 : 29,766,379
ss715592720	12.07	6	10,851,807	24	GmDof28	10,834,411 : 10,835,873
					AtHB7-like	11,001,256 : 11,002,690
					BT098823	11,087,031 : 11,088,904
ss715593866	11.68	6	20,940,014	4	GmFULa	19,584,069 : 19,590,900
					E1	20,207,077 : 20,207,940
					GmDr1	20,530,825 : 20,534,411
ss715595268	8.21	6	50,979,817	8	GmAGL11	51,206,358 : 51,214,568
ss715599060	7.32	7	9,552,345	8	GmPIN4	9,785,659 : 9,788,973
ss715602532	8.99	8	4,746,918	27	GmANR1	4,783,892 : 4,786,796
ss715606091	9.89	10	2,836,780	19	GmHSFA6b	2,576,457 : 2,577,884
ss715607376	7.66	10	44,420,445	22	DQ075204	44,392,255 : 44,401,221
					GmSTOP1	44,733,937 : 44,735,540
					E2 / GmGIa	45,294,735 : 45,316,121
ss715612745	8.56	12	37,306,504	30	GmWRKY16	37,131,311 : 37,133,954
ss715616725	7.91	13	16,852,674	9	GmRFP1	17,176,807 : 17,178,478
ss715614138	9.95	13	24,850,013	7	GSTL5	24,798,696 : 24,801,446
ss715614348	7.61	13	26,865,258	9	GmMYB173	26,688,360 : 26,690,995
ss715621196	7.48	15	194,375	1	GmWRKY27	290,660 : 292,140
ss715621150	7.93	15	19,649,749	8	GmRLPK3	19,977,272 : 19,982,301
ss715624069	10.76	16	29,256,979	17	Pdh1	29,960,568 : 29,967,155
					GmbZIP117	29,960,723 : 29,966,796
ss715624382	14.61	16	31,208,131	7	E9 / GmFT2a	31,109,978 : 31,114,934
ss715627971	7.87	17	4,779,185	27	GmZTL1	4,745,075 : 4,749,003
ss715628182	8.11	17	7,525,172	19	MYB173 / MYB175	7,393,424 : 7,395,444
ss715625789	9.38	17	11,482,490	17	GmZFP5	11,560,483 : 11,561,457
ss715627703	10.11	17	40,082,518	15	GmAOS2	40,223,942 : 40,225,689
ss715631394	9.337971	18	48,623,829	12	DREB2D;2	49,030,874 : 49,032,925
ss715635425	7.28	19	45,204,441	17	Dt1	45,183,675 : 45,184,980
ss715635641	11.92	19	47,194,890	23	GmMYB12B2	47,123,500 : 47,124,980
					GmMYB64	47,439,602 : 47,441,444
					E3	47,633,059 : 47,641,958
ss715637729	7.16	20	36,692,059	13	Gmpdiq-1a	36,705,455 : 36,711,658

*Naming according to `www.soybase.org` (Grant et al. 2010)

Construction of core collections

Since the environmentally defined precore comprised 3,663 accessions, we constructed core collections comprising 5% and 10% of the precore accessions to obtain manageable subsets for further evaluation by breeders and researchers.

Evaluation of Core Subset Optimization Strategies

We compared strategies that maximized allelic diversity within core collection accessions or maximized genetic distance between core entries, as well as a joint maximization of both parameters. As shown before (Thachuk et al. 2009), a combination of both parameters revealed a trade-off because the maximization of one parameter reduced the other (Fig. S17). Therefore, maximizing each parameter separately is the best approach for a given sampling objective (Fig. 4).

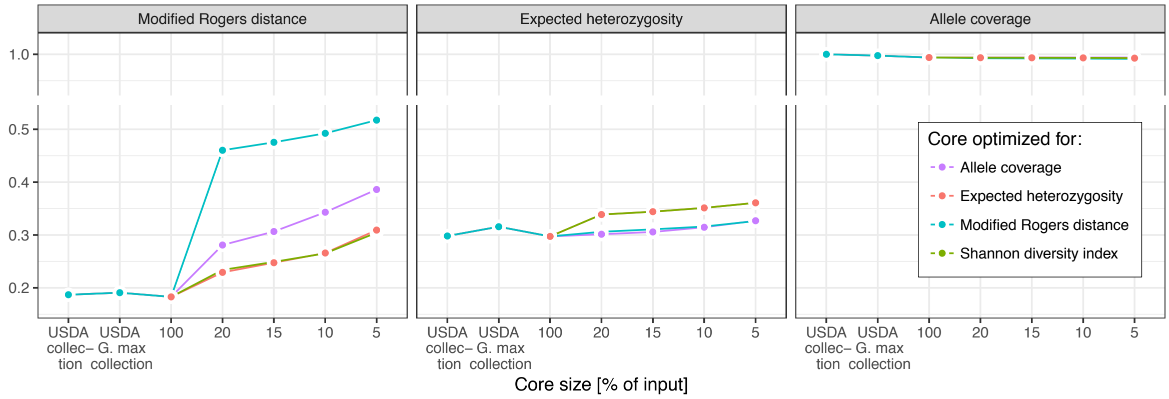

Fig. 4. Genetic distance and allelic diversity preserved in cores of varying size with varying optimization objectives. Core size '100' denotes the 3,663 precore accessions. Values on the y-axis refer to the respective facet label and all variables are defined in [0,1]. Core subset optimization based on Modified Rogers distance was most effective in forming cores consisting of distinct accessions.

Importantly, both approaches retained >99% of SNP alleles in all core collections. The loss of <1% of alleles resulted from the removal of *G. soja* specific alleles and the environmental stratification (Fig. S18). We compared two measures of allelic diversity, H_{exp} and Shannon's diversity index.

Both gave very similar results with respect to diversity within and distance between accessions, and were increased in core collections (Fig. 4). The genetic distance between accessions increased more strongly in core collections than allelic diversity and this behavior was independent of the optimization strategy. For example, in the most successful scenario the average minimum Modified Rogers distance increased 1.5-fold in the 5% core compared to the precore whereas H_{exp} differed only marginally (Fig. 4). Our estimates of allelic diversity (Table S3) might show an ascertainment bias caused by the design the SoySNP50K array if it includes SNPs that were preselected for high minor allele frequencies to guarantee high informativeness (Song et al. 2013). This most likely contributed to the modest increase of H_{exp} in core subset optimization (Malomane et al. 2018). By contrast, core subset optimization based on Modified Rogers distance purged highly similar accessions from subsets and was effective in forming cores consisting of distinct accessions (Fig. 4). We therefore based the core subset optimization exclusively on the maximization of genetic distance between core accessions.

Evaluation of Core Subset Sampling Strategies

The construction of core collections resulted in an underrepresentation of early maturing accessions (Fig. 5A and S19) because average redundancy levels were higher in early than within late maturing accessions (Fig. S20). Since early maturity is an important trait for CEU soybean breeding we explored sampling strategies based on a stratification by maturity group designation that favored the selection of early accessions. Any stratification prior to optimization represents a trade-off between accepting higher levels of redundancy in exchange for gaining improved representation in the stratification criterion. The best relative representation across all maturity groups was therefore observed for subsets sampled according to a classic stratification strategy (Fig. 5A) of independent sampling within three groups of accessions and aggregating selected accessions into a final core (Table S4). But it also resulted in the lowest level of realized genetic distance between accessions (Fig. 5B) and reflects that accessions in one subcore can be similar to accessions in another, which reduces

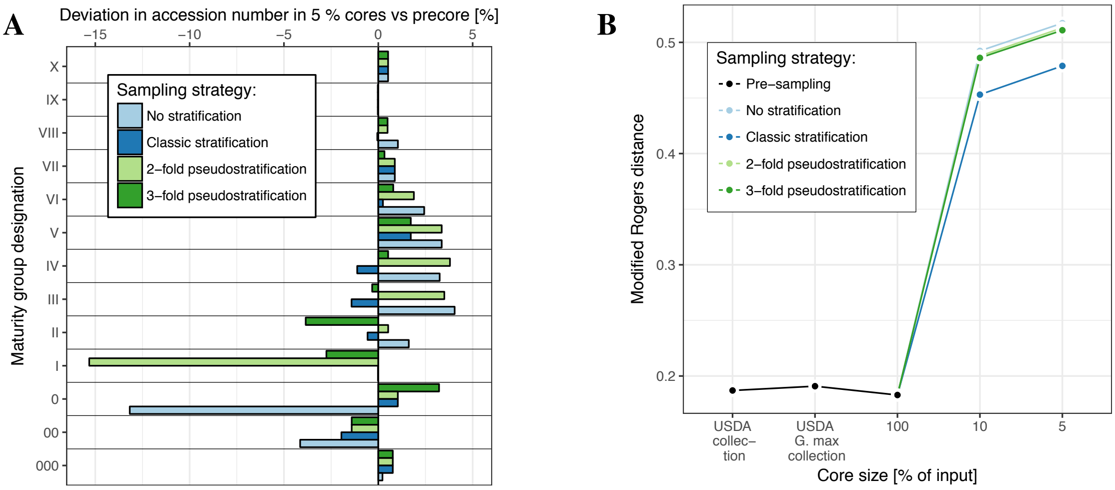

Fig. 5. A: Evaluation of different core sampling strategies with regards to the preservation of MG fractions in 5% cores relative to the precore. **B**: Increase in genetic distance between core accessions measured by Modified Roger's distance. Among different sampling strategies, the 3-fold pseudostratification* showed good results in terms of favouring early maturing material while maintaining high average minimum genetic distance between accessions.
*Consecutive sampling within groups of maturity categories *000 - 0*, *000 - I*, *000 - X* with fixation of already picked accessions (Tab. S4)

.

the overall diversity of the complete set. To increase genetic distance we adopted two pseudostratification sampling strategies that allowed to coordinate sub-core subset selection in a consecutive, step-wise fashion that ensured the selection of complementary sub-cores (Table S4). Sampling with pseudostratification resulted in cores that in terms of genetic dissimilarity between accessions were comparable to the no stratification benchmark. The 2-fold pseudostratification however failed in ensuring a good level of representation for accessions of maturity group *I*, presumably because accessions of this group on average are more similar to accessions of maturity groups *000 - 0* than to later accessions. The 3-fold pseudostratification resulted in a high average minimum genetic distance between accessions in the final set and also maintained acceptable levels of representation over all maturity groups. We therefore used the 3-fold pseudostratification to construct the final core collections with 5% and 10% of individuals from the precore collection.

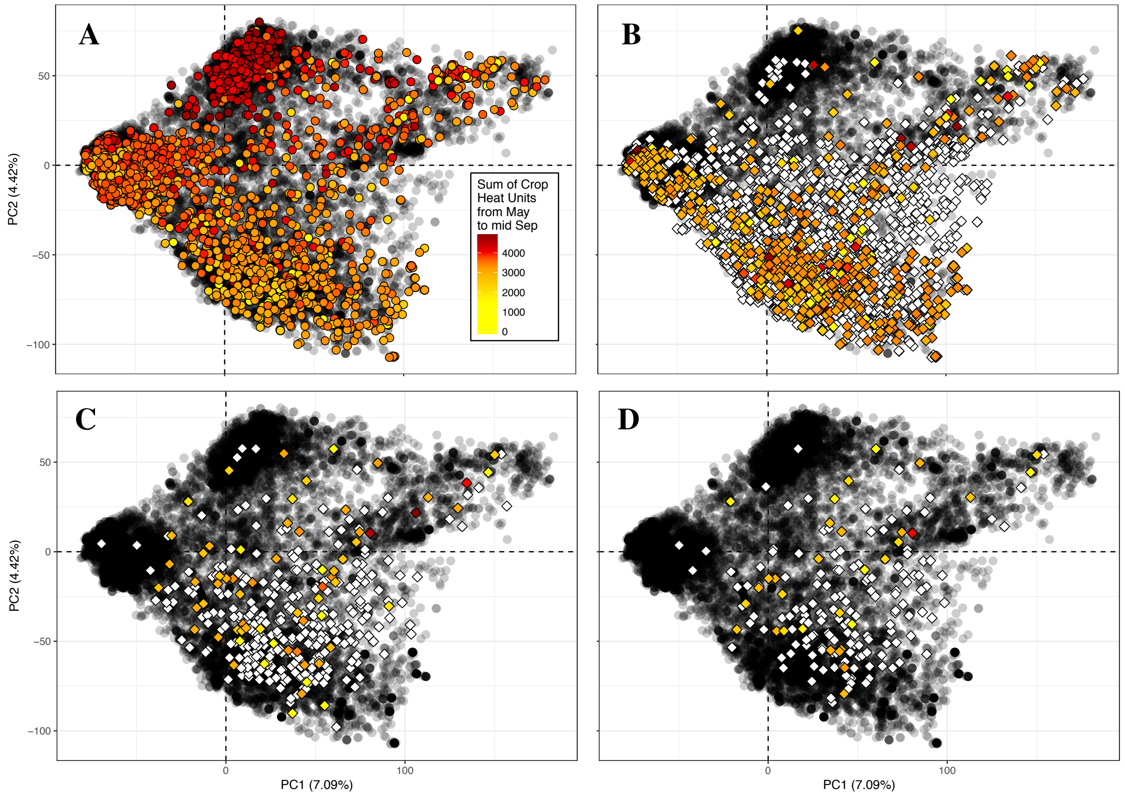

Fig. 6. Principal component analysis summarizing the genetic structure in > 17.000 *G. max* accessions conserved in the *Introduced G. max* subcollection of the USDA Soybean Germplasm Collection. Each dot represents one accession. **A**: Accessions with collection site information available in their passport data are highlighted according to temperature supply at origin. **B**: 3,663 accessions selected for the environmental precore based on environmental data (coloured) and / or early maturity group ratings (white) are highlighted. **C**: 366 entries of the 10% core are highlighted. **D**: 183 entries of the 5% core are highlighted.

Further Effects of Core Sampling: Shifts in Germplasm Composition and Allele Frequencies

The reduction of the number of accessions through our environmental stratification and the selection of core subsets maximized for genetic diversity was naturally accompanied by changes in the geographic and genomic composition of the respective accession population at each stage of the core formation process (Fig. 2, 6 and S21). The reduction of genetic redundancy among core entries furthermore resulted in genome wide changes in population-wise allele frequencies (Fig. S22). These mainly affected loci with low minor allele frequencies in the environmental precore by elevating the frequencies of the respective alleles in the cores (Fig. 8). Changes in MAFs between the 3,663 precore accessions and the core collection entries however never exceeded 0.25 (Fig. 22) and the groups showed a comparable distribution along the first two components of a PCA (Fig. 7).

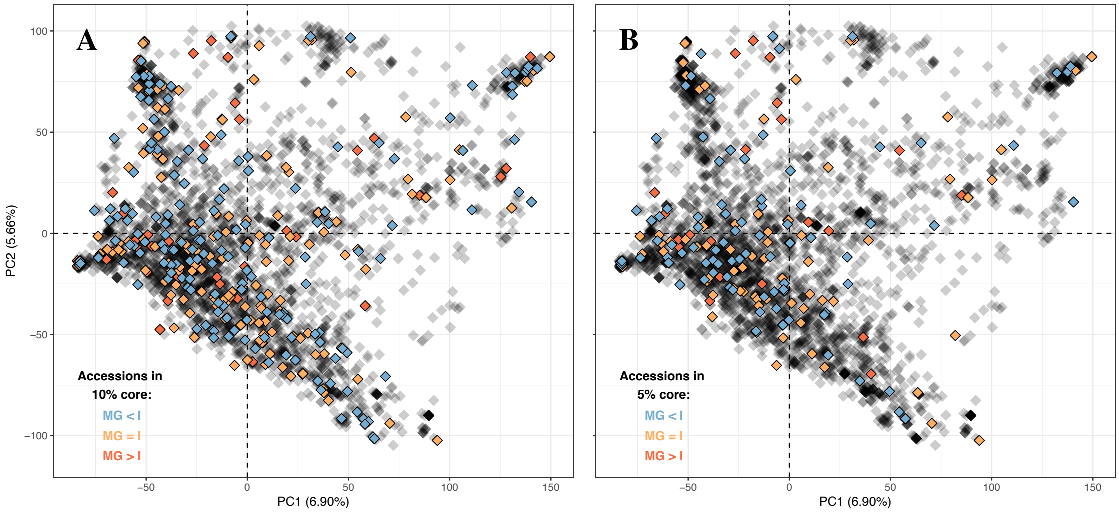

Fig. 7. Principal component analysis summarizing the genetic structure among 3,663 precore accessions shows an even spatial distribution of the core entries: core accessions are highlighted according to maturity group ratings. **A**: 366 entries of the 10% core. **B**: 183 entries of the 5% core.

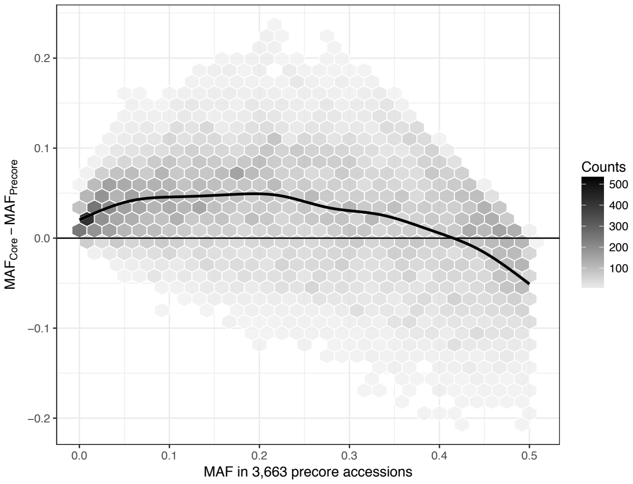

Fig. 8. Pairwise comparison of MAFs in the precore versus the MAF change from precore to 5% core: Core subset optimization resulted in genome wide MAF changes through enrichment of low frequency variants in the core. Fig. S22 provides a more detailed overview on MAF changes.

Also the genetic basis of the presumed environmental adaptation was unaffected by the core formation process and well preserved in the cores as can be observed in a comparison of genetic differentiation between core, precore and non-precore accessions (Fig. S16).

Robustness of Core Sampling

With 3,663 precore accessions, $\binom{3,663}{183} = 1.2 \times 10^{314}$ core subsets with 5% ($n = 183$) and 10% (n = 366) $\binom{3,663}{366} = 1.6 \times 10^{515}$ accessions are possible. To find the best core, stochastic algorithms to approximate the optimal solution as implemented in `Core Hunter` need to be used. Independent runs of these search algorithms may produce different results, but using a fixed seed to define the starting point ensures the reproducibility of the optimization solution as long as the termination criteria are kept constant. All results presented so far regarded core subset sampling solutions derived from optimization runs with a fixed seed of one. We verified the stability of our results from runs with fixed seeds by performing re-sampling with random seeds for five additional runs. Average levels of genetic distances between accessions in these runs were highly similar to the previous runs. The intersection of all 5% cores included the same 82 out of 183 (45%) and 223 out of 366 10%-core accessions (61%). Only a small proportion of accessions (9% / 5% in 5%/10% cores, respectively) was restricted to a single core subset and seemingly interchangeable through genetically similar entries (Table 2). Core subsets differing in sampling intensities were comparable on an individual level too, with 134 accessions of the 5% core having been also selected for the 10% core solution.

We followed Odong et al. (2013) and investigated how the formation of core collections depends on the marker set used by randomly dividing our marker set in a training set and an evaluation set of equal size (split-half analysis). The training set was used to sample cores with varying sampling intensities and the average minimum genetic distance between core entries was subsequently evaluated independently based on both marker sets and based on all available markers. Estimates of genetic distance were highly similar in all comparisons of the same sampling intensity and did not differ between marker sets ($\Delta MR \leq 0.00032$, data not shown). This suggests that our SNP marker set was sufficiently dense to warrant for the robust estimation of genetic distance and the presented core solutions can thus be considered stable within the disparities discussed above.

TABLE 2. Comparison of core collection sampling alternatives: Sampling intensity, core size, averaged minimum Modified Rogers distance between core accessions sampled with fixed seed, mean and standard deviation of averaged minimum Modified Rogers distance between core accessions of all sampled cores (including cores sampled with random seed), proportion of accessions that jointly occur in all sampled cores, proportion of accessions private to one core.

Core	N	MR	$\overline{MR}$	SD	$\bigcap_{i=1}^{6} Core_i$	$\bigcap_{i=1}^{6} Core_i = \emptyset$
5%	183	0.5110	0.5113	0.0003	0.45	0.09
10%	366	0.4860	0.4866	0.0003	0.61	0.05

Evaluation of phenotypic properties

We assessed changes in major phenotypic characteristics of accession sets throughout the core collection formation process, i.e. from the *Introduced G. max* subcollection over the environmental precore to the actual core subsets. The largest phenotypic changes were caused by the environmental stratification that removed most accessions of maturity group ≥ *II* and therefore the majority of the original collection (Table 3). As a consequence, variation in phenotypic traits of the precore and core subsets differed substantially from the *Introduced G. max* collection while the smaller core collections preserved the phenotypic variation of the precore (Table 4). Regarding the average phenotypic performance, seed weight was reduced in the cores compared to the precore while the seed yields did not differ significantly and seed components (protein and oil content) varied partially. A reduction in seed weight may result from the marker-based selection of more diverse exotic accessions instead of more recently adapted material with larger seeds. Cores sampled with random seeds produced similar results (Table S5).

TABLE 3. Composition of the *Introduced G. max* sub-collection with regards to maturity group ratings and the shift towards earlier maturity in precore accessions and core entries.

Group	N	*000*	*00*	*0*	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*	*X*
G. max collection	17,189	132	473	1,063	1,563	1,756	1,639	3,896	2,480	1,466	882	963	751	123
Precore	3,663	132	472	1,063	1,562	201	112	61	37	11	8	2	1	1
10% core	366	15	36	117	155	5	11	8	9	5	2	2	0	1
5% core	183	8	21	59	73	3	5	4	5	2	1	1	0	1

TABLE 4. Selected phenotypic properties of the *Introduced G. max* sub-collection, the environmental precore and the core collections. Differences between groups are reported as significant at a p-value < 0.05.

Group	N	$\overline{\text{Yield}}$ [Mg/ha]	Var	$\overline{\text{Seed weight}}$ [cg/seed]	Var	$\overline{\text{Protein}}$ [%]	Var	$\overline{\text{Oil}}$ [%]	Var
G. max collection	17,189	1.954^{a}	0.503^{a}	15.31^{a}	29.53^{a}	44.33^{a}	7.98^{a}	17.81^{a}	4.36^{a}
Precore	3,663	2.283^{b}	0.479^{b}	15.63^{b}	16.41^{b}	42.69^{b}	7.23^{b}	18.88^{b}	3.04^{b}
10% core	366	2.262^{b}	0.547^{ab}	14.67^{c}	13.36^{b}	43.18^{c}	6.88^{b}	18.53^{c}	3.30^{b}
5% core	183	2.149^{b}	0.578^{ab}	13.96^{c}	16.25^{b}	43.30^{bc}	10.05^{ab}	18.21^{ac}	4.77^{ab}

DISCUSSION

Creation of Core Collections

By combining the *Focused Identification of Germplasm Strategy* (FIGS) with core collection methodology, we identified subsets of genebank accessions from a large soybean germplasm collection (N > 17,000) that are likely pre-adapted to cultivation in Central Europe while simultaneously conserving a high level of genetic diversity. To establish these core collections, we first defined a precore of 3,663 accessions with potential abiotic adaptation to CEU based on environmental data. This precore was subsequently reduced to two core collections of 183 and 366 accessions (5% and 10% of the precore, please consult the supplementary information for further core accession details) by limiting genetic redundancies among core entries. We quantified the success of this approach at each step of the process and found strong evidence for the enrichment of adaptive genetic variation for abiotic adaptation to high-latitude cold regions throughout the genome (Fig. 3, S12 to S16, Tab. 1 and Tab. S2).

Limitations and Evidence for Enrichment of Environmental Adaptation

By inferring local adaptation from environmental data we followed FIGS methodology that is based on the assumption that landraces adapted to the environment in which they were originally cultivated by natural and artificial selection (Street et al. 2008; Sanders et al. 2013). Knowledge about the geographic origin of germplasm is therefore a key parameter in the successful application of FIGS for abiotic traits. For our implementation of FIGS in the USDA Soybean Germplasm Collection, the origin of accessions was approximated by the georeference for the respective collection sites recorded in the passport data. This information was only available for parts of the collection and in many cases seemed to be an approximate reconstruction of the historic collection sites, sometimes with limited accuracy. Inaccurate georeference information lead to false associations of accessions with environmental data and even though we removed the least reliable collection site - origin misspecifications, this certainly led to the inclusion of some non-adapted accessions into the environmental precore. Another limitation concerns the granularity of the environmental data: The spatial resolution and categorical scope of WorldClim data is sufficiently detailed, but the temporal resolution is restricted to monthly averages. Since records on local cropping practices such as the period of cultivation of landraces in their native environment are lacking too, a detailed reconstruction of the agro-ecological conditions that influenced local adaptation was not possible. Missing georeferences for the majority of accessions were even more limiting, which required us to use maturity group assignments of these accessions as sole proxy for environmental adaptation. Although field trials for the maturity group classification were conducted in North America, we consider using these information suitable for this purpose because of the high heritability of this trait (Xavier et al. 2018).

To test whether the selection of the precore accessions led to an enrichment of abiotic adaptation to CEU environments, we conducted a selection scan for genetic differentiation between precore accessions and non-precore accessions. Although the marker density of the SoySNP50K array was too low to pinpoint differentiation at the level of single genes, we found evidence that the selection based on climate data and maturity group assignments targets genomic regions involved in abiotic adaptation and enriches genetic variants which are advantageous for cultivation in CEU in the precore. Highly differentiated regions mainly indicated genomic regions that harbor well-known genes responsible for photoperiodic adaptation in soybean. This result is expected because the selection of adapted accessions was partially based on maturity groups that are conditioned by photoperiod genes. In addition, also new and promising candidate genes for adaptation to abiotic factors like heat, drought and cold stress were highlighted (Tab. 1 and Tab. S2). Most importantly, these signals of adaptation were preserved in the core collections (Fig. S16) which indicates that the genetic variation responsible for this presumed adaptation can be introgressed from core accessions into CEU elite breeding germplasm.

Towards a Characterization of Core Collections

To unlock the full potential of the core accessions, long term dedication of breeders and researchers for the detailed genotypic and phenotypic characterization of the cores will be necessary. A list of core entries is included as supplementary information for further reference. Environmental adaptation is complex with possibly manifold ways to adapt to the same stress which in a real world scenario will not be present as a single factor but as a combination of multiple factors. Therefore, multi-location field trials will be necessary to accurately estimate the level of abiotic adaptation to CEU environments, complemented by trials with controlled conditions in which the most promising accessions can be confronted with selected stresses. It should be noted that no accession is likely to outcompete current CEU varieties per se, but may contribute to improve breeding germplasm. Detailed phenotyping will identify beneficial variation in landraces for use in soybean prebreeding. The increasing availability of high-throughput aerial and field-based phenomics technologies holds great potential in this respect (Furbank and Tester 2011; Araus and Cairns 2014).

The efficient introduction of beneficial abiotic adaptation into breeding germplasm via marker assisted selection or genome editing requires the identification of causal genetic variants. The allele frequency of causal variants is a major determinant for the detection power in association-mapping panels (Spencer et al. 2009) and variants with MAF below 5% are usually not considered because of their low statistical power. Our optimization strategy for the selection of core accessions resulted in elevated MAF for rare alleles (Fig. 8), which will aid in the discovery of rare traits using modern phenotyping approaches and provides an increased power to map causal variants despite the fairly small core collection sizes. Accessions that harbor favorable alleles can be subsequently included in large multiparent populations for genetic fine-mapping and gene identification (Cockram and Mackay 2018).

Core Collections for a Future Climate

The time span required for the development of new varieties requires long-term planning in an era of climate change. We therefore selected precore accessions not only using current climate parameters of the TPE, but also considered climate projections for the year 2070, in which lower precipitation and higher temperatures during the soybean cropping season are expected for CEU (Fig. S7 and S8). These projections suggest that future varieties may not require cold adaptation, but will have to be highly drought tolerant. It is reasonable to argue that both types of adaptation will remain relevant because climate estimates are based on long-term averages, whereas short-term extreme events during the cropping season are frequently limiting crop yields. The future climate is expected to become more volatile and extreme events more frequent, which will include low temperatures in the early and high temperatures in the later growing season. The improvement of crop abiotic adaptation is therefore pivotal to equip agricultural systems with crop varieties for the future. The novel combination of FIGS and core collection methodology is a promising strategy to assemble relevant objective driven core collections and to increase the efficiency of germplasm characterization. Both will aid the identification of alleles for abiotic stress tolerance in order to complement modern breeding germplasm. Upon discovery, targeted crosses and combinations of marker-assisted selection and speed breeding or genome engineering can be employed to incorporate beneficial alleles into modern backgrounds in real-time (Varshney et al. 2018; Watson et al. 2018). Genomic selection has emerged as a strategy to predict complex traits also in genebank populations (Jarquin et al. 2016; Schnable et al. 2016), but its successful application is challenging in genetically diverse germplasm collections that are frequently lacking phenotypic characterization data for traits of interest (Langridge and Waugh 2019). Here, upon their sufficient characterization, objective driven core collections from FIGS selected germplasm groups have the potential to complement prediction efforts in genebank germplasm for abiotic adaptation by serving as training and validation populations.

AUTHOR CONTRIBUTION STATEMENT

MH and KS designed the study, MH analyzed the data and wrote the first version of the manuscript. MH and KS wrote the final version of the manuscript.

ACKNOWLEDGEMENTS

The authors thank the United States Department of Agriculture (USDA) and the USDA Soybean Germplasm Collection for their open data and free distribution of germplasm policies that made this work possible. This work was funded by BMEL Project Sojagenopath (FKZ 2814EPS012).

CONFLICT OF INTEREST STATEMENT

On behalf of all authors, the corresponding author states that there is no conflict of interest.

REFERENCES

Araus, J. L. and Cairns, J. E. (2014). “Field high-throughput phenotyping: The new crop breeding frontier.” *Trends in Plant Science*, 19(1), 52–61.

Balko, C., Hahn, V., and Ordon, F. (2014). “Chilling tolerance in soybeans (*Glycine max*) – a prerequisite for soybean cultivation in Germany.” *Journal of Cultivated Plants*, 66(11), 378–388.

Bari, A., Street, K., Mackay, M., Endresen, D. T. F., de Pauw, E., and Amri, A. (2012). “Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables.” *Genetic Resources and Crop Evolution*, 59(7), 1465–1481.

Bemer, M., Van Mourik, H., Muiño, J. M., Ferrándiz, C., Kaufmann, K., and Angenent, G. C. (2017). “*FRUITFULL* controls *SAUR10* expression and regulates Arabidopsis growth and architecture.” *Journal of Experimental Botany*, 68(13), 3391–3403.

Bernard, R. L. (1971). “Two Major Genes for Time of Flowering and Maturity in Soybeans.” *Crop Science*, 11, 242–244.

Beukelaer, H. D., Davenport, G. F., De Meyer, G., and Fack, V. (2017a). “JAMES: An objectoriented Java framework for discrete optimization using local search metaheuristics." *Software - Practice and Experience*, 47(6), 921–938.

Beukelaer, H. D., Davenport, G. F., and Fack, V. (2017b). "Multi-Purpose Core Subset Selection, <www.corehunter.org>.

Beukelaer, H. D., Smýkal, P., Davenport, G. F., and Fack, V. (2012). "Core Hunter II: fast core subset selection based on multiple genetic diversity measures using Mixed Replica search." *BMC Bioinformatics*, 13(1), 312.

Bhullar, N. K., Street, K., Mackay, M., Yahiaoui, N., and Keller, B. (2009). "Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the *Pm3* resistance locus." *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), 9519–9524.

BMEL (2016). "Ackerbohne, Erbse & Co., <`https://www.bmel.de/SharedDocs/Downloads/Broschueren/EiweisspflanzenstrategieBMEL.pdf?{_}{_}blob=publicationFile`>.

Bootsma, A., McKenney, D. W., Anderson, D., and Papadopol, P. (2007). "A re-evaluation of crop heat units in the maritime provinces of Canada." *Canadian Journal of Plant Science*, 87(2), 281–287.

Brown, D. and Bootsma, A. (1993). "Crop Heat Units for Corn and Other Warm Season Crops in Ontario." *Report No. Factsheet No. 93-119*, Ministry of Agriculture and Food Ontario, <https://www.sojafoerderring.de/wp-content/uploads/2014/02/Berechnung-CHU-Uni-Guelph-Ontario.pdf>.

Browning, B. L. and Browning, S. R. (2016). "Genotype Imputation with Millions of Reference Samples." *American Journal of Human Genetics*, 98(1), 116–126.

Buzzell, R. I. (1971). "Inheritance of a soybean flowering response to fluorescent-daylength conditions." *Canadian Journal of Genetics and Cytology*, 13(4), 703–707.

Cober, E. R., Molnar, S. J., Rai, S., Soper, J. F., and Voldeng, H. D. (2013). "Selection for cold tolerance during flowering in short-season soybean." *Crop Science*, 53(4), 1356–1365.

Cockram, J. and Mackay, I. (2018). "Genetic Mapping Populations for Conducting High-Resolution Trait Mapping in Plants." *Plant Genetics and Molecular Biology*, R. K. Varshney, M. K. Pandey, and A. Chitikineni, eds., Springer International Publishing, Cham, 109–138.

Diamond, J. (2002). "Evolution, consequences and future of plant and animal domestication." *Nature*, 418(6898), 700–707.

Dray, S. and Dufour, A.-B. (2007). "The ade4 Package: Implementing the Duality Diagram for Ecologists." *Journal of Statistical Software*, 22(4).

Du, Q. L., Cui, W. Z., Zhang, C. H., and Yu, D. Y. (2010). "*GmRFP1* encodes a previously unknown RING-type E3 ubiquitin ligase in Soybean (*Glycine max*)." *Molecular Biology Reports*, 37(2), 685–693.

Endresen, D. T. F. (2010). "Predictive association between trait data and ecogeographic data for Nordic barley landraces." *Crop Science*, 50(6), 2418–2430.

European Soya Declaration (2017). "Common Declaration of Austria, Croatia, Finland, France, Germany, Greece, Hungary, Italy, Luxemburg, the Netherlands, Poland, Romania, Slovakia and Slovenia European Soya Declaration – Enhancing soya and other legumes cultivation, <https://www.bmel.de/SharedDocs/Downloads/Landwirtschaft/Pflanze/SojaErklaerung.pdf?{_}{_}blob=publicationFile>.

Fick, S. and Hijmans, R. (2017). "Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas." *International Journal of Climatology*, 37(12), 4302–4315.

Frankel, O. (1984). "Genetic Perspectives of Germplasm Conservation." *Genetic Manipulation: Impact on Man and Society*, W. Arber, K. Llimensee, W. Peacock, and P. Stralinger, eds., Cambridge University Press, Cambridge, 161–170.

Funatsuki, H., Kawaguchi, K., Matsuba, S., Sato, Y., and Ishimoto, M. (2005). "Mapping of QTL associated with chilling tolerance during reproductive growth in soybean." *Theoretical and Applied Genetics*, 111(5), 851–861.

Funatsuki, H., Matsuba, S., Kawaguchi, K., Murakami, T., and Sato, Y. (2004). "Methods for evaluation of soybean chilling tolerance at the reproductive stage under artificial climatic conditions."

Plant Breeding, 123(6), 558–563.

Funatsuki, H., Suzuki, M., Hirose, A., Inaba, H., Yamada, T., Hajika, M., Komatsu, K., Katayama, T., Sayama, T., Ishimoto, M., and Fujino, K. (2014). "Molecular basis of a shattering resistance boosting global dissemination of soybean." *Proceedings of the National Academy of Sciences*, 111(50), 17797–17802.

Furbank, R. T. and Tester, M. (2011). "Phenomics - technologies to relieve the phenotyping bottleneck." *Trends in Plant Science*, 16(12), 635–644.

Gass, T., Schori, A., Fossati, A., Soldati, A., and Stamp, P. (1996). "Cold tolerance of soybean (*Glycine max* (L.) Merr.) during the reproductive phase." *European Journal of Agronomy*, 5(1-2), 71–88.

Gautier, M. (2015). "Genome-wide scan for adaptive divergence and association with population-specific covariates." *Genetics*, 201(4), 1555–1579.

Gizlice, Z., Carter, T. E., and Burton, J. W. (1994). "Genetic Base for North American Public Soybean Cultivars Released between 1947 and 1988." *Crop Science*, 34(5), 1143–1151.

Grant, D., Nelson, R. T., Cannon, S. B., Shoemaker, R. C., and (SoyBase) (2010). "SoyBase, the USDA-ARS soybean genetics and genomics database, <http://soybase.org>.

Günther, T. and Coop, G. (2013). "Robust identification of local adaptation from allele frequencies." *Genetics*, 195(1), 205–220.

Hahn, V. and Würschum, T. (2014). "Molecular genetic characterization of Central European soybean breeding germplasm." *Plant Breeding*, 133(6), 748–755.

Harlan, J. R. (1992). "Space, Time, and Variation." *Crops & Man*, American Society of Agronomy, Crop Science Society of America, Madison, WI, Chapter 7, 135 – 155.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). "Very high resolution interpolated climate surfaces for global land areas." *International Journal of Climatology*, 25(15), 1965–1978.

Hui, W., Gel, Y., and Gastwirth, J. (2008). "lawstat: an R package for law, public policy and biostatistics." *Journal of Statistical Software*, 28(3), 1–26.

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., Specht, J. E., Shoemaker, R. C., and Cregan, P. B. (2006). “Impacts of genetic bottlenecks on soybean genome diversity..” *PNAS*, 103(45), 16666–16671.

Iuchi, S., Kobayashi, Y., Koyama, H., and Kobayashi, M. (2014). “*STOP1*, a Cys2/His2 type zinc-finger protein, plays critical role in acid soil tolerance in Arabidopsis.” *Plant Signaling & Behavior*, 3(2), 128–130.

Iwasaki, K., Kamauchi, S., Wadahama, H., Ishimoto, M., Kawada, T., and Urade, R. (2009). “Molecular cloning and characterization of soybean protein disulfide isomerase family proteins with nonclassic active center motifs.” *FEBS Journal*, 276(15), 4130–4141.

Jarquin, D., Specht, J., and Lorenz, A. (2016). “Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions.” *G3*, 6(8), 2329–2341.

Jiang, B., Nan, H., Gao, Y., Tang, L., Yue, Y., Lu, S., Ma, L., Cao, D., Sun, S., Wang, J., Wu, C., Yuan, X., Hou, W., Kong, F., Han, T., and Liu, B. (2014). “Allelic combinations of soybean maturity Loci E1, E2, E3 and E4 result in diversity of maturity and adaptation to different latitudes..” *PlOS ONE*, 9(8), e106042.

Khazaei, H., Street, K., Bari, A., Mackay, M., and Stoddard, F. L. (2013). “The FIGS (Focused Identification of Germplasm Strategy) approach identifies traits related to drought adaptation in *Vicia faba* genetic resources.” *PLOS ONE*, 8(5).

Kongrit, D., Jisaka, M., Iwanga, C., Yokomichi, H., Katsube, T., Nishimura, K., Nagaya, T., and Yokota, K. (2007). “Molecular Cloning and Functional Expression of Soybean Allene Oxide Synthases.” *Bioscience, Biotechnology and Biochemistry*, 71(2), 491–498.

Kovinich, N., Saleem, A., Arnason, J. T., and Miki, B. (2012). “Identification of two anthocyanidin reductase genes and three red-brown soybean accessions with reduced anthocyanidin reductase 1 mRNA, activity, and seed coat proanthocyanidin amounts.” *Journal of Agricultural and Food Chemistry*, 60(2), 574–584.

Kurasch, A. K., Hahn, V., Leiser, W. L., Vollmann, J., Schori, A., Bétrix, C.-A., Mayr, B., Winkler, J., Mechtler, K., Aper, J., Sudaric, A., Pejic, I., Sarcevic, H., Jeanson, P., Balko, C., Signor, M., Miceli, F., Strijk, P., Rietman, H., Muresanu, E., Djordjevic, V., Pospišil, A., Barion, G., Weigold, P., Streng, S., Krön, M., and Würschum, T. (2017). "Identification of mega-environments in Europe and effect of allelic variations at maturity *E* loci on adaptation of European soybean." *Plant, Cell & Environment*, 0000, 765–778.

Langridge, P. and Waugh, R. (2019). "Harnessing the potential of germplasm collections." *Nature Genetics*, 51(2), 200–201.

Liao, Y., Zou, H. F., Wei, W., Hao, Y. J., Tian, A. G., Huang, J., Liu, Y. F., Zhang, J. S., and Chen, S. Y. (2008). "Soybean *GmbZIP44*, *GmbZIP62* and *GmbZIP78* genes function as negative regulator of ABA signaling and confer salt and freezing tolerance in transgenic Arabidopsis." *Planta*, 228(2), 225–240.

Littlejohns, D. a. and Tanner, J. W. (1976). "Preliminary Studies on the Cold Tolerance of Soybean Seedlings." *Canadian Journal of Plant Science*, 56(2), 371–375.

Liu, B., Watanabe, S., Uchiyama, T., Kong, F., Kanazawa, A., Xia, Z., Nagamatsu, A., Arai, M., Yamada, T., Kitamura, K., Masuta, C., Harada, K., and Abe, J. (2010). "The Soybean Stem Growth Habit Gene *Dt1* Is an Ortholog of Arabidopsis *TERMINAL FLOWER1*." *Plant Physiology*, 153(1), 198–210.

Ma, Y. Y., Zhang, L. W., Li, P. L., Gan, R., Li, X. P., Zhang, R., Wang, Y., and Wang, N. N. (2006). "Cloning and preliminary characterization of three receptor-like kinase genes in soybean." *Journal of Integrative Plant Biology*, 48(11), 1338–1347.

Malomane, D. K., Reimer, C., Weigend, S., Weigend, A., Sharifi, A. R., and Simianer, H. (2018). "Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies." *BMC genomics*, 19(1), 22.

McGonigle, B., Keeler, S. J., Lau, S.-M. C., Koeppe, M. K., and O'Keefe, D. P. (2000). "A Genomics Approach to the Comprehensive Analysis of the Glutathione *S* -Transferase Gene Family in Soybean and Maize." *Plant Physiology*, 124(3), 1105–1120.

Mizoi, J., Ohori, T., Moriwaki, T., Kidokoro, S., Todaka, D., Maruyama, K., Kusakabe, K., Osakabe, Y., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2013). "*GmDREB2A;2*, a Canonical DEHYDRATION-RESPONSIVE ELEMENT-BINDING PROTEIN2-Type Transcription Factor in Soybean, Is Posttranslationally Regulated and Mediates Dehydration-Responsive Element-Dependent Gene Expression." *Plant Physiology*, 161(1), 346–361.

Nelson, R. L. (2011). "Managing self-pollinated germplasm collections to maximize utilization." *Plant Genetic Resources*, 9(01), 123–133.

Odong, T. L., Jansen, J., van Eeuwijk, F. A., and van Hintum, T. J. (2013). "Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation." *Theoretical and Applied Genetics*, 126(2), 289–305.

Oliveira, M. F., Nelson, R. L., Geraldi, I. O., Cruz, C. D., and de Toledo, J. F. F. (2010). "Establishing a soybean germplasm core collection." *Field Crops Research*, 119(2-3), 277–289.

Olsson, A. S., Engström, P., and Söderman, E. (2004). "The homeobox genes *ATHB12* and *ATHB7* encode potential regulators of growth in response to water deficit in Arabidopsis." *Plant Molecular Biology*, 55(5), 663–677.

Park, S.-Y., Yu, J.-W., Park, J.-S., Li, J., Yoo, S.-C., Lee, N.-Y., Lee, S.-K., Jeong, S.-W., Seo, H. S., Koh, H.-J., Jeon, J.-S., Park, Y.-I., and Paek, N.-C. (2007). "The Senescence-Induced Staygreen Protein Regulates Chlorophyll Degradation." *the Plant Cell Online*, 19(5), 1649–1664.

Recknagel, J. D. S. (2015). "Sorten für Deutschland, <`https://www.sojafoerderring.de/anbauratgeber/sortenratgeber/deutschland/`>.

Sanders, R., Bari, A., Street, K., and Devlin, M. (2013). "A new approach to mining agricultural gene banks – to speed the pace of research innovation for food security. 'FIGS' - the Focused Identification of Germplasm Strategy." *Report No. 3*, International Center for Agricultural Research in the Dry Areas (ICARDA), Beirut.

Sasaki, K. and Imai, R. (2012). "Pleiotropic Roles of Cold Shock Domain Proteins in Plants." *Frontiers in Plant Science*, 2(January), 1–6.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H. T., Wing, R. a., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R. C., and Jackson, S. a. (2010). "Genome sequence of the palaeopolyploid soybean.." *Nature*, 463(7278), 178–183.

Schnable, P. S., Wu, Y., Wang, M. L., Guo, T., Yu, J., Tesso, T. T., Bernardo, R., Roozeboom, K. L., Wang, D., Yu, X., Li, X., Mitchell, S. E., Pederson, G. A., and Zhu, C. (2016). "Genomic prediction contributing to a promising global strategy to turbocharge gene banks." *Nature Plants*, 2(10), 16150.

Silva, P. A., Silva, J. C. F., Caetano, H. D., Machado, J. P. B., Mendes, G. C., Reis, P. A., Brustolini, O. J., Dal-Bianco, M., and Fontes, E. P. (2015). "Comprehensive analysis of the endoplasmic reticulum stress response in the soybean genome: Conserved and plant-specific features." *BMC Genomics*, 16(1), 1–20.

Somers, D. E., Schultz, T. F., Milnamow, M., and Kay, S. A. (2000). "*ZEITLUPE* encodes a novel clock-associated PAS protein from Arabidopsis." *Cell*, 101(3), 319–329.

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., and Cregan, P. B. (2013). "Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean." *PLoS ONE*, 8(1), 1–12.

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., and Cregan, P. B. (2015). "Fingerprinting Soybean Germplasm and Its Utility in Genomic Research.." *G3*, 5(10), 1999–2006.

Song, W., Solimeo, H., Rupert, R. A., Yadav, N. S., and Zhu, Q. (2002). "Functional dissection of a Rice Dr1/DrAp1 transcriptional repression complex." *Plant Cell*, 14(January), 181–195.

Spencer, C. C., Su, Z., Donnelly, P., and Marchini, J. (2009). "Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip." *PLoS Genetics*, 5(5).

Street, K., Mackay, M., Zuev, E., Kaul, N., El Bouhssini, M., Konopka, J., and Mitrofanova, O.

(2008). "Diving into the genepool : a rational system to access specific traits from large germplasm collections." *Proceedings of the 11th International wheat genetics symposium*, 28–31.

Su, L. T., Li, J. W., Liu, D. Q., Zhai, Y., Zhang, H. J., Li, X. W., Zhang, Q. L., Wang, Y., and Wang, Q. Y. (2014). "A novel MYB transcription factor, *GmMYBJ1*, from soybean confers drought and cold tolerance in *Arabidopsis thaliana*." *Gene*, 538(1), 46–55.

Sun, H., Jia, Z., Cao, D., Jiang, B., Wu, C., Hou, W., Liu, Y., Fei, Z., Zhao, D., and Han, T. (2011). "*GmFT2a*, a soybean homolog of flowering locus T, is involved in flowering transition and maintenance." *PLOS ONE*, 6(12), 18–20.

Tanksley, S. D. and McCouch, S. R. (1997). "Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild." *Science*, 277(5329), 1063–1066.

Thachuk, C., Crossa, J., Franco, J., Dreisigacker, S., Warburton, M., and Davenport, G. F. (2009). "Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures.." *BMC bioinformatics*, 10(243), 1–13.

Varshney, R. K., Singh, V. K., Kumar, A., Powell, W., and Sorrells, M. E. (2018). "Can genomics deliver climate-change ready crops?." *Current Opinion in Plant Biology*, 45, 205–211.

Wang, C., Hu, S., Gardner, C., and Lübberstedt, T. (2017a). "Emerging Avenues for Utilization of Exotic Germplasm." *Trends in Plant Science*, 22(7), 624–637.

Wang, H., Zhao, S., Gao, Y., and Yang, J. (2017b). "Characterization of dof transcription factors and their responses to osmotic stress in poplar (*Populus trichocarpa*)." *PLoS ONE*, 12(1), 1–19.

Wang, Y., Chai, C., Valliyodan, B., Maupin, C., Annen, B., and Nguyen, H. T. (2015). "Genome-wide analysis and expression profiling of the PIN auxin transporter gene family in soybean (Glycine max)." *BMC Genomics*, 16(1), 1–13.

Watanabe, S., Xia, Z., Hideshima, R., Tsubokura, Y., Sato, S., Yamanaka, N., Takahashi, R., Anai, T., Tabata, S., Kitamura, K., and Harada, K. (2011). "A map-based cloning strategy employing a residual heterozygous line reveals that the *GIGANTEA* gene is involved in soybean maturity and flowering." *Genetics*, 188(2), 395–407.

Watson, A., Ghosh, S., Williams, M. J., Cuddy, W. S., Simmonds, J., Rey, M. D., Asyraf Md Hatta, M., Hinchliffe, A., Steed, A., Reynolds, D., Adamski, N. M., Breakspear, A., Korolev, A., Rayner, T., Dixon, L. E., Riaz, A., Martin, W., Ryan, M., Edwards, D., Batley, J., Raman, H., Carter, J., Rogers, C., Domoney, C., Moore, G., Harwood, W., Nicholson, P., Dieters, M. J., Delacy, I. H., Zhou, J., Uauy, C., Boden, S. A., Park, R. F., Wulff, B. B., and Hickey, L. T. (2018). "Speed breeding is a powerful tool to accelerate crop research and breeding." *Nature Plants*, 4(1), 23–29.

Xavier, A., Thapa, R., Muir, W. M., and Rainey, K. M. (2018). "Population and quantitative genomic properties of the USDA soybean germplasm collection." *Plant Genetic Resources: Characterisation and Utilisation*, 16(6), 513–523.

Yamaguchi, N., Kurosaki, H., Ishimoto, M., and Kawasaki, M. (2015). "Early-Maturing and Chilling-Tolerant Soybean Lines Derived from Crosses between Japanese and Polish Cultivars." *Plant Production Science*, 18(August 2014), 234–239.

Yu, G. H., Jiang, L. L., Ma, X. F., Xu, Z. S., Liu, M. M., Shan, S. G., and Cheng, X. G. (2014). "A soybean C2H2-type zinc finger gene *GmZF1* enhanced cold tolerance in transgenic *Arabidopsis*." *PLOS ONE*, 9(10).

Zeng, X., Liu, H., Du, H., Wang, S., Yang, W., Chi, Y., Wang, J., Huang, F., and Yu, D. (2018). "Soybean MADS-box gene *GmAGL1* promotes flowering via the photoperiod pathway." *BMC Genomics*, 19(1), 1–17.

Zhao, C., Takeshima, R., Zhu, J., Xu, M., Sato, M., Watanabe, S., Kanazawa, A., Liu, B., Kong, F., Yamada, T., and Abe, J. (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 Biology*, 16(1), 1–15.

Zhou, Q. Y., Tian, A. G., Zou, H. F., Xie, Z. M., Lei, G., Huang, J., Wang, C. M., Wang, H. W., Zhang, J. S., and Chen, S. Y. (2008). "Soybean WRKY-type transcription factor genes, *GmWRKY13*, *GmWRKY21*, and *GmWRKY54*, confer differential tolerance to abiotic stresses in transgenic *Arabidopsis* plants." *Plant Biotechnology Journal*, 6(5), 486–503.