Geographic distribution of the E1 family of genes and their effects on reproductive timing in soybean

Nicholas Dietz 1, Rachel Combs-Giroir 1, Grace Cooper 1, Minviluz Stacey 1, Carrie Miranda 2 and Kristin Bilyeu 3*

Abstract

Background: Soybean is an economically important crop which flowers predominantly in response to photoperiod. Several major loci controlling the quantitative trait for reproductive timing have been identified, of which allelic combinations at three of these loci, E1, E2, and E3, are the dominant factors driving time to flower and reproductive period. However, functional genomics studies have identified additional loci which affect reproductive timing, many of which are less understood. A better characterization of these genes will enable fine-tuning of adaptation to various production environments. Two such genes, E1La and E1Lb, have been implicated in flowering by previous studies, but their effects have yet to be assessed under natural photoperiod regimes.

Results: Natural and induced variants of E1La and E1Lb were identified and introgressed into lines harboring either E1 or its early flowering variant, e1-as. Lines were evaluated for days to flower and maturity in a Maturity Group (MG) III production environment. These results revealed that variation in E1La and E1Lb promoted earlier flowering and maturity, with stronger effects in e1-as background than in an E1 background. The geographic distribution of E1La alleles among wild and cultivated soybean revealed that natural variation in E1La likely contributed to northern expansion of wild soybean, while breeding programs in North America exploited e1-as to develop cultivars adapted to northern latitudes.

Conclusion: This research identified novel alleles of the E1 paralogues, E1La and E1Lb, which promote flowering and maturity under natural photoperiods. These loci represent sources of genetic variation which have been underutilized in North American breeding programs to control reproductive timing, and which can be valuable additions to a breeder’s molecular toolbox.

Background

Soybean [Glycine max (L.) Merr.] is the world’s most economically important oilseed crop and was domesticated from its wild progenitor (Glycine soja [Sieb. & Zucc.]) more than 5000 years ago [1]. It is adapted to temperate latitudes and flowers in response to short day photoperiods. Reproductive timing is critical for optimizing plant yield in any production environment. Several genes to date have been shown to influence flowering and maturity in soybean, including E1-E4 [2–5], E6-E10 [6–10], Tof11 and Tof12 [11, 12], and J [13], while variation in E1, E2, and E3 are the predominant loci controlling flowering time in US-adapted varieties [14, 15]. However, further work is needed to identify additional mechanisms useful for fine-tuning reproductive timing in response to photoperiod and to improve adaptation of soybean to various production environments.

The major maturity gene E1 is a B3-related transcription factor which suppresses expression of key flowering genes, GmFT2a and GmFT5a, in the leaf under long...
days [16]. The dominant, functional allele, E1, delays flowering and maturity by 23 days and 18 days, respectively, compared to the partially functional e1-as allele; nonfunctional e1-fs and e1-nl alleles condition even earlier flowering and may contribute to photoperiod insensitivity [2]. E2, an orthologue of the Arabidopsis GIGANTEA gene, is a circadian clock gene that plays a role in modulating diurnal expression patterns of floral regulators [3]. The E3 and E4 genes are phytochrome molecules involved in perception of red and far-red light, respectively [5]. Phytochrome and circadian clock signals converge to promote transcription of E1, and thus inhibit flowering, under long days. This E1-mediated repression is relieved once day length shortens past a certain threshold, as determined predominantly by the allelic combination of E1, E2, and E3 [16].

Soybean has undergone two whole-genome duplication events in its evolutionary history, with subsequent fractionation back to diploid [17]. As a result, more than 50% of its genes are present as paralogous copies. The major maturity gene E1 has two paralogues, E1-like-a (E1La) and E1-like-b (E1Lb), both of which are located in the pericentromeric region of chromosome 04 [16]. Similar to E1, E1La and E1Lb are transcription factors which exhibit diurnal expression patterns and down-regulate transcription of GmFT2a and GmFT5α under long day photoperiods [18]. Using incandescent lamps to artificially extend day length, Zhu et al., 2019 showed that E1Lb inhibits flowering most strongly under far-red enriched long days, its magnitude similar to that of the E4 gene [19]. Despite this effect, natural variation in E1Lb appears to be very rare among soybean adapted to northern latitudes. Variation in E1La has not previously been explored. The gain of photoperiod insensitivity has been categorized into three genotypic groups: 1) disfunction of e3 and e4 (E1/e3/e4), 2) disfunction of e1 and e3 (e1/e3/E4), and 3) partial functionality of E1 with disfunction of e3 (e1-as/e3/E4), in combination with other unknown factors contributing to photoperiod insensitivity [20]. Subsequently, variation in GmFT5α [21], as well as disfunction in e1Lb, have both been implicated in photoperiod insensitivity in an e1-as/e3/E4 background [19]. Not surprisingly, co-silencing the entire E1 family of genes in an otherwise extremely late flowering landrace from Southern China led to an apparent complete photoperiod insensitivity under natural daylength and short day conditions [22].

Despite previous reports, the impact that E1La and E1Lb independently have on flowering and maturity under a natural light regime, as well as the contribution of E1La to the expansion of wild and adapted soybean to new production environments, have yet to be explored. In the present study, we show that variation in E1La and E1Lb each have significant effects on reproductive timing when E1 is partially functional (e1-as), but that the impact of E1Lb is abolished in a functional E1 background. Furthermore, we demonstrate that natural variation in the E1La gene has contributed to adaptation of wild soybean to northern latitudes but has been underutilized as a source of photoperiod insensitivity in cultivars released in North America.

Results
Natural and induced variation in the E1 paralogues, E1La and E1Lb

Although many genes impacting wild and domesticated soybean phenology have been identified, a subset of flowering time and reproductive period genes are relevant to this research including E1 and its homologs E1La and E1Lb as well as Tof11 and Tof12, the GIGANTEA gene E2, and the phytochrome E3 (Supplemental Table 1) [3, 4, 12, 16, 18]. E1Lb was originally positioned on chromosome 18, but the subsequent genome version (Williams 82.a2.v1) has both E1La and E1Lb positioned on chromosome 04 separated by about 10 million base pairs (Mb) (Supplemental Table 1). Compared to the characterized variant alleles, the functional versions of E1, E2, E3, Tof11 and Tof12 delay flowering and maturity and are the de facto alleles for G. soja (Table 1). Tof11 was not included in the Williams 82 reference genome Williams 82.a2.v1 annotation.

To determine the potential allelic variation present in E1La and E1Lb, we conducted a reverse genetics investigation for these genes from among a publicly available set of 302 whole genome re-sequenced accessions containing both G. max and G. soja accessions [23]. Using our SNPViz haplotype viewer tool [24], a single nonconservative missense mutation in the E1La gene was identified in ten G. soja accessions, leading to a lysine to glutamate substitution at amino acid position 82 (hereafter referred to as e1La:K82E). The K82E substitution is a positively to negatively charged amino acid change, and it falls within a relatively conserved region of the protein sequence (Fig. 1a; Supplemental Table 2).

Subsequent analysis of an expanded soybean resequencing dataset of 775 accessions distributed between 110 G. soja and 665 G. max revealed a total of 15 G. soja and 2 G. max accessions with e1La:K82E alleles [23, 25]. In our original analysis of the 302 soybean dataset, no variant alleles of the E1Lb gene were identified. In a later analysis of the 775 accessions data for the E1Lb gene, there were two G. soja accessions predicted to contain an S34R missense mutation (data not shown).

We utilized a reverse genetics approach to identify an induced mutant line with a ~2.6 Mbp deletion on chromosome 04 that included the E1Lb gene from a collection of Williams 82 fast-neutron mutant lines [26]. The boundaries of this lesion were approximated using
comparative genomic hybridization (CGH), revealing a deletion of 49 predicted genes in the G. max (v1) reference genome, including the E1Lb gene (hereafter referred to as e1lb:Del) (Fig. 1b; Supplemental Table 3).

Molecular breeding scheme to develop soybean germplasm with variant alleles of E1La and E1Lb
To directly investigate the impact E1La and E1Lb have on flowering time and maturity under natural light conditions, we developed and utilized lines selected by genotype from populations that had segregated for e1la:K82E or e1lb:Del. New molecular marker assays were developed to track the e1la:K82E and e1lb:Del mutant alleles. Because undomesticated G. soja was the initial source of the e1la:K82E alleles, a breeding scheme was devised to isolate those alleles from the confounding effects of Tof11 and Tof12 as well as other undesirable G. soja agronomic alleles (Supplemental Figures 1 & 2; Table 1). Seven populations were eventually utilized to develop lines with e1la:K82E alleles and other combinations of E1, E2, and E3 alleles, while two populations were used to develop lines with e1lb:Del alleles and either E1 or e1-as alleles (Tables 1 and 2).

Impact of E1La and E1Lb on reproductive timing in soybean under natural light conditions
Although the e1la:K82E alleles have not been previously assessed and e1lb:Del alleles were the result of induced mutation, the e1-as E2 E3 E1La E1Lb genotype is known to predominate in soybean cultivars adapted to MG III environments in the US [14]. Our adapted reference control line Williams 82 therefore contains the genotype E1La E1Lb e1-as E2 E3 (Table 1). A subset of the population parents or control lines, and test lines with mutant E1La or E1Lb were selected from the developed populations (Table 2; Supplemental Table 4) and grown in our MG III Missouri field environment during the 2018 and 2019 growing seasons. Plots were evaluated for

Table 1 Maturity gene alleles for soybean parents, controls, and test lines used in population development and characterization

| ID       | Name            | e1-as | E1La   | E1Lb   | E2 | E3 | Tof11-1 | Tof12-1 | Experiment* |
|----------|-----------------|-------|--------|--------|----|----|---------|---------|-------------|
| parent   | PI 547831       | E1    | e1laK82E | REF    | REF | REF | Tof11   | Tof12   | A           |
| parent   | PI 522226       | E1    | e1laK82E | REF    | REF | REF | Tof11   | Tof12   | A           |
| parent   | KB16-28#666     | e1-la | e1laK82E | REF    | REF | REF | REF     | REF     | A           |
| parent   | KB16-W2F3       | e1-as | e1laK82E | REF    | REF | REF | REF     | REF     | A           |
| parent and test | KB17-29514 | E1 | e1laK82E | REF    | REF | REF | REF     | A, B    |             |
| parent and test | KB17-19481 | e1-as | e1laK82E | REF    | REF | REF | REF     | A, B    |             |
| parent and test | W82 FN | e1-as | REF | e1lbDel | REF | REF | REF | REF | A, B, C |
| parent and control | Jake   | E1    | REF    | REF    | REF | REF | REF | REF | A, B, C |
| control   | Ellis HOLL      | E1    | REF    | REF    | REF | REF | REF     | C       |             |
| parent and control | Williams 82 | e1-as | REF | REF | REF | REF | REF | REF | A, B, C |
| parent and control | LGD4-6000 | e1-as | REF | REF | REF | REF | REF | REF | A, B, C |
| parent     | EXP e3          | e1-as | REF    | REF    | e3  | REF | REF     | A       |             |
| control   | Brookings       | e1-as | REF    | REF    | e3  | REF | REF     | C       |             |
| parent and control | Deuel | e1-as | REF | e2    | REF | REF | REF | A, C    |             |
| test      | Candor          | e1-as | REF    | e2    | REF | REF | REF     | C       |             |

*Breeding; B=2018/2019 field; C=2020 field

Fig. 1 Natural and induced variants of E1La and E1Lb. a Weblogo depicting conserved domain of E1La where lysine to glutamate missense change occurs (asterisk). b Comparative genomic hybridization depicting deleted region of Chr18 (moved to Chr04 in Wm82.a2.v1) containing the E1Lb gene
phenotypes for days to flower and days to maturity, and there were significant differences for the phenotypes and for the year effect (Figs. 2 and 3). In 2018, lines fixed for e1la:K82E/e1-as flowered 5 days earlier, and matured 24 days earlier, than lines with reference (REF) alleles for E1La and E1Lb (e1-as). Lines fixed for e1lb:Del/e1-as flowered 4 days earlier, and matured 9 days earlier, than reference lines for E1La and E1Lb (e1-as). There was no significant difference between days to flower for e1la:K82E and e1lb:Del lines, but e1la:K82E lines were significantly earlier for days to maturity than e1lb:Del lines in the e1-as background.

The absolute values were different in 2019 than 2018, but the results were similar; in 2019, lines fixed for e1la:K82E/e1-as flowered 8 days earlier, and matured 16 days earlier, than lines with reference alleles- E1La and E1Lb (e1-as). Lines fixed for e1lb:Del/e1-as flowered 6 days earlier, and matured 8 days earlier, than reference lines- E1La and E1Lb (e1-as) in 2019, and similar to 2018, e1lb:Del lines in the e1-as background were not significantly different than e1la:K82E/e1-as for days to flowering, but were significantly later than e1la:K82E/e1-as lines and earlier than reference lines for days to maturity (Fig. 2).

Soybean lines with functional versions of the E1 gene are not typically adapted to a MG III field environment [14], but we combined E1 with the mutant alleles of E1La or E1Lb (Table 2; Supplemental Table 4) to investigate their ability to influence photoperiod response (Fig. 3). The first frost in a MG III environment typically

Table 2 Soybean information for populations developed to select lines for characterization of the E1La and E1Lb impact on flowering time and maturity

| Population Code | Name 1 | Name 2 | Female 1 | Male 1 | Target 1 | Target 2 | G. soja % | Years | Experiment |
|-----------------|--------|--------|----------|--------|----------|----------|-----------|-------|------------|
| E1_e1la   | KB17-2 | Jake   | KB16-2B#666 | e1la:K82E | E1      | 12.5     | 2018, 2019 | A, B  |
| e1_e1la   | KB17-1 | W82    | KB16-2B#666 | e1la:K82E | e1-as   | 12.5     | 2018, 2019 | A, B  |
| E1_e1lb   | KB17-6 | W82    | Jake      | e1lb:Del | E1      | 0        | 2018, 2019 | A, B  |
| e1_e1lb   | KB16-5 | LG04-6000 | W82    | e1lb:Del | e1-as   | 0        | 2018, 2019 | A, B  |
| e1la_e2   | KB18-1 | Deuel  | KB17-1#481 | e1la:K82E | e2      | 6.3      | 2020     | C     |
| e1la_e3   | KB18-23 | EXP e3 | KB17-1#481 | e1la:K82E | e3      | 6.3      | 2020     | C     |
| e1_e1la   | KB18-16 | W82   | KB17-1#481 | e1la:K82E | e1-as   | 6.3      | 2020     | C     |
| e1_e1la   | KB18-17 | W82   | KB16-W2F3 | e1la:K82E | e1-as   | 12.5     | 2020     | C     |
| E1_e1la   | KB18-18 | Jake   | KB17-2#514 | e1la:K82E | E1      | 6.3      | 2020     | C     |

Fig. 2 For the e1-as background, days to flower and days to maturity for e1la:K82E test lines, e1lb:Del test lines, and E1La E1Lb reference (REF) controls. Means comparisons were conducted using an ANOVA, and significance groups were obtained using a Fisher’s LSD test with false discovery rate (FDR) correction (P = 0.05). Within each plot, categories that were not significantly different share the same significance letter. Lines were categorized by their selected genotypes, and n represents the number of plots of lines (each replicated three times) per genotype category. e1la:K82E is represented as e1la, e1lb:Del is represented as e1lb, and E1La and E1Lb indicate the REF alleles. Box plots show the mean values (diamonds), median (solid line), quartile span of the data (box), range (vertical lines), and outliers (dots). a Days to flower in 2018 experiment; b Days to flower in 2019 experiment; c Days to maturity in 2018 experiment; d Days to flower in 2019 experiment.
occurs before E1 lines have matured. Lines fixed for e1la:K82E/E1 flowered 5 days earlier compared to the reference controls-E1Lb/E1. Lines with e1la:K82E/E1 matured 16 days earlier than the killing frost, but reference controls- E1La and E1Lb (E1) and lines fixed for e1lb: Del/E1 were killed by frost. The only line fixed for e1lb: Del/E1 was significantly later for days to flowering compared to the reference controls- E1La and E1Lb (E1) in 2018. In 2019, lines fixed for e1la:K82E/E1 flowered 4 days earlier than reference controls-- E1La and E1Lb (E1), and matured 7 days before the killing frost. Lines fixed for e1lb:Del/E1 flowered the same day as reference controls- E1La and E1Lb (E1), and did not mature before first frost (Fig. 3).

The length of the reproductive cycle is a critical determinant of plant yield. Given that E1La and E1Lb pleiotropically affect both flowering time and maturity, we calculated the mean percentage of time spent in each phase of the life cycle (Fig. 4). The relative length of the reproductive phase between lines harboring e1la:K82E or e1lb:Del, compared to their reference controls, appeared to fluctuate between years. However, in each year, the length of the reproductive phase for genotype groups in the e1-as background were always within 3% of its respective control group. It should be noted that parental controls containing the E1/E1La/E1LB genotype, and lines containing E1/E1La/e1lb:Del, did not mature before the killing frost in either evaluation year. For lines containing these genotypes, the date of the frost was used as the maturation date, and thus the percentages for these genotypes do not represent the true length of the reproductive period. Most interestingly, lines fixed for e1-as/E1La/E1Lb, and lines fixed for E1/e1la:K82E/E1Lb, had significantly different reproductive lengths (average of 67 and 57%, respectively) (Welch’s two-sample t-test, t = 5.83, p < 0.001), highlighting a difference in the regulation of reproductive timing between E1 and E1La.

To understand the effects of the E1La alleles in a complex maturity background we created soybean lines that were segregating for two additional E genes and tested them in our MG III field environment. Further development of soybean germplasm targeted to MG III and MG V but selected for the e1la:K82E alleles was done to reduce the G. soja genetic background (Supplemental Figures 1 & 2). Soybean lines were developed that combined the e1la:K82E alleles with other maturity gene combinations present in MG I (e1-as e2 E3 E1La E1Lb) and MG II (e1-as E2 e3 E1a E1b) soybean varieties (Table 2 and Supplemental Table 4) [14]. A field experiment in our MG III environment for days to flower and days to maturity was conducted in 2020 with the new test lines and parents or controls (Table 3). All lines in the 2020 field experiment had functional E1Lb alleles. Similar to the 2018 and 2019 experiment, e1la:K82E lines in the MG III background (e1-as E2 E3 E1Lb) flowered about 8 days earlier and matured about 10 days earlier than the MG III background control lines (Table 3). The MG I and MG II lines with e1la:K82E alleles and either e2 or e3 alleles flowered and matured...
earlier than the control lines for MG I and MG II (Table 3). The new MG V (E1 E2 E3) lines fixed for e1la:K82E alleles flowered 4 days earlier and matured 8 days earlier than the MG V controls.

Geographic distribution of E1 and E1La Variant alleles of E1 and E1La are candidates for driving northern expansion of wild soybeans; we hypothesized that the geographic location of e1-as and e1la:K82E alleles in G. soja and G. max accessions would illuminate their origin and distribution. For a set of 92 G. soja Plant Introduction (PI) accessions from the Germplasm Resources Information Network (GRIN) categorized as Maturity Group II and earlier, we directly determined the alleles by Sanger sequencing; in addition, the alleles of E1 and E1La were assigned from re-sequencing data for the subset of 56 G. soja accessions from Zhou et al., 2015 for which latitude information could be obtained (Supplemental Table 5). The combined 148 G. soja accessions with their E1 and E1La genotypes were assessed for their geographic distribution across soybean’s center of

| Type   | E genotype | E1LA | n   | DTF | ΔDTF*a | DTM  | ΔDTM |
|--------|------------|------|-----|-----|--------|------|------|
| MG I   | e1-as      | e2   | 3   | 35.5|        | 110.5|      |
| MG I   | e1-asK82E  | e2   | 3   | 33.7| -1.8   | 98.5 | -12  |
| MG II  | e1-as      | e2   | 3   | 34.7| -3.3   | 105.3| -9.7 |
| MG III | e1-as      | e2   | 3   | 47.0|        | 126.0|      |
| MG V   | E1         | E2   | 2   | 39.1| -7.9   | 115.6| -10.4|
| MG V   | E1         | E2   | 2   | 59.0| -4     | 132.0| -8   |

*aDifference in days to flower (DTF) or days to maturity (DTM) for lines with contrasting E1La genotypes

Table 3 Days to flower and days to maturity for 2018 experimental test lines and their controls

Fig. 4 Number of days spent in the vegetative (lower/darker bar) and reproductive phases (upper/lighter bar) for e1laK82E test lines, e1lbDel test lines, and E1La E1Lb Reference (REF) controls in both e1-as and E1 backgrounds. Lines were categorized by their selected genotypes, and n represents the number of plots of lines (each replicated three times) per genotype category. e1laK82E is represented as e1la, e1lbDel is represented as e1lb, and E1La and E1Lb indicate the REF alleles. Bar charts represent the number of days after planting, with the percent of total life span for each phase indicated by (%). Error bars for each group represent the standard deviation. For days to maturity, the first frost in 2018 (152 days) and 2019 (134 days) is indicated, and all plots not matured at first frost were assigned that value for maturity. a Length of vegetative and reproductive phases of lines containing e1-as evaluated in 2018; b Length of vegetative and reproductive phases of lines containing e1-as evaluated in 2019; c Length of vegetative and reproductive phases of lines containing E1 evaluated in 2018; d Length of vegetative and reproductive phases of lines containing E1 evaluated in 2019.
origin in East Asia. The \( e1-as \) allele was somewhat rare and restricted geographically to far northern regions higher than 50° North latitude (Fig. 5a). The \( E1 \) with \( e1la:\text{K82E} \) allele combination was much more prevalent and spanned a larger latitudinal range, although it was almost entirely absent in accessions originating from below 40°N. Interestingly, northern \( G. \text{soja} \) accessions generally contained either \( e1-as \) or \( e1la:\text{K82E} \), but rarely both. Accessions containing the allele combination \( E1/E1La \) had the lowest mean latitude at 36.1°N, while accessions with the allele combinations \( E1/e1la:\text{K82E} \) and \( e1-as/E1La \) had higher mean latitudes at 51.4°N and 55°N, respectively (Fig. 5b).

To discern whether the \( e1la:\text{K82E} \) allele has been utilized in North American breeding programs, we conducted an expanded analysis of accessions contained in the GRIN using proxy SNPs from the SoySNP50k array determined to be in high association with either the \( e1-as \) causative mutation or the \( e1la:\text{K82E} \) causative mutation. The strength of association was estimated using a parameter called “combined pessimistic accuracy,” which is a pairwise calculation between each SoySNP50k marker and the causal mutation that determines the frequency of the Reference and Alternate haplotypes for each position (see Methods for additional details) [14, 27]. The SoySNP50k markers with the highest combined pessimistic accuracy to \( e1-as \) (ss715593865 – GM06: 20916554) and \( e1la:\text{K82E} \) (ss715587601 – GM04: 37750626) were used as proxy markers to assess geographic distribution among North American cultivars. The full list of all North American accessions available from the GRIN was filtered to contain only cultivars which had homozygous allele calls for the \( E1 \) and \( E1La \) proxy SNPs, and for which latitude and longitude information could be obtained. After filtering, the final

**Fig. 5** Geographic distribution of \( E1 \) and \( E1La \) haplotypes. \( a, b \) Subset of 56 \( G. \text{soja} \) accessions from Zhou et al., 2015, plus 92 additional Sanger sequenced \( G. \text{soja} \) accessions \( n = 148 \). \( c, d \) North American cultivars present in the GRIN for which \( E1 \) and \( E1La \) alleles were imputed using SNP50k proxy SNPs \( n = 592 \). Means comparisons for boxplots were conducted using an ANOVA, and significance groups were obtained using an LSD test with false discovery rate (FDR) correction \( P = 0.05 \). Maps of Asia and North America were obtained using the open-source R package ‘rnaturalearth’.
accession list contained a total of 594 cultivars (Supplemental Table 6). This analysis suggested that the e1la:K82E haplotype has likely been used only rarely in North American cultivar development, and that it was exclusive to lines adapted to the northern US and Canada (Fig. 5c). This is in contrast to the e1-as allele, which is used extensively in mid and northern MGs throughout the US. Mean latitudes for cultivars with allele combinations e1-as/e1la:K82E and E1/e1la:K82e had higher latitudes of 49.8°N and 48.8°N, respectively, when compared to e1-as/E1La (42.1°N) and E1/E1La-containing (37.8°N) cultivars (Fig. 5d).

To investigate whether variation in E1La may be present in soybean germplasm for which SoySNP50k genotype information was not available, we genotyped E1La from among a set of 26 natto and 19 tofu lines from the North Dakota State University breeding program. Interestingly, 24 of the 26 natto lines contained the e1La:K82E allele, however, all of the tofu germplasm possessed the Reference allele of E1La (Supplemental Tables 7 and 8). This is in contrast to our analysis of the geographic distribution of cultivars using a SoySNP50k proxy SNP, which suggested that the e1La:K82E allele was rare in North America. Expanding the use of the Proxy SNP for the e1la:K82E allele, we evaluated the frequency of accessions with imputed e1la:K82E alleles for G. soja and G. max GRIN accessions along with their country of origin. For G. max accessions, Japan was the origin for 46.8% of the imputed e1la:K82E alleles, while the distribution of e1la:K82E Proxy SNP was split between Russia, Japan, China, and South Korea for G. soja accessions (Supplemental Figure 3).

**Discussion**

Soybean is one of the most economically important crops worldwide, with adaptation to the correct photoperiod being critical for adequate yield. Several key genes controlling flowering time and maturity have been cloned and are being utilized extensively in breeding programs. However, many genes have been implicated in reproductive timing in soybean based on functional genomics, but the magnitude of their effects, and their prevalence in breeding programs, are not well understood. Further work is needed to characterize these lesser-known genes before they can be exploited to fine-tune the life cycle of soybean for different production environments.

E1 and its paralogues E1La and E1Lb exhibit similar expression patterns, principally a peak just after dawn and just before dusk under long days, and little or no expression under short days [18]. Lines with E1La and E1Lb down-regulated exhibited higher expression of the florigen promoting genes, FT2a and FT5a, and earlier flowering than control plants under artificial light, confirming that both functional genes inhibit flowering under long day conditions; however, this study was done in an e1-nl e2 E3 E4 genetic background [18]. In a study using incandescent lights to extend day length, a single-base deletion mapped to the E1Lb gene was shown to confer earlier flowering in a far-eastern Russian cultivar [19]. This E1Lb null allele was identified in a total of five Russian soybean cultivars that all had a maturity genotype of e1-as e2 e3 E4 [19]. RNAi suppression of E1 and its paralogues resulted in a near-complete loss of photoperiod sensitivity and was sufficient to convert an extremely late-flowering MG VIII cultivar to MG 000 [22]. Our research characterized the role that E1La and E1Lb each have independently on flowering time and maturity under a natural photoperiod. We identified a lysine to glutamate missense mutation in the E1La gene from among a set of publicly available re-sequenced accessions, and identified an induced deletion of the E1Lb gene, from which we developed lines in both e1-as and E1 backgrounds. Our results suggest that compared to their variant alleles, functional versions of each of the three members of the E1 gene family are together contributing to the repression of FT2a and FT5a; therefore, E1, E1La, and E1Lb suppress soybean flowering and maturity under natural long day photoperiod conditions, consistent with previous gene expression studies [18].

Our field experiments with natural light provided environments that represent soybean production scenarios for maturity group III that are optimized for the the variant e1-as alleles along with functional versions of the E2, and E3 genes [14]. The summer solstice at our field location provides 14 h and 54 min of daylight from sunrise to sunset. The daylength typically reaches its maximum and has begun to shorten prior to soybean plants flowering in the field. Taken together, the results demonstrated that, similar to its parologue E1, E1La functions to delay flowering and maturity under long day conditions, with the e1La:K82E allele having a stronger effect on promoting maturity in an e1-as background than in an E1 background. The e1la:K82E alleles also appeared to promote flowering and maturity in genetic backgrounds with additional defects in the major maturity genes E2 and E3 when e1-as alleles were present. Likewise, the E1Lb gene functions to delay flowering and maturity in a partially functional e1-as background; however, the ability of e1Lb:Del to promote flowering appears to be abolished in a fully functional E1 background. It appears that a natural null allele of E1Lb has contributed to adaptation of some soybean cultivars in Russian production environments [19]. While the magnitude of phenotypic effects of the E1L genes are different under natural light regimes than artificial light, our results show similar trends to those published in previous reports describing E1La and E1Lb, with deficiencies in...
**E1La, E1Lb** or both leading to earlier flowering and maturity [18, 19, 22]. Our experiments were intended to provide practical information and novel alleles of new maturity genes that could be used in the context of the established maturity gene combinations to fine-tune the timing of flowering and maturity to optimize photoperiod sensitivity for enhanced yield potential in existing soybean production environments.

In addition, we assessed the role that allelic combinations of *E1* and *E1La* have played in adaptation of wild and cultivated soybean to northern latitudes. This analysis revealed that the *e1la*:K82E allele is present in high frequency in *G. soja* accessions adapted to higher latitudes, and that the *e1*-as allele is relatively rare. An analysis of varieties released in North America using SoySNP50k proxy SNPs suggested a heavy reliance on *e1*-as to develop cultivars adapted to northern production environments, but little use of *e1la*:K82E. However, a direct genotyping analysis of the *E1La* gene in a specialty breeding program in North Dakota revealed that the *e1la*:K82E allele is being exploited to develop natto cultivars. Indeed, there was an apparent high frequency of the *e1la*:K82E allele based on the proxy SNP in wild and cultivated soybeans originating in Japan, where natto is a traditional soyfood [28]. Together, these results revealed that wild soybean and North American breeding programs have exploited different members of the *E1* gene family as the predominant source of reducing photoperiod sensitivity; however, variation in *E1La* may play an important role in adaptation of North American cultivars to far northern latitudes. In concert with the disparity in reproductive lengths we observed between lines fixed for *e1*-as and *e1la*:K82E, this also explains, at least in part, the shorter reproductive phase generally observed in *G. soja* accessions, when compared to *G. max*.

**Conclusions**

We identified natural and induced variation in the *E1* paralogues, *E1La* and *E1Lb*, and demonstrated that these variant alleles independently promoted earlier flowering and maturity. Initial efforts suggested that variation in these genes is rare in North American breeding programs; however, further investigation revealed that variation in *E1La* is being exploited in a specialty breeding program in North Dakota. These novel alleles of *E1La* and *E1Lb* constitute valuable resources in a breeder’s toolbox for better adaptation of germplasm to northern production environments.

**Methods**

**Natural and induced variation in *E1La* and *E1Lb***

**E1La**

The 302 soybean accessions with whole genome re-sequencing data [23] were evaluated in the haplotype visualization tool, SNPViz [24] for variant genomic sequence positions in *E1La* (Glyma04g24640; Wm82.a1.v1.1) on chromosome 04 in the region around position 28,293,933 to 28,294,806, and ten *G. soja* accessions contained a haplotype that included a nonsynonymous A/G variant at position 28,294,378. The ten *G. soja* accessions also contained a synonymous variant (G/T) at Gm04:28,294,356 that was present in an additional 14 *G. soja* accessions. The soybean allele catalog (http://soykb.org/GenescapeAnalysis/search.php) was used to assess the distribution of alleles of *E1La* and *E1Lb* in our curated data set of 775 whole genome re-sequenced soybean accessions [23, 25]. A protein blast at NCBI, using the Williams 82 reference peptide sequence of *E1La*, was used to obtain the orthologous sequences from 23 different legume species with the highest percent identity (Supplemental Table 2). Multiple sequence alignment was generated using the “msa” R package, and the weblogo (trimmed to 20 amino acids) was generated using the “ggplot2” R package.

**E1Lb**

Soybean seeds of cultivar Williams 82 were originally obtained from the GRIN and irradiated with fast neutrons (FN) at 20, 25, 30, and 35 Gy at the McClellan Nuclear Radiation Center (University of California, Davis). To determine the copy number variation (CNV) events induced in the mutagenized population, select mutants were analyzed by comparative genomic hybridization (CGH) using a Roche NimbleGen 696,139-feature soybean CGH microarray following previously published protocol [26, 29]. The oligonucleotide probes of 50- to 70-mers spaced at approximately 1.1 kb intervals were designed based on the Williams 82 genome sequence (Wm82.a1 version). Copy number variation events were called following previously set criteria [29]. Based on the detected CNVs from the CGH analysis, we identified a mutant (MO12) harboring ~2.6 Mbp deletion on chromosome 04 encoding the *E1Lb* gene [26].

As part of a separate project, seeds from the entire set of *G. soja* accessions were obtained from the GRIN, and a subset of 419 of the accessions phenotyped for maturity group II or earlier was selected for characterization. DNA was isolated from ground seed tissue using the DNeasy Plant Mini Kit (Qiagen, Inc., Valencia, CA) according to the manufacturer’s instructions. Samples were first evaluated for *E1* or *e1*-as alleles using our established SimpleProbe assay [14], and of those that were successfully genotyped, 57 were *e1*-as, and 244 accessions were *E1*. Subsequently, 84 *E1* and 20 *e1*-as accessions were evaluated for their *E1La* sequences. Sanger sequencing at the University of Missouri DNA Core Facility of 642 bp *E1La* PCR amplicons from *G. soja* accessions utilized PCR primers *E1La*-F1: 5′- A AACAC TCAAAGCCCGATCA-3′ and *E1La*-R2: 5′- GATTGTGAAAGTGAAAGCTAACACAG-3′ as described.
previously, except ampiclons were isolated by ethanol precipitation prior to sequencing with the primer E1La-F1 [30].

**Molecular marker assays** Plant samples for genotyping were tagged and leaf presses were prepared on FTA cards as described [31]. A new Simple probe genotyping assay was developed to track E1La or e1La:K82E alleles. Reactions were carried out in 20 μl containing template, primers, 0.2 μM final concentration of SimpleProbe, buffer (40 mM Tricine-KOH [pH 8.0] 16 mM MgCl2, 3.75 μg ml−1 BSA.), 5% DMSO, 200 μM dNTPs, and 0.2X Titanium Taq polymerase (BD Biosciences, Palo Alto, CA). The e1La:K82E antisense probe was 5′-Fluorescein-SPC-GGCAAATTTTGCTCCCTCCAAATC-Phosphate-3′. Originally, the primers E1La-F1 and E1La-R2 were used in the assay, but difficulties amplifying from FTA card samples led to the replacement with nE1La-F1 (5′-GGGAGTTTCAACAACACTGAAGC-3′) and nE1La-R2 (5′-GGTGTCATGTCCCCAAACTCTAAC-3′) that targeted a 233 bp product. In both cases, the forward primers (E1La-F1 or nE1La-F1) were used at 5 μM final concentration and the reverse primers (E1La-R2 or nE1La-R2) were used at 2 μM final concentration in the PCR. Genotyping reactions were performed using a Lightcycler 480 II real time PCR instrument (Roche), using the following PCR parameters: 95 °C for 5 min followed by 40 cycles of 95 °C for 20 s, 60 °C for 20 s, 72 °C for 20 s, and then a melting curve from 50 °C to 70 °C.

A Tm-shift assay [32], based on GC primer tails of differing lengths, was developed to discriminate between E1Lb and e1Lb:Del alleles. Because the e1Lb:Del allele is a deletion of the entire E1Lb gene, this assay was unable to distinguish between the REF E1Lb allele and lines that were heterozygous for E1Lb. Reactions were carried out in 20 μl containing template, primers, 0.063 μM final concentration of EvaGreen Fluorescent Dye (Biotium) was used as the dye at 0.063 μM final concentration [33], and primers for Tof11 were DTF1tailf: 5′-GCAAACACCTTGAACATAGAAT-3′, DTF1tailR: 5′-gccggccaggggccAGCCACATTGCCATTTC-3′, and DTF1tailGsr2 (5′-gccggcAGCCACATTGCCATTTC-3′). Primers for Tof12 were DTF2TailW82 (5′-gctgggagggcggcCA-TAAAAGCTGCAGTAGATACCT-3′), DTF2TailGs (5′-gctgggCATAAAGCTGCAGTAGATTC-3′), and DTF2TailR1 (5′-GCATTGATGATACACATTGC-3′).

**Development of mutant e1la and e1l populations**

**Plant materials**

Seeds of *G. soja* accessions PI52226 and PI547831 containing e1La:K82E alleles were obtained from the GRIN. A line from fast neutron mutagenesis of Williams 82 was the source of the W82 FN e1Lb:Del alleles. Jake is a MG V determinate cultivar provided with permission by the developer [34], Williams 82 is a MG III indeterminate cultivar obtained from the GRIN [35], Deuel is a MG I indeterminate cultivar released by South Dakota Agricultural Experiment Station, PVP 201000318 and provided with permission by the developer; Brookings is a late MG I cultivar released by South Dakota Agricultural Experiment Station, and provided with permission by the developer, LG04–6000 is a MG IV indeterminate cultivar provided with permission by the developer [36], Candor is an early MG II cultivar provided with permission from Sevita International, Ellis HOLL is an experimental seed composition MG V determinate line from the University of Tennessee provided with permission by the developer, and the EXP e3 line was an experimental seed composition line verified to contain e3-tr alleles developed by the authors [4, 24]. The genes relevant to this work as parent lines are classified as having the reference Williams 82 alleles (REF) or the indicated alternate alleles specific to each gene (Table 3).

**Breeding schemes**

Soybean populations were developed with different parent combinations (Table 2). Generally, F1 seeds were produced at the South Farm Research Center near Missouri, during the summer field season followed by a cycle of self-pollination that produced F2
seeds in the winter nursery near Upala, Costa Rica. During the second cycle in the winter nursery, the F₃ plants were sampled and underwent genotypic selection prior to single plant harvest of F₂:₃ seeds. To isolate the e₁la:K₈₂E alleles out of the G. soja genetic background, breeding efforts with PI₅₄₇₇₈₁ and PI₅₂₂₂₂₆ with G. max parents were directed at selecting for e₁la:K₈₂E alleles and the desired E₁ alleles leading to the parent line KB₁₆-2B #6₆₆₆, which was still segregating at E₁ and the parent line KB₁₆-W₂F₃, which was fixed for e₁-as (Table 3, Supplemental Figure 1, and Supplemental Figure 2). Desirable alleles of tof₁₁–₁ and tof₁₂–₁ were also confirmed by genotyping for the parent lines (Supplemental Table 4). Additional breeding was conducted to reduce the G. soja genetic background for new parent lines with the e₁la:K₈₂E alleles, (Supplemental Fig. 1B and 1C). The resulting lines KB₁₇–2#₅₁₄₅ and KB₁₇–1₄₄₈₁ were used both as new parent lines for populations 9 (E₁) and 5, 6, and 7 (e₁-as), respectively, as well as for phenotypic analysis as parts of the population 1 and 2 experimentally tested lines (Table 1; Table 2 and Supplemental Table 4).

A line carrying the e₁lb:Del alleles (W₈₂ FN) was identified from a set of fast neutron mutagenized Williams 8₂ lines, and lines with E₁ or e₁-as alleles along with e₁lb:Del alleles were selected by genotype from two populations. Population 3 lines were selected for e₁lb: Del/E₁ and population 4 lines were selected for e₁lb: Del/e₁-as and (Table 2).

Other populations were made following the general strategy breeding cycle of creating F₁ seeds, advancing to F₂ plants in winter nursery for genotypic selection of e₁la:K₈₂E with E₁, e₁-as, e₂, or e₃; harvest of F₂:₃ seeds; then one generation advance to F₃:₄ in Columbia, Missouri for seed supply for plots for the 2020 field experiment. The population information including the parents for each population is listed in Table 2, the parents and test lines are listed with their genotypes in Table 1, and the experimental categories and genotypes from the populations utilized for the field experiments are listed in Supplemental Table 4.

**Evaluation of e₁la and e₁lb populations for flowering time and maturity**

Two field experiments were conducted, one that included selections from populations 2–5 in 2018 and 2019 (1₈/1₉) and one that included selections from populations 6–1₀ in 2₀₂₀ (2₀) (Table 2; Supplemental Table 4). For the 1₈/1₉ experiment, F₃ lines fixed for e₁la:K₈₂E and e₁lb:Del, along with E₁La and E₁lb reference control lines, were planted in 3’ plots (1’ planted, 2’ alleys) on May 1₅th, 2₀₁₈ at the South Farm Research Center in Columbia, Missouri. The following year, F₄ lines fixed for e₁la:K₈₂E and e₁lb:Del, along with E₁La and E₁lb reference control lines, were planted in 5’ plots (3’ planted, 2’ alleys) on May 3₁st, 2₀₁₉ in Columbia, Missouri. In both years, lines were grown in a randomized complete block design with three replicates per line and were scored for flowering time (R₁) and maturity (R₈) as a function of days after planting (DAP). In 2₀₁₈, the R₁ dates for each plant were averaged to get the mean R₁ date for each plot. In 2₀₁₉, plots were marked as R₁ once flowers were observed on at least three plants in the plot. In both years, plots were marked as R₈ once 9₅% of pods on the main stem were mature. First frost occurred on October 1₆th in 2₀₁₈ (day 1₅₂) and on October 1₂th in 2₀₁₉ (day 1₃₄). For the statistical analysis, any plot that did not mature by this time was given an R₈ score of the day of the first frost. Means comparisons were conducted using an ANOVA in R, and significance groups were obtained using a Fisher’s LSD test with false discovery rate (FDR) correction (P = 0.₀₅). For the 2₀ experiment, 5₀-seed plots of F₃:₄ lines along with controls were planted in random order in 5’ plots (3’ planted, 2’ alleys) on June ₁st, 2₀₂₀ in Columbia, Missouri. Genotypes were replicated, but lines were not. Plots were marked as R₁ once flowers were observed on at least three plants in the plot, and maturity was estimated for 9₅% mature pods on the main stem, or maturity was forecasted 2–₃ days in advance. The first frost occurred on October 1₆th in 2₀₂₀ (day ₁₃₇).

**Geographic analysis of natural variation in E₁ and E₁La**

E₁ and E₁La genotypes for 5₆ re-sequenced Glycine soja accessions were obtained from Zhou et al., 2₀₁₅ analyzed in SNPViz [2₄]. E₁ and E₁La genotypes for 9₂ additional MG II or earlier Glycine soja accessions were obtained from Sanger sequencing. As North American cultivars derived from early maturity groups were under-represented in our resequencing panel, cultivars with SoySNP5₀k data were instead pulled from the GRIN. The E₁ and E₁La genotypes for each cultivar were estimated using a proxy SNP from the SoySNP5₀k array. E₁ and E₁La genotypes were first determined for a set of 7₇₅ resequenced accessions [2₃, 2₅]. Strength of association between the e₁-as and e₁la:K₈₂E causal mutations and all of the SoySNP5₀k variants within 1 Mb of the causal mutations on chromosomes 0₆ and 0₄, respectively, was calculated using a parameter called “combined pessimistic accuracy.” Accuracy for each SNP is calculated as the percentage of the 7₇₅ resequenced accessions with either the REF or ALT haplotype combinations between SNP and causal mutation.
This “combined pessimistic accuracy” equation is modified from Metz (1978) to capture combined sensitivity, specificity, and missing data [37]. The single SNP on chromosomes 06 (ss715593865 – GM06:20916554; Accuracy = 94.1%) and 04 (ss715587601 – GM04: 37750626; Accuracy = 92.0%) with the highest accuracy values were used to estimate $E1$ and $E1La$ genotypes for all North American cultivars pulled from the GRIN (only cultivars for which latitude and longitude coordinates could be obtained, and proxy SNP genotypes were homozygous, were used). To avoid overplotting on map figures, genotype groups containing more than 200 cultivars were randomly sampled. Latitude and longitude coordinates for plotting were obtained directly from the GRIN, where available. For accessions where coordinates were not available in GRIN, Google geocoding was used to obtain latitude and longitude coordinates for the state/province of origination. A small amount of “randomness” was introduced to plotting coordinates to prevent complete overlap of accessions originating from the same state/province.

For the frequency distribution of imputed $e1La$K82E alleles by country of origin, country of origin assignments for the 476 $G. \text{max}$ and 499 $G. \text{soja}$ accessions containing the $e1La:K82E$ SoySNP50k Proxy SNP (ss715587601 – GM04: 37750626) were obtained from the GRIN. Accessions missing country of origin information in the GRIN were assigned a value of “Unknown” for plotting. Histograms were generated in R v4.0.2 using the ggplot2 package, v3.3.2.

DNA was prepared as described previously from a single dry seed from 19 tofu and 26 natto experimental lines from the North Dakota State University soybean breeding program [14]. SimpleProbe genotyping assays for $E1/e1-as$ and $E1La/e1La$K82E were conducted as described above with the DNA samples and controls.

**Acknowledgements**

The authors acknowledge excellent technical assistance of Christine Cole and Mark Johnston. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

**Authors’ contributions**

ND and KB conceived and designed the experiments and analyzed data; ND, R C-G, GC, and KB performed experiments; CM and MS contributed materials; ND and KB wrote the manuscript with contributions from R C-G, MS, and CM. All authors have read and approved the manuscript.

**Funding**

Not applicable.

**Availability of data and materials**

The datasets supporting the conclusions of this article are included within the article (and its additional files). SoySNP50k genotype data for the $e1-as$ and $e1La$K82E proxy SNPs were obtained from Soybase (https://soybase.org/). Geographic coordinates and origin information for accessions were obtained from the Germplasm Resources Information Network search page (https://npgsweb.ars-grin.gov/npgsglobal/search). The data from both repositories are freely accessible to the public for download. Seeds were obtained with permission from the USDA National Plant Germplasm System, from line or cultivar developers, or developed as part of this research.

**Declarations**

**Ethics approval and consent to participate**

Research and field studies on plants (either cultivated or wild), including the collection of plant material, complied with relevant institutional, national, and international guidelines and legislation.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

**Author details**

1Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA. 2Department of Plant Sciences, North Dakota State University, Fargo, ND 58108, USA. 3USDA/ARS Plant Genetics Research Unit, Columbia, MO 65211, USA.

**Supplementary Information**

The online version contains supplementary material available at https://doi.org/10.1186/s12870-021-03197-x.
