Published March 28, 2014

ORIGINAL RESEARCH

Rapid Identification of Alleles at the Soybean Maturity Gene E3 using genotyping by Sequencing and a Haplotype-Based Approach

Aurélie Tardivel, Humira Sonah, François Belzile, and Louise S. O'Donoughue*

Abstract

In eastern Canada, earliness is an important trait for soybean given the short growing season. The aim of this work was to develop tools for breeders to rapidly identify alleles present in their germplasm at the recently cloned maturity locus E3 (*GmPhyA3*). The tremendous throughput of modern DNA sequencing technology has allowed the use of genotyping by sequencing (GBS) approaches to identify and genotype thousands of single nucleotide polymorphisms (SNPs) across the entire genome. We have used a GBS protocol and SNP-calling pipeline optimized for soybean to characterize 53 near-isogenic lines (NILs) contrasting for maturity loci. Results obtained clearly showed the suitability of GBS to provide a dense SNP coverage and very accurate information on the location and size of introgressed regions. We then developed a GBS haplotype method to characterize 91 plant introductions (PIs) as well as a set of 305 lines representative of the Eastern Canadian germplasm for their allelic status at the *GmPhyA3* gene. Six distinct haplotypes in and around the E3 locus were observed. Subsequent tests on two genotypes per haplotype (PCR test for a previously reported allele, sequencing entire gene), and validation on a subset of lines, allowed to determine that each of these corresponded to a different allele of this gene. We found that the functional allele *E3Ha* and the loss of function allele *e3-tr* were the two most prevalent in the Eastern Canadian germplasm, while the *e3-fs* allele was found at low frequency and *e3-ns* was absent. These results show that this approach is a powerful method for rapid allelic characterization, and its application to other maturity genes will be useful for breeding purposes.

SOYBEAN [*Glycine max* (L.) Merr.] is an economically important crop grown on a global scale. Cultivars belong to different maturity groups (MGs 000–X), which indicate adaptation to growth under a wide range of latitudes (Cober et al., 2001). Control of maturity is an important factor for geographic adaptation and optimization of agronomic performance. In Eastern Canada, earliness is an essential trait given the short growing season. However, the development of very early cultivars is limited by a negative correlation between high yield and early maturity (Cober and Morison, 2010). Understanding the genetic basis of maturity is key to the further expansion of soybean production in Canada. Maturity is controlled both by temperature and photoperiod (Cober et al., 1996). Soybean is basically a short-day plant, and the ability to grow soybeans under the long days typical of summer in relatively high latitudes is due to the presence of photoperiod insensitivity genes (Cober et al., 2001). However, little is known about the exact diversity and frequency of the alleles at these genes in the Canadian soybean germplasm.

Nine loci have been reported to control flowering and maturity in soybean: E1 on Linkage Group (LG) C2 (Gm06; Bernard, 1971; Molnar et al., 2003), E2 on LG O (Gm10; Bernard, 1971; Akkaya et al., 1995; Cregan et al., 1999), E3 on LG L (Gm19; Buzzell, 1971; Molnar et al., 2003), E4 on LG I (Gm20; Buzzell and Voldeng, 1980; Molnar et al., 2003), E5 (McBlain and Bernard, 1987), E6 (Bonato and

Published in The Plant Genome 7
doi: 10.3835/plantgenome2013.10.0034

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

A.Tardivel, H.Sonah, and F. Belzile, Département de Phytologie and Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec, QC, Canada G1V 0A6; L.S. O'Donoughue, CÉROM, Centre de recherche sur les grains inc., 740, chemin Trudeau, Saint-Mathieu-de-Beloeil, QC J3G 0E2 Canada. Received 18 Oct. 2013. *Corresponding author (louise.odonoughue@cerom.qc.ca).

Abbreviations: GBS, genotyping by sequencing; GM, *Glycine max*; LG, linkage group; MAF, minor allele frequency; MG, maturity group; NIL, near-isogenic line; PAV, presence–absence variant; Phy, phytochrome locus label; SNP, single nucleotide polymorphism.

Vello, 1999), E7 on LG C2 (Gm06; Cober and Voldeng, 2001; Molnar et al., 2003), E8 on LGs C1 (Gm04; Cober et al., 2010), and J which has not been mapped (Ray et al., 1995). Recent work has identified the genes underlying four of these maturity loci (E1 to E4). Using a positional cloning approach, the E1 locus has been found to correspond to the *E1* gene (Glyma06 g23040; Xia et al., 2012a) that encodes a protein thought to function as a flowering repressor (Thakare et al., 2010, 2011). The genes corresponding to the E2 and E3 loci were both identified using positional cloning on residual heterozygous lines (Watanabe et al., 2009, 2011). The E2 is a gene orthologous to the *Arabidopsis* flowering time gene GIGANTEA (*GmGIa*, Glyma10 g36600), and E3 has been found to correspond to a phytochrome A gene (*GmPhyA3*, Glyma19 g41210). The E4 locus has also been identified as a phytochrome A gene (*GmPhyA2*, Glyma20 g22160) using a candidate gene approach (Liu et al., 2008). These efforts resulted in major advances in the understanding of the molecular basis of maturity and a molecular network regulating flowering time in soybean has been proposed (Xia et al., 2012b). Different natural and independent mutations contributing to maturity date variation have been reported for the E1 locus (*E1*, *e1-nl*, *e1-fs*, *e1-as*; Xia et al., 2012a), the E2 locus (*E2*, *e2*; Watanabe et al., 2011; Shin and Lee, 2012), the E3 locus (*E3Ha*, *E3Mi*, *e3-tr*, *e3Mo*, *e3-fs*, *e3-ns*; Watanabe et al., 2009; Shin and Lee, 2012; Xu et al., 2013) and the E4 locus (*E4Ha*, *e4* [*SORE-1*], *e4-oto*, *e4-tsu*, *e4-kam*, *e4-kes*; Liu et al., 2008; Tsubokura et al., 2013). Allele-specific markers have been proposed for some of these alleles and are now available for breeding purposes (Tsubokura et al., 2013; Shin and Lee, 2012; Xu et al., 2013). However, in the majority of cases, the reported alleles have been found in East Asian landraces or cultivars and, for the E4 locus, the distribution is limited to relatively small geographical regions (Tsubokura et al., 2013). Consequently, information on the allelic status at these genes for different and larger germplasm panels is limited. Indeed, the alleles that have been reported may not represent all of the existing diversity, and there is a need to complete this information in larger germplasm collections.

The narrow genetic base of North American soybean breeding programs is well documented, with 10 ancestors estimated to account for 80% of the genes, and only five cultivars accounting for ~65% of all genes in the northern germplasm (Gizlice et al., 1994). Soybean production in Canada started in Southern Ontario in the early 20th century from a limited number of PIs. During the period between 1920 and 1970, breeding efforts were made to identify and develop soybean cultivars adapted to Canada and, cultivars released during this period were generally derived from the selection of Chinese PIs (Beversdorf et al., 1995). In the 1970s, Swedish germplasm contributed to the earlier maturing Canadian germplasm (Lambert, 1979) as these lines were both very early maturing (MG 000) and tolerant to midseason cool temperatures The so-called "Fiskeby" series was developed from different crosses using parental lines from the sea of Okhotsk region of Northern Asia (Holmberg, 1973). Canadian breeders have used the Swedish lines and elite Fiskeby varieties, and many of the cultivars released since 1970 have known relationships to 840-7-3 or Fiskeby V (Cober and Voldeng, 2012; Beversdorf et al., 1995; Voldeng et al., 1997). Incorporation of genes from the Swedish lines into the Canadian germplasm contributed to the expansion of soybean production north and east in Ontario, and into Quebec and the Maritime provinces of Canada (Voldeng et al., 1997).

Though pedigree information can indicate sources of early maturity genes in Canadian germplasm, the alleles present at the different maturity loci remain unknown. To obtain comprehensive information of this nature, sequencing the entire gene in all lines can be performed. However, this strategy can be expensive and time consuming. An alternative approach would be to use SNP haplotypes to distinguish different alleles of a given gene. For this to be feasible, a dense SNP coverage is required to ensure that a sufficient number of SNPs within and around the gene of interest can define distinct haplotypes. The tremendous throughput of modern DNA sequencing technology has allowed the use of GBS approaches to identify and genotype thousands of SNPs across the entire genome (Elshire et al., 2011; Sonah et al., 2013). Here we demonstrate that SNP haplotypes can be used to rapidly identify the alleles present in a given germplasm collection at the recently cloned maturity gene *GmPhyA3*.

MATERIALS AND METHODS

Plant Material

Three sets of soybean genotypes belonging to diverse origins and MGs were used in this study (Supplementary Table S1). Set A is composed of 53 NILs contrasting for various E loci (E1, E2, E3, E4, and E7) and developed in different genetic backgrounds using five recurrent parents (Molnar et al., 2003; Elroy Cober, personal communication, 2011). These lines were graciously provided by E. Cober (Agriculture and Agri-Food Canada, Ottawa). Set B contained 91 early PI lines obtained from the National Plant Germplasm System (USDA-ARS) and were selected on the basis of their MG (0 to 000) and diverse geographical origins (Europe, China, Japan, Russia, and North America). This collection also includes lines from the Fiskeby series. Set C consisted of 305 lines, 302 of which were representative of the Eastern Canadian soybean germplasm (MGs II to 000). Three additional late accessions from MG VI to VIII (PI 159925, Parana [PI 628879] and Paranagoiana [PI 628880]) were included in this set because they were of interest at other loci related to maturity. Seeds of several lines of Set C were graciously provided by E. Cober (Agriculture and Agri-Food Canada, Ottawa) and I. Rajcan (University of Guelph, Guelph). For lines of Sets B and C, the allelic status at maturity genes was largely unknown.

SNP Genotyping by Sequencing

DNA was extracted from 100 mg of fresh young leaves using the Qiagen DNeasy 96 Plant kit (Toronto, Canada) following the manufacturer's protocol. DNA

Table 1. Detailed results of genotyping-by-sequencing single nucleotide polymorphism (SNP) analysis for each set.

Set	Multiplexing (lines/lane)	Mean number of reads/line	Mapped reads	SNPs with MAF† > 0.002	SNPs with MAF > 0.02	Mean depth (reads/SNP)
			%			
A	48	2,963,554.5	86.3	6,739	6,368	25.8
B	48	3,232,322.1	88.0	12,976	11,594	27.3
C	96	1,597,415.9	87.5	13,655	8,953	13.0
A + B + C	–	2,090,025.8	87.5	15,910	10,157	17.4

†MAF = minor allele frequency.

concentrations were normalized to 10 ng/μL and subsequently used for GBS library preparation. Libraries were prepared at the Plate-Forme d'Analyses Génomiques (Université Laval, Quebec City, QC, Canada) essentially as described in Elshire et al. (2011) using *Ape*KI as the restriction enzyme. Single-end sequencing (100 bp reads) was performed on an Illumina HiSeq2000 (McGill University-Genome Quebec Innovation Center, Montreal, QC, Canada) on two scales: 48 genotypes/sequencing lane for Sets A and B and 96 genotypes/sequencing lane for Set C. After sequencing, the IBIS genotyping-by-sequencing tool (IGST)-GBS pipeline (Sonah et al., 2013) was used to obtain a catalog of SNPs. After correction steps and elimination of residual heterozygotes, SNPs were then filtered to retain those with <20% missing data and a minor allele frequency (MAF) > 0.01. Finally, imputation of missing data was performed using fastPhase (Scheet and Stephens, 2006). Variant annotation to describe the SNP location (intergenic, intron, or exon) and predict their functional effect (synonymous, nonsynonymous) was performed by SnpEff (http://snpeff.sourceforge.net, verified 4 Mar. 2014). After imputation, a working data set with MAF 0.02 was kept.

Haplotype Characterization

Pairs of NILs contrasting for the E3 locus (*GmPhyA3* gene, Glyma19 g41210) were examined using the Flapjack graphical genotype visualization tool (Milne et al., 2010) to define the SNP landscape in and around E3. The SNP genotypes were observed in two pairs of NILs. L62–667 and Harosoy, OT89–1 and Maple Arrow, developed by introgression of *e3* alleles from T204 and PI196529 into Harosoy (*E3*) and Maple Arrow (*E3*), respectively (Voldeng and Saindon, 1991; Bernard et al., 1991), were compared.

A simple visual inspection of the distribution of SNP markers around the *GmPhyA3* gene in the two remaining sets (B and C) was used to define a minimal number of SNPs defining haplotypes. In addition to SNPs, the absence of reads flanking six *Ape*KI restriction sites located within a 13 kb region that is deleted in the *e3-tr* allele (Watanabe et al., 2009) was used to identify lines carrying this deletion. To define the genetic distance represented by the haplotype window at the E3 locus, we identified the closest neighboring markers available on the consensus map in SoyBase (http://www.soybase.org/, accessed 1 Sept. 2013; verified 4 Mar. 2014). Graphical haplotypes were prepared using GGT software (Van Berloo, 2008).

Allele Characterization

Distinct haplotypes shared by at least two lines were subjected to a detailed analysis. Two lines representative of each haplotype were first analyzed by testing for previously described alleles (*E3Mi*, *e3Mo*, *e3-tr*) using the PCR assays described by Watanabe et al. (2009). Because the alleles *e3-fs* and *e3-ns* had not yet been reported, we did not test their presence with the allele-specific markers recently proposed by Xu et al. (2013). In the case of the haplotype corresponding to the *e3-tr* allele, the PCR assay designed to test this deletion (Watanabe and al. 2009) was performed on all individuals of Sets A and B exhibiting an absence of reads in this region. In cases where haplotypes did not carry the hallmarks of known alleles, two lines from each haplotype were further analyzed by sequencing the entire *GmPhyA3* gene using six overlapping amplicons. Primers used for sequencing the entire gene were designed using NCBI Primer-BLAST. These primers are also detailed in the Supplementary Table S2. In some cases, nested PCR were used to increase specificity of primers. If a nonsynonymous mutation was observed, the presence of this polymorphism was tested by sequencing the specific amplicon in the remaining individuals sharing this haplotype. All sequencing work was performed at the Plate-Forme d'Analyses Génomiques, Université Laval. The PCR reactions were performed in a 20-μL volume containing 20 ng of genomic DNA, 0.5 μ*M* each of forward and reverse primers, 0.2 m*M* deoxyribonucleotide triphosphate, 1 × PCR buffer, and 0.5 unit of *Taq* DNA polymerase. Amplifications were performed as follows: 1 cycle (94°C for 3 min); 35 cycles (94°C denaturation for 45 sec, 60°C annealing for 45 sec, 72°C for 2 min); 1 cycle (72°C for 5 min).

RESULTS

SNP Genotyping

For the 449 individuals composing Sets A, B, and C, over 938 million 100-bp reads were obtained and analyzed to call SNPs both within each set and across all three sets simultaneously. As can be seen in Table 1, the mean number of reads/line was ~3 million for Sets A and B, while it was ~1.6 million for Set C, as the GBS libraries were sequenced less extensively (96 genotypes/lane) in the latter case than for Sets A and B (48 genotypes/lane). One genotype belonging to Set C was removed from the dataset, as it yielded a very low number of reads (2537). On average,

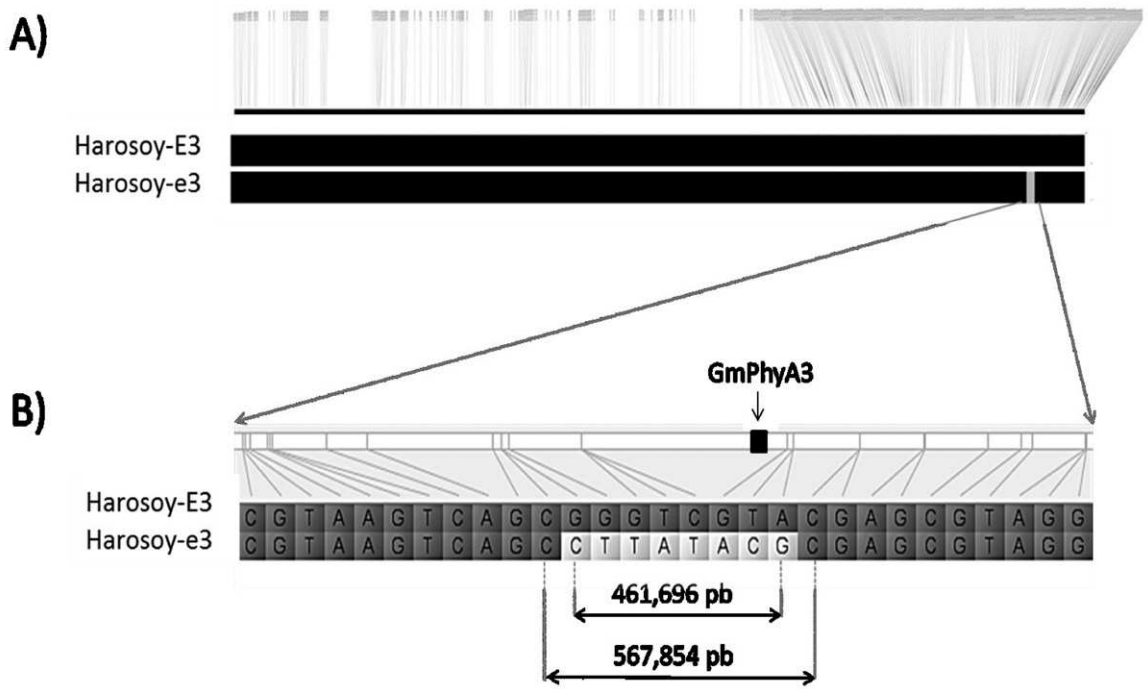

Figure 1. Harosoy isolines contrasting for single nucleotide polymorphisms (SNPs) in the vicinity of the E3 locus. (A) Physical distribution of 528 SNP markers found on chromosome Gm19 and localization of the introduced e3 region in the Harosoy background. Dark areas correspond to monomorphic markers, whereas grey corresponds to polymorphic markers. (B) Detailed view of the physical distribution of polymorphic markers found between the two isolines in comparison with the position of the GmPhyA3 gene.

87.5% of the reads were successfully mapped to a position in the reference genome and the number of SNPs that were successfully called from these data ranged from 6739 for Set A to 15,910 for the combined analysis (at MAF > 0.002). The mean number of reads supporting a SNP call (mean depth) was directly correlated to the scale of sequencing, ranging from 13 (96-plex) to ~26 (48-plex). When SNPs were derived from both sequencing scales as in the combined analysis, mean depth was 17.4 reads/SNP locus.

Using SnpEff, further analysis revealed that 55.4% of these SNPs resided in intergenic regions. Among the remaining SNPs, 14.0% were located in introns and 25.2% in exons. In coding regions, 53.7% of the SNPs induced nonsynonymous changes. For the ensuing work, a data set of 10,157 SNPs with a MAF > 0.02 was used.

Definition of the Size of the Introgressed Segments around E3 in NILs

To define the size of the genomic regions containing the introgressed *e3* alleles, we compared SNP genotypes in two pairs of NILs contrasting at this locus (Harosoy and Maple Arrow backgrounds). As can be seen in Fig. 1, of the 528 SNP markers located on Gm19 that were polymorphic among the broader collection of soybean lines, only eight differed between Harosoy and L62-667, and all of these were located within the vicinity of the *GmPhyA3* gene (47,511,095 to 47,520,052). None of the polymorphic SNPs were located within the gene itself, as the 100-bp reads flanking *Ape*KI sites located within the gene were identical in sequence. Based on this, we can estimate that the introgressed segment in L62-667 is at least 461,696 bp and at most 567,854 bp long. A similar analysis conducted between Maple Arrow and its isoline (OT89-1) revealed that 20 SNPs (all on Gm19) defined the extent of the introgressed region surrounding the E3 locus. The recombination breakpoints in this case were at least 529,517 bp, but no further than 665,622 apart (data not shown). Interestingly, the introgressed regions in L62-667, though of a similar size, did not overlap with the introgressed region in OT89-1.

In addition to the introgressed segment containing the *GmPhyA3* gene on Gm19 in the Harosoy isoline L62-667, six additional introgressed segments were found on chromosomes Gm01, Gm12, Gm16, and Gm20. Five of these regions were relatively small, ranging from 63 to 824 kb, while the largest introgressed segment (on Gm01) extended over 33 Mb.

Definition of Haplotypes in and around the E3 Locus

In total, four SNPs (Gm19_47,347,321, Gm19_47,558,095, Gm19_47,673,726, and Gm19_48,031,432) located near the *GmPhyA3* gene were found to be useful in defining distinct haplotypes within Sets B and C. A fifth informative polymorphism was obtained by analyzing the presence or absence of reads flanking three *Ape*KI sites located within the *GmPhyA3* gene. This presence–absence variant (PAV) is due to a 13-kb deletion that is diagnostic of the *e3-tr* allele. The two closest markers flanking the haplotype window were identified as BARC-021321-04035 (Gm19:47,258,164..47,259,043) and BARC-055773-13715 (Gm19:48,037,252..48,037,734). These markers mapped respectively at 86.2 cM and 90.4 cM, which indicates that the haplotype window was at most 4.2 cM in length.

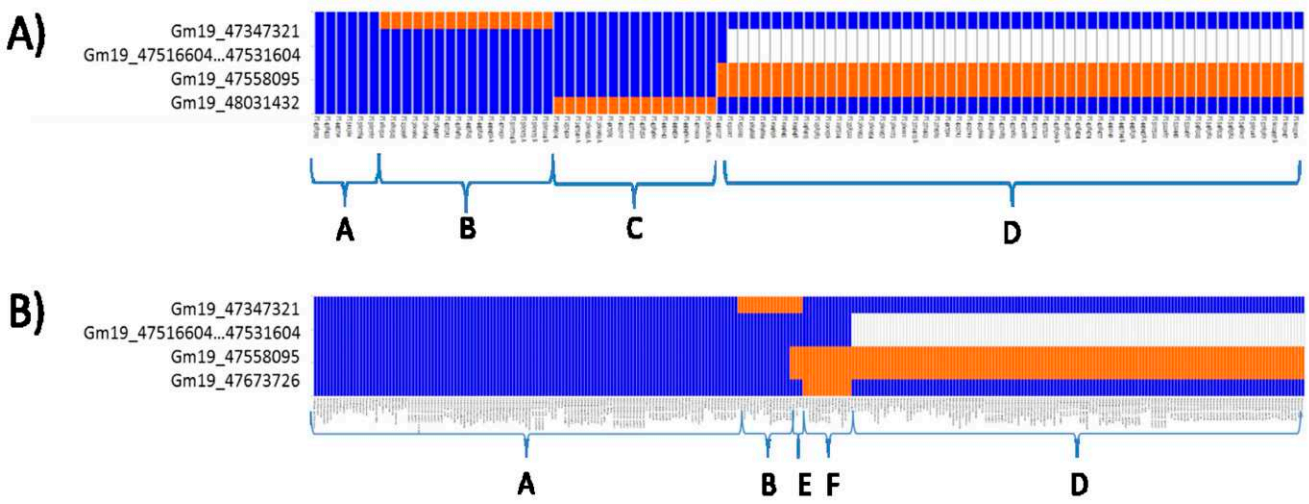

Figure 2. Graphical representation of single nucleotide polymorphism (SNP) haplotypes for (A) 91 genotypes composing Set B, (B) 304 genotypes composing Set C. Each vertical bar corresponds to one individual, each horizontal line corresponds to one marker. Blue represents the allele present in the reference genome (Williams 82) and orange the alternate allele. White is used to indicate an absence of reads mapping in the GmPhyA3 gene within a 13 kb segment that is deleted in the e3-tr allele (Watanabe et al., 2009).

Within Set B, three of the four SNPs (Gm19_47,347,321, Gm19_47,558,095, and Gm19_48,031,432), in addition to the PAV, allowed us to define a total of four haplotypes (A, B, C, and D) shared by at least two lines (Fig. 2A). A single line (PI 442037) exhibited an additional haplotype, but such haplotypes found in a single line were not investigated further. The fourth SNP (Gm19_47,673,726) was not used in this set, as it was monomorphic among all individuals. Similarly, in Set C, a different trio of SNPs (Gm19_47,347,321, Gm19_47,558,095, and Gm19_47,673,726) and the PAV were used to define five haplotypes (Fig. 2B). Of these haplotypes, three (A, B, and D) had been observed in Set B, and two (E and F) were novel. Again, one of the four SNPs (Gm19_48,031,432) was not found to be helpful in defining haplotypes in this collection of lines.

The frequency of the different haplotypes was quite different in the two sets of soybean lines (cf. Table 2). Three of the haplotypes (A, B, and D) were found in both sets, while the other three (C, E, and F) were observed in only one of the two sets. Haplotype A was found in six lines (out of 91, 6.6%) in Set B, while it was much more prevalent (128 of 304 lines, 41.9%) in Set C. Overall, haplotype D was the next most frequent haplotype, as it was found in 53 (58.2%) and 139 (45.7%) of the lines comprising Sets B and C, respectively. Haplotype B was found in an equal number of lines (16) in both sets representing 17.6% of Set B, but only 5.3% of Set C. Haplotype C was observed in 15 lines (16.5%) within Set B, but was totally absent from Set C. Conversely, Haplotypes E and F were unique to Set C and represented 1% (3 lines) and 3.9% (12 lines), respectively, of this collection of lines.

Characterization of Alleles at the E3 Locus

To determine if the different haplotypes defined above corresponded to different alleles of the *GmPhyA3* gene, we first examined the haplotypes found in lines carrying previously described alleles. The cultivar Harosoy is known to carry one of the functional alleles of *GmPhyA3* (*E3Ha*) and

Table 2. Frequency of the different haplotypes and the corresponding alleles in Sets B and C.

Haplotypes	Alleles	Set B (PI collection)	Set C (Canadian germplasm)
		%	
A	*E3Ha*	6.6	41.9
B	*e3-fs*	17.6	5.3
C	*e3-ns*	16.5	0.0
D	*e3-tr*	58.2	45.7
E	*E3Mi*	0.0	1.0
F	*E/e3p.Thr832Ala*	0.0	3.9

this line (present in Set C) was found to have Haplotype A. All lines sharing this haplotype were negative in PCR assays capable of detecting features diagnostic of the other three previously characterized alleles (*E3Mi*, *e3Mo*, and *e3-tr*). Finally, the entire *GmPhyA3* gene was sequenced in two of the Haplotype A lines and the resulting sequence was identical to the one reported for *E3Ha*. Thus, we conclude that Haplotype A corresponds to the *E3Ha* allele.

As all lines sharing haplotype B were negative in PCR assays for the three previously characterized alleles (*E3Mi*, *e3Mo* and *e3-tr*), the entire *GmPhyA3* gene was sequenced in two of the haplotype B lines; Traff is a Swedish cultivar and PI 180525 a German line released in 1948. Because of the lack of detailed or comprehensive pedigree information, it is not possible to know if these two lines are related. The results showed that these two lines shared a 1-bp deletion at position 2525 in Exon 1 (based on the Harosoy *GmPhyA3* sequence, accession AB468154) resulting in a frame shift mutation and a premature stop codon (Fig. 3). This allele corresponded to the *e3-fs* allele, another recently described allele (Xu et al., 2013). A targeted sequencing analysis was performed to confirm the presence of this mutation in the 14 remaining genotypes sharing this haplotype in Set B

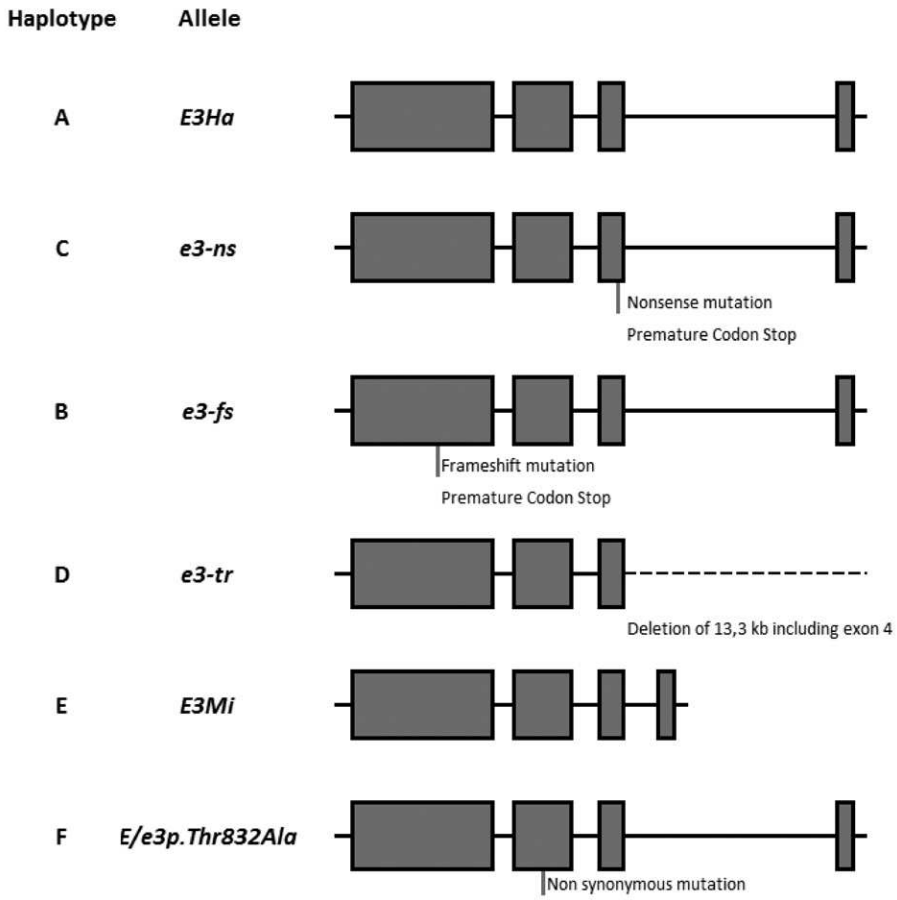

Figure 3. Alleles at the GmPhyA3 gene and the corresponding haplotypes. For each of the six haplotypes (A through F) observed in at least two lines, a different allele of the GmPhyA3 gene was identified. Two of these (E3Ha and E3Mi) result in a functional protein and a late maturity phenotype. Three alleles (e3-fs, e3-nl and e3 Tr) result in a truncated protein and an early maturity phenotype. The final allele (E/e3p.Thr832Ala) is characterized by an amino acid substitution in the second exon and it is not presently known if this results in a change in the phenotype.

and 2 genotypes in Set C; in all cases, the point mutation characteristic of the *e3-fs* allele was observed.

Haplotype C consisted of 15 lines also found to be negative for the three previously characterized alleles. The *GmPhyA3* gene was sequenced for two of the haplotype C lines in Set B; Fiskeby V is a Swedish cultivar and PI 257429 a German line with PI assigned in 1959. Because of the lack of pertinent pedigree information, it is also not possible to know if these two lines are related. Both lines shared a point mutation at position 5077 in Exon 3 (Fig. 3). This allele is characterized by a nucleotide substitution (C to T), introducing a premature stop codon (Gly to stop), thus resulting in a shorter protein (1046 amino acids), and it corresponds to the recently reported *e3-ns* allele (Xu et al., 2013). Targeted sequencing of the segment containing this mutation in the 13 remaining haplotype B lines confirmed the presence of this mutation among all these lines.

The lines exhibiting haplotype D were all characterized by an absence of reads in the 3′ end (Intron 3 and Exon 4) of the *GmPhyA3* gene consistent with the known *e3-tr* allele (Fig. 3). The PCR amplification tests of the 53 lines carrying this haplotype in Set B confirmed in all cases the presence of this deletion.

Allele-specific PCR amplification tests for the *E3Mi* allele indicated that the three lines of Set C exhibiting haplotype E (Fig. 3) carried the *E3Mi* allele. Further targeted sequencing confirmed the presence of this late-maturity allele.

Finally, the sequencing of two haplotype F lines in Set C (OAC Thames and OAC Oxford) and subsequent targeted sequencing of the 10 remaining lines, showed a common nucleotide substitution (from A to G) at position 4042 in the second exon (Fig. 3). This novel point mutation (named *E/e3p.Thr832Ala*) leads to a nonsynonymous amino acid substitution (Thr for Ala) at position 832 of the protein (sequence KF516231). Alignment with phytochrome A proteins from other species indicated that this mutation occurs in a stretch of amino acids that are not highly conserved among orthologous proteins (data not shown).

DISCUSSION

Dense SNP Coverage allows for a More Precise Characterization of NILs

The objectives of this portion of the work were twofold. First, we wanted to evaluate the suitability of the GBS method to provide a dense SNP coverage on pairs of isolines, genetic materials in which properly mapped SNPs are

expected to cluster in a few regions of introgression. This first objective was clearly achieved as, in isoline L62-667, all of the markers that were fixed for an allele differing from that contributed by the recurrent parent (Harosoy) were clearly clustered in seven introgressed regions ranging in size from 63 kb to 33 Mb. The size of this last introgressed segment is clearly exceptional, as it was the only one to exceed 1 Mb in size. The fact that it lies entirely in a heterochromatic pericentromeric region in which little recombination occurs could explain the size of this residual segment. Indeed, this segment is estimated to cover a genetic distance of only 5 cM (SoyBase).

The second objective was to more precisely define the extent of the introgressed segments in the isolines. Previously, Molnar et al. (2003) mapped the E3 locus on LG L (Gm19) using 430 SSR markers on the Harosoy NILs. They found that the E3 locus was mapped to a 32.4 cM interval, between Satt099 and the end of the LG; an interval of 5.7 Mb (SoyBase). In the present study, we were able to define the introgressed region containing the E3 locus much more precisely, as the introgressed segment was found to be greater than 462 kb but smaller than 568 kb. This represents a tenfold improvement in the precision of the size estimate of the introgressed segment. As for the additional introgressed segments, we were able to define six in this work, distributed on four chromosomes. Molnar et al. (2003) reported three such additional contrasting regions using six Harosoy NILs (including L62-667): Satt293 on LG H (Gm12), Satt367 on LG I (Gm20), and Sat119 on LGK (Gm09). The first two SSR markers map to introgressed segments that we also observed on Gm12 (663 kb) and Gm20 (824 kb), but for unknown reasons, we did not find polymorphic SNPs on Gm09, even though 31 of our SNPs covered a 1Mb region around Sat119. Of the four remaining introgressed segments that we detected using SNP markers, all went undetected with the set of SSR markers used by Molnar et al. (2003). The relatively small size of three of these intervals (from 63 to 360 kb) may explain this. However, the very large (33 Mb) introgressed segment on Gm01 also went undetected. This could be due to its location in a heterochromatic region in which little recombination occurs. Indeed, genetic markers flanking this segment are located only 5 cM apart (SoyBase), such that the detection of this large physical interval would require a dense coverage of markers on the genetic map.

The GBS haplotype method described here not only allowed us to define much more precisely the introgressed segment, but also allowed us to distinguish between two introgressed alleles. As the *e3* donor parent used in the two pairs of NILs were not the same (Bernard et al., 1991; Voldeng and Saindon, 1991), the observed SNP haplotypes were also different. Indeed, while the *e3* allele introduced into Harosoy from T204 (Bernard et al., 1991) was found to be the *e3-tr* allele (Watanabe et al., 2009), the allele introduced into Maple Arrow from PI 196529 (Voldeng and Saindon 1991) was *e3-ns*. The introgressed region in OT89-1 did not overlap with the region found in L62-667. Indeed, this region was located at the border of the gene but didn't include the gene GmPhyA3. As all other lines carrying the allele (*e3-ns*) introduced in OT89-1 also didn't show any polymorphic markers closer to the GmPhyA3 gene, we can conclude that it is due to a lack of polymorphic markers in the vicinity of the gene.

Identification of Existing and Novel Alleles at the E3 Locus through the Study of SNP Haplotypes

The second phase of the work aimed to explore the possibility of using SNP haplotypes to rapidly define and distinguish alleles at the E3 locus. In previous work, the identification of alleles at a maturity locus relied on the systematic sequencing of the entire gene in a small set of lines of interest. Once a set of alleles had been defined, allele-specific tests were designed to allow the identification of a diagnostic feature of each allele (Xu et al., 2013; Tsubokura et al., 2013). In practice, however, the characterization of a large set of lines (such as the 449 analyzed here) could require performing a very large number of allele-specific tests to identify previously-described alleles and could fail to correctly identify additional alleles. Here, we show that six different haplotypes were observed in our collection of lines and that each one corresponded to a distinct allele at the E3 locus. Five of these represented previously reported alleles (*E3Ha*, *e3-tr*, *E3Mi*, *e3-fs*, *e3-ns*; Watanabe et al., 2009; Xu et al., 2013) and one allele (*E/e3p.Thr832Ala*) is described here for the first time. The impact of the amino acid substitution in *E/e3p.Thr832Ala* allele on the functionality of the protein remains to be investigated.

From a breeding perspective, the identification of three different loss of function alleles (*e3-tr*, *e3-ns*, and *e3-fs*) at the E3 locus may prove useful. Even if all three are predicted to lead to a similar phenotype in regards to maturity, this can now be tested more formally in a broad range of genetic backgrounds. Furthermore, even if these *e3* alleles were equivalent in their impact on flowering and maturity, closely linked genes contributing to other traits could carry different alleles and result in differences in the phenotype of these lines as a consequence of linkage drag. Finally, a precise knowledge of the different alleles at the E3 locus, as well as their distribution among Canadian lines and the PIs, provides useful information for the selection of parents to develop new cultivars.

E3Ha and *e3-tr* are the Predominant Alleles in the Canadian Germplasm

The functional late *E3Ha* allele and the loss of function *e3-tr* allele were the two predominant alleles found in Set C representing the Canadian germplasm (respectively 41.9 and 45.7%). The Canadian soybean crop evolved from a foundation provided by a limited number of early selections from PIs originating generally from northern China. The cultivars released between 1940 and 1970 were mostly derived from this narrow parental germplasm base. Among these, the cultivars Mandarin (Ottawa) and A.K. Harrow are particularly important founders (Beversdorf et al., 1995). In our study, Mandarin was found to carry the *e3-tr* allele, whereas Harosoy (Mandarin × A.K.

Harrow) is known to carry the *E3Ha* allele. Harosoy is one of the two most common ancestors in North America (Gizlice et al., 1994; Beversdorf et al., 1995; Mikel et al., 2010). Among the lines derived from Harosoy, Maple Arrow (Harosoy (*E3Ha*) × 840-7-3 (*e3-tr*), a milestone cultivar in the 1970s) and Evans (Harosoy [*E3Ha*] × Merit [*e3-tr*]), are important parents in the breeding history of Ontario short-season cultivars (Cober and Voldeng 2012; Beversdorf et al., 1995; Voldeng et al., 1997). These two cultivars also carry a functional *E3* allele (Molnar et al., 2003). We found that they exhibit Haplotype A and therefore most likely also carry the *E3Ha* allele.

The high frequency of the *e3-tr* allele in the Canadian germplasm collection could be derived from a few key ancestors. Mandarin (Ottawa), Manitoba brown, and cultivars derived from these two founders were found to carry this early allele. Second, this loss of function allele is also found in other different key ancestor introductions (unrelated to Mandarin and Manitoba brown) and their progeny. For example, Merit (Johnson, 1960) and Flambeau, an early USSR introduction (Strommen et al., 1967), both carry the *e3-tr* allele. This suggests that the *e3-tr* allele was introduced more than once from different sources in the Canadian germplasm.

A Limited Contribution from Swedish Germplasm

The early alleles *e3-fs* and *e3-ns* were found in 17.6 and 16.5%, respectively, of Set B (PI lines), but in only 5.3 and 0% of the Canadian germplasm collection. In Set B, both alleles were found in cultivars from diverse geographical origins (Europe, China, Japan, and Russia). With the exception of a few Swedish (Fiskeby) varieties, the PIs of Set B carrying either of these two alleles are not clearly known to have been grown or used as parents in Canada. From the 1960s, very early introductions from Sweden were used as parents in Canadian soybean breeding and many of the cultivars released since 1970 have Swedish germplasm in their pedigrees (predominantly 840-7-3 and Fiskeby V; Cober and Voldeng, 2012; Beversdorf et al., 1995; Voldeng et al., 1997). In Set B, of the 12 Swedish lines, six were genotyped as *e3-tr* (including 840-7-3), two as *e3-fs*, and four as *e3-ns*. Given the limited or complex pedigree information available, we can only hypothesize that Traff can be a probable source of the *e3-fs* allele present in Canadian germplasm. With regards to the low frequency of this allele in the collection, it appears that the *e3-fs* allele from early Swedish cultivars was not extensively used or maintained in the breeding of early Canadian cultivars.

The *e3-ns* loss of function allele has been found in some Fiskeby varieties including Fiskeby V, but this allele is not present at all in the Canadian germplasm collection we sampled. While this cultivar was among the worldwide standards for adaptation to long days and cool summers (Lambert, 1979) and represent a key progenitor in Canadian soybean, it appears that Fiskeby V did not contribute significantly to the introduction of early e3 alleles into Canadian cultivars.

Minor Alleles and Haplotypes

In our study, the functional *E3Mi* allele described by Watanabe et al. (2009) was found in only three genotypes: PI 159925, 'Toki,' and its progeny OT11-09. Toki is a recently released cultivar (2007) and results from crosses involving AC Proteus (*e3-ns*), Ken Feng (a Chinese cultivar), and Tamba Black (a Japanese cultivar; Cober et al., 2007). Interestingly, PI 159925 was introduced in the United States from Peru in 1947 (Villarroel and Kilen, 2009), and is the source of the recessive *j* allele at the long juvenile maturity locus *J* that delays flowering in response to short days (Ray et al., 1995; Cober, 2011).

Unsurprisingly, we did not find the loss of function *e3Mo* early allele. Additional characterization work performed on 30 soybean accessions (Harada et al., 2011) showed that this allele is rare and has only been reported to date in the cultivar Moshido Gong 503.

In both Sets B and C, only a few genotypes showed unique haplotypes. These genotypes have not been further analyzed yet. These rare haplotypes could represent additional alleles, be explained by rare recombination events or, finally, could also be artifacts due to an inaccurate fastPhase imputation on diagnostic markers.

Conclusions

Genotyping by Sequencing combined with a haplotype based approach allowed us to rapidly, efficiently, and economically characterize a large number of soybean lines at the E3 locus. This approach could be used for any known gene of interest and, combined with the use of NILs, can also be used as an approach to narrow down the search of yet uncharacterized genes. Our haplotype-based approach consists of three steps: (i) use GBS to obtain a high density marker coverage, (ii) define haplotypes in the vicinity of a gene of interest, and (iii) relate haplotypes to previously reported alleles or novel alleles through the use of allele-specific PCR assays or sequencing of the entire gene. The results of the current study together with the characterizations of the other maturity genes currently underway will provide us marker tools for breeding and a better understanding of early maturity in the Canadian germplasm. This will allow a more targeted approach to early maturity breeding.

Supplemental Information Available

Supplemental information is included with this article.

Acknowledgments

This work was funded jointly by The Canadian Field Crop Research Alliance and Agriculture and Agri-Food Canada under the Developing Innovative Agri-Products Program (DIAP). The authors also wish to thank Dr. Istvan Rajcan (University of Guelph, Canada) and Dr. Elroy Cober (Agriculture and AgriFood Canada, Ottawa, Canada) for providing some of the soybean lines used in this research.

References

Akkaya, M.S., R.C. Shoemaker, J.E. Specht, A.A. Bhagwat, and P.B. Cregan. 1995. Integration of simple sequence repeat DNA markers into a soybean linkage map. Crop Sci. 35:1439–1445. doi:10.2135/cropsci1995.0011183X003500050030x

Berloo, R.V. 2008. GGT 2.0: Versatile software for visualization and analysis of genetic data. J. Hered. 992:232–236. doi:10.1093/jhered/esm109

Bernard, R.L. 1971. Two major genes for time of flowering and maturity in soybeans. Crop Sci. 11:242–244. doi:10.2135/cropsci1971.0011183X001100020022x

Bernard, R.L., R.L. Nelson, and C.R. Cremeens. 1991. USDA soybean genetic collection: Isoline collection. Soybean Genet. Newsl. 18:27–57.

Beversdorf, W.D., R.I. Buzzell, G.R. Ablett, and H.D. Voldeng. 1995. Soybean. In: A.E. Slinkard and D.R. Knott, editors, Harvest of gold, The history of field crop breeding in Canada. University of Saskatchewan Univ. Ext. Press, Saskatoon, SK. p. 1–16.

Bonato, E.R., and N.A. Vello. 1999. E6, a dominant gene conditioning early flowering. Genet. Mol. Biol. 22:229–232. doi:10.1590/S1415-47571999000200016

Buzzell, R.I. 1971. Inheritance of a soybean flowering response to fluorescent-daylength conditions. Can. J. Genet. Cytol. 13:703–707.

Buzzell, R.I., and H.D. Voldeng. 1980. Inheritance of insensitivity to long daylength. Soybean Genet. Newsl. 7:29–29.

Cober, E.R., and M.J. Morrison. 2010. Regulation of seed yield and agronomic characters by photoperiod sensitivity and growth habit genes in soybean. Theor. Appl. Genet. 120:1005–1012. doi:10.1007/s00122-009-1228-6

Cober, E.R. 2011. Long juvenile soybean flowering responses under very short photoperiods. Crop Sci. 51:140–145. doi:10.2135/cropsci2010.05.0262

Cober, E.R., R.J.D. Guillemette, and H.D. Voldeng. 2007. Toki soybean. Can. J. Plant Sci. 87:513–514. doi:10.4141/CJPS06004

Cober, E.R., S.J. Molnar, M. Charette, and H.D. Voldeng. 2010. A new locus for early maturity in soybean. Crop Sci. 50:524–527. doi:10.2135/cropsci2009.04.0174

Cober, E.R., D.W. Stewart, and H.D. Voldeng. 2001. Photoperiod and temperature responses in early-maturing, near-isogenic soybean lines. Crop Sci. 41:721–727. doi:10.2135/cropsci2001.413721x

Cober, E.R., J.W. Tanner, and H.D. Voldeng. 1996. Soybean photoperiod-sensitivity loci respond differentially to light quality. Crop Sci. 36:606–610. doi:10.2135/cropsci1996.0011183X003600030014x

Cober, E.R., and H.D. Voldeng. 2001. A new soybean maturity and photoperiod-sensitivity locus linked to E1 and T. Crop Sci. 41:698–701. doi:10.2135/cropsci2001.413698x

Cober, E.R., and H.D. Voldeng. 2012. A retrospective look at short-season soybean cultivar development in Ontario. Can. J. Plant Sci. 92:1239–1243. doi:10.4141/cjps2012-032

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, and J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464–1490. doi:10.2135/cropsci1999.3951464x

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):e19379. doi:10.1371/journal.pone.0019379

Gizlice, Z., T.E. Carter, and J.W. Burton. 1994. Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci. 34:1143–1151. doi:10.2135/cropsci1994.0011183X003400050001x

Harada, K., S. Watanabe, X. Zhengjun, Y. Tsubokura, N. Yamanaka, and T. Anai. 2011. Positional cloning of the responsible genes for maturity loci E1, E2 and E3 in soybean. In: P.D. Krezhova, editor, Soybean—Genetics and novel techniques for yield enhancement. InTech, Rijeka, Croatia. p. 51–76.

Holmberg, S.A. 1973. Soybeans for cool temperate climates. Agric. Hort. Genet. 31:1–20.

Johnson, H.W. 1960. Registration of soybean varieties, VII. Agron. J. 1:659–660. doi:10.2134/agronj1960.00021962005200110015x

Lambert, J.W. 1979. Holmberg, soybean pioneer in the far north. Soybean News 30(3):4.

Liu, B., A. Kanazawa, H. Matsumura, R. Takahashi, K. Harada, and J. Abe. 2008. Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. Genetics 180:995–1007. doi:10.1534/genetics.108.092742

McBlain, B.A., and R.L. Bernard. 1987. A new gene affecting the time of flowering and maturity in soybeans. J. Hered. 78:160–162.

Mikel, M.A., B.W. Diers, R.L. Nelson, and H.H. Smith. 2010. Genetic diversity and agronomic improvement of North American soybean germplasm. Crop Sci. 50:1219–1229. doi:10.2135/cropsci2009.08.0456

Milne, I., P. Shaw, G. Stephen, M. Bayer, L. Cardle, W.T.B. Thomas, A.J. Flavell, and D. Marshall. 2010. Flapjack—Graphical genotype visualization. Bioinformatics 26(24):3133–3134. doi:10.1093/bioinformatics/btq580

Molnar, S.J., S. Rai, M. Charette, and E.R. Cober. 2003. Simple sequence repeat (SSR) markers linked to E1, E3, E4, and E7 maturity genes in soybean. Genome 46:1024–1036. doi:10.1139/g03-079

Ray, J.D., K. Hinson, J.E.B. Mankono, and M.F. Malo. 1995. Genetic control of a long-juvenile trait in soybean. Crop Sci. 35:1001–1006. doi:10.2135/cropsci1995.0011183X003500040012x

Scheet, P., and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78:629–644. doi:10.1086/502802

Shin, J.H., and S.H. Lee. 2012. Molecular markers for the E2 and E3 genes controlling flowering and maturity in soybean. Mol. Breed. 30:1793–1798. doi:10.1007/s11032-012-9743-6

Sonah, H., M. Bastien, E. Iquira, A. Tardivel, G. Légaré, B. Boyle, É. Normandeau, J. Laroche, S. Larose, M. Jean, and F. Belzile. 2013. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE 8(1):1–9. doi:10.1371/journal.pone.0054603

Strommen, A.M., C.O. Rydberg, and J.H. Torrie. 1967. Registration of Flambeau soybeans. Crop Sci. 7(3):279. doi:10.2135/cropsci1967.0011183X000700030045x

Thakare, D., S. Kumudini, and R.D. Dinkins. 2010. Expression of flowering-time genes in soybean E1 near-isogenic lines under short and long day onditions. Planta 231:951–963. doi:10.1007/s00425-010-1100-6

Thakare, D., S. Kumudini, and R.D. Dinkins. 2011. The alleles at the E1 locus impact the expression pattern of two soybean FT-like genes shown to induce flowering in Arabidopsis. Planta 234:933–943. doi:10.1007/s00425-011-1450-8

Tsubokura, Y., H. Matsumura, M. Xu, B. Liu, H. Nakashima, T. Anai, F. Kong, X. Yuan, H. Kanamori, Y. Katayose, R. Takahashi, K. Harada, and J. Abe. 2013. Genetic variation in soybean at the maturity locus e4 is involved in adaptation to long days at high latitudes. Agronomy 3:117–134. doi:10.3390/agronomy3010117

Villarroel, D.A., and T.C. Kilen. 2009. Effect of a long-juvenile trait upon flowering in soybeans grown in short and long photoperiods. Crop Sci. 35:1001–1006.

Voldeng, H.D., E.R. Cober, D.J. Hume, C. Gillard, and M.J. Morrison. 1997. Fifty-eight years of genetic improvement of short-season soybean cultivars in Canada. Crop Sci. 37:428–431. doi:10.2135/cropsci1997.0011183X003700020020x

Voldeng, H.D., and G. Saindon. 1991. Registration of seven long-daylength insensitive soybean genetic stocks. Crop Sci. 31:1399. doi:10.2135/cropsci1991.0011183X003100050095x

Watanabe, S., R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, Y. Nakamoto, N. Yamanaka, R. Takahashi, M. Ishimoto, T. Anai, S. Tabata, and K. Harada. 2009. Map-based cloning of the gene associated with the soybean maturity locus E3. Genetics 182:1251–1262. doi:10.1534/genetics.108.098772

Watanabe, S., Z. Xia, R. Hideshima, Y. Tsubokura, S. Sato, N. Yamanaka, R. Takahashi, T. Anai, S. Tabata, K. Kitamura, and K. Harada. 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. doi:10.1534/genetics.110.125062

Xia, Z., S. Watanabe, T. Yamada, Y. Tsubokura, H. Nakashima, H. Zhai, T. Anai, S. Sato, T. Yamazaki, S. Lü, H. Wu, S. Tabata, and K. Harada. 2012a. Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proc. Natl. Acad. Sci. USA doi:10.1073/pnas.1117982109

Xia, Z., H. Zhai, B. Liu, F. Kong, X. Yuan, H. Wu, E.R. Cober, and K. Harada. 2012b. Molecular identification of genes controlling flowering time, maturity, and photoperiod response in soybean. Plant Syst. Evol. 298:1217–1227. doi:10.1007/s00606-012-0628-2

Xu, M., Z. Xu, B. Liu, F. Kong, Y. Tsubokura, S. Watanabe, Z. Xia, K. Harada, A. Kanazawa, T. Yamada, and J. Abe. 2013. Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. BMC Plant Biol. 13:91. doi:10.1186/1471-2229-13-91