Genetics and Molecular Biology, 35, 1 (suppl), 322-330 (2012)

www.sbg.org.br

Research Article

Cell wall, lignin and fatty acid-related transcriptome in soybean: Achieving gene expression patterns for bioenergy legume

Maria Clara Pestana-Calsa, Cinthya Mirella Pacheco, Renata Cruz de Castro, Renata Rodrigues de Almeida, Nayara Patrícia Vieira de Lira and Tercilio Calsa Junior

Laboratório de Genômica e Proteômica de Plantas, Departamento de Genética, Centro de Ciências Biológicas, Universidade Federal de Pernambuco, Recife, PE, Brazil.

Abstract

Increasing efforts to preserve environmental resources have included the development of more efficient technologies to produce energy from renewable sources such as plant biomass, notably through biofuels and cellulosic residues. The relevance of the soybean industry is due mostly to oil and protein production which, although interdependent, results from coordinated gene expression in primary metabolism. Concerning biomass and biodiesel, a comprehensive analysis of gene regulation associated with cell wall components (as polysaccharides and lignin) and fatty acid metabolism may be very useful for finding new strategies in soybean breeding for the expanding bioenergy industry. Searching the Genosoja transcriptional database for enzymes and proteins directly involved in cell wall, lignin and fatty acid metabolism provides gene expression datasets with frequency distribution and specific regulation that is shared among several cultivars and organs, and also in response to different biotic/abiotic stress treatments. These results may be useful as a starting point to depict the Genosoja database regarding gene expression directly associated with potential applications of soybean biomass and/or residues for bioenergy-producing technologies.

Key words: Glycine max, Genosoja platform, tissue-specific expression.

Introduction

Interest has grown in recent years for biodiesel fuel based on animal fat or plant oils due to its relative compatibility with automotive diesel engines. There are many plant species whose derived oils have been used for biodiesel production, such as sunflower (Pessoa *et al.*, 2010), peanut (Perez *et al.*, 2010), castor bean (Godoy *et al.*, 2011), physic nut (Kumar and Sharma, 2008) and soybean (Pessoa *et al.*, 2010), among others. Soybean (*Glycine max* (L.) Merrill) is one of the world's main protein and oil sources, representing about 30% of the plant oil used for food, feed and industry. Recently, soybean has also been used for biodiesel production. Plant oils used for such application are basically composed of triacylglycerides, glycerol esthers and fatty acids. The term mono- or diglyceride refers to the number of acids, and in soybean oil the predominant one is oleic acid (Pessoa *et al.*, 2010).

Ethanol production may derive from degradation of cell wall associated with cellulolytic processes on biomass. The cell wall of higher plants is the first structure that produces signaling molecules in response to biotic or abiotic stress signals. They are composed of polysaccharide, proteins and lignin, the last found in specific cell types (Cassab *et al.*, 1988). Lignin, after cellulose, is the second most abundant land polymer, essential for the structural integrity of cell wall and protection against pathogen action. In vascular plants, lignin also accounts for mechanical resistance and transport of nutrients, water and metabolites (Baucher *et al.*, 2003). Lignin is formed by amorphous, highly complex molecules, whose polymer is made of mainly aromatic phenylpropane units (Rowell *et al.*, 2005).

Soybean cultivated in association with other crops has been analyzed regarding inputs, outputs, requirement in non-renewable resources, higher energy use efficiency and cost/benefit ratio; although presenting high bioenergy output rates, soybean has been considered one of the most energy-intensive crops (Mandal *et al.*, 2002). One possible way, which has already been considered in several countries, is to increase soybean energy output through enhancing biofuel production from its cellulosic biomass (Comis, 2006; Siqueira *et al.*, 2008) or from industrial post-processed residues (Sensöz and Kaynar, 2006).

According to recently increasing demand for renewable biofuels world-wide, the viability of such an alterna-

Send correspondence to Tercilio Calsa Junior. Departamento de Genética, Centro de Ciências Biológicas, Universidade Federal de Pernambuco, Av. Prof. Moraes Rego 1235, Cidade Universitária, 50670-901 Recife, PE, Brazil. E-mail: tercilio@ufpe.br.

tive to petroleum depends on the verification of net energy gain, environmental benefits, economical competitiveness and large-scale production without reducing food supplies (Hill *et al.*, 2006). Until recent increases in petroleum prices, high production costs made biofuels unprofitable without subsidies. Biodiesel provides sufficient environmental advantages to merit subsidy. Transportation biofuels such as synfuel hydrocarbons or cellulosic ethanol, if produced from low-input biomass grown on agriculturally marginal land or from waste biomass, could provide much greater supplies and environmental benefits than food-based biofuels (Hill *et al.*, 2006).

The increasing demand for food combined with the relatively recent, but also growing, need for biofuel derived from legume species has justified scientific and technological efforts aimed at improving knowledge on legume biology, especially the symbiotic nitrogen fixation and assimilation which has been a specific target in Fabales research. Such relevance has proven evident after finishing (or nearly completing) and publishing the genome sequences of the legume species *Lotus japonicus* (Sato *et al.*, 2008), *Medicago truncatula* (barrel medic) and soybean (*Glycine max*; Schmutz *et al.*, 2010). Also of note, soybean transcriptome databases have been launched as comprehensive tools for gene expression studies on distinct cultivars and several experimental treatments, allowing *in silico* investigation of a wide range of molecular and physiological processes including growth and development as well as responses to biotic or abiotic stress factors. For example, the SoyXpress database (Cheng and Strömvik, 2008) harbors 380,095 soybean ESTs (expressed sequence tags) from conventional and transgenic cultivars, with associated gene ontology terms, metabolic pathways, SwissProt identifiers and Affymetrix microarray gene expression data from several experiments. More recently, a transcriptional atlas of soybean was published, where 69,145 putative genes are predicted, 46,430 of which with high confidence (Libault *et al.*, 2010); this database comprises cDNA-derived Illumina Solexa sequences from 14 conditions or tissues, and genes identified *in silico* with significant differential expression between libraries were experimentally validated via RT-qPCR.

In Brazil, landmark research has been carried out in the Brazilian Soybean Genome Consortium (Genosoja Project), established in 2008 and involving several research groups in soybean genomics, with the aim to investigate structural and functional aspects of soybean gene expression, under relevant agricultural stresses, mainly Asian rust, nematodes, drought and nitrogen fixation (Abdelnoor *et al.*, 2009). The Genosoja database includes public access data from the soybean genome (Schmutz *et al.*, 2010), so it constitutes a comprehensive molecular genetics base for *in silico* and physiological analyses whose applications may reach basic research and also strategies to circumvent agronomical constrains in the Brazilian soybean industry (Nascimento *et al.*, 2012).

Brazils sugarcane bioethanol produced in 2008 represented ca 37.3% of the worldwide production of 65.6 billion liters (Renewable Fuels Association). Although small in scale by Brazil standards, production of bioethanol from biomass sources other than sugarcane is in agreement with worldwide interest in bioethanol as an energy source and, in this scenario, potentially useful candidates include agroindustry by-products, such as soybean molasses and residual field biomass (Siqueira *et al.*, 2008). The Brazilian production of soybean ranks second worldwide, representing about 28% of global production (Soystat, 2009). Molasses is a co-product of protein-concentrate meal obtained after de-oiled soybean processing; quantitatively, one ton of soybean yields about 716 kg of de-oiled meal, from which are extracted 522 kg of protein concentrate (usually for the animal feed industry) and 190.8 kg molasses. This, if submitted to fermentation with specific *Saccharomyces* strains, will result in 18.4 kg ethanol plus 533.6 kg vinasse (80.5% moisture; Siqueira *et al.*, 2008; Figure 1).

Indeed, the relatively well-established production of biodiesel from soybean oil can be summed up by bioethanol obtained from the soybean industry waste and/or cellulosic biomass. In this direction, soybean molecular biology and functional genomics may be applied to aid in breeding programs aiming to increase not only oil content

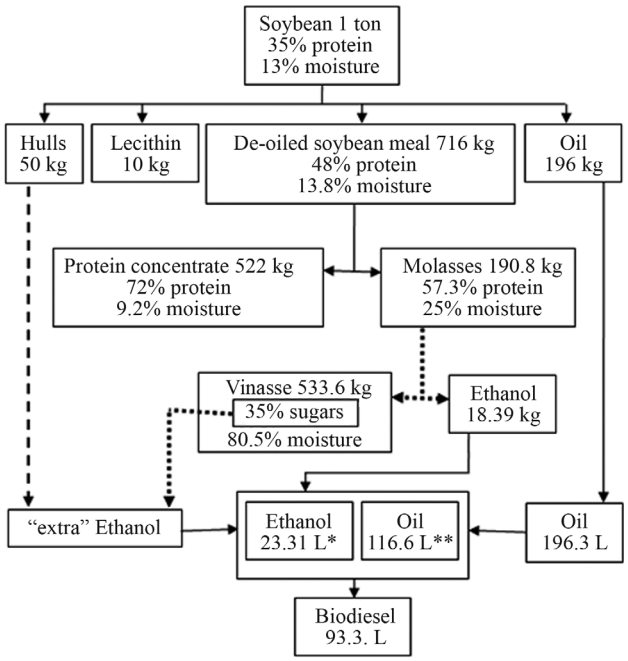

Figure 1 - Schematic flow-chart for biofuel production from soybean, modified from Siqueira *et al.* (2008). Final step amounts can be adjusted. Dotted lines refer to yeast-based fermentation processes; dashed line refers to cellulose hydrolytic process, in both cases aiming to increase ethanol yield from residual biomass. *Amount of ethanol obtained from molasses fermentation. **Amount of oil needed for biodiesel synthesis reaction with 23.3 L ethanol. "Extra" ethanol is not quantified here, although cellulolytic production from hulls is possible (Schirmer-Michel *et al.*, 2007), as well as from vinasse sugars redirected to conventional (hexoses) or specific (pentoses) fermentation.

in seeds, but also other polymeric components that may support technological strategies towards producing biofuels or bioenergy, such as cell wall polysaccharides and lignin.

Thus, the use of the Genosoja database for *in silico* predictions on soybean transcriptome related to enzymes and proteins involved in cell wall, lignin and fatty acids metabolism is expected to result in valuable information over the regulation of the coding gene expression in different cultivars, organs and also under stress factor treatments. In this work, gene and variety-specific transcriptional activity was found to be correlated with potential targets regarding soybean breeding for bioenergy.

Material and Methods

Genosoja, the Soybean Genome Project database (Nascimento *et al.*, 2012) was searched by keyword in order to identify sequences related to genes encoding products whose putative annotation is associated with fatty acids, cell wall polysaccharides (mostly cellulose or pectin) or lignin metabolism. Lignin was considered separately, despite being a cell wall component, because its structural features are different from polysaccharides. To be included in the analysis platform, ESTs assembled in contigs and singlets (unigenes), components of gene models also had to present Gene Ontology (GO) terms associated with fatty acids, cell wall polysaccharides or lignin metabolism. Annotation by BlastX against public protein databases (NR) and GO terms from the Genosoja database were considered for *in silico* analyses. Retrieved EST sequences were assessed and organized considering the standard Genosoja nomenclature for reads and cDNA libraries, derived from buds, cotyledons, endosperm, epicotyls, flowers, leaves, roots, shoots, seeds, pods or stems (Nascimento *et al.*, 2012). ESTs with extremely similar or identical putative annotations were counted, and their frequency was normalized for total number of ESTs in each corresponding library (expressed as a percentage), or for total number of ESTs derived from the same cultivar.

The normalized frequencies of contrasting libraries were compared to infer *in silico* gene expression variation among cDNA libraries and varieties, according to the Genosoja database (Nascimento *et al.*, 2012). Transcripts most directly annotated as involved in metabolism of fatty acids, lignin and cell wall main polysaccharides were separately analyzed for expression pattern by hierarchical clustering using the EPClust/EBI expression profiler (Brazma and Vilo, 2000) online tools, following standard parameters. Fatty acids and lignin data matrices were directly used in correlation measure-based (uncentered) and UPGMA (average) distances, while cell wall-related data matrix was $\log_2$ normalized prior to clustering. Statistical significance of pairwise comparisons was assessed, when necessary, from non-normalized transcript frequencies in different libraries, based on the Audic and Claverie (1997) test, with significant p-value ≤ 0.05.

Results

A keyword-based search on the NR/GO annotated Genosoja EST database (ESTs + GeneModels) resulted in a total subset of 4,094 transcript reads strictly assigned to bioenergy gene products. Among these, 1,934 were putatively related to fatty acids., 2,111 to cell wall polysaccharides, with 827 related to cellulose, 1,202 to pectin and 82 not directly associated with pectin or cellulose (NPC), and finally, 49 reads were associated with lignin terms. A normalized distribution of such ESTs in soybean cDNA libraries is summarized in Table S1 (Supplementary material), grouped by reproductive or vegetative organ-derived libraries and described according to tissue or treatment type.

General detection of fatty acids or cell-wall related transcripts in different soybean organs is shown in Figure 2, for distinct reproductive and vegetative ones. The frequency of reads putatively associated with fatty acids and cell wall components identified in cDNA libraries constructed from vegetative or reproductive soybean organs samples was comparatively analyzed, resulting in a more comprehensive picture of transcriptional response associated with different growth or developmental stages and stressing factors (Figure 3).

The transcript frequencies from specific genes, significantly annotated as coding for proteins directly involved in fatty acids, cell wall polysaccharides or lignin metabolism, were normalized as a percentage of the total number of reads in the library, and used for hierarchical clustering of such gene expression patterns (Figure 4), in an attempt to identify possible co-regulation. In a similar investigative approach, hierarchical clustering of mostly the same genes was then performed considering the soybean commercial varieties used for cDNA library construction (Figure 5); here, differences in the expression of these genes were potentially associated with genotypes.

Discussion

Out of the total number of soybean ESTs putatively found as related to metabolism regarding biofuels/bioenergy, 47.2% are involved in fatty acid biosynthesis/catabolism (FA), as expected for an oil-producing legume. The remaining ESTs found were assigned to cell wall components, with pectin-related (CWP) ones being are more abundant (29.3%) than the cellulose-related (CWC), these comprising 20.2%. Lignin-associated (CWL) transcript reads accounted for only 1.2%, which is proportionally less than the amount of reads annotated as linked to other cell wall polysaccharide biopolymers (NPCL) which accounted for 2.0%.

A direct analysis of the distribution of the so-called "bioenergy" genes dataset across the cDNA libraries re-

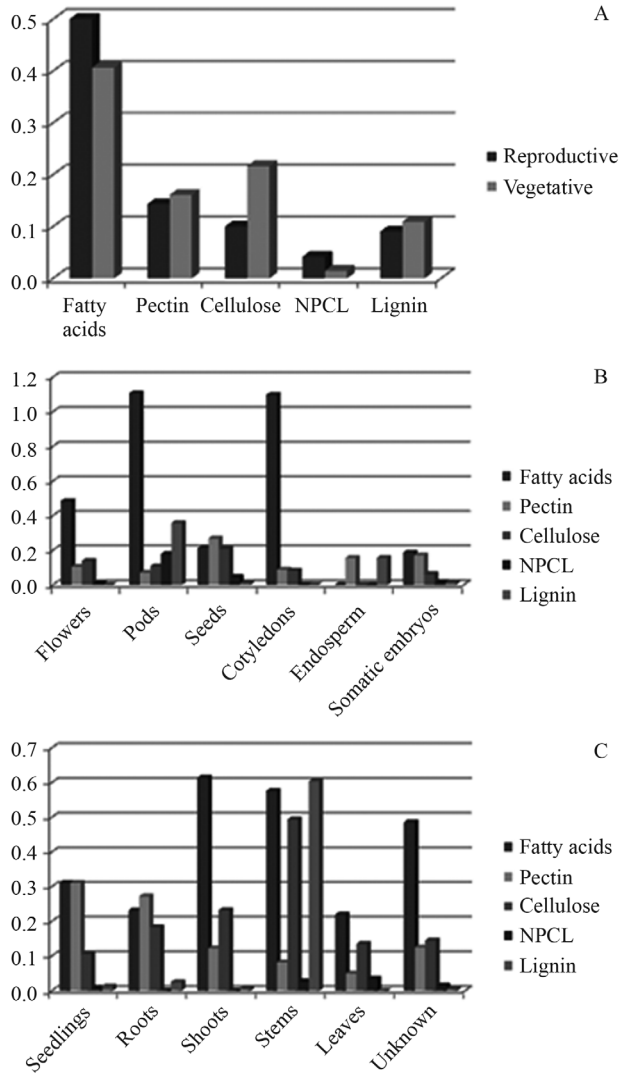

Figure 2 - Frequency distribution of reads putatively associated with fatty acids and cell wall components identified in reproductive or vegetative soybean organs (A). Detailed distribution in specific reproductive (B) and vegetative (C) organs is presented. NPCL: not directly associated with cellulose or pectin transcript. Y-axis in percentage.

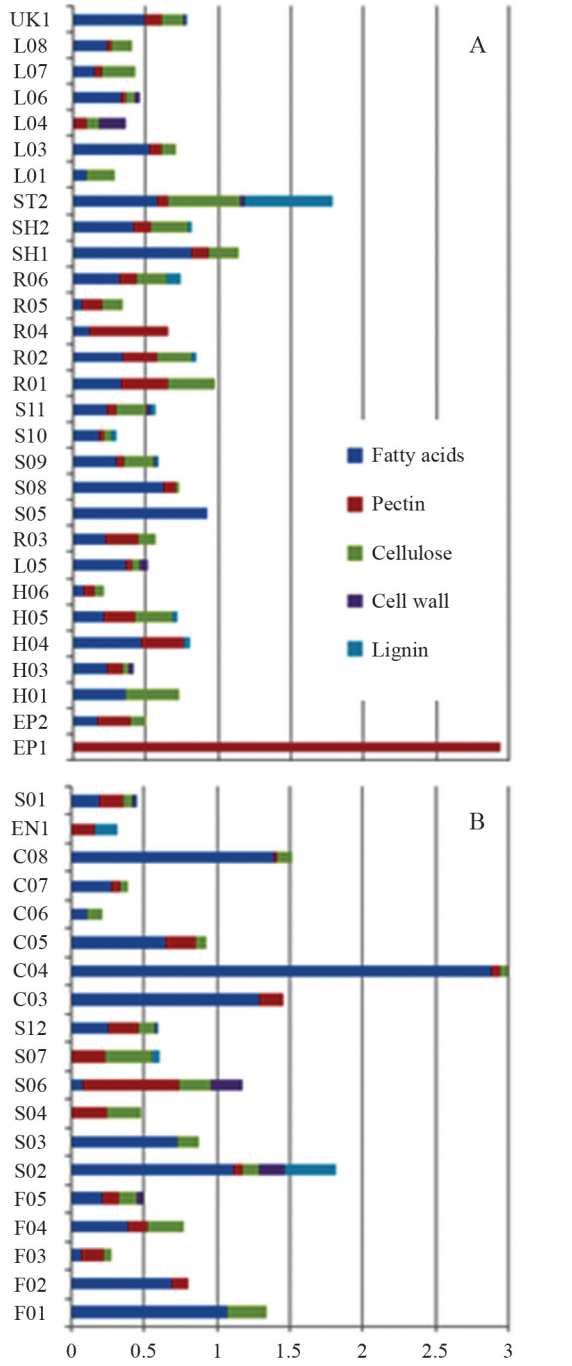

Figure 3 - Frequency distribution of reads putatively associated with fatty acids and cell wall components identified in cDNA libraries derived from vegetative (A) or reproductive (B) soybean organs. The cell wall category is represented by components not directly associated with cellulose, pectin or lignin related transcripts. X-axis in percentage. Library identification is the same as in Table S1.

vealed that transcription related to FA metabolism is most represented in cotyledon and pod libraries (Table S1 and Figure 3), as also expected for a lipid storage-specific organ in this legume. This was the case of libraries C04, C08 and C03, constructed from immature or very young cotyledons, where FA-related transcripts reached up to 2.8% of the total library (Figure 3B).

Within cell wall-related transcripts, the CWP fraction presented a higher frequency in epicotyls of 2-week-old seedling (almost 3%) and seed coat (ca 0.7%) libraries, respectively EP1 and S06 (Figure 3). On the other hand, CWC associated reads were more frequent in stems (0.5%) and root nodules (0.3%), corresponding to ST2 and R01 libraries. The reads putatively assigned to NPCL comprised on average no more than 0.2% of each library: observed maximum frequency for NPCL reads was 0.22% (in seed coats, S06) and 0.18% (in immature leaves, L04). The most frequently transcribed CWL sequences were found in stems (0.6%, ST2) and pods (0.36%, S02). Such general and putative transcriptional picture indicates, although theoretical and validation-dependent, some insights concerning gene activity were directly associated with the main components

of soybean biomass , other than lipids and proteins. For example, genes involved in pectin and other hemicelluloses metabolism seem to be more active in epicotyl and seed coat, which is of great interest since the hulls are a soybean industry residue potentially useful for novel strategies in biofuel production (Wu *et al.*, 2010). In addition, leaves have been shown to be a main site for expression for NPCL-related genes, which shall be considered when leaf material represents a higher proportion of plant biomass used in bioenergy production, but in its composition different hemicelluloses and other cell wall components are significant.

Stem-specific gene expression showed higher correlation to cellulose/lignin biosynthesis and rearrangement than in any other soybean organ, which is quite expected according to its structural support function, and very important to any further energetic usage of soybean stem-rich biomass. Noteworthy, root nodules also had high transcription level associated with CWC, which may point to unknown molecular mechanisms underlying the activities of cellulose metabolism that may be essential to nodule formation, including structural changes in host root cell wall (Sherrier *et al.*, 2005) and proper functioning, with effects on nitrogen fixation and consequently in general plant/biomass growth. Indeed, the observed transcriptional preference for CWL sequences in pods suggests that lignin may be even more structurally relevant to the fruit, perhaps in transpiration avoidance and water content maintenance (Anterola and Lewis, 2002).

In more detail, the reproductive organs of soybean presented higher transcriptional activity associated with FA and NPCL, while vegetative organs showed more frequent transcription of genes putatively annotated to CW(P/C/L) (Figure 2A). Among reproduction structures, the flowers, pods, seeds and cotyledons were the site of highest expression of FA genes, and a more intense expression of CW(P/C) and CWL genes was observed in seeds and pods, respectively (Figure 2B), which is relatively expected for lipid-storage organs. Regarding structures in developmental vegetation, FA gene expression did not presente marked differences in frequency levels compared to cell wall-related genes, except in shoots (Figure 2C). In fact, n most vegetative organs, FA gene transcription was in similar ranges as that observed for CW(P/C/L), and in stems the lignin-associated transcripts level were even higher than FA genes (Figure 2C).

Results from hierarchical clustering analyses provided clues to the transcriptional regulation of few but key-step genes whose products are essential in FA and CW metabolism. As a caveat one must note that the Genosoja database is not fully comprehensive, that this work included just EST/cDNA data only, and that equivalent libraries were not constructed from all soybean genotypes used. Nevertheless, comparative analysis of gene expression patterns may help to identify co-regulated genes concerning biomass/bioenergy-related gene expression, as well as to determine similarity levels among transcriptional profiles of libraries (*i.e.*, organs, tissues, treatments), or even among soybean varieties.

For cell wall associated transcripts, only CWL presented a relatively clear profile, possibly due to the fact that only 2 genes could be assessed, as these were the only matches in the keyword search. Actually, they are considered to code for unknown soybean proteins putatively assigned to the *O*-methyltransferase domain, thus referring to the lignin biosynthetic pathway. From the Genosoja database, their clustered expression pattern in cDNA libraries was very distinct (Figure 4C), with ACU22737.1 "isoform" transcripts being found in more similar levels in distinct organs, whereas ACU21012.1 expression was higher, although not statistically significant, in etiolated hypocotyls (H05) and germinating shoots (SH2). Nevertheless, ACU22737.1 appears to be significantly more transcribed than ACU21012.1 in roots ($p = 0.0069$, R06), seeds with globular-stage embryo ($p = 0.0436$, S12) and seedlings ($p = 0.0078$, S10). Furthermore, it appears to be more, though not statistically significant, expressed in *Phytophthora sojae*-infected hypocotyls (H04). Hence, it is inferred that these lignin-associated soybean genes may have rather different regulation pathways triggered by light (Su *et al.*, 2005), germination, organ-specific and defense-related cell wall component cross-linking.

The other categories of CW genes did not present a highly informative clustering of expression patterns, mostly because of the generally low transcription level after normalization (Figure 4A,B). However, some patterns emerged, such as the higher and statistically significant expression level of the expansin coding gene in the seedling epicotyl (EP1; $0.00027 < p < 0.03195$) library compared to other libraries (Figure 4A). Expansins are known to be essential to cell wall loosening in plant structures in elongation stages (Lee *et al.*, 2003), especially in stems. Herein, the expansin gene was detected to be significantly more expressed in etiolated seedling hypocotyls than in light-grown whole seedlings (H05 *vs.* S10, $p = 0.00596$). In addition, it was proportionally more expressed when cotyledons were excluded from the seedling sample, since a significantly higher transcription was observed in seedlings minus cotyledons when compared to whole seedlings (S11 *vs.* S10, $p = 0.03136$).

Grouping of the libraries based on expression profiles similarity was not as evident for CW genes as observed for FA genes (Figure 4B). Here, the several soybean fatty acid desaturases (FADs) were very similar with respect to frequency variation across the libraries, inferring co-regulation mechanisms for this group of key enzymes in fatty acid metabolism and lipid content manipulation (Pham *et al.*, 2010). More specifically, transcripts for a putative microsomal Ω6-FAD were represented in almost every library, being significantly higher ($p \sim 10^{-9}$) in those from cot-

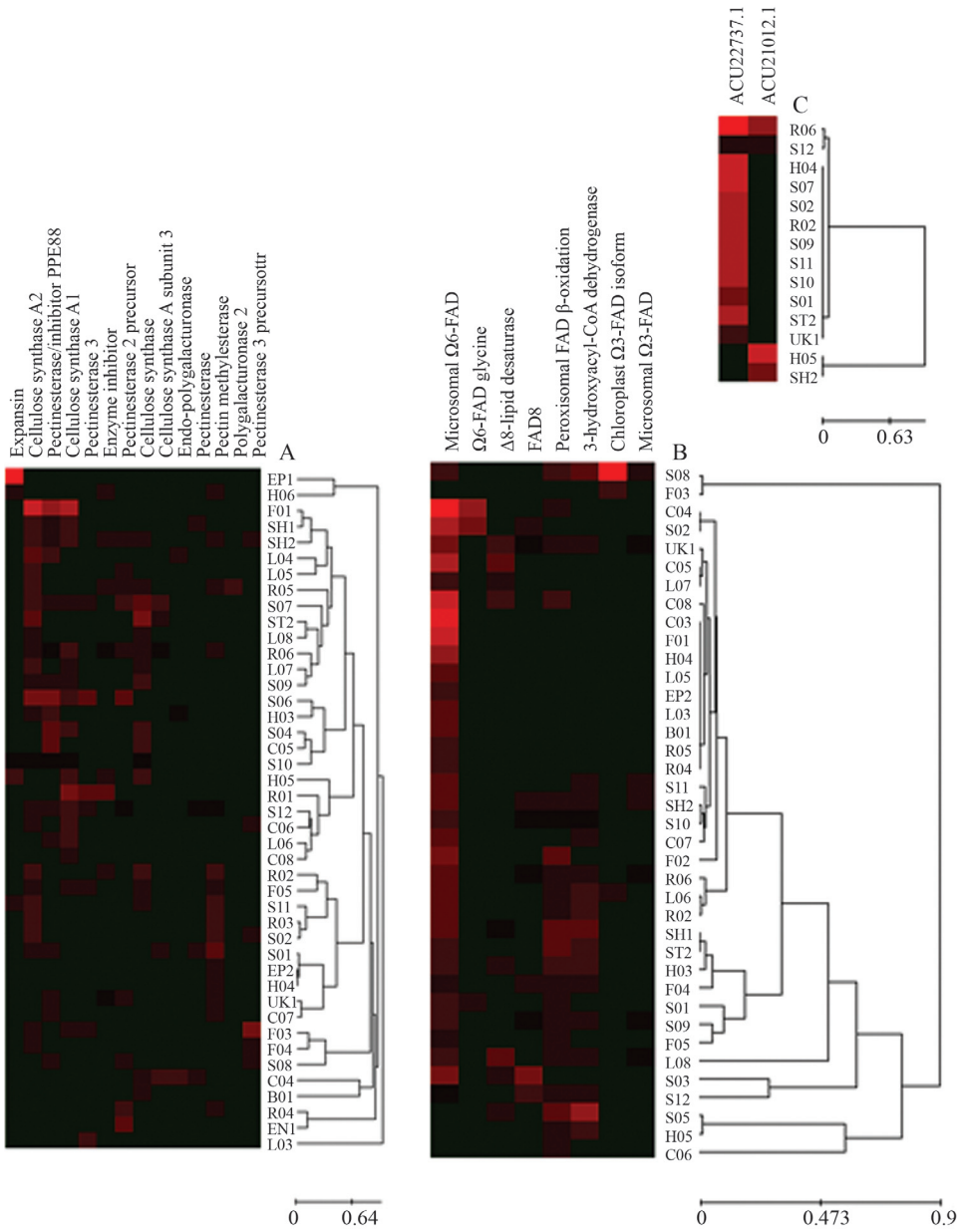

Figure 4 - Normalized transcriptional expression hierarchical clustering pattern of soybean genes associated with cell wall polysaccharides (A), fatty acids (B) and lignin (C) metabolism, in distinct cDNA libraries. Color scale ranges from black (no expression, 0) to bright red (maximum normalized expression frequency, in percentage, in A: 2.9412; B: 2.3107; C: 0.0730). Bars indicate distance by UPGMA.

yledons, seeds, roots, and floral meristem. It was the only FAD-associated transcript detected in *Phytophthora sojae*-infected hypocotyls (H04), indicating putative specific association to this patosystem. In seedlings (S09), it also was the most frequent significant FAD-associated transcript differentially detected (p = 0.0028). In contrast, transcription associated with chloroplast Ω3-FAD gene was higher in floral meristematic apices (F04, p~10^{-7}), and expression levels detected for FAD 8 were significantly and preferentially higher in seeds with globular-stage embryos (S12, p~10^{-7}).

The expression of the two putative lignin-related *O*-methyltransferases was profiled by hierarchical clustering among soybean varieties/cultivars used for library construction. This showed that ACU22737.1 was more transcribed in almost all genotypes than ACU21012.1, reaching significance for differential expression only in Asgrow A3237 (p = 0.0436). This did not apply to Corolla, PI567374, Delsoy 5710, Ogden and Minsoy x Noir RI, since in these genotypes, expression levels were practically undetected for both genes. In Williams genotypes both genes presented intermediate to low transcript levels (Figure 5B).

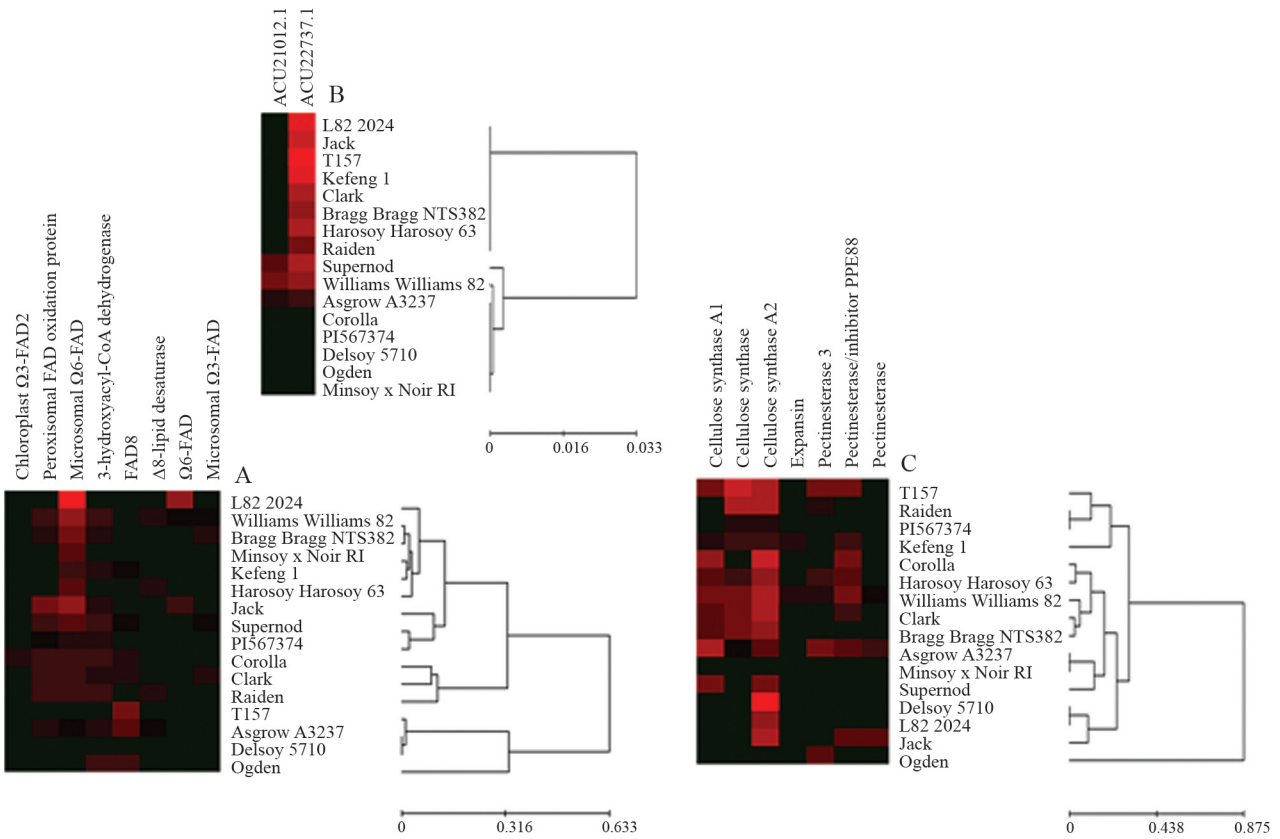

Figure 5 - Normalized transcriptional expression hierarchical clustering pattern of soybean genes associated with cell wall polysaccharides (A), fatty acids (B) and lignin (C) metabolism, in distinct varieties. Color scale ranges from black (no expression) to bright red (maximum normalized expression frequency, in percentage, in A: 0.7846; B: 0.0460; C: 0.2255). Bars indicate distance by UPGMA.

Varieties could be grouped following FA gene expression clustering. Transcripts for the microsomal Ω6-FAD were the most frequent ones in almost all genotypes, especially in L82 2024 (Figure 5A), except for Asgrow A3237, where they presented one of the lowest frequencies, and FAD 8 transcripts were the most frequent among FAD-coding genes (p~10^{-7}). Concerning CW genes, expression clustering revealed a generally simultaneous higher transcription for cellulose synthases and lower transcription for pectinesterases. Cellulose synthase-coding transcripts from *CesA1* and *CesA2* showed very distinct expression patterns in Delsoy 5710 and Jack genotypes, while in the Corolla variety both genes seemed to be more intensely transcribed. The expression difference in these genes was observed to be significant for Asgrow A3237 (p = 0.0008), Clark (p = 0.0175), Harosoy/Harosoy 63 (p = 0.0333), Williams/Williams 82 (p = 0.0020 and Raiden (p = 0.0001). This type of information, despite being specific and not comprehensive for the several varieties tested so far, should be useful for soybean breeding programs focusing on new strategies, by conventional crosses/selection and by transgeny, to obtain varieties with application focus on biomass, bioenergy and cellulosic-based biofuels production (Pham *et al.*, 2010). Further studies on linking the expression data of these genes to FA and polysaccharide composition of selected parental varieties should prove to be very helpful.

In conclusion, this work aimed at providing initial information from the Genosoja database regarding gene expression directly associated with potential uses of soybean biomass and/or waste for bioenergy-producing technologies. FA, CW polysaccharides and lignin metabolism and physiological properties *in planta* are ultimately governed by genomic features interacting with environmental factors, and are a first line of research focus towards improving soybean as feedstock for bioenergy, biodiesel, as well as for bioethanol and many other by-products and subproducts that could soon be commercially available due to intensive lignocellulosic chemistry and technologies. The soybean industry may soon benefit from investments in oil-protein-cellulose-lignin-biofuel from biomass and from conventional waste, not usually directed toward profitable and ecologically suitable destinations. Finally, the highly relevant position of Brazil as a soybean producer should further the development of this commodity as an additional bioenergy source in this country.

Acknowledgments

The authors are grateful to Eliseu Binneck, Alexandre Lima Nepomuceno, Ricardo Vilella Abdelnoor and Fran-

cismar Correa Marcelino Guimarães (EMBRAPA Soja, Londrina, PR, Brazil) for access and information interchange from the Genosoja Project and Database, as well as, to Marcelo Falsarella Carazzolle and Leandro Costa do Nascimento (Universidade Estadual de Campinas, Campinas, SP, Brazil) for help with Genosoja database managing software. Financial support; from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Universidade Federal de Pernambuco is acknowledged.

References

Abdelnoor RV, Nepomuceno AL, Barros EG, Sá MFG, Binneck E, Marcelino FC, Brommonschenkel SH, Almeida J, Benko-Iseppon AM, Schuster I, *et al.* (2009) GENOSOJA – A Brazilian Soybean Genome Consortium. Plant & Animal Genomes XVII Conference, San Diego, P034.

Anterola AM and Lewis NG (2002) Trends in lignin modification: A comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. Phytochemistry 61:221-294.

Audic S and Claverie JM (1997) The significance of digital gene expression profiles. Genome Res 7:986-995.

Baucher M, Halpin C, Petit-Conil M and Boerjan W (2003) Lignin: Genetic engineering and impact on pulping. Crit Rev Biochem Mol Biol 38:305-350.

Brazma A and Vilo J (2000) Gene expression data analysis. FEBS Lett 480:17-24.

Cassab GI, Lin JJ, Lin LS and Varner JE (1988) Ethylene effect on extensin and peroxidase distribution in the subapical region of pea epicotyls. Plant Physiol 88:522-524.

Cheng KCC and Strömvik MV (2008) SoyXpress: A database for exploring the soybean transcriptome. BMC Genomics 9:368-377.

Godoy MG, Gutarra ML, Castro AM, Machado OL and Freire DM (2011) Adding value to a toxic residue from the biodiesel industry: Production of two distinct pool of lipases from *Penicillium simplicissimum* in castor bean waste. J Ind Microbiol Biotechnol 38:945-953.

Hill J, Nelson E, Tilman D, Polasky S and Tiffany D (2006) Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. Proc Natl Acad Sci USA 103:11206-11210.

Kumar A and Sharma S (2008) An evaluation of multipurpose oil seed crop for industrial uses (*Jatropha curcas* L.): A review. Industr Crops Prod 28:1-10.

Lee DK, Ahn JH, Song SK, Choi YD and Lee JS. (2003) Expression of an expansin gene is correlated with root elongation in soybean. Plant Physiol 131:985-997.

Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G and Stacey G (2010) An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. Plant J 63:86-99.

Mandal KG, Sahab KP, Ghosha PH, Hatia KM and Bandyopadhyaya KK (2002) Bioenergy and economic analysis of soybean-based crop production systems in central India. Biomass and Bioenergy 23:337-345.

Nascimento LC, Costa GGL, Binneck E, Pereira GAG and Carazzolle MF (2012) A web-based bioinformatics interface applied to Genosoja Project: Databases and pipelines. Genet Mol Biol 35.

Pérez A, Casas A, Fernández CM, Ramos MJ and Rodríguez L (2010) Winterization of peanut biodiesel to improve the cold flow properties. Bioresour Technol 101:7375-7381.

Pessoa FL, Magalhães SP and Falcão PW (2010) Production of biodiesel via enzymatic ethanolysis of the sunflower and soybean oils: Modeling. Appl Biochem Biotechnol 161:238-244.

Pham AT, Lee JD, Shannon JG and Bilyeu KD (2010) Mutant alleles of FAD2-1A and FAD2-1B combine to produce soybeans with the high oleic acid seed oil trait. BMC Plant Biol 10:e195.

Rowell RM, Pettersen R, Han JS, Rowell JS and Tshabalala MA (2005) Cell wall chemistry. In: Rowell RM (ed) Handbook of Wood Chemistry and Wood Composites. CRC Press, Boca Raton, pp 35-72.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, *et al.* (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res 15:227-239.

Schirmer-Michel AC, Flôres SH, Hertz PF, Matos GS and Ayub MAZ (2007) Production of ethanol from soybean hull hydrolysate by osmotolerant *Candida guilliermondii* NRRL Y-2075. Bioresource Technology 99:2898-2904.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, *et al.* (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178-183.

Sensöz S and Kaynar I (2006) Bio-oil production from soybean (*Glycine max* L.); fuel properties of Bio-oil. Industr Crops Prod 23:99-105.

Sherrier DJ, Taylor GS, Silverstein KA, Gonzales MB and VandenBosch KA (2005) Accumulation of extracellular proteins bearing unique proline-rich motifs in intercellular spaces of the legume nodule parenchyma. Protoplasma 225:43-55.

Siqueira PF, Karp SG, Carvalho JC, Sturm W, Rodríguez-León JA, Tholozan J-L, Singhania RR, Pandey A and Soccol CR (2008) Production of bio-ethanol from soybean molasses by *Saccharomyces cerevisiae* at laboratory, pilot and industrial scales. Bioresour Technol 99:8156-8163.

Su G, An Z, Zhang W and Liu Y (2005) Light promotes the synthesis of lignin through the production of H_2O_2 mediated by diamine oxidases in soybean hypocotyls. J Plant Physiol 162:1297-1303.

Wu HC, Hsu SF, Luo DL, Chen SJ, Huang WD, Lur HS and Jinn TL (2010) Recovery of heat shock-triggered released apoplastic Ca^{2+} accompanied by pectin methylesterase activity is required for thermotolerance in soybean seedlings. J Exp Bot 61:2843-2852.

Internet Resources

Comis D (2006) Turning soybean plants into ethanol or particleboard. Agricultural Research, http://arsserv0.tamu.edu/is/ar/archive/nov06/soybean1106.pdf (August 7, 2011).

Lotus (*Lotus japonicas*) genome database, http://www.kazusa.or.jp/lotus (07.08.2011).

Barrel medic (*Medicago truncatula*) genome database, http://www.medicago.org/genome (August 7, 2011).
Soybean (*Glycine max*) genome database, http://www.phytozome.net/soybean (August 7, 2011).
SoyXpress database, http://soyxpress.agrenv.mcgill.ca (August 9, 2011).
Renewable Fuels Association – RFA, http://www.ethanolrfa.org/ (November 25, 2010).
Soystat – Soy bean statistics database (2009), http://www.soystats.com/2009 (November 25, 2010).
EPClust/EBI expression profiler, http://www.bioinf.ebc.ee/EP/EP/EPCLUST/index.cgi (August 10, 2011).

Supplementary Material

The following online material is available for this article:

Table S1 - Frequency distribution of reads (ESTs) from cDNA libraries.

This material is available as part of the online article from http://www.scielo.br/gmb.

Table S1. Normalized frequency distribution of reads (ESTs) from cDNA libraries derived from reproductive or vegetative organs / tissues, in percentage (%).

Organ / Tissue	Library	Total EST	Fatty acids (%)	Cell wall (%)			
				Pectin	*Cellulose*	*NPC**	*Lignin*
Reproductive							
Flower	F01	374	1.070	-	0.267	-	-
	F02	870	0.690	0.115	-	-	-
	F03	5,901	0.068	0.153	0.051	-	-
	F04	4,188	0.382	0.143	0.239	-	-
	F05	5,315	0.207	0.113	0.132	0.038	-
Pod	S02	2,804	1.106	0.071	0.107	0.178	0.357
Seed	S03	684	0.731	-	0.146	-	-
	S04	842	-	0.238	0.238	-	-
	S06	1,359	0.074	0.662	0.221	0.221	-
	S07	2,176	-	0.230	0.322	-	0.046
	S12	457,343	0.255	0.207	0.112	0.012	0.003
Cotyledon	C03	620	1.290	0.161	-	-	-
	C04	3,289	2.888	0.061	0.152	-	-
	C05	1,403	0.641	0.214	0.071	-	-
	C06	2,846	0.105	-	0.105	-	-
	C07	3,336	0.270	0.060	0.060	-	-
	C08	3,753	1.386	0.027	0.107	-	-
Endosperm	EN1	1,291	-	0.155	-	-	0.155
Somatic embryo	S01	6,462	0.186	0.170	0.062	0.015	0.015
Vegetative							
Seedling	EP1	34	-	2.941	-	-	-
	EP2	1,793	0.167	0.223	0.112	-	-
	H01	274	0.365	-	0.365	-	-
	H03	3,875	0.232	0.103	0.052	0.026	-
	H04	2,366	0.465	0.296	-	-	0.042
	H05	5,112	0.215	0.215	0.254	-	0.039
	H06	2,875	0.070	0.070	0.070	-	-

	L05	1,948	0.359	0.051	0.051	0.051	-
	R03	887	0.225	0.225	0.113	-	-
	S05	1,084	0.923	-	-	-	-
	S08	4,509	0.621	0.089	0.022	-	-
	S09	10,467	0.296	0.057	0.201	0.010	0.029
	S10	19,357	0.181	0.026	0.052	0.005	0.031
	S11	3,011	0.232	0.066	0.199	0.033	0.033
Root	R01	613	0.326	0.326	0.326	-	-
	R02	2,934	0.341	0.239	0.239	-	0.034
	R04	1,836	0.109	0.545	-	-	-
	R05	3,535	0.057	0.141	0.141	-	-
	R06	16,428	0.323	0.110	0.213	-	0.091
Shoot	SH1	2,453	0.815	0.122	0.204	-	-
	SH2	6,517	0.414	0.123	0.261	-	0.015
Stem	ST2	3,644	0.576	0.082	0.494	0.027	0.604
Leaf	L01	1,042	0.096	-	0.192	-	-
	L03	1,139	0.527	0.088	0.088	-	-
	L04	1,117	-	0.090	0.090	0.179	-
	L06	3,058	0.327	0.033	0.065	0.033	-
	L07	3,555	0.141	0.056	0.225	-	-
	L08	15,861	0.227	0.032	0.151	-	-
Unknown	UK1	19,962	0.486	0.125	0.145	0.015	0.005

**NPC: no directly associated to pectin, cellulose or lignin.* Library codes are according to the Genosoja database (Nascimento et al.; in this issue), as it follows: C03: young cotyledons of greenhouse grown plants; C04: immature cotyledons of greenhouse grown plants; C05: cotyledons of 8 days old plantlets; C06: wounded cotyledons; C07: degenerating cotyledons of 9-10 days etiolated seedlings; C08: cotyledons of 3 and 7 days; EN1: endosperm tissue in developing seeds; EP1: epicotyls of 2 weeks seedlings; EP2: seedling epicotyls; F01: floral meristem; F02: mature flowers; F03: mature flowers of field grown plants; F04: floral meristematic apices; F05: immature flowers of field grown plants; H01: hypocotyl and plumule of 3 days germinated seeds; H03: hypocotyl and plumule of germinating seeds; H04: *Phytophthora sojae*-infected hypocotyls; H05: etiolated hypocotyl tissue of 9-10 days seedlings; H06: etiolated hypocotyls; L01: senescing leaf tissue of mature greenhouse grown plants; L03: fully expanded leaves of

greenhouse grown plants; L04: immature leaves of greenhouse grown plants; L05: unexpanded leaves and shoot tips of 2 week seedlings; L06: drought stressed leaf tissue; L07: leaves from 3 weeks greenhouse grown plants; L08: leaves; R01: root nodules of greenhouse grown plants; R02: roots of 7 day old plants; R03: seedling roots; R04: roots of bulked plants; R05: roots of 8 day old plants; R06: roots; S01: somatic embryos cultured on MSD20 medium; S02: mature seed pods of greenhouse grown plants; S03: germinating seeds; S04: young seeds; S05: 18 day seedlings; S06: seed coats; S07: seed coats of greenhouse grown plants; S08: 11 day seedlings; S09: whole seedlings of greenhouse grown plants; S10: seedlings; S11: seedlings minus cotyledons; S12: seeds with globular-stage embryos; SH1: shoot 24 h post-germination; SH2: germinating shoots; ST2: stem tissue of greenhouse grown plants; UK1: unknown bulked organs and cultivars.