Whole-Genome Resequeuing Identifies the Molecular Genetic Cause for the Absence of a Gy5 Glycinin Protein in Soybean PI 603408

Jason D. Gillman,* Won-Seok Kim,* Bo Song,* Nathan W. Oehrle,* Nilesh R. Tawari,* Shanshan Liu,* and Hari B. Krishnan*†

*United States Department of Agriculture – Agricultural Research Service Plant Genetics Research Unit, Columbia, Missouri, †Plant Science Division, University of Missouri, Columbia, Missouri, ‡Key Laboratory of Soybean Biology at the Chinese Ministry of Education, Northeast Agricultural University, Harbin 150030, China, and §Computational and Systems Biology Group, Genome Institute of Singapore, Agency for Science, Technology and Research, 138672 Singapore

ORCID ID: 0000-0001-6437-3672 (H.B.K.)

ABSTRACT During ongoing proteomic analysis of the soybean (Glycine max (L.) Merr) germplasm collection, PI 603408 was identified as a landrace whose seeds lack accumulation of one of the major seed storage glycinin protein subunits. Whole genomic resequencing was used to identify a two-base deletion affecting glycinin 5. The newly discovered deletion was confirmed to be causative through immunological, genetic, and proteomic analysis, and no significant differences in total seed protein content were found to be due to the glycinin 5 loss-of-function mutation per se. In addition to focused studies on this one specific glycinin subunit-encoding gene, a total of 1,858,185 nucleotide variants were identified, of which 39,344 were predicted to affect protein coding regions. In order to semiautomate analysis of a large number of soybean gene variants, a new SIFT 4G (Sorting Intolerant From Tolerated 4 Genomes) database was designed to predict the impact of nonsynonymous single nucleotide soybean gene variants, potentially enabling more rapid analysis of soybean resequencing data in the future.

KEYWORDS soybean Glycine max glycinin genomic resequencing seed proteomics single nucleotide variants (SNV) amino acid changes (AAC)

Soybean (Glycine max (L.) Merr) is one of the most productive temperate crops, the seeds of which are a valuable source of protein and oil. Soybean seed proteins are largely (~70%) comprised of two ancestrally related multi-gene seed storage protein families: glycinins and β-conglycinin (Nielsen et al. 1989; Krishnan 2001). Glycinins are encoded by five primary seed-expressed gene family members (and two pseudogenes) (Nielsen et al. 1989; Beilinson et al. 2002). Glycinin is synthesized as a larger precursor protein, which is post-translationally cleaved into acidic and basic subunits that are held together by a disulfide bond (Staswick et al. 1984). Glycinins are classified into groups I and II on the basis of their size and sequence homology (Nielsen et al. 1989). Group I comprises three subunits, A₁bB₂, A₁b B₁b, and A₂B₁α, whereas group II includes two subunits, A₂B₄ and A₅ A₄B₃ (Nielsen et al. 1989). The β-conglycinin gene family, though quite large, contains only three proteins (α', α, and β protein subunits), which accumulate at appreciable levels in developing seeds (Thanh and Shibasaki 1978; Coates et al. 1985). The accumulation of the β-subunit of β-conglycinin is regulated by the availability of nitrogen and sulfur. Soybean plants grown in the presence of excess nitrogen have seeds that tend to have increased accumulation of the β-subunit of β-conglycinin (Gayler and Sykes 1985; Paek et al. 1997). Glycinin and β-conglycinin, on account of their abundance, are mainly responsible for the protein nutritive value of soybean seed. However, neither of these groups of proteins are rich in essential sulfur-containing amino acids (methionine and cysteine), though glycinins contain a relatively greater amount than β-conglycinins (Thanh and Shibasaki 1978).

Radiation-induced and spontaneous mutations that affect the accumulation of soybean seed proteins have been identified (Cho et al. 1989;
Kitamura and Kaizuma 1981; Ogawa et al. 1995; Scallon et al. 1985; Hayashi et al. 1998; Kim et al. 2008, 2013; Takahashi et al. 1994). A soybean line that is devoid of both glycinins and β-conglycinins has also been developed by integrating multiple, distinct mutations (Takahashi et al. 2003). Interestingly, such a mutant was able to grow and reproduce normally, indicating that glycinins and β-conglycinins are dispensable and function only as storage proteins (Takahashi et al. 2003).

Cultivated soybean has almost no outcrossing (< 0.5%) in the field (Yoshimura et al. 2006). Owing to this fact, and the extreme genetic bottleneck events that occurred during domestication and cultivar development (Gizlice et al. 1994, 1996), soybean breeding can introduce artificial impacts on genetic diversity as compared to G. max landraces and/or wild G. soja lines. For example, the ancestors of all high-yielding North American soybean commercial lines examined have intact, expressed glycinin and β-conglycinin genes (Kim et al. 2008), yet mutants for several seed storage proteins exist in G. max landraces (such as PI 60348). In contrast, direct targets of domestication can feature either elevated or repressed diversity in landraces as compared to wild progenitors; e.g., all known wild soybeans (G. soja) have black seedcoats, yet a range of seedcoat coloration is known to exist in G. max landraces (e.g., green, yellow, brown, black, and red-brown, etc.) (Gillman et al. 2011), and almost all high-yielding soybean cultivars have been selected to have yellow seedcoats, with the only coloration present restricted to the hilum (Vallyyodan et al. 2016).

During the course of routine screening of seed from the USDA-GRIN soybean germplasm collection with immunological and proteomic analyses, a soybean line (PI 603408) was identified that produces seeds that lack a major seed storage protein, which was preliminary identified as a glycinin subunit based on protein size via SDS-PAGE analysis. The effect of this novel gy5 loss-of-function allele was examined through a combination of proteomic and immunological analysis, NIRS spectroscopy, and classical genetic analysis. Very in-depth (~59-fold coverage) genomic resequencing revealed a two-base deletion introducing a frameshift mutation in exon three of one specific glycinin gene (gy5), which was completely correlated with the absence of the glycinin protein subunit. In addition to determining the cause of the seed protein alteration, the deep genomic resequencing provided the opportunity to catalog the genetic variation in this virtually unexamined landrace. A new SIFT4G database, which uses a phylogenetic approach to predict the effect of single nucleotide changes on protein function, was developed and validated. The new genetic polymorphism data and the new database will likely be of use in future gene-variant and gene-function studies in soybean.

### MATERIALS AND METHODS

#### Plant materials

Seed of line PI 603408 was obtained from the USDA-GRIN collection. PI 603408 was donated by the Chinese Academy of Agricultural Sciences to the USDA-GRIN collection in 1998, marked as originating from Liaoning province. Seed of line “Patriot” was obtained from David Sleper of the University of Missouri. PI 603408 was crossed during the summer of 2013 to Patriot and to line “10/81b,” which is a gy1a/gy1a and gy4/gy4 F4 line (PI 605781 B × Patriot) produced as part of a previous study (Kim et al. 2013). F2 seed from the two crosses were advanced two generations by single seed descent at a winter nursery. In 2014, F3 seeds from crosses were planted at South Farm Experimental farm (Columbia, MO, Latitude 38.908189, Longitude −92.78693, Mexico silt loam soil) and in 2015 at Hinkson Field (Columbia, MO, Latitude 38.928015, Longitude −92.351425, Raymond silt loam soil). Seeds were planted in 15 ft rows with a 3 ft gap between plots, a spacing of ~2 in between plants, and row spacing of 30 in between rows. In 2014, individual plants were tagged and leaf tissue from individual plants sampled, freeze-dried, and crude DNA was isolated as previously described (Xin et al. 2003) or high-quality DNA was isolated with a DNeasy plant mini kit (QIAGEN, Valencia, CA). When plants had matured in 2014, individual plants were single plant threshed and seeds were stored at 4° and 39% relative humidity. Seed was harvested from individual plants in 2014 and 20 seeds were planted as five single plots in 2015, with spacing as described above. At maturity, each plot in 2015 was bulk harvested and stored at conditions described above until analysis.

#### Genomic DNA resequencing analysis

High-quality DNA was isolated from ~50 mg of lyophilized seedling leaf tissue from five PI 603408 plants using a DNeasy mini plant kit according to manufacturer’s recommendations (QIAGEN). Library construction and sequencing were performed by Global Biologics. (Columbia, MO). Briefly, genomic DNA was fragmented to 100–300 bp and two replicate Illumina genomic DNA resequencing libraries (100–300 bp insert) were prepared and sequenced using two entire lanes on a HiSeq2000 (2 × 100 bp, paired end). A total of 679,476,503 reads were collected. FASTQ files were imported into the CLC genomics workbench (version 9, QIAGEN), and each lane was separately mapped (due to memory limitations) to the Williams 82 reference genome W82.a2.v1 (Schmutz et al. 2010) using the following settings: mismatch cost of 2, insertion cost = 3, deletion cost = 3, length fraction = 0.5, similarity fraction = 0.8, auto-detection of
paired distances = yes, and nonspecific matches = ignored (due to the ancestrally polyploid nature of soybean). The two lanes of mappings were then merged within the CLC genomics workbench. Full details on read mapping are listed in Table 1.

**Variant calling and effect prediction**

Sequence variants relative to the Williams 82 a2.v1 assembly were called using the basic variant detection function of CLC genomics workbench, using the following settings: ploidy = 2, ignore positions with coverage above = 35 cycles of 95° for 20 sec, 60° for 20 sec, and 72° for 20 sec, followed by a melting curve from 70° to 95°, with 20 readings taken every 1°.

**Prediction of impact of nonsynonymous AAC in PI 603408**

SIFT 4G (Vaser et al. 2016) was used to predict the effect of nonsynonymous AAC on protein function. A custom database of predictions for all possible nonsynonymous SNPs was built using SIFT 4G for G. max, using the W82.a2.v1 assembly. SIFT outputs whether an AAC is deleterious or tolerated, and assigns a score. As described previously (Ng and Henikoff 2001), an amino acid substitution is predicted deleterious if the SIFT score is ≤ 0.05, and tolerated if the score is > 0.05; SIFT scores range from 0 to 1. However, for certain genes, no meaningful prediction could be made. On various datasets, SIFT’s accuracy ranges from 70.82 to 84.86% (Vaser et al. 2016). Complete details on the coding region polymorphisms identified, read count, and SIFT 4G predictions are present in Supplemental Material, File S1, and gene annotation information for W82.a2.v1 is located in File S2 (downloaded from www.soybase.org). CLC genomics identifies four classes of variant in coding regions: Single Nucleotide Variant substitutions (SNV), Multiple Nucleotide Variant substitutions (MNV), deletions, insertions, and “Replacements” (a term describing variants that combine deletions/insertions). Variants identified in PI 603408 resequencing are also provided in Variant Call Format files for all variants identified in PI603408 and for nonsynonymous substitutions in File S3 and File S4, respectively.

**DNA isolation Gy1, Gy4, and Gy5 genotyping assays**

Genomic DNA was used with Gy1 and Gy4 and genotyping assays as previously described (Kim et al. 2013).

A genotyping assay was designed to detect the two-base deletion affecting gy5 in PI 603408, which relies on the introduction of “GC-tails” to a primer specific to each allele. A genotyping reaction has equal concentrations (0.5 μM in PCR reaction) of three primers: (1) a common primer that amplifies both alleles (5’-ACCATGACTTCTGGTCTG-3’); (2) a primer specific for wild-type Gy5 (5’-GGGGGCGCTTGGGAGCCCAAGTATA-3’); and (3) one specific for the gy5 deletion (5’-GGGGCAGCGCCGCTTGGGAACCACCATAG-3’). Bold indicates the allelic difference and underlines indicates the GC-tail. In addition to primers, each reaction contained 10 μl of 2×QuantiTect SYBR Green (QIAGEN), 5–50 ng of genomic DNA, and water sufficient to bring the volume to 20 μl. Samples were amplified and analyzed on a Lightcycler 480 II instrument under the following conditions: 95° for 5 min followed by 35 cycles of 95° for 20 sec, 60° for 20 sec, and 72° for 20 sec, followed by a melting curve from 70° to 95°, with 20 readings taken every 1°.

**NIRS protein and oil determination**

Seed moisture, oil, and protein were determined using ~50 intact seed with a NIRS monochromator model FOSS 6500 (FOSS North America, Eden Prairie, MN) using the transport quarter cup (dimension 97 mm × 55 mm) and a calibration previously developed (La et al. 2014) by Andrew Scaboo of the University of Missouri. Seed oil and protein values were adjusted to 13% moisture content before statistical analysis.

**Statistical analysis**

JMP Version 11 software (SAS Institute, Cary, NC) was used for one-way ANOVA tests. For any ANOVA tests that displayed significant differences at the P < 0.05 level, means were then compared using t-test ad hoc tests (α = 0.05 significance level cutoff). Full details on protein and oil data for the lines are present in File S5. Correlation between Gy5 protein band presence/absence and Gy5/gy5 genotypes are present in File S6.

**1D gel electrophoresis**

All chemicals described under electrophoresis and immunoblot analysis were obtained from Sigma Aldrich (St. Louis, MO). Soybean seeds were ground into a fine powder with a mortar and pestle and extracted with 1 ml of SDS-PAGE sample buffer containing protease inhibitor cocktail (Plant ProteaseArrest, G-Biosciences). The solution was centrifuged for 10 min to remove insoluble material at 16,100 × g. Supernatant was removed to a new tube and protein/buffer mixture was boiled for 5 min. A 10 μl aliquot of the boiled solution was used for electrophoresis. 1D separation was performed following the protocol of Laemmli (Laemmli 1970) using 13–15% T gels in a Mini250 apparatus (GE Healthcare). A constant current of 20 mA/gel was run with a typical run time of 1.2 hr. Following separation, gels were removed from the cassette, placed immediately in Coomassie R-250 staining solution, and destained with a 10% acetic acid solution.

**Immunoblot analysis**

Seed proteins were resolved by SDS-PAGE as previously described, then electrophoretically transferred onto a 0.45 μm nitrocellulose
membrane. Membranes were then incubated with 5% nonfat dry milk/TBS buffer (10 mM Tris-Cl, pH 7.5 and 500 mM NaCl) for 1 hr at room temperature. Following this step, membranes were incubated overnight with antibodies that had been diluted 1:20,000 in TBST (TBS with 3% nonfat dry milk containing 0.2% Tween 20). The following day, membranes were washed 3× with TBST and incubated with goat anti-rabbit IgG-horseradish peroxidase conjugate that had been diluted 1:20,000 in TBST. Proteins that reacted with antibodies were detected using a SuperSignal West Pico chemiluminescence kit (Pierce).

### 2D electrophoresis

Soybean seeds were ground to a fine powder and a 250 mg subsample was used for extraction with a cold mortar/pestle and 5 ml of extraction buffer [0.9 M sucrose, 0.1 M Tris-Cl (pH 8.8), and 0.4% 2-mercaptoethanol] as well as 50 μl of Plant ProteaseArrest (G-Biosciences). Samples were ground for 5 min until a liquid consistency was reached and removed to a 15 ml tube. Next, 5 ml of Tris-equilibrated phenol was added and phase separation was achieved using centrifugation (5000 × g, 20 min) via a swing-bucket rotor. The upper phenolic phase was removed to a fresh tube and proteins were precipitated using 10 volumes of freshly prepared 100% methanol with 0.1 M ammonium acetate for 2 hr at −80°C, followed by centrifugation (12000 × g, 20 min) at 4°C. The resulting protein pellet was thoroughly resuspended in freshly prepared ice-cold solution (100% methanol, 0.1 M ammonium acetate, and 10 mM DTT). Washing was repeated 3× with the same solution and 3× with freshly prepared 100% acetone containing 10 mM DTT (ice cold). Incubations of 30 min at −20°C were followed by centrifugation at 12000 × g for 10 min at 4°C between each wash.

400 μg of protein sample was loaded per IEF strip using in-gel rehydration. Linear gradient, 13 cm IPG strips (GE Healthcare) were brought to a rehydration volume of 250 μl with 7 M urea, 2 M thiourea, 1% CHAPS, 2% C7Bzo, 5% glycerol, and 2.2% 2-HED, containing protein sample. IEF strips were equilibrated (post-IEF) with 5% SDS in a urea-based solution (0.05 M Tris-Cl pH 8.8, 6 M urea, 30% glycerol, and 0.1% bromophenol blue) containing 2% DTT for 10 min, and again in a urea-based solution (0.05 M Tris-Cl pH 8.8, 6 M urea, 30% glycerol, and 0.1% bromophenol blue) containing 2% DTT for 10 min, and again with 0.1 M ammonium acetate for 2 hr at −80°C, followed by centrifugation (12000 × g, 20 min) via a swing-bucket rotor. The resulting protein pellet was thoroughly resuspended in freshly prepared ice-cold solution (100% methanol containing 10 mM DTT) at 4°C. Incubations of 30 min at −20°C were followed by centrifugation at 12000 × g for 10 min at 4°C between each wash.

### 2D image acquisition and analysis

1DE and 2DE Coomassie-stained gels were destained with multiple changes of ultrapure H2O to remove nonspecific background. Gels were scanned separately using an Epson V700 perfection scanner under control of Adobe Photoshop. 1DE Images were analyzed using Phoretix-Quant (TotalLab, Newcastle upon Tyne, UK) for band identification, location, and Rf, and 2DE images were analyzed using Delta2D V3.6 (Decodon, Greifswald, Germany); spot location calibration, and normalized % spot volume data were obtained using a technique known as differential gel imaging and analysis.

### Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are either fully represented within this article, with the exception of the in-gel digestion, peptide analysis, and mass spectrometry data, which can be obtained from the authors. The authors also note that all data necessary for calibrating and normalizing% spot volume data are included in the published figures. The authors state that all data necessary for calibrating and normalizing% spot volume data are included in the published figures. The authors state that all data necessary for calibrating and normalizing% spot volume data are included in the published figures. The authors state that all data necessary for calibrating and normalizing% spot volume data are included in the published figures. The authors state that all data necessary for calibrating and normalizing% spot volume data are included in the published figures.
archived in a sequence read archive, or are present within manuscript supplemental files. FASTQ files and BAM files have been archived at the NCBI sequence read archive under project PRJNA343126 and accession SRP090021.

RESULTS AND DISCUSSION

SDS-PAGE analysis of seed From PI 603408 confirmed the 42 kDa protein to be a Gy5 subunit

As part of ongoing investigations to find soybean lines with altered seed protein composition and/or content, PI 603408 was identified as a line that lacked a single protein band (Figure 1, arrow), as determined by 1D SDS-PAGE analysis. The molecular weight of this protein band was estimated to be 42 kDa. It has previously been demonstrated that soybean seed proteins can be preferentially precipitated from total seed extracts by the addition of calcium (Krishnan et al. 2009). Based on previous proteomic studies, the protein absent in PI 603408 was tentatively identified as Glycinin 5 (Glyma13g18450.1/Glyma.13g123500.1). Further confirmation was obtained by immunoblot analysis using antibodies raised against Gy5 protein (Figure 1B). Western blot analysis revealed that anti-Gy5 antibodies reacted strongly against two proteins with molecular weights of 42 and 40 kDa. The identity of the 40 kDa protein most likely represents the closely related glycinin gene family member Gy4. In contrast, soybean PI 603408 failed to accumulate the 42 kDa protein (Figure 1B), indicating that the 42 kDa protein is the Gy5 subunit.

Establishing a SIFT 4G database to assist evaluation of mutations identified by genome resequencing of PI 603408

In order to determine the molecular genetic cause for the protein band absence, an attempt was made to clone the gy5 gene by simple PCR utilizing several primer combinations. Glycinin-specific PCR amplification products of the expected sizes were obtained, yet sequence analysis revealed that all amplified sequences corresponded only to the Gy4 gene. This result is not surprising given the very high sequence homology of Gy4 and Gy5 genes (Nielsen et al. 1989). PI 603408 represents a completely untapped reservoir of genetic diversity, as well as a novel gy5 mutation. For both these reasons, the genome of PI 603408 was resequenced using two distinct Illumina genomic DNA resequencing libraries, which were mapped to the Williams 82 a2.v1 assembly reference sequence (Schmutz et al. 2010) using a very high level of sequencing depth (~59-fold coverage). This revealed a total of 1,858,185 nucleotide variants, which included: 1,542,261 SNV, 44,162 MNV, 129,082 insertions, 134,457 deletions, and 8223 replacements. A subset of the newly identified polymorphisms were predicted to result in...
AAC and/or impact protein coding genomic regions: 34,470 SNV, 1490 MNV, 1576 insertions, 1708 deletions (including the gy5 two-base deletion), and 100 replacements (more complicated rearrangements or deletion/insertions). Collectively, an estimated 39,344 nonsynonymous alterations were found in 13,590 genes, including 2285 predicted frameshift mutations and 241 stop codon gains. As validation for the small read mapping results, two known mutations were also identified: (1) a photoperiod response gene E2 (Glyma.10G21500) in PI 603408 has a mutation [1582A > T, Lys528* (Watanabe et al. 2011)] which truncates the open reading frame by converting a lysine residue to a stop codon; and (2) a single-base nucleotide deletion affecting the Glyma.01G214600 (182delT, Leu61fs) GmSGR2/D1 gene, which has been shown to be associated with retention of chlorophyll (i.e., staygreen) in soybean (Fang et al. 2014; Nakano et al. 2014).

The effect of a frameshift event on protein function/accumulation, for example due to deletion or insertion, is relatively easy to predict. Unfortunately, these events are relatively rare in natural populations; SNV-based amino acid substitution events are far more common. Two major hurdles in whole-scale genomic resequencing studies are: (1) analysis of less extreme polymorphism-predicted effects in a phylogenetic context; and (2) the need for automation owing to the extremely large number of variants identified in such studies. Toward this end, a custom SIFT 4G database was created for the current version of the soybean genome (W82.a2.v1), which can evaluate and predict the effect of SNV. This database will be publically available for download and open use by anyone (http://sift.bii.a-star.edu.sg/sift4g/). Out of 22,877 SNVs predicted by SIFT 4G, 3316 (of which 975 have low-confidence predictions) were predicted to be damaging, whereas 16,837 were predicted to be tolerated. The effect of another 2724 nonsynonymous substitutions (mutations or deletion/insertions) were predicted to be damaging, whereas 16,837 were predicted to be tolerated.

**Identification of loss-of-function mutations in seed-expressed genes**

The major seed-expressed glycinin and conglycinin genes in the genomic resequencing mapping were found for allelic differences in PI 603408 (Table 2) and a number of putative variants were identified. A two-base deletion was found within exon 3 of Glyma13g18450/Glyma.13g123500 Glycinin 5 (584_585delTA), which introduces a frameshift mutation (Ile61956, within exon 3). We also identified a number of nonsynonymous variants in other glycinin-encoding genes, though none were predicted to affect protein accumulation, as determined by SIFT analysis. (Table 2, full details of all nonsynonymous variants are in File S1). *A priori*, it might have been anticipated that mapping of small 100–200 bp sequencing reads to the complicated polyploid soybean genome could result in difficulties, particularly with the large multi-gene glycinin and conglycinin families. However, this does not appear to have been a significant hindrance.

**Molecular marker assays confirmed gy5 mutation is causative for absence of Gy5 subunit in seed of PI 603408**

The presence of multiple glycinin mutations could have the potential to result in pleiotropic effects on seed composition, particularly seed protein and/or oil content. In order to track the novel two-nucleotide gy5 deletion, a GC-tail molecular marker (Wang et al. 2005) was developed and tested on a segregating population derived from a cross between PI 603408 and a public cultivar, Patriot. DNA from individual F3 plants...
were genotyped, F$_3$4 seed from single plant threshes were harvested, and a subset analyzed via SDS-PAGE (Figure 2 and Figure 3) and with an NIRS calibration able to predict soybean seed protein and seed oil. Seed constituents from F$_6$5 lines planted in a completely randomized field experiment in 2015 were also analyzed by SDS-PAGE analysis and NIRS.

The gy5 molecular marker assay was 95% accurate (69/74) in predicting the presence of the glycinin subunit using a crude DNA preparation (Xin et al. 2003); a small number of heterozygotes were mistakenly called as homozygous gy5/gy5. When DNA was reisolated using a DNeasy plant mini kit (QIAGEN), the gy5 assay was 100% accurate (73/73). No evidence for segregation distortion was observed in either cross (data not shown), and no significant differences were noted between any genotypic classes (gy5/gy5; gy5/WT; WT/WT) for protein in 2014 or 2015 (Table 3). A very slight (0.1%) reduction in seed oil was noted for the homozygote mutant line (gy5/gy5) relative to the homozygote (WT/WT) genotypic class for the cross of (Patriot × PI 605781 B); this may indicate slight linkage drag.

**2D comparative analysis of the gy1/gy4/gy5 homozygote line in comparison to Patriot revealed substantial changes in only Gy1/Gy4/Gy5 proteins**

In previous work, we had begun to integrate two distinct glycinin mutations (gy1/gy1 and gy4/gy4) in the background of soybean cultivar Patriot (Kim et al. 2013). To develop a soybean line lacking Gy1/Gy4/Gy5 proteins, a second cross was made with PI 603408 crossed to a line derived from Patriot × PI 605781 B homozygous for two glycinin mutations (gy1/gy1 and gy4/gy4, ~50% genome from PI 605781 B and ~50% genome from Patriot). Due to the presence of three segregating genes and a smaller-sized population, the second set of RILs was only evaluated using F$_3$4 seed produced in 2015. No significant differences in seed oil or seed protein were noted between these lines in 2015 (Table 3); however, the loss of three different glycinin subunits has the potential to have proteome rebalancing pleiotropic effects on seed proteome composition (Schmidt and Herman 2008; Schmidt et al. 2011; Herman 2014). To evaluate this possibility, seed proteins were separated in two-dimensions (isoelectric focusing and SDS-PAGE) and compared with seed of a gy1/gy4/gy5 homozygous mutant in relation to Williams 82 seed (Figure 4). The absence of three glycinin protein subunits (as well as their precursors) was confirmed. Aside from the absence of Gy1/Gy4/Gy5 protein spots and slight differential protein band migration (presumably due to different isoforms), the seed proteomes were very comparable; little or no significant changes in β-conglycinin, Bowman–Birk proteinase inhibitors, Kunitz Trypsin Inhibitors, or Gy2 and Gy3 protein levels were noted (Figure 4).

Proteome rebalancing has been demonstrated to dramatically alter protein levels in lines with multiple genes whose expression has been reduced through RNAi (Schmidt et al. 2011). We saw no evidence of this phenomenon in the triple mutant genetic material. However, the lines in this study are not isogenic as is the case in RNAi studies; a substantial backcrossing effort would be required to generate true near-isogenic lines. Relatively large variances were noted for seed protein and oil gy1/gy4/gy5 genotypic RIL categories in comparison to parental line seeds (File S5), which we attribute to multiple independent genetic loci controlling seed traits between our parental lines. Although there appears to be no significant effect of the gy5 mutation alone (Table 3), there may be a small decrease in total protein in the triple mutant materials that is hidden by other genetic factors (or alternatively, small increases in other seed storage proteins). As a result, we cannot conclusively confirm or refute the proteome rebalancing hypothesis (Schmidt et al. 2011) with our genetic material at present.

**Conclusions**

Through proteomic analysis, an unimproved soybean landrace, PI 603408, was identified whose seeds lacked a glycinin protein subunit. Through whole genomic resequencing at a very high coverage depth (~59-fold), the molecular genetic cause was determined to be a two-base deletion that introduces a frameshift mutation in Glycinin 5. This was confirmed by cosegregation of the mutation with the absence of Gy5 protein in two independent segregating populations. The two-nucleotide deletion was found to have no significant effect on seed protein in field experiments over 2 yr. In addition, a total of 1,858,185 nucleotide variants were predicted by resequencing PI 603408, as compared to the reference genome Williams 82, and 39,344 variants were predicted to result in coding region changes, affecting 13,590 genes. A newly developed SIFT 4G database was
used to predict the effect of the SNV using ancestral conservation scoring across a range of diverse species. We anticipate that the new SIFT 4G database, as well as the extremely high coverage depth (average 59.1-fold) resequencing information for PI 603408, will prove useful in future soybean gene diversity and gene function studies.

ACKNOWLEDGMENTS

The authors would like to acknowledge the expert technical assistance of US Department of Agriculture, Agricultural Research Service (USDA-ARS) technicians Jeremy Mullis and Alexandria Berghaus. Mention of a trademark, vendor, or proprietary product does not constitute a guarantee or warranty of the product by the USDA and does not imply its approval to the exclusion of other products or vendors that may also be suitable. The USDA-ARS, Midwest Area, is an equal opportunity, affirmative action employer and all agency services are available without discrimination.

LITERATURE CITED


La, T. C., S. M. Pathan, T. Vuong, I.-D. Lee, A. M. Scaboo et al., 2014 Effect of high-oleic acid soybean on seed oil, protein concentration, and yield. Crop Sci. 54(5): 2054–2062.


Xin, Z., J. P. Velten, M. J. Oliver, and J. J. Burke, 2003 High-throughput DNA extraction method suitable for PCR. Biotechniques 34: 820–826.


Communicating editor: J. Ma