Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Low-coverage massively parallel pyrosequencing of cDNAs enables proteomics in non-model species: Comparison of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes

Andrea Bräutigam[a,b], Roshan P. Shrestha[c], Doug Whitten[d], Curtis G. Wilkerson[c,d], Kevin M. Carr[d], John E. Froehlich[e], Andreas P.M. Weber[a,c,*]

[a] *Institut für Biochemie der Pflanzen, Heinrich-Heine-Universität, Universitätsstraße 1, D-40225 Düsseldorf, Germany*
[b] *Graduate Program in Genetics, Michigan State University, East Lansing, MI 48824, USA*
[c] *Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA*
[d] *Research Technology Support Facility, Michigan State University, East Lansing, MI 48824, USA*
[e] *Department of Energy Plant Research Laboratory, Michigan State University, East Lansing, MI 48824 USA*

ARTICLE INFO

Article history:
Received 3 October 2007
Received in revised form 22 January 2008
Accepted 7 February 2008

Keywords:
Pyrosequencing
Massively parallel sequencing
Expressed sequence tags
Transcript profiling
Proteomics
Pisum sativum

ABSTRACT

Proteomics is a valuable tool for establishing and comparing the protein content of defined tissues, cell types, or subcellular structures. Its use in non-model species is currently limited because the identification of peptides critically depends on sequence databases. In this study, we explored the potential of a preliminary cDNA database for the non-model species *Pisum sativum* created by a small number of massively parallel pyrosequencing (MPSS) runs for its use in proteomics and compared it to comprehensive cDNA databases from *Medicago truncatula* and *Arabidopsis thaliana* created by Sanger sequencing. Each database was used to identify proteins from a pea leaf chloroplast envelope preparation. It is shown that the pea database identified more proteins with higher accuracy, although the sequence quality was low and the sequence contigs were short compared to databases from model species. Although the number of identified proteins in non-species-specific databases could potentially be increased by lowering the threshold for successful protein identifications, this strategy markedly increases the number of wrongly identified proteins. The identification rate with non-species-specific databases correlated with spectral abundance but not with the predicted membrane helix content, and strong conservation is necessary but not sufficient for protein identification with a non-species-specific database. It is concluded that massively parallel sequencing of cDNAs substantially increases the power of proteomics in non-model species.

1. Introduction

The first sequenced plant model species, *Arabidopsis thaliana* (thale cress), was chosen not only for its relatively small genome, but also for its small size and rapid life cycle that make it amenable to genetics (Meinke et al., 1998; Somerville and Somerville, 1999; TAGI, 2000). In addition, a large collection of mutants is available, including sequence-indexed insertion mutants that facilitate both forward and reverse genetics approaches (Jander et al., 2002; Parinov and Sundaresan, 2000). However, due to its small size, the presence of a range of secondary metabolites, and the lack of established protocols for isolation of subcellular organelles, *Arabidopsis* does not always represent the ideal model system for, e.g., organellar proteomics. Proteomics is as a valuable tool for establishing the protein complement of cells and subcellular structures (Baginsky and Gruissem, 2006; Baginsky et al., 2004; Dunkley et al., 2006; Heazlewood et al., 2004; Ito et al., 2007; Kleffmann et al., 2006; Lilley and Dupree, 2006; Peltier et al., 2004; von Zychlinski et al., 2005; Ytterberg et al., 2006), especially since the prediction capability of bioinformatics approaches proved insufficient for large scale annotation of organelle proteomes (Jarvis, 2004; Millar et al., 2006; Reyes-Prieto et al., 2007). In addition, multiple targeting of proteins has been documented frequently (Duchene et al., 2005; Millar et al., 2006; Taira et al., 2004) and recently also non-canonical targeting of proteins to, e.g., chloroplast via the secretory system (Miras et al., 2002, 2007; Radhamony and Theg, 2006; Villarejo et al., 2005). In contrast to *Arabidopsis*, the garden pea (*Pisum sativum*) is excellently suited for organelle isolation

* Corresponding author at: Institut für Biochemie der Pflanzen, Heinrich-Heine-Universität, Universitätsstraße 1, D-40225 Düsseldorf, Germany.
Tel.: +49 211 81 12347; fax: +49 211 81 13706.
E-mail address: andreas.weber@uni-duesseldorf.de (A.P.M. Weber).

0168-1656/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2008.02.007

and biochemical studies of enzymes and established protocols for organelle isolation are available in the literature (e.g., Corpas et al., 1999; Miflin and Beevers, 1974; Tobin, 1996). Unfortunately, little is known about the power of proteomics in non-model species for which no extensive sequence database is available. Current peptide identification technology relies on the generation of ideal mass spectra from theoretical libraries. A sequence database is translated in six frames and the resulting protein sequences are *in silico* digested with trypsin. The resulting peptides are used to calculate an ideal mass spectrum. If an observed spectrum matches a theoretically predicted spectrum with a certain probability the corresponding peptide is called "identified". This method of identification demands a perfect sequence match between the sample peptide and the database peptide, although some programs, such as implementations of the X!-Tandem software (Craig and Beavis, 2004), allow the inclusion of single amino acid mismatches. Allowing more than one mismatch increases the error rate and the time required for the search. With increasing evolutionary distance, perfect matches become less likely even between highly conserved proteins, in particular since conservative changes, such as aspartate to glutamate will already cause a spectral mismatch. In contrast, low quality databases, such as the one discussed in this communication, limit the identification of peptides either by sequencing and assembling errors causing amino acid changes in predicted peptides or by not providing enough peptide coverage for correct identifications due to short contigs. *De novo* sequencing of peptides is considered too slow and limited by computing time for high throughput applications (Baginsky and Gruissem, 2006; Pevtsov et al., 2006). Currently, the identification of proteins from non-model species with limited sequence coverage frequently relies on databases generated from closely related species (Schmidt et al., 2007) or indeed all sequences that are available in public databases (Taylor et al., 2005) although this method will especially limit the identification of less conserved proteins.

It is has been recently proposed to use massively parallel pyrosequencing to fully explore the potential of proteomics in non-model species such as pea (Weber et al., 2007). In this study, we systematically assessed the potential and limitations of massively parallel pyrosequencing to support proteomics applications. To this end, we compared proteomics based on a low-coverage transcriptome sequence database of the garden pea consisting of many short sequence contigs with frequent frameshift errors with a conventionally created and fairly comprehensive cDNA database of a closely related model species (*Medicago truncatula*), and with a high-quality, virtually error-free database generated from a completely sequenced model species (*Arabidopsis thaliana*). We established the limitations of each database and we tested how the degree of conservation, the abundance of mass spectra generated from a particular protein, and the number of transmembrane domains influence the odds for successful protein identification using a non-species-specific database. Finally we discuss the consequences of interpreting the proteomics sample based on the different database results.

2. Material and methods

2.1. Massively parallel pyrosequencing and generation of sequence databases

Three different databases were generated for proteome analyses. For the generation of the pea transcriptome database, one non-normalized and several normalized libraries were generated and sequenced using massively parallel pyrosequencing technology (Margulies et al., 2005). The preparation of cDNA libraries was conducted as described previously (Weber et al., 2007), with the exception that some libraries were normalized to decrease the proportion of highly abundant transcripts. To this end, 1 μg of double-stranded cDNA was normalized using a commercial kit (Trimmer-kit, Evrogen, Moscow, Russia) that is based on Kamchatka crab duplex-specific nuclease (Zhulidov et al., 2004). Following normalization, the cDNA populations were amplified by PCR and 3 μg of the resulting PCR-amplified cDNA was used per sequencing reaction. One cDNA library was generated from leaves and one from hypocotyl. Five preparations generated from the leaf library (four normalized, one non-normalized) and one preparation from the hypocotyl library (normalized) were sequenced using a Roche/454 GS20 instrument. This technology delivered an average read length of 100 nts. Two additional libraries from etiolated and de-etiolated (4 h light) leaves were sequenced on a half plate each, using a GS FLX instrument, which allowed for average read lengths of 250 nts. The preliminary pea database used in the reported work was assembled from approximately 2 million pyrosequencing reads with a total of about 230 million nucleotides. The pyrosequencing reads were combined with publicly available pea cDNA sequences from Genbank and the IPK Gatersleben EST database. All sequences, except the full-length and partial mRNA sequences from GenBank, were subjected to QA using the SeqClean EST trimming and validation tool cDNAs and ESTs were clustered and assembled using the TGI Clustering Tools (TGICL) in a multi stage pipeline. The sequence reads were loosely clustered with a modified version of megablast (Zhang et al., 2000) and subsequently the clusters were submitted to the CAP3 sequence assembling program (Huang and Madan, 1999). All programs are available at http://compbio.dfci.harvard.edu/tgi/software. 1,570,251 reads (80% of the total) were assembled into 135,250 contigs. For the proteome analysis, only contigs longer than 100 basepairs were used. The pea cDNA database is currently undergoing further development and a detailed description of the pea cDNA sequence database and its assembly will be reported in an upcoming manuscript (Shrestha et al., in preparation). The Medicago database was based on the *M. truncatula* gene index of tentative consensus sequences assembled by TIGR, currently maintained at the Dana-Faber Cancer Institute and Harvard Medical School of Health (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago). For the Arabidopsis database the latest Arabidopsis proteome annotation from the TAIR7 (www.arabidopsis.org) genome release was used. All pea cDNA contigs matching one or multiple mass spectra are provided as supplementary material to this article.

2.2. Processing of proteomics samples

As a proteome sample, chloroplast envelope membranes were isolated from 10 to 14 days old pea plants as described previously (Douce and Joyard, 1979; Keegstra and Yousif, 1986). Envelope membrane samples (approx. 100 μg of protein) were mixed with SDS–PAGE loading buffer, incubated for 20 min on a reaction tube shaker at 15 ° C, and subsequently separated by 12.5% SDS-PAGE. After staining with Coomassie Brilliant Blue, each gel lane was cut into ten equally sized slices. Proteins contained in the gel slices were subjected to tryptic cleavage as described by Shevchenko et al. (1996). Extracted peptides were automatically loaded onto a Waters Symmetry C18 peptide trap (5 μm, 180 μm × 20 mm) at a flow rate of 4 μL/min in 2% acetonitrile/0.1% formic acid for 5 min by a Waters nanoAcquity Sample Manager. The peptides were eluted onto a Waters BEH C18 nanoAcquity column (1.7 μm, 100 μm × 100 mm) and eluted over 90 min using a Waters nanoAcquity UPLC into a ThermoElectron LTQ-FTICR mass spectrometer with a flow rate of 300 nL/min (Buffer *A* = 99.9% water/0.1% formic acid, Buffer *B* = 99.9% acetonitrile/0.1% formic acid: gradient of 5%

B to 40% B from 0 to 63 min, 40% B to 90% B from 63 to 71 min and 5% B from 71 to 90 min). The top ten ions of each survey scan, which were taken at a resolution of 50,000, were subjected to automated low energy collision induced dissociation. The resulting MS/MS spectra were converted to a peak list using BioWorks Browser v 3.2. This empirical mass spectra library of 27,443 queries was compared to three databases with sequences from *Arabidopsis thaliana*, *Medicago truncatula* and *Pisum sativum*, respectively, using the Mascot searching algorithm, v 2.2 (www.matrixscience.com). Carbamidomethyl cysteine was set as a fixed peptide modification and oxidation of methionine was allowed. Up to two missed tryptic sites were allowed. The peptide tolerance was set to ±10 ppm and the MS/MS tolerance to 0.8 kDa. All proteomics data has been submitted to the PRIDE database (http://www.ebi.ac.uk/pride/) and is accessible under the accession number 3227.

2.3. Analysis of proteomics data

The Mascot output was loaded into the program Scaffold (www.proteomesoftware.com) which calculates protein identification probabilities based on PeptideProphet and ProteinProphet (Keller et al., 2002; Nesvizhskii et al., 2003). The protein identification threshold was set to 99% probability for all databases and at least two unique peptides for each protein ("stringent criteria") or to 99% probability for at least one database with lesser probabilities allowed for the other databases on the same protein and one unique peptide for each protein ("relaxed criteria"). For further analysis the data was exported to MS Excel. For each protein identified, the sequences in both the pea database as well as the Medicago database were mapped to the Arabidopsis proteome using BlastX (Altschul et al., 1997) and the Arabidopsis gene identifier (AGI) of the closest homologue was noted. Based on its closest Arabidopsis homologue, the protein was annotated and the predicted number of membrane spanning helices was retrieved from TAIR. In cases where the AGIs obtained from the three different identification strategies did not match, the identifications were manually inspected and priority was given to the identification with the highest probability score. For identical probability scores the highest number of unique peptides mapped to a protein was declared the correct identification. When multiple sequences were matched to the same spectra, identifications were either called "mistaken" if none of the matches were identical to the protein identified correctly in the other databases or "multiple" if one of the proteins was identical to the correct match determined as outlined earlier. The complete list of proteins was collapsed into a non-redundant list based on the AGIs with sequences yielding the same hit in the BlastX hit being summed as the same protein.

3. Results

3.1. Properties of the cDNA sequence databases

The characteristics of a low-coverage pea cDNA database generated by limited pyrosequencing were compared to two databases generated with conventional sequencing technology. The pea cDNA sequence database contained more than 31,000 relatively short contigs. About 29,000 contigs were between 300 and 1000 nts in length, about twice as many in this length category as in the Medicago (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago) and an Arabidopsis transcriptome databases (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=arab) (Fig. 1). In contrast, contigs between 1 and 2 kb were massively underrepresented (1697 contigs) in the pea database in comparison to the Medicago database (6224 long contigs) and the Arabidopsis database (13,033 long contigs) and contigs longer than 2000 nts were almost absent from the pea database (Fig. 1). In addition to having shorter contigs, many pea contigs, when translated in all six frames, apparently contained frameshifts that are likely due to assembly or base calling errors (see Supplementary File of pea contigs).

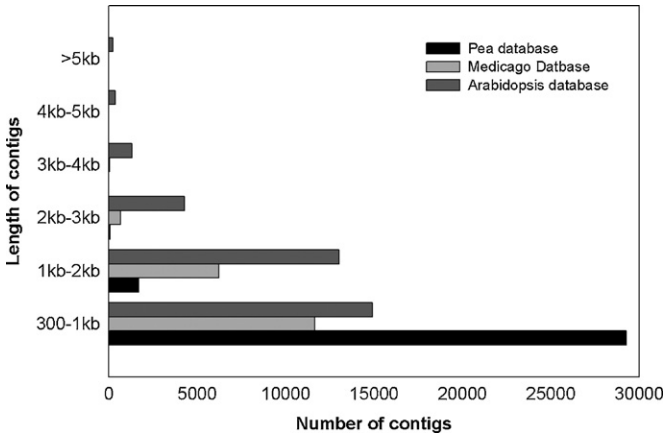

Fig. 1. The length distribution of contigs (assembled transcripts) in databases used for proteome analyses.

The Medicago database consisted of the tentative consensus sequences (TCs) assembled from EST projects from *M. truncatula*. *Medicago* is being developed as a model legume species and is hypothesized to serve as a research template for the garden pea. The Medicago TCs were comprised of about 36,000 unique sequences. About one half of these sequences were tentative consensus sequences that represent contigs consisting of several ESTs and the other half were singletons. The length distribution of the Medicago TCs was more similar to that of the Arabidopsis transcriptome database and the sequence quality judged from open reading frame annotation software was high (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago). While both the Arabidopsis and Medicago databases contain similar numbers of cDNA sequences that are shorter than 1 kb, the Medicago database contains only half as many sequences between 1 and 2 kb, and only one-tenth for transcripts longer than 2 kb (Fig. 1). The Arabidopsis proteome derived from the TAIR7 build of the completely sequenced genome served as an example for a completely sequenced genome. The capability of the new pea database to reliably identify proteins despite its obvious shortcomings was compared with the capabilities of the databases of *M. truncatula* and *A. thaliana*.

3.2. Performance of the databases in proteomics

A pea chloroplast envelope membrane proteome sample was analyzed using the sequence databases described above. To identify the most advantageous database and program parameters for protein identification, we employed a factorial approach. An empirical library of uninterpreted mass spectra was generated from a single proteomics experiment of pea chloroplast envelope membranes. This particular sample was chosen because it represents a relatively minor share of the total cellular proteome and because it contains a large number of highly hydrophobic membrane proteins and thus represents a challenging target for protein identification. In addition, several proteomic studies of chloroplast envelope membranes from a range of plant species have been published, thus providing good references for comparison. All three databases described above were matched to the empirical spectra library using either stringent or relaxed criteria. For both criteria, the largest number of proteins could be identified with the pea database, followed

Table 1
Summary of protein identifications

Database	Stringent criteria			Relaxed criteria		
	Number of identifications	Total number of spectra	Avg. number of spectra per ID	Number of identifications	Total number of spectra	Avg. number of spectra per ID
Combined	255	8222	32	283	8892	31
Pisum sativum	221	5012	23	255	5139	20
Medicago truncatula	125	1977	16	203	2198	11
Arabidopsis thaliana	82	1233	15	165	1555	9

by *Medicago* and *Arabidopsis*. Applying stringent criteria, a total of 8222 spectra were matched to 255 non-redundant proteins using a combination of all three databases. A list of all identifications can be found in Supplementary Table 1 and a non-redundant list of identified proteins in Supplementary Table 2. Under stringent conditions the pea database allowed matches of 5012 spectra against 221 non-redundant proteins (86% of the total). On average each protein was identified by 23 spectra with up to 362 spectra per protein (Table 1). The Medicago database yielded 1977 matched spectra on 125 proteins (49% of the total). With only 32% or 82 proteins the Arabidopsis database allowed the fewest matches. With relaxed criteria the total number of proteins identified using a combination of all three databases increased to 283 and 8892 spectra were matched. In the pea database, the relaxed criteria allowed for 5139 matching spectra, but a large number of proteins that were identified with only a few spectra caused the average number of spectra per protein to drop to 20. The Medicago TC database yielded 2198 spectra mapping to 203 proteins, almost 72% of the total. The *A. thaliana* database yielded the lowest number of identified proteins, with only 165 proteins or 58% of the total (Table 1).

There were 36 proteins which could not be identified with the pea database but yielded significant identifications with at least one of the other databases. Eight of these proteins are encoded on an organelle genome (Supplemental Table 2). Six proteins were very similar to closely related proteins, which were present in the pea database and were identified. For nine of these proteins corresponding sequences were absent from the pea database. The remaining 12 proteins remained unidentified with the pea database although corresponding sequences were contained in the database, as indicated by Blast searches against the database.

3.3. Abundance of mass spectra is positively correlated with correct protein identification in non-species-specific databases

With both non-species-specific databases 51% or 68% of the proteins could not be identified when stringent criteria were applied. We next investigated whether there are protein characteristics that might serve as predictors of identification probability in the non-species-specific databases. Although the correlation between the spectral count and the abundance of a protein in not absolute, the spectral count has been used successfully to estimate protein abundance (Liu et al., 2004; Lu et al., 2007; Zybailov et al., 2005). The proteins in the pea database were arranged into five groups based on the abundance of their matching mass spectra. More than half of the protein identifications using the pea database were based on two to ten spectra, whereas only 3.6% of the proteins were identified with more than 100 spectra each. For each abundance class, the percentage of proteins that could also be identified with the Medicago and the Arabidopsis databases were plotted (Fig. 2). The abundance of spectra as determined by the spectral count obtained with the pea database correlated with the identification rate with non-species-specific databases ($R^2 > 0.95$). With the Medicago database almost all of the proteins represented by more than 100 mass spectra could be identified whereas only 38% of the proteins with a spectral count below 10 were identified. With the Arabidopsis database about two thirds of the proteins with a high spectral count were identified and the identification rate dropped to 17% for proteins that had spectral counts of less than 10 (Fig. 2).

3.4. Protein hydrophobicity is a poor predictor for correct protein identification

We also tested whether the presence of transmembrane helices adversely affected the probability that a protein was identified using a non-species-specific database as has been proposed by (Eichacker et al., 2004). Proteins that were identified using the pea database were grouped according to the number of predicted membrane spanning helices (none, one, two, three, four to six, more than seven) and the presence of beta sheets. More than half of the proteins identified with the pea database were not predicted to contain transmembrane helices and the other groups contained 2.3–16.8% of the proteins. For each group the percentage of proteins that could also be identified with the non-species-specific databases was plotted (Fig. 3). Unlike protein abundance as deduced from spectral counts, the number of predicted transmembrane helices is of less predictive value for identification rate since no correlation was observed between the number of transmembrane domains and the probability for identification using a non-species-specific database. However, among the proteins with more than 7 predicted membrane spanning helices only 1 (Medicago database) or none (Arabidopsis database) was identified out of 10 proteins identified with the pea database.

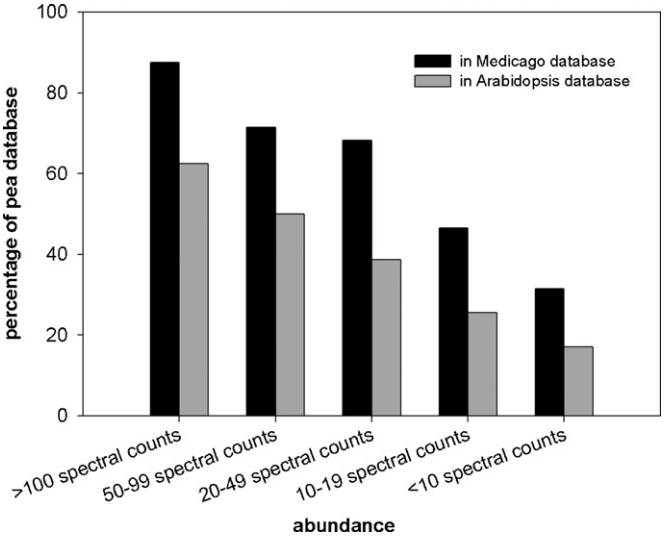

Fig. 2. The number of spectral counts detected for a protein is correlated with the rate of identification. Proteins in the pea database were grouped by spectral abundance and the percentage of proteins also identified with the non-species-specific databases was plotted. The following absolute numbers equal 100%: >100 spectral counts eight proteins, 50–99 spectral counts 14 proteins, 20–49 spectral counts 44 proteins, 10–19 spectral counts 43 proteins and <10 spectral counts 111 proteins.

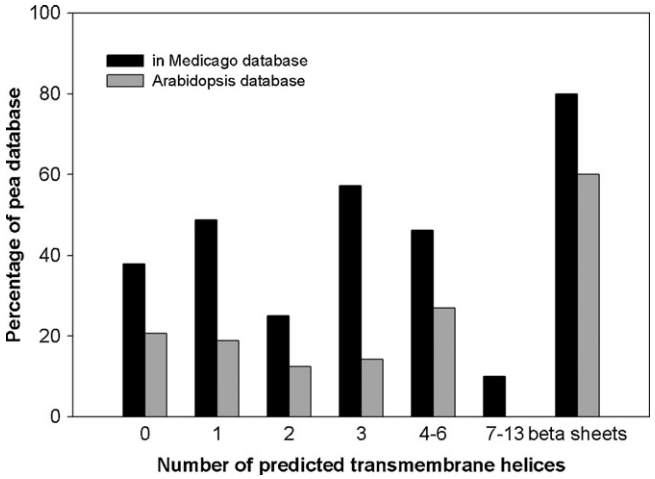

Fig. 3. The number of predicted membrane spanning helices for a protein is not correlated with the rate of identification. Proteins in the pea database were grouped by predicted membrane spanning helices and the percentage of proteins also identified with the non-species-specific databases was plotted. The following absolute numbers equal 100%: no predicted transmembrane helices 111 proteins, 1 predicted transmembrane helix 37 proteins, 2 predicted transmembrane helices 24 proteins, 3 predicted transmembrane helices 7 proteins, 4–6 predicted transmembrane helices 26 proteins, 7–13 predicted transmembrane helices 10 proteins, predicted beta sheets 5 proteins.

3.5. Analysis of abundant non-identified and of wrongly identified proteins

Although proteins with more matching spectra in the pea database had a higher probability to be also identified with the non-species-specific databases, many of the 15 proteins with the highest number of spectra could only be identified with the species-specific database (Table 2). The Medicago database did not allow the identification of the small subunit of RubisCO and a chaperonin, and a hydroperoxide lyase and the triosephosphate phosphate translocator (TPT) did not pass the stringent threshold for identification probability ($p < 0.01$). With the Arabidopsis database both the hydroperoxide lyase and the small subunit of RubisCO remained unidentified whereas the TPT, two components of the protein import complex, Toc159 and Toc34, a carbonic anhydrase and an unknown protein did not pass the probability threshold (Table 2). The sequences of TPT, which was not identified with the Medicago and the Arabidopsis databases, and of malate dehydrogenase (NAD-MDH), which was identified with all three databases, were aligned (Fig. 4) and the tryptic fragments were determined. PsTPT is highly similar to both AtTPT (77.5% identity and 14.5% similarity) and MtTPT (92.8% identity and 14.5% similarity). The same is true for PsNAD-MDH and AtNAD-MDH (73% identity and 18.6% similarity) and PsNAD-MDH and MtNAD-MDH (91.7% identity and 3.9% similarity). Despite the high degree of sequence identity AtTPT shares only one and MtTPT only six tryptic peptides with the pea protein. The TPT peptides from the empirical spectra library were mapped onto the aligned sequences. All but one of the TPT peptides that generated mass spectra contained at least one amino acid exchange with respect to *Medicago* or *Arabidopsis* (Fig. 4). In contrast, 16 peptides of the Medicago NAD-MDH exactly matched the pea peptides. For 13 of those 16 theoretically predicted peptides mass spectra were detected experimentally, thus permitting reliable identification of the protein using the Medicago database. Only four tryptic peptides of the Arabidopsis NAD-MDH matched the pea protein sequence, but since for two of them mass spectra were experimentally detected the protein could still be reliably identified (Fig. 4).

When the successful protein identifications were compared between the three databases, several non-matching identifications were revealed (Fig. 5). Manual inspection of the peptide sequences, the peptide probability scores, and the protein coverage indicated that one of the identifications is likely correct. With stringent criteria none of the identifications with the pea database appeared to be invalid compared to the non-species-specific databases. Using relaxed criteria, two proteins were erroneously identified. Peptides for the large subunit of RubisCO were mistakenly annotated as part of a RubisCO-like protein of unknown function encoded on the mitochondrial genome. A second misidentification is of the root glutamate synthase (GLU2) instead of the leaf glutamate synthase (GLU1). Using the Medicago database, 24 proteins were erroneously identified with relaxed criteria and in three cases the peptide with matching spectra is part of more than one protein and therefore yields multiple identifications of proteins only one of which is correct. Sixteen of the misidentified proteins are closely related to the correct protein and eight identifications corresponded to completely different protein. Unlike with the pea database, even

Table 2
Databases of a different species yield weaker IDs even with abundant proteins; ID prob (identification probability calculated with Proteinprophet as implemented in Scaffold; AGI corresponding Arabidopsis protein)

Annotation	*Pisum sativum*			*Arabidopsis thaliana*			*Medicago truncatula*		
	Number of spectra	ID prob. (%)	Pea contig	Number of spectra	ID prob. (%)	AGI	Number of spectra	ID prob. (%)	Tentative consensus sequence
Tic110	362	100	CL546Contig1	11	100	multiple	24	99	TC103049
Tic55	185	100	CL4007Contig1	44	100	AT2G24820	102	100	TC101323
TPT	183	100	CL63Contig5	29	92	AT5G11210, AT5G46110	18	94	TC100560
Toc75	177	100	CL979Contig1	23	100	AT3G46740	42	100	TC99006
ClpC protease	146	100	CL20018Contig1, CL699Contig1	115	100	AT3G48870, AT5G50920	110	100	TC94369, TC94370
Toc159	114	100	CL600Contig1	11	92	AT4G02510	40	100	TC112083, TC96006
Unknown	112	100	CL1491Contig2	4	90	AT4G17840	69	100	TC95492
Carbonic anhydrase	78	100	SCL17Contig25, SCL63Contig2	9	96	AT3G01500	27	100	TC94246
RubisCO activase	71	100	SCL4Contig186	18	100	AT1G73110	39	100	TC94141
RubisCO small subunit	67	100	SCL2Contig158	No identification			No identification		
Chaperonin	61	100	SCL57Contig13	46	100	AT3G13470, AT5G56500	No identification		
NAD-MDH	61	100	CL460Contig4	12	100	AT3G47520	40	100	TC107189
Toc34	57	100	CL969Contig1	4	92	AT1G02280	18	100	TC94636, TC94638
Hydroxyperoxide lyase	52	100	CL3726Contig2	No identification			4	85	TC104771
ABC transporter	51	100	CL6175Contig1	12	100	AT4G25450	17	100	BG587938, TC103874

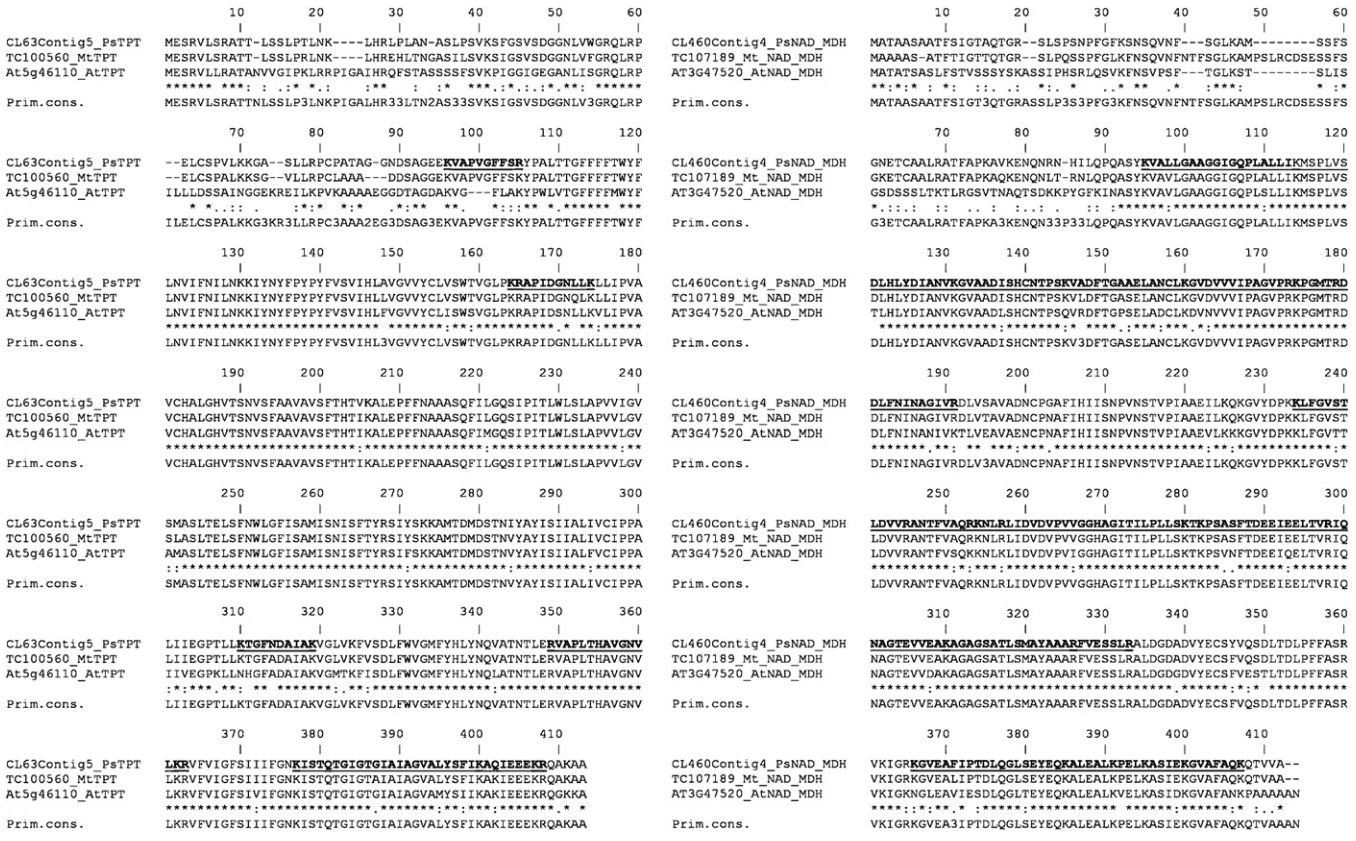

Fig. 4. The alignment of protein sequences for the TPT (identified only with the pea database) and MDH (identified in all databases) indicates that a high degree of conservation is necessary but not sufficient for successful identification with a non-species-specific database; peptides identified by the pea database are bold and underlined; predicted cut sites for trypsin are shaded in grey in the pea sequence; theoretical cut sites for the *Arabidopsis* and *Medicago* proteins are in conserved positions except for one amino acid substitution which destroys the cut site in AtMDH.

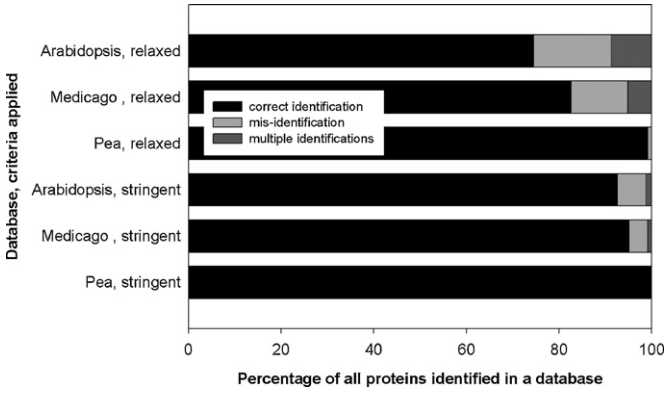

Fig. 5. The non-species-specific databases yield a significant amount of mis-identifications and multiple identifications.

under stringent conditions, 15 misidentifications persisted. With the Arabidopsis database 29 proteins were mistakenly identified when relaxed criteria were used and eight multiple identifications existed. More stringent criteria still resulted in five misidentifications and in one peptide, which identified multiple proteins, but all persisting misidentifications identify closely related proteins (for details see Supplementary Table 2).

4. Discussion

The results presented in this paper indicate that the prospects for identifying proteins from a species with limited sequence resources by proteomics can be massively increased by generating a species-specific transcriptome database by MPSS, even if the resulting database is of low quality, compared to sequence databases generated by conventional sequencing. When non-species-specific databases are used, the odds for protein discovery are limited, and the probability to identify a protein can be predicted by its abundance but not by its content of membrane spanning helices. Strong sequence conservation is necessary but not sufficient to identify a protein with a non-species-specific database. Especially when the identification criteria are relaxed to allow imperfect matches and therefore more protein identifications with non-species-specific databases, the identifications are more prone to erroneous identifications.

Our initial expectation was that both short contig length and the relatively high rate of sequencing errors that are characteristic to low-coverage MPSS projects would severely limit the prospects for successful protein identification. Unexpectedly, however, the pea database developed in this study was superior to the tested non-species-specific databases with regard to the rate of protein discovery and the quality of identifications despite being of low quality and low coverage, as compared to conventional databases. Not only more proteins were identified but also the average number of spectra mapping to each protein was higher (Table 1). The advantage of species specificity clearly outweighs the quality issues of the database. Only 10–14% of the total proteins identified remained unidentified with this database. Detailed analysis of these non-identified proteins revealed that organelle-encoded proteins are overrepresented (22%) among those not identified with the pea database, as compared to the overall proportion of identified proteins that are encoded in organelles (4%) (Supplemental Table 2).

Since the mRNA-isolation protocol used in this study included two consecutive rounds of poly-A+ purification and because reverse translation of mRNA into first strand cDNA was primed by oligo-dT, it is likely that transcripts from organellar genomes are underrepresented in the sample, as reported previously (Weber et al., 2007). For nuclear encoded proteins, sequences could not be identified for nine proteins in the pea transcriptome database and for seven additional proteins the corresponding nucleotide sequences were either short or fragmented into multiple unassembled short contigs, thus demonstrating how an unfinished MPSS-generated sequence database limits proteomic identification technology if the sequence contigs are too short. Incidentally, most of the contigs in the pea database are shorter than 1 kb with the majority being shorter than 400 nts (Fig. 1). This translates into a stretch of approximately 140 amino acids, assuming that no 5′ or 3′-UTRs are contained in the sequence. Depending on how the tryptic fragments map onto the short sequence, the identification probability can probably be too low to identify proteins reliably. This problem is currently being addressed by generating additional sequence coverage and new assembling methods for the short and midrange sequence reads obtained by the MPPS technology. The recently released GS FLX instrument produces sequence reads that are 2.5-times longer as the sequence reads of the GS 20. In addition, preliminary tests showed that including a second, less stringent clustering step in the assembly pipeline might serve to produce longer contigs that are more suitable for proteomics applications although this may aggravate the problem of frameshift errors during assembly. In conclusion, the analysis of the proteome sample with several databases served to determine the quality of the sequence database for proteomics. Since it was less likely for proteins to be identified with non-species-specific databases the presence of such identifications indicated that the species-specific database could still be improved.

Considering the multitude of short contigs (Fig. 1) and the presence of sequencing errors that resulted in frameshifts (Supplementary Data File) it was surprising that, with both databases from model species, fewer proteins were identified than with the novel pea database (Table 1). The Arabidopsis database represents a completely sequenced genome that is well annotated and presumably complete. Although the evolutionary split between the genome of *P. sativum* and *A. thaliana* occurred about 100 million years ago (Wikström et al., 2001), we hypothesized that the degree of conservation might be sufficient to identify most proteins, albeit with a lower peptide count, especially given that the database is complete. Surprisingly, the Arabidopsis database only allowed the identification of 32% of the proteins relative to all databases combined. Notably, in a previous chloroplast envelope proteomics study that used spinach chloroplasts and mostly non-species-specific sequences for identification only 50 proteins (25% compared to this study and 15% compared to (Froehlich et al., 2003)) were identified in total, amongst them 21 membrane proteins with more than 4 membrane helices (Ferro et al., 2002). The phosphate translocators that were identified in this previous study were limited to those for which species-specific sequences were available in public databases at that time. In addition to the Arabidopsis database a more closely related model species database was tested for its performance in a proteomics application. *M. truncatula* and *P. sativum* are close relatives in the subfamily Papilionoideae which separated about 25 million years ago (Lavin et al., 2005), which is equivalent to the evolutionary distance between *Arabidopsis* and the *Brassica* species (Yang et al., 1999), and both microsynteny and macrosynteny between the genomes have been demonstrated (Gualtieri et al., 2002; Kalo et al., 2004). *Medicago* is being developed as the model legume species, also serving as a research template for pea. Since the evolutionary distance between *Medicago* and pea is smaller than between *Arabidopsis* and pea, we hypothesized that the Medicago database would allow for the identification of more proteins. Indeed, the Medicago database allowed the detection of half of the proteins under stringent conditions, although the average number of spectra detected for each protein was only 50% of those obtained with the pea database. Although the number of sequence contigs longer than 1 kb is significantly lower in the Medicago database than in the Arabidopsis database (Fig. 1), indicating that a large portion of the available sequences do not represent full length cDNAs, the Medicago database is more successful than the Arabidopsis database in identifying chloroplast envelope proteins (compare to Table 1). Based on the quality of the databases and the number of proteins that could be identified, we conclude that evolutionary distance imposes a higher penalty for protein identification than does the quality of the sequence database. This underscores the inability of current peptide identification software for proteomics applications to tolerate amino acid mismatches between the theoretically and experimentally generated spectra especially when stringent identification criteria are applied.

We hypothesized that relaxed identification criteria may allow imperfect matches for peptides and might thus enhance the identification rate for both the Medicago and the Arabidopsis database. Relative to the total number of proteins identified, 58% of proteins could be identified with relaxed criteria compared to 32% with stringent criteria with the Arabidopsis database and 72% compared to 49% with the Medicago database. Unfortunately, the relaxed identification criteria also lead to a high number of mistakenly identified proteins and the identification of multiple family members that shared a single conserved peptide. In conclusion, although allowing protein identifications with a single peptide match and low scoring peptides could increase the coverage, the high number of mistaken and multiple identifications make this strategy inadvisable for increasing the number of protein identifications.

The experiment clearly established the value of a species-specific database, even if it was of low quality (Fig. 1), for proteomics identification technologies. Since many projects will have to rely on sequence databases of related species, we tested several parameters to determine what limits the identification. Theoretically, for any successful identification under stringent criteria, a protein must yield at the least two fragmented peptides whose spectra match theoretical spectra from the library. Frequently a higher number of unique peptides are identified for more abundant proteins and the likelihood, that a least two completely conserved peptides are present among them, increases. Accordingly, the abundance of spectra and the identification rate were indeed closely correlated (Fig. 2). The correlation was not absolute though; we found that extremely abundant proteins, such as the TPT, were only identified with the pea database (Table 2) whereas about one third (Medicago database) to one-fifth (Arabidopsis database) of low abundance proteins could still be identified with a non-species-specific database (Fig. 2). Conserved structural features of proteins may also limit identification. Proteins with comparatively high numbers of membrane spanning helices and underrepresented recognition sites for trypsin are harder to identify than soluble proteins since there are frequently less spectra available for matching to the respective database (Eichacker et al., 2004). When the proteins identified in this study were grouped according to their predicted membrane helix content and the ratio of proteins identified with the non-species-specific databases relative to the pea database was plotted, however, no correlation was observed (Fig. 3). Possibly, the sample size with two databases and a limited number of membrane proteins was too small to reveal a correlation. Since highly hydrophobic proteins with more than seven transmembrane helices, such as the TPT (Weber et al., 2005), could not be identified with the non-species-specific databases, any analysis of a mem-

brane proteome will likely critically depend on the availability of a species-specific sequence database.

Finally, we studied several of the abundant proteins that were identified in all three databases and the proteins that remained unidentified in the Medicago and the Arabidopsis database. As an example for an abundant membrane protein, TPT, which remained unidentified in the non-species-specific databases, was compared with the soluble protein NAD-MDH that was identified with all three databases. Although the TPT orthologues are slightly more conserved than the NAD-MDH orthologues, the mass spectra generated from TPT could not be mapped onto the sequences provided by the Medicago and Arabidopsis databases. For *Arabidopsis*, only a single tryptic peptide is completely conserved between the corresponding pea and *Arabidopsis* proteins; hence the protein could not be identified with confidence. Although the Medicago sequence shares six tryptic peptides with the pea sequence, only one of these peptides was matched to the empirical spectra library, which is not sufficient for a significant identification with stringent criteria. Two of the conserved peptides are longer than 23 amino acids and two are very short, which might be the reason why spectra corresponding to these peptides were not experimentally detected. NAD-MDH could be identified with all databases since matching spectra were generated from the protein. A high degree of conservation is necessary but not sufficient for a successful identification of a protein with a non-species-specific database since a single amino acid exchange will preclude a peptide match.

The reduced information caused by using a non-species-specific database hampers the interpretation of the data. To visualize this fact, proteins identified with each database were drawn in identical positions onto a schematic representation of a chloroplast. The likely mitochondrial and endomembrane system contaminants were also drawn on the corresponding structures (Fig. 6). As expected from the reduced number of proteins identified with the Arabidopsis and Medicago databases, the total number of membrane proteins and contaminating proteins is also lower. The Arabidopsis database identified no protein likely to reside in the endomembrane system. Hence one would conclude that there is no contamination from this source, although both the Medicago and the pea database reveal that there is at least one protein present that is believed to reside in the endomembrane system. In the extreme case of the Arabidopsis database used to analyze the pea chloroplast envelope proteome, none of the transport proteins catalyzing the major metabolite fluxes across the envelope (see Weber, 2004; Weber and Fischer, 2007, for recent reviews) could be identified. With the pea database all transporters for phosphorylated sugars that are predicted to reside in the inner envelope of chloroplasts were identified: the triosephosphate, the phosphoenolpyruvate, and the pentosephosphate/phosphate translocators (Flügge et al., 2003; Weber et al., 2004) whereas only the pentose phosphate translocator could be identified with the Medicago database. The two translocator system for importing 2-oxoglutarate and exporting glutamate could also be identified only with the pea database (Renné et al., 2003; Reumann and Weber, 2006; Schneidereit et al., 2006; Weber and Flugge, 2002; Weber et al., 1995). A plastidic ADP/ATP translocator (Möhlmann et al., 1998; Neuhaus et al., 1997; Reiser et al., 2004) could be identified with all three databases as well as several members of the mitochondrial carrier family that have previously been shown to be targeted to chloro-

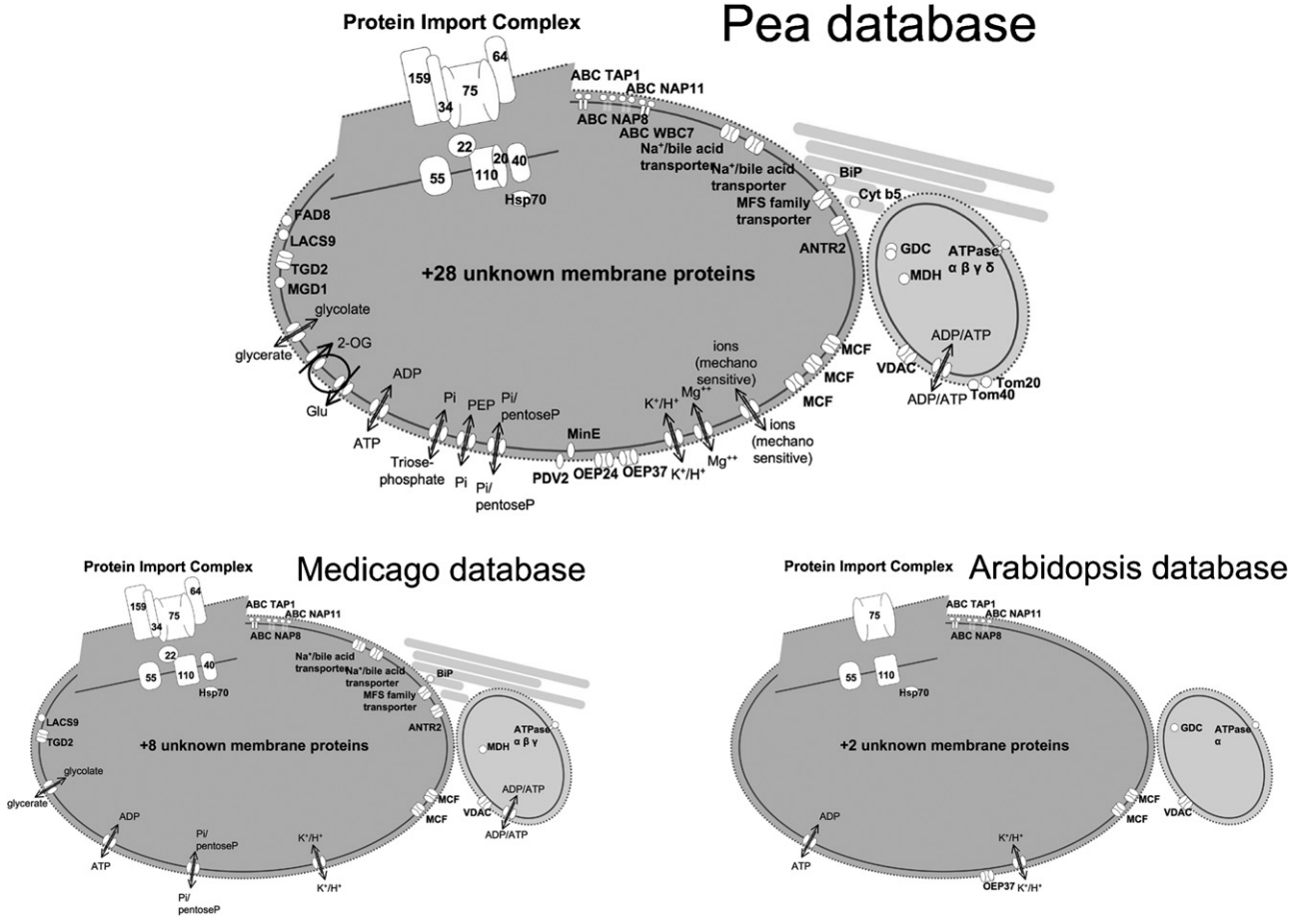

Fig. 6. Non-species-specific databases limit the discovery of new proteins and the interpretation of known proteins in the pea chloroplast envelope proteome sample.

plasts (Bedhomme et al., 2005; Bouvier et al., 2006; Picault et al., 2004). At least one member of the potassium proton exchanger family (Maser et al., 2001) was identified with all databases but a magnesium transporter (Li et al., 2001) and a putative mechanosensitive channel (Haswell and Meyerowitz, 2006) were only seen in the pea database. Two ABC transporters, PAA1 (Shikanai et al., 2003) and HMA1 (Seigneurin-Berny et al., 2006), which import copper and possibly other metal ions into the chloroplast, could be identified with the pea database but not with either of the non-species-specific databases. Three other ABC type transporters were sufficiently conserved for identification with all three databases. For the import apparatus, all canonical components that are known to date (Gutensohn et al., 2006) were identified with the pea database except for Tic21 (Teng et al., 2006), which was recently assigned the function of an iron transporter (Duy et al., 2007). With the Medicago database the majority of import complex components could be identified but with the Arabidopsis database only four components were identified. Of the plastid division machinery, PDV2 and MinE were found with the pea database (Glynn et al., 2007; Miyagishima et al., 2006), whereas another component, Arc 6 that has been identified previously (Froehlich et al., 2003), was not identified, but was also only present as a highly fragmented sequence in the pea database. Several likely membrane associated proteins involved in fatty acid and membrane lipid metabolism were identified. For fatty acid synthesis and modification, acetyl co-enzyme A carboxylase, a fatty acid desaturase (McConn et al., 1994) and a long chain fatty acid coenzyme A ligase (Schnurr et al., 2002) were identified as well as proteins involved in lipid metabolism such as 1,2-diacylglycerol 3-beta-galactosyltransferase for the synthesis of galacto- and UDP-sulfoquinovose:DAG sulfoquinovosyltransferase for synthesis of sulfolipids (SQD2) (Jarvis et al., 2000; Yu et al., 2002). Most proteins were identified with the pea database but SQD2 was only identified with the non-species-specific databases. SQD2 was likely not identified because its sequence is fragmented into several short contigs in the pea database (for a complete list of identified proteins and the capabilities of each database please refer to Supplementary Table 1). Based on the analysis we conclude that any proteome analysis relying on protein identifications based on a non-species-specific database is limited in its conclusions about the presence of proteins and that generating a species-specific database even if it is of low quality can massively enhance protein discovery.

Acknowledgements

We thank Shari Tjugum-Holland and Jeff Landgraff of the Michigan State University Research Technology Support Facility for assistance with RNA and DNA analysis and DNA sequencing. This work was supported by a Strategic Partnership Grant (Next Generation Sequencing Center) of the Michigan State University Foundation (to A.P.M.W), NSF-grants IOB-0548610 (to A.P.M.W) and MCB-0519740 (to A.P.M.W), and by an Arabidopsis Functional Genomics Network (WE 2231/4-1) award of the Deutsche Forschungsgemeinschaft (to A.P.M.W).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbiotec.2008.02.007.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Baginsky, S., Gruissem, W., 2006. *Arabidopsis thaliana* proteomics: from proteome to genome. J. Exp. Bot. 57, 1485–1491.

Baginsky, S., Siddique, A., Gruissem, W., 2004. Proteome Analysis of Tobacco Bright Yellow-2 (BY-2) Cell Culture Plastids as a Model for Undifferentiated Heterotrophic Plastids. J. Proteome Res. 3, 1128–1137.

Bedhomme, M., Hoffmann, M., McCarthy, E.A., Gambonnet, B., Moran, R.G., Rebeille, F., Ravanel, S., 2005. Folate metabolism in plants – an Arabidopsis homolog of the mammalian mitochondrial folate transporter mediates folate import into chloroplasts. J. Biol. Chem. 280, 34823–34831.

Bouvier, F., Linka, N., Isner, J.C., Mutterer, J., Weber, A.P.M., Camara, B., 2006. Arabidopsis SAMT1 defines a plastid transporter regulating plastid biogenesis and plant development. Plant Cell 18, 3088–3105.

Corpas, F.J., Palma, J.M., Sandalio, L.M., Lopez-Huertas, E., Romero-Puertas, M.C., Barroso, J.B., Del Rio, L.A., 1999. Purification of catalase from pea leaf peroxisomes: identification of five different isoforms. Free Radic. Res. 31 (Suppl.), S235–S241.

Craig, R., Beavis, R.C., 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1467.

Duchene, A.-M., Giritch, A., Hoffmann, B., Cognat, V., Lancelin, D., Peeters, N.M., Zaepfel, M., Marechal-Drouard, L., Small, I.D., 2005. Dual targeting is the rule for organellar aminoacyl-tRNA synthetases in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U.S.A. 102, 16484–16489.

Dunkley, T.P., Hester, S., Shadforth, I.P., Runions, J., Weimar, T., Hanton, S.L., Griffin, J.L., Bessant, C., Brandizzi, F., Hawes, C., Watson, R.B., Dupree, P., Lilley, K.S., 2006. Mapping the Arabidopsis organelle proteome. Proc. Natl. Acad. Sci. U.S.A. 103, 6518–6523.

Douce, R., Joyard, J., 1979. Isolation and properties of the envelope of spinach chloroplasts. In: Reid, E. (Ed.). Plant Organelles. Ellis Horwood Publishers, pp. 47–59.

Duy, D., Wanner, G., Meda, A.R., von Wiren, N., Soll, J., Philippar, K., 2007. PIC1, an Ancient permease in *Arabidopsis* chloroplasts, mediates iron transport. Plant Cell 19, 986–1006.

Eichacker, L.A., Granvogl, B., Mirus, O., Muller, B.C., Miess, C., Schleiff, E., 2004. Hiding behind hydrophobicity: transmembrane segments in mass spectrometry. J. Biol. Chem. 279, 50915–50922.

Ferro, M., Salvi, D., Riviere-Rolland, H., Vermat, T., Seigneurin-Berny, D., Grunwald, D., Garin, J., Joyard, J., Rolland, N., 2002. Integral membrane proteins of the chloroplast envelope: identification and subcellular localization of new transporters. Proc. Natl. Acad. Sci. U.S.A. 99, 11487–11492.

Flügge, U.I., Hausler, R.E., Ludewig, F., Fischer, K., 2003. Functional genomics of phosphate antiport systems of plastids. Physiol. Plant 118, 475–482.

Froehlich, J.E., Wilkerson, C.G., Ray, W.K., McAndrew, R.S., Osteryoung, K.W., Gage, D.A., Phinney, B.S., 2003. Proteomic study of the *Arabidopsis thaliana* chloroplastic envelope membrane utilizing alternatives to traditional two-dimensional electrophoresis. J. Proteome Res. 2, 413–425.

Glynn, J.M., Miyagishima, S.Y., Yoder, D.W., Osteryoung, K.W., Vitha, S., 2007. Chloroplast division. Traffic 8, 451–461.

Gualtieri, G., Kulikova, O., Limpens, E., Kim, D.J., Cook, D.R., Bisseling, T., Geurts, R., 2002. Microsynteny between pea and *Medicago truncatula* in the SYM2 region. Plant Mol. Biol. 50, 225–235.

Gutensohn, M., Fan, E., Frielingsdorf, S., Hanner, P., Hou, B., Hust, B., Klösgen, R.B., 2006. Toc, Tic, Tat et al.: structure and function of protein transport machineries in chloroplasts. J. Plant Physiol. 163, 333–347.

Haswell, E.S., Meyerowitz, E.M., 2006. MscS-like proteins control plastid size and shape in *Arabidopsis thaliana*. Curr. Biol. 16, 1–11.

Heazlewood, J.L., Tonti-Filippini, J.S., Gout, A.M., Day, D.A., Whelan, J., Millar, A.H., 2004. Experimental analysis of the Arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. Plant Cell 16, 241–256.

Huang, X.Q., Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome Res. 9, 868–877.

Ito, J., Heazlewood, J.L., Millar, A.H., 2007. The plant mitochondrial proteome and the challenge of defining the posttranslational modifications responsible for signalling and stress effects on respiratory functions. Physiol. Plant. 129, 207–224.

Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M., Last, R.L., 2002. Arabidopsis map-based cloning in the post-genome era. Plant Physiol. 129, 440–450.

Jarvis, P., Dörmann, P., Peto, C.A., Lutes, J., Benning, C., Chory, J., 2000. Galactolipid deficiency and abnormal chloroplast development in the Arabidopsis MGD synthase 1 mutant. Proc. Natl. Acad. Sci. U.S.A. 97, 8175–8179.

Jarvis, P., 2004. Organellar proteomics: chloroplasts in the spotlight. Curr. Biol. 14, R317–R319.

Kalo, P., Seres, A., Taylor, S.A., Jakab, J., Kevei, Z., Kereszt, A., Endre, G., Ellis, T.H.N., Kiss, G.B., 2004. Comparative mapping between *Medicago sativa* and *Pisum sativum*. Mol. Genet. Genomics 272, 235–246.

Keegstra, K., Yousif, A.E., 1986. Isolation and characterization of chloroplast envelope membranes. Meth. Enzymol. 118, 316–325.

Keller, A., Nesvizhskii, A.I., Kolker, E., Aebersold, R., 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392.

Kleffmann, T., Hirsch-Hoffmann, M., Gruissem, W., Baginsky, S., 2006. plprot: a comprehensive proteome database for different plastid types. Plant Cell Physiol. 47, 432–436.

Lavin, M., Herendeen, P.S., Wojciechowski, M.F., 2005. Evolutionary rates analysis of leguminosae implicates a rapid diversification of lineages during the tertiary. Syst. Biol. 54, 575–594.

Li, L., Tutone, A.F., Drummond, R.S., Gardner, R.C., Luan, S., 2001. A novel family of magnesium transport genes in *Arabidopsis*. Plant Cell 13, 2761–2775.
Lilley, K.S., Dupree, P., 2006. Methods of quantitative proteomics and their application to plant organelle characterization. J. Exp. Bot. 57, 1493–1499.
Liu, H., Sadygov, R.G., Yates, J.R., 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. 76, 4193–4201.
Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E.M., 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 25, 117–124.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.
Maser, P., Thomine, S., Schroeder, J.I., Ward, J.M., Hirschi, K., Sze, H., Talke, I.N., Amtmann, A., Maathuis, F.J.M., Sanders, D., Harper, J.F., Tchieu, J., Gribskov, M., Persans, M.W., Salt, D.E., Kim, S.A., Guerinot, M.L., 2001. Phylogenetic relationships within cation transporter families of *Arabidopsis*. Plant Physiol. 126, 1646–1667.
McConn, M., Hugly, S., Browse, J., Somerville, C., 1994. A mutation at the fad8 locus of *Arabidopsis* identifies a second chloroplast [omega]-3 desaturase. Plant Physiol. 106, 1609–1614.
Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D., Koornneef, M., 1998. *Arabidopsis thaliana*: a model plant for genome analysis. Science 282, 679–682.
Miflin, B.J., Beevers, H., 1974. Isolation of Intact Plastids from a Range of Plant Tissues. Plant Physiol. 53, 870–874.
Millar, A.H., Whelan, J., Small, I., 2006. Recent surprises in protein targeting to mitochondria and plastids. Curr. Opin. Plant Biol. 9, 610–615.
Miras, S., Salvi, D., Ferro, M., Grunwald, D., Garin, J., Joyard, J., Rolland, N., 2002. Non-canonical transit peptide for import into the chloroplast. J. Biol. Chem. 277, 47770–47778.
Miras, S., Salvi, D., Piette, L., Seigneurin-Berny, D., Grunwald, D., Reinbothe, C., Joyard, J., Reinbothe, S., Rolland, N., 2007. TOC159- and TOC75-independent import of a transit sequence-less precursor into the inner envelope of chloroplasts. J. Biol. Chem. 282, 29482–29492.
Miyagishima, S.Y., Froehlich, J.E., Osteryoung, K.W., 2006. PDV1 and PDV2 mediate recruitment of the dynamin-related protein ARC5 to the plastid division site. Plant Cell 18, 2517–2530.
Möhlmann, T., Tjaden, J., Schwoppe, C., Winkler, H.H., Kampfenkel, K., Neuhaus, H.E., 1998. Occurrence of two plastidic ATP/ADP transporters in *Arabidopsis thaliana* L.–molecular characterisation and comparative structural analysis of similar ATP/ADP translocators from plastids and *Rickettsia prowazekii*. Eur. J. Biochem. 252, 353–359.
Nesvizhskii, A.I., Keller, A., Kolker, E., Aebersold, R., 2003. A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 4646–4658.
Neuhaus, H.E., Thom, E., Mohlmann, T., Steup, M., Kampfenkel, K., 1997. Characterization of a novel eukaryotic ATP/ADP translocator located in the plastid envelope of *Arabidopsis thaliana* L. Plant J. 11, 73–82.
Parinov, S., Sundaresan, V., 2000. Functional genomics in *Arabidopsis*: large-scale insertional mutagenesis complements the genome sequencing project. Curr. Opin. Biotechnol. 11, 157–161.
Peltier, J.B., Ytterberg, A.J., Sun, Q., van Wijk, K.J., 2004. New functions of the thylakoid membrane proteome of *Arabidopsis thaliana* revealed by a simple, fast, and versatile fractionation strategy. J. Biol. Chem. 279, 49367–49383.
Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C., Zhang, X., 2006. Performance evaluation of existing de novo sequencing algorithms. J. Proteome Res. 5, 3018–3028.
Picault, N., Hodges, M., Paimieri, L., Palmieri, F., 2004. The growing family of mitochondrial carriers in *Arabidopsis*. Trends Plant Sci. 9, 138–146.
Radhamony, R.N., Theg, S.M., 2006. Evidence for an ER to Golgi to chloroplast protein transport pathway. Trends Cell Biol. 16, 385–387.
Reiser, J., Linka, N., Lemke, L., Jeblick, W., Neuhaus, H.E., 2004. Molecular physiological analysis of the two plastidic ATP/ADP transporters from *Arabidopsis*. Plant Physiol. 136, 3524–3536.
Renné, P., Dreßen, U., Hebbeker, U., Hille, D., Flügge, U.I., Westhoff, P., Weber, A.P.M., 2003. The *Arabidopsis* mutant *dct* is deficient in the plastidic glutamate/malate translocator DiT2. Plant J. 35, 316–331.
Reumann, S., Weber, A.P.M., 2006. Plant peroxisomes respire in the light: Some gaps of the photorespiratory C2 cycle have become filled - Others remain. Biochim. Biophys. Acta 1763, 1496–1510.
Reyes-Prieto, A., Weber, A.P.M., Bhattacharya, D., 2007. The Origin and Establishment of the Plastid in Algae and Plants. Annu. Rev. Genet. 41, 147–168.
Schmidt, U.G., Endler, A., Schelbert, S., Brunner, A., Schnell, M., Neuhaus, H.E., Marty-Mazars, D., Marty, F., Baginsky, S., Martinoia, E., 2007. Novel tonoplast transporters identified using a proteomic approach with vacuoles isolated from cauliflower buds. Plant Physiol. 145, 216–229.
Schneidereit, J., Häusler, R.E., Fiene, G., Kaiser, W.M., Weber, A.P.M., 2006. Antisense repression reveals a crucial role of the plastidic 2-oxoglutarate/malate translocator DiT1 at the interface between carbon and nitrogen metabolism. Plant J. 45, 206–224.
Schnurr, J.A., Shockey, J.M., de Boer, G.-J., Browse, J.A., 2002. Fatty acid export from the chloroplast. molecular characterization of a major plastidial acyl-coenzyme a synthetase from *Arabidopsis*. Plant Physiol. 129, 1700–1709.
Seigneurin-Berny, D., Gravot, A., Auroy, P., Mazard, C., Kraut, A., Finazzi, G., Grunwald, D., Rappaport, F., Vavasseur, A., Joyard, J., Richaud, P., Rolland, N., 2006. HMA1, a new Cu-ATPase of the chloroplast envelope, is essential for growth under adverse light conditions. J. Biol. Chem. 281, 2882–2892.
Shevchenko, A., Wilm, M., Vorm, O., Mann, M., 1996. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. Anal. Chem. 68, 850–858.
Shikanai, T., Muller-Moule, P., Munekage, Y., Niyogi, K.K., Pilon, M., 2003. PAA1, a P-Type ATPase of *Arabidopsis*, functions in copper transport in chloroplasts. Plant Cell 15, 1333–1346.
Somerville, C., Somerville, S., 1999. Plant functional genomics. Science 285, 380–383.
The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.
Taira, M., Valtersson, U., Burkhardt, B., Ludwig, R.A., 2004. *Arabidopsis thaliana* GLN2-encoded glutamine synthetase is dual targeted to leaf mitochondria and chloroplasts. Plant Cell 16, 2048–2058.
Taylor, N.L., Heazlewood, J.L., Day, D.A., Millar, A.H., 2005. Differential impact of environmental stresses on the pea mitochondrial proteome. Mol. Cell. Proteomics 4, 1122–1133.
Teng, Y.S., Su, Y.S., Chen, L.J., Lee, Y.J., Hwang, I., Li, H.M., 2006. Tic21 is an essential translocon component for protein translocation across the chloroplast inner envelope membrane. Plant Cell 18, 2247–2257.
Tobin, A.K., 1996. Subcellular fractionation of plant tissues. Isolation of chloroplasts and mitochondria from leaves, Methods Mol. Biol. 59, 57–68.
Villarejo, A., Buren, S., Larsson, S., Dejardin, A., Monne, M., Rudhe, C., Karlsson, J., Jansson, S., Lerouge, P., Rolland, N., von Heijne, G., Grebe, M., Bako, L., Samuelsson, G., 2005. Evidence for a protein transported through the secretory pathway en route to the higher plant chloroplast. Nat. Cell. Biol. 7, 1224.
von Zychlinski, A., Kleffmann, T., Krishnamurthy, N., Sjolander, K., Baginsky, S., Gruissem, W., 2005. Proteome analysis of the rice etioplast - Metabolic and regulatory networks and novel protein functions. Mol. Cell. Prot. 4, 1072–1084.
Weber, A., Flugge, U.I., 2002. Interaction of cytosolic and plastidic nitrogen metabolism in plants. J. Exp. Bot. 53, 865–874.
Weber, A., Menzlaff, E., Arbinger, B., Gutensohn, M., Eckerskorn, C., Flugge, U.I., 1995. The 2-oxoglutarate/malate translocator of chloroplast envelope membranes: molecular cloning of a transporter containing a 12-helix motif and expression of the functional protein in yeast cells. Biochemistry 34, 2621–2627.
Weber, A.P.M., 2004. Solute transporters as connecting elements between cytosol and plastid stroma. Curr. Opin. Plant Biol. 7, 247–253.
Weber, A.P.M., Fischer, K., 2007. Making the connections – the crucial role of metabolite transporters at the interface between chloroplast and cytosol. FEBS Lett. 581, 2215–2222.
Weber, A.P.M., Schneidereit, J., Voll, L.M., 2004. Using mutants to probe the in vivo function of plastid envelope membrane metabolite transporters. J. Exp. Bot. 55, 1231–1244.
Weber, A.P.M., Schwacke, R., Flügge, U.I., 2005. Solute transporters of the plastid envelope membrane. Annu. Rev. Plant Biol. 56, 133–164.
Weber, A.P.M., Weber, K.L., Carr, K., Wilkerson, C., Ohlrogge, J.B., 2007. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. Plant Physiol. 144, 32–42.
Wikström, N., Savolainen, V., Chase, M.W., 2001. Evolution of the angiosperms: calibrating the family tree. Proc. R. Soc. Lond., B, Biol. Sci. 268, 2211–2220.
Yang, Y.W., Lai, K.N., Tai, P.Y., Li, W.H., 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. J. Mol. Evol. 48, 597–604.
Ytterberg, A.J., Peltier, J.B., van Wijk, K.J., 2006. Protein profiling of plastoglobules in chloroplasts and chromoplasts. A surprising site for differential accumulation of metabolic enzymes, Plant Physiol. 140, 984–997.
Yu, B., Xu, C.C., Benning, C., 2002. Arabidopsis disrupted in SQD2 encoding sulfolipid synthase is impaired in phosphate-limited growth. Proc. Natl. Acad. Sci. U.S.A. 99, 5732–5737.
Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000. A greedy algorithm for aligning DNA sequences. J. Comp. Biol. 7, 203–214.
Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S., Vagner, L.L., Khaspekov, G.L., Kozhemyako, V.B., Matz, M.V., Meleshkevitch, E., Moroz, L.L., Lukyanov, S.A., Shagin, D.A., 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res. 32, e37.
Zybailov, B., Coleman, M.K., Florens, L., Washburn, M.P., 2005. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. Anal. Chem. 77, 6218–6224.