frontiers in Plant Science

REVIEW
published: 24 November 2015
doi: 10.3389/fpls.2015.01021

Expanding Omics Resources for Improvement of Soybean Seed Composition Traits

*Juhi Chaudhary, Gunvant B. Patil, Humira Sonah †, Rupesh K. Deshmukh †, Tri D. Vuong, Babu Valliyodan and Henry T. Nguyen **

Division of Plant Sciences, National Center for Soybean Biotechnology, University of Missouri, Columbia, MO, USA

Food resources of the modern world are strained due to the increasing population. There is an urgent need for innovative methods and approaches to augment food production. Legume seeds are major resources of human food and animal feed with their unique nutrient compositions including oil, protein, carbohydrates, and other beneficial nutrients. Recent advances in next-generation sequencing (NGS) together with "omics" technologies have considerably strengthened soybean research. The availability of well annotated soybean genome sequence along with hundreds of identified quantitative trait loci (QTL) associated with different seed traits can be used for gene discovery and molecular marker development for breeding applications. Despite the remarkable progress in these technologies, the analysis and mining of existing seed genomics data are still challenging due to the complexity of genetic inheritance, metabolic partitioning, and developmental regulations. Integration of "omics tools" is an effective strategy to discover key regulators of various seed traits. In this review, recent advances in "omics" approaches and their use in soybean seed trait investigations are presented along with the available databases and technological platforms and their applicability in the improvement of soybean. This article also highlights the use of modern breeding approaches, such as genome-wide association studies (GWAS), genomic selection (GS), and marker-assisted recurrent selection (MARS) for developing superior cultivars. A catalog of available important resources for major seed composition traits, such as seed oil, protein, carbohydrates, and yield traits are provided to improve the knowledge base and future utilization of this information in the soybean crop improvement programs.

Keywords: legumes, soybean, seed traits, omics, genomics, next-generation sequencing (NGS), quantitative trait loci (QTL), genome-wide association study (GWAS)

OPEN ACCESS

Edited by:
Rajeev K. Varshney, International Crops Research Institute for the Semi-Arid Tropics, India

Reviewed by:
Ramanjulu Sunkar, Oklahoma State University, USA
Swarup Kumar Parida, National Institute of Plant Genome Research, India

***Correspondence:**
Henry T. Nguyen
nguyenhenry@missouri.edu

†Present Address:
Humira Sonah and Rupesh K. Deshmukh, Division of Plant Science, University Laval, QC, Canada

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

Received: *31 July 2015*
Accepted: *05 November 2015*
Published: *24 November 2015*

Citation:
Chaudhary J, Patil GB, Sonah H, Deshmukh RK, Vuong TD, Valliyodan B and Nguyen HT (2015) Expanding Omics Resources for Improvement of Soybean Seed Composition Traits. Front. Plant Sci. 6:1021. doi: 10.3389/fpls.2015.01021

INTRODUCTION

In view of the increasing world population, production of sustainable food supplies will be a critical challenge in the twenty-first century. The world population is projected to cross 9 billion by 2050, indicating that food supplies must be doubled to meet the requirement of the expanding population (Varshney et al., 2013a; Zhou et al., 2015). Apart from the quantity of food, quality is also a critical issue to maintain nutritive values with increased potential for yield. Seeds are an important part of the plant due to their role in reproduction and storing food reserves in the embryonic cotyledons. Legume seeds are an essential source of food, feed, minerals, and also

provides biological nitrogen fixation by forming a symbiotic relationship with rhizobia (Gepts et al., 2005). Soybeans are unique in legumes with a seed content of about 40% protein and 21% oil on a dry matter basis. It is the most widely grown oil seed crop in the world and represented 56% of the world's vegetable oil seed production in 2013. The United States is the leading soybean producer with 34% [108 Million Metric Tons (MMT)], followed by Brazil with 30% (94.5 MMT), and Argentina with 18% (56 MMT) of the world production (SoyStats 2014, www.soystats.com). Besides the total seed oil and protein content, oil components (fatty acids) and protein components (amino acids) are also desirable for long term shelf life and nutrition. The animal feed industry uses about 70% of soybean meal due to it being high in protein with a good amino acid balance. Soybean meal provides more energy than any other plant protein source (Cromwell, 2012). Furthermore, soybean is also used as sources of industrial and pharmaceutical applications as well as in the production of biodiesel (Goldberg and Stacey, 2008). Several international and domestic soybean processors prefer soybean with different combinations of seed composition. Soybean fatty acids include palmitic acid (10%), stearic acid (4%), oleic acid (18%), linoleic acid (55%), and linolenic acid (13%). Higher oleic acid and lower linolenic acid without generating *trans*-fats are desirable for oil stability and addressing health concerns of soybean oil (Clemente and Cahoon, 2009; Lee et al., 2012). In addition, manipulation of the amino acid profile [methionine (Met), lysine (Lys), and threonine (Thr)], is desired to improve seed protein quality since the animal feed industry uses about 77% of soybean meal as a source of protein and amino acids (Warrington et al., 2015).

Another important component of soybean seeds are carbohydrates. Their concentration is very crucial for determining animal feed quality (Wilson, 2004). Soybean seeds contain about 12% soluble carbohydrates, which includes sucrose, raffinose, and stachyose. Among the three carbohydrates, only higher sucrose content is desirable as metabolizable energy in the animal feed. Raffinose and stachyose accumulation causes indigestibility and flatulence which ultimately results in reduced economic and dietary value of soybean seed (Hagely et al., 2013). Therefore, it is a prerequisite to improve seed utility by adjusting the concentration of the above-mentioned components in a desired proportion.

The soybean seed composition traits are considered complex due to multiple gene control, environmental factors, and the interaction between molecular, biochemical, and genetic mechanisms within the seed. The negative correlation of protein with oil and yield is poorly understood and breaking this relationship is another challenge in soybean improvement (Clemente and Cahoon, 2009). Additionally, the variation in seed composition is largely affected by environment; for example, cooler temperature negatively affects seed storage protein content (Bellaloui et al., 2015). Therefore, the identification of environmentally stable germplasm with desirable seed composition and elucidating their genetic control and complex metabolic interactions during seed development are needed for soybean improvement. Such a multi-tier investigation requires extensive experimental efforts, which involves genomic, transcriptomic, metabolomic, ionomic, and phenomic tools (**Figure 1**).

The development of soybean seed with high quality and improved yield can be accelerated using modern breeding techniques, such as marker-assisted selection (MAS) (Xu and Crouch, 2008), genomic selection (Desta and Ortiz, 2014), and genome editing approaches (Voytas and Gao, 2014). In recent years, transcriptome analysis (Severin et al., 2010), proteomics (Eldakak et al., 2013), metabolomics (Saito and Matsuda, 2010) phenomics (Zhu et al., 2012), and ionomics (Singh et al., 2013b) have progressed at a rapid pace. A comprehensive overview summarizing increasing research efforts in soybean genomics, transcriptomics, and proteomics during the last decade is highlighted in the Supplementary Figure 1.

The soybean genome was the first published legume reference genome (Schmutz et al., 2010) and was followed by other legume genome sequences (Varshney et al., 2012a, 2013b; Schmutz et al., 2014). Until recently, with the advancement in next-generation sequencing (NGS) technology, whole genome sequencing (WGS)

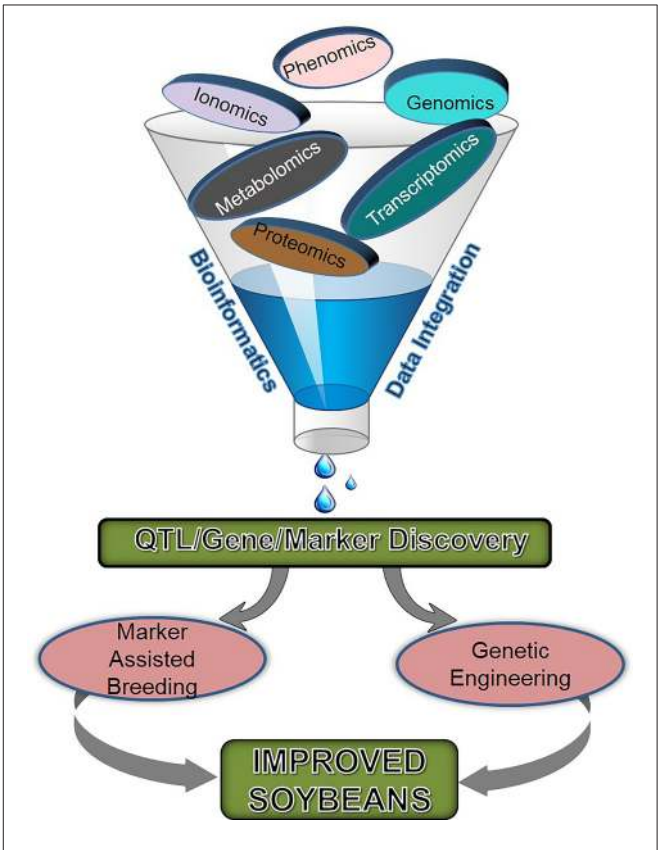

FIGURE 1 | Different omics approaches and their integrated tools being used for soybean breeding.

Abbreviations: SSR, Simple Sequence Repeats; SNP, Single Nucleotide Polymorphism; WGS, Whole Genome Sequencing; WGRS, Whole Genome Re-Sequencing; GWAS, Genome-Wide Association Studies; GS, Genomic Selection; MAS, Marker-Assisted Selection, QTL, Quantitative Trait Loci; GBS, Genotyping-by-Sequencing; RIL, Recombinant Inbred Lines; Chr., Chromosome.

has become possible for all major crops, including soybean and is also being utilized for several orphan legume species (Varshney et al., 2012b). Due to low sequencing cost, NGS has been widely used in various *de novo* sequencing, whole genome re-sequencing (WGRS), genotyping-by-sequencing (GBS), and transcriptome analysis. This has made a significant impact in molecular breeding programs through marker development and agronomic traits mapping (Metzker, 2010; Peterson et al., 2012; Poland and Rife, 2012; Varshney et al., 2012b; Sonah et al., 2013).

Although rapid progresses in the use of omics tools have been demonstrated, data mining and analyses are still challenging tasks. There is a wide range of genetic variation in oil and protein content among soybean accessions of the USDA Soybean Germplasm Collection, but it is extremely rare to find an accession with higher protein and oil content (Wilson, 2004). For decades, geneticists have used a quantitative trait loci (QTL) mapping approach to identify major genes responsible for seed composition traits, yielding several putative candidate genes, but currently, there are no precise genomic loci identified for these traits in soybean. Technological advances in sensitivity, resolution, high-throughput, and reduced costs of the "omics" based assays have provided a doorway for the applications of complex trait studies. The resulting data includes molecular markers, transcript sequences, genetic linkage maps, and physical maps; all of which would help in the elucidation of complex traits. Therefore, the integration of several "omics" platforms can be an excellent approach for the assessment of various seed composition traits. This review aims to highlight significant studies using omics approaches such as genomics, transcriptomics, metabolomics, proteomics, and phenomics applied to soybean seed composition improvement.

GENOMICS DEVELOPMENT

Molecular Mapping of Seed Composition Traits in Soybean

Molecular markers allow precise, cost effective, and high-throughput identification of genetic variants for different traits. Markers are important in breeding applications for developing genetic linkage maps, germplasm evaluation, phylogenetic and evolutionary analysis, selection of desired alleles and mapping of genes/QTL. Simple sequence repeat (SSR) markers have been extensively utilized to study seed composition traits in soybean (Wang et al., 2014b; Warrington et al., 2015); for example, seed oil, protein, and seed size QTL (Hyten et al., 2004), fine-mapping of soybean protein QTL on chromosome (Chr.) 20 (Nichols et al., 2006). A publicly available SSR marker database, containing about 33,000 markers was developed from WGS information (Song et al., 2010). Eskandari et al. (2013) utilized SSR markers and identified QTL for oil content on Chr. 9, which also had a significant positive effect on seed protein composition. For the improvement of soybean meal, Pathan et al. (2013) detected QTL using both SSR and single nucleotide polymorphism (SNP) markers for seed protein, oil, and seed weight across genetic backgrounds and environments on Chrs. 5 and 6. The SSR markers are less abundant in the genome and has limitations in high-throughput applicability as compared to SNP markers to be utilized in large breeding programs (Singh et al., 2013a).

The availability of a well-annotated soybean genome sequence has facilitated the development of SNP markers and is being utilized in crop improvement (**Table 1**). The genotyping approaches include GBS (Elshire et al., 2011; Sonah et al., 2013), restriction site associated DNA sequencing (Baird et al., 2008), SoySNP50K iSelect BeadChip (Song et al., 2013), SoySNP6K Infinium BeadChip (Akond et al., 2013), and the Axiom SoyaSNP array for approximately 180,000 SNPs (Lee et al., 2015). Furthermore, NGS has also facilitated the development of various SNP genotyping assays, such as KASPar, GoldenGate, and Infinium Chips, which can also be applied in genome-wide marker development (Varshney et al., 2015).

Researchers have shown that fatty acid composition and protein content of soybean seed are largely affected by environmental factors (more specifically temperature), which challenges the efforts in phenotyping (Bellaloui et al., 2015). To overcome this problem, MAS for fatty acid has been successfully employed in several breeding programs (Pham et al., 2010, 2012, 2014). Precise mapping of QTL by high-throughput genotyping platforms has revolutionized MAS for soybean seed trait improvement (Supplementary Tables 1, 2; Supplementary Figure 2). In a recent study, Warrington et al. (2015) identified a major QTL on Chr. 20 (55% phenotypic variation) for seed protein and amino acid content in a Benning × Danbaekkong population; however, a negative correlation between total protein and amino acid (especially for Thr) was observed. Information about genomic loci governing seed composition traits collected from QTL, genome-wide association study (GWAS) and WGRS studies from the past few years is summarized to locate genomic hot-spots in chromosome locations. Interestingly, a majority of identified QTL for seed composition traits were found to be on Chrs. 20, 15, 6, and 5 (Supplementary Table 1; **Figure 2**). The QTL on Chr. 20 explained 12–55% of phenotypic variation associated with increased protein content. Compilation of data showed that several genomic loci for oil and protein from different studies were found to be co-localized. This confounding region could provide an entry point to investigate the basis of correlation between various seed composition components and could be useful for gene pyramiding strategies.

Several efforts of meta-QTL analysis have been performed for seed composition traits in soybean (Zhaoming et al., 2009; Sun et al., 2012). QTL meta-analysis combines datasets from independent studies to detect consensus QTL and to shrink the QTL confidence intervals making them more useful for MAS (Rudner et al., 2002). In soybean, there are several meta-QTL studies that have been conducted for seed traits; for example, 17 meta-QTL have been identified for hundred-seed weight (HSW) using 65 QTL from 12 studies (Zhaoming et al., 2009). In another study, targeting HSW by multi-environmental mapping followed by meta-analysis, 15 consensus QTL were identified (Sun et al., 2012). In addition, seed oil content was also examined through meta-QTL analysis (Qi et al., 2011a,b). Similarly, Zhao-Ming et al. (2011) performed meta-QTL analysis for seed protein content and reported 23 consensus QTL by integrating 107 QTL. In summary, these studies identified hotspots for seed traits and

TABLE 1 | Whole genome re-sequencing efforts performed in soybean.

Sr. No.	No. of lines used	Genotypes	Sequencing depth	SNP calling method	No. of SNPs	References
1	1	*G. soja*	∼52.07X	De novo	∼2.5 Million	Kim et al., 2010
2	31	17 *G. soja* and 14 *G. max* (cultivated soybean)	×5 depth	SOAP	6,318,109	Lam et al., 2010
3	25	8 *G. soja*, 17 *G. max* (8 landraces, and 9 elite lines/cultivars)	–	SOAP	5,102,244	Li et al., 2013
4	16	10 *G. max* and 6 *G. soja*	>14x	GATK	3,871,469	Chung et al., 2014
5	7	*G. soja*	∼111.9X	SOAP	3.62–4.72 M SNP per line	Li et al., 2014
6	11	10 Semi-wild and 1 *G. soja*	9 Semi-wild at ∼3X while 1 Semi-wild at ∼41X, and 1 Wild at ∼55X	SOAP	7,704,637	Qiu et al., 2014
7	302	62 *G. soja*, 240 *G. max* (130 landraces, and 110 improved cultivars)	>11X	GATK	9,790,744	Zhou et al., 2015

this could be helpful to improve seed composition traits. Meta-QTL analysis gives the basis for gene mining and also facilitates refining soybean genetic maps. However, meta-QTL may indicate presence of pleotropic traits by creating QTL groups or clusters for several traits which necessitates confirmation of hotspots by integration with other approaches.

Association Mapping

QTL mapping is routinely conducted and requires mapping populations derived from bi-parental crosses, such as F_2, backcross (BC), or recombinant inbred lines (RILs). Despite the success of QTL mapping, it is limited by a narrow range of allelic diversity and roughly estimated QTL intervals, which refers to large genomic regions (Borevitz and Nordborg, 2003). Therefore, to address this issue, a statistically more advanced method, GWAS has been developed. GWAS utilizes allelic diversity in large sets of germplasm and the recombination events accumulated over hundreds of years during the evolution and domestication. A few merits of the GWAS approach includes high resolution, often to the gene level, and the use of previously well studied populations that provide a correlation between genetic and phenotypic variations. In addition, GWAS provides statistical power to reconnect the phenotype to its underlying genetics (Brachi et al., 2011). A number of GWAS have been performed for various traits in soybean with great success in identifying loci with high mapping precision (Supplementary Table 3). Vaughn et al. (2014) used GWAS to identify genetic loci with <1 Mbp resolution governing the soybean seed composition traits. In this study, the seed oil, protein, amino acids, sucrose, and oligosaccharide (RFO's) content were dissected to identify the precise position of genes. A study conducted to explore the genome wide association between seed protein and oil content in soybean revealed 40 SNPs associated with protein and 25 SNPs associated with oil content (Hwang et al., 2014). In addition, Zhou et al. (2015) used high density SNP data with WGRS and identified more precise regions for seed oil content, seed weight, seed coat color, and domestication traits. GWAS can be used in place of QTL mapping to eliminate the limitation of bi-parental mapping populations. However, GWAS has its own merits and demerits (Korte and Farlow, 2013). Combining GWAS with QTL mapping will provide advantages for precise mapping and understanding of genetic architecture of seed traits (**Figure 2**). Recently, Sonah et al. (2015) demonstrated the integrated approach of QTL mapping and GWAS for the identification of loci/candidate genes for various traits including seed weight, oil, and protein content in soybean. A total of 25 loci were identified, including three loci associated with seed weight and eight each for oil and protein content.

Despite GWAS's merits, it has some limitations in the detection of rare alleles that are associated with natural genetic variation, which may lead to obscure analysis and synthetic associations (Dickson et al., 2010; Platt et al., 2010). Even though a considerable amount of false positives in GWAS can be reduced by using a correction of population and kinship approach, such errors cannot be fully eliminated. This is particularly true when the subpopulations are extensively diverse. However, association studies are genuinely promising to identify significant genotype-phenotype correlations comprehensively for complex traits (Hwang et al., 2014; Vaughn et al., 2014). It is foreseeable that the initial success in candidate gene identification using an association approach would greatly help in the advancement of the deeper understanding of complex soybean seed traits.

Genomic Selection

Association studies and MAS have been used in plant improvement programs for several years. Nevertheless, they have some limitations, such as long selection cycles and the identification of significant marker-QTL associations which is unable to capture "minor" gene effects (Desta and Ortiz, 2014). These limitations can be effectively addressed by using a promising approach, known as genomic selection (GS) (Meuwissen et al., 2001). Unlike MAS, GS utilizes the entire set of markers to predict genomic estimated breeding values (GEBVs) for all genotyped individuals within a breeding population by

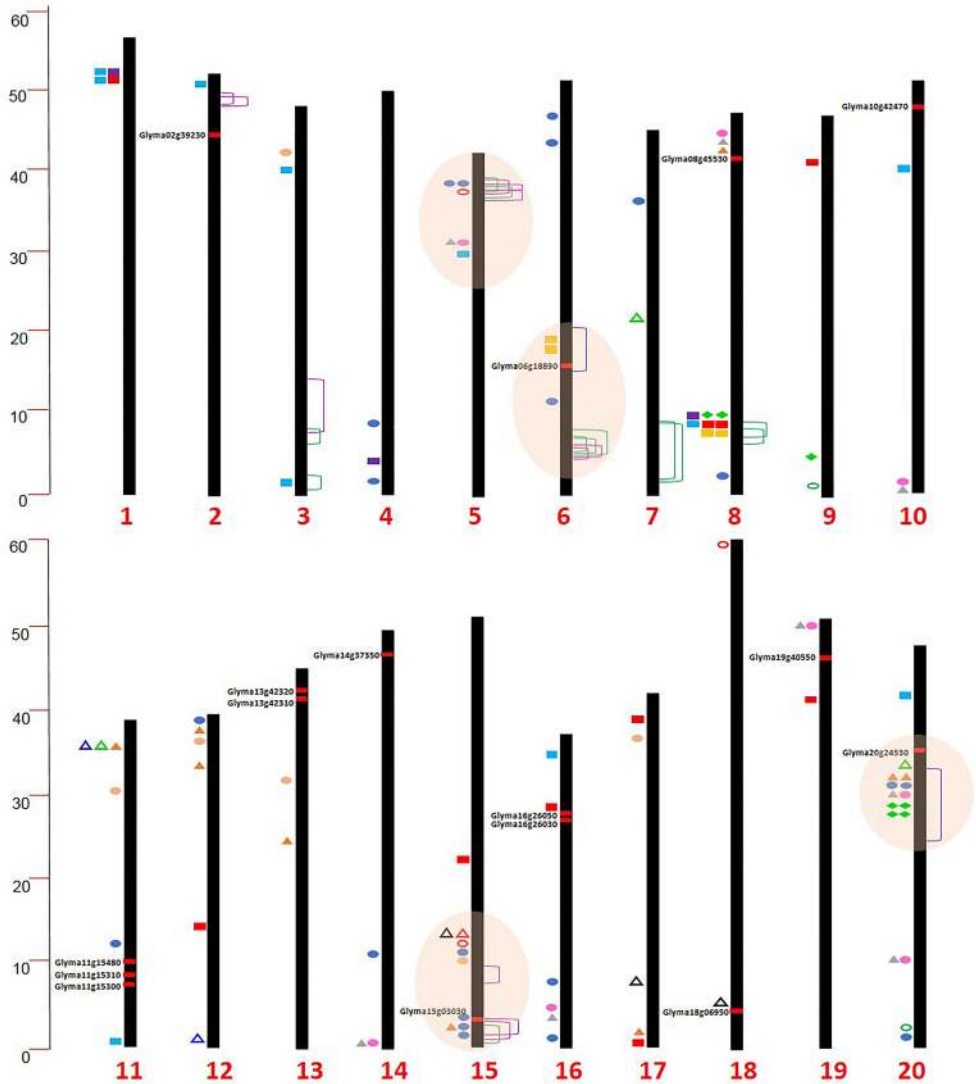

FIGURE 2 | Chromosomal locations of genomic hot-spots, promising genes, QTL, GWAS, and linked markers for soybean seed composition from several studies. Protein, Oil, Cysteine, Lysine, Methionine, Threonine, Sucrose, Stachyose (Vaughn et al., 2014); Protein and oil (Hwang et al., 2014); Protein, Oil (Sonah et al., 2015); Glucose, Sucrose, Fructose, Stachyose (Wang et al., 2014b); Protein, Oil (Pathan et al., 2013), Protein, Methionine (Warrington et al., 2015).

capturing the total additive genetic variance for a particular trait of interest (Heffner et al., 2009, 2011). GS requires a training population (both genotyped and phenotyped) to calculate breeding values by using all marker information and avoiding biased marker effects simultaneously (Heffner et al., 2009). The primary benefit of GS is that selection can be imposed at a very early stage in the breeding process, thus accelerating the breeding cycles without phenotyping. However, optimizing the constituents of a training population is very challenging and is influenced by a number of parameters, such as choice of model, size of training data, trait heritability, span of linkage disequilibrium (LD), marker density, and strength of genetic relationships between training and validation populations (Bentley et al., 2014).

Due to the advantages of GS over conventional methods, it has been successfully applied to a variety of crops, such as wheat and maize (Crossa et al., 2010), pear (Iwata et al., 2013), sugar beet (Würschum and Kraft, 2014), and is constantly being used for other crop species. There are only a few reports for the application of GS in soybean (Hu et al., 2011; Shu et al., 2012; Bao et al., 2014; Jarquín et al., 2014). Hu et al. (2011) studied primary embryogenesis capacity in soybean using 126 RILs and 80 SSRs and reported a strong correlation (r^2 = 0.78) between GEBVs and phenotypic data (Hu et al., 2011).

A study was performed on soybean HSW using 79 sequence-characterized amplified region (SCAR) markers for 288 varieties. This study reported a correlation coefficient of 0.904 amongst the GEBVs and phenotypic values (Shu et al., 2012). Recently, Bao et al. (2014) investigated soybean cyst nematode resistance via genotyping of 282 accessions employing the 1536 SNP array and demonstrated a significant prediction accuracy (0.59–0.67) for soybean cyst nematode resistance in soybean. Another study reported a prediction accuracy of 0.64 for grain yield, which indicates a good potential of GS utilization strategy in soybean (Jarquín et al., 2014). It is noticeable that implementation of a GS approach is needed in soybean seed related traits. Moreover, with the declining genotyping costs and increasing phenotyping costs, GS can be helpful to mitigate many of the selections associated with phenotyping and ultimately speed up the breeding cycles. Hence, the GS statistical method is more feasible and more emphasis should be given to employ this approach in the improvement of seed composition with available genotypic and phenotypic data.

SOYBEAN SEED TRANSCRIPTOMICS ADVANCEMENTS

Regulation of gene expression occupies a central role in the flow of genetic information. Location and level of gene expression gives the insight in the functional regulation, thus, by collecting and comparing transcriptome of different tissue-types, stages or development, researchers can gain a deeper understanding of how changes in transcription may affect the phenotype. There are several tools for generating and mining the transcriptomes including gene-by-gene and global methods for quantification of expression levels (Wirta, 2006). Global methods allow for a nearly comprehensive analysis of the transcriptome, which comprise of hybridization-based (microarrays and GeneChips); sequence tag-based [ESTs sequencing, cDNA deep sequencing, serial analysis of gene expression (SAGE), and massive parallel signature sequencing (MPSS)], and RNA-sequencing (RNA-seq) approaches. Chip-based technologies involving microarrays have become a dominant platform after ESTs sequencing and genome sequencing in several plant species.

The RNA-seq approach provides information of large-scale sequences of coding and non-coding RNAs without prior genomic information. RNA-seq data captures transcriptome dynamics across different tissues without data set normalization and has been applied in the identification of genes and regulatory networks associated with soybean seed composition (Severin et al., 2010; Goettel et al., 2014). It also has considerable advantages in examining novel transcript, splicing events, and allele-specific expression. Due to these advantages, RNA-Seq is considered a valuable technology in understanding transcriptomic dynamics during developmental and physiological changes (Severin et al., 2010).

To date, several studies have been performed to track gene expression changes from fertilization to maturity in soybean seed development events utilizing RNA-Seq (Severin et al., 2010). For instance, a genome-wide expression analysis was performed for soybean MADS gene family using publicly available RNA-seq data from 17 tissues (Fan et al., 2013). Similarly, the public RNA-seq database was utilized to study sugar transporter genes (SWEET) in soybean seed and other tissues (Patil et al., 2015). This study identified several SWEET genes, which are highly expressed, and plays an important role in nutrient unloading during seed development and seed filling stages. Significant efforts have been made in understanding gene function and regulation using transcriptome profiling (**Table 2**). Recently, O'Rourke et al. (2014) used RNA-seq to analyze oil biosynthesis, nitrogen assimilation, and transcription factors affecting oil, carbohydrate, and protein deposition during seed fill. In another study, Goettel et al. (2014) investigated the transcript polymorphism in soybean lines varying in oil composition and content. They demonstrated a high correlation of transcript and genetic variation associated with oil quality traits. In addition, Yin et al. (2014) compared soybean seed specific genes from microarray, DDD (differential digital display), and RNA-seq databases and identified 184 seed development specific genes. Most of the identified genes were found to be related to nutrient reservoir activity, lipid binding, enzyme inhibitor activity, peptidase regulator activity, hydrolase activity, embryo development, lipid transport, proteolysis, vacuoles, or lipid particles. Due to the large number of gene expression studies performed on soybean seed, a huge amount of data is available in the Gene Expression Omnibus (GEO) at National Center for Biotechnology Information (NCBI) database (Supplementary Figure 3). The large amount of microarray and deep sequencing transcriptomic data have allowed the development of a soybean co-expression network database containing 23,267 genes, 1873 miRNA-target pairs, and a group of acyl-lipid pathways containing 221 enzymes and more than 1550 genes (Yu et al., 2014).

The wild soybean (*G. soja*) is the unique resource to study the regulation of seed composition traits (specifically oil and protein) because *G. max* produces nearly twice as much oil and less protein than *G. soja*. The difference in seed oil content and composition within soybean germplasm is largely affected by genomic variation and expression profile of the genes involved in fatty acid biosynthesis and other unknown regulators. Until now, several wild soybean accessions have been re-sequenced (Kim et al., 2010; Joshi et al., 2013; Li et al., 2014); however, they have not been studied at the transcriptome level, specifically during the seed developmental stages. Sequence variants and expression polymorphism associated with gene function can help in dissecting the underlying causes of phenotypic variation.

PROTEOMICS ADVANCES

Proteomics is the large-scale study of structural and functional features of a set of proteins present in an organism. Proteomics has gained ample popularity over the genome-based technologies as it directly deals with biochemical processes. Moreover, post-translational modifications of protein can also be studied through proteomic techniques (Eldakak et al., 2013). Soybean seed proteome is dominated by two major storage proteins,

TABLE 2 | List of significant seed related gene expression studies in soybean.

Purpose of the study	Tissue stages/condition	Approach	References
Gene expression during embryo development and isoflavonoid biosynthesis	30, 40, 50, 60, and 70 DAP	Microarray	Dhaubhadel et al., 2007
Expression profile of cotyledon during seedling development	Imbibed underground seed for 48 h; radicle 10–15, 16–25 mm long; hypocotyl 40–50 mm; green and yellow cotyledon above the ground, cotyledon mostly green, cotyledon fully green	Microarray	Gonzalez and Vodkin, 2007
Identification of differentially expressed genes in soybean seeds differing in oil content	15, 22, 29, 36, and 43 DAF	Microarray	Wei et al., 2008
Seed developmental stages from mid-maturation to full maturation	25–50, 75–100, 100–200, 400–500 mg at green stage; 200–300 mg at yellow; 100–200 mg at dry seed	Microarray	Jones et al., 2010
High quality gene expression in 14 diverse tissues (aerial, underground and seed tissues)	Pod 7, 10–13, 14–17 DAF; Seed 21, 25, 28, 35 DAF, Roots after 12 DAI; Nodules at 20–25 DAI	RNA-seq	Severin et al., 2010
Seed developmental stages and gene expression profiles	10–50 DAF (with 5 day interval)	Microarray	Sha et al., 2012a
Explore genes involved with lipid biosynthesis during seed development	15 DAF, and then every 5 days until 70 DAF	RNA-seq	Chen et al., 2012
Change in gene expression pattern from the beginning of seed formation	Pod 1 WAF, bean 2-mm and 5-mm, and 12–14 mm	Microarray	Asakura et al., 2012
Tissue and seed developmental stage specific MicroRNAs (miRNAs) identification	25–50 mg, 100–200 mg green; 300–400 mg yellow seed	MiRNA sequencing (Degradome Sequencing	Shamimuzzaman and Vodkin, 2012
Seed development from fertilization to maturity	4, 12–14, 22–24 DAF whole seed; 5–6 mg whole seed; 100–200, 400–500 mg cotyledon, 100–200 mg whole seed, dry	RNA-seq	Jones and Vodkin, 2013
Seed transcriptomes from nine soybean genotypes varying in oil composition and content, and to identify sequence variation in seeds at gene, pathway and systems levels	S6 stage of seed	RNA-seq	Goettel et al., 2014
Gene expression patterns for seed protein and oil synthesis during seed development	1 and 2 WAF	RNA-seq	Jang et al., 2015

DAF, Days After Flowering; WAF, Week After Flowering; DAI, Days After Inoculation; DAP, Days After Pollination.

glycinin (11S legumin type) and conglycinin (7S vicilin type), and it also includes many moderately abundant proteins, such as the kunitz trypsin inhibitors, lectin, P34 allergen, sucrose binding protein, urease, oleosins, and several thousand low abundance proteins (Herman, 2014). Due to the complexity of soybean seed proteins, several proteomic studies have been conducted to understand the protein expression, functions, and interactions. Gel-based, mass-spectrometry based methods or a combination of techniques have been tested in soybean to study global changes during seed development (Nouri and Komatsu, 2010; Oh et al., 2014). However, gel-based methods, such as, two-dimensional gel electrophoresis (2-DGE) is considered a time-consuming technique with poor sensitivity and reproducibility as compared to gel-free methods. Recently, a combinatorial approach, NanoUPLC-MS^E (liquid chromatography combined with mass spectrometry) procedure was used for the assessment of peptide profile in soybean seeds and compared with the Uniprot database (Murad and Rech, 2012). In another study, the high yielding cultivar, Jidou17 and its parental lines were used to examine differentially expressed proteins using an iTRAQ-based (isobaric tags for relative and absolute quantitation) method (Qin et al., 2013). All of these methods are considered to be high-throughput, accurate, sensitive, and more robust than gel-based techniques.

Soybean seed traits have been exploited using proteome profiling with gel-based and gel-free techniques, e.g., peptide mass fingerprinting of seed proteins was performed using 2-DGE and resulted in detection of approximately 150 seed proteins. Most of them were found to be related to seed storage proteins (Mooney and Thelen, 2004). A seed filling protein profile was analyzed at 2, 3, 4, 5, and 6 weeks after flowering using 2-DGE and MALDI-TOF (matrix-assisted laser desorption ionization time-of-flight) mass spectrometry. This study identified 422 proteins including 216 non-redundant proteins. Most abundant proteins were found to be involved in metabolism, protein destination and storage, metabolite transport, and disease/defense (Hajduch et al., 2005). Agrawal et al. (2008) studied the protein profile associated

with seed filling using 2-DGE and semi-continuous multi-dimensional protein identification technology (Sec-MudPIT) coupled with liquid chromatography-mass spectrometry. A total of 478 non-redundant proteins were identified, which were mainly involved in metabolism, protein destination and storage (Agrawal et al., 2008). In another study, Barbosa et al. (2012) compared the expression patterns of seed proteins of transgenic and non-transgenic soybean using 2-D DIGE (difference gel electrophoresis), MALDI quadrupole time-of-flight (QTOF) and electrospray ionization (ESI) QTOF. The phosphoproteomic profile of soybean, rapeseed, and Arabidopsis seed at five developmental stages was analyzed. A total of 2001 phosphopeptides containing 1026 unambiguous phosphorylation sites from 956 proteins were identified. In comparison with other large-scale phosphoproteomic studies, 652 of the phosphoproteins were found to be novel. The unique proteins fall into several gene ontology categories, some of which were found to be involved with metabolic processes, RNA binding, and embryonic development (Meyer et al., 2012). Stored nutrients and their mobilization in soybean seed determines the early plant vigor and these reserve components can be determined by proteomic profiling. A proteome based study revealed different mechanisms for reserve mobilization in soybean and rice during germination (Han et al., 2013).

Technological advances in proteomics methods offer more sensitivity, greater rapidity, and proteome coverage. These techniques and generated data can potentially help to acquire a comprehensive understanding of the physiology of seed reserves. However, the data processing and analysis are still a bottleneck in proteomics studies. To overcome this, it is important to integrate publicly available data and bioinformatics tools into a more robust linear pipeline for soybean seed trait improvement.

METABOLOMICS ADVANCES

Metabolomic studies provide a high-throughput assessment of all metabolites, which represent the complete set of small molecules in the target tissue or organism. Small molecule metabolites in plant composition can be analyzed using metabolomics in combination with sophisticated statistical and computational methods. A metabolomic approach has been widely used to dissect biochemical composition and regulation in plant seeds, such as in rice (Matsuda et al., 2012), maize (Rao et al., 2014), and tomato (Toubiana et al., 2015). Soybean seed contains flavonoids, isoflavones, saponins, phytosterols, and several other metabolites that have a considerable impact on human health. Metabolomics is an emerging field, which has been used to assist in the biochemical analysis of complex mixtures and considered a robust, sensitive, and powerful technology (Nakabayashi and Saito, 2013). Fukusaki and Kobayashi (2005) explained the technical elements, statistical analysis, and practical applications while Putri et al. (2013) elaborated on the latest developments in analytical methods and data analysis in the metabolomics area.

García-Villalba et al. (2008) developed a metabolic profiling method for genetically modified (GM) and conventional soybean lines. This method identified more than 40 compounds and interestingly showed significant quantitative differences between conventional and GM soybean lines. In another study, the soybean seed metabolome (169 metabolites) was assessed in GM and conventional lines. Interestingly, no significant variation was observed between the seed metabolomes except in the engineered targeted pathway (Clarke et al., 2013). In addition, metabolite-metabolite interaction was studied in seed; and a seed metabolic network map was constructed based on 169 metabolites from 29 soybean cultivars. This might be helpful for metabolic engineering to enhance seed quality in soybean (Lin et al., 2014).

Advances in database development and bioinformatics tools are still lagging behind for seed metabolomics area. A database model, ArMet (an architecture for metabolomics) was designed for Arabidopsis and *Solanum tuberosum* metabolomics studies (Jenkins et al., 2004). SoyMetDB (http://soymetdb.org) has been developed for integrating, mining and visualizing metabolomic data from soybeans (Joshi et al., 2010). The potential findings of various metabolic studies performed on different soybean tissues will provide a basis to improve soybean seed quality and perhaps yield. To date, studies focusing on seed specific metabolites are limited; therefore, there is a dire need to create a comprehensive metabolome of the soybean seed.

PHENOMICS DEVELOPMENT

Plant phenotyping is a process of recording quantitative and qualitative plant traits and has been the backbone of breeding programs to improve desired traits. Phenotype includes a set of morphological, structural, physiological, and biochemical traits that characterize a genotype at a given stage, date, or environment. High-throughput genotyping technologies have provided a quantum of genomic information for individual lines and large mapping populations. However, the phenotypic information is not as extensive and provides limited contribution to the advancement of crops due to the labor intensive, time-consuming task of collecting phenotypic data. Plant breeding strategies call for high-throughput, rapid, and accurate phenotyping developments particularly due to the utilization of more crosses, replications, and environments (Araus and Cairns, 2014). Genotypic information is a selection criteria in advanced statistical models for trait improvement but phenotypic data is required at the initial selection of lines which necessitates the development of phenotypic data collection methods. The genomic prediction model is one such example that uses genotypic information for selections; however, the phenotypic information is needed to train a prediction model (Lorenz et al., 2011). The advancements of high-throughput measurement methods of plant traits have helped breeding in various ways. For example, integration of high-throughput phenotyping (HTP) with association studies will be helpful in trait discovery and phenotypic predictions.

HTP platforms have seen advances in non-destructive and time-series based methods. Some examples include image-based computer vision phenotyping, image processing, data extraction tools, and the availability of public datasets. The techniques available due to recent advances provide platforms for data

measurement during different growth stages of individual plants or large mapping populations. Attempts to exploit phenotypic measurements, new software and hardware tools are available such as BreedVision (Busemeyer et al., 2013), TraitCapture (Brown et al., 2014), and Pheno-Copter (Chapman et al., 2014). The major seed composition traits in soybean including total oil, protein, fatty acids, carbohydrate, ash, and moisture can be measured using high resolution nuclear magnetic resonance (HR-NMR) and near infrared (NIR) methods (Baianu et al., 2012). A soybean seed composition database for approximately 15,000 accessions is available with protein, amino acid, oil, fatty acid, isoflavone, carbohydrate, fiber, and moisture data. This data was collected from over 80,000 spectroscopic NIR and NMR measurements from both bulk and single soybean seed samples (Baianu, 2011).

Baianu et al. (2012) first reported the HR-NMR for the determination of amino acids from whole soybean seeds without seed protein extraction and they also reported the calibration models, methodologies, and validation procedures for the measurement. The latest NIR grain analyzer (FOSS) is evolving with the modern electronics and precision optical components for quality data. In a recent study, crude protein and amino acid were determined by utilizing NIR to identify QTL for these two traits (Warrington et al., 2015). Other than seed oil and protein composition, the other components, such as lipoxygenases and secondary metabolites can be measured using low throughput methods. Seed lipoxygenase produces an unpleasant beany flavor and this can be measured with colorimetric assays and single-dimension sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (Suda et al., 1995).

Besides the seed composition traits, another study used a laser light back-scattering imaging technology to identify specific soybean cultivars through charge-coupled device (CCD) camera images (Zhu et al., 2012). In this approach, soybean seed surface was directly illuminated by laser light, and later using image processing technology to categorize different cultivar seeds. Progress in phenotypic characterization falls far behind when compared to the progress made in the high-throughput and automated genotyping techniques. Therefore, there is a need for advancing phenotyping tools and databases development to eliminate the setbacks in the research programs aiming to develop better quality soybeans.

IONOMICS IMPROVEMENT

The advances in the areas of genomics, proteomics, metabolomics, and phenomics have led to the development of elemental profile analysis as well, known as "ionomics." Ionomics can be used for the analysis of physiological, biochemical, elemental, and mineral profiling in living systems with a high-throughput and cost effective method (Baxter, 2010). The elemental composition of soybean seed is an important component of their overall nutritional value and is controlled by element availability during seed development. Hence, understanding the ionome of soybean seed and its correlation with genetic factors has the potential to improve seed composition and nutrient values. Typically, the conventional methodologies for elemental analysis are based on either the electronic properties of an atom (emission, absorption, and fluorescence spectroscopy) or nuclear properties (radioactivity or atomic number; Salt et al., 2008). In addition, advances in mass spectroscopy technology, such as inductively coupled plasma (ICP), has helped in multiple element analysis and enables complete ionome instead of just individual elements.

Ionomics has been used to compare transgenic and non-transgenic soybeans (Yan et al., 2007; Mataveli et al., 2010, 2012). Similarly, another study proposed a combination of microwave-induced combustion (MIC) and ICP-MS to analyze bromine, chlorine, iodine, and the associated products in soybeans (Barbosa et al., 2013). Recently, a high-throughput approach was utilized to identify seed elemental composition in 947 mutagenized lines (Ziegler et al., 2013). In this study, they identified mutants with modified seed element profiles. Sha et al. (2012b), performed ionome analysis of soybean seed and concluded that the seed ionome is affected by the cropping system and manure applications. In addition to the essential elements, many crop plants including soybean, may accumulate the non-essential or toxic elements such as cadmium and lead (Sugiyama et al., 2011). With the advantages of low cost, high-throughput capabilities compared to the proteomics and metabolomics, ionomics became a powerful approach to understand complex biological systems controlling elemental accumulation in plants. Thus, ionome study can be utilized as an effective tool to build connections not only with genome, metabolome, and physiological processes of plant, but also with environment and ecology for elemental variation between genotypes, loci/genes identification.

GENETIC ENGINEERING

The availability of the sequence information provides new avenues for the engineering of soybean seed composition. The new "omics" approaches offers a potential resource to comprehend the metabolic regulatory network governing seed storage compound accumulation. To accommodate the global food demand and growing population needs, several genetic engineering and genome editing strategies need to be employed to obtain the next-generation crop at faster rate than the conventional breeding. Studies have reported improved soybean seed oil content via X-ray, and EMS-induced mutations in the desired genes (Dierking and Bilyeu, 2009; Pham et al., 2010, 2011). For example, deletion of a 100 kb sequence on Chr. 10 in a mutant line M23 increased oleic acid content but caused yield drag (Pham et al., 2011). Similarly, the SACPD-C deletion in an A6 mutant line showed elevated stearic acid content; however, simultaneously a reduced seed yield was also observed (Gillman et al., 2014). Nevertheless, conventional methods require labor intensive screening procedure to identify germplasm accessions with desired phenotypes and stable performance in multiple generations and environments.

Genetic engineering entails a range of activities for the benefit of agriculture, such as creation of genetically modified (GM) or transgenic crops. A number of transgenic approaches have been explored for modifying the seed composition traits,

including total oil (Lardizabal et al., 2008), protein (Schmidt et al., 2011), oleic acid (Haun et al., 2014), phytosterols (Neelakandan et al., 2012; Nguyen et al., 2013), tocopherols (Karunanandaa et al., 2005), and isoflavone (Yu et al., 2003). Marginal enhancement of total oil content was achieved by seed-specific expression of DGAT2 enzyme from oil accumulating fungi (Lardizabal et al., 2008). In another study, isoflavone levels in soybean seeds were increased via metabolic engineering of the complex phenylpropanoid biosynthesis pathway (Yu et al., 2003). Various techniques, such as transgene-mediated RNA interference (RNAi), has been successfully employed by researchers in enhancing soybean seed oil (Buhr et al., 2002; Wagner et al., 2011).

The genetically modified organisms (GMOs) have quite a few limitations, such as social acceptance, cost-related concerns (regulatory and licensing), limited resources, and time. To overcome these drawbacks, genome editing has arisen as a modern approach for specifically targeting and modifying DNA sequences for crop improvement and is considered a non-GM approach (Voytas and Gao, 2014). Several methodologies are available for genome engineering, such as engineered homing endonucleases or meganucleases, zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs). These are being used to mutagenize genomes at precise locations (Cermak et al., 2011; Sander et al., 2011). An optimized mutagenesis approach is required for the simultaneous recovery of plants with single or multiple mutations in a gene of interest without interfering with the rest of the genetic background. Therefore, site-directed mutagenesis could be a powerful method for producing the desired results. Moreover, CRISPR (clustered regularly interspaced short palindromic repeats)/Cas (CRISPR-associated) has emerged as a simpler method that permits introduction of desired SNP into a gene of interest without a need for extensive design and time-consuming assembly of individual DNA-binding proteins (Belhaj et al., 2013; Bortesi and Fischer, 2014; Voytas and Gao, 2014). In addition, Cas9 has been reported as a successful method for genome editing in tobacco (Nekrasov et al., 2013), rice, wheat, Arabidopsis, tobacco, and sorghum (Jiang et al., 2013), and sweet orange (Jia and Wang, 2014). Targeted genome editing has been tested for soybean seed composition improvement. The context-dependent assembly (CoDA) was used to create ZFNs for target specific mutagenesis (*DICER-LIKE* and other genes involved in RNA silencing) in soybean. This approach showed successful heritable transmission of the ZFN-induced mutation in subsequent generations and could be useful for making mutations in duplicated genes efficiently (Curtin et al., 2011). While in another study, mutations were induced in two fatty acid desaturase-2 genes (*FAD2-1A* and *FAD2-1B*) using TALENs. The changed fatty acid profile of the seed with increased oleic acid (20–80%) and decreased linoleic acid (from 50% to less than 4%) was reported for *FAD2-1A* and *FAD2-1B* mutations (Haun et al., 2014). In comparison to other transgenic technologies, the CRISPR system is considered a non-GM technology, since the CAS9 can be deleted from the host plant via backcrossing in subsequent generations, once the mutation is accomplished. The usefulness of the CRISPR system being a simple and cost effective technique for genome editing in soybean can't be denied. This approach can be applied to confirm candidate genes, novel alleles/phenotypes, and engineer soybeans with high quality seed traits.

INTEGRATED "OMICS" APPROACHES

With the improvements in different "omics" approaches and the development of computational tools; this has provided information related to gene function, genome structures, biological pathways, metabolic and regulatory networks and has greatly contributed to the understanding of plant systems. Although several biological network models are available, they do not always provide a complete depiction of cellular and molecular networks solely based on genome, transcriptome, or metabolome data. In the soybean research community, several groups have re-sequenced soybean germplasm (Chung et al., 2014; Li et al., 2014; Qiu et al., 2014; Zhou et al., 2015). Their studies were focused on population genetics and also to understand the domestication process of cultivated soybeans (**Table 1**). Secondly, integrated approaches have not been explored to study the interacting networks between genomic sequences and transcript profiles. Thus, an integrated data analysis approach is essential to investigate and fully understand the physiological, biochemical, and molecular interactions. For instance, QTL mapping and GWAS are helpful tools in the identification of chromosomal regions that relate to phenotypic traits; however, underlying genes in that region are in large numbers and the candidate gene is not identified with QTL or GWAS alone (Deshmukh et al., 2014). The integrated genomic approach, combined with WGRS data and seed development related transcriptome for diverse germplasm and RILs could identify a novel/uncharacterized variant-linked co-expression network associated with seed composition traits.

In soybeans, a combination of these approaches has led to successful discoveries, for instance, Kovinich et al. (2011) combined gene expression and metabolite data to elucidate the control of the R locus identification of pigment biosynthesis genes (Kovinich et al., 2011). In a similar study, metabolic and transcriptional changes were assessed in developing soybean seeds to identify metabolic engineering targets. The study concluded that transcriptional activation and signaling involvement was much higher during seed maturation and dormancy (Collakova et al., 2013). The integration of expression QTL (eQTL) and phenotypic QTL (pQTL) was employed in the identification of genes for isoflavone content in soybean seeds and 11 potential candidate genes were identified (Wang et al., 2014a). Recently, ionomics and metabolomics were coupled for a comprehensive assessment of GM and non-GM soybean lines (Kusano et al., 2015). Li et al. (2015) found the highest metabolic flux during early seed fill by integrating metabolomics and transcriptomics analysis. Furthermore, the metabolic flux was found to be consistent with regard to the transcript and metabolite level changes during the seed development stages. All of these studies clearly illustrated that an integrated "omics" approach needs to be applied for better understanding of seed composition traits in soybean.

AVAILABLE ONLINE RESOURCES

Recent advancements in the "omics" methodologies and approaches in the past decade has provided a large amount of data with the objective of discovering key genes associated with phenotypic traits. The acquired data and information are available at several public databases, which facilitates sharing of the generated information. The availability of integrated and focused databases from different dimensions have allowed phenotypic trait improvement in soybean as well. The online available databases include user-friendly interfaces, such as chromosomal visualizer, omics datasets, computational comparisons, and search tools that allow easy data analysis for a particular objective. A considerable number of soybean information resources are freely accessible and are summarized in **Table 3**. The available data in the public databases for various applications will facilitate and accelerate molecular elucidation of cellular system linked to agronomically important traits.

CONCLUDING REMARKS

The advancements in NGS technologies have made huge amounts of sequence information publicly available and now it is time to integrate and utilize them for the improvement of seed composition and other traits (**Table 1**; **Figures 1**, **2**). Hundreds of QTL for seed oil, protein, and other seed component traits have been identified and reported; however, very few studies were incorporated in breeding programs (Supplementary Table 2). The negative correlation, marginal improvement, and low stability across different geographical locations have undermined the development of cultivars with high meal quality, oil and protein content, and yield. However, for protein improvement with better yield and maintained oil content, attempts are being made by utilizing identified QTL. Most of the commercial soybean cultivars are fixed for the low protein allele at the major QTL on Chr. 20, suggesting that introgression of desired allele at Chr. 20 into an existing elite background would enhance protein content. Also, since there is no single commercial soybean cultivar with the FAO standard for total sulfur containing amino acids; therefore, it presents an opportunity for soybean researchers to improve nutrient values of soybean seed using integrated approaches. It is essential that breeding efforts should be made considering the overall improvement of protein, oil, and yield traits in soybeans. As evident from the discussion above, advances in "omics" need to be integrated, particularly in genomics, proteomics, metabolomics, phenomics, and ionomics for soybean seed composition. Strategies to dissect the inverse relationship and environmental stability of seed storage protein with oil and yield needs to be designed. This could be achieved by uniting "omics" and conventional approaches.

A number of potential strategies have been outlined, including combining diverse and wild germplasm associated with specific seed traits and studying them in various "omics" dimensions to enhance our knowledge. It is projected that the combined

TABLE 3 | List of wide-ranging available online resources for soybean.

Web Name	URL	Description/Tools/Applications	References
Soybean Genomics and Microarray Database (SGMD)	http://bioinformatics.towson.edu/	Genomic, EST, and microarray database	Alkharouf and Matthews, 2004
Soybean transcriptome database (SoyXpress)	http://soyxpress.agrenv.mcgill.ca	Metabolic pathways, EST sequences, Microarray and Affymetrix gene expression data	Cheng and Strömvik, 2008
Soybean Proteome Database	http://proteome.dc.affrc.go.jp/Soybean/	Proteome, Metabolome, Transcriptome datasets, 2D-PAGE and proteomics information, comparative proteomics under flooding, drought and salt stress.	Sakata et al., 2009
SoyBase	http://www.soybase.org	Genetic and physical maps, Genome sequence, Transposable elements, Annotations, Graphical chromosome visualizer	Grant et al., 2010
Soybean Knowledge Base (SoyKB)	http://soykb.org/	Graphical chromosome visualizer, Genes/proteins, miRNAs/sRNAs, Metabolite profiling, Molecular markers, Plant Introduction and traits information	Joshi et al., 2012
Soybean proteins database (SoyProDB)	http://bioinformatics.towson.edu/Soybean_Seed_Proteins_2D_Gel_DB/Home.aspx	Seed protein identification, 2D gel image data	Tavakolan et al., 2013
Soybean Cyst Nematode proteins database (SCNProDB)	http://bioinformatics.towson.edu/	SCN protein identification, 2D gel images data	Natarajan et al., 2014
Soybean Functional Network (SoyFN)	http://nclab.hit.edu.cn/SoyFN	Functional gene network, microRNA functional network, gene annotation, genome browser	Xu et al., 2014
Soybean Functional Genomics Database (SFGD)	http://bioinformatics.cau.edu.cn/SFGD/	Gbrowse, microarray expression profiling, transcriptome data, gene co-expression regulatory network, acyl-lipid metabolism pathways, cis-element significance analysis	Yu et al., 2014

soybean seed content of protein and oil should be increased by 10% by 2025 (http://unitedsoybean.org/about-usb/strategic-planning/). Most of the US soybeans have 59–62% combined protein and oil content. It is a great challenge to develop soybeans with >70% combined protein and oil. In addition, there is a need to develop soybean germplasm with increased meal protein (48–50%) by improving amino acid balance (10% increase in methionine, cysteine, and threonine) and environmental stability without reducing the seed oil or yield. Therefore, it is essential to identify and validate independent novel QTL for high protein content and to breakdown the negative correlation with yield and oil content. To meet this challenge, it is necessary to bridge the gap between biological and computational system by integrating multiple "omics" approaches, statistical genetic models, and bioinformatics tools, such as data pipelines, cloud computing, and user-friendly script developments. Despite the challenges, advances in "omics" technologies offers a promising potential to create next-generation soybeans with the desired seed composition traits. With constant developments in breeding technologies in conjunction with "omics" approaches, it is foreseeable in the future that a high yielding soybean cultivar with balanced amino acid, increased oil and protein content can be subsequently developed. In summary, this review provides a glimpse of advances made in improvement of soybean seed traits using integrated omics approaches. The discussed information would be a useful resource to accelerate desired seed composition improvement and, in turn, meet the global soybean demand in coming years.

ACKNOWLEDGMENTS

Authors acknowledge Missouri Soybean Merchandising Council and United Soybean Board, USA, for the funding support to our seed composition research program. The authors also would like to thank Theresa A. Musket for the language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2015.01021

REFERENCES

Agrawal, G. K., Hajduch, M., Graham, K., and Thelen, J. J. (2008). In-depth investigation of the soybean seed-filling proteome and comparison with a parallel study of rapeseed. *Plant Physiol.* 148, 504–518. doi: 10.1104/pp.108.119222

Akond, M., Liu, S., Schoener, L., Anderson, J. A., Kantartzi, S. K., Meksem, K., et al. (2013). SNP-Based genetic linkage map of soybean using the SoySNP6K Illumina Infinium BeadChip genotyping array. *J. Plant Genom. Sci.* 1, 80–89. doi: 10.5147/jpgs.2013.0090

Alkharouf, N. W., and Matthews, B. F. (2004). SGMD: the soybean genomics and microarray database. *Nucleic Acids Res.* 32, D398–D400. doi: 10.1093/nar/gkh126

Araus, J. L., and Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends Plant Sci.* 19, 52–61. doi: 10.1016/j.tplants.2013.09.008

Asakura, T., Tamura, T., Terauchi, K., Narikawa, T., Yagasaki, K., Ishimaru, Y., et al. (2012). Global gene expression profiles in developing soybean seeds. *Plant Physiol. Biochem.* 52, 147–153. doi: 10.1016/j.plaphy.2011.12.007

Baianu, I. (2011). Soybean composition database from NIR, NMR and GC-MS analyses. *Nat. Proc.* 1–60. doi: 10.1038/npre.2011.6201.3

Baianu, I., You, T., Costescu, D., Lozano, P., Prisecaru, V., and Nelson, R. L. (2012). Determination of soybean oil, protein and amino acid residues in soybean seeds by high resolution nuclear magnetic resonance (NMRS) and near infrared (NIRS). *Nat. Proc.* 1–62. doi: 10.1038/npre.2012.7053.1

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3:e3376. doi: 10.1371/journal.pone.0003376

Bao, Y., Vuong, T., Meinhardt, C., Tiffin, P., Denny, R., Chen, S., et al. (2014). Potential of association mapping and genomic selection to explore PI 88788 derived soybean cyst nematode resistance. *Plant Genome* 7, 1–13. doi: 10.3835/plantgenome2013.11.0039

Barbosa, H. S., Arruda, S. C., Azevedo, R. A., and Arruda, M. A. (2012). New insights on proteomics of transgenic soybean seeds: evaluation of differential expressions of enzymes and proteins. *Anal. Bioanal. Chem.* 402, 299–314. doi: 10.1007/s00216-011-5409-1

Barbosa, J. T. P., Santos, C. M. M., Dos Santos Bispo, L., Lyra, F. H., David, J. M., Korn, M. D. G. A., et al. (2013). Bromine, chlorine, and iodine determination in soybean and its products by ICP-MS after digestion using microwave-induced combustion. *Food Anal. Methods* 6, 1065–1070. doi: 10.1007/s12161-012-9511-6

Baxter, I. (2010). Ionomics: the functional genomics of elements. *Brief. Funct. Genomics* 9, 149–156. doi: 10.1093/bfgp/elp055

Belhaj, K., Chaparro-Garcia, A., Kamoun, S., and Nekrasov, V. (2013). Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. *Plant Methods* 9:39. doi: 10.1186/1746-4811-9-39

Bellaloui, N., Bruns, H. A., Abbas, H. K., Mengistu, A., Fisher, D. K., and Reddy, K. N. (2015). Agricultural practices altered soybean seed protein, oil, fatty acids, sugars, and minerals in the Midsouth USA. *Front. Plant Sci.* 6:31. doi: 10.3389/fpls.2015.00031

Bentley, A. R., Scutari, M., Gosman, N., Faure, S., Bedford, F., Howell, P., et al. (2014). Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. *Theor. Appl. Genet.* 127, 2619–2633. doi: 10.1007/s00122-014-2403-y

Borevitz, J. O., and Nordborg, M. (2003). The impact of genomics on the study of natural variation in Arabidopsis. *Plant Physiol.* 132, 718–725. doi: 10.1104/pp.103.023549

Bortesi, L., and Fischer, R. (2014). The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnol. Adv.* 33, 41–52. doi: 10.1016/j.biotechadv.2014.12.006

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* 12:232. doi: 10.1186/gb-2011-12-10-232

Brown, T. B., Cheng, R., Sirault, X. R., Rungrat, T., Murray, K. D., Trtilek, M., et al. (2014). TraitCapture: genomic and environment modelling of plant phenomic data. *Curr. Opin. Plant Biol.* 18, 73–79. doi: 10.1016/j.pbi.2014.02.002

Buhr, T., Sato, S., Ebrahim, F., Xing, A., Zhou, Y., Mathiesen, M., et al. (2002). Ribozyme termination of RNA transcripts down−regulate seed fatty acid genes in transgenic soybean. *Plant J.* 30, 155–163. doi: 10.1046/j.1365-313X.2002.01283.x

Busemeyer, L., Mentrup, D., Möller, K., Wunder, E., Alheit, K., Hahn, V., et al. (2013). BreedVision—A multi-sensor platform for non-destructive field-based phenotyping in plant breeding. *Sensors* 13, 2830–2847. doi: 10.3390/s130302830

Cermak, T., Doyle, E. L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., et al. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* 39:e82. doi: 10.1093/nar/gkr218

Chapman, S. C., Merz, T., Chan, A., Jackway, P., Hrabar, S., Dreccer, M. F., et al. (2014). Pheno-copter: a low-altitude, autonomous remote-sensing robotic helicopter for high-throughput field-based phenotyping. *Agronomy* 4, 279–301. doi: 10.3390/agronomy4020279

Chen, H., Wang, F.-W., Dong, Y.-Y., Wang, N., Sun, Y.-P., Li, X.-Y., et al. (2012). Sequence mining and transcript profiling to explore differentially expressed genes associated with lipid biosynthesis during soybean seed development. *BMC Plant Biol.* 12:122. doi: 10.1186/1471-2229-12-122

Cheng, K. C., and Strömvik, M. V. (2008). SoyXpress: a database for exploring the soybean transcriptome. *BMC Genomics* 9:368. doi: 10.1186/1471-2164-9-368

Chung, W.-H., Jeong, N., Kim, J., Lee, W. K., Lee, Y.-G., Lee, S.-H., et al. (2014). Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. *DNA Res.* 21, 153–167. doi: 10.1093/dnares/dst047

Clarke, J. D., Alexander, D. C., Ward, D. P., Ryals, J. A., Mitchell, M. W., Wulff, J. E., et al. (2013). Assessment of genetically modified soybean in relation to natural variation in the soybean seed metabolome. *Sci. Rep.* 3:3082. doi: 10.1038/srep03082

Clemente, T. E., and Cahoon, E. B. (2009). Soybean oil: genetic approaches for modification of functionality and total content. *Plant Physiol.* 151, 1030–1040. doi: 10.1104/pp.109.146282

Collakova, E., Aghamirzaie, D., Fang, Y., Klumas, C., Tabataba, F., Kakumanu, A., et al. (2013). Metabolic and transcriptional reprogramming in developing soybean (Glycine max) embryos. *Metabolites* 3, 347–372. doi: 10.3390/metabo3020347

Cromwell, D. (2012). *Soybean Meal–An Exceptional Protein Source*. Lexington, KY: Animal and food sciences department university of Kentucky.

Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Curtin, S. J., Zhang, F., Sander, J. D., Haun, W. J., Starker, C., Baltes, N. J., et al. (2011). Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. *Plant Physiol.* 156, 466–473. doi: 10.1104/pp.111.172981

Deshmukh, R., Sonah, H., Patil, G., Chen, W., Prince, S., Mutava, R., et al. (2014). Integrating omic approaches for abiotic stress tolerance in soybean. *Front. Plant Sci.* 5:244. doi: 10.3389/fpls.2014.00244

Desta, Z. A., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* 19, 592–601. doi: 10.1016/j.tplants.2014.05.006

Dhaubhadel, S., Gijzen, M., Moy, P., and Farhangkhoee, M. (2007). Transcriptome analysis reveals a critical role of CHS7 and CHS8 genes for isoflavonoid synthesis in soybean seeds. *Plant Physiol.* 143, 326–338. doi: 10.1104/pp.106.086306

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol.* 8:e1000294. doi: 10.1371/journal.pbio.1000294

Dierking, E. C., and Bilyeu, K. D. (2009). New sources of soybean seed meal and oil composition traits identified through TILLING. *BMC Plant Biol.* 9:89. doi: 10.1186/1471-2229-9-89

Eldakak, M., Milad, S. I., Nawar, A. I., and Rohila, J. S. (2013). Proteomics: a biotechnology tool for crop improvement. *Front. Plant Sci.* 4:35. doi: 10.3389/fpls.2013.00035

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

Eskandari, M., Cober, E. R., and Rajcan, I. (2013). Genetic control of soybean seed oil: I. QTL and genes associated with seed oil concentration in RIL populations derived from crossing moderately high-oil parents. *Theor. Appl. Genet.* 126, 483–495. doi: 10.1007/s00122-012-1995-3

Fan, C.-M., Wang, X., Wang, Y.-W., Hu, R.-B., Zhang, X.-M., Chen, J.-X., et al. (2013). Genome-wide expression analysis of soybean MADS genes showing potential function in the seed development. *PLoS ONE* 8:e62288. doi: 10.1371/journal.pone.0062288

Fukusaki, E., and Kobayashi, A. (2005). Plant metabolomics: potential for practical operation. *J. Biosci. Bioeng.* 100, 347–354. doi: 10.1263/jbb.100.347

García-Villalba, R., León, C., Dinelli, G., Segura-Carretero, A., Fernández-Gutiérrez, A., Garcia-Cañas, V., et al. (2008). Comparative metabolomic study of transgenic versus conventional soybean using capillary electrophoresis–time-of-flight mass spectrometry. *J. Chromatogr. A* 1195, 164–173. doi: 10.1016/j.chroma.2008.05.018

Gepts, P., Beavis, W. D., Brummer, E. C., Shoemaker, R. C., Stalker, H. T., Weeden, N. F., et al. (2005). Legumes as a model plant family. Genomics for food and feed report of the cross-legume advances through genomics conference. *Plant Physiol.* 137, 1228–1235. doi: 10.1104/pp.105.060871

Gillman, J. D., Stacey, M. G., Cui, Y., Berg, H. R., and Stacey, G. (2014). Deletions of the SACPD-C locus elevate seed stearic acid levels but also result in fatty acid and morphological alterations in nitrogen fixing nodules. *BMC Plant Biol.* 14:143. doi: 10.1186/1471-2229-14-143

Goettel, W., Xia, E., Upchurch, R., Wang, M.-L., Chen, P., and An, Y.-Q. C. (2014). Identification and characterization of transcript polymorphisms in soybean lines varying in oil composition and content. *BMC Genomics* 15:299. doi: 10.1186/1471-2164-15-299

Goldberg, B., and Stacey, G. (2008). *Genetics and Genomics of Soybean*. New York, NY: Springer Science & Business Media.

Gonzalez, D. O., and Vodkin, L. O. (2007). Specific elements of the glyoxylate pathway play a significant role in the functional transition of the soybean cotyledon during seedling development. *BMC Genomics* 8:468. doi: 10.1186/1471-2164-8-468

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, D843–D846. doi: 10.1093/nar/gkp798

Hagely, K. B., Palmquist, D., and Bilyeu, K. D. (2013). Classification of distinct seed carbohydrate profiles in soybean. *J. Agric. Food Chem.* 61, 1105–1111. doi: 10.1021/jf303985q

Hajduch, M., Ganapathy, A., Stein, J. W., and Thelen, J. J. (2005). A systematic proteomic study of seed filling in soybean. Establishment of high-resolution two-dimensional reference maps, expression profiles, and an interactive proteome database. *Plant Physiol.* 137, 1397–1419. doi: 10.1104/pp.104.056614

Han, C., Yin, X., He, D., and Yang, P. (2013). Analysis of proteome profile in germinating soybean seed, and its comparison with rice showing the styles of reserves mobilization in different crops. *PLoS ONE* 8:e56947. doi: 10.1371/journal.pone.0056947

Haun, W., Coffman, A., Clasen, B. M., Demorest, Z. L., Lowy, A., Ray, E., et al. (2014). Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. *Plant Biotech. J.* 12, 934–940. doi: 10.1111/pbi.12201

Heffner, E. L., Jannink, J.-L., and Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* 4, 65–75. doi: 10.3835/plantgenome.2010.12.0029

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.2135/cropsci2008.08.0512

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

Hu, Z., Li, Y., Song, X., Han, Y., Cai, X., Xu, S., et al. (2011). Genomic value prediction for quantitative traits under the epistatic model. *BMC Genet.* 12:15. doi: 10.1186/1471-2156-12-15

Hwang, E.-Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., et al. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15:1. doi: 10.1186/1471-2164-15-1

Hyten, D. L., Pantalone, V. R., Sams, C. E., Saxton, A. M., Landau-Ellis, D., Stefaniak, T. R., et al. (2004). Seed quality QTL in a prominent soybean population. *Theor. Appl. Genet.* 109, 552–561. doi: 10.1007/s00122-004-1661-5

Iwata, H., Hayashi, T., Terakami, S., Takada, N., Sawamura, Y., and Yamamoto, T. (2013). Potential assessment of genome-wide association study and genomic selection in Japanese pear Pyrus pyrifolia. *Breed. Sci.* 63, 125. doi: 10.1270/jsbbs.63.125

Jang, Y. E., Kim, M. Y., Shim, S., Lee, J., and Lee, S.-H. (2015). Gene expression profiling for seed protein and oil synthesis during early seed development in soybean. *Genes Genomics* 37, 409–418. doi: 10.1007/s13258-015-0269-2

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., et al. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15:740. doi: 10.1186/1471-2164-15-740

Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A. R., Taylor, J., et al. (2004). A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.* 22, 1601–1606. doi: 10.1038/nbt1041

Jia, H., and Wang, N. (2014). Targeted genome editing of sweet orange using Cas9/sgRNA. *PLoS ONE* 9:e93806. doi: 10.1371/journal.pone.0093806

Jiang, W., Zhou, H., Bi, H., Fromm, M., Yang, B., and Weeks, D. P. (2013). Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in *Arabidopsis*, tobacco, sorghum and rice. *Nucleic Acids Res.* 41:e188. doi: 10.1093/nar/gkt780

Jones, S. I., Gonzalez, D. O., and Vodkin, L. O. (2010). Flux of transcript patterns during soybean seed development. *BMC Genomics* 11:136. doi: 10.1186/1471-2164-11-136

Jones, S. I., and Vodkin, L. O. (2013). Using RNA-Seq to profile soybean seed development from fertilization to maturity. *PLoS ONE* 8:e59270. doi: 10.1371/journal.pone.0059270

Joshi, T., Patil, K., Fitzpatrick, M. R., Franklin, L. D., Yao, Q., Cook, J. R., et al. (2012). Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. *BMC Genomics* 13:S15. doi: 10.1186/1471-2164-13-S1-S15

Joshi, T., Valliyodan, B., Wu, J.-H., Lee, S.-H., Xu, D., and Nguyen, H. T. (2013). Genomic differences between cultivated soybean, G. max and its wild relative G. soja. *BMC Genomics* 14:S5. doi: 10.1186/1471-2164-14-S1-S5

Joshi, T., Yao, Q., Franklin, L. D., Brechenmacher, L., Valliyodan, B., Stacey, G., et al. (2010). "SoyMetDB: The Soybean Metabolome Database. Bioinformatics and Biomedicine (BIBM)," in *2010 EEE International Conference on Bioinformatics and Biomedicine* (Hong Kong), 203–208.

Karunanandaa, B., Qi, Q., Hao, M., Baszis, S. R., Jensen, P. K., Wong, Y.-H. H., et al. (2005). Metabolically engineered oilseed crops with enhanced seed tocopherol. *Metab. Eng.* 7, 384–400. doi: 10.1016/j.ymben.2005.05.005

Kim, M. Y., Lee, S., Van, K., Kim, T.-H., Jeong, S.-C., Choi, I.-Y., et al. (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. *Proc. Natl. Acad. Sci. U.S.A.* 107, 22032–22037. doi: 10.1073/pnas.1009526107

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. doi: 10.1186/1746-4811-9-29

Kovinich, N., Saleem, A., Arnason, J. T., and Miki, B. (2011). Combined analysis of transcriptome and metabolite data reveals extensive differences between black and brown nearly-isogenic soybean (Glycine max) seed coats enabling the identification of pigment isogenes. *BMC Genomics* 12:381. doi: 10.1186/1471-2164-12-381

Kusano, M., Baxter, I., Fukushima, A., Oikawa, A., Okazaki, Y., Nakabayashi, R., et al. (2015). Assessing metabolomic and chemical diversity of a soybean lineage representing 35 years of breeding. *Metabolomics* 11, 261–270. doi: 10.1007/s11306-014-0702-6

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42, 1053–1059. doi: 10.1038/ng.715

Lardizabal, K., Effertz, R., Levering, C., Mai, J., Pedroso, M. C., Jury, T., et al. (2008). Expression of Umbelopsis ramanniana DGAT2A in seed increases oil in soybean. *Plant Physiol.* 148, 89–96. doi: 10.1104/pp.108.123042

Lee, J.-D., Bilyeu, K. D., Pantalone, V. R., Gillen, A. M., So, Y.-S., and Shannon, J. G. (2012). Environmental stability of oleic acid concentration in seed oil for soybean lines with and mutant genes. *Crop Sci.* 52, 1290–1297. doi: 10.2135/cropsci2011.07.0345

Lee, Y. G., Jeong, N., Kim, J. H., Lee, K., Kim, K. H., Pirani, A., et al. (2015). Development, validation, and genetic analysis of a large soybean SNP genotyping array. *Plant J.* 8, 625–636. doi: 10.1111/tpj.12755

Li, L., Hur, M., Lee, J.-Y., Zhou, W., Song, Z., Ransom, N., et al. (2015). A systems biology approach toward understanding seed composition in soybean. *BMC Genomics* 16:S9. doi: 10.1186/1471-2164-16-s3-s9

Li, Y.-H., Zhao, S.-C., Ma, J.-X., Li, D., Yan, L., Li, J., et al. (2013). Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* 14:579. doi: 10.1186/1471-2164-14-579

Li, Y.-H., Zhou, G., Ma, J., Jiang, W., Jin, L.-G., Zhang, Z., et al. (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32, 1045–1052. doi: 10.1038/nbt.2979

Lin, H., Rao, J., Shi, J., Hu, C., Cheng, F., Wilson, Z. A., et al. (2014). Seed metabolomic study reveals significant metabolite variations and correlations among different soybean cultivars. *J. Integr. Plant Biol.* 56, 826–836. doi: 10.1111/jipb.12228

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., et al. (2011). 2 genomic selection in plant breeding: knowledge and prospects. *Adv. Agron.* 110, 77. doi: 10.1016/B978-0-12-385531-2.00002-5

Mataveli, L. R. V., Fioramonte, M., Gozzo, F. C., and Arruda, M. A. Z. (2012). Improving metallomics information related to transgenic and non-transgenic soybean seeds using 2D-HPLC-ICP-MS and ESI-MS/MS. *Metallomics* 4, 373–378. doi: 10.1039/c2mt00186a

Mataveli, L. R. V., Pohl, P., Mounicou, S., Arruda, M. A. Z., and Szpunar, J. (2010). A comparative study of element concentrations and binding in transgenic and non-transgenic soybean seeds. *Metallomics* 2, 800–805. doi: 10.1039/c0mt00040j

Matsuda, F., Okazaki, Y., Oikawa, A., Kusano, M., Nakabayashi, R., Kikuchi, J., et al. (2012). Dissection of genotype–phenotype associations in rice grains using metabolome quantitative trait loci analysis. *Plant J.* 70, 624–636. doi: 10.1111/j.1365-313X.2012.04903.x

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nat. Rev. Genetics* 11, 31–46. doi: 10.1038/nrg2626

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Meyer, L. J., Gao, J., Xu, D., and Thelen, J. J. (2012). Phosphoproteomic analysis of seed maturation in Arabidopsis, rapeseed, and soybean. *Plant Physiol.* 159, 517–528. doi: 10.1104/pp.111.191700

Mooney, B. P., and Thelen, J. J. (2004). High-throughput peptide mass fingerprinting of soybean seed proteins: automated workflow and utility of UniGene expressed sequence tag databases for protein identification. *Phytochemistry* 65, 1733–1744. doi: 10.1016/j.phytochem.2004.04.011

Murad, A. M., and Rech, E. L. (2012). NanoUPLC-MSE proteomic data assessment of soybean seeds using the Uniprot database. *BMC Biotechnol.* 12:82. doi: 10.1186/1472-6750-12-82

Nakabayashi, R., and Saito, K. (2013). Metabolomics for unknown plant metabolites. *Anal. Bioanal. Chem.* 405, 5005–5011. doi: 10.1007/s00216-013-6869-2

Natarajan, S., Tavakolan, M., Alkharouf, N. W., and Matthews, B. F. (2014). SCNProDB: a database for the identification of soybean cyst nematode proteins. *Bioinformation* 10:387. doi: 10.6026/97320630010387

Neelakandan, A. K., Chamala, S., Valliyodan, B., Nes, W. D., and Nguyen, H. T. (2012). Metabolic engineering of soybean affords improved phytosterol seed traits. *Plant Biotech. J.* 10, 12–19. doi: 10.1111/j.1467-7652.2011.00623.x

Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J. D., and Kamoun, S. (2013). Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* 31, 691–693. doi: 10.1038/nbt.2655

Nguyen, H. T., Neelakadan, A. K., Quach, T. N., Valliyodan, B., Kumar, R., Zhang, Z., et al. (2013). Molecular characterization of Glycine max squalene synthase genes in seed phytosterol biosynthesis. *Plant Physiol. Biochem.* 73, 23–32. doi: 10.1016/j.plaphy.2013.07.018

Nichols, D., Glover, K., Carlson, S., Specht, J., and Diers, B. (2006). Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. *Crop Sci.* 46, 834–839. doi: 10.2135/cropsci2005.05-0168

Nouri, M. Z., and Komatsu, S. (2010). Comparative analysis of soybean plasma membrane proteins under osmotic stress using gel—based and LC MS/MS—based proteomics approaches. *Proteomics* 10, 1930–1945. doi: 10.1002/pmic.200900632

Oh, M., Nanjo, Y., and Komatsu, S. (2014). Gel-free proteomic analysis of soybean root proteins affected by calcium under flooding stress. *Front. Plant Sci.* 5:559. doi: 10.3389/fpls.2014.00559

O'Rourke, J. A., Bolon, Y.-T., Bucciarelli, B., and Vance, C. P. (2014). Legume genomics: understanding biology through DNA and RNA sequencing. *Ann. Bot.* 113, 1107–1120. doi: 10.1093/aob/mcu072

Pathan, S. M., Vuong, T., Clark, K., Lee, J.-D., Shannon, J. G., Roberts, C. A., et al. (2013). Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. *Crop Sci.* 53, 765–774. doi: 10.2135/cropsci2012.03.0153

Patil, G., Valliyodan, B., Deshmukh, R., Prince, S., Nicander, B., Zhao, M., et al. (2015). Soybean (Glycine max) SWEET gene family: insights through comparative genomics, transcriptome profiling and whole genome re-sequence analysis. *BMC Genomics* 16:520. doi: 10.1186/s12864-015-1730-y

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* 7:e37135. doi: 10.1371/journal.pone.0037135

Pham, A.-T., Bilyeu, K., Chen, P., Boerma, H. R., and Li, Z. (2014). Characterization of the fan1 locus in soybean line A5 and development of molecular assays for high-throughput genotyping of FAD3 genes. *Mol. Breeding* 33, 895–907. doi: 10.1007/s11032-013-0003-1

Pham, A.-T., Lee, J.-D., Shannon, J. G., and Bilyeu, K. D. (2010). Mutant alleles of FAD2-1A and FAD2-1B combine to produce soybeans with the high oleic acid seed oil trait. *BMC Plant Biol.* 10:195. doi: 10.1186/1471-2229-10-195

Pham, A.-T., Lee, J.-D., Shannon, J. G., and Bilyeu, K. D. (2011). A novel FAD2-1 A allele in a soybean plant introduction offers an alternate means to produce soybean seed oil with 85% oleic acid content. *Theor. Appl. Genet.* 123, 793–802. doi: 10.1007/s00122-011-1627-3

Pham, A.-T., Shannon, J. G., and Bilyeu, K. D. (2012). Combinations of mutant FAD2 and FAD3 genes to produce high oleic acid and low linolenic acid soybean oil. *Theor. Appl. Genet.* 125, 503–515. doi: 10.1007/s00122-012-1849-z

Platt, A., Vilhjálmsson, B. J., and Nordborg, M. (2010). Conditions under which genome-wide association studies will be positively misleading. *Genetics* 186, 1045–1052. doi: 10.1534/genetics.110.121665

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* 5, 92–102. doi: 10.3835/plantgenome2012.05.0005

Putri, S. P., Yamamoto, S., Tsugawa, H., and Fukusaki, E. (2013). Current metabolomics: technological advances. *J. Biosci. Bioeng.* 116, 9–16. doi: 10.1016/j.jbiosc.2013.01.004

Qi, Z.-M., Wu, Q., Han, X., Sun, Y.-N., Du, X.-Y., Liu, C.-Y., et al. (2011a). Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. *Euphytica* 179, 499–514. doi: 10.1007/s10681-011-0386-1

Qi, Z.-M., Xue, H., Sun, Y.-N., Qiong, W., Shan, D.-P., Du, X.-Y., et al. (2011b). An integrated quantitative trait locus map of oil content in soybean, Glycine max (L.) Merr., generated using a meta-analysis method for mining genes. *Agric. Sci. China* 10, 1681–1692. doi: 10.1016/S1671-2927(11)60166-1

Qin, J., Gu, F., Liu, D., Yin, C., Zhao, S., Chen, H., et al. (2013). Proteomic analysis of elite soybean Jidou17 and its parents using iTRAQ-based quantitative approaches. *Proteome Sci.* 11, 12. doi: 10.1186/1477-5956-11-12

Qiu, J., Wang, Y., Wu, S., Wang, Y.-Y., Ye, C.-Y., Bai, X., et al. (2014). Genome Re-Sequencing of Semi-Wild Soybean Reveals a Complex Soja Population Structure and Deep Introgression. *PLoS ONE* 9:e108479. doi: 10.1371/journal.pone.0108479

Rao, J., Cheng, F., Hu, C., Quan, S., Lin, H., Wang, J., et al. (2014). Metabolic map of mature maize kernels. *Metabolomics* 10, 775–787. doi: 10.1007/s11306-014-0624-3

Rudner, L., Glass, G. V., Evartt, D. L., and Emery, P. J. (2002). *A User's Guide to the Meta-analysis of Research Studies*. College Park: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland. 37.

Saito, K., and Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annu. Rev. Plant Biol.* 61, 463–489. doi: 10.1146/annurev.arplant.043008.092035

Sakata, K., Ohyanagi, H., Nobori, H., Nakamura, T., Hashiguchi, A., Nanjo, Y., et al. (2009). Soybean proteome database: a data resource for plant differential omics. *J. Proteome Res.* 8, 3539–3548. doi: 10.1021/pr900229k

Salt, D. E., Baxter, I., and Lahner, B. (2008). Ionomics and the study of the plant ionome. *Annu. Rev. Plant Biol.* 59, 709–733. doi: 10.1146/annurev.arplant.59.032607.092942

Sander, J. D., Dahlborg, E. J., Goodwin, M. J., Cade, L., Zhang, F., Cifuentes, D., et al. (2011). Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nat. Methods* 8, 67–69. doi: 10.1038/nmeth.1542

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

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

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

Severin, A. J., Woody, J. L., Bolon, Y.-T., Joseph, B., Diers, B. W., Farmer, A. D., et al. (2010). RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol* 10:160. doi: 10.1186/1471-2229-10-160

Sha, A.-H., Li, C., Yan, X.-H., Shan, Z.-H., Zhou, X.-A., Jiang, M.-L., et al. (2012a). Large-scale sequencing of normalized full-length cDNA library of soybean seed at different developmental stages and analysis of the gene expression profiles based on ESTs. *Mol Biol Rep* 39, 2867–2874. doi: 10.1007/s11033-011-1046-1

Sha, Z., Oka, N., Watanabe, T., Tampubolon, B. D., Okazaki, K., Osaki, M., et al. (2012b). Ionome of soybean seed affected by previous cropping with mycorrhizal plant and manure application. *J. Agric. Food Chem.* 60, 9543–9552. doi: 10.1021/jf3024744

Shamimuzzaman, M., and Vodkin, L. (2012). Identification of soybean seed developmental stage-specific and tissue-specific miRNA targets by degradome sequencing. *BMC Genomics* 13:310. doi: 10.1186/1471-2164-13-310

Shu, Y., Yu, D., Wang, D., Bai, X., Zhu, Y., and Guo, C. (2012). Genomic selection of seed weight based on low-density SCAR markers in soybean. *Genet. Mol. Res* 12, 2178–2188. doi: 10.4238/2013.July.3.2

Singh, N., Choudhury, D. R., Singh, A. K., Kumar, S., Srinivasan, K., Tyagi, R., et al. (2013a). Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PLoS ONE* 8:e84136. doi: 10.1371/journal.pone.0084136

Singh, U. M., Sareen, P., Sengar, R. S., and Kumar, A. (2013b). Plant ionomics: a newer approach to study mineral transport and its regulation. *Acta Physiol. Plant.* 35, 2641–2653. doi: 10.1007/s11738-013-1316-8

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., et al. (2013). An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS ONE* 8:e54603. doi: 10.1371/journal.pone.0054603

Sonah, H., O'donoughue, L., Cober, E., Rajcan, I., and Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS−GWAS approach and validation by QTL mapping in soya bean. *Plant Biotech. J.* 3, 10. doi: 10.1111/pbi.12249

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE* 8:e54985. doi: 10.1371/journal.pone.0054985

Song, Q., Jia, G., Zhu, Y., Grant, D., Nelson, R. T., Hwang, E.-Y., et al. (2010). Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1. 0) in soybean. *Crop Sci.* 50, 1950–1960. doi: 10.2135/cropsci2009.10.0607

Suda, I., Hajika, M., Nishiba, Y., Furuta, S., and Igita, K. (1995). Simple and rapid method for the selective detection of individual lipoxygenase isoenzymes in soybean seeds. *J. Agric. Food Chem.* 43, 742–747. doi: 10.1021/jf00051a034

Sugiyama, M., Ae, N., and Hajika, M. (2011). Developing of a simple method for screening soybean seedling cadmium accumulation to select soybean genotypes with low seed cadmium. *Plant Soil* 341, 413–422. doi: 10.1007/s11104-010-0654-1

Sun, Y.-N., Pan, J.-B., Shi, X.-L., Du, X.-Y., Wu, Q., Qi, Z.-M., et al. (2012). Multi-environment mapping and meta-analysis of 100-seed weight in soybean. *Mol. Biol. Rep.* 39, 9435–9443. doi: 10.1007/s11033-012-1808-4

Tavakolan, M., Alkharouf, N. W., Khan, F. H., and Natarajan, S. (2013). SoyProDB: a database for the identification of soybean seed proteins. *Bioinformation* 9:165. doi: 10.6026/97320630009165

Toubiana, D., Batushansky, A., Tzfadia, O., Scossa, F., Khan, A., Barak, S., et al. (2015). Combined correlation−based network and mQTL analyses efficiently identified loci for branched−chain amino acid, serine to threonine, and proline metabolism in tomato seeds. *Plant J.* 81, 121–133. doi: 10.1111/tpj.12717

Varshney, R. K., Chen, W., Li, Y., Bharti, A. K., Saxena, R. K., Schlueter, J. A., et al. (2012a). Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* 30, 83–89. doi: 10.1038/nbt.2022

Varshney, R. K., Kudapa, H., Pazhamala, L., Chitikineni, A., Thudi, M., Bohra, A., et al. (2015). Translational genomics in agriculture: some examples in grain legumes. *Crit. Rev. Plant Sci.* 34, 169–194. doi: 10.1080/07352689.2014.897909

Varshney, R. K., Ribaut, J.-M., Buckler, E. S., Tuberosa, R., Rafalski, J. A., and Langridge, P. (2012b). Can genomics boost productivity of orphan crops? *Nat. Biotechnol.* 30, 1172–1176. doi: 10.1038/nbt.2440

Varshney, R. K., Roorkiwal, M., and Nguyen, T. (2013a). Legume genomics: from genomic resources to molecular breeding. *Plant Genome* 6, 1–7. doi: 10.3835/plantgenome2013.12.0002in

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., et al. (2013b). Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat. Biotechnol.* 31, 240–246. doi: 10.1038/nbt.2491

Vaughn, J. N., Nelson, R. L., Song, Q., Cregan, P. B., and Li, Z. (2014). The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3 4, 2283–2294. doi: 10.1534/g3.114.013433

Voytas, D. F., and Gao, C. (2014). Precision genome engineering and agriculture: opportunities and regulatory challenges. *PLoS Biol.* 12:e1001877. doi: 10.1371/journal.pbio.1001877

Wagner, N., Mroczka, A., Roberts, P. D., Schreckengost, W., and Voelker, T. (2011). RNAi trigger fragment truncation attenuates soybean FAD2−1 transcript suppression and yields intermediate oil phenotypes. *Plant Biotech. J.* 9, 723–728. doi: 10.1111/j.1467-7652.2010.00573.x

Wang, Y., Han, Y., Teng, W., Zhao, X., Li, Y., Wu, L., et al. (2014a). Expression quantitative trait loci infer the regulation of isoflavone accumulation in soybean (Glycine max L. Merr.) seed. *BMC Genomics* 15:680. doi: 10.1186/1471-2164-15-680

Wang, Y. Q., Chen, P. Y., and Zhang, B. (2014b). Quantitative trait loci analysis of soluble sugar contents in soybean. *Plant Breeding* 133, 493–498. doi: 10.1111/pbr.12178

Warrington, C. V., Abdel-Haleem, H., Hyten, D. L., Cregan, P. B., Orf, J. H., Killam, A., et al. (2015). QTL for seed protein and amino acids in the Benning× Danbaekkong soybean population. *Theor. Appl. Genet.* 128, 839–850. doi: 10.1007/s00122-015-2474-4

Wei, W.-H., Chen, B., Yan, X.-H., Wang, L.-J., Zhang, H.-F., Cheng, J.-P., et al. (2008). Identification of differentially expressed genes in soybean seeds differing in oil content. *Plant Sci.* 175, 663–673. doi: 10.1016/j.plantsci.2008.06.018

Wilson, R. F. (2004). "Seed composition," in *Soybeans: Improvement, Production, and Uses*, eds H. R. Boerma and J. E. Specht (Madison, WI: ASA, CSSA, and SSSA), 621–677.

Wirta, V. (2006). *Mining the Transcriptome-Methods and Applications*. Estocolmo: Royal Institute of Technology, School of Biotechnology.

Würschum, T., and Kraft, T. (2014). Cross-validation in association mapping and its relevance for the estimation of QTL parameters of complex traits. *Heredity* 112, 463–468. doi: 10.1038/hdy.2013.126

Xu, Y., and Crouch, J. H. (2008). Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci.* 48, 391–407. doi: 10.2135/cropsci2007.04.0191

Xu, Y., Guo, M., Liu, X., Wang, C., and Liu, Y. (2014). SoyFN: a knowledge database of soybean functional networks. *Database* 2014:bau019. doi: 10.1093/database/bau019

Yan, P.-M., Wang, W.-Y., Rui, Y.-K., Zhang, F.-S., and Jin, Y.-H. (2007). Application of ICP-MS/ICP-AES to the detection of wholesome elements and heavy metals in soybean from Northeastern China. *Spectrosc. Spect. Anal.* 27, 1629.

Yin, G., Xu, H., Liu, J., Gao, C., Sun, J., Yan, Y., et al. (2014). Screening and identification of soybean seed-specific genes by using integrated bioinformatics of digital differential display, microarray, and RNA-seq data. *Gene* 546, 177–186. doi: 10.1016/j.gene.2014.06.021

Yu, J., Zhang, Z., Wei, J., Ling, Y., Xu, W., and Su, Z. (2014). SFGD: a comprehensive platform for mining functional information from soybean transcriptome data and its use in identifying acyl-lipid metabolism pathways. *BMC Genomics* 15:271. doi: 10.1186/1471-2164-15-271

Yu, O., Shi, J., Hession, A. O., Maxwell, C. A., McGonigle, B., and Odell, J. T. (2003). Metabolic engineering to increase isoflavone biosynthesis in soybean seed. *Phytochemistry* 63, 753–763. doi: 10.1016/S0031-9422(03)00345-5

Zhaoming, Q., Yanan, S., Lijun, C., Qiang, G., Chunyan, L., Guohua, H., et al. (2009). Meta-analysis of 100-seed weight QTLs in soybean. *Sci. Agric. Sin.* 42, 3795–3803.

Zhao-Ming, Q., Ya-Nan, S., Qiong, W., Chun-Yan, L., Guo-Hua, H., and Qing-Shan, C. (2011). A meta-analysis of seed protein concentration QTL in soybean. *Can. J. Plant Sci.* 91, 221–230. doi: 10.4141/cjps09193

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33, 408–414. doi: 10.1038/nbt.3096

Zhu, D., Li, Y., Wang, D., Wu, Q., Zhang, D., and Wang, C. (2012). The identification of single soybean seed variety by laser light backscattering imaging. *Sensor Lett.* 10, 399–404. doi: 10.1166/sl.2012.1836

Ziegler, G., Terauchi, A., Becker, A., Armstrong, P., Hudson, K., and Baxter, I. (2013). Ionomic screening of field-grown soybean identifies mutants with altered seed elemental composition. *Plant Genome* 6, 1–9. doi: 10.3835/plantgenome2012.07.0012

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.