In-Depth Investigation of the Soybean Seed-Filling Proteome and Comparison with a Parallel Study of Rapeseed[1][W][OA]

Ganesh Kumar Agrawal*, Martin Hajduch[2], Katherine Graham, and Jay J. Thelen

Department of Biochemistry, Life Sciences Center, University of Missouri, Columbia, Missouri 65211 (G.K.A., M.H., K.G., J.J.T.); and Research Laboratory for Biotechnology and Biochemistry, Kathmandu, Nepal (G.K.A.)

To better understand the metabolic processes of seed filling in soybean (*Glycine max*), two complementary proteomic approaches, two-dimensional gel electrophoresis (2-DGE) and semicontinuous multidimensional protein identification technology (Sec-MudPIT) coupled with liquid chromatography-mass spectrometry, were employed to analyze whole seed proteins at five developmental stages. 2-DGE and Sec-MudPIT analyses collectively identified 478 nonredundant proteins with only 70 proteins common to both datasets. 2-DGE data revealed that 38% of identified proteins were represented by multiple 2-DGE species. Identified proteins belonged to 13 (2-DGE) and 15 (Sec-MudPIT) functional classes. Proteins involved in metabolism, protein destination and storage, and energy were highly represented, collectively accounting for 61.1% (2-DGE) and 42.2% (Sec-MudPIT) of total identified proteins. Membrane proteins, based upon transmembrane predictions, were 3-fold more prominent in Sec-MudPIT than 2-DGE. Data were integrated into an existing soybean proteome database (www.oilseedproteomics.missouri.edu). The integrated quantitative soybean database was compared to a parallel study of rapeseed (*Brassica napus*) to further understand the regulation of intermediary metabolism in protein-rich versus oil-rich seeds. Comparative analyses revealed (1) up to 3-fold higher expression of fatty acid biosynthetic proteins during seed filling in rapeseed compared to soybean; and (2) approximately a 48% higher number of protein species and a net 80% higher protein abundance for carbon assimilatory and glycolytic pathways leading to fatty acid synthesis in rapeseed versus soybean. Increased expression of glycolytic and fatty acid biosynthetic proteins in rapeseed compared to soybean suggests that a possible mechanistic basis for higher oil in rapeseed involves the concerted commitment of hexoses to glycolysis and eventual de novo fatty acid synthesis pathways.

Plant seeds accumulate proteins, oils, and carbohydrates because these nitrogen and carbon reserves are necessary for early seed germination and seedling growth (for review, see Weber et al., 2005). These reserve components are synthesized during an extended phase of seed development, loosely termed seed filling. Seed filling is the period when rapid metabolic and morphological (size, weight, and color) changes occur, encompassing cellular processes that include cell expansion and the early stage of desiccation (Rubel et al., 1972; Mienke et al., 1981; Agrawal and Thelen, 2006). Seed filling is also the period that largely determines the relative levels of storage reserves in seeds. The relative proportion of storage components in seeds varies dramatically among different plant species. For example, soybean (*Glycine max*) seed contains approximately 40% protein and 20% oil (Hill and Breidenbach, 1974; Ohlrogge and Kuo, 1984). In contrast, seed of oilseed rape (*Brassica napus*; also called rapeseed or canola) contains approximately 15% protein and 40% oil (Norton and Harris, 1975). To gain insight into the complex process of seed development, identification of genes and proteins and their dynamic expression profiles during seed filling are beginning to provide a framework for more in-depth comparative studies.

Due to remarkable progress in the area of high-throughput transcriptomics (for review, see Brady et al., 2006) and proteomics (for review, see Agrawal et al., 2005), large-scale investigations of genes and proteins in plants are currently achievable. These technologies are providing a global picture of mechanisms that govern plant growth and development. In recent years, transcriptomics have been applied to investigate the changes in gene expression during seed development in cereal (maize [*Zea mays*; Lee et al., 2002], rice [*Oryza sativa*; Zhu et al., 2003; Yamakawa et al., 2007]), legume (soybean [Dhaubhadel et al., 2007; Le et al., 2007]), and oilseed (Arabidopsis [*Arabidopsis thaliana*; Ruuska et al., 2002; Hennig et al., 2004]) plants. These studies have provided a global view of gene activity and transcrip-

[1] This work was supported by the National Science Foundation-Plant Genome Research Program Young Investigator Award (grant no. DBI–0332418 to J.J.T.).

[2] Present address: Institute of Plant Genetics and Biotechnology, Slovak Academy of Sciences, Akademicka 2, Sk–950 07 Nitra, Slovak Republic.

* Corresponding author; e-mail agrawalg@missouri.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Ganesh Kumar Agrawal (agrawalg@missouri.edu).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.108.119222

tional networks linking metabolic and regulatory responses to seed developmental processes.

Proteomic analyses have also been performed to analyze seed filling in *Medicago truncatula* (Gallardo et al., 2003, 2007), soybean (Hajduch et al., 2005), and rapeseed (Agrawal and Thelen, 2006; Hajduch et al., 2006) using high-resolution two-dimensional gel electrophoresis (2-DGE) in combination with either matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) and/or nano-electrospray ionization liquid chromatography-tandem mass spectrometry (nESI-LC-MS/MS). Gallardo et al. (2003) profiled and identified 84 proteins at various stages of *M. truncatula* seed development. Hajduch et al. (2006) established confident quantitative expression profiles for 794 rapeseed protein spots on high-resolution 2-D gel reference maps. Of 794 protein spots, 517 protein spots representing 289 nonredundant (NR) proteins were successfully identified with an identification efficiency of 65.1% mainly due to utilization of a multiparallel protein identification strategy. Hajduch et al. (2006) performed both MALDI-TOF-MS and nESI-LC-MS/MS to analyze trypsin-digested peptides (from 2-DGE spots) followed by mass spectral search queries against both National Center for Biotechnology Information (NCBI) and The Institute for Genomic Research plant databases. MALDI-TOF-MS and nESI-LC-MS/MS were found to be complementary approaches to increase proteome coverage. Additionally, a parallel quantitative phosphoproteomics analysis revealed a surprisingly large number of phosphoproteins residing in developing rapeseed (Agrawal and Thelen, 2006). More than 44% of the identified phosphoproteins were enzymes involved in various metabolic pathways, suggesting that we are probably only beginning to understand the complexities of metabolic regulation in plants. In addition to these metabolic enzymes, tubulin β-8 chain, luminal binding protein, heat shock proteins, proteasome proteins, 14-3-3 proteins, annexins, and cruciferin subunits were also identified.

These proteomic studies have successfully established a quantitative 2-DGE-based workflow from protein isolation to protein assignment for developing soybean and rapeseed. However, 2-DGE-based proteomic strategies often underrepresent proteins with extreme hydrophobicity, mass, or pIs (for review, see Gorg et al., 2004) because of solubility (for review, see Santoni et al., 2000) or technical limitations to the approach. To identify these extreme proteins, gel-free proteomic approaches, such as multidimensional protein identification technology (MudPIT; Link et al., 1999; Washburn et al., 2001), are available. MudPIT separates peptides by a strong cation exchange phase in the first dimension followed by reverse-phase chromatography in the second dimension. MudPIT and 2-DGE have been performed comparatively in only a few plant investigations and the results indicated these were complementary approaches (Koller et al., 2002; Katavic et al., 2006). Given the physical diversity and broad dynamic range of expressed proteins in seed proteomes, it is advantageous to use multiple separation techniques like 2-DGE and MudPIT to expand proteome coverage.

In this study, we performed an in-depth investigation of proteins expressed during seed filling in soybean using 2-DGE and semicontinuous MudPIT (Sec-MudPIT) in combination with nESI-LC-MS/MS. The main objectives of this study are to (1) expand coverage of the soybean seed proteome and build resources for dissecting biological processes involved in seed filling; (2) compare the soybean study here with a parallel study of rapeseed to provide a global view on intermediary metabolism and its regulation in protein-rich (soybean) versus oil-rich (rapeseed) seeds; and (3) disseminate the integrated datasets of 2-DGE and Sec-MudPIT from soybean to the scientific community. 2-DGE and Sec-MudPIT analyses conducted on five sequential seed stages (2, 3, 4, 5, and 6 weeks after flowering [WAF]) collectively identified 478 NR proteins with only 70 proteins in common. An integrated 2-DGE database allowed us to map the protein identification and quantitative expression profile for 675 protein spots on high-resolution 2-D gel reference maps. High-quality and integrated datasets were then used to map protein components on metabolic biosynthetic pathways of carbohydrates, fatty acids, and proteins in soybean and compared with a parallel study of rapeseed. Significant differences were observed with respect to protein abundance, 2-DGE species (2-DGES), and expression trends reflective of differences in protein and oil content in these two economically important crop plants.

RESULTS

Parallel Proteomic Strategies Identify 478 NR Proteins from Developing Soybean Seed

Proteomic workflows illustrating the stages of seed development, utilization of 2-DGE and Sec-MudPIT separation techniques, MS, and associated bioinformatics tools for establishing a detailed soybean seed-filling proteome database are schematically depicted in Figure 1. To examine protein expression throughout seed filling, developing seed at five sequential stages (2, 3, 4, 5, and 6 WAF) was analyzed. Rapid increases in whole seed and embryo mass, total protein, and fatty acids were previously observed during these stages in soybean as well as rapeseed (Hajduch et al., 2005, 2006; Agrawal and Thelen, 2006). Previous proteomics analysis suggests that seed developmental stages of 2 through 6 WAF include most physiological and cellular processes occurring during seed filling (Hajduch et al., 2005).

2-DGE and Sec-MudPIT Are Complementary Approaches for Investigating the Seed-Filling Proteome of Soybean

Collectively, 2-DGE and Sec-MudPIT analyses resulted in the unambiguous identification of 478 NR proteins (i.e. unique accession nos.). 2-DGE and Sec-

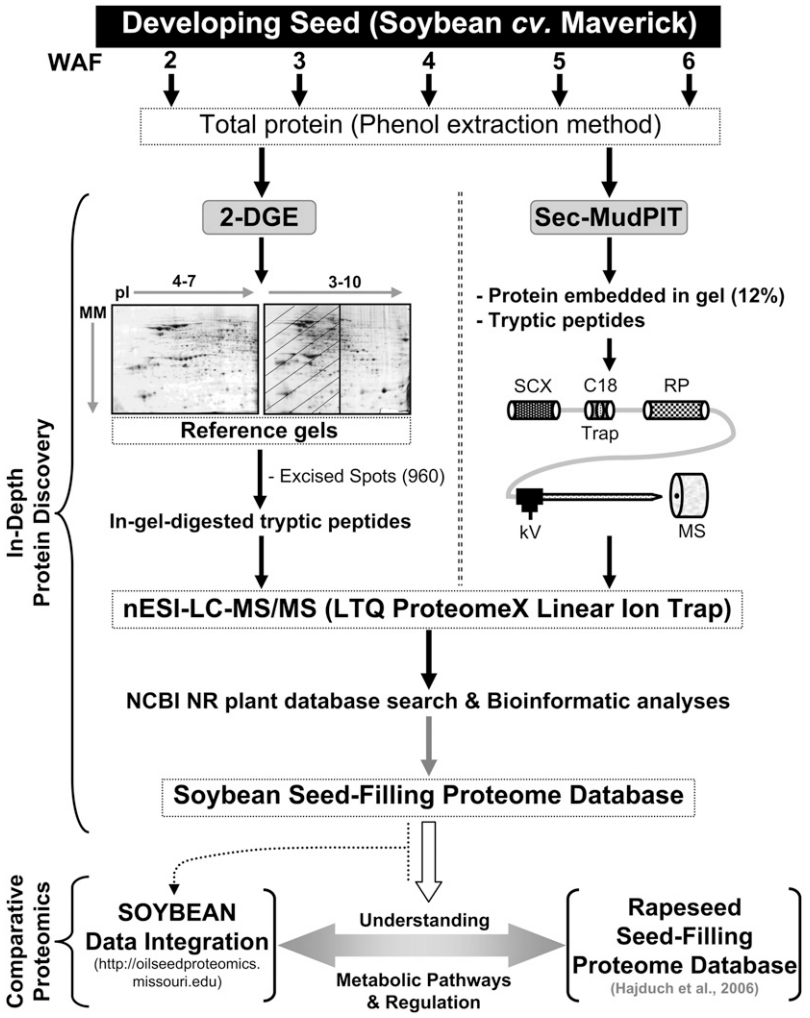

Figure 1. Schematic illustration of proteomics strategy includes establishment of a detailed soybean proteome database and comparison with a parallel study of rapeseed. 2-DGE and Sec-MudPIT complementary approaches were used for in-depth investigation of proteins, both quantification and identification, expressed during seed filling in soybean. 2-D reference gels were obtained by pooling equal amounts (0.2 mg) of protein sample from each of the five developmental stages studied. In the case of reference gel (pH 3–10), protein spots were excised only from the pH range 7 to 10. In Sec-MudPIT analyses, whole seed proteins isolated from three independent biological samples of the same developmental stage were used to carry out three independent Sec-MudPIT analyses. The nESI-LC-MS/MS-acquired data were searched against NCBI NR plant databases. Integrated quantitative soybean proteomic datasets were used to perform a comparative proteomics analysis with a parallel study of rapeseed. SCX, Strong cation exchange chromatography.

MudPIT separately identified 229 (representing 531 protein spots) and 319 NR proteins, respectively. Interestingly, only 70 NR proteins were common to both datasets (Supplemental Tables S1, S2, and S10; highlighted in light yellow). These results demonstrate that 2-DGE and Sec-MudPIT are truly complementary approaches. Abundance of 2-D protein spots at all five seed developmental stages are given after normalization with a correction constant (to merge pH 4–7 and 3–10 gels; Hajduch et al., 2005; Supplemental Table S3) and also after subtraction of seed storage proteins (SSPs; Supplemental Table S4). For Sec-MudPIT, the total number of NR peptides identified for each protein assignment along with other associated information is provided for all biological replications and seed developmental stages in Supplemental Table S10. Moreover, searches of Sec-MudPIT raw data against the plant database II did not reveal acrylamide modifications for any Cys-containing peptide, indicating that the conditions for embedding proteins in polyacrylamide were compatible with downstream MS.

Identified Proteins of the Seed-Filling Proteome Possess Diverse Biochemical Properties

Sec-MudPIT Is More Representative of Extreme Proteins of High Mass and pIs

2-DGE and Sec-MudPIT together identified proteins with a wide range of masses and pIs without any observable gaps (Fig. 2, A and B). Proteins within the 20- to 60-kD mass range were predominant in the seed-filling proteome representing 73% (2-DGE) and 47% (Sec-MudPIT) of total identified proteins (Fig. 2A). 2-DGE-identified proteins below 20 kD and above 60 kD were 8% (18 proteins) and 19% (44 proteins), respectively, whereas Sec-MudPIT-identified proteins were 8% (26 proteins) and 45% (143 proteins), respectively. Therefore, Sec-MudPIT identified 26% more high mass proteins over the 2-DGE approach. Proteins in the pI 4 to 7 range dominated the pI spectrum, accounting for 77% (2-DGE) and 65% (Sec-MudPIT) of total identified proteins from soybean seed (Fig. 2B). Sec-MudPIT better represented proteins of high pI (above 7), accounting for 35% (111 proteins) versus 23% (53 proteins) of 2-DGE analysis. Neither approach

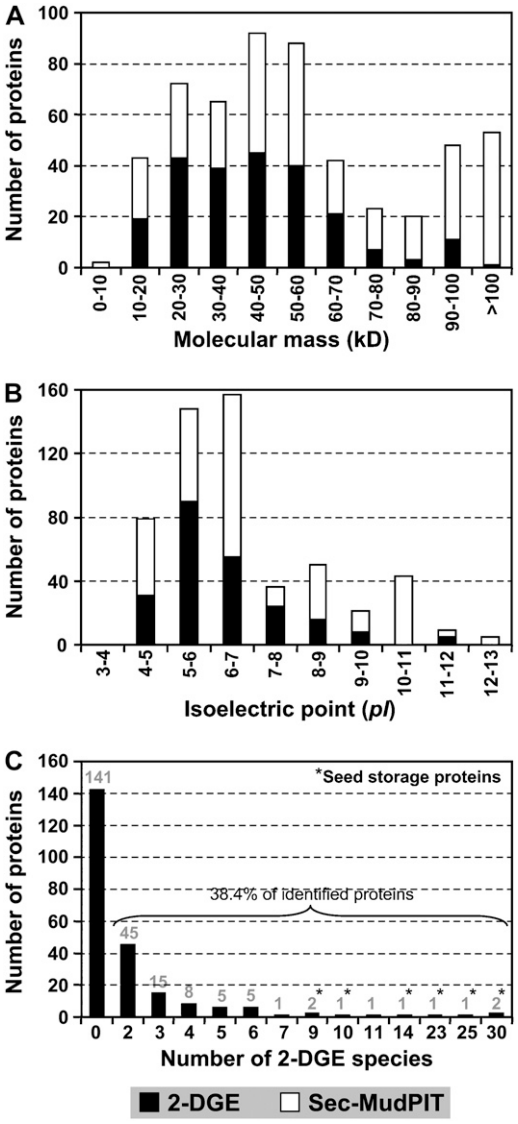

Figure 2. Biochemical properties of soybean proteins expressed during seed development. Theoretical molecular masses and pIs of proteins identified by 2-DGE and Sec-MudPIT were used for comparative analysis. Results presented in A and B are based on NR proteins. A, Molecular mass distribution over 10-kD increments of proteins identified by 2-DGE and Sec-MudPIT. B, Distribution of proteins versus pI. C, Distribution of 2-DGES of proteins identified by 2-DGE. Total number of proteins possessing multiple 2-DGES is mentioned on the top of each bar. SSPs are marked by asterisks.

identified proteins with pI values below 4. A similar protein mass and pI distribution was also seen in the established vegetative vacuole proteome in Arabidopsis using multiple proteomic approaches—multidimensional LC-MS/MS and 1-DGE (also called GelC-MS) coupled with nESI-LC-MS/MS (Carter et al., 2004). In that study, proteins within the 20- to 60-kD mass range and the 4 to 7 pI range represented 58% and 63%, respectively, of total identified proteins (402 proteins). Similar distribution patterns of 2-DGE-identified protein masses and pIs were also found in Arabidopsis tissues using the 2-DGE approach (Giavalisco et al., 2005); nearly 65% of identified proteins mapped to the pI 5 to 7 range. Altogether, Sec-MudPIT is the superior approach to characterize high mass and pI proteins.

2-DGE Data Analysis Reveals Approximately 38% of the Identified NR Proteins as Multiple 2-DGE Species

One advantage of 2-DGE over Sec-MudPIT is the ability to detect fractional pI or mass changes. It is well known that a single gene can manifest itself as multiple protein spots on a 2-D gel due to alternative splicing (for review, see Smith et al., 1989; Godovac-Zimmermann et al., 2005) and/or posttranslational modifications (PTMs; Agrawal and Thelen, 2006; for review, see Kersten et al., 2006). For simplicity, 2-D gel protein spots produced either from a single gene or from paralogous genes have collectively been called 2-DGES throughout the text because this study does not attempt to define the basis for each redundant 2-DGE spot. Nevertheless, we do understand that 2-DGE spots derived from a single gene (not from paralogous genes) are frequently termed isoforms and will be mentioned wherever appropriate. Distribution of 2-DGES in the soybean seed-filling proteome is shown in Figure 2C. Of 229 NR proteins, 88 were represented by multiple 2-DGES varying from 2 to as many as 30, suggesting that 38.4% of total identified proteins are possibly under the influence of different regulatory mechanisms, including PTMs and alternative splicing. Based on these results, it appears that each protein, on average, could form approximately 4.4 species of 2-DGE in the proteome of soybean seed filling. The production of 2-DGES was found to be widespread in the seed-filling proteome and associated with proteins of diverse function (Supplemental Fig. S1). Results also suggest that the frequency of 2-DGES is not random. For example, the protein destination and storage functional class possessed 29 proteins with 2-DGES, the highest number of proteins identified in any functional class. These 29 proteins were related to folding and stability, proteolysis, and storage.

Of 88 proteins, eight proteins together produced 148 species of 2-DGE ranging from seven to 30 per protein. Interestingly, all these proteins were identified as SSPs (Fig. 2C). These SSPs were glycinin (seven, 10, and 14 total 2-DGES for accession nos. 6015515, 90186615, and 4249566, respectively), soybean agglutinin in A complex (nine total 2-DGES for accession no. 14719778), β-conglycinin α-prime subunit (23 and 30 for accession nos. 68264915 and 74271741), glycinin G1 subunit (25 for accession no. 255221), and glycinin G2 subunit (30 for accession no. 18609). These results suggest that soybean SSPs, on average, produce approximately 18.5 protein spots of 2-DGE per protein. The only non-SSPs that generated more than seven species of 2-DGE were seed maturation protein (nine for accession no. 9622153) and Suc-binding protein (SBP; 11 for accession no. 6179947). Interestingly, these proteins are the most prominently expressed proteins in soybean seed.

To determine whether SSPs also produce a high number of 2-DGES in other plant species during seed development, the most systematic 2-DGE proteome study available to date for rapeseed (Hajduch et al., 2006) was surveyed. As shown in Supplemental Figure S2, 33.4% of the identified NR proteins contained more than one 2-DGES. SSPs were again the proteins (cruciferins) that produced the most forms (10.4 protein spots/protein), ranging from six to 24. Therefore, SSPs produce the highest number of 2-DGES in both soybean and rapeseed, which suggests that this might be a general phenomenon in developing seeds. Alternatively, the observation of multiple forms may simply reflect the prominence of SSPs.

Sec-MudPIT Approach Identifies a 3-Fold Higher Number of Membrane Proteins Than 2-DGE

Membrane proteins remain underrepresented in global 2-DGE-based analyses despite attempts to improve representation on 2-D gels, including solubilization efficiency prior to and during isoelectric focusing (IEF; for review, see Santoni et al., 2000; Gorg et al., 2004). Sec-MudPIT was therefore included to increase membrane proteins in the seed-filling proteome of soybean. Usually two parameters are used to describe the properties of membrane proteins: (1) the grand average of hydrophobicity (GRAVY) based on the empirical values of Kyte and Doolittle (1982); and (2) the number of predicted transmembrane domains (TMDs). GRAVY scores shown in Figure 3A provide an indication of protein hydrophobicity; the higher the protein GRAVY, the greater the probability of membrane association. To our surprise, and as can be seen from Figure 3A, only 8.7% (20 of 229 proteins) and 9.4% (30 of 319 proteins) of the total identified protein in 2-DGE and Sec-MudPIT analyses, respectively, showed predicted GRAVY scores between 0 and 1.0. Sec-MudPIT was able to identify some highly hydrophobic membrane proteins (e.g. exhibiting a positive GRAVY score in the range of 0.25–1.0), whereas 2-DGE failed to detect any of these proteins. The large fraction of identified proteins by both 2-DGE and Sec-MudPIT approaches harbored GRAVY scores below +0, suggesting that these proteins are hydrophilic. However, prediction of TMDs in identified proteins using the THUMBUP prediction tool manifested opposite, but expected, results (Fig. 3B). A total of 44 (of 229 in 2-DGE) and 144 (of 319 in Sec-MudPIT) proteins were predicted as likely candidates containing at least one TMD (Fig. 3B). Therefore, results obtained from the GRAVY and TMD prediction tools suggest that <10% of the proteins identified by either 2-DGE or Sec-MudPIT are hydrophobic, but 19% (2-DGE) and 45% (Sec-MudPIT) of identified proteins possess at least one TMD, implying that the majority of these predicted membrane proteins are largely hydrophilic. Moreover, proteomic studies on membrane proteins of other organisms have also reported differences between GRAVY and TMD prediction tools for prediction of membrane proteins as observed in this study (Fountoulakis and Gasser, 2003; for review, see Santoni et al., 2000).

Predicted membrane proteins from 2-DGE had a maximum of three TMDs; 91% of membrane proteins from 2-DGE (40 proteins) contained only one TMD (Fig. 3B). Similar results have been reported in 2-DGE-based studies of other organisms, even though those studies contained technical differences like extraction of membrane proteins with Triton X-114 to enrich membrane proteins (Klein et al., 2005; Zhang et al., 2005; Mattow et al., 2007). For example, the very recent 2-DGE analysis of the plasma membrane proteome of mycobacteria identified a maximal number of three TMDs in a protein (Mattow et al., 2007). In contrast, predicted membrane proteins identified in the Sec-MudPIT analysis harbored up to 13 TMDs and 37% of all membrane proteins possessed two or more TMDs (Fig. 3B). Sec-MudPIT identified 15 predicted membrane proteins (10.4%) with four or more TMDs, including PSII chlorophyll apoproteins (six TMDs), ATP-binding cassette transporter (six TMDs), unknown proteins (nine TMDs), and inorganic pyrophosphatase (13 TMDs). Furthermore, predicted membrane proteins of Sec-MudPIT were found to have TMDs up to 30 amino acids in length (Supplemental Fig.

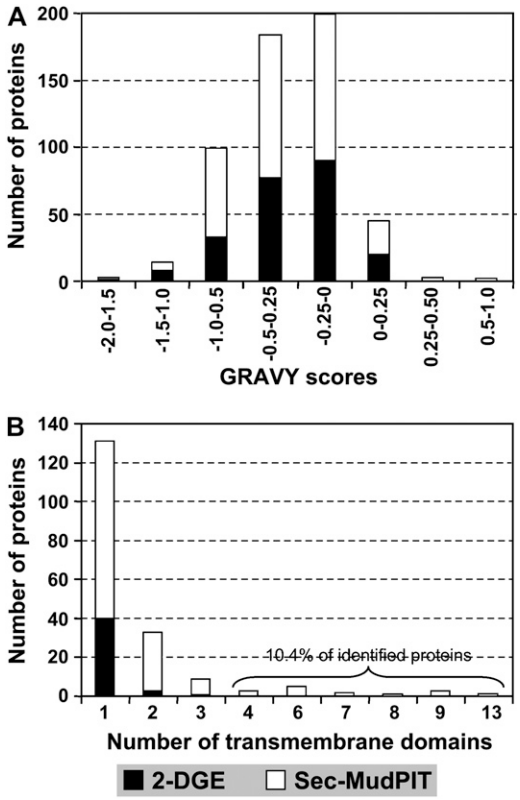

Figure 3. Distribution of hydrophobicity and TMDs of identified soybean proteins. Distribution is based on NR proteins. A, Distribution of proteins by hydrophobicity. The GRAVY value (http://bioinformatics.org/sms2/protein_gravy.html) was calculated as described previously (Kyte and Doolittle, 1982). B, Number of predicted TMDs identified in proteins as determined by the THUMBUP method (http://sparks.informatics.iupui.edu/Softwares-Services_files/thumbup.html).

S3). TMDs ranging from 16 to 18 and 20 to 30 amino acids were only observed with predicted membrane proteins of Sec-MudPIT. In contrast, predicted membrane proteins of 2-DGE had TMDs ranging from six to 15 and one with 19 amino acids (Supplemental Fig. S3). The high coverage of predicted membrane proteins, TMD frequency, and TMD length together indicate the unbiased nature of the modified Sec-MudPIT strategy employed. The low number of predicted membrane proteins identified in the 2-DGE analyses could be due to several factors (for review, see Santoni et al., 2000; Gorg et al., 2004), most notably, the tendency of membrane proteins to aggregate during IEF. Taken together, Sec-MudPIT in combination with the polyacrylamide-embedded protein in-gel digestion system identifies 3-fold higher putative membrane proteins with maxima of 13 TMD up to 30 amino acids in length.

Presence of Functionally Diverse Proteins in Developing Soybean Seed

Sorted NR proteins within the functional class and subclass are listed in Supplemental Tables S1 and S2 for 2-DGE and S10 for Sec-MudPIT. 2-DGE- and Sec-MudPIT-identified proteins belonged to 13 and 15 functional classes, respectively (Table I). Major functional classes of proteins that emerged from both 2-DGE and Sec-MudPIT analyses were primary metabolism, protein destination and storage, and energy, collectively accounting for 61.1% (140 of 229 proteins) and 42.2% (135 of 319 proteins) of identified NR proteins, respectively. Proteins of unclassified and transposon classes were only present in the Sec-MudPIT analyses.

The 44 and 144 predicted membrane proteins of 2-DGE and Sec-MudPIT, respectively, were also organized into their functional classes (Table I). Predicted membrane proteins of 2-DGE and Sec-MudPIT belonged to eight and 15 functional classes, respectively. Primary metabolism and protein destination and storage were the most abundant class of membrane proteins, jointly representing 61.3% (27 proteins) and 35.4% (51 proteins) of predicted membrane proteins for 2-DGE and Sec-MudPIT, respectively. Predicted membrane proteins involved in secondary metabolism, intracellular traffic, cell/growth and division, protein synthesis, transposons, and unclassified functional classes were identified only by Sec-MudPIT, highlighting the usefulness of a Sec-MudPIT approach for expanding the proteome by way of membrane proteins. Looking at data on total protein and predicted membrane proteins as per functional classes, it appears that predicted membrane proteins cover a significant proportion of proteins involved in different functional classes.

Integrated Soybean Seed-Filling Proteome Database Represents a Rich Resource of Quantitative Protein Data

One of the goals of this study was to further refine the extant soybean seed-filling proteome database (Fig. 4). This database has been appended to the Web portal for proteomics research on oilseed plants (http://oilseedproteomics.missouri.edu) that is freely accessible to the scientific community. The integrated database includes composite datasets derived independently from 2-DGE and Sec-MudPIT proteomic approaches. The 2-DGE dataset contains the 2-DGE data collected in this study and by Hajduch et al. (2005). Hajduch et al. (2005) used 2-DGE coupled with MALDI-TOF-MS to identify expressed proteins during seed filling in soybean. Prior to 2-DGE data integration, protein spots of reference gels were matched to protein spots of the reference gel developed by Hajduch et al. (2005) to use the same group spot ID assigned previously by Hajduch et al. (2005) using ImageMaster 2D platinum software, version 5 (called ImageMaster software; GE Healthcare). This step facilitated data integration and avoided confusion on spot ID from this and the previous study. Merging of two datasets resulted in unambiguous assignment of 675 of 960 total protein spots, representing 70% coverage; 253 and 144 were exclusively identified using nESI-LC-MS/MS and MALDI-TOF-MS, respectively (Supplemental Fig. S4). Therefore, 2-D gel reference maps (pH 4–7 and 3–10) provide densely mapped landmarks of protein spots associated with protein assignments and quantitative expression profile data. As schematically depicted in Figure 4, the 2-D gel reference map offers two viewing options. The first option is through protein spots on the reference map (Fig. 4; 2-DGE). Protein spots identified using nESI-LC-MS/MS and/or MALDI-TOF-MS were turned into an active, hyperlinked protein spot. The presence of the cursor on any of these spots automatically displays spot group ID and protein name. If the protein spot is clicked, it leads users to another Web page displaying the quantitative expression profile and the hyperlinked MS data—MS/MS data (nESI-LC-MS/MS) and peptide mass fingerprint data (MALDI-TOF-MS). The second option to view the quantitative expression profile and identification data is tables of identified proteins for nESI-LC-MS/MS and MALDI-TOF-MS. In these tables, protein IDs assigned by ImageMaster software are hyperlinked to respective expression profiles. In the case of Sec-MudPIT, all output results for each biological replication in addition to a composite table are included as Supplemental Data (Supplemental Table S10).

DISCUSSION

An in-depth understanding of metabolic events that determine overall components of storage reserves in seeds of crop plants will be important for improving seed quality and yield. Given the multiplicity of metabolic events, an integrated proteomics approach was used to provide a global perspective on this complex system. Keeping the objectives in mind, results will be discussed within two major headings: (1) seed development—

Table I. *Functional classification of soybean NR proteins identified by 2-DGE and Sec-MudPIT*

Functional classification was carried out according to the classification established for Arabidopsis by Bevan et al. (1998) with modification as discussed in "Materials and Methods."

Functional Class	2-DGE		Sec-MudPIT	
	Total Protein (229)	Membrane Protein (44)	Total Protein (319)	Membrane Protein (144)
Primary metabolism	54	17	55	37
Energy	34	4	33	13
Cell growth/division	9	0	5	2
Transcription	4	0	20	7
Protein synthesis	18	0	33	1
Protein destination and storage	52	10	47	14
Transporters	6	2	7	6
Intracellular traffic	5	0	5	3
Cell structure	6	3	10	7
Signal transduction	12	1	15	7
Disease/defense	15	5	7	5
Secondary metabolism	7	0	9	6
Unclear classification	7	2	24	12
Unclassified	0	0	36	16
Transposons	0	0	13	8

establishment and dissemination of integrated soybean proteome database; and (2) characterizing seed metabolic pathways and regulation: comparative parallel proteomics of soybean and rapeseed.

Seed Development—Establishment and Dissemination of Integrated Soybean Proteome Database

Seed development has been studied in *M. truncatula* (Gallardo et al., 2003, 2007), soybean (Hajduch et al., 2005), wheat (*Triticum aestivum*; Vensel et al., 2005), rapeseed (Agrawal and Thelen, 2006; Hajduch et al., 2006), and maize (Mechin et al., 2007) using quantitative proteomics. These proteomic analyses applied only 2-DGE-based approaches in combination with either MALDI-TOF-MS and/or nESI-LC-MS/MS for protein identification. In this study, two independent separation but complementary proteomic techniques, 2-DGE and Sec-MudPIT, were used to ensure in-depth protein discovery of the seed-filling proteome in soybean. For MudPIT analyses, two changes were made to collect unbiased data, specifically for membrane, acidic, and basic proteins. The first change was the use of Sec-MudPIT over traditional MudPIT. Sec-MudPIT involves an optimized semicontinuous pumped salt gradient versus the injected salt step with traditional MudPIT. In our experience, and as reported by Nagele et al. (2004), Sec-MudPIT improves the overall identification of peptides and proteins compared to traditional MudPIT by optimizing the performance of strong cation exchange chromatography in the first dimension. The second change was performing in-gel trypsin digestion by embedding SDS-solubilized protein into a polyacrylamide gel matrix without electrophoresis to avoid in-solution digestion. SDS is an ionic detergent that is proven for solubilizing membrane proteins. One advantage of in-gel trypsin digestion of embedded

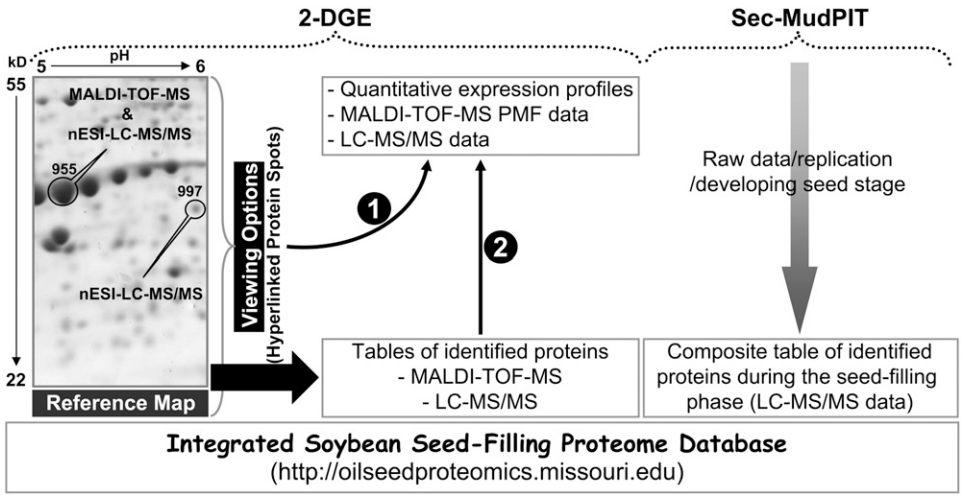

Figure 4. Architecture of the expanded soybean seed-filling proteome Web database. The displayed 2-D gel is an expanded area of the high-resolution reference map (pH 4–7) corresponding to pH 5 to 6 and molecular mass 22 to 55 kD, where protein spots 955 and 997 identified by MALDI-TOF-MS and nESI-LC-MS/MS and nESI-LC-MS/MS, respectively, are displayed. Details are described in the text.

proteins is that washing steps effectively remove interfering substances, such as SDS, which interferes with trypsin activity and also with MS analyses resulting in poor sensitivity (Yu et al., 2003). This modified approach in combination with nESI-LC-MS/MS analysis has recently been shown to identify a large number of membrane proteins in a complex protein mixture prepared from prostate cancer cells (Lu and Zhu, 2005), suggesting that membrane and other extreme proteins from developing seeds might be more accurately presented to the mass spectrometer.

A combination of 2-DGE and Sec-MudPIT identified 478 total NR proteins with an overlap of only 70 proteins, demonstrating that these two approaches are complementary and useful for in-depth protein discovery in developing seed. Sec-MudPIT, in particular, better represented extreme proteins in the dataset, such as membrane proteins, which were underrepresented in all previous 2-DGE-based proteomic studies on seed development, including this one. 2-DGE has been advantageous in identifying the diversity of 2-DGES, which accounted for 38% of total identified proteins by 2-DGE analysis. Based on this result, it is tempting to speculate that a significant portion of the soybean seed-filling proteome is perhaps under post-translational control. One good example is the SSPs. SSPs produced seven to 30 species of 2-DGE per gene in soybean and six to 24 in rapeseed. Interestingly, SSPs, specifically cruciferin, were previously identified as phosphoproteins (Agrawal and Thelen, 2006) and the phosphorylation sites in some cruciferins were recently mapped (Wan et al., 2007).

Due to the low efficiency of protein identification by peptide mass fingerprinting, previous 2-DGE analysis of seed filling in soybean (Hajduch et al., 2005) failed to identify a significant portion of proteins that were identified here. For example, a large number of soybean proteins mapped on metabolic pathways in Figure 7 was exclusively identified in this study. This difference is likely due to the use of nESI-LC-MS/MS analysis instead of MALDI-TOF-MS/peptide mass fingerprinting. Indeed, it was recently shown that nESI-LC-MS/MS coupled with 2-DGE increases protein identification efficiency approximately 3-fold compared to MALDI-TOF-MS in rapeseed (Hajduch et al., 2006). These examples demonstrate the importance of a multiparallel experimental design to expand proteome coverage.

Studies on seed development in different plants, including this study, have revealed that, although proteins of diverse function are expressed during seed filling, proteins involved in metabolism are highly represented. One of the aims of such large-scale proteomics analysis has been to gain insight into the complex metabolic network, to predict metabolic flow, and to identify key enzymatic steps. This study has integrated data derived from 2-DGE (both this study and previous study by Hajduch et al. [2005]) and Sec-MudPIT techniques to update a soybean seed-filling proteome Web database. This database displays interactive high-resolution 2-D gel reference maps in which 70% of the detected protein spots are characterized. This high-quality database may be used for comparative proteomics, evaluation of seed quality, and as a complement to the ongoing soybean genome annotation.

Characterizing Seed Metabolic Pathways and Regulation: Comparative Parallel Proteomics of Soybean and Rapeseed

The integrated soybean seed-filling proteome database developed in this study was used for comparison with a parallel study of rapeseed (Hajduch et al., 2006) to provide a global view on metabolic pathways involved in seed-filling processes in two distinct oilseeds. These parallel studies from the same lab using the same quantitative proteomic approaches allow for a controlled comparison to shed light on the fundamental metabolic differences responsible for protein-rich versus oil-rich seeds. It is worth mentioning at least two parameters that were used to present the integrated quantitative proteomics data. First, to characterize expression trends for proteins involved in either a particular enzymatic reaction or specific metabolic pathways, composite expression analysis (Hajduch et al., 2005) was performed by summing normalized relative spot volumes for those 2-DGE spot groups that map to the same cognate gene. Second, if identification of a spot group by nESI-LC-MS/MS did not match with the MALDI-TOF-MS assignment, MALDI-TOF-MS-identified protein was considered confident only if total NR peptides were either equal to or greater than 3-fold of total NR peptides identified by nESI-LC-MS/MS.

Lipoxygenase and SBPs Are Highly Represented in the Seed of Soybean But Not in Rapeseed

One of the interesting findings of this comparative parallel proteomics analysis is the high representation of lipoxygenase (LOX) and SBP in developing soybean seed (Supplemental Tables S1, S2, and S10). In total, 12 and 14 protein spots corresponded to eight and one unique LOX and SBP accessions were identified, respectively. Prominent expression of multiple paralogs/isoforms of LOX and SBP implies functional evolution and significance in developing soybean seed. LOX produces unsaturated fatty acids collectively called oxylipins, including jasmonic acid, which are critically important for plant growth and development (for review, see Siedow, 1991; Porta and Rocha-Sosa, 2002). Although the physiological roles of LOX in mature or developing seed have not yet been firmly established with respect to seed quality, LOXs play active roles in several physiological processes during plant life, including seed defense and storage, germination, vegetative growth, and abiotic and biotic stresses (for review, see Siedow, 1991; Porta and Rocha-Sosa, 2002). To gain insight into the physiological roles of

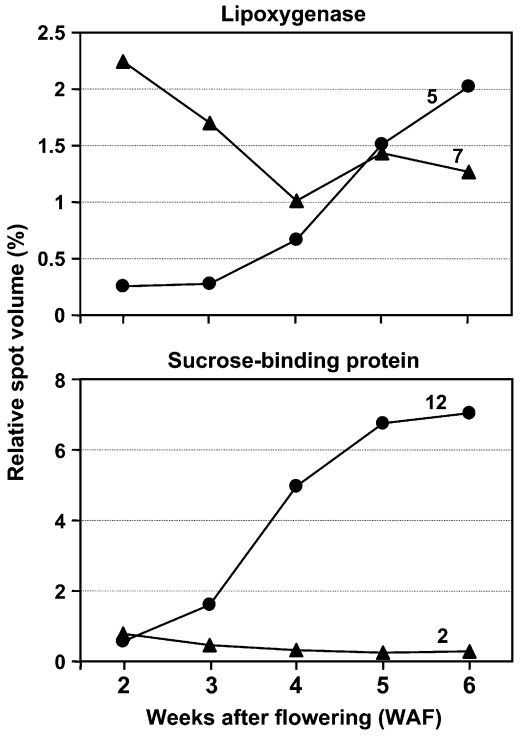

Figure 5. Nonhierarchical clustering of LOX and SBP composite expression profiles. These composite expression profiles were generated using the 2-DGE derived total protein spot abundance dataset after normalization and subtraction of SSPs. A total of 12 and 14 protein spots identified as LOXs or SBPs, respectively, by nESI-LC-MS/MS (this study) and MALDI-TOF-MS (previous study; Hajduch et al., 2005) were used to generate composite expression profiles. Total number of protein spots grouped together to form an expression profile is given. The composite expression profile is the sum of the relative spot volume at each of the five developmental stages of seed filling of all 2-DGES.

LOX in developing seed, nonhierarchical cluster analysis of 12 protein spot expression profiles generated two distinct composite expression trends, reflecting multiple functions of LOX proteins during seed filling (Fig. 5). The total relative abundance of expression trend composed of seven expression profiles decreased sharply from 2 to 4 WAF and maintained nearly at the same level until 6 WAF. This result suggests possible involvement of LOX proteins in active metabolic and cellular processes like cell expansion. In contrast, total abundance of expression trend composed of five expression profiles increased as seed filling progressed and reached a maximum at 6 WAF.

In contrast to soybean, no LOX proteins were detected in the 2-DGE-based proteomics study of seed filling in rapeseed (Hajduch et al., 2006). SecMudPIT analysis of total protein isolated from rapeseed of 4 WAF showed no identification of LOX protein (G.K. Agrawal and J.J. Thelen, unpublished data). Furthermore, the 2-DGE-based proteomic studies did not find LOX protein expressed during seed filling in developing wheat (Vensel et al., 2005) or maize endosperm (Mechin et al., 2007). However, LOX proteins were detected in developing seed of *M. truncatula* (Gallardo et al., 2003). Thus, LOX proteins are highly represented in developing seed of soybean and *M. truncatula*, and possibly other legumes. Although the functional significance of LOX in seed is yet unknown, LOX proteins might also be associated with seed quality, such as taste and flavor, because LOX proteins metabolize polyunsaturated fatty acids and give rise to compounds that can produce desirable or undesirable tastes and odors (Hildebrand, 1989).

Like LOX, proteomic studies suggest that SBPs are also underrepresented in developing seed of nonlegume plants like rapeseed (Hajduch et al., 2006), wheat (Vensel et al., 2005), and maize (Mechin et al., 2007). It is worth mentioning that SBPs were first identified from soybean and implicated in Suc translocation-dependent physiological processes (Ripp et al., 1988). The expression patterns of 12 isoforms of SBP strengthen the notion that SBP may be involved in Suc transport during seed development (Fig. 5). Of 14 SBP isoforms, the total abundance of 12 isoforms at 2 WAF increased almost exponentially until 6 WAF at which point it represented 7% of the total soybean protein (Fig. 5). A similar expression pattern for SBPs at the transcript level has been reported in developing seed of Arabidopsis (Ruuska et al., 2002), although overall abundance was much lower than in soybean. Increased expression of SBP during seed filling is in line with an increased demand of carbon for synthesis of storage reserve components in seed. The expression profile of the other two SBP isoforms did not show significant change during seed filling.

Expression Profiling of Metabolic Proteins Involved in Storage Reserve Synthesis and Metabolism in Soybean and Rapeseed Reveals Substantial Up-Regulation of Fatty Acid Synthesis in Rapeseed

Metabolic pathways responsible for synthesis and storage of reserve components are mostly known (Hill, 2004; Weber et al., 2005). However, the regulation of individual enzymes and net changes in abundance with the progression of seed filling is only recently being discovered. To gain insight into the connection between seed filling-specific processes and metabolic pathways involved therein, composite expression profiles of metabolic proteins were compiled for polysaccharide synthesis, amino acid synthesis, glycolysis, fatty acid and lipid synthesis, seed maturation, and storage proteins (Fig. 6).

Overall expression trends and abundance of proteins involved in amino acid and storage protein syntheses closely resembled one another in soybean and rapeseed. Proteins involved in polysaccharide synthesis and glycolysis also exhibited similar trends in soybean and rapeseed, although overall levels of expression were 20% to 50% lower in soybean. In general, soybean and rapeseed proteins were predominantly expressed in the early phase of seed filling (i.e. 2 WAF) for polysaccharide synthesis, amino acid syn-

thesis, and glycolysis. Interestingly, their total abundance decreased sharply after 3 WAF with seed filling, which coincided with an increase in total abundance of storage proteins from 3 to 6 WAF. The importance of glycolytic processes in carbon assimilation cannot be overestimated because they comprised up to 7.4% and 10.3% of total protein expression in soybean and rapeseed, respectively.

Proteins involved in fatty acid synthesis and seed maturation exhibited dramatically different metabolic profiles between soybean and rapeseed. The abundance of rapeseed proteins involved in fatty acid synthesis increased sharply from 2 WAF, reaching a maximum at 4 WAF (approximately 6% of total protein expression), followed by a sharp decrease until 6 WAF, whereas soybean proteins remained at almost the same level from 2 through 4 WAF with a slight decline at 5 WAF. This stark difference in expression is undoubtedly reflective of low versus high oil content in seeds of soybean and rapeseed, respectively. Although the abundance of seed maturation proteins did not increase with seed filling in soybean, in rapeseed these proteins increased almost 9-fold between 2 and 6 WAF. Taken together, these results suggest that total abundance of proteins associated with amino acid and storage protein syntheses are regulated in an inverse coordinated manner in both soybean and rapeseed; that is, a decrease in de novo amino acid synthesis is followed by a sharp increase in total storage protein abundance after 3 WAF. However, it should be noted that de novo amino acid synthesis remains at 50% of the 2 to 3 WAF levels by 6 WAF, indicating de novo synthesis has not completely shut down. Protein abundances of polysaccharide synthesis and glycolysis are also inversely correlated with storage protein synthesis, and their turnover may supplement the de novo-synthesized pools of free amino acids for storage protein synthesis. Interestingly, expression of these proteins is higher and sustained longer during seed filling in rapeseed versus soybean.

Mapping of Soybean and Rapeseed Metabolic Proteins on Suc Assimilatory and Glycolytic Pathways Leading to Oil Synthesis Identifies Potential Metabolic Differences

Monosaccharide breakdown is a complex process in plants due to the existence of a complete or nearly complete glycolytic pathway in plastids parallel to the cytosolic pathway (Miernyk and Dennis, 1983; for review, see Plaxton, 1996; Plaxton and Podesta, 2006). Despite its central role in carbon assimilation during seed development, little is known about the regulation of glycolysis. Recently, proteomic data collected on rapeseed seed filling were exploited to examine the redundancy of glycolytic pathways between cytosol and plastid (Hajduch et al., 2006). It was reported that carbon flow from Suc appears to primarily follow a cytosolic glycolytic track until phospho*enol*pyruvate (PEP), at which point carbon is likely imported into the plastid and converted into pyruvate and acetyl-CoA for de novo fatty acid synthesis (Hajduch et al., 2006). Soybean seed contains only 20% oil compared to 40% oil in rapeseed; therefore, mapping the carbon metabolic network with proteins identified in soybean and rapeseed datasets may help in identifying key enzymatic reactions based on differences in protein abun-

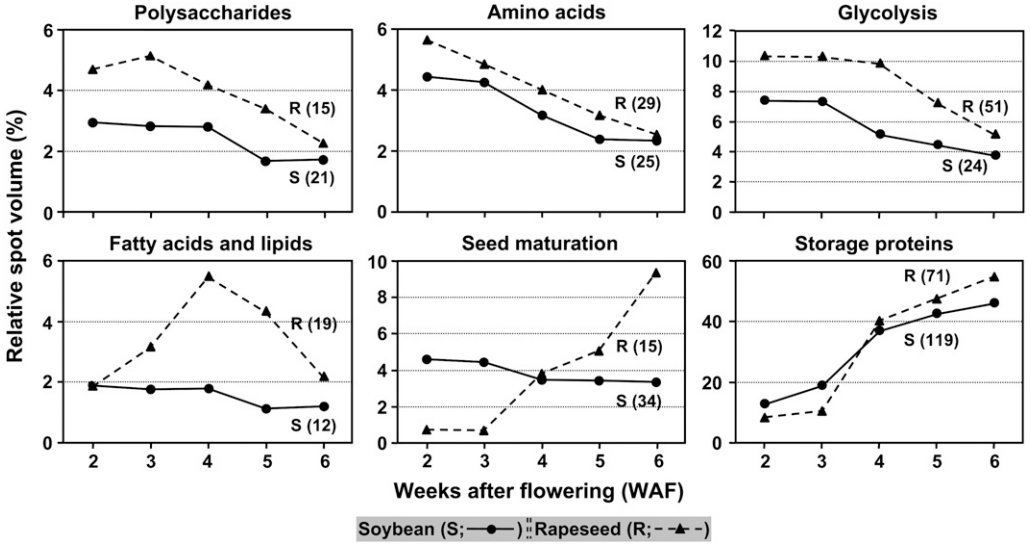

Figure 6. A general overview of 2-DGE composite expression trends of a particular set of proteins involved in synthesis of carbohydrate, protein, and fatty acid in soybean and comparison with a parallel study of rapeseed. Composite expression profiles were created using the 2-DGE-derived total protein spot abundance dataset after normalization and subtraction of SSPs, except for storage proteins where no subtraction of SSPs was performed. S and R represent soybean (solid line) and rapeseed (dashed line), respectively. Number in parentheses is the total number of 2-DGE protein spots observed by nESI-LC-MS/MS (this study) and MALDI-TOF-MS (previous study; Hajduch et al., 2005) and summed for the expression trend.

dance and localization. Suc assimilatory and glycolytic pathways from Suc to pyruvate, to TCA cycle, and to fatty acid synthesis have been schematically depicted (Fig. 7).

In total, 54 and 80 2-D protein spots of soybean and rapeseed, respectively, were identified and mapped along Suc assimilation and glycolytic pathways, indicating a concerted action of approximately 48% more proteins for fatty acid synthesis in rapeseed than soybean. Moreover, a net protein abundance of 80% was also found higher in rapeseed than soybean. This difference is significant and might affect carbon flow into de novo fatty acid synthesis and eventual triacylglycerol accumulation in soybean because most of the carbon is derived from monosaccharide breakdown. This is perhaps the reason that significant differences in metabolic proteins of nine enzymatic steps were observed. These include Suc synthase, cytosolic phosphoglucomutase, fructokinase, Fru-bisP aldolase, triose-P isomerase, phosphoglycerate kinase, 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, pyruvate kinase (PK), and pyruvate dehydrogenase complex (PDC) proteins for enzymatic steps 1, 3, 5, 8, 9, 11, 12, 14, and 18, respectively. These metabolic proteins differed in their total abundance, expression profiling, or localization. For example, triose-P isomerase showed a strong decline in total expression in soybean versus an increase in rapeseed as seed filling progressed (step 9). PK was detected as a cytosolic form in soybean and as a plastidial form in rapeseed (step 14).

Other enzymes that were substantially different between soybean and rapeseed were Rubisco, PEP car-

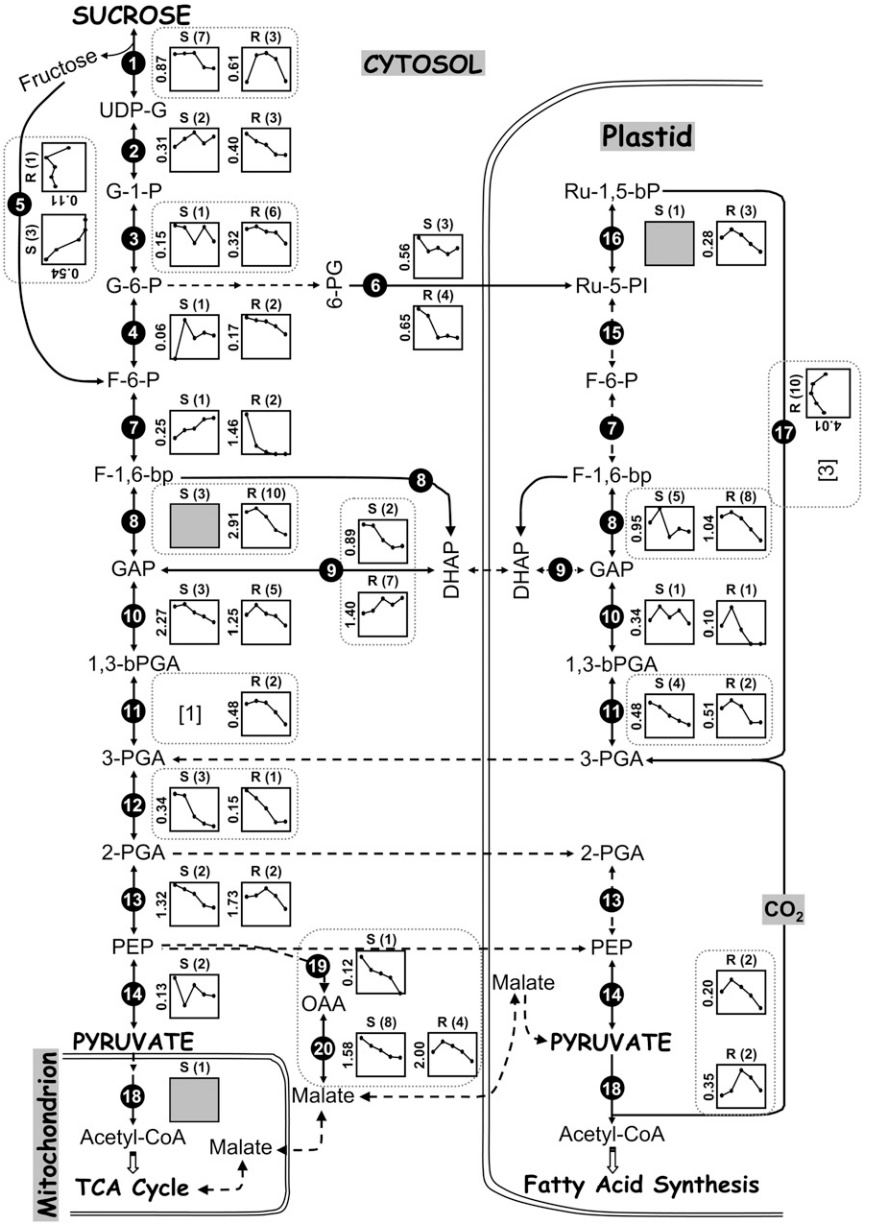

Figure 7. Comparative proteomics view of carbon assimilatory and glycolytic pathways leading to fatty acid synthesis in soybean and rapeseed reveals apparent differences in enzyme abundance, number of 2-DGES, and expression trends. Composite expression profiles of identified enzymes in soybean (S) and rapeseed (R) are presented after normalization and subtraction of SSPs. The *y*-axis value is the total relative abundance of spot volume at the peak of expression. Each filled circle within the expression profile represents the developmental stage of seed filling starting at 2 WAF through 6 WAF. The total number of protein spots identified by nESI-LC-MS/MS (this study) and MALDI-TOF-MS (previous study; Hajduch et al., 2005) for a given enzyme is noted in parentheses. Absence of protein expression profiles for a given enzymatic step indicates no identification of that enzyme in either soybean or rapeseed by 2-DGE analysis. In this situation, Sec-MudPIT datasets were searched and the total number of identified enzymes is given in the bracket near each enzyme number for soybean. Shaded squares indicate that expression profiles were not obtained for those identified proteins due to either very low abundance or presence in less than three developmental stages. Broken arrows denote no identification of enzymes in soybean and rapeseed datasets. Broken boxes highlight abundance and expression profile differences between soybean and rapeseed. Abbreviations for metabolites are as defined in the text and as follows: UDP-G, uridine diphospho-Glc; G-1-P, Glc-1-P; G-6-P, Glc-6-P; 6-PG, 6-phospho-D-gluconate; F-6-P, Fru-6-P; F-1,6-bP, Fru-1,6-bisP; GAP, glyceraldehyde-3-P; DHAP, dihydroxyacetone phosphate; 1,3-bPGA, 1,3-bisphosphoglyceric acid (or 1,3-bisphosphoglycerate); PGA, phosphoglyceric acid; Ru-5-PI, ribulose-5-P; Ru-1,5-bP, ribulose-1,5-bisP.

boxylase (PEPC), and malate dehydrogenase (MDH) for enzymatic steps 17, 19, and 20, respectively. Rubisco was recently shown to play an essential role in recycling carbon dioxide released by plastidial PDC to maintain the efficiency of oil production in the embryo (Schwender et al., 2004a). Rubisco is expressed at very low levels in developing soybean seed and was only detected by Sec-MudPIT, whereas it is prominent in rapeseed with 10 species of 2-DGE (approximately 4% of the total protein expression at the peak). Perhaps this difference in Rubisco abundance could contribute to low carbon efficiency and therefore reduced oil in soybean.

Further investigation of these differential proteins revealed two interesting features: (1) most of them are expressed as multiple 2-DGES in soybean and rapeseed; and (2) except Rubisco, all of the 11 metabolic proteins have previously been shown to be phosphoproteins (Thelen et al., 1998; Hardin et al., 2003; Tang et al., 2003; Tripodi et al., 2005; Agrawal and Thelen, 2006). Based on these findings, it is possible that the majority of Suc assimilatory enzymes undergo PTMs during seed filling, most likely protein phosphorylation. Consistent with this view, a large number of phosphoproteins involved in metabolism and energy have been reported to be expressed in a dynamic manner during seed filling in rapeseed (Agrawal and Thelen, 2006). Moreover, some of these key enzymes have already been implicated in the regulation of Suc assimilation or glycolysis, including Suc synthase (for review, see Hill, 2004), PK (Andre et al., 2007), and PEPC (for review, see Jeanneau et al., 2002).

Previous studies on oil synthesis in plant seeds have proposed two major routes for carbon flow to de novo fatty acid synthesis (for review, see Jeanneau et al., 2002; Schwender et al., 2004b; Plaxton and Podesta, 2006). The first, and likely ubiquitous, route involves cytosolic glycolysis to PEP followed by PEP import to plastids and production of acetyl-CoA by the sequential action of plastid PK and PDC. An alternative shunt proposes cytosolic PEP is converted to oxaloacetic acid by PEPC (step 19) and then malate by MDH (step 20), followed by malate import into plastids and successive decarboxylations to pyruvate (by a plastid NADP malic enzyme) and acetyl-CoA (plastid PDC) for de novo fatty acid synthesis. Diverse approaches, including transcriptomics, stable isotopic labeling, and proteomic analyses of developing seed from Arabidopsis and rapeseed, support the first route (White et al., 2000; Ruuska et al., 2002; Schwender et al., 2004a; Hajduch et al., 2006). In contrast, biochemical studies of developing seed and isolated leucoplasts support the second route for castor bean (*Ricinus communis*; for review, see Plaxton and Podesta, 2006). With soybean, the abundance and expression patterns of metabolic proteins in the cytosol and plastid do not unequivocally support either carbon route (Fig. 7). For example, the two PK forms identified were mapped to a cytosolic form and PEPC was detected in soybean, but not in rapeseed; additionally, cytosolic MDH was very prominent in soybean. Collectively, these data suggest that the second route could also be functioning in developing seed of soybean. Further investigation of key enzymatic steps for these two carbon routes may help in defining the major route of carbon flow for de novo fatty acid synthesis in soybean and perhaps in other plants.

In conclusion, we show that 2-DGE and Sec-MudPIT are complementary proteomic approaches for characterizing the soybean seed-filling proteome. This study (1) initiates comparative parallel proteomics analyses of developing oilseeds by comparing the quantitative seed-filling proteome of soybean and rapeseed; and (2) provides a global proteomics perspective on accumulation of storage reserves during seed filling, highlighting the importance of proteomics study in answering the fundamental biological questions of what properties distinguish high versus low oilseed.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

Soybean (*Glycine max* L. Maverick) seeds were grown in soil (Promix) in a greenhouse (light/dark cycles of 16 h [26°C]/8 h [21°C], and 48% humidity) in Columbia, MO. Plants were fertilized with fertilizer (15:30:15::nitrogen:phosphorus:potassium) at 2-week intervals. Flowers were tagged immediately after opening of buds (between 1 and 3 PM CST). Developing pods were collected between 1 and 3 PM at precisely 2, 3, 4, 5, and 6 WAF (i.e. 14, 21, 28, 35, and 42 d after flowering). Seeds were excised from collected pods, frozen immediately with liquid nitrogen, and stored at −80°C until use. Four biological sample pools were collected for each developmental stage. A total of four biological replications were performed for 2-DGE and three for Sec-MudPIT analyses.

2-DGE and Image Analyses

Total seed protein was isolated from seeds (0.5 g) of each developmental stage according to a modified phenol-based procedure as described previously (Hajduch et al., 2005). Protein pellets were resuspended in IEF extraction solution (8 M [w/v] urea, 2 M [w/v] thiourea, 4% [w/v] CHAPS, 2% [v/v] Triton X-100, and 50 mM [w/v] dithiothreitol) and subjected to 2-DGE analysis followed by image analysis using ImageMaster software as described previously (Hajduch et al., 2005). Briefly, ImageMaster software analyses of acquired 2-D gel images resulted in two independent datasets from pH 4 to 7 and pH 3 to 10 2-DGE gels (Fig. 1; 2-DGE). Because spot redundancy was eliminated by analyzing only the 7 to 10 region of the 3 to 10 gels, the two datasets could be statistically merged to enable direct comparison of spot group abundances from both sets of gels. Normalization was performed by calculating the correction constants for gels pH 4 to 7 and pH 3 to 10 using a formula described previously (Hajduch et al., 2005). Constants were calculated for each experimental stage (statistical average of biological replicates) and employed by multiplying individual relative volumes with the correction constant for particular dataset (i.e. pH 4–7 and 3–10).

Identification of 2-DGE Protein Spots by nESI-LC-MS/MS

In-gel digestion of 2-D protein spots and analysis of peptides by nESI-LC-MS/MS on an LTQ ProteomeX instrument was performed as described previously (Agrawal and Thelen, 2006; Hajduch et al., 2006). The only change introduced into the method file was to acquire data-dependent triggered MS/MS scans for the five most intense parent ions following each full scan (400–2,000 *m/z*).

Sec-MudPIT Analysis of Complex Protein Mixture from Developing Soybean Seed

Protein pellets obtained from three biological samples for each developmental stage were separately resuspended in IEF extraction solution plus 2%

(w/v) SDS, vortexed at low speed for 30 min at room temperature (RT), and centrifuged at 14,000 rpm for 15 min to remove insoluble material. Supernatant was used to measure protein concentration in triplicate using RediPlate EZQ protein quantification kit (Molecular Probes) and ovalbumin as standard according to the manufacturer's protocol. Five-hundred-microgram total protein was embedded in 12% polyacrylamide in 1.5-mL microcentrifuge tube to perform in-gel digestion with trypsin. A 50-μL polyacrylamide gel was typically made of 28 μL of the protein solution, 15 μL of acrylamide solution (40% acrylamide stock; 39% [w/v] acrylamide and 1% [w/v] bis-acrylamide), 1.5 μL of 1% (w/v) ammonium persulfate, and 0.7 μL of *N,N,N′,N′*-tetramethylethylenediamine. The polymerization reaction was carried out for 40 min at RT. The polymerized gel plug was diced into approximately 1-mm cubes and transferred into new 1.5-mL microcentrifuge tubes. The gel pieces were washed with ammonium bicarbonate solution (100 mM, pH 8.0), reduced with 10 mM dithiothreitol at RT for 1 h, alkylated with 40 mM iodoacetamide in the dark for 1 h at RT, and in-gel digested with sequencing-grade modified trypsin for 20 h at 37°C as described previously (Hajduch et al., 2005).

All samples were reconstituted in 50 μL of 0.1% (v/v) formic acid (FA) in water. Peptides (15 μL; 150 μg) were analyzed on a ProteomeX LTQ workstation (Thermo-Finnigan) using the Sec-MudPIT configuration according to the manufacturer's instructions. Peptides were loaded onto a strong cation exchange resin (BioBasic SCX, 100 × 0.32 mm, 300 Å, 5 μm; Thermo-Finnigan) and eluted with 12 semicontinuous pumped salt (ammonium chloride) solution gradients (5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 125, and 250 mM, respectively) onto peptide traps (C18, 5 × 1 mm; Thermo-Finnigan) for concentrating and desalting prior to final separation by reversed-phase capillary column (BioBasic C18, 100 × 0.18 mm, 300Å, 5 μm; Thermo-Finnigan) using an acetonitrile gradient (0% to 90% [v/v] solvent B in solvent A for a duration of 75 min; solvent A = 0.1% [v/v] FA in water; solvent B = 100% acetonitrile containing 0.1% [v/v] FA). Eluted peptides were ionized with a fused-silica PicoTip emitter (12 cm, 360 μm OD, 75 μm ID, 30-μm tip; New Objective) at ion spray 3.5 kV with 250 nL/min flow rate. Five data-dependent MS/MS scans (isolation width 2 amu, 35% normalized collision energy, minimum signal threshold 500 counts, dynamic exclusion [repeat count, 2; repeat duration, 30 s; exclusion duration, 180 s]) of the five most intense parent ions were acquired in a positive acquisition mode following each full scan (mass range *m/z* 400–2,000).

Database Customization and Indexing

The NCBI (ftp://ftp.ncbi.nih.gov/blast) NR database (as of June 9, 2006) was used for querying all nESI-LC-MS/MS acquired data. The FASTA database utilities and indexer of the BioWorks 3.2SR1 software was used to create a plant database (keywords: Arabidopsis; *Brassica*; *Glycine*; *Medicago*; *Oryza*; and *Zea*) extracted from NCBI NR database. Because BioWorks selects sequences based on the presence of mentioned keywords in their sequence description, some entries were of non-plant origin. Hence, the BioWorks-selected plant database was then subjected to an in-house-built program to filter non-plant sequences from the database. Two types of indexed plant databases were created against trypsin enzyme with static and differential modifications using the FASTA database utilities and indexer of the BioWorks 3.2SR1 software. The indexed plant database I contained Cys as static (carboxyamidomethylation; +57 D) and Met as differential (oxidation; +16 D) modifications. The indexed plant database II contained differential modification on Cys (propionamide; +71.07790).

Database Search

2-DGE- or Sec-MudPIT-based acquired nESI-LC-MS/MS raw data (MS/MS spectra) were searched against the indexed plant database I using BioWorks 3.2SR1 software, which utilizes the SEQUEST algorithm (Eng et al., 1994; Yates et al., 1995) for processing the raw data. The search parameters for this database were set as follows: enzyme: trypsin; number of internal cleavage sites: 2; mass range: 400 to 4,000; threshold: 500; minimum ion count: 35; and peptide mass tolerance: 1.5. To obtain high-confidence and unambiguous protein assignments, the following criteria were applied: (1) a minimum of two NR and nonoverlapping peptides; (2) a peptide correlation score (X_{Corr}) of at least 1.5, 2.0, and 2.5 for +1, +2, and +3 charged ions, respectively; and (3) a peptide probability of equal or lower than 0.05. A peptide probability of 0.05 provides a minimum of 98.5% confidence on the peptide assignment. The BioWorks 3.2SR1 software considers a differential modification of Met on the same peptide sequence as a NR peptide; however, in this study, such peptides have been considered a single NR peptide for the confident assignments of proteins and therefore such protein with a single NR peptide was ambiguous and manually deleted from the output results table. The same search parameters and criteria for protein assignments were also applied for searches against the plant database II. Because a total of 13 raw files were collected for each trypsin-digested complex sample due to an application of 12 pumped salt gradients plus one flow through in the Sec-MudPIT method, 13 Search Results Files resulted from the database search. These 13 Search Results Files were opened together to obtain a single-output multiconsensus Results File using the parameter Load MultiConsensus Results of BioWorks 3.2SR1 software. During this process, all the proteins, peptides, and related information are grouped into one list for the sample. BLAST searching was performed against the current NCBI NR database to update the annotation of unknown or hypothetical proteins from both 2-DGE and Sec-MudPIT studies.

Functional Classification

Functional classifications of identified proteins into main classes were performed as established for Arabidopsis (*Arabidopsis thaliana*; Bevan et al., 1998). However, a number of subclasses were added (or modified) to the main functional classes to present the function of proteins more accurately and to begin developing a more appropriate classification scheme for plant seed. These changes are amino acids and polyamines (in place of amino acids), fatty acids and lipids (in place of lipid and sterol), polysaccharide catabolism, and others within the primary metabolism class; light reactions and Calvin cycle within the photosynthesis subclass, and others within the energy class; seed maturation within the cell growth/division class; allergens within the protein destination and storage class; oil body within the cell structure class; and flavonoids within the secondary metabolism class (Supplemental Tables S1, S2, and S10).

Subcellular Localization

Subcellular localizations of proteins mapped onto metabolic pathways were predicted using three independent programs: TargetP (http://www.cbs.dtu.dk/services/TargetP; Emanuelsson et al., 2000), iPSORT (http://hc.ims.u-tokyo.ac.jp/iPSORT; Bannai et al., 2002), and Predotar, version 1.03 (http://genoplante-info.infobiogen.fr/predotar/predotar.html; Small et al., 2004).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Distribution of soybean proteins without and with 2-DGES according to their main functional categories.

Supplemental Figure S2. Distribution of 2-DGES in the seed-filling proteome of rapeseed.

Supplemental Figure S3. Length of TMD identified in proteins as determined by the THUMBUP method.

Supplemental Figure S4. 2-DGE proteomics approach has unambiguously identified and mapped 675 protein spots on two high-resolution 2-D reference maps.

Supplemental Table S1. Soybean proteins identified during seed development by a combination of 2-DGE and nESI-LC-MS/MS and their functional classification.

Supplemental Table S2. NR soybean proteins identified during seed development using a combination of 2-DGE and nESI-LC-MS/MS according to their functional classification.

Supplemental Table S3. Total protein abundance dataset after normalization to correction constant.

Supplemental Table S4. Total protein abundance dataset after normalization to correction constant and subtraction of SSPs.

Supplemental Table S5. Sec-MudPIT analyses of total protein isolated from 2 WAF (biological replications 1, 2, and 3).

Supplemental Table S6. Sec-MudPIT analyses of total protein isolated from 3 WAF (biological replications 1, 2, and 3).

Supplemental Table S7. Sec-MudPIT analyses of total protein isolated from 4 WAF (biological replications 1, 2, and 3).

Supplemental Table S8. Sec-MudPIT analyses of total protein isolated from 5 WAF (biological replications 1, 2, and 3).

Supplemental Table S9. Sec-MudPIT analyses of total protein isolated from 6 WAF (biological replications 1, 2, and 3).

Supplemental Table S10. Composite table of NR proteins expressed during all five developmental stages of seed development and identified using a combination of Sec-MudPIT and LC-MS/MS and their functional classification.

ACKNOWLEDGMENT

We thank Jianjiong Gao, a graduate student of Prof. Dong Xu (Computer Science Department, University of Missouri, Columbia, MO), for updating the soybean Web site.

Received March 18, 2008; accepted June 12, 2008; published July 3, 2008.

LITERATURE CITED

Agrawal GK, Thelen JJ (2006) Large scale identification and quantification profiling of phosphoproteins expressed during seed filling in oilseed rape. Mol Cell Proteomics **5:** 2044–2059

Agrawal GK, Yonekura M, Iwahashi Y, Iwahashi H, Rakwal R (2005) System, trends and perspectives of proteomics in dicot plants. Part I. Technologies in proteome establishment. J Chromatogr B Analyt Technol Biomed Life Sci **815:** 109–123

Andre C, Froehlich JE, Moll MR, Benning C (2007) A heteromeric plastidic pyruvate kinase complex involved in seed oil biosynthesis in *Arabidopsis*. Plant Cell **19:** 2006–2022

Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S (2002) Extensive feature detection of N-terminal protein sorting signals. Bioinformatics **18:** 298–305

Bevan M, Bancroft I, Bent E, Love K, Goodman H, Dean C, Bergkamp R, Dirkse W, Van Staveren M, Stiekema W, et al (1998) Analysis of 19 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. Nature **391:** 485–488

Brady SM, Long TA, Benfey PN (2006) Unraveling the dynamic transcriptome. Plant Cell **18:** 2101–2111

Carter C, Pan S, Zouhar J, Avila EL, Girke T, Raikhel NV (2004) The vegetative vacuole proteome of *Arabidopsis thaliana* reveals predicted and unexpected proteins. Plant Cell **16:** 3285–3303

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

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol **300:** 1005–1016

Eng J, McCormack AL, Yates JR III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom **5:** 976–989

Fountoulakis M, Gasser R (2003) Proteomic analysis of the cell envelop fraction of *Escherichia coli*. Amino Acids **24:** 19–41

Gallardo K, Firnhaber C, Zuber H, Hericher D, Belghazi M, Henry C, Kuster H, Thompson R (2007) A combined proteome and transcriptome analysis of developing *Medicago truncatula* seeds. Mol Cell Proteomics **6:** 2165–2179

Gallardo K, Le Signor C, Vandekerckhove J, Thompson RD, Burstin J (2003) Proteomics of *Medicago truncatula* seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. Plant Physiol **133:** 664–682

Giavalisco P, Nordhoff E, Kreitler T, Kloppel KD, Lehrach H, Klose J, Gobom J (2005) Proteome analysis of *Arabidopsis thaliana* by two-dimensional gel electrophoresis and matrix-assisted laser desorption/ionization-time of flight mass spectrometry. Proteomics **5:** 1902–1913

Godovac-Zimmermann J, Kleiner O, Brown LR, Drukier AK (2005) Perspectives in spicing up proteomics with splicing. Proteomics **5:** 699–709

Gorg A, Weiss W, Dunn MJ (2004) Current two-dimensional electrophoresis technology for proteomics. Proteomics **4:** 3665–3685

Hajduch M, Casteel JE, Hurrelmeyer KE, Song Z, Agrawal GK, Thelen JJ (2006) Proteomic analysis of seed filling in *Brassica napus*. Developmental characterization of metabolic isozymes using high-resolution two-dimensional gel electrophoresis. Plant Physiol **141:** 32–46

Hajduch M, Ganapathy A, Stein JW, Thelen JJ (2005) A systematic proteomic study of seed filling in soybean: establishment of high-resolution two-dimensional reference maps, expression profiles, and an interactive proteome database. Plant Physiol **137:** 1397–1419

Hardin SC, Tang GQ, Scholz A, Holtgraewe D, Winter H, Huber SC (2003) Phosphorylation of sucrose synthase at serine 170: occurrence and possible role as a signal for proteolysis. Plant J **35:** 588–603

Hennig L, Gruissem W, Grossniklaus U, Kohler C (2004) Transcriptional programs of early reproductive stages in Arabidopsis. Plant Physiol **135:** 1765–1775

Hildebrand DF (1989) Lipoxygenases. Physiol Plant **76:** 249–253

Hill JE, Breidenbach RW (1974) Proteins of soybean seeds. II. Accumulation of the major protein components during seed development and maturation. Plant Physiol **53:** 747–751

Hill MJ (2004) Control of storage-product synthesis in seeds. Curr Opin Plant Biol **7:** 302–308

Jeanneau M, Vidal J, Gousset-Dupont A, Lebouteiller B, Hodges M, Gerentes D, Perez P (2002) Manipulating PEPC levels in plants. J Exp Bot **53:** 1837–1845

Katavic V, Agrawal GK, Hajduch M, Harris SL, Thelen JJ (2006) Protein and lipid composition analysis of oil bodies from two *Brassica napus* cultivars. Proteomics **6:** 4586–4598

Kersten B, Agrawal GK, Iwahashi H, Rakwal R (2006) Plant phosphoproteomics: a long road ahead. Proteomics **6:** 5517–5528

Klein C, Garcia-Rizo C, Bisle B, Scheffer B, Zischka H, Pfeiffer F, Siedler F, Oesterhelt D (2005) The membrane proteome of *Halobacterium salinarum*. Proteomics **5:** 180–197

Koller A, Washburn MP, Lange BM, Andon NL, Deciu C, Haynes PA, Hays L, Schieltz D, Ulaszek R, Wei J, et al (2002) Proteomic survey of metabolic pathways in rice. Proc Natl Acad Sci USA **99:** 11969–11974

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol **157:** 105–132

Le BH, Wagmaister JA, Kawashima T, Bui AQ, Harada JJ, Goldberg RB (2007) Using genomics to study legume seed development. Plant Physiol **144:** 562–574

Lee JM, Williams ME, Tingey SV, Rafalski JA (2002) DNA array profiling of gene expression changes during maize embryo development. Funct Integr Genomics **2:** 13–27

Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR III (1999) Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol **17:** 676–682

Lu X, Zhu H (2005) Tube-gel digestion: a novel proteomics approach for high throughput analysis of membrane proteins. Mol Cell Proteomics **4:** 1948–1958

Mattow J, Siejak F, Hagens K, Schmidt F, Koehler C, Treumann A, Schaible UE, Kaufmann SH (2007) An improved strategy for selective and efficient enrichment of integral plasma membrane proteins of mycobacteria. Proteomics **7:** 1687–1701

Mechin V, Thevenot C, Li Guilloux M, Prioul JL, Damerval C (2007) Developmental analysis of maize endosperm proteome suggests a pivotal role for pyruvate orthophosphate dikinase. Plant Physiol **143:** 1203–1219

Mienke DW, Chen J, Beachy RN (1981) Expression of storage-protein genes during soybean seed development. Planta **153:** 130–139

Miernyk JA, Dennis DT (1983) Mitochondrial, plastid, and cytosolic isozymes of hexokinase from developing endosperm of Ricinus communis. Arch Biochem Biophys **226:** 458–468

Nagele E, Vollmer M, Horth P (2004) Improved 2D nano-LC/MS for proteomics applications: a comparative analysis using yeast proteome. J Biomol Tech **15:** 134–143

Norton G, Harris JF (1975) Compositional changes in developing rape seed (*Brassica napus* L.). Planta **123:** 163–174

Ohlrogge JB, Kuo TM (1984) Control of lipid synthesis during soybean seed development: enzymic and immunochemical assay of acyl carrier protein. Plant Physiol **74:** 622–625

Plaxton WC (1996) The organization and regulation of plant glycolysis. Annu Rev Plant Physiol Plant Mol Biol **47:** 185–214

Plaxton WC, Podesta FE (2006) The functional organization and control of plant respiration. Crit Rev Plant Sci **25:** 159–198

Porta H, Rocha-Sosa M (2002) Plant lipoxygenases. Physiological and molecular features. Plant Physiol **130:** 15–21

Ripp KG, Viitanen PV, Hitz WD, Franceschi VR (1988) Identification of membrane protein associated with sucrose transport into cells of developing soybean cotyledons. Plant Physiol **88:** 1435–1445

Rubel A, Rinne RW, Canvin DT (1972) Protein, oil, and fatty-acid in developing soybean seeds. Crop Sci **12:** 739–741

Ruuska SS, Girke T, Benning C, Ohlrogge JB (2002) Contrapuntal networks of gene expression during *Arabidopsis* seed filling. Plant Cell **14:** 1191–1206

Santoni V, Molloy M, Rabilloud T (2000) Membrane proteins and proteomics: un amour impossible? Electrophoresis **21:** 1054–1070

Schwender J, Goffman F, Ohlrogge JB, Schachar-Hill Y (2004a) RuBisCO without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature **432:** 779–782

Schwender J, Ohlrogge JB, Schachar-Hill Y (2004b) Understanding flux in plant metabolic networks. Curr Opin Plant Biol **7:** 309–317

Siedow JN (1991) Plant lipoxygenase: structure and function. Annu Rev Plant Physiol Plant Mol Biol **42:** 145–188

Small I, Peeters N, Legeai F, Lurin C (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics **4:** 1581–1590

Smith CWJ, Patton JG, Nadal-Ginard B (1989) Alternative splicing in the control of gene expression. Annu Rev Genet **23:** 527–577

Tang GQ, Hardin SC, Dewey R, Huber SC (2003) A novel C-terminal proteolytic processing of cytosolic pyruvate kinase, its phosphorylation and degradation by the proteasome in developing soybean seeds. Plant J **34:** 77–93

Thelen JJ, Muszynski MG, Miernyk JA, Randall DD (1998) Molecular analysis of two pyruvate dehydrogenase kinases from maize. J Biol Chem **273:** 26618–26623

Tripodi KE, Turner WL, Gennidakis S, Plaxton WC (2005) In vivo regulatory phosphorylation of novel phospho*enol*pyruvate carboxylase isoforms in endosperm of developing castor oil seeds. Plant Physiol **139:** 967–978

Vensel WH, Tanaka CK, Cai N, Wong JH, Buchanan BB, Hurkman WJ (2005) Developmental changes in the metabolic protein profiles of wheat endosperm. Proteomics **5:** 1594–1611

Wan L, Ross ARL, Yang J, Hegedus DD, Kermode AR (2007) Phosphorylation of the 12S globulin cruciferin in wild-type and abi1-1 mutant *Arabidopsis thaliana* (thale cress) seeds. Biochem J **404:** 247–256

Washburn MP, Wolters D, Yates JR III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol **19:** 242–247

Weber H, Borisjuk L, Wobus U (2005) Molecular physiology of legume seed development. Annu Rev Plant Biol **56:** 253–279

White JA, Todd J, Newman T, Focks N, Girke T, de Ilárduya OM, Jaworski JG, Ohlrogge JB, Benning C (2000) A new set of *Arabidopsis* expressed sequence tags from developing seeds. The metabolic pathway from carbohydrates to seed oil. Plant Physiol **124:** 1582–1594

Yamakawa H, Hirose T, Kuroda M, Yamaguchi T (2007) Comprehensive expression profiling of rice grain filling-related genes under high temperature using DNA microarray. Plant Physiol **144:** 258–277

Yates JR III, Eng JK, McCormack AL, Schieltz D (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem **67:** 1426–1436

Yu YQ, Gilar M, Lee PJ, Bouvier ESP, Gebler JC (2003) Enzyme-friendly, mass spectrometry-compatible surfactant for in-solution enzymatic digestion of proteins. Anal Chem **75:** 6023–6028

Zhang L, Xie J, Wang X, Liu X, Tang X, Cao R, Hu W, Nie S, Fan C, Liang S (2005) Proteomic analysis of mouse liver plasma membrane: use of differential extraction to enrich hydrophobic membrane proteins. Proteomics **5:** 4510–4524

Zhu T, Budworth P, Chen W, Nicholas P, Chang HS, Guimil S, Su W, Estes B, Zou G, Wang X (2003) Transcriptional control of nutrient partitioning during rice grain filling. Plant Biotechnol J **1:** 59–70