BMC Genomics

Transcriptome profiling reveals the spatial-temporal dynamics of gene expression essential for soybean seed development

Hengyou Zhang[1,2†], Zhenbin Hu[3†], Yuming Yang[1], Xiaoqian Liu[1], Haiyan Lv[1], Bao-Hua Song[4], Yong-qiang Charles An[5], Zhimin Li[1*] and Dan Zhang[1*]

Abstract

Background: Seeds are the economic basis of oilseed crops, especially soybeans, the most widely cultivated oilseed crop worldwide. Seed development is accompanied by a multitude of diverse cellular processes, and revealing the underlying regulatory activities is critical for seed improvement.

Results: In this study, we profiled the transcriptomes of developing seeds at 20, 25, 30, and 40 days after flowering (DAF), as these stages represent critical time points of seed development from early to full development. We identified a set of highly abundant genes and highlighted the importance of these genes in supporting nutrient accumulation and transcriptional regulation for seed development. We identified 8925 differentially expressed genes (DEGs) that exhibited temporal expression patterns over the course and expression specificities in distinct tissues, including seeds and nonseed tissues (roots, stems, and leaves). Genes specific to nonseed tissues might have tissue-associated roles, with relatively low transcript abundance in developing seeds, suggesting their spatially supportive roles in seed development. Coexpression network analysis identified several underexplored genes in soybeans that bridge tissue-specific gene modules.

Conclusions: Our study provides a global view of gene activities and biological processes critical for seed formation in soybeans and prioritizes a set of genes for further study. The results of this study help to elucidate the mechanism controlling seed development and storage reserves.

Keywords: Soybean, *Glycine max*, Seed development, Transcriptome, Spatial/temporal gene expression

Background

Soybean [*Glycine max* (L.) Merr.] is the most widely cultivated oilseed crop worldwide, and soybean seeds are a major source of highly valuable protein and edible oil for human and animal consumption [1]. Soybean production (bushels per acre) has nearly doubled since 1987 [2] and accounts for approximately 68 and 27% of world plant meal and oil production, respectively (www.fas. usda.gov/data/oilseeds), reflecting the key role of soybean seeds in providing food for the global population. Furthermore, it is estimated that a 70% increase in soybean production is needed to meet the increasing demand for plant-based protein in the next decades [3, 4]. It is important to understand the mechanism of seed development with the goal of continuously increasing soybean yield production and maintaining its sustainable role as a major source of protein and oil worldwide.

Soybean seeds are composed of functionally distinct tissues, including the embryo, embryonic cotyledon, and seed coat. These tissues are successively differentiated

* Correspondence: lzmzdb@126.com; zhangd@henau.edu.cn
†Hengyou Zhang and Zhenbin Hu contributed equally to this work.
[1]Collaborative Innovation Center of Henan Grain Crops, College of Agronomy, Henan Agricultural University, Zhengzhou 450002, China
Full list of author information is available at the end of the article

and grown during the maturation phase [5]. The filial embryo from double fertilization eventually grows into cotyledons that serve as storage organs for food reserves, mainly oil and protein. The seed coat is maternal tissue surrounding the embryo and cotyledon to provide protection and nutrient delivery in the form of photoassemblies to support the developing embryo [5, 6]. Ontogenesis and development of these tissues are molecularly programmed and correlate with parallel changes in nutrient levels in seeds. Seed storage reserves, such as stored proteins and lipids, are synthesized and accumulate during seed filling mainly from early to later maturation stages [7]. Overall, seed development is a complex developmental process accompanied by a multitude of molecular activities simultaneously occurring in coordinating tissue development and storage reserve metabolism.

Thus far, several genes involved in seed nutrient accumulation have been identified, such as bZIP123 involved in lipid accumulation and GA20OX associated with seed weight and oil accumulation [8–11], as well as genes associated with seed development [12]. Conservation of several regulatory processes of seed development between soybeans and the model plant, *Arabidopsis,* has been revealed [7, 13], thereby enabling homolog-based cloning and functional validation of related genes in soybeans [14, 15]. However, the functions of many soybean genes in seed development and nutrient accumulation have been modified in palaeopolyploid soybeans. For example, *ABSCISIC ACID INSENSITIVE 3b* (*ABI3b*) functions like LEAFY COTYLEDON 2 (*LEC2*) [16], and WRINKLED1 (WRI1) is associated with plant architecture instead of its homologous role in affecting oil production in *Arabidopsis* [14]. In contrast, each of these major regulators associated with seed development, such as LEC1, ABI3, ABA-RESPONSIVE ELEMENT BINDING PROTEIN3 (AREB3), and BASIC LEUCINE ZIPPER67 (bZIP67), putatively regulate thousands of targets involved in the control of diverse developmental processes through mosaic combinations of regulators [17]. These studies suggest that soybeans have evolved a complex regulatory network during seed development that is, in part, different from that in *Arabidopsis,* in which the current knowledge of plant seed development has been mostly obtained [18, 19]. Hence, the regulatory mechanism of seed development in soybeans is largely unclear and needs to be comprehensively explored.

Sequencing-based transcriptome profiling (RNA-seq) has been demonstrated to be an effective approach in dissecting the regulatory mechanisms of complex traits [20] and has been applied to investigate the changes in gene expression during seed development in legumes and oilseed plants [5]. For example, analysis of transcriptomes in developing soybean seeds has identified hub genes from a set of regulatory genes putatively participating in oil and protein accumulation [21, 22]. In the present study, we profiled the transcriptomes of progressively developing seeds at 20, 25, 30, and 40 days after flowering (DAF) to extend our knowledge of the molecular mechanism that controls seed development and nutrient production. We uncovered a variety of gene activities during the dynamic processes of seed development and identified multiple gene sets with spatial and temporal expression during the period. The results provide a global view of gene activity essential to the developmental process in soybeans, providing insight into the regulatory mechanism of seed formation.

Results and discussion

Dramatic morphology and reserve changes during seed development

Whole developing soybean seeds at four sequential time points (20, 25, 30, and 40 DAF) covering the major stages of seed development were sampled in triplicate (Fig. 1A). This period of seed development is coincident with seed growth stages from R5 (beginning seed) to R6 (full-size seed) of soybean development [23], which are morphologically comparable to early and mid-maturation stages as previously described [7]. This period also co-occurs with the rapid accumulation of lipids and storage proteins [20, 24]; thus, the seed size and seed weight increased significantly as it grew toward a fully developed seed (Fig. 1A). Two major storage reserves in soybean seeds, namely, protein and oil, were quantified during this period. As shown in Fig. 1B, the accumulation of stored oil was approximately 14.11% at D20 and continuously increased to 17.85% at D40 with the greatest increase of 1.85% occurring in the period from D20 to D25 relative to the later periods of D25-D30 by 1.11% and D30-D40 by 0.78%. In contrast, as the seeds matured, the stored protein content was 50.87% at D20, which decreased to 45.83% (5.04% decrease) at D25 and continuously declined to 42.26% at D30 and to 41.77% at D40. Thus, the oil content in the developing seeds increased more dramatically in the early period and less dramatically in the late period of seed filling, and vice versa for the protein decrease. The reduction in the relative percentage of protein content over the period is due to its negative correlation with the oil content in soybean seeds [25]. In contrast, the actual protein levels increased rapidly between 10 and 40 DAF [26]. Hence, this period of seed development presents rapid morphological changes (such as size and weight) and major metabolic changes (such as oil and protein), which should be governed by a variety of cellular processes/gene activities in seeds [17, 20].

Transcriptomic abundance shifts during seed development

An in-depth understanding of biological processes associated with seed development and reserve accumulation
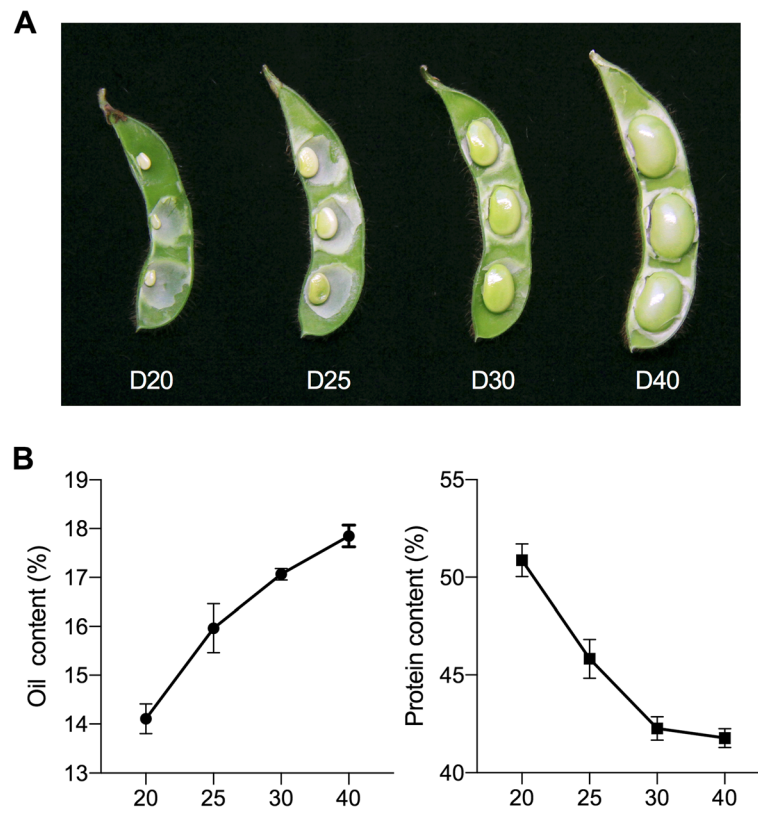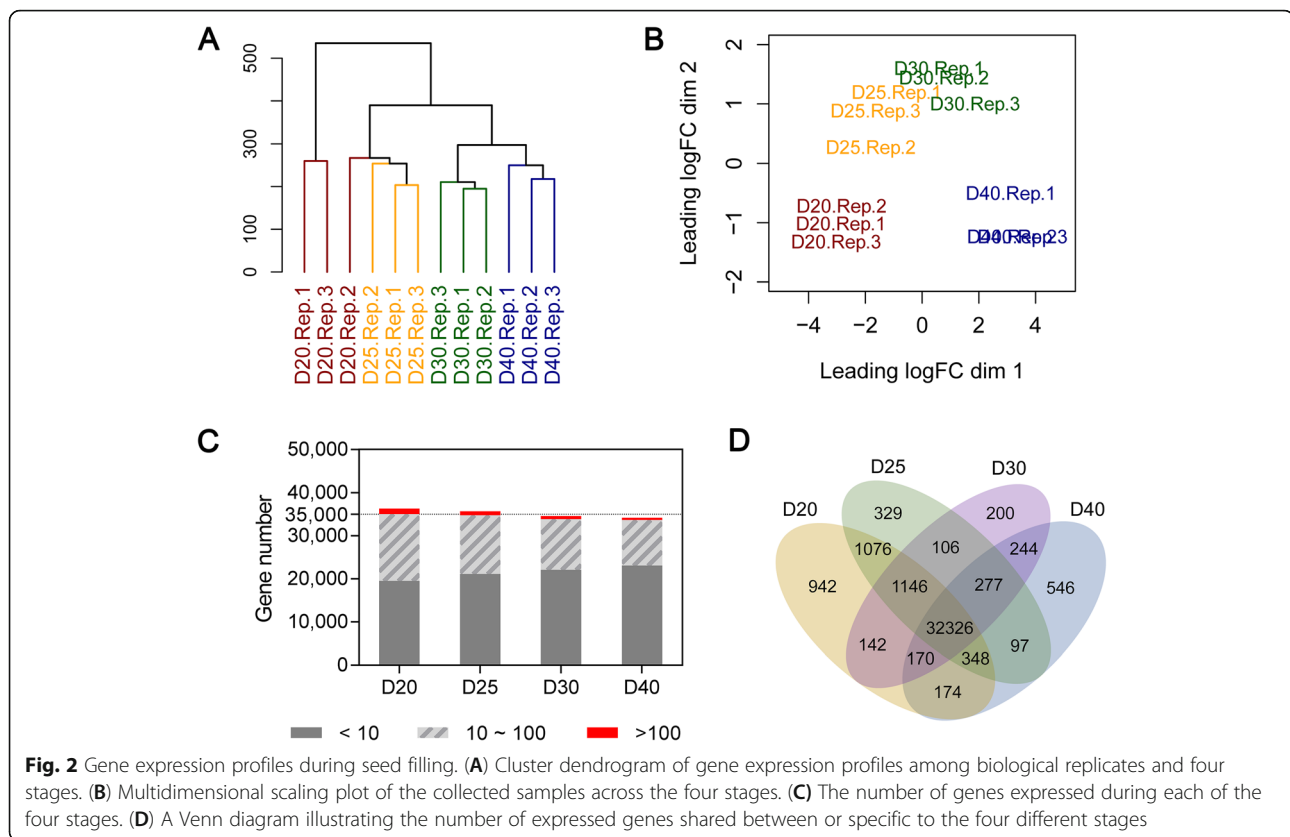
Fig. 1 Morphology of developing seeds at 20, 25, 30, and 40 DAF (**A**) and changes in oil and protein during the selected stages (**B**)

in soybean seeds is important for improving seed quality and yield potential. Given the multiplicity of genes controlling seed development coupled with seed filling, time-course transcriptomic profiling of four-stage seeds that covered the representative period of seed development was performed with the aim of obtaining a global view of the regulatory mechanism underlying this complex development process. In total, ~ 313.4 million sequencing reads from all sampled seeds were generated, with an average of 26.1 million reads per sample (Table S1). On average, over 97.43% (25.4 million) of reads were kept after trimming the low-quality reads and used for read alignment. Uniquely mapped reads were used to quantify gene expression abundance as normalized by counts per million (cpm). After removing reads/transcripts < 5, we determined that, on average, 35,000 (62.5%) out of 56,044 gene models were expressed in developing seeds. We observed a gradient reduction (approximately 700 genes, 2% of the total) in the number of expressed genes over the course of seed development. The reduction could be, in part, related to the shutdown of many transcription processes as the seeds grow toward dormancy [27]. A previous study has reported a programmed increase in methylation levels during soybean seed development [7]. Thus, it would be interesting to determine whether the increased methylation level is related to the reduction of the seed transcriptome.

Both clustering dendrogram analysis and multidimensional scale analysis indicated that gene expression levels among the replicated samples were generally highly related to those between time-point samples (Fig. 2A, B), indicating gene expression specificity to the time points. For the transcript abundance, approximately 61.07% of expressed genes were in the range of 0–10 cpm, 36.41% were in the range of 10–100 cpm, and only 2.53% of genes were highly expressed, at cpm ≥ 100 (Fig. 2C). As the seeds enlarged, the number of genes with an abundance ≥10 cpm generally decreased, and the number of genes with an abundance ≤10 cpm gradually increased. Among the expressed genes, 32,326 were expressed across all four stages of developing seeds. There were 942, 329, 200, and 546 genes specifically expressed at D20, D25, D30, and D40, respectively (Fig. 2D). These results suggest a temporal shift in transcript abundance for many highly expressed genes as seeds developed, which might reflect the foregoing morphological and/or metabolic changes during seed development.

A cluster dendrogram analysis showed that replicates from each condition were closely clustered as supported by a similar separation pattern in a multidimensional scaling (MDS) plot of log-FC values. We further

Fig. 2 Gene expression profiles during seed filling. (**A**) Cluster dendrogram of gene expression profiles among biological replicates and four stages. (**B**) Multidimensional scaling plot of the collected samples across the four stages. (**C**) The number of genes expressed during each of the four stages. (**D**) A Venn diagram illustrating the number of expressed genes shared between or specific to the four different stages

performed *q*RT-PCR on 16 randomly selected DEGs (Table S4) for validation, and a high correlation (r = 0.87) between the RNA-seq and qRT-PCR results indicated the robustness of our RNA-seq results.

Genes with high transcript abundance during seed development and filling

We next characterized the genes that were constitutively highly expressed across all four time points. In total, 443 genes with a mRNA abundance cpm ≥ 100 were identified [28] (Table S3). GO term enrichment analysis indicated that enriched biosynthetic processes (BPs) essential for nutrient accumulation were amino acid metabolism, carbohydrate metabolism, glycolysis, and hexose metabolism (Table S3). This gene set was also enriched for the molecular function (MF) term nutrient reservoir activity and cellular component (CC) terms related to lipid storage bodies, endoplasmic reticulum (ER), and ribosomes, sites where fatty acids are primarily synthesized and stored [29, 30]. When we relaxed the threshold to cpm ≥ 50, the number of newly defined highly expressed genes increased by threefold to 1386 genes (Table S3). Accordingly, the number of expressed genes in the majority of enriched terms was approximately doubled or tripled, such as hexose metabolic process (8 to 22), glycolysis (7 to 18), cellular nitrogen compound metabolic process (18 to 36), cellular amino acid

biosynthetic process (15 to 32), and ribosomal proteins (47 to 132). It should be noted that no change in the number of genes with terms related to monolayer-surrounded lipid storage bodies (7), nutrient reservoir activity (11), and the sulfur amino acid metabolic process (6 to 7) was observed, suggesting the importance of maintaining high transcript abundances of these genes for seed filling. Additional enriched GO terms for the 1386 genes identified were as follows: protein localization, glutamine family amino acid biosynthetic process, intracellular transport, gene silencing by RNA, and vesicle-related components (COP-II-, COP-I, coated vesicles, transport vesicles, ER to Golgi transport vesicles, and membrane-bound vesicles).

A detailed investigation of the 443 genes (cpm ≥ 100) revealed several with demonstrated roles in the accumulation of storage reserves, such as oil, protein, and carbohydrates. Not surprisingly, some genes participating in the metabolism of protein and oil were identified. However, several of the highly abundant genes detected during seed filling have rarely been reported (Table S3), such as the oil accumulation-related genes, *FAD2* (467.83–2000.69) catalyzing oleic acid accumulation [31], OLEOSIN proteins (373.03–6190.57 cpm) encoding the oil body surrounding lipid storage bodies [30, 32], and metabolic genes for the biosynthesis of leucine, thiamin (vitamin B1), and ribosomal proteins. Genes

Zhang *et al. BMC Genomics* (2021) 22:453

Page 5 of 13

encoding storage proteins, such as cupin family proteins, showed the highest transcript abundance, ranging from 34,000–100,000 cpm, in developing seeds followed by xylem bark cysteine peptidase 3 (*Glyma.08G116300* with 28,399.81 cpm), trypsin inhibitor 1 (222.79 ~ 18,559.51 cpm), and oleosin family proteins, which were highly abundant (373.03–6190.57 cpm). These proteins in soybeans appear to be redundant and serve as a surfactant for the oil body [33, 34]. A recent study on oleosin on chr20 [32] has indicated that it is likely regulated by LEC2 [15], by which oil production is enhanced through increasing oil body turnover. Investigation of the cis-regulatory sequences in the promoter region could be helpful in using this set of genes to achieve oil content improvement. In contrast, a set of highly abundant trypsin inhibitor genes (222.79–18,559.51 cpm) could be silenced by gene editing to inhibit their expression or mutate them to achieve better seed quality because trypsin inhibitors have an antinutritional effect [35].

More attention should be given to nutrient transporters that mediate nutrient flux into the developing embryo, such as the *SWEET* sugar transporters (274.98–943.19 cpm), proton pump interactor 1 and *SUCROSE SYNTHASEs* (121.5–1,186.21 cpm) controlling starch synthesis. Recent studies have highlighted the essential role of SWEET proteins in transporting sugar from maternal to embryonic tissues to support seed development and oil accumulation [12, 36, 37], indicating the agriculturally critical role of causing high oil accumulation in cultivated soybean seeds relative to the low oil content in wild soybeans [11]. High abundances of these transporters, together with related studies, suggest that the rate of sugar transport and sugar accumulation is important for nutrient reserve accumulation in seeds. Further investigation of downstream affected steps/components may help in identifying the major route of carbon flow for oil accumulation or carbon partitioning between oil and proteins in soybeans and, perhaps, other plant species, such as maize and rice [38, 39]. Several *LOX1* genes (541.27–16,226.05 cpm) had high transcript abundances in the present study, but homologous proteins have not been detected during seed filling in wheat and maize endosperms [40, 41], suggesting a lineage-specific role of LOX proteins in soybean seeds, possibly related to seed storage proteins and lipids, such as unsaturated fatty acids [42, 43], or special tastes and odors [44].

Transcription factors, such as ABA INSENSITIVE 3 (ABI3; *Glyma.08G357600*; *Glyma.18G176100*) and its interacting regulator, bZIP67, were identified in high abundance. The high abundance of both regulators may be due to the increasing need for their involvement in cell growth and the accumulation of storage reserves as seeds enlarge. Indeed, both regulators have common functions and are essential for initiating gene expression

changes for biological processes critical for seed development, such as gibberellic acid signaling, photosynthesis, and maturation-related processes [16, 17, 45, 46]. LEC2s (cpm = 0) were undetectable in this study, possibly because they are highly specific to certain seed compartments and barely detected in whole seeds [15]. In addition, well-documented genes homologous to LEC1, which is a central regulator of seed development [17], were not in high abundance but showed decreased expression over time. Other high-abundance regulators with demonstrated roles in *Arabidopsis* that are less explored in soybean seeds were also identified, such as PLASTID TRANSCRIPTIONALLY ACTIVE 9 (PTAC9), TANDEM CCCH ZINC FINGER PROTEIN 4 (TCZF4) and circadian rhythm regulators, including TIME FOR COFFEE (TIC), LATE ELONGATED HYPOCOTYL 1 (LHY1), and CONSTANS-LIKE 4 (COL4). Determination of the functions of the regulators and the cognate cis-regulatory elements may facilitate the identification of seed regulatory networks that have contributing roles in the regulation of genes temporally and spatially in developing seeds [17, 18, 47]. All highly expressed genes and the functional annotations are listed in Table S2.

Genes are temporally regulated during seed development

To comprehensively identify genes involved in developmental processes during seed filling, we conducted a pairwise comparison of transcriptomic profiles among the different developmental stages of seeds. We observed a gradient increase in the number of DEGs between the two neighboring time points as the seeds developed, with 523 in D25 vs. D20, 1486 in D30 vs. D25, and 2485 in D40 vs. D30 (Fig. 3). When comparisons were made across nonadjacent time points (over 10-day intervals), including 4079 in D30 vs. D20, 5617 in D40 vs. D25, and 7670 in D40 vs. D20, the numbers of DEGs increased. In total, the pairwise comparison identified 8925 nonredundant DEGs that are required to program soybean seed development from early to full seed development (Table S2).

Clustering analysis was performed for all 8925 DEGs based on expression and is shown in a heat map (Fig. 4). The analysis classified the DEGs into seven groups (A to G) exhibiting different gene expression patterns. The difference in the temporal expression patterns suggested stage-specific roles of these grouped DEGs throughout seed development. Groups A and G represented the two largest clusters (2182 and 4945 DEGs, respectively) while exhibiting contrasting expression patterns. The DEGs in cluster A displayed the predominant and highest expression levels at D40, and those in group G showed a continuous decrease in gene expression as seeds matured. Enrichment analysis for cluster A genes showed that D40-specific genes were enriched for cell communication (GO: 0007154) and transcription (GO:0006351), and the
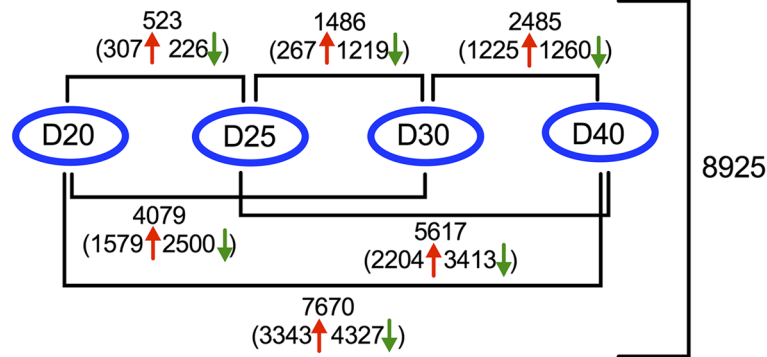
Fig. 3 Number of DEGs estimated by pairwise comparison of the four stages

enriched biological processes occurred in the cell periphery and plasma membrane (Fig. 4B). This stage represented the later stage of reserve production and storage as well as the active period of the shift to maturation in which multiple programming might be initiated, such as cessation of organ differentiation, cessation of cell division, and preparation for seed desiccation [48]. A representative gene with specifically high expression at D40 is ABI3 (Fig.

4C) [16], which is an essential regulator in abscisic acid (ABA) signaling primarily participating in seed maturation and dormancy in *Arabidopsis* [49]. The high abundance of AB13 during D20-D30 and dramatic increase toward D40 suggested its critical role during seed development, especially in the later stage of maturation-related processes. ABI3 has recently been demonstrated to be a key transcription factor modulating the transition from

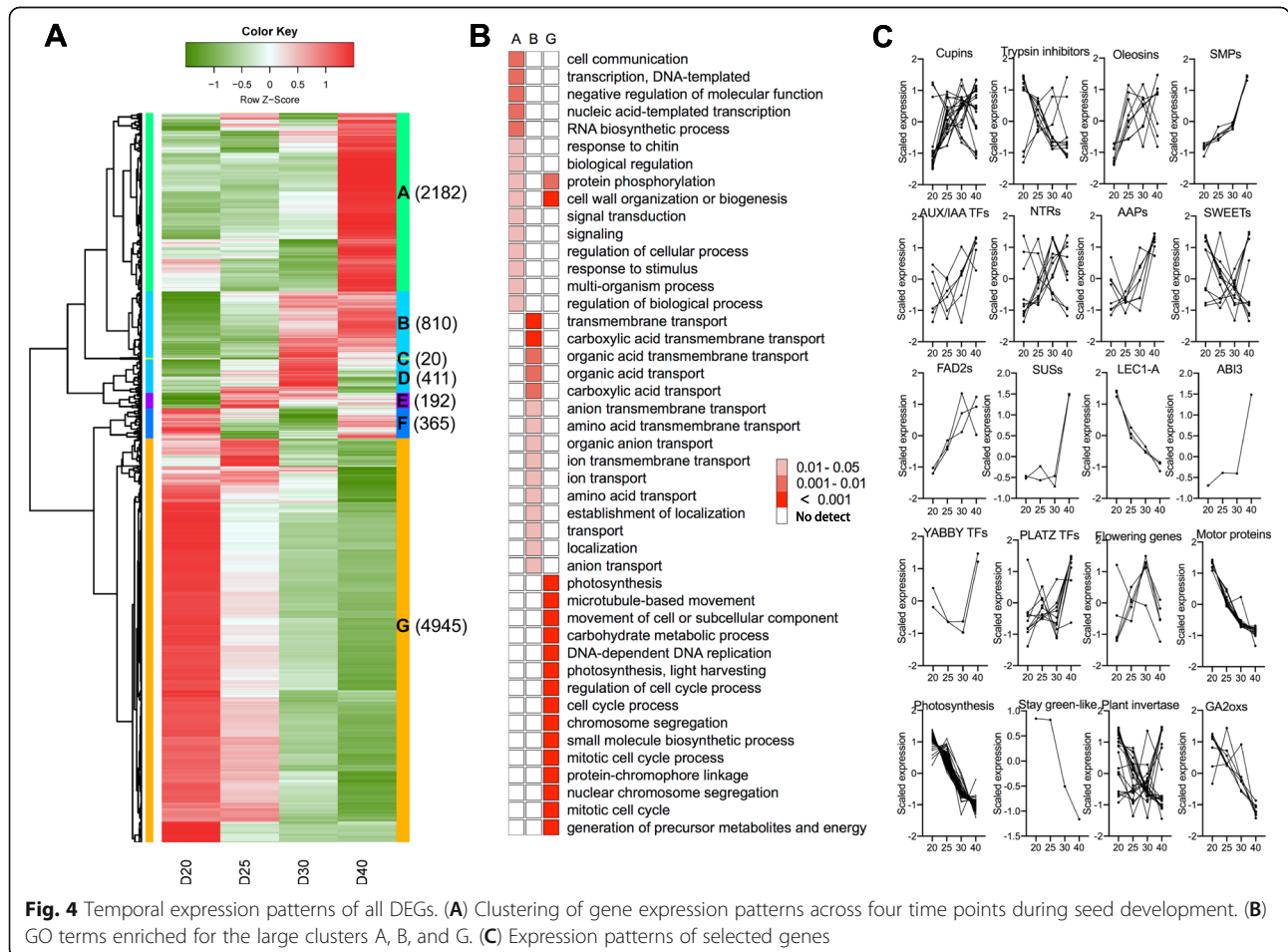**Fig. 4** Temporal expression patterns of all DEGs. (**A**) Clustering of gene expression patterns across four time points during seed development. (**B**) GO terms enriched for the large clusters A, B, and G. (**C**) Expression patterns of selected genes

morphogenesis to maturation by combinational inter-action with other seed transcription factors [17]. Further study of the coexpressed genes or binding sites [17, 21] would provide an opportunity to understand regulation in soybean seeds. Identification of the key seed development regulators (such as ABI3) indicated the robustness of our analysis and suggested that other seed development-associated genes identified in our study deserve further attention. For example, a similar expression pattern was also observed for two sucrose synthases (SUSs) involved in cellulose and starch biosynthesis in seeds [50].

Genes classified in cluster B (810) were highly expressed at D40 and had a similar expression pattern to that in cluster A, except for those genes that were upregulated beginning at D30. Enriched GO terms for cluster B were transmembrane transport (GO:0055085) and nutrient reservoir activity (GO:0045735), suggesting that biological processes associated with nutrient transport, such as sugar transporters, amino acid transporters, and nutrient accumulation, were active beginning at D30 and maintained at D40. The developmental period of D30-D40 is approximately coincident with the middle-maturation stage, which represents the major period of accumulation for storage, such as that of protein and lipids [7]. Expression pattern analysis showed that despite the varying expression patterns observed, genes involved in the aforementioned biological processes showed obvious upregulation during seed development and filling (Fig. 4C), such as cupin proteins, oleosins, amino acid permeases (APPs), FAD2, nitrate transporters, and seed maturation proteins (SMPs). Such stark upregulation in expression undoubtedly reflects the critical roles of the related biological processes in storage reserve accumulation. Transcription factors involved in seed development were also upregulated in cluster B, such as the AUS/IAA transcription factor, YABBY, and uncharacterized PLATZ. Interestingly, the genes related to flowering (TFLs, Tof12, and PHYA) showed an upregulation/downregulation shift during the course, suggesting expanded roles during seed development, but this idea has yet to be explored. This finding is partially supported by a recent study demonstrating that *Arabidopsis TFL1* not only functions as a shoot identity gene but also serves as a signaling molecule essential in determining seed size, and the *tfl1–20* loss-of-function mutant has been shown to produce larger developing seeds than wild type through regulation of endosperm cellularization [51]. The expression patterns for these DEGs during seed development are illustrated in Fig. 4C.
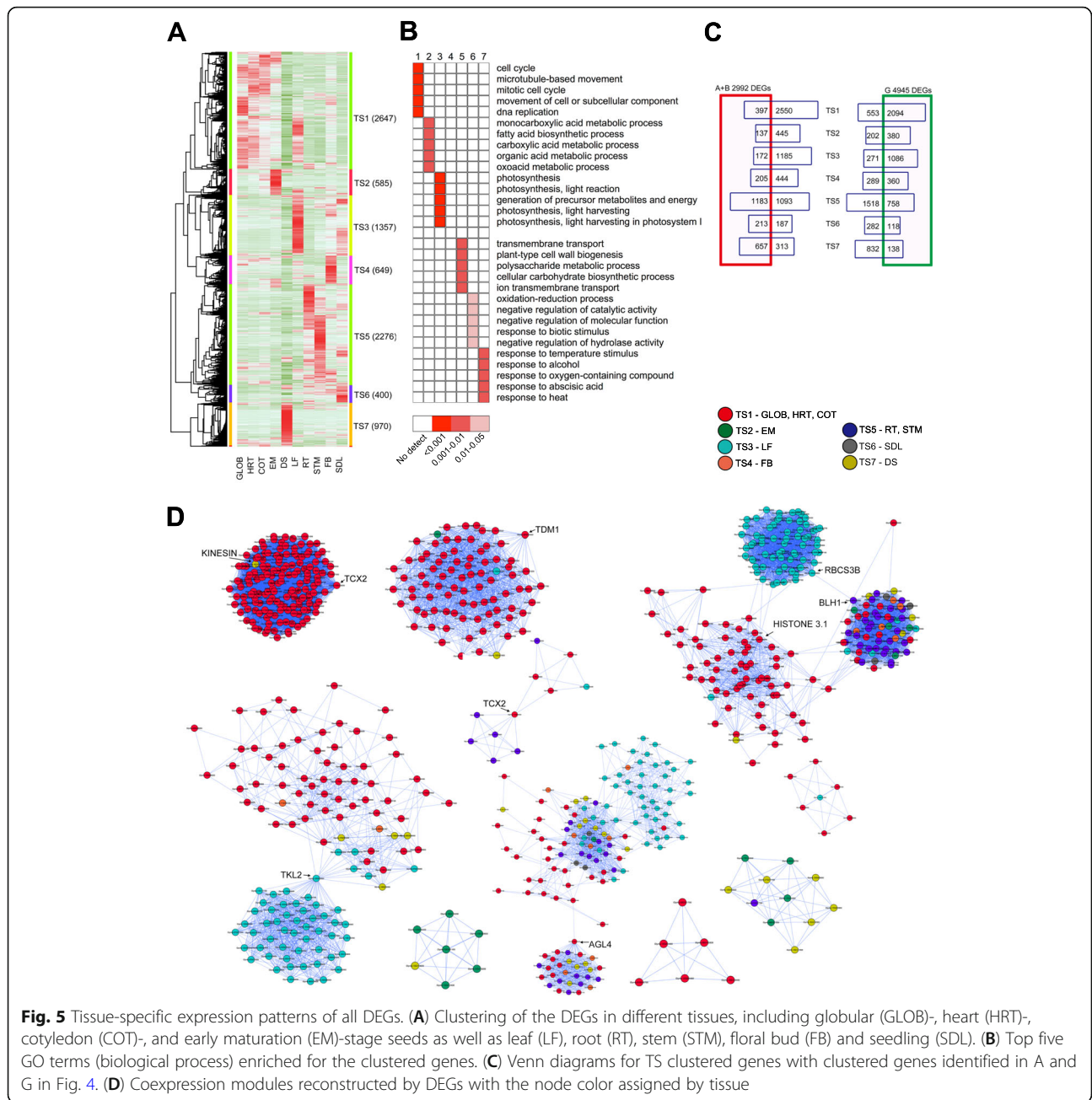
In contrast, the top GO terms enriched for the genes in cluster G are mainly related to photosynthesis (GO: 0015979) and microtubule-based processes (GO:0007017). Downregulation of these gene sets mainly reflect the reduction of the photosynthetic process as seeds developed, which as evidenced by downregulation of the bulk of genes encoding varying processes of photosynthesis, such as chlorophyll A-B binding family proteins, high-chlorophyll fluorescent proteins, and light-harvesting complex photosystem II (Table S1). Other family genes exhibiting downregulated expression patterns were also identified, such as GA2OXs, motor proteins, LACS9s, stay green-like genes, and LEC1-like genes (Table S1). The decrease in expression of these genes with diverse roles associated with phytohormones (such as GA) [52], chlorophyll breakdown during senescence [53], and seed developmental processes [47] might correspond to multiple processes of gradual cessation of cell division or commencement of maturation as seeds develop and mature. Clusters C-F comprised relatively fewer DEGs that displayed relatively high expression levels at two of the four selected time points, and no GO term was significantly enriched.

DEGs exhibit spatial expression patterns in seed compartments and nonseed tissues

We next examined the spatial expression profiles of all DEGs in different developmental tissues using the Harada-Goldberg soybean RNA-seq dataset (http://seedgenenetwork.net). This dataset contains 10 tissues of Williams 82, which is the same accession used in our study, including 6 reproductive tissues [floral buds (FB) and five stages of developing seeds, including whole-mount globular (GLOB), heart (HRT), cotyledon (COT), early-maturation (EM), and dry seed (DS)] and four vegetative tissues (leaf (LF), root (RT), stem (STM), and seedling (SDL)). Although this study was performed independently, it was complementary to ours, and integration of our results with spatial analysis using multiple tissues may provide additional insight into the mechanism.

The DEGs exhibited clear tissue-specific (TS) patterns, as indicated by the higher expression in certain tissues compared to others. This analysis grouped the DEGs into seven large (TS1-TS7) groups with each containing over 400 genes, and the remaining three groups (not shown) comprised less than 30 genes (Table S2). The seven large groups, which were further investigated, are shown in Fig. 5A. Enrichment analysis for the seven TS groups revealed that no common top five enriched GO terms, indicating possible certain roles in the highly expressed tissues. Interestingly, genes that were highly expressed in developing seeds were divided into three groups: TS1 grouped genes highly expressed in three early-stage developing seeds, such as GLOB, HRT, and COT; TS2 grouped genes highly expressed in the EM-stage seeds; and TS3 grouped genes highly expressed in the later-stage seeds and DS-stage seeds. Approximately 2647 (29.7%) of the DEGs were clustered into TS1, consisting of genes primarily expressed in GLOB-, HRT-,

Fig. 5 Tissue-specific expression patterns of all DEGs. (**A**) Clustering of the DEGs in different tissues, including globular (GLOB)-, heart (HRT)-, cotyledon (COT)-, and early maturation (EM)-stage seeds as well as leaf (LF), root (RT), stem (STM), floral bud (FB) and seedling (SDL). (**B**) Top five GO terms (biological process) enriched for the clustered genes. (**C**) Venn diagrams for TS clustered genes with clustered genes identified in A and G in Fig. 4. (**D**) Coexpression modules reconstructed by DEGs with the node color assigned by tissue

and COT-stage seeds, and these genes showed continuous expression with either upregulation or downregulation. The majority (79.1%, 2094 of 2647) of TS1 was classified into the G group (downregulation group) (Fig. 5A, C), exhibiting downregulation as seeds grew. Genes (585, 6.6% of total) in TS2 were predominantly expressed in the EM-stage seeds. In total, 970 (10.9%) DEGs were exclusively expressed in DS-state seeds in TS7. This spatial difference in expression among the different stages of seeds suggested the possible different roles of the two gene sets in seeds. Indeed, TS1 genes are mainly involved in processes related to the cell cycle

and DNA replication, which might account for early seed compartment morphogenesis and growth (Fig. 5B), and TS2 (EM) was more specialized in metabolic processes, such as fatty acid biosynthesis and metabolism of carboxylic acids and organic acids. In contrast, genes highly expressed in DS-stage seeds participate in the response to environmental stimuli, such as temperature, heat, alcohol, and plant hormones, which regulate plant growth, such as abscisic acid, and over half (67.7%) of this gene set was upregulated (Fig. 5C). For example, LEA14 (LATE EMBRYOGENESIS ABUNDANT 14) from TS7 is involved in wounding and light stress as

well as possible protection against desiccation [54, 55]. Thus, the DEGs highly expressed in DS-stage seeds might be partially related to preparing transcripts for dormancy and, possibly, seed germination under favorable conditions.

Recent studies have indicated that the expression of many of these seed-expressed genes is highly spatially specific in seeds, and this specificity represents functional relevance. For example, seed coat-specific *SWEET39* is involved in sugar delivery from maternal plants to developing embryos [56] and endosperm-specific *TFL1* in endosperm cellularization, which affects seed size [51]. In addition, endosperm-exclusive SWEET15 participates in seed development by regulating sugar delivery into developing embryos [12]. These studies suggest that many of the genes highly expressed in seed compartments, as identified here (Table S2), can be prioritized for detailed investigations to gain a more comprehensive understanding, including YABBA, GA3OX1, NF-YB6, NF-YA9, FAD8, KAS I/III, SWEETs, OLEOSINs, AATs (amino acid transporters), and uncharacterized LEC1-LIKE (Glyma.07G268100, Glyma.17G005600), which serves as a key regulator of fatty acid biosynthesis in *Arabidopsis* [57].

In total, 1357 (15.2%) DEGs clustered into TS3 and were mainly expressed in leaves; they are involved in many aspects of photosynthesis with over 80% of the DEGs downregulated as seeds matured (Fig. 5C). These results indicated that this gene set maintain photosynthetic roles and is also associated with seed developmental processes. It is likely that the expression of many of these photosynthesis-related genes in seeds might be influenced by mosaic interactions of different transcription factors, such as LEC1, ABI3, bZIP67, and AREB3 [17, 47]. Genes that were highly expressed in the RT and STM shared many DEGs in common and clustered in TS5, which agreed with the tissue functionality of both being involved in nutrient translocation [58]. These genes are mainly involved in carbohydrate and sugar metabolism as well as nutrient and ion transport to meet the large demand for nutrient uptake and transport (Fig. 5B). TS6 was the smallest group (400), with genes highly expressed in SDLs and participating in processes such as oxidation-reduction. These results indicated that genes that have relatively high expression in nonseed tissues, such as LF, FB, and RT, might also maintain tissue-represented functions in developing seeds (Fig. 5B); however, their expression is relatively low in developing seeds, suggesting their indispensable roles in seed development. For example, AAP2 from TS6 (stem and leaf) affects the levels of nitrogen and carbon assimilation as well as seed yield by altering xylem-phloem transfer in amino acid transfer [59], and FRD3 (MATE efflux family protein) from TS6 functions in the root xylem for

efficient iron uptake from the xylem into leaf cells [60]. Therefore, the identification of these genes with nonseed tissue specificities in developing seeds was reflective of the necessity of their basic functionalities in maintaining the needed processes to support seed development. Further investigation of these genes may provide deep insight into their supportive roles during seed development.

As indicated in this and other studies [17, 61], multiple layers of gene activities are involved in seed development, and the expression of the genes is also subjected to environmental effects. Despite promising genes identified from integrative analysis, it is ideal to compare the transcriptomes in different seed compartments and nonseed tissues at each of the time points in the same environment, which would offer a more comprehensive understanding of gene activities in developing seeds over the period.

Coexpression network-inferred seed regulatory network

We next extracted the coexpression connections of all DEGs from a global SoyNet database (http://www.inetbio.org/soynet/) constructed with transcriptomes from different seed compartments at different developmental stages, along with other tissues [47, 62]. Only a small fraction of the DEGs could be reconstructed into modules. In total, 134 modules were identified with the numbers of nodes (genes) ranging from 2 to 151. After color assignment based on the aforementioned TS groups, 11 modules consisting of relatively tissue-specific patterns were identified (Fig. 5D). It is interesting that many genes from TS1 (GLOB, HRT, and COT) and TS3 (LF) were clustered together as individual TS-specific modules compared to the other TS groups. Two modules specific for TS1 were identified. TCX2 (TESMIN-LIKE CXC2) controls stem cell division by regulating stem cell type-specific networks [63]. TCX2 slightly deviated from the main module, suggesting its role in orchestrating the connected genes of this module for coordinated division of different cell types for seed development, which is in agreement with the function of GLOB/HRT/COT, where ontogenesis of different seed compartments occurs during this stage (http://seedgenenetwork.net/). Similarly, TDM1 plays an important role in meiosis termination [64], and it was strongly associated with genes in the TS1-dominated modules. We determined that uncharacterized TKL2 (TRANSKETOLASE 2) might be a key connector bridging two modules comprising LF and GLOB/HRT/COT genes. It is likely that the link between RBCS3B (RUBISCO SMALL SUBUNIT 3B) and BLH1 (EMBRYO SAC DEVELOPMENT ARREST 29) connected the two individual modules. These results suggest a functional relevance of these modules in seed development, and

Zhang *et al. BMC Genomics*   (2021) 22:453

Page 10 of 13

these genes that are underexplored in soybeans deserve more attention.

Conclusions

In conclusion, we profiled the transcriptomes of developing seeds at four representative stages and provided a global transcriptomic view of the developmental process of storage reserve accumulation and seed development. The results highlight a set of highly abundant genes (storage genes, transporters, and transcription factors) that are important for soybean development and reveal that seed formation involves a variety of genes with tissue-associated roles to facilitate energy source transport, cell division, enlargement, and synthesis of highly abundant storage proteins and oil. Our study provides a global view of how a large number of genes are coordinated during seed development and may provide a framework for further investigations. Certainly, the role that these region-specific and period-specific genes and processes play in seed formation remains to be determined.

Methods

Plant growth and sample collection

The Williams 82 soybean variety was used in this study. Seeds were planted at the Maozhuang Experimental Station of Henan Agricultural University in 2018 in a three-row plot with row lengths of 200 cm and row spacing of 50 cm. A randomized block design was used for the field planting design. A flower on each node was tagged to indicate the flowering time. Developing seeds at 20 DAF (at the R5 stage) and at 25, 30, and 40 DAF (at the R6 stage) were collected [23], flash-frozen with liquid nitrogen, and stored in a − 80 °C freezer for RNA extraction. For RNA extraction, two to three developing seeds per plant were sampled and pooled as one biological replicate, and three biological replicates were used per time point. For protein and oil quantification, developing seeds from five plants (3–5 seeds per plant) were pooled as one biological replication; three biological replications were used. The crude extract and protein were quantified using the micro-Kjeldahl method and Soxhlet extractor, respectively, as previously described [65, 66].

RNA isolation, library construction, and transcriptome sequencing

Total RNA was isolated using TRIzol reagent (Invitrogen, CA, USA) following the manufacturer's protocol. Prior to library construction, total RNA was treated with RNase-free DNase I (New England Biolabs, Ipswich, MA, USA) to remove any contamination of genomic DNA. Library construction for RNA-seq was performed following a previously described protocol [67]. The quantified library was sequenced on the Illumina HiSeq 2500 sequencing platform (Illumina Inc., San Diego, CA, USA) at Biomarker Technologies (Beijing, China) and produced 200-bp paired-end reads.

Data processing and differential expression

Raw sequencing data quality was evaluated with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) followed by removal of low-quality reads (quality value < 20) using Trimmomatic v0.33 [68]. The trimmed reads were aligned to the Williams 82 soybean reference genome (Wm82.a2.v1) [69] with TopHat (v2.1.1) using the minimum intron size (−i) parameter 30 and the maximum intron size (−I) of 15,000. The number of reads mapped to annotated genes was counted using featureCounts [70]. Analysis to determine differentially expressed genes (DEGs) was performed using EdgeR [71], and genes with an *fdr* ≤ 0.01 and a fold change ≥1.5 were considered significantly differentially expressed between the two conditions. Enrichment analyses for GO terms and KEGG pathways were performed in Gene Ontology (http://geneontology.org/docs/go-enrichment-analysis/) and KOBAS 3.0 (http://kobas.cbi.pku.edu.cn/), respectively, as previously described [72].

Quantitative real-time PCR

Gene expression was examined in the same tissues used for RNA-seq by qRT-PCR as previously described [32]. Briefly, total RNA was isolated from the tissues using the RNA Simple Total RNA Kit (TaKaRa, Japan), and 1 μg of RNA was treated with RNase-free DNase I (TaKaRa, Japan) before cDNA synthesis. First-strand cDNA was synthesized using the SuperScript III First-Strand Synthesis System (Invitrogen, USA) following the manufacturer's instructions. Gene expression was determined using the Bio-Rad CFX96 Touch Real-Time PCR System (Bio-Rad, CA, USA). The PCR mixture contained 5 μL of cDNA, 0.5 μL of 10 μmol L^{-1} gene-specific primers, and 10 μL of Real-Time PCR SYBR Mix. The PCR conditions were as follows: 94 °C for 3 min and 40 cycles at 94 °C for 15 s and 60 °C for 15 s. The soybean *TUBULIN* gene (GenBank: AY907703.1) was amplified as an internal reference, and a negative control reaction was performed using water instead of cDNA. Three biological replicates per sample were used, and each reaction was performed in triplicate.

Coexpression network construction

DEGs were extracted from the coexpression network in SoyNet that was preconstructed using transcriptomes from diverse tissues and seed compartments [62]. The coexpressed genes were reconstructed into modules using Molecular Complex Detection (MCODE) [73] using the default parameters. The module network was visualized using Cytoscape [74].

Zhang *et al. BMC Genomics*   (2021) 22:453

Page 11 of 13

Abbreviations
BP: Biosynthetic processes; CC: Cellular component; COT: Cotyledon; cpm: Counts per million; DAF: Day after flowering; DEGs: Differentially expressed genes; DS: Dry seed; EM: Early-maturation; ER: Endoplasmic reticulum; GLOB: Globular; GO: Gene Ontology; HRT: Heart; LF: Leaf; MDS: Multidimensional scaling; MF: Molecular function; MCODE: Molecular Complex Detection; RT: Root; SDL: Seedling; SMPs: Seed maturation proteins; STM: Stem; TS: Tissue-specific

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07783-z.

> **Additional file 1.**

Authors' contributions
DZ and HZ conceived and designed the experiments. HZ, ZH, YY and XL conducted the experiment. HZ and HL performed data analysis. BS and YA assisted data analysis. HZ and DZ wrote the manuscript. ZL provided experimental materials. All authors read and approved the manuscript.

Availability of data and materials
All data generated or analyzed during this study are included in this article and its additional files.
The datasets generated and/or analysed during the current study are available in the Sequence Read Archive of the National Center for Biotechnology Information, "SRR9613714 - SRR9613721, SRR9613704, SRR13810418 - SRR13810420" under the BioProject: PRJNA551573. The Williams 82 soybean reference genome sequence was downloaded from the Phytozome v12 (https://phytozome.jgi.doe.gov/pz/portal.html). Data that was used for co-expression construction was retrieved from SoyNet (https://www.inetbio.org/soynet/).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Collaborative Innovation Center of Henan Grain Crops, College of Agronomy, Henan Agricultural University, Zhengzhou 450002, China. [2]The Innovative Academy of Seed Design, Key Laboratory of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin 150081, China. [3]Department of Biology, Saint Louis University, St. Louis, MO, USA. [4]Department of Biological Sciences, University of North Carolina at Charlotte, Charlotte, NC 28223, USA. [5]US Department of Agriculture, Agricultural Research Service, Midwest Area, Plant Genetics Research Unit at Donald Danforth Plant Science Center, St. Louis, MO 63132, USA.

References
1. Anderson EJ, Ali MD, Beavis WD, Chen P, Clemente TE, Diers BW, et al. Soybean [*Glycine max* (L.) Merr.] breeding: History, improvement, production and future opportunities. In: Al-Khayri J, Jain S, Johnson D, editors. Advances in Plant Breeding Strategies: Legumes. Cham: Springer; 2019. p. 431–516.
2. USDA-NASS: United State Department of Agriculture, National Agricultural Statistics Service. https://www.nass.usda.gov/Charts_and_Maps/Field_Crops/index.php. 2019.
3. Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, et al. Food security: the challenge of feeding 9 billion people. Science. 2010; 327(5967):812–8. https://doi.org/10.1126/science.1185383.
4. Tilman D, Balzer C, Hill J, Befort BL. Global food demand and the sustainable intensification of agriculture. P Natl Acad Sci USA. 2011;108(50):20260–4. https://doi.org/10.1073/pnas.1116437108.
5. Le BH, Wagmaister JA, Kawashima T, Bui AQ, Harada JJ, Goldberg RB. Using genomics to study legume seed development. Plant Physiol. 2007;144(2): 562–74. https://doi.org/10.1104/pp.107.100362.
6. Weber H, Borisjuk L, Wobus U. Molecular physiology of legume seed development. Annu Rev Plant Biol. 2005;56(1):253–79. https://doi.org/10.1146/annurev.arplant.56.032604.144201.
7. Lin JY, Le BH, Chen M, Henry KF, Hur J, Hsieh TF, et al. Similarity between soybean and *Arabidopsis* seed methylomes and loss of non-CG methylation does not affect seed development. Proc Natl Acad Sci U S A. 2017;114(45): E9730–9. https://doi.org/10.1073/pnas.1716758114.
8. Song QX, Li QT, Liu YF, Zhang FX, Ma B, Zhang WK, et al. Soybean GmbZIP123 gene enhances lipid content in the seeds of transgenic Arabidopsis plants. J Exp Bot. 2013;64(14):4329–41. https://doi.org/10.1093/jxb/ert238.
9. Lu X, Li QT, Xiong Q, Li W, Bi YD, Lai YC, et al. The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. Plant J. 2016;86(6):530–44. https://doi.org/10.1111/tpj.13181.
10. Danzer J, Mellott E, Bui AQ, Le BH, Martin P, Hashimoto M, et al. Down-regulating the expression of 53 soybean transcription factor genes uncovers a role for SPEECHLESS in initiating stomatal cell lineages during embryo development. Plant Physiol. 2015;168(3):1025–35. https://doi.org/10.1104/pp.15.00432.
11. Wang S, Liu S, Wang J, Yokosho K, Zhou B, Yu Y, Liu Z, Frommer WB, Ma J, Chen L *et al*: Simultaneous changes in seed size, oil content, and protein content driven by selection of SWEET homologues during soybean domestication. National Science Review 2020:nwaa110.
12. Wang S, Yokosho K, Guo R, Whelan J, Ruan YL, Ma JF, Shou H: The soybean sugar transporter GmSWEET15 mediates sucrose export from endosperm to early embryo. Plant Physiol 2019:https://doi.org/10.1104/pp.1119.00641.
13. Jo L, Pelletier JM, Harada JJ. Central role of the LEAFY COTYLEDON1 transcription factor in seed development. J Integr Plant Biol. 2019;61(5):564–80. https://doi.org/10.1111/jipb.12806.
14. Guo W, Chen LM, Chen HF, Yang HL, You QB, Bao AL, et al. Overexpression of GmWRI1b in soybean stably improves plant architecture and associated yield parameters, and increases total seed oil production under field conditions. Plant Biotechnol J. 2020;18(8):1639–41. https://doi.org/10.1111/pbi.13324.
15. Manan S, Ahmad MZ, Zhang GY, Chen BB, Haq BU, Yang JH, et al. Soybean LEC2 regulates subsets of genes involved in controlling the biosynthesis and catabolism of seed storage substances and seed development. Front Plant Sci. 2017;8:1604. https://doi.org/10.3389/fpls.2017.01604.
16. Zhang D, Zhao M, Li S, Sun L, Wang W, Cai C, et al. Plasticity and innovation of regulatory mechanisms underlying seed oil content mediated by duplicated genes in the palaeopolyploid soybean. Plant J. 2017;90(6):1120–33. https://doi.org/10.1111/tpj.13533.
17. Jo L, Pelletier JM, Hsu SW, Baden R, Goldberg RB, Harada JJ. Combinatorial interactions of the LEC1 transcription factor specify diverse developmental programs during soybean seed development. Proc Natl Acad Sci U S A. 2020;117(2):1223–32. https://doi.org/10.1073/pnas.1918441117.
18. Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, et al. Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci U S A. 2010;107(18):8063–70. https://doi.org/10.1073/pnas.1003530107.

19. Garg R, Singh VK, Rajkumar MS, Kumar V, Jain M. Global transcriptome and coexpression network analyses reveal cultivar-specific molecular signatures associated with seed development and seed size/weight determination in chickpea. Plant J. 2017;91(6):1088–107. https://doi.org/10.1111/tpj.13621.

20. Du J, Wang S, He C, Zhou B, Ruan YL, Shou H. Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. J Exp Bot. 2017;68(8):1955–72. https://doi.org/10.1093/jxb/erw460.

21. Qi ZM, Zhang ZG, Wang ZY, Yu JY, Qin HT, Mao XR, et al. Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development. Plant Cell and Environment. 2018;41(9):2109–27.

22. Yang SN, Miao L, He JB, Zhang K, Li Y, Gai JY. Dynamic transcriptome changes related to oil accumulation in developing soybean seeds. Int J Mol Sci. 2019;20(9):18.

23. Fehr WR, Caviness CE: Stages of soybean development. Iova Coop. Ext. Serv. Spec. Rep. 80. Iowa State Univ., Ames, IA. 1977.

24. Richie SW, Hanway JJ, Thomas HE, Benson GO. How a soybean plant develops。 special report NO.53. Ames: Iowa State University of Science and Technology Cooperative Extension Service; 1996.

25. Assefa Y, Purcell LC, Salmeron M, Naeve S, Casteel SN, Kovacs P, et al. Assessing variation in US soybean seed composition (protein and oil). Front Plant Sci. 2019;10:298. https://doi.org/10.3389/fpls.2019.00298.

26. Hill JE, Breidenbach RW. Proteins of soybean seeds: II. Accumulation of the major protein components during seed development and maturation. Plant Physiol. 1974;53(5):741–51.

27. Considine MJ, Considine JA. On the language and physiology of dormancy and quiescence in plants. J Exp Bot. 2016;67(11):3189–203. https://doi.org/10.1093/jxb/erw138.

28. Feng N, Song G, Guan J, Chen K, Jia M, Huang D, et al. Transcriptome profiling of wheat inflorescence development from spikelet initiation to foral patterning identified stage-specific regulatory genes. Plant Physiol. 2017;174(3):1779–94. https://doi.org/10.1104/pp.17.00310.

29. Huang AHC. Plant lipid droplets and their associated proteins: potential for rapid advances. Plant Physiol. 2018;176(3):1894–918. https://doi.org/10.1104/pp.17.01677.

30. Ohlrogge J, Browse J. Lipid biosynthesis. Plant Cell. 1995;7(7):957–70. https://doi.org/10.1105/tpc.7.7.957.

31. Du C, Chen Y, Wang K, Yang Z, Zhao C, Jia Q, et al. Strong co-suppression impedes an increase in polyunsaturated fatty acids in seeds overexpressing *FAD2*. J Exp Bot. 2019;70(3):985–94. https://doi.org/10.1093/jxb/ery378.

32. Zhang D, Zhang H, Hu Z, Chu S, Yu K, Lv L, et al. Artificial selection on *GmOLEO1* contributes to the increase in seed oil during soybean domestication. PLoS Genet. 2019;15(7):e1008267. https://doi.org/10.1371/journal.pgen.1008267.

33. Schmidt MA, Herman EM. Suppression of soybean oleosin produces micro-oil bodies that aggregate into oil body/ER complexes. Mol Plant. 2008;1(6):910–24. https://doi.org/10.1093/mp/ssn049.

34. Sarmiento C, Ross JHE, Murphy DJ. Expression and subcellular targeting of a soybean oleosin in transgenic rapeseed. Implications for the mechanism of oil-body formation in seeds. Plant J. 1997;11(4):783–96. https://doi.org/10.1046/j.1365-313X.1997.11040783.x.

35. Gillman JD, Kim WS, Krishnan HB. Identification of a new soybean kunitz trypsin inhibitor mutation and its effect on bowman-birk protease inhibitor content in soybean seed. J Agric Food Chem. 2015;63(5):1352–9. https://doi.org/10.1021/jf505220p.

36. Chen LQ, Qu XQ, Hou BH, Sosso D, Osorio S, Fernie AR, et al. Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. Science. 2012;335(6065):207–11. https://doi.org/10.1126/science.1213351.

37. Zhang HY, Goettel W, Song QJ, Jiang H, Hu ZB, Wang ML, An YQC. Selection of GmSWEET39 for oil and protein improvement in soybean. PLoS Genet. 2020;16(11):e1009114. https://doi.org/10.1371/journal.pgen.1009114.

38. Sosso D, Luo DP, Li QB, Sasse J, Yang JL, Gendrot G, et al. Seed filling in domesticated maize and rice depends on SWEET-mediated hexose transport. Nat Genet. 2015;47(12):1489–93. https://doi.org/10.1038/ng.3422.

39. Yang JL, Luo DP, Yang B, Frommer WB, Eom JS. SWEET11 and 15 as key players in seed filling in rice. New Phytol. 2018;218(2):604–15. https://doi.org/10.1111/nph.15004.

40. Mechin V, Thevenot C, Le Guilloux M, Prioul JL, Damerval C. Developmental analysis of maize endosperm proteome suggests a pivotal role for pyruvate orthophosphate dikinase. Plant Physiol. 2007;143(3):1203–19. https://doi.org/10.1104/pp.106.092148.

41. Vensel WH, Tanaka CK, Cai N, Wong JH, Buchanan BB, Hurkman WJ. Developmental changes in the metabolic protein profiles of wheat endosperm. Proteomics. 2005;5(6):1594–611. https://doi.org/10.1002/pmic.200401034.

42. Porta H, Rocha-Sosa M. Plant lipoxygenases. Physiological and molecular features. Plant Physiol. 2002;130(1):15–21. https://doi.org/10.1104/pp.010787.

43. Fischer AM, Dubbs WE, Baker RA, Fuller MA, Stephenson LC, Grimes HD. Protein dynamics, activity and cellular localization of soybean lipoxygenases indicate distinct functional roles for individual isoforms. Plant J. 1999;19(5):543–54. https://doi.org/10.1046/j.1365-313X.1999.00550.x.

44. King JM, Svendsen LK, Fehr WR, Narvel JM, White PJ. Oxidative and flavor stability of oil from lipoxygenase-free soybeans. J Am Oil Chemists Soc. 1998;75(9):1121–6. https://doi.org/10.1007/s11746-998-0123-7.

45. Kroj T, Savino G, Valon C, Giraudat J, Parcy F. Regulation of storage protein gene expression in *Arabidopsis*. Development. 2003;130(24):6065–73. https://doi.org/10.1242/dev.00814.

46. Lara P, Onate-Sanchez L, Abraham Z, Ferrandiz C, Diaz I, Carbonero P, et al. Synergistic activation of seed storage protein gene expression in Arabidopsis by ABI3 and two bZIPs related to OPAQUE2. J Biol Chem. 2003;278(23):21003–11. https://doi.org/10.1074/jbc.M210538200.

47. Pelletier JM, Kwong RW, Park S, Le BH, Baden R, Cagliaria A, et al. LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development. Proc Natl Acad Sci U S A. 2017;114(32):E6710–9. https://doi.org/10.1073/pnas.1707957114.

48. Angelovici R, Galili G, Fernie AR, Fait A. Seed desiccation: a bridge between maturation and germination. Trends Plant Sci. 2010;15(4):211–8. https://doi.org/10.1016/j.tplants.2010.01.003.

49. Nambara E, Hayama R, Tsuchiya Y, Nishimura M, Kawaide H, Kamiya Y, et al. The role of ABI3 and FUS3 loci in *Arabidopsis thaliana* on phase transition from late embryo development to germination. Dev Biol. 2000;220(2):412–23. https://doi.org/10.1006/dbio.2000.9632.

50. Baroja-Fernandez E, Munoz FJ, Li J, Bahaji A, Almagro G, Montero M, et al. Sucrose synthase activity in the sus1/sus2/sus3/sus4 *Arabidopsis* mutant is sufficient to support normal cellulose and starch production. Proc Natl Acad Sci U S A. 2012;109(1):321–6. https://doi.org/10.1073/pnas.1117099109.

51. Zhang B, Li CX, Li Y, Yu H. Mobile TERMINAL FLOWER1 determines seed size in *Arabidopsis*. Nature Plants. 2020;6(9):1146–57. https://doi.org/10.1038/s41477-020-0749-5.

52. Rieu I, Eriksson S, Powers SJ, Gong F, Griffiths J, Woolley L, et al. Genetic analysis reveals that C-19-GA 2-oxidation is a major gibberellin inactivation pathway in *Arabidopsis*. Plant Cell. 2008;20(9):2420–36. https://doi.org/10.1105/tpc.108.058818.

53. Park SY, Yu JW, Park JS, Li J, Yoo SC, Lee NY, et al. The senescence-induced staygreen protein regulates chlorophyll degradation. Plant Cell. 2007;19(5):1649–64. https://doi.org/10.1105/tpc.106.044891.

54. Bray EA. Genes commonly regulated by water-deficit stress in Arabidopsis thaliana. J Exp Bot. 2004;55(407):2331–41. https://doi.org/10.1093/jxb/erh270.

55. Dunaeva M, Adamska W. Identification of genes expressed in response to light stress in leaves of *Arabidopsis thaliana* using RNA differential display. Eur J Biochem. 2001;268(21):5521–9. https://doi.org/10.1046/j.1432-1033.2001.02471.x.

56. Miao L, Yang SN, Zhang K, He JB, Wu CH, Ren YH, et al. Natural variation and selection in GmSWEET39 affect soybean seed oil content. New Phytol. 2020;225(4):1651–66. https://doi.org/10.1111/nph.16250.

57. Tan HL, Yang XH, Zhang FX, Zheng X, Qu CM, Mu JY, et al. Enhanced seed oil production in canola by conditional expression of *Brassica napus LEAFY COTYLEDON1* and *LEC1-LIKE* in developing seeds. Plant Physiol. 2011;156(3):1577–88. https://doi.org/10.1104/pp.111.175000.

58. Atkins CA, Smith PMC. Translocation in legumes: assimilates, nutrients, and signaling molecules. Plant Physiol. 2007;144(2):550–61. https://doi.org/10.1104/pp.107.098046.

59. Zhang L, Tan Q, Lee R, Trethewy A, Lee YH, Tegeder M. Altered xylem-phloem transfer of amino acids affects metabolism and leads to increased seed yield and oil content in *Arabidopsis*. Plant Cell. 2010;22(11):3603–20. https://doi.org/10.1105/tpc.110.073833.

60. Rogers EE, Guerinot ML. FRD3, a member of the multidrug and toxin efflux family, controls iron deficiency responses in *Arabidopsis*. Plant Cell. 2002;14(8):1787–99. https://doi.org/10.1105/tpc.001495.

61. Chen LQ, Lin IWN, Qu XQ, Sosso D, McFarlane HE, Londono A, et al. A cascade of sequentially expressed sucrose transporters in the seed coat and endosperm provides nutrition for the *Arabidopsis* embryo. Plant Cell. 2015;27(3):607–19. https://doi.org/10.1105/tpc.114.134585.

Zhang *et al. BMC Genomics*   (2021) 22:453

Page 13 of 13

62. Kim E, Hwang S, Lee I. SoyNet: a database of co-functional networks for soybean *Glycine max*. Nucleic Acids Res. 2017;45(D1):D1082–9. https://doi.org/10.1093/nar/gkw704.

63. Clark NM, Buckner E, Fisher AP, Nelson EC, Nguyen TT, Simmons AR, et al. Stem-cell-ubiquitous genes spatiotemporally coordinate division through regulation of stem-cell-specific gene networks. Nat Commun. 2019;10(1):5574. https://doi.org/10.1038/s41467-019-13132-2.

64. Cifuentes M, Jolivet S, Cromer L, Harashima H, Bulankova P, Renne C, et al. TDM1 regulation determines the number of meiotic divisions. PLoS Genet. 2016;12(2):e1005856. https://doi.org/10.1371/journal.pgen.1005856.

65. Yang XS, Wang LL, Zhou XR, Shuang SM, Zhu ZH, Li N, et al. Determination of protein, fat, starch, and amino acids in foxtail millet [*Setaria italica* (L.) Beauv.] by fourier transform near-infrared reflectance spectroscopy. Food Sci Biotechnol. 2013;22(6):1495–500. https://doi.org/10.1007/s10068-013-0243-1.

66. Yang YB, Jia GQ, Deng LG, Qin L, Chen EY, Cong XJ, et al. Genetic variation of yellow pigment and its components in foxtail millet (*Setaria italica* (L.) P. Beauv.) from different eco-regions in China. J Integr Agric. 2017;16(11):2459–69. https://doi.org/10.1016/S2095-3119(16)61598-8.

67. Zhang D, Zhang H, Chu S, Li H, Chi Y, Triebwasser-Freese D, et al. Integrating QTL mapping and transcriptomics identifies candidate genes underlying QTLs associated with soybean tolerance to low-phosphorus stress. Plant Mol Biol. 2017;93(1–2):137–50. https://doi.org/10.1007/s11103-016-0552-x.

68. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.

69. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463(7278):178–83. https://doi.org/10.1038/nature08670.

70. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656.

71. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.

72. Zhang HY, Kjemtrup-Lovelace S, Li CB, Luo Y, Chen LP, Song BH. Comparative RNA-seq analysis uncovers a complex regulatory network for soybean cyst nematode resistance in wild soybean (*Glycine soja*). Sci Rep-Uk. 2017;7(1):9699. https://doi.org/10.1038/s41598-017-09945-0.

73. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4(1):2. https://doi.org/10.1186/1471-2105-4-2.

74. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc. 2007;2(10):2366–82. https://doi.org/10.1038/nprot.2007.324.

Publisher's Note