Available online at www.sciencedirect.com
SCIENCE DIRECT®
Phytochemistry 65 (2004) 1517–1530
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

Review

Untangling multi-gene families in plants by integrating proteomics into functional genomics

Pia G. Sappl, Joshua L. Heazlewood, A. Harvey Millar *

Plant Molecular Biology Group, School of Biomedical and Chemical Sciences, The University of Western Australia, M310, Biochemistry, 35 Stirling Highway, Crawley, Perth 6009, WA, Australia

Received 9 January 2004; received in revised form 1 April 2004
Available online 24 May 2004

Abstract

The classification and study of gene families is emerging as a constructive tool for fast tracking the elucidation of gene function. A multitude of technologies can be employed to undertake this task including comparative genomics, gene expression studies, sub-cellular localisation studies and proteomic analysis. Here we focus on the growing role of proteomics in untangling gene families in model plant species. Proteomics can specifically identify the products of closely related genes, can determine their abundance, and coupled to affinity chromatography and sub-cellular fractionation studies, it can even provide location within cells and functional assessment of specific proteins. Furthermore global gene expression analysis can then be used to place a specific family member in the context of a cohort of co-expressed genes. In model plants with established reverse genetic resources, such as catalogued T-DNA insertion lines, this gene specific information can also be readily used for a wider assessment of specific protein function or its capacity for compensation through assessing whole plant phenotypes. In combination, these resources can explore partitioning of function between members and assess the level of redundancy within gene families.

Keywords: Gene families; Proteomics; *Arabidopsis*; Organelles

Contents

* Corresponding author. Tel.: +61-8-9380-7245/6488-7245; fax: +61-8-9380-1148/6488-1148.
E-mail address: hmillar@cyllene.uwa.edu.au (A.H. Millar).

0031-9422/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2004.04.021

1. Introduction

Historically, and for largely technical reasons, genes and their products have generally been studied as single entities. In this post-genomic era, with publicly available data from genome and EST sequencing projects, it is evident that most genes are not singletons but exist as members of gene families. Consequently, the polypeptide products of genes operate alongside similar proteins in cells that are encoded by other members of their gene families. The development of gene and protein specific technologies has paved the way for studying the expression and contribution of the individual members of such families and allows us to begin to answer questions about both partitioning of functions between members and redundancy within gene families. With the completions of both the *Arabidopsis* and rice genomes (Kaul et al., 2000; Goff et al., 2002) recent publications have utilised these genomic resources to build a picture of particular gene families. By taking a genomic perspective, researchers are able to define a gene family in the context of their gene of interest. The study of gene families will not only simplify the task of elucidating the function of every type of protein, but will also help us appreciate the evolutionary pressures leading to the expansion and preservation of gene duplicates within plant genomes. In this review we overview the scale of the gene family issue in model plants and consider research contributions that attempt to tackle it. We have not tried to systematically review all the literature, but give references to illustrate key aspects of gene family function and redundancy that have been discovered through experimentation to date. We especially consider the place of proteomic approaches, both applied and potential, in determining the expression, location and function of specific gene family products in plants.

2. Plant functional genomics – why study gene families?

John Donne placed individual people in the context of their society in the early 17th century when he wrote "no man is an island, entire of itself, every man is a piece of the continent, a part of the main" (Donne, 1999). The same concept can also be applied to our understanding of genes and proteins. Genes have their place and purpose in genomes and proteins their location and functional significance in the context of proteomes. The immediate influence of the wider genome and proteome on the individual gene and protein are thus seen in the operation and evolution of gene families.

2.1. Defining gene families

A gene family can be defined as 'a set of genes coding for diverse proteins which, by virtue of their high degree of sequence similarity, are believed to have evolved from a single ancestral gene' (Lackie and Dow, 1999). Grouping genes into their respective families in this manner is an important step towards the comprehensive annotation of genomes (Dayhoff, 1976). It is important to note that this is not simply an anthropomorphic exercise of society building in genomes. Rather, it is an attempt to understand function, sub-functionalisation and gene redundancy within the context of the best and most complete set of information available. At present this information is largely raw genome sequence and predicted open reading frames, hence the emphasis on primary sequence in gene family definition.

A variety of parameters could in theory be utilised to define families but those most commonly used are degrees of primary sequence identity and shared functional motifs or domains (Wu et al., 2003). Many of the larger families or 'superfamilies' are grouped together based on a shared motif or domain and consequently may be comprised of members with disparate functions, despite a common ancestry. The MYB family of DNA-binding proteins (Stracke et al., 2001), the protein kinase superfamilies (eg. Hrabak et al., 2003; http://plantsp.sdsc.edu/) and the glutathione S-transferase (GST) family of detoxification proteins (Wagner et al., 2002) in *Arabidopsis* are good examples of such families. The MYB family is comprised of over 130 proteins that have significant sequence identity in the conserved DNA-binding domain but vary greatly outside this region. Other than acting as transcriptional regulators, their gene targets and specific roles differ substantially. The protein kinase superfamily contains about 1000 members in-

volved in many aspects of structural and constitutive phosphorylation through to rapid signalling in defence responses (http://plantsp.sdsc.edu/). The GST superfamily of 53 members collectively exhibit a variety of enzymatic activities ranging from their GSH-transferase activity to glutathione peroxidase, isomerase and dehydroascorbate reductase activity (Wagner et al., 2002). Such large families can be considered as a whole, but can also be grouped into smaller sub-classes or clades using methods such as sequence similarity or intron:exon structure, which further aids the overall analyses by simplifying the larger families into potentially more related functional groupings. Fig. 1 shows the example of the GST superfamily sub-divided into six classes based on amino acid identity and intron:exon placement. This type of analysis shows that the plant specific phi and tau classes clearly dominate both in terms of GST gene numbers and expression levels and reveals highly expressed members within each class. Consideration of the spatial arrangement of genes within the genome may also be informative. Superfamilies may arise from both local duplication events, leading to clusters of closely related genes being located side by side, and also large-scale duplication events, leading to closely related genes being located on different chromosomes or on different regions of the same chromosome. In the case of the GST family, 35 of the 53 GST genes occur in 12 clusters, each cluster containing members of a single class and pointing to a multitude of local duplication events. Approximately half of the GST genes also lie in segmentally duplicated regions of the genome (Fig. 1(B)). For example a cluster of four phi GSTs: F4, F5, F6 and F7 reside in a segmentally duplicated region common to chromosomes 1 and 4, where GSTF2 is found. These duplications result in gene family members, often encoding highly similar protein products, being encoded in different chromosomal environments and often having very different expression patterns. Thus defining the exact gene responsible for a protein is vital since it allows knowledge of the complexity of gene families and the origin of family members to be considered in the interpretation of data.

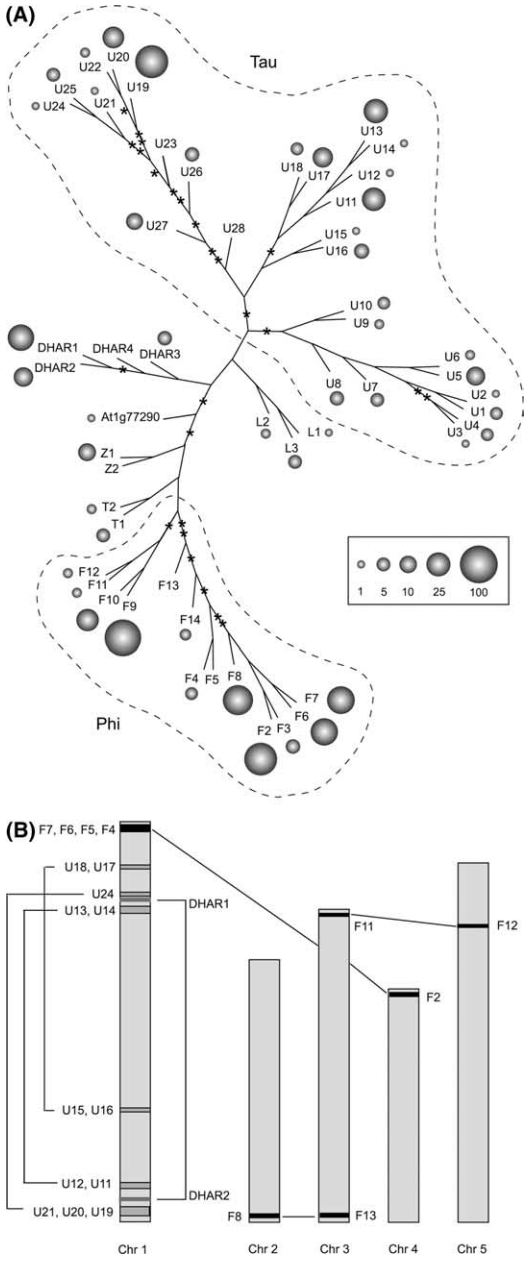


Fig. 1. (A) A diagrammatic representation of the *Arabidopsis thaliana* GST super-family. Unrooted bootstrapped ($n = 1000$) phylogenetic tree based on a multiple sequence alignment using 53 full-length GST protein sequences. The number of ESTs (TIGR gene indicies) for each individual GST is represented by the volume of the corresponding sphere. Such phylogenetic trees obtained from sequence alignment analysis are one of the more common methods for visualising intra-family relationships. The more closely related members form clades and potentially share some similar functions within the context of an entire superfamily. Computer based alignments and the analysis of protein or nucleotide sequences can be undertaken through a variety of packages, including CLUSTAL W (Thompson et al., 1994), PHYLIP (Felsenstein, 1993) and PAUP (Swofford, 2003). (B). Distribution of duplicated GST genes over the five *Arabidopsis* chromosomes. Close to half of the GST genes have arisen via large-scale duplication events and 23 are located within segmentally duplicated regions of the *Arabidopsis* genome (information obtained from TIGR). The distribution of the GSTs residing in segmental duplications is indicated with connecting lines.

2.2. Scale of families in model plants

The post-genomic analyses of *Arabidopsis* and rice have revealed an unexpected predominance of gene families. In fact 65% of predicted proteins from *Arabidopsis* are now considered to belong to multi-gene families (Fig. 2). The specific functional analysis of each gene within every gene family remains an enormous task for researchers to consider. However, the identification and appreciation of gene families has allowed researchers to target their resources towards specific members or subgroups as representatives of the whole. By taking such an

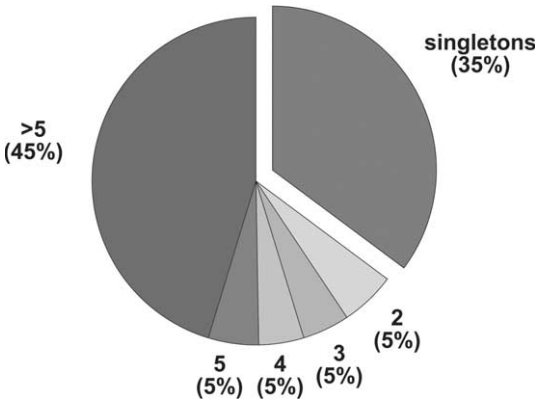


Fig. 2. The distribution of gene singletons and paralogous family members within the *Arabidopsis* genome. Sixty-five percent of genes belong to a gene family. The gene family data were obtained from The Institute for Genomic Research (TIGR: (Wortman et al., 2003) and was created using sequence identity and protein domain matches on the ≈28,000 predicted proteins from *Arabidopsis* compiled by TIGR.

approach, 25,498 predicted genes from *Arabidopsis* can be reduced to 11,601 distinct groups (Kaul et al., 2000) while a subset of ≈40,000 predicted rice genes reduced to ≈15,000 discrete families (Goff et al., 2002). With this knowledge, several routes can be taken for analysis. Firstly, researchers can work on a family set in order to dissect its function or functions. To undertake such a task, gene and protein specific technologies must be employed for the analysis of each individual member. Alternatively, gene families can be considered as functionally redundant sets. This route should be approached with caution, but in the context of the significant task of dealing with a multitude of permutations it may be the best route for a global appreciation of genome function, even if some subtleties are lost. This approach can potentially fast-track functional determination by allowing knowledge to be extrapolated for similar genes and will be especially valuable for members with high sequence identity. These closely related members may represent functional homologs with co-ordinated responses or duplicates that act redundantly. In *Arabidopsis*, 50% of all proteins belong to families with at least five members (Fig. 2). It appears that tandem gene duplications and segmental duplications have been major driving forces or at least the mechanistic processes in the explosion of many of these gene families (Kaul et al., 2000). For example, segmental duplication is responsible for 6303 gene duplications while 1582 arrays of tandem duplications contain 4140 genes in *Arabidopsis* (Kaul et al., 2000). But these processes do not account for all of the apparent duplications found in *Arabidopsis* and many genes with high levels of sequence similarity are scattered throughout the genome. Comparing the chromosomal locations of family members can reveal whether gene duplications were a result of tandem or chromosomal segmental duplications. Thus genomic resources can be extremely useful in building an informative picture of gene families that can then help to direct experimental work either to provide a list for reductionist work on a single family, or to provide a framework for a simplification of the gene loci list into putative functional groups.

2.3. Resources for studying plant gene families

Building a genomic picture of a gene family can be a time consuming task, especially for researchers not familiar with the specific class of proteins it encodes. With the finalization of the *Arabidopsis* genome sequencing and annotation projects, a number of research articles have attempted to define many prominent gene families in this model species (Table 1). Multiple web-based resources have also appeared such as the 'gene families list' at The *Arabidopsis* Information Resource (TAIR), which relies on researchers submitting lists of genes that belong to families (http://www.arabidopsis.org/info/genefamily/genefamily.html). Currently the list comprises 5369 genes and 740 families but while many

Table 1
A list of some of the major characterised *Arabidopsis* superfamilies

Family	Members	Function	References
Protein kinase	~1000	Signalling	PlantsP: http://plantsp.sdsc.edu/; Hrabak et al. (2003)
F-box proteins	568–694	Miscellaneous	Gagne et al. (2002); Kuroda et al. (2002)
Pentatricopeptide (PPR) repeat	~450	RNA processing	http://www.evry.inra.fr/public/projects/ppr/ppr.html
Cytochrome P450	272	Metabolism	http://arabidopsis-p450.biotec.uiuc.edu/
bHLH factors	162	Transcription	Bailey et al. (2003)
MYB factors	133	Transcription	http://www.mpiz-koeln.mpg.de/myb/
Phosphatases	131	Signalling	PlantsP: http://plantsp.sdsc.edu/
ABC proteins	129	Transport	Sanchez-Fernandez et al. (2001)
Zinc finger (C3HC4)	125	Miscellaneous	http://arabidopsis.med.ohio-state.edu/AtTFDB/index.jsp
UDP-glucosyltransferase	121	Metabolism	http://www.biobase.dk/P450/
AP2/EREBP factors	120	Transcription	http://arabidopsis.med.ohio-state.edu/AtTFDB/index.jsp
MADS box factors	104	Transcription	Martinez-Castilla and Alvarez-Buylla (2003)

Gene families compiled using combinations of sequence identity and by the presence of defined functional domains.

Table 2
The top 10 most abundant gene families in *Arabidopsis* based on the number of protein matches per InterPro domain (Mulder et al., 2003)

Family	Members	Function	InterPro domain
Protein kinase	1055	Signalling	IPR000719
Ser/Thr protein kinase	799	Signalling	IPR008271
F-box domain	650	Miscellaneous	IPR001810
TPR domain	575	Miscellaneous	IPR008941
Leucine-rich repeat	544	Miscellaneous	IPR001611
Zinc finger (C3HC4)	492	Miscellaneous	IPR001841
Pentatricopeptide (PPR) repeat	457	RNA processing	IPR002885
HMG-I/HMG-Y (A + T-hook)	385	DNA processing	IPR000637
MYB DNA-binding domain	344	Transcription	IPR001005

families are well defined, the annotation of others appears somewhat preliminary. The Plant Specific Database has further extended these families at TAIR and integrated them with EST counts, sub-cellular localisation predictions, membrane/soluble scores and an automatic functional assignment from the Munich Information Center for Protein Sequences (MIPS) to add an extra dimension of functionality to these family lists (http://genomics.msu.edu/plant_specific/index. html). An alternative approach to the methodical assessment of families based on whole sequence protein similarity is simply the automated use of functional motifs or domains to define families (Table 2). Use of this approach in *Arabidopsis* provides a complementary view of superfamilies, which sometimes confirms the results from whole sequence similarity studies, and sometimes provides overlapping family structures (compare Tables 1 and 2). Ultimately, links to phylogenetic trees and expression analyses could be incorporated into these family description structures. These types of resources are likely to become an important way of obtaining some idea of the size and phylogeny of gene families and utilising the detailed knowledge of other researchers.

3. Making the most of gene specific technologies

The experimental study of gene families relies on the ability to readily distinguish between related members. Traditional methods for following gene expression, such as northern hybridisation, often failed to distinguish between the similarly sized members of gene families and suffered from cross-hybridisation between similar sequences. Only by separately following and identifying individual members of gene families and their products can we really address questions of differential expression, localisation and ultimately gene function within families. The uses of technologies such as semi-quantitative RT-PCR and more recently real-time PCR and oligomeric microarrays can be used to address some of these issues. These technologies can be readily designed to deal with the confounding issue of sequence redundancies within gene families.

3.1. Profiling transcript expression of gene families

A fundamental question relating to gene families is whether gene explosion to form a particular family is accompanied more by divergence of function or by divergence of expression. In many gene families members are expressed simultaneously in multiple tissues. Rather than simply asking: ''Is gene X expressed in tissue Y?'' we need more subtle measures of expression to tease apart differences between related genes. This is aided by gene specific transcript expression profiling under a range of tissue types and conditions. An appreciation of the subtleties of transcriptional control (e.g., dose response, kinetics, magnitude of response) will aid in uncovering similarities and differences amongst gene members. This kind of analysis is important for understanding why highly similar genes might be preserved in the genome and the data also forms the basis for further functional studies.

Recent transcript profiling of gene families has revealed that in most families all members are expressed at some point in time. Gene specific analysis of the 14 multidrug-resistance related protein (MRP) ABC transporter family members (Kolukisaoglu et al., 2002), the 10 xyloglucan fucosyltransferase (*FUT*) genes (Sarria et al., 2001) and 33 xyloglucan endotransglucosylase (*XTH*) genes (Yokoyama and Nishitani, 2001) has revealed that all members of these families were expressed. However, individual members displayed differential expression patterns. Even highly similar genes belonging to very tight clades appear to be expressed very differently. Other differences include tissue specificities, cell specific and treatment specific induction patterns. It appears that the diversification by duplication of the XTH family has allowed for specialisation of organ-specific expression and hormone induction rather than variation in enzyme function (Yokoyama and Nishitani, 2001). Similar conclusions were also reached for the five sub-families of the plasma membrane proton pump

ATPases where duplication appears to have also allowed for a diversification of regulation (Arango et al., 2003). Defining the combination of transcription factors that regulate the expression of individual members of gene families will be very useful in order to understand how diversification of regulation has been achieved.

Within gene families some members are often highly expressed, perhaps providing activity at a constitutive level, while other members are lowly expressed, possibly expressed only in specific tissues or under more specific conditions. Comparing EST numbers within gene families can often identify highly expressed members (Rafalski et al., 1998; Andrews et al., 2000). Thus far, this approach appears to hold true in studies of *Arabidopsis* gene families (Yokoyama and Nishitani, 2001; Dixon et al., 2002; Wagner et al., 2002; Millar and Heazlewood, 2003). It may be that these highly expressed genes are conserved across species and perform a 'default' role. For instance an *Arabidopsis* guaiacol peroxidase expressed with abundant expression levels in most tissues, that is also highly represented in EST libraries has a specific homolog in cotton, soybean and tobacco that are also highly expressed (Tognolli et al., 2002).

Lowly expressed members of gene families are often overlooked and sometimes discounted as pseudogenes. However in well annotated genomic projects, such as that of *Arabidopsis*, there appear few cases where annotated ORFs designated as members of gene families are non-expressed pseudogenes. Lowly expressed genes with few or no ESTs could potentially be induced under specific conditions or in particular developmental stages or tissues. Sometimes these genes appear to be more responsive to environmental cues than the highly expressed constitutive isoforms. For example the lowly expressed members of the gene families encoding the mitochondrial import apparatus in *Arabidopsis* are specially induced under stress conditions (Lister et al., 2004).

3.2. Reverse genetics and complementation to address functionality

Another approach to studying gene families involves addressing the question of functional redundancy through reverse genetics. There are many limitations and challenges in this approach including the difficulty of identifying phenotypic changes. Determining the expression patterns of family members may help in the analysis of gene knockouts such that attention may be focussed on particular developmental stages or organs where closely related members are not simultaneously expressed. A number of studies have used reverse genetic approaches in this way to study function in the context of multi-gene families. Sarria et al. (2001) analysed the nine *Arabidopsis FUT* genes involved in xyloglucan biosynthesis, overexpressed three different members, and analysed the sugar composition of cell walls in *Arabidopsis*. This study highlighted the difficulties in identifying phenotypic changes associated with particular genes and the authors pointed to the need for a targeted analysis of cell wall matrix components in order to differentiate the role of these three members. Individual knockouts of the *Arabidopsis* plasma membrane proton pumping ATPases, in contrast, had noticeable phenotypes and double knockouts had a lethal phenotype, suggesting some redundancy exists in this family, a homozygous individual knockout of a related ATPase was found to be lethal (Arango et al., 2003). Gene function and hence gene redundancy is likely a function of both biochemical function and the spatiotemporal expression of the gene. Thus, sequence similarity amongst family members alone will not always allow for predictions of functional redundancy. Knockouts of single members of the syntaxin gene SYP2 and SYP4 sub-families resulted in lethality of the male gametophyte in *Arabidopsis* (Sanderfoot et al., 2001). Thus the proteins involved in the fusion of transport vesicles to target membranes, despite extensive duplication in *Arabidopsis* as compared to model eukaryotes such as yeast, clearly serve essential non-redundant functions (Sanderfoot et al., 2001).

Complementation is often used to show redundancy or equivalence of function. However this may also be a very crude measure of function. In their study of *Arabidopsis* proton pumping ATPases, Arango et al. (2003) show that although PMA2 and PMA4 can replace the yeast H^+-ATPase gene, they do not produce identical phenotypes. Furthermore while different sub-family members can functionally replace the yeast H^+-ATPase gene there are subtle differences in the degree of complementation. Therefore these methods of measuring redundancy are probably insufficient to determine the roles of individual members without detailed characterisation of the complementation in planta. The more knowledge we have of the properties and functions of different families the more chance we have of finding phenotypes in knockouts.

4. Making the most of protein specific technologies

The annotation of genome sequences coupled with developments in mass spectrometry has meant that proteomics can now also be fully integrated into a genomic context and protein families can be explored. Rather than relying on a battery of specific antibodies to try and identify protein products, single mass spectrometers yield peptide mass fingerprints and tandem mass spectrometers deliver peptide sequence information to provide a high level of downstream specificity (Graves and Haystead, 2002). Consequently, mass spectrometry analysis provides protein-for-gene identifications such that a protein product can be mapped back to a single

genomic locus in a sequenced organism. This level of identification is critical in the study of gene families and their individual members, providing a level of specificity necessary for following individual gene products. However, due to limitations in the dynamic resolution of mass spectrometry and its non-quantitative nature, pre-fractionation of cellular protein is still generally required to study more than the most abundant set of proteins in complex cell lysates. This fractionation may come in the form of polyacrylamide gel electrophoresis, sub-cellular fractionation, or through column chromatography, culminating in online fractionation of proteins or peptides by HPLC (Whitelegge et al., 1998; Herbert and Righetti, 2000; Jonsson, 2001). In the following sections we describe in detail some proteomic approaches that have proved useful for the study of gene family products.

4.1. Two-dimensional gel separations for proteomics of protein families

2D-PAGE is a very useful method of separating cellular proteins (Herbert, 1999). Typically 1D-PAGE followed by western blotting and immunodetection fails to distinguish between closely related members of gene families unless antibodies have been carefully raised against specific regions of the protein of interest. However this antibody cross-reactivity can be exploited in the location of multiple family members on 2D gels. Western blots of 2D gels can be successfully probed to locate immunoreactive proteins (Petersen et al., 1993; Suty et al., 2003). The corresponding protein spot on a duplicate Coomassie stained 2D gel can be excised, trypsin digested and individual protein members identified by mass spectrometry. The combination of immunodetection with the high resolution of 2D-PAGE means that closely related proteins (e.g. up to 95% identical at the amino acid level) can be distinguished due to differences in their isoelectric points and variation in amino acid sequence in peptides derived by proteolysis (Millar et al., 2001). A series of plant thioredoxins (Maeda et al., 2003) and a strange variety of proteins that reacted with mammalian NO synthase antibodies have been identified by this method (Butt et al., 2003). We have used this approach to separate the single immunodetected band of GSTs in *Arabidopsis* on 1D-PAGE into a series of specific family members detected with the same antibody on 2D-PAGE (Fig. 3(A); Sappl et al., 2004). 2D-PAGE has the advantage of providing quantitative information about the abundance of different family members and may also reveal post-translational modifications (Peck et al., 2001). Unfortunately this analysis is time consuming, technically demanding and its application is limited by difficulties in automation allowing for high throughput.

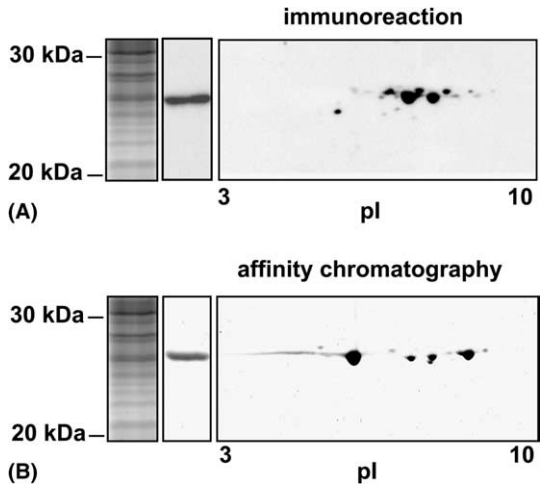


Fig. 3. Comparison of 1D and 2D PAGE analysis of expressed members of the GST family from *Arabidopsis*. (A) Anti-GST antibodies detected a single band on 1D PAGE, but can be separated to 6 major and several minor protein features by 2D PAGE that are encoded by at least five different GST genes. (B) Coomassie blue stained single GST elution band from GSH affinity media can be separated into an overlapping set of protein spot features by 2D PAGE comprising a wider set of GST isoforms (Sappl et al., 2004).

4.2. One-dimensional gel separations for proteomics of protein families

The similar molecular mass of gene family members, rather than being a limitation, can also be exploited as a useful approach for honing in on a sub-group of proteins of interest. For example, gel slices from a single lane and corresponding to the MW range of the gene family products can be excised, trypsin digested, and analysed by mass spectrometry to identify multiple members simultaneously. This approach has a higher throughput but provides only semi-quantitative data. We have used it to identify the major mitochondrial carrier family members that are present in *Arabidopsis* mitochondria (Millar and Heazlewood, 2003) and a set of 20 GSTs expressed in *Arabidopsis* cell culture (Sappl et al., 2004).

4.3. Non-gel undirected chromatography for proteomics of protein families

An alternative proteomic approach involves the large-scale identification of hundreds or thousands of proteins. Although undirected this approach has the potential to generate a huge amount of data, allowing for the study of many families of gene products. Specifically this involves LC-MS/MS or LC/LC-MS/MS analysis of protein mixtures. The major analysis of the rice proteome by this methodology revealed multiple components of storage protein gene families, with 13 different glutelins, 10 prolamins and 7 globulins (Koller et al., 2002). Extensive work using these techniques has

also been carried out to identify protein sets in *Arabidopsis* (VerBerkmoes et al., 2002) and in mitochondria (Heazlewood et al., 2004) and chloroplasts (Froehlich et al., 2003). LC-MS/MS separations of intact proteins from photosynthetic protein complexes have also provided broad identification of the subunits as well as information on post-translational modification and presequence cleavage points (Sharma et al., 1997; Gomez et al., 2002; Whitelegge et al., 2002; Zolla et al., 2002). Unfortunately these approaches in themselves are not typically quantitative but techniques such as isotope coded affinity tagging (ICAT) could be incorporated to provide some quantitative comparison between samples (Gygi et al., 1999).

4.4. Non-gel directed chromatography for proteomics of protein families

A wide array of affinity chromatography media are available for protein purification based on substrate mimicking or other specific binding characteristics (Burgess and Thompson, 2002; Bauer and Kuster, 2003). These can be exploited for proteomic analysis of gene family members both to assess family member expression and to provide a measure of functionality to defining families in the context of the physical properties of the protein products rather than just protein sequence or gene structure similarities. Immobilised metal chromatography is being exploited to identify metal binding proteins involved in transport or chelation of the metal or proteins that use the metal in catalysis (Lopez et al., 2000; Gupta et al., 2002). Similar techniques have also been used to isolate over 130 lectins with a binding affinity for mannose from rice resulting in the identification of a series of mannose binding proteins (Andon et al., 2003). Methodology for affinity isolation of polysaccharide binding proteins in plants has also been reported (Eckermann et al., 2002), as has the enrichment and identification of phosphoproteins, which as low abundance components in plant cells require affinity purification for identification (Romeis, 2001; Hansson and Vener, 2003). Nucleic acid binding proteins can be affinity-purified using a variety of media and metabolic enzymes using ADP, ATP, NAD and NADP resins (Schott, 1984). Bound glutathione also provides an avenue to purify and assess the products of protein families that utilise this tripeptide. We have used this media to isolate a specific subset of 8 *Arabidopsis* GSTs that bind this media and then separated these on 2D gels (Fig. 3(B)). It should of course be noted that association of proteins with complexes or apparent affinity of proteins for a ligand can be opportunistic and final assignments of functional association requires complementary data such as genetic evidence.

4.5. Distribution of gene family products between cellular sub-compartments

Often redundant functions are required in eukaryotes to provide operations in separated compartments of the cell. In such instances identical enzymatic functions would not translate to redundancy. Two main avenues have been presented for considering the sub-cellular location of gene family products in this context. Firstly, epitope tagging has been used to localise each member of a family to a sub-cellular compartment (Kumar et al., 2002; Hilson et al., 2003). Secondly, the results of sub-cellular proteomic studies have been combined to retrospectively define localisation of gene family products (Jung et al., 2000). The most popular route for epitope tagging is the generation of GFP chimeras with either the entire protein sequence or the predicted targeting sequence. This has been used to localise isoforms of the branched chain aminotransferase required in amino acid synthesis and metabolism to different organelles in *Arabidopsis* by fluorescence imaging (Diebold et al., 2002). Similar approaches have been used to localise five geranylgeranyl diphosphate synthases to chloroplasts, mitochondria and the ER (Okada et al., 2000) and nine calcium-dependent protein kinases to a host of intracellular locations (Dammann et al., 2003). Sub-cellular proteomic studies have now been performed in chloroplasts (Peltier et al., 2002; Schubert et al., 2002; Ferro et al., 2003; Froehlich et al., 2003), mitochondria (Heazlewood et al., 2004), peroxisomes (Fukao et al., 2002), nuclei (Bae et al., 2003; Calikowski et al., 2003), ER and plasma membrane (Prime et al., 2000) and the cell wall (Chivasa et al., 2002) of *Arabidopsis* (Fig. 4(A)). When these sets are compared, discrete proteome complements can be viewed for each sub-compartment. Proteins that span multiple compartments, shown as the intersections in the Venn diagram depicted, include contaminants in organelle preparations but also authentically multi-targeted proteins (Fig. 4(B)). These experimentally defined sets can also be analysed for the sub-cellular distributions of gene families to dissect their specific functions. As an example, the gene families encoding malate dehydrogenase and ascorbate peroxidase are localised in multiple compartments according to this analysis (Table 3). Similar studies have helped to define the sub-cellular location of products from other defence enzyme and protein import pathway families (Chew et al., 2003; Milla et al., 2003; Chew et al., 2004). A significant advantage of sub-fractionation is the ability to produce proteomic depth. Sub-cellular proteomics can more reliably examine low abundance isoforms in low-abundance proteomes within the cell. Such methods have been successfully employed to illustrate that low-abundance isoforms of particular families are actively expressed and do not represent pseudogenes (Friso et al., 2004; Heazlewood et al., 2004).

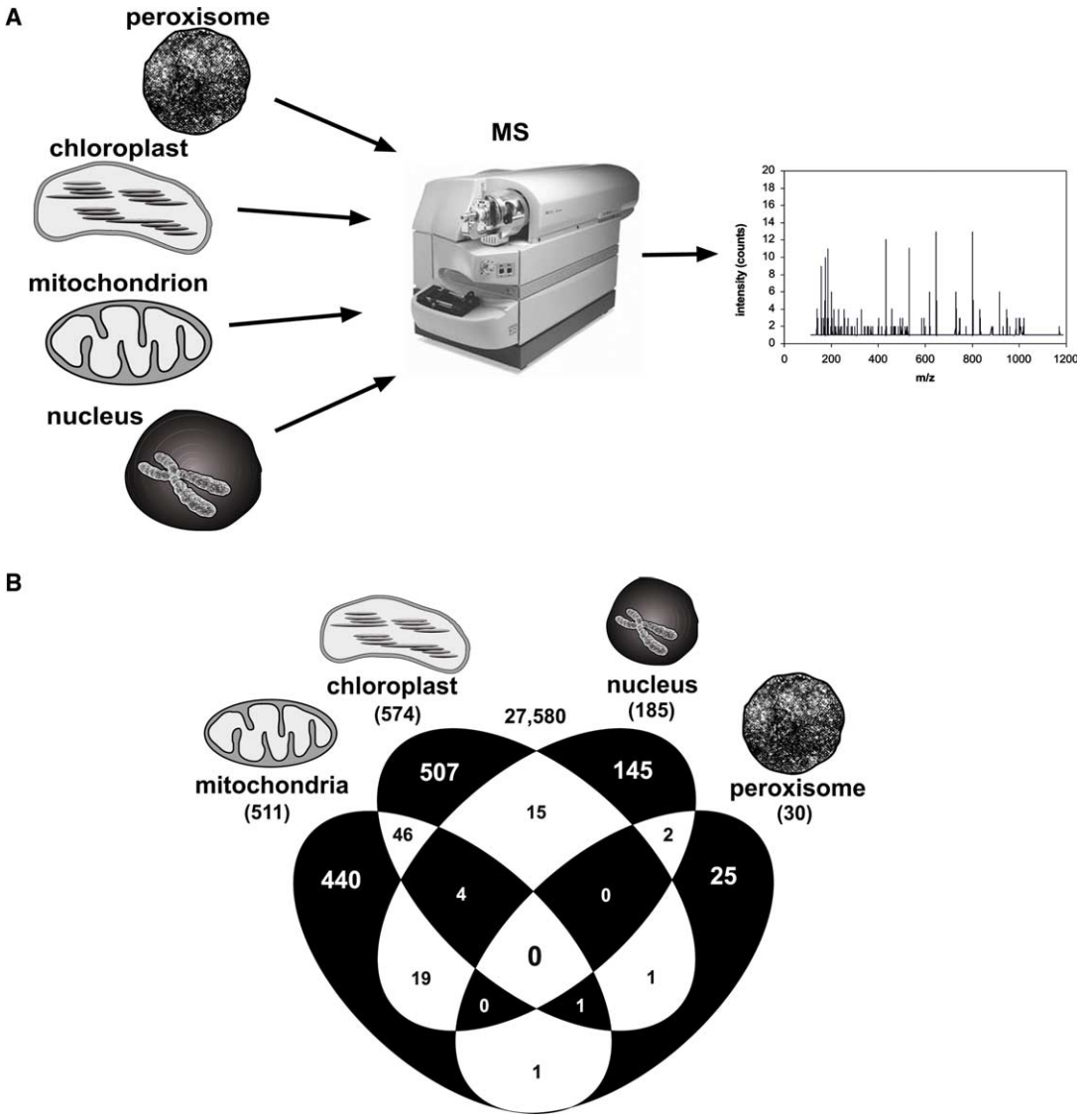


Fig. 4. Proteomic analyses of *Arabidopsis* sub-cellular compartments. (A). Isolated sub-compartments were analysed by mass spectrometry using a variety of techniques including 2D PAGE, LC-MS/MS and 1D PAGE. The mass spectrometer pictured is the Applied Biosystems Q TRAP™ LC/MS/MS System (http://www.appliedbiosystems.com). (B). A Venn diagram comparing the identified overlaps between experimentally defined proteomes of four sub-compartments of *Arabidopsis*. Mitochondria: (Heazlewood et al., 2004; Heazlewood and Millar unpublished data), peroxisome: (Fukao et al., 2002), nucleus: (Bae et al., 2003; Calikowski et al., 2003), plastid: (Peltier et al., 2002; Schubert et al., 2002; Ferro et al., 2003; Froehlich et al., 2003; Friso et al., 2004). The numbers in brackets represent the total sets identified from each organelle while the number outside the Venn diagram represents the null set.

4.6. Pitfalls in gel and non-gel assessment of protein families

The major disadvantage of gel-based approaches is the loss of proteins, and even whole protein families, due to the relatively stringent requirements of isoelectric focusing. Notably, low abundance proteins, highly basic and even moderately hydrophobic proteins are usually under-represented (Santoni et al., 1999; Gygi et al., 2002). Further, the gel environment tends to lead to more artefactual protein modifications such as methionine oxidation and cysteine modification by acrylamide radicals (Swiderek et al., 1998), potentially masking in vivo modifications such as glycosylation and phosphorylation that are of most interest to researchers (Sickmann et al., 2002). Clearly the major disadvantage of using a non-gel approach is the difficulty in readily quantifying protein member abundances. Incomplete digests of samples and differential ionization of peptides are two of the main problems associated with quantification when using these techniques (Link et al., 1999). Arraying samples on a 2D gel enables an assessment of quantitation of various gene family members as has been demonstrated for

Table 3
Sub-cellular localisation of *Arabidopsis* gene families using proteomic data

Family	Proteome location	References
Malate dehydrogenase		
At3g47520	Chloroplast	Ferro et al. (2003)
At5g58330	Chloroplast	Froehlich et al. (2003)
At1g53240	Mitochondrion	Millar et al. (2001)
At3g15020	Mitochondrion	Heazlewood et al. (2004)
At2g22780	Peroxisome	Fukao et al. (2002)
At5g09660	Peroxisome	Fukao et al. (2002)
At1g04410	Cytosol	Inferred
At5g43330	Cytosol	Inferred
At5g56720	Cytosol	Inferred
Ascorbate peroxidase		
At4g35000	cpt/perox	Fukao et al. (2002), Ferro et al. (2003) and Froehlich et al. (2003)
At4g09010	Chloroplast	Peltier et al. (2002), Schubert et al. (2002) and Friso et al. (2004)
At4g08390	mito/cpt	Heazlewood et al. (2004)
At1g77490	Cytosol	Prediction
At3g09640	Cytosol	Prediction
At1g07890	Unknown	
At4g32320	Unknown	
At4g35970	Unknown	

Gene identifiers correspond to *Arabidopsis* Genome Initiative (AGI) chromosomal loci which are accessible at TAIR.

components of the outer membrane mitochondrial import complex (Werhahn and Braun, 2002). A further significant issue is that in some cases during analysis of mass spectrometric data there is a failure to assign a peptide as the product of a single genomic locus. This will occur if the tryptic peptide comes from a conserved region of the protein and is therefore not unique to a single gene family member. If the sample was from a single spot on a 2D gel then the peptide can be more confidently assigned to a single gene based on other matching peptides from this sample. However if the sample is a complex mix of proteins (e.g. MW gel slice or an LC-MS/MS sample) then the peptide cannot be assigned with confidence. More recently several software programs have been developed in an attempt to deal with issues associated with the significant matching of MS derived spectra derived from such complex mixtures (Keller et al., 2002; Tabb et al., 2002; Nesvizhskii et al., 2003). Tryptic peptides can also arise from alternative splice forms, so an intimate knowledge of the gene family and EST alignments is important to be sure that the actual proteins under study correspond to the predicted gene models used for the mass spectrometry matches (Lisacek et al., 2001). In some cases, the use of intact protein mass determinations by LC-MS/MS (Whitelegge et al., 2002) could further complement peptide analysis to define the gene family member responsible for peptide products in complex samples. Such an approach was recently utilised to confirm the presence of the two parologous genes that code for the large subunit of the oxygen-evolving enzyme (PsbO) from an *Arabidopsis* thylakoid preparation (Gomez et al., 2003).

5. Where to now? Challenges for the future in gene families

To date research of gene families has tended to focus on genomic organization of genes and some basic expression analysis. The next challenge is 'functional' characterisation and it is probable that an intimate knowledge of the expression of gene families will greatly aid this process. This next stage will involve understanding the transcriptional regulation of related genes and appreciating the real and apparent levels of functional redundancy within defined genomes in plants. Detailed and sophisticated expression profiling of entire gene families will shed light on the degree of differential expression. Ultimately an intimate knowledge of the promoter elements regulating individual genes will be necessary to determine how members are differentially regulated. For example, there is evidence that tandemly duplicated polygalacturonase-inhibiting protein (PGIP) genes encoding proteins with similar activities are coordinately and differentially induced during fungal infection through differing pathways (Ferrari et al., 2003). It is proposed that differential regulation of functionally redundant proteins might be important to ensure that PGIPs are expressed following activation of different pathways which will be important for antagonistic cross-talk in signalling pathways (Ferrari et al., 2003).

A further dimension to the issue of dissecting gene families and functionality is the structure of the protein product. Subtle differences in primary sequence that differentiate many members of gene families may become far more obvious at this level. Currently there are only 57 structures specifically available for *Arabidopsis* proteins,

and 12 available for rice (http://www.ncbi.nlm.nih.gov/), although some extrapolations could be undertaken from work being carried out on universally conserved domain structures from other organisms such as the protein interacting WD-repeat structure (Smith et al., 1999), of which there are ≈60 members present in *Arabidopsis*. Global analysis of protein structures by a combination of fold recognition and domain threading has begun with the recent release of the iGAP genome annotation of *Arabidopsis* (Li et al., 2003), but the value of this resource for multi-gene family studies remains undetermined.

The central question still remains as to why some genes are prone to duplication and are subsequently retained while others are not. It is possible that multigene families allow for a finer control of expression and protein function through sub-functionalisation. By partitioning functions among multiple genes there is thus simply more capacity for specialisation. But what is the force preventing all genes from forming complex families? Potentially it is difficult to co-ordinate the expression of large gene families efficiently and this acts as a negative balance against genome expansion. We also need to ask, what properties or functions do large gene families have in common? A perusal of the major families indicates that for the sequenced genomes of *Arabidopsis* and rice many are involved in plant defence, secondary metabolism and signalling (Kaul et al., 2000; Goff et al., 2002). It may be that genes required to undertake quick, large transcription responses are prone to becoming gene families. Such a development may help provide a readily tuneable on/off switch for a variety of specialised functions in response to a host of stimuli. In contrast, metabolic functions, house-keeping components and structural proteins which are more constitutively expressed, tend to be encoded in smaller gene families and more significantly these classes contain many singleton genes in plants.

Acknowledgements

Research grants from the Australian Research Council Discovery Programme to A.H.M are greatly acknowledged. P.G.S is a recipient of a Grains Research and Development Corporation PhD scholarship and A.H.M is an Australian Research Council QEII Research Fellow.

References

Andon, N.L., Eckert, D., Yates 3rd, J.R., Haynes, P.A., 2003. High-throughput functional affinity purification of mannose binding proteins from *Oryza sativa*. Proteomics 3, 1270–1278.

Andrews, J., Bouffard, G.G., Cheadle, C., Lu, J.N., Becker, K.G., Oliver, B., 2000. Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. Genome Research 10, 2030–2043.

Arango, M., Gevaudant, F., Oufattole, M., Boutry, M., 2003. The plasma membrane proton pump ATPase: the significance of gene subfamilies. Planta 216, 355–365.

Bae, M.S., Cho, E.J., Choi, E.Y., Park, O.K., 2003. Analysis of the *Arabidopsis* nuclear proteome and its response to cold stress. Plant Journal 36, 652–663.

Bailey, P.C., Martin, C., Toledo-Ortiz, G., Quail, P.H., Huq, E., Heim, M.A., Jakoby, M., Werber, M., Weisshaar, B., 2003. Update on the basic helix-loop-helix transcription factor gene family in *Arabidopsis thaliana*. Plant Cell 15, 2497–2502.

Bauer, A., Kuster, B., 2003. Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes. European Journal of Biochemistry 270, 570–578.

Burgess, R.R., Thompson, N.E., 2002. Advances in gentle immunoaffinity chromatography. Current Opinions in Biotechnology 13, 304–308.

Butt, Y.K., Lum, J.H., Lo, S.C., 2003. Proteomic identification of plant proteins probed by mammalian nitric oxide synthase antibodies. Planta 216, 762–771.

Calikowski, T.T., Meulia, T., Meier, I., 2003. A proteomic study of the *Arabidopsis* nuclear matrix. Journal of Cellular Biochemistry 90, 361–378.

Chew, O., Whelan, J., Millar, A.H., 2003. Molecular definition of the ascorbate-glutathione cycle in *Arabidopsis* mitochondria reveals dual targeting of antioxidant defenses in plants. Journal of Biological Chemistry 278, 46869–46877.

Chew, O., Lister, R., Qbadou, S., Heazlewood, J.L., Soll, J., Schleiff, E., Millar, A.H., Whelan, J., 2004. A plant outer mitochondrial membrane protein with high amino acid sequence identity to a chloroplast protein import receptor. FEBS Letters 557, 109–114.

Chivasa, S., Ndimba, B.K., Simon, W.J., Robertson, D., Yu, X.L., Knox, J.P., Bolwell, P., Slabas, A.R., 2002. Proteomic analysis of the *Arabidopsis thaliana* cell wall. Electrophoresis 23, 1754–1765.

Dammann, C., Ichida, A., Hong, B., Romanowsky, S.M., Hrabak, E.M., Harmon, A.C., Pickard, B.G., Harper, J.F., 2003. Subcellular targeting of nine calcium-dependent protein kinase isoforms from *Arabidopsis*. Plant Physiology 132, 1840–1848.

Dayoff, M.O., 1976. The origin and evolution of protein superfamilies. Federation Proceedings 35, 2132–2138.

Diebold, R., Schuster, J., Daschner, K., Binder, S., 2002. The branched-chain amino acid transaminase gene family in *Arabidopsis* encodes plastid and mitochondrial proteins. Plant Physiology 129, 540–550.

Dixon, D.P., Lapthorn, A., Edwards, R., 2002. Plant glutathione transferases. Genome Biology 3, reviews3004.1–3004.10.

Donne, J., 1999. Devotions Upon Emergent Occasions and Death's Duel. Random House, Toronto.

Eckermann, N., Fettke, J., Steup, M., 2002. Identification of polysaccharide binding proteins by affinity electrophoresis in inhomogeneous polyacrylamide gels and subsequent SDS-PAGE/matrix-assisted laser desorption ionization-time of flight analysis. Analytical Biochemistry 304, 180–192.

Felsenstein, J., 1993. PHYLIP (Phylogeny Inference Package) version 3.5c Distributed by the author. Department of Genetics, University of Washington, Seattle.

Ferrari, S., Vairo, D., Ausubel, F.M., Cervone, F., De Lorenzo, G., 2003. Tandemly duplicated *Arabidopsis* genes that encode polygalacturonase-inhibiting proteins are regulated coordinately by different signal transduction pathways in response to fungal infection. Plant Cell 15, 93–106.

Ferro, M., Salvi, D., Brugiere, S., Miras, S., Kowalski, S., Louwagie, M., Garin, J., Joyard, J., Rolland, N., 2003. Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana*. Molecular & Cellular Proteomics 2, 325–345.

Friso, G., Ytterberg, A.J., Giacomelli, L., Peltier, J.B., Rudella, A., Sun, Q., van Wijk, K.J., 2004. In-depth analysis of the thylakoid membrane proteome of *Arabidopsis thaliana* chloroplasts; new proteins, functions and a plastid proteome database. The Plant Cell 16, 478–499.

Froehlich, J.E., Wilkerson, C.G., Ray, 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. Journal of Proteome Research 2, 413–425.

Fukao, Y., Hayashi, M., Nishimura, M., 2002. Proteomic analysis of leaf peroxisomal proteins in greening cotyledons of *Arabidopsis thaliana*. Plant & Cell Physiology 43, 689–696.

Gagne, J.M., Downes, B.P., Shiu, S.H., Durski, A.M., Vierstra, R.D., 2002. The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in *Arabidopsis*. Proceedings of the National Academy of Sciences of the United States of America 99, 11519–11524.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., Briggs, S., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296, 92–100.

Gomez, S.M., Nishio, J.N., Faull, K.F., Whitelegge, J.P., 2002. The chloroplast grana proteome defined by intact mass measurements from liquid chromatography mass spectrometry. Molecular & Cellular Proteomics 1, 46–59.

Gomez, S.M., Bil, K.Y., Aguilera, R., Nishio, J.N., Faull, K.F., Whitelegge, J.P., 2003. Transit peptide cleavage sites of integral thylakoid membrane proteins. Molecular & Cellular Proteomics 2, 1068–1085.

Graves, P.R., Haystead, T.A., 2002. Molecular biologist's guide to proteomics. Microbiology and Molecular Biology Reviews 66, 39–63.

Gupta, R.K., Dobritsa, S.V., Stiles, C.A., Essington, M.E., Liu, Z., Chen, C.H., Serpersu, E.H., Mullin, B.C., 2002. Metallohistins: a new class of plant metal-binding proteins. Journal of Protein Chemistry 21, 529–536.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R., 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature Biotechnology 17, 994–999.

Gygi, S.P., Rist, B., Griffin, T.J., Eng, J., Aebersold, R., 2002. Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags. Journal of Proteome Research 1, 47–54.

Hansson, M., Vener, A.V., 2003. Identification of three previously unknown in vivo protein phosphorylation sites in thylakoid membranes of *Arabidopsis thaliana*. Molecular & Cellular Proteomics 2, 550–559.

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 signalling and regulatory components, provides assessment of targeting prediction programs and points to plant specific mitochondrial proteins. Plant Cell 16, 241–256.

Herbert, B., 1999. Advances in protein solubilisation for two-dimensional electrophoresis. Electrophoresis 20, 660–663.

Herbert, B., Righetti, P.G., 2000. A turning point in proteome analysis: sample prefractionation via multicompartment electrolyzers with isoelectric membranes. Electrophoresis 21, 3639–3648.

Hilson, P., Small, I., Kuiper, M.T., 2003. European consortia building integrated resources for *Arabidopsis* functional genomics. Current Opinion in Plant Biology 6, 426–429.

Hrabak, E.M., Chan, C.W., Gribskov, M., Harper, J.F., Choi, J.H., Halford, N., Kudla, J., Luan, S., Nimmo, H.G., Sussman, M.R., Thomas, M., Walker-Simmons, K., Zhu, J.K., Harmon, A.C., 2003. The *Arabidopsis* CDPK-SnRK superfamily of protein kinases. Plant Physiology 132, 666–680.

Jonsson, A.P., 2001. Mass spectrometry for protein and peptide characterisation. Cellular & Molecular Life Sciences 58, 868–884.

Jung, E., Heller, M., Sanchez, J.C., Hochstrasser, D.F., 2000. Proteomics meets cell biology: the establishment of subcellular proteomes. Electrophoresis 21, 3369–3377.

Kaul, S., Koo, H.L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L.J., Feldblyum, T., Nierman, W., Benito, M.I., Lin, X.Y., Town, C.D., Venter, J.C., Fraser, C.M., Tabata, S., Nakamura, Y., Kaneko, T., Sato, S., Asamizu, E., Kato, T., Kotani, H., Sasamoto, S., Ecker, J.R., Theologis, A., Federspiel, N.A., Palm, C.J., et al., 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

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. Analytical Chemistry 74, 5383–5392.

Koller, A., Washburn, M.P., Lange, B.M., Andon, N.L., Deciu, C., Haynes, P.A., Hays, L., Schieltz, D., Ulaszek, R., Wei, J., Wolters, D., Yates 3rd, J.R., 2002. Proteomic survey of metabolic pathways in rice. Proceedings of the National Academy of Sciences of the United States of America 99, 11969–11974.

Kolukisaoglu, H.U., Bovet, L., Klein, M., Eggmann, T., Geisler, M., Wanke, D., Martinoia, E., Schulz, B., 2002. Family business: the multidrug-resistance related protein (MRP) ABC transporter genes in *Arabidopsis thaliana*. Planta 216, 107–119.

Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K.H., Miller, P., Gerstein, M., Roeder, G.S., Snyder, M., 2002. Subcellular localization of the yeast proteome. Genes & Development 16, 707–719.

Kuroda, H., Takahashi, N., Shimada, H., Seki, M., Shinozaki, K., Matsui, M., 2002. Classification and expression analysis of *Arabidopsis* F-box-containing protein genes. Plant & Cell Physiology 43, 1073–1085.

Lackie, J.M., Dow, J.A.T., 1999. The Dictionary of Cell & Molecular Biology, third ed. Academic Press, London.

Li, W.W., Quinn, G.B., Alexandrov, N.N., Bourne, P.E., Shindyalov, I.N., 2003. A comparative proteomics resource: proteins of *Arabidopsis thaliana*. Genome Biology 4, R51.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., Yates, J.R., 1999. Direct analysis of protein complexes using mass spectrometry. Nature Biotechnology 17, 676–682.

Lisacek, F.C., Traini, M.D., Sexton, D., Harry, J.L., Wilkins, M.R., 2001. Strategy for protein isoform identification from expressed sequence tags and its application to peptide mass fingerprinting. Proteomics 1, 186–193.

Lister, R., Chew, O., Lee, M.-N., Heazlewood, J., Clifton, R., Parker, K., Millar, A., Whelan, J., 2004. A transcriptomic and proteomic characterisation of the *Arabidopsis* mitochondrial protein import apparatus and its response to mitochondrial dysfunction. Plant Physiology 134, 777–789.

Lopez, M.F., Kristal, B.S., Chernokalskaya, E., Lazarev, A., Shestopalov, A.I., Bogdanova, A., Robinson, M., 2000. High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. Electrophoresis 21, 3427–3440.

Maeda, K., Finnie, C., Stergaard, O.S., Svensson, B., 2003. Identification, cloning and characterization of two thioredoxin h isoforms,

HvTrxh1 and HvTrxh2, from the barley seed proteome. European Journal of Biochemistry 270, 2633–2643.

Martinez-Castilla, L.P., Alvarez-Buylla, E.R., 2003. Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. Proceedings of the National Academy of Sciences of the United States of America 100, 13407–13412.

Milla, M.A., Maurer, A., Huete, A.R., Gustafson, J.P., 2003. Glutathione peroxidase genes in *Arabidopsis* are ubiquitous and regulated by abiotic stresses through diverse signaling pathways. Plant Journal 36, 602–615.

Millar, A.H., Heazlewood, J.L., 2003. Genomic and proteomic analysis of mitochondrial carrier proteins in *Arabidopsis*. Plant Physiology 131, 443–453.

Millar, A.H., Sweetlove, L.J., Giege, P., Leaver, C.J., 2001. Analysis of the *Arabidopsis* mitochondrial proteome. Plant Physiology 127, 1711–1727.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J.A., Vaughan, R., Zdobnov, E.M., 2003. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Research 31, 315–318.

Nesvizhskii, A.I., Keller, A., Kolker, E., Aebersold, R., 2003. A statistical model for identifying proteins by tandem mass spectrometry. Analytical Chemistry 75, 4646–4658.

Okada, K., Saito, T., Nakagawa, T., Kawamukai, M., Kamiya, Y., 2000. Five geranylgeranyl diphosphate synthases expressed in different organs are localized into three subcellular compartments in *Arabidopsis*. Plant Physiology 122, 1045–1056.

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

Peltier, J.B., Emanuelsson, O., Kalume, D.E., Ytterberg, J., Friso, G., Rudella, A., Liberles, D.A., Soderberg, L., Roepstorff, P., von Heijne, G., van Wijk, K.J., 2002. Central functions of the lumenal and peripheral thylakoid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction. Plant Cell 14, 211–236.

Petersen, A., Schramm, G., Becker, W.M., Schlaak, M., 1993. Comparison of four grass pollen species concerning their allergens of grass group V by 2D immunoblotting and microsequencing. Biol Chem Hoppe Seyler 374, 855–861.

Prime, T.A., Sherrier, D.J., Mahon, P., Packman, L.C., Dupree, P., 2000. A proteomic analysis of organelles from *Arabidopsis thaliana*. Electrophoresis 21, 3488–3499.

Pruess, M., Fleischmann, W., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E., Mittard, V., Mulder, N., Phan, I., Servant, F., Apweiler, R., 2003. The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. Nucleic Acids Research 31, 414–417.

Rafalski, J.A., Hanafey, M., Miao, G.H., Ching, A., Lee, J.M., Dolan, M., Tingey, S., 1998. New experimental and computational approaches to the analysis of gene expression. Acta Biochimica Polonica 45, 929–934.

Romeis, T., 2001. Protein kinases in the plant defence response. Current Opinion in Plant Biology 4, 407–414.

Sanchez-Fernandez, R., Davies, T.G., Coleman, J.O., Rea, P.A., 2001. The *Arabidopsis thaliana* ABC protein superfamily, a complete inventory. Journal of Biological Chemistry 276, 30231–30244.

Sanderfoot, A.A., Pilgrim, M., Adam, L., Raikhel, N.V., 2001. Disruption of individual members of *Arabidopsis* syntaxin gene families indicates each has essential functions. Plant Cell 13, 659–666.

Santoni, V., Rabilloud, T., Doumas, P., Rouquie, D., Mansion, M., Kieffer, S., Garin, J., Rossignol, M., 1999. Towards the recovery of hydrophobic proteins on two-dimensional electrophoresis gels. Electrophoresis 20, 705–711.

Sappl, P.G., Onãte-Sanchez, L., Singh, K., Millar, A.H., 2004. Proteomic analysis of glutathione *S*-transferases of *Arabidopsis thaliana* reveals differential salicylic acid induced expression of the plant specific Phi and Tau classes. Plant Molecular Biology, in press.

Sarria, R., Wagner, T.A., O'Neill, M.A., Faik, A., Wilkerson, C.G., Keegstra, K., Raikhel, N.V., 2001. Characterization of a family of *Arabidopsis* genes related to xyloglucan fucosyltransferase1. Plant Physiology 127, 1595–1606.

Schott, H., 1984. Affinity Chromatography: Template Chromatography of Nucleic Acids and Proteins. Dekker, New York.

Schubert, M., Petersson, U.A., Haas, B.J., Funk, C., Schroder, W.P., Kieselbach, T., 2002. Proteome map of the chloroplast lumen of *Arabidopsis thaliana*. Journal of Biological Chemistry 277, 8354–8365.

Sharma, J., Panico, M., Barber, J., Morris, H.R., 1997. Purification and determination of intact molecular mass by electrospray ionization mass spectrometry of the photosystem II reaction center subunits. Journal of Biological Chemistry 272, 33153–33157.

Sickmann, A., Mreyen, M., Meyer, H.E., 2002. Identification of modified proteins by mass spectrometry. IUBMB Life 54, 51–57.

Smith, T.F., Gaitatzes, C., Saxena, K., Neer, E.J., 1999. The WD repeat: a common architecture for diverse functions. Trends in Biochemical Sciences 24, 181–185.

Stracke, R., Werber, M., Weisshaar, B., 2001. The R2R3-MYB gene family in *Arabidopsis thaliana*. Current Opinion in Plant Biology 4, 447–456.

Suty, L., Lequeu, J., Lancon, A., Etienne, P., Petitot, A.S., Blein, J.P., 2003. Preferential induction of 20S proteasome subunits during elicitation of plant defense reactions: towards the characterization of plant defense proteasomes. International Journal of Biochemistry & Cell Biology 35, 637–650.

Swiderek, K.M., Davis, M.T., Lee, T.D., 1998. The identification of peptide modifications derived from gel-separated proteins using electrospray triple quadrupole and ion trap analyses. Electrophoresis 19, 989–997.

Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4. Sinauer Associates, Sunderland, MA.

Tabb, D.L., McDonald, W.H., Yates 3rd, J.R., 2002. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. Journal of Proteome Research 1, 21–26.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research 22, 4673–4680.

Tognolli, M., Penel, C., Greppin, H., Simon, P., 2002. Analysis and expression of the class III peroxidase large gene family in *Arabidopsis thaliana*. Gene 288, 129–138.

VerBerkmoes, N.C., Hettich, R.L., Bruce, B.D., Nguyen, R., Savage, T.L., 2002. One- and two-dimensional LC/MS/MS analysis of *Arabidopsis thaliana* proteome. LC GC North America.

Wagner, U., Edwards, R., Dixon, D.P., Mauch, F., 2002. Probing the diversity of the *Arabidopsis* glutathione *S*-transferase gene family. Plant Molecular Biology 49, 515–532.

Werhahn, W., Braun, H.P., 2002. Biochemical dissection of the mitochondrial proteome from *Arabidopsis thaliana* by three-dimensional gel electrophoresis. Electrophoresis 23, 640–646.

Whitelegge, J.P., Gundersen, C.B., Faull, K.F., 1998. Electrospray-ionization mass spectrometry of intact intrinsic membrane proteins. Protein Science 7, 1423–1430.

Whitelegge, J.P., Zhang, H., Aguilera, R., Taylor, R.M., Cramer, W.A., 2002. Full subunit coverage liquid chromatography electrospray ionization mass spectrometry (LCMS+) of an oligomeric membrane protein: cytochrome b(6)f complex from spinach and the cyanobacterium *Mastigocladus laminosus*. Molecular & Cellular Proteomics 1, 816–827.

Wortman, J.R., Haas, B.J., Hannick, L.I., Smith Jr., R.K., Maiti, R., Ronning, C.M., Chan, A.P., Yu, C., Ayele, M., Whitelaw, C.A., White, O.R., Town, C.D., 2003. Annotation of the *Arabidopsis* genome. Plant Physiology 132, 461–468.

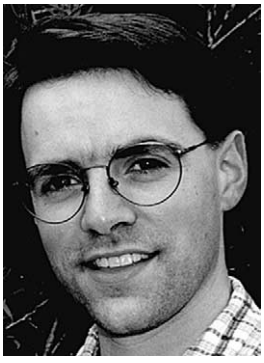
Wu, C.H., Huang, H., Yeh, L.S., Barker, W.C., 2003. Protein family classification and functional annotation. Computational Biology and Chemistry 27, 37–47.

Yokoyama, R., Nishitani, K., 2001. A comprehensive expression analysis of all members of a gene family encoding cell-wall enzymes allowed us to predict cis-regulatory regions involved in cell-wall construction in specific organs of *Arabidopsis*. Plant & Cell Physiology 42, 1025–1033.

Zolla, L., Rinalducci, S., Timperio, A.M., Huber, C.G., 2002. Proteomics of light-harvesting proteins in different plant species. Analysis and comparison by liquid chromatography-electrospray ionization mass spectrometry. Photosystem I. Plant Physiology 130, 1938–1950.

Pia G. Sappl is a PhD student supported by a scholarship from the Australian Grain's Research and Development Corporation in the Plant Molecular Biology Group, School of Biomedical and Chemical Sciences at The University of Western Australia. She obtained her BSc (Hons 1st class) from The University of Western Australia, and was the recipient of the Lugg Medal for Biochemistry for her undergraduate degree. Her PhD research focuses on the glutathione *S*-transferase gene family in *Arabidopsis*, using proteomic and genomic studies of the structure and expression of this family and T-DNA knock-out resources to probe gene family redundancy.

Joshua L. Heazlewood is Post-Doctoral Research Associate in the Plant Molecular Biology Group, School of Biomedical and Chemical Sciences at The University of Western Australia. He obtained his PhD in Plant Molecular Biology from La Trobe University, Australia investigating the MYB gene family in *Arabidopsis* using reverse genetic techniques. In 2001 he obtained a Post-Doctoral position with Dr. Millar's group at The University of Western Australia where he has been using liquid chromatography and gel electrophoresis coupled to tandem mass spectrometry to investigate the mitochondrial proteomes of model plants.



A. Harvey Millar is an Australian Research Council Queen Elizabeth II Research Fellow in the Plant Molecular Biology Group, School of Biomedical and Chemical Sciences at the University of Western Australia. He obtained his PhD in Biochemistry from the Australian National University, Canberra, Australia. He then worked as a Human Frontier Science Programme Long-Term Fellow in the Department of Plant Sciences in Oxford, UK, before returning to Australia in 1999 via a series of research fellowships held at The University of Western Australia. Dr. Millar's group is focussed on proteomic analysis in the model plants *Arabidopsis* and rice, with a special emphasis on mitochondrial proteomes and plant stress/defence strategies. His work aims to integrate proteomic data into biochemical analysis of plants, and also into the increasing information available in model plants relating to gene families, gene expression patterns and genetic resources.