DNA Research, 2019, 26(3), 243–260
doi: 10.1093/dnares/dsz005
Advance Access Publication Date: 16 April 2019
Full Paper

Full Paper

Highly multiplexed AmpliSeq technology identifies novel variation of flowering time-related genes in soybean (*Glycine max*)

Eri Ogiso-Tanaka, Takehiko Shimizu, Makita Hajika, Akito Kaga*, and Masao Ishimoto

Institute of Crop Science (NICS), NARO (National Agriculture and Food Research Organization), 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan

*To whom correspondence should be addressed. Tel./Fax. +81 29 838 7452. Email: kaga@affrc.go.jp

Edited by Prof. Kazuhiro Sato

Received 17 December 2018; Editorial decision 1 March 2019; Accepted 11 March 2019

Abstract

Whole-genome re-sequencing is a powerful approach to detect gene variants, but it is expensive to analyse only the target genes. To circumvent this problem, we attempted to detect novel variants of flowering time-related genes and their homologues in soybean mini-core collection by target re-sequencing using AmpliSeq technology. The average depth of 382 amplicons targeting 29 genes was 1,237 with 99.85% of the sequence data mapped to the reference genome. Totally, 461 variants were detected, of which 150 sites were novel and not registered in dbSNP. Known and novel variants were detected in the classical maturity loci—*E1*, *E2*, *E3*, and *E4*. Additionally, large indel alleles, *E1-nl* and *E3-tr*, were successfully identified. Novel loss-of-function and missense variants were found in *FT2a*, *MADS-box*, *WDR61*, *phytochromes*, and *two-component response regulators*. The multiple regression analysis showed that four genes—*E2*, *E3*, *Dt1*, and *two-component response regulator*—can explain 51.1–52.3% of the variation in flowering time of the mini-core collection. Among them, the two-component response regulator with a premature stop codon is a novel gene that has not been reported as a soybean flowering time-related gene. These data suggest that the AmpliSeq technology is a powerful tool to identify novel alleles.

Key words: AmpliSeq, target re-sequencing, genotyping, flowering time-related gene, soybean

1. Introduction

Flowering time is critical for successful seed production by plants. Flowering time and maturity are the most important traits to determine the adaptability of soybean [*Glycine max* (L.) Merr.] cultivation. These not only restrict the cultivation area but also greatly affect plant architecture and yield.[1,2] Therefore, it is necessary to clarify the genetic factors affecting flowering time and maturity and control them using a combination of alleles with different genetic effects on flowering time. To combine such alleles freely based on DNA marker-assisted selection, a catalogue of alleles for breeding materials will be necessary. Soybean is a typical short-day plant. Several functional nucleotide polymorphisms responsible for diversity in flowering time among cultivars are already known.[3] Classical maturity loci designated as *E* loci have been characterized, including *E1* and *E2*,[4] *E3*,[5] *E4*,[6] *E7*,[7] *E8*,[1] *E9*,[8] and *E10*.[9] Of these, *E1*,[10] *E2*,[11] *E3*,[12] *E4*,[13] and *E9*[8] have been isolated as flowering time-related genes. *E1* encodes putative transcriptional factor containing plant-specific B3 domain.[10] *E2* encodes a homologue of

GIGANTEA.[11] *E3* and *E4* encode a homologue of the photoreceptor phytochrome A (PHYA).[12] *E9* encodes the florigen protein FT2a.[8]

In the *E1* gene, three alleles, namely *e1-as* (=*e1* designated by Bernard[4]), *e1-fs*, and *e1-nl*, have been reported as early flowering phenotype under long-day conditions.[10] The *e1-as* allele has a single missense mutation (Arg15Thr) in the coding region. The *e1-as* genotype promotes flowering for ∼10 days compared with that by the *E1* genotype under natural day-length conditions at Matsudo, Japan (35°78′N, 139°90′E). The *e1-fs* allele has a 1-bp deletion, resulting in a premature stop codon in the cultivar Sakamotowase.[10] *e1-nl* is a null allele in which ∼142 kb, including the entire *E1* gene, is deleted in some early flowering cultivars.[10] In contrast, only one *e2* allele has been reported in the *E2* gene. The *e2* allele has one premature stop codon mutation due to single nucleotide polymorphism (SNP) in the 10th exon.[11] The *e2* genotype promotes flowering for ∼9 days under natural day-length conditions at Tsukuba, Japan (36°03′N, 140°04′E).[11] In the *E3* gene, *e3-Mo*, *e3-fs*, *e3-tr* (=*e3* designated by Buzzell[5]), and *e3-ns* have been reported as nonfunctional alleles.[12,14] The *e3-Mo* alleles have SNP for a non-synonymous amino acid substitution (G1050R) in the third exon.[3,12] The *e3-fs* allele has a single base insertion in the exon, resulting in frameshift mutation.[14] The *e3-ns* allele has a nonsense mutation in which a single nucleotide substitution in exon 3 creates a stop codon.[14] The *e3-tr* alleles lack a 13.33-kb genomic region including a part of exons 3 and 4.[3,14] These nonfunctional alleles promote flowering under long-day conditions. In addition, *E3-Mi* and *E3-Ha* also have been reported as functional alleles.[12,14] The *E3-Mi* alleles have a 2.633-kb deletion in the third intron. As for the *E4* gene, five nonfunctional alleles, viz., *e4-SORE-1*, *e4-kam*, *e4-kes*, *e4-oto*, and *e4-tsu*, have been reported.[12,14,15] The *e4-SORE-1* (=*e4* designated by Buzzell and Voldeng[6]) alleles have a *Ty1/copia*-like retrotransposon (*SORE-1*) insertion in first exon, resulting in a nonfunctional allele.

These *E1–E4* genes can result in variation in flowering time by controlling the expression of *FLOWERING LOCUS T* (*FT*) genes, *FT2a* and *FT5a*.[10,11,14,16] The florigen protein FT2a is encoded by *E9*. The *e9* allele has a *SORE-1* insertion in the first intron. Although eight SNPs and six InDels in the *E9* have been reported, the influence on gene function is unknown.[17] The *FT5a* gene was identified as *qDTF-J*, which promotes the flowering time for ∼5 days under natural day-length conditions at Hokkaido, Japan (43°07′N, 141°35′E).[18] In the *FT5a* gene, 13 SNPs and 3 InDels are reported only in the promoter and untranslated regions (UTRs) in 439 cultivated and wild soybean accessions.

The functional nucleotide polymorphisms of the four *E* genes (*E1–E4*) are useful to predict the flowering time and could explain ∼62–66% of the phenotypic variation in flowering time among 63 Japanese accessions under long-day conditions.[3] However, prediction of flowering time will be difficult if the breeding materials have unknown alleles affecting flowering phenotype. Therefore, development of a sequencing system that can easily capture as many alleles as possible is required. Recently, it became possible to obtain whole-genome information easily with the development of next-generation sequencing (NGS) technologies. However, it is still expensive for resequencing large genomes. Moreover, it is necessary to have analytical and storage environments to deal with enormous amounts of whole-genome sequence data of genetic resources. Target resequence is one of the alternative sequencing methods to obtain sequence data of a limited region, which can minimize cost and time for data analysis and decrease data storage. The AmpliSeq technology (Thermo Fisher Scientific, Waltham, MA, USA) is one of the target re-sequencing technologies, a multiplex polymerase chain reaction (PCR)-based assay targeting regions of interest. The AmpliSeq Designer[19] designs primer set that amplifies PCR products ranging from 75 to 375 bp in the target region, and multiplex-PCR products are sequenced by NGS. The method enables amplification of ∼6,000 amplicons by ultra-high multiplex PCR and constructs a targeted sequencing library in 10 h.[20] In routine genotyping of crop breeding, NGS-based techniques need to meet several criteria. The processing time between sample collection and interpretation of sequencing should be short. Furthermore, it is necessary to construct libraries using limited amount of input DNA including partially degraded DNA sample and the read depth must be deep enough to detect variant accurately. The Ion Torrent platform[19,20] in combination with the AmpliSeq multiplex PCR can use DNA input of as low as 10 ng, and the processing time between sample collection and sequence analysis can be finished within 5 days.[20] The AmpliSeq technology is frequently used for studying human inherited cancer, but it can also be applied to plant and agronomic research.

In this study, we applied AmpliSeq technology to clarify the alleles of flowering time-related genes and their homologues in diverse soybean germplasm to identify novel and known variations associated with flowering time.

2. Materials and methods

2.1. Plant materials and DNA extraction

DNA was extracted from 192 accessions of a soybean mini-core collection, provided by Genebank, NARO[21] (Supplementary Table S1). Of these, 122 accessions were sown and germinated in plastic pots on rock wool material ‘grodan’ (Nittobo, Tokyo, Japan) moistened with water. After 10 days under 12 h light/12 h dark conditions at 25°C, the first leaf was collected in a 2-ml tube. The leaf tissue of 38 samples was ground in liquid nitrogen and CTAB buffer, and then immediately used for DNA extraction manually.[22] The remaining 84 samples were dried in a freeze dryer (FDU-2100, EYELA, Tokyo, Japan). These samples were lyophilized at −80°C for 12 h and stored at 4°C. The dried leaves were crushed using a ShakeMaster (Bio Medical Science Inc. Tokyo, Japan) and the leaf powder was used to extract DNA. DNA from 44 samples was extracted using the CTAB DNA extraction kit (NR-502, KURABO, Osaka, Japan) and DNA extraction robot PE-480 (GENE PREP STAR, KURABO). DNA from another 40 samples was extracted by the bead-based method of the BioSprint 96 DNA Plant Kit on robotic workstation (QIAGEN, Hilden, Germany). From the other 70 samples, DNA was extracted from the seed tissue using the BioSprint 96 DNA Plant Kit on robotic workstation, according to the manufacturer's instruction (QIAGEN). The seed tissue samples were obtained by scraping dried seed and crushing using Zirconia beads and TissueLyser II (QIAGEN). The quality of extracted DNA was evaluated based on the DNA integrity number, which is an index showing the fragmentation degree of DNA using TapeStation (Agilent Technologies, Santa Clara, CA, USA). The DNA concentration was measured using the Qubit Fluorometer (Thermo Fisher Scientific) by exciting at 485 nm and measuring the fluorescence intensity at 520 nm. The instrument was calibrated with the Quant-iT dsDNA BR Assay kit (Thermo Fisher Scientific), according to the manufacturer's instructions.

2.2. Ion AmpliSeq custom panel design

A custom panel targeting 29 genes was designed based on Soybean reference genome version 1.1[23] using the Ion AmpliSeq Designer tool[19] version 1.2.9 using the standard DNA (125–275 bp amplicon target sizes) option. Two primer pools were designed to amplify 382 amplicons, covering 29 target genes of total length 64.98 kb (Table 1 and Supplementary Tables S2 and S3). These included the coding regions of *B3 domain containing genes* (*E1* and homologue), *Phytochrome A genes* (*E3 and E4*), *Phytochrome B genes*, *FT/TERMINAL FLOWER 1* (*TFL1*) *family* genes (including *FT2a*, *FT5a*, *and Dt1*), *two-component response regulator*-like genes, MADS box gene, WD repeat-containing gene (*WDR61*), Achaete-scute transcription factor gene, and a part of the exon of *GIGANTEA* (*E2*).

2.3. Library preparation and sequencing

The Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific) was used to quantify DNA for NGS library construction. The NGS library was constructed using the Ion AmpliSeq Library Kit 2.0 (Thermo Fisher Scientific), according to the manufacturer's protocol (Japanese version corresponding to Manual 2014.7 rev.B.0 version: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0013432_Ion_AmpliSeq_Library_Prep_on_Ion_Chef_UG.pdf, 25 March 2019, date last accessed). For the multiplex-PCR amplification, 1–10 ng of each DNA was amplified using one primer pool (191 amplicon primer pairs) per reaction. This was performed using 4 µl of 5× Ion AmpliSeq HiFi Master Mix, 10 µl of 2× AmpliSeq Custom Primer Pool, 1–10 ng of DNA, and the volume was made up to 20 µl with nuclease-free water. The reaction mix was heated for 2 min at 99°C for enzyme activation, followed by 18 two-step cycles at 99°C for 15 s and at 60°C for 4 min, and ending with a holding period at 10°C. As for low-quality DNA samples, 21 cycles were subjected under similar conditions. The primers of amplicons were digested and phosphorylated for adapter ligation using 2 µl of FuPa enzyme per sample at 55°C for 10 min, followed by enzyme inactivation at 60°C for 20 min. To enable multiple libraries to be loaded per chip, 2 µl of a unique diluted mix, including Ion Xpress Barcode and Ion P1 Adapters at standard volumes, was ligated to the end of the digested amplicons using 2 µl of DNA ligase at 22°C for 30 min, followed by ligase inactivation for 10 min at 72°C. The resulting un-amplified adapter-ligated library was purified using 45 µl of Agencourt AMPure XP Reagent (Beckman Coulter, Brea, CA, USA), followed by washing using 150 µl of freshly prepared 70% ethanol. After purification, 50 µl of Platinum PCR SuperMix High Fidelity and 2 µl of Library Amplification Primer Mix of the Ion AmpliSeq Library Kit 2.0 were added to the dried AMPure XP beads, and then the reaction plate was placed on a magnetic rack to separate the beads from the supernatant. The amplicon library in the supernatant was further amplified to enrich the material for accurate quantification at 98°C for 2 min, followed by five two-step cycles at 98°C for 15 s and at 60°C for 1 min. The amplified amplicon library was then purified using 25 µl of AMPure XP, followed by a second purification step with 60 µl of AMPure XP and 150 µl of freshly prepared 70% ethanol. The concentration and size distribution of amplicons in the library were then determined using an Agilent BioAnalyzer DNA High-Sensitivity chip or TapeStation 4200 D1000 chip (Agilent Technologies), according to the instruction of the manufacturer. After quantification, each library was diluted to a concentration of 100 pM prior to template preparation. Subsequently, the libraries were pooled in equimolar amounts prior to further processing. Emulsion PCR, emulsion breaking, and enrichment for template preparation of ion sphere particles were performed using the Ion 520 & 530 and 540 Kit-Chef (Thermo Fisher Scientific) according to the instruction of the manufacturer. After the preparation of ion sphere particles, sequencing was performed with an Ion Torrent Ion S5 or S5XL system using Ion 520 and 540 Chip (Thermo Fisher Scientific), according to the instruction of the manufacturer.

2.4. Data analysis

The Ion S5/S5XL sequence data were mapped to the soybean genome reference version 2.0 (Gmax275: https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?%20organism=Phytozome, 25 March 2019, date last accessed) using Ion Torrent Suite version 5.2.1 software. In typical genome databases of soybean (Williams 82), such as, Phytozome and Soybase, Gmax275 is widely used instead of Gmax189.[23] The assembly size and number of predicted protein-coding loci of Gmax275 are 978 Mb and 56,044, which are higher than 969.6 Mb and 46,430 of Gmax189, respectively. In the Gmax275 assembly, several genes are located on chromosomes/scaffolds different from those of Gmax189. For example, 238 genes on the chromosomes of Gmax189 are located on the scaffolds of Gmax275, whereas 100 genes on the scaffolds of Gmax189 are located on the chromosomes of Gmax275 (http://www.soybase.org/correspondence/methods.txt, 25 March 2019, date last accessed). In this study, one of the target genes, *Glyma18g22670* (B3 domain-containing protein), on chromosome 18 of Gmax189 was located on chromosome 4 (*Glyma.04G143300*) in Gmax275. However, the structure of all target genes used in this study was the same between Gmax189 and Gmax275. Ion Torrent Suite software was optimized for Ion Torrent raw data analysis; alignment using Torrent Mapping Alignment Program (TMAP) version 5.2.25 and Coverage Analysis v.5.8.0.8, and variant calling using Torrent Variant Caller version 5.2.1.38 and plug-in version 5.2.25. To evaluate PCR amplification efficiency of each amplicon, the amplicons per 100k reads mapped (APKM) as the count scaled by the total number of amplicons sequenced *N* times per 100k reads as follows:

$$\text{APKM} = \frac{X_i}{\frac{N}{10^5}} = \frac{X_i}{N}.10^5,$$

where X_i represents the read coverage *X* of target amplicon *i*.

Variant calling was performed using the default (low stringency) and custom parameters (Supplementary Table S4). All accessions used in this study were propagated by the single seed descent method; therefore, all variants should be detected as homozygous theoretically. The parameter of TMAP in Torrent Variant Caller was changed to loosen the judgment condition of homozygous by setting 'snp_min_allele_freq' from 0.15 (default) to 0.3. In this condition, the SNP was detected with allele frequency of >70% as homozygous. The InDels were detected as homozygous when allele frequency was >75% (default parameter). Sequence variants detected as heterozygous under these conditions were excluded. The vcf files obtained were annotated and filtered using the snpEff version 4.0e.[24]

2.5. Detection of known and novel SNPs and InDels

The known alleles *E1*, *E2*, *E3*, and *E4* were investigated as described above.[3] The polymorphism information from the Single Nucleotide Polymorphism Database[25] (dbSNP: https://www.ncbi.nlm.nih.gov/snp, 25 March 2019, date last accessed, data downloaded on 29 April 2016) was used to investigate whether the detected polymorphism has already been identified in the whole-genome sequence of soybean. As the position of soybean genome in the National Center

Table 1. Genomic regions or SNP targeted by the AmpliSeq design

	Gmax275 (ver. 2.0)				Gmax189 (ver. 1.1)			Description[a]	Gene name[b]	Target	Total designed target length of amplicons[c]	Target size[d]	Number of amplicon[e]	Coverage[f] (%)	Reference
Gene ID	Chr.	Start	End	Gene ID	Chr.	Start	End								
Glyma.02G069500	2	6116379	6117379	Glyma02g07650	2	6041770	6042409	FLOWERING LOCUS T	*GmFTL7*	Exon + UTR	861	639	6	100.0	37, 38
Glyma.03G194700	3	40522597	40525110	Glyma03g35250	3	42534078	42535865	TERMINAL FLOWER 1	*GmTFL2*	Exon + UTR	1,577	1,003	9	100.0	37, 38
Glyma.03G227300	3	42918771	42923401	Glyma03g38620	3	44925730	44930360	Phytochrome A	*GmPHYA4*	Exon + UTR	3,688	2,876	20	98.6	
Glyma.04G143300[g]	4	26120011	26120532	Glyma18g22670	18	25739929	25740831	B3 domain-containing protein[h]	*E1Lb*	Exon + UTR	1,037	902	5	100.0	11, 54
Glyma.04G156400	4	36758125	36758770	Glyma04g24640	4	28293933	28294806	B3 domain-containing protein[h]	*E1La*	Exon + UTR	967	873	5	100.0	11, 54
Glyma.06G207800	6	20207077	20207940	Glyma06g23026	6	20006928	20007814	B3 domain-containing protein[h]	*E1*	Exon + UTR	1,059	886	6	100.0	3, 4, 11
Glyma.08G363100	8	47458142	47459829	Glyma08g47810	8	46606934	46608654	FLOWERING LOCUS T	*GmFT4*	Exon + UTR	1,300	746	8	100.0	20
Glyma.08G363200	8	47472881	47473362	Glyma08g47823	8	46621704	46622185	FLOWERING LOCUS T	*GmFT6*	Exon + UTR	612	265	3	100.0	37, 38
Glyma.09G035500	9	2960395	2967229	Glyma09g03990	9	2919887	2926740	Phytochrome B	*GmPHYB1*	Exon + UTR	5,150	4,336	27	99.7	32
Glyma.09G143500	9	35652219	35653967	Glyma09g26550	9	33049107	33050904	BROTHER OF FT AND TFL 1	*GmTFL4*	Gene	1,813	1,122	11	98.7	37
Glyma.10G141400	10	37489560	37495624	Glyma10g28170	10	36962521	36968813	Phytochrome A		Exon + UTR	5,212	4,065	29	94.6	14
Glyma.10G221500	10	45294735	45316121	Glyma10g36600	10	44732730		GIGANTEA	*E2*	SNP	222	1	1	100.0	3, 4, 12
Glyma.12G073900	12	5508365	5522772	Glyma12g07861	12	5496565	5511828	Two-component response regulator-like		Exon + UTR	4,116	2,871	24	100.0	
Glyma.15G140000	15	11435551	11442683	Glyma15g14980	15	11415495	11422656	Phytochrome B		Exon + UTR	5,445	4,542	28	98.1	32
Glyma.16G044100	16	4135885	4137742	Glyma16g04830	16	4115033	4116923	FLOWERING LOCUS T	*GmFT5a/GmFTL4*	Exon + UTR	1,683	1,109	9	98.0	15, 19, 37
Glyma.16G044200	16	4162525	4164824	Glyma16g04840	16	4141774	4144073	FLOWERING LOCUS T	*GmFT3a/GmFTL1*	Exon + UTR	1,031	486	6	92.7	15, 37
Glyma.16G150700	16	31109999	31114963	Glyma16g26660	16	30741660	30746677	FLOWERING LOCUS T	*GmFT2a/GmFTL3*	Exon + UTR	1,515	935	9	99.4	9, 15, 18, 33, 37
Glyma.16G151000	16	31148829	31151842	Glyma16g26690	16	30780496	30783509	FLOWERING LOCUS T	*GmFT2b/GmFTL5*	Exon + UTR	860	464	5	88.0	15, 37
Glyma.16G196300	16	35777815	35779317	Glyma16g32080	16	35274147	35275762	BROTHER OF FT AND TFL 1	*GmTFL3*	Exon + UTR	1,554	1,038	6	99.0	37
Glyma.16G200700	16	36179891	36187469	Glyma16g32540	16	35676581	35684221	MADS box protein		Exon + UTR	2,534	1,095	15	98.8	48, 49
Glyma.17G052100	17	3955518	3958432	Glyma17g05990	17	4225839	4228888	WD repeat-containing protein 61	*WDR61*	Exon + UTR	1,743	1,318	9	100.0	50
Glyma.17G090500	17	7052506	7053858	Glyma17g09810	17	7317271	7318756	Achaete-scute transcription factor-related[i]		Exon + UTR	1,594	1,166	9	100.0	
Glyma.19G108100	19	36030632	36032867	Glyma19g28390	19	35849981	35852216	FLOWERING LOCUS T	*GmFT3b/GmFTL2*	Exon + UTR	952	463	5	100.0	15, 37
Glyma.19G108200	19	36049111	36051851	Glyma19g28400	19	35868460	35871203	FLOWERING LOCUS T	*GmFT5b/GmFTL6*	Exon + UTR	1,113	560	6	90.0	15, 37

Continued

Table 1 continued

Gene ID	Gmax275 (ver. 2.0)			Gene ID	Gmax189 (ver. 1.1)			Description[a]	Gene name[b]	Target	Total designed target length of amplicons[c]	Target size[d]	Number of amplicon[e]	Coverage[f] (%)	Reference
	Chr.	Start	End		Chr.	Start	End								
Glyma.19G194300	19	45183357	45185175	Glyma19g37890	19	44979743	44981657	TERMINAL FLOWER 1	*Dt1/GmTFL1*	Exon + UTR	1,411	1,078	8	100.0	37, 53
Glyma.19G224200	19	47633059	47641958	Glyma19g41210	19	47511095	47520052	Phytochrome A	*E3*	Exon + UTR	5,235	4,400	27	98.6	3
Glyma.19G260400	19	50364718	50369677	Glyma19g44970	19	50244046	50249070	Pseudo-response regulator 5		Gene	5,673	4,941	31	98.3	
Glyma.20G090000	20	33236018	33241692	Glyma20g22160	20	32087412	32093306	Phytochrome A	*E4*	Exon + UTR	4,871	4,076	25	99.2	3, 14
Glyma.U034500[g]	Scaffold_32	197150	220019	Glyma11g15580	11	11232271	11255186	Two-component response regulator-like		Exon + UTR	5,352	3,833	30	98.1	

[a] Gene description from Phytozome 12.

[b] Gene name refers to Kong et al.,[8] Wu et al.,[29] Fan et al.,[52] and Cao et al.[53]

[c] Total size of amplified region by designed amplicon primers.

[d] Target size (bp) of amplicon based on soybean genome version 1.1 (Gmax189).

[e] Total number of designed amplicons on target gene.

[f] Percentage of target region covered by amplicon.

[g] Glyma.04G143300 and Glyma.U034500on Gmax275 genome version were different chromosome positions on Gmax189.

[h] The gene annotation manually curated.

[i] 'Achaete-scute transcription factor related' gene was included as control for variant detection.

for Biotechnology Information (NCBI) and Phytozome v12.1 are not consistent, we converted the position of SNP in the dbSNP from the NCBI to that of Gmax275 of Phytozome v12.1. The SNP ID number (rs; refSNP cluster) was used for the SNP name.

2.6. Validation of SNPs and InDels

The detected variants of the *Phytochrome A* (*E3*, *E4*) and *FT* genes (*FT2a* and *FT5a*) were further confirmed by Sanger sequencing. The exon containing the novel variants was amplified by PCR using the primers shown in Supplementary Table S5. The PCR product was purified using Affymetrix ExoSap-IT regent (ExoSap-IT, USB Corporation, Staufen, Germany) and directly sequenced for both sense and antisense strands using Big Dye Terminator version 3.1 (Applied Biosystems, Foster City, CA, USA) in an ABI 3500 Genetic Analyzer (Applied Biosystems), according to the manufacturer's protocol. The sequences were analysed using Genetics software version 10.0.8 (GENETYX Corp., Japan).

2.7. Gene-based multiple regression association testing for flowering time

Flowering time was evaluated from 2011 to 2013 at the National Institute of Crop Science (36°02′N, 140°11′E), Tsukuba, Japan. Seeds were sown on 12 July 2011, and 10 July 2012 and 2013. A starter fertilizer containing 3, 10, and 10 g m^{-2} of N, P_2O_5, and K_2O, respectively, was applied. Each accession was planted in single-row plots. Each row was paved 0.7 m apart and each plot comprised 12 plants that were spaced 0.13 m apart. The average days to flowering in each plot were used for analysis. Association between days to flowering and each polymorphic SNPs/InDels was assessed using linear regression, where the simulated trait values across the 190 individuals were regressed onto the numeric code of each SNP and InDel genotype; this tested the null hypothesis of the additive allelic effect on the trait. Regression analyses were performed using 'lm()' in R.[26] First, simple linear regression analysis was performed to assess the influence of the detected variant on flowering time at the significance level of $P < 0.05$. Subsequently, multiple linear regression analysis was performed using the significantly representative variants after removing redundant variants at the significance level of $P < 0.05$.

3. Results and discussion

3.1. Amplicon design and comparison of library quality using DNA samples derived from the leaf and seed

To evaluate the performance of AmpliSeq, we focused on gene region of 29 genes (Table 1) selected from known genes related to flowering time and their homologues in soybean. A total of 382 amplicon primer pairs consisting of two primer pools (Supplementary Table S3) were designed for the 64.98-kb target region using the AmpliSeq designer tool. These primer pairs covered 98.4% of the target region, ranging from 89.8% to 100%, by overlapping PCR products of total length 70,180 bp (Supplementary Table S2). The average amplicon size including primer region was 237.1 bp ranging from 125 to 275 bp (target region was 65–232 bp). The target gene with the lowest coverage (89.8%) was *Glyma.16g044200* (*FT*-like gene).

We examined the DNA quality necessary for AmpliSeq library construction because genotyping is commonly performed using low-quality DNA especially that derived from the seed of soybean for

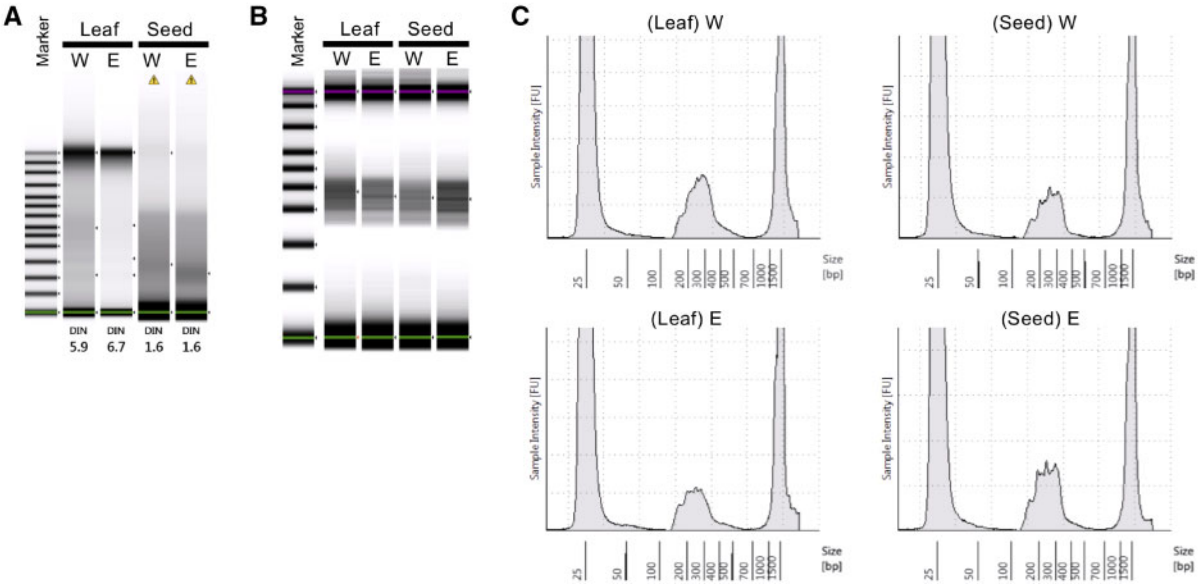

Figure 1. Evaluation of quality of DNA and AmpliSeq library prepared from the DNA using the Agilent 2200 TapeStation system. The AmpliSeq libraries were evaluated using the D1000 screen tape. (A) Quality of the DNA derived from the leaf and seeds. W and E indicate Williams 82 and Enrei, respectively. The numerical assessment of DNA quality ranged from 1 to 10 based on the DNA integrity number (DIN). A high DIN indicates highly intact DNA, whereas a low DIN indicates degraded DNA. (B) Distribution of amplicons in the AmpliSeq library shown as a gel image. (C) Electropherogram of the same AmpliSeq library as shown in (B). Lower (25 bp) and upper (1,500 bp) peak are the standard markers. The middle peak indicates the library.

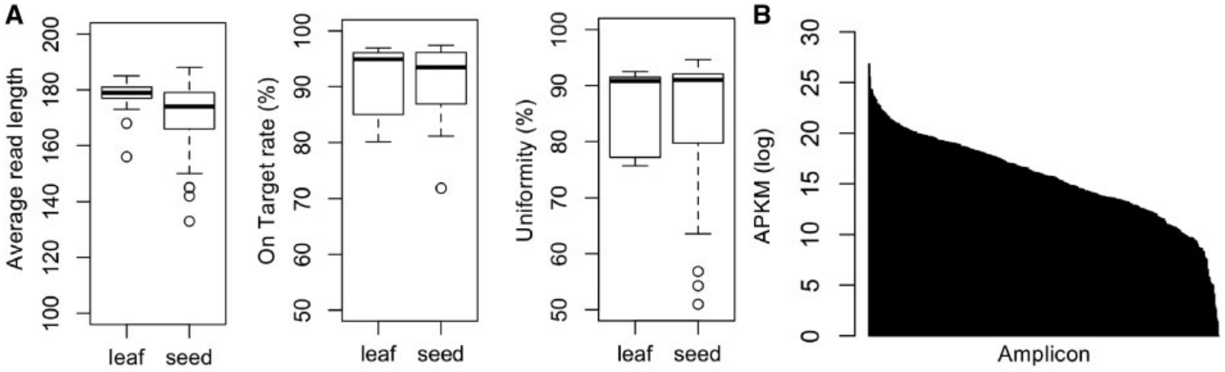

Figure 2. Comparison of sequence performance metrics classified by the plant materials used for DNA extraction. (A) The box plot of average read length, on-target rate, and uniformity. The on-target rate is on-target percent of the aligned reads. Uniformity is the percent of bases in all the amplicon-targeted regions covered by at least 0.2× the mean base read depth. (B) Average normalized reads (APKM) per sample across 382 amplicons generated from the 192 mini-core collection. The *X* and *Y* axes indicate 382 amplicons sorted in their read coverage and APKM shown as the mean on a log scale, respectively.

marker-assisted selection. Low-quality DNA derived from the seed was obtained at concentrations of 0.5–3 ng/µl, whereas high-quality DNA from the leaf was obtained at concentrations of 30–50 ng/µl (Fig. 1A). We used 1–10 ng of seed-derived DNA and 10 ng of leaf-derived DNA for preparing AmpliSeq library. To confirm whether low-quality DNA can produce a library of sufficient quality, distribution range of amplicons in the libraries prepared using low-quality DNA was compared with that of high-quality DNA, which is recommended for sequencing using the Agilent 2100 Bioanalyzer or TapeStation 4200 (Fig. 1A and B). No difference was observed between low-quality DNA from the seed and high-quality DNA from the leaf in the size range of amplicons (130–370 bp) or maximum peak amplitude (Fig. 1B and C). These results reveal that the AmpliSeq library of sufficient yield and quality can be prepared from low-quality DNA. We then prepared sufficient amount of library using DNA derived from the leaf or seed of the soybean mini-core collection.

3.2. Performance of NGS and uniformity of amplicon coverage

Among 105,761,267 reads obtained, 105,603,249 (99.85%) reads were mapped to Williams 82 reference genome Gmax275 using TMAP and the average read depth across the target region was 1,237× (Supplementary Table S6). According to the on-target rate, 94.12% of the reads was mapped to the targeted regions. The average read length, average on-target rate, and uniformity (percent of reads >0.2× of mean coverage in the sample) of the leaf and seed samples were similar, but a few seed samples showed lower average read length and uniformity (Fig. 2A). The highly fragmented DNA sample showed low amplification of long amplicons (> 200 bp) and

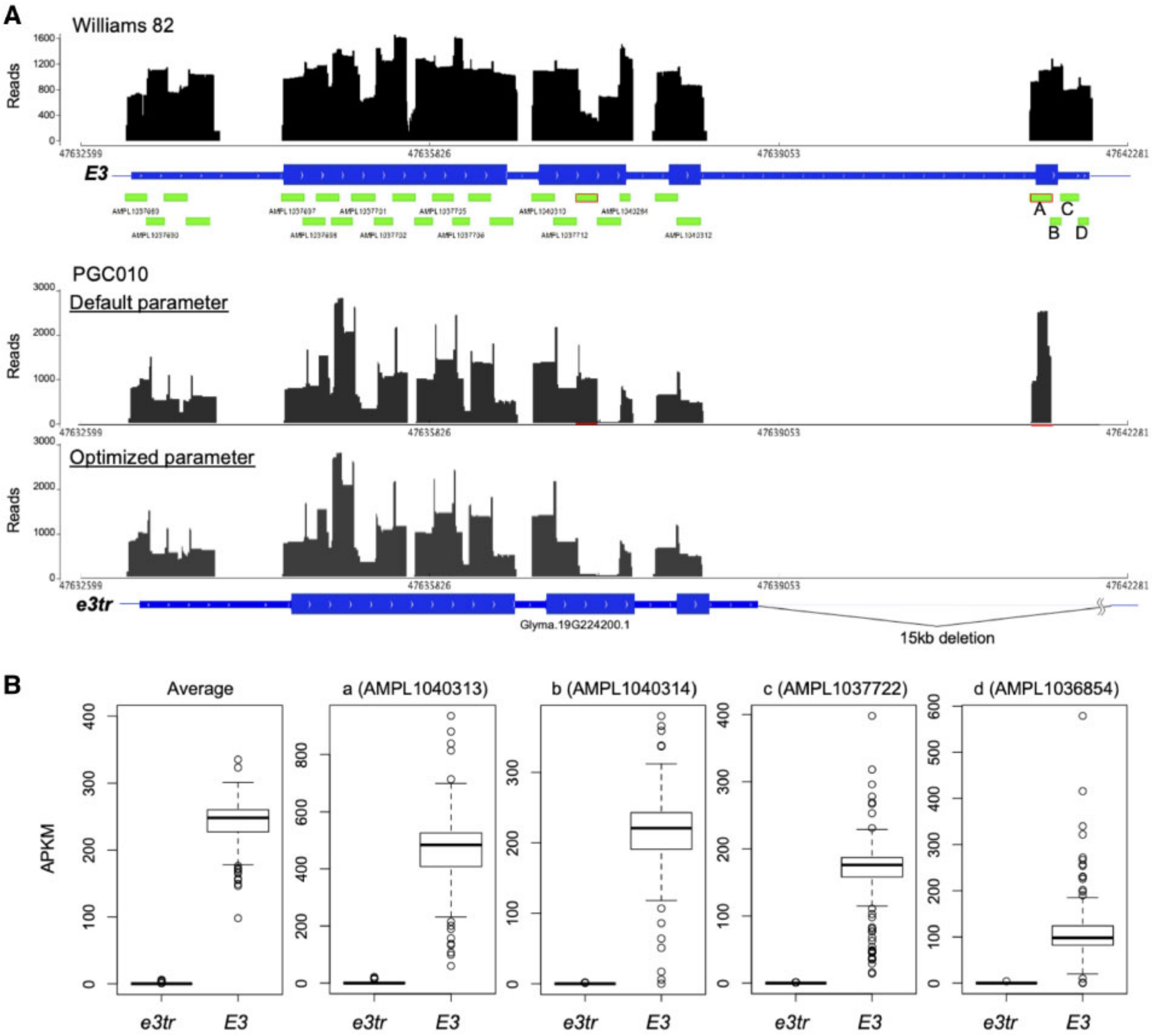

Figure 3. Distribution of mapped sequence reads of two different alleles in the *E3* gene region. (A) Top: Read coverage of Williams 82 with the default parameters of TMAP. Bottom: Read coverage of PGC010 with the default and optimized parameters of TMAP. The *E3* gene of PGC010 has a large deletion in the fourth exon. (B) APKM of amplicons from variant (*e3T*) and wild type (WT) on the fourth exon of *E3*. Average APKM of four amplicons (left most) and APKM of each amplicon—(a) AMPL1040313, (b) AMPL1040314, (c) AMPL1037722, and (d) AMPL1036854.

low uniformity (Supplementary Fig. S1A and B). Low average amplicon length or low uniformity of the samples might be caused by DNA fragmentation or contamination of the DNA solution.[27]

To compare the efficiency of PCR amplification of each amplicon, the APKM was calculated for each amplicon. The magnitude of APKM was similar irrespective of the type of DNA sample between the leaf and seed (Supplementary Fig. S2A). In contrast, there was no relationship between the APKM and read length of amplicons (Supplementary Fig. S2B). The APKM ranged from 0 to 3,333. Only one amplicon (AMPL1040290) had zero read. As the primer pair of AMPL1040290 was designed for the region flanking the TA-repeat microsatellite, it is difficult to amplify by multiplex PCR. We could amplify amplicons of 400–450 bp using the single primer pair of AMPL1040290.

We evaluated known alleles as an example to verify whether the reads were correctly mapped. Among two alleles with a large deletion, *e3-Mo* and *e3-tr*, at the *E3* gene on Chr19,[12] the read mapping status of the *e3tr* allele, with a 15-kb deletion including the fourth exon, was examined by designing four amplicons, viz., AMPL1040313, AMPL1040314, AMPL1037722, and AMPL1036854 (Fig. 3). The mapped reads from Williams 82 covered the entire fourth exon by the four amplicons (Fig. 3A). In contrast, the reads from PGC010 were mapped to a part of the fourth exon, which could not be mapped because of absence of the fourth exon in PGC010. When we confirmed the sequence of the mapped reads, these are found to contain several polymorphic sites originated from another region (Supplementary Fig. S3A). The primer pair of AMPL1040313 was found to have a similar (2- and 1-bp mismatches in the forward and reverse primers) sequence to that of the *E3* homologous gene on Chr3 (Supplementary Fig. S3B). A comparison of sequence of the original amplicon designed for the *E3* gene on Chr19 with that designed for the *phyA* gene on Chr3 revealed that the amplicon of *phyA* could be mapped preferentially to *E3* in the absence of sequence information (Supplementary Fig. S3C). To preferentially output the alignment containing indel, we changed the penalty parameter by using the option '-A 10 -M 60 -O

50 -E 1' to TMAP. The default parameters of TMAP option are as follows: '-A', score for a match [default = 1]; '-M', the mismatch penalty [3]; '-O', the indel start penalty [5]; and '-E', the indel extension penalty [2]. By increasing the penalty values related to base match and InDels, the miss-mapped reads on the fourth exon can be reduced from 2,426 (default parameter) to 3 reads (Fig. 3A). By optimizing these parameters, the miss-mapped reads on the second exon of *E3* were also mapped to the correct position, *phyA* on Chr3 (second exon; Fig. 3A and Supplementary Fig. S4). We also investigated whether the 15-kb deletion can be detected using the read coverage. The APKM of the four amplicons located at the fourth exon of *E3* was compared between the *E3* and *e3tr* alleles (Fig. 3B). The number of accessions classified as *E3* and *e3tr* by additional marker analysis was 152 and 36, respectively (Supplementary Table S7). The APKM of four amplicons on the fourth exon of the *e3tr* allele was almost zero, whereas the APKM of the *E3* allele varied depending on the accessions, and it was difficult to judge the presence or absence of deletion from the APKM of each amplicon. However, it was possible to classify the presence or absence of deletion clearly (Fig. 3B and Supplementary Table S7) when the average APKM of the four amplicons was used instead, because differences in amplification efficiency due to sequence variation at the priming site can be cancelled using the APKM of multiple amplicons.

As described above, appropriate parameters are required to map short amplicon reads to the correct genomic region. As the AmpliSeq technology has been mainly used in animals in which palaeopolyploidy is considerably rare, no such limitation has been reported. Soybean is an ancient tetraploid, which underwent two whole-genome duplications (palaeopolyploidy); most of the genes have paralogous genes with multiple copies.[23] The information provided above would be useful when the AmpliSeq technology is applied to plant species, which have experienced whole-genome duplication or triplications.

3.3. Detected variants of flowering time-related genes

A total of 192 soybean mini-core collection was analysed to detect novel variants in flowering time-related genes by AmpliSeq. Among the 461 variants (SNPs or InDels) detected in the target regions, 311 (67.5%) sites have already been reported or registered in dbSNP,[25] whereas 150 sites (32.5%) were novel (Table 2 and Supplementary Table S8). The variants detected were compared in depth with information of flowering time-related genes, *E1*, *E2*, *E3*, *E4*, *FT2a*, *FT5a*, and their homologues.[3,17,18] Further, we performed linear regression analysis to detect responsible variants associated with flowering time under long-day field conditions. Among the 461 variants, 207, 206, and 219 were found to be potentially associated with flowering time in 2011, 2012, and 2013 by the simple linear regression analysis, respectively (Table 3 and Supplementary Table S8). The most significantly associated variant with flowering time was SNP (rs124971350) at *E2* (*e2* allele) ($P < 2.0e^{-16}$ for 3 yrs) (Table 3 and Supplementary Table S8). The second most significant variant was a large deletion in *E3* (*e3-tr* allele) ($P < 1.9e^{-13}$, $1.6e^{-13}$, and $5.2e^{-14}$ for 2011, 2012, and 2013, respectively). The other seven genes, *WD repeat-containing protein 61*, *Dt1*, *MADS-box protein*, two genes of *two-component response regulator*-like genes, *FT2a*, and *PhyB*, showed highly significant association with flowering time ($P < 0.0001$) (Table 3 and Supplementary Table S8).

E1 and *E1*-like genes

Five alleles, *E1*, *e1-as*, *e1-nl*, *e1-fs*, and one novel missense (Chr06_20207355) were identified at *E1* (Fig. 4, Supplementary Fig. S5A, and Supplementary Tables S8 and S9). In the other two B3 domain containing *E1*-like genes, only one synonymous variant (rs123097808) was detected in *Glyma.04G156400/E1La*, whereas no variant was detected in the coding region of *Glyma.04G143300/E1Lb* (Supplementary Fig. S5B and C). Among the five *E1* alleles, the frequency of *e1-as* allele (Chr06:20207322 C: Williams 82 type) was 0.09 and Chr06:20207322 C to G nucleotide change (rs123612969) was 0.91 among the soybean mini-core collection. In contrast, *e1-nl*, which lacks the entire *E1* gene,[10] was only found in Swedish cultivar FiskebyV (PGC001) (Supplementary Table S9). This allele was determined by the read coverage at the *E1* genomic region. The average normalized read coverage of all six amplicons (AMPL1037682–AMPL1037687) was three in PGC001, whereas that of these amplicons in the other accessions was 181 (ranging from 36 to 430). Additional experiments to confirm the deletion in the *E1* genomic region by PCR amplification revealed that only PGC001 lacks the *E1* genomic region among all accessions and possesses the *e1-nl* allele. Another allele, *e1-fs*, which had 1-bp deletion variant (Chr06_20207323) was also found in one accession PGC002 (Fig. 4, Supplementary Fig. S5A, and Supplementary Table S9). The deletion (Chr06_20207323) causes a frameshift and introduces a premature stop codon at Lys76. These loss-of-function alleles were not included in the association analysis due to very low allele frequency (only one accession each), but might explain very early flowering of PGC001 and PGC002 under long-day-length field condition. PGC002 (Wase kuro daizu) is originated from the southern part of Japan and classified as the summer-type soybean, early-maturity group in low-latitude regions of Japan. The summer-type soybean has low photoperiod sensitivity, and *e1-fs* can explain this characteristic. In contrast, the novel missense (Chr06_20207355) from PGC139 and PGC147 (Supplementary Fig. S5A and Supplementary Table S9) did not show large effect on flowering time and might not significantly affect the E1 function.

E2

Two alleles, *E2* and *e2*, and one novel SNP variant (Chr10_45310686) were detected at *E2* (Fig. 4, Supplementary Fig. S5D, and Supplementary Table S8). Among them, functional defective *e2* allele had A to T nucleotide change (Table 3, K528*, rs124971350) and the allele frequency among the soybean mini-core collection was 0.42 (Supplementary Table S8). A novel SNP (Chr10_45310686), which causes missense variant of Ile490Met, was detected only in PGC086 with the *e2* allele (Supplementary Table S9 and Supplementary Fig. S5D).

E3

Among two alleles, *e3-tr* and *e3-Mo*, detected at *E3*, the frequency of *e3-tr* allele, which has a large deletion in the fourth exon, was 0.19 (Fig. 4, Supplementary Fig. S5E, and Supplementary Tables S7 and S8). The missense variant of *e3-Mo* (Chr19_47638302: G to A, Gly1050Arg) in the third exon of *E3* was detected in PGC019 and PGC042 (Moshidou Gong 503)[3] derived from Korean Peninsula and China, respectively. The *e3-Mo* variant is not registered in dbSNP, but we found in one Chinese landrace Ni Ding Hua Mei Dou from 302 soybean re-sequence data[28] (SRR1533240 in NCBI SRA).

Table 2. Classification of detected variants of 28 flowering time-related genes by snpEff analysis

Glyma ID	Description	Gene symbol	Number of variants in total	Number of novel variants	Number of known variants	Number of missense variants	Number of 'high' effect variants to gene function[a]
Glyma.02G069500	FLOWERING LOCUS T	*GmFTL7*	10	2	8	0	0
Glyma.03G194700	TERMINAL FLOWER 1	*GmTFL2*	2	0	2	0	0
Glyma.03G227300	Phytochrome A	*GmPHYA4*	70	5	65	35	5
Glyma.04G143300	B3 domain-containing protein	*E1Lb*	3	1	2	0	0
Glyma.04G156400	B3 domain-containing protein	*E1La*	4	0	4	0	0
Glyma.06G207800	B3 domain-containing protein	*E1*	5	2	3	3	2
Glyma.08G363100	FLOWERING LOCUS T	*GmFT4*	9	2	7	0	0
Glyma.08G363200	FLOWERING LOCUS T	*GmFT6*	6	1	5	0	0
Glyma.09G035500	Phytochrome B	*GmPHYB1*	10	6	4	2	0
Glyma.09G143500	BROTHER OF FT AND TFL 1	*GmTFL4*	1	1	0	0	0
Glyma.10G141400	Phytochrome A		20	8	12	3	1
Glyma.10G221500	GIGANTEA	*E2*	2	1	1	1	1
Glyma.12G073900	Two-component response regulator-like		18	3	15	5	1
Glyma.15G140000	Phytochrome B		27	9	18	7	2
Glyma.16G044100	FLOWERING LOCUS T	*GmFT5a/GmFTL4*	14	5	9	0	0
Glyma.16G044200	FLOWERING LOCUS T	*GmFT3a/GmFTL1*	8	2	6	1	0
Glyma.16G150700	FLOWERING LOCUS T	*GmFT2a/GmFTL3*	22	13	9	2	1
Glyma.16G151000	FLOWERING LOCUS T	*GmFT2b/GmFTL5*	9	3	6	1	0
Glyma.16G196300	BROTHER OF FT AND TFL 1	*GmTFL3*	20	5	15	1	0
Glyma.16G200700	MADS box protein		45	15	30	7	0
Glyma.17G052100	WD repeat-containing protein 61	*WDR61*	15	1	14	0	1
Glyma.17G090500	Achaete-scute transcription factor-related		17	10	7	0	6
Glyma.19G108100	FLOWERING LOCUS T	*GmFT3b/GmFTL2*	0	0	0	0	0
Glyma.19G108200	FLOWERING LOCUS T	*GmFT5b/GmFTL6*	11	3	8	1	0
Glyma.19G194300	TERMINAL FLOWER 1	*Dt1/GmTFL1*	14	5	9	5	0
Glyma.19G224200	Phytochrome A	*E3*	16	8	8	5	3
Glyma.19G260400	Pseudo-response regulator 5		42	11	31	8	0
Glyma.20G090000	Phytochrome A	*E4*	16	2	14	1	0
Glyma.U034500	Two-component response regulator-like		26	26	0	5	7

[a] High-impact variant to gene function includes criteria of stop lost, stop gained, and frameshift.

Table 3. DNA polymorphisms in nine genes significantly associated with flowering time variation in the mini-core collection as revealed by the single linear regression analysis

Glyma ID	Gene description	Gene symbol[a]	Chromosome	Position (bp)	Reference sequence	Alternative sequence	SNP name	*P* value of association tests			Functional effect	Amino acid change	Functional class
								2011	2012	2013			
Glyma.06G207800	B3 domain-containing protein	*E1*	Chr06	20207322	C	G	rs123612969 (*e1-as*)	0.0226	0.0153	0.0333	Missense_variant	p.Thr75Arg/c.224C>G	MISSENSE
Glyma.10G221500	GIGANTEA	*E2*	Chr10	45310798	A	T	rs124971350 (*e2*)	2E-16	2E-16	2E-16	Stop-gained	p.Lys527*/c.1582A>T	HIGH
Glyma.12G073900	Two-component response regulator-like	-	Chr12	5508242	T	G	rs125308101	0.00258	0.0082	0.0019	Upstream_gene_variant		
			Chr12	5508672	CT	C	rs745192414	0.00155	0.00261	0.00503	Intron_variant	c.-120 + 219delT	
			Chr12	5508702	T	C	rs743387935	0.0211	0.0228	0.0109	Intron_variant	c.-119-198T>C	
			Chr12	5509310	G	A	rs388619068	0.0146	0.0119	0.00907	Missense_variant	p.Asp98Asn/c.292G>A	MISSENSE
			Chr12	5509317	C	T	rs125308103	0.00107	0.00148	0.00319	Missense_variant	p.Ser100Leu/c.299C>T	MISSENSE
			Chr12	5519728	A	C	rs125308115	0.00229	0.00749	0.00169	Missense_variant	p.Lys378Gln/c.1132A>C	MISSENSE
			Chr12	5520578	A	G	rs389122657	0.014	0.0112	0.00865	Synonymous_variant	p.Ala504Ala/c.1512A>G	SILENT
			Chr12	5520945	T	C	rs125308117	8.96E-10	1.76E-08	7.74E-11	Stop_lost	p.Ter627Glnext*?/c.1879T>C	MISSENSE
HIGH			Chr12	5521029	G	T	rs743179814	0.0244	0.0203	0.0089	3_Prime_UTR_variant	c.*82G>T	
Glyma.15G140000	Phytochrome B	-	Chr15	11436193	G	A	rs388435741	5.17E-06	7.93E-06	8.47E-06	3_Prime_UTR_variant	c.*218C>T	
			Chr15	11441207	C	T	rs126279495	5.17E-06	7.93E-06	8.47E-06	Missense_variant	p.Val394Ile/c.1180G>A	MISSENSE
			Chr15	11442400	A	C	rs126279502	0.0753	0.0769	0.0431	5_Prime_UTR_variant	c.-14T>G	
Glyma.16G150700	FLOWERING LOCUS T	*GmFT2a/GmFTL3*	Chr16	31110004	TATAAGAAAGC	T	rs392064733_1	0.0397	0.117	0.0266	5_Prime_UTR_variant	c.-50_-59delATAAGAAAGC	
			Chr16	31110004	TATAAGAAAGCA	TA	rs392064733_2	0.0397	0.117	0.0266	5_Prime_UTR_variant	c.-49_-58delTAAGAAAGCA	
			Chr16	31110991	A	AATAT	rs864598505_1	6.24E-05	0.000274	0.000111	Intron_variant	c.202-56_202-55insATAT	
			Chr16	31111033	T	A	rs126829817	6.24E-05	0.000274	0.000111	Intron_variant	c.202-15T>A	
			Chr16	31111042	T	C	rs126829818	0.0697	0.21	0.0419	Splice_region_variant	c.202T>C	
			Chr16	31111349	A	G	rs126829819	6.24E-05	0.000274	0.000111	Intron_variant	c.304 + 72A>G	
			Chr16	31114633	G	A	Chr16_31114633	5.46E-06	4.14E-06	3.05E-07	Missense_variant	p.Gly169Asp/c.506G>A	MISSENSE
			Chr16	31114658	AGA	AA	Chr16_31114658_2	0.0765	0.0538	0.043	3_Prime_UTR_variant	c.*1delG	
			Chr16	31114930	G	T	rs126829846	6.24E-05	0.000274	0.000111	3_Prime_UTR_variant	c.*272G>T	
Glyma.16G200700	MADS box protein	-	Chr16	36179909	T	G	rs126888526	1.13E-07	3.18E-06	3.35E-07	3_Prime_UTR_variant	c.*218A>C	
			Chr16	36180002	G	A	rs126888527	1.09E-07	1.63E-06	1.98E-07	3_Prime_UTR_variant	c.*125C>T	
			Chr16	36180087	T	TGA	Chr16_36180087	8.8E-08	2.16E-06	2.07E-07	3_Prime_UTR_variant	c.*39_*40insTC	
			Chr16	36180122	A	C	rs126888528	4.19E-07	9.01E-06	1.03E-06	3_Prime_UTR_variant	c.*5T>G	
			Chr16	36180183	T	A	rs389022394	4.96E-07	1.03E-05	1.32E-06	Missense_variant	p.Thr219Ser/c.655A>T	MISSENSE
			Chr16	36182854	T	TA	Chr16_36182854	6.15E-06	4.11E-05	3.79E-06	Intron_variant	c.511-49_511-50insT	

Continued

Table 3 continued

Glyma ID	Gene description	Gene symbol[a]	Chromosome	Position (bp)	Reference sequence	Alternative sequence	SNP name	*P* value of association tests			Functional effect	Amino acid change	Functional class
								2011	2012	2013			
			Chr16	36182857	T	A	rs126888574	8.41E-07	7.3E-06	5.2E-07	Intron_variant	c.511-52A>T	
			Chr16	36183423	T	G	rs126888580_2	6.9E-06	5.27E-06	2.61E-06	Intron_variant	c.427-19A>C	
			Chr16	36183426	G	C	rs126888581_1	0.00286	0.00159	0.0018	Intron_variant	c.427-22C>G	
			Chr16	36183435	G	C	rs126888582	6.8E-06	5.07E-06	2.66E-06	Intron_variant	c.427-31C>G	
			Chr16	36183450	G	A	rs126888583	0.0104	0.00721	0.00804	Intron_variant	c.427-46C>T	
			Chr16	36183510	C	T	rs126888584	0.000219	0.00275	0.000461	Intron_variant	c.427-106G>A	
			Chr16	36183518	A	G	rs126888585	3.22E-10	9.24E-09	3.46E-10	Intron_variant	c.427-114T>C	
			Chr16	36183541	C	A	rs126888586	2.98E-10	8.83E-09	3.27E-10	Intron_variant	c.427-137G>T	
			Chr16	36183556	T	C	Chr16_36183556	0.0507	0.0363	0.048	Intron_variant	c.427-152A>G	
			Chr16	36183568	T	A	rs126888588	0.00193	0.0172	0.00328	Intron_variant	c.427-164A>T	
			Chr16	36183573	T	A	rs126888589	0.00193	0.0172	0.00328	Intron_variant	c.427-169A>T	
			Chr16	36184165	T	C	rs126888605	4.71E-05	3.54E-05	0.000019	Intron_variant	c.327-41A>G	
			Chr16	36184729	T	C	rs126888609	2.18E-09	2.34E-08	6.84E-10	Missense_variant	p.Thr79Ala/c.235A>G	MISSENSE
			Chr16	36184733	C	T	rs126888610	2.18E-09	2.34E-08	6.84E-10	Synonymous_ variant	p.Ser77Ser/c.231G>A	SILENT
			Chr16	36184819	A	T	rs126888611	0.0341	0.0817	0.0244	Intron_variant	c.183-38T>A	
			Chr16	36187211	A	C	rs744233319	0.000287	0.000744	0.00118	Synonymous_ variant	p.Ser36Ser/c.108T>G	SILENT
			Chr16	36187552	T	G	rs126888636	0.0383	0.0315	0.0302	Upstream_gene_ variant		
			Chr16	36187600	A	G	rs126888637	0.0383	0.0315	0.0302	Upstream_gene_ variant		
Glyma.17G052100	WD repeat-containing protein 61	*WDR61*	Chr17	3955280	A	AT	rs126942305_1	0.00214	0.00191	0.00197	Downstream_gene_ variant		
			Chr17	3955411	A	G	rs126942313	2.56E-10	3.12E-10	1.07E-10	Downstream_gene_ variant		
			Chr17	3955475	TC	AA	rs126942314_1	0.00359	0.00645	0.00279	Downstream_gene_ variant		
			Chr17	3955476	C	A	rs126942315	0.00359	0.00645	0.00279	Downstream_gene_ variant		
			Chr17	3955546	T	A	rs126942316	0.00374	0.00157	0.00264	3_Prime_UTR_ variant	c.*125A>T	
			Chr17	3955554	T	C	rs126942317	0.031	0.0299	0.0358	3_Prime_UTR_ variant	c.*117A>G	
			Chr17	3955716	A	G	rs388258139	0.00418	0.00761	0.00338	Synonymous_ variant	p.Ala307Ala/c.921T>C	SILENT
			Chr17	3955763	C	CG	Chr17_3955763	0.00444	0.00194	0.00318	Frameshift_ variant	p.Val291_Ala292fs/ c.873_874insC	HIGH
			Chr17	3955764	A	G	rs126942318	0.00374	0.00157	0.00264	Synonymous_ variant	p.Val291Val/c.873T>C	SILENT
			Chr17	3955884	G	C	rs126942319	0.00374	0.00157	0.00264	Synonymous_ variant	p.Val251Val/c.753C>G	SILENT
			Chr17	3956142	T	C	rs126942321	6.36E-08	3.72E-08	1.25E-08	Synonymous_ variant	p.Ala165Ala/c.495A>G	SILENT
			Chr17	3956163	T	C	rs126942322	0.00644	0.00295	0.00466	Synonymous_ variant	p.Lys158Lys/c.474A>G	SILENT
			Chr17	3958319	C	G	rs126942340	9.03E-10	4.86E-10	1.98E-10	Synonymous_ variant	p.Ser16Ser/c.48G>C	SILENT

Continued

Table 3 continued

Glyma ID	Gene description	Gene symbol[a]	Chromosome	Position (bp)	Reference sequence	Alternative sequence	SNP name	P value of association tests			Functional effect	Amino acid change	Functional class
								2011	2012	2013			
			Chr17	3958374	T	G	rs126942341	2.32E-10	9.96E-11	5.12E-11	5_Prime_UTR_ variant	c.-8A>C	
Glyma.19G194300	TERMINAL FLOWER 1	*Dt1/GmTFL1*	Chr19	45183701	T	A	rs127928573 (*dt1*)	0.00122	0.00587	0.000401	Missense_variant	p.Arg166Trp/c.496A>T	MISSENSE
			Chr19	45183808	C	T	rs745009806	0.0319	0.0184	0.0153	Missense_variant	p.Arg130Lys/c.389G>A	MISSENSE
			Chr19	45183859	G	A	rs127928574	4.29E-11	3.71E-11	6.02E-10	Missense_variant	p.Pro113Leu/c.338C>T	MISSENSE
			Chr19	45184581	G	T	Chr19_45184581	0.014	0.153	0.0428	Splice_region_ variant	c.202C>A	
			Chr19	45185131	C	T	Chr19_45185131	0.00805	0.00351	0.00764	5_Prime_UTR_ variant	c.-142G>A	
Glyma.19G224200	Phytochrome A	*E3*	Chr19	47633086	T	TA	Chr19_47633086	0.000763	0.000353	0.00085	5_Prime_UTR_ variant	c.-693_-694insA	
			Chr19	47634596	A	G	rs127944661	0.0052	0.00108	0.00429	Synonymous_ variant	p.Ser43Ser/c.129A>G	SILENT
			Chr19	47635025	C	T	rs388644281	3.18E-05	3.43E-05	3.69E-05	Synonymous_ variant	p.Ile186Ile/c.558C>T	SILENT
			Chr19	47635737	C	A	rs389001110	0.0534	0.052	0.048	Missense_variant	p.Leu424Ile/c.1270C>A	MISSENSE
			Chr19	47636564	G	A	rs127944664	0.0098	0.00232	0.00888	Intron_variant	c.2074 + 23G>A	
			Chr19	47636607	G	T	rs127944665	0.0127	0.0031	0.0113	Intron_variant	c.2074 + 66G>T	
			Chr19	47637258	A	G	rs393405985_1	0.044	0.0447	0.0427	Missense_variant	p.Thr832Ala/c.2494A>G	MISSENSE
			Chr19	47638302	G	A	Chr19_47638302 (*e3-Mo*)	0.0433	0.0306	0.0485	Missense_variant	p.Gly1050Arg/c.3148G>A	MISSENSE
			Chr19	47638344	A	T	rs389636522	0.00508	0.00113	0.00433	Splice_region_ variant	c.3183A>T	
			Chr19	47641562		15 kb deletion	*e3-tr*	1.9E-13	1.61E-13	5.15E-14	Loss of exon 4		HIGH
Glyma.U034500	Two-component response regulator-like	-	Scaffold_32	197169	C	T	Scaffold_32_197169	0.000318	0.000771	0.000612	3_Prime_UTR_ variant	c.*958G>A	
			Scaffold_32	197421	T	C	Scaffold_32_197421	0.000034	0.00007	8.54E-05	3_Prime_UTR_ variant	c.*706A>G	
			Scaffold_32	197459	A	T	Scaffold_32_197459	1.23E-09	1.94E-08	2.71E-09	3_Prime_UTR_ variant	c.*668T>A	
			Scaffold_32	198053	T	A	Scaffold_32_198053	2.16E-05	4.99E-05	0.000048	Intron_variant	c.*133 + 29A>T	
			Scaffold_32	198773	A	AT	Scaffold_32_198773	1.85E-06	3.75E-06	2.85E-06	Frameshift_variant	p.Met737_Ala738fs/ c.2209_2210insA	HIGH
			Scaffold_32	199551	A	C	Scaffold_32_199551	4.88E-09	6.54E-08	1.23E-08	Synonymous_ variant	p.Ala529Ala/c.1587T>G	SILENT
			Scaffold_32	199605	T	G	Scaffold_32_199605	2.16E-05	4.99E-05	0.000048	Intron_variant	c.1575-42A>C	
			Scaffold_32	199656	TA	T	Scaffold_32_199656	0.0096	0.0104	0.011	Intron_variant	c.1574 + 2delT	
			Scaffold_32	199722	C	G	Scaffold_32_199722	0.000318	0.000771	0.000612	Missense_variant	p.Gly504Ala/c.1511G>C	MISSENSE
			Scaffold_32	202754	G	A	Scaffold_32_202754	0.0142	0.01	0.0105	Stop_gained	p.Arg308*/c.922C>T	
HIGH			Scaffold_32	206717	A	G	Scaffold_32_206717	1.23E-09	1.94E-08	2.71E-09	Intron_variant	c.792 + 27T>C	
			Scaffold_32	216494	C	T	Scaffold_32_216494	4.88E-09	6.54E-08	1.23E-08	Synonymous_ variant	p.Val168Val/c.504G>A	SILENT
			Scaffold_32	218380	A	T	Scaffold_32_218380	3.29E-09	4.29E-08	8.4E-09	Synonymous_ variant	p.Pro73Pro/c.219T>A	SILENT

[a] Gene names refer to those in Kong et al.,[8] Fan et al.,[52] and Cao et al.[53]

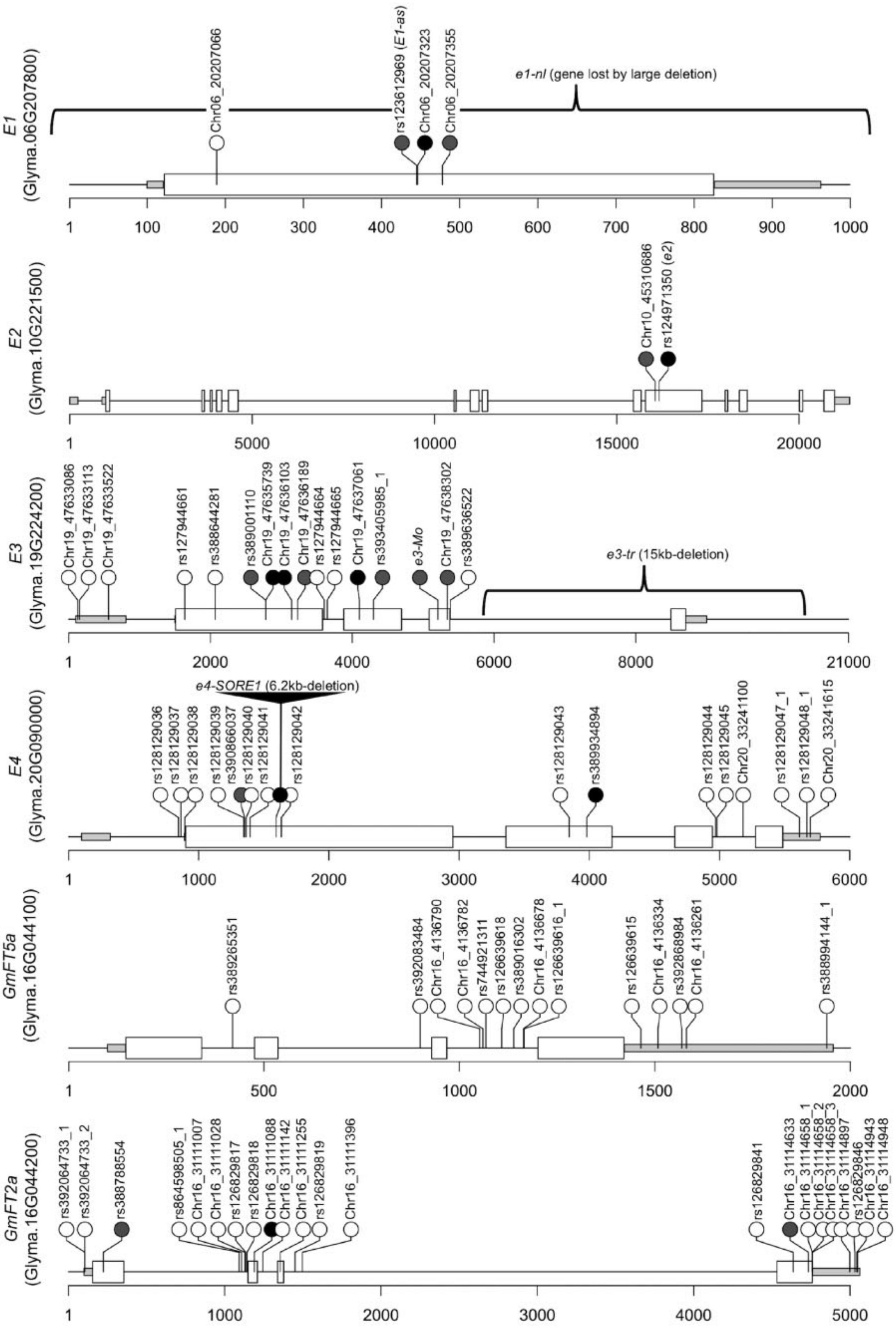

Figure 4. Detected variants in *E1, E3, E4, FT5a*, and *FT2a* from 192 mini-core collection. The grey and white boxes indicate UTRs and exons. The solid lines indicate 5′-uptream and intron regions. The black and grey circles indicate loss-of-function (described as 'HIGH' impact on the gene function in Supplementary Table S8) and missense variants. Braces indicate known large InDels. These InDels can be detected by read depth.

E4

The *e4-SORE-1* allele has a 6.2-kb insertion in the first exon. It was difficult to estimate this insertion from the read coverage of amplicon in the region. However, the presence or absence of a large insertion could be estimated from the read coverage of amplicon (AMPL1037734) at the break point of a large insertion (Supplementary Fig. S5F). The average APKM of break point was 325 in the reference type sequence, whereas it was zero in the insertion type sequence of PGC001 and PGC021 derived from Sweden and Japan, respectively. This insertion was also confirmed by the PCR.

Most SNPs (11 of 13 sites) in *E4* were found from PGC123 and PGC134 derived from Nepal and China (Supplementary Table S9), but there was only one missense variant (rs390866037: Leu151Ser), which likely affects gene function. As these variants were detected as homozygous, they are considered to be real variants, not detected by the miss-mapped reads. A frameshift variant in the second exon was only found in PGC005 (Supplementary Table S9). This accession flowered earlier than Williams 82 under field conditions in spite of the same gene combination for all other flowering-related genes (Supplementary Table S1).

Other *Phytochrome A* genes

Five variants, three frameshifts and two splice site variants, were identified to be high-impact variant to another *PhyA* gene, *Glyma.03G227300/GmPHYA4* (Supplementary Fig. S5G and Supplementary Table S9). This *PhyA* gene consisted of two main haplotypes, namely, reference type (Hap1–Hap5) and pseudogene type (Hap6–Hap11), which had various loss-of-function sites. The other *PhyA* gene, *Glyma.10G141400*, had only one novel frameshift variant (Chr10_37491867) from two accessions, PGC045 and PGC189 derived from Korea and East Timor (Supplementary Tables S6 and S9 and Supplementary Fig. S5H).

Phytochrome B genes

There has been no report of natural variation in the *PhyB* genes affecting flowering, but the overexpression of *GmPHYB1* accelerates flowering under short-day conditions in *Arabidopsis*.[29] Only one missense variant (rs124458274) was found from *GmPHYB1* (*Glyma.09G035500*) (Supplementary Table S8 and Supplementary Fig. S5I). rs124458274 was a common variation in the mini-core collection (allele frequency = 0.69). Two novel frameshifts and six missense variants (one was novel) were found in the other *PhyB* gene *Glyma.15G140000* (Supplementary Table S8 and Supplementary Fig. S5J). Although the frameshift variant (Chr15_11442094) was identified in the 19 accessions (Hap8, Supplementary Table S9), no association with flowering time under the examined field conditions was observed.

FLOWERING LOCUS T

Two florigen genes *FT5a* and *FT2a* in the soybean genome play a major role in the induction of flowering.[17,18,30] As no variant was detected in the exon of *FT5a*, it appears that *FT5a* is highly conserved under the evolutionary constraint (Fig. 4 and Supplementary Fig. S5K). Nine and five variants were detected in the intron and 3′-UTR, respectively. Of these, four variants were associated with flowering time determined by simple linear regression analysis (Supplementary Tables S3 and S8). Two variants (rs126630615 in 3′-UTR and rs126639618 in the third intron) were reported by Takeshima et al.[18] This *FT5a* region has been reported to be one of the flowering time quantitative trait loci (QTLs) in the chromosomal segment substitution lines (CSSLs) derived from a cross between Peking and Enrei.[31] In this study, the nucleotide differences between Enrei and Peking were identified as rs126639616_1 (Fig. 4) in the intron and rs388994144_1 (Fig. 4) in the 3′-UTR region (Supplementary Fig. S5K and Supplementary Table S8). As natural variants in 5′- and 3′-UTR of the *FT*-like gene affect gene expression and flowering time in rice,[32,33] the variant rs388994144_1 in 3′-UTR region might be involved in gene expression of *FT5a* and regulation of flowering time in soybean.

The *FT2a* is a paralogue of *FT5a* and has been named as *E9*.[8] *E9* is a leaky allele that is caused by allele-specific transcriptional repression due to the insertion of *SORE-1* into the first intron. The presence of *SORE-1* (*FT2a-TO* allele) delays flowering for 10 days under natural day-length conditions at Harbin, China (45°43′N, 126°45′E).[8] Zhao et al.[17] also reported a difference of 10 days or more in flowering time between *E9* and *e9* in Sapporo, Japan (43°07′N, 141°35′E). As we did not design the primers on the intron of *FT2a*, the presence or absence of *SORE-1* is unknown. In this study, one frameshift and two missense variants were found in *FT2a* (*Glyma.16G150700*) (Fig. 4 and Supplementary Fig. S5L). In the first exon, missense SNP rs388788554 (Glu23Asp) was detected in PGC066 (Hap13, Supplementary Table S9). In the fourth exon, missense novel SNP (Chr16_31114633) was detected in seven accessions (Hap10, Supplementary Table S9). Another novel frameshift variant (Chr16_31111088) in the fourth exon was detected in PGC166 (Hap14, Supplementary Table S9). Enrei had four known and two novel variants in the intron and 3′-UTR of *FT2a*, whereas there was no variant in Peking. As QTL is not reported in the *FT2a* region of the CSSL between Peking and Enrei,[31] these variants might be not involved in the regulation of flowering time under the evaluation conditions of CSSLs.

FT-like genes

Four missense variants were detected in the other three *FT* homologues, rs126830445 in *Glyma.16G151000* (*GmFT2b/GmFTL5*, Supplementary Fig. S5M), rs127848197 in *Glyma.19G108200* (*GmFT5b/GmFTL6*, Supplementary Fig. S5O), a novel SNP (Chr16_35778390) in *Glyma.16G196300* (*GmTFL3*, Supplementary Fig. S5N), and Chr16_4162554 in *Glyma.16G044200* (*FT3a/GmFTL1*, Supplementary Fig. S5P). The frequency of rs126830445 in *GmFT2b/GmFTL5* was 0.45, whereas that of rs127848197 in *GmFT5b/GmFTL6* was 0.94. A novel missense (Val98Ile) SNP (Chr16_35778390) in *GmTFL3* was only found in PGC037. This accession (YAKUMO MEAKA) is a landrace from Hokkaido, northern part of Japan. In contrast, a novel missense variant (Chr16_4162554) in *FT3a/GmFTL1* was only found in PGC134 (Hap7, Supplementary Table S9), which is a medium-maturing accession. No functional defect or missense variant in the other *FT*-like genes, *Glyma.02G069500* (*GmFTL7*, Supplementary Fig. S5Q), *Glyma.08G363100* (*GmFT4*, Supplementary Fig. S5R), and *Glyma.08G363200* (*GmFTL6*, Supplementary Fig. S5S), was found. No variants were detected in *Glyma.19G108100* (*GmFT3b/GmFTL2*). The information of alleles identified in these *FT*-like genes will be useful to clarify the influence of these variants on flowering regulation.

TFL1-like genes

Two *TFL1*-like genes, *GmTFL2* (*Glyma.03G194700*) and *GmTFL1* (*Glyma.19G194300*), exist in the soybean genome. No loss-of-function or missense variant was found in *GmTFL2* (Supplementary Fig.

S5T), whereas five missense variants were found in *GmTFL1*, which determine the growth habit of soybean, classically named as *Dt1* locus[34] (Supplementary Fig. S5U and Supplementary Table S8). These sites are located where amino acids are highly conserved across *TFL1* orthologues: GmTFL2, GmTFL1/Dt1, *Lotus japonicas* CEN/TFL1, pea TFL1a, *Arabidopsis* TFL1, *Arabidopsis* ATC, and *Antirrhinum majus* CEN.[35–39] The variant site of rs745009806 (Arg130Lys) was conserved in *TFL*, but not in ATC and CEN. Another four variant sites, rs127928577 (Arg62Ser), rs392653457 (Leu67Gln), rs127928574 (Pro113Leu), and rs127928573 (Arg166Trp), exist at a highly conserved amino acid site. Of these, rs127928573 (Arg166Trp) is known as *dt1* allele in soybean.[34] As the loss-of-function *Sidt1* allele has been reported at S79N in *Sesamum indium* L.,[40] the other three missense variants should be examined to verify whether they are new defective *dt1* alleles or not.

Two-component response regulator-like genes

Among three *two-component response regulator*-like genes screened, the stop-lost variant (rs125308117) and five missense variants were found in *Glyma.12G073900* (Supplementary Table S8 and Supplementary Fig. S5V). The allele frequency of the stop-lost variant (rs125308117) was 0.31. In *Glyma.19G260400*, only seven missense variants were found (Supplementary Table S8 and Supplementary Fig. S5W). Among seven variants with high impact on gene function in *Glyma.U034500* on scaffold 32 (Supplementary Table S8 and Supplementary Fig. S5), four were frameshift variants due to InDels. Frameshift variant (A > AT, M737I, Scaffold32: 198773) was the major allele, and 82% of the mini-core collection possesses this allele. In contrast, other frameshift variants of insertion (C > CT, Q759M, scaffold_32: 198706) from PGC044, deletion (TTGCC > -, G409D, scaffold_32: 200001) from 16 accessions, and deletion (AC > A, V355L, scaffold_32: 200258) from three accessions, PGC005, PGC094, and PGC174, were rare alleles in the mini-core collection (Supplementary Table S9). The remaining three variants (scaffold_32_199043, scaffold_32_202754, and scaffold_32_218486) with high impact on gene function were stop-gained variant. These results indicate that *Glyma.U034500* of most soybean accessions, except for Hap1, Hap2, Hap3, and Hap4 (Supplementary Table S9), losses its function. Among these three genes, *Glyma.12G073900* and *Glyma.U034500* showed high similarity (91%) at the amino acid sequence level. The fact that length of the amino acid sequence of *Glyma.12G073900* of Williams 82 is shorter (92 aa) than that of *Glyma.U034500* (765 aa) at the C terminal indicates that *Glyma.12G073900* encodes truncated protein. Although flowering control by two-component response regulator-like genes has been reported in various species,[41–43] the role of this gene and its variant in soybean flowering are unknown. Among the *two-component response regulator*-like genes, only variants in *Glyma.12G073900* and *Glyma.U034500* were associated with flowering time, determined by simple linear regression analysis (Table 3 and Supplementary Table S8). *Glyma.U034500* (Chr11 11.23-11.26Mb on Gmax189) is located near a previously reported QTL as *qFT-B1* (nearest marker: Satt519 74.7cM, Chr11 13.98Mb on Gmax189) in the 96 from the cross between Tokei 780 and the soja accession Hidaka 4.[44] Although they reported the effect of *qFT-B1* is 3.4–10.8 days, the genotype of this frameshift site (Scaffold 32:198773) in the parent of recombinant inbred lines (RILs) is unknown. It can be confirmed using the detected variants as a DNA marker whether detected stop-lost and stop-gained variants in *two-component response regulator*-like genes are responsible genes for the flowering time.

Other genes

Seven missense variants were identified in *Glyma.16G200700* encoding MADS box protein, whereas functional defect variant was not found (Supplementary Table S8 and Supplementary Fig. S5Y). MADS-domain transcription factor of the *AGL6* gene is known to be a factor responsible for the regulation of lateral organ development, flowering time, and circadian clock in *Arabidopsis*.[45,46] *AGL6* regulates flowering through the *FLC* family genes and *FT*.[46] Two missense variants, rs389022394 and rs126888609, on *Glyma.16G200700* are significantly associated with the flowering time, determined by the simple linear regression analysis (Table 3 and Supplementary Table S8). As there is no report for *FLC*-like genes in soybean, it will be important to examine whether the detected two variants from *Glyma.16G200700* have an effect on the flowering time.

One novel frameshift variant (Chr17_3955763) of WD repeat-containing protein 61 (*Glyma.17g52100*) was significantly associated with flowering time, determined by the simple linear regression analysis (Table 3 and Supplementary Table S8). Although the allele frequency of the frameshift variant was 0.52 (Supplementary Table S8), no QTL has been reported in this region. In *Arabidopsis*, WD repeat-containing protein VIP3 regulates flowering time via the vernalization pathway;[47] however, the vernalization pathway is not known in soybean and it is difficult to infer the role. As a large proportion of the mini-core collection has the novel frameshift variant for *Glyma.17G052100* and missense variant for *Glyma.16G200700*, it is necessary to confirm genetically whether these novel variants really affect the flowering time. The transcription factor gene, *Glyma.17G090500,* was sequenced as the control for variant detection. All known variants were detected correctly (data not shown).

3.4. Gene-based association test for flowering time

To refine responsible variants associated with variation in flowering time in the mini-core collection, we performed multiple linear regression analysis using variants significantly associated with flowering time in the simple linear regression analysis. The variants of *e2*, *e3-tr* and stop-lost variant (rs125308117) of *two-component response regulator*-like gene on Chr12 were significant in 3 yrs, and rs127928573 in *Dt1* was significant only in 2013 (Table 4). These genes could explain 51.82%, 51.13%, and 52.83% of the phenotypic variation of flowering time among the mini-core collection in 3 yrs, respectively. In this study, the variants of *E1* and *E4* could not be incorporated into the association analysis due to the low frequency of *e1-nl* (0.5%) and *e4-SORE1* (1%) alleles in the mini-core collection. The extent of variation explained in this study was ∼10% lower than 62–66% reported by Zhai et al.[48] Even though the allele frequency of *e1-as* was relatively high (9%), *e1-as* was not significant in the multiple linear regression analysis. This is probably because the genetic effect of *e1-as* is smaller than that of *E2, E3,* and *two-component response regulator*-like gene. In the simple linear regression analysis, the *P*-value of *E2* ($2.0e^{-16}$), *E3* ($5.2e^{-14}$–$1.9e^{-13}$), and *two-component response regulator*-like gene ($7.7e^{-11}$–$1.8e^{-8}$) was considerably lower than that of *e1-as* (0.015–0.033) (Table 3). Further experiment using a larger population size is required to examine the remaining variation that could not be explained by the three genes with *e1-as*.

Table 4. DNA polymorphisms responsible for flowering time variation in the mini-core collection as revealed by multiple regression analysis

Data set	SNP No.	SNP name	Physical position[a]		Glyma ID	Description[b]	Alleles	Effect[c]	MAF	Parameters estimated by linear regression			Contribution rate (%)
			Chromosome	bp						β[d]	S.E.	P-value[e]	
2011													
	E2	rs124971350 (*e2*)	Chr10	45310798	Glyma.10G221500	GIGANTEA (E2)	A/T	Stop_gained	0.42	−5.21	1.51	2.5E-03**	21
	SNP379	*e3tr*[f]	Chr19	47641562	Glyma.19G224200	Phytochrome A (E3)	Large deletion	Loss of exon 4	0.17	−9.88	3.05	3.9E-03**	19
	SNP156	rs125308117	Chr12	5520945	Glyma.12G073900	Two-component response regulator-like	T/C	Stop_lost	0.31	11.24	3.53	4.5E-03**	12
2012													
	E2	rs124971350 (*e2*)	Chr10	45310798	Glyma.10G221500	GIGANTEA (E2)	A/T	Stop_gained	0.42	−5.44	1.47	1.3E-03**	22
	SNP379	*e3tr*[f]	Chr19	47641562	Glyma.19G224200	Phytochrome A (E3)	Large deletion	Loss of exon 4	0.17	−8.50	2.95	9.0E-03**	19
	SNP156	rs125308117	Chr12	5520945	Glyma.12G073900	Two-component response regulator-like	T/C	Stop_lost	0.31	9.65	3.42	1.0E-02*	10
2013													
	E2	rs124971350 (*e2*)	Chr10	45310798	Glyma.10G221500	GIGANTEA (E2)	A/T	Stop_gained	0.42	−5.36	1.66	4.0E-03**	20
	SNP379	*e3tr*[f]	Chr19	47641562	Glyma.19G224200	Phytochrome A (E3)	Large deletion	Loss of exon 4	0.17	−12.18	3.34	1.5E-03**	13
	SNP156	rs125308117	Chr12	5520945	Glyma.12G073900	Two-component response regulator-like	T/C	Stop_lost	0.31	12.69	3.86	3.5E-03**	19
	SNP349	rs127928573	Chr19	45183701	Glyma.19G194300	TERMINAL FLOWER 1	T/A	Missense_variant	0.09	11.23	5.11	3.9E-02*	0.5

MAF: minor allele frequency; S.E.: standard error.

[a] Physical position on Gmax275.

[b] Gene description was obtained from Phytozome 12.

[c] Effect to gene function annotated by snpEff. Effect of AMPL1040314 was defined by manually.

[d] Standardized regression coefficients.

[e] Adjusted P-value was obtained from multivariate models days to flowering and genotype as covariates. Signification codes: '**' 0.01 '*' 0.05.

[f] Large deletion on E3 estimated by coverage of four amplicon on 4th exon.

The other five genes, namely, WD repeat-containing protein 61 (*Glyma.17G052100*), MADS-box protein (*Glyma.16G200700*), PhyB (*Glyma.15G140000*), two-component response regulator-like gene (*Glyma.U034500*), and *FT2a/GmFTL3* (*Glyma.16G150700*), were significant in the simple linear regression analysis ($P < 0.0001$) but not significant in the multiple linear regression analysis (Table 3 and Supplementary Table S8). Variants that differ between Enrei and Peking can be used to confirm allele effect on flowering time using the phenotypic data of CSSLs.[31] Peking had a novel frameshift variant (Chr17_3955763) in WD repeat-containing protein 61 (*Glyma.17G052100*), two missense variants (rs389022394 and rs126888609) in MADS-box protein (*Glyma.16G200700*), and one frameshift variant (Chr15_11442094) in PhyB (*Glyma.15G140000*), and no variant in Enrei (Supplementary Table S8). However, no flowering time QTL has been reported to Chr17, Chr16, and Chr15; these genes may not be involved in flowering time regulation under the evaluation conditions of CSSLs.[31]

It was the stop-gain allele *E2* that showed the highest association with flowering time. The effect of this variant promotes flowering about 5 days (Table 4). Watanabe et al.[11] reported that the difference in days to flowering between *E2/E2* and *e2/e2* was ~9 days, which is consistent with the result of this study. The next strong association with flowering time was observed at *E3*. The *e3-tr* allele (Horosy-*e3*) has been reported to promote flowering for ~17 days,[13] but it was estimated as 9–13 days in this study. The smaller estimation at *E3* can be explained by the absence of *e3-Mo* allele. As there are only two accessions, PGC019 and PGC042 (Supplementary Table S9), the *e3-Mo* allele could not be included in the association analysis.

The effect of the missense variant (rs127928573, Arg166Trp) of *Dt1* was detected only for 2013 data set; it delayed flowering by ~11 days compared with that by the *Dt1* allele (Table 4). *Dt1* is reported as the locus strongly associated with days to maturity and plant height.[49] Zhang et al.[49] identified the *Dt1* gene at 18.6-kb upstream of the peak SNP, which was associated with days to maturity and plant height. *Dt1* plays a primary role in not only stem termination but also floral transition.[50,51] As no visible influence on the flowering time has been reported with *dt1* VIGS-induced suppression,[34] the detected SNP on *Dt1* in this study suggests the presence of other gene in the surrounding region related to the flowering time.

The effect of stop-lost variant (rs125308117) in the *two-component response regulator*-like gene (*Glyma.12G073900*) was significant (P = 4.5 e^{-3} in 2011, P = 1.0 e^{-2} in 2012, P = 3.5 e^{-3} in 2013), and the plant flowers ~10–13 days later. Involvement of the *two-component response regulator*-like gene in flowering time has been reported in *Arabidopsis* and rice; it may be functionally preserved as a flowering time-related gene in soybean. Williams 82 (reference genome) has C-terminal truncated protein as described above, whereas the rs125308117 variant has longer amino acid sequence and allele effect of delayed flowering for 4.7 days (Table 4). Although *Glyma.12G073900* is located near a previously reported QTL as *qFT-H* (nearest marker: Satt442 on Chr12: 6,390,806–6,391,062) in RILs with the *E1* allele from the cross between Tokei 780 and the *soja* accession Hidaka 4,[44] the allele type of *Glyma.12G073900* in both accessions is unknown. The genomic region surrounding *Glyma.12G073900* has been reported to include flowering time QTL *qDFF-Gm12* in CSSLs.[31] *Glyma.12G073900* of Peking (PGC084) is the stop-lost type (longer protein), whereas that of Enrei (PGC025) is reference type (truncated protein). Similar to the present study, Peking allele delayed flowering by ~3.7 days (LOD score is 36.3, flanking markers: C12-BARC- 015603-02006 and s024200450).[31] These data suggest that *Glyma.12G073900* is one of the candidate gene for *qDFF-Gm12*.

4. Conclusions

Flowering time and maturity are the most important factors affecting adaptability and yield. To increase the yield of soybean, it is necessary to control flowering time at an appropriate time using a combination of flowering time-related genes or alleles. Preparing a catalogue of flowering time-related genes makes it possible to freely combine alleles with various effects using the DNA markers. Our results indicate that novel alleles and accessions with such novel alleles can be rapidly detected using the AmpliSeq technology. Although multiple defective alleles were identified, we could not include all of them in the association study of flowering time due to low allele frequency. Nevertheless, the variants detected in this study could explain 51.1–52.3% of the flowering time variation in the soybean mini-core collection. These variants consisted of a novel two-component response regulator gene besides known flowering time-related genes. Therefore, the AmpliSeq technology is useful for discovering novel variants in the target genes.

Data availability

All sequences analysed in this study have been deposited in the DDBJ database under the BioProject Accession number: PRJDB7633.

Acknowledgements

We thank Naoki Kamei and Osamu Takahashi (Thermo Fisher Scientific, Life Technologies, Japan Ltd.) for sequencing using Ion S5 XL and offering technical support in using the Ion S5 sequencer. We also thank Yuko Yokota (NICS, NARO) and Takashi Sayama (NICS, NARO) for providing the materials necessary for the study. This study was supported in part by the Special Scheme Project on Advanced Research and Development for Next-Generation Technology from the Ministry of Agriculture, Forestry and Fisheries of Japan.

Accession number

All sequences analyzed in the present study have been deposited in the DDBJ database under the BioProject Accession number: PRJDB7633.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Cober, E.R. and Morrison, M.J. 2010, Regulation of seed yield and agronomic characters by photoperiod sensitivity and growth habit genes in soybean, *Theor. Appl. Genet.*, **120**, 1005–12.
2. Liu, W.X., Kim, M.Y., Van, K., et al. 2011, QTL identification of yield-related traits and their association with flowering and maturity in soybean, *J. Crop Sci. Biotechnol.*, **14**, 65–70.
3. Tsubokura, Y., Watanabe, S., Xia, Z., et al. 2014, Natural variation in the genes responsible for maturity loci *E1*, *E2*, *E3* and *E4* in soybean, *Ann. Bot.*, **113**, 429–41.
4. Bernard, R.L. 1971, Two major genes for time of flowering and maturity in soybean, *Crop Sci.*, **11**, 242–4.
5. Buzzell, R.I. 1971, Inheritance of a soybean flowering response to fluorescent daylength conditions, *Can. J. Genet. Cytol.*, **13**, 703–7.
6. Buzzell, R.I. and Voldeng, H.D. 1980, Inheritance of insensitivity to long daylength, *Soyb. Genet. Newsl.*, **7**, 26–9.
7. Cober, E.R. and Voldeng, H.D. 2001, A new soybean maturity and photoperiod-sensitivity locus linked to E1 and T, *Crop Sci.*, **41**, 698–701.

8. Kong, F.J., Nan, H.Y., Cao, D., et al. 2014, A new dominant gene E9 conditions early flowering and maturity in soybean, *Crop Sci.*, **54**, 2529–35.
9. Samanfar, B., Molnar, S.J., Charette, M., et al. 2017, Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean, *Theor. Appl. Genet.*, **130**, 377–90.
10. Xia, Z., Watanabe, S., Yamada, T., et al. 2012, Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering, *Proc. Natl Acad. Sci. USA*, **109**, E2155–64.
11. Watanabe, S., Xia, Z., Hideshima, R., et al. 2011, A map-based cloning strategy employing a residual heterozygous line reveals that the *GIGANTEA* gene is involved in soybean maturity and flowering, *Genetics*, **188**, 395–407.
12. Watanabe, S., Hideshima, R., Xia, Z., et al. 2009, Map-based cloning of the gene associated with the soybean maturity locus *E3*, *Genetics*, **182**, 1251–62.
13. Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K. and Abe, J. 2008, Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene, *Genetics*, **180**, 995–1007.
14. Xu, M., Xu, Z., Liu, B., et al. 2013, Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean, *BMC Plant Biol.*, **13**, 91.
15. Tsubokura, Y., Matsumura, H., Xu, M., et al. 2013, Genetic variation in soybean at the maturity locus *E4* is involved in adaptation to long days at high latitudes, *Agronomy*, **3**, 117–34.
16. Kong, F., Liu, B., Xia, Z., et al. 2010, Two coordinately regulated homologs of *FLOWERING LOCUS T* are involved in the control of photoperiodic flowering in soybean, *Plant Physiol.*, **154**, 1220–31.
17. Zhao, C., Takeshima, R., Zhu, J., et al. 2016, A recessive allele for delayed flowering at the soybean maturity locus E9 is a leaky allele of FT2a, a FLOWERING LOCUS T ortholog, *BMC Plant Biol.*, **16**, 20.
18. Takeshima, R., Hayashi, T., Zhu, J., et al. 2016, A soybean quantitative trait locus that promotes flowering under long days is identified as FT5a, a FLOWERING LOCUS T ortholog, *J. Exp. Bot.*, **67**, 5247–58.
19. Ion AmpliSeq Designer. https://ampliseq.com/help/startDesign.action (25 March 2019, date last accessed)
20. AmpliSeq protocol. https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0006735_AmpliSeq_DNA_RNA_LibPrep_UG.pdf.
21. Kaga, A., Shimizu, T., Watanabe, S., et al. 2012, Evaluation of soybean germplasm conserved in NIAS genebank and development of mini core collections, *Breed. Sci.*, **61**, 566–92.
22. Chankaew, S., Isemura, T., Naito, K., et al. 2014, QTL mapping for salt tolerance and domestication-related traits in *Vigna marina* subsp. *oblonga*, a halophytic species, *Theor. Appl. Genet.*, **127**, 691–702.
23. Schmutz, J., Cannon, S.B., Schlueter, J., et al. 2010, Genome sequence of the palaeopolyploid soybean, *Nature*, **463**, 178–83.
24. Cingolani, P., Platts, A., Wang, L.L., et al. 2012, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3, *Fly (Austin)*, **6**, 80–92.
25. Sherry, S.T., Ward, M.H., Kholodov, M., et al. 2001, dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res.*, **29**, 308–11.
26. R Core Team. 2017, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. https://www.R-project.org/.
27. *Troubleshooting Guides*. https://www.ampliseq.com/help/troubleshooting.action.
28. Zhou, Z., Jiang, Y., Wang, Z., et al. 2015, Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean, *Nat. Biotechnol.*, **33**, 408–14.
29. Wu, F., Zhang, X., Li, D. and Fu, Y.F. 2011, Ectopic expression reveals a conserved PHYB homolog in soybean, *PLoS One*, **6**, e27737.
30. Sun, H., Jia, Z., Cao, D., et al. 2011, GmFT2a, a soybean homolog of FLOWERING LOCUS T, is involved in flowering transition and maintenance, *PLoS One*, **6**, e29238.
31. Watanabe, S., Shimizu, T., Machita, K., et al. 2018, Development of a high-density linkage map and chromosome segment substitution lines for Japanese soybean cultivar Enrei, *DNA Res.*, **25**, 123–36.
32. Kojima, S., Takahashi, Y., Kobayashi, Y., et al. 2002, *Hd3a*, a rice ortholog of the *Arabidopsis FT* gene, promotes transition to flowering downstream of Hd1 under short-day conditions, *Plant Cell Physiol.*, **43**, 1096–105.
33. Ogiso-Tanaka, E., Matsubara, K., Yamamoto, S., et al. 2013, Natural variation of the RICE FLOWERING LOCUS T 1 contributes to flowering time divergence in rice, *PLoS One*, **8**, e75959.
34. Liu, B., Watanabe, S., Uchiyama, T., et al. 2010, The soybean stem growth habit gene *Dt1* is an ortholog of *Arabidopsis terminal flower1*, *Plant Physiol.*, **153**, 198–210.
35. Guo, X., Zhao, Z., Chen, J., et al. 2006, A putative CENTRORADIALIS/TERMINAL FLOWER 1-like gene, Ljcen1, plays a role in phase transition in *Lotus japonicus*, *J. Plant Physiol.*, **163**, 436–44.
36. Foucher, F., Morin, J., Courtiade, J., et al. 2003, DETERMINATE and LATE FLOWERING are two TERMINAL FLOWER1/CENTRORADIALIS homologs that control two distinct phases of flowering initiation and development in pea, *Plant Cell*, **15**, 2742–54.
37. Bradley, D., Ratcliffe, O., Vincent, C., Carpenter, R. and Coen, E. 1997, Inflorescence commitment and architecture in *Arabidopsis*, *Science*, **275**, 80–3.
38. Mimida, N., Goto, K., Kobayashi, Y., et al. 2001, Functional divergence of the TFL1-like gene family in *Arabidopsis* revealed by characterization of a novel homologue, *Genes Cells*, **6**, 327–36.
39. Bradley, D., Carpenter, R., Copsey, L., Vincent, C., Rothstein, S. and Coen, E. 1996, Control of inflorescence architecture in *Antirrhinum*, *Nature*, **379**, 791–7.
40. Zhang, H., Miao, H., Li, C., et al. 2016, Ultra-dense SNP genetic map construction and identification of SiDt gene controlling the determinate growth habit in *Sesamum indicum* L, *Sci. Rep.*, **6**, 31556.
41. Nakamichi, N., Kita, M., Niinuma, K., et al. 2007, *Arabidopsis* clock-associated pseudo-response regulators PRR9, PRR7 and PRR5 coordinately and positively regulate flowering time through the canonical CONSTANS-dependent photoperiodic pathway, *Plant Cell Physiol.*, **48**, 822–32.
42. Kwon, C.T., Koo, B.H., Kim, D., Yoo, S.C. and Paek, N.C. 2015, Casein kinases I and 2α phosphorylate *Oryza sativa* pseudo-response regulator 37 (OsPRR37) in photoperiodic flowering in rice, *Mol. Cells*, **38**, 81–8.
43. Turner, A., Beales, J., Faure, S., Dunford, R.P. and Laurie, D.A. 2005, The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in 799 barley, *Science*, **310**, 1031–4.
44. Lu, S.J., Li, Y., Wang, J.L., et al. 2016, Identification of additional QTLs for flowering time by removing the effect of the maturity gene *E1* in soybean, *J. Integr. Agric.*, **15**, 42–9.
45. Koo, S.C., Bracko, O., Park, M.S., et al. 2010, Control of lateral organ development and flowering time by the *Arabidopsis thaliana* MADS-box gene AGAMOUS-LIKE6, *Plant J.*, **62**, 807–16.
46. Yoo, S.K., Wu, X., Lee, J.S. and Ahn, J.H. 2011, AGAMOUS-LIKE 6 is a floral promoter that negatively regulates the FLC/MAF clade genes and positively regulates FT in *Arabidopsis*, *Plant J.*, **65**, 62–76.
47. van Nocker, S. and Ludwig, P. 2003, The WD-repeat protein superfamily in *Arabidopsis*: conservation and divergence in structure and function, *BMC Genomics*, **4**, 50.
48. Zhai, H., Lü, S., Wang, Y., et al. 2014, Allelic variations at four major maturity *E* genes and transcriptional abundance of the *E1* gene are associated with flowering time and maturity of soybean cultivars, *PLoS One*, **9**, e97636.
49. Zhang, J.P., Song, Q.J., Cregan, P.B., et al. 2015, Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm, *BMC Genomics*, **16**, 217.
50. Bernard, R. 1972, Two genes affecting stem termination in soybeans, *Crop Sci.*, **12**, 235–9.
51. Tian, Z., Wang, X., Lee, R., et al. 2010, Artificial selection for determinate growth habit in soybean, *Proc. Natl Acad. Sci. USA*, **107**, 8563–8.
52. Fan, C., Hu, R., Zhang, X., et al. 2014, Conserved CO-FT regulons contribute to the photoperiod flowering control in soybean, *BMC Plant Biol.*, **14**, 9.
53. Cao, D., Takeshima, R., Zhao, C., et al. 2017, Molecular mechanisms of flowering under long days and stem growth habit in soybean, *J. Exp. Bot.*, **68**, 1873–84.