Article

Genetic Variation in Soybean at the Maturity Locus \textit{E4} Is Involved in Adaptation to Long Days at High Latitudes

Yasutaka Tsubokura ¹,†, Hisakazu Matsumura ²,†, Meilan Xu ³,†, Baohui Liu ⁴,*; Hiroko Nakashima ⁵, Toyoaki Anai ⁶, Fanjiang Kong ⁴, Xiaohui Yuan ⁴, Hiroyuki Kanamori ⁶, Yuichi Katayose ¹, Ryoji Takahashi ²,⁷, Kyuya Harada ¹ and Jun Abe ³,*

¹ National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan; E-Mails: Tsubokura@snowseed.co.jp (Y.T.); katayose@affrc.go.jp (Y.K.);
haradaq@nifty.com (K.H.)
² Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki, 305-8518, Japan; E-Mails: sm2193@cc.saga-u.ac.jp (H.M.); masako@affrc.go.jp (R.T.)
³ Research Faculty of Agriculture, Hokkaido University, Sapporo, 060-8589, Japan; E-Mail: xumeilan_1984@yahoo.co.jp
⁴ Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin, 150040, China; E-Mails: kongfj@neigaehrb.ac.cn (F.K.); yuanxh@neigaehrb.ac.cn (X.Y.)
⁵ Saga University, Saga, 840-8502, Japan; E-Mails: nakashima-hiroko@miyajima-soy.co.jp (H.N.);
anai@cc.saga-u.ac.jp (T.A.)
⁶ Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki 305-0854, Japan; E-Mail: kanamo@affrc.go.jp
⁷ National Institute of Crop Science, Tsukuba, Ibaraki, 305-8518, Japan

† These authors contributed equally to this work.

* Authors to whom correspondence should be addressed; E-Mails: jabe@res.agr.hokudai.ac.jp (J.A.);
liubh@neigaehrb.ac.cn (B.L.); Tel.: +81-11-706-2442 (J.A.); +86-451-86685735 (B.L.);
Fax: +81-11-706-4933 (J.A.); +86-451-86603736 (B.L.).

Received: 11 December 2012; in revised form: 10 January 2013 / Accepted: 25 January 2013 / Published: 4 February 2013

Abstract: Soybean (\textit{Glycine max}) cultivars adapted to high latitudes have a weakened or absent sensitivity to photoperiod. The purposes of this study were to determine the molecular basis for photoperiod insensitivity in various soybean accessions, focusing on the sequence diversity of the \textit{E4} (\textit{GmphyA2}) gene, which encodes a phytochrome A (phyA) protein, and
its homoeolog (GmphyA1), and to disclose the evolutionary consequences of two phyA homoeologs after gene duplication. We detected four new single-base deletions in the exons of E4, all of which result in prematurely truncated proteins. A survey of 191 cultivated accessions sourced from various regions of East Asia with allele-specific molecular markers reliably determined that the accessions with dysfunctional alleles were limited to small geographical regions, suggesting the alleles’ recent and independent origins from functional E4 alleles. Comparison of nucleotide diversity values revealed lower nucleotide diversity at non-synonymous sites in GmphyA1 than in E4, although both have accumulated mutations at almost the same rate in synonymous and non-coding regions. Natural mutations have repeatedly generated loss-of-function alleles at the E4 locus, and these have accumulated in local populations. The E4 locus is a key player in the adaptation of soybean to high-latitude environments under diverse cropping systems.

**Keywords:** genetic diversity; soybean; flowering; phytochrome A; photoperiod insensitivity

### 1. Introduction

Flowering time determines the adaptability of plant species to diverse environments. Molecular dissection of flowering behavior in Arabidopsis thaliana has revealed that the transition from vegetative to reproductive growth is under the control of a complicated network involving more than 60 genes [1]. Natural allelic variation has been surveyed to explore the molecular mechanisms underlying the adaptation of Arabidopsis to diverse environments, but has been identified in only a few of these flowering genes [2]. Among the genes, FRIGIDA (FRI) and FLOWERING LOCUS C (FLC), which are pivotal regulators of the vernalization pathway, each exhibit a high degree of functional polymorphism, which likely underlies the extensive natural variation in flowering time [3–7]. More than 25 independent loss-of-function alleles have so far been described at the FRI locus [2]. Most early-flowering spring annual ecotypes have evolved multiple times from late-flowering winter annual ancestors through independent loss-of-function mutations in one or both genes.

Natural variations in flowering time have been explored in another model plant species, rice (Oryza sativa). In contrast to Arabidopsis, rice is a short-day plant with no response to vernalization. The association of variation in flowering time with DNA polymorphisms of major flowering genes has been assayed in a core collection of 64 rice cultivars [8]. Flowering time in rice is closely correlated with expression levels of Hd3a, a rice ortholog of Arabidopsis FLOWERING LOCUS T (FT). Variability in the expression of Hd3a is due in part to sequence variations in the Hd3a promoter region, but is also likely to be affected by different alleles of Hd1, a rice ortholog of Arabidopsis CONSTANS (CO). Sequencing Hd1 in the core collection identified 17 haplotypes, 9 of which are nonfunctional owing to frame-shift and nonsense mutations; the presence of these mutations suggests that polymorphism in Hd1 is one of the main causes of the diversity of flowering time in rice [8]. The findings in Arabidopsis and rice thus suggest that different genes underlie the natural variation in the control of flowering in these two species, and that independently induced mutations at a few key loci have repeatedly contributed to the natural variation in flowering of the two species.
Soybean (*Glycine max*) is cultivated over a wide range of latitudes, from the equator to high latitudes of at least 50° N. However, each cultivar is restricted to a relatively narrow range of latitudes. The wide adaptability of soybean has thus been created by natural variation in a number of major genes and quantitative trait loci (QTL) that control flowering behavior. Soybean is a short-day plant, and flowering is induced when the day length is shorter than a critical length. This sensitivity to photoperiod is weak or absent in soybean cultivars adapted to high latitudes, which should initiate flowering under long-day (LD) conditions of early summer to mature within limited frost-free seasons. Four major maturity loci—*E1*, *E3*, *E4*, and *E7*—have so far been reported to be involved in the control of this insensitivity [9–15, reviewed in 16]. Recent molecular analyses have revealed that *E3* and *E4* encode the phytochrome A (phyA) proteins *GmphyA3* and *GmphyA2*, respectively [17,18]; and that *E1* encodes a protein that contains a putative bipartite nuclear localization signal and a region distantly related to a B3 domain, and controls time to flowering by suppressing the expression of two soybean orthologs of *FT*, *GmFT2a* and *GmFT5a*, under the regulation of *E3* and *E4* [19]. A phyA-regulated *E1-GmFT* pathway is thus a key determinant of the adaptation of soybean to long-daylength environments.

The genetic mechanism of the photoperiod insensitivity varies among cultivars [16]. Genetic analyses have revealed that soybean cultivars and landraces that are adapted to the cool summers of northern Japan possess recessive genotypes at the *E3* and *E4* loci, namely *e3e3e4e4* [20,21]. Another group of photoperiod-insensitive cultivars are grown mainly as a short-season crop across Japan and the Korean peninsula. One of these cultivars, Sakamotowase, has the *e3e3E4E4* genotype [20,21] and an allele at or a gene tightly linked to the *E1* locus that controls the insensitivity in the presence of *E4* [21]. Xia *et al.* [19] analyzed the *E1* sequence of Sakamotowase, and found that it possessed a dysfunctional allele, *e1-fs*, that produced a truncated protein that was unable to suppress the function of *GmFTs* owing to a premature stop codon due to a frame shift caused by a single-base deletion. The photoperiod insensitivity of Sakamotowase is thus most likely controlled by a dysfunctional allele at the *E1* locus under the genetic background of the *e3e3E4E4* genotype. Accordingly, at least two genetic mechanisms are known so far to be involved in the insensitivity to photoperiod of soybean.

In addition to these two cultivar groups, various other landraces and cultivars that are adapted to LD conditions of high latitudes are also insensitive to photoperiod, but the genetic mechanisms involved are unknown. Here, we report that most of these cultivars possess independently induced dysfunctional alleles at the *E4* locus. Our data suggest that independent mutations at this locus have contributed to the adaptation of soybean to LD conditions of high latitudes.

2. Results

2.1. Classification of Photoperiod-Insensitive Soybean Accessions

The genetic variation underlying photoperiod insensitivity was surveyed for the 27 accessions collected from various regions of East Asia (Figure 1A). When sown in the late of May, these accessions flowered in the middle to late of July in Sapporo, Japan (43°06’ N, 141°35’ E) in which the natural daylength including twilight reached a maximum of 16.5 h, and exhibited no marked delay in flowering in artificially-induced LD conditions of 20 h generated by incandescent lamps. They were classified into three distinct groups (I–III), two singletons (IV, V), and a separate group (VI),
by means of UPGMA cluster analysis of the combined data for 11 isozymes and 9 SSRs (Figure 1B). Similar results were obtained in the analyses using isozymes and SSRs separately (data not shown). Group I consisted mainly of the landraces from Hokkaido, Japan, and far-eastern Russia (accessions 1–8), including Miharudaizu, whose genotype at the $E1$, $E3$, and $E4$ loci was determined as $E1E1e3e3e4e4$ [20]. The Group I accessions from northern Japan (accessions 1–7) possessed the same genotype at all of the marker loci tested, although they differed in their time to flowering and their seed coat colors. Group II consisted of landraces that are grown as a short-season crop across Japan and the Korean peninsula (9–15), including Sakamotowase, which has the genotype $e3e3E4E4$ and a dysfunctional allele ($e1-fs$) at the $E1$ locus [19–21]. Group III consisted of landraces collected in northern Honshu, Japan (16–21). Tsukue-4 (group IV, accession 22) and Otomewase (group V, accession 23) together formed a loose clade. The accessions from northeastern China and far-eastern Russia (24–27) formed a loose clade that was separate from the other five groups. The genotypes at the $E1$, $E3$, and $E4$ maturity loci of the accessions in groups III to VI have not yet been determined.

**Figure 1.** Photoperiod-insensitive early-maturing soybean accessions analyzed in this study. (A) Geographical distribution of the photoperiod-insensitive accessions from East Asia; (B) Classification of the photoperiod-insensitive accessions by means of UPGMA, based on the similarity of 20 polymorphic isozyme and SSR marker loci.

### 2.2. Sequence and DNA Marker Analyses of $E4$

A crossing experiment between Miharudaizu (Group I) and Kamaishi-17 (Group III) exhibited no transgressive segregation in flowering time under the artificially-induced LD conditions (our unpublished data). This suggests that Kamaishi-17 possesses the same genotype ($e3e3e4e4$) at the $E3$ and $E4$ loci as is the case in Miharudaizu. The dysfunction of $e4$ allele in Miharudaizu is caused by an insertion of a Ty1/copia-like retrotransposon, SORE-1, in exon 1 [17]. An analysis with allele-specific
markers, which detect the presence or absence of the insertion, however, revealed that Kamaishi-17 did not have the e4 allele in which SORE-1 had been inserted [22]. This prompted us to analyze the E4 (GmphyA2) sequences of Kamaishi-17 and the other photoperiod-insensitive accessions.

The sequence analysis revealed that Kamaishi-17 had a single-base deletion at position 3085 from the adenine of the start codon in exon 2 (Figure 2A). This deletion resulted in a frame shift that led to premature termination of translation, and the gene was thus predicted to produce a truncated protein of 894 amino acids (AA) in length (Figure 2B). The result obtained from the sequence analysis was thus in good agreement with our expectation from the crossing experiment, indicating that Kamaishi-17 possessed a dysfunctional e4 allele, as was the case in Miharudaizu. We then extended the sequence analysis to the other three accessions, which were selected from each of groups IV to VI (Tsukue-4, Otomewase, and Keshuang). Interestingly, all had single-base deletions at different sites in exons 1 (Otomewase) or 2 (Tsukue-4 and Keshuang), and these variants were predicted to produce truncated proteins of different lengths: 456 AA in Otomewase, 759 AA in Tsukue-4, and 979 AA in Keshuang (Figure 2A,B). The SORE-1-inserted e4 allele produced a truncated protein of 237 AA [17]. The predicted AA sequences produced in Kamaishi-17 and Keshuang lacked a histidine-kinase domain required for phosphorylation, but retained the two PAS domains (PAS1 and PAS2) that are important for downstream signaling, whereas the Otomewase variant lacked all three domains, and the Tsukue-4 variant lacked both the PAS2 and histidine-kinase domains. No other DNA polymorphism was detected in the sequences, other than these deletions, among the accessions we tested or between those and Williams 82, a cultivar that was used for whole-genome sequencing (Glyma20g22160). We designated these variant alleles after the names of cultivars: e4-oto in Otomewase, e4-tsu in Tsukue-4, e4-kam in Kamaishi-17, and e4-kes in Keshuang.

Figure 2. Independent single-nucleotide deletions in the E4 gene that result in premature stop codons that produce dysfunctional truncated proteins. (A) Sites of deletions and of the insertion of a Ty1/copia-like retrotransposon, SORE-1, in the GmphyA2 exons. The deleted nucleotides are presented in parentheses; (B) Premature stop codons (*) generated by the deletions or by the insertion of SORE-1.
We then developed markers to reliably determine which alleles the remaining photoperiod-insensitive accessions possessed (Figure 3). The five other accessions from Group III had a PCR product with the same digestion pattern as Kamaishi-17: when digested by *Afl*II, the amplified 494-bp product was separated into fragments of 286 and 208 bp. Similarly, the three accessions of Group VI from far-eastern Russia (Zeya-2, Oktyabr-70, and Severnaya-4) had a PCR product with the same digestion pattern as Keshuang: when digested by *Bsp*HI, the amplified 494-bp product was separated into fragments of 399 and 95 bp. The digestion patterns observed in e4-oto and e4-tsu were not detected in the rest of the collection of 27 accessions. Furthermore, the marker analyses for the four alleles and the e4 allele containing the SORE-1 insert revealed that the accessions in Group II all possessed the dominant E4 allele, like Sakamotowase, whereas those in Group I all possessed the e4 allele containing the SORE-1 insert, like Miharudaizu. Accordingly, all of the photoperiod-insensitive accessions except for the Group II accessions had different loss-of-function alleles due to single-base deletions or the insertion of SORE-1.
Figure 3. Allele-specific DNA markers for the loss-of-function alleles at the E4 locus. Harosoy possesses a dominant E4 allele. The accessions, Otomewase, Tsukue-4, Kamaishi-17, and Keshuang possess loss-of-function alleles, e4-oto, e4-tsu, e4-kam, and e4-kes, respectively.

2.3. Survey of Genetic Variation Using Allele-Specific DNA Markers

To determine the geographical distributions of the newly detected dysfunctional alleles at the E4 locus, we extended the marker analysis to a total of 164 cultivated soybean accessions sourced from East Asia (64 from China, 30 from Korea, and 70 from Japan; Supplemental Table S1), in addition to the 27 photoperiod-insensitive accessions. The digestion patterns at each of the four markers indicated that all of the accessions except for a landrace from northern Japan (Wasekeburi) possessed the dominant E4 allele. Wasekeburi possessed the e4-kam allele, which was distributed mainly in northern Honshu (Figure 1A). The other loss-of-function alleles were not detected in this collection. Therefore, these loss-of-function alleles appear to be rare in the cultivated soybean germplasm.
2.4. Comparison of Nucleotide Diversity between E4 and GmphyA1

Sequencing and DNA marker analyses revealed that photoperiod-insensitive accessions, except for those in Group II, possessed dysfunctional alleles at the E4 (GmphyA2) locus. The E4 gene possesses a homoeologous copy, GmphyA1 (Glyma10g28170) [17], owing to the paleopolyploid nature of the soybean genome [23]. This raises questions about whether there are dysfunctional mutations responsible for earlier flowering in GmphyA1, and about the function, if any, of GmphyA1. To answer these questions, we sequenced E4 and GmphyA1 in wild and cultivated accessions collected from various regions of East Asia (Supplemental Tables S2 and S3).

Characteristics of the DNA polymorphisms in the E4 and GmphyA1 regions are summarized in Table 1. Recently released sequence data for 31 wild and cultivated soybeans [24], excluding those with missing or obscure data, were also included in the nucleotide diversity analysis. The E4 region comprised a total of 6341 aligned base pairs; across this region, 44 sites were polymorphic, comprising 33 SNPs, 9 single- or multiple-base insertion–deletions (indels), and 2 SSRs (Supplemental Figure S1). In addition to the 4 single-base deletions and the insertion of SORE-1, 10 SNPs occurred in exons, of which 4 generated amino acid substitutions. On the other hand, the GmphyA1 region comprised a total of 5517 aligned base pairs; across this region, 20 sites were polymorphic. Of these, 7 were detected in exons, of which only 1 SNP caused an amino acid substitution (Supplemental Figure S2). The analysis of GmphyA1 included 9 photoperiod-insensitive accessions that were analyzed for the E4 sequences. No sequence variation causing a dysfunction of GmphyA1 was detected in the 9 accessions; all had the same amino acid sequence as Williams 82, a photoperiod-sensitive cultivar (Supplemental Table S3 and Supplemental Figure S2).

Table 1. DNA polymorphisms in two homoeologous phytochrome A genes, GmphyA1 and E4 (GmphyA2) in cultivated and wild soybeans.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>S</th>
<th>Hap</th>
<th>$\pi$ (× $10^{-3}$)</th>
<th>$\theta$ (× $10^{-3}$)</th>
<th>$\pi$ (× $10^{-3}$)</th>
<th>$\theta$ (× $10^{-3}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GmphyA1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cultivated soybean</td>
<td>35</td>
<td>17</td>
<td>6</td>
<td>1.35</td>
<td>1.03</td>
<td>0.02</td>
<td>0.09</td>
</tr>
<tr>
<td>Wild soybean</td>
<td>24</td>
<td>18</td>
<td>12</td>
<td>1.31</td>
<td>1.40</td>
<td>0.03</td>
<td>0.10</td>
</tr>
<tr>
<td>Combined</td>
<td>59</td>
<td>20</td>
<td>15</td>
<td>1.35</td>
<td>1.36</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>E4 (GmphyA2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cultivated soybean</td>
<td>52</td>
<td>18</td>
<td>7</td>
<td>0.13</td>
<td>0.66</td>
<td>0.13</td>
<td>0.34</td>
</tr>
<tr>
<td>Wild soybean</td>
<td>25</td>
<td>39</td>
<td>13</td>
<td>2.11</td>
<td>2.01</td>
<td>0.56</td>
<td>0.61</td>
</tr>
<tr>
<td>Combined</td>
<td>77</td>
<td>44</td>
<td>19</td>
<td>1.12</td>
<td>1.65</td>
<td>0.39</td>
<td>0.70</td>
</tr>
</tbody>
</table>

n, number of accessions compared; S, total number of segregating sites; Hap, number of haplotypes; $\pi$, Nei’s nucleotide diversity; $\theta$, Watterson’s estimator; s, nucleotide diversity at synonymous sites and non-coding regions; a, nucleotide diversity at nonsynonymous sites.
Two common measures of nucleotide diversity, Tajima’s estimator of diversity $\pi$ [25] and Watterson’s estimator $\theta$ [26], were calculated for synonymous and non-coding regions ($s$) and for non-synonymous sites ($a$) (Table 1). All of the mutations except the SSRs were collectively considered to be SNPs and were subjected to nucleotide diversity analysis. For the synonymous sites and non-coding regions, the two homoeologs showed similar nucleotide diversities for all accessions combined: $\pi = 1.35 \times 10^{-3}$ for GmphyA1 and $1.12 \times 10^{-3}$ for E4, and $\theta = 1.36 \times 10^{-3}$ and $1.65 \times 10^{-3}$, respectively. Accordingly, mutations appear to have accumulated at almost the same rate in the two homoeologs since gene duplication occurred. On the other hand, the nucleotide diversity values for non-synonymous sites in GmphyA1 ($\pi = 0.03 \times 10^{-3}$, $\theta = 0.16 \times 10^{-3}$) were only 8% and 23%, respectively, of the corresponding values in E4 ($\pi = 0.39 \times 10^{-3}$, $\theta = 0.70 \times 10^{-3}$). Comparison of nucleotide diversity in the cultivated and wild soybeans further produced different results between the two genes: GmphyA1 had similar diversity in the cultivated and wild soybeans in all diversity parameters, whereas E4 (GmphyA2) had lower diversity in the cultivated soybean than in the wild soybean in all diversity parameters; cultivated soybean retained only 6% ($\pi(s)$) to 33% ($\theta(s)$) or 23% ($\pi(a)$) to 56% ($\theta(a)$) of the diversity present in the wild soybean population.

2.5. Haplotype Networks

Minimum-span haplotype networks were constructed using all of the observed polymorphisms to determine the origins of the dysfunctional alleles and to elucidate the structure of the variations observed in the cultivated and wild soybeans (Figure 4; Supplemental Figures S1 and S2). The haplotype network for the E4 region consisted of 17 haplotypes, including five non-functional alleles, four alleles detected in this study and the allele containing SORE-1 [17], with 2 putative unmapped recombinants. All of the loss-of-function alleles appear to have derived from haplotype 14. Wild soybeans possessed 12 haplotypes that were not found in cultivated soybeans. Only haplotype 14 was common to both wild and cultivated soybeans. All of the cultivated accessions that we tested except for a Chinese one (in haplotype 5) possessed haplotype 14 or non-functional alleles that were derived from the former haplotype. On the other hand, the haplotype network for the GmphyA1 region consisted of 13 haplotypes with 2 putative unmapped recombinants. Of these, 9 were specific to wild soybeans, 3 (haplotypes 3, 4, and 12) were common to both wild and cultivated soybeans, and 3 (haplotypes 1, 13, and 15) were specific to cultivated soybeans. The haplotypes for the GmphyA1 region that were observed in the cultivated soybeans were divided into two clusters (haplotypes 1 to 11 and haplotypes 12 and 13) that differed by at least 7 SNPs. The existence of such distantly related haplotypes in cultivated soybeans resulted in a higher nucleotide diversity in GmphyA1 than in wild soybeans (Table 1). In contrast, the reduction of nucleotide diversity in E4 in cultivated soybeans (Table 1) may be attributable to the predominant distribution of haplotype 14.
Figure 4. Haplotype networks for two phytochrome A homoeologs, GmphyA1 and E4 (GmphyA2). Light blue circles, haplotypes specific to wild soybean (WS); orange, those common to both wild and cultivated soybeans; green, those specific to cultivated soybean (CS). Black arrows, insertions; white arrows, deletions. Black dots represent differences (e.g., SNPs) separating the haplotypes. The number of accessions that belong to each haplotype is given in parentheses.

3. Discussion

Sequencing and DNA marker analyses of E4 revealed that all of the photoperiod-insensitive accessions analyzed, except for those in Group II (Figure 1), have a dysfunctional allele at the E4 locus. We detected four new, independent dysfunctional alleles (e4-oto, e4-tsu, e4-kam, and e4-kes), all of which exhibited single-base deletions in the first or second exon. These deletions generated premature stop codons as a result of a frame shift, resulting in truncated proteins of different lengths. It is therefore likely that the loss of function in the E4 gene played an important role in the evolution of insensitivity to photoperiod in early-flowering, photoperiod-insensitive accessions that are adapted to high latitudes. Like the natural variations observed in FLC and FRI in Arabidopsis [3–7] and in Hd1 in rice [8], natural mutations have repeatedly generated loss-of-function alleles at the E4 locus in soybean. The survey of
191 cultivated soybean accessions collected from various regions of East Asia further revealed that the 4 dysfunctional alleles were limited to relatively small geographical regions, as was the case in the e4 allele containing the SORE-1 insert, which was detected only in northern Japan out of 332 cultivated and 85 wild soybean accessions that were surveyed [22]. The loss-of-function alleles at the E4 locus may therefore have originated relatively recently and independently in different soybean landraces that possess the functional E4 allele of haplotype 14. Mutations leading to early flowering and the resultant early maturity would have permitted the use of diverse cropping systems and would consequently have extended the soybean production season. Under human selection, the loss-of-function alleles at the E4 locus may have accumulated multiple times in local populations. The E4 locus is therefore a key player in the adaptation of soybean to LD conditions of high latitudes and diverse cropping systems.

Soybean is a paleopolyploid species with a complex genome, which is estimated to have become duplicated both 59 and 13 million years ago [23, reviewed in [27]]. Approximately 75% of the genes are present as multiple copies, some of which have diverged in their functions, as suggested by different expression patterns between homoeologs [28,29]. phyA is one such example, and consists of 2 sets of homoeologous partners, GmphyA1/GmphyA2(E4) and GmphyA3(E3)/GmphyA4 [16–18]. The presence of multiple copies of soybean phyA contrasts sharply with other legume species such as pea (Pisum sativum), Medicago truncatula, and Lotus japonicus, which all possess a single phyA gene [30]. Of the four copies, GmphyA2 and GmphyA3 correspond to the soybean maturity genes E4 and E3, respectively; however, neither a major gene nor a QTL controlling flowering time has so far been reported near the genomic positions of GmphyA1 and GmphyA4. In particular, GmphyA4 is most likely dysfunctional in Williams 82, a cultivar that has been used for whole-genome sequencing, because of a deletion in the third exon [16,18].

E3 and E4 were originally identified by different responses of flowering to LD conditions induced by light with a high red (R) to far-red (FR) quantum ratio generated by R-enriched fluorescent lamps and by light with a low R:FR ratio generated FR-enriched incandescent lamps [9–12]. E3 controls flowering under LD conditions with a high R:FR ratio; e3e3 recessive homozygous plants can initiate flowering under these conditions [9]. E4 is involved in flowering under LD conditions with a low R:FR ratio; a recessive e4 allele is necessary for plants homozygous for the e3 allele to flower under these conditions [10–13]. Both genes thus control flowering under LD conditions with a wide range of R:FR ratios, but in a non-additive manner. phyA is an FR sensor that is involved, directly or via interactions with other photoreceptors, in various developmental processes [31]. It also acts as a red-light photoreceptor, particularly under R light with a high photon irradiance; in Arabidopsis, quadruple-null mutants for the phytochrome family (phyBphyCphyDphyE) that only contain functional phyA were able to respond to the R-mediated de-etiolation of seedlings and survive until flowering under continuous R light with a high photon irradiance [32,33]. The different responses of E3 and E4 to LD conditions with different light qualities may therefore indicate that the two genes participate in different aspects of phyA functions.

On the other hand, the function of GmphyA1, a homoeolog of E4, remains undetermined, because no genetic variants producing any phenotypic differences have been available at this locus. However, two findings suggest that like E4, GmphyA1 is also involved in both the de-etiolation response and flowering under FR-enriched LD conditions [13,17]. First, the e4 allele partially impaired the de-etiolation response to continuous FR light [17]. This is in sharp contrast to the phyA null mutants of Arabidopsis,
pea, and rice, which show a complete loss of the de-etiolation response under continuous FR light [34–37]. E3 is not involved in the de-etiolation responses under either continuous R or FR light, suggesting that the redundancy in the de-etiolation response of the e4 allele may be attributable to GmphyA1 [17]. Second, when combined with a dominant allele at the E1 locus, a double-recessive genotype for the E3 and E4 loci retains the photoperiod sensitivity, particularly to LD conditions with a low R:FR ratio (<1.0), although it is insensitive to LD conditions with a relatively high R:FR ratio (1.0–5.0) [13]. These findings suggest that the homoeolog of E4, GmphyA1, itself functions redundantly with E4 in both de-etiolation responses and photoperiod responses under FR-enriched light.

The results obtained from our sequencing analyses of a diverse collection that included both wild and cultivated accessions introduced mainly from various regions of East Asia reveal that E4 and GmphyA1 exhibit almost the same nucleotide diversities at synonymous sites and in non-coding regions among all accessions combined, suggesting that the two phyA genes have accumulated mutations at almost the same rate since gene duplication. However, the nucleotide diversity at non-synonymous sites, as a whole, was lower in GmphyA1 than in E4. In particular, the dysfunctional mutations were concentrated in only E4, despite their predicted redundant functions in both photoperiod sensitivity and the de-etiolation response [13,17]. The low diversity in non-synonymous sites at GmphyA1 may therefore indicate that there are some differences in phyA functions between the two homoeologs, and that GmphyA1 might have been more amenable than E4 to purifying selection. Further understanding of the function of GmphyA1 will be needed before we can explain why the mutations are concentrated in only one of the two homoeologs.

Nucleotide diversity in a homoeologous gene pair has also been evaluated in the soybean orthologs of Arabidopsis TERMINAL FLOWER 1 (TFL1), a gene involved in the phase transition in the shoot apical meristem (SAM) [38]. The soybean TFL1 ortholog consists of two homoeologs, GmTFL1a and GmTFL1b, the latter of which is the determinate growth habit gene Dt1 [38,39]. The two homoeologs are expressed differently: GmTFL1b is expressed mainly in the vegetative SAM and the roots, whereas GmTFL1a is expressed mainly in the stem tip after flowering and in the immature cotyledons [39]. Arabidopsis TFL1 is highly expressed in the shoot apex and roots and weakly in the seeds and siliques (Arabidopsis eFP Browser [40,41]). Therefore, the different expression profiles of GmTFL1a and GmTFL1b may reflect the subfunctionalization of the Arabidopsis TFL1 gene. Tian et al. [38] found that at least four allelic variants at the Dt1 locus in the cultivated soybean population caused stem termination in the SAM as a result of single amino acid substitutions, whereas no non-synonymous mutation in GmTFL1a was detected in either wild or cultivated soybeans. Subfunctionalization following duplication of the multifunctional ancestral gene may have enabled one of the homoeologs to accumulate functional mutations under human selection without any constraints imposed by the other functions of the ancestral gene. The asymmetrical accumulation of dysfunctional mutations observed in the maturity gene E4 and its homoeolog may therefore reflect their subfunctionalization as well.
4. Experimental Section

4.1. Materials

The 27 photoperiod-insensitive accessions used in this study included 19 accessions from northern Japan (9 from Hokkaido and 10 from the Tohoku region), 3 from the Korean Peninsula, and 5 from northeastern China and far-eastern Russia. cv. Harosoy was also included in the analysis because it carries a dominant $E4$ allele [12]. The $E4$ sequences were analyzed for 36 cultivated and 15 wild soybean accessions (Supplemental Tables S2 and S3), including nine photoperiod-insensitive accessions. We also analyzed the sequences of $GmphyA1$ (Glyma10g28170), a homoeolog of $E4$, from 26 cultivated and 13 wild soybean accessions to compare the molecular diversity of these two homoeologs (Supplemental Tables S2 and S3). The DNA marker analysis was carried out for a total of 191 accessions, including the 27 photoperiod-insensitive accessions and 164 cultivated soybean accessions (Supplemental Table S1), to determine the geographical distribution of the loss-of-function alleles of $E4$.

4.2. Methods

4.2.1. Classification of Photoperiod-Insensitive Accessions by Isozymes and SSRs

We genotyped 11 isozyme and 9 simple sequence repeat (SSR) markers to classify the 27 accessions, as described in [42] for isozymes and in [43] for SSRs. The isozyme loci that we tested were $Aco1$, $Aco2$, $Aco4$, $Aph$, $Enp$, $Est1$, $Dia1$, $Idh1$, $Idh2$, $Lap$, $Mpi$, and $Pgm1$. The SSR markers were Satt002, Satt0038, Satt063, Satt156, Satt180, Satt197, Satt228, Satt262, and Satt600, and were selected from 20 SSRs tested in 131 Asian soybean accessions and shown to have high levels of genetic diversity [43]. Total genomic DNA was extracted from young trifoliate leaves, as described in [44]. SSR analysis used 6% denatured polyacrylamide gel electrophoresis with fluorescent-labeled primers, and was performed using an ABI 377 sequencer (Perkin Elmer/Applied Biosystems, Foster City, CA, USA). We used the GeneScan software (v. 3.1) to score the observed polymorphisms. The genetic distance between each pair of the 27 accessions was calculated as $1 - P$, where $P$ is the proportion of shared alleles for the loci tested. The genetic distance matrix was subjected to cluster analysis using the unweighted pair-group method with an arithmetic average (UPGMA), and was performed using the PHYLIP software [45].

4.2.2. Sequence Analyses

The region sequenced covered the full coding sequences, 4 introns, the 5'-untranslated region (UTR) and the 3'-UTR, and their flanking regions of $E4$; and the full coding sequences, 3 introns, 3'-UTR, and the 3'-flanking region of $GmphyA1$. Two overlapping fragments were amplified from the total genomic DNA for each homoeolog using $ExTaq$ DNA polymerase (Takara, Ohtsu, Shiga, Japan) with the following homoeolog-specific primers: 5'-CACGTAGATTCTCTCATTCAACGACACT-3' and 5'-CAATGCTTACCTTGATGCCCCACATAAGATT-3' for $GmphyA2$; 5'-AGCGAGATGGCTGACACTTGGGAGTGC-3' and 5'-CTCGGTGCTGACCAGATAC-3' and 5'-CTCGGTGCTGACCAGATAC-3' for $GmphyA1$. They were then purified with the ExoSAP-IT enzyme kit (GE Life Sciences Japan, Tokyo Japan). The purified PCR products were used as templates for forward and reverse sequencing reactions,
and sequenced with an ABI PRISM 3100 Avant Genetic Analyzer using a BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems Japan, Tokyo, Japan). The sequences for novel dysfunctional alleles were further confirmed by cloning with the pGEN-T easy vector system (Promega K. K. Japan, Tokyo, Japan), followed by sequencing as described above.

4.2.3. Analysis of the Distribution of Loss-of-Function Alleles Using DNA Markers

Allele-specific DNA markers were developed from sequences flanking the mutation sites. We used cleaved amplified polymorphic sequence (CAPS) markers and derived CAPS (dCAPS) markers. The targeted region for each mutation was amplified from the DNA preparations using ExTaq polymerase with primers specific to each mutation. The primers used were 5’-CCCAGACACTCTCTTGTGTGAT-3’ and 5’-CCATACTCTCGGTATCTTTG-3’ for e4-oto; 5’-CACCCTAGGAGTTGTGGTGTGTT-3’ and 5’-GCGGTTTCTGACAATTGCCTGATA-3’ for e4-tsu; 5’-CTTAATAAAGCCATGACTGTGTTTGTGTTGATGATGGTATCTTTG-3’ and 5’-CTTGAGTTTCAATGAGGTTTCAAC-3’ for e4-kam and e4-kes. A marker analysis for the e4 allele containing inserted SORE-I to detect amplification products of different lengths was carried out as described in [17], using a common forward primer, 5’-AGACGTAGTGCTAGGGCTAT-3’ and two allele-specific primers, 5’-GCATCTCGCATCACCAGATCA-3’ for E4 and 5’-GCTCATCCCTTCGAATTCAAC-3’ for e4. The PCR products were digested with appropriate restriction enzymes for all of the alleles except for the SORE-I-inserted e4 (Sacl for e4-oto, EcoRV for e4-tsu, AflII for e4-kam, and BspHI for e4-kes). The PCR products or digestion products were separated by electrophoresis in 0.8% or 3% agarose gel, and visualized under UV light.

4.2.4. Statistical Analyses

Sequence alignment for GmphyA1 and GmphyA2 was done using the CLUSTALW algorithm [46]. Sequence variability was estimated using the DnaSP software (v. 5.0) [47]. Using this software, we calculated the number of segregating sites (S), the number of haplotypes (Hap), Tajima’s estimator of diversity (π) [25], and Watterson’s estimator (θ) [26] for synonymous sites and non-coding regions (s) and for non-synonymous sites (a). Haplotype networks were constructed from informative DNA polymorphisms, and then adjusted for haplotype-specific polymorphisms.

5. Conclusions

We detected four novel dysfunctional alleles at the E4 locus in early-maturing, photoperiod-insensitive soybean accessions from various geographical origins in East Asia. These alleles have accumulated independently and repeatedly in local populations of northern Japan and northeastern China. The E4 locus may therefore be a key player in the adaptation of soybean to LD conditions of high latitudes and diverse cropping systems. The allele-specific markers developed in this study will be useful tools to assess the genotypes and facilitate marker-assisted selection in breeding of cultivars adapted to higher latitudes. Comparison of the two phyA sequences may provide insights into why the mutations have accumulated only in E4, and not in its homoeolog GmphyA1. The lower nucleotide diversity at non-synonymous sites in GmphyA1 relative to E4 suggests an unknown functional divergence between the two homoeologs despite their redundant functions in photoperiodic
flowering and photomorphogenesis, as have been indicated by previous genetic analyses [13,17]. Further molecular dissection of the functions of the phyA gene copies may facilitate our understanding of the evolutionary consequences of duplicated genes and adaptation to higher latitudes in soybean.

Acknowledgments

The authors are grateful to Alexander Y. Ala (All Russian Research Institute of Soybean, Russia) and Helmut Knupffer (Leibniz Institute for Plant Genetics and Crop Plant Research, Germany) for providing us seeds of Russian and North Korean cultivars, respectively. This work was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (23380001) for Jun Abe, a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Genomics for Agricultural Innovation, DD-2040) for Kyuya Harada, the Natural Science Foundation of China (30971813), the Program of “One Hundred Talented People” of Chinese Academy of Sciences (KZCX2-YW-BR-11), Grant No. 2009ZX08009-013B, and Grant No.JC200919 for Baohui Liu, and the Natural Science Foundation of China (31071445), the Natural Science Foundation of Heilongjiang Province (ZD201001), the Program of "One Hundred Talented People" of Chinese Academy of Sciences for Fanjiang Kong.

References


Accession Numbers

Sequence data from this article can be found in the GenBank/EMBL/DDBJ data libraries under the following accession number: Kamaishi-17 GmphyA2 allele (e4-kam, AB643573), Keshuang GmphyA2 allele (e4-kes, AB643574), Otowase GmphyA2 allele (e4-oto, AB643575), Tsukue-4 GmphyA2 allele (e4-tsu, AB643576), Karafuto-1 SORE-I-inserted GmphyA2 allele (e4, AB643577), 12 GmphyA1 haplotypes (AB643550 to AB643561), and 11 GmphyA2 haplotypes (AB643562 to AB643572).