Journal of Integrative Plant Biology 2009, **51** (2): 207–223

Analysis of the *Arabidopsis* Floral Proteome: Detection of over 2 000 Proteins and Evidence for Posttranslational Modifications

Baomin Feng[1], Lianchao Li[2], Xiaofan Zhou[3], Bruce Stanley[4] and Hong Ma[1*]

([1]*Department of Biology, the Pennsylvania State University*, University Park, PA 16802, USA;
[2]*Proteomics and Mass Spectrometry Core Facility, 003 Althouse Laboratory, the Pennsylvania State University*, University Park, PA 16802, USA;
[3]*The Intercollege Graduate Program in Cell and Developmental Biology, the Huck Institutes of the Life Sciences, the Pennsylvania State University*, University Park, PA 16802, USA;
[4]*Section of Research Resources H093, the College of Medicine, the Pennsylvania State University*, Hershey, PA 17033, USA)

Abstract

The proteome of the *Arabidopsis* flower has not been extensively studied previously. Here, we report a proteomic analysis of the wild type *Arabidopsis* flower. Using both two-dimensional electrophoresis/mass spectrometry (2-DGE/MS) and multi-dimensional protein identification technology (MudPIT) approaches, we identified 2 446 proteins. Although a single experiment or analysis uncovered only a subset of the proteins we identified, a combination of multiple experiments and analyses facilitated the detection of a greater number of proteins. When proteins are grouped according to RNA expression levels revealed by microarray experiments, we found that proteins encoded by genes with relatively high levels of expression were detected with greater frequencies. On the other hand, at the level of the individual gene/protein, there was not a good correlation between protein spot intensity and microarray values. We also obtained strong evidence for post-translational modification from 2-DGE and MudPIT data. We detected proteins that are annotated to function in protein synthesis, folding, modification, and degradation, as well as the presence of regulatory proteins such as transcription factors and protein kinases. Finally, sequence and evolutionary analysis of genes for active methyl group metabolisms suggests that these genes are highly conserved. Our results allow the formulation of hypotheses regarding post-translational regulation of proteins in the flower, providing new understanding about *Arabidopsis* flower development and physiology.

Key words: *Arabidopsis*; floral; methylation; multi-dimensional protein identification technology; protein modification; proteome.

Feng B, Li L, Zhou X, Stanley B, Ma H (2009). Analysis of the *Arabidopsis* floral proteome: detection of over 2 000 proteins and evidence for posttranslational modifications. *J. Integr. Plant Biol.* **51**(2), 207–223.

Available online at www.jipb.net

The proteome represents the complete set of proteins that are present in an entire organism, a specific organ, tissue, cells or even sub-cellular compartments (Jorrin et al. 2007). Proteins play crucial roles in most cellular processes, as fundamental structural components and catalysts of most biochemical reactions, including metabolism, signaling, and regulation. Meanwhile their modifications provide a very dynamic and flexible mechanism for regulating biological processes. Therefore, investigating the proteome is a key to better understanding the mechanism of cellular processes. However, due to the variable physical and chemical properties of proteins, the need for sufficient samples, and other difficulties of protein characterization, the analysis of the proteome has been more challenging than that of the transcriptome.

Recently, rapid advances in techniques for protein separation and identification have created opportunities for deeper investigation of the proteome. Two-dimensional gel electrophoresis (2-DGE) separates proteins according to molecular weight and iso-electric point with relatively high precision (Jorrin et al. 2007). In addition, the abundance of proteins analyzed using 2-DGE

Received 9 Jun. 2008 Accepted 11 Jun. 2008

Supported by a Grant from the US Department of Energy (DE-FG02-02ER15332) to H. Ma, and by funds from the Huck Institutes of the Life Sciences and the Biology Department at the Pennsylvania State University.

*Author for correspondence.

Tel: +1 814 863 6414;

Fax: +1 814 863 1357;

E-mail: <hxm16@psu.edu>.

doi: 10.1111/j.1744-7909.2008.00787.x

can be estimated using staining intensity on the gel. Proteins separated by 2-DGE can be further analyzed by mass spectrometry (MS) for identification. Other separation techniques such as liquid chromatography (LC) have been used to separate small peptides (Washburn et al. 2001; Delahunty and Yates 2007). This is the basis for a whole-proteome analysis called multidimensional protein identification technology (MudPIT). In this method, peptides generated by trypsin digestion are separated by two- or multiple-dimension LC and their masses determined by MS (Florens and Washburn 2006; Ru et al. 2006; Lee et al. 2007; Maor et al. 2007). Tandem MS/MS methods such as those carried out by mass spectrometers called matrix assisted laser desorption/ionization-time of flight (MALDI-TOF/TOF) or quadrupole TOF (Q-TOF) can determine both peptide masses and the specific fragmentation patterns of peptides, which can reliably identify proteins when combined with proper statistical estimates of false discovery rates. Today with the complete genome sequences of model organisms, large scale protein identification can be achieved by matching the peptide mass spectra with those predicted from the sequence database, including protein modifications (Arabidopsis Genome Initiative [AGI] 2000). These techniques and resources allow new opportunities to investigate the proteome, such as studies of the plant proteome from various tissues and cells or after treatment (Gallardo et al. 2001; Peck et al. 2001; Carter et al. 2004; Jarvis 2004; Nuhse et al. 2004; Holmes-Davis et al. 2005; Majeran et al. 2005; Peck 2005; Bayer et al. 2006; Chibani et al. 2006; Rajjou et al. 2006; Lee et al. 2007; Thelen and Peck 2007). However, the *Arabidopsis* floral proteome has not been extensively studied previously.

The *Arabidopsis* flower has four types of floral organs (sepals, petals, stamens and pistils). The *Arabidopsis* flower development has been divided into 12 stages from the initiation of the floral meristem to the opening of the flower (Smyth et al. 1990). From stages 1–9, a number of key developmental events take place, such as the initiation of floral organ primordia, specification of organ type, meiosis and gametophyte initiation. At stage 12, all floral organs are mature, followed by later stages with pollen germination, double fertilization, and the early embryo development (Smyth et al. 1990). Our previous transcript profiling using microarray indicated that there are approximately 14 000 genes expressed in the stage-12 flower (Zhang et al. 2005b). In the last two decades, dozens of genes, such as those determining flowering time and floral organ identity, have been characterized by molecular genetic studies (Ma 2005; Zahn et al. 2006). Those genetically identified genes represent only a small portion of the genes that are believed to be involved in flower development.

After the completion of mRNA biogenesis, there are additional levels of regulation that affect the quantity and activity of the protein products. Translational control, protein modification and turnover can all contribute to the regulation of protein abundance and activities. Such information can only be obtained by studies at the protein level. In yeast and mammals, although there is a general correlation between mRNA and protein levels, a substantial number of genes exhibit a discrepancy between mRNA and protein abundance (Futcher et al. 1999; Gygi et al. 1999; Pradet-Balade et al. 2001; Griffin et al. 2002; Schmidt et al. 2007). Because little is known about protein levels and protein modifications in the flower, we are interested in obtaining information about floral proteins using a proteomics approach.

Furthermore, we would like to investigate protein modifications that are present in the flower. The critical roles of protein modifications have been suggested by previous genetic studies. For example, the *CLV1* gene is required to determine the size of floral meristem and encodes a receptor-like protein kinase, suggesting that it functions by phosphorylating downstream proteins (Clark et al. 1996; Brand et al. 2000; Schoof et al. 2000; DeYoung et al. 2006). Also, protein ubiquitination can affect the stability and localization of proteins, thus affecting their biological functions. Defects in ubiquitination pathways can result in abnormal floral organ identity and meiosis as suggested by the phenotypes caused by a mutation in *ASK1*, which encodes a key component of SKP1/Cullin/F-box (SCF) complexes that facilitate ubiquitination of target proteins (Samach et al. 1999; Yang et al. 1999a; Zhao et al. 1999; Zhao et al. 2001; Ni et al. 2004). However, there is little knowledge about protein modification in the flower.

In this study, we used 2-DGE and MS to identify 216 proteins in the flower proteome at stage 12. We also found some proteins at two or more spots by 2-DGE, suggesting that these proteins had alternative forms possibly due to modifications. In addition, we identified nearly 2 400 floral proteins from two MudPIT experiments. Moreover, we recovered further evidence for modified peptides using the MudPIT analyses. Additionally, our proteomic data suggest that methylation pathways play an important role in the flower. In general, relatively large percentages of protein products were identified for genes with high detected mRNA levels, particularly by two or more experiments. Although low percentages of proteins encoded by genes with relatively low levels of mRNA were identified by proteomics, we were able to detect proteins that correspond to genes with very low microarray intensities. Our analyses indicate that proteomics is an informative approach to study protein properties in a high throughput manner.

Results

Identification of floral proteins by 2-DGE/MS and MudPIT

To investigate the *Arabidopsis* floral proteome, we first used 2-DGE electrophoresis and mass spectrometry to analyze proteins extracted with trichloroacetic acid (TCA) from stage 12 floral buds, as defined by Smyth et al. (1990), in three biological

replicates. Two dimension gel electrophoresis was carried out in two steps: isoelectric focusing (first dimension) and sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) (second dimension). After preliminary 2-DGE experiments, we chose to use pH 4–7 immobilized pH gradient (IPG) (gel strips for the first dimension because this range can resolve most protein spots than can be resolved by a pH 3–10 IPG strip and other single strips). On the 2-DGE, we were able to detect more than 300 spots by staining with Coomassie blue R-250 (Figure 1; other gels not shown); all three replicates showed very similar spot patterns. A total of 265 spots were cut from the 2-DGE for protein identification by mass spectrometry and database search using the Mascot software (Perkins et al. 1999). Among these spots, at least one protein was identified from each of 185 spots by Q-TOF MS analysis (Supplementary Table 1). A subset of identified proteins was further tested and confirmed by MALDI-TOF/TOF MS analysis. In total, 216 non-redundant proteins were identified from the 185 spots on 2-DGE by using MALDI-TOF/TOF and/or Q-TOF MS/MS experiments.

Although 2-DGE provided estimates of relative protein abundance, molecular weight, and pI, the number of proteins identified using 2-DGE was relatively small. To sample the flower proteome more deeply, we also used the LC-MALDI MudPIT approach. We carried out two independent MudPIT experiments using two biological replicates of stage 12 flowers. The flowers were ground and extracted with TCA/acetone, then about 1 mg of the TCA/acetone extracted crude protein powder was digested with trypsin and analyzed by using MudPIT. The procedure for MudPIT analysis involves the generation, separation, and detection of a large collection of peptides from the entire proteome. This process results in the detection of highly abundant proteins with relatively high probability. Therefore, MudPIT samples a subset of all peptides generated from the proteome, rather similar to EST (expressed sequence tag) analysis. In contrast, MudPIT is unlike microarray analysis, which can reproducibly detect gene expression above some threshold levels. To detect a maximum number of proteins with limited flower sample and resources, we analyzed MS data from

Figure 1. A 2-D gel image for identification of 216 proteins from *Arabidopsis* floral tissues.

Proteins were separated in pH range of 4–7. Each spot identified is labeled with spot number; the spots were analyzed after either hand cutting (such as 1, 2, 3) or robot cutting (such as 1A01, 2C05, 1D11).

the two MudPIT experiments using two databases (SwissProt or National Center for Biotechnology Information (NCBI)) and two algorithms (Mascot or ProteinPilot) (Shilov et al. 2007). The combination of all of the analyses of both MudPIT experiments resulted in the identification of 2 397 proteins (Supplementary Table 2). Among these, 1 552 proteins were identified with the data from the first MudPIT experiment and 2 217 proteins were identified from the second experiment, with an overlap of 1 372 proteins identified by both experiments.

Multiple sampling and bioinformatic analyses increase protein detection

Among the 216 proteins identified by 2-DGE/MS analysis, 167 (77%) were identified by at least one of the MudPIT experiments (Figure 2), indicating that as few as two MudPIT experiments can detect most, albeit not all, proteins uncovered by 2-DGE analysis. As mentioned above, many proteins were detected by only one of the two MudPIT experiments, as observed by other scientists, including researchers in the John Yates III lab where original MudPIT procedures were developed (Washburn et al. 2001; Washburn et al. 2003; Delahunty and Yates 2007). Although identifications of the most abundant proteins tend to show good overlap between replicates, each experiment is only able to sample a fraction of the peptides that are in a complex mixture, and the less abundant proteins that are identified tend to differ between experiments. This has led to the estimation that one would have to run at least nine replicates to have identified 95% of all peptides detectable by MS at least once, with each additional replicate experiment identifying additional peptides/proteins (Washburn et al. 2003; Delahunty and Yates 2007).

To obtain further understanding of the MudPIT results, we sorted the identified proteins on the basis of whether a protein was detected (i) in one or the other of the MudPIT experiments; (ii) in searches with either the SwissProt or NCBI databases; and (iii) using either the Mascot or Paragon (ProteinPilot) algorithms, resulting in eight sets of protein identifications (Supplementary Table 2). When we compared these eight sets of proteins, we found that similar to the partial overlap between the results of the two MudPIT experiments, the two databases also supported the detection of partially overlapping protein sets, as did the two algorithms (Figure 2). Therefore, the combination of multiple analyses of the same MS/MS data by different algorithms yielded a greater number of protein identifications than any single analysis of one database with one algorithm. In fact, such multiple-algorithm – searches have been advocated as a better approach to MS/MS database searching by other investigators (Resing et al. 2004).

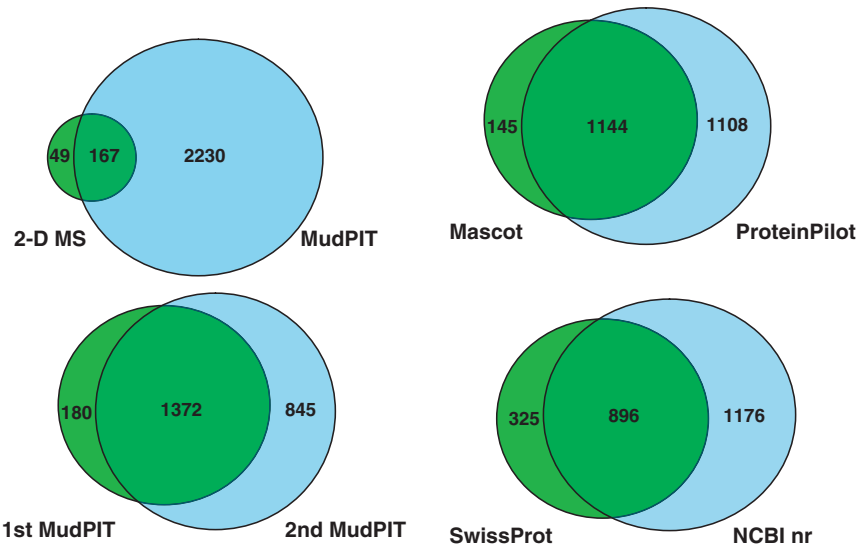

Figure 2. Comparisons of proteins identified by different approaches and different analyses.

The number at each area represents the number of proteins identified. Upper left, overlap between the proteins identified by two-dimensional electrophoresis/mass spectrometry (2-DGE/MS) and those from the combined set of two multi-dimensional protein identification technology (MudPIT) experiments. Lower left, a comparison between two MudPIT experiments. Upper right, protein identifications by the Mascot algorithm from both MudPIT experiments using two protein databases (SwissProt and National Center for Biotechnology Information (NCBI)) was compared with the proteins identified by the ProteinPilot algorithm with the same two experiments using the same two databases. Lower right, protein identifications by searching the SwissProt database with data from two MudPIT experiments using two algorithms (Mascot and ProteinPilot) was compared with the same analysis by searching the NCBInr database.

Because the proteins identified by MudPIT analyses are likely to result from a sampling process, in which more abundant proteins are represented by more peptide spectra than are less abundant proteins, we wondered what might be the relationship between the level of gene expression and the frequency of detection of the corresponding protein. We divided the *Arabidopsis* genes into categories according to the microarray data that we previously obtained for expression in the flower using the Affymetrix ATH1 gene chip (Zhang et al. 2005b). We identified those proteins that are encoded by the genes in each category, and calculated the percentage of proteins detected by our proteomics experiments. We found that the general trend is that protein products of genes with higher expression levels tend to be detected at a higher frequency, particularly by two or more experiments (Figure 3). However, even for genes expressed at very high levels, not all encoded proteins were identified by the 2-DGE/MS and MudPIT experiments. At the same time, for the very large number of genes with relatively low values of microarray data, even a small percentage represented the detection of many proteins. Specifically, some of the identified proteins correspond to genes whose expression was below the level of reliable detection by the microarray experiment. This might be in part due to a lack of a good correlation between mRNA level and protein levels (see below), as previously observed for yeast and in human kidney cell lines, where observed changes in mRNA level can account for no more than 40% of the observed changes (up or down) in protein levels (Gygi et al. 1999; Griffin et al. 2002).

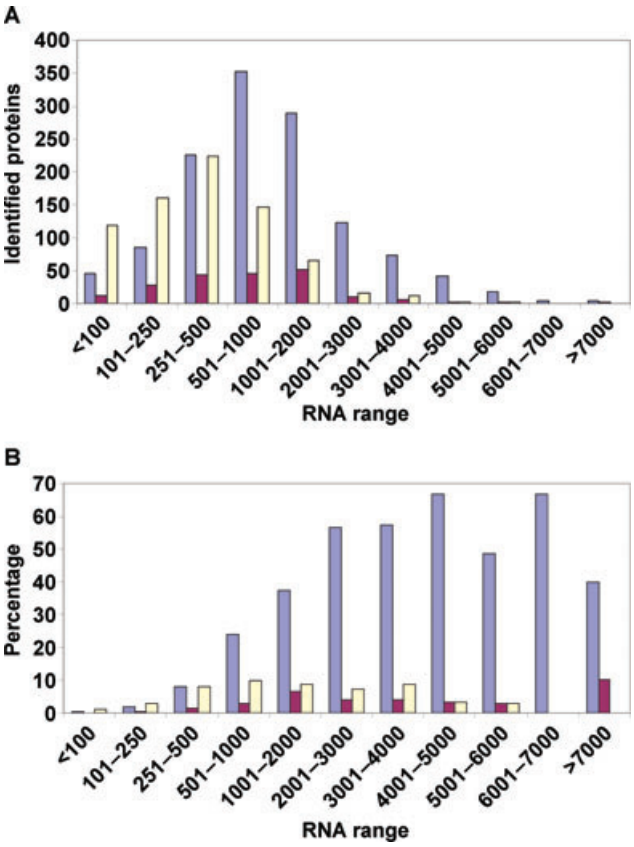

Figure 3. The distribution of proteins identified by multi-dimensional protein identification technology (MudPIT) at different RNA ranges. The RNA abundance was determined by ATH1 microarray experiments.

(**A**) The actual numbers of proteins identified at different RNA ranges were compared among three types of identifications: the identifications by both MudPIT experiments (blue bar), by only one MudPIT and two or more analysis (purple bar), and by only one analysis (yellow bar).
(**B**) The percentage of proteins identified at different RNA ranges were compared as above (colors indicate the same analyses). The percentage is taken by dividing the number of identified protein with the total number of genes of which the mRNA is present as suggested by microarray.

Comparison between protein and transcript levels

Although genome-wide mRNA profiling techniques such as microarray experiments can estimate relative gene expression levels, they do not provide information regarding the protein level. In addition, the above described relatively high protein detection frequencies corresponding to genes with high microarray values suggest that genes with high transcript levels tend to have high protein levels. However, it is not clear at the individual gene/protein level, whether there is a good correlation between mRNA and protein levels. Therefore, we have examined the 2-DGE spot intensity using the PDQuest software, and compared the spot intensity with the microarray value of the corresponding gene.

As mentioned before, sometimes two or more proteins were identified from the same 2-DGE spot. In those cases, we used the staining intensity from the PDQuest software as a crude estimate for the protein with the highest ion score and peptides number detected by MS. If one protein was detected in two or more spots, we used the sum of the intensities from all spots with the protein for an estimate of the total quantity of the protein. We compared the estimated protein amounts with microarray data and found a poor correlation ($R^2 = 0.064\ 2$) between protein and RNA. Because the estimates for proteins found in the same spots were clearly inaccurate, we also compared the result for 2-DGE spots that only had a single detected protein identified; the resulting correlation was still very poor ($R^2 = 0.033\ 9$) (Figure 4). A possible explanation for poor correlation is the presence of closely related proteins, which might also contribute to the same spots.

Protein modifications were suggested by 2-D image and MudPIT

Protein modifications can affect the conformation of proteins, thereby affecting protein activities and stability. To obtain

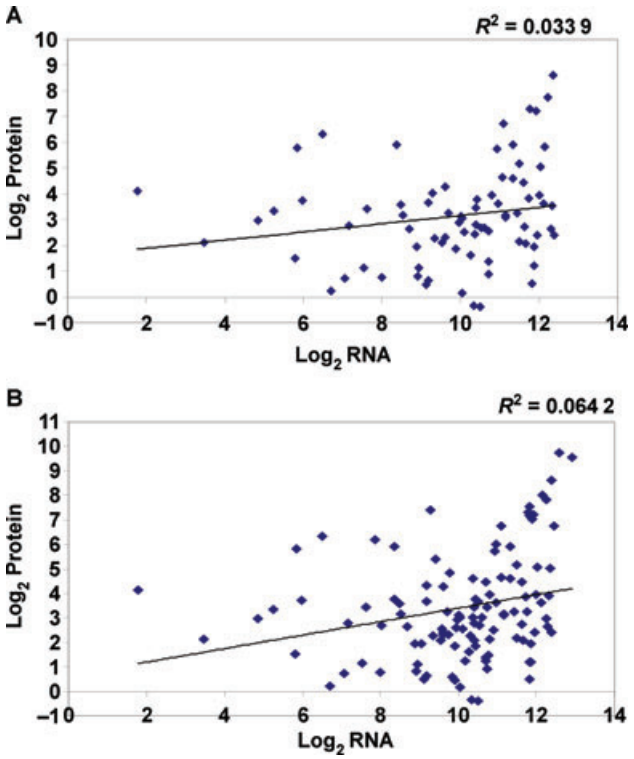

Figure 4. Comparison between the RNA and protein abundances.

(**A**) The log ratio of RNA (detected by microarray) and the log ratio of protein abundance (spot intensity detected by 2-D gel and normalized with molecular weight) were compared for the spots with single protein identified.

(**B**) A similar comparison was done by using all spots from 2-D gel image. For the spots with multiple proteins identified, the spot intensity was used as an estimation of the highest represented protein in that spot (highest identified peptide number and ion score value).

evidence for protein modification at a global scale, we took two approaches. First, we examined the proteins identified from 2-DGE/MS and found that 30 proteins were detected in two or more 2-DGE spots (Table 1), suggesting that, for each of these proteins, at least one spot potentially represented a modified form of the protein, although changes *in vitro* cannot be ruled out without further analysis. Among these 30 proteins, some were found at two spots with similar molecular weights but different pI values. For other proteins, the two spots had close pI values but different molecular weights; others corresponded to spots that differed in both molecular weight and pI. In addition, some spots yielded multiple proteins, indicating that several proteins migrated to a small area of the 2-DGE. This is not likely due to a failure to focus proteins in general, because the majority of the protein spots resulted in the identification of a single protein per spot; however, with an estimated 15 000 to 20 000 proteins expressed in any cell type, and even more protein forms expected when posttranslational modifications are included, it is little wonder that at least a good number of the 1 000–2 000 spots visible on a 2D gel would contain more than one protein. Although some proteins from the same spot showed expected molecular weight and pI, in more cases, proteins with different expected molecular weights and pI values migrated to a same position on the 2-DGE, again suggesting that some of these proteins were modified (Table 1).

Our second approach to obtaining evidence for protein modifications relied on the MudPIT results, which revealed many types of peptide modifications with statistical significance (C=confidence> 99%, estimated false discovery rate below 5%). Among these, several could have occurred *in vivo*: including methylation, acetylation, formylation, and palmitylation (Table 2); we have also found some phosphor-peptides, but these had low confidence values (ion scores) and is not further described here. Other modifications, such as oxidation, are likely to have occurred during sample preparation and are not represented here. Altogether, 237 modified peptides representing 119 proteins were detected with confidence values of >99% (Table 2), suggesting that these proteins were modified. This provides valuable information for hypothesizing potential regulation of the proteins. Interestingly, among the 30 proteins that were detected at two or more spots on the 2-DGE, 12 proteins also had peptide modifications from the MudPIT analysis, further supporting the idea that these proteins were modified.

Putative functions of the detected floral proteins

To obtain clues about possible functions of the proteins identified by our proteomic experiments, we examined their GO (gene ontology) annotations (Figure 5) (Rhee et al. 2003; Poole 2007). Compared with the GO distributions of the predicted *Arabidopsis* proteome, we found that the percentage of hydrolases, structural proteins and proteins with other enzyme activities (not clearly predicted) are significantly increased. On the other hand, transcription factors, receptors and other molecular functions were under-represented compared with the whole genome. Nevertheless, our results indicate that the proteomic approaches were able to detect some of these putative regulatory proteins.

To further analyze the proteomic data, we looked for biochemical pathways that are significantly represented by the proteomic data using the Aracyc database on the the Arabidopsis Information Resource (TAIR) website (Table 3) (Zhang et al. 2005a). Many of the detected pathways are for amino acid synthesis, the TCA cycle, and photosynthesis. In addition, components of methylation pathways and one-carbon unit metabolism are significantly enriched, including a complete set of the SAM (S-adenosyl methionine) pathway components, and many methyltransferases. The importance of methylation is also supported by the detection of many methylated peptides. Furthermore, enzymes for farnesylation and geranlygeranylation were

Table 1. Protein with multiple spots on 2-D and modifications detected by multi-dimensional protein identification technology (MudPIT)

AGI	Spot	Modification shown by MudPIT
AT1G01050	22, 23	
AT1G04410	14, 1C03	Methyl(E)@2
AT1G07890	23, 1C01	
AT1G19570	38, 1B07	
AT1G53240	10, 1G04	Trimethyl(K)@3; Dethiomethyl(M)@10
AT1G67090	68, 29, 67, 79	Formyl@N-term; Methyl(L)@5; Methyl(S)@6; Methyl(Q)@2
AT2G21660	60, 1B05	Methyl(S)@8; Methyl(S)@13, Protein Terminal Acetyl@N-term; Dimethyl(R)@16, Formyl@N-term; Dimethyl(R)@9
AT2G24270	55-1, 11, 55-3, 1F11	Protein Terminal Acetyl@N-term; Formyl@N-term
AT2G37220	17, 18	Acetyl(S)@20; Arg->GluSA(R)@23; Formyl@N-term
AT2G39730	1A02, 3, 56-58, 1E08	Formyl@N-term
AT2G46860	22, 23	
AT2G47470	23, 1B11	
AT2G47730	65, 1G11	
AT3G09200	61, 2C01	
AT3G25050	15, 1A04	Formyl@N-term; Phospho(S)@8;
AT3G52930	1C04, 2B07	
AT3G53460	60, 61	
AT3G55440	23, 28, 2C02	
AT4G03280	48, 1A06	Formyl@N-term
AT4G04460	64, 1G10	
AT4G08390	63-1, 63-2	
AT4G18480	79, 56-60	
AT4G31300	25, 80	
AT4G38680	28, 29, 80	
AT4g38970	5, 1F03	Trimethyl(K)@6
AT5G08690	52, 56-60, 1B08	Dimethyl(K)@18, OxidAtion(P)@2; Methyl(S)@3; Formyl(K)@12
AT5G38410	67, 50, 73, 68	Formyl(K)@1, Formyl@N-term
AT5G48480	36, 60	Protein Terminal Acetyl@N-term
AT5G66570	62, 1A11, 2B08, 2B09	Formyl@N-term
ATCG00490	12, 16, 19, 78, 11, 23, 31, 32, 60, 61, 1A05, 1B02, 1B12, 1C02, 1D10, 1F06, 2B11, 1E04	Methyl(L)@10; Methyl(H)@2; Methyl(D)@5 Methyl(H)@7 Acetyl@N-term; Methyl(T)@3; Methyl(H)@3 Methyl(D)@6; Hex(1)HexNAc(1)(T)@9; Methyl(L)@5; Formyl@N-term; Methyl(R)@11 Methyl(N)@6; Methyl(S)@7 Methyl(T)@1

also found, suggesting that lipid modifications of proteins might be important for flower development (Supplementary Table 3) (Pei et al. 1998; Yalovsky et al. 2000; Caldelari et al. 2001; Brady et al. 2003).

Our results also showed enrichment for proteins involved in protein synthesis, folding and degradation. A large number of ribosomal proteins were identified, including subunits of cytosolic and chloroplast ribosomes. In addition, chaperones important for proper folding and unfolding of proteins were uncovered, suggesting that they might be important in floral tissues. The components of ubiquitin-dependent proteasome pathway for protein degradation were also highly represented. Besides proteasome subunits, floral proteins involved in ubiquitination included ubiquitin and ubiquitin-like proteins, ubiquitin-specific proteases, E1-ubiquitin activation enzyme, E2-ubiquitin conjugation enzyme, and subunits of ubiquitin ligase such as SKP1 proteins (ASK1 and ASK2) and some F-box proteins.

We also detected several types of structural proteins, including cytoskeleton and motor proteins such as tubulins, actins, dynamin, and kinesins. It was previously thought that membrane proteins are hard to extract and identify. Our results showed that the proteomic procedures were able to detect membrane associated transporters, proteins in the endo-membrane systems and associated with vesicle (SNARE, clathrin etc.). Moreover, proteins with putative functions in signaling processes, such as receptors, GTPases and kinases/phosphotases, and in protein–protein interactions were also found. Another group of proteins are those involved in the synthesis and signal transduction of hormones such as auxin, brassinosteroids, jasmonic acid (JA) and ethylene.

Table 2. Number of peptide modifications detected by multi-dimensional protein identification technology (MudPIT) with 99% confidence

Type of peptide modifications	Number of peptides modifications
Formylation	52
Acetylation	7
Trimethylation	7
Dimethylation	24
Methylation	141
HexNAC	5
Myristoylation	3

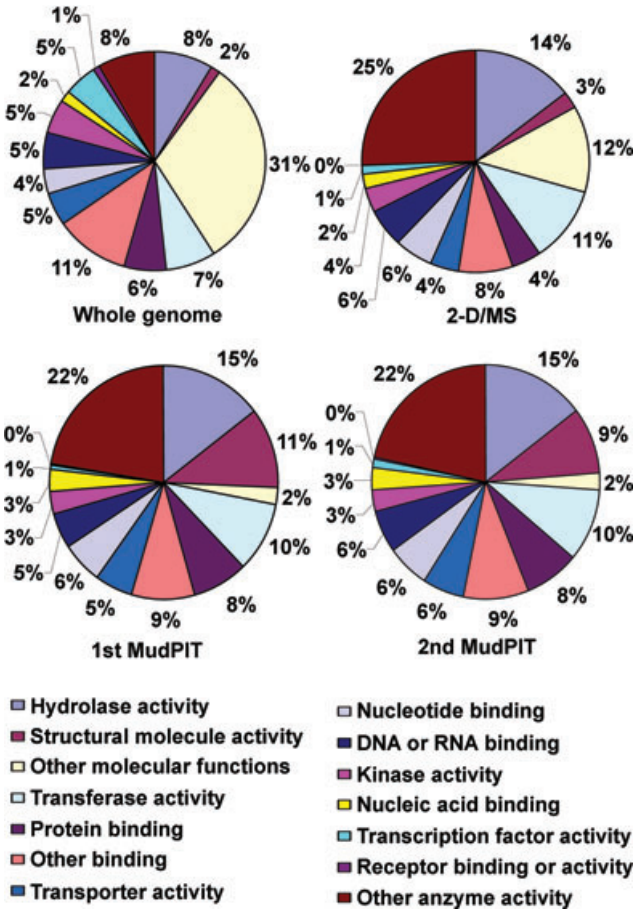

Figure 5. Gene ontology (GO) categories of the proteins identified by two-dimensional electrophoresis/mass spectrometry (2-DGE/MS) and multi-dimensional protein identification technology (MudPIT) .

The GO categories were based on molecular function as suggested by The Arabidopsis Information Resource (TAIR) website.

Table 3. Important categories of proteins identified by multi-dimensional protein identification technology (MudPIT)

Category	Molecular function	Number of proteins identified
Proteolysis	UBQ	8
	E1	4
	E2	5
	F-box	2
	SKIP	2
	Cullin	1
	20S proteasome	7
	26S proteasome	16
	Proteases	21
Ribosome	30S subunits	4
	50S subunits	4
	40S subunits	58
	60S subunits	91
Membrane	Plasma and other membrane transporters/intergral proteins (ABC transporter, aquaporin, ion changel or transporter (H+, K+, Ca2+), organic molecue transporter etc.)	58
	Nucler pore, importin etc.	4
	Vacuole	6
	Golgi, Vesicle transport	19
Cytoskeleton and motors	Actin	9
	Tubulin	13
	Dynamin	6
	Kinesin	3
	Myosin	5
	Villin	2
Kinase/phosphotase	Receptor like	5
	MAP Kinases	4
	PI pathway kinases	3
	Other kinases	14
	Phosphotases	8
Hormone related	Auxin response	4
	ABA	2
	Ethylene	2
	JA	1
Important known floral regulators	Transcriptional factor	5
	Small RNA biogenesis	2
	Light signalling/protein degradation	3
	Receptor like kinase	1

ABA, abscisic acid; JA, jasmonic acid; MAP, mitogen-activating protein; PI, phosphatidylinositol signal transduction; SKIP1, suppressor of kinetochore protein 1; UBQ, ubiquitin.

Finally, it is worth noting that we detected a number of proteins with genetically defined functions in controlling light signaling and flowering time (CONSTANS, COP1, COP8) (Deng et al. 1991; Chamovitz et al. 1996; Valverde et al. 2004; Liu et al. 2008), floral organ identity (AP2, SUP) (Kunst et al. 1989; Drews et al. 1991; Sakai et al. 1995; Jacobsen and Meyerowitz 1997; Wurschum et al. 2006), meristem size and floral organ number (ASK1, CLV1, DCL1, SUP) (Clark et al. 1997; Yang et al. 1999a; Brand et al. 2000; Zhao et al. 2001; Park et al. 2002; Zhao et al. 2003; Ni et al. 2004; Ogawa et al. 2008), organ shape and polarity (SERRATE) (Lobbes et al. 2006; Yang et al. 2006) and early anther development (SPL) (Schiefthaler et al. 1999; Yang et al. 1999b; Ito et al. 2004). Because mutations in many of these genes dramatically alter early flower development, their functions at late floral stages have not been carefully studied. The proteomic results here indicate that these proteins are still expressed at significant levels during later stages of floral development, suggesting that they function at these stages as well, and further studies are warranted.

Phylogenetic analysis of SAM synthesis pathway components

Our 2-DGE/MS analysis identified a number of proteins involved in the SAM pathway. In particular, we detected all four SAM synthetases encoded by the *Arabidopsis* genome, each of which was supported by isoform-specific peptides. SAM synthetase (SAMS) is a key enzyme in the active methyl cycle, catalyzing the formation of SAM, the major donor of methyl groups for numerous methylation reactions (Loenen 2006). Although the SAMS enzyme activity is expected to be important and conserved, it is not clear whether specific *SAMS* genes are conserved or divergent among plants. It is possible that the four *Arabidopsis SAMS* genes might have different functions. To investigate the evolutionary history of the *SAMS* genes, we carried out phylogenetic analysis of the *SAMS* genes and their homologs. We used the *Arabidopsis* SAMS protein sequences as queries to search against other plant genomes, including *Populus trichocarpa*, rice, maize, *Selaginella moellendorffii* and *Physcomitrella patens*, and the expressed sequence tag (EST) sequences of *Pinus pinaster*. We identified six *SAMS* genes in *Populus*, three in rice, two in maize, three in *Selaginella*, two in *Physcomitrella* and one in *Chlamydomonas*.

The alignment of protein sequences showed that members of this family have highly similar sequences, with the lowest observed identity of 85% and similarity of 91% (between At2g36880 and Os01g18860). A preliminary neighbor-Joining (NJ) tree (not shown) was constructed by using amino acid sequences, but the resolution of the tree is poor due to the highly conserved protein sequences. Thus the cDNA sequences of these genes were retrieved and used for further analysis. Both the NJ and maximum likelihood (ML) tree generated have the same topology (Figure 6). The tree topology indicates that genes from eudicots, monocots, pine, fern and moss all form well supported lineage-specific clades, respectively, suggesting that the most recent common ancestors of angiosperms, seed plants, vascular plants, and land plants all had a single copy of SAMS, which was derived from a single copy in the common ancestor of green plants. A number of duplications are supported by the topology, in eudicots, *Selaginella*, and *Physcomitrella*. In addition, recent duplication events have occurred after the divergence of *Arabidopsis* and *Populus*.

Therefore, the multiple copies in different species have had relatively short history, suggesting that they are still functionally related, consistent with the observed high levels of sequence similarity. At the same time, the two clades in the eudicots have been maintained since before the divergence of *Arabidopsis* and poplar, estimated to be approximately 100 million years or more, suggesting that these two clades might have diverged functionally, although they still have very highly similar coding regions. In addition, we examine our microarray data (Zhang et al. 2005b) and publicly available microarray data for expression patterns of the four *Arabidopsis* genes and found that they are all highly expressed in flowers and also other organs.

S-adenosyl methionine is a methyl donor and we also detected two other enzymes in the active methyl cycle, homocysteinase (Zhu et al. 2003) and methioine synthase (Eichel et al. 1995; Ravanel et al. 2004) in our proteomic experiments. To investigate whether genes encoding these two enzymes are also conserved, we carried out phylogenetic studies of the gene families for these two enzymes. Database searches identified, respectively, two homocysteinase and three methionine synthase encoding genes in *Arabidopsis*, two and four in *Populus*, one and two in rice, one and three in maize, one and one in *Asparagus*, two and one in pine, one and two in *Selaginella*, two and three in *Physcomitrella* and one of each in *Chlamydomonas*. Similar to SAMS, these two protein families are also quite well conserved. The NJ and ML trees of both families showed similar topologies to the SAM synthetase family, but the homocysteinase family has fewer duplication events within monocots and eudicots compared with the other two. The microarray data also indicated a high expression level of all the *Arabidopsis* genes in the two families except the methionine synthase-encoding gene At5g20980, which was expressed at a very low level. The phylogenetic tree showed that At5g20980 has an obvious longer branch compared with other genes, suggesting that it has evolved relatively rapidly.

In summary, our phylogenetic analyses indicate that genes encoding key enzymes in the active methyl cycle have been conserved through much of the history of land plants; it is likely that these genes have important functions that have been maintained during land plant evolution. Although there also have been some recent duplications within the eudicots and/or monocots lineages, most of the duplicated genes are highly similar in sequence and expression, suggesting that they still have similar functions.

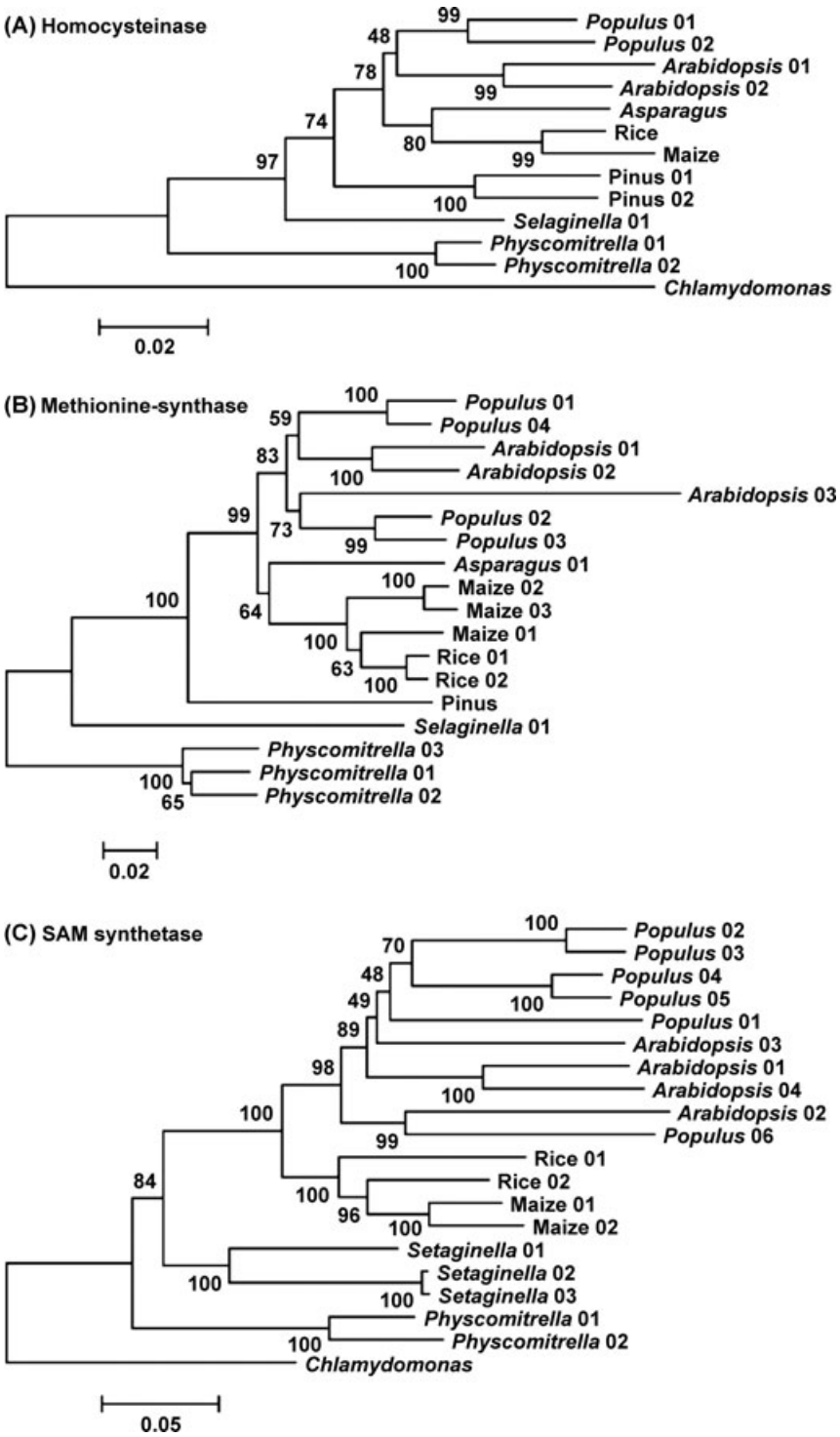

Figure 6. Phylogenetic trees of S-adenosyl methionine (SAM) synthetase (**A**), homocysteinase (**B**), and methionine synthase (**C**) gene families. The trees are constructed by neighbor-joining (NJ) method based on the alignment of CoDing Sequences (CDS) sequences.

Discussion

A combination of proteomics methods and analyses allows increased detection of floral proteins

We showed here that 2-DGE and MudPIT experiments can uncover a portion of the floral proteome and that the identification results from these proteomic approaches are partially overlapping. The results indicate that both proteomic methods sample the proteome, but cannot detect all or even most proteins in one experiment (it is estimated that any one cell type in higher eukaryotic cells expresses about 15 000–20 000 different proteins, not counting post-translational modifications). Although 2-DGE is thought to favor abundant proteins, it is intriguing to note that some of the proteins detected by the 2-DGE analysis were missed by both of the MudPIT experiments, suggesting that they might be present at moderate or low levels. This is supported by the observation that even proteins of very faint 2-DGE spots can be detected by MS, indicating that this approach is quite sensitive. One advantage of 2-DGE is that it provides information about the size and pI of the whole protein, as well as a rough estimate of the abundance. However, on a per protein basis, 2-DGE is more laborious and expensive than MudPIT, which allows the detection and identification of more than a thousand proteins in one experiment, albeit without the whole protein level information, such as pI, approximate molecular weight, and abundance. In addition, 2-DGE requires the proteins to be extracted from the cells and separated on the 2-DGE; in contrast, MutPIT experiments allow crude extracts to be digested with trypsin and the total peptide pools are then analyzed by 2-D LC and MS. The procedural difference might also contribute to the difference in the identified proteins.

The MudPIT approach is quite similar in a general way to the analysis of the transcriptome using EST analysis, which samples from a cDNA library by sequencing random clones and genes with larger transcript numbers are sampled more frequently. Therefore, the fact that there is considerable difference in the sets of proteins identified by the two MudPIT experiments strongly suggests that the total floral proteome is much larger than the $\sim$2 400 proteins we have uncovered. From microarray studies, it was estimated that more than 14 000 genes are expressed in the *Arabidopsis* flower. Many of these genes are expressed at low levels; our results indicate that even the products of genes with very low mRNA levels can be detected but at much lower frequencies than the products of highly expressed genes. Therefore, additional fractionation and enrichment procedures, as well as multiple MudPIT experiments with the different enriched fractions, are likely needed to detect the lower abundance proteins.

We also found that analysis of the MS data from MudPIT experiments using two different algorithms (ProteinPilot and Mascot) with two protein databases (SwissProt and NCBI) yielded overlapping but non-identical protein sets. With either algorithm, significantly greater numbers of non-redundant protein identifications were possible when searching against the *A. thaliana* entries in the NCBInr database than when searching against the *A. thaliana* entries in the SwissProt database, suggesting that *A. thaliana* sequences may be under-represented in the SwissProt database; there may also be some difference in the way that information regarding the predicted *Arabidopsis* proteome is organized in the databases, and of course the two algorithms differ in the way the MS/MS data are analyzed and scored. Our experience from the analysis of the *Arabidopsis* floral proteome indicates that when carrying out a proteome-wide survey, it is beneficial to analyze the data using multiple algorithms/databases, as previously suggested by Resing et al. (2004), in order to maximize the detection of the proteins present in the tissue. Further studies will likely provide better understanding of both the experimental and the bioinformatic components of proteomic efforts.

To learn whether the proteins we detected are also expressed in other parts of the plant, we compared our results with previous proteomic data from *Arabidopsis*. Holmes-Davis et al. (2005) reported the identification of 135 proteins from mature pollen using 2-DG/MS; 82 of these proteins were detected by our MudPIT experiments, consistent with the fact that our samples of stage-12 flowers contained mature pollen. In a recently study, 2 342 proteins were identified from leaves (Lee et al. 2007). Among these, 1 109 were also identified by our MudPIT experiments, indicating that many proteins are shared between leaves and flowers. In addition, more than a third (145 out of 402; Carter et al. 2004) of detected vegetative vacuolar proteins and about one sixth (119/689; Chibani et al. 2006) of seed dormancy-related proteins were found in our studies. Therefore, the floral proteome observed in our studies showed substantial overlap with proteins uncovered from other organs, suggesting that many proteins carry out functions that are common in the flower and other parts of the plant, in agreement with mRNA expression data (Zhang et al. 2005b).

Evidence for post-translational modifications in the *Arabidopsis* floral proteome

Protein modifications are important for cellular processes and cannot be detected by mRNA level analysis. We have obtained strong evidence for multiple protein modifications in the *Arabidopsis* flower. First, the distribution pattern of some of the proteins on 2-DGE suggests that they had multiple forms showing different molecular weights and pI values. Second, the MS/MS analysis of the peptides released from these proteins directly suggest the type and position of the covalent modifications (Tables 1, 2). The MudPIT data revealed many modifications that potentially had occurred *in vivo*, such as formylation, methylation, myristoylation, and acetylation, all with strong confidence scores ($>$99%) indicated in the modified peptides. Thus, proteomic methods can directly detect such

modifications and provide information regarding the possible regulation of modified proteins, which might play important roles in many pathways.

As an example of one observed post-translational modification, methylation is a covalent attachment of a methyl group to target molecules. In plants, the primary donor of the methyl group is SAM, which is generated by SAM cycle (Loenen 2006). A broad range of methyltransferases recognizes different target molecules and transfer the methyl group from SAM to the target sites. In plants, methylation on DNA, RNA, protein or small molecules has been shown to be important. DNA methylation is the most well-studied and is implicated in the remodeling of chromatin structure and repression of gene expression. In the *Arabidopsis* flower, DNA methylation has been shown to be involved in the vernalization pathway for flowering (Amasino 2004; Schmitz et al. 2008), the regulation of the *SUPERMAN* gene (Sakai et al. 1995; Jacobsen and Meyerowitz 1997; Lindroth et al. 2001; Cao and Jacobsen 2002; Jackson et al. 2002), genomic imprinting and embryo developments (Grossniklaus et al. 2001; Jullien et al. 2006; Baroux et al. 2007; Huh et al. 2007). Reduction in DNA methylation can cause complex abnormal phenotypes (Jacobsen and Meyerowitz 1997; Jacobsen et al. 2000). RNA methylation is related to RNA stability. For example, microRNA stability is regulated by the *HEN1* RNA methyltransferase (Li et al. 2005; Yu et al. 2005). Protein methylation such as histone methylation is also involved in many regulatory processes (Johnson et al. 2002; Schmitz et al. 2008). In addition, small molecules such as plant hormones (auxin, giberilline, salicylic acid and jasmonic acid) are also modified by methylation (Seo et al. 2001; Cheong and Choi 2003; Zubieta et al. 2003; Qin et al. 2005; Varbanova et al. 2007).

We have detected several SAM cycle enzymes, including all four SAMS. Sequence similarity and phylogenetic analysis suggest that these proteins have highly conserved functions during plant evolution. At the same time, there are several duplications in major angiosperm groups, including one that occurred before the split of *Arabidopsis* and poplar, suggesting that they might have non-identical functions. Similarly, the genes encoding methionine synthases also have diverged prior to the *Arabidopsis*-poplar divergence, but homocysteinase genes only have more recent duplications. It is thought that duplicated genes allow functional divergence, which in turn is important for the maintenance of duplicated genes. However, the duplicated SAM cycle genes remain highly similar in sequence and expression, suggesting that they might still have the same or very similar functions. One reason multiple genes with very similar functions might be maintained is that additional copies can promote very high level expression, as supported by the observation that they are expressed at high levels in different organs. If future experiments can show that these SAM cycle enzymes indeed have the same or very similar functions, and multiple genes are needed to achieve high level expression, this would be an excellent example of maintenance of functionally similar duplicated genes.

We have also detected different types of methyltransferases. The presence of multiple methyltransferases suggests that many types of molecules are potentially modified by methylation. Moreover, we have also detected many methylated peptides, with one, two, or three methyl groups, suggesting that protein methylation in the flower is quite complex. Studies of the biological significance of the methylations of specific proteins will likely lead to new insights about the regulation of protein activities.

In conclusion, we present a proteomic study of the *Arabidopsis* flower. We illustrate that multiple experiments and analyses can lead to relatively deep sampling of the proteome, and that proteins with varying abundance as estimated from staining intensity can be detected. We further present strong evidence for protein modification, including methylation, and suggest that they are important for normal flower development and physiology. Our study opens a new door for the understanding of the flower and should stimulate future research in this fertile area.

Materials and Methods

Plant materials and growth conditions

The *Arabidopsis* wild type plants of the Columbia ecotype were used in this study. All plants were grown at 23 °C under long day conditions of 16 h in light and 8 h in darkness. Flower buds at stage 12 were collected and frozen immediately in liquid nitrogen.

Protein isolation and separation by 2-D electrophoresis

To separate and identify floral proteins, we used 2-DGE gels with MALDI-TOF/TOF and/or Q-TOF identification of trypsin-digested gel-spots. Total protein was isolated from the flower buds using acetone/TCA precipitation methods. Briefly, about 0.5–1.0 g flower buds were ground into fine powder in liquid nitrogen. The powder was re-suspended in 15% TCA with 0.07% β-mecaptoethanol in acetone (pre-chilled at -70 °C), and precipitated at -20 °C for at least 45 min. The precipitated proteins were collected by spinning at 11 000*g* at 4 °C for 20 min. The pellet was washed twice with cold acetone and dried under vacuum to form crude protein powder, which was then dissolved in rehydration buffer (8 mol/L Urea, 2% CHAPS, 0.002% bromophenol blue) containing 0.5% Bio-Lyte and 50 mmol/L dithiothreitol (DTT). Protein concentration was measured by using the Bradford assay (Bio-Rad Protein Assay reagents, Hercules, CA, USA).

The proteins were then focused on Bio-Rad IEF (isoelectric focusing) instruments. The SDS-PAGE electrophoresis was done according to the user's manual from Bio-Rad, with the Bio-Rad

broad range protein markers used as molecular weight standards. The 2-D PAGE gels were then stained with Coomassie Brilliant Blue and scanned with a GC-800 densitometer. The acquired images were analyzed using Bio-Rad PDQuest software.

For the first dimension –IEF, we used Readystrip IPG strips (pH 4–7, 11 cm) and IPG buffer (pH 3–10) from Bio-Rad Laboratories (Hercules, CA, USA). The gels were rehydrated for 16 h with the rehydration buffer containing 200 μg protein and then were focused in a PROTEAN IEF cell apparatus (Bio-Rad Laboratories) under the following conditions: running temperature: 20 °C; maximum current: 50 μA/gel; Step 1: 200 V for 30 min (linear ramp); Step 2: 300 V for 30 min (rapid ramp); Step 3: 8 000 V for 150 min (linear ramp); Step 4: 8 000 V for 55 000 Vh (linear ramp). After completion of IEF, the gels were incubated for 15 min in SDS equilibration buffer (50 mM Tris-HCl, pH 8.8, 6 mol/L urea, 30% glycerol, 2% SDS, and 0.002% bromophenol blue) with 10 mg/mL dithiothreitol and then switched into the same buffer containing 25 mg/mL iodoacetamide for another 15 min. For the second dimension, proteins were separated by discontinuous SDS-PAGE in a Criterion Dodeca Cell apparatus (Bio-Rad Laboratories) using 12.5% precast gels. 2-D gels were stained for protein with Coomassie Blue R-250 and scanned by using a GS-800 calibrated densitometer (Bio-Rad Laboratories). 2-D images were analyzed with PDQuest software (Version 7.4 from Bio-Rad Laboratories).

Identification of proteins by MALDI-TOF MS and/or Q-TOF MS

To identify the proteins, 265 visible protein spots on 2-D gel were cut out, digested by trypsin and put to MS analysis. The protein spots of interest were cut out from the 2-D gel either manually or using the Spot Cutter from Bio-Rad. Gel spots were subjected to in-gel trypsin digestion according to the user's manual (Trypsin Gold, Promega, Madison, WI, USA). Briefly, the gel slices were washed three times with 100 μL of 25 mmol/L ammonium bicarbonate and dehydrated with 100 μL of 50% acetonitrile, and dried in a Speed-Vac. They were then incubated overnight at 37 °C with trypsin (12.5 ng/μL in 25 mmol/L ammonium bicarbonate). Peptides were then extracted twice with 25 μL of 5% formic acid for 20 min. The extracts were dried in a Speed-Vac and resuspended in 35 μL of 5% acetonitrile with formic acid (0.1%).

For Q-TOF (nano-LC-ESI-MS/MS) analysis, tryptic digests were analyzed by capillary liquid chromatography–nanoelectrospray ionization-tandem mass spectrometry (CapLC–ESI-MS/MS). A Micromass Q-TOF Premier mass spectrometer coupled with a Waters CapLC HPLC unit was used for the analysis. Peptides (1–5 μL) were injected into solvent A (acetonitrile/water/formic acid, 5/95/0.1) supplied by the auxiliary pump of the capillary HPLC unit and trapped in a Dionex μ-Precolumn (C_{18} PepMap 100, 5 μm, 300 μm × 5 mm) for on-line desalting and concentration. After washing for 5 min with solvent A at 10 μL/min, trapped peptides were then back-flushed with the gradient solvent flow on to the analytical column, a Dionex PepMap fused silica capillary column (C_{18}, 5 μm, 75 μm × 150 mm), using a 10-port switching valve. The analytical column was run with a gradient (0–100% solvent B; acetonitrile/ water/formic acid; 95/5/0.1; in 45 min). The mass spectrometry was calibrated using Glu-Fib product ion.

The Q-TOF mass spectrometer was operated to acquire MS/MS of tryptic peptides in data-dependent acquisition mode for precursor ion selection using charge-state recognition and intensity threshold as selection criteria using MassLynx 4.0 SP1. In order to carry out the tandem mass spectrometric data acquisition, a survey scan (2 s) over the *m*/*z* of 400–1 500 was carried out. From each survey scan, up to four most intense precursor ions based on the selection criteria were selected for tandem mass spectrometry to obtain the production spectra resulting from collision-induced dissociation in the presence of argon. The product ion spectra (6–8 s) collected were processed using Protein Lynx Global Server 2.1 and were converted to peak list text files for database searching. For protein identification, the Internet-based MASCOT MS/MS Ions Search tool (http://www.matrixscience.com/cgi/searchform.pl?FORMVER=2&SEARCH=MIS) was used to manually search against the NCBI non-redundant protein database. Search parameters were set as follows: taxonomy, all entries; enzyme, trypsin; variable modifications, acetyl (N-term), oxidation (M), phospho (ST), phospho (Y), pyro-glu (N-term E), pyro-glu (N-term Q); mass values, monoisotopic; protein mass, unrestricted peptide; mass tolerance, 150 ppm; fragment mass tolerance, 0.2 D; max missed cleavages, 1. Proteins with a minimum of one peptide match (with a significant ion score of 25 according to MASCOT) and containing a sequence tag with at least three amino acids in a row in either Y- or B-ion series, but not matching to any other proteins, were considered as positive and listed in the final summary. All proteins with only a single peptide match were verified by manual inspection of Y- and B-ion series to identify the minimal sequence tag of three consecutive ions. Each of these proteins had multiple numbers of two amino acid sequence tags (significant Mascot scores >25). For matches representing protein families, only one member is presented. All proteins identified through the NCBI database (and corresponding annotations) will be verified by BLAST searches against the public available database. To obtain the corresponding AGI identifier from SwissProt protein identifier, a C script is used to search the NCBI database. Corresponding microarray data from the Stage 12 flower tissue was used to evaluate the corresponding genes' transcripts level (Zhang et al. 2005b).

Multi-dimensional protein identification

For MudPIT experiments, 4.5 mg dried protein powder was suspended in 1 mL of 100 mmol/L NH_4HCO_3, then mixed by

vigorous vortexing and sonication to improve the dissolution of proteins. The amount of proteins in this sample was estimated by the Bradford assay. Approximately 800–1 000 μg of protein was digested with about 20–30 μg of Trypsin Gold (Promega). DTT was added to the digestion reaction to a final concentration 10 mM, and the sample was kept at 37 °C for 1 h. Then iodoacetamide (IDA) was added to a final concentration of 50 mmol/L, and the sample was kept in dark for 1 h. About 20–30 μg of Trypsin Gold (Promega) was added to the sample. It was then incubated at 37 °C for 18 h. After clearing the sample by centrifugation at top speed for 10 min, the supernatant was adjusted to pH 3–4 by adding 10% trifluoroacetic acid (TFA). This mixture of tryptic peptides was sent to the Penn State College of Medicine Mass Spec/Proteomics facility for MudPIT analysis. After 2D-LC separation of the tryptic peptide mix, the spectra from each spot were analyzed by using Mascot or ProteinPilot software searching against either the SwissProt or NCBInr databases. The SwissProt identifier and NCBI accession number were converted to the *Arabidopsis* genome identifier by using a script that can search the online NCBI databases. Some *Arabidopsis* genome identifiers were obtained by doing a local batch blast against *Arabidopsis* whole genome sequences (Altschul et al. 1997; AGI 2000).

Go annotation and pathway analysis

To categorize the proteins identified by different approaches, and find the important pathways for flower development, we searched the GO annotation of each protein on The Arabidopsis Information Resource (TAIR) website and the pathway information from a database from the TAIR website. The percentage of each category is calculated and represented in a pie graph. A script was written for searching the pathways from the database Aracyc. Manually sorting and finding pathways were also undertaken for some important pathways such as the methylation pathway.

Phylogenetic analysis

The protein sequences of the genes selected from the list were used as queries to carry out TBLASTN search against plant genomes on NCBI and JGI (Joint Genome Institute), or The Institute for Genomic Research (TIGR) Plant Transcript Assemblies database for those species without complete genome sequences. An *e*-value cut-off at 1*e*-05 was used. Each new result was used to carry out a second round of search until no new result was found. For each gene family, the alignment was carried out by MUSCLE version 3.6 (Edgar 2004). The alignment results were manually adjusted by using GeneDoc version 2.6.002. The neighbor-joining trees were constructed by using MEGA 4.0 with a bootstrap analysis of 1 000 replicates (Tamura et al. 2007). The maximum likelihood trees were constructed by using PHYML 2.4.4 with a bootstrap analysis of 100 replicates (Guindon and Gascuel 2003). The substitution model and parameters for each family was chosen according to the test results of ProtTest 1.4 (Abascal et al. 2005) for protein tree or MrModelTest 1.32 (Posada and Crandall 1998) for DNA tree. Only the NJ trees were shown with bootstrap values for NJ.

Acknowledgements

We thank Jason Ma (Stanford University, CA, USA) for his assistance with C scripts for analyzing the data. We thank Jiong Wang (technician, Ma's Lab, the Pennsylvania State University (PSU)) for help with plant care, Dr John Diller and Dihong Lu (Ma's Lab, PSU) for their comments on the manuscript, and Anne Stanley (the College of Medicine, PSU, Hershey, PA, USA) for assistance in processing the MudPIT samples. We also appreciate suggestions on proteomics procedures from Zhixin Zhao (Assmann's Lab, PSU).

References

Abascal F, Zardoya R, Posada D (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Amasino R (2004). Vernalization, competence, and the epigenetic memory of winter. *Plant Cell* **16**, 2553–2559.

***Arabidopsis* Genome Initiative** (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.

Baroux C, Pien S, Grossniklaus U (2007). Chromatin modification and remodeling during early seed development. *Curr. Opin. Genet. Dev.* **17**, 473–479.

Bayer EM, Bottrill AR, Walshaw J, Vigouroux M, Naldrett MJ, Thomas CL et al. (2006). *Arabidopsis* cell wall proteome defined using multidimensional protein identification technology. *Proteomics* **6**, 301–311.

Brady SM, Sarkar SF, Bonetta D, McCourt P (2003). The *ABSCISIC ACID INSENSITIVE 3* (*ABI3*) gene is modulated by farnesylation and is involved in auxin signaling and lateral root development in *Arabidopsis*. *Plant J.* **34**, 67–75.

Brand U, Fletcher JC, Hobe M, Meyerowitz EM, Simon R (2000). Dependence of stem cell fate in *Arabidopsis* on a feedback loop regulated by CLV3 activity. *Science* **289**, 617–619.

Caldelari D, Sternberg H, Rodriguez-Concepcion M, Gruissem W, Yalovsky S (2001). Efficient prenylation by a plant geranylgeranyltransferase-I requires a functional CaaL box motif and a proximal polybasic domain. *Plant Physiol.* **126**, 1416–1429.

Cao X, Jacobsen SE (2002). Role of the *Arabidopsis* DRM methyltransferases in de novo DNA methylation and gene silencing. *Curr. Biol.* **12**, 1138–1144.

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.

Chamovitz DA, Wei N, Osterlund MT, von Arnim AG, Staub JM, Matsui M et al. (1996). The COP9 complex, a novel multisubunit nuclear regulator involved in light control of a plant developmental switch. *Cell* **86**, 115–121.

Cheong JJ, Choi YD (2003). Methyl jasmonate as a vital substance in plants. *Trends Genet.* **19**, 409–413.

Chibani K, Ali-Rachedi S, Job C, Job D, Jullien M, Grappin P (2006). Proteomic analysis of seed dormancy in *Arabidopsis*. *Plant Physiol.* **142**, 1493–1510.

Clark SE, Jacobsen SE, Levin JZ, Meyerowitz EM (1996). The *CLAVATA* and *SHOOT MERISTEMLESS* loci competitively regulate meristem activity in *Arabidopsis*. *Development* **122**, 1567–1575.

Clark SE, Williams RW, Meyerowitz EM (1997). The *CLAVATA1* gene encodes a putative receptor kinase that controls shoot and floral meristem size in *Arabidopsis*. *Cell* **89**, 575–585.

Delahunty CM, Yates JR 3rd (2007). MudPIT: multidimensional protein identification technology. *Biotechniques* **43**, 563, 565, 567 passim.

Deng XW, Caspar T, Quail PH (1991). *cop1*: a regulatory locus involved in light-controlled development and gene expression in *Arabidopsis*. *Genes Dev.* **5**, 1172–1182.

DeYoung BJ, Bickle KL, Schrage KJ, Muskett P, Patel K, Clark SE (2006). The CLAVATA1-related BAM1, BAM2 and BAM3 receptor kinase-like proteins are required for meristem function in *Arabidopsis*. *Plant J.* **45**, 1–16.

Drews GN, Bowman JL, Meyerowitz EM (1991). Negative regulation of the *Arabidopsis* homeotic gene *AGAMOUS* by the *APETALA2* product. *Cell* **65**, 991–1002.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.

Eichel J, Gonzalez JC, Hotze M, Matthews RG, Schroder J (1995). Vitamin-B12-independent methionine synthase from a higher plant (*Catharanthus roseus*). Molecular characterization, regulation, heterologous expression, and enzyme properties. *Eur. J. Biochem.* **230**, 1053–1058.

Florens L, Washburn MP (2006). Proteomic analysis by multidimensional protein identification technology. *Methods Mol. Biol.* **328**, 159–175.

Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI (1999). A sampling of the yeast proteome. *Mol. Cell. Biol.* **19**, 7357–7368.

Gallardo K, Job C, Groot SP, Puype M, Demol H, Vandekerckhove J et al. (2001). Proteomic analysis of *Arabidopsis* seed germination and priming. *Plant Physiol.* **126**, 835–848.

Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L et al. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **1**, 323–333.

Grossniklaus U, Spillane C, Page DR, Kohler C (2001). Genomic imprinting and seed development: endosperm formation with and without sex. *Curr. Opin. Plant Biol.* **4**, 21–27.

Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704.

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730.

Holmes-Davis R, Tanaka CK, Vensel WH, Hurkman WJ, McCormick S (2005). Proteome mapping of mature pollen of *Arabidopsis thaliana*. *Proteomics* **5**, 4864–4884.

Huh JH, Bauer MJ, Hsieh TF, Fischer R (2007). Endosperm gene imprinting and seed development. *Curr. Opin. Genet. Dev.* **17**, 480–485.

Ito T, Wellmer F, Yu H, Das P, Ito N, Alves-Ferreira M et al. (2004). The homeotic protein AGAMOUS controls microsporogenesis by regulation of *SPOROCYTELESS*. *Nature* **430**, 356–360.

Jackson JP, Lindroth AM, Cao X, Jacobsen SE (2002). Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416**, 556–560.

Jacobsen SE, Meyerowitz EM (1997). Hypermethylated *SUPERMAN* epigenetic alleles in *Arabidopsis*. *Science* **277**, 1100–1103.

Jacobsen SE, Sakai H, Finnegan EJ, Cao X, Meyerowitz EM (2000). Ectopic hypermethylation of flower-specific genes in *Arabidopsis*. *Curr. Biol.* **10**, 179–186.

Jarvis P (2004). Organellar proteomics: chloroplasts in the spotlight. *Curr. Biol.* **14**, R317–319.

Johnson L, Cao X, Jacobsen S (2002). Interplay between two epigenetic marks. DNA methylation and histone H3 lysine 9 methylation. *Curr. Biol.* **12**, 1360–1367.

Jorrin JV, Maldonado AM, Castillejo MA (2007). Plant proteome analysis: a 2006 update. *Proteomics* **7**, 2947–2962.

Jullien PE, Katz A, Oliva M, Ohad N, Berger F (2006). Polycomb group complexes self-regulate imprinting of the Polycomb group gene *MEDEA* in *Arabidopsis*. *Curr. Biol.* **16**, 486–492.

Kunst L, Klenz JE, Martinez-Zapater J, Haughn GW (1989). *AP2* gene fdetermines the identity of perianth organs in flowers of *Arabidopsis thaliana*. *Plant Cell* **1**, 1195–1208.

Lee J, Garrett WM, Cooper B (2007). Shotgun proteomic analysis of *Arabidopsis thaliana* leaves. *J. Sep. Sci.* **30**, 2225–2230.

Li J, Yang Z, Yu B, Liu J, Chen X (2005). Methylation protects miRNAs and siRNAs from a 3′-end uridylation activity in *Arabidopsis*. *Curr. Biol.* **15**, 1501–1507.

Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S et al. (2001). Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **292**, 2077–2080.

Liu LJ, Zhang YC, Li QH, Sang Y, Mao J, Lian HL et al. (2008). COP1-mediated ubiquitination of CONSTANS is implicated in cryptochrome regulation of flowering in *Arabidopsis*. *Plant Cell* **20**, 292–306.

Lobbes D, Rallapalli G, Schmidt DD, Martin C, Clarke J (2006). SERRATE: a new player on the plant microRNA scene. *EMBO Rep.* **7**, 1052–1058.

Loenen WA (2006). S-adenosylmethionine: jack of all trades and master of everything? *Biochem. Soc. Trans.* **34**, 330–333.

Ma H (2005). Molecular genetic analyses of microsporogenesis and microgametogenesis in flowering plants. *Annu. Rev. Plant Biol.* **56**, 393–434.

Majeran W, Cai Y, Sun Q, van Wijk KJ (2005). Functional differentiation of bundle sheath and mesophyll maize chloroplasts determined by comparative proteomics. *Plant Cell* **17**, 3111–3140.

Maor R, Jones A, Nuhse TS, Studholme DJ, Peck SC, Shirasu K (2007). Multidimensional protein identification technology (MudPIT) analysis of ubiquitinated proteins in plants. *Mol. Cell. Proteomics* **6**, 601–610.

Ni W, Xie D, Hobbie L, Feng B, Zhao D, Akkara J et al. (2004). Regulation of flower development in *Arabidopsis* by SCF complexes. *Plant Physiol.* **134**, 1574–1585.

Nuhse TS, Stensballe A, Jensen ON, Peck SC (2004). Phosphoproteomics of the *Arabidopsis* plasma membrane and a new phosphorylation site database. *Plant Cell* **16**, 2394–2405.

Ogawa M, Shinohara H, Sakagami Y, Matsubayashi Y (2008). *Arabidopsis* CLV3 peptide directly binds CLV1 ectodomain. *Science* **319**, 294.

Park W, Li J, Song R, Messing J, Chen X (2002). CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12**, 1484–1495.

Peck SC, Nuhse TS, Hess D, Iglesias A, Meins F, Boller T (2001). Directed proteomics identifies a plant-specific protein rapidly phosphorylated in response to bacterial and fungal elicitors. *Plant Cell* **13**, 1467–1475.

Peck SC (2005). Update on proteomics in *Arabidopsis*. Where do we go from here? *Plant Physiol.* **138**, 591–599.

Pei ZM, Ghassemian M, Kwak CM, McCourt P, Schroeder JI (1998). Role of farnesyltransferase in ABA regulation of guard cell anion channels and plant water loss. *Science* **282**, 287–290.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567.

Poole RL (2007). The TAIR database. *Methods Mol. Biol.* **406**, 179–212.

Posada D, Crandall KA (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.

Pradet-Balade B, Boulme F, Beug H, Mullner EW, Garcia-Sanz JA (2001). Translation control: bridging the gap between genomics and proteomics? *Trends Biochem. Sci.* **26**, 225–229.

Qin G, Gu H, Zhao Y, Ma Z, Shi G, Yang Y et al. (2005). An indole-3-acetic acid carboxyl methyltransferase regulates *Arabidopsis* leaf development. *Plant Cell* **17**, 2693–2704.

Rajjou L, Belghazi M, Huguet R, Robin C, Moreau A, Job C et al. (2006). Proteomic investigation of the effect of salicylic acid on *Arabidopsis* seed germination and establishment of early defense mechanisms. *Plant Physiol.* **141**, 910–923.

Ravanel S, Block MA, Rippert P, Jabrin S, Curien G, Rebeille F et al. (2004). Methionine metabolism in plants: chloroplasts are autonomous for *de novo* methionine synthesis and can import S-adenosylmethionine from the cytosol. *J. Biol. Chem.* **279**, 22 548–22 557.

Resing KA, Meyer-Arendt K, Mendoza AM, Aveline-Wolf LD, Jonscher KR, Pierce KG et al. (2004). Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **76**, 3556–3568.

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A et al. (2003). The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* **31**, 224–228.

Ru QC, Zhu LA, Katenhusen RA, Silberman J, Brzeski H, Liebman M et al. (2006). Exploring human plasma proteome strategies: high efficiency in-solution digestion protocol for multi-dimensional protein identification technology. *J. Chromatogr. A.* **1111**, 175–191.

Sakai H, Medrano LJ, Meyerowitz EM (1995). Role of SUPERMAN in maintaining *Arabidopsis* floral whorl boundaries. *Nature* **378**, 199–203.

Samach A, Klenz JE, Kohalmi SE, Risseeuw E, Haughn GW, Crosby WL (1999). The *UNUSUAL FLORAL ORGANS* gene of *Arabidopsis thaliana* is an F-box protein required for normal patterning and growth in the floral meristem. *Plant J.* **20**, 433–445.

Schiefthaler U, Balasubramanian S, Sieber P, Chevalier D, Wisman E, Schneitz K (1999). Molecular analysis of *NOZZLE*, a gene involved in pattern formation and early sporogenesis during sex organ development in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **96**, 11664–11669.

Schmidt MW, Houseman A, Ivanov AR, Wolf DA (2007). Comparative proteomic and transcriptomic profiling of the fission yeast *Schizosaccharomyces pombe*. *Mol. Syst. Biol.* **3**, 79.

Schmitz RJ, Sung S, Amasino RM (2008). Histone arginine methylation is required for vernalization-induced epigenetic silencing of FLC in winter-annual *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **105**, 411–416.

Schoof H, Lenhard M, Haecker A, Mayer KF, Jurgens G, Laux T (2000). The stem cell population of *Arabidopsis* shoot meristems is maintained by a regulatory loop between the *CLAVATA* and *WUSCHEL* genes. *Cell* **100**, 635–644.

Seo HS, Song JT, Cheong JJ, Lee YH, Lee YW, Hwang I et al. (2001). Jasmonic acid carboxyl methyltransferase: a key enzyme for jasmonate-regulated plant responses. *Proc. Natl. Acad. Sci. USA* **98**, 4788–4793.

Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP et al. (2007). The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics* **6**, 1638–1655.

Smyth DR, Bowman JL, Meyerowitz EM (1990). Early flower development in *Arabidopsis*. *Plant Cell* **2**, 755–767.

Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599.

Thelen JJ, Peck SC (2007). Quantitative proteomics in plants: choices in abundance. *Plant Cell* **19**, 3339–3346.

Valverde F, Mouradov A, Soppe W, Ravenscroft D, Samach A, Coupland G (2004). Photoreceptor regulation of CONSTANS protein in photoperiodic flowering. *Science* **303**, 1003–1006.

Varbanova M, Yamaguchi S, Yang Y, McKelvey K, Hanada A, Borochov R et al. (2007). Methylation of gibberellins by *Arabidopsis* GAMT1 and GAMT2. *Plant Cell* **19**, 32–45.

Washburn MP, Ulaszek RR, Yates JR 3rd (2003). Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology. *Anal. Chem.* **75**, 5054–5061.

Washburn MP, Wolters D, Yates JR 3rd (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247.

Wurschum T, Gross-Hardt R, Laux T (2006). APETALA2 regulates the stem cell niche in the *Arabidopsis* shoot meristem. *Plant Cell* **18**, 295–307.

Yalovsky S, Rodriguez-Concepcion M, Bracha K, Toledo-Ortiz G, Gruissem W (2000). Prenylation of the floral transcription factor *APETALA1* modulates its function. *Plant Cell* **12**, 1257–1266.

Yang L, Liu Z, Lu F, Dong A, Huang H (2006). SERRATE is a novel nuclear regulator in primary microRNA processing in *Arabidopsis*. *Plant J.* **47**, 841–850.

Yang M, Hu Y, Lodhi M, McCombie WR, Ma H (1999a). The *Arabidopsis SKP1-LIKE1* gene is essential for male meiosis and may control homologue separation. *Proc. Natl. Acad. Sci. USA* **96**, 11416–11421.

Yang WC, Ye D, Xu J, Sundaresan V (1999b). The *SPOROCYTELESS* gene of *Arabidopsis* is required for initiation of sporogenesis and encodes a novel nuclear protein. *Genes Dev.* **13**, 2108–2117.

Yu B, Yang Z, Li J, Minakhina S, Yang M, Padgett RW et al. (2005). Methylation as a crucial step in plant microRNA biogenesis. *Science* **307**, 932–935.

Zahn L, Feng BM, Ma H (2006). Beyond the ABC model: regulation of floral homeotic genes. In: Soltis DE, Leebens-Mack JH, Soltis PS, eds. *Advances in Botanical Research*. Vol. 44. Elsevier. pp. 161–205.

Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD et al. (2005a). MetaCyc and AraCyc: metabolic pathway databases for plant research. *Plant Physiol.* **138**, 27–37.

Zhang X, Feng B, Zhang Q, Zhang D, Altman N, Ma H (2005b). Genome-wide expression profiling and identification of gene activities during early flower development in *Arabidopsis*. *Plant Mol. Biol.* **58**, 401–419.

Zhao D, Yang M, Solava J, Ma H (1999). The *ASK1* gene regulates development and interacts with the *UFO* gene to control floral organ identity in *Arabidopsis*. *Dev. Genet.* **25**, 209–223.

Zhao D, Yu Q, Chen M, Ma H (2001). The *ASK1* gene regulates B function gene expression in cooperation with *UFO* and *LEAFY* in *Arabidopsis*. *Development* **128**, 2735–2746.

Zhao D, Ni W, Feng B, Han T, Petrasek MG, Ma H (2003). Members of the *Arabidopsis-SKP1-like* gene family exhibit a variety of expression patterns and may play diverse roles in *Arabidopsis*. *Plant Physiol.* **133**, 203–217.

Zhu J, Hu X, Dizin E, Pei D (2003). Catalytic mechanism of S-ribosylhomocysteinase (LuxS): direct observation of ketone intermediates by 13C NMR spectroscopy. *J. Am. Chem. Soc.* **125**, 13379–13381.

Zubieta C, Ross JR, Koscheski P, Yang Y, Pichersky E, Noel JP (2003). Structural basis for substrate recognition in the salicylic acid carboxyl methyltransferase family. *Plant Cell* **15**, 1704–1716.

Note: While this paper was being prepared for submission, the following paper reporting an independent study was published: **Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S et al.** (2008). Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* **320**, 938–941. It reports large scale proteome analysis of wild-type Arabidopsis organs, including flowers.

Supporting Information

The following Supporting Information is available for this article:

Supplementary table 1. Proteins detected by 2-DGE and MS.
Supplementary table 2. Proteins detected by MudPIT experiments.
Supplementary table 3. Pathways represented by detected proteins.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1744-7909.2008.00787.x.

(This link will take you to the article abstract).

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

(Handling editor: Chun-Ming Liu)