2016

Genomic Analysis of Storage Protein Deficiency in Genetically Related Lines of Common Bean (Phaseolus vulgaris)

Sudhakar Pandurangan

Marwan Diapari

Fuqiang Yin

Seth Munholland

Gregory E. Perry

See next page for additional authors

frontiers in Plant Science

ORIGINAL RESEARCH
published: 31 March 2016
doi: 10.3389/fpls.2016.00389

Genomic Analysis of Storage Protein Deficiency in Genetically Related Lines of Common Bean (*Phaseolus vulgaris*)

*Sudhakar Pandurangan[1,2†], Marwan Diapari[2], Fuqiang Yin[2,3], Seth Munholland[4], Gregory E. Perry[5], B. Patrick Chapman[2], Shangzhi Huang[3], Francesca Sparvoli[6], Roberto Bollini[6], William L. Crosby[4], Karl P. Pauls[5] and Frédéric Marsolais[1,2]**

[1] Department of Biology, University of Western Ontario, London, ON, Canada, [2] Genomics and Biotechnology, London Research and Development Centre, Agriculture and Agri-Food Canada, London, ON, Canada, [3] Department of Bioscience and Biotechnology, School of Life Sciences, Sun Yat-sen University, Guangzhou, China, [4] Department of Biological Sciences, University of Windsor, Windsor, ON, Canada, [5] Department of Plant Agriculture, University of Guelph, Guelph, ON, Canada, [6] Institute of Agricultural Biology and Biotechnology, National Research Council, Milan, Italy

OPEN ACCESS

Edited by:
Paul Gepts,
University of California, Davis, USA

Reviewed by:
Alejandra A. Covarrubias,
Universidad Nacional Autónoma de México, Mexico
Eliot Herman,
University of Arizona, USA

***Correspondence:**
Frédéric Marsolais
frederic.marsolais@agr.gc.ca

†Present address:
Sudhakar Pandurangan, Breeding and Agronomy, Brandon Research and Development Centre, Agriculture and Agri-Food Canada, Brandon, MB, Canada

Specialty section:
This article was submitted to Plant Genetics and Genomics, a section of the journal Frontiers in Plant Science

Received: *14 January 2016*
Accepted: *14 March 2016*
Published: *31 March 2016*

Citation:
Pandurangan S, Diapari M, Yin F, Munholland S, Perry GE, Chapman BP, Huang S, Sparvoli F, Bollini R, Crosby WL, Pauls KP and Marsolais F (2016) Genomic Analysis of Storage Protein Deficiency in Genetically Related Lines of Common Bean (Phaseolus vulgaris). Front. Plant Sci. 7:389. doi: 10.3389/fpls.2016.00389

A series of genetically related lines of common bean (*Phaseolus vulgaris* L.) integrate a progressive deficiency in major storage proteins, the 7S globulin phaseolin and lectins. SARC1 integrates a lectin-like protein, arcelin-1 from a wild common bean accession. SMARC1N-PN1 is deficient in major lectins, including erythroagglutinating phytohemagglutinin (PHA-E) but not α-amylase inhibitor, and incorporates also a deficiency in phaseolin. SMARC1-PN1 is intermediate and shares the phaseolin deficiency. Sanilac is the parental background. To understand the genomic basis for variations in protein profiles previously determined by proteomics, the genotypes were submitted to short-fragment genome sequencing using an Illumina HiSeq 2000/2500 platform. Reads were aligned to reference sequences and subjected to *de novo* assembly. The results of the analyses identified polymorphisms responsible for the lack of specific storage proteins, as well as those associated with large differences in storage protein expression. SMARC1N-PN1 lacks the lectin genes *pha-E* and *lec4-B17*, and has the pseudogene *pdlec1* in place of the functional *pha-L* gene. While the α-phaseolin gene appears absent, an approximately 20-fold decrease in β-phaseolin accumulation is associated with a single nucleotide polymorphism converting a G-box to an ACGT motif in the proximal promoter. Among residual lectins compensating for storage protein deficiency, mannose lectin FRIL and α-amylase inhibitor 1 genes are uniquely present in SMARC1N-PN1. An approximately 50-fold increase in α-amylase inhibitor like protein accumulation is associated with multiple polymorphisms introducing up to eight potential positive *cis*-regulatory elements in the proximal promoter specific to SMARC1N-PN1. An approximately 7-fold increase in accumulation of 11S globulin legumin is not associated with variation in proximal promoter sequence, suggesting that the identity of individual proteins involved in proteome rebalancing might also be determined at the translational level.

Keywords: genome sequencing, introgression, deletion, lectin, phaseolin, common bean, *Phaseolus vulgaris*

INTRODUCTION

Storage protein deficiency in crops is compensated through a mechanism of proteome rebalancing, whereby seed protein concentration is maintained at its normal level (Herman, 2014; Wu and Messing, 2014). This property has been used to express foreign recombinant protein (Schmidt and Herman, 2008; Lin et al., 2013; Hegedus et al., 2014) and for protein quality improvement (Kita et al., 2009; Wu et al., 2013; Kim et al., 2014). In soybean, because seed protein composition influences tofu quality, research has been performed to identify genetic variants for major seed proteins (Liu et al., 2006; Hayashi et al., 2009; Tsubokura et al., 2012; Kim et al., 2013; Wang et al., 2014) and to mobilize this genetic variation into cultivated varieties through marker-assisted selection (Jegadeesan et al., 2012; Song et al., 2014).

Common bean (dry bean, *Phaseolus vulgaris*) is the most important food legume for direct human consumption. A set of genetically related lines integrating a progressive deficiency in major storage proteins has been described (Osborn et al., 2003). The 7S globulin and major lectins are encoded at two unique loci. The major lectin or arcelin-phytohemagglutinin-α-amylase inhibitor (APA) locus in SARC1 is derived from the wild accession G12882 and includes the insecticidal lectin arcelin-1. SMARC1-PN1 and SMARC1N-PN1 integrate a deficiency in phaseolin introduced from a *Phaseolus coccineus* accession. SMARC1N-PN1 further integrates a lectin deficiency from the cultivar Great Northern 1140. The three lines share a common genetic background from the cultivar Sanilac. The deficiency in phaseolin and lectins is associated with an increased concentration of sulfur amino acids, cysteine and methionine, primarily at the expense of the non-protein amino acid, *S*-methylcysteine, and increased levels of sulfur-rich proteins (Taylor et al., 2008; Marsolais et al., 2010; Yin et al., 2011; Liao et al., 2012). This property is of interest to improve protein quality and relevant to nutritional claims on protein content. The changes in protein composition are associated with increased protein solubility (Taylor et al., 2008).

The objective of the present study was to characterize the genetic polymorphisms responsible for differences in phaseolin and lectin expression between SARC1 and SMARC1N-PN1. To do so, a combination of approaches was used, including re-analysis of a quantitative proteomic dataset coupled with Western blotting and affinity purification, genomic PCR and genomic sequencing. The results identify several polymorphisms associated with storage protein deficiency and shed light on the process of proteome rebalancing in crop seeds.

MATERIALS AND METHODS

Plant Material and Growth

Common bean (*Phaseolus vulgaris* L.) genotypes were grown in a growth cabinet (Environmental Growth Chambers, Chagrin Falls, OH, USA) under 16 h light (300–400 μmol photons m^{-2} s^{-1}) and a temperature cycling between 18 and 24°C (Pandurangan et al., 2012). The generation of SARC1, SMARC1-PN1 and SMARC1N-PN1 genetic stocks was described by Osborn et al. (2003). Seeds from parents G12882 and Great Northern 1140 were obtained from the Germplasm Resources Information Network of the United States Department of Agriculture-Agricultural Research Service, Western Regional Plant Introduction Station, Pullman, WS, USA. A number of *Phaseolus coccineus* seeds lacking phaseolin, originally characterized at the CNR in Pisa, Italy (Durante et al., 1989), that are commonly found in local markets in Tuscany, Italy and were kindly provided by Luccarini, were confirmed by examining seed protein profiles for the absence of phaseolin. Mature seed tissue (100 mg) was homogenized in 0.5 × Sample Buffer [4% SDS, 25 mM Tris–HCl pH 6.8, 2.5% (v/v) glycerol] to extract total protein. The extracts were boiled immediately for 5 min, centrifuged for 15 min at room temperature and supernatants were saved. Protein concentration was determined using the Bio-Rad Protein Assay solution (Mississauga, ON, Canada) and bovine serum albumin as standard. Equal amount of protein was separated by SDS-PAGE on a 10% polyacrylamide gel.

Protein Analysis by Spectral Counting

Quantitative proteomic data (Marsolais et al., 2010) were re-analyzed using Scaffold 2 software (Proteome Software Inc., Portland, OR, USA) against the UniProt database, section *Viridiplantae* (as of March 31, 2009).

Purification of Mannose Lectin FRIL

Mannose lectin FRIL was purified by affinity chromatography on D-mannose agarose and eluted competitively with methyl α-D-mannopyranoside (Sigma–Aldrich, Oakville, ON, Canada) as described by Colucci et al. (1999). The identity of protein bands was confirmed by LC-MS after tryptic digestion as described in (Marsolais et al., 2010). The peak list was searched against NCBInr/Other green plants using Mascot[1].

Analysis of α-Amylase Inhibitor 1 by Western Blot

Mature seed tissue (100 mg) was extracted and protein quantified as described above. Equal amount (2 μg) separated by SDS-PAGE on a 15% polyacrylamide gel was transferred to a nitrocellulose membrane (9 cm × 6 cm) at 15 V for 20 min using a semi-dry transfer apparatus (Bio-Rad Laboratories, Inc.). The membrane was blocked with Odyssey Blocking Buffer (LI-COR Biosciences, Lincoln, NE, USA) at room temperature for 1 h. The membrane was incubated with 1:2000 dilution of anti-α-amylase inhibitor antibodies (Lioi et al., 2007) for 1 h, followed by 1:10,000 dilution of goat IRDye800R Conjugated Affinity Purified Anti-Rabbit IgG (Rockland Immunochemicals Inc., Limerick, PA, USA) for 1 h. Immunodetection was achieved

[1]http://www.matrixscience.com

by scanning with an Odyssey Infrared Imaging System (LI-COR). Bands were quantified using ImageStudio ver. 3.1 software (LI-COR).

DNA Isolation and PCR Genotyping

Leaf tissue was frozen in liquid nitrogen and ground to a fine powder using a mortar and pestle. Genomic DNA was isolated using the GenElute Plant Genomic DNA Miniprep Kit (Sigma–Aldrich) following the manufacturer's protocol. PCR was carried out using 50 ng of genomic DNA as template for 35 cycles with Taq DNA polymerase and the following gene specific primers: for *ARC1*, F: 5′-AGCAACGACGCCTCCTTCAACG-3′ and R: 5′-CCTTTAAGTTTGGGCCGAGAGCCG-3′; for *arc3-II*, F: 5′-ACTAGCTTCCACCAAGGCGATCC-3′ and R: 5′-TTCTGTC ATAGCGGAGGGTGTAGC-3′; for *arc4-I*, F: 5′-AGTATCCG CCCATACAGTAACAATG-3′ and R: 5′-CACGCTGCTGGTA GAGAAGTTG-3′; for *pha-E*, F: 5′-CGCACACACTTGC AACATCCC-3′ and R: 5′-GGTTTGGGGTCCCAGTGAACG T-3′; for α-amylase inhibitor 1, F: 5′-GAAACCTCCTTCAAC ATCGATGG-3′ and R: 5′-CCCTCACCCAGTCGTAAACTTCT-3′; and for mannose lectin FRIL, F: 5′-GTGGAGGAAACCC TGTGGGTGC-3′ and R: 5′-CGGCTCCTTCACCTCGTTGTTC T-3′. The following primers were used to confirm the presence of *pdlec1* in SMARC1N-PN1 and Great Northern 1140: PhaL-F165, 5′-CTCCTCTTCTCACTATGACAC-3′ and PhaL-R838, 5′-GACTCCAAACTCCACCTTCC-3′. The following primers were used to amplify the β-phaseolin promoter: pBetaPhsF, 5′-CCTTTCTTGGTATGTAAGTCCG-3′ and pAlphaBetaPhsR, 5′-AGTAGAGTAGTATTGAATATGAGTTG-3′. PCR products were sequenced using a 3130XL Genetic Analyzer (Life Technologies, Burlington, ON, Canada).

Next Generation Sequencing

DNA was isolated using a Qiagen DNeasy Plant Mini Kit (Toronto, ON, Canada). Care was taken to isolate intact genomic DNA. To minimize shearing, the samples were not vortexed and wide-bore pipette tips were used for handling. After the final wash, the samples were eluted in 100 µl of 10 mM Tris-HCl pH 8.0. DNA samples were visualized on a 1% agarose gel. DNA concentration and purity was determined using a Nanodrop 1000 (Thermo Scientific, Wilmington, DE, USA). Genomic DNA samples from Sanilac, SARC1, SMARC1-PN1 and SMARC1N-PN1 were submitted for paired-end read sequencing on an Illumina HiSeq 2000/2500 platform (San Diego, CA, USA) at the Clinical Genomics Centre, Toronto, ON, Canada, following recommended guidelines. Low quality reads were filtered out, resulting in approximately 925–1025 million reads per sample. In addition to paired-end read sequencing, three mate-pair libraries were prepared, short (3.5–4.5 kb), medium (5–7 kb), and large (8–11 kb) using Illumina's protocol to obtain 50 base pair reads. Samples from each size were multiplexed and run on a single lane. Sequencing data can be found in the short read archive at the National Center for Biotechnology Information, with the following accession numbers: for Sanilac, SRP055506; for SARC1, SRP055509; for SMARC1N-PN1, SRP055510; and for SMARC1-PN1, SRP055511.

Paired-end reads were aligned to the *P. vulgaris* G19833 genome sequence (Schmutz et al., 2014), scaffold assemblies of the BAT-93 genome (Vlasova et al., 2016) and of the OAC-Rex genome (Perry et al., unpublished results[2]), a BAC clone for the APA locus of an arcelin-5 genotype (Kami et al., 2006) and α-phaseolin gene sequences from Sanilac (Anthony et al., 1990; Diniz et al., 2014), as described in O'Rourke et al. (2013). Reads were aligned with BWA using default parameters (Li and Durbin, 2009). Sequence Alignment/Map (SAM) files generated were converted to sorted indexed Binary Alignment/Map (BAM) files using SAMtools (Li et al., 2009). Alignments were visualized with IGV (Robinson et al., 2011).

To assemble paired-end and mate-pair read data, sequencing reads were analyzed by performing FastQC[3] to get the read profile and evaluate possible sequence contamination. A modified blosum filter was applied to remove duplicate reads using pybloomfaster[4]. Low quality reads were removed using fastq_quality_filter[5]. A custom adapter trimmer was performed to remove any contaminating sequence identified by FastQC. Reads were reorganized by title for the synchronizer using a custom sort script. A custom synchronizer script was run to ensure that the R1 and R2 files contained the same reads in the same order after filtering. Orphans were saved, but not used. Trimmed files were archived. Paired end reads were assembled into contigs using Ray (Boisvert et al., 2010). A custom contig fractionator script was run to generate a set of artificial paired end reads with controlled overlap for use in the scaffolding. Scaffolding was performed using ALLPATHS-LG (Gnerre et al., 2011). Assembly stats were cross checked with assemblathon_stats.pl. CEGMA was used to identify 248 Core Eukaryotic Genes (CEGs) as an indirect measure of functional completeness of the assembly (Parra et al., 2009).

Promoter Analysis

Proximal promoter sequences were analyzed and compared using a database of plant *cis*-acting regulatory DNA elements, PLACE[6] (Higo et al., 1999).

Accession Numbers

Additional nucleotide sequence data from this article have been deposited in the GenBank database under accession numbers: [GenBank ID: KU258848] for mannose lectin FRIL from SMARC1N-PN1; [GenBank ID: KU258849] from G12882; [GenBank ID: KU258850] from *P. coccineus*; and [GenBank ID: KU258846] for α-amylase inhibitor like protein from SARC1; [GenBank ID: KU258847] from SMARC1-PN1; and [GenBank ID: KU258845] from SMARC1N-PN1.

[2]http://www.beangenomics.ca

[3]http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[4]https://github.com/brentp/pybloomfaster

[5]http://hannonlab.cshl.edu/fastx_toolkit/

[6]https://sogo.dna.affrc.go.jp/cgi-bin/sogo.cgi?sid=&lang=en&pj=640&action=page&page=newplace

RESULTS

Proteomic Analysis of Phaseolin and Lectin Composition in SARC1 and SMARC1N-PN1

To understand the effect of storage protein deficiency on the composition of phaseolins and lectins, shotgun proteomic data from total protein extracts from SARC1 and SMARC1N-PN1 (Marsolais et al., 2010) was re-analyzed and quantified with SCAFFOLD software. The results are presented in **Table 1**. SCAFFOLD is particularly adept at assigning spectra to a given accession among a group of closely related proteins, although in all cases, the algorithm reported protein grouping ambiguity except for arcelin-like protein 4 and α-amylase inhibitor like protein. The results confirmed the absence of α-phaseolin and the residual levels of β-phaseolin (Phaseolin precursor, encoded by *Phs*) present in SMARC1N-PN1. In prior analyses, the presence of β-phaseolin in SMARC1N-PN1 had been inferred from the results of two-dimensional gel electrophoresis based proteomics (Marsolais et al., 2010). For lectins, the present results suggest that there are three distinct arcelins as well as arcelin-like protein 4 in SARC1. This new analysis confirms the absence of lectins encoded by *lec4-B17* and *pha-E* in SMARC1N-PN1. Partial compensation by a leucoagglutinating phytohemagglutinin, encoded by *PDLEC2* (Voelker et al., 1986), α-amylase inhibitor like protein, α-amylase inhibitor 1 and mannose lectin FRIL are also apparent in these data.

Mannose Lectin FRIL and α-Amylase Inhibitor 1 Are Uniquely Present in SMARC1N-PN1

For lectins detected at relatively low levels by spectral counting, it was not clear whether they are truly present or whether they are detected based on their high sequence similarity with other lectins. This was further investigated for mannose lectin FRIL and α-amylase inhibitor 1. Mannose lectin FRIL was affinity purified from mature seed of Sanilac, SARC1, SMARC1-PN1 and SMARC1N-PN1 on mannose-agarose and the purified protein analyzed by SDS-PAGE. Protein bands corresponding to mannose lectin FRIL were uniquely present in SMARC1N-PN1 (**Figure 1A**). Three bands were observed having apparent molecular masses of 20, 17, and 16 kDa, respectively. The first one constitutes the N-terminal subunit and the other two the C-terminal subunit (Moore et al., 2000). This was confirmed by a proteomics approach, based on the coverage of each subunit by identified tryptic peptides (**Table 2**).

α-Amylase inhibitor 1 was immunodetected in mature seed protein extracts of Sanilac, SARC1, SMARC1-PN1 and SMARC1N-PN1 using polyclonal antibodies raised against recombinant α-amylase inhibitor. Two major bands of approximately 14.5 and 13 kDa were detected (**Figure 1B**), corresponding to chain 2 and chain 1 of α-amylase inhibitor 1, respectively (Moreno and Chrispeels, 1989; Yamaguchi, 1991). No signal was detected in SARC1 and SMARC1-PN1. Protein levels were higher in SMARC1N-PN1 than in Sanilac, as determined by quantification of the main protein band corresponding to chain 2, by approximately 2.5-fold.

Analysis and Validation of Lectin Gene Composition by Genomic PCR

Based on the above results, analysis of lectin gene composition was conducted by genomic PCR using primers complementary to the coding sequence or, where possible, to the 5′-untranslated region. Samples included the three genetically related lines as well as the parental background Sanilac, the two other parents, G12882 and Great Northern 1140 and a *P. coccineus* phaseolin deficient genotype, supposed to bear the same *phs* null allele of the SMARC1N-PN1 line. The genomic PCR results confirmed the presence of three different arcelin genes in SARC1, SMARC1-PN1, and G12882, the source of arcelin genes in the two lines (**Figure 2A**). In addition, no amplification of

TABLE 1 | Differentially expressed phaseolins and lectins in mature seeds of SARC1 and SMARC1N-PN1 quantified by spectral counting, as unweighted spectrum count, with a minimum of 1 peptide identified with 95% probability (average ± standard deviation); *n* = 3; n.s., not significant.

Protein	Gene name	UniProt accession	SARC1	SMARC1N-PN1	*t*-test *p*-value
Phaseolin precursor	*Phs*	Q43632	826 ± 175	40 ± 34	0.002
α-Phaseolin		Q41115	799 ± 152	0	0.0008
Phaseolin, α-type		P07219	713 ± 122	0	0.0005
Arcelin-1 precursor	*ARC1*	P19329	955 ± 273	0	0.004
Arcelin	*arc3-II*	Q8RVY3	67 ± 11	0	0.0004
Arcelin-like protein 4	*arl4*	Q8RVX7	90 ± 8	0	0.00004
Arcelin	*arc4-I*	Q8RVX4	73 ± 3	0	0.000003
Lectin precursor	*lec4-B17*	Q8RVX5	27 ± 11	0	0.015
Phytohemagglutinin	*pha-E*	Q8RVX6	83 ± 41	0	0.03
Phytohemagglutinin	*pha-L*	Q8RVH2	72 ± 59	7 ± 1	n.s.
Leucoagglutinating phytohemagglutinin	*PDLEC2*	P15231	41 ± 28	121 ± 38	0.04
α-Amylase inhibitor like protein		Q9SMH0	3 ± 1	150 ± 11	0.00002
α-Amylase inhibitor 1		A0T2V3	17 ± 17	311 ± 32	0.0001
Mannose lectin FRIL		Q9M7M4	3 ± 2	116 ± 16	0.0002

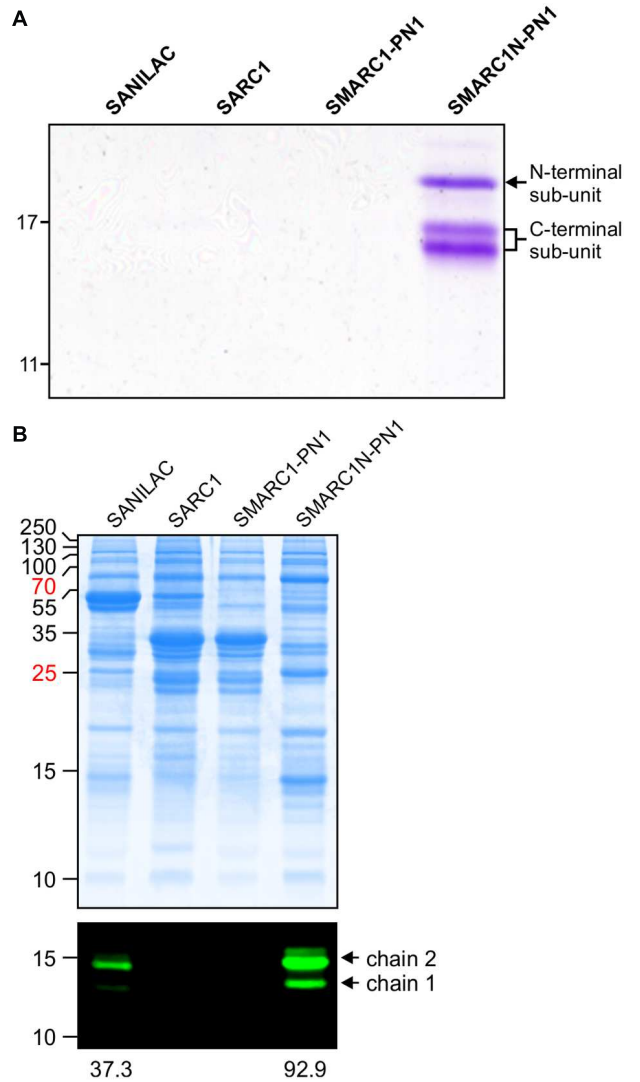

FIGURE 1 | (A) Affinity purification of mannose lectin FRIL. SDS-PAGE of mature seed extracts after purification on mannose agarose. The band of 20 kDa corresponds to the N-terminal subunit of mannose lectin FRIL. The bands of 17 and 16 kDa correspond to the C-terminal subunit. The identity of protein bands was confirmed by proteomics (see **Table 2**). **(B)** Immunoblotting of total protein extracts from mature seed with polyclonal antibodies raised against recombinant α-amylase inhibitor. Bands of 14.5 and 13 kDa correspond to chain 2 and chain 1 of α-amylase inhibitor 1, respectively.

pha-E, encoding erythroagglutinating phytohemagglutinin, was observed in SMARC1N-PN1 and in Great Northern 1140, the source of lectin deficiency. The genomic PCR data confirmed the presence of the α-amylase inhibitor 1 gene in SMARC1N-PN1 and its absence in SARC1, SMARC1-PN1 and G12882. The α-amylase inhibitor 1 gene was also detected in Sanilac. Mannose lectin is encoded on chromosome 7 and not in the APA locus which is situated chromosome 4. The mannose lectin gene was found to be present in SMARC1N-PN1, G12882 and the *P. coccineus* genotype. Alignment of conceptual translations of PCR products indicated that mannose lectin FRIL originates from G12882 in SMARC1N-PN1, and was likely lost during crossing and propagation of the lines that led to SARC1 and SMARC1-PN1 (**Figure 2B**).

Genome Sequencing

To gain more insight into the polymorphisms associated with storage protein deficiency, the genomes of the three genetically related lines, SARC1, SMARC1-PN1 and SMARC1N-PN1 and of the recurrent parent, Sanilac were sequenced using a whole-genome shotgun sequencing approach which combined Illumina sequenced fragment libraries to obtain 100 bp paired end reads along with mate-pair libraries of fragments of three different lengths to assist *de novo* assembly, with a sequence read coverage of the estimated genome size greater than 150-fold (Supplementary Table S1). Two different approaches were used to analyze the data. In the first approach, paired end reads were mapped to a reference sequence using Burrows–Wheeler Aligner software. In the second approach, scaffold assemblies of the four genomes were generated using ALLPATHS-LG and analyzed for the genes of interest (Supplementary Table S2).

Absence of *lec4-B17* and *pha-E* and Presence of the Pseudogene *pdlec1* in SMARC1N-PN1

For analysis of the APA locus, BAT-93, a Mesoamerican genotype, was most similar to SMARC1N-PN1. **Figure 3A** shows the alignment of the paired end reads to the part of the BAT-93 scaffold00141 containing the APA locus, visualized with IGV. Sequences in gray are identical. Color highlights variant bases. Peak height indicates the number of reads aligned. BAT-93 and other genomic templates were annotated manually after blastn against NCBInr and blastx of individual APA coding sequences against UniProt, based on highest sequence identity to a known lectin accession. In order to annotate the genes in the alignments, reads were joined manually to generate a coding sequence which was used for blastx against UniProt. The gene order was found to be conserved across reference genotypes (BAT-93, G19833, OAC-Rex and the BAC-71F18 from the arcelin-5 genotype). However, the composition of APA genes varied. For the phytohemagglutinin gene located between *pha-E* and the α-amylase inhibitor like protein gene (**Figure 3A**), different alleles were found to be present. G19833 and OAC-Rex have *pha-L*, as do Sanilac, SARC1 and SMARC1-PN1. BAT93 and SMARC1N-PN1 have the *pdlec1* pseudogene, previously characterized from Pinto UI111 (Voelker et al., 1986). The presence of the *pdlec1* pseudogene in SMARC1N-PN1 and in Great Northern 1140 was confirmed by PCR amplification and sequencing of the PCR products. The sequences isolated were 100% identical to that reported by Voelker et al. (1986). The *pdlec1* allele is characterized by a deletion of a single nucleotide, cytosine, after position 32 of the coding sequence, resulting in a premature stop codon at position 132. BAT93 and SMARC1N-PN1 also share the *PDLEC2* gene, coding for a leucoagglutinating phytohemagglutinin isoform, further extending the homology with Pinto UI111 (Voelker et al., 1986). G02771, a wild, arcelin-5 genotype, has the arcelin-5 phytohemagglutinin (Kami et al., 2006). Arcelin-5

TABLE 2 | Identification of protein bands from SMARC1N-PN1 in Figure 1A by LC-MS-MS and Mascot search following trypsin digestion.

Protein band apparent molecular mass (kDa)	Name	GI number	Score	Matches	Number of peptides matching N-terminal subunit	Number of peptides matching C-terminal subunit	Predicted mass of respective subunit (Da)	Coverage of respective subunit (%)
20	Mannose lectin FRIL	6822274	284	35	35	0	14242.7	33
17	Hypothetical protein	593675212	520	22	2	20	14992.7	74
16	Mannose lectin FRIL	6822274	566	34	4	30	16877.8	41

The accession named hypothetical protein represents mannose lectin FRIL in the reference G19833 genome (Phytozome accession number PHAVU_007G070100g).

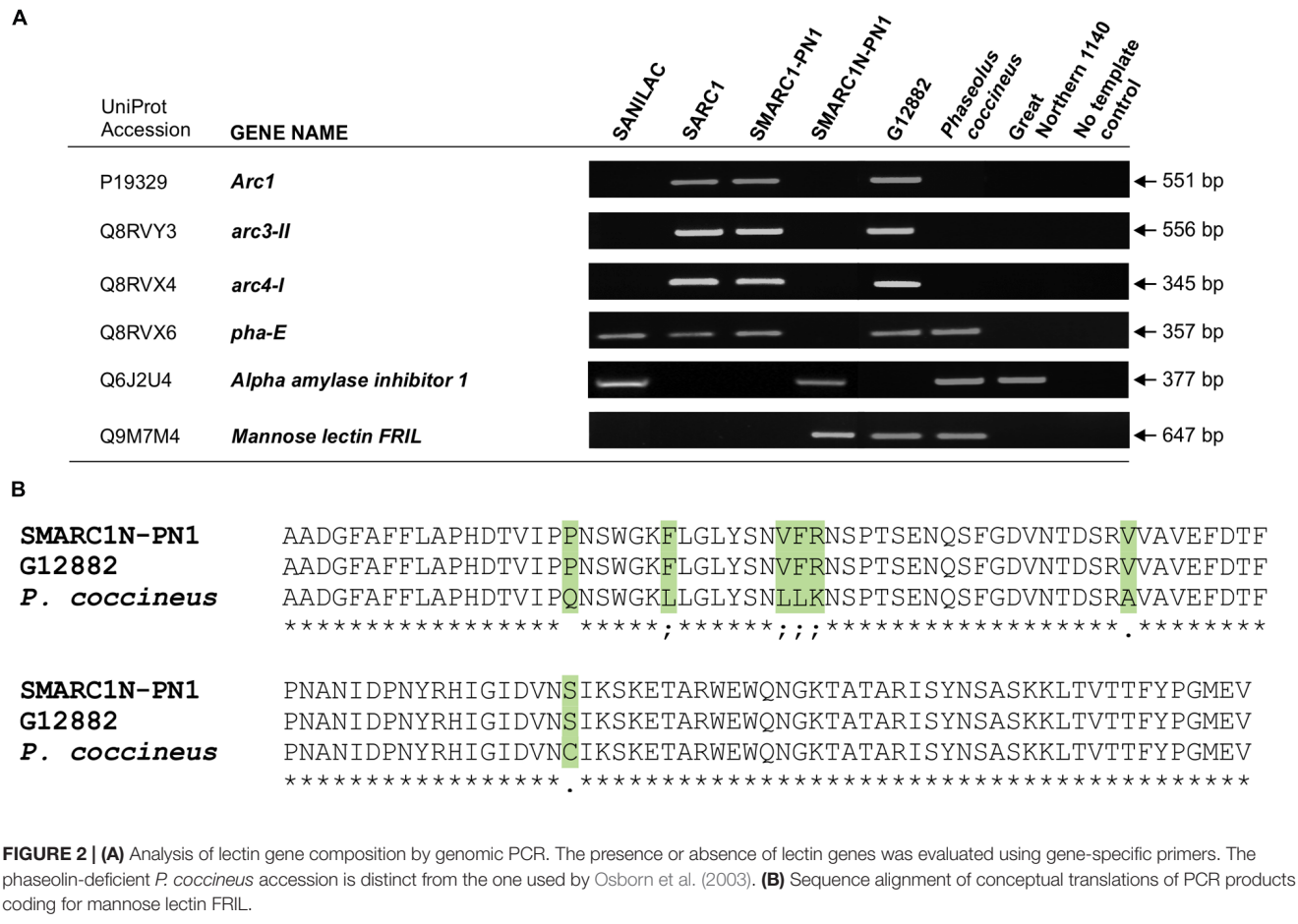

FIGURE 2 | (A) Analysis of lectin gene composition by genomic PCR. The presence or absence of lectin genes was evaluated using gene-specific primers. The phaseolin-deficient *P. coccineus* accession is distinct from the one used by Osborn et al. (2003). **(B)** Sequence alignment of conceptual translations of PCR products coding for mannose lectin FRIL.

phytohemagglutinin is 99% identical to *pdlec1*, but is not a pseudogene. The alignment in **Figure 3A** confirmed the absence of *lec4-B17* and *pha-E* in SMARC1N-PN1. The alignment also suggested the absence of *PDLEC2* and α-amylase inhibitor like protein gene in Sanilac. *PDLEC2* and the α-amylase inhibitor 1 gene appeared only partially covered in SARC1 and SMARC1-PN1 suggesting their absence in these genotypes. This conclusion is supported by the Western blotting and genomic PCR data for α-amylase inhibitor 1 (**Table 1, Figures 1** and **2**). It was not possible to verify the presence of *PDLEC2* by genomic PCR due to high degree of sequence identity between leucoagglutinating phytohemagglutinin genes.

Multiple Polymorphisms in the Promoter of α-Amylase Inhibitor Like Protein Are Associated with Increased Expression in SMARC1N-PN1

The scaffold assemblies of SARC1, SMARC1-PN1 and SMARC1N-PN1 contained a full length coding sequence for α-amylase inhibitor like protein (Supplementary Table S3). This is consistent with alignments of paired end reads (**Figure 3A**). Promoter sequences were aligned. Polymorphic sites were searched for differences in *cis* regulatory motifs between SMARC1N-PN1 versus SARC1 and SMARC1-PN1 using the PLACE database (Higo et al., 1999). This analysis revealed the presence of multiple individual positive *cis*-regulatory motifs

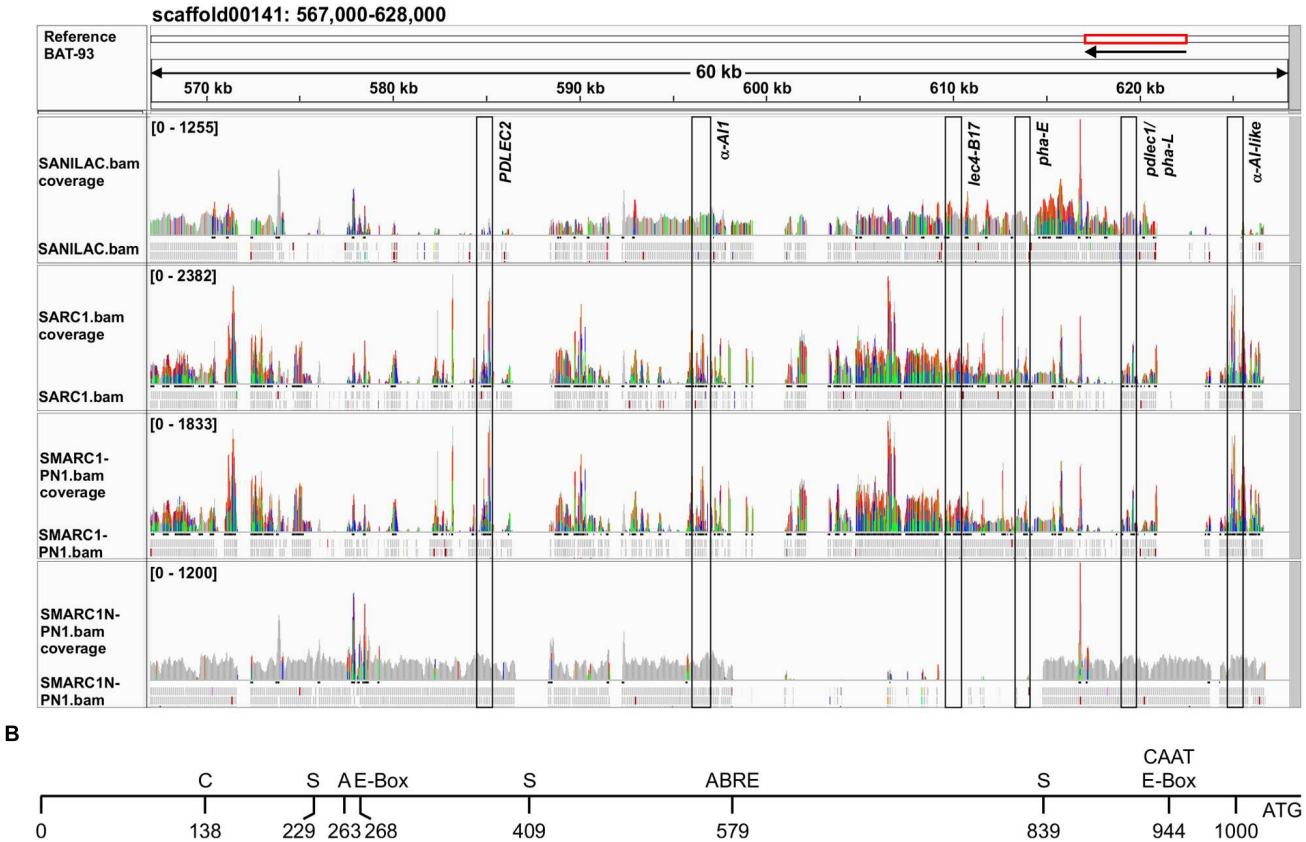

FIGURE 3 | (A) Sequence reads of lectin locus aligned to the reference BAT-93 genome. Paired end reads from the four genotypes sequenced by Illumina HiSeq2000/2500 were aligned and visualized with IGV. The positions of the lectin genes are indicated by boxes. Gray indicates sequence identity. Color highlights variant bases compared to the reference sequence; Green for A; Red for T; Orange for G; and Blue for C. *PDLEC2* and α-amylase inhibitor like protein gene (α-*AI-like*) are not covered in Sanilac; likewise α-amylase inhibitor 1 (α-*AI1*) in SARC1 and SMARC1-PN1. BAT-93 and SMARC1N-PN1 have the *pdlec1* allele in place of *pha-L*. **(B)** Schematic view of polymorphisms in the promoter of α-amylase inhibitor like protein between SMARC1N-PN1 versus SARC1 and SMARC1-PN1 giving rise to unique *cis*-regulatory elements in SMARC1N-PN1. Analysis of *cis*-regulatory elements was performed using the Place database. C, CAAACAC element; S, soybean embryo factor 4 binding motif; A, core AACA motif; ABRE, abscisic acid related element.

that are unique to SMARC1N-PN1 (**Figure 3B**). These include a CAAACAC element characterized in the napin promoter of *Brassica napus* (Stålberg et al., 1996), three instances of the soybean embryo factor 4 binding motif characterized by Lessard et al. (1991), a core AACA motif (5′-AACAAAC-3′) present in the rice glutelin promoter (Wu et al., 2000), and an abscisic acid related element (5′-ACGTGGC-3′) required for *RD29B* expression in *Arabidopsis* seed (Nakashima et al., 2006). There are also two instances of the E-box (Stålberg et al., 1996), the second overlapping with a CAAT box, proximal to the start codon (Shirsat et al., 1989).

Differences in β-Phaseolin Accumulation Correlate with a Single Nucleotide Polymorphism Converting a G-Box Motif into an ACGT Motif in the Promoter of SMARC1N-PN1

Several functional regions within the β-phaseolin promoter have been defined by deletion analyses (Bustos et al., 1991; van der Geest and Hall, 1996; Chandrasekharan et al., 2003). These include four RY repeat motifs (5′-CATGC/TA-3′) (Bobb et al., 1997), a G-box binding motif (5′-CACGTG-3′) and E-box motif (5′-CACCTG-3′) (Kawagoe et al., 1994), CACA element (Li et al., 1999), vicillin box (Chern et al., 1996a,b), ACGT motif and CAAT box (5′-CCAAAT-3′ in *Phs* promoter) (Li et al., 1999) (**Figure 4A**). Deletion analysis studies previously showed that the G-box, RY motifs, E-box and CAAT box are required for high level expression of a reporter in transgenic *Arabidopsis* seed (Chandrasekharan et al., 2003). Binding of the B3-domain containing VP1/ABI3 member PvAlf is required for β-phaseolin expression (Bobb et al., 1995). Gene activation is a two-step process, requiring PvAlf and abscisic acid (Li et al., 1999). Each of these two steps is associated with specific chromatin modifications (Ng et al., 2006) resulting in nucleosomal displacement over the three phased TATA boxes (Li et al., 1998). PvAlf binds to the promoter via the RY repeat motifs (Carranco et al., 2004). ABI5, a bZIP transcription factor, acts downstream from abscisic acid in β-phaseolin expression (Ng and Hall, 2008). Deletion analysis results implicated the G-box

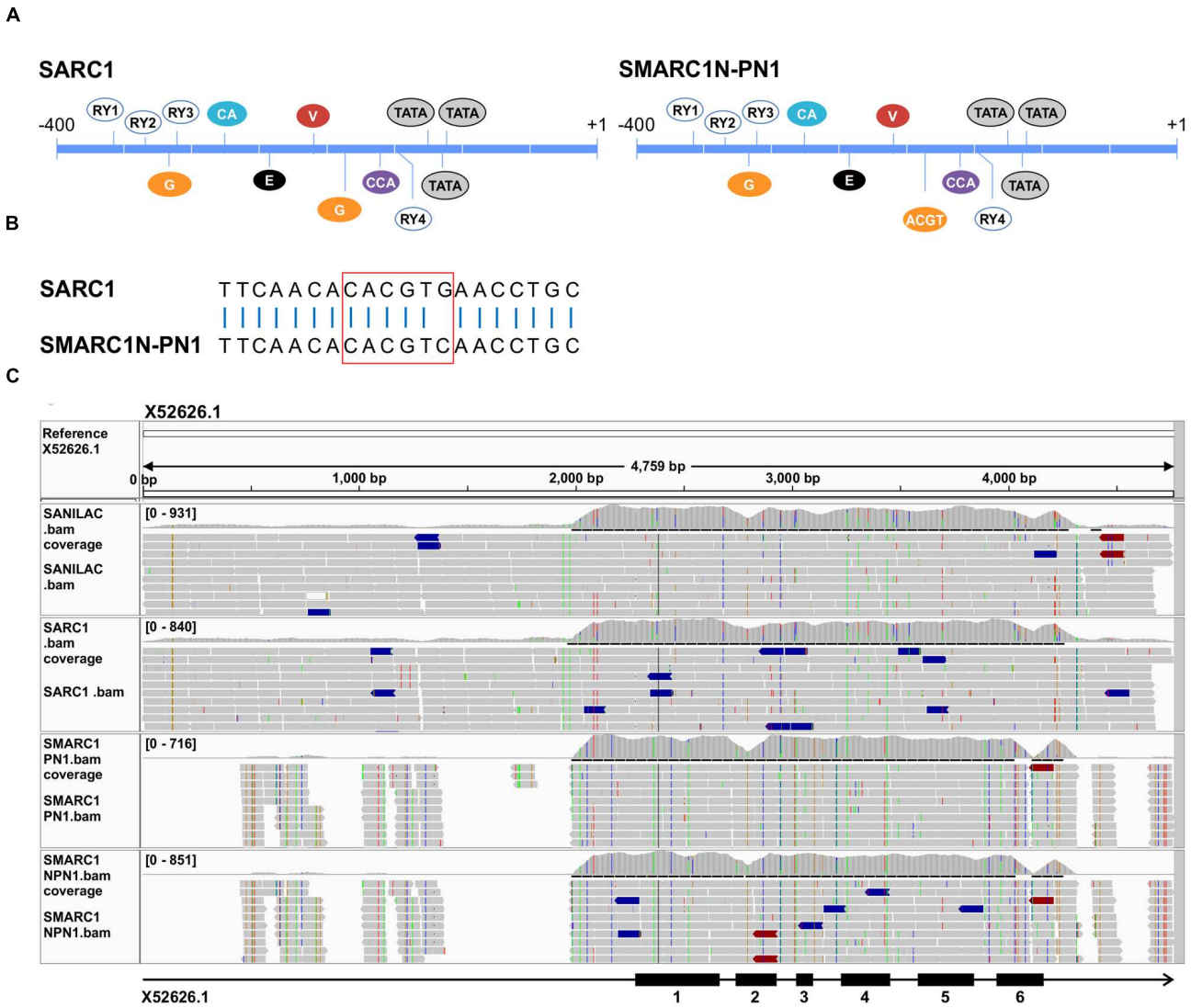

FIGURE 4 | (A) Schematic view of *cis*-regulatory elements present in the proximal promoter of β-phaseolin in SARC1 and Sanilac versus SMARC1N-PN1 and SMARC1-PN1. Regulatory motifs are designated as in Chandrasekharan et al. (2003). A single nucleotide polymorphism converts a G-box in SARC1 and Sanilac to an ACGT motif in SMARC1N-PN1 and SMARC1-PN1 **(B)**. G, G-box; E, E-box; C, CACA box; V, vicillin box; CCA, CCAAT box. **(C)** Sequence read alignments to the α-phaseolin gene from Sanilac (accession number X52626). The bar below shows the intron-exon structure with exons numbered as black boxes. The alignment shows incomplete coverage of the promoter region in SMARC1-PN1 and SMARC1N-PN1.

as the major abscisic acid responsive element in the β-phaseolin promoter.

In SMARC1N-PN1, β-phaseolin accumulates at lower levels than in SARC1, by approximately 20-fold (**Table 1**). Analysis of paired end read alignments to the reference genome G19833 revealed complete coverage of the β-phaseolin gene in the four genotypes. Focusing on the proximal promoter, all of the *cis*-regulatory elements described above except one were conserved (**Figures 4A,B**). Sanilac and SARC1 have a second G-box downstream from the first element. A single nucleotide polymorphism converts the ACGT motif present in SMARC1-PN1 and SMARC1N-PN1 into this second G-box motif (**Figures 4A,B**). This polymorphism was confirmed by genomic PCR and sequencing of the PCR products. In addition, the same fragment was amplified from phaseolin-containing and phaseolin-deficient *P. coccineus* genotypes. Both had the ACGT motif present. These results suggest that the single nucleotide polymorphism was introduced from *P. coccineus* into SMARC1-PN1 and SMARC1N-PN1, abrogating the second G-box motif. The present study associates this single nucleotide polymorphism with the genotypic difference in β-phaseolin accumulation.

For α-phaseolin, read alignments to reference sequences from Sanilac (Anthony et al., 1990; Diniz et al., 2014) showed a complete coverage of the coding section of the gene in all four genotypes (**Figure 4C**). Surprisingly, some polymorphisms

were observed between the alignment of reads from Sanilac and the reference sequences. Polymorphisms in phaseolin exons and introns clustered in pairs between Sanilac/SARC1 and SMARC1-PN1/SMARC1N-PN1, as expected. The presence of polymorphisms in SMARC1-PN1/SMARC1N-PN1 did not introduce premature stop codons, or affect intron splicing as predicted by GeneSeqer (Usuka et al., 2000). Promoter sequences were poorly covered in the alignment with SMARC1-PN1/SMARC1N-PN1, with large gaps upstream of the proximal promoter. Although scaffold assemblies contained sequences having similarity to phaseolin in Sanilac and SARC1 (Supplementary Table S4), these sequences were too fragmentary to reach a definitive conclusion on the nature of the polymorphism(s) responsible for the absence of α-phaseolin accumulation in SMARC1N-PN1.

Differences in Legumin Accumulation Are Not Associated with Genetic Polymorphisms

In SMARC1N-PN1, the most abundant protein in mature seed is the 11S globulin legumin (Marsolais et al., 2010). Blastx search of scaffold assemblies with the conceptual translation of legumin (Yin et al., 2011) identified one major scaffold per genotype. Sequences were extracted from the scaffold 247 for SARC1 and scaffold 972 for SMARC1N-PN1 and aligned. The alignment revealed the absence of polymorphism between the proximal 670 bp promoter sequences from the two genotypes (data not shown).

DISCUSSION

The goal of this study was to identify genetic polymorphisms associated with storage protein deficiency and proteome rebalancing in common bean, using the genetically related lines SARC1, SMARC1-PN1 and SMARC1N-PN1 and their parental background Sanilac. The three lines are genetic stocks exhibiting a similar percentage of the Sanilac background (83.6–87.5) (Osborn et al., 2003). They are expected to contain significant genetic variability coming from other parents, which include G12882, *Phaseolus coccineus* and Great Northern 1140. The re-analysis of proteomic data confirmed the identity of phaseolin and lectin isoforms which are affected by, or compensate for, seed storage protein deficiency. Arcelin genotypes are classified into types which are generally considered to contain a specific arcelin allele (Osborn et al., 1986; Lioi et al., 2003). Although arcelin-1 was the major arcelin quantified in SARC1, the proteomic and PCR genotyping data confirmed the presence of two other arcelin genes beside *Arc1*, *arc3-II* and *arc4-I*. Hartweck et al. (1991) had previously noted the presence of different arcelin variants in SARC1, differing in subunit composition (dimer vs. tetramer) and N-terminal sequence. According to the results of read alignments and genomic PCR, the deficiency in erythroagglutinating phytohemagglutinin and lectin appears due to the absence of the corresponding genes, *pha-E* and *lec4-B17*, respectively, in SMARC1N-PN1. SMARC1N-PN1 integrates a distinct allele substituting for *pha-L*, the pseudogene *pdlec1*, and *PDLEC2*. These had been identified from another genetic source of lectin deficiency, Pinto UI111 (Voelker et al., 1986). The present results suggest that Great Northern 1140 and Pinto UI111 share the same APA locus (Osborn and Bliss, 1985). These two genotypes are representative of market classes belonging both to the Durango land race, derived from the Middle American center of domestication (Singh et al., 1991; Mensack et al., 2010). While the genetic relationship between these two genotypes is unknown, the results suggest that they share a common origin (McClean and Myers, 1990). The reference genotype BAT-93, also a Middle American genotype, shares the *pdlec1* allele and *PDLEC2*, although it contains functional copies of *pha-E* and *lec4-B17*.

The deficiency in α-phaseolin is likely to be due to the partial or complete absence of the gene in SMARC1-PN1 and SMARC1N-PN1. Notably, the promoter sequence was poorly covered in read alignments. The high degree of sequence identity between phaseolin genes precluded the design of primers specific to α-phaseolin. The quality of the *de novo* genome assemblies was insufficient to reach a definitive conclusion regarding this gene. The high degree of sequence identity between phaseolin or lectin coding sequences hampers the assembly process. In future, the addition of long reads may facilitate gap closing and scaffold joining in the assemblies. This may help to clarify the status of the α-phaseolin gene in SMARC1-PN1 and SMARC1N-PN1. The large decrease in β-phaseolin accumulation in SMARC1N-PN1 as compared with SARC1, of approximately 20-fold, was associated with a single nucleotide polymorphism converting a G-box motif into an ACGT motif in the proximal promoter. The originally characterized sequence from Tendergreen (Slightom et al., 1983), an Andean genotype, as well as the reference Andean genome G19833, have the ACGT motif like SMARC1-PN1 and SMARC1N-PN1. This ACGT motif was shown to have little influence on the levels of reporter gene expression in transgenic *Arabidopsis* seeds (Chandrasekharan et al., 2003). However, an upstream G-box motif was required for high level expression. This single nucleotide polymorphism was likely introgressed from the *P. coccineus* accession, as it was present in *P. coccineus* genotypes examined in this study.

The results of this study also shed light on the mechanisms leading to compensation by residual lectins. Mannose lectin FRIL and α-amylase inhibitor 1 genes are absent from SARC1 and present in SMARC1N-PN1. The levels of α-amylase inhibitor 1 are slightly higher in SMARC1N-PN1 than in Sanilac, by 2.5-fold, the gene being present in a different genomic context. α-Amylase inhibitor-like protein is of particular interest. Unlike β-phaseolin, the large difference in protein accumulation, of approximately 50-fold, is associated with multiple polymorphisms in the proximal promoter, introducing eight potential positive *cis*-regulatory elements related to seed expression specific to SMARC1N-PN1, including a CAAT box in the right location, important for high level expression of tissue-specific genes. For legumin, the major storage protein in SMARC1N-PN1, accounting for close to 20% of total protein, the proximal promoters of 670 bp in size were identical between SARC1 and

SMARC1N-PN1. While legumin levels are raised by close to 7-fold in SMARC1N-PN1 relative to SARC1, its transcripts levels were elevated by approximately 2-fold during seed development (Liao et al., 2012). These results are consistent with those obtained with soybeans in which expression of major seed storage proteins was down-regulated by RNAi (Schmidt et al., 2011). The authors concluded that while seed protein concentration appears to be genetically determined, the identity of proteins compensating for storage protein deficiency in these lines is determined at the post-transcriptional level, in the absence of genetic polymorphisms (Herman, 2014). The identification of variants at the phaseolin and APA loci in the present study may also be useful for genetic diversity analyses and marker-assisted breeding in common bean.

AUTHOR CONTRIBUTIONS

SP, FY, BC, SH, WC, KP, and FM designed the research; SP, MD, FY, SM, FM analyzed data; GP contributed data; FS and RB contributed germplasm and reagents; all authors contributed to writing of the manuscript.

FUNDING

Funding is acknowledged from the Ontario Research Fund, Research Excellence Program, ORF-RE-043 project "*Phaseolus* genomics for improved bio-product development."

ACKNOWLEDGMENTS

FY was supported by the China Scholarship Council. We thank Xuejiang Shi, from the Clinical Genomics Centre, Mount Sinai Hospital, Toronto for Illumina sequencing. We are indebted to the McGill University and Génome Québec Innovation Centre for technical help with proteomic experiments. We thank Vi Nguyen, London Research and Development Centre, for assistance with Sanger sequencing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2016.00389

REFERENCES

Anthony, J. L., Vonder Haar, R. A., and Hall, T. C. (1990). Nucleotide sequence of an α-phaseolin gene from *Phaseolus vulgaris*. *Nucleic Acids Res.* 18:3396. doi: 10.1093/nar/18.11.3396

Bobb, A. J., Chern, M. S., and Bustos, M. M. (1997). Conserved RY-repeats mediate transactivation of seed-specific promoters by the developmental regulator PvALF. *Nucleic Acids Res.* 25, 641–647. doi: 10.1093/nar/25.3.641

Bobb, A. J., Eiben, H. G., and Bustos, M. M. (1995). PvAlf, an embryo-specific acidic transcriptional activator enhances gene expression from phaseolin and phytohemagglutinin promoters. *Plant J.* 8, 331–343. doi: 10.1046/j.1365-313X.1995.08030331.x

Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol.* 17, 1519–1533. doi: 10.1089/cmb.2009.0238

Bustos, M. M., Begum, D., Kalkan, F. A., Battraw, M. J., and Hall, T. C. (1991). Positive and negative cis-acting DNA domains are required for spatial and temporal regulation of gene expression by a seed storage protein promoter. *EMBO J.* 10, 1469–1479.

Carranco, R., Chandrasekharan, M. B., Townsend, J. C., and Hall, T. C. (2004). Interaction of PvALF and VP1 B3 domains with the β-phaseolin promoter. *Plant Mol. Biol.* 55, 221–237. doi: 10.1007/s11103-004-0512-8

Chandrasekharan, M. B., Bishop, K. J., and Hall, T. C. (2003). Module-specific regulation of the β-phaseolin promoter during embryogenesis. *Plant J.* 33, 853–866. doi: 10.1046/j.1365-313X.2003.01678.x

Chern, M. S., Bobb, A. J., and Bustos, M. M. (1996a). The regulator of MAT2 (ROM2) protein binds to early maturation promoters and represses PvALF-activated transcription. *Plant Cell* 8, 305–321. doi: 10.1105/tpc.8.2.305

Chern, M. S., Eiben, H. G., and Bustos, M. M. (1996b). The developmentally regulated bZIP factor ROM1 modulates transcription from lectin and storage protein genes in bean embryos. *Plant J.* 10, 135–148. doi: 10.1046/j.1365-313X.1996.10010135.x

Colucci, G., Moore, J. G., Feldman, M., and Chrispeels, M. J. (1999). cDNA cloning of FRIL, a lectin from *Dolichos lablab*, that preserves hematopoietic progenitors in suspension culture. *Proc. Natl. Acad. Sci. U.S.A.* 96, 646–650. doi: 10.1073/pnas.96.2.646

Diniz, A. L., Zucchi, M. I., Santini, L., Benchimol-Reis, L. L., Fungaro, M. H., and Vieira, M. L. (2014). Nucleotide diversity based on phaseolin and iron reductase genes in common bean accessions of different geographical origins. *Genome* 57, 69–77. doi: 10.1139/gen-2013-0183

Durante, M., Bernardi, R., Lupi, M. C., and Pini, S. (1989). *Phaseolus coccineus* storage proteins. II. Electrophoretic analysis and erythroagglutinating activity in various cultivars. *Plant Breed.* 102, 58–65. doi: 10.1111/j.1439-0523.1989.tb00315.x

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1513–1518. doi: 10.1073/pnas.1017351108

Hartweck, L. M., Vogelzang, R. D., and Osborn, T. C. (1991). Characterization and comparison of arcelin seed protein variants from common bean. *Plant Physiol.* 97, 204–211. doi: 10.1104/pp.97.1.204

Hayashi, M., Kitamura, K., and Harada, K. (2009). Genetic mapping of Cgdef gene controlling accumulation of 7S globulin (β-conglycinin) subunits in soybean seeds. *J. Hered.* 100, 802–806. doi: 10.1093/jhered/esp046

Hegedus, D. D., Baron, M., Labbe, N., Coutu, C., Lydiate, D., Lui, H., et al. (2014). A strategy for targeting recombinant proteins to protein storage vacuoles by fusion to *Brassica napus* napin in napin-depleted seeds. *Protein Expr. Purif.* 95, 162–168. doi: 10.1016/j.pep.2013.12.009

Herman, E. M. (2014). Soybean seed proteome rebalancing. *Front. Plant Sci.* 5:437. doi: 10.3389/fpls.2014.00437

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27, 297–300. doi: 10.1093/nar/27.1.297

Jegadeesan, S., Yu, K., Woodrow, L., Wang, Y., Shi, C., and Poysa, V. (2012). Molecular analysis of glycinin genes in soybean mutants for development of gene-specific markers. *Theor. Appl. Genet.* 124, 365–372. doi: 10.1007/s00122-011-1711-8

Kami, J., Poncet, V., Geffroy, V., and Gepts, P. (2006). Development of four phylogenetically-arrayed BAC libraries and sequence of the APA locus in *Phaseolus vulgaris*. *Theor. Appl. Genet.* 112, 987–998. doi: 10.1007/s00122-005-0201-2

Kawagoe, Y., Campbell, B. R., and Murai, N. (1994). Synergism between CACGTG (G-box) and CACCTG cis-elements is required for activation of the bean seed storage protein β-phaseolin gene. *Plant J.* 5, 885–890. doi: 10.1046/j.1365-313X.1994.5060885.x

Kim, W.-S., Gillman, J. D., and Krishnan, H. B. (2013). Identification of a plant introduction soybean line with genetic lesions affecting two distinct glycinin subunits and evaluation of impacts on protein content and composition. *Mol. Breed.* 32, 291–298. doi: 10.1007/s11032-013-9870-8

Kim, W.-S., Jez, J. M., and Krishnan, H. B. (2014). Effects of proteome rebalancing and sulfur nutrition on the accumulation of methionine rich δ-zein in transgenic soybeans. *Front. Plant Sci.* 5:633. doi: 10.3389/fpls.2014.00633

Kita, Y., Nakamoto, Y., Takahashi, M., Kitamura, K., Wakasa, K., and Ishimoto, M. (2009). Manipulation of amino acid composition in soybean seeds by the combination of deregulated tryptophan biosynthesis and storage protein deficiency. *Plant Cell Rep.* 29, 87–95. doi: 10.1007/s00299-009-0800-5

Lessard, P., Allen, R., Bernier, F., Crispino, J., Fujiwara, T., and Beachy, R. (1991). Multiple nuclear factors interact with upstream sequences of differentially regulated β-conglycinin genes. *Plant Mol. Biol.* 16, 397–413. doi: 10.1007/BF00023991

Li, G., Bishop, K. J., Chandrasekharan, M. B., and Hall, T. C. (1999). β-Phaseolin gene activation is a two-step process: PvALF-facilitated chromatin modification followed by abscisic acid-mediated gene activation. *Proc. Natl. Acad. Sci. U.S.A.* 96, 7104–7109. doi: 10.1073/pnas.96.12.7104

Li, G., Chandler, S. P., Wolffe, A. P., and Hall, T. C. (1998). Architectural specificity in chromatin structure at the TATA box in vivo: nucleosome displacement upon β-phaseolin gene activation. *Proc. Natl. Acad. Sci. U.S.A.* 95, 4772–4777. doi: 10.1073/pnas.95.8.4772

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liao, D., Pajak, A., Karcz, S. R., Chapman, B. P., Sharpe, A. G., Austin, R. S., et al. (2012). Transcripts of sulphur metabolic genes are co-ordinately regulated in developing seeds of common bean lacking phaseolin and major lectins. *J. Exp. Bot.* 63, 6283–6295. doi: 10.1093/jxb/ers280

Lin, Y., Pajak, A., Marsolais, F., McCourt, P., and Riggs, C. D. (2013). Characterization of a cruciferin deficient mutant of *Arabidopsis* and its utility for overexpression of foreign proteins in plants. *PLoS ONE* 8:e64980. doi: 10.1371/journal.pone.0064980

Lioi, L., Galasso, I., Lanave, C., Daminati, M. G., Bollini, R., and Sparvoli, F. (2007). Evolutionary analysis of the APA genes in the *Phaseolus* genus: wild and cultivated bean species as sources of lectin-related resistance factors? *Theor. Appl. Genet.* 115, 959–970. doi: 10.1007/s00122-007-0622-1

Lioi, L., Sparvoli, F., Galasso, I., Lanave, C., and Bollini, R. (2003). Lectin-related resistance factors against bruchids evolved through a number of duplication events. *Theor. Appl. Genet.* 107, 814–822. doi: 10.1007/s00122-003-1343-8

Liu, S., Ohta, K., Dong, C., Thanh, V. C., Ishimoto, M., Qin, Z., et al. (2006). Genetic diversity of soybean (*Glycine max* (L.) Merrill) 7S globulin protein subunits. *Genet. Resour. Crop Evol.* 53, 1209–1219. doi: 10.1007/s10722-005-2434-y

Marsolais, F., Pajak, A., Yin, F., Taylor, M., Gabriel, M., Merino, D. M., et al. (2010). Proteomic analysis of common bean seed with storage protein deficiency reveals up-regulation of sulfur-rich proteins and starch and raffinose metabolic enzymes, and down-regulation of the secretory pathway. *J. Proteomics* 73, 1587–1600. doi: 10.1016/j.jprot.2010.03.013

McClean, P., and Myers, J. (1990). Pedigrees of dry bean cultivars, lines and PIs. *Annu. Rep. Bean Improv. Coop.* 33, xxv–xxx.

Mensack, M. M., Fitzgerald, V. K., Ryan, E. P., Lewis, M. R., Thompson, H. J., and Brick, M. A. (2010). Evaluation of diversity among common beans (*Phaseolus vulgaris* L.) from two centers of domestication using 'omics' technologies. *BMC Genomics* 11:686. doi: 10.1186/1471-2164-11-686

Moore, J. G., Fuchs, C. A., Hata, Y. S., Hicklin, D. J., Colucci, G., Chrispeels, M. J., et al. (2000). A new lectin in red kidney beans called PvFRIL stimulates proliferation of NIH 3T3 cells expressing the Flt3 receptor. *Biochim. Biophys. Acta* 1475, 216–224. doi: 10.1016/S0304-4165(00)00067-2

Moreno, J., and Chrispeels, M. J. (1989). A lectin gene encodes the α-amylase inhibitor of the common bean. *Proc. Natl. Acad. Sci. U.S.A.* 86, 7885–7889. doi: 10.1073/pnas.86.20.7885

Nakashima, K., Fujita, Y., Katsura, K., Maruyama, K., Narusaka, Y., Seki, M., et al. (2006). Transcriptional regulation of ABI3- and ABA-responsive genes including RD29B and RD29A in seeds, germinating embryos, and seedlings of *Arabidopsis*. *Plant Mol. Biol.* 60, 51–68. doi: 10.1007/s11103-005-2418-5

Ng, D. W., Chandrasekharan, M. B., and Hall, T. C. (2006). Ordered histone modifications are associated with transcriptional poising and activation of the phaseolin promoter. *Plant Cell* 18, 119–132. doi: 10.1105/tpc.105.037010

Ng, D. W. K., and Hall, T. C. (2008). PvALF and FUS3 activate expression from the phaseolin promoter by different mechanisms. *Plant Mol. Biol.* 66, 233–244. doi: 10.1007/s11103-007-9265-5

O'Rourke, J. A., Iniguez, L. P., Bucciarelli, B., Roessler, J., Schmutz, J., McClean, P. E., et al. (2013). A re-sequencing based assessment of genomic heterogeneity and fast neutron-induced deletions in a common bean cultivar. *Front. Plant Sci.* 4:210. doi: 10.3389/fpls.2013.00210

Osborn, T. C., Blake, T., Gepts, P., and Bliss, F. A. (1986). Bean arcelin 2. Genetic variation inheritance and linkage relationships of a novel seed protein of *Phaseolus vulgaris* L. *Theor. Appl. Genet.* 71, 847–855. doi: 10.1007/BF00276428

Osborn, T. C., and Bliss, F. A. (1985). Effects of genetically removing lectin seed protein on horticultural and seed characteristics of common bean. *J. Am. Soc. Hortic. Sci.* 110, 484–488.

Osborn, T. C., Hartweck, L. M., Harmsen, R. H., Vogelzang, R. D., Kmiecik, K. A., and Bliss, F. A. (2003). Registration of *Phaseolus vulgaris* genetic stocks with altered seed protein compositions. *Crop Sci.* 43, 1570–1571. doi: 10.2135/cropsci2003.1570

Pandurangan, S., Pajak, A., Molnar, S. J., Cober, E. R., Dhaubhadel, S., Hernández-Sebastià, C., et al. (2012). Relationship between asparagine metabolism and protein concentration in soybean seed. *J. Exp. Bot.* 63, 3173–3184. doi: 10.1093/jxb/ers039

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. *Nucleic Acids Res.* 37, 289–297. doi: 10.1093/nar/gkn916

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Schmidt, M. A., Barbazuk, W. B., Sandford, M., May, G., Song, Z., Zhou, W., et al. (2011). Silencing of soybean seed storage proteins results in a rebalanced protein composition preserving seed protein content without major collateral changes in the metabolome and transcriptome. *Plant Physiol.* 156, 330–345. doi: 10.1104/pp.111.173807

Schmidt, M. A., and Herman, E. M. (2008). Proteome rebalancing in soybean seeds can be exploited to enhance foreign protein accumulation. *Plant Biotechnol. J.* 6, 832–842. doi: 10.1111/j.1467-7652.2008.00364.x

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* 46, 707–713. doi: 10.1038/ng.3008

Shirsat, A., Wilford, N., Croy, R., and Boulter, D. (1989). Sequences responsible for the tissue specific promoter activity of a pea legumin gene in tobacco. *Mol. Gen. Genet.* 215, 326–331. doi: 10.1007/BF00339737

Singh, S. P., Gepts, P., and Debouck, D. G. (1991). Races of common bean (*Phaseolus vulgaris*, Fabaceae). *Econ. Bot.* 45, 379–396. doi: 10.1007/bf02887079

Slightom, J. L., Sun, S. M., and Hall, T. C. (1983). Complete nucleotide sequence of a French bean storage protein gene: phaseolin. *Proc. Natl. Acad. Sci. U.S.A.* 80, 1897–1901. doi: 10.1073/pnas.80.7.1897

Song, B., Shen, L., Wei, X., Guo, B., Tuo, Y., Tian, F., et al. (2014). Marker-assisted backcrossing of a null allele of the α-subunit of soybean (*Glycine max*) β-conglycinin with a Chinese soybean cultivar (a). The development of improved lines. *Plant Breed.* 133, 638–648. doi: 10.1111/pbr.12203

Stålberg, K., Ellerström, M., Ezcurra, I., Ablov, S., and Rask, L. (1996). Disruption of an overlapping E-box/ABRE motif abolished high transcription of the napA storage-protein promoter in transgenic *Brassica napus* seeds. *Planta* 199, 515–519.

Taylor, M., Chapman, R., Beyaert, R., Hernández-Sebastià, C., and Marsolais, F. (2008). Seed storage protein deficiency improves sulfur amino acid content in common bean (*Phaseolus vulgaris* L.): redirection of sulfur from γ-glutamyl-S-methyl-cysteine. *J. Agric. Food Chem.* 56, 5647–5654. doi: 10.1021/jf800787y

Tsubokura, Y., Hajika, M., Kanamori, H., Xia, Z., Watanabe, S., Kaga, A., et al. (2012). The β-conglycinin deficiency in wild soybean is associated with the tail-to-tail inverted repeat of the α-subunit genes. *Plant Mol. Biol.* 78, 301–309. doi: 10.1007/s11103-011-9865-y

Usuka, J., Zhu, W., and Brendel, V. (2000). Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* 16, 203–211. doi: 10.1093/bioinformatics/16.3.203

van der Geest, A. H., and Hall, T. C. (1996). A 68 bp element of the β-phaseolin promoter functions as a seed-specific enhancer. *Plant Mol. Biol.* 32, 579–588. doi: 10.1007/BF00020199

Vlasova, A., Capella-Gutierrez, S., Rendon-Anaya, M., Hernandez-Onate, M., Minoche, A. E., Erb, I., et al. (2016). Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. *Genome Biol.* 17, 32. doi: 10.1186/s13059-016-0883-6

Voelker, T. A., Staswick, P., and Chrispeels, M. J. (1986). Molecular analysis of two phytohemagglutinin genes and their expression in *Phaseolus vulgaris* cv. Pinto, a lectin-deficient cultivar of the bean. *EMBO J.* 5, 3075–3082.

Wang, J., Liu, L., Guo, Y., Wang, Y. H., Zhang, L., Jin, L. G., et al. (2014). A dominant locus, qBSC-1, controls β subunit content of seed storage protein in soybean (*Glycine max* (L.) Merri.). *J. Integr. Agric.* 13, 1854–1864. doi: 10.1016/S2095-3119(13)60579-1

Wu, C., Washida, H., Onodera, Y., Harada, K., and Takaiwa, F. (2000). Quantitative nature of the Prolamin-box, ACGT and AACA motifs in a rice glutelin gene promoter: minimal cis-element requirements for endosperm-specific gene expression. *Plant J.* 23, 415–421. doi: 10.1046/j.1365-313x.2000.00797.x

Wu, Y., and Messing, J. (2014). Proteome balancing of the maize seed for higher nutritional value. *Front. Plant Sci.* 5:240. doi: 10.3389/fpls.2014.00240

Wu, Y., Yuan, L., Guo, X., Holding, D. R., and Messing, J. (2013). Mutation in the seed storage protein kafirin creates a high-value food trait in sorghum. *Nat. Commun.* 4:2217. doi: 10.1038/ncomms3217

Yamaguchi, H. (1991). Isolation and characterization of the subunits of *Phaseolus vulgaris* α-amylase inhibitor. *J. Biochem.* 110, 785–789.

Yin, F., Pajak, A., Chapman, R., Sharpe, A., Huang, S., and Marsolais, F. (2011). Analysis of common bean expressed sequence tags identifies sulfur metabolic pathways active in seed and sulfur-rich proteins highly expressed in the absence of phaseolin and major lectins. *BMC Genomics* 12:268. doi: 10.1186/1471-2164-12-268

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.