Mol Breeding (2018) 38: 93
https://doi.org/10.1007/s11032-018-0851-9

Allelic differentiation at the *Sg-1* locus for the terminal sugar of the C-22 position of group A saponin in Chinese wild soybean (*Glycine soja* Sieb. & Zucc.)

Yuya Takahashi · Xianghua Li · Chigen Tsukamoto · Kejing Wang

Received: 13 February 2018 / Accepted: 19 June 2018 / Published online: 1 July 2018

Abstract Group A saponins are thought to be the cause of bitter and astringent tastes in processed foods of soybean (*Glycine max*), and the elimination of group A saponins is an important breeding objective. The group A saponins include two main Aa and Ab types, controlled by codominant alleles at the *Sg-1* locus that is one of several key loci responsible for saponin biosynthesis in the subgenus *Glycine soja*. However, A0 mutant lacking group A saponin is a useful gene resource for soybean quality breeding. Here, eight Chinese wild soybean A0 accessions were sequenced to reveal the mutational mechanisms, and the results showed that these mutants were caused by at least three kinds of mechanisms involving four allelic variants (*sg-1*$^{0\text{-}b2}$, *sg-1*$^{0\text{-}b3}$, *Sg-1*$^{b\text{-}0}$, and *Sg-1*$^{b\text{-}01}$). The *sg-1*$^{0\text{-}b2}$ had two nucleotide deletions at positions + 72 and + 73 involving in the 24th and 25th amino acids. The *sg-1*$^{0\text{-}b3}$ contained a stop codon (TGA) at the 254th residue. The *Sg-1*$^{b\text{-}0}$ and *Sg-1*$^{b\text{-}01}$ were two novel A0-type mutants, which likely carried normal structural alleles, and nevertheless did not encode group A saponin due to unknown mutations beyond the normal coding regions. In addition, to reveal the structural features, allelic polymorphism, and mechanisms of the abiogenetic absence of group A (i.e., A0 phenotype), nucleotide sequence analysis was performed for the *Sg-1* locus in wild soybean (*Glycine soja*). The results showed that *Sg-1* alleles had a lower conservatism in the coding region; as high as 18 sequences were found in Chinese wild soybeans in addition to the *Sg-1*^{a} (Aa) and *Sg-1*^{b} (Ab) alleles. *Sg-1*^{a} and *Sg-1*^{b} alleles were characterized by eight synonymous codons and nine amino acid substitutions. Two evolutionarily transitional allelic sequences (*Sg-1*^{a7} and *Sg-1*^{b2}) from *Sg-1*^{a} toward *Sg-1*^{b} were detected.

Keywords Saponin A0-αg · Group A saponin · Soybean saponins · Nucleotide sequence · Wild soybean · *Glycine soja*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11032-018-0851-9) contains supplementary material, which is available to authorized users.

Y. Takahashi · X. Li · K. Wang (✉)
Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: wangkj@caas.net.cn

Y. Takahashi
Iwate Agricultural Research Center, Kitakami, Iwate 024-003, Japan

C. Tsukamoto (✉)
Faculty of Agriculture, Iwate University, Morioka, Iwate 020-8550, Japan
e-mail: chigen@iwate-u.ac.jp

Introduction

Soybean [*Glycine max* (L.) Merr.] has a number of nutritional and functional components such as proteins, lipids, vitamins, dietary fibers, isoflavones, phytic acid, sterols, lectins, and saponins (Tsukamoto and Yoshiki 2006). Soybean saponins are secondary metabolites with a triterpenoid (C_{30}) aglycone with one or two sugar chain(s) (Price et al. 1987). They have become the

subject of breeding because their chemical characteristics affect soybean food properties including taste and health benefits (Tsukamoto and Yoshiki 2006). Triterpenoid saponins represent 0.6–6.2% of dry weight of soybean seeds (Shiraiwa et al. 1991b), and wild soybean (*Glycine soja* Sieb. and Zucc.) seeds contain at least three times more saponin than cultivated soybean (Tsukamoto et al. 1994). Soybean saponins, depending on the chemical structural characteristics of their aglycones, can be divided into six groups: group A, DDMP, group B, group E, group α, and group β (Shiraiwa et al. 1991a, c, Kudou et al. 1992, 1993, Tsukamoto et al. 1993, Krishnamurthy et al. 2014a, b, Takahashi et al. 2016a, b, 2017). Group A saponins have soyasapogenol A as their aglycone with one or two sugar chain(s) attached at the C-3 and C-22 position(s) (except A-series) (Fig. 1). DDMP, group B, and E saponins show beneficial activities for health (Fenwick et al. 1991, Kuzuhara et al. 2000, Rowlands et al. 2002, Murata et al. 2006, Ellington et al. 2005 2006, Kang et al. 2005, Lee et al. 2005, Ishii and Tanizawa 2006) and group A saponins have functions of preventing memory impairment (Hong et al. 2014) and inhibiting adipocyte differentiation (Yang et al. 2015). The health functionalities of group α saponins are unknown (Itabashi et al. 2016).

According to the terminal sugar moieties at the C-22 position of soyasapogenol A, group A saponins are classified into four subgroups: Aa-, Ab-, A0-, and A-series. The C-22 hydroxyl position is attached by triacetyl xylose (AcXyl)-arabinose (Ara) in Aa-series, by tetraacetyl glucose (AcGlc)-Ara in Ab-series, by Ara in A0-series, and by no sugar (OH) in A-series (Fig. 1). The C-22 terminal sugar moieties (AcXyl and AcGlc) are controlled by the codominant alleles (*Sg-1*^{a} and *Sg-1*^{b}) in a single *Sg-1* locus in chromosome no. 7 (Takada et al. 2010; Sayama et al. 2012), and the recessive allele *sg-1*^{0} makes the terminal sugar (glucose or xylose) incapable of binding to the secondary sugar (arabinose) at the C-22 position, and consequently, no group A acetyl saponin Aa- or Ab-series occurs but instead results in an A0 component. Recently, the *Sg-1* locus was cloned and sequenced and nine amino acid differences were shown between *Sg-1*^{a} and *Sg-1*^{b} alleles (Sayama et al. 2012). Interestingly, the allelic frequencies differ between *Sg-1*^{a} and *Sg-1*^{b}; South Korean wild soybeans had 98.4% *Sg-1*^{a} allele and 1.2% *Sg-1*^{b} allele in 3720 wild soybean plants (Krishnamurthy et al. 2014a, b) and Chinese wild soybean had 78.7% *Sg-1*^{a} allele and 20.8% *Sg-1*^{b} allele in 3795 accessions (Takahashi et al. 2016a, b). The such high frequency of *Sg-1*^{a} in the wild species suggests that the *Sg-1*^{a} allele could be primordial and the *Sg-1*^{b} be acquired from *Sg-1*^{a}.

The group A Aa- and Ab-series saponins cause bitter and astringent tastes in soybean seeds because of acetylation of the terminal sugar moieties at the C-22 position (Okubo et al. 1992). Thus, the genetic reduction of group A acetyl saponins is an important subject in soybean breeding. A wild soybean (CWS2133) and an ethyl methane sulfonate (EMS)-treated soybean mutant (PE1515) in South Korea, and a wild soybean (JP 36121) and a soybean variety (Kinusayaka) in Japan, were reported to be A0 type caused by the recessive *sg-1*^{0} gene, where four respective different mutations (deletions and termination codons) led to lack of Aa or Ab saponin in these accessions (Sayama et al. 2012; Krishnamurthy et al. 2015; Park et al. 2016). The occurrence of many Aa- or Ab-lacking mutants implies that the biological activity or function of Aa- and Ab-type saponins could be substituted by other saponins in plants.

Interestingly, the allelic frequencies differ between *Sg-1*^{a} and *Sg-1*^{b}; South Korean wild soybeans had 98.4% *Sg-1*^{a} allele and 1.2% *Sg-1*^{b} allele in 3720 wild soybean plants (Krishnamurthy et al. 2014a, b) and Chinese wild soybean had 78.7% *Sg-1*^{a} allele and 20.8% *Sg-1*^{b} allele in 3795 accessions (Takahashi et al. 2016a, b). Such high frequency of *Sg-1*^{a} in the wild species suggests that the *Sg-1*^{a} allele could be primordial and the *Sg-1*^{b} be acquired from *Sg-1*^{a}. In this way, another interesting issue is that how many amino acid substitutions in *Sg-1*^{a} allele could encode Ab saponin but it does not have to reach at the complete nine amino acid substitutions in *Sg-1*^{b} allele, i.e., whether there are transitional allelic sequences between *Sg-1*^{a} and *Sg-1*^{b} alleles. In addition, little is known about whether there are amino acid mutations or codon (nucleotide) mutations characteristic of *Sg-1*^{a} or *Sg-1*^{b} allele in Aa and Ab saponins. Wang et al. (2008) observed two transitional sequences the *Tib* and *Tia* alleles in soybean Kunitz trypsin inhibitor protein (SKTI) We think that there is the possibility of having different allelic sequences for Aa- or Ab-type saponin in different accessions in spite of the similar and same mobility on the TLC. To demonstrate this possibility, we examined the structural features and nucleotide mutations in the *Sg-1* locus through gene sequencing. In this work, we report the mutational mechanisms of the A0 accession variants detected in Chinese wild soybeans and the allelic variation and differentiation at the *Sg-1* locus, the characteristic single

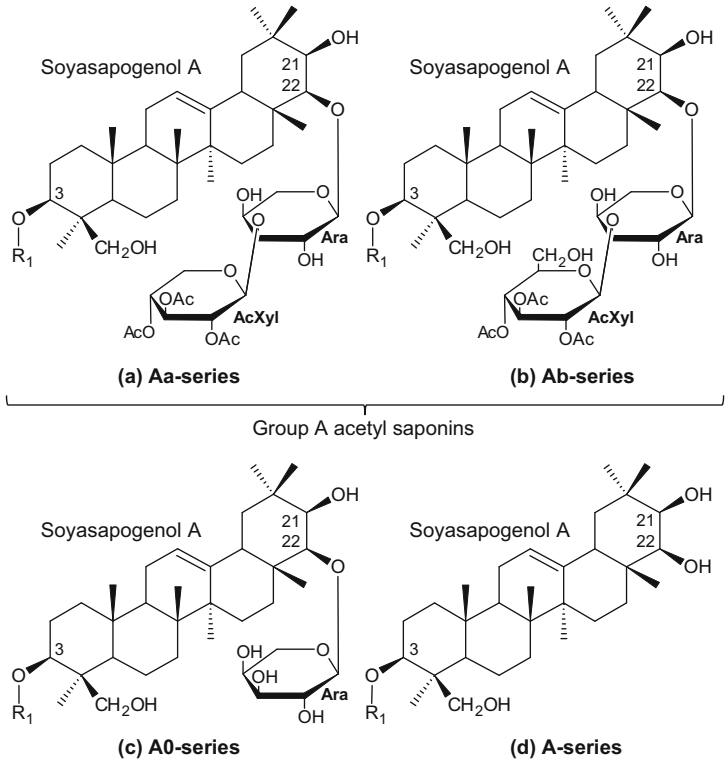

Fig. 1 Structure and nomenclature of group A saponin components. Group A saponins have a soyasapogenol A as the aglycone with one or two sugar chain(s) attached at the C-3 (and the C-22) position(s). The terminal sugar moiety at the C-22 position is genetically controlled by codominant alleles at the *Sg-1* locus

	C-22 position (R_2)			
	Aa-series	Ab-series	A0-series	A-series
	Sg-1^a	*Sg-1*^{b}	*sg-1*^{0}	- (*sg-7*)
C-3 position (R_1)	O-(AcXyl-Ara-)	O-(AcGlc-Ara-)	O-Ara	OH
Glc-Gal-GlcUA-	Aa	Ab	A0-αg	A-αg
Rha-Gal-GlcUA-	Au	Ac	A0-βg	A-βg
Gal-GlcUA-	Ae	Af	A0-γg	A-γg
Glc-Ara-GlcUA-	Ax	Ad	A0-αa	A-αa
Rha-Ara-GlcUA-	Ay	Az	A0-βa	A-βa
Ara-GlcUA-	Ag	Ah	A0-γa	A-γa

nucleotide polymorphisms (SNPs) between *Sg-1*^{a} and *Sg-1*^{b} alleles, and the evolutionary relationship among polymorphic sequences in *G. soja* is also discussed.

Materials and methods

Materials and chemicals

A total of 3805 Chinese wild soybean accessions were randomly taken from the Chinese wild soybean collection were identified for saponin composition using TLC analysis prior to sequencing. All the eight A0-type wild soybeans (nos. 0115, 0262, 0676-1, 0676-2 1168, 1842, 5026, and 4278) lacking Aa or Ab group A saponin were taken and 72 Aa- and 74 Ab-type accessions were randomly selected for gene sequencing at the *Sg-1* locus (Table S1). These A0-type accessions were collected at different places in northeast China, except for no. 4278 from Henan Province. Two Japanese Aa- and Ab-type soybean varieties "Shirosennari" and "Ohsuzu" were sequenced as standards because these two varieties had been reported to carry *Sg-1*^{a} and *Sg-1*^{b} alleles, respectively (Sayama et al. 2012). All chemicals (methanol, chloroform, sulfuric acid, acetonitrile, and formic acid) of analytical grade were purchased from Beijing Chemical Works (Beijing, China, http://www.beijingchemworks.com/).

Extraction of saponin components

Soybean seeds were divided into hypocotyls, cotyledons, and seed coats with a utility knife from mature seeds. Saponin components were extracted from intact seed hypocotyls with a tenfold volume (*w*/*v*) of 80% methanol at room temperature (25 °C) for 12 h. They were stored at − 20 °C until analysis.

TLC analysis for preliminary screening

TLC analysis was performed according to Krishnamurthy et al. (2012) with minor revisions to identify saponin composition (Takahashi et al. 2016a, b). The 10-μL extracts were directly applied on silica gel coated TLC plates (Merck Millipore, Darmstadt, Germany). The plates were developed with the lower phase mixed chloroform, methanol, and water (65:35:10, *v*/*v*) in a glass chamber for 22 min and dried at 115 °C for 15 min. After the TLC plates were cooled to room temperature, they were developed with 10% (*v*/*v*) of dilute sulfuric acid for 12 min and dried at 115 °C for 12 min to visualize saponin components on the plates. These results were recorded by scanning with an Epson Perfection 2400 photo (Seiko Epson Corporation, Nagano, Japan).

Liquid chromatography-mass spectroscopy analysis

To identify the facticity for TLC-detected A0-type accessions, liquid chromatography-mass spectroscopy (LC-MS analysis was performed using a LC-MS-2020 system (Shimadzu Corporation, Kyoto, Japan) with a photodiode array (PDA) detector and a reverse-phase column (Inertsil ODS-4, 2.1 mm I.D. × 150 mm, 3 μm; GL Sciences, Tokyo, Japan). Saponin extracts were diluted with five times (*v*/*v*) with 80% methanol and 5 μL was injected. Saponins were eluted with 0.1% (*v*/*v*) formic acid (solvent A) and acetonitrile including 0.1% (*v*/*v*) formic acid (solvent B) at a flow rate of 0.15 mL/min. Linear gradient elution was carried out as follows: solvent B was initiated at 10% (*v*/*v*) and increased to 90% (*v*/*v*) over 80 min. Saponin components were monitored at 205 and 292 nm using a mass spectrometer in the positive ion mode of the electrospray ionization (ESI[+]). Data of UV chromatograms and MS spectra were analyzed using the dedicated application LabSolutions version 5.42 SP6 (Shimadzu Corporation).

Nucleotide sequences analysis of *Sg-1* gene

Gene sequence analysis was carried out to clarify the allelic structural features and the nucleotide variation of A0 type at the *Sg-1* locus. Genomic DNA was extracted from a seed of each accession using the NuClean Plant Genomic DNA Kit (Cat. CW0531M; CWBIO Co., Beijing, China). The *Sg-1* gene was amplified by PCR using two sets of primers (set 1, forward: 5′-ATGG ATCTTCAACAACGACCACT-3′, reverse: 5′-CTCT TCTCGCCCCTCTCTTG-3′; and set 2, forward: 5′-TCAAGAGAGGGGCGAGAAGA-3′, reverse: 5′-TCAGGTGGCCGACTTAGAGT-3′) designed on the basis of the issued sequences of $Sg\text{-}1^a$ and $Sg\text{-}1^b$ (Sayama et al. 2012). The amplified fragment lengths were 731 and 721 bp, respectively. Twenty microliter of mixture for PCR was subjected to 94 °C for 5 min for initial denaturation, followed by 35 cycles composed of 94 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s, and 72 °C for 8 min to complete elongation. The amplified products were cloned using the vector PEASY-T1 (Trans Gene Co., Beijing, China) and were sequenced with an ABI 370XL Genetic Analyzer with a BigDye3.1 Terminator Cycle Sequencing Kit (Applied Biosystems). Nucleotide sequences were submitted to phylogenetic analysis using MEGA version 6.0 (Tamura et al. 2013), with the neighbor-joining (NJ) method. Bootstrapping with 1000 replications for the NJ analysis was carried out.

Results

Phenotyping saponin composition at *Sg-1* locus

The saponin composition analysis of 3805 wild soybean accessions determined by TLC analysis indicated that the *Sg-1* locus produced three phenotypes of group A saponins: Aa, Ab and A0 types (Fig. 2a). About 78.8% of wild soybeans were Aa type and 20.9% were Ab type. Eight variant accessions (nos. 0115, 0262, 0676-1, 0676-2, 1168, 1842, 4278, and 5026) were determined to lack group A saponin Aa or Ab, and these accessions simultaneously occurred with saponin A0-αg and/or A0-αa, instead of group A acetyl saponins Aa and Ab (Fig. 2a). The LC-MS analysis confirmed that the eight A0 variant accessions did not contain any group A saponin Aa or Ab component, and that they contained A0-αg and/or A0-αa (nos. 0676-2, 1842, and 4278) (Fig. 2b). Thus, all eight

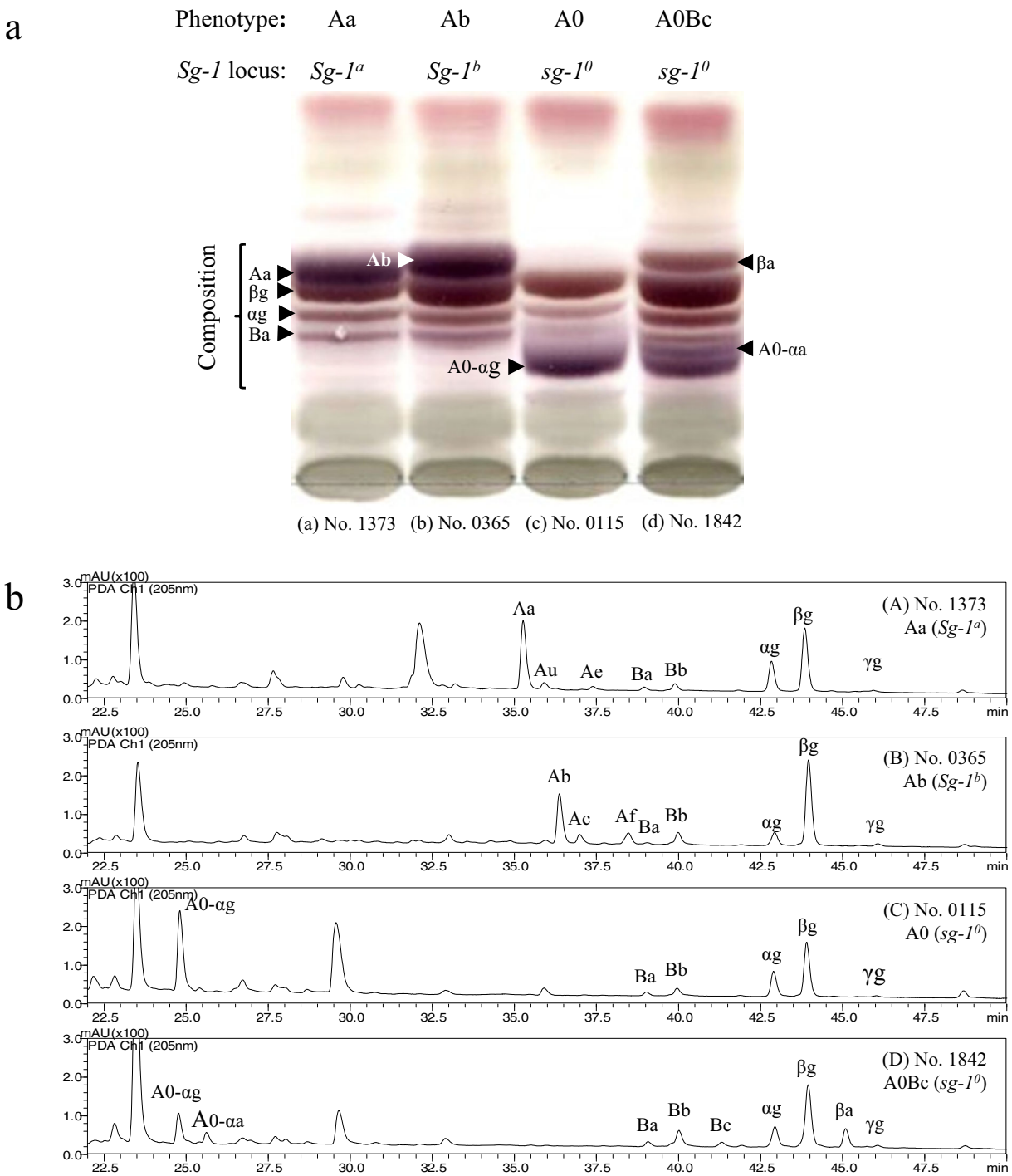

Fig. 2 TLC and LC-PDA/MS analyses of saponin composition in seed hypocotyls. **a** TLC patterns of saponin composition; accession nos. 1373 and 0365 were normal Aa (allele *Sg-1*^{a}) and Ab (allele *Sg-1*^{b}) types and 0115 and 1842 were two A0 variants. **b** LC-PDA/MS patterns of group A acetyl saponin components in the four accessions under 205 nm (UV); the Aa and Ab saponins appeared at 35.0–38.5 min in accessions nos. 1373 and 0365 but were absent in the phenotype mutants nos. 0115 and 1842, as shown in (**a**). Two components (A0-αg and A0-αa) of group A acetyl saponin were eluted at 24.5–26.0 min from the A0 mutants

accessions were determined as A0 phenotype mutants of group A saponins.

Sequence polymorphism of the *Sg-1* gene

A total of 146 randomly selected wild soybean accessions (72 TLC-Aa type and 74 TLC-Ab type), eight novel TLC-A0-type wild accessions, and two standard soybean cultivars (Shirosennari and Ohsuzu) were used to determine the nucleotide sequences of the *Sg-1* alleles. Analysis of the amplified gene sequences showed that the normal Aa and Ab types had the same length of 1431 bp, encoding 476 amino acids. A total of up to 18 allelic sequences were detected from normal Aa and Ab types, characterized by a higher level of non-conservatism: eight allelic sequences (*Sg-1*^{a} and *Sg-1*$^{a1-7}$) were distinguished from normal Aa type and ten were recognized from Ab-type accessions (*Sg-1*^{b} and *Sg-1*$^{b1-9}$) (Table 1). Forty-three Aa-type and 33 Ab-type accessions had the same gene sequences as those of *Sg-1*^{a} and *Sg-1*^{b} alleles from standard cultivars Shirosennari for Aa type and Ohsuzu for Ab type, respectively (Sayama et al. 2012). However, two alleles (*Sg-1*^{a} and *Sg-1*^{b}) differed by nine amino acids at residues 99, 102, 128, 138, 143, 144, 145, 149, and 292 and eight nucleotides (eight nonsense codons) at nucleotide positions + 114, 156, 387, 456, 495, 561, 978, and 1035 (Table 1).

In the 72 Aa-type wild soybean accessions, eight allelic sequence variants (*Sg-1*$^{a1-7}$) occurred (Fig. S3, Table 1). Of which, one transitional sequence (*Sg-1*^{a7}) from *Sg-1*^{a} toward *Sg-1*^{b} with two base characteristics of *Sg-1*^{b} was first detected in wild soybean (Table S2); it evolved from *Sg-1*^{a} by altering an amino acid at residue 292 (Pro → Ser) and providing a characteristic synonymous mutation (A → G) of the *Sg-1*^{b} allele at position + 978. The *Sg-1*^{a7} also evolved into two correlative mutant sequences, starting from a *Sg-1*^{a6} sequence with an amino acid change (Gln → Lys) at residue 248 into *Sg-1*^{a5} through a synonymous mutation (G → C) at position + 978 and a synonymous mutation (C → A) at + 1311 (Figs. 3 and S3, Table 1).

Some copies of *Sg-1*^{a} allele also mutated into *Sg-1*^{a4} by an amino acid change at residue 453 (Gly → Ala) and into *Sg-1*^{a3} through a synonymous mutation (at + 222, C → T) and further into *Sg-1*^{a2} through another synonymous mutation (at + 873, G → A) from *Sg-1*^{a3} (Figs. 3 and S3, Table 1).

In the 74 Ab-type wild soybean accessions, there were nine allelic sequence variants (*Sg-1*$^{b1-9}$) (Table 1). Of these, one (*Sg-1*^{b2}) was a transitional sequence from *Sg-1*^{a} toward *Sg-1*^{b} (Table S2). The *Sg-1*^{b1} had the complete nine substituted amino acid residues and five of the eight synonymous codons of *Sg-1*^{b} from *Sg-1*^{a}, and only three characteristics of *Sg-1*^{a} remained at position + 495 (ATT), + 561 (GAC) and + 1035 (CCT); however, it also carried an accidental amino acid mutation at residue 248 (Gln → Lys). It was deduced that a *Sg-1*$^{b1'}$ sequence should exist according existence of the *Sg-1*^{b1} sequence, which had the normal residue 248 (Gln) (Table S2).

Another *Sg-1*^{b2} transitional sequence is derived from the precursor (*Sg-1*$^{b1'}$) of *Sg-1*^{b1}, which carried only one characteristic nucleotide for *Sg-1*^{a} at + 1035 (CCT). However, the unchanged base was an essential gene characteristic of *Sg-1*^{b} despite being only one nucleotide (Table 1). The complete *Sg-1*^{b} allele was formed by a final one synonymous mutation at position + 1035 (T → G) from the transitional *Sg-1*^{b2} sequence. The *Sg-1*^{b2} also separately evolved into three allelic sequences by a synonymous mutation (G → A) at + 389 for *Sg-1*^{b3}, by a synonymous mutation at + 457 (C → T) for *Sg-1*^{b4}, and by an amino acid change at residue 172 for *Sg-1*^{b5} (Figs. 3 and S4, Table 1).

In addition, the *Sg-1*^{b} allele also independently mutated into four variant sequences by two amino acid substitutions at residues 306 and 356 for *Sg-1*^{b8} and by three respective synonymous mutations for *Sg-1*^{b6} (at + 397 C → A), *Sg-1*^{b7} (at + 531 A → G), and *Sg-1*^{b9} (at + 1240 A → C).

Diverse mechanisms for A0-type wild soybeans

Eight A0 phenotype wild soybean accessions with no group A acetyl saponin Aa or Ab shown on TLC were revealed to be caused by at least three kinds of mechanisms in four new allelic sequence variants: *sg-1*$^{0-b2}$, *sg-1*$^{0-b3}$, *Sg-1*$^{b-0}$, and *Sg-1*$^{b-01}$ (Fig. S5, Table 2).

First, the *sg-1*$^{0-b2}$ induced A0 variation by deletion mutations in three accessions (nos. 0115, 0262, and 1168), where two nucleotide deletions occurred at positions + 72 and + 73 involving in two amino acid absences at residues 24 and 25 in the mutant allele, with a characteristic codon (CCT) of *Sg-1*^{a} at position + 1035.

Second, the *sg-1*$^{0-b3}$ incurred A0 variation by a stop codon in two accessions (nos. 1842 and 5026). The stop codon occurred through a nonsense mutation by transition at position + 762 (TGG → TGA), with an amino

Table 1 Nucleotide and amino acid differences among different allelic sequences at *Sg-1* locus for normal Aa and Ab types in Chinese wild soybean

Type	Seq.	Sample no.	nt. AA	114 38	156 52	222 74	295 99	304 102	384 128	387 129	389 130	397 133	412–3 138	427 143	431 144	433–4 145	446 149
Aa	*Sg-1*^{a}	Shirosennari		GTC Val	AAA Lys	CGC Arg	GGA Gly	GCC Ala	AAG Lys	CTG Leu	CGC Arg	CGA Arg	TCA Ser	GCC Ala	GTC Val	AGC Ser	TCC Ser
Aa	*Sg-1*^{a}	43		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	*Sg-1*^{a1}	11		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	*Sg-1*^{a2}	1		–	–	- - T	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	*Sg-1*^{a3}	5		–	–	- - T	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	*Sg-1*^{a4}	1		–	–	- - C	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	*Sg-1*^{a5}	2		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	*Sg-1*^{a6}	8		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	*Sg-1*^{a7}	1		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*Sg-1*^{b}	Ohsuzu		*- - G*	*- - G*	–	*A- -*	*T- -*	*- - C*	*- - C*	–	–	*GG -*	*T- -*	*- G -*	*GC -*	*-G-*
				–	–	–	*Arg*	*Ser*	*Asn*	–	–	–	*Gly*	*Ser*	*Gly*	*Ala*	*Cys*
Ab	*Sg-1*^{b}	33		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*Sg-1*^{b1}	8		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*Sg-1*^{b2}	10		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*Sg-1*^{b3}	1		–	–	–	–	–	–	–	- A -	–	–	–	–	–	–
				–	–	–	–	–	–	–	His	–	–	–	–	–	–
Ab	*Sg-1*^{b4}	2		–	–	–	–	–	–	–	- G -	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*Sg-1*^{b5}	2		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*Sg-1*^{b6}	1		–	–	–	–	–	–	–	–	A - -	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*Sg-1*^{b7}	15		–	–	–	–	–	–	–	–	C - -	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*Sg-1*^{b8}	1		–	–	–	–	–	–	–	–	–	–	–	–	–	–
				–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*Sg-1*^{b9}	1		–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 1 (continued)

Type	Seq.	Sample no.	nt. AA	114 38	156 52	222 74	295 99	304 102	384 128	387 129	389 130	397 133	412–3 138	427 143	431 144	433–4 145	446 149
				–	–	–	–	–	–	–	–	–	–	–	–	–	–

Type	456 152	457 153	495 165	514 172	531 177	561 187	570 190	742 248	873 291	874 292	917 306	978 326	1035 345	1066 356	1240 414	1311 437	1358 453
Aa	TCT	CAC	ATT	GTC	AGA	GAC	CTC	CAG	AAG	CCA	GAG	AGA	CCT	TGC	AGA	CGC	GGT
	Ser	His	Ile	Val	Arg	Asp	Leu	Gln	Lys	Pro	Glu	Arg	Pro	Trp	Arg	Arg	Gly
Aa	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	–	–	–	–	–	–	- - T	–	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	–	–	–	–	–	–	- - C	–	- - A	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	–	–	–	–	–	–	–	–	- - G	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	- C -
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	Ala
Aa	–	–	–	–	–	–	–	A - -	–	*-T - -*	–	- - C	–	–	–	- - A	- G -
	–	–	–	–	–	–	–	Lys	–	*Ser*	–	–	–	–	–	–	–
Aa	–	–	–	–	–	–	–	–	–	–	–	*- - G*	–	–	–	- - C	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Aa	–	–	–	–	–	–	–	C - -	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	*- - A*	–	*- - C*	–	–	*- - T*	–	–	–	–	–	–	*- - G*	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	–	–	- - T	–	–	- - C	–	A - -	–	–	–	–	- - T	–	–	–	–
	–	–	–	–	–	–	–	Lys	–	–	–	–	–	–	–	–	–
Ab	–	–	*- - C*	–	–	*- - T*	–	C - -	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 1 (continued)

Type	456 152	457 153	495 165	514 172	531 177	561 187	570 190	742 248	873 291	874 292	917 306	978 326	1035 345	1066 356	1240 414	1311 437	1358 453
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	–	T - -	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	–	Tyr	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	–	C- -	–	A - -	–	–	–	–	–	–	–	–	–	–	–	–	–
	–	–	–	Ile	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	–	–	–	G - -	–	–	–	–	–	–	–	–	- - G	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	–	–	–	–	- - G	–	–	–	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ab	–	–	–	–	- - A	–	–	–	–	–	-C-	–	–	A - -	–	–	–
	–	–	–	–	–	–	–	–	–	–	Ala	–	–	Ser	–	–	–
Ab	–	–	–	–	–	–	–	–	–	–	-A-	–	–	T - -	C - -	–	–
	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

nt mutational nucleotide positions, *AA* amino acid positions. Italicized characters are nucleotides or amino acids that differ between the standard *Sg-1*^{a} and *Sg-1*^{b} alleles

Shirosennari is a standard variety of *Sg-1*^{a} allele (Sayama et al. 2012). Ohsuzu is a standard variety of *Sg-1*^{b} allele (Sayama et al. 2012)

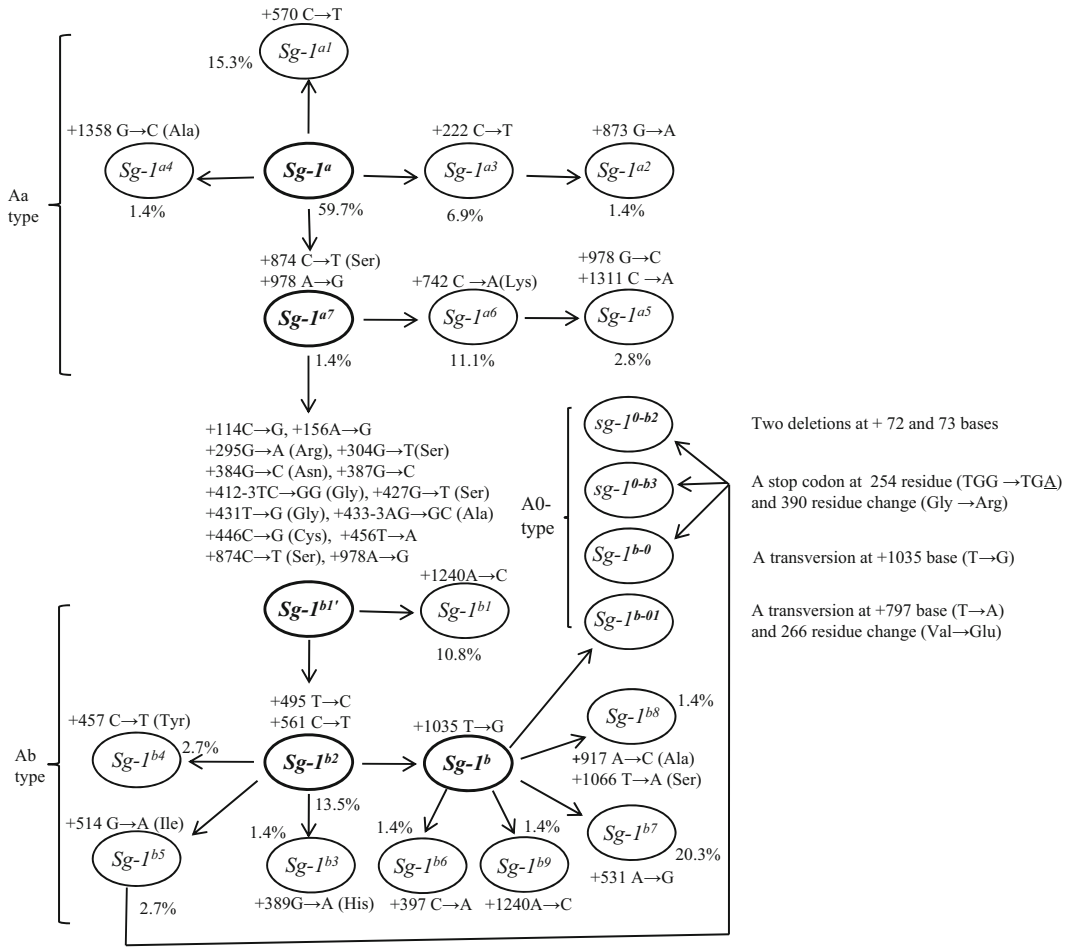

Fig. 3 Phylogenetic relationships of 22 functional polymorphic sequences (and deduced one *Sg-1*$^{b1'}$) detected at the *Sg-1* locus. The number of variant sequence accessions and their percentages in the determined 146 Aa or Ab accessions are indicated. Boldface letters (symbols) indicate the route of the evolutionary differentiation of *Sg-1* alleles from *Sg-1*^{a} to *Sg-1*^{b}. The *Sg-1* locus was highly polymorphic. Six allelic sequences *Sg-1*^{a1}, *Sg-1*^{a3}, *Sg-1*^{a6}, *Sg-1*^{b1}, *Sg-1*^{b2}, and *Sg-1*^{b7} had relatively high frequencies. The percentages (%) were the frequency of variant sequences in the sequenced Aa or Ab samples

acid substitution (Gly → Arg) at residue 390 and a characteristic synonymous codon (CCT) of *Sg-1*^{a} at position + 1035.

Third, the *Sg-1*$^{b-01}$ led to the occurrence of A0 phenotype by unknown causes in one accession (no. 4278), where only a nucleotide transversion mutation existed at position + 797 (GTA → GAA) and accordingly led to an amino acid change (Val → Glu) at residue 266 in the mutant allele, with the characteristic codon (CCT) of *Sg-1*^{a} at position + 1035.

Finally, the *Sg-1*$^{b-0}$ led to occurrence of A0 phenotype by similarly unknown causes beyond the reading region of the gene in two accessions (nos. 0676-1 and 0676-2). In this case, an amino acid changed from Val (GTC) to Ile (ATC) at residue 172 in the mutant allele, with the characteristic codon (CCT) of *Sg-1*^{a} at position + 1035 in one accession (no. 4278).

Structural features of the *Sg-1* locus

A total of 22 allelic sequences, of which 18 normally expressed the biosynthesis of Aa- and Ab-type saponins (Table 1) and four did not (Table 2), were sequenced and aligned for comparison. The NJ tree based on sequence structures (Fig. S1) grouped these sequences into two large groups, Aa-type (Aa allele and its derived variants) and Ab-type saponin (Ab allele and its derived variants) (Table 1) or three categories, Aa-type, Ab-type and transitional-type variants (*Sg-1*^{a7} and *Sg-1*^{b1} or *Sg-1*$^{b1'}$ and *Sg-1*^{b2}).

Table 2 Nucleotide and amino acid differences at the *Sg-1* locus among variants lacking group A acetyl saponins detected in Chinese wild soybean

Sample	Species (Country)	Type (Allele)	nt AA	27 ~ 56 9 ~ 19 ↓ ↓	72 24	73 25	114 38	156 52	295 99	304 102	384 128	387 129	412-3 138	427 143	431 144	433-4 145
Shirosennari	*G. max* (Japan)	Aa ($Sg\text{-}1^{a}$)		AAA~GGG - - - ~ - - -	CTC Leu	TGC Cys	GTC Val	AAA Lys	GGA Gly	GCC Ala	AAG Lys	CTG Leu	TCA Ser	GCC Ala	GTC Val	AGC Ser
JP-36121	*G. soja* (Japan)	A0 ($sg\text{-}1^{0\text{-}a}$)		**Deletion** **Deletion**	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
CWS2133	*G. soja* (Korea)	A0 ($sg\text{-}1^{0\text{-}a1}$)		- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
Ohsuzu	*G. max* (Japan)	Ab ($sg\text{-}1^{b}$)		- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	- - G - - -	- - G - - -	A - - Arg	T - - Ser	- - C Asn	- - C - - -	GG- Gly	T - - Ser	GGC Gly	GCC Ala
Kinusayaka	*G. max* (Japan)	A0 ($sg\text{-}1^{0\text{-}b}$)		- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
PE1515	*G. max* (Korea)	A0 ($sg\text{-}1^{0\text{-}b1}$)		- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
0542, 2161	*G. soja* (China)	Ab ($Sg\text{-}1^{b5}$)		- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
676-1, 676-2	*G. soja* (China)	A0 ($Sg\text{-}1^{b\text{-}0}$)		- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
0115, 0262, 1168	*G. soja* (China)	A0 ($sg\text{-}1^{0\text{-}b2}$)		- - - ~ - - - - - - ~ - - -	CT **Del**	GC **Del**	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
1842, 5026	*G. soja* (China)	A0 ($sg\text{-}1^{0\text{-}b3}$)		- - - ~ - - - - - - ~ - - -	- - C - - -	T - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
4278	*G. soja* (China)	A0 ($Sg\text{-}1^{b\text{-}01}$)		- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -

Sample	446 149	456 152	495 165	514 172	561 187	598 ~ 645 200 ~ 215 ↓ ↓	762 254	797 266	874 292	948 316	978 326	1004 335	1035 345	1168 390
Shirosennari	TCC Ser	TCT Ser	ATT Ile	GTC Val	GAC Asp	ATC~CAC - - - ~ - - -	TGG Trp	GTA Val	CCA Pro	TGG Trp	AGA Arg	TGG Trp	CCT Pro	GGG Gly
JP-36121	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
CWS2133	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	- - - - - -	- - A **Stop**	- - - - - -	- - - - - -	- - - - - -	- - - - - -
Ohsuzu	-G- Cys	- - A - - -	- - C - - -	- - - - - -	- - T - - -	- - - ~ - - - - - - ~ - - -	- - - - - -	- - - - - -	T - - Ser	- - G - - -	- - G - - -	- - - - - -	- - G - - -	- - - - - -
Kinusayaka	- - -	- - -	- - -	- - -	- - -	**Deletion**	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -

Table 2 (continued)

Sample	446 149	456 152	495 165	514 172	561 187	598 ~ 645 200 ~ 215 ↓ ↓	762 254	797 266	874 292	948 316	978 326	1004 335	1035 345	1168 390
	- - -	- - -	- - -	- - -	- - -	**Deletion**	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -
PE1515	- - -	- - -	- - -	- - -	- - -	- - - ~ - - -	- - -	- - -	- - -	- - -	- - -	- A -	- - -	- - -
	- - -	- - -	- - -	- - -	- - -	- - - ~ - - -	- - -	- - -	- - -	- - -	- - -	**Stop**	- - -	- - -
0542, 2161	- - -	- - -	- - -	A - -	- - -	- - - ~ - - -	- - -	- - -	- - -	- - -	- - -	- G -	- - T	- - -
	- - -	- - -	- - -	Iie	- - -	- - - ~ - - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -
676-1, 676-2	- - -	- - -	- - -	- - -	- - -	- - - ~ - - -	- - -	- - -	- - -	- - -	- - -	- - -	- - G	- - -
	- - -	- - -	- - -	- - -	- - -	- - - ~ - - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -
0115, 0262,	- - -	- - -	- - -	- - -	- - -	- - - ~ - - -	- - -	- - -	- - -	- - -	- - -	- - -	- - T	- - -
1168	- - -	- - -	- - -	- - -	- - -	- - - ~ - - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -
1842, 5026	- - -	- - -	- - -	- - -	- - -	- - - ~ - - -	- - A	- - -	- - -	- - -	- - -	- - -	- - -	A - -
	- - -	- - -	- - -	- - -	- - -	- - - ~ - - -	**Stop**	- - -	- - -	- - -	- - -	- - -	- - -	Arg
4278	- - -	- - -	- - -	G - -	- - -	- - - ~ - - -	- - G	- A -	- - -	- - -	- - -	- - -	- - -	G - -
	- - -	- - -	- - -	- - -	- - -	- - - ~ - - -	- - -	Glu	- - -	- - -	- - -	- - -	- - -	- - -

nt nucleotide positions, *AA* amino acid positions. ↓ (arrowhead) indicates the deleted first and end nucleotides

Shirosennari standard variety of *Sg-1*^{a} allele (Sayama et al. 2012). *Ohsuzu* standard variety of *Sg-1*^{b} allele (Sayama et al. 2012)

Sample no.0542 and 2161 are normal Ab phenotype encoded by the *Sg-1*^{b5} variant sequence with an amino acid mutation from Val to Ile (GTC → ATC) at 172th residue

The *sg-1*$^{0\text{-}a1}$ (CWS2133) was reported by Krishnamurthy et al. (2015); *sg-1*$^{0\text{-}a}$ (JP-36121) and *sg-1*$^{0\text{-}b}$ (Kinusayaka) by Sayama et al. (2012); *sg-1*$^{0\text{-}b1}$ (PE1515) by Park et al. (2016)

The differences in allelic sequences between both *Sg-1^a* and *Sg-1^b* types were characterized by 18 nucleotides involving nine amino acid substitutions and eight synonymous codons (nucleotides) (Table 1). These nine amino acid changes all occurred in 295–874 bp (at residues 99–292), particularly concentrated in the anterior region of 295–456 bp (at residues 99–149), where there were eight amino acid substitutions (Table 1). This suggests that the differentiation of *Sg-1^a* from the *Sg-1^b* allele had undergone intense nucleotide arrangement in the small specific region, particularly in the three consecutive residues of positions 143, 144, and 145.

Allelic sequences for Aa type had *Sg-1^b*-characters or vice versa. There were shared codons in four residue positions in the Aa and Ab types. *Sg-1^a* and *Sg-1^b* shared codons CAG (Gln) and AAG (Lys) at residue 248, codon TCA (Ser) at residue 292, codon AGG (Arg) at the residue 326, and codon CCT (Pro) at residue 345 (Table 1).

Nine substituted amino acids between *Sg-1^a* and *Sg-1^b* alleles involved 11 nucleotide mutations, seven of which were transversions. The eight synonymous codons between *Sg-1^a* and *Sg-1^b* alleles involved four transversion mutations of nucleotides.

There were a total of 15 nucleotide mutations in these allelic sequence variants except *Sg-1^a* and *Sg-1^b* alleles; eight were transversions and seven were transition mutations (Table 1). Of these, seven base mutations engendered seven respective amino acid changes in two Aa-type accessions, *Sg-1^{a4}* (at residue 453, Gly → Ala) and *Sg-1^{a5}* (at residue 248, Gln → Lys) and four Ab-type ones, *Sg-1^{b3}* (at residue 130, Arg → His), *Sg-1^{b4}* (at residue 153, His→Tyr), *Sg-1^{b5}* (at residue 172, Val → Ile), and *Sg-1^{b8}* (at residue 306, Glu → Ala, and residue 356, Trp → Ser). These single accidental amino acid changes did not affect the synthesis of Aa- and Ab-type saponins.

Discussion

Characteristics and phylogenetic relationships of the polymorphic sequences between *Sg-1^a* and *Sg-1^b* alleles

The present research confirmed by the large number of sequences analyzed that the *Sg-1^a* and *Sg-1^b* alleles differed by nine characteristic amino acid substations and eight synonymous codons as reported by Sayama et al. (2012). It was also shown that there was sequence polymorphism in the *Sg-1* locus; 18 allelic sequences were identified and their phylogenetic relationships were inferred (Fig. S1), of which eight allelic sequences were detected from Aa-type and 10 from Ab-type Chinese wild soybeans in this investigation (Table 1), which suggested that the *Sg-1* locus had a lower conservatism or high mutability in the coding region, as also supported by findings of many mutations in soybean and wild soybean (Sayama et al. 2012; Krishnamurthy et al. 2015; Park et al. 2016).

As expected, evolutionary transitional sequences remained in the wild species. Two transitional allelic sequences (*Sg-1^{a7}* and *Sg-1^{b2}*) were detected from *Sg-1^a* toward *Sg-1^b* in these 18 polymorphic allelic sequences. According to the existing polymorphic sequences, the differentiation route between *Sg-1^a* and *Sg-1^b* alleles was deduced (Table S2). However, a transitional precursor *Sg-1$^{b1'}$* sequence with a normal residue 248 (Gln) was deduced to exist, based on the existence of *Sg-1^{b1}* sequence with an amino acid mutation (at residue 248, Gln → Lys) (Table S2).

We also noted that the mutation of amino acid (Gln → Lys) at residue 248 in *Sg-1^{a6}* and *Sg-1^{b1}* was likely two independent and coincidental events, in which the same point mutation was repeated twice. The two events were probably separated by a very large time interval in two completely different stages of allelic differentiation, the second vent (in *Sg-1^{b1}* variant sequence) occurring after the institution of the nine amino acids between *Sg-1^a* and *Sg-1^b* (Fig. 3, Table S2).

The present analysis revealed that the nine amino acid substitutions between *Sg-1^a* and *Sg-1^b* alleles occurred in a region of 295–874 bp (at residues 99–292) of the amplified length of 1431 bp, particularly concentrating in the anterior region of 295–456 bp (at residues 99–149), where there were eight amino acid substitutions (Table 1). This suggests that the differentiation of *Sg-1^a* from *Sg-1^b* underwent intense nucleotide arrangement in a small specific region, particularly in the consecutive three residues of positions 143, 144, and 145. Thus, not all the nine amino acid substitutions in *Sg-1^b* from *Sg-1^a* occurred gradually, and some were possibly simultaneously replaced.

Diverse mechanisms for A0 type

Four sequence variants (*sg-1$^{0\text{-}a}$*, *sg-1$^{0\text{-}a1}$*, *sg-1$^{0\text{-}b}$*, and *sg-1$^{0\text{-}b1}$*) of A0 phenotype lacking group A acetyl

saponins have been reported in soybean and wild soybean, caused by the recessive *sg-1*^{0} gene, where respective different mutations (deletions and termination codons) led to deficiency of Aa- or Ab-series saponins (Sayama et al. 2012; Krishnamurthy et al. 2015; Park et al. 2016). The *sg-1*$^{0\text{-}a}$ (JP-36121, *G. soja*) and *sg-1*$^{0\text{-}b}$ (Kinusayaka, *G. max*) were both deletion mutations of relative long base fragments (Table 2) (Sayama et al. 2012). The *sg-1*$^{0\text{-}a1}$ (CWS2133, *G. soja*) and *sg-1*$^{0\text{-}b1}$ (PE1515, *G. max*) were both nonsense mutations leading to stop codons (Table 2) (Krishnamurthy et al. 2015; Park et al. 2016).

Our analysis showed eight A0-type wild soybean accessions that lacked group A saponins attributed to four new sequence variants: *sg-1*$^{0\text{-}b2}$ (accession nos. 0115, 0262, and 1168), *sg-1*$^{0\text{-}b3}$ (nos. 1842 and 5026), *Sg-1*$^{b\text{-}0}$ (nos. 0676–1 and 0676–2), and *Sg-1*$^{b\text{-}01}$ (no. 4278) (Table 2). We also found a base deletion mutation (*sg-1*$^{0\text{-}b2}$) and a nonsense mutation (*sg-1*$^{0\text{-}b3}$); however, *sg-1*$^{0\text{-}b2}$ only deleted two nucleotides involving neighboring codons, compared with more nucleotide deletions for *sg-1*$^{0\text{-}a}$ and *sg-1*$^{0\text{-}b}$ (Table 2).

We first found two novel A0-type mutants that had normal allelic sequences (*Sg-1*$^{b\text{-}0}$ and *Sg-1*$^{b\text{-}01}$). However, the two mutants (*Sg-1*$^{b\text{-}0}$ and *Sg-1*$^{b\text{-}01}$) were caused by non-gene structural changes on *Sg-1*. There were two reasons for this: (1) although variant *Sg-1*$^{b\text{-}0}$ had a specific amino acid mutation (isoleucine, Ile) at residue 172, the *Sg-1*^{b5} variant (nos. 0542 and 2161) also had this amino acid at the same residue position and could produce normal Ab saponin (Table 2), and (2) many single amino acid mutations (*Sg-1*^{a4}, *Sg-1*^{a5}, *Sg-1*^{b3}, and *Sg-1*^{b4}) and even double amino acid mutations (*Sg-1*^{b4}) did not influence function of the *Sg-1* enzyme (Table 1). Therefore, the *Sg-1*$^{b\text{-}01}$ A0 mutant also impossibly lost its function in virtue of only the one amino acid mutation (Val → Glu) (G<u>T</u>A → G<u>A</u>A) at residue 266 (Tables 1 and 2). Consequently, we assigned dominant symbols *Sg-1*$^{b\text{-}0}$ and *Sg-1*$^{b\text{-}01}$ to the two kinds of novel A0 phenotype accessions nos. 0676-1 and 0676-2, and no. 4278, respectively (Table 2).

There could be other unknown genetic variations beyond the normal coding regions making the *Sg-1*$^{b\text{-}0}$ and *Sg-1*$^{b\text{-}01}$ variants to have the A0 phenotype. We analyzed the upstream operon sequences (2000 bp) of the coding regions for *Sg-1*$^{b\text{-}01}$ and *Sg-1*$^{b\text{-}0}$ (data not shown) and the two mutants showed no abnormity in their promoter sequences against the standard varieties "Shirosennari" (*Sg-1*^{a}) and "Ohsuzu" (*Sg-1*^{b}). We surmise that the absence of group A saponins in the two A0 variant accessions was possibly due to (a) epigenetics, such as gene methylation, and (b) mutation in other genes that regulate and control expression of the *Sg-1* alleles. These hypothetical mutations would affect normal expression of *Sg-1* alleles. Our conjectures on the reasons for the absence of group A saponins in the two A0 accessions need to be resolved in future studies.

The *sg-1*$^{0\text{-}b2}$, *sg-1*$^{0\text{-}b3}$, and *Sg-1*$^{b\text{-}0}$ all had the specific amino acid mutation (isoleucine, Ile) at residue 172 and the same characteristic nonsense codon (CC<u>T</u>) of *Sg-1*^{a} at position + 1035, which suggested that the three defective variants were derivatives of the *Sg-1*^{b5}, and *Sg-1*$^{b\text{-}01}$ was derived from *Sg-1*^{b} by an amino acid mutation at residue 797 (Val → Glu) (Figs. S2 and S5, Table 2). All these A0 variant accessions (*sg-1*^{0} or *Sg-1*^{0}) might be utilized to eliminate the unpleasant tastes in soybean foods by genetic breeding.

Spread and distribution of the *Sg-1* alleles in Chinese wild soybean

All phylogenetic relationships of 18 polymorphic functional sequences detected at the *Sg-1* locus are shown in Figs. 3 and S1. In Aa-type saponin, sequence *Sg-1*^{a4} from the original *Sg-1*^{a} allele, *Sg-1*^{a2} (synonymous) from *Sg-1*^{a3}, all with single base alterations, were detected in one accession, respectively. It is not certain whether these sequence changes occurred long ago during evolution. The *Sg-1*^{a5} from *Sg-1*^{a6} sequence appeared in two accession; it should have existed for a long period because it has two synonymous mutations and the two accessions also existed in relatively far interval space (Fig. S2). The *Sg-1*^{a7} was a very ancient transitional allele between *Sg-1*^{a} and *Sg-1*^{b}, and possessed the earliest two nucleotide substitutions from *Sg-1*^{a} toward *Sg-1*^{b}, and may have arisen not long after naissance of the *Sg-1*^{a} allele although it was detected in one accession from Jiangxi Province in southeast China (Table S2). Three allelic sequences *Sg-1*^{a1} and *Sg-1*^{a3} (synonymous) from the original *Sg-1*^{a} allele, and the *Sg-1*^{a6} (synonymous) from *Sg-1*^{a7}, had relatively higher frequencies with 15.3, 6.9, and 11.1%, respectively.

The *Sg-1*^{a1} sequence was spread mostly over the vast area along the Yellow River valley and its north in China, with the exception of a *Sg-1*^{a1} accession that also appeared in Sichuan Province of southwest China. The *Sg-1*^{a6} sequence was mainly distributed in northeast China, with the exception of one accession that

appeared in the northwestern Ningxia area (Fig. S2). The *Sg-1*^{a3} was mainly distributed in the middle and lower reaches of the Changjiang River and the southeast coast (Fig. S2).

Likewise, in Ab-type saponin, for three sequences—*Sg-1*^{b6} (synonymous), *Sg-1*^{b8} (two base changes), and *Sg-1*^{b9} (synonymous) from the second original *Sg-1*^{b} allele—it was not certain whether they occurred long ago as they were detected in only one accession. The *Sg-1*^{b4} in two accessions was also not certain whether it occurred long ago as the accessions grew closer (Fig. S2). However, the *Sg-1*^{b5} sequence was probably older owing to the relatively longer geographic interval between the two *Sg-1*^{b5} accessions. Two sequences, *Sg-1*^{b7} (synonymous) from *Sg-1*^{b} and *Sg-1*^{b1} from *Sg-1*$^{b1'}$, surprisingly had relatively high frequencies of 20.3 and 10.8%, respectively, and were geographically regionally distributed along the Yellow River valley and along the Changjiang River valley and in its southern areas, respectively, with the exception of one *Sg-1*^{b1} accession that also occurred in Hebei Province of north China (Fig. S2). The *Sg-1*^{b1} should be regarded as derived from the ancient *Sg-1*$^{b1'}$ transitional sequence from which another detected ancient transitional sequence *Sg-1*^{b2} originated. The transitional *Sg-1*^{b2} allele from *Sg-1*^{a} toward *Sg-1*^{b} possessed the same sequences of nucleotides with the *Sg-1*^{b} allele except for a synonymous base transversion mutation (+ 1035 T → G) (Table S2) and was distributed through the vast northern areas of the Changjiang River, with the exception of one *Sg-1*^{b2} accession that also occurred in far Guangxi of southwest China (Fig. S2). As an ancient gene, it was unsurprising that this *Sg-1*^{b2} allele had a relatively high frequency of 13.5%.

Wild soybean is a self-pollinating plant. It was reported that there are positive genetic relationships among individuals in a small space of diameter of 30 m in field due to seed or pollen dispersal (Jin et al. 2003). Seed dispersal can occur via animals or water (Cain et al. 2000) and is one factor that influences the spatial pattern of variation and population genetic structure of wild soybean (Abe 2000). Kuroda et al. (2008) reported a strong positive genetic correlation between individuals in neighboring populations within a range of 400 m in wild soybean. Even in a range of 200 km, there are close genetic relationships between populations (Kuroda et al. 2006). Kuroda et al. (2006) found that wild soybean seeds can spread as far as 12.4 km from an original population and we also found a long-distance dispersal of 1.5 km for wild soybean seeds (Wang and Li 2012).

The present study demonstrated that not only had the original *Sg-1* alleles continued to independently mutate but also the transitional sequences or mutant sequences continuously resulted in new derivative or mutant sequences (Fig. 3, Table 1). The data showed high polymorphism in *Sg-1* alleles (Table 1) and some allelic sequences spread over large areas (Fig. S2). The *Sg-1*^{b2} allelic sequence spread through a long distance of about 2400 km in a farthest straight line, *Sg-1*^{a1} across about 2600 km, and *Sg-1*^{b1} over about 1600 km (Fig. S2). The *Sg-1*^{a6} accessions were distributed across a spread of about 1800 km in approximately longitude within the northern areas (Fig. S2) and they could survive. However, like *Sg-1*^{b1} and *Sg-1*^{b2}, such long-distance spread of genes in latitude could not be established through animals and birds carrying seeds, because wild soybeans could hardly survive under such long-distance migrations in latitude since wild soybean is a short-day plant. The unique possibility is that some polymorphic alleles that occurred in ancient times were gradually disseminated throughout this species by concomitant species spread of wild soybean in China. Such high polymorphism of the *Sg-1* locus has potential as an important molecular indicator to explore the specific area of origin of soybeans in China.

Author contributions Y.T. performed experiments and wrote the article; X.L. performed experiments; C.T. designed this research; K.J.W. designed the entire research and wrote the article.

Funding information This work was supported by the National Natural Science Foundation of China (Grant No. 31571697), the Sci & Tech Innovation Program of Chinese Academy of Agricultural Sciences, and the National Basic Research Program from the Ministry of Science and Technology of the People's Republic of China (Item No. 2011FY110200).

References

Abe J (2000) The genetic structure of national population of wild soybeans revealed by isozymes and RFLPs of mitochondrial DNAs: possible influence of seed dispersal cross-pollination and demography. In: Proceeding 7th MAFF International Workshop Genetic Resources Part 1. Wild Legumes. AFFRC and NIAR, Japan

Cain ML, Milligan BC, Sterand AE (2000) Long-distance seed dispersal in plant populations. Am J Bot 87:1217–1227

Ellington AA, Berhow MA, Singletary KW (2005) Induction of macroautophagy in human colon cancer cells by soybean B-group triterpenoid saponins. Carcinogenesis 26:159–167

Ellington AA, Berhow MA, Singletary KW (2006) Inhibition of Akt signaling and enhanced ERK1/2 activity are involved in induction of macroautophagy by triterpenoid B-group soyasaponins in colon cancer cells. Carcinogenesis 27:298–306

Fenwick GR, Price KR, Tsukamoto C, Okubo K (1991) Saponins. In: D'Mello JPF, Duffus CM, Duffus JH (eds) Toxic substances in crop plants. The Royal Society of Chemistry, Cambridge, pp 285–327

Hong SW, Yoo DH, Woo JY, Jeong JJ, Yang JH, Kim DH (2014) Soyasaponins Ab and Bb prevent scopolamine-induced memory impairment in mice without the inhibition of acetylcholinesterase. J Agric Food Chem 62:2062–2068

Ishii Y, Tanizawa H (2006) Effects of soyasaponins on lipid peroxidation through the secretion of thyroid hormones. Biol Pharm Bull 29:1759–1763

Itabashi M, Tsukamoto C, Kurosaka A, Krishnamurthy P, Shin TS, Yang SH, Son E, Chung G (2016) Efficient method for large-scale preparation of two components H and I of Sg-6 saponins from whole seeds of wild soybean (*Glycine soja* Sieb. and Zucc.). J Liq Chromatogr Relat Technol 39(14):640–646

Jin Y, He TH, Lu BR (2003) Fine scale genetic structure in a wild soybean (*Glycine soja*) population and the implication for conservation. New Phytol 159:513–519

Kang JH, Sung MK, Kawada T, Yoo H, Kim YK, Kim JS, Yu R (2005) Soybean saponins suppress the release of proinflammatory mediators by LPS-stimulated peritoneal macrophages. Cancer Lett 230:219–227

Krishnamurthy P, Tsukamoto C, Yang SH, Lee JD, Chung G (2012) An improved method to resolve plant saponins and sugars by TLC. Chromatographia 75:1445–1449

Krishnamurthy P, Tsukamoto C, Singh RJ, Lee JD, Kim HS, Yang SH, Chung G (2014a) The Sg-6 saponins, new components in wild soybean (*Glycine soja* Sieb. and Zucc.): polymorphism, geographical distribution and inheritance. Euphytica 198:413–424

Krishnamurthy P, Lee CM, Tsukamoto C, takahashi Y, Singh RJ, Lee JD, Chung G (2014b) Evaluation of genetic structure of Korean wild soybean (*Glycine soja*) based on saponin allele polymorphism. Genet Resour Crop Evol 61:1121–1130

Krishnamurthy P, Lee JD, Ha BK, Chae JH, Song JT, Tsukamoto C, Singh RJ, Chung G (2015) Genetic characterization of group A acetylsaponin-deficient mutants from wild soybean (*Glycine soja* Sieb. and Zucc.). Plant Breed 321:316–321

Kudou S, Tonomura M, Tsukamoto C, Shimoyamada M, uchida T, Okubo K (1992) Isolation and structural elucidation of the major genuine soybean saponin. Biosci Biotech Biochem 56: 142–143

Kudou S, Tonomura M, Tsukamoto C, Uchida T, Sakabe T, Tamura N, Okubo K (1993) Isolation and structural elucidation of DDMP-conjugated soyasaponins as genuine saponins from soybean seeds. Biosci Biotechnol Biochem 57:546–550

Kuroda Y, Kaga A, Tomooka N, Vaughan DA (2006) Population genetic structure of Japanese wild soybean (*Glycine soja*) based on microsatellite variation. Mol Ecol 15:959–974

Kuroda Y, Kaga A, Tomooka N, Vaughan AD (2008) Gene flow and genetic structure of wild soybean (*Glycine soja*) in Japan. Crop Sci 48:1071–1079

Kuzuhara H, Nishiyama S, Minowa N, sasaki K, Omoto S (2000) Protective effects of soyasapogenol A on liver injury mediated by immune response in a concanavalin A-induced hepatitis model. Eur J Pharmacol 391:175–181

Lee SO, Simons AL, Murphy PA, Hendrich S (2005) Soyasaponins lowered plasma cholesterol and increased fecal bile acids in female golden Syrian hamsters. Exp Biol Med 230:472–478

Murata M, Houdai T, Yamamoto H, Matsumori N, Oishi T (2006) Membrane interaction of soyasaponins in association with their antioxidation effect—analysis of biomembrane interaction. Soy Protein Res 9:82–86 (in Japanese)

Okubo K, Iijima M, Kobayashi Y, Yoshikoshi M, Uchida T, Kudou S (1992) Components responsible for the undesirable taste of soybean seeds. Biosci Biotechnol Biochem 56:99–103

Park J, Kim JH, Krishnamurthy P, Tsukamoto C, Song JT, Chung G, Shannon G, Lee JD (2016) Characterization of a new allele of the saponin-synthesizing gene *Sg-1* in soybean. Crop Sci 56:385–391

Price KR, Johnson IT, Fenwick GR (1987) The chemistry and biological significance of saponins in foods and feedstuffs. Crit Rev Food Sci Nutr 26:27–135

Rowlands JC, Berhow MA, Badger TM (2002) Estrogenic and antiproliferative properties of soy sapogenols in human breast cancer cells in vitro. Food Chem Toxicol 40:1767–1774

Sayama T, Ono E, Takagi K, Takada Y, Horikawa M, Nakamoto Y, Hirose A, Sasama H, Ohashi M, Hasegawa H, Terakawa T, Kikuchi A, Kato S, Tatsuzaki N, Tsukamoto C, Ishimoto M (2012) The *Sg-1* glycosyltransferase locus regulates structural diversity of triterpenoid saponins of soybean. Plant Cell 24: 2123–2138

Shiraiwa M, Kudo S, Shimoyamada M, Harada K, Okubo K (1991a) Composition and structure of group A saponin in soybean seed. Agric Biol Chem 55:315–322

Shiraiwa M, Harada K, Okubo K (1991b) Composition and content of saponins in soybean seed according to variety, cultivation year and maturity. Agric Biol Chem 55:323–331

Shiraiwa M, Harada K, Okubo K (1991c) Composition and structure of group B saponin in soy seed. Agric Biol Chem 55: 911–917

Takada Y, Sayama T, Kikuchi A, Kato S, Tatsuzaki N, Nakamoto Y, Suzuki A, Tsukamoto C, Ishimoto M (2010) Genetic analysis of variation in sugar chain composition at the C-22 position of group A saponin in soybean, *Glycine max* (L.) Merrill. Breed Sci 60:3–8

Takahashi Y, Li XH, Tsukamoto C, Wang KJ (2016a) Identification of a novel variant lacking group-A soyasaponin in a Chinese wild soybean (*Glycine soja* Sieb. & Zucc.): implications for breeding significance. Plant Breed 135(5): 607–613

Takahashi Y, Tsukamoto C, Li XH, Wang KJ (2016b) Polymorphic studies of soybean saponin components in Chinese wild soybean accessions (*Glycine soja* Sieb. and Zucc.). In: Abstracts of 7th International Crop Sci Congress. Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing, pp 109

Takahashi Y, Li XH, Tsukamoto C, Wang KJ (2017) Categories and components of soyasaponin in the Chinese wild soybean (*Glycine soja* Sieb. & Zucc.) genetic resource collection. Genet Resour Crop Evol 64:2161–2171

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729

Tsukamoto C, Yoshiki Y (2006) Soy saponin. In: Sugano M (ed) Soy in health and disease prevention. CRC Press, Taylor and Francis group, New York, pp 155–172

Tsukamoto C, Kikuchi A, Harada K, Kitamura K, Okubo K (1993) Genetic and chemical polymorphisms of saponins in soybean seed. Phytochemistry 34:1351–1356

Tsukamoto C, Kikuchi A, Kudou S, Harada K, Iwasaki T, Okubo K (1994) Genetic improvement of saponin components in soybean. In: Huang MT, Osawa T, Ho CT, Rosen RT (eds) Food phytochemicals for cancer prevention I. American Chemical Society, Washington, DC, pp 372–379

Wang KJ, Li XH (2012) Genetic characterization and gene flow in different geographical-distance neighboring natural populations of wild soybean (*Glycine soja* Sieb. & Zucc.) and implications for protection from GM soybeans. Euphytica 186:817–830

Wang KJ, Takahata Y, Kono Y, Kaizuma N (2008) Allelic differentiation of Kunitz trypsin inhibitor in wild soybean (*Glycine soja*). Theor Appl Genet 117:565–573

Yang SH, Ahn EK, Lee JA, Shin TS, Tsukamoto C, Suh JW, Itabashi M, Chung G (2015) Soyasaponins Aa and Ab exert an anti-obesity effect in 3T3-L1 adipocytes through down-regulation of PPARγ. Phytother Res 29:281–287