Genomics Data 6 (2015) 92–98

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Identification of differentially expressed genes between developing seeds of different soybean cultivars

Rongshuang Lin [a], Jane Glazebrook [a], Fumiaki Katagiri [a], James H. Orf [b], Susan I. Gibson [a,*]

[a] Department of Plant Biology, University of Minnesota, 1500 Gortner Ave., Saint Paul, MN 55108, USA
[b] Department of Agronomy & Plant Genetics, University of Minnesota, 1991 Upper Buford Circle, Saint Paul, MN 55108, USA

ARTICLE INFO

Article history:
Received 4 May 2015
Received in revised form 5 August 2015
Accepted 7 August 2015
Available online 13 August 2015

Keywords:
Soybean
Seed
Composition
Yield
Transcriptional profiling

ABSTRACT

Soybean is a major source of protein and oil and a primary feedstock for biodiesel production. Research on soybean seed composition and yield has revealed that protein, oil and yield are controlled quantitatively and quantitative trait loci (QTL) have been identified for each of these traits. However, very limited information is available regarding the genetic mechanisms controlling seed composition and yield. To help address this deficiency, we used Affymetrix Soybean GeneChips® to identify genes that are differentially expressed between developing seeds of the Minsoy and Archer soybean cultivars, which differ in seed weight, yield, protein content and oil content. A total of 700 probe sets were found to be expressed at significantly different (defined as having an adjusted *p*-value below or equal to 0.05 and an at least 2-fold difference) levels between the two cultivars at one or more of the three developmental stages and in at least one of the two years assayed. Comparison of data from soybeans collected in two different years revealed that 97 probe sets were expressed at significantly different levels in both years. Functional annotations were assigned to 78% of these 97 probe sets based on the SoyBase Affymetrix™ GeneChip® Soybean Genome Array Annotation. Genes involved in receptor binding/activity and protein binding are overrepresented among the group of 97 probe sets that were differentially expressed in both years assayed. Probe sets involved in growth/development, signal transduction, transcription, defense/stress response and protein and lipid metabolism were also identified among the 97 probe sets and their possible implications in the regulation of agronomic traits are discussed. As the Minsoy and Archer soybean cultivars differ with respect to seed size, yield, protein content and lipid content, some of the differentially expressed probe sets identified in this study may thus play important roles in controlling these traits. Others of these probe sets may be involved in regulation of general seed development or metabolism. All microarray data and expression values after GCRMA are available at the Gene Expression Omnibus (GEO) at NCBI (http://www.ncbi.nlm.nih.gov/geo), under accession number GSE21598.

Specifications table	
Subject area	Biology
More specific subject area	Plant biology, agriculture
Type of data	Table of expression values after GCRMA, CEL files with raw GeneChip data
How data was acquired	Affymetrix Soybean GeneChips®
Data format	Raw data and expression values after GCRMA
Experimental factors	None
Experimental features	Microarray expression profiling to identify genes that are differentially expressed in developing seeds of the Minsoy and Archer soybean varieties or in two recombinant inbred lines derived from a Minsoy X Archer cross.
Data source location	Saint Paul, MN, USA.
Data accessibility	All microarray data and expression values after GCRMA are available at the Gene Expression Omnibus (GEO) at NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21598), under accession number GSE21598.

Value of the data

- Although soybean is a major source of protein and oil and a primary feedstock for biodiesel production, little is known about the genetic mechanisms controlling variations in yield and seed composition between different soybean varieties.
- This study provides expression-profiling data for developing seeds from the Minsoy and Archer soybean varieties, which differ with respect to seed yield, composition and size. These data may thus aid in studies aimed at determining how alterations in gene expression affect these critical agronomic traits.

Abbreviations: FDR, false discovery rate; GEO, Gene Expression Omnibus; GCRMA, Guanine Cytosine Robust Multi-Array analysis; GO, Gene Ontology; GPI, glycosylphosphatidylinositol; HY, high yield; LY, low yield; QTL, quantitative trait locus; RCB, randomized complete block; RIL, recombinant inbred line; TAG, triacylglycerol.

* Corresponding author.
E-mail addresses: rsglorialin@gmail.com (R. Lin), jglazebr@umn.edu (J. Glazebrook), katagiri@umn.edu (F. Katagiri), orfxx001@umn.edu (J.H. Orf), gibso043@umn.edu (S.I. Gibson).

http://dx.doi.org/10.1016/j.gdata.2015.08.005

- Expression profiling data is also provided for developing seeds of two recombinant inbred lines from a Minsoy X Archer cross that differ with respect to seed yield. These data may thus aid in studies aimed at determining how alterations in gene expression affect seed yield.
- A total of 700 probe sets (roughly corresponding to 700 genes) that exhibit significantly different expression values between the Minsoy and Archer soybean varieties during at least one of the three developmental stages assayed are identified in this study.
- Information about the expression levels of genes at different stages of soybean seed development is expected to aid studies on seed development.

1. Introduction

Soybean (*Glycine max* (L.) Merrill) is a rich source of protein and oil and is one of the most important crop plants worldwide. Traditionally, soybeans have mainly been used for vegetable oil production, animal feed, and direct human consumption. In recent years there has been an increase in the use of soybean for biodiesel production, with approximately 10% of the U.S. soybean acres being used for biodiesel production [2]. Expanded production and use of biodiesel depends partly on the price of the feedstock. In order to make biodiesel production from soybean more economically competitive, improvements in soybean seed composition and/or yield are needed [3].

Dehulled soybean seeds (embryos) contain an average 41% protein, 28% carbohydrates, 25% oil and 5% ash [4]. Extensive research has been carried out regarding the regulation of soybean seed composition. Both seed oil and protein content are quantitatively inherited, and many quantitative trait loci (QTL) associated with the two traits have been identified [5–10]. However, although many components of the biochemical pathways for protein and oil biosynthesis in developing seeds have been identified [11–14], little is known about the genetic mechanisms that control protein or oil content [15,16]. Even though it is possible to increase seed protein and oil content simultaneously to some extent, the negative correlation between protein and oil content makes it very difficult to increase the content of one without decreasing the other [6,7]. Research into yield determinants has also been performed in soybean [17–19]; yet, the basis of yield improvement remains unclear. The fact that seed protein and oil content as well as yield are greatly affected by environmental factors makes molecular dissection of these important agronomic traits more difficult [7,17,18,20–24].

To gain more information regarding the genes controlling seed protein, oil, and yield in soybean, we initiated a genomics approach to identify genetic factors that control variations in gene expression. The theory underlying this approach is that regulation of gene expression plays a critical role in determination of agronomic traits. In brief, a major reason why two cultivars differ with respect to agronomic traits is likely to be differences in the expression levels of key genes. Thus, it is of interest to identify those genes that are differentially expressed between cultivars that vary with respect to the agronomic traits of interest. Towards this end we used Affymetrix Soybean GeneChips® to perform transcriptional profiling experiments on developing seeds from the Minsoy and Archer soybean cultivars to identify differentially expressed genes. The Minsoy and Archer cultivars differ with respect to seed yield and size and also exhibit minor differences in seed composition (Table 1) [7,9,18]. These experiments resulted in identification of approximately 700 probe sets that are differentially expressed between seeds of the Minsoy and Archer soybean cultivars at one or more of the three developmental stages tested in at least one of two years assayed. Some of these probe sets may thus be involved in helping regulate seed composition or yield, while other probe sets may be involved in helping regulate general seed development or metabolism.

Table 1
Comparison of seed composition, weight and yield between the Minsoy and Archer cultivars and two recombinant inbred lines (RILs). Data for the two parental lines were taken from published sources [7,9,18] and data for the two RILs are based on combined data collected from plants grown during 2000 in Becker, MN and during 2001 in Waseca, MN (unpublished data).

Trait	RIL 6049-273	RIL 6049-32	Archer	Minsoy
Yield (bu/a)	66.1	33.5	49	29
Oil (g/kg)	166	162	187	178
Protein (g/kg)	344	366	340	353
Seed weight (mg/seed)	188	142	161	123

2. Results and discussion

2.1. Identification of differentially expressed probe sets

The Minsoy and Archer soybean cultivars were chosen for these experiments. Minsoy exhibits high seed protein content, and Archer is an elite cultivar adapted to northern U.S. areas with high seed oil content, seed weight and yield [9]. Previously, a recombinant inbred line (RIL) population was constructed using these two cultivars and QTL associated with seed protein, oil and yield were identified [9,18]. The parental lines were grown and developing seeds were collected in the summers of both 2007 and 2008. In addition, in 2007 developing seeds were also collected from two RILs that were derived from a cross between Minsoy and Archer (Table 1). Seeds were harvested at three developmental stages that are critical for seed number, seed size and seed nutrient accumulation [25]. Transcriptional profiling experiments were performed using Affymetrix GeneChip® Soybean Genome Arrays. Data analyses were performed as described in the Materials and methods section.

The 37,744 probe sets specific to the soybean genome were extracted. Data preprocessing and quality assessment of the hybridization data revealed that the data are of good quality (Supplementary files 1–3). The percentages of probe sets identified as “present” were 74.5–77.9%, which is very similar to the results obtained by Alvord et al. [26] using the same soybean Affymetrix array for research on fungal infection. After removing transcripts with very low expression values (background noise), about 25,000 transcripts were determined to be expressed at significant levels in soybean seeds and were thus used for subsequent statistical analysis. Based on cutoffs of an adjusted *p*-value (q-value) of less than or equal to 0.05 and two-fold or greater differences in expression, the number of differentially expressed probe sets between the two parental lines and the RILs were determined and are listed in Table 2.

Table 2
Comparison of number of differentially expressed transcripts between Minsoy and Archer and two recombinant inbred lines (RILs). The significant cutoff value is q-value = 0.05 and 2-fold change in expression values. The two RILs are LY (lower yield) line 6049–32 and HY (higher yield) line 6049-273.

Affymetrix probe sets	Stage 1	Stage 2	Stage 3	Total in all stages combined[a]
Higher in Minsoy in 2007	117	120	167	273
Higher in Archer in 2007	172	140	135	301
Total different between Minsoy and Archer in 2007	289	260	302	574
Higher in Minsoy in 2008	62	67	73	104
Higher in Archer in 2008	89	76	97	128
Total different between Minsoy and Archer in 2008	151	143	170	232
Higher in LY RIL in 2007	74	6	0	78
Higher in HY RIL in 2007	55	14	0	66
Total different between LY and HY RILs in 2007	129	20	0	144

[a] The total number of differentially expressed transcripts in all stages combined is less than the sum of the numbers of differentially expressed transcripts in each of the individual stages as some transcripts were differentially expressed at multiple developmental stages.

Using the cutoffs described above, a total of 574 probe sets were found to be differentially expressed between the two parental cultivars in 2007 and 232 probe sets were found to be differentially expressed between the two parental cultivars in 2008. These differentially expressed probe sets are distributed approximately equally among the three developmental stages tested and a substantial number of these probe sets were found to be differentially expressed at more than one developmental stage (Fig. 1). The higher number of differentially expressed probe sets in 2007 compared with 2008 may be related to the fact that replicate samples were collected from a single row in 2007 and thus may have had lower variability than the replicate samples collected from multiple rows in 2008. About 48% and 45% of the differentially expressed probe sets were expressed at higher levels in Minsoy than in Archer in 2007 and 2008, respectively. Thus, the number of probe sets that are expressed at higher levels in Minsoy is approximately equal to the number of probe sets that are expressed at higher levels in Archer. The union of the differentially expressed probe sets identified in 2007 and 2008 is 700; data for expression values, fold changes and q-values for these genes are available in Supplementary file 4.

Transcript levels were also compared between tissue samples collected in 2007 from the two RILs 6049–273 and 6049–32 (Table 1). Using the same cutoffs described above, a total of 144 probe sets were found to be expressed at significantly different levels between these two RILs in at least one developmental stage (Table 2). Interestingly, no probe sets were found to be expressed at significantly different levels between seeds of the two RILs at developmental stage 3, the most mature stage analyzed. Among the 144 differentially expressed probe sets, 78 were expressed at higher levels in Line 6049-32 while 66 were expressed at higher levels in Line 6049-273; therefore, an approximately equal number of probe sets are expressed at higher levels in each of the two lines. Information regarding expression of these differentially expressed probe sets is provided in Supplementary file 5.

For the two parental cultivars, the probe sets that were differentially expressed in both of the years analyzed may represent genes that play more important roles in controlling differences in seed development or metabolism, potentially including differences in some of the agronomic traits, such as yield, observed between the parental cultivars. We therefore compared the transcripts that were differentially expressed between the parental cultivars in 2007 and 2008 and identified probe sets that were differentially expressed in both years (Fig. 1). There are 75, 56, and 41 transcripts that were differentially expressed in both years for stages 1, 2 and 3, respectively. As some of these transcripts were consistently differentially expressed at more than one developmental stage, the total number of transcripts that were differentially expressed in both years in at least one developmental stage is 97. In all cases where a transcript was found to be consistently differentially expressed in more than one developmental stage, the transcript was found to be higher in the same parental line at all stages in which that transcript was differentially expressed. Information regarding these 97 transcripts is provided in Supplementary file 6 and discussed below.

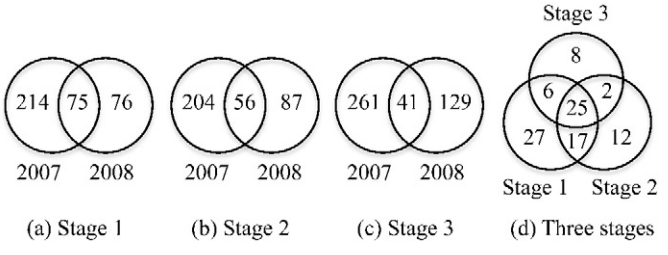

Fig. 1. Comparison of differentially expressed probe sets between Minsoy and Archer. A comparison was made for each developmental stage to identify probe sets differentially expressed in both 2007 and 2008 (panels a, b, c); then the numbers from the three stages were combined in panel (d). In panels (a), (b) and (c), the number in the left circle represents probe sets unique to 2007, the number in the right circle represents probe sets unique to 2008, and the number in the center area represents probe sets common to 2007 and 2008. Panel (d) illustrates the distribution of probe sets common to 2007 and 2008 among the three stages.

2.2. Functional categories of differentially expressed probe sets

As the 97 transcripts found to be differentially expressed in both years analyzed may be more likely to play important roles in controlling differences in seed development and metabolism, including potentially differences in some agronomic traits, than genes that were found to be differentially expressed in only one of the years analyzed, further analyses and discussion are focused on these 97 transcripts. Annotation information for all probe sets that were differentially expressed in either year between either the parental cultivars or the two RILs analyzed is also provided in Supplementary Files 4 and 5. Approximately 78% of the 97 probe sets differentially expressed in Minsoy and Archer in both years analyzed were annotated based on their homology to *Arabidopsis thaliana* gene models or to protein sequences from other organisms (Supplementary File 6). Probe sets were classified based on their GOslim descriptions and the percentages of probe sets assigned to each GOslim category were calculated (Fig. 2). As shown in Fig. 2, many probe sets fall into the "unknown" or "other" categories for molecular functions or biological processes. Probe sets involved in

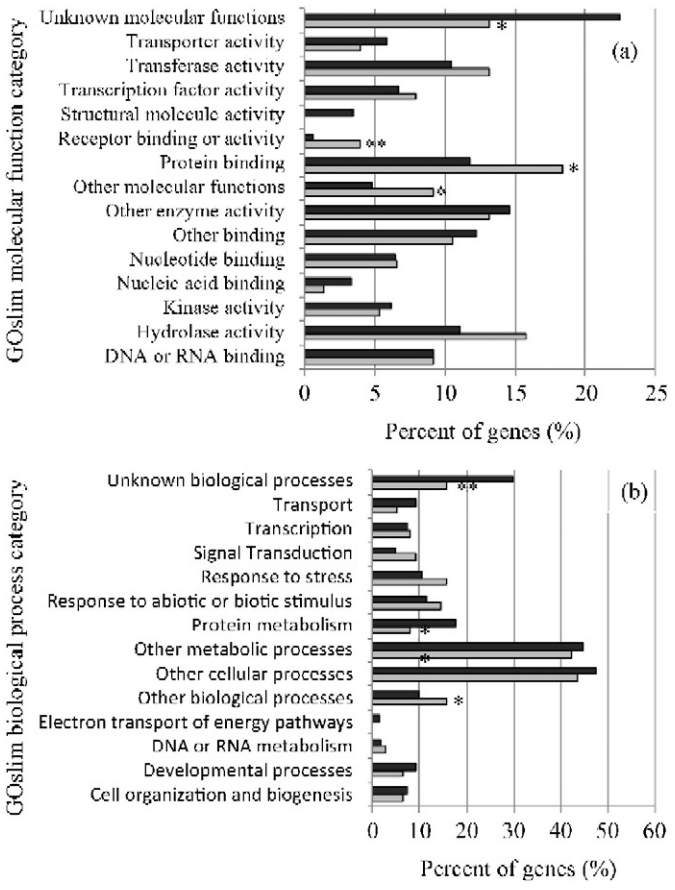

Fig. 2. Gene Ontology (GO)slim categorization of differentially expressed probe sets. Panel (a) describes the percentage of probe sets annotated to different GOslim functional categories. Panel (b) describes the percentage of probe sets annotated to different GOslim biological process categories The percentage of probe sets annotated to a GOslim category is equal to the number of probe sets annotated to terms in that GOslim category/total number of probe sets annotated to any term in this ontology times 100. The light-gray bars illustrate data from the 97 probe sets differentially expressed between Minsoy and Archer and the dark-gray bars represent all the other probe sets on the array whose expression was detected in 2007 and 2008. * indicates that the results of the 97 differentially expressed probe sets differed from those of all other detected probe sets on the array with a p-value <0.1 and ** indicates the results differed with a p-value <0.05 using Fisher's exact test.

Table 3
Probe sets potentially involved in seed growth/development and signal transduction. Probe sets that were found to be expressed at significantly different levels between Minsoy and Archer in one or more developmental stages during 2007 and 2008 were searched for those predicted by their GO assignments to be involved in seed growth/development/signal transduction. Significant differences in transcript levels are defined as having a q-value ≤0.05 and fold difference ≥ 2. The developmental stage(s) and year(s) in which significant differences in transcript levels were detected for a particular gene are indicated. The cultivar that has higher expression values is also indicated. S1, S2, S3 represent developmental stage 1 (2 mm seeds), stage 2 (3.5 mm seeds) and stage 3 (5–6 mm seeds), respectively. (07) and (08) refer to 2007 and 2008, respectively.

Affymetrix probe set ID	Gene description	Stage (year)	Higher expression
Gma.1471.2.S1_at	Suppressor of PHYA-105 1 (SPA1)	S1,2 (07);S1,2,3 (08)	Archer
Gma.12643.1.A1_at	Light repressible receptor protein kinase	S1 (07);S1,2,3 (08)	Minsoy
GmaAffx.88161.1.S1_at	Phytochrome (PHYA)	S1 (07);S1,2,3 (08)	Archer
Gma.7387.1.A1_at	Pseudo-response regulator 7 (PRR7)	S1,2,3 (07,08)	Minsoy
GmaAffx.77896.1.S1_at	Suppressor of auxin resistance 3 (SAR3)	S1,2,3 (07,08)	Archer
GmaAffx.90343.1.S1_at	GAST1 protein homolog 4 (GASA4)	S1,2 (07);S2,3 (08)	Minsoy
GmaAffx.90343.1.S1_s_at	GAST1 protein homolog 4 (GASA4)	S1,2 (07);S2,3 (08)	Minsoy
Gma.473.1.A1_at	Retinoblastoma-related 1 (RBR1)	S1 (07);S1,2,3 (08)	Archer
GmaAffx.21668.1.A1_at	COBRA (COB)	S1,2,3 (07,08)	Archer
GmaAffx.28349.1.S1_at	Ankyrin repeat_containing 2B (AKR2B)	S1,2 (07);S1,2,3 (08)	Archer
GmaAffx.7290.1.A1_at	alpha-N-acetylglucosaminidase family/NAGLU family	S3 (07,08)	Archer

receptor binding/activity and protein binding were overrepresented compared with the remaining expressed transcripts. The finding that probe sets involved in receptor binding/activity or protein binding are overrepresented among the differentially expressed probe sets is interesting as genes with these activities may be more likely than most genes to play important roles in signal transduction pathways. For biological process, probe sets in the protein metabolism category were underrepresented among the differentially expressed probe sets.

The list of 97 probe sets was also searched for those predicted to be involved in processes that might be expected to affect seed yield or composition. These processes include seed growth/development/signal transduction (Table 3), defense and stress response (Table 4), protein metabolism (Table 5) and lipid metabolism (Table 6).

Among the probe sets potentially involved in seed growth and development (Table 3), one homolog of the gibberellic acid-stimulated *Arabidopsis* (GASA) gene *GASA4* was identified in the differentially expressed genes. *GASA4* is one of 14 *GASA* family members in *Arabidopsis* and research has shown that that overexpression of GASA4 in *Arabidopsis* increased seed size and total seed yield [27]. A homolog of COBRA was also differentially expressed. COBRA is a glycosylated protein anchored to the plasma membrane through glycosylphosphatidylinositol (GPI) and was initially characterized as a regulator of orientated cell expansion in *Arabidopsis* roots [28]. It was shown subsequently that COBRA plays an important role in regulating anisotropic expansion during postembryonic development in plants via its involvement in cellulose microfibril orientation [29]. The difference in COBRA expression throughout the three developmental stages may be associated with the difference in seed size between the two cultivars. Archer has larger seeds and 3.7-fold higher expression of the COBRA-related gene on average than Minsoy, based on data from three developmental stages and two years assayed.

A total of 20 probe sets were found to be associated with defense and stress responses (Table 4). The fact that many genes involved in stress/defense responses were expressed differently between the two cultivars may imply the importance of stress/defense responses in seed development and probably in seed size and yield. Studies have shown that defense genes and pathogenesis-related genes are up-regulated in barley embryos during seed development [30]. The authors proposed that this “developmental defense activation” might provide protection to both the developing embryo and dormant seed.

Among the probe sets implicated in lipid metabolism (Table 6), GmaAffx.37987.1.S1_at, which is predicted to encode a diacylglycerol acyltransferase (DGAT), may be of particular interest. Studies have shown that DGAT may play a decisive role in triacylglycerol (TAG) biosynthesis [12] and it has been proposed to catalyze one of the rate-limiting steps in TAG formation during seed development in soybean

Table 4
Probe sets potentially involved in defense and stress responses. Probe sets predicted to be involved in defense and/or stress response and that exhibit significant differences in transcript levels between Minsoy and Archer were identified. See Table 3 legend for additional information.

Affymetrix probe set ID	Gene description	Stage (year)	Higher expression
Gma.1034.4.S1_s_at[b]	Caffeoyl-CoA 3-O-methyltransferase, putative	S3 (07);S1,3 (08)	Minsoy
Gma.11082.2.A1_s_at[b]	Caffeoyl-CoA 3-O-methyltransferase, putative	S3 (07);S1,3 (08)	Minsoy
Gma.11344.1.S1_at	Disease resistance gene (R gene)	S1,2,3 (07,08)	Archer
Gma.13012.1.A1_at	Disease resistance protein (TIR-NBS-LRR class), putative	S1,2,3 (07,08)	Archer
Gma.17871.1.S1_at	Disease resistance protein (TIR-NBS-LRR class), putative	S1 (07);S1,2,3 (08)	Archer
Gma.13036.1.S1_at	Inducer of CBF expression 1 (ICE1)	S1,2,3 (07,08)	Minsoy
Gma.1464.1.S1_at	Rotamase FKBP 1 (ROF1)	S1,2 (07);S1,2,3 (08)	Archer
Gma.1571.1.S1_at	ALDH10A8, 3-chloroallyl aldehyde dehydrogenase/oxidoreductase	S1,2 (07);S2,3 (08)	Minsoy
Gma.16586.1.S1_at	Harpin-induced 1	S1,2,3 (07);S3 (08)	Minsoy
Gma.2628.1.A1_at	*Arabidopsis thaliana* peptide-N-glycanase 1 (AtPNG1)	S1,2 (07,08)	Archer
Gma.4155.1.S1_at[a]	RAP2.12,ERF/AP2 transcription factor family	S1,2,3 (07,08)	Minsoy
Gma.4155.2.S1_s_at[a]	RAP2.12,ERF/AP2 transcription factor family	S1,2,3 (07,08)	Minsoy
GmaAffx.14960.1.S1_s_at[a]	RAP2.12,ERF/AP2 transcription factor family	S1 (07);S1,2,3 (08)	Minsoy
Gma.7716.1.A1_at	Zinc induced facilitator 1 (ZIF1)	S2 (07);S1,2,3 (08)	Minsoy
Gma.8522.1.S1_at	Major latex protein-related/MLP-related	S1,2 (07);S1,2,3 (08)	Archer
GmaAffx.21668.1.A1_at	COBRA (COB)	S1,2,3 (07,08)	Archer
GmaAffx.59014.1.S1_s_at	Dehydrin family protein	S1,2 (07);S1 (08)	Archer
GmaAffx.66290.1.S1_at	Disease resistance protein (CC-NBS class), putative	S1,2 (07);S1,2,3 (08)	Archer
GmaAffx.77896.1.S1_at	Suppressor of auxin resistance 3 (SAR3), nucleoporin	S1,2,3 (07,08)	Archer
GmaAffx.93642.1.S1_s_at	Secretory protein, putative	S2 (07);S1,2 (08)	Archer

[a] These probe sets may represent the same gene as they all show the highest match to the same soybean gene model.
[b] Similarly, these two probe sets may represent the same gene.

Table 5
Probe sets potentially related to protein metabolism. Probe sets predicted to be involved in protein metabolism and that exhibit significant differences in transcript levels between Minsoy and Archer were identified. See Table 3 legend for additional information.

Affymetrix probe set ID	Gene description	Stage (year)	Higher expression
Gma.13070.2.A1_at	Ribosomal protein L25	S1,2,3 (07,08)	Minsoy
Gma.2628.1.A1_at	*Arabidopsis thaliana* peptide-N-glycanase 1 (AtPNG1)	S1,2 (07)	Archer
Gma.7719.1.A1_at	TIF3H1, translation initiation factor	S1 (07);S1,2,3 (08)	Archer
GmaAffx.16360.1.S1_at	Glyoxylate/hydroxypyruvate reductase	S2 (07);S1,2,3 (08)	Archer
GmaAffx.20624.1.S1_at	PUB14, ubiquitin-protein ligase	S1 (07,08)	Archer
GmaAffx.93183.1.S1_at	S-adenosylmethionine decarboxylase (SAMDC)	S1,2 (07);S1,2,3 (08)	Archer

and other species [13]. For example, Settlage et al. found that DGAT activity was positively correlated with the oil content of mature seeds and suggested that DGAT activity may be "an indicator of coordinated genetic expression of gene-products in the entire glycerolipid synthetic pathway" [13]. In addition, expression of a fungal DGAT gene in transgenic soybean plants increased seed oil 1.5% by weight [31]. Interestingly, in our study the soybean cultivar with slightly higher oil content (Archer) also has 3.2-fold higher transcript levels (averaged across the three developmental stages and the two years assayed) for this gene than does Minsoy (Supplementary file 6). Further analyses of GmaAffx.37987.1.S1_at and the other differentially expressed genes and their expression patterns will be necessary to determine which of them are most likely to affect seed yield or composition, or to play more generalized roles in seed development or metabolism.

3. Conclusions

This manuscript reports the identification of genes that are expressed at significantly different levels in developing seeds from two soybean cultivars, Minsoy and Archer, that differ with respect to seed yield and size and also exhibit minor differences in seed composition. Further study of these genes and their regulation should prove useful to researchers studying the basic biology of seed development and may provide new insights into the genetic mechanisms that regulate soybean seed yield, size and composition.

4. Materials and methods

4.1. Plant materials

The Minsoy and Archer soybean cultivars were used for these experiments. A recombinant inbred line (RIL) population was constructed from these two cultivars and QTL associated with seed protein, oil, and yield were identified in previous research [9,18]. Two of these RILs (Line 6049–273 and Line 6049–32) that are similar in maturity but differ in yield were also used in these experiments. The Minsoy and Archer parental lines were grown in St. Paul, Minnesota during the summers of 2007 and 2008. The two RILs were also grown in 2007. In 2007, each line was planted as a single row. In 2008, a randomized complete block (RCB) design was used, with the field being divided into three sections (blocks) that were each sub-divided into four sub-sections. Minsoy and Archer seeds were each planted in two random sub-sections of each block. Seeds were planted in the three blocks, which represent the three replicates used in these experiments, one to two weeks apart. The plantings were staggered as collecting developing seeds was quite time consuming. Sowing the replicates one to two weeks apart thus allowed collection of the seeds from each replicate at approximately the same time of day and at the same age. Seeds were harvested at three developmental stages, namely, seed length = 2 mm, 3.5 mm, and 5–6 mm, which correspond approximately to soybean reproductive stages R4, R5 and early R6, respectively. In 2007 three independent samples were collected for each line and developmental stage. For each sample, seeds were collected from multiple plants and then pooled. In 2008, two seed samples (one from each row) were collected for each line at each stage within each replicate. For each sample, seeds were collected from multiple plants. The pairs of seed samples were then pooled. Thus, three sets of independent tissue samples were collected for Minsoy and Archer for each of three developmental stages in both 2007 and 2008. Similar sets of samples were collected in parallel for the two RILs in 2007. All tissue collections were done during the same time of the day (from 10 am to 2 pm) to minimize diurnal effects. Collected seeds were placed immediately in dry ice and then transferred to −80 °C until RNA extraction.

4.2. RNA isolation and microarray hybridization

Seed tissues were ground to a fine powder in liquid nitrogen using a mortar and pestle and total RNA was extracted according to the protocol of the Sigma Spectrum™ Plant Total RNA Kit (Sigma-Aldrich, St. Louis, MO, USA). RNA purity and quality were checked using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). For expression profiling, cRNA synthesis and labeling was conducted according to the Affymetrix GeneChip Expression Analysis Technical Manual. The Affymetrix GeneChip® Soybean Genome Arrays were hybridized, washed, scanned, and checked for quality according to the manufacturer's protocols [32].

4.3. Analysis of microarray data

The Affymetrix Soybean GeneChip® contains a total of 61,170 probe sets, of which 37,744 are soybean probe sets. The remainders are probe sets for two soybean pathogens [33]. The 37,744 probe sets specific to

Table 6
Probe sets potentially involved in lipid metabolism. Probe sets predicted to be involved in lipid metabolism and that exhibit significant differences in transcript levels between Minsoy and Archer were identified. Gene function was predicted both on the basis of SoyBase Affymetrix ™ GeneChip® Soybean Genome Array Annotation and the *Arabidopsis* lipid gene database [14] (http://www.plantbiology.msu.edu/lipids/genesurvey/index.htm). See Table 3 legend for additional information.

AffyID	Gene description	Stage (year)	Higher expression
Gma.13043.1.S1_at	*Arabidopsis thaliana* carboxyesterase 20 (AtCXE20)	S1,2,3 (07);S3 (08)	Minsoy
Gma.13216.2.S1_at[a]	Ceramidase family protein	S1,2,3 (07,08)	Archer
Gma.4488.1.A1_s_at[a]	Ceramidase family protein	S1,2,3 (07,08)	Archer
Gma.6802.1.A1_at	GLIP5, carboxylesterase/lipase	S1,3 (07);S1,2,3 (08)	Minsoy
GmaAffx.18905.1.S1_at	Esterase/lipase/thioesterase	S1,2 (07,08)	Minsoy
GmaAffx.37987.1.S1_at	Diacylglycerol acyltransferase family	S1 (07,08)	Archer
GmaAffx.6790.1.S1_at	Epoxide hydrolase, putative	S1,2,3 (07);S1 (08)	Minsoy

[a] These two probe sets may represent the same gene as they both show the highest match to the same soybean gene model.

the soybean genome were extracted. The raw fluorescence data in CEL files generated by the Affymetrix software were then preprocessed for quality assessment using the R programming environment and Bioconductor software according to the methods of Alvord et al. [26].

Standard quality control analyses were performed to check the quality of the data obtained from each of the microarrays. The distribution of the fluorescence signal intensity across the probes was checked for each of the 54 arrays. All were found to exhibit similar distributions (Supplementary File 1). The affyPLM package was used to compare the log-transformed expression levels of each gene in each array to the median value of the gene across all of the arrays. All were found to be close to zero (Supplementary File 2). The affyPLM package was used to calculate the normalized unscaled standard error (i.e. to compare the distribution of the standard error for each gene in each array with the median value of the standard errors for the gene across all of the arrays). All were found to be near one (Supplementary file 3).

After preprocessing, a final expression value in a $\log_2$ scale was obtained for each probe set on each array using the Guanine Cytosine Robust Multi-Array analysis (GCRMA) function in the gcrma package [34]. Probe sets with average expression values higher than 4.0 were deemed expressed in soybean seeds. Accordingly, probe sets with average expression values smaller than 4.0 were excluded from the analysis as they are likely to represent experimental noise. The lmFit function (limma package) [35] was used to fit a linear model to identify differentially expressed genes between the two parental lines and the two RILs at each stage. The empirical Bayes (eBayes) method was used to adjust variances, and Benjamini-Hochberg's False Discovery Rate (FDR) was calculated for multiple test correction [36]. Probe sets with an adjusted p-value (q-value) ≤ 0.05 and a fold change ≥ 2 were considered differentially expressed between the two lines. The 2007 and 2008 data were analyzed separately and differentially expressed probe sets were identified for each developmental stage; the common probe sets between the two years for each stage were then identified and used for subsequent analysis. All microarray data and expression values after GCRMA were submitted to Gene Expression Omnibus (GEO) at NCBI http://www.ncbi.nlm.nih.gov/geo under accession GSE21598.

4.4. Functional categorization of differentially expressed probe sets

Annotations were extracted primarily from two sources. First, annotations for all transcripts were downloaded from the SoyBase Affymetrix™ GeneChip® Soybean Genome Array Annotation [1], where Gene Ontology (GO) terms were assigned to each Affymetrix Soybean GeneChip probe set using the GO terms associated with the best *Arabidopsis* sequence match. Next, probe sets with no clear annotations from the annotations mentioned above were searched against UniProt, the *Arabidopsis* genome, and the *Medicago truncatula* genome using HarvEST:SoyChip (version 1.08) [37] software. The best blastX hits with matches to 11 perfect probes and an e value $<10^{-6}$ were extracted by the software. Finally, the probe sets for which *Arabidopsis* gene models were identified were further classified. For this classification, the corresponding *Arabidopsis* gene models were searched against the *Arabidopsis* lipid data base [14] to identify genes potentially involved in lipid metabolism. The percentage of probe sets assigned to each GOslim category was calculated as: number of probe sets associated with the GOslim category/total number of probe sets for which annotations were available for this ontology X 100. As some *Arabidopsis* proteins have been assigned to more than one category, the sum of the percentages assigned to each category is greater than 100%. Fisher's exact test was used to test for over-representation or under-representation of each GOslim-defined biological process or molecular function in the differentially expressed probe sets compared with the remaining probe sets that were expressed in soybean seeds [38]. A GOslim term was judged as significantly overrepresented or underrepresented when the p-value <0.1.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.08.005.

Transparency document

The Transparency document associated with this article can be found, in the online version.

Acknowledgments

We thank the University of Minnesota Discovery Grant program, the Minnesota Soybean Research & Promotion Council (5-08N, 15-09C, 667307-17-10C and 8-11C), the University of Minnesota Initiative for Renewable Energy and the Environment (RM-0001-09) and the Consortium for Plant Biotechnology Research (GO12026-304/DE-FG36-02GO12026-008) for financial support of this project. None of these sponsors played any role in the design of the experiments, the collection, analysis or interpretation of the data, or the writing or submission of this manuscript. We thank Mr. Phil Schaus for aiding in planting and caring for the soybean lines used in these experiments. We also thank the BioMedical Genomics Center-MicroArray Facility, University of Minnesota for conducting the Affymetrix GeneChip hybridization experiments and the University of Minnesota Supercomputing Institute for assisting in data analysis. We thank Drs. Steve Wanamaker and Timothy J. Close for helping with HarVEST:SoyChip software and Dr. Michelle Graham for providing the Affymetrix Soychip annotations.

References

[1] SoyBase Affymetrix &trade GeneChip ® Soybean Genome Array Annotation Version 2 Page [http://soybase.org/AffyChip/index.php]
[2] Soybean Oil and Biodiesel Usage Projections & Balance Sheet [http://www.extension.iastate.edu/agdm/crops/outlook/biodieselbalancesheet.pdf]
[3] R.F. Wilson, Soybean: Market Driven Research Needs. in: G. Stacey (Ed.), Plant Genetics and Genomics: Crops and Models. Volume 2, 1st editionSpringer, New York 2008, pp. 3–15.
[4] Y.D. Daveby, P. Aman, Chemical composition of certain dehulled seeds and their hulls with special reference to carbohydrates. Swed. J. Agric. Res. 23 (1993) 133–139.
[5] W.K. Zhang, Y.J. Wang, G.Z. Luo, J.S. Zhang, C.Y. He, X.L. Wu, J.Y. Gai, S.Y. Chen, QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. Theor. Appl. Genet. 108 (2004) 1131–1139.
[6] D.L. Hyten, V.R. Pantalone, C.E. Sams, A.M. Saxton, D. Landau-Ellis, T.R. Stefaniak, M.E. Schmidt, Seed quality QTL in a prominent soybean population. Theor. Appl. Genet. 109 (2004) 552–561.
[7] E.C. Brummer, G.L. Graef, J.H. Orf, J.R. Wilcox, R.C. Shoemaker, Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci. 37 (1997) 370–378.
[8] S.K. Stombaugh, J.H. Orf, H.G. Jung, K. Chase, K.G. Lark, D.A. Somers, Quantitative trait loci associated with cell wall polysaccharides in soybean seeds. Crop Sci. 44 (2004) 2101–2106.
[9] J.H. Orf, K. Chase, T. Jarvik, L.M. Mansur, P.B. Cregan, F.R. Adler, K.G. Lark, Genetics of soybean agronomic traits. I. Comparison of three related recombinant inbred populations. Crop Sci. 39 (1999) 1642–1651.
[10] B.W. Diers, P. Keim, W.R. Fehr, R.C. Shoemaker, RFLP analysis of soybean seed protein and oil content. Theor. Appl. Genet. 83 (1992) 608–612.
[11] M.J. Hills, Control of storage-product synthesis in seeds. Curr. Opin. Plant Biol. 7 (2004) 302–308.
[12] S.C. Lung, R.J. Weselake, Diacylglycerol acyltransferase: a key mediator of plant triacylglycerol synthesis. Lipids 41 (2006) 1073–1088.
[13] S.B. Settlage, P. Kwanyuen, R.F. Wilson, Relation between diacylglycerol acyltransferase activity and oil concentration in soybean. J. Am. Oil Chem. Soc. 75 (1998) 775–781.
[14] F. Beisson, A.J.K. Koo, S. Ruuska, J. Schwender, M. Pollard, J.J. Thelen, T. Paddock, J.J. Salas, L. Savage, A. Milcamps, V.B. Mhaske, Y. Cho, J.B. Ohlrogge, Arabidopsis genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. Plant Physiol. 132 (2003) 681–697.
[15] H.B. Krishnan, S.S. Natarajan, A.A. Mahmoud, R.L. Nelson, Identification of glycinin and beta-conglycinin subunits that contribute to the increased protein content of high-protein soybean lines. J. Agr. Food Chem. 55 (2007) 1839–1845.
[16] A.R. Slabas, J.W. Simon, A.P. Brown, Biosynthesis and regulation of fatty acids and triglycerides in oil seed rape. Current status and future trends. Eur. J. Lipid Sci. Tech. 103 (2001) 455–466.
[17] J. Yuan, V.N. Njiti, K. Meksem, M.J. Iqbal, K. Triwitayakorn, M.A. Kassem, G.T. Davis, M.E. Schmidt, D.A. Lightfoot, Quantitative trait loci in two soybean recombinant inbred line populations segregating for yield and disease resistance. Crop Sci. 42 (2002) 271–277.

[18] J.H. Orf, K. Chase, F.R. Adler, L.M. Mansur, K.G. Lark, Genetics of soybean agronomic traits: II. Interactions between yield quantitative trait loci in soybean. Crop Sci. 39 (1999) 1652–1657.
[19] D. Wang, G.L. Graef, A.M. Procopiuk, B.W. Diers, Identification of putative QTL that underlie yield in interspecific soybean backcross populations. Theor. Appl. Genet. 108 (2004) 458–467.
[20] W.M. Breene, S. Lin, L. Hardman, J. Orf, Protein and oil content of soybeans from different geographic locations. J. Am. Oil Chem. Soc. 65 (1988) 1927–1931.
[21] M.U. Haq, A.P. Mallarino, Response of soybean grain oil and protein concentrations to foliar and soil fertilization. Agron. J. 97 (2005) 910–918.
[22] J.M.G. Thomas, K.J. Boote, L.H. Allen, M. Gallo-Meagher, J.M. Davis, Elevated temperature and carbon dioxide effects on soybean seed composition and transcript abundance. Crop Sci. 43 (2003) 1548–1557.
[23] S.J. Britz, J.F. Cavins, Spectral quality during pod development modulates soybean seed fatty-acid desaturation. Plant Cell Environ. 16 (1993) 719–725.
[24] L.R. Gibson, R.E. Mullen, Soybean seed composition under high day and night growth temperatures. J. Am. Oil Chem. Soc. 73 (1996) 733–737.
[25] D.A. McWilliams, D.R. Berglund, G.J. Endres, Soybean growth and management quick guide. North Dakota State University Extension Service, 1999 A-1174.
[26] W.G. Alvord, J.A. Roayaei, O.A. Quiñones, K.T. Schneider, A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R. Brief. Bioinform. 8 (2007) 415–431.
[27] I. Roxrud, S.E. Lid, J.C. Fletcher, E.D.L. Schmidt, H.G. Opsahl-Sorteberg, GASA4, one of the 14-member Arabidopsis GASA family of small polypeptides, regulates flowering and seed development. Plant Cell Physiol. 48 (2007) 471–483.
[28] G. Schindelman, A. Morikami, J. Jung, T.I. Baskin, N.C. Carpita, P. Derbyshire, M.C. McCann, P.N. Benfey, COBRA encodes a putative GPI-anchored protein, which is polarly localized and necessary for oriented cell expansion in *Arabidopsis*. Genes Dev. 15 (2001) 1115–1127.
[29] F. Roudier, A.G. Fernandez, M. Fujita, R. Himmelspach, G.H.H. Borner, G. Schindelman, S. Song, T.I. Baskin, P. Dupree, G.O. Wasteneys, P.N. Benfey, COBRA, an Arabidopsis extracellular glycosyl-phosphatidyl inositol-anchored protein, specifically controls highly anisotropic expansion through its involvement in cellulose microfibril orientation. Plant Cell 17 (2005) 1749–1763.
[30] M.E. Nielsen, F. Lok, H.B. Nielsen, Distinct developmental defense activations in barley embryos identified by transcriptome profiling. Plant Mol. Biol. 61 (2006) 589–601.
[31] K. Lardizabal, R. Effertz, C. Levering, J. Mai, M.C. Pedroso, T. Jury, E. Aasen, K. Gruys, K. Bennett, Expression of *Umbelopsis ramanniana DGAT2A* in seed increases oil in soybean. Plant Physiol. 148 (2008) 89–96.
[32] GeneChip Expression Analysis Technical Manual [http://media.affymetrix.com/support/downloads/manuals/expression_analysis_technical_manual.pdf]
[33] GeneChip Soybean Genome Array [http://www.affymetrix.com/support/technical/datasheets/soybean_datasheet.pdf]
[34] Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, F. Spencer, A model-based background adjustment for oligonucleotide expression arrays. J. Am. Stat. Assoc. 99 (2004) 909–917.
[35] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3 (2004) (Article 3).
[36] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc., Ser B 57 (1995) 289–300.
[37] HarvEST:SoyChip [http://harvest.ucr.edu/]
[38] S. Siegel, Nonparametric Statistics for the Behavioral Sciences. McGraw Publishers, New York, 1956.