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.