Analysis of the Arabidopsis Cytosolic Proteome Highlights Subcellular Partitioning of Central Plant Metabolism.

Jun Ito¹, Tanveer S. Batth¹, Christopher J. Petzold¹, Alyssa M. Redding¹, Robert Verboom²,
Etienne H. Meyer^{2,3}, A. Harvey Millar² and Joshua L. Heazlewood*¹

¹ Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, One Cyclotron Road MS 978-4101, Berkeley, California, 94720, USA.

² Australian Research Council (ARC) Centre of Excellence in Plant Energy Biology and Centre for Comparative Analysis of Biomolecular Networks, The University of Western Australia, Crawley 6009, Western Australia, Australia

³ Institut de Biologie Moléculaire des Plantes, 12 rue du général Zimmer, 67084 Strasbourg cedex, France

***Corresponding Author**

Joshua L. Heazlewood

Joint BioEnergy Institute

Lawrence Berkeley National Laboratory

One Cyclotron Road MS 978-4466

Berkeley, CA 94720, USA.

Ph +1 510 495 2694

Fax +1 510 495 2437

jlheazlewood@lbl.gov

Running Title: Arabidopsis cytosolic proteome

Abstract

30 The plant cell cytosol is a dynamic and complex intracellular matrix, by definition it contains no compartmentation by lipid bilayers, but still maintains a wide variety of biochemical networks and often links metabolic pathways across multiple organelles. There have been numerous detailed proteomic studies of organelles in the model plant *Arabidopsis thaliana* although no such analysis has been undertaken on the cytosol. We have isolated the cytosolic fraction from
35 cell suspensions of *Arabidopsis* by using a gentle homogenization and employing offline multidimensional protein identification technology on three replicates have robustly identified 1071 cytosolic proteins. Functional annotation of the cytosolic proteome revealed its significant roles in protein synthesis and degradation, RNA metabolism and basic sugar metabolism. This included an array of important cytosol-related functions in particular: the ribosome, the set of
40 tRNA catabolic enzymes, the ubiquitin-proteasome pathway, glycolysis and associated sugar metabolism pathways, phenylpropanoid biosynthesis, vitamin metabolism, nucleotide metabolism, an array of signaling and stress-responsive molecules and NDP-sugar biosynthesis. This set of cytosolic proteins provides for the first time a direct and extensive analysis of enzymes responsible for the myriad of reactions in the *Arabidopsis* cytosol and defines an
45 experimental set of plant protein sequences that are not targeted to subcellular locations following translation and folding in the cytosol.

Keywords

Arabidopsis, cytosol, MudPIT, SUBA, plant proteomics, SRM

Introduction

The cytosol is the intracellular fluid containing all eukaryotic cellular components and importantly is the conduit allowing interactions between partitioned metabolic processes {Asaad, 2003 #321}. The cytosol is a crowded environment of a highly complex mixture of dissolved ionic solutes, small molecule metabolites and macromolecules including proteins and nucleic acids {Cayley, 1991 #331; Clegg, 1984 #334}, where over 50 % of total cellular protein in eukaryotes is believed to reside {Lahav, 1982 #361}. The effect of crowding can significantly alter the rates and state of chemical equilibrium of biological reactions {Asaad, 2003 #321}, reduce the rates of diffusion particularly with larger molecules {Verkman, 2002 #398}, encourage protein folding and self-association {Morar, 2001 #372}, and promote the association of macromolecules, for example, when proteins form complexes {Zhou, 2008 #408}. Despite the highly complex molecular environment in the cytosol, distinct levels of organization exist, which can be observed at defined areas in the cytosol. This includes small molecule concentration gradients i.e. Na^+ , K^+ and Ca^{2+} in osmoregulation and signal transduction {Berridge, 1997 #326; Lang, 2007 #362; Lecourieux, 2006 #363}, association of related proteins in distinct complexes; such as in the ribosome and proteasome machinery {Dinman, 2009 #339; Schrader, 2009 #391} and cytoskeletal sieving, where actin fibers restrict ribosomes and organelles from entry to certain areas in the cytosol {Luby-Phelps, 1987 #365; Provance, 1993 #382}.

Over 200,000 metabolites are estimated to be present inside plant cells, reflecting the innumerable cellular reactions taking place within this structure {Weckwerth, 2003 #403}. While cellular organelles undertake specialist functions in plants, many fundamental processes in the

plant cell take place in the cytosol including glycolysis{Plaxton, 1996 #378}, part of the pentose
75 phosphate pathway{Schnarrenberger, 1995 #390}, protein biosynthesis and degradation{Bailey-
Serres, 2009 #322; Vierstra, 2009 #399}, signal transduction{Klimecka, 2007 #354; Lecourieux,
2006 #363}, primary and secondary metabolite biosynthesis and transportation{Lunn, 2007
#367; Martinoia, 2007 #368; Krueger, 2009 #359; Weber, 2007 #402; Lundmark, 2006 #366},
stress response signaling{Cazale, 2009 #332; Yamada, 2008 #406; Sugio, 2009 #396} and
80 accumulation of enzymes for defense and detoxification (e.g. GSTs, isoprenoid biosynthesis
{Dixon, 2009 #340; Sappl, 2009 #416; Laule, 2003 #415}). Nuclear-encoded organellar proteins
are synthesized in the cytosol by ribosomes and delivered to compartments by their N-terminal
signal peptides {Jarvis, 2008 #350; Prassinos, 2008 #381; Huang, 2009 #349}. Thus far, the only
reported characterization of the plant cytosolic proteome to our knowledge has been in soybean
85 root nodules. The study identified a total of 69 proteins with roles in carbon metabolism (28 %),
nitrogen metabolism (12 %), reactive oxygen metabolism (12 %) and vesicular trafficking
(11 %) {Oehrle, 2008 #375}. In the model plant *Arabidopsis thaliana*, proteomic analyses have
only targeted major cytosolic protein complexes including the characterizations of the cytosolic
ribosome {Chang, 2005 #333; Giavalisco, 2005 #344; Carroll, 2008 #330} and the 26S
90 proteasome {Yang, 2004 #407}. Alternative characterization methods such as the use of
fluorescent protein localization have been more piecemeal. However a recent study investigating
reports of glutathione-S-transferases (GSTs) members in various *Arabidopsis* organelle
proteomes, confirmed the cytosolic localization of phi family members (F2, F6, F8, F9, and F12)
and tau family members (U2, U7, U9, U11, U12, U19, and U28) {Dixon, 2009 #340}. It was
95 concluded the GSTs in mitochondrial, plastid and vacuolar samples were more likely highly
abundant cytosolic contaminants rather than genuine functional residents of these organelles

{Dixon, 2009 #340}. The plethora of reactions in the plant cytosol and the availability of the well-annotated *Arabidopsis* genome {Initiative, 2000 #316} deserve an in-depth analysis of the *Arabidopsis* cytosolic proteome. Such a set can assist in defining the subcellular partitioning of multi-gene families, provide experimental evidence for the lack of targeting of particular proteins to organelle structures as well as provide a reference to delineate major cytosolic protein contaminants from various organelle proteomes.

This study outlines the isolation and characterization of three independent cytosolic fractions from cell suspensions of *Arabidopsis thaliana*. Organelle contamination of the cytosolic fractions was analyzed by Western blotting and mass spectrometry-based selected reaction monitoring. A total of 1071 cytosolic proteins were rigorously characterized and from these we describe the interconnection of central plant metabolic reactions in the context of subcellular partitioning of the cytosol from organelles.

Experimental Procedures

Arabidopsis thaliana Suspension Cell Culture

A heterotrophic *Arabidopsis* cell culture (Landsberg *erecta*) established from callus of stem explants was maintained by weekly subculture as outlined in {May, 1993 #369}, with the exception of the carbon source, which was 2 % (w/v) glucose. The cell cultures were maintained at 22°C and light intensity of 90 $\mu\text{mol m}^{-2} \text{s}^{-1}$ in an orbital shaker (120 rpm). After seven days, each flask (120 mL) contained 8–12 g of cells (fresh weight) and growth was approximately in the middle of the log phase.

Cytosol Isolation from Arabidopsis Cell Culture

Protoplasts were prepared from 7 day-old *Arabidopsis* suspension-cultured cells according to procedures outlined in {Meyer, 2008 #422}. Typically, 30 g of cells (FW) were collected from and incubated 3 h with gentle orbital rotation (85 rpm) in the dark at 22 °C. Protoplasts were resuspended in 60 mL of (0.4 M sucrose, 50 mM Tris-HCl, pH 7.5, 3 mM EDTA and 2 mM DTT) and disrupted by five strokes in a Potter-Elvehjem homogenizer at 4 °C. The homogenate was centrifuged at $800 \times g$ for 15 min at 4 °C; the supernatant retained and centrifuged at $10,000 \times g$ for 15 min at 4 °C, the supernatant was retained again and centrifuged at $100,000 \times g$ for 1 h at 4 °C. The supernatant (cytosolic fraction) was collected and stored at -80°C .

Gel Electrophoresis and Immunoblotting

Western blotting was carried out with 30 μg protein in precast gels (8–16 % (w/v) acrylamide) in a Tris-Glycine buffering system (Invitrogen, Carlsbad CA). Gel electrophoresis was performed

at 125 V/gel for 1.5 h. Polyacrylamide gels were incubated in transfer solution for 1 h at room
135 temperature. Protein transfer onto a HybondTM-ECL nitrocellulose membrane (GE Healthcare,
Piscataway, NJ) was performed with an Owl HEP-1 semi-dry electroblotting system (Thermo
Fisher Scientific, Rockford, IL), according to manufacturer instructions. Transferred proteins
were probed with primary antibodies targeted at organelle proteins VDAC-1, PsbA, H⁺ATPase,
fructose-1,6-biphosphatase (Agrisera, Vännäs, Sweden), histone H3 and calreticulin (Abcam,
140 Cambridge, UK). Detection was undertaken using the peroxidase linked secondary antibody
Anti-Rabbit IgG (Sigma-Aldrich, St. Louis, MO) with SuperSignal West Dura Extended
Duration Substrate (Thermo Fisher Scientific, Rockford, IL) and recorded on a BioSpectrumAC
Imaging System (UVP, Upland, CA).

145 ***Total Soluble Protein Extraction from Arabidopsis Cell Culture***

Seven day-old Arabidopsis suspension-cultured cells were filtered through one layer of
Miracloth (Calbiochem/EMD Chemicals, Gibbstown, NJ) to remove medium. Approximately 10
g of cells (FW) were homogenized with a Polytron blender (Kinematica, Kriens, Switzerland) in
30 mL of 100 mM Tris-HCl, 3 mM EDTA, 20 mM ascorbic acid, pH 8 and Complete Protease
150 Inhibitor Cocktail (Roche, Mannheim, Germany) at 4 °C. The homogenate was filtered through 3
layers of Miracloth and centrifuged at 20,000 × g for 30 min at 4 °C. The supernatant (total
soluble protein) was desalted and concentrated with 5 kDa Ultrafree centrifugal filter devices
(Millipore, Billerica, MA) and stored at −80 °C.

155 ***Selected Reaction Monitoring Analysis***

Three technical replicates of cytosolic and total soluble protein tryptic digests were analyzed using an Applied Biosystems 4000 Q-Trap mass spectrometer coupled to a TEMPO nanoLC-2D (Eskigent, Dublin, CA). One μg of each sample along with 50 fmol of BSA digest was loaded onto a Pepmap100 μ -guard column (Dionex/LC Packings, Sunnyvale, CA) via a Tempo Nano Autosampler and washed with buffer A (98 % H_2O , 2 % ACN, 0.1 % formic acid) for 20 min (15 $\mu\text{l}/\text{min}$ flow rate) before injection onto a Dionex Pepmap100 analytical column (75 μm i.d., 150 mm length, 100 $\text{\AA}$, 3 μm) at a flow rate of 300 nl/min.. The LC gradient, using buffers A and B; (98 % ACN, 2 % H_2O , 0.1 % formic acid) was as follows: 5-30 % B over 2 min, 30-50 % B in 15 min, 50-80 % B in 3 minutes. MS/MS spectra were collected for 2 s over a mass range of 100–1,600 m/z with Q1 resolution = LOW and rolling collision energy. A list of target peptides, from which SRM transitions were chosen were generated by processing the IDA results with both Mascot 2.1 (Matrix Science, Boston, MA) and Protein Pilot 2.0 (Applied Biosystems, Foster City, CA) search programs (Table 1). SRM data was collected by Analyst 1.5 (Applied Biosystems) operating in Multiple Reaction Monitoring (MRM) mode. Peptides were eluted off the LC by using the same gradient as the IDA method with the dwell time for the SRM transitions set to 100 ms and the full cycle time of less than 3 s. Peak area quantification was determined using MultiQuant 1.2TM (Applied Biosystems) with gaussian smooth width set to 3.

Strong Cation-Exchange (SCX) Chromatography

Protein lysates representing the cytosolic fractions were diafiltrated and concentrated with 5 kDa Ultrafree centrifugal filter devices (Millipore, Billerica, MA). Trypsin (Invitrogen; Carlsbad, CA) was added to a final ratio of 1:10 (w/w) and incubated overnight at 37 °C. Samples were concentrated in a SpeedVac, dissolved in buffer A (5 mM KH_2PO_4 , pH 3 and 25 % ACN), and

loaded onto a 4.6 × 200-mm polySULFOETHYL aspartamide A column (PolyLC, Columbia,
MD) on an UltiMate HPLC system (Dionex/LC Packings, Sunnyvale, CA). Peptides were eluted
with an increasing KCl gradient (25–68 min, 0 %–30 % buffer B; 69–105 min, 30 %–100 %
buffer B; where buffer B consisted of 5 mM KH₂PO₄, pH 3, 800 mM KCl and 25 % ACN). The
eluate was collected in 15-16 fractions, desalted and concentrated with C18 Macro SpinColumns
(Harvard Apparatus, Holliston, MA) and further concentrated in a SpeedVac.

Quadrupole Time-of-Flight (Q-TOF) Mass Spectrometry

Concentrated peptide fractions were resuspended in 3 % ACN and injected onto a Pepmap100 μ-
guard column (Dionex/LC Packings, Sunnyvale, CA) via Tempo Nano Autosampler and washed
with buffer A (98 % H₂O, 2 % ACN, 0.1 % formic acid) for 20 minutes. Peptides were then
separated on a Dionex Pepmap100 analytical column (75 μm i.d., 150 mm length, 100 Å, 3 μm)
using a TEMPO nanoLC-2D (Eskigent) system coupled to a QSTAR Elite Q-TOF mass
spectrometer (Applied Biosystems, Foster City, CA). Separated peptides were dispersed by
nano-electrospray ionization (ESI) into the QSTAR Elite. The LC gradient consisted of buffers A
and B (98 % ACN, 2 % H₂O, 0.1 % formic acid) with: 2 % to 37 % B over 88 minutes, 37 % to
80 % in 10 minutes where it was held for 10 minutes, 80 % B to 90 % B in 8 minutes, followed
by a sharp decline to 2 % B in 2 minutes where it is held for 30 minutes. Three product ion scans
were collected from each cycle with a maximum 2 second accumulation time depending on
intensities of fragment ions. A threshold of 50 counts was required for ions to be selected for
fragmentation. Parent ions and their isotopes were excluded from further selection for 1 min,
with a mass tolerance of 100 ppm.

LC-MS/MS Data Interpretation and Analysis

Data produced on the QSTAR Elite MS/MS system were exported as .mgf files from Analyst QS v1.1 using the script available from Matrix Sciences (version 1.6b23). The script was set to
205 centroid the survey scan ions (TOF MS) at a height percentage of 50 % and a merge distance of 0.1 amu (for charge-state determination), centroid MS/MS data at a height percentage of 50 % and a merge distance of 2 amu, reject a CID if less than ten peaks, discard ions with charge equal and greater than 5+. All MS/MS data from each independent cytosol preparation (total three) were merged into a single data file. Prior to database interrogation parent spectra were calibrated
210 using the DTARefinery utility (version 1.2) {Petyuk, 2010 #415}. Briefly, .mgf data files were converted to .dta data files and used to query the latest *Arabidopsis thaliana* protein set from The Arabidopsis Information Resource (TAIR version 9, need sequence no and total residue numbers??) by DTARefinery which employs X!Tandem (0.01 Maximum Valid Expectation Value, ± 200 ppm Parent Monoisotopic Mass Error Tolerance and dynamic PTMs: 15.9949@M).
215 Recalibration of parent ions employed Additive Regressions (Median and Lowess) using default settings with only the m/z dimension selected. Recalibrated data were interrogated with the Mascot search engine version 2.2.04 (Matrix Science, Boston, MA) with a peptide tolerance of ± 100 ppm and MS/MS tolerance of ± 0.3 Da; variable modification was Oxidation (M); up to one missed cleavage for trypsin was selected and the instrument type was set to ESI-QUAD-TOF.
220 Searches were performed against an in-house database comprising of *Arabidopsis thaliana* proteins (TAIR9, need sequence no and total residue numbers??) and also included human keratin, BSA and trypsin sequences. A false discovery rate and ions score or expected cut-off was calculated for each experiment by combining all data (each experimental set of SCX runs) and interrogating Mascot using the Decoy feature on the MS/MS Ions Search interface. A

225 significance threshold was selected to produce a false discovery rate of $\leq 5\%$ and to determine
the ions score or expected cut-off to be employed. Specific false discovery rates were 2.71 % for
AtCyto-1, 2.9 % for AtCyto-2 and 2.28 % for AtCyto-3 employing Ions score or expect cut-off
for peptides of 29 ($p < 0.05$). (this is saying ion score cut off of 29? What is an expect cut-off?)

230 *Characterization of the Arabidopsis Cytosol*

Functional assignments of identified proteins were initially performed using the MapMan
BinCodes {Thimm, 2004 #432} and then manually reassigned using AraCyc metabolic pathways
{Zhang, 2005 #417} and homology-based comparisons with Basic Local Alignment Search Tool
(BLAST) {Altschul, 1990 #319}. Transmembrane domain predictions were pre-calculated by the
235 HMMTOP program {Tusnady, 1998 #414} and expressed sequence tags numbers (ESTs) were
retrieved from TAIR {Swarbreck, 2008 #418}. Reports of experimental or predicted subcellular
location were obtained from the Arabidopsis Subcellular database (SUBA) {Heazlewood, 2007
#345}. Further prediction information were obtained from BaCelLo {Pierleoni, 2006 #426},
WoLF PSORT {Horton, 2007 #427}, LOCTree {Nair, 2005 #428}, SLP-Local {Matsuda, 2005
240 #429} and SubLoc {Hua, 2001 #431}.

Results

Enrichment of the Arabidopsis Cytosol

245 The major challenge in isolating the cytosol from plant cells is the need to physically disrupt the plant cell wall but to maintain organelle integrity, in order to avoid gross contamination of the cytosol by soluble protein content from organelles. The strategy utilized to isolate the cytosol from Arabidopsis cell cultures was developed from previous approaches isolating intact mitochondria and peroxisomes from Arabidopsis cell cultures. These approaches indicated that

250 organelle integrity was best maintained by forming protoplasts through enzymatic removal of cell walls followed by gentle pressure disruption of protoplasts with a Potter homogenizer {Meyer, 2008 #422; Eubel, 2008 #419}. Successive centrifugation steps of $800 \times g$, $10,000 \times g$ and $100,000 \times g$ facilitated the sequential removal of unbroken cells, large cellular debris, organelles and small vesicular components from the sample. To ascertain the distribution of

255 organelles and assess purity of the cytosolic fraction during the enrichment process, samples were analyzed by Western blotting with antibodies raised against protein markers for mitochondria, plastids, nucleus, the endomembrane system, plasma membranes, and the cytosol (Figure 1A). The marker for the cytosolic glycolytic enzyme fructose-1,6-bisphosphatase (cFBPase) was present throughout the enrichment process and present at significant quantities in

260 the cytosolic fraction. The presence of this soluble cytosolic enzyme in insoluble enrichment fractions ($10,000 \times g$ and $100,000 \times g$ pellets) is likely due to the association of the glycolytic enzymes with organelles such as mitochondria {Giege, 2003 #419}. In contrast, the signals for organelle markers VDAC-1 (mitochondria) and calreticulin (endomembrane system) were minimal in the cytosolic fraction and absent for PsbA (plastid), histone H3 (nucleus), and

265 H^+ ATPase (plasma membrane) (Figure 1A). These results indicated only slight mitochondrial
and endomembrane system contamination in the cytosolic fraction, with virtually no detection of
plastid, plasma membrane or nuclei. Analysis of the enriched cytosolic fraction by SDS-PAGE
indicated the integrity of the sample during the isolation procedure (Figure 1B). Similar analyses
were undertaken on a further two cytosolic enrichment procedures to produce three highly
270 enriched Arabidopsis cytosolic fractions (Supplementary Figure 1).

Selected Reaction Monitoring Analysis

A mass spectrometry-based method for detecting organelle contamination in the cytosolic
fractions was developed to supplement Western blotting results. Selected Reaction Monitoring
275 (SRM) was used here to measure the intensities of select organelle and cytosolic marker
precursor peptide/fragment ion pairs (transitions) as indicative of their relative parent protein
abundances in Arabidopsis total soluble protein and cytosolic fractions. Candidate peptides from
organelle proteins were selected for SRM profiling if they were identified in the cytosolic
fraction when analyzed by LC-MS/MS. Peptides were chosen from well-documented marker
280 proteins (Table 1). A total soluble protein lysate and a cytosolic fraction from Arabidopsis cell
culture were both analyzed in triplicate and the relative average transition signal intensities of the
six markers were compared in each sample (Figure 2). In the total soluble protein fraction, the
average transition intensity of the cytosolic marker LOS1 was 4 to 10 times greater than the four
organelle markers ($5 \times \text{ATP-}\beta$, $10 \times \text{CAT}$, $6 \times \text{GAPA}$ and $4 \times \text{HIS-H4}$), but in the cytosolic
285 fraction this was significantly increased by around 11 fold with values 55 to 105 times greater
($105 \times \text{ATP-}\beta$, $55 \times \text{CAT}$, $71 \times \text{GAPA}$ and $78 \times \text{HIS-H4}$) (Figure 2). Similarly, the average
transition intensity of the second cytosolic marker GAPC in the total soluble protein fraction was

2 to 5 times greater than the four organelle markers ($3 \times \text{ATP-}\beta$, $5 \times \text{CAT}$, $3 \times \text{GAPC}$ and $2 \times \text{HIS-H4}$), but in the cytosolic fraction this was also considerably increased by around 9 fold with values from 18 to 37 times greater ($37 \times \text{ATP-}\beta$, $18 \times \text{CAT}$, $25 \times \text{GAPC}$ and $27 \times \text{HIS-H4}$) (Figure 2). These results were indicative of marked reduction in the abundance of the four organelle proteins in the cytosolic fraction relative to cytosolic LOS1 and GAPC, when compared with their relative abundances in the total soluble protein fraction. This supported our Western blotting results indicating the majority of organelles were removed during our isolation of the cytosolic fraction.

Characterization of the Arabidopsis Cytosol by LC-MS/MS

Cytosolic fractions from the three enrichment procedures were fractionated by preparative SCX-HPLC into 15-16 peptide fractions and each analyzed by LC-MS/MS. The calibrated and concatenated MS/MS data files for each sample (AtCyto-1, AtCyto-2 and AtCyto-3) were analyzed using the Mascot search engine against the latest Arabidopsis protein release (TAIR9). The number of non-redundant proteins identified in each sample using stringent MS/MS matching conditions was 1,375 (AtCyto-1), 1,252 (AtCyto-2) and 1,734 (AtCyto-3) resulting in a total number of 2,264 protein identifications by all three experiments combined (Supplementary Table 1). The reproducibility of both the cytosolic preparations and subsequent mass spectrometry is illustrated by the minimal number of exclusive identifications from each experiment. Only 214 proteins or 9.5 % of the overall total of 2,264 protein identifications for both AtCyto-1 and AtCyto-2 were unique identifications and 506 proteins or 22.3 % of the overall total for AtCyto-3 were unique (Supplementary Figure 2).

In order to improve confidence levels for defining the Arabidopsis cytosolic proteome and to account for deviations in the preparations and MS/MS data matching; a final list of proteins was constructed that required the identification of a protein in at least two of the three independent experiments. The total number of proteins that fulfilled this criterion was 1,364 proteins and

315 represents the stringent set of proteins reproducibly identified in this study (Supplementary Table 2a). A number of proteins identified in the cytosolic samples by mass spectrometry were identified only by redundant peptides matches and were excluded. After combining the three cytosolic preparations with the above criteria, a total of 624 proteins could not be conclusively identified in any of the three preparations as they shared redundant matches to the 1,364. While
320 nearly 80% of the 624 redundant proteins were predicted to arise by alternate splicing, they still represent potentially valid identifications arising from this analysis (Supplementary Table 2b).

Defining the Arabidopsis Cytosolic Proteome

While both Western and SRM analyses with organelle markers demonstrated the level of
325 contamination in the preparation was minimal, we wanted to define a stringent proteome that best represented the Arabidopsis cytosolic proteome. Our approach to define this set of proteins and the level of organelle contamination involved interrogation of the cytosolic fraction (1364 proteins) against SUBA, the Arabidopsis SUB-cellular Database {Heazlewood, 2007 #345}. SUBA contains localization evidence for 5689 proteins by MS originating from 23 published
330 reports. SUBA also contains 1953 individual protein localizations by fluorescent targeting (FP). Given this coverage, proteins were considered contaminants of the cytosol preparation if they were previously reported in mitochondria, chloroplast, plasma membrane, cell wall, vacuole or peroxisome at least two times by separate proteome analysis (MS) and protein fusion study (FP)

or at least three times by the same method (MS or FP) in the same subcellular location

335 (Supplementary Table 3). Proteins that met these criteria, but were also shown by FP to localize to the cytosol or showed unclear FP localization were not included as contaminants. The limited number of proteomic analyses and their relatively small defined protein sets from the nucleus (3 reports; 36, 158 and 217 proteins), ER (2 reports; 10 and 182 proteins) and Golgi (2 reports; 10 and 89 proteins) indicated they were poorly represented and only consisted of abundant proteins
340 from these proteomes. Consequently we used less stringent cut-offs, utilizing two localizations by either both or the same method (MS and/or FP) for the nucleus and at least one location by either method in ER and Golgi. This initial process produced a collection of prominent contaminants (180 proteins) from various cellular organelles.

345 A large number of conflicting subcellular localization information (for both FP and MS) still remained in the list of cytosolic proteins. To further eliminate potentially contaminating proteins and construct a highly robust Arabidopsis cytosolic proteome we analyzed the Mascot score distributions of these prominent contaminating proteins identified above. Since the Mascot output provides a relative quantitation value through the Exponentially Modified Protein
350 Abundance Index (emPAI) for each protein based on peptide matches {Ishihama, 2005 #425}, and previous Western and SRM analyses indicated low contamination levels in the preparation, we examined the score distribution of the major contaminants to identify a representative value. The emPAI values for the major contaminating proteins showed an extremely skewed distribution, with relative content (mol %) ranging from 0.0024 to 0.9934. As a result, we
355 employed the median value to represent a contamination score for each cytosolic preparation; 0.03795 (AtCyto-1), 0.03880 (AtCyto-2) and 0.02928 (AtCyto-3). These values were

subsequently employed as cut-offs to remove any protein with conflicting experimental localizations (FP or MS) and all emPAI scores below the above determined values. A further 113 proteins were removed from the cytosolic analysis producing a total of 287 contaminants and represents 21.0 % of all proteins identified in the defined cytosolic fraction (total 1364). Of the 293 contaminants, 52 (3.8 %) were plastid proteins, 28 (2 %) plasma membrane, 22 (1.6 %) nuclear, 20 (1.5 %) peroxisomal, 17 (1.2 %) extracellular, 14 (1.0 %) mitochondrial, 9 (0.8 %) ER, 10 (0.7 %) vacuolar, 8 (0.6 %) Golgi proteins and 108 (8.3 %) were defined as miscellaneous. Thus the defined Arabidopsis cytosolic proteome from this study comprises 1071 proteins.

Functional Classification of the Arabidopsis Cytosol

To further confirm the integrity of the newly defined cytosolic proteome, an analysis of transmembrane domain numbers was undertaken. The cytosolic proteome will likely represent a set of soluble proteins that are unlikely to contain transmembrane domains. While both Western blots and SRM analysis indicated that membrane systems and hence membrane proteins were clearly removed during the cytosolic enrichment procedure; we decided to further corroborate these findings by analyzing the presence of predicted transmembrane domains (TMD). Analysis of the cytosolic proteome outlined above (1071) with the HMMTOP TMD prediction program {Tusnady, 1998 #414} identified 1034 (96.5 %) proteins with no predicted TMD. This was considerably higher than the 76.7 % of proteins with no predicted TMD in the entire Arabidopsis proteome (Figure 3). A further 32 (3.0 %) were predicted with a single TMD, compared with 12.3 % for the TAIR9 protein set and even more strikingly, only 5 (0.5 %) had two or more predicted TMDs, compared with 11 % for all proteins in TAIR9 (Figure 3). The removal of

380 contaminants from the cytosolic preparations defined by LC-MS/MS (1364) reduced the number of proteins with TMD by 39 proteins, producing a small change in the soluble (no TMD) fraction of the cytosolic proteome (1071), adding some validation to this process. Thus the lack of predicted TMDs for the vast majority of the defined cytosolic proteome supported results from Western blot and SRM organelle marker analyses that both the cytosolic fraction defined by LC-
385 MS/MS (1364) and the cytosolic proteome (1071) were largely free of membrane associated contamination.

To understand the functional roles of the cytosol in cellular metabolism the defined cytosolic proteome were assigned to functional categories. Using the broad characterizations outlined by
390 MapMan a total of twenty eight functional groups were assembled (Figure 4). Unsurprisingly, the vast majority of the cytosolic proteome is associated with the regulation of protein synthesis (32 %), involving the translational machinery, degradation, modifications and folding. Other functional categories associated with protein biosynthetic processes include the RNA functional group which is comprised of RNA interaction proteins. Thus the total proportion of proteins in
395 the cytosolic proteome involved in RNA and protein processing is around 40 %. While only 11 % of the proteome is designated as unknown (not assigned), many of the functional assignments that fall within these broad categories are made based on the presence of a known functional domains rather than experimental information. Nonetheless well characterized processes were identified including glycolysis, cell wall precursor biosynthesis, S-adenosyl-
400 methionine cycle (Amino Acid), phenylpropanoid biosynthesis (Secondary), protein kinases / phosphatases and 14-3-3 proteins (Signaling) as well as actins and tubulins (Cell).

Bioinformatic Profiling the Arabidopsis Cytosol

The prediction of subcellular location has been widely used to determine the subcellular location
405 of proteins within the cell. Currently few prediction algorithms attempt to determine cytosolic
localization of a protein. Since this study represents the first large scale characterization of the
cytosol in plants we decided to analyze the performance of current prediction to this subcellular
location. A number of publicly available programs attempt to identify cytosolic localizations in
plants including WoLF PSORT, LOCTree, BaCelLo, SLP-Local and SubLoc. A performance
410 analysis of these algorithms in assessing cytosol localization was conducted with the
experimentally determined (1071) cytosolic proteome (Table 2). The accurate assignment rate
amongst the five predictors varied considerably with a range between 25.8 % (BaCelLo) and
72.0 % (SLP-Local). Importantly these values should be viewed in the context of global
predictions of cytosol localization in Arabidopsis where over prediction can be assessed
415 {Heazlewood, 2004 #423}. While global values for SLP-Local were unavailable, it should be
noted that this particular program does not discriminate between the nucleus and the cytosol
which likely greatly inflates its positive rate. Consequently the performances of both LOCTree
and WoLF PSORT in positively assigning cytosolic localizations at 36.6 % and 47.1 %
respectively at low global prediction levels (10.7 % and 18.1 % respectively) are the better
420 performing algorithm in this analysis.

A comparison of the defined cytosolic proteome (1071) against the SUBA database identified
only 94 proteins with evidence of cytosolic localization by tagged fluorescent protein studies
(Supplementary Table 3). The SUBA database reports that the total number of proteins localized
425 to the cytosol by fluorescent protein experiments is 397. This poor intersection between

proteomic studies and fluorescent tagged protein localizations is common due to proteomic studies identifying more abundant proteins while fluorescent protein localization experiments often arise from studies of specific proteins of interest (e.g. studies of the localizations for all members of a multigene family). In the case of the cytosolic proteins, only 218 proteins (55 %) of the 397 proteins were represented by 10 or more ESTs in transcript sequence databases. In contrast 83 of the 94 cytosolic proteins (88.3 %), which were both identified by this study and confirmed by fluorescent tagged protein studies, are represented by more than 10 ESTs in sequence databases.

By further employing the Boolean search options available at the SUBA database we were also able to perform an analysis of the false positive and false negative rates of WOLF PSORT, SubLoc and LocTree in predicting cytosolic protein localization. This is possible by assessing the proportions of cytosolic predicted versus experimentally verified cytosolic proteins, and the cytosolic predicted versus experimental proven non-cytosolic protein sets for each predictor (Supplementary Table 4). This analysis shows these predictors have a 0.53 to 0.71 false positive rate and a 0.50 to 0.66 false negative rate against experimental data, indicating the need for improved prediction or more experimental verification. These data can also be utilized to predict the size of the cytosolic proteome in Arabidopsis in the same way they were used to predict the size of the mitochondrial and chloroplast proteomes {Millar, 2006 #424}. Based on this analysis (Supplementary Table 4), the Arabidopsis cytosolic proteome is predicted to contain $\sim 5400 \pm 650$ proteins, thus the experimental set reported in this manuscript would be $\sim 20\%$ of the total cytosolic proteome.

Discussion

450

A large proportion (~ 90 %) of the 1071 proteins defined as the Arabidopsis cytosolic proteome were allocated to functional categories through gene annotation information, MapMan Bins, sequence homology, functional domains and annotations from Arabidopsis metabolic pathways.

Not surprisingly, components related to protein biosynthesis and degradation machinery

455 dominates the cytosol with nearly a third of the identified proteins involved in these processes.

Nonetheless we were also able to characterize an array of metabolic pathways located in the cytosol including glycolysis, phenylpropanoid and isoprenoid biosynthesis, the S-

adenosylmethionine cycle, Vitamin B₆ metabolism, nucleotide metabolism and nucleotide-sugar biosynthesis.

460

Contaminants of the Cytosol Preparations

The large number of stringent protein identifications (1364) from our in-depth analysis of the Arabidopsis cytosolic fraction presented the problem of identifying low-abundant cytosolic proteins from organelle contaminants. This dilemma was similarly highlighted by two recent

465 large-scale proteomic studies of Arabidopsis chloroplasts where 25 to 30 % of the identified proteins could not be verified by previous experimental data or subcellular prediction programs {Zybailov, 2008 #417; Ferro, #418}. Further complicating our efforts to define the Arabidopsis

cytosolic proteome were dealing with the cytosolic proteins that functionally associate with

organelles, such as glycolytic enzymes with mitochondria {Giege, 2003 #419}, members of the

470 ubiquitin-proteasome pathway with the nucleus {Hotton, 2008 #348; Smalle, 2004 #393} and

polysomes (ribosomes bound to mRNA) with various membrane surfaces {de Jong, 2006 #420}

as prominent examples. We also found that computational programs developed for predicting protein targeting to organelles could not provide us with confident cytosolic predictions by analyzing their false positive and false negative rates against experimental data. We therefore
475 relied on the large experimental sets of MS (5689 proteins) and FP (1953 proteins) localizations in the SUBA database {Heazlewood, 2007 #345} to develop a method of elimination. After utilizing this approach we were confident that we had produced a robust set of proteins that best reflect the major constituent of the cytosolic proteome.

480 ***Protein Biosynthesis and Degradation in the Cytosol***

The major components driving protein synthesis in the plant cytosol; ribosomes, aminoacyl-tRNA synthetases and translation factors, were all identified in this study. Cytosolic ribosomes are large ribonucleoprotein complexes mediating the peptidyl transferase reaction of polypeptide synthesis, fundamental for translating proteins from transcripts encoded in eukaryotic nuclear
485 genomes {Bailey-Serres, 2009 #322; Carroll, 2008 #330; Chang, 2005 #333}. Plant 80S cytosolic ribosomes are ~3.2-MDa in size, slightly smaller than their mammalian counterparts and consist of two subcomplexes-the small 40S subunit and large 60S subunit {Chang, 2005 #333; Cammarano, 1972 #329}. In this broad survey of the cytosolic proteome, 93 previously confirmed members from 60 ribosomal gene families were identified. Additionally, other central
490 components of plant protein synthesis were also identified. Aminoacyl-tRNA synthetases facilitate the direct attachment of a specific amino acid to its corresponding tRNA to form aminoacylated tRNAs (aa-tRNAs), where they are covalently attached to the growing peptide chain by ribosomes using the mRNA as the template {Pujol, 2008 #383}. Aminoacyl-tRNA synthetases for 19 out of 20 amino acids were identified (tyrosine tRNA synthetase was not

495 identified). In conjunction with a multitude of translation initiation factors, elongation factors
and a release factor that mediate the initiation, elongation and termination steps of protein
synthesis, these data underscore the prevalence of the protein synthesis machinery in the cytosol.

In both the plant cytosol and nucleus a selective degradation process occurs that eliminates
500 damaged, misfolded and/or regulatory proteins no longer required. This process is largely
mediated by another major multi-enzyme complex, namely the ubiquitin/26S proteasome
pathway {Smalle, 2004 #393; Hotton, 2008 #348}. Proteins targeted for degradation are
covalently tagged with the highly conserved small protein ubiquitin (Ub) by a cascade of three
enzyme groups: E1 (ubiquitin-activating enzyme), E2 (ubiquitin-conjugating enzyme) and E3
505 (ubiquitin ligase). Following ubiquitination, most Ub-protein conjugates are recognized and
degraded by the ATP-dependent protein complex-26S proteasome {Yang, 2004 #407}. The
Arabidopsis 26S proteasome comprises of 31 primary protein subunits in two subcomplexes, the
20S Core Protease (CP) and the 19S Regulatory Particle (RP) {Vierstra, 2009 #399}. A previous
proteomic characterization of highly purified 26S proteasome from Arabidopsis seedlings
510 identified most of the CP and RP subunits {Yang, 2004 #407}. In our analysis of the cytosolic
proteome we identified many of the same 26S proteasome components, including 14 out of 24
CP subunits, 7 of the 11 RPT subunits and 13 of the 18 RPN subunits. Major proteins involved
in the ubiquitin conjugation cascade were also identified here, including members of E1, E2 and
E3 and a number of ubiquitin related proteins.

Carbohydrate Metabolism in the Cytosol

Glycolysis is the fundamental metabolic pathway found in virtually all living organisms where hexose sugars are converted to ATP, pyruvate and substrates for various anabolic reactions {Plaxton, 1996 #378}. Plant glycolysis utilizes sucrose and starch as principal substrates, taking
520 place in either the plastid or the cytosol {Plaxton, 1996 #378}. A hallmark of plant cytosolic glycolysis is its flexibility to switch between alternative enzymatic reactions using ATP or pyrophosphate (PPi) as energy donors. This is believed to be modulated by factors such as tissue type, the developmental stage of the plant and various environmental stresses {Plaxton, 1996 #378}. In Arabidopsis cell cultures, we identified all the cytosolic enzymes in the glycolytic PPi-
525 dependant alternate pathway, beginning with sucrose synthase catalyzing the reaction of sucrose to UDP-glucose and ending with pyruvate kinase converting phosphoenolpyruvate to pyruvate {Plaxton, 1996 #378}. This included enzymes mediating the two PPi-dependent glycolytic steps: UDP-glucose pyrophosphorylase and two genes encoding phosphofructokinase (PFP). In addition the cytosolic enzymes phosphoenolpyruvate carboxylase and malate dehydrogenase,
530 along with mitochondria-localized malic enzyme (not identified in this study) are thought to form a glycolytic bypass to the reaction of phosphoenolpyruvate to pyruvate by cytosolic pyruvate kinase {Plaxton, 2004 #379}.

As with glycolysis the pentose phosphate pathway (PPP) is a related and central metabolic
535 pathway found in most organisms generating reductant (NADPH) and pentose sugars by two respective stages; oxidative (OPPP) and non-oxidative {Kruger, 2003 #360}. NADPH is used by plants for reductive biosynthetic reactions including fatty acid synthesis and the assimilation of inorganic nitrogen and to protect against oxidative stress {Neuhaus, 2000 #373; Juhnke, 1996 #353}. Pentose sugars are utilized as carbon skeletons for the synthesis of many important

540 molecules including nucleotides, aromatic amino acids, phenylpropanoids and lignin {Allen, 2009 #317; Herrmann, 1999 #346}. In plants it has been shown with castor bean, soybean, cauliflower, tobacco, pea and spinach oxidative and non-oxidative stages are in the plastid and the oxidative stage is in the cytosol, but it is not clear if the non-oxidative stage does take place in the cytosol {Debnam, 1999 #338; Schnarrenberger, 1995 #390; Nishimura, 1979 #374; 545 Journet, 1985 #352; Hong, 1990 #347}. Intermediates of PPP can be exchanged between the cytosol and plastid through a family of pentose phosphate translocators across the plastid inner envelope, which may compensate for any absence of the non-oxidative stage in the cytosol {Eicks, 2002 #342}. In Arabidopsis we have identified the two NADPH-producing enzymes of oxidative PPP that also generate glucono- δ -lactone-6'-phosphate and ribulose-5'-phosphate 550 respectively; glucose-6'-phosphate dehydrogenase (G6PD) and 6'-phosphogluconate dehydrogenase. This included two cytosolic G6PDs; G6PD5 and 6 {Wakao, 2008 #401} and all three 6'-phosphogluconate dehydrogenase genes with one (At3g02360) previously shown by GFP to localize in the cytosol {Reumann, 2007 #385}. We have also identified the two enzymes in the first branch step of non-oxidative PPP, a ribulose 5'-phosphate 3'-epimerase (RPE) 555 (At1g63290) and a ribose 5'-phosphate isomerase (RPI) (At1g71100). This was in agreement with a previous study investigating the likely subcellular locations of the four Arabidopsis enzymes of non-oxidative PPP with computational N-terminal targeting sequence analysis. Three RPEase isoforms, including At1g63290 and two RPI isoforms including At1g71100 were predicted as cytosolic {Howles, 2006 #411}. RPI converts ribulose 5'-phosphate from OPPP to 560 ribose 5'-phosphate, which serves as a substrate for Vitamin B₆ and nucleotide biosynthesis.

Vitamin B₆ Biosynthesis

Vitamins are essential organic compounds with varying molecular structures required only in small amounts to exert their effects across a broad range of cellular reactions. Only recently have several pathways of vitamin biosynthesis in plants been unraveled, with the biosynthetic pathways of Vitamins B₅, C and H taking place in the cytosol and mitochondria, B₉ in the cytosol, plastid and mitochondria and B₆ believed to take place entirely in the cytosol {Baldet, 1997 #323; Tambasco-Studart, 2005 #397; Pinon, 2005 #377; Smirnoff, 2001 #394; Coxon, 2005 #335; Sahr, 2005 #388}. Vitamin B₆ is a coenzyme for many metabolic enzymes and possessing antioxidant properties. Two distinct pathways involved in its metabolism are known in Arabidopsis; the *de novo* biosynthetic and salvage pathways {Roje, 2007 #387; Tambasco-Studart, 2005 #397}. The active form of Vitamin B₆; pyridoxal 5'-phosphate is synthesized from pentose phosphate pathway intermediates ribose 5'-phosphate or ribulose 5-phosphate and glycolytic intermediates glyceraldehyde 3'-phosphate or dihydroxyacetone phosphate by two genes (*PDX1* and *PDX2*) {Tambasco-Studart, 2005 #397}. Two functional Arabidopsis homologs of *PDX1* (PDX1.1 and 1.3) along with the single *PDX2*, co-localize to the cytosol {Tambasco-Studart, 2005 #397}. We identified both PDX1.1 and 1.3 and interestingly the claimed to be non-functional PDX1.2 {Tambasco-Studart, 2005 #397}, but we did not identify PDX2.

Nucleotide Biosynthesis

Nucleotides are central components for many important biological processes; they form the structural backbones of RNA and DNA, they provide cellular energy as adenosine-5'-triphosphate (ATP) and guanosine-5'-triphosphate (GTP), they can act as cofactors in metabolic reactions through flavin adenine dinucleotide (FAD), Coenzyme A and Nicotinamide adenine

dinucleotide phosphate (NADP) and as nucleotide-sugar substrates for the plant cell wall biosynthesis. Nucleotide biosynthesis can occur via the energy-consuming *de novo* biosynthesis and energy-conserving salvage pathways. Both these pathways for purine, pyrimidine and pyridine nucleotide biosynthesis begin with the conversion of ribose 5'-phosphate to

590 phosphoribosyl- α -1-diphosphate (PRPP) by PRPP synthase {Krath, 1999 #357}. Arabidopsis contains five isoforms of PRPP synthase, where isoforms 1 and 2 require Pi for maximal activity and isoforms 3 and 4 are Pi-independent {Krath, 1999 #357}. We have identified the cytosolic PRPP synthase; Pi-independent isoform 4 {Krath, 1999 #358; Koroleva, 2005 #355}. The *de novo* biosynthetic pathway of the purines adenosine 5'-monophosphate (AMP), inosine 5'-

595 monophosphate (IMP) and guanosine 5'-monophosphate (GMP) follow a 14 step process, mostly mediated by single copy genes {Zrenner, 2006 #410}. The steps beginning with PRPP to the synthesis of IMP and AMP are thought to take place in the plastid, based on sequence analysis indicating N-terminal plastid targeting for all enzymes {Zrenner, 2006 #410}. It is not clear how IMP and AMP are exported from the plastid into the cytosol. However, recent evidence points to

600 IMP possibly being converted to AMP in the plastid, exported to the cytosol by a plastidic adenine nucleotide uniporter and converted back to IMP for GMP synthesis by AMP deaminase- a central enzyme of purine biosynthesis and catabolism {Leroch, 2005 #364; Zrenner, 2006 #410}. AMP deaminase and the two enzymes of GMP biosynthesis; IMP dehydrogenase (IMPDH) and GMP synthase (GMPS) are believed to reside in the cytosol because they all lack

605 predicted N-terminal plastid targeting sequences {Zrenner, 2006 #410}. Indeed, we provide for the first time experimental evidence of cytosolic localization for the two genes encoding IMPDH, the single gene GMPS and the single gene AMP deaminase (FAC1) in Arabidopsis. Furthermore, we have localized the central enzymes of the adenosine salvage pathway; an

adenine phosphoribosyltransferase (APT1) and the two genes encoding adenosine kinase (ADK1
610 and 2). APT and ADK convert adenine and adenosine, respectively to AMP. Previous cDNA
analysis of ADK1 and 2 predicted their locations in the cytosol and the three genes encoding
APT 1-3 were shown by immunolocalization to be cytosolic {Moffatt, 2000 #370; Allen, 2002
#318}.

615 The *de novo* pyrimidine nucleotide biosynthetic pathway involves the synthesis of uridine 5'-
monophosphate (UMP) from carbamoylphosphate, aspartate, and PRPP in six enzymatic steps
{Zrenner, 2006 #410}. All but one of these steps takes place in the plastid, with the exception
being dihydroorotate dehydrogenase which converts orotate from dihydroorotate in the
mitochondria {Zrenner, 2006 #410}. Alternatively, UMP can also be recycled via the less-
620 understood pyrimidine salvage pathway. First, pyrimidine nucleotides are sequentially
catabolized by pyrimidine 5'-nucleotidase (UMPH) and uridine nucleosidase (URH) into the
nucleoside uridine and the base uracil. Arabidopsis has a single homologous gene to human
UMPH-1 and five homologues to inosine-uridine-preferring nucleoside hydrolase of *Leishmania*
major {Zrenner, 2006 #410}. As yet none of these genes have been cloned or characterized to
625 confirm their activities. Following this, the bi-functional uracil phosphoribosyltransferase
(UPRT)/uridine kinase (UK) converts PRPP + uracil and uridine + ATP, respectively into UMP
{Zrenner, 2006 #410}. Gene expression data suggest the UMP salvage pathway may take place
in the cytosol and plastid {Yamada, 2003 #405; Schmid, 2005 #389}. We have experimental
evidence to show the main components of the pyrimidine catabolism and salvage pathway are
630 located in the cytosol: three members of bi-functional UK/UPRTs, a UK, UMPH-1 and an
inosine-uridine preferring URH. In addition, we have evidence of the successive steps of UMP

conversion to UDP, uridine 5'-triphosphate (UTP) and cytidine 5'-triphosphate (CTP) in the cytosol with the respective identifications of a UMP kinase, a broad-acting nucleoside diphosphate kinase (NDPK-1) and a CTP synthase, which has not yet been characterized in plants.

Conclusions

Our extensive analysis of the Arabidopsis cytosolic proteome produced 1071 identifications, which was a significant improvement over the previous soybean root nodule cytosolic set of 69 proteins on 2-DE. We described the components of protein synthesis and degradation and the related pathways of glycolysis, incomplete pentose phosphate pathway, Vitamin B₆, nucleotide and NDP-sugar biosynthesis to highlight the central role of the cytosol as the common stage for essential plant cellular processes. As a whole, this expanded list of 1071 plant cytosolic proteins will be important to better understand the dynamic and complex reactions of the cytosol within the plant cell.

Acknowledgments

This work was part of the DOE Joint BioEnergy Institute (<http://www.jbei.org>) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy. AHM is supported by the Australian Research Council (ARC) as an Australian Professorial Fellow and by the ARC Centre of Excellence in Plant Energy Biology. We are grateful to the Edinburgh Cell Wall Group (Prof. Stephen Fry) for providing the Arabidopsis cell culture.

Tables

	Protein	Proteotypic Peptide	Peptide Mass (amu)	Peptide m/z	Fragment Ion m/z and Series (y#)
660	LOS1	IMGPNYIPGEK	1217.61	609.81	430.23 (y4)
	GAPC	GILGYTEDDVVSTDFVGDNR	2171.01	1086.51	461.19 (y4)
	ATP- β	DAPALVDLATGQEILATGIK	1995.09	998.55	1029.54 (y10)
		VLNTGAPITVPVGR	1392.68	697.34	838.7 (y8)
	CAT	EGNFDLVGNNFPVFFVR	1969.96	985.99	764.48 (y6)
665		GPILLEDYHLLK	1538.92	513.95	1046.48 (y8)
	GAPA	DSPLDIIAINDTGGVK	1626.84	814.43	987.51 (y10)
	HIS H4	ISGLIYEETR	1179.61	590.81	697.5 (y5)

Table 1. Transitions of cytosolic and organelle marker peptides monitored by SRM.

670 Cytosol: LOS1 (translation elongation factor) and GAPC (glyceraldehyde-3-phosphate dehydrogenase C-subunit); plastid: GAPA (glyceraldehyde-3-phosphate dehydrogenase A-subunit); mitochondrion ATP- β (ATP synthase β -subunit); nucleus HIS-H4 (histone H4) and peroxisome CAT (catalase).

675	Prediction Program	Cytosol	Proteome	TAIR9	TAIR9
		Proteome	(%)	Proteome	(%)
	WoLF PSORT	504	47.1 %	5,760	18.1 %
	LOCtree	392	36.6 %	3,462	10.9 %
	BaCelLo	276	25.8 %	6,012	19.8 %
680	SLP-Local	771	72.0 %	n/a	n/a
	SubLoc	592	55.3 %	8,773	27.5 %

Table 2. Assessment of various algorithms for the prediction of cytosolic localizations in Arabidopsis.

685 Cytosol Proteome is number of proteins predicted as cytosolic from the defined Arabidopsis cytosolic proteome (1071). Proteome (%) is the percentage this represents of the 1071. TAIR9 Proteome is the number of proteins predicted as cytosolic from Arabidopsis. TAIR9 (%) is the percentage that the prediction represents to the total proteome potentially coded by Arabidopsis. Global values were unavailable for SLP-Local.

690

Figure Legends

Figure 1. Immunological analysis of cytosolic enrichment procedure.

695 (A) Protein lysate (30 μ g) from *Arabidopsis* cell culture protoplasts (protoplast), 10,000 $\times$ g crude mixed organelle pellet (10K pellet), 100,000 $\times$ g crude mixed organelle pellet (100K pellet) and cytosolic fraction (cytosol) were separated by SDS-PAGE and analyzed by Western blotting. Polyclonal antibodies were used to detect antigens of *Arabidopsis* cFBPase (cytosol), H⁺-ATPase (plasma membrane), calreticulin (endomembrane system), histone H3 (nucleus),
700 VDAC-1 (mitochondria) and PsbA (plastid).

(B) Approximately 200 μ g of protein from the *Arabidopsis* cytosolic fraction separated on a 12 % SDS-PAGE gel and stained with Coomassie Brilliant Blue. Molecular weight markers are marked on the left of the gel (kDa).

705 Figure 2. Assessment of organelle contamination by Selected Reaction Monitoring (SRM).

Comparisons of average transition intensities of marker peptides for the cytosol (LOS1 and GAPC), mitochondrion (ATP- β), peroxisome (CAT), plastid (GAPA) and nucleus (HIS-H4) in the total soluble protein lysate (Total Soluble Prot.) and cytosolic fraction (Cytosol). Both samples were analyzed in triplicate and the relative average transition signal intensities of the six
710 markers and standard deviations (error bars) are displayed. CPS (normalized) is the signal intensity measured in counts per second by a 4000 Q-Trap mass spectrometer and normalized with a 50 fmol BSA digest added to each run.

Figure3. Predicted transmembrane domains (TMD) frequencies in the cytosolic sample compared to the entire Arabidopsis proteome (TAIR9).

Pre-computed transmembrane predictions by HMMTOP for Arabidopsis proteins (TAIR9) were downloaded from The Arabidopsis Information Resource (TAIR). Frequency distributions of the number of predicted transmembrane domains were analyzed using a histogram. A total of seven bins representing number of predicted transmembrane domains per protein were employed.

Figure 4. Analysis of the cytosolic fractions by functional category.

A comparison of the defined Arabidopsis cytosolic proteome (1071) by functional categories outlined by MapMan {Thimm, 2004 #432}. Proteins involved in the biosynthesis and degradation of protein biosynthesis dominate the proteome (Protein, RNA).

Figure 1

730

Figure 2.

Figure 3.

735

Figure 4.

Predictor	Predicted Cytosol Arabidopsis	Expt. any location (SUBA + 1071)	Expt. in cytosol (SUBA + 1071)	Expt. non-cytosolic	FPR cytosol prediction	Est. correct predictions	FNR cytosol prediction	Predicted cytosol	Non-predictable expt. cytosol
WoLF-PSORT	5760	1472	705	767	0.52	2759	0.55	6170.87	3412
LOCTree	3462	1110	546	564	0.51	1703	0.65	4918.535	3216
SubLoc	8773	2666	781	1885	0.71	2570	0.50	5189.43	2619

740 **Supplementary Table 4. Estimated size of the Arabidopsis cytosolic proteome utilizing data in the SUBA Database and the newly defined proteome (1071) employing the abilities of three subcellular prediction algorithms.**

Predicted Cytosol Arabidopsis: All proteins predicted to be cytosolic in Arabidopsis

Expt. any location (SUBA + 1071): Predicted cytosol and experimentally determined any location (MS or FP)

Expt. in cytosol (SUBA + 1071): Predicted cytosol and experimentally determined to be cytosolic (MS or FP)

745 **Expt. non-cytosolic:** Predicted cytosol but experimentally non-cytosolic

FPR cytosol prediction: False positive rate for cytosol prediction

Est. correct predictions: Estimation of correct predictions from total cytosol predictions in Arabidopsis.

FNR cytosol prediction: False negative rate for cytosol prediction

Predicted cytosol: The predicted size of the proteome based on validated performance.

750 **Non-predictable expt. cytosol:** Size of the unpredictable cytosolic proteome

Supplementary Figure 1. Repeat immunological analysis of cytosolic enrichments (AtCyto-2 and AtCyto-3).

Thirty micrograms of Arabidopsis cell culture protoplasts (Protoplast), $10,000 \times g$ crude mixed organelle pellet (10K pellet), $100,000 \times g$ crude mixed organelle pellet (100K pellet) and cytosolic fraction (Cytosol) were separated by SDS-PAGE and transferred onto nitrocellulose membranes. Polyclonal antibodies were used to detect antigens of Arabidopsis cFBPase

(cytosol), H+ATPase (PM-plasma membrane), calreticulin (endomembrane system), histone H3

(nucleus), VDAC-1 (mitochondria) and PsbA (plastid). Second and third experiments are annotated as -2 and -3, respectively.

Supplementary Figure 2. Venn diagram outlining the protein identification overlap and unique identifications by LC-MS/MS between the three cytosolic preparations AtCyto-1, AtCyto-2 and AtCyto-3.

References